1. Introduction

1.1 Scope of the Course

1.1.1 General Remarks

1.1.2 Exercises

1.2 Introduction to the Course

1.2.1 Some General Remarks

1.2.2 Electronic Materials and Electronic Products

1.3 Required Reading and Exercises

1.3.1 General Remarks

1.3.2 Ohms Law and Materials Properties

1.3.3 Ohms Law and Classical Physics

1.3.4 The Hall Effect

1.3.5 Averaging Vectors

Electrical Engineers usually have little specialized knowledge of chemical reactions or solid state mechanics. And yet, our ability to fabricate semiconductor devices depends critically on our understanding of these items.

Murarka and Peckerar
Preface to their text book

1. Introduction

1.1 Scope of the Course

1.1.1 General Remarks

Some Important Links

- For a detailed *table of contents* use the link
- The *organization, timetable, etc.* for the running term can be found in the link.
- If you like to read *prefaces*, just click.
- For book recommendations: Consult the *list of books*

What is Special About this Course

- The lecture course "Electronic Materials" has a somewhat *special status* for two reasons:
  - **1.** It is far to short to really cover the topic appropriately, but yet, *it overlaps with other courses*. The reason for this is the mix of students who are required to take this course (see below).
  - **2.** It *had* a special format for the exercise part [1)].
  - Unfortunately, in the fall term of **2004**, this exercise format had to be abandoned for various reasons in favor of the more classical format.

Relation to Other Courses

- This graduate course "**Electronic Materials**" (in German: Werkstoffe der Elektrotechnik und Sensorik I) is a *required* course for

Study Course
1. *All* Materials Science Diploma students
2. *All* Master of Mat. Science and Engineering students.
3. *All* Electrical Engineering Diploma students.
4. *All* "Wirtschafts-Ingenieur ET&IT" Diploma students.

- Exactly what "required" means depends on your study course - look up your "Prüfungsordnung". Essentially the following rules obtain:
  - The first three study courses must pass the written examination, the last one must obtain the "Schein" for the exercise class
  - Even if you are *not* required to obtain the exercise "Schein" or the 1.5 ECTS "Credit Points", it is highly recommended to participate in the exercise class since it is a preparation for the examination!
- It interacts with several other courses in the materials science and electrical engineering curriculum. There is considerable overlap with the following courses
- **Silicon Technology I + II** (In German: Halbleitertechnologie I + II)

- This course is required for Matwiss students in the Diploma track and electrical engineers specializing in solid state electronics.
- It contains everything taught in the **Si**-technology section of "Electronic Materials". However, since the bulk of the electrical engineers will not be exposed to **Si**-technology anywhere else, "Electronic Materials" will cover the subject briefly. For all others, this part can be seen as an introduction to "Silicon Technology I + II"

Solid State Physics for Engineers II

- This course is required for Matwiss students in the Diploma and Master track and electrical engineers specializing in solid state electronics.
- Dielectrics and magnetic materials will be covered in depth and from a more theoretical background. Again, the relevant chapters in "Electronic Materials" may be seen as introduction by those students enrolling in "Solid State II"; for the others it is an essential component of electrical engineering.

The course has a very special relation to "**Introduction to Materials Science I + II**", which is a required course for all engineering (undergraduate) students.

- "Electronic Materials" can be seen as part III of this series, because it covers the two major subjects left open in "Introduction to Materials Science I + II": *dielectrics* and *magnetic materials*. Moreover, the **Si**-technology part begins where the semiconductor part of "Introduction to Materials Science I + II" ends.
- However, "Electronic Materials" is fully self-contained and can be taken by itself, provided the basic requirements are met.
- For details of the contents of "Introduction to "Materials Science I + II" refer to the Hyperscripts (in German)
  MaWi I
  MaWi II

Sensors I (In German: "Werkstoffe der Elektrotechnik und Sensorik II")

- Required for all Materials Science students in the diploma track.
  (Used to be required for all electrical engineers).
- Continues "Electronic Materials" with emphasize on sensor applications and ionic materials, but is self-contained and can be taken by itself.
- "Electronic Materials" will include a brief chapter concerning ionic materials for those who do not take "Sensors I"

Semiconductors

- This course overlaps a little bit with "Electronic Materials", but essentially continues where Electronic Materials ends for Semiconductors.

Background Knowledge

Mathematics

- The course does not employ a lot of math. You should be familiar, however, with complex numbers, Fourier transforms and differential equations.

General Physics and Chemistry

- A general undergraduate level of basic physics should be sufficient. You should be comfortable with units and conversion between units.

General Materials Science

- You must know basic crystallography, quantum theory and thermodynamics.

[1)] Conventional exercises were abandoned in favor of "professional" presentations including a paper to topics that are within the scope of the course but will not be covered in a regular class. A list of the topics is given in the "Running Term" folder; the rules for the seminar will be found in the link. The contents and the style of the presentation will be discussed. For details use the link.

1.1.2 Exercises

Format of Exercises

- Exercises will consist of two parts:
  - A "classical" exercise part (where the student is supposed to work out exercises questions handed out a week ahead of ecercise class) .
  - A "Multiple Choice" part, where questionaires have to be filled in.
  - For details see the "Running term" and the News
- The old "Seminar" format has been abandoned.
  - However, you may still find helpful hints for **presentation techniques** in the link.

1.2 Introduction to the Course

1.2.1 Some General Remarks

- So what are "Electronic Materials"? Ask Google and you get an answer!

- Progress in Electrical Engineering was always dependent on progress in materials. For quite some time, electrical engineering meant electro*mechanical* engineering, and electrical products were made from "trivial" materials, as seen from a modern point of view. What was needed were cables, insulators, ferromagnetic sheet metal for transformers and generators, and a lot of metal for the general mechanics. A few applications centered around some mysterious materials - out of that grew *electronics* and electronic materials. But even then there were key materials:
  - *Cu wires* of all kinds. Not so trivial - how do you make a insulated but still flexible wire?
  - Insulating materials - plastics didn't quite exist yet. ***Mica*** was one of the key materials - there were mines for it!
  - *Graphite* and ***tungsten*** were important, whenever things got hot, like the filament in the light bulb or in a **vacuum tube**.
  - The "tube of **Braun**" - the "*Braunsche Röhre*" as it was known in Europe - the first **cathode ray tube** (***CRT)*** in other words - needed complicated glass work and some *ZnS* as electroluminescent material
  - Strange compounds like "*phosphor bronze*" were developed for contacts.
  - And **Selenium** (**Se**) was important for rectifiers, although nobody quite understood how it worked.
- The essential break through in the thirties was the **vacuum tube**; with it came electronics: Rectifiers, amplifiers, radio, black-and white **TV**, colour **TV**. It's not that long ago, but obviously long enough for some not to remember!
- The next break-through was called **transistor**; it happened in **1947**. **Integrated circuits** followed around **1970**, and since then we witness exponential growth with growth rates in the complexity of electronics (at constant prices) of up to **40%** a year!
  - A good (german) book covering this development in some detail is Hans **Queissers** "*Kristallne Krisen*".

1.2.2 Electronic Materials and Electronic Products

Electronic Products

- *Electronic Materials* are what you find inside the *components* making up *electronic products*. They consist of some stuff that you cannot easily exchange with something else - not even in principle - without losing the function.
  - What you *can* change easily for example, is the material for the box, the housing. Use *Al* instead of plastic or vice versa for your video recorder - it would still work, needing at most some minor adjustments.
  - You also may change (in principle) the metal for real wires. Using *Au*, *Ag*, or *Al* instead of - let's say - *Cu*, makes little difference for the function.
  - But exchange *any* material in a "chip" (i.e. in an **integrated circuit**) with something else (even allowing for minor adjustments) - and that definitely will be the end of your product.
- Let's look at some typical products or product groups that contain electronic materials:
  - *Electronics* in general (Computer, **TV**, Radio, Radar, Microwave, ...).
  - **Flat panel displays** (***FPD***).
  - **Micromechanics** and **Microsystems** (***MEMS***).
  - **Solar cells**.
  - **Lasers** (in particular semiconductor Lasers).
  - **Batteries**, **Accumulators**; energy storage systems in general.
  - **Sensors**, in particular solid state sensors, that convert whatever they sense directly into a current or a voltage.
  - **Fuel Cells**.
  - **Magnetic Memories**.

Looking at Components

- Consider, e.g., a **laptop** or **notebook** in more detail. If you take it apart, you find the "high tech" stuff:
  - Any number of *chips*, i.e. integrated circuits.
  - Some **quartz oscillators**.
  - A **hard disc**, i.e. a magnetic memory for the bulk memory.
  - A reading head for the hard disc that uses the "**giant magnetoresistance effect**"
  - A ***CD ROM***, i.e. an optical memory and a semiconductor ***Laser***
  - A **flat-panel display** (**FPD**) using "liqiud crystals", whichis pretty big as a single component, but cannot be subdivided in smaller pieces.
- But there is also "low tech" - or so it seems:
  - **Capacitors** and **inductors**.
  - *Switches*, connectors, the keyboard as a unit.
  - *Insulation*.
  - Mechanical stuff like the disk drive, but also the *housing.*
- Some components betray their key material in their name ("*quartz*" oscillator) or by common knowledge (who, besides some so-called intellectuals, does not know that the word "chip" is almost a synonym for Silicon?), but for most components we have to look deeper - we must open them up (which will almost always destroy them). What do we find?

Electronic Materials

- Lets open up a chip. We find
  - **Packaging material** - either some *polymer* blend or *ceramics*.
  - A "chip" mostly consisting of **Si**, but interlaced in an intricate pattern with other materials like **P**, **B**, **As**, **SiO_2**, **Si_3N_4**, **$MoSi_2$**, **W**, **TiN**, **Al**, **Cu**….
  - A **lead frame** - the little pins sticking out of the package - made of some *metal alloys*.
  - Tiny **wires** connecting the leads to the chip or some pretty sophisticated stuff doing this job.
- Now open up the **FPD**. You will find many materials, the most suspicious beyond what we already found in chips are:
  - **Liquid crystals**, some strange liquid stuff.
  - **Amorphous Si**.
  - **Indium tin oxide *ITO***, a transparent electrode.
  - *Plastic foils* acting as *polarizers*.
  - A plastic (or *glass*) front and end plate.
- Now lets look at the **Laser** coming with the **CD** drive :
  - You find a complex mix of **GaAs**, **GaAlAs**, some other elements, as well as wires and packaging materials.
  - And all of this is quite different from what you find in the **Si** chips!
  - Soon you would find **GaN** in your Laser diode - and the capacity of your **CD** memory will quadruple!
- We could continue this, *but by now you got the idea*:

Progress in Electronic and Communication Technology is driven by
Progress in Material Science
(and almost nothing else)

1.3 Required Reading and Exercises

1.3.1 General Remarks

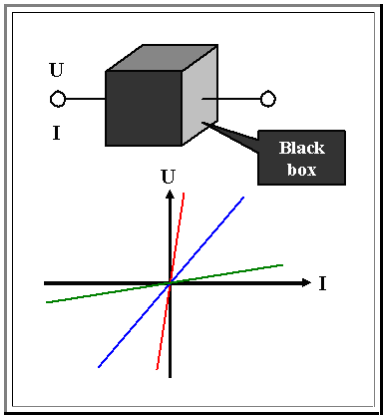
- This course is now (fall term **2004**) a required course in four different study programs:
  - Study program "Diplom-Ingenieur Elektrotechnik".
  - Study program "Diplom-Wirtschaftsingenieur Elektrotechnik".
  - Study program "Diplom-Ingenieur Materialwissenschaft".
  - Study program "Master of Materials Science and Engineering".
- Students in this programs have different backgrounds and different expectations. In order to provide a common platform to start from, there wil be some *required reading*.
  - The required reading modules are from the Hyperscript "Introduction to Materials Science II" from the Study program "Diplom-Ingenieur Materialwissenschaft".
  - They are reproduced in this Hyperscript and, wherever necessary, translated into English.
- Exercises in any format and possibly questions in the exam will assume that you are thoroughly familiar with the topics contained in the required reading modules.

Required reading Modules are shown in
Purple
in the Matrix of Modules

This Module comes from "Introduction to Materials Science II"; where you will find the most recent version. The link takes you there.

Required Reading

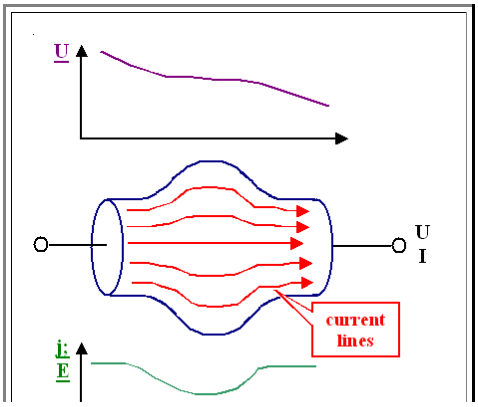
1.3.2 Ohms Law and Materials Properties



- In this subchapter we will give an outline of how to progress from the simple version of **Ohms "Law"**, which is a kind of "electrical" definition for a black box, to a formulation of the same law from a *materials point of view* employing (almost) first principles.
  - In other words: The *electrical engineering* point of view is: If a "black box" ***exhibits*** a linear relation between the (dc) current ***I*** flowing through it and the voltage ***U*** applied to it, it is an **ohmic resistor**.
  - That is illustrated in the picture: As long as the voltage-current characteristic you measure between two terminals of the black box is linear, the black box is called an (ohmic) resistor).
  - Neither the slope of the ***I-U***-characteristics matters, nor the material content of the box.
- The*Materials Science* point of view is quite different. Taken to the extreme, it is:
  - Tell me what kind of material is in the black box, and I tell you:
    1. If it really is an *ohmic* resistor, ***i.e.*** if the current relates *linearly* to the voltage for reasonable voltages and *both* polarities.
    2. What its (specific) resistance will be, including its temperature dependence.
    3. And everything else of interest.
- In what follows we will see, what we have to do for this approach. We will proceed in **3** steps
  - In the first two steps, contained in this sub-chapter we simply reformulate Ohms law in physical quantities that are related to material properties. In other words, we look at the properties of the moving charges that produce an electrical current. But we only *define* the necessary quantities; we do not calculate their numerical values.
  - In the third step - which is the content of many chapters - we will find ways to actually *calculate* the important quantities, in particular for semiconductors. As it turns out, this is not just difficult with classical physics, but simply impossible. We will need a good dose of quantum mechanics and statistical thermodynamics to get results.

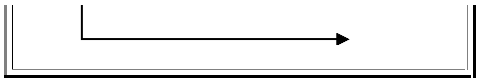
1. Step: Move to *specific* quantities

- First we switch from current ***I*** and voltage ***U*** to the **current density *j*** and the **field strength *E***, which are not only independent of the (uninteresting) size and shape of the body, but, since they are *vectors*, carry much more information about the system.
  - This is easily seen in the schematic drawing below.



- *Current density* ***j*** and *field strength* ***E*** may depend on the coordinates, because ***U*** and ***I*** depend on the coordinates, e.g. in the way schematically shown in the picture to the left. However, for a homogeneous material with constant cross section, we may write

$$j = \frac{I}{F}$$



- with F = cross sectional area. The direction of the *vector* $\underline{j}$ would be parallel to the normal vector $\underline{f}$ of the reference area considered: it also may differ locally. So in full splendor we must write

$$\underline{j}(x,y,z) = \frac{I(x,y,z)}{F} \cdot \underline{f}$$

The "global" field strength is

$$E = \frac{U}{I}$$

- With I = length of the body. If we want the *local* field strength $\underline{E}(x,y,z)$ as a vector, we have, in principle, to solve the Poisson equation

$$\nabla \cdot \underline{E}(x,y,z) = \frac{\rho(x,y,z)}{\epsilon\epsilon_0}$$

- With $\rho(x,y,z)$ = ***charge density***. For a homogeneous materisl with constant cross section, however, $\underline{E}$ is parallel to $\underline{f}$ and constant everywhere, again which is clear without calculation.

So. to make things easy, for a homogenous material of length I with constant cross-sectional area F, the field strength $\underline{E}$ and the current density $\underline{j}$ do not depend on position - they have the same numerical value everywhere.

- For this case we can now write down Ohms law with the new quantities and obtain

$$j \cdot F = I = \frac{1}{R} \cdot U = \frac{1}{R} \cdot E \cdot I$$

$$\underline{j} = \frac{I}{F \cdot R} \cdot \underline{E}$$

The fraction $I/F \cdot R$ *obviously* (think about it!) has the *same numerical value* for *any* homogeneous cube (or homogeneous whatever) of a given material; it is, of course, the **specific conductivity** σ

$$\sigma = \frac{1}{\rho} = \frac{I}{F \cdot R}$$

- and ρ is the **specific resistivity**. In words: A $\mathbf{1\ cm^3}$ cube of homogeneous material having the specific resistivity ρ has the resistance $R = (\rho \cdot I)/F$
- Of course, we will never mix up the *specific resistivity* ρ with the *charge density* ρ or *general densities* ρ, because we know from the context what is meant!
- The *specific resistivity* obtained in this way is necessarily identical to what you would define as specific resistivity by looking at some rectangular body with cross-sectional area F and length I.
- The *specific conductivity* has the dimension $[\sigma] = \Omega^{-1}\mathbf{cm}^{-1}$, the dimension of the *specific resistivity* is $[\rho] = \Omega\mathbf{cm}$. The latter is more prominent and you should at least have a feeling for representative numbers by remembering

ρ (metal)	≈	**2 μΩcm**
ρ (semicoductor)	≈	**1 Ωcm**
ρ (insulator)	≈	**1 GΩcm**

- Restricting ourselves to isotropic and homogenoeus materials, restricts σ and ρ to being *scalars* with the *same numerical value* everywhere, and Ohms law now can be formulated for any material with weird shapes and being quite inhomogeneous; we "simply" have

$$\underline{\boldsymbol{j}} = \sigma \cdot \underline{\boldsymbol{E}}$$

- Ohms law in this *vector form* is now valid at *any point* of a body, since we do not have to make assumptions about the shape of the body.
  - Take an arbitrarily shaped body with current flowing through it, cut out a little cube (with your "mathematical" knife) at the coordinates **(*x,y,z*)** without changing the flow of current, and you must find that the local current density and the local field strength obey the equation given above *locally*.

$$\underline{\boldsymbol{j}}(x,y,z) = \sigma \cdot \underline{\boldsymbol{E}}(x,y,z)$$

  - Of course, obtaining the external current ***I*** flowing for the external voltage ***U*** now needs summing up the contributions of all the little cubes, i.e. integration over the whole volume, which may not be an easy thing to do.
- Still, we have now a much more powerful version of Ohms law! But we should now harbor a certain suspicion:
  - There is no good reason why $\underline{\boldsymbol{j}}$ must always be *parallel* to $\underline{\boldsymbol{E}}$. This means that for the most general case σ is not a *scalar* quantity, but a *tensor*; $\sigma = \sigma_{ij}$.
    (There is no good way to write tensors in html; we use the $_{ij}$ index to indicate tensor properties.
  - Ohms law then writes

$$j_x = \sigma_{xx} \cdot E_x + \sigma_{xy} \cdot E_y + \sigma_{xz} \cdot E_z$$

$$j_y = \sigma_{yx} \cdot E_x + \sigma_{yy} \cdot E_y + \sigma_{yz} \cdot E_z$$

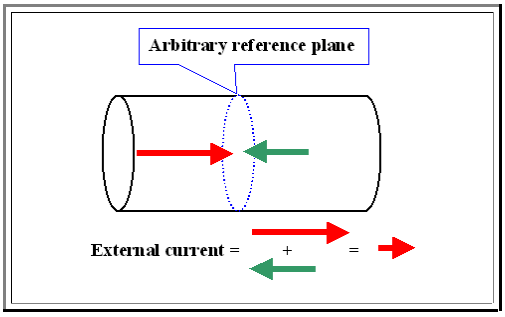
$$j_z = \sigma_{zx} \cdot E_x + \sigma_{zy} \cdot E_y + \sigma_{zz} \cdot E_z$$

- For anisotropic inhomogeneous materials you have to take the tensor, and its components will all depend on the coordinates - that is the most general version of Ohms law.
  - Note that this is *not* so general as to be meaningless: We still have the basic property of Ohms law: The local current density is directly proprotional to the local field strength (and not, for example, to **exp– [const. · $\underline{\boldsymbol{E}}$]**).
- Our goal now is to find a relation that allows to calculate σ_{ij} for a given material (or material composite); i.e. we are looking for
  - $\sigma_{ij} = \sigma_{ij}$**(material, temperature, pressure, defects...)**

2. Step: Describe σ_{ij} in Terms of the Carrier Properties

- Electrical current needs *mobile* charged "things" or *carriers* that are *mobile*. Note that we do not automatically assume that the charged "things" are *always* electrons. *Anything* charged and mobile will do.
- What we want to do now is to express σ_{ij} in terms of the *properties* of the carriers present in the material under investigation.
  - To do this, we will express an electrical current as a "mechanical" stream or current of (charged) particles, and compare the result we get with Ohms law.
- First, lets define an electrical current in a wire in terms of the carriers flowing through that wire. There are *three* crucial points to consider
- **1.** The external electrical current as measured in an Ampèremeter is the result of the *net* current flow through any cross section of an (uniform) wire.
  - In other words, the measured current is proportional to the *difference* of the number of carriers of the same charge sign moving from the *left to right* through a given cross sectional area *minus* the number of carriers moving from the *right to the left*.

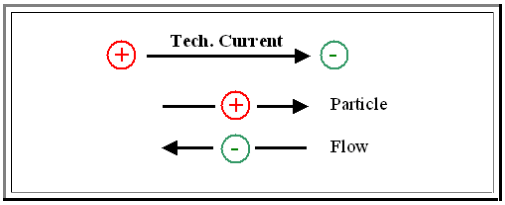
- In short: the *net* current is the difference of two *partial* currents flowing in opposite directions:



- Do not take this point as something simple! We will encounter cases where we have to sum up **8** partial currents to arrive at the externally flowing current, so keep this in mind!

2. In summing up the individual current contributions, *make sure the signs are correct*. The rule is simple:

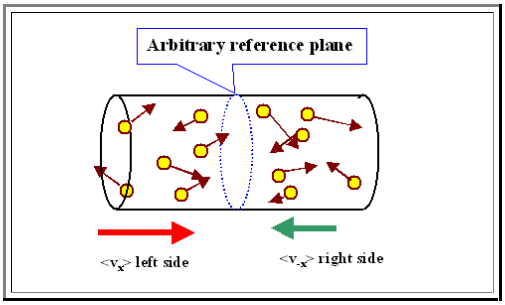
- The *electrical* current is (for historical reasons) defined as flowing from **+** to **–**. For a *particle* current this means:



- In words: A technical current ***I*** flowing from **+** to **–** may be obtained by *negatively* charged carriers flowing in the *opposite* direction (from **–** to **+**), by *positively* charged carriers flowing in the *same* direction, or from both kinds of carriers flowing at the same time in the proper directions.
- The particle currents of *differently* charged particles then must be *added*! Conversely, if negatively charged carriers flow in the same directions as positively charged carriers, the value of the partial current flowing in the "wrong" direction must be subtracted to obtain the external current.

3. The flow of particles through a reference surface as symbolized by one of arrows above, say the arrow in the ***+x*** -direction, must be seen as an *average* over the ***x*** -component of the velocity of the individual particles in the wire.

- Instead of *one* arrow, we must consider as many arrows as there are particles and take their *average*. A more detailed picture of a wire at *a given instant* thus looks like this



- An instant later it looks entirely different *in detail*, but exactly the same *on average*!
- If we want to obtain the net flow of *particles* through the wire (which is obviously proportional to the net *current* flow), we could take the average of the velocity components $\langle v_{+x} \rangle$ pointing in the ***+x*** direction (to the right) on the left hand side, and subtract from this the average $\langle v_{-x} \rangle$ of the velocity components pointing in the ***–x*** direction (to the left) on the right hand side.
- We call this *difference* in velocities the **drift velocity** $\underline{v}_D$ of the *ensemble* of carriers.
- If there is no driving force, e.g. an electrical field, the velocity vectors are randomly distributed and $\langle v_{+x} \rangle = \langle v_{-x} \rangle$; the drift velocity and thus net current is zero as it should be.

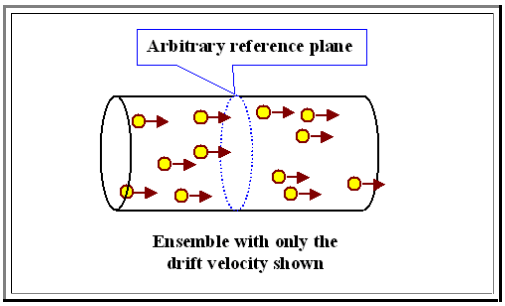
Average properties of ensembles can be a bit tricky. Lets look at some properties by considering the analogy to a localized **swarm of summer flies** "circling" around like crazy, so that the ensemble looks like a small cloud of smoke. A more detailed treatment can be found in 1.3.5.

- First we notice that while the *individual* fly moves around quite fast, its *vector* velocity $\underline{v}_i$ averaged over time ***t***, $\langle \underline{v}_i \rangle_t$, must be zero as long as the swarm as an ensemble doesn't move.
- In other words, the flies, *on average*, move just as often to the left as to the right, etc. The net current produced by *all* flies at any given instance or by *one* individual fly after sufficient time is obviously zero for *any* reference surface.

In real life, however, the fly swarm "cloud" often moves *slowly* around - it has a finite *drift velocity* which must be just the difference between the average movement in drift direction minus the average movement in the opposite direction.

- The *drift velocity* thus can be identified as the proper average that gives the net current through a reference plane perpendicular to the direction of the drift velocity.
- This drift velocity is usually much smaller than the average magnitude of the velocity $\langle \mathbf{v} \rangle$ of the individual flies. Its value is the difference of two large numbers - the average velocity of the *individual* flies in the drift direction minus the average velocity of the *individual* flies in the direction opposite to the drift direction.

Since we are only interested in the drift velocity of the ensemble of flies (or in our case, carriers) we may now simplify our picture as follows:



We now equate the *current density* with the *particle flux density* by the basic law of current flow:

- Current density $\underline{j}$ = Number N of particles carrying the charge q flowing through the cross sectional area F (with the normal vector $\underline{f}$ and $|\underline{f}| = 1$) during the time interval t, or

$$\underline{j} = \frac{q \cdot N}{F \cdot t} \cdot \underline{f}$$

- In scalar notation, because the direction of the current flow is clear, we have

$$j = \frac{q \cdot N}{F \cdot t}$$

The problem with this formula is N, the *number* of carriers flowing through the cross section F every second.

- N is not a basic property of the material; we certainly would much prefer the carrier *density* $n = N/V$ of carriers. The problem now is that we have to chose the volume $V = F \cdot l$ in such a way that it contains just the right number N of carriers.
- Since the cross section F is given, this means that we have to pick the length l in such a way, that all carriers contained in that length of material will have moved across the internal interface after **1** second.
- This is easy! The trick is to give l just that particular length that allows *every* carrier in the defined portion of the wire to reach the reference plane, i.e.

$$l = v_D \cdot t$$

- This makes sure that *all* carriers contained in this length, will have reached F after the time t has passed, and thus *all* carriers contained in the volume $V = F \cdot v_D \cdot t$ will contribute to the current density. We can now write the current equation as follows:

$$j = \frac{q \cdot N}{F \cdot t} = \frac{q \cdot n \cdot V}{F \cdot t} = \frac{q \cdot n \cdot F \cdot l}{F \cdot t} = \frac{q \cdot n \cdot F \cdot v_D \cdot t}{F \cdot t}$$

This was shown in excessive detail because now we have the *fundamental law of electrical conductivity* (in obvious vector form)

$$\underline{j} = q \cdot n \cdot \underline{v}_D$$

- This is a very general equation relating a *particle current* (density) via its *drift velocity* to an *electrical current* (density) via the charge ***q*** carried by the particles.
  - Note that it does not matter at all, *why* an ensemble of charged particles moves on average. You do not need an electrical field as driving force anymore. If a concentration gradient induces a particle flow via diffusion, you have an electrical current too, if the particles are charged.
  - Note also that electrical current flow *without* an electrical field as primary driving force as outlined above is *not* some odd special case, but at the root of most electronic devices that are more sophisticated than a simple resistor.
  - Of course, if you have different particles, with different density drift velocity and charge, you simply sum up the individual contributions as pointed out above.
- All we have to do now is to compare our equation from above to Ohms law:

$$\underline{j} = q \cdot n \cdot \underline{v_D} := \sigma \cdot \underline{E}$$

  - We then obtain

$$\sigma = \frac{q \cdot n \cdot v_D}{E} := \text{constant}$$

- If Ohms law holds, σ must be a constant, and this implies *by necessity*

$$\frac{v_D}{E} = \text{constant}$$

  - And this is a simple, but far reaching equation saying something about the driving force of electrical currents (= electrical field strength ***E***) and the drift velocity of the particles in the material.
  - What this means is that *if* $v_D/E = \text{const.}$ holds for *any* (reasonable) field ***E***, the material will show ohmic behavior. *We have a first condition for ohmic behavior expressed in terms of material properties*.
  - If, however, v_D/E is constant (in time) for a *given* field, but with a value that depends on **E**, we have $\sigma = \sigma(E)$; the behavior will *not be ohmic*!
- The requirement $v_D/E = \text{const.}$ for *any* electrical field thus requires a drift velocity in field direction for the particle, which is directly proportional to ***E***. This leads to a simple conclusion:
  - This is actually a rather strange result! A charged particle in an electrical field experiences a constant force, and Newtons first law tells us that this will induce a constant accelerations, i.e. its velocity should increase all the time! Its velocity therefore would grow to infinity - if there wouldn't be some kind of friction.
  - We thus conclude that there *must* exist some mechanism that acts like a frictional force on all accelerated particles, and that this frictional force in the case of ohmic behavior must be in a form where *the average drift velocity obtained is proportional to the driving force*.
- Since $v_D/E = \text{constant}$ must obtain for all (ohmic) materials under investigation, we may give it a name:

$$\frac{v_D}{E} = \mu = \text{Mobility} = \text{Material constant}$$

  - The **mobility** ***μ*** of the carriers has the unit
    [μ] = (m/s)/(V/m) = $m^2/V \cdot s$.
  - The ***mobility*** **μ** (Deutsch: **Beweglichkeit**) then is a *material constant*; it is determined by the "friction", i.e. the processes that determine the average velocity for carriers in different materials subjected to the same force $q \cdot \underline{E}$.
  - **Friction**, as we (should) know, is a rather unspecified term, but always describing energy transfer from some moving body to the environment.
  - Thinking ahead a little bit, we might realize that **μ** is a basic material constant *even in the absence of electrical fields*. Since it is tied to the "friction" a moving carrier experiences in its environment - the material under consideration - it simply expresses how fast carriers give up surplus energy to the lattice; and it must not matter how they got the surplus energy. It is therefore no suprise if **μ** pops up in all kinds of relations, e.g. in the famous Einstein - Smoluchowski equation linking *diffusion coefficients* and *mobility* of particles.

- We now can write down the *most general form of Ohms law* applying to all materials meeting the two requirements: ***n* = const.** and **μ = const.** everywhere. It is expressed completely in particle (= material) properties.

$$\sigma = q \cdot n \cdot \mu$$

  - The task is now to calculate ***n*** and **μ** from first priciples, i.e. from only knowing what atoms we are dealing with in what kind of structure (e.g. crystal + crystal defects)
  - This is a rather formidable task since σ varies over a extremely wide range, cf. a short table with some relevant numbers.
- In order to get acquainted with the new entity "mobility", we do a little exercise:

Exercise 1.3-1

Derive and dicuss numbers for **μ**

- Since we like to give σ as a positive number, we always take only the magnitude of the charge ***q*** carried by a particle.
  - However, if we keep the sign, e.g. write $\sigma = -e \cdot n \cdot \mu_e$ for electrons carrying the charge $q = -e$; **e =** elementary charge, we now have an indication if the particle current and the electrical current have the *same* direction ($\sigma > 0$) or opposite directions ($\sigma < 0$) as in the case of electrons.
  - But it is entirely a matter of taste if you like to ***schlepp*** along the signs all the time, or if you like to fill 'em in at the end.
- Everything more detailed then this is no longer universal but specific for certain materials. The remaining task is to calculate ***n*** and **μ** for given materials (or groups of materials).
  - This is not too difficult for simple materials like metals, where we know that there is one (or a few) free electrons per atom in the sample - so we know ***n*** to a sufficient approximation. Only **μ** needs to be determined.
  - This is fairly easily done with classical physics; the results, however, are flawed beyond repair: They just do not match the observations and the unavoidable conclusion is that classical physics must not be applied when looking at the behavior of electrons in simple metal crystals or in any other structure - we will show this in the immediately following subchapter **2.1.3**.
- We obviously need to resort to **quantum theory** and solve the **Schrödinger equation** for the problem.
  - This, surprisingly, is also fairly easy in a simple approximation. The math is not too complicated; the really difficult part is to figure out what the (mathematical) solutions actually *mean*. This will occupy us for quite some time.

Questionaire

Multiple Choice Questions to 1.3.2

Required Reading

1.3.3 Ohms Law and Classical Physics

- In this subchapter we will look at the *classical* treatment of the movement of electrons inside a material in an electrical field.
  - In the preceding subchapter we obtained the most basic formulation of *Ohms law*, linking the specific conductivity to two fundamental material parameters:

$$\sigma = q \cdot n \cdot \mu$$

- For a homogeneous and isotropic material (***e.g.*** polycrystalline metals or single crystal of cubic semiconductors), the *concentration* of carriers n and their *mobility* μ have the same value everywhere in the material, and the specific conductivity σ is a *scalar*.
  - This is boring, however. So let's look at useful complications:
- In general terms, we may have more than one kind of carrier (this is the common situation in semiconductors) and n and μ could be functions of the temperature T, the *local* field strength E_{loc} resulting from an applied *external* voltage, the detailed structure of the material (e.g. the defects in the lattice), and so on.
  - We will see that these complications are the essence of advanced electronic materials (especially semiconductors), but in order to make life easy we first will restrict ourselves to the special class of **ohmic materials**.
  - We have seen before that this requires n and μ to be independent of the local field strength. However, we still may have a temperature dependence of σ; even commercial ohmic resistors, after all, do show a more or less pronounced temperature dependence - their resistance increases roughly linearly with T.
- In short, we are treating *metals*, characterized by a *constant* density of *one* kind of carriers (**=** electrons) in the order of **1 ...3** electrons per atom in the metal.

Basic Equations and the Nature of the "Frictional Force"

- We consider the electrons in the metal to be "free", i.e. they can move freely in any direction - the atoms of the lattice thus *by definition* do not impede their movement
  - The (local) electrical field $\underline{E}$ then exerts a force $\underline{F} = -e \cdot \underline{E}_{loc}$ on any given electron and thus accelerates the electrons in the field direction (more precisely, opposite to the field direction because the field vector points from **+** to **–** whereas the electron moves from **–** to **+**).
  - In the fly swarm analogy, the electrical field would correspond to a steady airflow - some wind - that moves the swarm about with constant drift velocity.
- Now, if a single electron with the (constant) mass **m** and momentum $\underline{p}$ is subjected to a force $\underline{F}$, the equation of motion from basic mechanics is

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

  - Note that $\underline{p}$ does not have to be zero when the field is switched on.
- If this would be all, the velocity of a given electron would acquire an ever increasing component in field direction and eventually approach infinity. This is obviously not possible, so we have to bring in a mechanism that destroys an unlimited increase in $\underline{v}$.
  - In classical mechanics this is done by introducing a **frictional force** $\underline{F}_{fr}$ that is proportional to the velocity.

$$\underline{F}_{fr} = -k_{fr} \cdot \underline{v}$$

- with k_{fr} being some friction constant. But this, while mathematically sufficient, is *devoid of any physical meaning* with regard to the moving electrons.
- There is no "*friction*" on an atomic scale! Think about it! Where should a friction force come from? An electron feels *only* forces from two kinds of fields - electromagnetic and gravitational (neglecting strange stuff from particle physics). So we have to look for another approach.

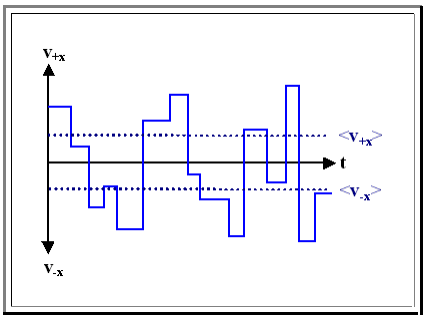
- What friction does to big classical bodies is to ***dissipate*** ordered kinetic energy of the moving body to the environment. Any ordered movement gets slowed down to zero (surplus) speed, and the environment gets somewhat hotter instead, i.e. *unordered* movement has increased.
  - This is called **energy dissipation**, and that is what we need: Mechanisms that take kinetic energy away from an electron and "give" it to the crystal at large. The science behind that is called *(Statistical) Thermodynamics* - we have encounterd it before.
- The best way to think about this, is to assume that the electron, flying along with increasing velocity, will *hit something else* along its way every now and then; it has a *collision* with something else, it will be **scattered** at something else.
  - This collision or scattering event will change its *momentum*, i.e. the magnitude and the direction of $\underline{\mathbf{v}}$, and thus also its kinetic energy E_{kin}, which is always given by

$$E_{kin} = \frac{m \cdot \underline{v}^2}{2} = \frac{\underline{p} \cdot v}{2}$$

- In other words, we consider collisions with something else, i.e. other particles (including "pseudo" particles), where the total energy and momentum of all the particles is preserved, but the individual particle looses its "memory" with respect to its velocity before the collision, and starts with a new momentum after every collision.
- What are the "partners" for collisions of an electron, or put in standard language, what are the **scattering mechanisms**? There are several possibilities:
  - Other ***electrons***. While this happens, it is not the important process in most cases. It also does not decrease the energy contained in the electron movement - the losses of some electron are the gains of others.
  - ***Defects***, e.g. foreign atoms, point defects or dislocations. This is a more important scattering mechanism and moreover a mechanism where the electron can transfer its surplus energy (obtained through acceleration in the electrical field) to the lattice, which means that the material heats up
  - ***Phonons***, i.e. *"quantized" lattice vibrations* traveling through the crystal. *This is the most important scattering mechanism*.
- Now that is a bit strange. While we (hopefully) have no problem imagining a crystal lattice with all atoms vibrating merrily, there is no immediate reason to consider these vibrations as being *localized* (whatever this means) and *particle-like*.
  - You are right - but nevertheless: The lattice vibrations indeed are best described by a bunch of particle-like **phonons** careening through the crystal.
  - This follows from a quantum mechanical treatment of lattice vibrations. Then it can be shown that these vibrations, which contain the thermal energy of the crystal, are quantized and show typical properties of (quantum) particles: They have a *momentum*, and an *energy* given by $\mathbf{h}\nu$ (**h =** Plancks constant, ν **=** frequency of the vibration).
- *Phonons* are a first example of "pseudo" particles; but there is no more "pseudo" to phonons than there is to photons.
  - We will not go into more details here. All we need to know is that a hot crystal has more phonons *and* more energetic phonons than a cold crystal, and treating the interaction of an electron with the lattice vibration as a collision with a phonon gives not only *correct* results, it is the *only* way to get results at all.
- At this point comes a crucial insight: It would be far from the truth to assume that only ***accelerated*** electrons scatter; scattering happens all the time to all the electrons moving randomly about because they all have some thermal energy. Generally, scattering is the mechanism to achieve thermal equilibrium and equidistribution of the energy of the crystal.
  - If electrons are accelerated in an electrical field and thus gain energy in excess of thermal equilibrium, scattering is the way to transfer this surplus energy to the lattice which then will heat up. If the crystal is heated up from the outside, scattering is the mechanism to turn heat energy contained in lattice vibrations to kinetic energy of the electrons.
  - Again: Even without an electrical field, scattering is the mechanism to transfer thermal energy from the lattice to the electrons (and back). Generally, scattering is the mechanism to achieve *thermal equilibrium* and equidistribution of the energy of the crystal.
  - Our free electrons in metals behave very much like a gas in a closed container. They ***careen*** around with some average velocity that depends on the energy contained in the **electron gas**, which is - in classical terms- a direct function of the temperature.

Averaging over Random Scattering Events

Lets look at some figures illustrating the scattering processes.



- Shown here is the *magnitude* of the velocity $\underline{v}_{\pm x}$ of an electron in ***+x*** and ***–x*** direction *without* an external field. The electron moves with constant velocity until it is scattered, then it continues with some new velocity.
- The scattering processes, though unpredictable at single events, must lead to the averages of the velocity, which is characteristic for the material and its conditions.
- The plural in "average*s*" is intentional: there *are* different averages of the velocity
- Whereas $\langle \underline{v} \rangle = 0$, $\langle v \rangle$ has a finite value; this is also true for $\langle \underline{v}_x \rangle = -\langle \underline{v}_{-x} \rangle$. Consult the "fly swarm modul" if you are unsure about this.

- From classical thermodynamics we know that the (classical) electron gas in thermal equilibrium with the environment contains the energy $E_{kin} = (1/2)kT$ per particle and degree of freedom, with k = Boltzmanns constant and T = absolute temperature. The three degrees of freedom are the velocities in ***x***-, ***y***- and ***z***-direction, so we must have

$$E_{kin,x} = \tfrac{1}{2} \cdot m \cdot \langle v_x^2 \rangle = \tfrac{1}{2} \cdot kT$$

$$\langle v_x \rangle = \left(\frac{kT}{m} \right)^{1/2}$$

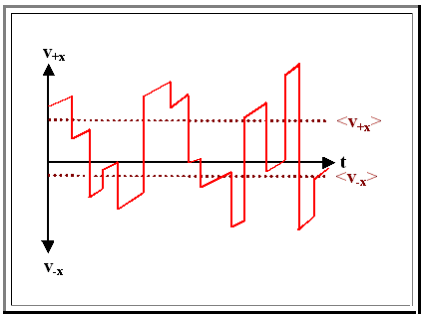
- For the other directions we have exactly the same relations, of course. For the total energy we obtain

$$E_{kin} = \frac{m \cdot \langle v_x^2 \rangle}{2} + \frac{m \cdot \langle v_y^2 \rangle}{2} + \frac{m \cdot \langle v_z^2 \rangle}{2} = \frac{m \cdot \langle v^2 \rangle}{2} = \frac{m \cdot (v_0)^2}{2} = \frac{3kT}{2}$$

- with $v_0 = \langle v \rangle$. v_0 is thus the **average velocity of a carrier** careening around in a crystal.
- At this point you should stop a moment and think about just how fast those electrons will be careening around at room temperature (**300K**) without plugging numbers in the equation. Got a feeling for it? Probably not. So look at the exercise question (and the solution) further down!.
- Now you should stop another moment and become very aware of the fact that this equation is from purely *classical* physics. It is absolutely true for *classical* particles - which electrons are actually not. Electrons obey the Pauli principle, i.e. they behave about as non-classical behavior as it is possible. This should make you feel a bit uncomfortable. Maybe the equation from above is not correct for electrons then? Indeed - it isn't. Why, we will see later; also how we can "repair" the situation!

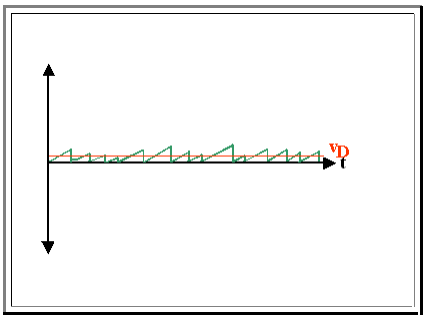
Now lets *turn on an electrical field*. It will accelerate the electrons *between* the collisions. Their velocity in field direction then increases linearly from whatever value it had right after a collision to some larger value right before the next collision.

- In our diagram from above this looks like this:



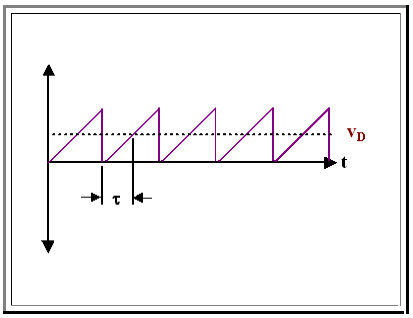
- Here we have an electrical field that accelerates electrons in in ***x***-direction (and "brakes" in ***–x*** direction). Between collisions, the electron gains velocity in ***+x***-direction at a constant rate (= identical slope).
- The average velocity in ***+x*** directions, $\langle v_{+x} \rangle$, is now larger than in ***–x*** direction, $\langle v_{-x} \rangle$.
- However, beware of the pitfalls of schematic drawings: For real electrons the difference is very small as we shall see shortly; the slope in the drawing is very exaggerated.

- The drift velocity is contained in the difference $\langle v_{+x} \rangle - \langle v_{-x} \rangle$; it is completely described by the velocity gain between collisions. For obtaining a value, we may neglect the instantaneous velocity right after a scattering event because they average to zero anyway and just plot the *velocity gain* in a simplified picture; always starting from zero after a collision.



- The picture now looks quite simple; but remember that it contains some not so simple averaging.
- At this point it is time to define a very meaningful *new* average quantity:
- The *mean time between collisions*, or more conventional, the mean time τ for reaching the drift velocity **v** in the simplified diagram. We also call τ the **mean scattering time** or just *scattering time* for short.

This is most easily illustrated by simplifying the scattering diagram once more: We simply use just *one* time - the average - for the time that elapses between scattering events and obtain:



- This is the *standard diagram* illustrating the scattering of electrons in a crystal usually found in text books; the definition of the scattering time τ is included
- It is highly idealized, if not to say just wrong if you compare it to the correct picture above. Of course, the average velocity of both pictures will give the same value, but that's like saying that the average speed v_a of all real cars driving around in a city is the same as the average speed of ideal model cars all going at v_a all the time.
- Note that τ is only *half* of the average time between collisions.

So, while this diagram is not wrong, it is a highly abstract rendering of the underlying processes obtained after several averaging procedures. From this diagram only, no conclusion whatsoever can be drawn as to the average velocities of the electrons without the electrical field!

New Material Parameters and Classical Conductivity

With the scattering concept, we now have two new (closely related) material parameters:

- The **mean (scattering) time** τ between two collisions as defined before, and a directly related quantity:
- The **mean free path** l between collisions; i.e. the distance travelled by an electron (on average) before it collides with something else and changes its momentum. We have

$$l = 2\tau \cdot (v_0 + v_D)$$

- Note that v_0 enters the defining equation for l, and that we have to take twice the scattering time τ because it only refers to half the time between collisions!

After we have come to this point, we now can go on: Using τ as a new parameter, we can rewrite Newtons equation from above:

$$\frac{dv}{dt} = \frac{\Delta v}{\Delta t} = \frac{v_D}{\tau}$$

- It is possible to equate the *differential* quotient with the *difference* quotient, because the velocity change is constant. From this we obtain

$$\frac{v_D}{\tau} = -\frac{E \cdot e}{m}$$

$$\Rightarrow \quad v_D = -\frac{E \cdot e \cdot \tau}{m}$$

- Inserting this equation for v_D in the old definition of the current density $j = -n \cdot e \cdot v_D$ and invoking the general version of Ohms law, $j = \sigma \cdot E$, ***yields***

$$j = \frac{n \cdot e^2 \cdot \tau}{m} \cdot E \quad := \quad \sigma \cdot E$$

  - This gives us the final result

$$\sigma = \frac{n \cdot e^2 \cdot \tau}{m}$$

- This is the *classical* formula for the conductivity of a classical "electron gas" material; i.e. metals. The conductivity contains the density n of the free electrons and their mean scattering time τ as material parameters.
  - We have a good idea about n, but we do not yet know τ_{class}, the mean *classical* scattering time for classical electrons. However, since we know the order of magnitude for the conductivity of metals, we may turn the equation around and use it to calculate the order of magnitude of τ_{class}. If you do the exercise farther down, you will see that the result is:

$$\tau_{class} = \frac{\sigma \cdot m}{n \cdot e^2} \approx (10^{-13} \ldots 10^{-15}) \text{ sec}$$

  - "Obviously" (as stated in many text books), this is a value that is *far too small* and thus the classical approach must be *wrong*. But is it really too small? How can *you* tell without knowing a lot more about electrons in metals?
  - Let's face it: *you can't !!*. So let's look at the ***mean free path l*** instead. We have

$$l = 2 \cdot \tau \cdot (v_0 + v_D)$$

and

$$(v_0)^2 = \frac{3kT}{m}$$

  - The last equation gives us a value $v_0 \approx 10^4$ **m/s** at room temperature! Now we need v_D, and this we can estimate from the equation given above to $v_D = -E \cdot \tau \cdot e/m \approx 1$ **mm/sec**, *if* we use the value for τ dictated by the measured conductivities. It is much smaller than v_0 and can be safely neglected in calculating l.
- We thus can rewrite the equation for the conductivity and obtain

$$\sigma = \frac{n \cdot e^2 \cdot l}{2 \cdot m \cdot (v_0 + v_D)}$$

- Knowing σ from experiments, but not l, allows to determine l. The smallest possible mean free path l_{min} between collisions (for $v_D = 0$) thus is

$$l_{min} = \frac{2 \cdot m \cdot v_0 \cdot \sigma}{n \cdot e^2} = 2 \cdot v_0 \cdot \tau \approx (10^{-1} - 10^{1})\ \text{nm}$$

  - *And this is certainly too small!*.

- But before we discuss these results, let's see if they are actually true by doing an exercise:

Exercise 1.3-2

Derive numbers for v_0, τ, v_D, and l

- Now to the important question: *Why* is a mean free path in the order of the size of an atom too small?
  - Well, think about the scattering mechanisms. The distance between lattice defects is certainly much larger, and a phonon itself is "larger", too.
  - Moreover, consider what happens at temperatures below room temperatures: l would become even smaller since v_0 decreases - somehow this makes *no sense*.
- It does not pay to spend more time on this. Whichever way you look at it, whatever tricky devices you introduce to make the approximations better (and physicists have tried very hard!), you will *not* be able to solve the problem: The mean free paths are never even coming close to what they need to be, and the conclusion which we will reach - maybe reluctantly, but unavoidably - must be:

There is no way to describe conductivity (in metals)
with *classical* physics!

Scattering and Mobility

- Somewhere on the way, we have also indirectly found that the **mobility μ** as defined before is just another way to look at scattering mechanisms. Let's see why.
  - All we have to do is to compare the equation for the conductivity from above with the master equation $\sigma = q \cdot n \cdot \mu$.
- This gives us immediately

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

$$\mu \approx \frac{e \cdot l}{2 \cdot m \cdot v_0}$$

  - In other words:

The *decisive* material property determining the mobility μ is the average time between scattering events or the mean free path between those events.

- The mobility **μ** thus is a basic material property, well-defined even without electrical fields, and just another way to express the scattering processes taken place by a number.

In the equations above slumbers an extremely important aspect of semicoductor technology.

- In all electronic devices carriers have to travel some distance before a signal can be produced. A **MOS** transistor, for example, switches currents on or off between its "Source" and "Drain" terminal depending on what voltage is applied to its "Gate". Source and drain are separated by some distance l_{SD}, and the "Drain" only "feels" the "on" state after the time it takes the carriers to run the distacne l_{SD}.
- How long does that take if the voltage between Source and Drain is U_{SD}?
- Easy. If we know the mobility **μ** of the carriers, we now their (average) velocity v_{SD} in the source-drain region, which by definition is $v_{SD} = \mu \cdot U_{SD}/l_{SD}$.
- The traveling time t_{SD} between source and drain for obvious reasons defines roughly the maximum frequency f_{max} the transistor can hand, we have $t_{SD} = l_{SD} / v_{SD}$ or

$$t_{SD} = \frac{l_{SD}^2}{\mu \cdot U_{SD}} \approx \frac{1}{f_{max}}$$

- The maximum frequency of a **MOS** transistor thus is directly proportional to the mobiltiy of the carriers in the material it is made from (always provided there are no other limiting factors). And since we used a rather general argument, we should not be surprised that pretty much the same relation is also true for most electronic devices, not just MOS transistors.
- This is a momentous statement: We linked a prime material parameter, the *material constant* **μ**, to one of the most important parameters of electronic circuits. We would like **μ** to be as large as possible, of course, and now we know what to do about it!

A simple exercise is in order to see the power of this knowlegde:

Exercise 1.3-3

What does it take to build a **4 GHz** microprocessor?

Questionaire

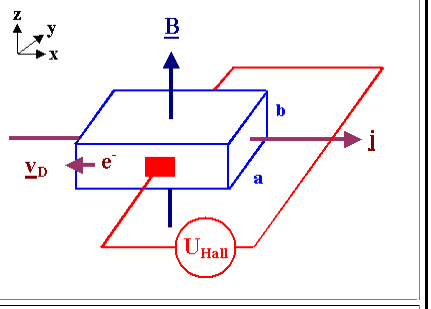
Multiple Choice Questions to 1.3.3

This Module comes from "Introduction to Materials Science II"; where you will find the most recent version. The link takes you there.

Required Reading

1.3.4 The Hall Effect

- This subchapter introduces *two* important topics: The **Hall effect** as an important observation in materials science and at the same time another irrefutable proof that classical physics just can't hack it when it comes to electrons in crystals.
  - The Hall effect describes what happens to current flowing through a conducting material - a metal, a semiconductor - if it is exposed to a magnetic field $\underline{\boldsymbol{B}}$.
  - We will look at this in *classical* terms; again we will encounter a fundamental problem.
- The standard geometry for doing an experiment in its most simple form is as follows:



Relevant vectors; the current is carried by electrons

  - A magnetic field $\underline{\boldsymbol{B}}$ is employed perpendicular to the current direction $\underline{\boldsymbol{j}}$, as a consequence a *potential difference* (i.e. a *voltage*) develops at right angles to both vectors.
  - In other words: A **Hall voltage** $\boldsymbol{U_{Hall}}$ will be measured perpendicular to $\underline{\boldsymbol{B}}$ and $\underline{\boldsymbol{j}}$.
  - In yet other words: An electrical field $\underline{\boldsymbol{E}}_{\mathbf{Hall}}$ develops in $\boldsymbol{y}$-direction
  - That is already the essence of the Hall effect.
- It is relatively easy to calculate the magnitude of the *Hall voltage* $\boldsymbol{U_{Hall}}$ that is induced by the magnetic field $\underline{\boldsymbol{B}}$.
  - First we note that we must also have an electrical field $\underline{\boldsymbol{E}}$ parallel to $\underline{\boldsymbol{j}}$ because it is the driving force for the current.
  - Second, we know that a magnetic field at right angles to a current causes a force on the moving carriers, the so-called **Lorentz force** $\boldsymbol{F_L}$, that is given by

$$\underline{\boldsymbol{F}}_L = q \cdot (\underline{\boldsymbol{v}}_D \times \underline{\boldsymbol{B}})$$

  - We have to take the drift velocity $\underline{\boldsymbol{v}}_D$ of the carriers, because the other velocities (and the forces caused by these components) cancel to zero on average. The vector product assures that $\underline{\boldsymbol{F}}_L$ is perpendicular to $\underline{\boldsymbol{v}}_D$ and $\underline{\boldsymbol{B}}$.
  - Note that instead the usual word "electron" the neutral term *carrier* is used, because in principle an electrical current could also be carried by charged particles other then electrons, e.g. positively charged ions. Remember a simple but important picture given before!
- For the geometry above, the Lorentz force $\underline{\boldsymbol{F}}_L$ has only a component in $\boldsymbol{y}$ - direction and we can use a scalar equation for it. $\boldsymbol{F_y}$ is given by

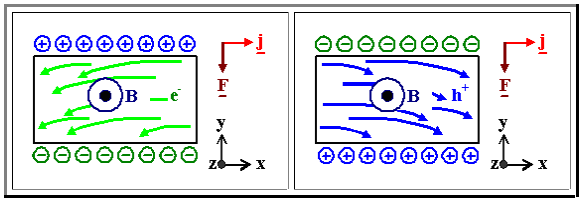
$$F_y = -q \cdot v_D \cdot B_z$$

  - We have to be a bit careful, however: We know that the force is in $\boldsymbol{y}$-direction, but we do longer know the sign. It changes if either $\boldsymbol{q}$, $\boldsymbol{v_D}$, or $\boldsymbol{B_z}$ changes direction and we have to be aware of that.
- With $\underline{\boldsymbol{v}}_D = \mu \cdot \underline{\boldsymbol{E}}$ and $\mu =$ mobility of the carriers, we obtain a rather simple equation for the force

$$F_y = -q \cdot \mu \cdot E_x \cdot B_z$$

  - It is important to note that for a fixed current density $\boldsymbol{j_x}$ the direction of the Lorentz force is independent of the sign of the charge carriers (the sign of the charge and the sign of the drift velocity just cancel each other).

This means that the current of carriers will be deflected from a straight line in ***y***-direction. In other words, there is a component of the velocity in ***y***-direction and the surfaces perpendicular to the ***y***-direction will become charged as soon as the current (or the magnetic field) is switched on. The flow-lines of the carriers will look like this:



- The charging of the surfaces is unavoidable, because some of the carriers eventually will end up at the surface where they are "stuck".
- Notice that the sign of the charge for a given surface depends on the sign of the charge of the carriers. Negatively charged electrons (**e^-** in the picture) end up on the surface opposite to positively charged carriers (called **h^+** in the picture).
- Notice, too, that the direction of the force $\underline{F}_y$ is the same for both types of carriers, simply because both q and $\underline{v}_D$ change signs in the force formula

The surface charge then induces an electrical field E_y in ***y***-direction which opposes the Lorentz force; it tries to move the carriers back.

- In *equilibrium*, the Lorentz force $\underline{F}_y$ and the force from the electrical field E_y in ***y***-direction (which is of course simply $q \cdot E_y$) must be equal with opposite signs. We obtain

$$q \cdot E_y = -q \cdot \mu \cdot E_x \cdot B_z$$

$$E_y = -\mu \cdot E_x \cdot B_z$$

The Hall voltage U_{Hall} now is simply the field in ***y***-direction multiplied by the dimension d_y in ***y***-direction.

- It is clear then that the (easily measured) Hall voltage is a *direct measure* of the mobility **μ** of the carriers involved, and that its **sign** or polarity will change if the sign of the charges flowing changes.

It is customary to define a **Hall coefficient** R_{Hall} for a given material.

- This can be done in different, but equivalent ways. In the link we look at a definition that is particularly suited for measurements. Here we use the following definition:

$$R_{Hall} = \frac{E_y}{B_z \cdot j_x}$$

In other words, we expect that the Hall voltage $E_y \cdot d_y$ (with d_y = dimension in ***y***-direction) is proportional to the current(density) j and the magnetic field strength B, which are, after all, the main experimental parameters (besides the trivial dimensions of the specimen):

$$E_y = R_{Hall} \cdot B_z \cdot j_x$$

The Hall coefficient is a material parameter, indeed, because we will get different numbers for R_{Hall} if we do experiments with identical magnetic fields and current densities, but different materials. The Hall coefficient, as mentioned before, has interesting properties:

- R_{Hall} will change its sign, if the sign of the carriers is changed because then E_y changes its sign, too. It thus indicates in the most unambiguous way imaginable if positive or negative charges carry the current.
- R_{Hall} allows to obtain the mobility **μ** of the carriers, too, as we will see immediately

R_{Hall} is easily calculated: Using the equation for E_y from above, and the basic equation $j_x = \sigma \cdot E_x$, we obtain for *negatively* charged carriers:

$$R_{Hall} = -\frac{\mu \cdot E_x \cdot B_z}{\sigma \cdot E_x \cdot B_z} = -\frac{\mu}{\sigma}$$

Measurements of the Hall coefficient of materials with a *known* conductivity thus give us *directly* the mobility of the carriers responsible for the conductance.

- The **–** sign above is obtained for *electrons*, i.e. *negative* charges.
- If positively charged carriers would be involved, the Hall constant would be positive.
- Note that while it is not always easy to measure the numerical value of the Hall voltage and thus of ***R*** with good precision, it is the easiest thing in the world to measure the *polarity* of a voltage.

Lets look at a few experimental data:

Material	**Li**	**Cu**	**Ag**	**Au**	**Al**	**Be**	**In**	Semiconductors (e.g. **Si**, **Ge**, **GaAs**, **InP**,...)
R ($\times 10^{-24}$) cgs units	–1,89	–0,6	–1,0	–0,8	+1,136	+2,7	+1,774	*positive* or *negative* values, depending on "doping"

Comments:
1. the *positive* values for the metals were measured under somewhat special conditions (low temperatures; single crystals with special orientations), for other conditions negative values can be obtained, too.
2. The units are not important in the case, but multiplying with **$9 \cdot 10^{13}$** yields the value in **m^3/Coulomb**

Whichever way we look at this, one conclusion is unavoidable:

- In certain materials including *metals*, the particles carrying the electrical current are *positively charged* under certain conditions. And this is *positively not possible* in a classical model that knows only *negatively charged electrons* as carriers of electrical current in solids!

Again we are forced to conclude:

There is no way to describe conductivity in metals and semiconductors with *classical* physics!

Questionaire

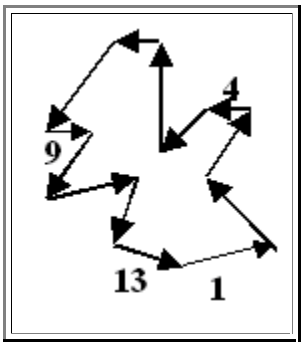
Multiple Choice Fragen zu 1.3.4

This Module comes from "Introduction to Materials Science II"; where you will find the most recent version. The link takes you there.

Required Reading

1.3.5 Averaging Vectors

- Lets look at bit closer at the averages one can take by considering a (localized) swarm of summer flies "circling" around like crazy, so that the ensemble looks like a small cloud of smoke.
  - "Localized" means that the swarm maintains a defined form (on average), so all flies are always inside this defined volume in space.
  - In the case of the charge carriers in a piece of metal, it simple implies that the carriers stay inside the piece of metal.
- First we notice again that while the *individual* fly moves around quite fast, its average *vector velocity* $\langle \underline{\mathbf{v}}_i \rangle_t$, averaged over time t, must be zero as long as the swarm as an ensemble doesn't move.
  - In other words, the flies, on average, move just as often to the left as to the right etc. The net current flowing through some surface produced by *all flies at any given instance*, or by *one individual fly after sufficient time* is obviously zero for *any* reference surface you care to chose. This is illustrated schematically below.



  - Shown are **13** velocity vectors of an individual fly; the chain of vectors closes so $\langle \underline{\mathbf{v}}_i \rangle_t = \mathbf{0}$.

- The average of the *magnitude* of the velocity of an *individual* fly, $\langle |\underline{\mathbf{v}}_i| \rangle_T = \langle \mathbf{v}_i \rangle_t$, however, is obviously *not* zero - the fly, after all, is buzzing around at high (average) speed. *Note the details in the equation above:* Only the underlining of **v** is different!
  - If we define $\langle \mathbf{v}_i \rangle_t$ as follows, we have a simple way of obtaining the average of the magnitude (we take only the positive root, of course) .

$$\langle \mathbf{v}_i \rangle_t = + \langle (\underline{\mathbf{v}}^2_i)^{1/2} \rangle_t$$

  - $\underline{\mathbf{v}}^2$ is a scalar, and the (positive) square root of $\underline{\mathbf{v}}^2$ gives always the (positive) magnitude of $\underline{\mathbf{v}}$; i.e. $|\underline{\mathbf{v}}|$
  - This is an elegant and workable definition, but beware:
    $\langle (\underline{\mathbf{v}}^2)^{1/2} \rangle$ is *not* the same as
    $(\langle \underline{\mathbf{v}}^2 \rangle)^{1/2}$!
    Lets try it with a few arbitrary numbers ⇒

$\lvert\underline{\mathbf{v}}\rvert$ =	3	4	6
$\langle (\underline{\mathbf{v}}^2)^{1/2} \rangle$ =	(3 + 4 + 6)/3 = 13/3 = 4,333...		
$(\langle \underline{\mathbf{v}}^2 \rangle)^{1/2}$ =	$[(9 + 16 + 36)/3]^{1/2} = 20{,}33^{1/2} = 4{,}51$		

- If we have $\langle \mathbf{v} \rangle_t = \langle (\underline{\mathbf{v}}^2)^{1/2} \rangle_t$, we may also calculate the average (over time) of the velocity *components* in x, y, and z - direction, $\langle \mathbf{v}_x \rangle_t$, $\langle \mathbf{v}_y \rangle_t$, $\langle \mathbf{v}_z \rangle_t$, of an individual fly *for a truly random movement*. (*We drop the index "i" now to make life easier*).
  - Again, the vector averages $\langle \underline{\mathbf{v}}_x \rangle$ and so on of the *vector* components must be **= 0** because in a truly random movement the components in **+** x and **-** x direction and so on must cancel on average.
  - Since the magnitude $|\underline{\boldsymbol{A}}|$ of a vector $\underline{\boldsymbol{A}}$ is given by the square root of the scalar product of the vector with itself . We have

$$\underline{A} \cdot \underline{A} = A_x \cdot A_x + A_y \cdot A_y + A_z \cdot A_z = A^2$$

$$A = |\underline{A}| = (\underline{A}^2)^{1/2}$$

- Since

$$\langle v^2 \rangle_t = \langle v_x^2 \rangle_t + \langle v_y^2 \rangle_t + \langle v_z^2 \rangle_t \ ,$$

- and since in a *truly random movement* we have

$$\langle v_x \rangle_t = \langle v_y \rangle_t = \langle v_z \rangle_t \ ,$$

- we end up with

$$\langle v^2 \rangle_t = 3 \langle v_x^2 \rangle$$

$$\langle v_x^2 \rangle_t = 1/3 \langle v^2 \rangle \ .$$

- From this we finally get

$$\langle v_x \rangle_t = \langle (v_x^2)^{1/2} \rangle_t = (1/3)^{1/2} \cdot \langle (v^2)^{1/2} \rangle_t = \frac{\langle v \rangle_t}{3^{1/2}}$$

In real life, however, the fly swarm "cloud" often moves slowly around - it has a finite **drift velocity** $\underline{v}_D$.

$$\underline{v}_D = \langle \underline{v}_i \rangle_t$$

- In consequence, $\langle \underline{v}_i \rangle_t$ is not zero, and $\langle v_{i,+x} \rangle_t$ (= average velocity component in **+*x*** direction) in general is different from $\langle v_{i,-x} \rangle_t$.
- Note that the drift velocity by definition is an average over vectors; we do not use the $\langle\rangle$brackets to signify that anymore. Note also, that the drift velocity of the *fly swarm* and the drift velocity of an *individual fly* must be identical if the swarm is to stay together.
- Without prove, it is evident that $v_{D,i,x} = \langle v_{i,+x} \rangle_t - \langle v_{i,-x} \rangle_t$ and so on. In words: The magnitude of the component of the average drift velocity of fly number **i** in ***x***-direction is given by the difference of the average velocity components in **+*x*** and **–*x*** direction.

This induces us to look now at the *ensemble*, the swarm of flies. What can we learn about the averages taken for the *ensemble* from the known averages of *individual* flies?

- As long as every fly does - on average - the same thing, the *vector* average over time of the ensemble is identical to that of an individual fly - if we sum up a few thousand vectors for *one* fly, or a few million for *lots* of flies does not make any difference. However, we also may obtain this average in a different way:
- We do not average *one fly in time* obtaining $\langle \underline{v}_i \rangle_t$, but at any given time *all flies in space*.
- This means, we just add up the velocity vectors of all flies at some moment in time and obtain $\langle \underline{v}_e \rangle_r$, the **ensemble average**. It is evident (but not easy to prove for general cases) that

$$\langle \underline{v}_i \rangle_t = \langle \underline{v}_e \rangle_r$$

- i.e. *time average = ensemble average*. The new subscripts "**e**" and "**r**" denote ensemble and space, respectively. This is a simple version of a very far reaching concept in stochastic physics known under the catch word "**ergodic hypothesis**".

This means that in "normal" cases, it doesn't matter how averages are taken. This is the reason why text books are often a bit unspecific at this point: It is intuitively clear what a drift velocity is and we don't have to worry about how it is obtained. It also allows us to drop all indices from now on whenever they are not really needed.

- In our fly swarm example, the drift velocity $\langle v_D \rangle = \langle \underline{v}_i \rangle$is usually much smaller than the average $\langle v_i \rangle$of the velocity magnitudes of an individual fly.

- The magnitude of $\langle \mathbf{v_D} \rangle$ is the difference of two large numbers - the average velocity of the *individual* flies in the drift direction *minus* the average velocity of the *individual* flies in the direction opposite to the drift direction.
- This induces an *asymmetry*: From a knowledge of the drift velocity *only*, *no inference whatsoever* can be made with regard to $\langle \mathbf{v_{i,+x}} \rangle$, $\langle \mathbf{v_{i,-x}} \rangle$ or $\langle \mathbf{v_i} \rangle$ whereas knowlegde of $\langle \mathbf{v_{i,+x}} \rangle$ and $\langle \mathbf{v_{i,-x}} \rangle$ tells us all there is to know in ***x***-direction

This teaches us a few things:

- **1.** Don't confuse $\langle \underline{\mathbf{v}} \rangle$ with $\langle \mathbf{v} \rangle$. The first quantity - for our flies - is zero or small, whereas the second quantity is large; they are totally different "animals".
- **2.** This means in other words: Don't confuse the property of the *ensemble* - the drift velocity $\mathbf{v_D}$ of the ensemble or swarm - with the properties of the *individuals* making up the ensemble.

2. Conductors

2.1 Definitions and General Properties

2.1.1 Metals

- *A few words before you start:*
  - Conductors *in general* are a bit boring, whereas conductors *in particular applications* are often hot topics (take the recent switch from **Al** to **Cu** in chip technology, for example).
  - There is a large number of highly optimized materials which are used as conductors nowadays. Just enumerating them is tiresome and not very interesting. Still, some knowledge of this issue is a must for materials scientists in the context of electronic materials.
  - As far as "theory" goes, there is either not much that goes beyond a basic knowledge of solid state physics (which, it is assumed, you already have), or very involved special theory (e.g. for superconductors or conducting polymers) - for which there is no time.
- In conclusion, we will only touch the issue, trying to present all major facets of the topic. In particular, the long list of applications for conductors (much longer than you probably would imagine) will be covered. This chapter, however, will be brief and mostly lists topics and key words.

The Basics

- The essential parameters of interest for conductors are:
- **1. Specific resistivity** ρ or **specific conductivity** $\sigma = 1/\rho$.
  - The defining "master" equation is

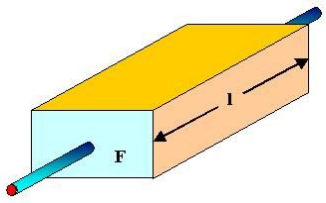
$$\sigma = |q| \cdot n \cdot \mu$$

  - With q = magnitude of the *charge* of the current carrying particles; n = *concentration* of current carrying particles (usually electrons in conductors) ; μ = **mobility** of the current carrying particles.
  - The units are

$$[\rho] = \Omega m$$

$$[\sigma] = (\Omega m)^{-1} = S/m$$

  - Note that ***S*** = "**Siemens**" = $1/\Omega$ = **A/V** is a bit old fashioned, but still in use. Note, too, that while the **SI** standard units call for the *meter* (**m**), you will find many values given in Ω***cm***.
  - A homogeneous material with a constant cross-sectional area ***F*** and a length ***l*** thus has a resistance of $R = (\rho \cdot l)/F$



$$R = \frac{\rho \cdot l}{F}$$

  - Or, in other words, a cube with **1 cm** length has a resistance ***R*** given in Ω that is numerically equal to ist specific resistance ρ given in Ω**cm**.
- If *electrons* are carrying the current, we have $q = -e$ = elementary charge $= 1.602 \cdot 10^{-19}$ **C**.
  - For units, conversions, and so on consult the link!

2. Ohm's law. Ohm's law (which was not a "law", but an empirical observation) formulated for the specific quantities writes

$$\underline{j} = \sigma \cdot \underline{E}$$

- With $\underline{j}$ = current density (a *vector*); $\underline{E}$ = electrical field strength (a vector); σ = specific conductivity, in general a *tensor* of **2nd** rank and, most important, *not* a function of the field strength $\underline{E}$ if not specifically noted. In other words, if the specific conductivity of a material is a constant, i.e. a fixed number with respect to $\underline{E}$, the material obeys Ohm's law.
- Ohm's law thus means that the $\underline{E}$ - $\underline{j}$ characteristics or the easily measured voltage - current characteristics are always *straight lines* through the origin! Within reasonable values of $\underline{E}$, or U, of course.

If you have *any* problem with these equations, perhaps because you feel Ohm's law should read $R = U/I$, or if you are not sure about the the meaning of the basic quantities, as e.g., *mobility*, *you have a problem*. Turn to the required reading module and other modules accessible from there.

- More about Ohm's law and the failure of classical physics in explaining the conductivity of metals can be found in a second required reading module. Add to this the required reading module for averaging vector quantities and you are ready for this chapter and others to come.

A remark to the ***mathematical notation***: HTML possibilities are limited and it is difficult to adhere to all rules of notation. In case of doubt, clarity and easy reading will have preference to formal correctness. This means:

- Whenever sensible, *cursive* symbols will be used for variables. It is *not* sensible, e.g., to use cursive letters for the velocity **v**, because the cursive v is easily mixed up with the Greek nu ν.
- All equations and the quantities used in equations are always **bold** - this greatly improves readability. However, it leaves little room for symbolizing vectors by **bold** lettering, and since underlining is cumbersome and not particularly helpful, we simply will mostly not use special notation for vectors. If you are able to understand this lecture course at all, you will know the vector (or tensor) quantities anyway.
- There are not enough letters in the alphabet to give every physical quantity an unambiguous symbol. One and the same symbol thus traditionally has several meanings, usually quite clear from the context. Occasionally, however, the danger of mix-up occurs. An example in case is the traditional use of the letter E for electrical field strength and for energies (and for Young's modulus in German). While in conventional texts one must give a different letter to these quantities, we will use the advantage of HTML and use *color coding* whenever the possibility of negative symbol interference raises its ugly head.

The *density* and *mobility* of *mobile charged* carriers thus determines the *conductivity*.

- The *carrier density* is a function of bonding (metallic, covalent in semiconductor, etc.), defects (doping in semiconductors) and temperature in general. In metals, however, n_e is nearly constant.
- The *mobility* is a function of collisions between carriers (e.g. electrons and holes) and/or between carriers and obstacles (e.g. **phonons** and crystal lattice defects).

Carrier concentration and mobility are, in general, hard to calculate from first principles. In semiconductors, the carrier density is easy to obtain, mobility is somewhat harder. In metals, the carrier density is rather fixed, but mobility is quite difficult to calculate, especially for "real" i.e. rather imperfect crystals. There are however, empirical rules or "laws".

- *Ohm's "law"* asserting that σ is *not* a function of $\underline{E}$ but only of some material property that can be expressed as a number.
- **Matthiesen's rule**, stating that

$$\rho = \rho_{Lattice}(T) + \rho_{defect}(N)$$

- With N = some measure of defect density.
- A "*rule of thumb*": ρ is proportional to T for $T >$ some T_{crit}

$$\Delta\rho = \alpha_\rho \cdot \rho \cdot \Delta T \approx \frac{0{,}4\%}{^oC}$$

- With **Temperature coefficient** $\alpha_\rho = 1/\rho \cdot d\rho / dT$.
- Then we have the **Wiedemann-Franz** "law", linking electrical conductivity to thermal conductivity, and so on.

The links give some graphs and numbers for representative metals.

- **Table of some metal properties**

- **ρ(*T*) for different defect densities in Na**
- **ρ(*T*) for different metals**

Some Values and Comments

The range of resistivity values (at room temperature) for metals is rather limited; here are some values as well as a first and last reminder that σ and ρ, while closely related, are quite different parameters with a numerical value that depends on the choice of the units! Do not mix up **cm** and **m**!

Metal	**Ag**	**Cu**	**Au**	**Al**	**Na**	**Zn**	**Ni**	**Fe**	**Sn**	**Pb**	**Hg**
ρ [μΩcm]	**1,6**	**1,7**	**2,2**	**2,7**	**4,2**	**5,9**	**6,8**	**9,7**	**12**	**21**	**97**
σ = 1/ρ [$10^6 \cdot \Omega^{-1}cm^{-1}$]	**0.625**	**0.588**	**0,455**	**0.37**	**0.238**	**0.169**	**0.147**	**0.103**	**0.083**	**0.046**	**0.01**
σ = 1/ρ [$10^6 \cdot \Omega^{-1}m^{-1}$]	**62,5**	**58.8**	**45.5**	**37**	**23.8**	**16.9**	**14.7**	**10.3**	**8.3**	**4.6**	**1**

- The temperature dependence, expressed e.g. in **ρ(300K)/ρ(100K)** may be a factor of **5 ...10**, so it is *not* a small factor. It may be used and is used, for measuring temperatures, e.g. with well-known **Pt** resistivity thermometers.
- This is something you should be aware of; cf. the anecdote in the link.

The specific resistivity, however, is not the only property that counts. In selecting a metal, important design parameters might also be:

- *Weight*, *mechanical strength*, *corrosion resistance*, *prize*, *compatibility* with other materials, ...

Sometimes it is advisable to look at "**figures of merit**", i.e. the numerical value coming out of a self-made formula that contains your important criteria in a suitable way.

- One very simple example: Lets say, *weight* is important. Define a *figure of merit* = ***F* = ρ*l* *d***, with ***d* =** density. The bigger ***F***, the better.
- You now get the following ranking (normalized to $F_{Na} = 1$):

Metal	**Na**	**K**	**Ca**	**Al**	**Mg**	**Cu**	**Ag**
F	**1**	**0,77**	**0,69**	**0,56**	**0,52**	**0,28**	**0,25**

The winner sodium! So you are going to use *Sodium* - **Na** for wiring?

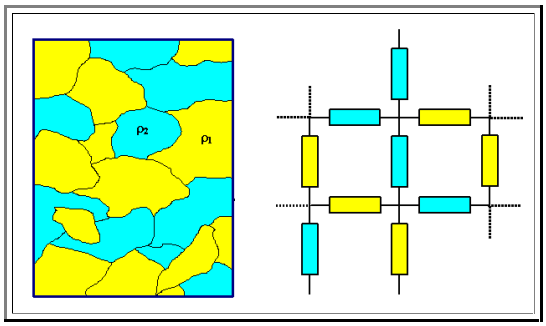
- Certainly not. Because now you will either include chemical stability ***C*** in your figure of merit (just multiply with ***C*** and assign values ***C* = 1** for great stability (e.g. **Au**, **Al**,), ***C* = 2** for medium stability (**Cu**, **Mg**) and ***C* = 5** for unstable stuff (**Na**, **K**, **Ca**). Or any other number reflecting the importance you put on this parameter. There is no ready made recipe - if precise numbers are not existing, you take your personal grading scale.
- And while you are at it, divide by some price index ***P***. Use the price per **kg**, or just a rough grade scale.
- In this simple example, you will get a surprising result: No matter what you do, the winner will be **Al**. It is the (base) material of choice for heavy duty applications when weight matters. In not so simple questions, you really may benefit from using the figure of merit approach.

Questionaire

Multiple Choice questions to 2.1.1

2.1.2 Alloys

- Pure metals are rarely used - in the real world you use **alloys**.
  - In principle, the specific resistivity ρ of an alloy can be obtained from the phase diagram and the ρ - values of the phases involved. Lets look at the extremes:
- **1.** Complete ***immiscibility***, e.g. in the case of **Au/Si**, or **Cu/W**. We may treat the resulting mix of metal particles as a *network of resistors* being linked in series and parallel. The volume fractions of the phases would constitute the weights - the treatment is not unlike the elastic modulus of compounds.



  - But no matter what kind of volume fraction you use and how you treat the resistor network - the resulting resistivity will *never* be smaller than that of the ingredient with the smallest resistivity.
- **2.** Complete *miscibility* (e.g. **Au/Ag**, **Cu/Ni**). Experimentally we find for small amounts (some %) of **B** in **A** (with **[B]** = concentration of **B**)

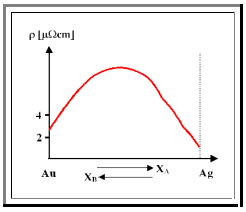
$$\rho \approx \rho_A + \text{const.} \cdot [B]$$

  - This formula is a special case of **Nordheims rule** which states .

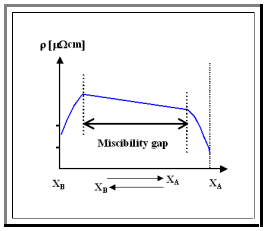
$$\rho \approx X_A \cdot \rho_A + X_B \cdot \rho_B + \text{const.} \cdot X_A \cdot X_B$$

  - This is pretty much an empirical law, it does not pay to justify it theoretically. Again, it is not possible to produce an alloy with a *resistivity smaller than one of its components*.
- If you have *intermetallic compounds* in your phase diagram, use Nordheim's rule with the intermetallic phases as X_A and X_B.
  - This leaves open the possibility that some intermetallic phase, i.e. a defined compound with its own crystal lattice, might have a lower resistivity than its constituents. While this is unlikely (if not outright impossible?) on theoretical grounds, no such intermetallics have been found so far.
- The sad fact then is that unleashing the full power of metallurgy and chemistry on mixing conductors (i.e. metals), will not give you a conductor with a specific conductivity better than **Ag**.
  - You will have to turn to **superconductors** (forgetting about cost considerations), if you can't live with **Ag**.
- Lets look at some examples:

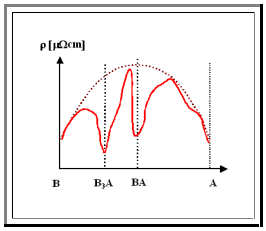
Complete miscibility
(e.g. **Au/Ag**)



Miscibility gap
(e.g. **Ag/Cu**); phase diagram in the link.



Intermetallic phases
(e.g. **Cu/Au**)



What do we learn? Something simple:

- Again: The resistivity *always* goes up in alloys or mixed materials as compared to the pure materials.
- Things are generally complicated, but *full of potential* for custom tailored properties!

Questionaire

Multiple Choice questions to 2.1.2

2.1.3 Non-Metalic Conductors

- We will just give a brief look at some especially important or useful **non-metallic conductors**:

Conducting Polymers

- That polymers, usually associated with insulators, can be very good conductors was a quite unexpected discovery some **20** years ago (Noble prize **2001**). They always need some "**doping**" with ionic components, however.
  - The resistivity can be ***exceedingly*** low. e.g. for *Iodine (I) doped poly-acethylene (**pAc**)* we may have. $\rho \leq$ **6,7 μΩcm**.
  - Or in other words: If you divide by the density for some figure of merit, it beats everything else, since **{ρ/density} (pAc) > {ρ/density} (Na)**!
  - More typical, however, are resistivities around (**10 .... 1000) μΩcm**.
- The conduction mechanism is along **–C=C–C=C–C=** chains, it is not yet totally clear. In fact, the first question is why this kind of chain is *not* generally highly conducting. Use the link for the answer.
  - The conductivity is strongly dependent on "doping" (in the **%** range!) with ions, and on many other parameters, the link gives an example.
  - So do not confuse this with the doping of semiconductors, where we typically add far less than **1 %** of a dopant!
- A new object of hot contemporary research are now ***semi*conducting polymers** which have been discovered about **10** years ago.

Transparent conductors

- **Indium Tin Oxide (*ITO*)** (including some variations) is the only really usable *transparent* conductor with reasonable conductivity (around **1** Ω**cm**)! It consists of **SnO_2** doped with **In_2O_3**.
  - **ITO** is technically *very important*, especially for:
    - flat panel displays, e.g. ***LCD***s .
    - solar cells.
    - research (e.g. for the electrical measurements of light-induced phenomena).
  - **ITO** is one example of conducting oxides, others are **TiO**, **NiO**, or **ZnO**. The field is growing rapidly and known as "***TCO***" = Transparent Conducting Oxides
- If you can find a transparent conductor much better than **ITO** (which leaves a lot to be desired), you may not get the Nobel prize, but you will become a rich person rather quickly.
  - Since **In** is rare, and the demand is exploding since the advent of **LCD**s, you also would be a rich person of you invested in **In** some years ago.

Ionic conductors

- Solid *Ionic conductors* are the materials behind " **Ionics**", including key technologies and products like
  - Primary batteries.
  - Rechargeable (secondary) batteries.
  - Fuel cells.
  - Sensors.
  - High temperature processes, especially smelting, refining, reduction of raw materials (e.g. **Al**-production).
  - There is an extra module devoted to the **Li** ion battery. This is important for you if you are interested in driving an affordable a car in **20** years or so.
  - They are also on occasion the unwanted materials causing problems, e.g. in corrosion or in the degradation of dielectrics.
  - See Chapter 2.4 for some details about ionic conductors.

Specialities : Intermetallics, Silicides, Nitrides etc.

Silicides, i.e. metal - silicon compounds, are important for microelectronics (***ME***) technology, but also in some more mundane applications, e.g. in heating elements. Some resistivity examples for silicides:

Silicide	$MoSi_2$	$TaSi_2$	$TiSi_2$	$CoSi_2$	$NiSi_2$	PtSi	Pd_2Si
ρ (μΩcm)	40 ...100	38...50	13..16	10...18	≈ 50	28...35	30...35

- It looks like the winner is **$CoSi_2$**. Yes, but it is difficult to handle and was only introduced more recently, like **$NiSi_2$**. In the earlier days (and at present) the other silicides given above were (and still are) used.

Some more examples of **special conductors** which find uses out there:

Material	HfN	TiN	TiC	TiB_2	C (Graphite)
ρ (μΩcm)	30...100	40...150	ca. 100	6 ...10	1000

Superconductors

Superconductors are in a class of their own. All kinds of materials may become superconducting at low temperatures, and there are neither general rules telling you *if* a material will become superconducting, nor at which *temperature*.

- There will be an advanced module some time in the future.

Why do we need those "exotic" materials?. There are two general reasons:

- **1.** Because, if just *one* specific requirement exists for your application that is not met by common materials, you simply have *no choice*. For example, if you need a conductor usable at **3000 K** - you take *graphite*. No other choice. It's as simple as that.
- **2.** Because *many* requirements must be met *simultaneously*. Consider e.g. **Al** for integrated circuits - there are plenty of important requirements; **see the link**. Since *no* material meets *all* of many requirements, an optimization process for finding an optimum material is needed.
- **Al** won the race for chip metallization for many years, but now is crowded out by **Cu**, because in some figure of merit the importance of low resistivity in the list of requirements is much larger now than it was in the past. It essentially overwhelms *almost* all other concerns (if there would not be an almost, we would have **Ag**!).

Questionaire

Multiple Choice questions to 2.1.3

2.1.4 Summary to: Conductors - Definitions and General Properties

What counts are the *specific* quantities:

- Conductivity σ (or the specific resistivity ρ = **1/** σ.
- current density $\underline{\boldsymbol{j}}$.
- (Electrical) field strength · $\underline{\boldsymbol{E}}$.

[ρ] = Ωm
[σ] = (Ωm)$^{-1}$ = S/m;
S = "Siemens"

The basic equation for σ is:
n **=** concentration of carriers,
μ = mobility of carriers.

$$\sigma = |q| \cdot n \cdot \mu$$

Ohm's law states:
It is valid for metals, but not for all materials.

$$\underline{\boldsymbol{j}} = \sigma \cdot \underline{\boldsymbol{E}}$$

σ (of conductors / metals) obeys (more or less) several rules; all understandable by looking at ***n*** and particularly ***μ***.

Matthiesen rule:
Reason: Scattering of electrons at defects (including phonons) decreases ***μ***.

$$\rho = \rho_{Lattice}(T) + \rho_{defect}(N)$$

"ρ**(*T*)** rule":
about **0,04 %** increase in resistivity per **K**
Reason: Scattering of electrons at phonons decreases ***μ***.

$$\Delta\rho = \alpha_\rho \cdot \rho \cdot \Delta T \approx \frac{0{,}4\%}{^oC}$$

Nordheim's rule:
Reason: Scattering of electrons at **B** atoms decreases ***μ***.

$$\rho \approx \rho_A + \text{const.} \cdot [B]$$

Major consequence: You can't beat the conductivity of pure **Ag** by "tricks" like alloying or by using other materials
(Not considering superconductors).

Non-metallic conductors are *extremely* important.

Transparent conductors (TCO's)
("ITO", typically oxides).

No flat panels displays = no notebooks etc. without **ITO**!

Ionic conductors (liquid and solid).

Batteries, fuel cells, sensors, ...

Conductors for high temperature applications; corrosive environments, ..
(Graphite, Silicides, Nitrides, ...).

Example: **$MoSi_2$** for heating elements in corrosive environments (dishwasher!).

Organic conductors (and semiconductors).

The future High-Tech key materials?

Numbers to know (order of magnitude accuracy sufficient)

ρ(decent metals) about **2 μΩcm**.
ρ(technical semiconductors)
around **1 Ωcm**.
ρ(insulators) > 1 GΩcm.

Questionaire

All Multiple Choice questions to 2.1

2.2. General Applications

2.2.1 Normal Conductors

A world without *conductors* is even harder to imagine than a world without *semiconductors*. Examples for applications include

- High-voltage free-air power transmission lines.
- High voltage wires for trains (getting "scratched" all the time).
- In-house wiring.
- Low-voltage wiring (car systems).
- High current wiring (machines).
- System on-board wiring.
- Bond wires for ***IC***'s (diameter **< 30µm**).
- Metallization on chips.
- Screening electrical or magnetic fields.
- Avoidance of electrostatic charging.
- Electrodes for batteries, chemical reactors etc.
- Antennas.

Each use has special requirements which should be met by the conducting material.

Some examples for requirements

- *Money* (Use of **Au**, **Ag**, **Pt** etc. may be critical).
- *Chemistry* (general stability and reactivity; esentiall excludes **Na**, **K**, **Hg** etc. for most applications; corrosion properties, ...).
- *Mechanical* properties (Pure metals are often too soft, but alloys have higher resistivity).
- *Thermal* properties (temperature coefficient; no metal usable beyond ca. **1000 K**) .
- *Compatibility with other materials* (contact corrosion, solderability, thermoelectric and thermomechanical properties, general chip compatibility, ...).
- *Compatibility with production technologies* (e.g. thin film deposition methods, wire making (try this with a brittle superconductor!),...).

Whole families of conductors, fine-tuned for a specific applications, were developed; below are some examples.

***Cu** based conductors*

There are many precisely specified **Cu**-based conductors for all kind of specific applications, examples are given in the link.

***Al** based conductors*

This family is primarily used for high-voltage free-air cables (in combination with a steel core) because of best fitting in terms of conductivity - price - mech. ***strength*** - corrosion requirements; cf. the illustration in the link.

Others

In *one* **IC** you may find the following conductor materials:

- Poly crystalline highly doped **Si**.
- Silicides; i.e. **Si** - metal compounds like **$NiSi_2$**.
- **Al** with $\leq$ **1%** of **Si** and **Cu** if the chip was made before, say, **2000**.
- **Cu** with some additions instead of **Al** if the chip was made after **2000**.
- **W**.
- **TiN**.

because one material simply does not meet the specific requirements for conductor on chips.

2.2.2 Contacts

- **Contacts**, meaning **mechanical contacts** here, are a major part of most electronic products. Even if there is no mechanical switch anymore, you still have the contact between the plug and the outlet, and/or the contact springs for the batteries.
- *Contacts* include the following items:
  - Switches, ***plugs***, relays, connections to removable parts (batteries, light bulbs, ...), ***pantographs*** (the thing on top of a locomotive), "brushes" (for motors), and so on.
  - Contacts are also the components or materials that often cause *trouble*. Contacts or switches are often the first components to break, and thus a nuisance to consumers like you and me.
- There are many specific requirements for **contact materials**:
  - Small contact resistance (it is never zero).
  - No sticking or welding under load.
  - No ***abrasion*** under load.
  - No intermixing of materials.
  - No wearing and tearing.
  - Suitable mechanical properties, e.g. good elasticity (forever) for switches.
- There are specific materials and group of materials generally favored for contacts:
  - **C** (graphite in many forms) for pantographs and whenever you want to draw a big current.
  - **Cu**, **Ag**, **Au**.
  - **Ru**, **Rh**, **Pd**, **Os**, **Ir**, **Pt**.
  - **Mo**, **W**.
  - ….
  - An example of **Ag**-based contact materials can be found in the link.
  - For contact applications we find *expensive* materials, because in many applications only small quantities are needed and the inertness of noble metals is what counts.

2.2.3 Resistors and Heating

Resistors

- Basic requirements for **resistors** (still one of the most numerous component in circuits) are:
  - Large region of ***R*** values (= device resistance in Ω) within *one* production technology.
  - Small (ideally vanishing) temperature coefficient .
  - Minimal noise.
  - Small dependence of ρ on production parameters (good repeatability).
  - No Ageing.
  - Small thermoelectrical coefficients to **Cu** (you want a resistor, not a thermoelement).
- Materials of choice include
  - **Ta**, **Ta** based alloys, and in particular "**Constantan**" (**55% Cu, 44% Ni, 1% Mn**), a resistor material with an especially small temperature coefficient α_ρ, but a large thermoelectric coefficient).
  - Strange mixtures of conductors and insulators including "**Cermet**" (short for Ceramics - Metals), e.g. **Cr - SiO_2**.
- Details and data in the *(future)* link.

Heating

- Basic requirements for **heating elements** are:
  - High melting point.
  - Chemical stability at high temperatures and in potentially corrosive environments.
  - Mechanical strength at high temperatures.
- The choice of a materials depends significantly on the range of temperatures envisioned. We have:
  - **FeNiCr**, **FeNiAl** alloys.
  - **Pt**, **W**, **Ta**, **Mo** - stable elements with a high melting point.
  - **$MoSi_2$** Among more industrial applications also used as heaters in dish washers - this is very aggressive environment!
  - Graphite (up to **3000 K** in non-oxidizing gas).
- Some details and data can be found in the links.
  - **Overview of resistivity and temperature range for some materials**
  - **Maximum temperatures for some materials**

2.2.4 Summary to: Conductors - General Applications

- No electrical engineering without conductors! Hundreds of specialized metal alloys exist just for "wires" because besides σ, other demands must be met, too:
  - Example for unexpected conductors being "best" compromise:

  Money, Chemistry (try Na!), Mechanical and Thermal properties, Compatibility with other materials, Compatibility with production technologies, ...

  Poly Si, Silicides, **TiN**, **W** in integrated circuits.

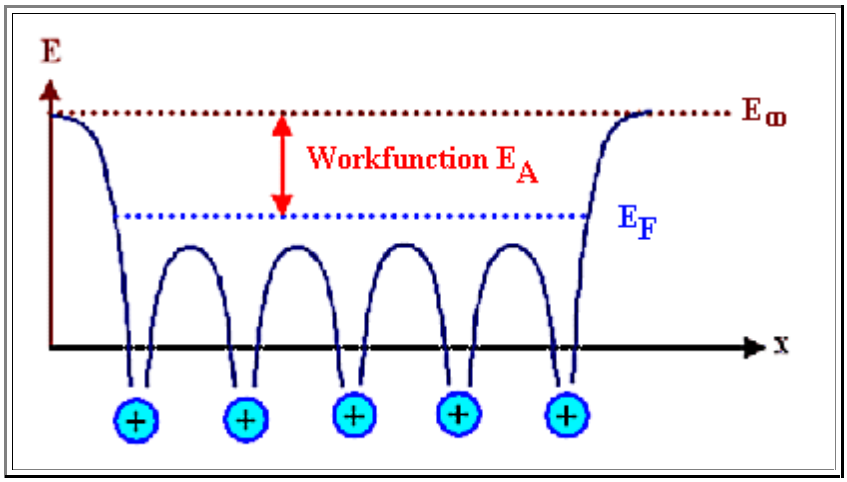
- Don't forget Special Applications:

  Contacts (switches, plugs, ...);
  Resistors;
  Heating elements; ...

2.3. Special Applications

2.3.1 Thermionic Emission

- **Cathodes** in **cathode ray tubes** (***CRT***), in regular **electron tubes** (still used for special applications), but also in all electron beam tools as e.g. *electron microscopes* or *electron beam welding*, are one example of *special conductors*. We need to have free electrons in the material *and* we need to extract them from the material.
  - For good cathodes we wish for some specific properties: First we want to extract lots of electrons *easily* and in *large quantities* (i.e. we want high current densities for little money).
  - Second, we want to extract them from a very *small area* (for high brightness), so that we can consider the electron beam to come from a *point source* which makes (electron) optics a lot less complicated to handle!
- Lets look at the free electron gas model and see how we can extract electrons in general.



  - For a metal, there are lots of electrons in the last band at all energies up to the Fermi energy, and at very low temperatures it takes at least the energy E_A to push an electron up the energy scale to E_∞, where it would be free to go wherever it wants - it is no longer "bound" to the crystal. We call that particular energy the ***work function*** of the material.
- The **work function** E_A of the material is thus the decisive quantity; it is the difference between the **Fermi energy** and the potential at infinity E_∞.

$$E_A = E_F - E_\infty$$

  - If we let $E_\infty = 0$ and have the energy scale going "down", we simply have .

$$E_A = E_F$$

- The current density for thermionic emission is given by the well-known **Richardson equation**, which we obtain by calculating how many electrons will have sufficient energy to overcome the energy barrier to the outside world from the energy distribution of electrons in a free electron gas model.
  - The necessary calculation is a meaningful mathematical exercise but we skip it to save time
- The Richardson equation for the current density j from a hot surface states writes:

$$j = A \cdot T^2 \cdot \exp - \frac{E_A}{kT}$$

  - From measuring $j = j(T)$ we expect (almost) **Arrhenius** behavior; E_A then follows from the slope of the plot, the constant A from its intersection with the j - axis.
  - If you are unsure about what this function looks like, use the function generator and play a bit.
  - The pre-exponential factor A can be calculated from the free electron gas model, but than it is only a crude approximation for real materials. Its free-electron-gas value is:
    $A_{theo} = 120\ A \cdot cm^{-2} \cdot K^{-2}$.

Lets compare that to some measured values (and bear in mind that ***A*** may depend on the Miller indices of the crystallographic plane from which we extract electrons, too - so numbers vary):

Material	**Fe**	**Ni**	**Pt**	**Ta**	**W**	**Cs**	**LaB_6**
A [$Acm^{-2}K^{-2}$]	26	30	32	55	60	162	25
E_A [eV]	4,5 - 4,8	5,15 - 5,35	5,65	4,15 - 4,8	4,2	1,8 - 2,14	2,6
T_m [^oC]	1 535	1 452	1 755	2 850	3 410	28,4	2 210

Cs has the lowest work function, but its melting point is so low that it is of no use. Optimizing everything, the winners are:

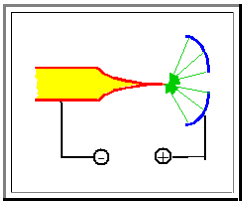
- **W**, the workhorse for cathode materials.
- **LaB_6**, a rather exotic material, because single crystals with very fine tips can be made that provide high current densities from a very small area. This is important whenever you want to focus the electron beam on a "point", e.g. in scanning electron microscopes. The focal point cannot be smaller than the area from where the electron beam was extracted from - *and you like it to be in the **nm** region*. The price one has to pay for this (besides for the **LaB_6** cathode, which is not cheap), is that the cathode has to be run in ultra high vacuum (***UHV***), because the fine tip would otherwise soon be damaged by ion bombardment resulting from ions extracted out of the residual gas atmosphere.

Questionaire

Multiple Choice questions to 2.3.1 (and 2.3.2)

2.3.2 Field Enhanced Emission and Tunnelling Effects

- If you run a **cathode**, emitting an electron beam, with *large* electrical fields between the cathode and the anode, you will find that your workfunction E_A seems to change to smaller values as the field strength increases.
  - This is called **Schottky effect**; it is observed at large field values of **$(10^5 - 10^8)$V/cm**.
- If you apply even higher field strengths (and remember: $E = U/d$; you do not need high voltages U, only small dimensions d), E_A seems to vanish altogether.

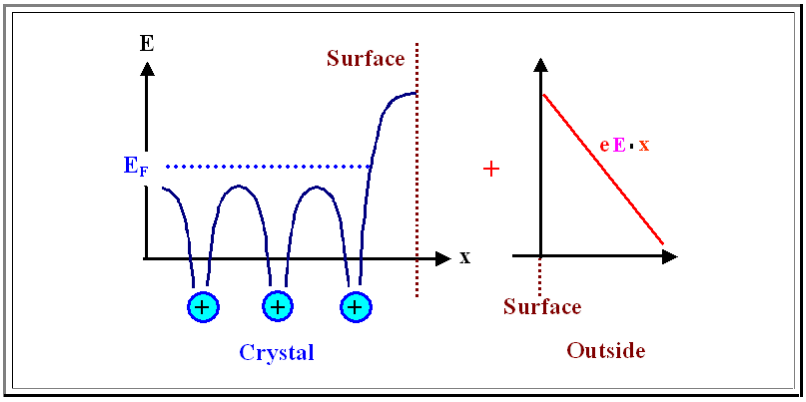


  - This effect is is called **field emission**. It works even at room temperature and is barely temperature dependent, so it can not be a temperature activated process.
  - Field emission is rather easy to obtain: all you have to do, is to make a very fine tip with a curvature of the tip in the **nm** - range as shown on the left.
  - Field emission might then occur with a few Volts between the anode and the tip, because all the field lines will have to converge on the small tip.

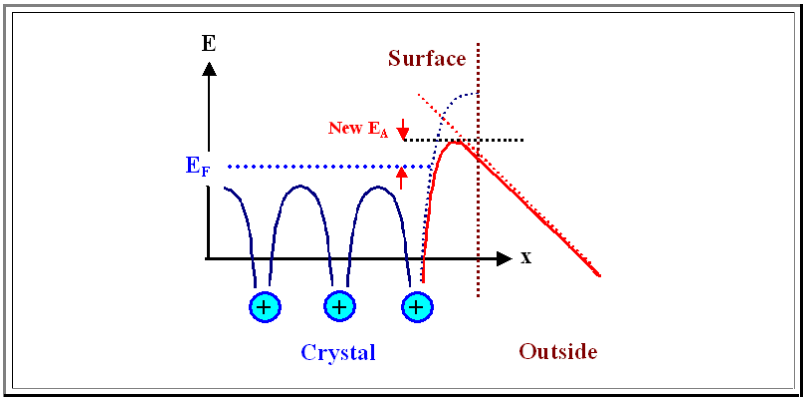
- How can we understand these effects? Whereas the *Schottky effect* is relatively straight forward, *field emission* is a manifestation of the *tunnelling effect*, a purely quantum mechanical phenomenon.
- Lets look at how the **free electron gas model** must be modified at high field strengths - and we will be able to account for *both* effects.
  - The potential energy E outside of the material is such that electrons are to be extracted - it is not constant, but varies with the field strength E simply as

$$E = e \cdot E \cdot x$$

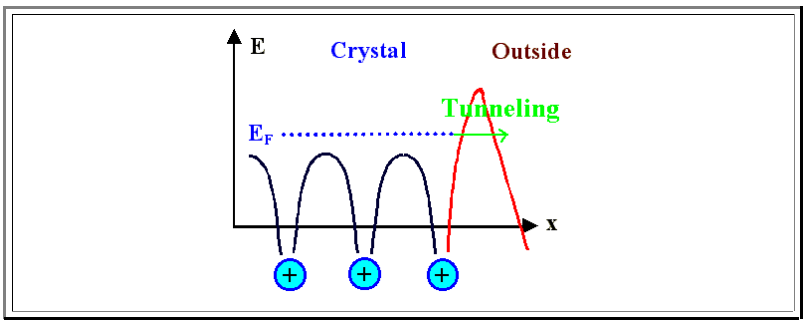
  - E, the (constant) applied field strength (written in mauve to make sure that we do not mix it up with the energy E). We have the following situation:



  - Simply summing up the energies graphically yields the qualitative energy curve for an electron at the edge of a crystal as shown below.



  - Whichever way you superimpose the potential energies, the potential barrier to the outside world will always be reduced. This explains qualitatively the *Schottky effect*.
- The *field emission effect* requires a somewhat different consideration.
  - Lets look at the *extremes* of the Schottky effect. For really high field strengths the potential barrier gets even lower and thinner, it may look somewhat like this:



- Now the **tunneling effect** may occur. It is a phenomenon inherent in quantum mechanics and allows electron "waves" to "*tunnel*" through a potential barrier.
  - In other words, the value of the **wave function** ψ for an electron does not got to zero abruptly at a potential barrier, but decays exponentially. There is then a finite amplitude for ψ *on the other side* of the potential barrier, an effect that is felt if the barrier is "thin" and low - as in the picture above. If the field strength is high enough, large quantities of electrons can directly tunnel to the outside world. More about tunnelling in the link.
- Field emission thus is a purely quantum mechanical effect; there is no classical counterpart whatsoever. It is used in a growing number of applications:
  - **Electron microscopes** for special purposes (e.g. scanning electron microscopes with high resolution at low beam voltage, a must for the chip industry) are usually equipped with **field emission "guns"**.
  - "**Scanning Tunnelling Microscopes**" (***STM***) which are used to view surfaces with atomic resolution, directly employ tunnelling effects.
  - Large efforts are being made to construct flat panel displays with millions of miniature field emission cathodes - at least one per pixel.
  - Some semiconductor devices (e.g. the "**tunnel diode**") depend on tunnelling effects through space charge regions.
- In other contexts, tunnelling is not useful, but may *limit* what you can do. Most notorious, perhaps, is the effect that *very thin* insulators - say **5 nm** and below - are insulating no more, a growing problem for the chip industry.

Questionaire

Multiple Choice questions to 2.3.1 *and* 2.3.2

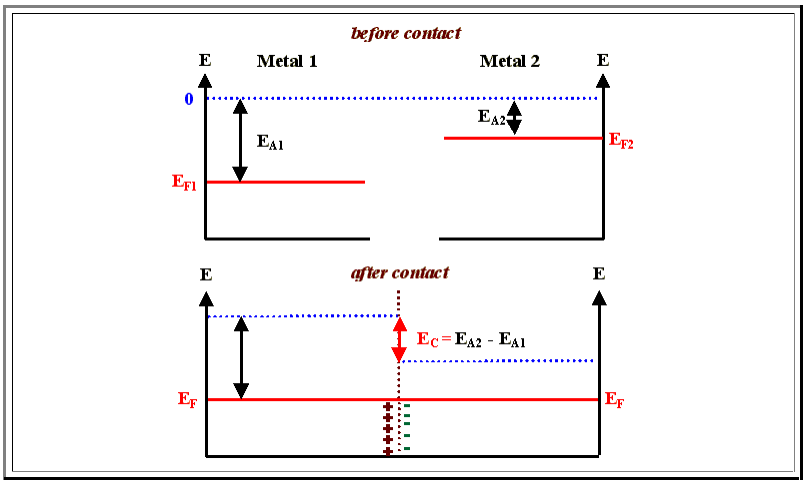
Note: :

A version translated into Ukrainian (by Fix Gerald) can be found in this link

2.3.3 Thermoelectric Effects

General Consideration

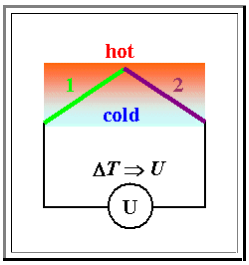
- So far we have only considered *one* conducting material; the unavoidable **contacts** between conductors, implicitly always required, were seemingly devoid of special properties.
  - We know that this *is not true* for many other contacts; e.g. combinations of
    - semiconductor - semiconductor.
    - semiconductor - conductor.
    - ionic conductor - conductor.
  - What about *metal - metal contacts*?
- We routinely solder wires of different conductors together or join them in any way, and do not worry about the contacts. Besides, maybe, a certain (usually small) **contact resistance** which is a property of the interface and must be added to the resistance of the two materials, there seems to be no other specific property of the contact.
  - But that is *only true* as long as *the temperature is constant* in the whole system of at least two conductors.
  - The reason for this is that we always get a **contact voltage**, as in the case of semiconductors, but the extension of the charged layers at the interface (the Debye lengths) is so short that no specific phenomena result from this.
  - Consider the band diagrams before and after joining two metals



- We have a dipole layer of charges at the interface which, owing to the large carrier density, is extremely thin and does not hinder current flow (it is easy for electrons to tunnel through the potential barrier).
  - We also have a *contact potential*, which is called the **Volta potential**. Since in any closed circuit (containing, e.g., the wires to the voltmeter), the *sum of the Volta potentials must be zero* in **thermal equilibrium**, it therefore can not be measured directly.
  - *If*, however, *one* of the at least two contacts needed for a closed circuit is at a temperature T_2 that is different from the temperature T_1 of the first contact, we have *non-equilibrium* and now a voltage may be measured. We observe the **Seebeck effect**, one of several **thermoelectric effects**.
- We will not go into details here (consult the link for this) but will only mention some applications and related effects.

Seebeck Effect

The Seebeck effect is the base for **thermoelements** or **thermocouples**, the standard device for measuring temperatures (the good old mercury thermometer is virtually nonexistent in technical applications, especially at high temperatures).



- Lets look at a typical situation: We have a thermocouple mader with a material **1** and a material **2**. It's "contacted" by whatever (material **3**, black lines). The junction of material**1** and material **2** is hot, the rest is cold (and has the same temperature).
- The voltmeter will show a **thermovoltage** that depends on ΔT and the two materials forming the thermocouple.

- Generally, the thermovoltage should be larger for couples of conductors with very different Fermi energies or carrier densities, since then the Volta potential is larger.
- Being more specific, the Volta potential should follow **Nernsts law**. But here we are only interested in the practical aspects of thermocouples.

For technically important materials, it is convenient to construct a voltage scale for thermocouples given in **mV/100K**.

- The voltage measured for a temperature difference of **100 K** is then the difference of the two values given on that scale for the two materials joined in a thermocouple. The zero point was arbitrarily chosen for **Pt**.

Bi	Ni	Pd	Pt	Hg	PtRh	Cu	Mo	Fe	NiCr	Sb
-7,7	-1,5	-0,3	0	0	0,7	0,77	1,2	1,92	2,6	4,8

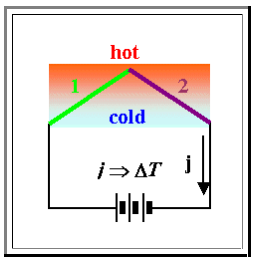
- Useful couples are, e.g. **Ni/NiCr**, with a thermovoltage of ca. **4 mV/100K** and a usable temperature range up to **1000 K**.

The Seebeck effect, for many years extensively used for measuring temperatures, can also be used to convert heat energy directly into electrical energy. **Thermoelectric generators** are becoming an exciting field of materials science, because optimized materials, based on a thorough understanding of the requirements for power generation and the concomitant requirements for the materials, are becoming available.

Other Thermoelectric Effects

There are several thermoelectrical effects which are deeply routed in **non-equilibrium thermodynamics**. Essentially, there is a "*reciprocating*" coupling of *gradients in driving forces* and *currents of any kind* (not just electrical currents but also, e.g. particle currents, heat currents, or even entropy currents).

- *Reciprocating* means, that if a *gradient* - e.g. in the *temperature* - induces an electric *current* across a junction (the Seebeck effect), than an electric current induced by *some other means* must produce a temperature gradient. And this does *not* address the heating simply due to ohmic heating!
- The "reversed" Seebeck effect does indeed exist, it is called the **Peltier effect**. In our schematic diagram it looks like this:



- An electrical current, in other words, that is driven through the system by a battery, would lead to a "heat" current, transporting thermal energy from one junction to the other one. One junction then goes down in temperature, the other one goes up.
- This effect would also occur in hypothetical materials with zero resistivities (we do not mean superconductors here). If there is some resistance R , the current will always lead to some heating of the wires everywhere which is superimposed on the Peltier effect.

- The temperature difference ΔT between the two junctions due to the external current density j induced by the battery and the Peltier effect then is approximately given by

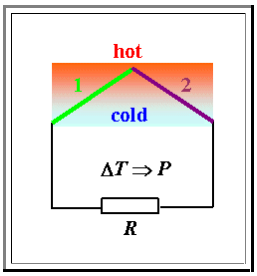
$$\Delta T \approx \text{const.} \cdot j$$

- The removal of heat or thermal energy thus is linear with the current density
- But there is always heating due to by ohmic losses, too. This is proportional to j^2, so it may easily overwhelm the Peltier effect and no net cooling is observed in this case.

The Peltier effect is not useful for *heating* - that is much easier done with resistance heating - but for *cooling*!

- With optimized materials, you can lower the temperature considerably at one junction by simply passing current through the device! The Peltier effect actually has been used for refrigerators, but now is mainly applied for controlling the temperature of specimens (e.g. chips) while measurements are being made.

One can do a third thing with thermoelements: Generate power. You have a voltage coupled to a temperaturr difference, and that can drive a current through a load in the form of a resistor.



- Invariably the question of the efficiency η of power generation comes up. How much of the thermal energy in the system is converted to electrical energy?
- This is not eays to calculate. It is, however, easy to guess to what η will be proportional:
  - η ∝ κ.
  - η ∝
  - η ∝ κ

There is one more effect worthwhile to mention: If you have an external current *and* an external temperature gradient at the same time, you have the **Thomson effect**. But we mention that only for completeness; so far the Thomson effect does not seem to be of technical importance. Again, more information is contained in the link.

Questionaire

Multiple Choice questions to 2.3.3

2.3.4 Summary to: Conductors - Special Applications

- Thermionic emission provides electron beams.
  The electron beam current (density) is given by the *Richardson equation*:
  - $A_{theo} = 120\ A \cdot cm^{-2} \cdot K^{-2}$ for free electron gas model
    $A_{exp} \approx (20 - 160)\ A \cdot cm^{-2} \cdot K^{-2}$
  - E_A = work function $\approx$ **(2 - >6) eV**
  - Materials of choice: **W**, **LaB_6** single crystal

$$j = A \cdot T^2 \cdot \exp -\frac{E_A}{kT}$$

- High field effects (tunneling, barrier lowering) allow large currents at low ***T*** from small (nm) size emitter

Needs **UHV**!

- There are several thermoelectric effects for metal junctions; always encountered in non-equilibrium.
  - *Seebeck effect:*
    Thermovoltage develops if a metal A-metal B junction is at a temperature different form the "rest", i.e. if there is a temperature gradeient

Essential for measuring (high) temperatures with a "thermoelement"
Future use for efficient conversion of heat to electricity ???

Questionaire

All Multiple Choice questions to 2.3

2.4 Ionic Conductors

2.4.1 General Remarks

- In **ionic conductors**, the current is transported by *ions* moving around (and possibly electrons and holes, too). Electrical current transport via **ions**, or *ions and electrons/holes*, is found in:
  - Conducting *liquids* called **electrolytes**.
  - **Ion conducting solids**, also called **solid electrolytes**.
- **Ionic conductivity** is important for many products:
  - Type **I** and type **II batteries** (i.e. regular and rechargeable).
  - **Fuel cells**.
  - **Electrochromic** windows and displays.
  - Solid state **sensors**, especially for reactive gases.
- In contrast to purely electronic current transport, there is *always* a chemical reaction tied to the *current flow* that takes place wherever the ionic current is converted to an electronic current - i.e. at the contacts or electrodes. There may be, however, a measurable potential difference *without* current flow in ionic systems, and therefore applications *not* involving chemical reactions.
  - This is a big difference to current flow with electrons (or holes), where no chemical reaction is needed for current flow across contacts since "chemical reactions " simply means that the system changes with time.
- If we look at the conductivity of solid ionic conductors, we look at the movement of ions in the crystal lattice - e.g. the movement (= diffusion) of O^{-} or H^{+} ions either as interstitials or as lattice ions.
  - In other words, we look at the diffusion of (ionized) atoms in some crstal lattice, described by a diffusion coefficient ***D***.
  - Since a diffusion coefficient ***D*** and a mobility **μ** describe essentially the same thing, it is small wonder that they are closely correlated - by the **Einstein-Smoluchowski relation** (the link leads you to the semiconductor Hyperscript with a derivation of the equation).

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

  - The conductivity of a solid-state ionic conductor thus becomes

$$\sigma = e \cdot c \cdot \mu = \frac{e^2 \cdot c \cdot D}{kT} = \frac{e^2 \cdot c \cdot D_0}{kT} \cdot \exp - \frac{H^m}{kT}$$

  - with the normal Arrhenius behaviour of the diffusion coefficient and H^m = migration enthalpy of an ion, carrying one elementary charge. In other words: we must expect complex and strongly temperature dependent behaviour; in particular if ***c*** is also a strong function of ***T***.

- **Ionics** is the topic of dedicated lecture courses, here we will only deal with two of the fundamental properties and equations - the **Debye length** and the **Nernst equation** - in a very simplified way.
  - The most general and most simple situation that we have to consider is a contact between two materials, at least one of which is a *solid ionic conductor* or solid electrolyte. Junctions with liquid electrolytes, while somewhat more complicated, essentially follow the same line of reasoning.
- Since this involves that some kind of ion can move, or, in other words, *diffuse* in the solid electrolyte, the *local concentration* ***c*** of the mobile ion can respond to two types of driving forces:
  - **1.** *Concentration gradients*, leading to particle currents j_{diff} (and, for particles with charge ***q***, automatically to an electrical current $j_{elect} = q \cdot j_{diff}$) given by **Ficks laws**

$$j_{diff} = -D \cdot \mathrm{grad}(c)$$

  - With ***D*** **= diffusion coefficient** of the diffusing particle.

- **2.** *Electrical fields* ***E***, inducing electrical current according to *Ohms law* (or whatever current - voltage - characteristics applies to the particular case), e.g.

$$j_{field} = \sigma \cdot E = q \cdot c \cdot \mu \cdot E$$

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

Both driving forces may be present *simultaneously*; the total current flow or voltage drop then results from the combined action of the two driving forces.

- Note that in one equation the current is proportional to the *gradient* of the concentration whereas in the other equation the proportionality is to the concentration *directly*. This has immediate and far reaching consequences for all cases where in equilibrium the two components must cancel each other as we will see in the next sub-chapter.

In general, the two partial currents will not be zero and some *net* current flow is observed. Under equilibrium conditions, however, there is no net current, this requires that the partial currents either are all zero, or that they must have the same magnitude (and opposite signs), so that they *cancel each other*.

- The equilibrium condition thus is

$$q \cdot j_{diff} = j_{field}$$

- The importance of this equation cannot be over emphasized. It imposes some general conditions on the *steady state concentration profile* of the diffusing ion and thus the charge density. Knowing the charge density distribution, the potential distribution can be obtained with the **Poisson** **equation**, and this leads to the *Debye length* and *Nernsts law* which we will discuss in the next paragraphs.

Questionaire

Multiple Choice questions to all of 2.4

2.4.2 Debye Length

Equilibrium of Diffusion and Field Currents

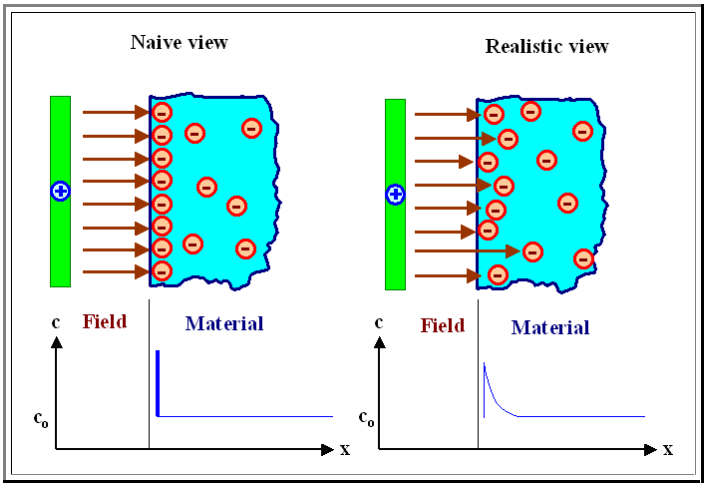
Nernst law is a special answer to the general and important question:

> How do *charged* and *mobile* particles redistribute themselves in an *electrical potential*
> if there are some restrictions to the obvious solution
> that they all move to one or the other pole of the field?

- It is the answer to this question that governs not only **pn**-junctions, but also batteries, fuel cells, or gas sensors, and, if you like, simply *all junctions*.

Let us consider a material that essentially contains mobile carriers of only *one* kind, i.e. a metal (electrons) , a (doped) semiconductor (electrons *or* holes, depending on doping), or a suitable ionic conductor (one kind of mobile ion).

- We imagine that we hold a positively charged plate at some (small) distance to the surface of a material having mobile negative charges (a metal, a suitable ionic conductor, a **n**-doped semiconductor, ...). In other words, the positively charged plate and the material are *insulated*, and no currents of any kind can flow between the two. However, there will be an electrical field, with field lines starting at the positive charges on the plate and ending on the negative charges inside the material. We have the following situation:



In a *naive* (and *wrong*) view, enough negatively charged carriers in the material would move to the surface to screen the field completely, i.e. prevent its penetration into the material. "Enough", to be more precise, means just the right number so that every field line originating from some charge in the positively charged plate ends on a negatively charged carrier inside the material.

- But that would mean that the concentration of carriers at the surface would be pretty much a δ- function, or at least a function with a very steep slope. That does not seem to be physically sensible. We certainly would expect that the concentration varies smoothly within a certain distance, and this distance we call **Debye length** right away.
- As you might know, the Debye length is a crucial material parameter not only in all questions concerning ionic conducitvity (the field of "**Ionics** "), but whenever the carrier concentration is not extremely large (i.e. comparable to the concenetration of atoms, i.e in metals).

We will now derive a simple formula for the *Debye length* . We start from the "naive" view given above and consider its ***ramifications***:

- If all (necessarily mobile) carriers would pile up at the interface, we would have a large concentration gradient and **Ficks law** would induce a very large *particle* current *away* from the interface, and, since the particles are charged, an *electrical* current at the same time! Since this **electrical *diffusion* current** $j_{el,\,Diff}$ is proportional to the concentration *gradient* **–grad (*c*(*x*))**, we have:

$$j_{el,\,Diff}(x) = -q \cdot D \cdot \text{grad}\,(c(x))$$

- With ***D*** = diffusion coefficient. Be clear about the fact that whenever you have a concentration gradient of mobile carriers, you will always have an electrical current by necessity. You may not notice that current because it might be cancelled by some other current, but it exists nevertheless.

The *electrical field* ***E*(*x*)**, that caused the concentration gradient in the first place, however, will also induce an electrical ***field*** **current** (also called **drift current**) $j_{field}(x)$, obeying Ohms law in the most simple case, which flows in the *opposite* direction of the electrical diffusion current. We have:

$$j_{field}(x) = q \cdot c \cdot \mu \cdot E(x)$$

- With μ = mobility, q = charge of the particle (usually a multiple of the elementary charge **e** of either sign); $q \cdot c \cdot \mu$, of course, is just the conductivity σ
- The *total* electrical current will then be the *sum* of the electrical field and diffusion current.

In *equilibrium*, both electrical currents obviously must be *identical* in magnitude and *opposite* in sign for every x, leading for one dimension to

$$q \cdot c(x) \cdot \mu \cdot E(x) = q \cdot D \cdot \frac{dc(x)}{dx}$$

Great, but too many unknowns. But, as we know (????), there is a relation between the diffusion coefficient D and the mobility **μ** that we can use; it is the **Einstein-Smoluchowski relation** (the link leads you to the semiconductor Hyperscript).

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

- We also can substitute the electrical Field $E(x)$ by $-dU(x)/dx$, with $U(x)$ = potential (or, if you like, voltage) across the system. After some reshuffling we obtain

$$-e\frac{dU(x)}{dx} = \frac{kT}{c(x)} \cdot \frac{dc(x)}{dx} = kT \cdot \frac{d\,[\ln c(x)]}{dx}$$

- We used the simple relation that $d(\ln c(x))/dx = 1/c(x) \cdot dc(x)/dx$. This little trick makes clear, why we always find relations between a voltage and the *logarithm* of a concentration.
- This is a kind of basic property of ionic devices. It results from the difference of the driving forces for the two opposing currents as noted before: The diffusion current is proportional to the *gradient* of the concentration whereas the field current is directly proportional to the concentration.

Integrating this simple differential equation once gives

$$U(x) + \frac{kT}{e} \cdot \ln c(x) = \text{const.}$$

- Quite interesting: the sum of two functions of x must be constant for any x and for any functions conceivable; the above sum is obviously a kind of *conserved quantity*.
- That's why we give it a name and call it the **electrochemical potential** V_{ec} (after muliplying with **e** so we have energy dimensions). While its two factors will be functions of the coordinates, its total value for any (x,y,z) coordinate in equilibrium is a *constant* (the three dimensional generalization is trivial). In other words we have

$$V_{ec} = V(x) + kT \cdot \ln c(x)$$

- with $V(x) = e \cdot U(x)$ = electrostatic potential energy.
- The electrochemical potential thus is a real energy like the potential energy or kinetic energy.

Obviously, *in equilibrium* (which means that nowhere in the material do we have a *net* current flow) the *electrochemical potential must have the same value anywhere in the material*.

- This reminds us of the **Fermi energy**. In fact, the electrochemical potential is nothing *but* the Fermi energy and the Fermi distribution in disguise.
- However, since we are considering *classical* particles here, we get the classical approximation to the Fermi distribution which is, of course, the **Boltzmann distribution** for E_F or V_{ec}, respectively, defining the zero point of the energy scale.

This is easy to see: Just rewriting the equation from above for $c(x)$ yields

$$c(x) = \exp - \frac{(Vx) - V_{ec}}{kT}$$

- What we have is the simple *Boltzmann distribution* for classical particles with the energy $(Vx) - V_{electrochem}$.

Calculating the Debye Length

- First we realize that the voltage or potential distribution (voltage times **e**) in the interior of a material *in equilibrium* can only be caused by *concentration distributions of carriers that obey equilibrium statistics*, i.e. the Boltzmann or the Fermi distribution.
  - This is simply what the equation above tells us.
- What we still need in order to calculate the Debye length is a linkage between potentials $e \cdot U(x) = V(x)$ and concentrations $c(x)$.
  - This is of course what the **Poisson equation**, the main equation for electrostatics, is all about. We will only look at the one-dimensional case here. The Poisson equation than states

$$-\frac{d^2 U}{dx^2} = \frac{dE}{dx} = \frac{e \cdot c(x)}{\epsilon\epsilon_0}$$

  - Now, for *good conductors* (i.e. **c(carriers) ≈ density of atoms ≈ 10^{22} cm^{-3}**), only a few of the carriers (a very small percentage) are needed to screen *any* reasonable electrical field. If you do not see this, do the exercise!

Exercise 2.4.2

Field Screening

- We may thus assume within a very good approximation that the carrier density at any point is given by the constant volume density c_0 of the field free material, *plus* a rather small space dependent addition $c_1(x)$; i.e.

$$c(x) = c_0 + c_1(x)$$

  - Obviously, only $c_1(x)$ is important for Poissons equation.
- From Boltzmanns distribution we know that

$$\frac{c(x)}{c_0} = 1 + \frac{c_1(x)}{c_0} = \exp\left(-\frac{\Delta(energy)}{kT}\right) = \exp\left(-\frac{V(x)}{kT}\right)$$

  - because the difference in energy of a carrier in the field free volume (i.e. where we have c_0) is simply the electrostatic energy associated with the electrical field.
  - Since we assumed $c_1 << c_0$, we may with ***impunity*** express the exponential function as a *Taylor series* of which we only retain the first term, obtaining:

$$1 + \frac{c_1(x)}{c_0} \approx 1 + \frac{V(x)}{kT}$$

$$c_1(x) = c_0 \cdot \frac{V(x)}{kT}$$

This is a simple trick, but important. Feeding the result back into Poissons equation yields:

$$\frac{d^2\,[c_1(x)]}{dx^2} = \frac{e^2 \cdot c_0 \cdot c_1(x)}{\epsilon \cdot \epsilon_0 \cdot kT}$$

For a simple one-dimensional case with a surface at $x = 0$ we obtain the final solution

$$c_1(x) = c_1(x = 0) \cdot \exp - \frac{x}{d}$$

The quantity d is the Debye length we were after, it is obviously given by

$$d = \textit{Debye length} = \left(\frac{\epsilon \cdot \epsilon_0 \cdot kT}{e^2 \cdot c_0} \right)^{1/2}$$

The *Debye length* is sometimes also called Debye-**Hückel** length (which is historically correct and just).

- $c_1(x = 0)$, of course, is given by the boundary condition, which for our simple case is:

$$c_1\,(x = 0) = c_0 \cdot \frac{V\,(x = 0)}{kT}$$

What is the meaning of the Debye length? Well, generalizing a bit, we look at the general case of a material having some *surplus charge* at a definite position somewhere in a material

- Consider for example the phase boundary of a (charged) precipitate, a charged grain boundary in some crystal, or simply a (point) charge somehow held at a fixed position *somewhere* in *some* material. The treatment would be quite similar to the one-dimensional case given here.

What we know now is quite important:

- If you are some Debye lengths away from these fixed charges, you will not "see" them anymore; their effect on the equilibrium carrier distribution then is vanishingly small.
- The Debye length resulting in *any* one of these situations thus is nothing but the *typical distance* needed for **screening** the surplus charge by the mobile carriers present in the material.
- In other words, after you moved about one Debye length away from the surplus charge, its effects on the mobile charges of the material are no longer felt.

More about the Debye length can be found in the Hyperscript "Semiconductors".

Questionaire

Multiple Choice questions to all of 2.4

2.4.3 Nernst's Equation

- **Nernst's equation** gives the *voltage* between two materials in close contact, i.e. the *potential difference* between the two materials. From the foregoing discussion, we know already two important facts about this potential:
  - It will change from one value to the other over a distance across the junction that is given by the (two) **Debye lengths** of the system.
  - The corresponding carrier concentrations are equilibrium concentrations and thus governed by the **Boltzmann distribution** (considering only classical particles at this point).
- If the potential difference is ΔU, we thus, using the Boltzmann distribution, obtain for the concentration of the carriers c_1 in material **1**, and c_2 in material **2**:

$$\frac{c_1}{c_2} = \exp -\frac{e \cdot \Delta U}{kT}$$

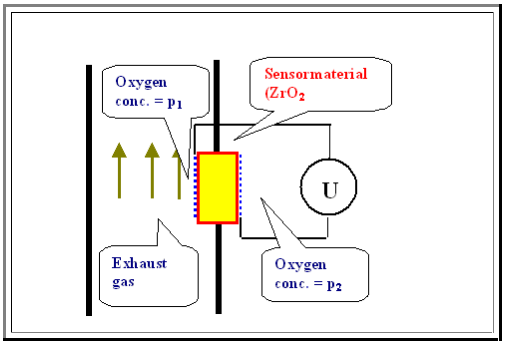
  - This is already *Nernst's equation* (or law) - *in a somewhat unusual way of writing*.
- Usually we (and everybody else) use the Boltzmann distribution to compute *concentrations* as a function of some other *known* parameters - the energy in this case. But this is *not* the only way for using a general equation!
  - Like any equation, it also works in *in reverse*: If we *know* the concentrations, we can calculate the energy difference that must go with them!
  - The important point now is that the concentrations of electrons in metals, but also of ions in ionic conductors, or holes in semiconductors, or any mobile carrier *a few Debye lengths away from the junction*, are fixed - there is no need to compute them!
  - What is *not fixed* is the *potential difference* $e \cdot \Delta U$ *a few Debye lengths away from the junction*, and that is what we now can obtain from the above equation by rewriting it for ΔU:

$$\Delta U = -\frac{kT}{e} \cdot \ln \frac{c_1}{c_2}$$

- This is **Nernst's equation** in its usual, but somewhat simplified form. We may briefly consider two complications:
  - **1.** If the particles carry z elementary charges, the first factor will now obviously write $kT/(z \cdot e)$.
  - **2.** If the *interaction* between particles is *not* negligible (which would mean, e.g., that Ficks law in its simple form would not be usable), the concentrations have to be replaced by the **activities** a of the particles.
  - If you want to know in detail what activities are - use the link. But all you have to know at this point is that activities are the particle concentrations *corrected for interaction effects*. To give an example: If a particle concentration is 10^{19} cm^{-3}, the activity might only be $5 \cdot 10^{18}$ cm^{-3}. If you use this factually wrong number, simple equations like the Boltzmann distribution that do not take into account particle interactions can still be used.
  - If the activity numbers are very different from the real concentration numbers, you are no longer doing Materials Science, but chemistry.
  - Using this, we obtain the *general* version of Nernst's law

$$\Delta U = -\frac{kT}{z \cdot e} \cdot \ln \frac{a_1}{a_2}$$

- Nernsts law, being the Boltzmann distribution in disguise, is of course extremely general. It gives the potential difference and thus the voltage of *any* contact between two materials that have sufficiently large concentrations of mobile carriers so that an equilibrium distribution can develop. It describes, among other things
  - The **contact voltage** (**Volta potential**) between two metals (i.e. **thermocouples**).
  - The built-in potential in **pn-junctions**
  - The voltage of any battery or accumulator.
  - The voltage of fuel cells.
  - The voltage produced by certain kinds of sensors.
- The last issue ***merits*** some brief explanation. Let's assume a material with a sufficiently large concentration of mobile O^{-} ions at interstitial sites (in other word, mobile interstitial point defects) at the working temperature - take Y_2O_3 stabilized ZrO_2 as an example (whatever that may be).
  - Use it to measure the amount of oxygen in a given gas mixture with the following **oxygen sensor** device:



- The sensor material experiences two different oxygen concentrations on its two surfaces, one of which is known (oxygen in air, a constant for all practical purposes), the other one is the concentration in the exhaust gas of a car which is supposed to be measured by the voltmeter
  - Two gas-permeable electrodes have been supplied which allow oxygen on both sides to react with the sensor material.
- In equilibrium, we will have some reaction between the oxygen in the gas and the oxygen in the crystal in the sense that oxygen will either come out, or diffuse into the material.
  - What we might expect is that the concentration of interstitial oxygen in the crystal will be larger near to the surface with the large oxygen gas concentration (air) compared to the surface exposed to a lower oxygen concentration (exhaust).
  - The gradient in the (negatively charged) oxygen concentration inside the material then will be determined by the **Debye length** of the system (in the real thing, which is ZrO_2, it will be just a few **nm**).
  - In total, the concentration $[O]_S$ of mobile **O**-interstitials right at the surface will be *somehow* tied to the partial pressure p_O of the oxygen on both sides; lets say we have a general relation like

$$[O]_S = \left(\text{const.} \cdot p_O \right)^n$$

  - But any other (reasonable) relation you can think of will be just as good.
- Nernst's law then tells us immediately, how the voltage between the two electrodes depends on the oxygen concentration or partial pressure in the exhaust: For the assumed relation we have

$$\Delta U = -\frac{kT}{e} \cdot \ln \frac{c_1}{c_2}$$

$$\Delta U = -\frac{kT}{e} \cdot \ln \frac{(p_1)^n}{(p_2)^n}$$

$$\Delta U = -\frac{n \cdot kT}{e} \cdot \ln \frac{p_1}{p_2}$$

- This is quite remarkable: We have an equation for the voltage that develops across *some* sensor as a function of the difference of the oxygen concentration on two sides of the sensor *without knowing much* about the details of the sensor! All we have to assume is that there is *some* mobile $\mathbf{O^{-}}$, *no* other free carriers, and that establishing equilibrium does not take forever.
- Only if you want to know the *precise* value of ***n*** do you have to delve into the detailed reactions at the interfaces.

This is essentially the working principle of not only the oxygen sensor in the exhaust system of any modern car ("**λ - Sonde**"), but of most, if not all, solid state sensors.

Questionaire
Multiple Choice questions to all of 2.4

2.4.4 Summary to: Ionic Conductors

- Electrical current can conducted by *ions* in
  - Liquid electrolytes (like H_2SO_4 in your "lead - acid" car battery); including gels
  - Solid electrolytes (= ion-conducting crystals). Mandatory for fuel cells and sensors
  - Ion beams. Used in (expensive) machinery for "nanoprocessing".

Challenge: Find / design a material with a "good" ion conductivity at room temperature

- Basic principle
  - *Diffusion current* j_{diff} driven by concentration gradients **grad(*c*)** of the charged particles (= ions here) equilibrates with the
  - *Field current* j_{field} caused by the internal field always associated to concentration gradients of charged particles plus the field coming from the outside
  - Diffusion coefficient ***D*** and mobility **μ** are linked via theEinstein relation;
    concentration ***c*(*x*)** and potential ***U*(*x*)** or field ***E*(*x*) = –d*U*/d*x*** by the Poisson equation.

$$j_{diff} = -D \cdot \mathrm{grad}(c)$$

$$j_{field} = \sigma \cdot E = q \cdot c \cdot \mu \cdot E$$

$$\mu = eD/kT$$

$$-\frac{d^2U}{dx^2} = \frac{dE}{dx} = \frac{e \cdot c(x)}{\epsilon\epsilon_0}$$

- Immediate results of the equations from above are:
  - In equilibrium we find a preserved quantity, i.e. a quantity independent of ***x*** - the electrochemical potential V_{ec}:
  - If you rewrite the equaiton for ***c*(*x*)**, it simply asserts that the particles are distributed on the energy scale according to the Boltzmann distrubution:

$$V_{ec} = \text{const.} = e \cdot U(x) + kT \cdot \ln c(x)$$

$$c(x) = \exp -\frac{(Vx) - V_{ec}}{kT}$$

  - Electrical field *gradients* and concentration *gradients* at "contacts" are coupled and non-zero on a length scale given by the **Debye length** d_{Debye}
  - The Debye length is an extremely important material parameter in "**ionics**" (akin to the space charge region width in semiconductors); it depends on temperature ***T*** and in particular on the (bulk) concentration c_0 of the (ionic) carriers.
  - The Debye length is not an important material parameter in metals since it is so small that it doesn't matter much.

$$d_{Debye} = \left(\frac{\epsilon \cdot \epsilon_0 \cdot kT}{e^2 \cdot c_0} \right)^{1/2}$$

The potential difference between two materials (her ionic conductors) in close contact thus...

- ... extends over a length given (approximately) by :

$$d_{Debye}(1) + d_{Debye}(2)$$

- ... is directly given by the Boltzmann distribution written for the energy:
(with the c_i =equilibrium conc. far away from the contact.

$$\frac{c_1}{c_2} = \exp -\frac{e \cdot \Delta U}{kT}$$ **Boltz-mann**

- The famous *Nernst equation*, fundamental to ionics, is thus just the Boltzmann distribution in disguise!

$$\Delta U = -\frac{kT}{e} \cdot \ln \frac{c_1}{c_2}$$ **Nernst's equation**

"Ionic" sensors (most famous the ZrO_2 - based O_2 sensor in your car exhaust system) produce a voltage according to the Nernst equation because the concentration of ions on the exposed side depends somehow on the concentration of the species to be measured.

Questionaire

Multiple Choice questions to all of 2.4

2.5 Summary: Conductors

What counts are the *specific* quantities:

- Conductivity σ (or the specific resistivity $\rho = 1/\sigma$
- current density $\vec{j}$
- (Electrical) field strength · $\vec{E}$

$$[\sigma] = (\Omega m)^{-1} = S/m;$$
$$S = 1/\Omega = \text{"Siemens"}$$
$$[\rho] = \Omega m$$

The basic equation for σ is:
n = concentration of carriers
μ = mobility of carriers

$$\sigma = |q| \cdot n \cdot \mu$$

Ohm's law states:
It is valid for metals, but not for all materials

$$\vec{j} = \sigma \cdot \vec{E}$$

σ (of conductors / metals) obeys (more or less) several rules; all understandable by looking at n and particularly μ.

Matthiesen rule
Reason: Scattering of electrons at defects (including phonons) decreases μ.

$$\rho = \rho_{Lattice}(T) + \rho_{defect}(N)$$

"$\rho(T)$ rule":
about **0,04 %** increase in resistivity per **K**
Reason: Scattering of electrons at phonons decreases μ

$$\Delta\rho = \alpha_\rho \cdot \rho \cdot \Delta T \approx 0{,}4\%\ /\ ^oC$$

Nordheim's rule:
Reason: Scattering of electrons at **B** atoms decreases μ

$$\rho \approx \rho_A + \text{const.} \cdot [B]$$

Major consequence: You can't beat the conductivity of pure **Ag** by "tricks" like alloying or by using other materials.
(Not considering superconductors).

Non-metallic conductors are *extremely* important.

Transparent conductors (TCO's)
("ITO", typically oxides)

No flat panels displays = no notebooks etc. without **ITO**!

Ionic conductors (liquid and solid)

Batteries, fuel cells, sensors, ...

Conductors for high temperature applications; corrosive environments, ..
(Graphite, Silicides, Nitrides, ...)

Example: **$MoSi_2$** for heating elements in corrosive environments (dishwasher!).

Organic conductors (and semiconductors)

The future High-Tech key materials?

Numbers to know (order of magnitude accuracy sufficient)

ρ(decent metals) about **2 μΩcm**.
ρ(technical semiconductors) around **1 Ωcm**.
ρ(insulators) > 1 GΩcm.

- No electrical engineering without conductors! Hundreds of specialized metal alloys exist just for "wires" because besides σ, other demands must be met, too:

  Money, Chemistry (try **Na**!), Mechanical and Thermal properties, Compatibility with other materials, Compatibility with production technologies, ...

  - Example for unexpected conductors being "best" compromise:

    Poly Si, Silicides, **TiN**, **W** in integrated circuits

- Don't forget Special Applications:

  Contacts (switches, plugs, ...);
  Resistors;
  Heating elements; ...

- Thermionic emission provides electron beams.
  The electron beam current (density) is given by the *Richardson equation*:

  $$j = A \cdot T^2 \cdot \exp - \frac{E_A}{kT}$$

  - $A_{theo} = 120\ A \cdot cm^{-2} \cdot K^{-2}$ for free electron gas model
    $A_{exp} \approx (20 - 160)\ A \cdot cm^{-2} \cdot K^{-2}$
  - E_A = work function $\approx$ **(2 - >6) eV**
  - Materials of choice: **W**, LaB_6 single crystal

- High field effects (tunneling, barrier lowering) allow large currents at low ***T*** from small (nm) size emitter

  Needs **UHV**!

- There are several thermoelectric effects for metal junctions; always encountered in non-equilibrium.

  - *Seebeck effect:*
    Thermovoltage develops if a metal A-metal B junction is at a temperature different form the "rest", i.e. if there is a temperature gradeient

    Essential for measuring (high) temperatures with a "thermoelement"
    Future use for efficient conversion of heat to electricity ???

  - *Peltier effect:*
    Electrical current ***I*** through a metal - metal (or metal - semiconductor) junction induces a temperature gradient $\propto I$, i.e. one of the junction may "cool down".

    Used for electrical cooling of (relatively small) devices. Only big effect if electrical heating ($\propto I^2$) is small.

- Electrical current can conducted by *ions* in

  - Liquid electrolytes (like H_2SO_4 in your "lead - acid" car battery); including gels
  - Solid electrolytes (= ion-conducting crystals). Mandatory for fuel cells and sensors
  - Ion beams. Used in (expensive) machinery for "nanoprocessing".

  Challenge: Find / design a material with a "good" ion conductivity at room temperature

- Basic principle

  - *Diffusion current* j_{diff} driven by concentration gradients **grad(*c*)** of the charged particles (= ions here) equilibrates with the
  - *Field current* j_{field} caused by the internal field always associated to concentration gradients of charged particles plus the field coming from the outside

  $$j_{diff} = -D \cdot \mathrm{grad}(c)$$

  $$j_{field} = \sigma \cdot E = q \cdot c \cdot \mu \cdot E$$

- Diffusion coefficient ***D*** and mobility **μ** are linked via theEinstein relation;
  concentration ***c*(*x*)** and potential ***U*(*x*)** or field ***E*(*x*) = –d*U*/d*x*** by the Poisson equation.

$$\mu = eD/kT$$

$$-\frac{d^2U}{dx^2} = \frac{dE}{dx} = \frac{e \cdot c(x)}{\epsilon\epsilon_0}$$

Immediate results of the equations from above are:

- In equilibrium we find a preserved quantity, i.e. a quantity independent of ***x*** - the electrochemical potential V_{ec}:

$$V_{ec} = \text{const.} = e \cdot U(x) + kT \cdot \ln c(x)$$

- If you rewrite the equaiton for ***c*(*x*)**, it simply asserts that the particles are distributed on the energy scale according to the Boltzmann distrubution:

$$c(x) = \exp -\frac{(Vx) - V_{ec}}{kT}$$

- Electrical field *gradients* and concentration *gradients* at "contacts" are coupled and non-zero on a length scale given by the **Debye length** d_{Debye}

$$d_{Debye} = \left(\frac{\epsilon \cdot \epsilon_0 \cdot kT}{e^2 \cdot c_0} \right)^{1/2}$$

- The Debye length is an extremely important material parameter in "**ionics**" (akin to the space charge region width in semiconductors); it depends on temperature ***T*** and in particular on the (bulk) concentration c_0 of the (ionic) carriers.
- The Debye length is not an important material parameter in metals since it is so small that it doesn't matter much.

The potential difference between two materials (her ionic conductors) in close contact thus...

- ... extends over a length given (approximately) by :

$$d_{Debye}(1) + d_{Debye}(2)$$

- ... is directly given by the Boltzmann distribution written for the energy:
  (with the c_i =equilibrium conc. far away from the contact.

$$\frac{c_1}{c_2} = \exp -\frac{e \cdot \Delta U}{kT} \quad \text{Boltz-mann}$$

- The famous *Nernst equation*, fundamental to ionics, is thus just the Boltzmann distribution in disguise!

$$\Delta U = -\frac{kT}{e} \cdot \ln \frac{c_1}{c_2} \quad \text{Nernst's equation}$$

"Ionic" sensors (most famous the ZrO_2 - based O_2 sensor in your car exhaust system) produce a voltage according to the Nernst equation because the concentration of ions on the exposed side depends somehow on the concentration of the species to be measured.

Questionaire

All multiple choice questions zu 2. Conductors

3. Dielectrics

3.1 Definitions and General Relations

3.1.1 Polarization and Dielectric Constant

3.1.2 Summary to: Polarization and Dielectric Constant

3.2 Mechanisms of Polarization

3.2.1 General Remarks

3.2.2 Electronic Polarization

3.2.3 Ionic Polarization

3.2.4 Orientation Polarization

3.2.5 Summary and Generalization

3.2.6 Local Field and Clausius - Mosotti Equation

3.2.7 Summary to: Polarization Mechanisms

3.3 Frequency Dependence of the Dielectric Constant

3.3.1 General Remarks

3.3.2 Dipole Relaxation

3.3.3 Resonance for Ionic and Atomic Polarization

3.3.4 Complete Frequency Dependence of a Model Material

3.3.5 Summary to: Frequency Dependence of the Dielectric Constant

3.4. Dynamic Properties

3.4.1 Dielectric Losses

3.4.2 Summary to: Dynamic Properties - Dielectric Losses

3.5 Electrical Breakdown and Failure

3.5.1 Observation of Electrical Breakdown and Failure

3.5.3 Summary to: Electrical Breakdown and Failure

3.6 Special Dielectrics

3.6.1 Piezo Electricity and Related Effects

3.6.2 Ferro Electricity

3.6.3 Summary to: Special Dielectrics

3.7 Dielectrics and Optics

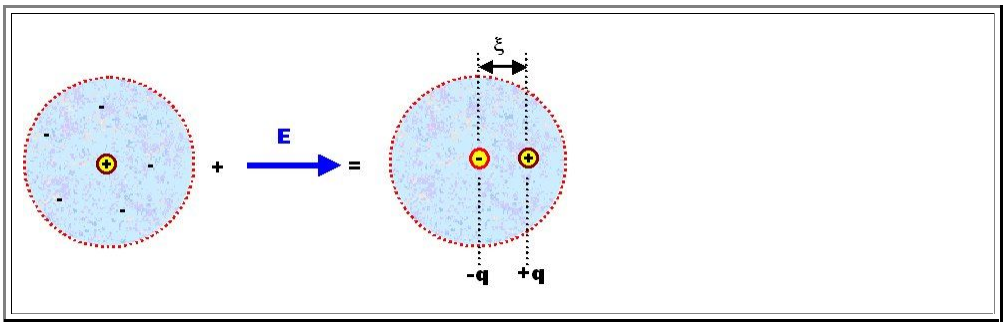
3.8 Summary: Dielectrics

3. Dielectrics

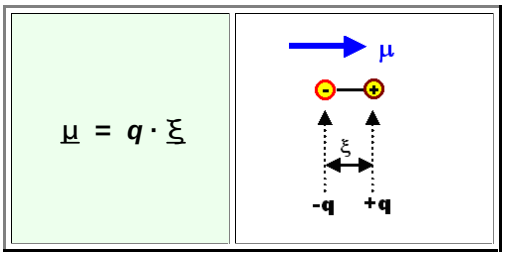
3.1 Definitions and General Relations

3.1.1 Polarization and Dielectric Constant

- For the sake of simplicity we assume that **dielectric materials** are *perfect* ***insulators***. In other words, there are no mobile charged particles.
- We want now to be able to answer *three questions*:
  - **1.** Given the atomic structure of the material - What is its **dielectric constant** (***DK***)?
  - **2.** How does the **DK** depend on the *frequency* of the external field?
  - **3.** How large is the *maximum field strength* a ***dielectric*** can take? Remember, no material can take arbitrarily large loads - mechanical, electical, whatever.
- For starters, we look at some general descriptions, definitions and general relations of the quantities necessary in this context.
- The dielectric constant of solids is an interesting material parameter only if the material is exposed to an electrical field (and this includes the electrical field of an **electromagnetic wave**). The effect of the *electrical field* (or just *field* for short from now on) can be twofold:
  - **1.** It *induces electrical dipoles* in the material and tries to align them in the field direction. In other words, *with* a field, dipoles come into being that do not exist *without* a field.
  - **2.** It tries to *align dipoles* that are already present in the material. In other words, the material *contains electric dipoles even without a field*.
- Of course we also may have a combination of both effects: The electrical field may change the distribution of existing dipoles while trying to align them, *and* it may generate new dipoles in addition.
- The *total* effect of an electrical field on a dielectric material is called the **polarization** of the material.
- To understand that better, lets look at the most simple object we have: A single *atom* (we do not even consider molecules at this point).
  - We have a positively charged nucleus and the electron "cloud". The smeared-out negative charge associated with the electron cloud can be averaged in space and time, and its charge center of gravity than will be at a point in space that coincides *exactly* with the location of the nucleus, because we must have spherical symmetry for atoms.
  - If we now apply an electrical field, the centers of charge will be separated. The electron cloud will be pulled in the direction of the positive pole of the field, the nucleus to the negative one. We may visualize that (ridiculously exaggerated) as follows:



  - The center of the positive and negative charges ***q* (= *z* · e)** are now separated by a distance ξ, and we thus induced a **dipole moment** μ which is defined by



  - It is important to understand that μ is a *vector* because ξ is a vector. The way we define it, its *tip will always point towards the positive charge*. For schematic drawings we simply draw a little arrow for μ.
- The magnitude of this *induced* ***dipole moment*** is a property of our particular atom, or, if we generalize somewhat, of the "particles" or building blocks of the material we are studying.

- In order to describe the ***bulk*** material - the sum of the particles - we *sum up all individual dipole moments* contained in the given volume of the material and divide this sum by the volume ***V***. This gives us the (volume independent) **polarization *P*** of the material. *Notice that we have a vector sum!*

$$\underline{P} = \frac{\Sigma\, \underline{\mu}}{V} = \langle \underline{\mu} \rangle \cdot N_V$$

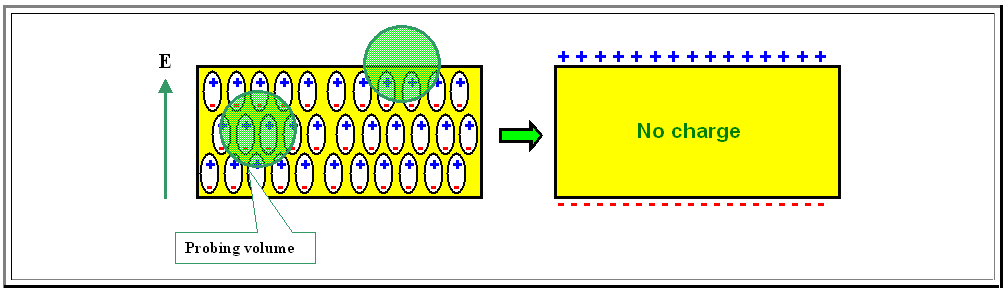
- With $\langle \underline{\mu} \rangle$ = average vector dipole moment; N_V = density of dipoles (per m^3).
- $\underline{P}$ thus points from the *negative* to the *positive* charge, too - a convention *opposite* to that used for the electrical field.
- The physical dimension of the polarization thus is C/m^2; (Coulomb per square meter). i.e. the polarization has the dimension of an *area charge*, and since $\underline{\mu}$ is a vector, $\underline{P}$ is a vector, too.

It is important to realize that a polarization $\underline{P} = 0$ does *not* mean that the material does *not* contain dipole moments, but only that the vector sum of all dipole moments is zero.

- This will always be the case if the dipole moment vectors are *randomly distributed* with respect to their directions. Look at the picture in one of the next subchapters if you have problems visualizing this. But it will also happen if there is an ordered distribution with pairs of opposing dipole moments; again contemplate a picture in one of the following subchapters if that statment is not directly obvious.

That $\underline{P}$ has the dimension of C/cm^2, i.e. that of an area charge, is *not* accidental but has an immediate interpretation.

- To see this, let us consider a simple plate capacitor or condenser with a *homogeneously polarized material* inside its plates. More generally, this describes an isotropic dielectric slab of material in a homogeneous electrical field. We have the following idealized situation:



- For sake of simplicity, all dipole moments have the same direction, but the subsequent reasoning will not change if there is only an average component of $\underline{P}$ in field direction. If we want to know the *charge density* ρ inside a small probing volume, it is clearly zero in the volume of the material (if averaged over probing volumes slightly larger than the atomic size), because there are just as many positive as negative charges.
- We are thus left with the *surfaces*, where there is indeed some charge as indicated in the schematic drawing. At one surface, the charges have effectively moved out a distance ξ , at the other surface they moved in by the same amount. We thus have a **surface charge**, called a *surface polarization charge*, that we can calculate.
- The number N_c of charges appearing at a surface with area ***A*** is equal to the number of dipoles contained in the surface "volume" $V_S = A \cdot \xi$ times the relevant charge **q** of the dipole. Using ξ to define the volume makes sure that we only have *one* layer of dipoles in the volume considered.
- Since we assume a homogeneous dipole distribution, we have the same polarization in any volume and thus $\underline{P} = \Sigma \underline{\mu} / V = \Sigma \underline{\mu}_S / V_S$ obtains. Therfore we can write

$$\underline{P} = \frac{\Sigma_V\, \underline{\mu}}{V} = \frac{\Sigma_S\, \underline{\mu}}{V_S} = \frac{\xi \cdot \Sigma_S\, q}{V_S} = \frac{\xi \cdot \Sigma_S\, q}{\xi \cdot A} = \frac{\Sigma_S\, q}{A}$$

$$\Sigma_S\, q = N_c = P \cdot A$$

- Σ_V or Σ_S denotes that the summation covers the total volume or the "surface" volume. Somewhere we "lost" the vector property of $\underline{P}$, but that only happens because we automatically recognize ξ as being perpendicular to the surface in question.
- *While this is a certain degree of **sloppiness**, it makes life much easier and we will start to drop the underlining of the vectors from now on whenever it is sufficiently clear what is meant.*

- The **area density** σ_{pol} of the charge on the surface is then

$$\sigma_{pol} = \frac{N_c}{A} = \frac{P \cdot A}{A} = |\underline{P}|$$

- Of course, σ_{pol} is a *scalar*, which we obtain if we consider $\underline{P} \cdot \underline{A}$ to be the scalar product of the vector $\underline{P}$ and the vector $\underline{A}$; the ***latter*** being perpendicular to the surface A with magnitude $|\underline{A}| = A$.

In purely electrical terms we thus can always replace a material with a *homogeneous* polarization $\underline{P}$ by two surfaces perpendicular to some direction, - lets say z - with a surface charge density of P_z (with, of course, different signs on the two different surfaces).

- If the polarization vector is *not* perpendicular to the surface we chose, we must take the *component* of the polarization vector *parallel to the normal vector* of the surface considered . This is automatically taken care of if we use the vector formulation for A.

A dielectric material now quite generally reacts to the presence of an electrical field by becoming polarized and this is expressed by the direction and magnitude of $\underline{P}$.

- $\underline{P}$ is a *measurable quantity* tied to the specific material under investigation. We now need a *material law* that connects cause and effect, i.e. a relation between the the electrical field causing the polarization and the amount of polarization produced.
- Finding this law is of course the task of basic theory. But long before the proper theory was found, experiments supplied a simple and rather (but not quite) *empirical* "law":

If we *measure* the polarization of a material, we usually find a linear relationship between the applied field E and P, i.e.

$$\underline{P} = \epsilon_0 \cdot \chi \cdot \underline{E}$$

- With the proportionality constant chosen to contain ϵ_0, the **permittivity constant** (of vacuum), times a material parameter χ ("kee"), the **dielectric susceptibility**.
- Note that including ϵ_0 in the relation is a convention which is useful in the **SI** system, where charges are always coupled to electrical fields via ϵ_0. There are other systems, however, (usually variants of the **cgs** system), which are still used by many and provide an unending source of confusion and error.
- This equation is to dielectric material what **Ohms** law is to conductors. It is no more a real "*law of nature*" than Ohms law, but a description of many experimental observations for which we will find deeper reasons forthwith.

Our task thus is to *calculate* χ from basic material parameters.

Connection between the Polarization $\underline{P}$ and the Electrical Displacement $\underline{D}$

Next we need the connection between the polarization P, or the dielectric susceptibility χ, with some older quantities often used in connection with **Maxwells equations**.

- Historically, *inside materials*, the electrical field strength E was (and still is) replaced by a vector $\underline{D}$ called the **electrical displacement** or **electrical flux density**, which is defined as

$$\underline{D} = \epsilon_r \cdot \epsilon_0 \cdot \underline{E}$$

- and ϵ_r was (and still is) called the **(relative) dielectric constant** (**DK**) of the material (the product $\epsilon_r \cdot \epsilon_0$ is called the **permittivity**).
- Note that in the English literature often the abbreviation κ ("Kappa") is used; in proper microelectronics slang one than talks of "low **k** materials" (pronounced "low khe" as in (O)K) when one actually means "low kappa" or "low epsilon relative".
- D is supposed to give the "acting" *flux* inside the material.

While this was a smart thing to do for Maxwell and his contemporaries, who couldn't know anything about materials (atoms had not been "invented" then); it is a bit unfortunate in retrospect because the *basic quantity* is the *polarization*, based on the elementary dipoles in the material, and the *material parameter* χ describing this - and not some changed "electrical flux density" and the relative dielectric constant of the material.

- It is, however, easy (if slightly confusing) to make the necessary connections. This is most easily done by looking at a simple plate capacitor. A full treatise is found in a basic module, here we just give the results.

The ***electric displacement*** $\underline{D}$ in a dielectric caused by some external field $\underline{E}_{ex}$ is the displacement D_0 in vacuum plus the *polarization* $\underline{P}$ of the material, i.e.

$$\underline{D} = \underline{D_0} + \underline{P} = \epsilon_0 \cdot \underline{E} + \underline{P}$$

Inserting everything we see that the relative dielectric constant ϵ_r is simply the *dielectric susceptibility* χ plus **1**.

$$\epsilon_r = 1 + \chi$$

For this "translations" we have used the relation $\boldsymbol{P = \epsilon_0 \cdot \chi \cdot E}$, which is *not* an ***a priori*** law of nature, but an empirical relation. However, we are going to prove this relation for specific, but very general cases forthwith and thus justify the equations above.

- We have also simply implied that $\underline{\boldsymbol{P}}$ is parallel to $\underline{\boldsymbol{E}}$, which is only reasonable for *isotropic materials*.
- In *anisotropic* media, e.g. non-cubic crystals, $\underline{\boldsymbol{P}}$ does not have to be parallel to $\underline{\boldsymbol{E}}$, the scalar quantities ϵ_r and χ then are *tensors*.

The basic task in the materials science of dielectrics is now to calculate (the tensor) χ from "first principles", i.e. from basic structural knowledge of the material considered. This we will do in the following paragraphs.

Questionaire

Multiple Choice questions to 3.1.1

3.1.2 Summary to: Polarization and Dielectric Constant

- The dielectric constant ϵ_r "somehow" describes the interaction of dielectric (i.e. more or less insulating) materials and electrical fields; e.g. via the equations ⇒
  - $\underline{D}$ is the **electrical displacement** or **electrical flux density**, sort of replacing $\underline{E}$ in the Maxwell equations whenever materials are encountered.
  - ***C*** is the capacity of a parallel plate capacitor (plate area ***A***, distance ***d***) that is "filled" with a dielectric with ϵ_r
  - ***n*** is the index of refraction; a quantity that "somehow" describes how electromagnetic fields with extremely high frequency interact with matter.
    in this equaiton it is assumed that the material has no magnetic properties at the frequency of light.

$$\underline{D} = \epsilon_0 \cdot \epsilon_r \cdot \underline{E}$$

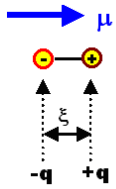
$$C = \frac{\epsilon_0 \cdot \epsilon_r \cdot A}{d}$$

$$n = (\epsilon_r)^{1/2}$$

- Electrical fields inside dielectrics polarize the material, meaning that the vector sum of electrical dipoles inside the material is no longer zero.
  - The decisive quantities are the dipole moment $\underline{\mu}$, a vector, and the Polarization ***P***, a vector, too.
  - Note: The dipole moment vector points from the negative to the positive charge - contrary to the electrical field vector!
  - The dipoles to be polarized are either already present in the material (e.g. in H_2O or in ionic crystals) or are induced by the electrical field (e.g. in single atoms or covalently bonded crystals like **Si**)
  - The dimension of the polarization $\underline{P}$ is **[C/cm^2]** and is indeed identical to the net charge found on unit area ion the surface of a polarized dielectric.

$$\underline{\mu} = q \cdot \underline{\xi}$$

$$\underline{P} = \frac{\Sigma \underline{\mu}}{V}$$



- The equivalent of "Ohm's law", linking current density to field strength in conductors is the Polarization law:
  - The decisive material parameter is χ ("kee"), the **dielectric susceptibility**
  - The "classical" flux density ***D*** and the Polarization are linked as shown. In essence, $\underline{P}$ only considers what happens in the material, while ***D*** looks at the total effect: material plus the field that induces the polarization.

$$\underline{P} = \epsilon_0 \cdot \chi \cdot \underline{E}$$

$$\epsilon_r = 1 + \chi$$

$$\underline{D} = \underline{D_0} + \underline{P} = \epsilon_0 \cdot \underline{E} + \underline{P}$$

- Polarization by necessity moves masses (electrons and / or atoms) around, this will not happen arbitrarily fast.
  - ϵ_r or χ thus must be functions of the frequency of the applied electrical field, and we want to consider the whole frequency range from **RF** via **HF** to light and beyond.

$\epsilon_r(\omega)$ is called the "**dielectric function**" of the material.

- The tasks are:
  - Identify and (quantitatively) describe the major mechanisms of polarization.
  - Justify the assumed linear relationship between ***P*** and χ.
  - Derive the dielectric function for a given material.

Questionaire

Multiple Choice questions to all of 3.1

3.2 Mechanisms of Polarization

3.2.1 General Remarks

- We have a material and we want to know its dielectric constant ϵ_r or dielectric susceptibility χ. We would want to have those quantities as functions of various variables for the same basic materials, e.g.
  - $\chi = \chi(\omega)$,
    i.e. χ as a function of the ***angular frequency*** ω of the electrical field.
  - $\chi = \chi(T)$;
    i.e. the dependence on the temperature ***T***.
  - $\chi = \chi$**(structure)**,
    i.e. the dependence of χ on the structure of a material including the kind and density of defects in the material. As an example we may ask how χ differs from *amorphous* to *crystalline* quartz (**SiO_2**).
- The answers to all of these questions must be contained in the *mechanisms* with which atoms and molecules respond to an electrical field, i.e. in the mechanisms leading to the formation and/or orientation of dipoles. These mechanisms are called **polarization mechanisms**.
  - We want a general *theory of polarization*. This is a complex task as well it must be, given the ***plethora*** of dielectric phenomena. However, the *basic principles* are rather simple, and we are only going to look at these.
- There are essentially *four basic kinds of polarization mechanisms*:
  - **Interface polarization**.
    Surfaces, grain boundaries, interphase boundaries (including the surface of precipitates) may be *charged*, i.e. they contain dipoles which may become oriented to some degree in an external field and thus contribute to the polarization of the material.
  - ***Electronic polarization,***
    also called atom or **atomic polarization**. An electrical field will always displace the center of charge of the electrons with respect to the nucleus and thus induce a dipole moment as discussed before. The ***paradigmatic*** materials for the simple case of *atoms with a spherical symmetry* are the noble gases in all aggregate forms.
  - **Ionic polarization**.
    In this case a (solid) material must have some ionic character. It then automatically has internal dipoles, but these built-in dipoles exactly cancel each other and are unable to rotate. The external field then *induces net dipoles* by slightly displacing the ions from their rest position. The paradigmatic materials are all simple ionic crystals like **NaCl**.
  - **Orientation polarization**.
    Here the (usually liquid or gaseous) material must have *natural dipoles* which can rotate freely. In thermal equilibrium, the dipoles will be randomly oriented and thus carry no net polarization. The external field aligns these dipoles to some extent and thus induces a polarization of the material. The paradigmatic material is water, i.e. **H_2O** in its liquid form.
- Some or all of these mechanisms may act simultaneously. Atomic polarization, e.g., is always present in any material and thus becomes superimposed on whatever other mechanism there might be.
  - Real materials thus can be very complicated in their dielectric behavior. In particular, *non-spherical atoms* (as, e.g., **Si** in a crystal with its four **sp^3** orbitals) may show complex electronic polarization, and mixtures of ionic and covalent bonding (e.g. in **SiO_2**, which has about equal ionic and covalent bonding contributions) makes calculations even more difficult. But the basic mechanisms are still the ones described above.
- The last three mechanisms are ***amenable*** to basic considerations and calculations; ***interface polarization***, however, defies basic treatment. There is simply no *general* way to calculate the charges on ***interfaces*** nor their contribution to the total polarization of a material.
  - Interface polarization is therefore often omitted from the discussion of dielectric properties. Here, too, we will not pursue this matter much further.
  - It would be totally wrong, however, to conclude that *interface polarization is technically not important* because, on the one hand, many dielectrics in real capacitors rely on interface polarization while, on the other hand, interface polarization, if present, may "kill" many electronic devices, e.g. the **MOS** transistor!
  - Let's look at this in an exercise:

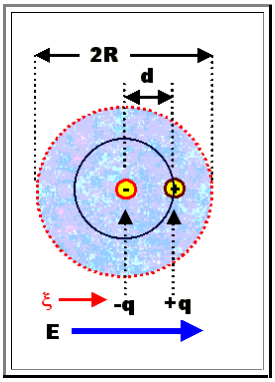
Exercise 3.2-2
Maximum polarization of water

Questionaire

Multiple Choice questions to 3.2.1

3.2.2 Electronic Polarization

For calculating the effect of **electronic polarization**, We consider an idealized atom with *perfect* ***spherical symmetry***.



- It has a point like charge **+ *ze*** in the nucleus, and
- The exact opposite charge **– *ze*** homogeneously distributed in the volume of the atom, which is

$$V = \frac{4}{3}\pi R^3$$

- With ***R*** **=** radius of the atom

The charge density ρ of the electrons then is

$$\rho = -\frac{3z \cdot e}{4\pi \cdot R^3}$$

In an electrical field ***E*** a force F_1 acts on charges given by

$$F_1 = z \cdot e \cdot E$$

- We will now drop the underlining for vectors and the mauve color for the electrical field strength ***E*** for easier readability.

The positive charge in the nucleus and the center of the negative charges from the electron "cloud" will thus experience forces in different direction and will become separated. We have the idealized situation shown in the image above.

The separation distance ***d*** will have a finite value because the separating force of the external field is exactly balanced by the attractive force between the centers of charge at the distance ***d***.

- How large is his attractive force? It is not obvious because we have to take into account the attraction between a *point* charge and *homogeneously distributed* charge.
- The problem is exactly analogous to the classical mechanical problem of a body with mass ***m*** falling through a hypothetical hole going all the way from one side of the globe to the other.
- We know the solution to that problem: The attractive force between the point mass and the earth is equal to the attractive force between two point masses if one takes *only* the mass of the volume inside the sphere given by the distance between the center of the spread-out mass and the position of the point mass.
- Knowing electrostatics, it is even easier to see why this is so. We may divide the force on a charged particles on any place inside a homogeneously charged sphere into the force from the "*inside*" sphere and the force from the hollow "*outside*" sphere. Electrostatics teaches us, that a sphere charged on the outside has *no field* in the *inside* , and therefore *no force* (the principle behind a Faraday cage). Thus we indeed only have to consider the "charge *inside* the sphere.

For our problem, the attractive force F_2 thus is given by

$$F_2 = \frac{\text{q(Nucleus)} \cdot \text{q(e in } d\text{)}}{4\pi \epsilon_0 \cdot d^2}$$

- with **q(Nucleus) = ze** and **q(e in *d*)** = the fraction of the charge of the electrons contained in the sphere with radius ***d***, which is just the relation of the volume of the sphere with radius ***d*** to the *total* volume . We have

$$q(\text{e in } d) = ze \cdot \frac{(4/3)\,\pi \cdot d^3}{(4/3)\,\pi \cdot R^3} = \frac{ze \cdot d^3}{R^3}$$

- and obtain for F_2:

$$F_2 = \left(\frac{(ze)^2}{4\,\pi\epsilon_0 \cdot R^3} \right) \cdot d$$

We have a linear force law akin to a spring; the expression in brackets is the "spring constant". Equating F_1 with F_2 gives the equilibrium distance d_E.

$$d_E = \frac{4\,\pi\epsilon_0 \cdot R^3 \cdot E}{ze}$$

Now we can calculate the **induced dipole moment** μ, it is

$$\mu = ze \cdot d_E = 4\,\pi\epsilon_0 \cdot R^3 \cdot E$$

The polarization P finally is given by multiplying with N, the density of the dipoles; we obtain

$$P = 4\,\pi \cdot N \cdot \epsilon_0 \cdot R^3 \cdot E$$

- Using the definition $P = \epsilon_0 \cdot \chi \cdot E$ we obtain the *dielectric susceptibility resulting from atomic polarization*, χ_{atom}

$$\chi_{atom} = 4\,\pi \cdot N \cdot R^3$$

Let's get an idea about the numbers involved by doing a simple exercise:

Exercise 3.2-3

Some numbers for atomic polarization

This is our first basic result concerning the polarization of a material and its resulting susceptibility. There are a number of interesting points:

- We justified the "law" of a linear relationship between E and P for the electronic polarization mechanism (sometimes also called atomic polarization).
- We can easily extend the result to a mixture of different atoms: All we have to do is to sum over the relative densities of each kind of atom.
- We can easily get an order of magnitude for χ. Taking a typical density of $N \approx 3 \cdot 10^{19}\ cm^{-3}$ and $R \approx 6 \cdot 10^{-9}\ cm$, we obtain

$$\chi \approx 8{,}14 \cdot 10^{-5}, \quad \text{or}$$

$$\epsilon_r = 1,\,000\,0814$$

In words: the electronic polarization of *spherical* atoms, while existent, is *extremely weak*. The difference to vacuum is at best in the promille range.

Concluding now that electronic polarization is totally unimportant, would be ***premature***, however. Atoms in crystals or in any solids do not generally have *spherical* symmetry. Consider **the sp^3** orbital of **Si**, **Ge** or diamond.

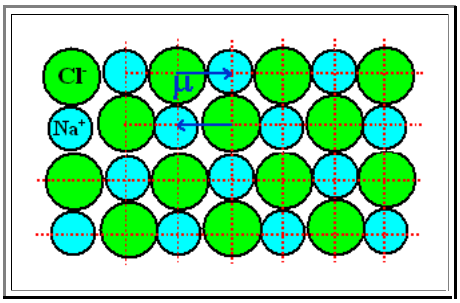
- Without a field, the center of the negative charge of the electron orbitals will still coincide with the core, but an external field breaks that symmetry, producing a dipole momentum.
- The effect can be *large* compared to spherical **s**-orbitals: **Si** has a dielectric constant (**DK**) of **12**, which comes exclusively from electronic polarization! Some values for semiconductors are given in the link.

Questionaire

Multiple Choice questions to 3.2.2

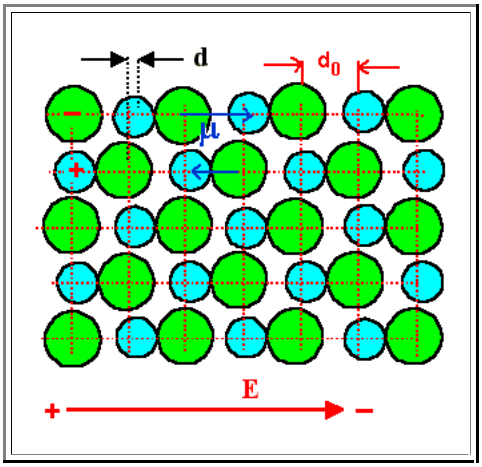
3.2.3 Ionic Polarization

- Consider a simple ionic crystal, e.g. **NaCl**.
  - The lattice can be considered to consist of **Na^+ - Cl^-** dipoles as shown below.



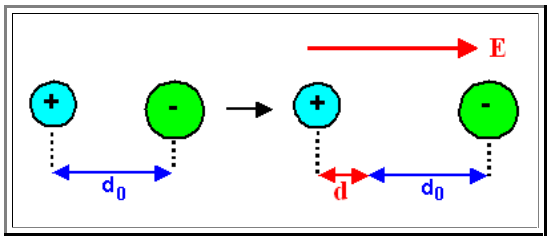
  - Each **Na^+ - Cl^-** pair is a *natural dipole*, no matter how you pair up two atoms.
  - The polarization of a given volume, however, is *exactly zero* because for every dipole moment there is a neighboring one with exactly the same magnitude, but opposite sign.
  - Note that the dipoles *can not rotate*; their direction is fixed.

- In an electric field, the ions feel forces in opposite directions. For a field acting as shown, the lattice distorts a little bit (hugely exaggerated in the drawing)



  - The **Na^+** ions moved a bit to the right, the **Cl^-** ions to the left.
  - The dipole moments between adjacent **NaCl** - pairs in field direction are now different and there is a *net dipole moment* in a finite volume now.

- From the picture it can be seen that it is sufficient to consider *one* dipole in field direction. We have the following situation:



  - Shown is the situation where the distance between the ions *increases* by ***d***; the symmetrical situation, where the distance *decreases* by ***d***, is obvious.
- How large is ***d***? That is easy to calculate:
  - The force F_1 increasing the distance is given by

$$F_1 = q \cdot E$$

  - With ***q*** = net charge of the ion.
  - The restoring force F_2 comes from the binding force, it is given as the derivative of the binding potential. Assuming a *linear relation* between binding force and deviation from the equilibrium distance d_0, which is a good approximation for $d << d_0$, we can write

$$F_2 = k_{IP} \cdot d$$

- With k_{IP} being the "*spring constant*" of the bond. k_{IP} can be calculated from the bond structure, it may also be expressed in terms of other constants that are directly related to the shape of the interatomic potential, e.g. the *modulus of elasticity* or *Youngs modulus*.
- If we do that we simply find

$$k_{IP} = Y \cdot d_0$$

- With Y = Youngs Modulus, and d_0 = equilibrium distance between atoms.

From force equilibrium. i.e. $F_1 - F_2 = 0$, we immediately obtain the following relations:

- *Equilibrium distance* d

$$d = \frac{q \cdot E}{Y \cdot d_0}$$

- *Induced dipole moment* μ (on top of the existing one)

$$\mu = \frac{q^2 \cdot E}{Y \cdot d_0}$$

- *Polarization* P

$$P = \frac{N \cdot q^2 \cdot E}{Y \cdot d_0}$$

Of course, this is only a very rough approximation for an *idealized* material and just for the case of increasing the distance. Adding up the various moments - some larger, some smaller - will introduce a factor **2** or so; but here we only go for the principle.

For *real* ionic crystals we also may have to consider:

- More complicated geometries (e.g. CaF_2, with ions carrying different amount of charge).
- This example was deliberatly chosen: The dielectric constant of CaF_2 is of paramount interest to the semiconductor industry of the **21**st century, because CaF_2 is pretty much the only usable material with an index of refraction n (which is directly tied to the **DK** via $\epsilon_r = n^2$) that can be used for making lenses for lithography machines enabling dimensions of about **0,1 µm**.
- If the field is not parallel to a major axis of the crystal (this is automatically the case in polycrystals), you have to look at the components of μ in the field direction and average over the ensemble.

Still, the basic effects is the same and ionic polarization can lead to respectable dielectric constants ϵ_r or susceptibilities χ.

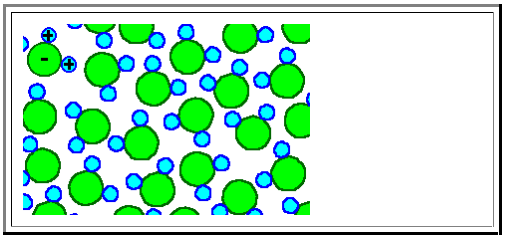
- Some values are given in the link.

Questionaire

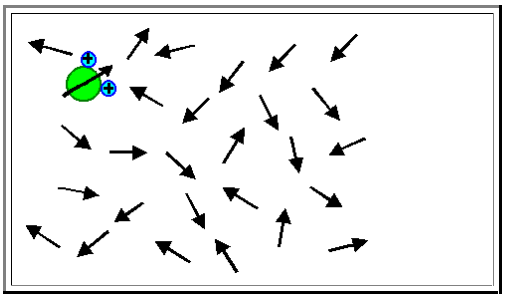
Multiple Choice questions to 3.2.3

3.2.4 Orientation Polarization

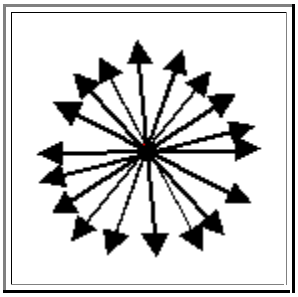
- In the case of **orientation polarization** we have a material with *built-in dipoles* that are *independent of each other*, i.e. they can *rotate freely* - in sharp contrast to ionic polarization.
  - The prime example is *liquid water*, where every water molecule is a little dipole that can have any orientation with respect to the other molecules. Moreover, the orientation changes *all the time* because the *molecules moves*! Orientation polarization for dielectric dipoles thus is pretty much limited to liquids - but we will encounter it in a major way again for *magnetic dipoles*.
  - A two-dimensional "piece of water" may - very graphically - look somewhat like the picture below that captures *one particular moment in time*. It is like a snapshot with a very, very short exposure time. A few nanoseconds later the same piece of water may look totally different in detail, but pretty much the same in general.
  - In a three-dimensional piece of water the blue and red circles would not have to be in the same plane; but that is easy to imagine and difficut to draw.



  - Shown is a bunch of water molecules that form natural dipoles because the negatively charged oxygen atom and the two positively charged **H** - atoms have different centers of charge. Each molecule carries a dipole moment which can be drawn as a vector of constant length. If we only draw a vector denoting the dipole moment, we get - in two dimensions - a picture like this:



  - Again, remember that both pictures are "*snap shots*" that only appear unblurred for very small exposure times, say picoseconds, because the dipoles wiggle, rotate, and move around rather fast, and that in *three* dimensions the vectors would also point out of the drawing plane.
- The total dipole moment is the *vector sum* of the individual dipole moments.

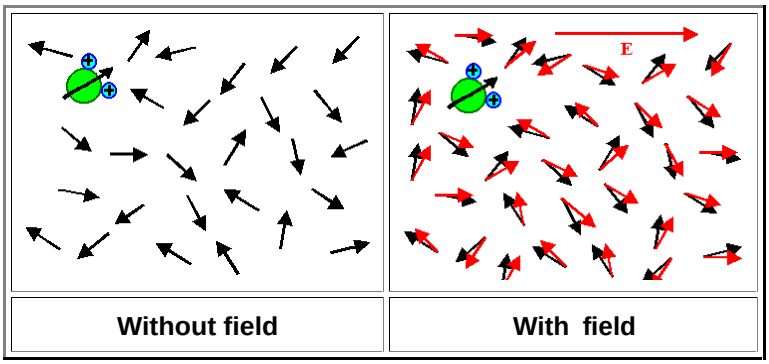


  - For dipoles oriented *at random*, at any given moment this looks like the picture below if we draw all vectors from a common origin: The sum of all dipole moments will be zero, if the dipoles are randomly oriented.
  - We can see this most easily if we have all dipoles start at the same origin. The picture, of course, is two-dimensional and crossly simplified. There would be a lot more (like $\mathbf{10}^{\mathbf{20}}$) dipoles for any appreciable amount of water - you really will average them to zero pretty well.
- If we now introduce a field ***E***, the dipoles would have a tendency to turn into the field because that would lower their energy.
  - If you have problems with this statement, just imagine the electrostatic interaction, which will always try to move the positive pole of the dipole towards the negative pole of the field, and vice versa for the negative pole - the dipole would align itself exactly along a field line of the external field for minimum energy.
  - Naively, we would then expect a *perfect orientation into the field* and a ***concomitantly*** large polarization because that would lead to the *minimum of the dipole energy*.
  - Well, water does have a pretty large **DK** of **81**, so there is obviously *some* orientation into the field, but it is easy (not really) to show (in an exercise) that this **DK** is several orders of magnitude too small for *fully* oriented dipole moments at some normal field strengths.

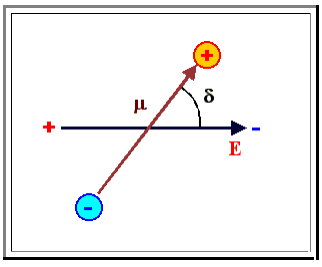
Exercise 3.2-1

Maximum polarization of water

- In reality, the orientation into the field direction will be *counteracted by random collisions* with other dipoles, and this process is energized by the *thermal energy* "**k*T***" contained in the water.
  - Again, the dipoles are not sitting still, but moving around and rotating all the time - because they contain *thermal energy* and thus also some **entropy**.
  - Whenever two molecules collide, their new orientation is *random* - all memory of an orientation that they might have had in the electrical field is lost. This is analogous to what happens to electrons carrying an electrical current in an electrical field.
  - The electrical field only induces a little bit of *average* orientation in field direction - most of the time an individual dipole points in all kinds of directions. This is the simple truth even so some (undergraduate) text books show pictures to the contrary. The "real" picture (in the sense of a snapshot with a very short exposure time) looks like this:



  - The orientation of all dipoles is just a little bit shifted so that an average orientation in field direction results. In the picture, the effect is even exaggerated!
- In fact, the state of being *liquid* by necessity implies quite a bit of **entropy**, and entropy means *disorder*.
  - Perfectly aligned dipoles would be in *perfect order* without any entropy - this is only possible at extremely low temperatures (and even there quantum theory would not allow it) where we will not have liquids any more, or more generally, dipoles that are able to rotate freely.
  - In other words, we must look for the minimum of the **free enthalpy *G*** and not for the minimum of the **internal energy *U***. At finite temperatures the minimum of the free enthalpy requires some **entropy *S***, i.e. randomness in the dipole orientation, so we should not expect perfect orientation.
- If you are not familiar with the basics of thermodynamics, you have a problem at this point. If you do know your thermodynamics, but are a bit insecure, turn to the basic module "Thermodynamics" (in the "Defects" Hyperscript) to refresh your memory.
- We obviously need to calculate the free enthalpy ***G=U – TS*** to see what kind of average orientation will result in a given field. Note that we use ***U***, the common symbol for the (internal) energy instead of ***H***, the common symbol for the enthalpy, because ***U*** and ***H*** are practically identical for solids and liquids anyway.
  - Moreover, a mix up with the magnetic field strength usually designated by ***H***, too, would be unavoidable otherwise. (The possible mix-up between internal energy ***U*** and voltage ***U*** is not quite so dangerous in this context).
- The internal energy od a dipole is clearly a function of its orientation with respect to the field. It must be minimal, when the dipole is aligned with the field and thedipole moment has the same direction as the electrical field, and maximal if the direction is reversed.
  - This is the easy part: The energy ***U*(δ)** of a dipole with dipole moment μ in a field ***E*** as a function of the angle δ ("delta") between the dipole moment direction and the field direction.



  - From basic electrostatics we have have

$$U(\delta) = -\ \underline{\mu} \cdot \underline{E} = -\ |\underline{\mu}| \cdot |\underline{E}| \cdot \cos\delta$$

  - The *minimum energy* ***U*** thus would occur for **δ=0°**, i.e. for perfect alignment in proper field direction (*note the minus sign!*); the maximum energy for **δ=180°**, i.e. for alignment the wrong way around.
  - That was for two dimensions - now we must look at this in *three* dimensions.

- In **3D** we see that all dipoles with the same angle δ between their axis and the field still have the same energy - and this means now all dipoles on a *cone* with opening angle **2δ** around the field axis if we consider possible orientations out of the plane of drawing.
- In order to obtain the *total internal energy* U_{total} of a *bunch* of dipoles having *all kinds* of angles δ with the field axis, we will have to sum up all cones.
- This means we take the number of dipoles $N(\delta)$ having a particular orientation δ times the energy belonging to that δ, and integrate the resulting function over δ from 0^o to 180^o. This is something that we could do - if we would know $N(\delta)$.

However, just calcuating U_{total} will not be of much use. We also *must* consider the entropy term $-TS$, because we do not want to calculate the total *internal energy* U_{total}, but the total *free enthalpy* $G=U_{total}-TS$.

- We need to consider that term as a function of all possible angle distributions and then see for which distribution we can minimize G.

But what is the entropy $S(N(\delta))$ of an ensemble of dipoles containing $N(\delta)$ members at the angle δ as a function of the many possible distribution $N(\delta)$? Not an easy question to answer from just looking at the dipoles.

Fortunately, we do not have to calculate ***S*** *explicitly!*

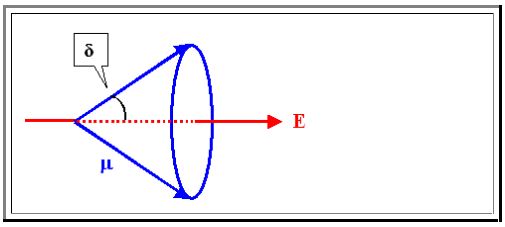
We *know* a formula for the distribution of (classical) particles on available energy levels that *automatically* gives the minimum of the free enthalpy!

- We have a *classical* system where a number of independent particles (the dipoles) can occupy a number of energy levels (between U_{min} and U_{max}) as defined by $\delta=0^o$ or $\delta=180^o$, respectively.
- Basic thermodynamics asserts that *in equilibrium*, the distribution of the particles on the available energy levels is given by the proper *distribution function* which is defined in such a way that it *always* gives the minimum of the free enthalpy.
- Since we deal with classical particles in this approach, we have to use the **Boltzmann distribution**. We obtain for $N(U)=$ number of dipoles with the energy U

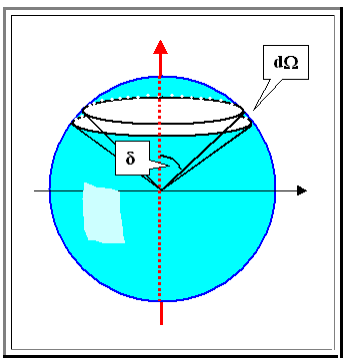
$$N(U) = A \cdot \exp - \frac{U(\delta)}{kT}$$

- With a constant A that has yet to be determined.

This Boltzmann distribution equation gives us the number of dipoles with a certain angle relative to the field direction, i.e. the number of dipoles that have their tips on a circle with an opening angle **2δ** relative to the field directions as shown below.



- We are, however, only interested in the *component of the dipole moment parallel to the field*. For this we look at the ***solid angle*** increment $d\Omega$ defined on the unit sphere as the segment between δ and $\delta + d\delta$.



- The number of dipoles lying in the cone angle increment defined by δ and $\delta+\Delta\delta$ is the same as the number of dipoles with tips ending on the surface of the unit sphere in the incremental angle $d\Omega$. It is given by $N(U(\delta)) \cdot d\Omega$.
- Note that $d\Omega$ is a measure of an incremental *area*; a kind of ribbon once around the unit sphere.
- The sum of the components μ_F of the dipole moments *in field direction* is then .

$$\mu_F = (N \cdot d\Omega) \cdot (\mu \cdot \cos \delta)$$

- If you are not familiar with spherical coordinates, this (and what we will do with it), looks a bit like magic. Since we do not want to learn Math in this lecture, the ***essentials to spherical coordinates*** are explained in detail in a basic module.

The *average dipole moment*, which is what we want to calculate, will now be obtained by summing up the contributions from all the **d**Ωs

$$\langle \mu_F \rangle = \frac{\int_0^{\pi} N(U(\delta)) \cdot \mu \cdot \cos\delta \cdot d\Omega}{\int_0^{\pi} N(U(\delta)) \cdot d\Omega}$$

- And the integrals have to be taken from the "top" of the sphere to the "bottom" , i.e. from **0** to π.

dΩ and δ are of course closely related, we simply have

$$d\Omega = 2\pi \cdot \sin\delta \cdot d\delta$$

Putting everything together, we obtain a pretty horrifying integral for μ_F that runs from **0** to π

$$\langle \mu_F \rangle = \frac{\mu \cdot \int_0^{\pi} \sin\delta \cdot \cos\delta \cdot \exp\frac{\mu \cdot E \cdot \cos\delta}{kT} \cdot d\delta}{\int_0^{\pi} \sin\delta \cdot \exp\frac{\mu \cdot E \cdot \cos\delta}{kT} \cdot d\delta}$$

One advantage is that we got rid of the undetermined constant ***A***. The integral, being a determined integral, is now simply a *number* depending on the parameters of the system, i.e. the temperature ***T***, the dipole moment μ and the field strength ***E***.

- *The problem has been reduced to a mathematical exercise in solving integrals*.

Since we are not interested at doing math, we just show the general direction toward a solution:

- Use the substitutions

$$\beta = \frac{\mu \cdot E}{kT}$$

$$x = \cos\delta$$

- The integral reduces to

$$\langle \mu_F \rangle = \frac{\mu \cdot \int_{+1}^{-1} x \cdot \exp(\beta \cdot x) \cdot dx}{\int_{+1}^{-1} \exp(\beta \cdot x) \cdot dx}$$

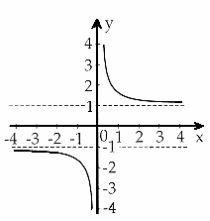
The final result after quite a bit of fiddling around is

$$\langle \mu_F \rangle = \mu \cdot L(\beta)$$

With **L(β)=Langevin function**, named after Paul Langevin, and defined as
.

$$L(\beta) = \coth(\beta) - \frac{1}{\beta}$$

$$\beta = \frac{\mu \cdot E}{kT}$$



y*=coth *x

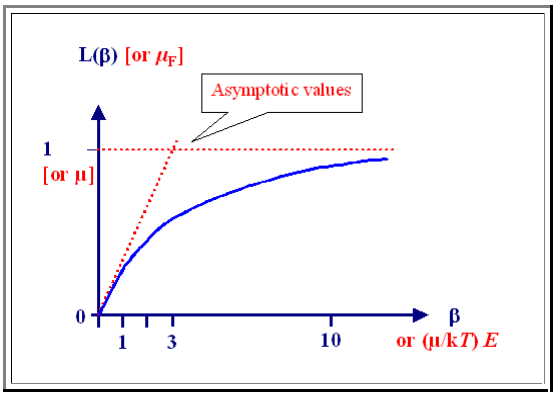
The "**coth**" is the *hyperbolic cotangent*, defined as **coth *x*=(e^x + e^{-x})/(e^x – e^{-x})=1/tanh *x***.

L(β) is a tricky function, because the **coth *x*** part looks pretty much like a hyperbola, from which the real hyperbola **1/*x*** is subtracted. What's left is almost nothing - **L(*x*)** values are between **0** and **1**

The polarization (always on average, too) is accordingly
.

$$P = N \cdot \langle \mu \rangle$$

This is a definite result, but it does not help much. We need to discuss the mathematical construct "*Langevin function* **L(β)**" to get some idea of what we obtained. We look at the graph in general units and in units of the dipole moment and electrical field (in red).



Since β is proportional to the field strength ***E***, we see that the dipole moment and the polarization increases monotonically with ***E***, eventually saturating and giving $\langle \mu_F \rangle = \mu$ which is what we must expect.

The question is, what *range* of β values is accessible for real materials. i.e. how close to the saturation limit can we get?

For that we look at some simple approximations.

If we develop **L(β)** into a series (consult a math textbook), we get

$$L(\beta) = \frac{\beta}{3} - \frac{\beta^3}{45} + \frac{2\beta^5}{945} - \ldots\ldots$$

- For large values of β we have **L(β) ≈ 1**, while for small values of β (**β < 1**), the Langevin function can be approximated by .

$$L(\beta) \approx 1/3 \cdot \beta$$

- The slope thus is **1/3** for **β → 0**.
- For "normal" circumstances, we always have **β << 1** (see below), and we obtain as final result for the **induced dipole moment** the **Langevin - Debye** equation

$$\langle \mu \rangle = \frac{\mu^2 \cdot E}{3kT}$$

$$\langle P \rangle = \frac{N \cdot \mu^2 \cdot E}{3kT}$$

- These equations will be rather good approximation for small values of μ and **E** and/or large values of ***T***. For very large fields and very small temperatures the average dipole moment would be equal to the built in dipole moment, i.e. all dipoles would be strictly parallel to the field. This is, however, not observed in "normal" ranges of fields and temperatures.

Lets see that in an example. We take

- ***E*=10^8 V/cm** which is about the highest field strength imaginable before we have electrical breakdown, **μ=10$^{-2\ 9}$ Asm**, which is a large dipole moment for a strongly polarized molecule, e.g. for **HCl**, and ***T*=300 K**.
- This gives us
  β=0,24 - the approximation is still valid. You may want to consult exercise 3.2-1 again (or for the first time) at this point and look at the same question from a different angle.

At ***T*=30 K**, however, we have **β=2,4** and now we must think twice:

- **1.** The approximation would no longer be good. *But*
- **2.** We no longer would have *liquid* **HCl** (or **H_2O**, or liquid whatever with a dipole moment), but solid **HCl** (or whatever) , and we now look at ionic polarization and no longer at orientation polarization!

You may now feel that this was a rather useless exercise - after all, who is interested in the **DK** of liquids? But consider: This treatment is *not* restricted to electric dipoles. It is valid for all kinds of dipoles that can rotate freely, in particular for the *magnetic dipoles* in paramagnetic materials responding to a magnetic field.

- Again, you may react with stating "Who is interested in paramagnets? Not an electrical engineer!" Right - but the path to *ferromagnets*, which definitely are of interest, starts exactly where orientation polarization ends; you cannot avoid it.

It is important to be aware of the basic condition that we made at the beginning: *there is no interaction between the dipoles*! This will not be true in general.

- Two water molecules coming in close contact will of course "feel" each other and they may have preferred orientations of their dipole moments relative to each other. In this case we will have to modify the calculations; the above equations may no longer be a good approximation.

- On the other hand, if there is a *strong* interaction, we automatically have some bonding and obtain a solid - ice in the case of water. The dipoles most likely cannot orientate themselves freely; we have a different situation (usually ionic polarization). There are, however, some solids where dipoles exist that can rotate to some extent - we will get very special effects, e.g. "ferroelectricity".

Questionaire

Multiple Choice questions to 3.2.4

3.2.5 Summary and Generalization

- For all three cases of polarization mechanisms, we had a *linear* relationship between the electrical field and the dipole moment (for fields that are not excessively large):

Electronic polarization	Ionic polarization	Orientation polarization
$\mu_{EP}=4\pi \cdot \epsilon_0 \cdot R^3 \cdot E$	$\mu_{IP} = \frac{q^2}{k_{IP}} \cdot E$	$\mu_{op} = \frac{\mu^2}{3kT} \cdot E$

- It seems on a first glance that we have justified the "*law*" $P=\chi \cdot E$.
  - However, that is not quite true at this point. In the "law" given by equation above, E refers to the *external* field, i.e. to the field that would be present in our capacitor *without* a material inside.
  - We have $E_{ex}=U / d$ for our plate capacitor held at a voltage U and a spacing between the plates of d.
  - On the other hand, the induced dipole moment that we calculated, always referred to the *field at the place of the dipole*, i.e. the *local* field E_{loc}. And if you think about it, you should at least feel a bit uneasy in assuming that the two fields are identical. We will see about this in the next paragraph.
- Here we can only define a factor that relates μ and E_{loc}; it is called the **polarizability** α. It is rarely used with a number attached, but if you run across it, be careful if ϵ_0 is included or not; in other words what kind of unit system is used.
  - We now can reformulate the three equations on top of this paragraph into one equation

$$\mu = \alpha \cdot E_{loc}$$

  - The ***polarizability*** α is a material parameter which depends on the polarization mechanism: For our three paradigmatic cases they are are given by

$$\alpha_{EP} = 4\pi \cdot \epsilon_0 \cdot R^3$$

$$\alpha_{IP} = \frac{q^2}{k_{IP}}$$

$$\alpha_{OP} = \frac{\mu^2}{3kT}$$

  - This does not add anything new but emphasizes the proportionality to E.
- So we *almost* answered our first basic question about dielectrics - but for a full answer we need a relation between the *local* field and the *external* field. This, unfortunately, is *not a particularly easy problem*
  - One reason for this is: Whenever we talk about electrical fields, we always have a certain scale in mind - without necessarily being aware of this. Consider: In a metal, as we learn from electrostatics, there is *no field at all*, but that is *only true* if we do not look too closely. If we look on an *atomic scale*, there are tremendous fields between the nucleus and the electrons. At a somewhat larger scale, however, they disappear or perfectly balance each other (e.g. in ionic crystals) to give no field on somewhat larger dimensions.
  - The scale we need here, however, is the *atomic scale*. In the electronic polarization mechanism, we actually "looked" *inside* the atom - so we shouldn't just stay on a "rough" scale and neglect the fine details.
- Nevertheless, that is what we are going to do in the next paragraph: *Neglect the details*. The approach may not be beyond reproach, but it works and gives simple relations.

Questionaire

Multiple Choice questions to all of 3.2

3.2.6 Local Field and Clausius - Mosotti Equation

- "Particles", i.e. atoms or molecules in a liquid or solid are basking in electrical fields - the external field that we apply from the outside is not necessarily all they "see" in terms of fields.
  - First, of course, there is a tremendous electrical field *inside* any atom. We have after all, positive charges and negative charges separated by a distance roughly given by the diameter of the atom.
  - Second, we have fields *between* atoms, quite evident for ionic crystals, but also possible for other cases of bonding.
- However, if you look at the materials at a scale somewhat larger than the atomic scale, all these fields must *average to zero*. Only then do we have a field-free interior as we always assume in ***electrical engineering*** ("no electrical field can penetrate a metal").
- Here, however, we are looking at the effect an external field has on *atoms* and *molecules*, and it would be ***preposterous*** to assume that what an atom "sees" as *local* electrical field is identical to what we apply from the outside.
  - Since all our equations obtained so far always concerned the *local* electrical field - even if we did not point that out in detail before - we now must find a relation between the external field and the local field, if we want to use the insights we gained for understanding the behavior of dielectrics on a macroscopic scale.
- We define the **local field $\boldsymbol{E_{loc}}$** to be the field felt by *one* particle (mostly an atom) of the material at its position **(x,y,z)**.
  - Since the superposition principal for fields always holds, we may express $\boldsymbol{E_{loc}}$ as a superposition of the external field $\boldsymbol{E_{ex}}$ and some field $\boldsymbol{E_{mat}}$ introduced by the surrounding material. We thus have

$$\underline{E}_{loc} = \underline{E}_{ex} + \underline{E}_{mat}$$

- All electrical fields can, in principle, be calculated from looking at the charge distribution $\rho(x, y, z)$ in the material, and then solving the Poisson equation (which you should know). The Poisson equation couples the charge distribution and the potential $V(x, y, z)$ as follows:

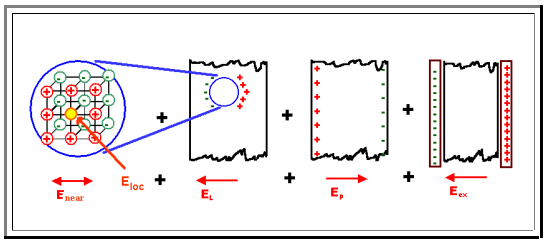
$$\Delta V = -\frac{\rho(x, y, z)}{\epsilon \cdot \epsilon_0}$$

$$\Delta = \text{Delta operator} = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2}$$

  - The electrical field then is just the (negative) gradient of the potential; $\underline{E} = -\nabla V$.
- Doing this is pretty tricky, however. We can obtain usable results in a good approximation in a much simpler way, by using the time-honored **Lorentz** approach or the **Lorentz model**.
  - In this approach we decompose the total field into *four* components.
  - For doing this, we *imagine* that we remove a small sphere containing a few **10** atoms from the material. We want to know the local field in the center of this sphere while it is still in the material; this is the local field $\underline{\boldsymbol{E}}_{mat}$ we are after. We define that field by the force it exerts on a charge at the center of the sphere that acts as a "probe".
- The essential trick is to calculate the field produced from the atoms inside the sphere and the field inside the now empty sphere in the material. The total local field then is simple the sum of both.
  - Like always, we do not consider the charge of the "probe" in computing the field that it probes. *The cut-out sphere thus must not contain the charge we use as the field probe!*
  - The cut-out material, in general, could produce an electrical field at its center since it is composed of charges. This is the ***1st*** component of the field, $\underline{\boldsymbol{E}}_{near}$ which takes into account the contributions of the atoms or ions inside the sphere. We will consider that field in an approximation where we average over the volume of the small sphere. To make things clear, we look at an ionic crystal where we definitely have charges in our sphere.
- $\underline{\boldsymbol{E}}_{near}$, however, is not the *only* field that acts on our probe. We must include the field that all the other atoms of the crystal produce *in* the hollow sphere left after we cut out some material. This field now fills the "empty" void left by taking out our sphere.
  - This field is called $\underline{\boldsymbol{E}}_{L}$ (the "***L***" stands for Lorentz); it compensates for the cut-out part - and that provides our ***2nd*** component.
- Now we only have to add the "macroscopic" fields from **1.** the polarization of the material and **2.** the external field that causes everything:
  - The field $\underline{\boldsymbol{E}}_{pol}$ is induced by the macroscopic polarization (i.e. by area charges equal to the polarization); it is the ***3rd*** component.

- The external field $E_{ex} = U/d$ from the applied voltage at our capacitor which supplies the *4th* component.

In a visualization, this looks like this:



The blue "sphere" cuts through the lattice (this is hard to draw). The yellow "point" is where we consider the local field; we have to omit the contribution of the charged atom there. We now have

$$E_{loc} = E_{ex} + E_{pol} + E_L + E_{near}$$

How large are those fields? We know the external field and also the field from the polarization (always assuming that the material completely fills the space inside the capacitor).

$$E_{ex} = \frac{U}{d} \qquad E_{pol} = -\frac{P}{\epsilon_0}$$

We do not know the other two fields, and it is not all that easy to find out how large they are. The results one obtains, however, are quite simple:

- Lorentz showed that $E_{near} = 0$ for *isotropic materials*, which is easy to imagine. Thus for cubic crystals (or polycrystals, or amorphous materials), we only have to calculate E_L.

E_L needs some thought. It is, however, a standard problem from electrostatics in a slightly different form.

- In the standard problem one calculates the field in a materials with a **DK** given by ϵ_r that does *not* fill a rectangular capacitor totally, but is in the shape of an ellipsoid including the extreme cases of a *pure sphere*, a *thin plate* or a *thin needle*. The result is always

$$E_{ellipse} = N_P \cdot \frac{P}{\epsilon_r \cdot \epsilon_0}$$

In words: The field inside a dielectric in the shape of an ellipsoid (of any shape whatsoever) that is put between the parallel plates of a typical capacitor arrangement, is whatever it would be if the dielectric fills the space between the plates completely times a *number* N_P, the value of which depends on the geometry.

- N_P is the so-called **depolarization factor**, a *pure number*, that only depends on the shape of the ellipsoid. For the extreme cases of the ellipsoid we have fixed and well-known ***depolarization*** factors:

  - **Thin plate: $N = 1$**
  - **Needle: $N = 0$**
  - **Sphere: $N = 1/3$**

.

Our case consists of having a sphere with $\epsilon_r = 1$. We thus obtain

$$E_L = \frac{P}{3\epsilon_0}$$

We have now all components and obtain

$$E_{loc} = \frac{U}{d} - \frac{P}{\epsilon_0} + \frac{P}{3\epsilon_o}$$

- $U/d - P/\epsilon_0$ is just the field we would use in the Maxwell equations, we call it E_0. It is the *homogeneous field* averaged over the whole volume of the homogeneous material
- The *local field* finally becomes

$$E_{loc} = E_0 + \frac{P}{3\epsilon_o}$$

This seems a bit odd? How can the *local* field be different from the *average* field?

- This is one of the tougher questions one can ask. The answer, not extremely satisfying, comes from the basic fact that *all* dipoles contribute to E_0, whereas for the *local field* you discount the effect of *one* charge - the charge you use for probing the field (the field of which must not be added to the rest!).
- If you feel somewhat uneasy about this, you are perfectly right. What we are excluding here is the action of a charge on itself. While we may do that because that was *one* way of defining electrical fields (the other one is Maxwells equation defining a field as directly resulting from charges), we can not so easily do away with the *energy* contained in the field of a single charge. And if we look at this, the whole theory of electromagnetism blows up! If the charge is a point charge, we get infinite energy, and if it is not a point charge, we get other major contradictions.
- Not that it matters in everyday aspects - it is more like a philosophical aspect. If you want to know more about this, read chapter **28** in the "Feynman lectures, Vol. 2"

But do not get confused now! The relation given above is perfectly valid for everyday circumstances and ordinary matter. Don't worry - be happy that a relatively complex issue has such a simple final formula!

We now can relate the *macroscopic* and *microscopic* parameters. With the old relations and the new equation we have a grand total of:

$$\mu = \alpha \cdot E_{loc}$$

$$P = N \cdot \alpha \cdot E_{loc}$$

$$E_{loc} = E_o + \frac{P}{3\epsilon_o}$$

From this we obtain quite easily

$$P = N \cdot \alpha \cdot \left(E_o + \frac{P}{3\epsilon_o} \right)$$

$$P = \frac{N \cdot \alpha \cdot E_o}{1 - N \cdot \alpha / 3\epsilon_o}$$

- With N = density of dipoles

Using the definition of P

$$P = \epsilon_o \cdot \chi \cdot E = \epsilon_o \cdot (\epsilon_r - 1) \cdot E$$

- and inserting it into the equations above gives as *final result* the connection between the polarizability α (the microscopic quantity) and the relative dielectric constant ϵ_r (the macroscopic quantity):

$$\frac{N \cdot \alpha}{3\,\epsilon_0} = \frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{\chi}{\chi + 3}$$

- This is the **Clausius - Mosotti equation**, it relates the *microscopic* quantity α on the left hand side to the *macroscopic* quantity ϵ_r (or, if you like that better, $\chi = \epsilon_r - 1$) on the right hand side of the equation. This has two far reaching consequences
  - We now can *calculate* (at least in principle) the dielectric constants of all materials, because we know how to calculate α.
  - We have an *instrument* to measure *microscopic* properties like the polarizability α, by measuring *macroscopic* properties like the dielectric constant and converting the numbers with the Clausius-Mosotti equation.
- You must also see this in an historical context: With the Clausius-Mosotti equation the dielectric properties of materials were essentially reduced to known electrical properties. There was nothing mysterious anymore about the relative dielectric constant. The next logical step now would be to apply quantum theory to dielectric properties.

3.2.7 Summary to: Polarization Mechanisms

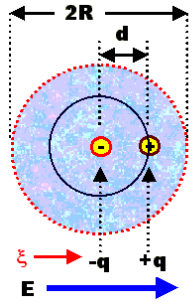
(Dielectric) polarization mechanisms in dielectrics are all mechanisms that

1. Induce dipoles at all (always with $\boldsymbol{\mu}$ in field direction)
   ⇒ Electronic polarization.
2. Induce dipoles already present in the material to "point" to some extent in field direction.
   ⇒ Interface polarization.
   ⇒ Ionic polarization.
   ⇒ Orientation polarization.

Quantitative considerations of polarization mechanisms yield

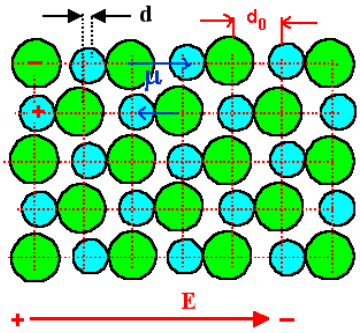
- **Justification (and limits) to the $\underline{P} \propto \underline{E}$ "law"**
- **Values for χ**
- **$\chi = \chi(\omega)$**
- **$\chi = \chi$(structure)**

Electronic polarization describes the separation of the centers of "gravity" of the electron charges in orbitals and the positive charge in the nucleus and the dipoles formed this way. it is always present



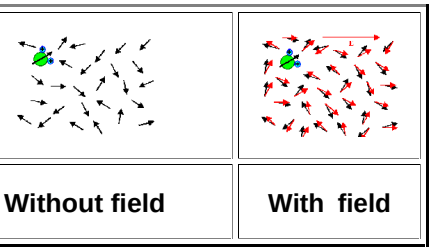
- It is a very weak effect in (more or less isolated) atoms or ions with spherical symmetry (and easily calculated).
- It can be a strong effect in e.g. covalently bonded materials like **Si** (and not so easily calculated) or generally, in solids.

Ionic polarization describes the net effect of changing the distance between neighboring ions in an ionic crystal like **NaCl** (or in crystals with some ionic component like **SiO_2**) by the electric field



- Polarization is linked to bonding strength, i.e. Young's modulus **Y**. The effect is smaller for "stiff" materials, i.e. **$P \propto 1/Y$**

Orientation polarization results from minimizing the free enthalpy of an ensemble of (molecular) dipoles that can move and rotate freely, i.e. polar liquids.



- It is possible to calculate the effect, the result invokes the Langevin function

$$L(\beta) = \coth(\beta) - \frac{1}{\beta}$$

- In a good approximation the polarization is given by ⇒

$$\langle P \rangle = \frac{N \cdot \mu^2 \cdot E}{3kT}$$

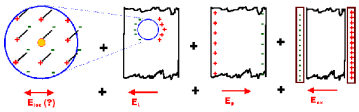
The induced dipole moment **μ** in all mechanisms is proportional to the field (for reasonable field strengths) at the location of the atoms / molecules considered.

$$\underline{\mu} = \alpha \cdot E_{loc}$$

- The proportionality constant is called polarizability α; it is a microscopic quantity describing what atoms or molecules "do" in a field.

- The local field, however, is not identical to the macroscopic or external field, but can be obtained from this by the Lorentz approach
- For isotropic materials (e.g. cubic crystals) one obtains

$$E_L = \frac{P}{3\epsilon_0}$$



$$E_{loc} = E_{ex} + E_{pol} + E_L + E_{near}$$

Knowing the local field, it is now possible to relate the microscopic quantity α to the macroscopic quantity ϵ or ϵ_r via the Clausius - Mosotti equations ⇒

- While this is not overly important in the engineering practice, it is a momentous achievement. With the Clausius - Mosotti equations and what went into them, it was possible for the first time to understand most electronic and optical properties of dielectrics in terms of their constituents (= atoms) and their structure (bonding, crystal lattices etc.)
- Quite a bit of the formalism used can be carried over to other systems with dipoles involved, in particular magnetism = behavior of magnetic dipoles in magnetic fields.

$$\frac{N \cdot \alpha}{3\,\epsilon_0} = \frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{X}{X + 3}$$

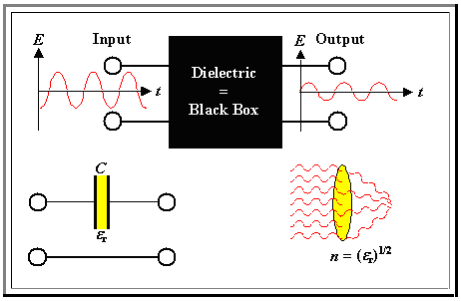
Questionaire

Multiple Choice questions to all of 3.1

3.3 Frequency Dependence of the Dielectric Constant

3.3.1 General Remarks

- All polarization mechanisms respond to an electrical field by shifting masses around. This means that masses must be accelerated and de-accelerated, and this will always take some time. So we *must* expect that the (mechanical) response to a field will depend on the *frequency* ν of the electrical field; on how often per second it changes its sign.
  - If the frequency is very large, no mechanical system will be able to follow. We thus expect that at *very large frequencies all polarization mechanisms will "die out"*, i.e. there is no response to an extremely high frequency field. This means that the dielectric constant ϵ_r will approach **1** for $\nu \Rightarrow \infty$.
- It is best to consider our dielectric now as a "**black box**". A signal in the form of an alternating electrical field ***E*** goes in at the input, and something comes out at the output, as shown below. Besides the Black Box scheme, two possible real expressions of such an abstract system are shown: A parallel-plate capacitor containing a dielectric, and an optical lens with an index of refraction $n = \epsilon_r$. The input would be a simple alternating voltage in the capacitor case, and a light wave in the lens case.



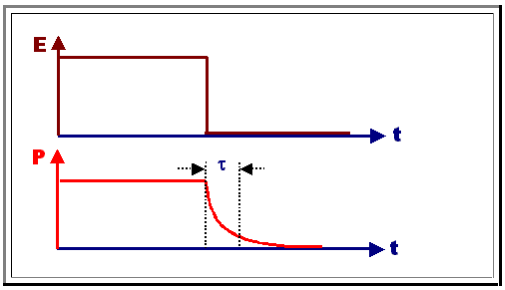
- As long as our system is linear ("twice the input $\Rightarrow$ twice the output), a sinewave going in will produce a sinewave coming out, i.e. the frequency does not change.
  - If a sinewave goes in, the output then can only be a sinewave with an amplitude and a phase different from the input, as schematically shown above.
  - If a complicated signal goes in, we decompose it into its Fourier components, consider the output for all frequencies separately, and then do a Fourier synthesis.
- With complex notation, the input will be something like $E_{in} = E_{in} \cdot \exp(i\omega t)$; the output then will be $E_{out} = E_{out} \cdot \exp i(\omega t + \phi)$.
  - We just as well could write $E_{out} = f(\omega) \cdot E_{in}$ with $f(\omega) =$ complex number for a given ω or complex function of ω.
  - $f(\omega)$ is what we are after. We call this function that relates the output of a dielectric material to its input the **dielectric function** of the material. As we will see, the dielectric function is a well-defined and very powerful entity for any material - even if we cannot calculate it from scratch. We can however, calculate dielectric functions for some model materials, and that will give us a very good idea of what it is all about.
- Since the **index of refraction** ***n*** is directly given by $\epsilon_r^{1/2}$ (assuming that the material has no magnetic properties), we have a *first very general statement*:
  - There exist no microscopes with "optical" lenses for very high frequencies of electrical fields, which means electromagnetic radiation in the *deep ultraviolet* or soft **X-rays**. And indeed, there are no **X-ray microscopes** with lenses[1)] (however, we still have mirrors!) because there are no materials with $\epsilon_r > 1$ for the frequencies of **X**-rays.
- Looking at the polarization mechanisms discussed, we see that there is a fundamental difference in the *dynamics* of the mechanisms with regard to the response to changing forces:
  - In *two* cases (*electron and ionic polarization*), the electrical field will try to change the distance between the charges involved. In response, there is a restoring force that is (in our approximation) directly proportional to the separation distance of the dipole charges. We have, in mechanical terms, an *oscillator*.
  - The characteristic property of *any* such oscillating system is the phenomena of **resonance** at a specific frequency.
  - In the case of the *orientation polarization*, there is no direct mechanical force that "pulls" the dipoles back to random orientation. Instead we have many statistical events, that respond in their *average results* to the driving forces of electrical fields.
  - In other words, if a driving force is present, there is an equilibrium state with an (average) net dipole moment. If the driving force were to disappear suddenly, the ensemble of dipoles will assume a new equilibrium state (random distribution of the dipoles) within some characteristic time called **relaxation time**. The process knows no resonance phenomena, it is characterized by its **relaxation time** instead of a resonance frequency.
- We thus have to consider just the two basic situations: **Dipole relaxation** and **dipole resonance**. Every specific mechanism in real materials will fit one of the two cases.

[1)] Well, never say never. Lenses for X-rays *do* exist for a few years by now. However, if you would see the contraption, you most likely wouldn't recognize it as a *lens*. If you want to know more, turn to the resarch of Prof. Lengeler and his group: http://2b.physik.rtwh-aachen.de

3.3.2 Dipole Relaxation

From Time Dependence to Frequency Dependence

- The easiest way to look at *relaxation phenomena* is to consider what happens if the driving force - the electrical field in our case - is suddenly switched off, after it has been constant for a sufficiently long time so that an equilibrium distribution of dipoles could be obtained.
  - We expect then that the dipoles will randomize, i.e. their dipole moment or their polarization will go to zero.
  - However, that cannot happen instantaneously. A specific dipole will have a certain orientation at the time the field will be switched off, and it will change that orientation only by some interaction with other dipoles (or, in a solid, with phonons), in other words upon collisions or other "violent" encounters. It will take a characteristic time, roughly the time between collisions, before the dipole moment will have disappeared.
  - Since we are discussing statistical events in this case, the individual characteristic time for a given dipole will be small for some, and large for others. But there will be an *average* value which we will call the **relaxation time** τ of the system. We thus expect a smooth change over from the polarization with field to zero within the relaxation time τ, or a behavior as shown below



- In formulas, we expect that ***P*** decays starting at the time of the switch-off according to

$$P(t) = P_0 \cdot \exp - \frac{t}{\tau}$$

- This simple equation describes the behavior of a simple system like our "ideal" dipoles very well. It is, however, not easy to derive from first principles, because we would have to look at the development of an ensemble of interacting particles in time, a classical task of non-classical, i.e. statistical mechanics, but beyond our ***ken*** at this point.
  - Nevertheless, we know that a relation like that comes up whenever we look at the decay of some ensemble of particles or objects, where some have more (or less) energy than required by equilibrium conditions, and the change-over from the excited state to the base state needs "help", i.e. has to overcome some energy barrier.
  - All we have to assume is that the number of particles or objects decaying from the excited to the base state is proportional to the number of excited objects. In other words, we have a relation as follows:

$$\frac{dn}{dt} \propto n = -\frac{1}{\tau} \cdot n$$

$$n = n_0 \cdot \exp - \frac{t}{\tau}$$

  - This covers for example radioactive decay, cooling of any material, and the decay of the foam or froth on top of your ***beer***: Bubbles are an energetically excited state of beer because of the additional surface energy as compared to a droplet. If you measure the height of the head on your beer as a function of time, you will find the exponential law.
- When we turn on an electrical field, our dipole system with random distribution of orientations has too much energy relative to what it could have for a better orientation distribution.
  - The "decay" to the lower (free) energy state and the concomitant built-up of polarization when we switch on the field, will follow our universal law from above, and so will the decay of the polarization when we turn it off.

We are, however, not so interested in the *time* dependence ***P*(*t*)** of the polarization when we apply some disturbance or *input* to the system (the switching on or off of the electrical field). We rather would like to know its *frequency* dependence ***P*(ω)** with **ω = 2πν =** angular frequency, i.e. the output to a periodic harmonic input, i.e. to a field like $E = E_0 \cdot \sin\omega t$.

- Since *any* signal can be expressed as a **Fourier series or Fourier integral** of sin functions as the one above, by knowing ***P*(ω)** we can express the response to *any* signal just as well.

In other words: We can switch back and forth between ***P*(τ)** and ***P*(ω)** via a **Fourier transformation**.

- We already know the time dependence ***P*(τ)** for a switch-on / switch-off signal, and from that we can - in principle - derive ***P*(ω)**.
- We thus have to consider the **Fourier transform** of ***P*(*t*)**. However, while clear in principle, details can become nasty. While some details are given in an advanced module, here it must suffice to say that our Fouriertransform is given by

$$P(\omega) = \int_0^\infty P_0 \cdot \exp -\frac{t}{\tau} \cdot \exp -(i\omega t) \cdot dt$$

- P_0 is the static polarization, i.e. the value of ***P*(ω)** for **ω = 0 Hz** , and $i = (-1)^{1/2}$ is the imaginary unit (note that in electrical engineering usually the symbol **j** is used instead of **i**).

This is an easy integral, we obtain

$$P(\omega) = \frac{P_0}{\omega_0 + i \cdot \omega}$$

$$\omega_0 = \frac{1}{\tau}$$

- Note that ω_0 is *not* $2\pi/\tau$, as usual, but just $1/\tau$. That does not mean anything except that it makes writing the formulas somewhat easier.

The ***P*(ω)** then are the *Fourier coefficients* if you describe the ***P*(*t*)** curve by a Fourier integral (or series, if you like that better, with infinitesimally closely spaced frequency intervals).

- ***P*(ω)** thus is the polarization response of the system if you jiggle it with an electrical field given by $E = E_0 \cdot \exp(i\omega t)$ that contains just one frequency ω.
- However, our Fourier coefficients are *complex numbers*, and we have to discuss what that means now.

Usimg Complex Numbers and Functions

Using the powerful math of complex numbers, we end up with a *complex* polarization. That need not bother us since by convention we would only consider the *real part* of ***P*** when we are in need of real numbers.

- Essentially, we are done. If we know the Amplitude (**= E_0**) and (circle) frequency ω of the electrical field in the material (taking into account possible "local field" effects), we know the polarization.

However, there is a smarter way to describe that relationship than the equation above, with the added benefit that this "smart" way can be generalized to all frequency dependent polarization phenomena. Let's see how it is done:

What we want to do, is to keep our basic equation that couples polarization and field strength for alternating fields, too. This requires that the susceptibility χ becomes frequency dependent. We then have

$$P(\omega) = \epsilon_0 \cdot \chi(\omega) \cdot E(\omega)$$

- and the decisive factor, giving the *amplitude* of ***P*(ω)**, is **χ(ω)**.
- The time dependence of ***P*(ω)** is trivial. It is either given by **exp i(ω*t* – ϕ)**, with ϕ accounting for a possible phase shift, or simply by **exp i(ω*t*)** if we include the phase shift in **χ(ω)**, which means it must be complex.
- The second possibility is more powerful, so that is what we will do. If we then move from the polarization ***P*** to the more conventional electrical displacement ***D***; the relation between ***D*(ω)** and ***E*(ω)** will require a *complex dielectric function* instead of a complex susceptibiltiy, and that is the quantity we will be after from now on.

- It goes without saying that for more complex time dependencies of the electrical field, the equation above holds for every for every **sin** component of the Fourier series of an arbitrary periodic function.

Extracting a *frequency dependent susceptibility* $\chi(\omega)$ from our equation for the polarization is fairly easy: Using the basic equation we have

$$\epsilon_0 \cdot \chi(\omega) = \frac{P(\omega)}{E(\omega)} = \frac{P_0}{E_0} \cdot \frac{1}{\omega_0 + i \cdot \omega} = \chi_s \cdot \frac{1}{1 + i \cdot \omega/\omega_0}$$

- $\chi_s = P_0 / E_0$ is the *static susceptibility*, i.e. the value for zero frequency.

Presently, we are only interested in the *real* part of the complex susceptibility thus obtained. As any complex number, we can decompose $\chi(\omega)$ in a real and a imaginary part, i.e. write it as

$$\chi(\omega) = \chi'(\omega) + i \cdot \chi''(\omega)$$

- with χ' and χ'' being the real and the imaginary part of the complex susceptibility χ . *We drop the* (ω) *by now, because whenever we discuss real and imaginary parts it is clear that we discuss frequency dependence).*
- All we have to do in order to obtain χ' and χ'' is to expand the fraction by $1 - i \cdot \omega/\omega_0$ which gives us

$$\epsilon_0 \cdot \chi(\omega) = \frac{\chi_s}{1 + (\omega/\omega_0)^2} - i \cdot \frac{\chi_s \cdot (\omega/\omega_0)}{1 + (\omega/\omega_0)^2}$$

- We thus have for the real and imaginary part of $\epsilon_0 \cdot \chi(\omega)$, which is almost, but not yet quite the *dielectric function* that we are trying to establish:

$$\epsilon_0 \cdot \chi' = \frac{\chi_s}{1 + (\omega/\omega_0)^2}$$

$$-\epsilon_0 \cdot \chi'' = \frac{\chi_s \cdot (\omega/\omega_0)}{1 + (\omega/\omega_0)^2}$$

This is pretty good, because, as we will see, the real *and* imaginary part of the complex susceptibility contain an unexpected wealth of material properties. Not only the dielectric behavior, but also (almost) all optical properties and essentially also the conductivity of non-perfect dielectrics.

Before we proceed to the *dielectric function* which is what we really want to obtain, we have to makes things a tiny bit more complicated - in three easy steps.

- **1.** People in general like the dielectric constant ϵ_r as a material parameter far better than the susceptibility χ - history just cannot be ignored, even in physics. Everything we did above for the polarization ***P***, we could also have done for the dielectric flux density ***D*** - just replace the letter "***P***" by "***D***" and "χ" by "ϵ_r" and we obtain a complex frequency dependent dielectric constant $\epsilon_r(\omega) = \chi(\omega) + 1$ with, of course, ϵ_s instead of χ_s as the zero frequency static case.
- **2.** So far we assumed that at very large frequencies the polarization is essentially zero - the dipole cannot follow and $\chi(\omega \rightarrow \infty) = 0$That is not necessarily true in the most general case - there might be, after all, other mechanisms that still "work" at frequencies far larger than what orientation polarization can take. If we take that into account, we have to change our consideration of relaxation somewhat and introduce the new, but simple parameter $\chi(\omega >> \omega_0) = \chi_\infty$ or, as we prefer, the same thing for the dielectric "constant", i.e. we introduce $\epsilon_r(\omega >> \omega_0) = \epsilon_\infty$.
- **3.** Since we always have either $\epsilon_0 \cdot \chi(\omega)$ or $\epsilon_0 \cdot \epsilon(\omega)$, and the ϵ_0 is becoming cumbersome, we may just include it in what we now call the **dielectric function** $\epsilon(\omega)$ of the material. This simply means that all the ϵ_i are what they are as the relative dielectric "constant" and multiplied with ϵ_0

This reasoning follows **Debye**, who by doing this expanded our knowledge of materials in a major way. Going through the points **1. - 3.** (which we will not do here), produces the final result for the frequency dependence of the *orientation polarization*, the so-called **Debye equations**:

- In general notation we have pretty much the same equation as for the susceptibility χ; the only real difference is the introduction of ϵ_∞ for the high frequency limit:

$$D(\omega) = \epsilon(\omega) \cdot E(\omega) = \left(\frac{\epsilon_S - \epsilon_\infty}{1 + i\,(\omega/\omega_0)} + \epsilon_\infty \right) \cdot E(\omega)$$

- The complex function $\epsilon(\omega)$ is the **dielectric function**. In the equation above it is given in a closed form for the dipole relaxation mechanism.

Again, we write the complex function as a sum of a real part and a complex part, i.e. as $\epsilon(\omega) = \epsilon'(\omega) - i \cdot \epsilon''(\omega)$. We use a "–" sign, as a matter of taste; it makes some follow-up equations easier. But you may just as well define it with a + sign and in some books that is what you will find. For the dielectric function from above we now obtain

$$\epsilon' = \epsilon_\infty + \frac{\epsilon_S - \epsilon_\infty}{1 + (\omega/\omega_0)^2}$$

$$\epsilon'' = \frac{(\omega/\omega_0)(\epsilon_S - \epsilon_\infty)}{1 + (\omega/\omega_0)^2}$$

- As it *must* be, we have

$$\epsilon'(\omega = 0) = \epsilon_S \qquad \epsilon''(\omega = 0) = 0$$

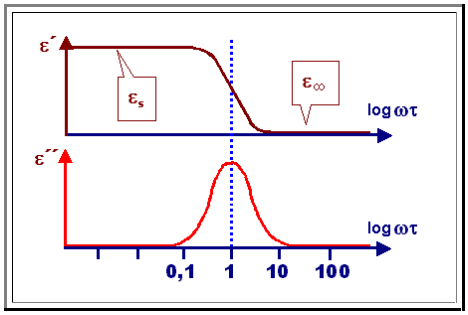
$$\epsilon'(\omega \to \infty) = \epsilon_\infty$$

From working with the complex notation for sin- and cosin-functions we also know that

- ϵ', the *real part* of a complex amplitude, gives the amplitude of the response that is *in phase* with the driving force, ϵ'', the *imaginary part*, gives the amplitude of the response that is *phase-shifted by* 90^o.

Finally, we can ask ourselves: What does it look like? What are the graphs of ϵ' and ϵ''?

- Relatively simple curves, actually, They always look like the graphs shown below, the three numbers that define a particular material (ϵ_S, ϵ_∞, and $\tau = 2\pi/\omega_0$) only change the numbers on the scales.



- Note that ω for curves like this one is always on a *logarithmic scale!*

What the dielectric function for orientation polarization looks like for real systems can be tried out with the JAVA applet below - compare that with the measured curves for water. We have a theory for the frequency dependence which is *pretty good*!

Since ϵ_∞ must be **= 1** (or some value determined by some *other* mechanism that also exists) if we go to frequencies high enough, the essential parameters that characterize a material with orientation polarization are ϵ_S and τ (or ω_0).

- ϵ_S we can get from the polarization mechanism for the materials being considered. If we know the dipole moments of the particles and their density, the Langevin function gives the (static) polarization and thus ϵ_S.
- We will *not*, however, obtain τ from the theory of the polarization considered so far. Here we have to know more about the system; for liquids, e.g., the mean time before two dipoles collide and "loose" all their memory about their previous orientation. This will be expressed in some kind of diffusion terminology, and we have to know something about the random walk of the dipoles in the liquid. This, however, will go far beyond the scope of this course.

- Suffice it to say that typical relaxation times are around **10^{-11} s**; this corresponds to *frequencies in the **GHz** range* , i.e. "**cm** -waves". We must therefore expect that typical materials exhibiting orientation polarization (e.g. water), will show some peculiar behavior in the microwave range of the electromagnetic spectrum.
  - In mixtures of materials, or in complicated materials with several different dipoles and several different relaxation times, things get more complicated. The smooth curves shown above may be no longer smooth, because they now result from a superposition of several smooth curves.
  - Finally, it is also clear that τ may vary quite a bit, depending on the material and the temperature. If heavy atoms are involved, τ tends to be larger and ***vice versa***. If movements speed up because of temperature, τ will get smaller.

3.3.3 Resonance for Ionic and Atomic Polarization

The frequency dependence of the electronic and ionic polarization mechanisms are mathematically identical - we have a driven oscillating system with a linear force law and some damping. In the simple classical approximation used so far, we may use the universal equation describing an oscillating system driven by a force with a **sin(ω*t*)** time dependence

$$m \cdot \frac{d^2x}{dt^2} + k_F \cdot m \cdot \frac{d\,x}{d\,t} + k_S \cdot x = q \cdot E_0 \cdot \exp(i\,\omega\,t)$$

- With ***m*** = mass, k_F = **friction** coefficient; describing damping, k_S = "spring" coefficient or constant; describing the restoring force, $q \cdot E_0$ = amplitude times charge to give a force , **E = E_0· exp (iω *t*)** is the time dependence of electrical field in complex notation.
- This is of course a gross simplification: In the equation above we look at *one* mass ***m*** hooked up to *one* spring, whereas a crystal consists of a hell of a lot of masses (= atoms), all coupled by plenty of springs (= bonds). Nevertheless, the analysis of just one oscillating mass provides the basically correct answer to our quest for the frequency dependence of the ionic and atomic polarization. More to that in link.

We know the *"spring" coefficient* for the electronic and ionic polarization mechanism; however, we do not know from our simple consideration of these two mechanisms the *"friction" term*.

- So lets just consider the general solution to the differential equation given above in terms of the general constants k_S and k_F and see what kind of general conclusions we can draw.

From classical mechanics we know that the system has a resonance frequency ω_0, the frequency with the maximum amplitude of the oscillation, that is (for undamped oscillators) always given by

$$\omega_0 = \left(\frac{k_S}{m}\right)^{1/2}$$

- The general solution of the differential equation is

$$x(\omega, t) = x(\omega) \cdot \exp(i\omega t + \phi)$$

- The angle ϕ is necessary because there might be some phase shift. This phase shift, however, is automatically taken care of if we use a complex amplitude. The complex ***x*(ω)** is given by

$$x(\omega) = \frac{q \cdot E_0}{m}\left(\left(\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + k_F^2\,\omega^2}\right) - i \cdot \left(\frac{k_F\,\omega}{(\omega_0^2 - \omega^2)^2 + k_F^2\,\omega^2}\right)\right)$$

- ***x*(ω)** indeed is a *complex function*, which means that the amplitude is not in phase with the driving force if the imaginary part is not zero.

Again, we are interested in a relation between the **sin** components of the polarization ***P*(ω)** and the **sin** components of the driving field $E = E_0 \cdot \exp(i\omega t)$ or the dielectric flux ***D*(ω)** and the field. We have

$$P = N \cdot q \cdot x(\omega)$$

$$D = \epsilon_0 \cdot \epsilon_r \cdot E = \epsilon_0 \cdot E + P = \epsilon_0 \cdot E + N \cdot q \cdot x(\omega)$$

If we insert ***x*(ω)** from the solution given above, we obtain a complex relationship between ***D*** and ***E***

$$D = \left(\epsilon_0 + \frac{N \cdot q^2}{m}\left(\left(\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + k_F^2\,\omega^2}\right) - i\left(\frac{k_F\,\omega}{(\omega_0^2 - \omega^2)^2 + k_F^2\,\omega^2}\right)\right)\right) \cdot E$$

- This looks pretty awful, but it encodes basic everyday knowledge!

This equation can be rewritten using the **dielectric function** defined before with the added generalization that we now define it for the permittivity, i.e, for

$$\epsilon(\omega) = \epsilon_r(\omega) \cdot \epsilon_0 = \epsilon'(\omega) - i \cdot \epsilon''(\omega)$$

- For the dielectric flux $\boldsymbol{D}$, which we prefer in this case to the polarization $\boldsymbol{P}$, we have as always

$$D(\omega, t) = [\epsilon'(\omega) - i \cdot \epsilon''(\omega)] \cdot E_0 \cdot \exp(i\omega t)$$

- The time dependence of $\boldsymbol{D}$ is simple given by $\exp(i\omega t)$, so the interesting part is only the ω - dependent factor.

Rewriting the equations for the real and imaginary part of ϵ we obtain the *general dielectric function for resonant polarization mechanisms*:

$$\epsilon' = \epsilon_0 + \frac{N \cdot q^2}{m} \left(\frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + k_F^2 \cdot \omega^2} \right)$$

$$\epsilon'' = \frac{N \cdot q^2}{m} \left(\frac{k_F \cdot \omega}{(\omega_0^2 - \omega^2)^2 + k_F^2 \cdot \omega^2} \right)$$

These formula describe the frequency dependence of the dielectric constant of *any* material where the polarization mechanism is given by separating charges with mass $\boldsymbol{m}$ by an electrical field against a *linear* restoring force.

- For the *limiting cases* we obtain for the real and imaginary part

$$\epsilon'(\omega = 0) = \left(\epsilon_0 + \frac{N \cdot q^2}{m} \right) \frac{1}{\omega_0^2} = \left(\epsilon_0 + \frac{N \cdot q^2}{m} \right) \frac{m}{k_S}$$

$$\epsilon'(\omega = \infty) = \epsilon_0$$

- For $\epsilon'(\omega = \infty)$ we thus have $\epsilon_r = \epsilon'/\epsilon_0 = 1$ as must be.

The most important material parameters for dielectric constants at the low frequency limit, i.e. $\omega \Rightarrow 0$, are therefore the *masses* $\boldsymbol{m}$ of the oscillating charges, their *"spring" constants* k_S, their *density* $\boldsymbol{N}$, and the *charge* $\boldsymbol{q}$ on the ion considered.

- We have no big problem with these parameters, and that includes the *"spring" constants*. It is a direct property of the bonding situation and in principle we know how to calculate its value.
- The friction constant k_F does not appear in the limit values of ϵ. As we will see below, it is only important for frequencies around the resonance frequency.

For this intermediate case k_F is the difficult parameter. On the atomic level, "friction" in a classical sense is *not defined*, instead we have to resort to *energy dispersion mechanisms*. While it is easy to see how this works, it is difficult to calculate numbers for k_F.

- Imagine a single oscillating dipole in an ionic crystal. Since the vibrating ions are coupled to their neighbours via binding forces, they will induce this atoms to vibrate, too - in time the whole crystal vibrates. The ordered energy originally contained in the vibration of *one* dipole (ordered, because it vibrated in field direction) is now dispersed as *unordered* thermal energy throughout the crystal.
- Since the energy contained in the original vibration is constant, the net effect on the single oscillating dipole is that of *damping* because its original energy is now spread out over many atoms. Formally, damping or energy dispersion can be described by some fictional "friction" force.
- Keeping that in mind it is easy to see that all mechanisms, especially interaction with phonons, that convert the energy in an *ordered vibration in field direction* to *unordered thermal energy* always appears as a kind of friction force on a particular oscillator. Putting a number on this fictional friction force, however, is clearly a different (and difficult) business.
- However, as soon as you realize that the dimension of k_F is **1/s** and that $1/k_F$ simply is about the time that it takes for an oscillation to "die", you can start to have some ideas - or you check the link.

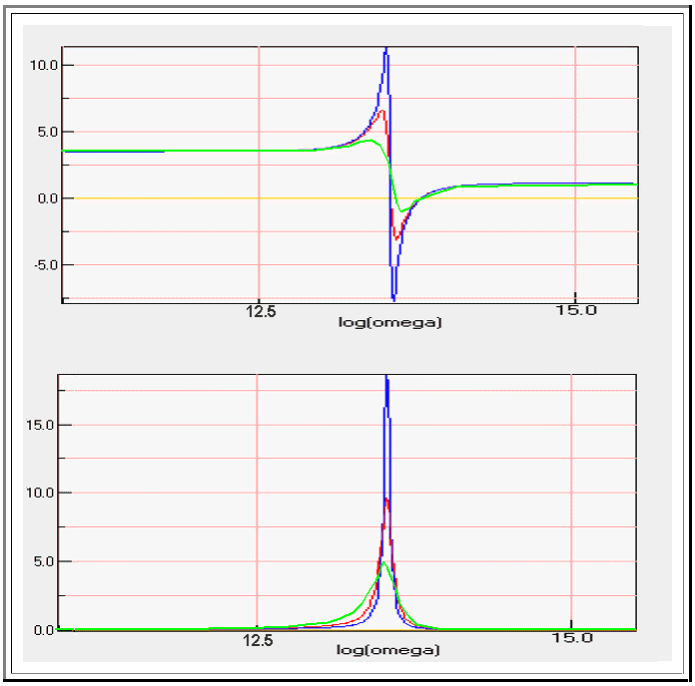
- Now lets look at some characteristic behavior and some numbers as far as we can derive them in full generality.
  - For the electronic polarization mechanism, we know the force constant, it is

$$k_S = \frac{(ze)^2}{4\pi \cdot \epsilon_0 \cdot R^3}$$

  - With the proper numbers for a hydrogen atom we obtain

$$\omega_0 \approx 5 \cdot 10^{16}\ \text{Hz}$$

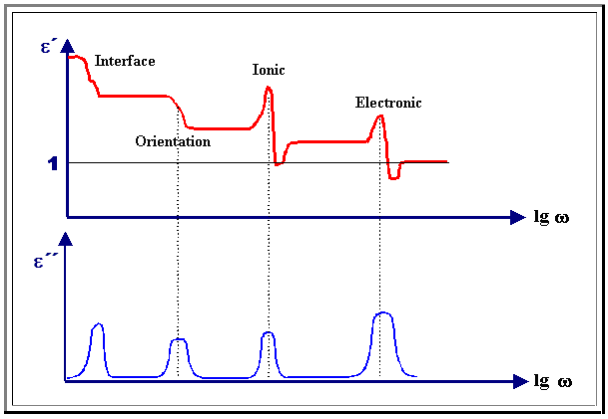
  - This is in the *ultraviolet region of electromagnetic radiation*. For all other materials we would expect similar values because the larger force constants ($(ze)^2$ overcompensates the increasing size R) is balanced to some extent by the larger mass.
  - For the *ionic polarization mechanism*, the masses are several thousand times higher, the resonance frequency thus will be considerably lower. It is, of course simply the frequency of the general lattice vibrations which, as we know, is in the 10^{13} Hz range
- This has an important consequence:
  - The dielectric constant at frequencies higher than about the frequency corresponding to the **UV** part of the spectrum is always **1**. And since the optical index of refraction n is directly given by the DK ($n = \epsilon^{1/2}$), there are no optical lenses beyond the **UV** part of the spectrum.
  - In other words: You can not built a deep-**UV** or **X**-ray microscope with lenses, nor - unfortunately - lithography machines for chips with smallest dimension below about **0,2 µm**. For the exception to this rule see the footnote from before.
- If we now look at the characteristic behavior of ω **'** and ω **''** we obtain quantitatively the following curves (by using the JAVA module provided for in the link):



  - Note that ω is again on a *logarithmic scale!*
- The colors denote different friction coefficients k_F. If k_F would be zero, the amplitude and therefore ϵ' would be ∞ at the resonance point, and ϵ'' would be zero everywhere, and infinity at the resonance frequency; i.e. ϵ'' is the Delta function.
  - While this can never happen in reality, we still may expect significantly larger ϵ values around the resonance frequency than in any other frequency region.
- That the maximum value of ϵ'' *decreases* with increasing damping might be a bit counter-intuitive at first (in fact it was shown the wrong way in earlier versions of this Hyperscript), but for that it extends over ever increasing regions in frequency.

3.3.4 Complete Frequency Dependence of a Model Material

- The frequency dependence of a given material is superposition of the various mechanisms at work in this material. In the *idealized* case of a model material containing *all four basic mechanisms in their pure form* (a non-existent material in the real world), we would expect the following curve.



  - Note that ω is once more on a *logarithmic scale!*

- This is *highly idealized* - there is no material that comes even close! Still, there is a clear structure. Especially there seems to be a correlation between the real and imaginary part of the curve. That is indeed the case; *one* curve contains *all the information* about the other.
  - *Real* dielectric functions usually are only interesting for a small part of the spectrum. They may contain fine structures that reflect the fact that there may be *more than one* mechanism working at the same time, that the oscillating or relaxing particles may have to be treated by *quantum mechanical* methods, that the material is a *mix* of several components, and so on.
  - In the link a real dielectric function for a more complicated molecule is shown. While there is a lot of fine structure, the basic resonance function and the accompanying peak for **ϵ''** is still clearly visible.
- It is a general property of complex functions describing physical reality that under certain very general conditions, the real and imaginary part are directly related. The relation is called **Kramers-Kronig relation**; it is a *mathematical*, not a *physical* property, that only demands two very general conditions to be met:
  - Since two functions with a time or frequency dependence are to be correlated, one of the requirements is *causality*, the other one *linearity*.
  - The Kramers-Kronig relation can be most easily thought of as a *transformation* from one function to another by a black box; the functions being inputs and outputs. *Causality* means that there is no output before an input; *linearity* means that twice the input produces twice the output. Otherwise, the transformation can be anything.
- The Kramers-Kronig relation can be written as follows: For any complex function, e.g. $\epsilon(\omega) = \epsilon'(\omega) + i\epsilon''(\omega)$, we have the relations

$$\epsilon'(\omega) = \frac{-2\omega}{\pi} \int_0^\infty \frac{\omega^* \cdot \epsilon''(\omega^*)}{\omega^{*2} - \omega^2} \cdot d\omega^*$$

$$\epsilon''(\omega) = \frac{2\omega}{\pi} \int_0^\infty \frac{\epsilon'(\omega^*)}{\omega^{*2} - \omega^2} \cdot d\omega^*$$

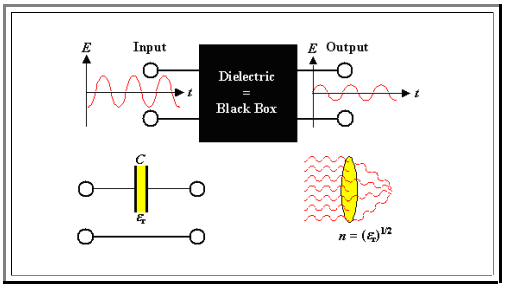
- The Kramers-Kronig relation can be very useful for experimental work. If you want to have the dielectric function of some materials, you only have to measure one component, the other one can be calculated.

Questionaire

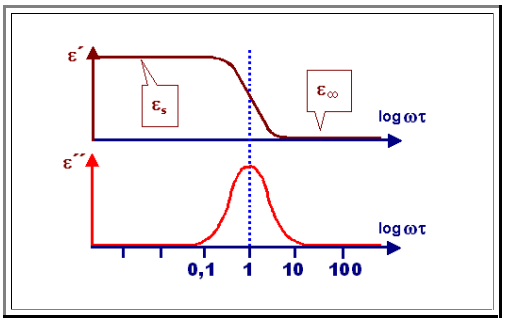
Multiple Choice questions to all of 3.3

3.3.5 Summary to: Frequency Dependence of the Dielectric Constant

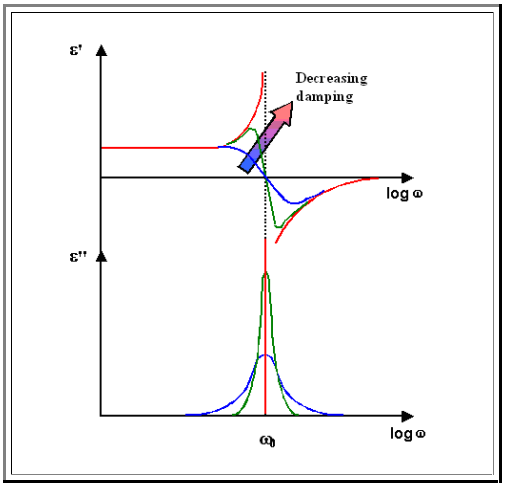
- Alternating electrical fields induce alternating forces for dielectric dipoles. Since in all polarization mechanisms the dipole response to a field involves the movement of masses, inertia will prevent arbitrarily fast movements.
  - Above certain limiting frequencies of the electrical field, the polarization mechanisms will "die out", i.e. not respond to the fields anymore.
  - This might happen at rather high (= optical) frequencies, limiting the index of refraction $n = (\epsilon_r)^{1/2}$



- The (only) two physical mechanisms governing the movement of charged masses experiencing alternating fields are relaxation and resonance.
- *Relaxation* describes the decay of excited states to the ground state; it describes, e.g., what happens for orientation polarization after the field has been switched off.
  - From the "easy to conceive" time behavior we deduce the frequency behavior by a Fourier transformation
  - The dielectric function describing relaxation has a typical frequency dependence in its real and imaginary part ⇒

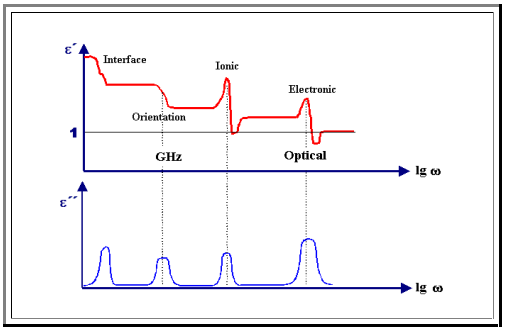


- Resonance describes anything that can be modeled as a mass on a spring - i.e. electronic polarization and ionic polarization.
  - The decisive quantity is the (undamped) resonance frequency $\omega_0 = (k_S/m)^{1/2}$ and the "friction" or damping constant k_F
  - The "spring" constant is directly given by the restoring forces between charges, i.e. Coulombs law, or (same thing) the bonding. In the case of bonding (ionic polarization) the spring constant is also easily expressed in terms of Young's modulus Y. The masses are electron or atom masses for electronic or ionic polarization, respectively.
  - The damping constant describes the time for funneling off ("dispersing") the energy contained in one oscillating mass to the whole crystal lattice. Since this will only take a few oscillations, damping is generally large.
  - The dielectric function describing relaxation has a typical frequency dependence in its real and imaginary part ⇒ The green curve would be about right for crystals.



- The complete frequency dependence of the dielectric behavior of a material, i.e. its dielectric function, contains all mechanisms "operating" in that material.
  - As a rule of thumb, the critical frequencies for relaxation mechanisms are in the **GHz** region, electronic polarization still "works" at optical (10^{15} **Hz**) frequencies (and thus is mainly responsible for the index of refraction).
  - Ionic polarization has resonance frequencies in between.
  - Interface polarization may "die out" already a low frequencies.
- A widely used diagram with all mechanisms shows this, but keep in mind that there is no real material with all 4 major mechanisms strongly present!

⇒



A general mathematical theorem asserts that the real and imaginary part of the dielectric function cannot be completely independent

- If you know the complete frequency dependence of either the real or the imaginary part, you can calculate the complete frequency dependence of the other.
- This is done via the Kramers-Kronig relations; very useful and important equations in material practice.

⇒

$$\epsilon'(\omega) = \frac{-2\omega}{\pi} \int_0^{\infty} \frac{\omega^* \cdot \epsilon''(\omega^*)}{\omega^{*2} - \omega^2} \cdot d\omega^*$$

$$\epsilon''(\omega) = \frac{2\omega}{\pi} \int_0^{\infty} \frac{\epsilon'(\omega^*)}{\omega^{*2} - \omega^2} \cdot d\omega^*$$

Questionaire

Multiple Choice questions to all of 3.3

3.4. Dynamic Properties

3.4.1 Dielectric Losses

The electric power (density) L lost per volume unit in any material as heat is *always* given by

$$L = j \cdot E$$

- With j = current density, and E = electrical field strength.

In our ideal dielectrics there is *no direct current*, only **displacement currents** $j(\omega) = dD/dt$ may occur for alternating voltages or electrical fields. We thus have

$$j(\omega) = \frac{dD}{dt} = \epsilon(\omega) \cdot \frac{dE}{dt} = \epsilon(\omega) \cdot \frac{d[E_0 \exp(i\omega t)]}{dt} = \epsilon(\omega) \cdot i \cdot \omega \cdot E_0 \cdot \exp(i\omega t) = \epsilon(\omega) \cdot i \cdot \omega \cdot E(\omega)$$

- (Remember that the dielectric function $\epsilon(\omega)$ includes ϵ_0).

With the dielectric function written out as $\epsilon(\omega) = \epsilon'(\omega) - i \cdot \epsilon''(\omega)$ we obtain

$$j(\omega) = \omega \cdot \epsilon'' \cdot E(\omega) + i \cdot \omega \cdot \epsilon' \cdot E(\omega)$$

real part of $j(\omega)$; in phase	imaginary part of $j(\omega)$ 90^o out of phase

.

That part of the displacement current that is *in phase* with the electrical field is given by ϵ'', the *imaginary* part of the dielectric function; that part that is **90^o** *out of phase* is given by the *real* part of $\epsilon(\omega)$. The power losses thus have two components

Active power[1] .

$$L_A = \text{power really lost, turned into heat} = \omega \cdot |\epsilon''| \cdot E^2$$

Reactive power

$$L_R = \text{power extended and recovered each cycle} = \omega \cdot |\epsilon'| \cdot E^2$$

[1] Other possible expressions are:
actual power, effective power, real power, true power

- Remember that active, or effective, or true power is energy deposited in your system, or, in other words, it is the power that heats *up your material*! The reactive power is just cycling back and forth, so it is not heating up anything or otherwise leaving direct traces of its existence.

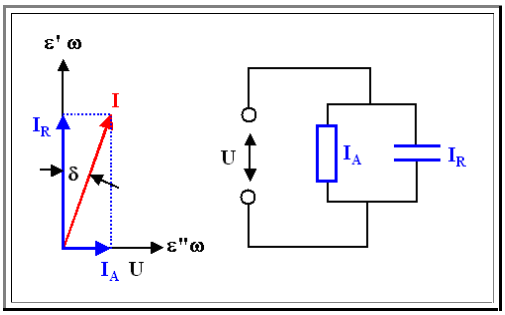
The first important consequence is clear:

- We can heat up even a "perfect" (= perfectly none **DC**-conducting material) by an **AC** voltage; most effectively at frequencies around its resonance or relaxation frequency, when ϵ'' is always maximal.
- Since ϵ'' for the resonance mechanisms is directly proportional to the friction coefficient k_R, the amount of power lost in these cases thus is directly given by the amount of "friction", or power dissipation, which is as it should be.

It is conventional, for reason we will see immediately, to use the quotient of L_A / L_R as a measure of the "quality" of a dielectric: this quotient is called "**tangens delta**" (**tg** δ) and we have

$$\frac{L_A}{L_R} := \text{tg}\,\delta = \frac{I_A}{I_R} = \frac{\epsilon''}{\epsilon'}$$

- Why this somewhat peculiar name was chosen will become clear when we look at a pointer representation of the voltages and currents and its corresponding equivalent circuit. This is a perfectly legal thing to do: We always can represent the current from above this way; in other words we can always model the behaviour of a real dielectric onto an **equivalent circuit diagram** consisting of an *ideal* capacitor with **C(ω)** and an *ideal* resistor with **R(ω)**.



- The current I_A flowing through the *ohmic resistor* of the equivalent circuit diagram is in phase with the voltage ***U***; it corresponds to the imaginary part ϵ'' of the dielectric function times ω.
  - The **90°** out-of-phase current I_R flowing through the *"perfect" capacitor* is given by the real part ϵ' of the dielectric function times ω.
  - The numerical values of both elements must depend on the frequency, of course - for $\omega = 0$, ***R*** would be infinite for an ideal (non-conducting) dielectric.
  - The smaller the angle δ or **tg** δ, the better with respect to power losses.
- Using such an **equivalent circuit diagram** (with always "ideal" elements), we see that a *real* dielectric may be modeled by a fictitious "ideal" dielectric having no losses (something that does not exist!) with an ohmic resistor in parallel that represents the losses. The value of the ohmic resistor (and of the capacitor) must depend on the frequency; but we can easily derive the necessary relations.
  - How large is ***R***, the more interesting quantity, or better, the *conductivity* σ of the material that corresponds to ***R***? Easy, we just have to look at the equation for the current from above.
  - For the in-phase component we simply have

$$\boldsymbol{j}(\omega) = \omega \cdot \epsilon'' \cdot \boldsymbol{E}(\omega)$$

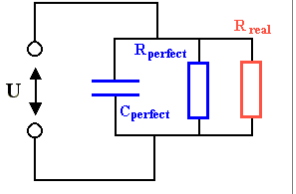
  - Since we *always* can express an in-phase current by the conductivity σ via

$$\boldsymbol{j}(\omega) := \sigma(\omega) \cdot \boldsymbol{E}(\omega)$$

  - we have

$$\sigma_{DK}(\omega) = \omega \cdot \epsilon''(\omega)$$

- In other words: The dielectric losses occuring in a perfect dielectric are completely contained in the imaginary part of the dielectric function and express themselves as if the material would have a frequency dependent conductivity σ_{DK} as given by the formula above.
  - This applies to the case where our dielectric is still a *perfect* insulator at **DC ($\omega = 0$ Hz)**, or, a bit more general, at low frequencies; i.e. for $\sigma_{DK}(\omega \rightarrow 0) = 0$.
- However, nobody is perfect! There is no *perfect* insulator, at best we have *good* insulators. But now it is easy to see what we have to do if a *real* dielectric is *not* a perfect insulator at low frequencies, but has some finite conductivity σ_0 even for $\omega = 0$. Take water with some dissolved salt for a simple and relevant example.
  - In this case we simple *add* σ_0 to σ_{DK} to obtain the total conductivity responsible for power loss



$$\sigma_{total} = \sigma_{perfect} + \sigma_{real}$$

$$= \sigma_{DK} + \sigma_0$$

- Remember: For resistors in parallel, you *add* the conductivities (or **1/R**'s) ; it is with *resistivities* that you do the **$1/R_{total} = 1/R_1 + 1/R_2$** procedure.

Since it is often difficult to separate σ_{DK} and σ_0, it is convenient (if somewhat confusing the issue), to use σ_{total} in the imaginary part of the dielectric function. We have

$$\epsilon'' = \frac{\sigma_{total}}{\omega}$$

- We also have a completely general way now, to describe the response of *any* material to an electrical field, because we now can combine dielectric behavior and conductivity in the complete dielectric function of the material.
- Powerful, but only important at high frequencies; as soon as the imaginary part of the "perfect" dielectric becomes noticeable. But high frequencies is where the action is. As soon as we hit the high **THz** region and beyond, we start to call what we do "*Optics*", or "*Photonics*", but the material roots of those disciplines we have right here.

In classical electrical engineering at not too large frequencies, we are particularly interested in the relative magnitude of both current contributions, i.e in **tgδ**. From the pointer diagram we see directly that we have

$$\frac{I_A}{I_R} = \text{tg}\ \delta$$

We may get an expression for **tg** δ by using for example the Debye equations for **ϵ'** and **ϵ''** derived for the dipole relaxation mechanism:

$$\text{tg}\ \delta = \frac{\epsilon''}{\epsilon'} = \frac{(\epsilon_S - \epsilon_\infty) \cdot \omega / \omega_0}{\epsilon_S + \epsilon_\infty \cdot \omega^2 / \omega_0^2}$$

- or, for the normal case of $\epsilon_\infty = 1$ (or , more correctly ϵ_0)

$$\text{tg}\ \delta = \frac{(\epsilon_S - 1) \cdot \omega / \omega_0}{\epsilon_S + \omega^2 / \omega_0^2}$$

- This is, of course, only applicable to real *perfect* dielectrics, i.e. for real dielectrics with $\sigma_0 = 0$.

The total power loss, the *really interesting quantity*, then becomes (using **ϵ'' = ϵ' · tgδ**, because **tgδ** is now seen as a *material parameter*) .

$$L_A = \omega \cdot \epsilon' \cdot E^2 \cdot \text{tg}\ \delta$$

This is a useful relation for a dielectric with a given **tg** δ (which, for the range of frequencies encountered in "normal" electrical engineering is approximately constant). It not only gives an idea of the electrical losses, but also a very rough estimate of the break-down strength of the material. If the losses are large, it will heat up and this always helps to induce immediate or (much worse) eventual breakdown.

We also can see now what happens if the dielectric is *not ideal* (i.e. totally insulating), but slightly conducting:

- We simply include σ_0 in the definition of **tgδ** (and then automatically in the value of **ϵ''**).
- **tg** δ is then non-zero even for low frequencies - there is a constant loss of power into the dielectric. This may be of some consequence even for small **tg** δ values, as the example will show:

The **tg** δ value for regular (cheap) insulation material as it was obtainable some **20** years ago at very low frequencies (**50 Hz**; essentially **DC**) was about **tg** δ **= 0,01**.

- Using it for a high-voltage line (***U* = 300 kV**) at moderate field strength in the dielectric (***E* = 15MV/m**; corresponding to a thickness of **20 mm**), we have a loss of **14 kW/m^3** of dielectric, which translates into about **800 m** high voltage line. So there is little wonder that high-voltage lines were not insulated by a dielectric, but by air until rather recently!

Finally, some examples for the **tg** δ values for commonly used materials (and low frequencies):

Material	ϵ_r	tg δ x 10^{-4}
Al_2O_3 (very good ceramic)	10	5....20
SiO_2	3,8	2
$BaTiO_3$	500 (!!)	150
Nylon	3,1	10...0,7
Poly...carbonate, ...ethylene ...styrol	about 3	
PVC	3	160

And now you understand how the microwave oven works and why it is essentially heating only the water contained in the food.

Questionaire

Multiple Choice questions to 3.2.1

3.4.2 Summary to: Dynamic Properties - Dielectric Losses

The frequency dependent current density ***j*** flowing through a dielectric is easily obtained. ⇒

$$j(\omega) = \frac{dD}{dt} = \epsilon(\omega) \cdot \frac{dE}{dt} = \omega \cdot \epsilon'' \cdot E(\omega) + i \cdot \omega \cdot \epsilon' \cdot E(\omega)$$

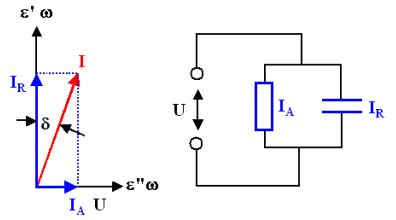
in phase | out of phase

- The in-phase part generates active power and thus heats up the dielectric, the out-of-phase part just produces reactive power
- The power losses caused by a dielectric are thus directly proportional to the imaginary component of the dielectric function

$$L_A = \text{power turned into heat} = \omega \cdot |\epsilon''| \cdot E^2$$

The relation between active and reactive power is called "tangens Delta" (**tg(δ)**); this is clear by looking at the usual pointer diagram of the current

$$\frac{L_A}{L_R} := \text{tg}\,\delta = \frac{I_A}{I_R} = \frac{\epsilon''}{\epsilon'}$$



- The pointer diagram for an *ideal* dielectric **σ(ω = 0) = 0**can always be obtained form an (ideal) resistor ***R*(ω)** in parallel to an (ideal) capacitor ***C*(ω)**.
- ***R*(ω)** expresses the apparent conductivity $\sigma_{DK}(\omega)$ of the dielectric, it follows that

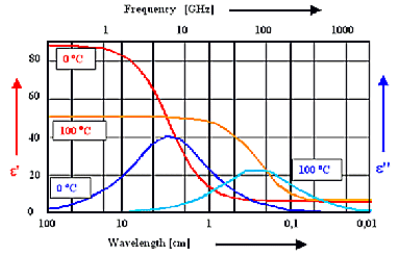
$$\sigma_{DK}(\omega) = \omega \cdot \epsilon''(\omega)$$

For a *real* dielectric with a non-vanishing conductivity at zero (or small) frequencies, we now just add another resistor in parallel. This allows to express *all* conductivity effects of a real dielectric in the imaginary part of its (usually measured) dielectric function via

$$\epsilon'' = \frac{\sigma_{total}}{\omega}$$

- We have no *all* materials covered with respect to their dielectric behavior - in principle even metals, but then resorting to a dielectric function would be overkill.

A good example for using the dielectric function is "dirty" water with a not-too-small (ionic) conductivity, commonly encountered in food.



- The polarization mechanism is orientation polarization, we expect large imaginary parts of the dielectric function in the **GHz** region.
- It follows that food can be heated by microwave (ovens)!

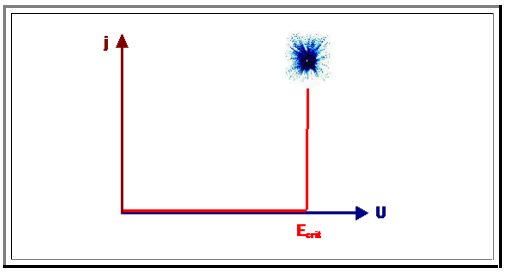
Questionaire

Multiple Choice questions to all of 3.4

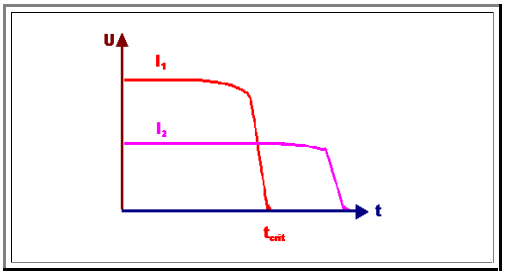
3.5 Electrical Breakdown and Failure

3.5.1 Observation of Electrical Breakdown and Failure

- As you know, the **first law of Materials science** is "Everything can be broken". Dielectrics are no exception to this rule. If you increase the voltage applied to a capacitor, eventually you will produce a big bang and a lot of smoke - the dielectric material inside the capacitor will have experienced "**electrical breakdown**" or electrical break-through, an irreversible and practically always destructive sudden flow of current.
  - The critical parameter is the field strength ***E*** in the dielectric. If it is too large, breakdown occurs. The (***DC***) current vs. field strength characteristic of a dielectric therefore may look look this:



  - After reaching $\boldsymbol{E_{crit}}$, a sudden flow of current may, within very short times ($\mathbf{10^{-8}}$ **s**) completely destroys the dielectric to a smoking hot mass of undefinable structure.
  - Unfortunately, $\boldsymbol{E_{crit}}$ is *not* a well defined material property, it depends on many parameters, the most notable (besides the basic material itself) being the production process, the thickness, the temperature, the internal structure (defects and the like), the age, the environment where it is used (especially humidity) and the time it experienced field stress.
- In the cases where time plays an essential role, the expression "**failure**" is used. Here we have a dielectric being used at nominal field strength well below its breakdown field-strength for some time (usually many years) when it more or less suddenly "goes up in smoke". Obviously the breakdown field strength decreases with operating time - we observe a failure of the material.
  - In this case the breakdown may not be explosive; but a leakage current may develop which grows over time until a sudden increase leads to total failure of the dielectric.
  - The effect can be most easily tested or simulated, by impressing a constant (*very small*) current in the dielectric and monitoring the voltage needed as a function of time. Remember that by definition you cannot have a large current flowing through an insulator = dielectric; but "ein bißchen was geht immer" - a tiny little current is always possible if you have enough voltage at your disposal. A typical voltage-time curve may then look like this:



  - The voltage needed to press your tiny test current through the dielectric starts to decrease rapidly after some time - hours, days, weeks, ..., and this is a clear indication that you dielectric becomes increasingly leaky, and will go up in smoke soon.
  - A typical result is that breakdown of a "good" dielectric occurs after - very roughly - **1 C** of charge has been passed.
- The following table gives a rough idea of critical field strengths for certain dielectric materials

Material	Critical Field Strength [kV/cm]
Oil	200
Glass, ceramics	200...400
Mica	200...700
Oiled paper	1800
Polymers	50...900
SiO_2 in **ICs**	> 10 000

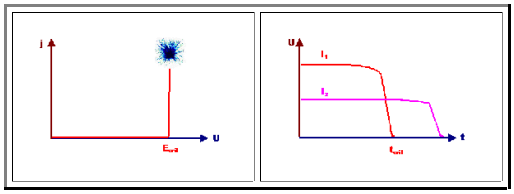
- The last examples serves to remind you that **field strength** is something *totally different from voltage*! Lets look at typical data from an integrated memory circuit, a so- called ***DRAM*** , short for **Dynamic Random Access Memory**. It contains a capacitor as the central storage device (no charge **= 1**; charge **= 0**). This capacitor has the following typical values:
  - *Capacity* ≈ **30 fF** (femtofarad)
    Dielectric: ***ONO***, short for three layers composed of Oxide (**SiO_2**), Nitride (**Si_3N_4**) and Oxide again - together about **8 nm** thick!
    Voltage: **5 V**, and consequently
    Field strength **E = 5/8 V/nm ≈ $6 \cdot 10^6$ V/cm**.
  - This is *far above the critical field strength* for practically all *bulk* materials! We see very graphically that high field strength and voltage have nothing to do with each other. We also see for the first time that materials in the form of a *thin film* may have properties quite different from their bulk behavior - fortunately they are usually much "better".
- Last, lets just note in passing, that electrical breakdown is *not* limited to insulators proper. Devices made from "*bad* " conductors - i.e. semiconductors or ionic conductors - may contain regions completely depleted of mobile carriers - space charge regions at junctions are one example.
  - These insulating regions can only take so much field strength before they break down, and this may severely limit their usage in products

Questionaire

Multiple Choice questions to 3.5.1

3.5.3 Summary to: Electrical Breakdown and Failure

- The first law of materials science obtains: At field strengths larger than some critical value, dielectrics will experience (destructive) electrical breakdown
  - This might happen suddenly (then calls break-down) , with a bang and smoke, or
  - it may take time - months or years - then called failure.
  - Critical field strength may vary from **< 100 kV/cm** to **> 10 MV / cm**.



- Highest field strengths in practical applications do not necessarily occur at high voltages, but e.g. in integrated circuits for very thin (a few **nm**) dielectric layers
  - Properties of thin films may be quite different (better!) than bulk properties!

Example 1: TV set, **20 kV** cable, thickness of insulation = **2 mm.** ⇒ ***E* = 100 kV/cm**
Example 2: Gate dielectric in transistor, **3.3 nm** thick, **3.3 V** operating voltage. ⇒ ***E* = 10 MV/cm**

- Electrical breakdown is a major source for failure of electronic products (i.e. one of the reasons why things go "kaputt" (= broke)), but there is no simple mechanism following some straight-forward theory. We have:
  - *Thermal breakdown*; due to small (field dependent) currents flowing through "weak" parts of the dielectric.
  - *Avalanche breakdown* due to occasional free electrons being accelerated in the field; eventually gaining enough energy to ionize atoms, producing more free electrons in a runaway avalanche.
  - *Local discharge* producing micro-plasmas in small cavities, leading to slow erosion of the material.
  - *Electrolytic breakdown* due to some ionic micro conduction leading to structural changes by, e.g., metal deposition.

Questionaire

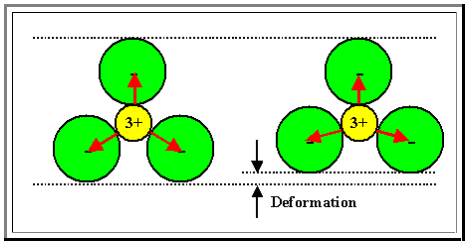
Multiple Choice questions to all of 3.5

3.6 Special Dielectrics

3.6.1 Piezo Electricity and Related Effects

Piezo Electricity

- The polarization of a material must not necessarily be an effect of electrical fields only; it may come about by other means, too.
  - Most prominent is the *inducement of polarization by mechanical deformation*, which is called *piezo electricity*. The reverse mechanism, the inducement of mechanical deformation by polarization, also falls under this heading.
- The principle of piezo electricity is easy to understand:



  - Let'as consider a crystal with ionic components and some arrangement of ions as shown (in parts) in the picture above. In the undeformed symmetrical arrangement, we have three dipole moments (red arrows) that exactly cancel in vector addition.
  - If we induce some elastic deformation as shown, the **symmetry is broken** and the three dipole moments no longer cancel - we have induced polarization by mechanical deformation.
  - We also realize that symmetry is somehow important. If we were to deform the "crystal" in a direction perpendicular to the drawing plane, nothing with respect to polarization would happen. This tells us:
    - Piezo electricity can be pronounced in single crystals if they are deformed in the "right" direction, while it may be absent or weak in polycrystals with randomly oriented grains.
    - Piezo electricity must be described by a tensor of second rank. What this means is that we must consider the full tensor properties of the susceptibility χ or the dielectric constant ϵ_r when dealing with piezoelectricity proper.
  - If one looks more closely at this, it turns out that the crystal symmetry must meet certain conditions. Most important is that it must not have an **inversion center**.
- We won't look into the tensor properties of piezoelectricity but just note that for piezo electric materials we have a general relation between polarization ***P*** and deformation ***e*** of the form

$$P = \text{const.} \cdot e$$

  - With ***e*** = mechanical ***strain*** = $\Delta l / l$ = relative change of length. (Strain is usually written as ϵ; but here we use ***e*** to avoid confusion with the dielectric constant).
- In piezo electric materials, mechanical deformation produced polarization, i.e an electrical field inside the material. The reverse then must be true, too:
  - Piezo electrical materials exposed to an electrical field will experience a force and therfore undergo mechanical deformation, i.e. get somewhat shorter or longer.
- So piezo electricity is restricted to crystals with relatively low symmetry (there must be no center of symmetry; i.e. no inversion symmetry) in single crystalline form (or at least strongly textured poly crystals). While that appears to be a rather limiting conditions, piezo electricity nevertheless has major technical uses:
  - Most prominent, perhaps, are the **quartz oscillators**, where suitable (and small) pieces of single crystals of quartz are given a very precisely and purely mechanically defined resonance frequency (as in tuning forks). Crystalline quartz happens to be strongly piezo electric; if it is polarized by an electrical field of the right frequency, it will vibrate vigorously, otherwise it will not respond. This can be used to control frequencies at a very high level of precision.
  - Probably just as prominent by now, although a rather recent big break-through, are **fuel injectors** for advanced ("common rail") Diesel engines. Makes for more fuel efficient and clean engines and is thus a good thing. The materials of choice for this mass application is ***PZT***, Lead zirconate titanate. This link gives a short description.
  - While for fuel injectors relatively large mechanical displacements are needed, the piezoelectric effect can just as well be used for precisely controlled very small movements in the order of fractions of **nm** to **µm**, as it is, e.g., needed for the scanning tunnelling microscope.

- There are many more applications (consult the links from above), e.g. for
  - Microphones.
  - Ultrasound generators.
  - Surface acoustic wave filters (**SAW**).
  - Sensors (e.g. for pressure or length).

Electrostriction

- An effect that must be kept separate from the piezo electricity is **electrostriction**, where again mechanical deformation leads to polarization.
  - It is an effect observed in many material, but usually much weaker than the piezo electric effect. Much simplified, the effect results if dipoles induced by electronic polarization are not exactly in field direction (e.g. in covalent bonds) and then experience a mechanical force (leading to deformation) that tries to rotate them more into the field direction.
  - The deformation **e** in this case depends on the *square of the electrical field* because the field induces the dipoles *and* acts on them. We have

$$e = \frac{\Delta l}{l} = \text{const} \cdot E^2$$

  - Because of the quadratic dependence, the sign of the field does not matter (in contrast to piezo electricity).
  - There is *no* inverse effect - a deformation does not produce an electric field.
- Electrostriction can be used to produce extremely small deformations in a controlled way; but it is not really much used.

Pyro Electricity

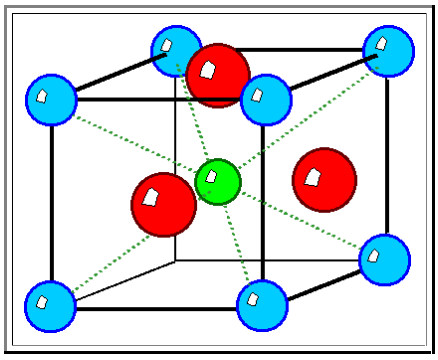
- Polarization can also be induced by sudden changes in the temperature, this effect is called **pyro electricity**; it is most notably found in natural **tourmalin** crystals.
  - The effect comes about because pyro electrical crystals are naturally polarized on surfaces, but this polarization is compensated by mobile ions in a "dirt" skin, so that no net polarization is observed.
  - Changes in temperature change the natural polarization, but because the compensation process may take a rather long time, an outside polarization is observed for some time.

Electrets

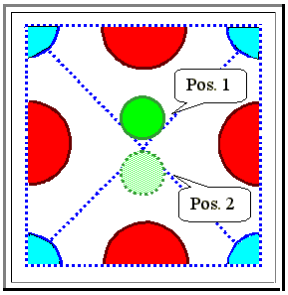
- The word "electret" is a combination of *electr*icity and magn*et* - and that tells it all:
  - Electrets are the electrical analog of (permanent) magnets: Materials that have a permanent macroscopic polarization or a permanent charge. Ferroelectric materials (see next sub-chapter) might be considered to be a sub-species of electrets with a permanent polarization that is "felt" if the internal domains do not cancel each other.
  - Electrets that contain surplus charge that is not easily lost (like the charge on your hair after brushing it on dry days) are mostly polymers, like fluoropolymers or polyproylene.
- Electrets have been a kind of scientific curiosity since the early 18th century (when people did a lot of rubbing things to generate electricity), their name was coined in 1885 by Oliver **Heaviside**
  - Lately, however, they were put to work. Cheap electret microphones are now quite ubiquitous; electrostatic filters and copy machines might employ electrets, too.
  - It is a safe bet that some of the "exotic" materials mentioned in this sub-chapter 3.6 (and some materials not even mentioned or maybe not yet discovered) will be turned into products within *your* career as an engineer, dear student!

3.6.2 Ferro Electricity

- The name, obviously, has nothing to do with "*Ferro*" (= Iron), but associates the analogy to ferro magnetism. It means that in some special materials, the electrical dipoles are not *randomly* distributed, but interact in such a way as to align themselves even *without* an external field.
  - We thus expect *spontaneous polarization* and a *very large* dielectric constant (**DK**).
  - This should be very useful - e.g. for making capacitors - but as in the case of ferro *magnetism*, there are not too many materials showing this behavior.
- The best known material used for many application is **$BaTiO_3$** (**Barium titanate**).
  - It has a simple lattice as far as materials with three different atoms can have a simple lattice at all. The doubly charged **Ba^{2+}** atoms sits on the corners of a cube, the **O^{2-}** ions on the face centers, and the **Ti^{4+}** ion in the center of the cube.
  - We have **8 Ba^{2+}** ions belonging to **1/8** to the elementary cell, **6 O^{2-}** ions belonging to **1/2** to the elementary cell, and one **Ti^{4+}** ion belonging in total to the cell, which gives us the **$BaTiO_3$** stoichiometry.
  - This kind of crystal structure is called a **Perovskite** structure; it is very common in nature and looks like the drawing below (only three of the six oxygen ions are shown for clarity):



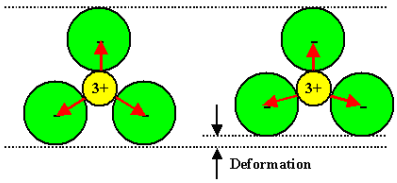
- Often, the lattice is not exactly cubic, but slightly distorted. In the case of **$BaTiO_3$** this is indeed the case: The **Ti** - ion does not sit in the *exact* center of the slightly distorted cube, but slightly off to one side. It thus has *two* symmetrical positions as schematically (and much exaggeratedly) shown below



  - Each elementary cell of **$BaTiO_3$** thus carries a dipole moment, and, what's more important, *the moments of neighbouring cells tend to line up*.
- The interactions between the dipoles that lead to a line-up can only be understood with quantum mechanics. It is not unlike the interactions of *spins* that lead to ferro magnetism.
  - We will not go into details of ferro electricity at this point. Suffice it to say that there are many uses. Traditionally, many capacitors use ferro-electric materials with high **DK** values. In recent years, a large interest in ferro-electrics for uses in integrated circuits has developed; we have yet to see if this will turn into new products.

3.6.3 Summary to: Special Dielectrics

- Polarization ***P*** of a dielectric material can also be induced by mechanical deformation ***e*** or by other means.
  - *Piezo electric materials* are anisotropic crystals meeting certain symmetry conditions like crystalline quartz (**SiO_2**): the effect is linear.
  - The effect also works in reverse: Electrical fields induce mechanical deformation
  - Piezo electric materials have many uses, most prominent are quartz oscillators and, recently, fuel injectors for Diesel engines.

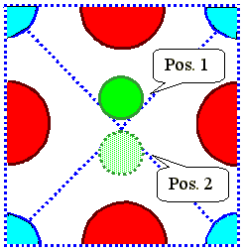


$$P = \text{const.} \cdot e$$

- *Electrostriction* also couples polarization and mechanical deformation, but in a quadratic way and only in the direction "electrical fields induce (very small) deformations".
  - The effect has little uses so far; it can be used to control very small movements, e.g. for manipulations in the **nm** region. Since it is coupled to electronic polarization, many materials show this effect.

$$e = \frac{\Delta l}{l} = \text{const} \cdot E^2$$

- *Ferro electric materials* posses a permanent dipole moment in any elementary cell that, moreover, are all aligned (below a critical temperature).
  - There are strong parallels to ferromagnetic materials (hence the strange name).
  - Ferroelectric materials have large or even very large (ϵ_r **> 1.000**) dielectric constants and thus are to be found inside capacitors with high capacities (but not-so-good high frequency performance)



$BaTiO_3$ unit cell

- *Pyro electricity* couples polarization to temperature changes; *electrets* are materials with permanent polarization, ... There are more "curiosities" along these lines, some of which have been made useful recently, or might be made useful - as material science and engineering progresses.

Questionaire

Multiple Choice questions to all of 3.6

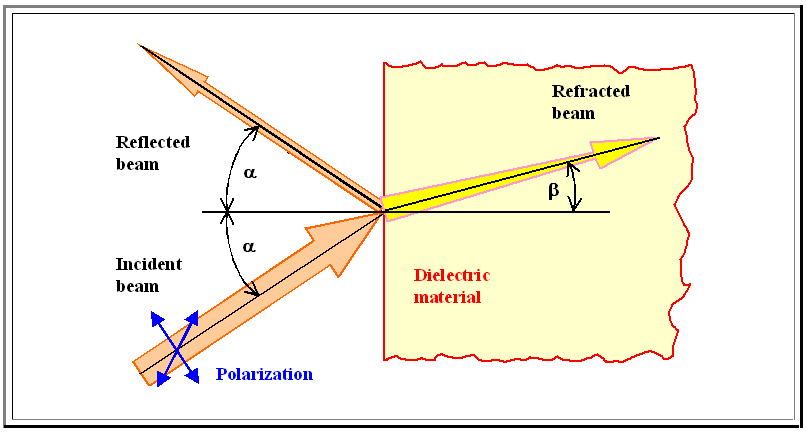
3.7 Dielectrics and Optics

3.7.1 Basics

- This subchapter can easily be turned into a whole lecture course, so it is impossible to derive all the interesting relations *and* to cover anything in depth. This subchapter therefore just tries to give a strong flavor of the topic.
- We know, of course, that the **index of refraction *n*** of a non-magnetic material is linked to the dielectric constant ϵ_r via a simple relation, which is a rather direct result of the Maxwell equations.

$$n = (\epsilon_r)^{1/2}$$

- But in learning about the origin of the dielectric constant, we have progressed from a simple constant ϵ_r to a complex **dielectric function** with frequency dependent real and imaginary parts.
  - What happens to ***n*** then? How do we transfer the wealth of additional information contained in the *dielectric function* to optical properties, which are to a large degreee encoded in the *index of refraction*?
- Well, you probably guessed it: We switch to a **complex index of refraction**!
  - But before we do this, let's ask ourselves what we actually want to find out. What are the optical properties that we like to know and that are *not* contained in a simple index of refraction?
  - Lets look at the *paradigmatic* experiment in optics and see what we *should* know, what we *already* know, and what we *do not* yet know.



- What we have is an electromagnetic wave, an *incident* beam (traveling in vacuum to keep things easy), which impinges on our dielectric material. As a result we obtain a *reflected* beam traveling in vacuum and a *refracted* beam which travels through the material. What do we know about the three beams?
  - The *incident beam* is characterized by its wavelength λ_i, its frequency ν_i and its velocity c_0, the direction of its **polarization** in some coordinate system of our choice, and the arbitrary angle of incidence α. We know, it is hoped, the simple **dispersion relation** for vacuum.

$$c_0 = \nu_i \cdot \lambda_i$$

  - c_0 is, of course, the velocity of light in vacuum, an absoute constant of nature.
- The *incident beam* also has a certain *amplitude* of the electric field (and of the magnetic field, of course) which we call E_0. The *intensity* I_i of the light that the incident beams embodies, i.e. the energy flow, is proportional to E_0^2 - never mix up the two!
- The *reflected beam* follows one of the basic laws of optics, i.e. *angle of incidence* = ***angle of emergence***, and its wavelength, frequency and magnitude of velocity are identical to that of the incident beam.
  - What we do *not know* is its *amplitude* and its *polarization*, and these two quantities must somehow depend on the properties of the incident beam *and* the properties of the dielectric.
- If we now consider the *refracted beam*, we know that it travels under an angle β, has the same frequency as the incident beam, but a wavelength λ_d and a velocity ***c*** that is different from λ_i and c_0.
  - Moreover, we must expect that it is *damped* or *attenuated* , i.e. that its amplitude decreases as a function of penetration depth (this is indicated by decreasing thickness of the arrow above). All parameters of the refracted beam may depend on the polarization of the incident beam.
  - Again, *basic* optics teaches that there are some simple relations. We have

$$\frac{\sin\alpha}{\sin\beta} = n$$ **Snellius law**

$$n = \frac{c_0}{c}$$ **From Maxwell equations**

$$c = \nu_i \cdot \lambda_d$$ **Always valid**

$$\lambda_d = \frac{1}{n} \cdot \lambda_i$$ **From the equations above**

- A bit more involved is another basic relations coming from the Maxwell equations. It is the equation linking ***c***, the speed of light in a material to the material "constants" ϵ_r and the corresponding magnetic permeability μ_0 of vacuum and μ_r of the material via

$$c = \frac{1}{(\mu_0 \cdot \mu_r \cdot \epsilon_0 \cdot \epsilon_r)^{1/2}}$$

- Since most optical materials are not magnetic, i.e. $\mu_r = 1$, we obtain for the index of refraction of a dielectric material our relation from above.

$$n = \frac{c_0}{c} = \frac{(\mu_0 \cdot \mu_r \cdot \epsilon_0 \cdot \epsilon_r)^{1/2}}{(\mu_0 \cdot \epsilon_0)^{1/2}} = \epsilon_r^{1/2}$$

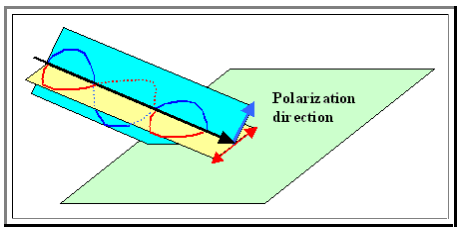
- Consult the basic optics module if you have problems so far.

If we now look at not-so-basic optics, we encounter the **Fresnel laws** of diffraction.

- Essentially, the *Fresnel laws* give the *intensity of the reflected beam* as a function of the *angle of incidence* , the *polarization* of the incident beam, and the *index of refraction* of the material.
- The Fresnel laws are not particularly easy to obtain (consult the basic module Fresnel laws), but the results are easy.

First, we must distinguish between the two basic polarization cases possible:

- The incident light might be polarized in such a way that the vector of the electrical field **E** lies either *in* the plane of the material, or *perpendicular* to it, as shown below. Anything in between than can be decomposed into the two basic cases.



- Lets call the amplitudes of the reflected beam A_{para} for the case of the polarization being *parallel* to the plane (= surface of the dielectric), and A_{perp} for the case of the polarization being *perpendicular* to the plane (blue case) as shown above. For a unit amplitude of the incident beam, the Fresnel laws then state

$$A_{para} = -\frac{\sin(\alpha - \beta)}{\sin(\alpha + \beta)} \qquad A_{perp} = -\frac{\tan(\alpha - \beta)}{\tan(\alpha + \beta)}$$

- We can substitute the angle β by using the relation from above and the resulting equations then give the intensity of the reflected light as a function of the material parameter ***n*** . Possible, but the resulting equations are no longer simple.
  - In order to stay simple and focus on the essentials, we will now consider only cases with about *perpendicular* incidence, i.e. $\alpha \approx 0^o$. This makes everything much easier.
  - At *small* angles we may substitute the argument of the **sin** or **tan** for the full function, and obtain for *both* polarizations

$$A \approx -\frac{(\alpha - \beta)}{(\alpha + \beta)}$$

  - Using the expression for ***n*** from above for small angles too, we obtain

$$n = \frac{\sin\alpha}{\sin\beta} \approx \frac{\alpha}{\beta}$$

  - Now we keep in mind that we are usually interested in intensities , and not in amplitudes. Putting everything together, we obtain for the reflectivity ***R*** , defined as the ratio of the intensity I_r of the reflected beam to the intensity I_i of the incident beam for almost perpendicular incidence

$$R = \frac{I_r}{I_i} = \frac{(n-1)^2}{(n+1)^2}$$

- The grand total of all of this is that if we know ***n*** and some basics about optics, we can answer most, *but not all* of the questions from above. But so far we also did not use a *complex* index of refraction either.
  - In essence, what is missing is any statement about the *attenuation* of the refracted beam, the **damping of the light** inside the dielectric - it is simply not contained in the equations presented so far.
  - This cannot be right. Electromagnetic radiation does not penetrate arbitrarily thick (and still perfect) dielectrics - it gets pretty dark, for example, in deep water even if it is perfectly clear.
  - In not answering the "damping" question, we even raise a new question: If we include damping in the consideration of wave propagation inside a dielectric, does it change the simple equations given above?
- The *bad* news is: It does! But relax: The *good* news is:
  - All we have to do is to exchange the "simple" refractive index ***n*** by a *complex refractive index* ***n**** that is directly tied to the complex dielectric function, and everything is taken care of.
  - We will see how this works in the next paragraph.

3.7.2 The Complex Index of Refraction

- In looking in detail at the polarization of dielectrics, we switched from a simple dielectric contant ϵ_r to a dielectric function $\epsilon_r(\omega) = \epsilon' + i\epsilon''$. This, after some getting used to, makes life much easier and provides for new insights not easily obtainable otherwise.
  - We now do exactly the same thing for the index of refraction, i.e. we replace ***n*** by a **complex index of refraction** ***n****.

$$n^* = n + i \cdot \kappa$$

  - We use the old symbol ***n*** for the real part instead of ***n'*** and κ instead of ***n''***, but that is simply to keep with tradition.
  - With the dielectric *constant* and a *constant* index of refraction we had the basic relation ,

$$n^2 = \epsilon_r$$

  - We simply use this relation now for defining the *complex* index of refraction. This gives us

$$(n + i\kappa)^2 = \epsilon' + i \cdot \epsilon''$$

  - With $n = n(\omega)$; $\kappa = \kappa(\omega)$, since ϵ' and ϵ'' are frequency dependent as discussed before.
- Re-arranging for ***n*** and κ yields somewhat unwieldy equations:

$$n^2 = \frac{1}{2}\left(\left(\epsilon'^2 + \epsilon''^2\right)^{1/2} + \epsilon'\right)$$

$$\kappa^2 = \frac{1}{2}\left(\left(\epsilon'^2 + \epsilon''^2\right)^{1/2} - \epsilon'\right)$$

- Anyway - That is all. *We now have optics covered*. An example of an real complex index of refraction is shown in the link.
  - So lets see how it works and what κ, the so far unspecified imaginary part of n_{com}, will give us.
- First, lets get some easier formula. In order to do this, we remember that ϵ'' was connected to the conductivity of the material and express ϵ'' in terms of the (total) conductivity as

$$\epsilon'' = \frac{\sigma_{DK}}{\epsilon_0 \cdot \omega}$$

  - Note that in contrast to the definition of ϵ'' given before in the context of the dielectric function, we have an ϵ_0 in the ϵ'' part. We had, for the sake of simplicity, made a convention that the ϵ in the dielectric function contain the ϵ_0, but here it more convenient to write it out, because then $\epsilon' = \epsilon_0 \cdot \epsilon_r$ is reduced to ϵ_r and directly related to the "simple" index of refraction ***n***
  - Using that in the expression $(n + i\kappa)^2$ gives

$$(n + i\kappa)^2 = n^2 - \kappa^2 + i \cdot 2n\kappa = \epsilon' + i \cdot \frac{\sigma_{DK}}{\epsilon_0 \cdot \omega}$$

  - We have a complex number on both sides of the equality sign, and this demands that the real and imaginary parts must be the same on both sides, i.e.

$$n^2 - \kappa^2 = \epsilon'$$

$$n\kappa = \frac{\sigma_{DK}}{2\epsilon_0\omega}$$

- Separating ***n*** and κ finally gives

$$n^2 = \frac{1}{2}\left(\epsilon' + \left(\epsilon'^2 + \frac{\sigma_{DK}^2}{4\epsilon_0^2\omega^2} \right)^{1/2} \right)$$

$$\kappa^2 = \frac{1}{2}\left(-\epsilon' + \left(\epsilon'^2 + \frac{\sigma_{DK}^2}{4\epsilon_0^2\omega^2} \right)^{1/2} \right)$$

- Similar to what we had above, but now with basic quantities like the "dielectric constant" $\epsilon' = \epsilon_r$ and the conductivity σ_{DK}^2.

- The equations above go beyond just describing the optical properties of (perfect) dielectrics because we can include all kinds of conduction mechanisms into σ, and all kinds of polarization mechanisms into ϵ'.
  - We can even use these equations for things like the reflectivity of metals, as we shall see.
- Keeping in mind that typical ***n***'s in the visible region are somewhere between **1.5 - 2.5** ($n \approx 2.5$ for diamond is one of the higher values as your girl friend knows), we can draw a few quick conclusions: From the simple but coupled equations for ***n*** and κ follows:
  - κ should be rather small for "common" optical materials, otherwise our old relation of $n = (\epsilon_r)^{1/2}$ would be not good.
  - κ should be rather small for "common" optical materials, because optical materials are commonly insulators, i.e. $\sigma_{DK} \approx 0$ applies.
  - For $\sigma_{DK} = 0$ (and, as we would assume as a matter of course, $\epsilon_r > 0$) we obtain immediately $n = (\epsilon_r)^{1/2}$ and $\kappa = 0$ - the old-fashioned simple relation between just ϵ_r and ***n***.
  - For large σ_{DK} values, both ***n*** and κ will become large. We don't know yet what κ means in physical terms, but very large ***n*** simply mean that the intensity of the reflected beam approaches **100 %**. Light that hits a good conductor thus will get reflected - well, that is exactly what happens between light and (polished) metals, as we know from everyday experience.
- But now we must look at some problems that can be solved with the complex index of refraction in order to understand what it encodes.

3.7.3 Using the Complex Index of Refraction

- Lets look at the physical meaning of **n** and κ, i.e. the real and complex part of the complex index of refraction, by looking at an electromagnetic wave traveling through a medium with such an index.
  - For that we simply use the general formula for the electrical field strength ***E*** of an electromagnetic wave traveling in a medium with refractive index **n***. For simplicities sake, we do it one-dimensional in the ***x***-direction (and use the index "**x**" only in the first equation). In the most general terms we have

$$E_x = E_{0,x} \cdot \exp i \cdot (k_x \cdot x - \omega \cdot t)$$

  - With k_x = component of the wave vector in **x**-direction = **k** = $2\pi/\lambda$, ω = circular frequency = $2\pi\nu$.
- No index of refraction in the formulas; but we know (it is hoped), what to do. We must introduce the velocity **v** of the elecromagnetic wave in the material and use the relation between frequency, wavelength, and velocity to get rid of ***k*** or λ, respectively.
  - In other words, we use

$$v = \frac{c}{n^*} \qquad v = \nu \cdot \lambda$$

$$k = \frac{2\pi}{\lambda} = \frac{\omega \cdot n^*}{c}$$

  - Of course, **c** is the speed of light in vacuum. Insertion yields

$$E_x = E_{0,x} \cdot \exp i \cdot \left(\frac{\omega \cdot n^*}{c} \cdot x - \omega \cdot t \right) = E_{0,x} \cdot \exp i \cdot \left(\frac{\omega \cdot (n + i \cdot \kappa)}{c} \cdot x - \omega \cdot t \right)$$

$$E_x = E_{0,x} \cdot \exp \cdot \left(\frac{i \cdot \omega \cdot n \cdot x}{c} - \frac{\omega \cdot \kappa \cdot x}{c} - i \cdot \omega \cdot t \right)$$

- The red expression is nothing but the wavevector, so we get a rather simple result:

$$E_x = \exp - \frac{\omega \cdot \kappa \cdot x}{c} \cdot \exp[\, i \cdot (k_x \cdot x - \omega \cdot t)]$$

- In words that means: if we use a complex index of refraction, the propagation of electromagnetic waves in a material is whatever it would be for a simple *real* index of refractions times a *damping factor* that decreases the amplitude exponentially as a function of ***x***.
  - Obviously, at a depth often called absorption length or penetration depth **W = c/**ω · κ, the intensity decreased by a factor **1/e**.
  - The imaginary part κ of the complex index of refraction thus describes rather directly the attenuation of electromagnetic waves in the material considered. It is known as **damping constant**, **attenuation index**, **extinction coefficient**, or (rather misleading) *absorption constant*. Misleading, because an absorption constant is usually the α in some exponential decay law of the form $I = I_0 \cdot \exp - \alpha \cdot x$.
  - Note: Words like "constant", "index", or "coefficient" are also misleading - because κ is not constant, but depends on the frequency just as much as the real and imaginary part of the dielectric function.

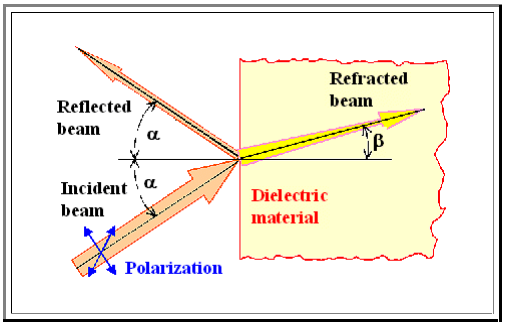
(Should be continued but won't)

3.7.4 Summary to: Dielectrics and Optics

The basic questions one would like to answer with respect to the optical behaviour of materials and with respect to the simple situation as illustrated are:

1. How large is the fraction ***R*** that is reflected? **1 – *R*** then will be going in the material.
2. How large is the angle β, i.e. how large is the refraction of the material?
3. How is the light in the material absorped, i.e. how large is the absorption coefficient?

Of course, we want to know that as a function of the wave length λ or the frequency **ν = c/**λ, the angle α, and the two basic directions of the polarization (



All the information listed above is contained in the complex index of refraction ***n**** as given ⇒

$n = (\epsilon_r)^{1/2}$	Basic definition of "normal" index of refraction ***n***
$n^* = n + i \cdot \kappa$	Terms used for complex index of refaction ***n**** ***n*** = real part κ = imaginary part
$n^{*2} = (n + i\kappa)^2 = \epsilon' + i \cdot \epsilon''$	Straight forward definition of ***n****

Working out the details gives the basic result that

- Knowing ***n*** = real part allows to answer question **1** and **2** from above via "Fresnel laws" (and "Snellius' law", a much simpler special version).
- Knowing κ = imaginary part allows to answer question **3** ⇒

$$E_x = \exp - \frac{\omega \cdot \kappa \cdot x}{c} \cdot \exp[\, i \cdot (k_x \cdot x - \omega \cdot t)]$$

Amplitude: Exponential decay with κ — **"Running" part of the wave**

- Knowing the dielectric function of a dielectric material (with the imaginary part expressed as conductivity σ_{DK}), we have (simple) optics completely covered!
  - If we would look at the tensor properties of ϵ, we would also have crystal optics (= anisotropic behaviour; things like birefringence) covered.
  - We must, however, dig deeper for e.g. non-linear optics ("red in - green (double frequency) out"), or new disciplines like quantum optics.

$$n^2 = \frac{1}{2}\left(\epsilon' + \left(\epsilon'^2 + \frac{\sigma_{DK}^2}{4\epsilon_0^2\omega^2}\right)^{1/2}\right)$$

$$\kappa^2 = \frac{1}{2}\left(-\epsilon' + \left(\epsilon'^2 + \frac{\sigma_{DK}^2}{4\epsilon_0^2\omega^2}\right)^{1/2}\right)$$

Questionaire

Multiple Choice questions to all of 3.7

3.8 Summary: Dielectrics

- The dielectric constant ϵ_r "somehow" describes the interaction of dielectric (i.e. more or less insulating) materials and electrical fields; e.g. via the equations ⇒
  - $\underline{D}$ is the **electrical displacement** or **electrical flux density**, sort of replacing $\underline{E}$ in the Maxwell equations whenever materials are encountered.
  - C is the capacity of a parallel plate capacitor (plate area A, distance d) that is "filled" with a dielectric with ϵ_r
  - n is the index of refraction; a quantity that "somehow" describes how electromagnetic fields with extremely high frequency interact with matter.
    in this equaiton it is assumed that the material has no magnetic properties at the frequency of light.

$$\underline{D} = \epsilon_0 \cdot \epsilon_r \cdot \underline{E}$$

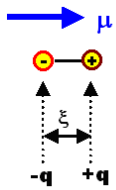
$$C = \frac{\epsilon_0 \cdot \epsilon_r \cdot A}{d}$$

$$n = (\epsilon_r)^{1/2}$$

- Electrical fields inside dielectrics polarize the material, meaning that the vector sum of electrical dipoles inside the material is no longer zero.
  - The decisive quantities are the dipole moment $\underline{\mu}$, a vector, and the Polarization P, a vector, too.
  - Note: The dipole moment vector points from the negative to the positive charge - contrary to the electrical field vector!
  - The dipoles to be polarized are either already present in the material (e.g. in H_2O or in ionic crystals) or are induced by the electrical field (e.g. in single atoms or covalently bonded crystals like **Si**)
  - The dimension of the polarization $\underline{P}$ is **[C/cm^2]** and is indeed identical to the net charge found on unit area ion the surface of a polarized dielectric.

$$\underline{\mu} = q \cdot \underline{\xi}$$

$$\underline{P} = \frac{\Sigma \underline{\mu}}{V}$$



- The equivalent of "Ohm's law", linking current density to field strength in conductors is the Polarization law:
  - The decisive material parameter is χ ("kee"), the **dielectric susceptibility**
  - The "classical" flux density D and the Polarization are linked as shown. In essence, $\underline{P}$ only considers what happens in the material, while D looks at the total effect: material plus the field that induces the polarization.

$$\underline{P} = \epsilon_0 \cdot \chi \cdot \underline{E}$$

$$\epsilon_r = 1 + \chi$$

$$\underline{D} = \underline{D_0} + \underline{P} = \epsilon_0 \cdot \underline{E} + \underline{P}$$

- Polarization by necessity moves masses (electrons and / or atoms) around, this will not happen arbitrarily fast.
  - ϵ_r or χ thus must be functions of the frequency of the applied electrical field, and we want to consider the whole frequency range from **RF** via **HF** to light and beyond.

$\epsilon_r(\omega)$ is called the "**dielectric function**" of the material.

- The tasks are:
  - Identify and (quantitatively) describe the major mechanisms of polarization.
  - Justify the assumed linear relationship between P and χ.
  - Derive the dielectric function for a given material.

(Dielectric) polarization mechanisms in dielectrics are all mechanisms that

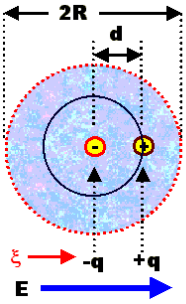
1. Induce dipoles at all (always with $\boldsymbol{\mu}$ in field direction)
   ⇒ Electronic polarization.
2. Induce dipoles already present in the material to "point" to some extent in field direction.
   ⇒ Interface polarization.
   ⇒ Ionic polarization.
   ⇒ Orientation polarization.

Quantitative considerations of polarization mechanisms yield

- **Justification (and limits) to the $\underline{P} \propto \underline{E}$ "law"**
- **Values for χ**
- **$\chi = \chi(\omega)$**
- **$\chi = \chi$(structure)**

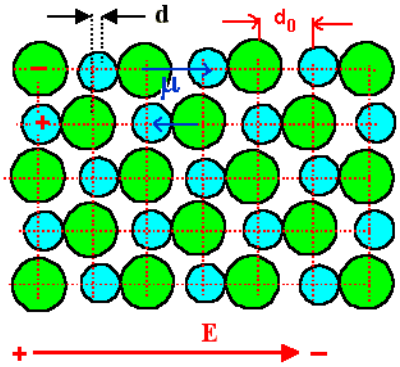
Electronic polarization describes the separation of the centers of "gravity" of the electron charges in orbitals and the positive charge in the nucleus and the dipoles formed this way. it is always present

- It is a very weak effect in (more or less isolated) atoms or ions with spherical symmetry (and easily calculated).
- It can be a strong effect in e.g. covalently bonded materials like **Si** (and not so easily calculated) or generally, in solids.



Ionic polarization describes the net effect of changing the distance between neighboring ions in an ionic crystal like **NaCl** (or in crystals with some ionic component like **SiO_2**) by the electric field

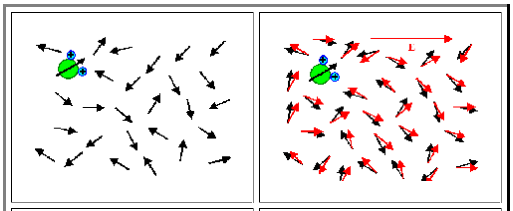
- Polarization is linked to bonding strength, i.e. Young's modulus **Y**. The effect is smaller for "stiff" materials, i.e.
  $P \propto 1/Y$



Orientation polarization results from minimizing the free enthalpy of an ensemble of (molecular) dipoles that can move and rotate freely, i.e. polar liquids.

- It is possible to calculate the effect, the result invokes the Langevin function

$$L(\beta) = \coth(\beta) - \frac{1}{\beta}$$



- In a good approximation the polarization is given by ⇒

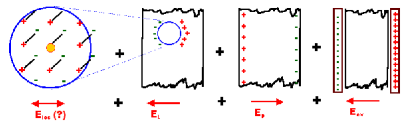
$$\langle P \rangle = \frac{N \cdot \mu^2 \cdot E}{3kT}$$

The induced dipole moment **μ** in all mechanisms is proportional to the field (for reasonable field strengths) at the location of the atoms / molecules considered.

$$\underline{\mu} = \alpha \cdot E_{loc}$$

- The proportionality constant is called polarizability α; it is a microscopic quantity describing what atoms or molecules "do" in a field.
- The local field, however, is not identical to the macroscopic or external field, but can be obtained from this by the Lorentz approach
- For isotropic materials (e.g. cubic crystals) one obtains

$$E_L = \frac{P}{3\epsilon_0}$$



$$E_{loc} = E_{ex} + E_{pol} + E_L + E_{near}$$

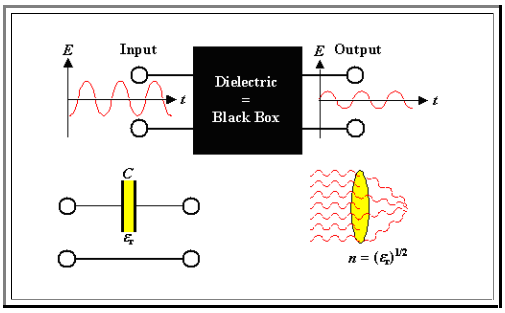
Knowing the local field, it is now possible to relate the microscopic quantity α to the macroscopic quantity ϵ or ϵ_r via the Clausius - Mosotti equations ⇒

- While this is not overly important in the engineering practice, it is a momentous achievement. With the Clausius - Mosotti equations and what went into them, it was possible for the first time to understand most electronic and optical properties of dielectrics in terms of their constituents (=atoms) and their structure (bonding, crystal lattices etc.)
- Quite a bit of the formalism used can be carried over to other systems with dipoles involved, in particular magnetism=behavior of magnetic dipoles in magnetic fields.

$$\frac{N \cdot \alpha}{3\,\epsilon_0} = \frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{X}{X + 3}$$

Alternating electrical fields induce alternating forces for dielectric dipoles. Since in all polarization mechanisms the dipole response to a field involves the movement of masses, inertia will prevent arbitrarily fast movements.

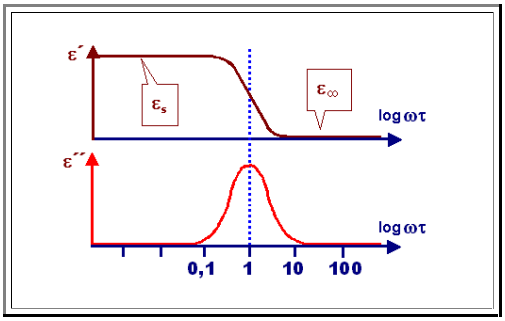
- Above certain limiting frequencies of the electrical field, the polarization mechanisms will "die out", i.e. not respond to the fields anymore.
- This might happen at rather high (=optical) frequencies, limiting the index of refraction $n=(\epsilon_r)^{1/2}$



The (only) two physical mechanisms governing the movement of charged masses experiencing alternating fields are relaxation and resonance.

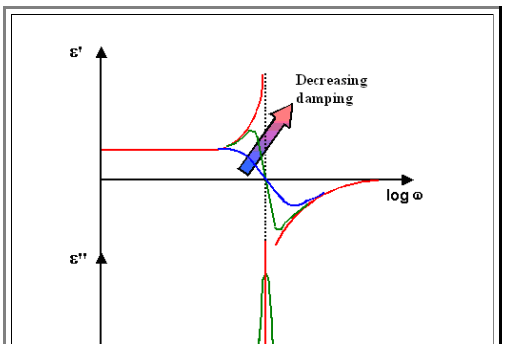
Relaxation describes the decay of excited states to the ground state; it describes, e.g., what happens for orientation polarization after the field has been switched off.

- From the "easy to conceive" time behavior we deduce the frequency behavior by a Fourier transformation
- The dielectric function describing relaxation has a typical frequency dependence in its real and imaginary part ⇒

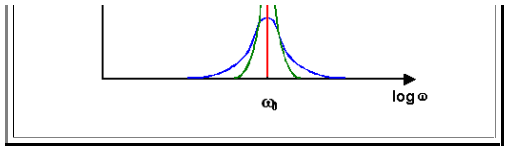


Resonance describes anything that can be modeled as a mass on a spring - i.e. electronic polarization and ionic polarization.

- The decisive quantity is the (undamped) resonance frequency $\omega_0=(k_S/m)^{1/2}$ and the "friction" or damping constant k_F
- The "spring" constant is directly given by the restoring forces between charges, i.e. Coulombs law, or (same thing) the bonding. In the case of bonding (ionic polarization) the spring constant is also easily expressed in terms of Young's modulus *Y*. The masses are electron or atom masses for electronic or ionic polarization, respectively.



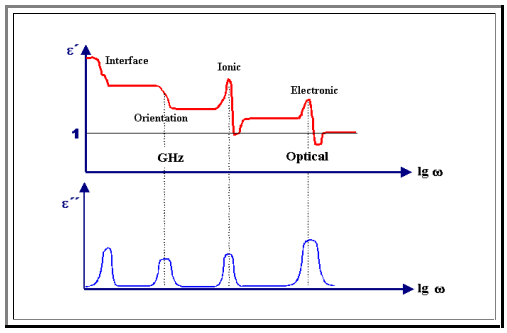
- The damping constant describes the time for funneling off ("dispersing") the energy contained in one oscillating mass to the whole crystal lattice. Since this will only take a few oscillations, damping is generally large.
- The dielectric function describing relaxation has a typical frequency dependence in its real and imaginary part ⇒
  The green curve would be about right for crystals.



The complete frequency dependence of the dielectric behavior of a material, i.e. its dielectric function, contains all mechanisms "operating" in that material.

- As a rule of thumb, the critical frequencies for relaxation mechanisms are in the **GHz** region, electronic polarization still "works" at optical (**10^{15} Hz**) frequencies (and thus is mainly responsible for the index of refraction).
- Ionic polarization has resonance frequencies in between.
- Interface polarization may "die out" already a low frequencies.

A widely used diagram with all mechanisms shows this, but keep in mind that there is no real material with all 4 major mechanisms strongly present!
⇒



A general mathematical theorem asserts that the real and imaginary part of the dielectric function cannot be completely independent

- If you know the complete frequency dependence of either the real or the imaginary part, you can calculate the complete frequency dependence of the other.
- This is done via the Kramers-Kronig relations; very useful and important equations in material practice.
  ⇒

$$\epsilon'(\omega) = \frac{-2\omega}{\pi} \int_0^\infty \frac{\omega^* \cdot \epsilon''(\omega^*)}{\omega^{*2} - \omega^2} \cdot d\omega^*$$

$$\epsilon''(\omega) = \frac{2\omega}{\pi} \int_0^\infty \frac{\epsilon'(\omega^*)}{\omega^{*2} - \omega^2} \cdot d\omega^*$$

The frequency dependent current density ***j*** flowing through a dielectric is easily obtained. ⇒

- The in-phase part generates active power and thus heats up the dielectric, the out-of-phase part just produces reactive power
- The power losses caused by a dielectric are thus directly proportional to the imaginary component of the dielectric function

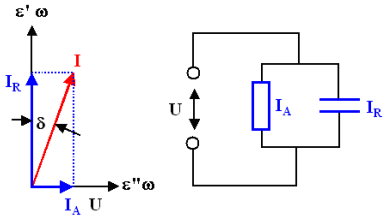
$$j(\omega) = \frac{dD}{dt} = \epsilon(\omega) \cdot \frac{dE}{dt} = \omega \cdot \epsilon'' \cdot E(\omega) + i \cdot \omega \cdot \epsilon' \cdot E(\omega)$$

in phase — out of phase

$$L_A = \text{power turned into heat} = \omega \cdot |\epsilon''| \cdot E^2$$

The relation between active and reactive power is called "tangens Delta" (**tg(δ)**); this is clear by looking at the usual pointer diagram of the current

$$\frac{L_A}{L_R} := \text{tg}\,\delta = \frac{I_A}{I_R} = \frac{\epsilon''}{\epsilon'}$$



- The pointer diagram for an *ideal* dielectric **σ(ω=0)=0**can always be obtained form an (ideal) resistor ***R*(ω)** in parallel to an (ideal) capacitor ***C*(ω)**.
- ***R*(ω)** expresses the apparent conductivity $\sigma_{DK}(\omega)$ of the dielectric, it follows that

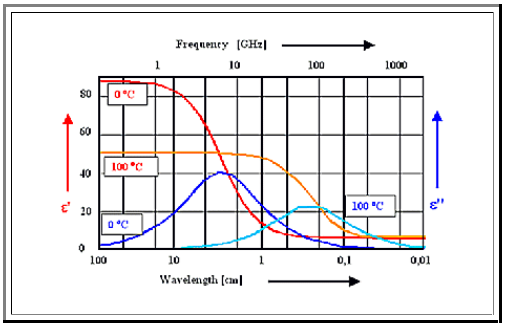
$$\sigma_{DK}(\omega) = \omega \cdot \epsilon''(\omega)$$

For a *real* dielectric with a non-vanishing conductivity at zero (or small) frequencies, we now just add another resistor in parallel. This allows to express *all* conductivity effects of a real dielectric in the imaginary part of its (usually measured) dielectric function via

$$\epsilon'' = \frac{\sigma_{total}}{\omega}$$

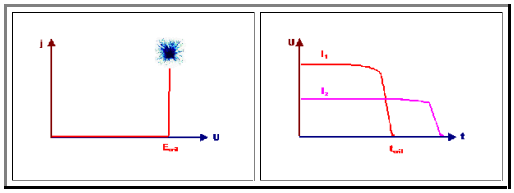
- We have no *all* materials covered with respect to their dielectric behavior - in principle even metals, but then resorting to a dielectric function would be overkill.

A good example for using the dielectric function is "dirty" water with a not-too-small (ionic) conductivity, commonly encountered in food.



- The polarization mechanism is orientation polarization, we expect large imaginary parts of the dielectric function in the **GHz** region.
- It follows that food can be heated by microwave (ovens)!

The first law of materials science obtains: At field strengths larger than some critical value, dielectrics will experience (destructive) electrical breakdown



- This might happen suddenly (then calls break-down) , with a bang and smoke, or
- it may take time - months or years - then called failure.
- Critical field strength may vary from **< 100 kV/cm** to **> 10 MV / cm**.

Highest field strengths in practical applications do not necessarily occur at high voltages, but e.g. in integrated circuits for very thin (a few **nm**) dielectric layers

Example 1: TV set, **20 kV** cable, thickness of insulation=**2 mm.** ⇒ ***E*=100 kV/cm**
Example 2: Gate dielectric in transistor, **3.3 nm** thick, **3.3 V** operating voltage. ⇒ ***E*=10 MV/cm**

- Properties of thin films may be quite different (better!) than bulk properties!

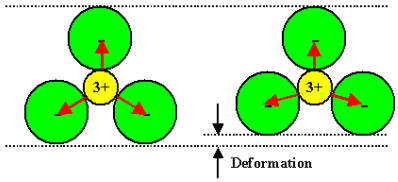
Electrical breakdown is a major source for failure of electronic products (i.e. one of the reasons why things go "kaputt" (=broke)), but there is no simple mechanism following some straight-forward theory. We have:

- *Thermal breakdown*; due to small (field dependent) currents flowing through "weak" parts of the dielectric.
- *Avalanche breakdown* due to occasional free electrons being accelerated in the field; eventually gaining enough energy to ionize atoms, producing more free electrons in a runaway avalanche.
- *Local discharge* producing micro-plasmas in small cavities, leading to slow erosion of the material.

- *Electrolytic breakdown* due to some ionic micro conduction leading to structural changes by, e.g., metal deposition.

Polarization ***P*** of a dielectric material can also be induced by mechanical deformation ***e*** or by other means.

- *Piezo electric materials* are anisotropic crystals meeting certain symmetry conditions like crystalline quartz (**SiO_2**): the effect is linear.
- The effect also works in reverse: Electrical fields induce mechanical deformation
- Piezo electric materials have many uses, most prominent are quartz oscillators and, recently, fuel injectors for Diesel engines.



$P = \text{const.} \cdot e$

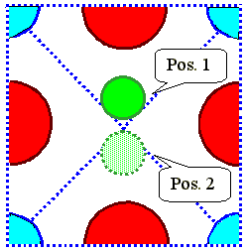
Electrostriction also couples polarization and mechanical deformation, but in a quadratic way and only in the direction "electrical fields induce (very small) deformations".

- The effect has little uses so far; it can be used to control very small movements, e.g. for manipulations in the **nm** region. Since it is coupled to electronic polarization, many materials show this effect.

$$e = \frac{\Delta l}{l} = \text{const} \cdot E^2$$

Ferro electric materials posses a permanent dipole moment in any elementary cell that, moreover, are all aligned (below a critical temperature).

- There are strong parallels to ferromagnetic materials (hence the strange name).
- Ferroelectric materials have large or even very large (ϵ_r **> 1.000**) dielectric constants and thus are to be found inside capacitors with high capacities (but not-so-good high frequency performance)

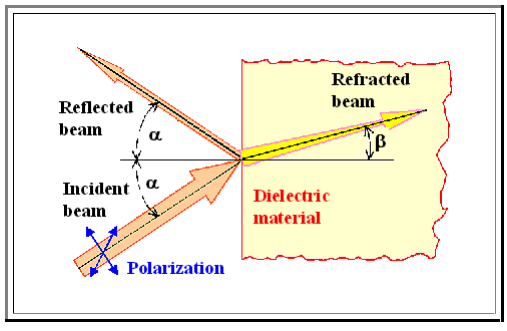


$BaTiO_3$ unit cell

Pyro electricity couples polarization to temperature changes; *electrets* are materials with permanent polarization, ... There are more "curiosities" along these lines, some of which have been made useful recently, or might be made useful - as material science and engineering progresses.

The basic questions one would like to answer with respect to the optical behaviour of materials and with respect to the simple situation as illustrated are:

1. How large is the fraction ***R*** that is reflected? **1 –** ***R*** then will be going in the material.
2. How large is the angle β, i.e. how large is the refraction of the material?
3. How is the light in the material absorped, i.e. how large is the absorption coefficient?



- Of course, we want to know that as a function of the wave length λ or the frequency **ν=c/**λ, the angle α, and the two basic directions of the polarization (

All the information listed above is contained in the complex index of refraction ***n**** as given ⇒

$n=(\epsilon_r)^{1/2}$	Basic definition of "normal" index of refraction ***n***
$n^* = n + i \cdot \kappa$	Terms used for complex index of refaction ***n**** ***n***=real part κ=imaginary part
$n^{*2}=(n + i\kappa)^2 = \epsilon' + i \cdot \epsilon''$	Straight forward definition of ***n****

Working out the details gives the basic result that

- Knowing ***n***=real part allows to answer question **1** and **2** from above via "Fresnel laws" (and "Snellius' law", a much simpler special version).
- Knowing κ=imaginary part allows to answer question **3** ⇒

$$E_x = \exp\left(-\frac{\omega \cdot \kappa \cdot x}{c}\right) \cdot \exp[\, i \cdot (k_x \cdot x - \omega \cdot t)]$$

Amplitude: Exponential decay with κ — **"Running" part of the wave**

Knowing the dielectric function of a dielectric material (with the imaginary part expressed as conductivity σ_{DK}), we have (simple) optics completely covered!

- If we would look at the tensor properties of ϵ, we would also have crystal optics (=anisotropic behaviour; things like birefringence) covered.
- We must, however, dig deeper for e.g. non-linear optics ("red in - green (double frequency) out"), or new disciplines like quantum optics.

$$n^2 = \frac{1}{2}\left(\epsilon' + \left(\epsilon'^2 + \frac{\sigma_{DK}^2}{4\epsilon_0^2\omega^2}\right)^{1/2}\right)$$

$$\kappa^2 = \frac{1}{2}\left(-\epsilon' + \left(\epsilon'^2 + \frac{\sigma_{DK}^2}{4\epsilon_0^2\omega^2}\right)^{1/2}\right)$$

Questionaire

Multiple Choice questions to all of 3

4. Magnetic Materials

4. Magnetic Materials

4.1 Definitions and General Relations

4.1.1 Fields, Fluxes and Permeability

- There are many analogies between dielectric and magnetic phenomena; the big difference being that (so far) there are *no magnetic "point charges"*, so-called magnetic monopoles, but only *magnetic dipoles*.
  - The first basic relation that we need is the relation between the magnetic flux density ***B*** and the magnetic field strength ***H*** *in vacuum*. It comes straight from the Maxwell equations:

$$B = \mu_o \cdot H$$

  - The symbols are:
    - ***B*** = **magnetic flux density** or **magnetic induction**,
    - μ_o = **magnetic permeability of the vacuum** = $4\pi \cdot 10^{-7}$ **Vs/Am** = $1{,}26 \cdot 10^{-6}$ **Vs/Am**
    - ***H*** = **magnetic field strength**

    .
  - The units of the magnetic field ***H*** and so on are
    - **[*H*] = A/m**
    - **[*B*]** = **Vs/m^2**, with **1Vs/m^2 = 1 Tesla**.

    B and ***H*** are vectors, of course.
  - **$10^3/4\pi$ A/m** used to be called **1 Oersted**, and **1 Tesla** equales **10^4 Gauss** in the old system.
  - Why the eminent mathematician and scientist *Gauss* was dropped in favor of the somewhat shady figure *Tesla* remains a mystery.
- If a material is present, the relation between magnetic field strength and magnetic flux density becomes

$$B = \mu_o \cdot \mu_r \cdot H$$

  - with μ_r = **relative permeability of the material** in complete analogy to the *electrical flux density* and the *dielectric constant*.
  - The relative permeability of the material μ_r is a material parameter without a dimension and thus a *pure number* (or several pure numbers if we consider it to be a *tensor* as before). It is the material property we are after.
- Again , it is useful and conventional to split ***B*** into the *flux density in the vacuum* plus the *part of the material* according to

$$B = \mu_o \cdot H + J$$

  - With ***J*** = **magnetic polarization** in analogy to the dielectric case.
- As a new thing, we now we define the **magnetization *M*** of the material as

$$M = \frac{J}{\mu_o}$$

  - That is only to avoid some labor with writing. This gives us

$$B = \mu_o \cdot (H + M)$$

- Using the independent definition of ***B*** finally yields

$$M = (\mu_r - 1) \cdot H$$

$$M := \chi_{mag} \cdot H$$

- With $\chi_{mag} = (\mu_r - 1) =$ **magnetic susceptibility**.
- It is really straight along the way we looked at dielectric behavior; for a direct comparison use the link

The magnetic susceptibility χ_{mag} is the *prime material parameter* we are after; it describes the response of a material to a magnetic field in exactly the same way as the dielectric susceptibility $\chi_{dielectr}$. We even chose the same abbreviation and will drop the suffix most of the time, believing in your intellectual power to keep the two apart.

- Of course, the four *vectors* ***H***, ***B***, ***J***, ***M*** are all parallel in isotropic homogeneous media (i.e. in amorphous materials and poly-crystals).
- In anisotropic materials the situation is more complicated; χ and μ_r then must be seen as tensors.

We are left with the question of the *origin of the magnetic susceptibility*. There are no **magnetic monopoles** that could be separated into magnetic dipoles as in the case of the dielectric susceptibility, there are only *magnetic dipoles* to start from.

- Why there are no magnetic monopoles (at least none have been discovered so far despite extensive search) is one of the tougher questions that you can ask a physicist; the ultimate answer seems not yet to be in. So just take it as a fact of life.
- In the next paragraph we will give some thought to the the origin of magnetic dipoles.

4.1.2 Origin of Magnetic Dipoles

- Where are magnetic dipoles coming from? The classical answer is simple: A **magnetic moment** ***m*** is generated whenever a *current flows in closed circle*.
  - *Of course, we will not mix up the letter* ***m*** *used for magnetic moments with the* m^*_e *, the mass of an electron, which we also need in some magnetic equations*.
  - For a current ***I*** flowing in a circle enclosing an area ***A***, ***m*** is defined to be

$$m = I \cdot A$$

  - This does not only apply to "regular" current flowing in a wire, but in the extreme also to a *single electron circling around an atom*.
- In the context of **Bohrs model** for an atom, the magnetic moment of such an electron is easily understood:
  - The current ***I*** carried by *one* electron orbiting the nucleus at the distance ***r*** with the frequency $\nu = \omega/2\pi$ is .

$$I = e \cdot \frac{\omega}{2\pi}$$

  - The area ***A*** is πr^2, so we have for the magnetic moment m_{orb} of the electron

$$m_{orb} = e \cdot \frac{\omega}{2\pi} \cdot \pi r^2 = \tfrac{1}{2} \cdot e \cdot \omega \cdot r^2$$

- Now the *mechanical* angular momentum ***L*** is given by

$$L = m^*_e \cdot \omega \cdot r^2$$

  - With m^*_e = mass of electron (*the * serves to distinguish the mass* m^*_e *from the magnetic moment* m^e *of the electron)*, and we have a simple relation between the mechanical angular momentum ***L*** of an electron (which, if you remember, was the decisive quantity in the Bohr atom model) and its magnetic moment ***m***.

$$\underline{m}_{orb} = -\frac{e}{2m^*_e} \cdot \underline{L}$$

  - The *minus sign* takes into account that mechanical angular momentum and magnetic moment are antiparallel; as before we note that this is a *vector equation* because both ***m*** and ***L*** are (polar) vectors.
  - The quantity $e/2m^*_e$ is called the **gyromagnetic relation** or quotient; *it should be a fixed constant* relating ***m*** and any given ***L***.
  - However, in real life it often deviates from the value given by the formula. How can that be?
  - Well, try to remember: Bohr's model is a mixture of classical physics and quantum physics and *far too simple* to account for everything. It is thus small wonder that conclusions based on this model will not be valid in all situations.
- In *proper quantum mechanics* (as in Bohr's semiclassical model) ***L*** comes in *discrete values only*. In particular, the fundamental assumption of Bohr's model was $L = n \cdot \hbar$, with ***n*** = quantum number **= 1, 2, 3, 4, ...**
  - It follows that m_{orb} *must be quantized, too*; it must come in multiples of

$$m_{orb} = \frac{h \cdot e}{4\pi \cdot m^*_e} = m_{Bohr} = 9.27 \cdot 10^{-24}\ Am^2$$

- This relation defines a **fundamental unit for magnetic dipole moments**, it has its own name and is called a **Bohr magneton**.
- It is for magnetism what an elementary charge is for electric effects.

But electrons orbiting around a nucleus are not the *only* source of magnetic moments.

- Electrons always have a **spin** ***s***, which, on the level of the Bohr model, can be seen as a built-in angular momentum with the value $\hbar \cdot s$. The spin quantum number ***s*** is ½, and this allows two directions of angular momentum and magnetic moment , always symbolically written as .

$$s = \begin{cases} +1/2 \\ -1/2 \end{cases}$$

It is possible, of course, to compute the circular current represented by a charged ball spinning around its axis if the distribution of charge in the sphere (or on the sphere), is known, and thus to obtain the magnetic moment of the spinning ball.

- Maybe that even helps us to understand the internal structure of the electron, because we know its magnetic moment and now can try to find out what kind of size and internal charge distribution goes with that value. Many of the best physicists have tried to do exactly that.
- However, as it turns out, whatever assumptions you make about the internal structure of the electron that will give the right magnetic moment will always get you into *deep trouble* with *other properties* of the electron. There simply is *no internal structure* of the electron that will explain its properties!
- We thus are forced to simply accept as a *fundamental property of an electron* that it always carries a magnetic moment of

$$m^e = \frac{2 \cdot h \cdot e \cdot s}{4\pi \cdot m^*_e} = \pm m_{Bohr}$$

- The factor **2** is a puzzle of sorts - not only because it appears at all, but because it is actually **= 2.00231928**. But pondering this peculiar fact leads straight to quantum electrodynamics (and several Nobel prizes), so we will not go into this here.

The total magnetic moment of an atom - *still within the Bohr model* - now is given by the (vector)sum of all the "orbital" moments and the "spin" moments of all electrons in the atom, taking into account all the quantization rules; i.e. the requirement that the angular momentums ***L*** cannot point in arbitrary directions, but only in fixed ones.

This is were it gets complicated - even in the context of the simple Bohr model. A bit more to that can be found in the link. But there are few rules we can easily use:

- All *completely filled orbitals* carry *no* magnetic moment because for every electron with spin ***s*** there is a one with spin ***-s***, and for every one going around " *clockwise*", one will circle "*counter clockwise* ". This means:
- *Forget the inner orbitals* - everything cancels!
- Spins on not completely filled orbitals tend to *maximize their contribution*; they will first fill all available energy states with spin up, before they team up and cancel each other with respect to magnetic momentum.
- The *chemical environment*, i.e. bonds to other atoms, incorporation into a crystal, etc., may strongly change the magnetic moments of an atom.

The net effect for a given (isolated) atom is simple. Either it has a magnetic moment in the order of a Bohr magneton because not all contributions cancel - or it has none. And it is possible, (if not terribly easy), to calculate what will be the case. A first simple result emerges: Elements with an *even* number of electrons have generally *no magnetic moment*.

We will leave the rules for getting the permanent magnetic moment of a single atom from the interaction of spin moments and orbital moments to the more advanced (quantum theory) textbooks, here we are going to look at the possible effects if you:

- bring atoms together to form a solid, or
- subject solids to an external magnetic field ***H***

A categorization will be given in the next paragraph.

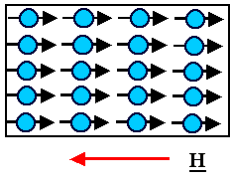
Questionaire

Multiple Choice questions to 4.1.2

4.1.3 Classifications of Interactions and Types of Magnetism

Dia-, Para-, and Ferromagnetism

- We want to get an idea of what happens to *materials* in external magnetic fields. "Material", in contrast to a single atom, means that we have plenty of (possibly different) atoms in close contact, i.e with some bonding. We can distinguish two basic cases:
  - **1.** The atoms of the material have *no magnetic moment* of their own. This is generally true for about one half of the elements; the ones with even atomic numbers and therefore an even number of electrons. The magnetic moments of the spins tends to cancel; the atoms will only have a magnetic moment if there is an orbital contribution. Of course, the situation may change if you look at *ions* in a crystal.
  - **2.** At least *some* of the atoms of the material have a magnetic moment. That covers the other half of the periodic table: All atoms with an odd number of electrons will have one spin moment left over. Again, the situation may be different if you look at ionic crystals.
- Lets see what can happen if you consider interactions of the magnetic moments with each other and with a magnetic field. First, we will treat the case of solids with *no magnetic moments* of their constituents, i.e. **diamagnetic materials**.
  - The following table lists the essentials

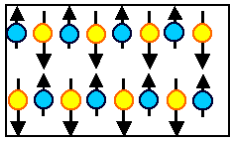
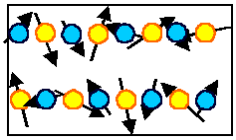
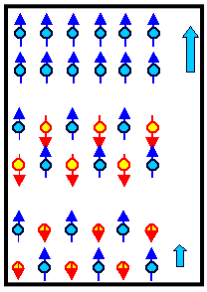
Diamagnetic Materials		
Magnetic moment?	No	
Internal magnetic interaction?	None	
Response to external field	Currents (and small magn. moments) are induced by turning on the field because the orbiting electrons are slightly disturbed. The induced magn. moments oppose the field. No temperature dependence Mechanism analogous to electronic polarisation in dielectrics,	 *The black arrows should be seen as being very short!!!!*
Value of μ_r	$\mu_r \lesssim 1$ in **diamagnetic** Small effect in "regular" materials	$\mu_r = 0$ in **superconductors** (ideal diamagnet)
Value of *B*	$B \lesssim \mu_0 \cdot H$	$B = 0$ in superconductors
Typical materials	All elements with filled shells (always even atomic number)	all noble gases, H_2, Cu, H_2O, NaCl, Bi, ... Alkali or halogene *ions*

- Since you cannot expose material to a magnetic field without encountering a changing field strength **d*H*/d*t*** (either by turning on the field on or by moving the specimen into a field), currents will be induced that produce a magnetic field of their own.
  - According to **Lenz's law**, the direction of the current and thus the field is always such as to oppose the generating forces. Accordingly, the **induced magnetic moment** will be *antiparallel* to the external field.
  - This is called **diamagnetism** and it is a weak effect in normal materials.
- There is an exception, however: **Superconductors**, i.e. materials with a **resistivity = 0** at low temperatures, will have their mobile charges responding without "resistance" to the external field and the induced magnetic moments will *exactly cancel* the external field.
  - Superconductors (at least the "normal" ones (or "type **I**" as they are called) therefore are always *perfectly field free* - a magnetic field cannot penetrate the superconducting material.
  - That is just as amazing as the zero resistance; in fact the magnetic properties of superconductors are just as characteristic for the superconducting state of matter as the resistive properties.
- There will be a backbone II module for superconductors in due time

If we now look at materials where at least *some of the atoms* carry a permanent magnetic moment, we have to look first at the possible *internal interactions* of the magnetic moments in the material and then at their interaction with an *external field*. Two limiting cases can be distinguished.

1. **1.** Strong internal interaction (i.e. interaction energies **» k*T***, the thermal energy). **Ferromagnetism** results
2. **2.** No or weak interaction. We have **paramagnetic materials**.

The first case of strong interaction will more or less turn into the second case at temperatures high enough so that **k*T* >>** interaction energy, so we expect a temperature dependence of possible effects. A first classification looks like this:

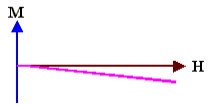
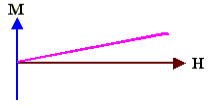
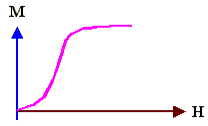
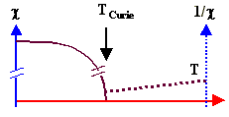
Paramagnetic and Ferromagnetic Materials		
Magnetic moment?	Yes	
Internal agnetic interaction?	Strong	Weak
Ordered regions?	Yes	No
	 This example shows a **ferrimagnetic material** Ordered magnetic structures that are stable in time. Permanent magnetization is obtained by the (vector) sum over the individual magnetic moments.	 Example for a **paramagnetic material** Unordered magnetic structure, fluctuating in time. Averaging over time yields no permanent magnetization
Response to external field	A large component of the magnetic moment may be in field direction	Small average orientation in field direction. Mechanism *fully* analogous to orientation polarization for dielectrics
Kinds of ordering	Many possibilities. Most common are *ferro*-, *antiferro*-, and *ferri*magnetism as in the self-explaining sequence below: 	
Value of μ_r	**μ_r >> 1** for ferromagnets **$\mu_r \approx 1$** for anti-ferromagnets **$\mu_r > 1$** for ferrimagnets	**$\mu_r \geq \approx 1$**
***T*-dependence**	Paramagnetic above **Curie Temperature**	Weak ***T***-dependence
Paramagnetic materials (at room temperature)		Mn, Al, Pt, O_2(gas and liquid), rare earth ions, ...

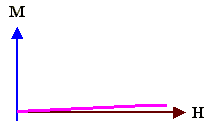
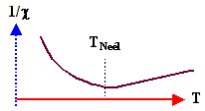
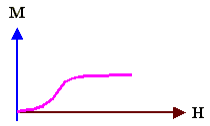
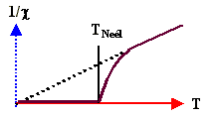
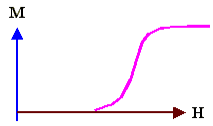
Ferromagnetic materials (with Curie- (or Neél) ***T***)	Ferro elements: Ferro technical: Ferri: Antiferro: (no technical uses)	Fe (770 ^{0}C), Co (1121 ^{0}C), Ni (358 ^{0}C), Gd (16 ^{0}C) "AlNiCo", Co_5Sm, $Co_{17}Sm_2$, "NdFeB" Fe_3O_4, MnO (116 ^{0}C), NiO (525 ^{0}C), Cr (308 ^{0}C)

- This table generated a lot of new names, definitions and question. It sets the stage for the dealing with the various aspects of *ferromagnetism* (including *ferri*- and *anti-ferro* magnetism as well as some more kinds of internal magnetic ordering. A few examples of ferromagnetic materials are given in the link.
  - There might be *many more types* of ordering: Any fixed relation between two vectors qualify. As an example, moment **2** might not be parallel to moment **1** but off by ***x*** degrees; and the succession of many moments might form a spiral pattern.
  - If you can think of some possible ordering (and it is not forbidden by some overruling law of nature), it is a safe bet that mother nature has already made it in some exotic substance. But, to quote Richard **Feynman**:
  - *"It is interesting to try to analyze what happens when a field is applied to such a spiral* (of magnetic ordering) *- all the twistings and turnings that must go on in all those atomic magnets. (Some people like to amuse themselves with the theory of these things!)"* (Lectures on Physics, Vol **II**, **37-13**; Feynmans emphasizes).
- Well, we don't, and just take notice of the fact that there is *some* kind of magnetic ordering for *some* materials.
  - As far as the element are concerned, the only ferromagnets are: **Fe**, **Ni**, and **Co**. (**Mn** almost is one, but not quite).
  - Examples for antiferromagnets include **Cr**, ...
  - And there are many, many *compounds*, often quite strange mixtures (e.g. **NdFeB** or $\mathbf{Sm_2Co_{17}}$), with *remarkable and often useful* ferro-, ferri, antiferro, or,..., properties.

Temperature Dependence of Magnetic Behavior

- How do we distinguish an *antiferromagnetic* material from a *paramagnet* or a *diamagnet*? They all appear not to be very "magnetic" if you probe them with a magnetic field.
  - We have to look at their behavior in a magnetic field *and* at the *temperature dependence* of that behavior. Ordering the atomic magnetic moments is, after all, a *thermodynamical effect* - it always has to compete with entropy - and thus should show some specific temperature dependence.
  - There are indeed quite characteristic curves of major properties with temperature as shown below.

Magnetization $M = M(H)$	Magnetic susceptibility $\chi_{mag} = \chi_{mag}(T)$	Remarks
		For **diamagnets** the susceptibility is negative and close to zero; and there is no temperature dependence.
		For **paramagnets**, the susceptibility is (barely) larger than zero and decreases with ***T***. Plotted as $\mathbf{1/\chi(T)}$ we find a *linear* relationship.
		For **ferromagnets** the susceptibility is large; the magnetization increases massively with **H**. Above a critical temperature $\mathbf{T_{Cu}}$, the *Curie temperature*, paramagnetic behavior is observed.

		**Antiferromagnets** are like paramagnets *above* a critical temperature T_{Ne} called **Neél temperature**. Below T_{Ne} the susceptibility is small, but with a ***T***-dependence quite different from paramagnets.
		**Ferrimagnets** behave pretty much like *ferro*magnets, except that the effect tends to be smaller. The **1/χ(*T*)** curve is very close to zero *below* a critical temperature, also called Neél temperature.
	Just for good measure, the behaviour of one of the more exotic magnetic materials. Shown is a **metamagnet**, behaving like a *ferro* magnet, but only *above* a critical magnetic field strength.	

- The question now will be if we can understand at least *some* of these observations within the framework of some simple theory, similar to what we did for dielectric materials
  - The answer is: *Yes*, we can - but only for the rather uninteresting (for engineering or applications) *dia*- and *para*magnets.
  - Ferro magnets, however, while *extremely interesting electronic materials* (try to imagine a world without them), are a different matter. A real understanding would need plenty of quantum theory (and has not even been fully achieved yet); it is far outside the scope of this lecture course. But a phenomenological theory, based on some assumptions that we do not try to justify, will come straight out from the theory of the orientation polarization for dielectrics, and that is what we are going to look at in the next subchapters.

Questionaire

Multiple Choice questions to 4.1.3

4.1.4 Summary to: Magnetic Materials - Definitions and General Relations

The **relative permeability μ_r** of a material "somehow" describes the interaction of magnetic (i.e. more or less all) materials and magnetic fields ***H***, e.g. vial the equations ⇒

- ***B*** is the **magnetic flux density** or **magnetic induction**, sort of replacing ***H*** in the Maxwell equations whenever materials are encountered.
- ***L*** is the inductivity of a linear solenoid (also called coil or inductor) with length ***l***, cross-sectional area ***A***, and number of turns ***t***, that is "filled" with a magnetic material with μ_r.
- ***n*** is *still* the index of refraction; a quantity that "somehow" describes how electromagnetic fields with extremely high frequency interact with matter.
  For all practical purposes, however, $\mu_r = 1$ for optical frequencies

$$B = \mu_0 \cdot \mu_r \cdot H$$

$$L = \frac{\mu_0 \cdot \mu_r \cdot A \cdot w^2}{l}$$

$$n = (\epsilon_r \cdot \mu_r)^{1/2}$$

Magnetic fields inside magnetic materials polarize the material, meaning that the vector sum of magnetic dipoles inside the material is no longer zero.

- The decisive quantities are the *magnetic* dipole moment $\underline{m}$, a vector, and the *magnetic* Polarization ***J***, a vector, too.
- Note: In contrast to dielectrics, we define an additional quantity, the **magnetization *M*** by simply including dividing ***J*** by μ_0.
- The magnetic dipoles to be polarized are either already present in the material (e.g. in **Fe, Ni or Co**, or more generally, in all **paramagnetic** materials, or are induced by the magnetic fields (e.g. in **diamagnetic** materials).
- The dimension of the magnetization $\underline{M}$ is **[A/m]**; i.e. the same as that of the magnetic field.

$$B = \mu_0 \cdot H + J$$

$$\underline{J} = \mu_0 \cdot \frac{\Sigma \underline{m}}{V}$$

$$M = \frac{J}{\mu_0}$$

The magnetic polarization ***J*** or the magnetization ***M*** are *not* given by some magnetic surface charge, because ⇒.

There is no such thing as a *magnetic monopole*, the (conceivable) counterpart of a negative or positive electric charge

The equivalent of "Ohm's law", linking current density to field strength in conductors is the *magnetic* Polarization law:

- The decisive material parameter is $\chi_{mag} = (\mu_r - 1)$ **= magnetic susceptibility**.
- The "classical" induction ***B*** and the magnetization are linked as shown. In essence, $\underline{M}$ only considers what happens in the material, while ***B*** looks at the total effect: material plus the field that induces the polarization.

$$M = (\mu_r - 1) \cdot H$$

$$M := \chi_{mag} \cdot H$$

$$B = \mu_0 \cdot (H + M)$$

Magnetic polarization mechanisms are formally similar to dielectric polarization mechanisms, but the physics can be entirely different.

Atomic mechanisms of magnetization are not directly analogous to the dielectric case

Magnetic moments originate from:

- The intrinsic magnetic dipole moments ***m*** of elementary particles with spin is measured in units of the Bohr magneton $\mathbf{m_{Bohr}}$.
- The magnetic moment $\boldsymbol{m^e}$ of the electron is ⇒
- Electrons "orbiting" in an atom can be described as a current running in a circle thus causing a magnetic dipole moment; too

$$m_{Bohr} = \frac{h \cdot e}{4\pi \cdot m^*_e} = 9.27 \cdot 10^{-24}\ Am^2$$

$$m^e = \frac{2 \cdot h \cdot e \cdot s}{4\pi \cdot m^*_e} = 2 \cdot s \cdot m_{Bohr} = \pm m_{Bohr}$$

The total magnetic moment of an atom in a crystal (or just solid) is a (tricky to obtain) sum of all contributions from the electrons, and their orbits (including bonding orbitals etc.), it is either:

- **Zero** - we then have a **diamagnetic material**.

**Magnetic field induces dipoles, somewhat analogous to elctronic polarization in dielectrics.
Always very weak effect (except for superconductors)
Unimportant for technical purposes**

- In the order of a few Bohr magnetons - we have a essentially a **paramagnetic material**.

**Magnetic field induces some order to dipoles; strictly analogous to "orientation polarization" of dielectrics.
Always very weak effect
Unimportant for technical purposes**

In some **ferromagnetic** materials spontaneous ordering of magnetic moments occurs below the Curie (or Neél) temperature. The important families are

- Ferromagnetic materials ⇑⇑⇑⇑⇑⇑⇑ large μ_r, **extremely important**.
- Ferrimagnetic materials ⇑⇓⇑⇓⇑⇓⇑ still large μ_r, **very important**.
- Antiferromagnetic materials ⇑⇓⇑⇓⇑⇓⇑ $\mu_r \approx 1$, unimportant

**Ferromagnetic materials:
Fe, Ni, Co, their alloys
"AlNiCo", Co_5Sm, $Co_{17}Sm_2$,
"NdFeB"**

There is characteristic temperature dependence of μ_r for all cases

Questionnaire

Multiple Choice questions to all of 4.1

4.2 Dia- and Paramagnetism

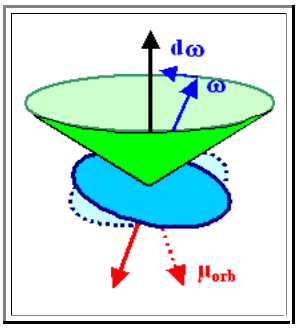
4.2.1 Diamagnetism

What is it Used for?

- It is customary in textbooks of electronic materials to treat dia- and paramagnetism in considerable detail. Considering that there is not a *single* practical case in electrical engineering where it is of any interest if a material is dia- or paramagnetic, there are only two justifications for doing this:
  - Dia- and paramagnetism lend themselves to *calculations* (and engineers like to calculate things).
  - It helps to *understand* the phenomena of magnetism in general, especially the quantum mechanical aspects of it.
- In this script we are going to keep the treatment of dia- and paramagnetism at a minimal level.

Diamagnetism - the Essentials

- The first thing to note about diamagnetism is that *all* atoms and therefore *all* materials show diamagnetic behavior.
  - Diamagnetism thus is always superimposed on all other forms of magnetism. Since it is a small effect, it is hardly noticed, however.
  - Diamagnetism results because all matter contains electrons - either "orbiting" the nuclei as in insulators or in the valence band (and lower bands) of semiconductors, or being "free", e.g. in metals or in the conduction band of semiconductors. All these electrons can respond to a (changing) magnetic field. Here we will only look at the (much simplified) case of a *bound electron orbiting a nucleus in a circular orbit*.
- The basic response of an *orbiting* electron to a changing magnetic field is a **precession** of the orbit, i.e. the polar vector describing the orbit now moves in a circle around the magnetic field vector ***H***:



  - The angular vector ω characterizing the blue orbit of the electron will experience a force from the (changing) magnetic field that forces it into a circular movement on the green cone.
  - Why do we emphasize "changing" magnetic fields? Because there is no way to bring matter into a magnetic field without changing it - either be switching it on or by moving the material into the field.
- What exactly happens to the *orbiting* electron? The reasoning given below follows the semi-classical approach contained within Bohr's atomic model. It gives essentially the right results *(in cgs units!)*.
  - The changing magnetic field, $\mathbf{d}\boldsymbol{H}/\mathbf{d}\boldsymbol{t}$, generates a force ***F*** on the orbiting electron via inducing a voltage and thus an electrical field ***E***. We can always express this as

$$\boldsymbol{F} = \mathbf{m}^*_\mathbf{e} \cdot \boldsymbol{a} = \mathbf{m}^*_\mathbf{e} \cdot \frac{\mathbf{dv}}{\mathbf{d}\boldsymbol{t}} := \mathbf{e} \cdot \boldsymbol{E}$$

  - With ***a*** = acceleration = $\mathbf{dv}/\mathbf{d}\boldsymbol{t} = \mathbf{e} \cdot \boldsymbol{E}/\mathbf{m}^*_\mathbf{e}$.
- Since $\mathbf{d}\boldsymbol{H}/\mathbf{d}\boldsymbol{t}$ primarily induces a voltage ***V***, we have to express the field strength ***E*** in terms of the induced voltage ***V***. Since the electron is orbiting and experiences the voltage during *one* orbit, we can write:

$$\boldsymbol{E} = \frac{\boldsymbol{V}}{\boldsymbol{L}}$$

  - With ***L*** = length of orbit = $\mathbf{2}\pi \cdot \boldsymbol{r}$, and ***r*** = radius of orbit.
  - ***V*** is given by the basic equations of induction, it is

$$V = -\frac{d\Phi}{dt}$$

- With Φ = magnetic flux = $\boldsymbol{H} \cdot \boldsymbol{A}$; and $\boldsymbol{A}$ = area of orbit = $\pi \cdot r^2$. The *minus sign* is important, it says that the *effect* of a changing magnetic fields will be opposing the *cause* in accordance with **Lenz's law**.

Putting everything together we obtain

$$\frac{dv}{dt} = \frac{e \cdot E}{m^*_e} = \frac{V \cdot e}{L \cdot m^*_e} = -\frac{e \cdot r}{2\, m^*_e} \cdot \frac{dH}{dt}$$

The total change in **v** will be given by integrating:

$$\Delta v = \int_{v_1}^{v_2} dv = -\frac{e \cdot r}{2m^*_e} \cdot \int_0^H dH = -\frac{e \cdot r \cdot H}{2\, m^*_e}$$

The magnetic moment m_{orb} of the undisturbed electron was $m_{orb} = ½ \cdot e \cdot v \cdot r$

- By changing **v** by Δ**v**, we change m_{orb} by Δm_{orb}, and obtain

$$\Delta m_{orb} = \frac{e \cdot r \cdot \Delta v}{2} = -\frac{e^2 \cdot r^2 \cdot H}{4m^*_e}$$

That is more or less the *equation for diamagnetism* in the primitive electron orbit model.

- What comes next is to take into account that the magnetic field does not have to be perpendicular to the orbit plane and that there are many electrons. We have to add up the single electrons and average the various effects.
- Averaging over all possible directions of $\boldsymbol{H}$ (taking into account that a field in the plane of the orbit produces zero effect) yields for the *average* induced magnetic moment almost the same formula:

$$\Delta m_{orb} = \langle \Delta m_{orb} \rangle = -\frac{e^2 \cdot \langle r \rangle^2 \cdot H}{6m^*_e}$$

- $\langle r \rangle$ denotes that we average over the orbit radii at the same time

Considering that not just *one*, but all $\boldsymbol{z}$ electrons of an atom participate, we get the final formula:

$$\Delta m = \langle \Delta m_{orb} \rangle = -\frac{e^2 \cdot z \cdot r^2 \cdot H}{6\, m^*_e}$$

The additional magnetization $\boldsymbol{M}$ caused by $\Delta \boldsymbol{m}$ is *all the magnetization there is for diamagnets*; we thus we can drop the Δ and get

$$M_{Dia} = \frac{\langle \Delta m \rangle}{V}$$

With the definition for the magnetic susceptibility $\chi = \boldsymbol{M}/\boldsymbol{H}$ we finally obtain for the relevant material parameter for diamagnetism

$$\chi_{dia} = -\frac{e^2 \cdot z \cdot \langle r \rangle^2}{6\, m^*_e \cdot V} = -\frac{e^2 \cdot z \cdot \langle r \rangle^2}{6\, m^*_e} \cdot \rho_{atom}$$

- With ρ_{atom} = number of atoms per unit volume

Plugging in numbers will yield χ values around $\mathbf{-(10^{-5} - 10^{-7})}$ in good agreement with experimental values.

4.2.2 Paramagnetism

- The treatment of paramagnetism in the most simple way is exactly identical to the treatment of orientation polarization. All you have to do is to replace the *electric dipoles* by magnetic dipoles, which we call *magnetic moments*.
  - We have permanent dipole moments in the material, they have no or negligible interaction between them, and they are free to point in any direction *even in solids*.
  - This is a major difference to electrical dipole moments which can *only* rotate if the whole atom or molecule rotates; i.e. only in liquids. This is why the treatment of magnetic materials focusses on ferromagnetic materials and why the underlying symmetry of the math is not so obvious in real materials.
  - In an external magnetic field the magnetic dipole moments have a tendency to orient themselves into the field direction, but this tendency is opposed by the thermal energy, or better entropy of the system.
- Using exactly the same line of argument as in the case of orientation polarization, we have for the potential energy W of a magnetic moment (or dipole) m in a magnetic field H

$$W(\varphi) = -\ \mu_0 \cdot \underline{m} \cdot \underline{H} = -\ \mu_0 \cdot m \cdot H \cdot \cos\varphi$$

  - With φ = angle between $\underline{H}$ and $\underline{m}$.
- In *thermal equilibrium*, the number of magnetic moments with the energy W will be $N[W(\varphi)]$, and that number is once more given by the Boltzmann factor:

$$N[W(\varphi)] = c \cdot \exp -(W/kT) = c \cdot \exp \frac{m \cdot \mu_0 \cdot H \cdot \cos\varphi}{kT} = N(\varphi)$$

  - As befoere, c is some as yet undetermined constant.
- As before, we have to take the component in field direction of all the moments having the same angle with H and integrate that over the unit sphere. The result for the induced magnetization m_{ind} and the total magnetization M is the same as before for the induced dielectric dipole moment:

$$m_{ind} = m \cdot \left(\coth\beta - \frac{1}{\beta} \right)$$

$$M = N \cdot m \cdot L(\beta)$$

$$\beta = \frac{\mu_0 \cdot m \cdot H}{kT}$$

  - With $L(\beta)$ = Langevin function = $\coth\beta - 1/\beta$

- The only interesting point is the *magnitude* of β. In the case of the orientation polarization it was ≤ 1 and we could use a simple approximation for the Langevin function.
  - We know that m will be of the order of magnitude of 1 Bohr magneton. For a rather large magnetic field strength of $5 \cdot 10^6$ A/m, we obtain as an estimate for an upper limit $\beta = 1.4 \cdot 10^{-2}$, meaning that the range of β is even smaller as in the case of the electrical dipoles.
  - We are thus justified to use the simple approximation $L(\beta) = \beta/3$ and obtain

$$M = N \cdot m \cdot (\beta/3) = \frac{N \cdot m^2 \cdot \mu_0 \cdot H}{3kT}$$

  - The paramagnetic susceptibility $\chi = M/H$, finally, is

$$\chi_{para} = \frac{N \cdot m^2 \cdot \mu_0}{3kT}$$

- Plugging in some typical numbers (A Bohr magneton for ***m*** and typical densities), we obtain $\chi_{para} \approx +10^{-3}$; i.e. an *exceedingly small* effect, but with certain characteristics that will carry over to ferromagnetic materials:
  - There is a strong temperature dependence and it follows the "**Curie law**":

$$\chi_{para} = \frac{const}{T}$$

  - Since ferromagnets of all types turn into paramagnets above the Curie temperature T_C, we may simply expand Curies law for this case to

$$\chi_{ferro}(T > T_C) = \frac{const^*}{T - T_C}$$

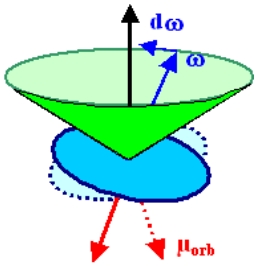
- In summary, paramagnetism, stemming from some (small) average alignement up of permanent magnetic dipoles associated with the atoms of the material, *is of no (electro)technical consequence*. It is, however, important for analytical purposes called "**Electron spin resonance**" (***ESR***) techniques.
- There are other types of paramagnetism, too. Most important is, e.g., the **paramagnetism of the free electron gas**. Here we have magnetic moments associated with spins of electrons, but in a *mobile* way - they are not fixed at the location of the atoms
  - But as it turns out, other kinds of paramagnetism (or more precisely: calculations taking into account that magnetic moments of atoms can not assume any orientation but only sone quantized ones) do not change the general picture: *Paramagnetism is a weak effect*.

4.2.3 Summary to: Dia- and Paramagnetism

- Dia- and Paramagentic propertis of materials are of no consequence whatsoever for products of electrical engineering (or anything else!)
  - Only their common denominator of being essentially "non-magnetic" is of interest (for a submarine, e.g., you want a non-magnetic steel)
  - For research tools, however, these forms of magnitc behavious can be highly interesting ("paramagentic resonance")

Normal diamagnetic materials: $\chi_{dia} \approx -(10^{-5} - 10^{-7})$
Superconductors (= ideal diamagnets): $\chi_{SC} = -1$
Paramagnetic materials: $\chi_{para} \approx +10^{-3}$

- Diamagnetism can be understood in a semiclassical (Bohr) model of the atoms as the response of the current ascribed to "circling" electrons to a changing magnetic field via classical induction ($\propto$ **dH/dt**).
  - The net effect is a precession of the circling electron, i.e. the normal vector of its orbit plane circles around on the green cone. ⇒
  - The "Lenz rule" ascertains that inductive effects oppose their source; diamagnetism thus weakens the magnetic field, $\chi_{dia} < 0$ must apply.



- Running through the equations gives a result that predicts a very small effect. ⇒ A proper quantum mechanical treatment does not change this very much.

$$\chi_{dia} = -\frac{e^2 \cdot z \cdot \langle r \rangle^2}{6\, m^*_e} \cdot \rho_{atom} \approx -(10^{-5} - 10^{-7})$$

- The formal treatment of paramagnetic materuials is mathematically completely identical to the case of orientation polarization
  - The range of realistc β values (given by largest ***H*** technically possible) is even smaller than in the case of orientation polarization. This allows tp approximate **L(β)** by **β/3**; we obtain:

$$\chi_{para} = \frac{N \cdot m^2 \cdot \mu_0}{3kT}$$

  - Insertig numbers we find that χ_{para} is indeed a number just slightly larger than **0**.

$$W(\varphi) = -\mu_0 \cdot \underline{m} \cdot \underline{H} = -\mu_0 \cdot m \cdot H \cdot \cos\varphi$$

Energy of magetic dipole in magnetic field

$$N[W(\varphi)] = c \cdot \exp -(W/kT) = c \cdot \exp \frac{m \cdot \mu_0 \cdot H \cdot \cos\varphi}{kT} = N(\varphi)$$

(Boltzmann) Distribution of dipoles on energy states

$$M = N \cdot m \cdot L(\beta)$$

$$\beta = \frac{\mu_0 \cdot m \cdot H}{kT}$$

Resulitn Magnetization with Langevin function **L(β)** and argument β

4.3 Ferromagnetism

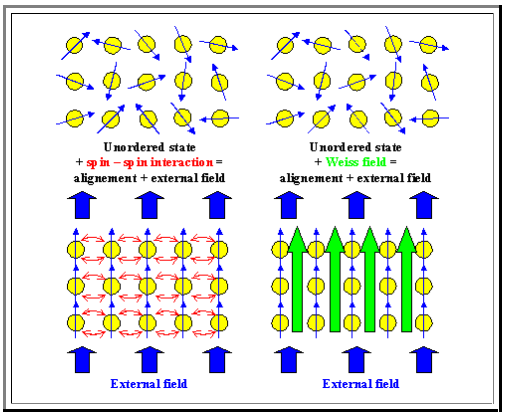
4.3.1 Mean Field Theory of Ferromagnetism

The Mean Field Approach

- In contrast to dia- and paramagnetism, *ferromagnetism* is of *prime importance* for electrical engineering. It is, however, one of the most difficult material properties to understand.
  - It is not unlike "ferro"electricity, in relying on strong interactions between neighbouring atoms having a permanent magnetic moment ***m*** stemming from the *spins* of electrons.
  - But while the interaction between electric dipoles can, at least in principle, be understood in classical and semi-classical ways, *the interaction between spins of electrons is an exclusively quantum mechanical effect with no classical analogon*. Moreover, a theoretical treatment of the three-dimensional case giving reliable results still eludes the theoretical physicists.
  - Originally I planned to present a very simplified theory of ferromagnetism based on spin-spin interaction - but forget it. It's rather complex but doesn't go very far. Just accept the fact that only **Fe**, **Co**, **Ni** (and some rare earth metals) show strong interactions between spins to result in (room temperature) ferromagnetism of elemental crystals.
  - In *compounds*, however, many more substances exist with spontaneous magnetization coming from the coupling of spins.
- There is, however, a relatively *simple theory of ferromagnetism*, that gives the proper relations, temperature dependences etc., - with one major drawback: It starts with an *unphysical assumption*.
  - This is the **mean field theory** or the **Weiss theory** of ferromagnetism. It is a phenomenological theory based on a central (wrong) assumption:

> **Substitute the elusive spin - spin interaction between electrons**
> **by the interaction of the spins with a very strong magnetic field.**

  - In other words, *pretend*, that in addition to your external field there is a *built-in magnetic field* which we will call the **Weiss field**. The Weiss field will tend to line up the magnetic moments - you are now treating ferromagnetism as an *extreme* case of paramagnetism. The scetch below illustrates this



- Of course, if the material you are looking at *is* a real ferromagnet, you don't have to *pretend* that there is a built-in magnetic field, because there *is* a large magnetic field, indeed. But this looks like mixing up cause and effect! What you want to result from a calculation is what you start the calculation with!
  - This is called a self-consistent approach. You may view it as a closed circle, where cause and effect loose their meaning to some extent, and where a calculation produces some results that are fed back to the beginning and repeated until some parameter doesn't change anymore.
  - Why are we doing this, considering that this approach is rather questionable? Well - it works! It gives the right relations, in particular the temperature dependence of the magnetization.
- The local magnetic field H_{loc} for an external field H_{ext} then will be

$$H_{loc} = H_{ext} + H_{Weiss}$$

- Note that this has not much to do with the local electrical field in the Lorentz treatment. We call it "local" field, too, because it is supposed to contain everything that acts *locally*, including the modifications we ought to make to account for effects as in the case of electrical fields. But since our fictitious "Weiss field" is so much larger than everything coming from real fields, we simply can forget about that.

- Since we treat this *fictive* field H_{Weiss} as an internal field, we write it as a superposition of the external field H and a field stemming from the internal magnetic polarization J:

$$H_{loc} = H_{ext} + w \cdot J$$

  - With J = magnetic polarization and w = **Weiss's factor**; a constant that now *contains the physics of the problem*.

- This is the decisive step. We now identify the Weiss field with the magnetic polarization that is caused by it. And, yes, as stated above, we now do mix up cause and effect to some degree: the fictitiuos Weiss field causes the alignments of the individual magnetic moments which than produce a magnetic polarization that causes the local field that we identify with the Weiss field and so on.
  - But that, after all, *is* what happens: the (magnetic moments of the) spins interact causing a field that causes the interaction, that ....and so on . If your mind boggles a bit, that is as it should be. The magnetic polarization caused by spin-spin interactions and mediating spin-spin interaction just *is* - asking for cause and effect is a futile question.
  - The Weiss factor w now contains *all the local effects* lumped together - in analogy to the Lorentz treatment of local fields, μ_0, and the interaction between the spins that leads to ferromagnetism as a result of some fictive field.
  - But lets be very clear: *There is no internal magnetic field H_{Weiss} in the material* before the spins become aligned. This completely fictive field just leads - within limits - to the same interactions you would get from a proper quantum mechanical treatment. Its big advantage is that it makes calculations possible if you determine the parameter w experimentally.

- All we have to do now is to repeat the calculations done for paramagnetism, substituting H_{loc} wherever we had H. Lets see where this gets us.

Orientation Polarization Math with the Weiss Field

- The potential energy W of a magnetic moment (or dipole) $\underline{\mathbf{m}}$ in an external magnetic field $\underline{H}$ now becomes

$$W = - m \cdot \mu_0 \cdot (H + H_{Weiss}) \cdot \cos\varphi$$

$$= - m \cdot \mu_0 \cdot (H + w \cdot J) \cdot \cos\varphi$$

- The Boltzmann distribution of the energies now reads

$$N(W) = c \cdot \exp - \frac{W}{kT} = c \cdot \exp \frac{m \cdot \mu_0 \cdot (H + w \cdot J) \cdot \cos\varphi}{kT}$$

- The Magnetization becomes

$$M = N \cdot m \cdot L(\beta)$$

$$= N \cdot m \cdot L\left(\frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}\right)$$

  - In the last equation the argument of $L(\beta)$ is spelled out; it is quite significant that β contains $w \cdot J$.

- The total polarization is $J = \mu_0 \cdot M$, so we obtain the final equation

$$J = N \cdot m \cdot \mu_0 \cdot L(\beta) = N \cdot m \cdot \mu_0 \cdot L\left(\frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}\right)$$

- Written out in full splendor this is

$$J = N \cdot m \cdot \mu_0 \cdot \coth\left(\frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}\right) - \frac{N \cdot kT}{(H + w \cdot J)}$$

- What we really want is the magnetic polarization ***J*** as a function of the external field ***H***. Unfortunately we have a *transcendental* equation for ***J*** which can not be written down directly without a "***J***" on the right-hand side.
  - What we also like to have is the value of the spontaneous magnetization ***J*** for no external field, i.e. for **H = 0**. Again, there is no analytical solution for this case.
  - There is an easy graphical solution, however: We actually have *two* equations for which must hold *at the same time*:
  - The argument β of the Langevin function is

$$\beta = \frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}$$

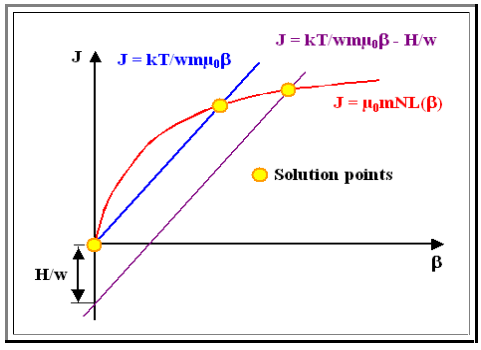
  - Rewritten for ***J***, we get our first equation:

$$J = \frac{kT \cdot \beta}{w \cdot m \cdot \mu_0} - \frac{H}{w}$$

  - This is simply a straight line with a slope and intercept value determined by the interesting variables ***H***, ***w***, and ***T***.
- On the other hand we have the equation for ***J***, and this is our second independent equation
.

$$J = N \cdot m \cdot \mu_0 \cdot L(\beta) = N \cdot m \cdot \mu_0 \cdot L\left(\frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}\right)$$

  - This is simply the Langevin function which we know for any numerical value for β
- All we have to do is to draw *both* functions in a ***J*** - β diagram
  - We can do that by simply putting in some number for β and calculating the results. The intersection of the two curves gives the solutions of the equation for ***J***.
  - This looks like this



- Without knowing anything about β, we can draw a definite conclusion:
- For $H = 0$ we have *two* solutions (or none at all, if the straight line is too steep): One for $J = 0$ and one for a rather large J.
- It can be shown that the solution for $J = 0$ is unstable (it disappears for an arbitrarily small field H) so we are left with a *spontaneous large magnetic polarization* without an external magnetic field as the first big result of the mean field theory.

We can do much more with the mean field theory, however.

- *First*, we note that switching on an *external magnetic field* does not have a large effect. J increases somewhat, but for realistic values of Hlw the change remains small.
- *Second*, we can look at the *temperature dependence* of J by looking at the straight lines. For $T \rightarrow 0$, the intersection point moves all the way out to infinity. This means that all dipoles are now lined up in the field and $L(\beta)$ becomes **1**. We obtain the *saturation value* J_{sat}

$$J_{sat} = N \cdot m \cdot \mu_0$$

- *Third*, we look at the effect of increasing *temperatures*. Raising T increases the slope of the straight line, and the two points of intersection move together. When the slope is equal to the slope of the Langevin function (which, as we know, is **1/3**), the two points of solution merge at $J = 0$; if we increase the slope for the straight line even more by increasing the temperature by an incremental amount, solutions do no longer exist and the spontaneous magnetization disappears.

This means, there is a *critical temperature* above which ferromagnetism disappears. This is, of course, the **Curie temperature** T_C.

- At the Curie temperature T_C, the slope of the straight line and the slope of the Langevin function for $\beta = 0$ must be identical. In formulas we obtain:

$$\frac{dJ}{d\beta} = \frac{kT_C}{w \cdot m \cdot \mu_0} = \text{slope of the straight line}$$

$$\left.\frac{dJ}{d\beta}\right|_{\beta = 0} = N \cdot m \cdot \mu_0 \cdot \frac{dL(\beta)}{d\beta} = \frac{N \cdot m \cdot \mu_0}{3}$$

- We made use of our old insight that the slope of the Langevin function for $\beta \rightarrow 0$ is **1/3**.

Equating both slopes yields for T_C

$$T_C = \frac{N \cdot m^2 \cdot \mu_0^2 \cdot w}{3k}$$

This is pretty cool. We did not solve an transcendental equation nor go into deep quantum physical calculations, but still could produce rather simple equations for prime material parameters like the Curie temerature.

- If we only would know w, the Weiss factor! Well, we do *not* know w, but now we can turn the equation around: If we know T_C, we can *calculate* the Weiss factor w and thus the *fictive magnetic field* that we need to keep the spins in line.
- In **Fe**, for example, we have $T_C = 1043$ K, $m = 2{,}2 \cdot m_{Bohr}$. It follows that

$$H_{Weiss} = w \cdot J = 1.7 \cdot 10^9 \text{ A/m}$$

- This is a truly *gigantic* field strength telling us that quantum mechanical spin interactions, if existent, are not to be laughed at.
- If you do not have a feeling of what this number means, consider the unit of H: A field of $1{,}7 \cdot 10^9$ **A/m** is produced if a current of $1{,}7 \cdot 10^9$ **A** flows through a loop (= coil) with **1 m²** area. Even if you make the loop to cover only **1 cm²**, you still need $1{,}7 \cdot 10^5$ **A**.

We can go one step further and approximate the Langevin function again for temperatures $> T_C$, i.e. for $\beta < 1$ by

$$L(\beta) \approx \frac{\beta}{3}$$

- This yields

$$J(T > T_C) \approx \frac{N \cdot m^2 \cdot \mu_0^2}{3kT} \cdot (H + w \cdot J)$$

- From the equation for T_C we can extract w and insert it, arriving at

$$J(T > T_C) \approx \frac{N \cdot m^2 \cdot \mu_0^2}{3k(T - T_C)} \cdot H$$

- Dividing by H gives the susceptibility χ for $T > T_C$ and the final formula

$$\chi = \frac{J}{H} = \frac{N \cdot m^2 \cdot \mu_0^2}{3k \cdot (T - T_C)} = \frac{\text{const.}}{T - T_C}$$

- This is the famous Curie law for the paramagnetic regime at high temperatures which was a phenomenological thing so far. Now we derived it with a theory and will therefore call it **Curie - Weiss law**.

In summary, the mean field approach ain't that bad! It can be used for attacking many more problems of ferromagnetism, but you have to keep in mind that it is only a description, and not based on sound principles.

4.3.2 Beyond Mean Field Theory

Some General Considerations

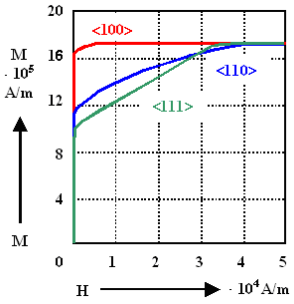
- According to the mean field theory, if a material is ferromagnetic, *all* magnetic moments of the atoms would be coupled and point in the same direction. We now ask a few questions:
  - **1.** *Which direction* is that going to be for a material just sitting there? Is there some preferred internal direction or are all directions equal? In other words: Do we have to make the fictitious Weiss field $\boldsymbol{H_{Weiss}}$ larger in some directions compared to other ones? Of course, we wonder if some crystallographic directions have "special status".
  - **2.** What happens if an external field is superimposed in some direction that does *not* coincide with a preferred internal direction?
  - **3.** What happens if it *does*? Or if the external field is *parallel* to the internal one, but pointing in the *opposite* direction?
- The (simple) mean field theory remains rather silent on those questions. With respect to the first one, the internal alignment direction would be determined by the direction of the fictive field $\boldsymbol{H_{Weiss}}$, but since this field does not really exist, each direction seems equally likely.
  - In real materials, however, we might expect that the direction of the magnetization is not totally random, but has some specific preferences. This is certainly what we must expect for *crystals*.
  - A specific direction in real ferromagnetic materials could be the result of crystal anisotropies, inhomogeneities, or external influences - none of which are contained within the mean field theory (which essentially treats perfectly isotropic and infinitely large materials).
- *Real* ferromagnetic materials thus are more complicated than suggested by the mean field theory - for a very general reason:
- Even if we can lower the internal *energy* $\boldsymbol{U}$ of a crystal by aligning magnetic moments, we still must keep in mind that the aim is always to minimize the *free enthalpy* $\boldsymbol{G = U - TS}$ of the *total* system.
  - While the *entropy* part coming from the degree of orderliness in the system of magnetic moments has been taken care of by the general treatment in the frame work of the orientation polarization, we must consider the enthalpy (or energy) $\boldsymbol{U}$ of the *system* in more detail. So far we only minimized $\boldsymbol{U}$ with respect to *single* magnetic moments in the Weiss field.
  - This is so because the mean field approach essentially relied on the fact that by aligning the spins relative to the (fictitious) Weiss field, we lower the energy of the individual spin or magnetic moments as treated before by some energy $\boldsymbol{W_{align}}$. We have

$$\boldsymbol{U_{align} = U_{random} - W_{align}}$$

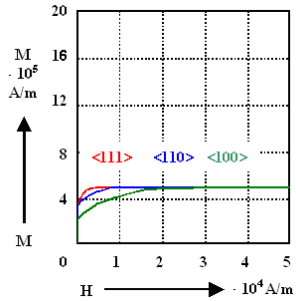
  - But, as discussed above, real materials are mostly (poly)crystals and we must expect that the real (quantum-mechanical) interaction between the magnetic moments of the atoms are different for different directions in the crystal. There is some anisotropy that must be considered in the $\boldsymbol{U_{align}}$ part of the free enthalpy.
  - Moreover, there are other contributions to $\boldsymbol{U}$ not contained in the mean field approach. Taken everything together makes quantitative answers to the questions above exceedingly difficult.
- There are, however, a few relatively simple general rules and experimental facts that help to understand what really happens if a ferromagnetic material is put into a magnetic field. Let's start by looking at the **crystal anisotropy**.

Crystal Anisotropy

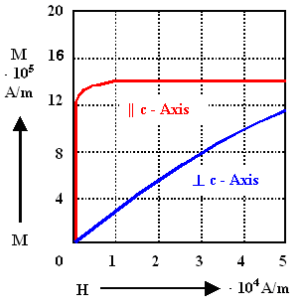
- Generally, we must expect that there is a preferred crystallographic direction for the spontaneous magnetization, the so-called "**easy directions**". If so, it would need some energy to change the magnetization direction into some other orientations; the "**hard directions**".
  - That effect, if existent, is easy to measure: Put a single crystal of the ferromagnetic material in a magnetic field $\boldsymbol{H}$ that is oriented in a certain crystal direction, and measure the magnetization of the material in that direction:
  - If it happens to be an *easy* direction, you should see a strong magnetization that reaches a **saturation** value - obtained when all magnetic moments point in the desired direction - already at low field strength $\boldsymbol{H}$. If, on the other hand, $\boldsymbol{H}$ happens to be in a *hard* direction, we would expect that the magnetization only turns into the $\boldsymbol{H}$ direction ***reluctantly***, i.e. only for large values of $\boldsymbol{H}$ will we find saturation.
- This is indeed what is observed, classical data for the elemental ferromagnets **Fe**, **Ni**, **Co** are shown below:



Anisotropy of magnetization in **Fe**.



Anisotropy of magnetization in **Ni**.



Anisotropy of magnetization in **Co**.

Anisotropy of the magnetization in **Fe** (***bcc*** lattice type), **Ni** (***fcc*** lattice type), and **Co** (***hcp*** lattice type).

- The curves are easy to interpret qualitatively along the lines stated above; consider, e.g., the **Fe** case:
  - For field directions not in **<100>**, the spins become aligned in the **<100>** directions pointing as closely as possible in the external field direction.
  - The magnetization thus is just the component of the **<100>** part in the field direction; it is obtained for arbitrarily small external fields.
  - Increasing the magnetization, however, means turning spins into a "hard" directions, and this will proceed reluctantly for large magnetic fields.
  - At sufficiently large fields, however, all spins are now aligned into the external field directions and we have the same magnetization as in the easy direction.
- The curves above contain the material for a simple little exercise:

Exercise 4.3-2

Magnetic moments of Fe, Ni, Co

Questionaire

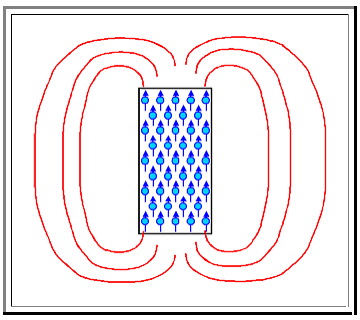
Multiple Choice questions to 4.3.2

4.3.3 Magnetic Domains

Reducing the External Magnetic Field

If we now turn back to the question of what you would observe for the magnetization of a single crystal of ferromagnetic material just sitting on your desk, you would now expect to find it completely magnetized in its *easy direction* - even in the presence of a not overly strong magnetic field.

- This would look somewhat like this:



- There would be a large *internal* magnetization *and* a large *external* magnetic field ***H*** - we would have an ideal **permanent magnet**.
- And we also would have a high energy situation, because the external magnetic field around the material contains magnetic field energy W_{field}.
- In order to make life easy, we do not care how large this energy is, even so we could of course calculate it. We only care about the general situation: We have a rather large energy outside the material caused by the perfect line up of the magnetic dipoles in the material
- How do we know that the field energy is rather large? Think about what will happen if you put a material as shown in the picture next to a piece of iron for example.
- What we have is obviously a strong **permanent magnet**, and as we know it will violently attract a piece of iron or just any ferromagnetic material. That means that the external magnetic field is strong enough to line up all the dipoles in an other ferromagnetic material, and that, as we have seen, takes a considerable amount of energy.

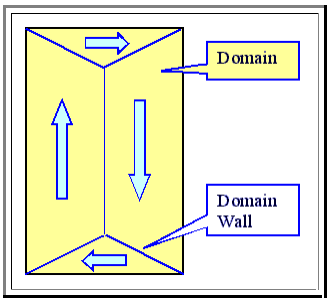
The **internal energy *U*** of the system thus must be written

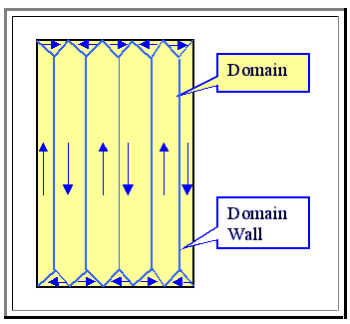
$$U_{align} = U_{random} - W_{align} + W_{field}$$

- The question is if we can somehow lower W_{field} substantially - possibly by spending some smaller amount of energy elsewhere. Our only choice is to *not* exploit the maximum alignment energy W_{align} as it comes from perfect alignment in *one* direction.
- In other words, are there non-perfect alignments patterns that only cost *a little bit* W_{align} energy, but safe *a lot* of W_{field} energy? Not to mention that we always gain a bit in entropy by not being perfect?

The answer is *yes* - we simply have to introduce **magnetic domains**.

- Magnetic domains are regions in a crystal with different directions of the magnetizations (but still pointing in one of the easy directions); they must by necessity be separated by **domain walls**. The following figures show some possible configurations





Both domain structures *decrease* the external field and thus W_{field} because the flux lines now can close inside the material. And we kept the alignment of the magnetic moments in most of the material, it only is disturbed in the domain walls. Now which one of the two configurations shown above is the better one?

- Not so easy to tell. With *many* domains, the magnetic flux can be confined better to the inside of the material, but the total domain wall area goes up - we loose plenty of W_{align}.
- The energy lost by non-perfect alignment in the domains walls can be expressed as a property of the domain wall, as a **domain wall energy**. A magnetic domain wall, by definition a two-dimensional defect in the otherwise perfect order, thus carries an energy (per cm^2) like any other two-dimensional defect.

- There must be an optimum balance between the energy gained by *reducing* the external field, and the energy lost in the domain wall energy. And all the time we must remember that the magnetization in a domain is always in an easy direction (without strong external fields).

We are now at the ***end of our tether***. While the ingredients for minimizing the system energy are perfectly clear, nobody can calculate exactly what kind of stew you will get for a given case.

- *Calculating* domain wall energies from first principles is already nearly hopeless, but even with experimental values and for perfect single crystals, it is not simple to deduce the *domain structure* taking into account the anisotropy of the crystal and the external field energy.

And, too make things even worse (for theoreticians) there are even *more* energetic effects that influence the domain structure. Some are important and we will give them a quick look.

Magnetostriction and Interaction with Crystal Lattice Defects

The interaction between the magnetic moments of the atoms that produces alignment of the moments - ferromagnetism, ferrimagnetism and so on - necessarily acts as a force between the atoms, i.e. the interaction energy can be seen as a potential and the (negative) derivative of this potential is a *force*.

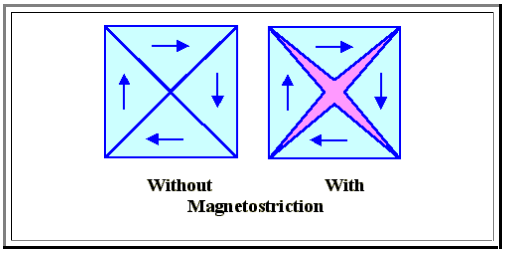
- This interaction force must be added to the general binding forces between the atoms.
- In general, we must expect it to be *anisotropic* - but not necessarily in the same way that the binding energy could be anisotropic, e.g. for covalent bonding forces.
- The total effect thus usually will be that the lattice constant is *slightly* different in the direction of the magnetic moment. A *cubic* crystal may become orthorhombic upon magnetization, and the crystal *changes dimension* if the direction of the magnetization changes.

A crystal "just lying there" will be magnetized in several directions because of its magnetic domains and the anisotropy of the lattice constants averages out: A cubic crystal is still - *on average* - cubic, but with a slightly changed lattice constant.

- However, if a large external field $\boldsymbol{H_{ex}}$ forces the internal magnetization to become oriented in field directions, the material now (usually) responds by some *contraction in field direction* (no more averaging out); this effect is called **magnetostriction**. This word is generally used for the description of the effect that the interatomic distances are different if magnetic moments are aligned.
- The *amount* of magnetostriction is different for different magnetic materials, again there are no straight forward calculations and experimental values are used. It is a complex phenomena.
- Magnetostriction is a useful property; especially since recently "giant magnetostriction" has been discovered. Technical uses seem to be just around the corner at present.

Magnetostriction also means that a piece of crystal that contains a magnetic domain would have a somewhat *different* dimension as compared to the same piece without magnetization.

- Lets illustrate that graphically with an (oversimplified, but essentially correct) picture:



- In this case the magnetostriction is perpendicular to the magnetization. The four domains given would assume the shape shown on the right hand side.

Since the crystal does not come apart, there is now some *mechanical strain and stress* in the system. This has two far reaching consequences

- **1.** We have to add the *mechanical energy* to the energy balance that determines the domain structure, making the whole thing even more complicated.
- **2.** We will have an *interaction* of domain walls with structural defects that introduce mechanical stress and strain in the crystal. If a domain wall moves across a dislocation, for example, it might relieve the stress introduced by the dislocation in one position and increase it in some other position. Depending on the signs, there is an attractive or repelling force. In any case, there is some interaction: *Crystal lattice defects attract or repulse domain walls*.

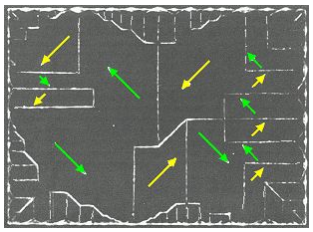
Generally speaking, both the domain structure and the movement of domain walls will be influenced by the internal structure of the material. A rather perfect single crystal may behave magnetically quite differently from a polycrystal full of dislocations.

- This might be hateful to the *fundamentalists* among the physicists: There is not much hope of calculating the domain structure of a given material from first principles and even less hope for calculating what happens if you deform it mechanically or do something else that changes its internal structure.
- However, we have the *engineering point of view*:

This is great

- The complicated situation with respect to domain formation and movement means that there are *many ways* to influence it.
- We do not have to live with a few materials and take them as they are, we have many options to tailor the material to specific needs. Granted, there is not always a systematic way for optimizing magnetic materials, and there might be much trial and error - but progress is being made.

What a real domain structure looks like is shown in the picture below. Some more can be found in the link.



- We see the domains on the surface of a single crystalline piece of **Ni**. How domains can be made visible is a long story - it is not easy! We will not go into details here.

Summarizing what we have seen so far, we note:

- **1.** The **domain structure** of a given magnetic material *in equilibrium* is the result of *minimizing the free enthalpy* mostly with respect to the *energy* term.
- **2.** There are several contributions to the energy; the most important ones being magnetic stray fields, magnetic anisotropy, magnetostriction and the interaction of the internal structure with these terms.
- **3.** The domain structure can be very complicated; it is practically impossible to calculate details. Moreover, as we will see, it is not necessarily always the equilibrium structure!

But this brings us to the next subchapter, the movement of *domain walls* and the *hysteresis curve*.

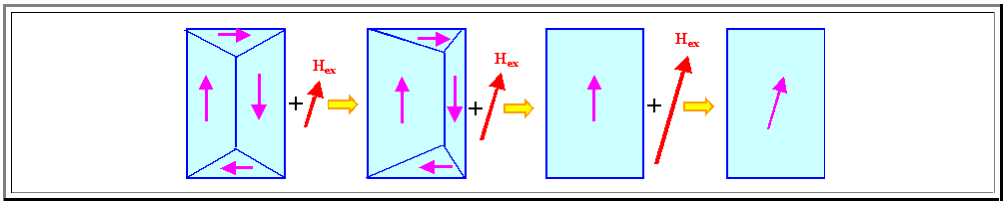
Questionaire

Multiple Choice questions to 4.3.3

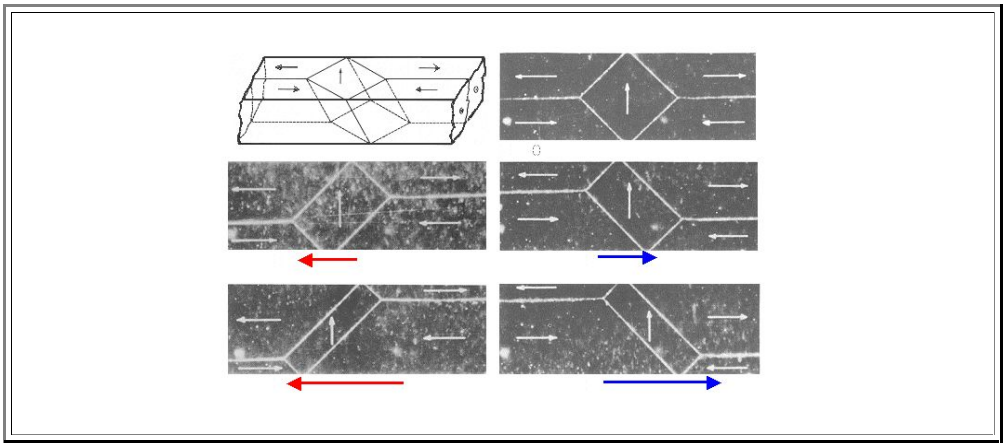
4.3.4 Domain Movement in External Fields

Domain Movement in External Fields

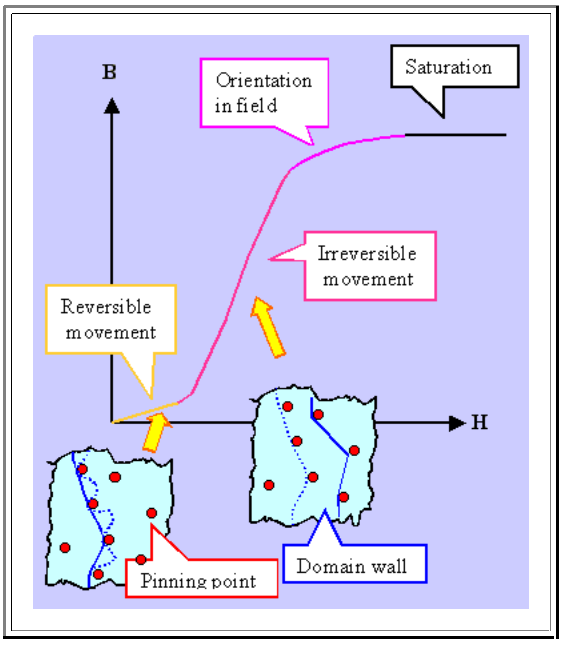
- What happens if we apply an external field to a ferromagnet with its equilibrium domain structure?
  - The domains oriented most closely in the direction of the external field will gain in energy, the other ones loose; always following the basic equation for the energy of a dipole in a field.
  - Minimizing the total energy of the system thus calls for increasing the size of favorably oriented domains and decreasing the size of unfavorably oriented ones. Stray field considerations still apply, but now we have an external field anyway and the stray field energy looses in importance.
  - We must expect that the most favorably oriented domain will win for large external fields and all other domains will disappear.
  - If we increase the external field beyond the point where we are left with only one domain, it may now even become favorable, to orient the atomic dipoles off their "easy" crystal direction and into the field.
  - After that has happened, all atomic dipoles are in field direction - more we cannot do. The magnetization than reaches a saturation value that cannot be increased anymore.
- Schematically, this looks like as shown below:



  - Obviously, domain walls have to move to allow the new domain structure in an external magnetic field.
- What this looks like in reality is shown below for a small single crystal of iron.



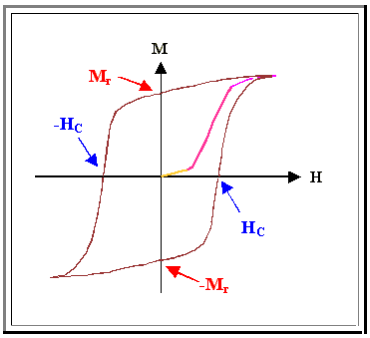
  - As noted before, domain walls interact with *stress* and strain in the lattice, i.e. with *defects* of all kinds. They will become "stuck" (the proper expression for things like that is "**pinned**") to defects, and it needs some force to ***pry*** them off and move them on. This force comes from the external magnetic field.
- The magnetization curve that goes with this looks like this:



- For small external fields, the domain walls, being pinned at some defects, just bulge out in the proper directions to increase favorably oriented domains and decrease the others. The magnetization (or the magnetic flux ***B***) increases about linearly with ***H***
- At larger external fields, the domain walls overcome the pinning and move in the right direction where they will become pinned by other defects. Turning the field of will not drive the walls back; the movement is irreversible.
- After just one domain is left over (or one big one and some little ones), increasing the field even more will turn the atomic dipoles in field direction. Since even under most unfavorable condition they were at most **45^{o}** off the external direction, the increase in magnetization is at most **1/cos(45^{o}) = 1.41**.
- Finally, saturation is reached. All magnetic dipoles are fully oriented in field direction, no further increase is possible.

If we switch off the external field anywhere in the irreversible region, the domain walls might relax back a little, but for achieving a magnetization of zero again, we must use force to move them back, i.e. an external magnetic field pointing in the opposite direction.

- In total we obtain the well known **hysteresis** behavior as shown in the **hysteresis curve** below.



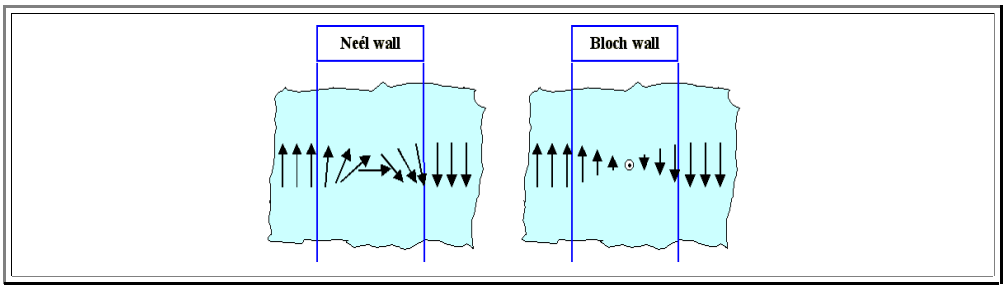
- The resulting hysteresis curve has two particular prominent features:
  - The remaining magnetization for zero external field, called the **remanence M_R**, and
  - the magnitude of the external field needed to bring the magnetization down to zero again. This is called **coercivity** or coercive field strength H_C.
- *Remanence* and *coercivity* are two numbers that describe the major properties of ferromagnets (and, of course, ferrimagnets, too). Because the exact shape of the hysteresis curve does not vary too much.
- Finally, we may also address the **saturation magnetization M_S** as a third property that is to some extent independent of the other two.

- Technical optimization of (ferro)magnetic materials first always focuses on these two numbers (plus, for reasons to become clear very soon, the resistivity).
- We now may also wonder about the dynamic behaviour, i.e. what happens if we change the external field with ever increasing frequency.

Domain Wall Structure

The properties of the domain walls, especially their interaction with defects (but also other domain walls) determine most of the magnetic properties of ferromagnets.

- What is the structure of a domain wall? How can the magnetization change from one direction to another one?
- There are two obvious geometric ways of achieving that goal - and that is also what really happens in practically all materials. This is shown below.



- What kind of wall will be found in real magnetic materials? The answer, like always is: Whichever one has the smallest (free) energy
  - In most bulk materials, we find the **Bloch wall**: the magnetization vector turns bit by bit like a screw out of the plane containing the magnetization to one side of the Bloch wall.
  - In thin layers (oft the same material), however, **Neél walls** will dominate. The reason is that Bloch walls would produce stray fields, while Neél walls can contain the magnetic flux in the material.
- Both basic types of domain walls come in many sub-types, e.g. if the magnetization changes by some defined angle other than **180^{o}**. In thin layers of some magnetic material, special domain structures may be observed, too.
- The interaction of domain walls with magnetic fields, defects in the crystal (or structural properties in amorphous magnetic materials), or intentionally produced structures (like "scratches", localized depositions of other materials, etc., can become fantastically complicated.
  - Since it is the domain structure together with the response of domain walls to these interactions that controls the hystereses curve and therefore the basic magnetic properties of the material, things are even more complicated as described before.
  - But do keep in mind: The underlying basic principles is the minimization of the free enthalpy, and there is nothing complicated about this. The fact that we can no easily write down the relevant equations, no to mention solving them, does not mean that we cannot understand what is going on. And the material has no problem in solving equations, it just assumes the proper structure, proving that there are solutions to the problem.

Questionaire

Multiple Choice questions to 4.3.4

4.3.5 Magnetic Losses and Frequency Behavior

General Remarks

- So far we have avoided to consider the frequency behavior of the magnetization, i.e. we did not discuss what happens if the external field oscillates!
  - The experience with electrical polarization can be carried over to some magnetic behaviour, of course. In particular, the frequency response of *paramagnetic* material will be quite similar to that of electric dipole orientation, and diamagnetic materials show close parallels to the electronic polarization frequency behaviour.
  - Unfortunately, this is of (almost) no interest whatsoever. The "almost" refers to **magnetic imaging** employing **magnetic resonance imaging** (***MRI***) or *nuclear spin resonance imaging* - i.e. some kind of "*computer tomography*". However, this applies to the paramagnetic behavior of the magnetic moments of the *nuclei*, something we haven't even discussed so far.
- What is of interest, however, is what happens in a *ferromagnetic* material if you have expose it to an *changing*, i.e. oscillating magnetic field. $\boldsymbol{H = H_0 \cdot \exp(i\omega t)}$
  - Nothing we discussed for dielectrics corresponds to this questions. Of course, the frequency behavior of ferroelectric materials would be comparable, but we have not discussed this topic.
  - Being wise from the case of dielectric materials, we suspect that the frequency behavior and some **magnetic energy losses** go in parallel, as indeed they do.
- In contrast to dielectric materials, we will start with looking at magnetic losses *first*.

Hystereses Losses

- If we consider a ferromagnetic material with a given hysteresis curve exposed to an oscillating magnetic field at low frequencies - so we can be sure that the internal magnetization can instantaneously follow the external field - we may consider *two* completely independent mechanisms causing losses.
  - **1.** The changing magnetic field induces currents wandering around in the material - so called ***eddy currents***. This is different from dielectrics, which we always took to be insulators: ferromagnetic materials are usually conductors.
  - **2.** The movement of domain walls needs (and disperses) some energy, these are the *intrinsic* magnetic losses or hystereses losses.
- Both effects add up; the energy lost is converted into heat. Without going into details, it is clear that the losses encountered increase with
  - **1.** The frequency f in *both* cases, because every time you change the field you incur the same losses per cycle.
  - **2.** The maximum magnetic flux B_{max} in both cases.
  - **3.** The conductivity $\sigma = 1/\rho$ for the eddy currents, and
  - **4.** The magnetic field strength H for the magnetic losses.
- More involved calculations (see the advanced module) give the following relation for the total ferromagnetic loss P_{Fe} per unit volume of the material

$$P_{Fe} \approx P_{eddy} + P_{hyst} \approx \frac{\pi^2 \cdot d^2}{6\rho} \cdot (f \cdot B_{max})^2 + 2f \cdot H_C \cdot B_{max}$$

  - With d = thickness of the material perpendicular to the field direction, H_C = coercivity.
- It is clear what you have to do to minimize the eddy current losses:
  - Pick a ferromagnetic material with a high resistivity - *if* you can find one. That is the point where *ferri***magnetic** materials come in. What you loose in terms of maximum magnetization, you may gain in reduced eddy losses, because many ferrimagnets are ceramics with a high resistivity.
  - Make d small by stacking insulated thin sheets of the (conducting) ferromagnetic material. This is, of course, what you will find in any run-of-the-mill transformer.
- We will not consider eddy current losses further, but now look at the remaining **hystereses losses** P_{hyst}
  - The term $H_C \cdot B_{max}$ is pretty much the area inside the hystereses curve. Multiply it with two times the frequency, and you have the hystereses losses in a good approximation.
  - In other words: There is nothing you can do - for a given material with its given hystereses curve.

- Your only choice is to select a material with a hystereses curve that is *just right*. That leads to several questions:
  - **1.** What kind of hystereses curve do I need for the application I have in mind?
  - **2.** What is available in terms of hystereses curves?
  - **3.** Can I change the hystereses curve of a given material in a defined way?
- The answer to these questions will occupy us in the next subchapter; here we will just finish with an extremely cursory look at the frequency behavior of ferromagnets.

Frequency Response of Ferromagnets

- As already mentioned, we only have to consider ferromagnetic materials - and that means the back-and-forth movement of domain walls in response to the changing magnetic field.
  - We do not have a direct feeling for how fast this process can happen; and we do not have any simplified equations, as in the case of dielectrics, for the forces acting on domain walls. Note that the atoms do *not* move if a domain wall moves - only the direction of the magnetic moment that they carry.
  - We know, however, from the bare fact that permanent magnets exist, or - in other words - that coercivities can be large, that it can take rather large forces to move domain walls - they might not shift easily.
  - This gives us at least a feeling: It will not be easy to move domain walls *fast* in materials with a large coercivity; and even for materials with low coercivity we must not expect that they can take large frequencies, e.g. in the optical region
  - There are materials, however, that still work in the **GHz** region.
- And that is where we stop. There simply is no general way to express the frequency dependence of domain wall movements.
  - That, however, does not mean that we cannot define a **complex magnetic permeability $\mu = \mu' + i\mu''$** for a particular magnetic material.
  - It can be done and it has been done. There simply is no *general* formula for it and that limits its general value.

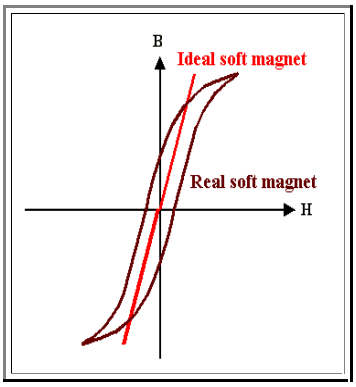
Questionaire

Multiple Choice questions to 4.3.5

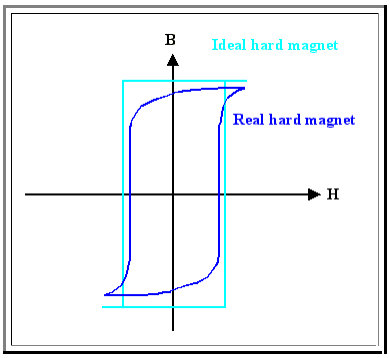
4.3.6 Hard and Soft Magnets

Definitions

- Lets quickly go over the three questions from the preceding sub-chapter
  - **1.** What kind of hystereses curve do I need for the application I have in mind?
- Lets look at two "paradigmatic" applications: A **transformer core** and a **magnetic memory**.
  - The transformer core is ferromagnetic in order to "transport" a large magnetic flux ***B*** produced by the primary coil to the secondary coil. What I want is that the induced flux ***B*** follows the primary field ***H*** as closely as possible.
  - In other words: There should be *no* hystereses loop - just a straight line, as shown below



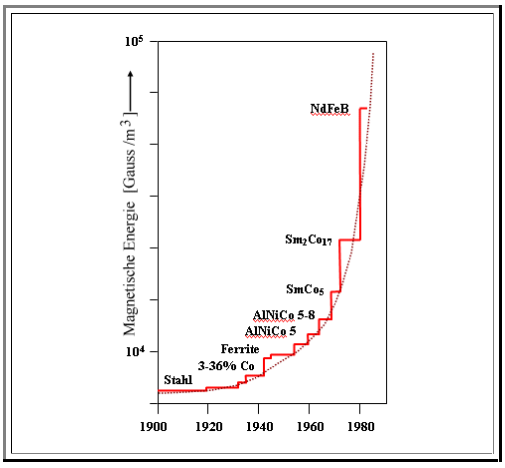
  - The ideal curve, without any hystereses, does not exist. What you get is something like the curve shown for a real **soft magnet** - because that is what we call a material with a kind of slender hystereses curve and thus small values of coercivity and remanence
  - If we switch on a positive field ***H*** and then go back to zero again, a little bit of magnetization is left. For a rather small reverse field, the magnetic flux reverses, too - the flux ***B*** follows ***H*** rather closely, if not exactly.
  - Hystereses losses are small, because the area enclosed in the hystereses loop is small.
  - But *some* losses remain, and the "transformer core" industry will be very happy if you can come up with a material that is just **1 %** or **2 %** "softer" than what they have now.
  - Beside losses, you have another problem: If you vary ***H*** sinusoidally, the output will be a somewhat distorted sinus, because ***B*** does not follow ***H*** linearly. This may be a problem when transforming *signals*.
- A soft magnetic material will obviously not make a good **permanent magnet**, because its remaining magnetization (its remanence) after switching off the magnetic field ***H*** is small.
  - But a permanent magnet is what we want for a **magnetic storage material**. Here we want to induce a large permanent magnetization by some external field (produced by the "writing head" of our storage device) that stays intact for many years if needs be. Some more information about magnetic storage can be found in an extra module
  - It should be *strong enough* - even so it is contained in a tiny area of the magnetic material on the tape or the storage disc - to produce a measurable effect if the reading head moves over it. It should not be *too strong*, however, because that would make it too difficult to erase it if we want to overwrite it with something else. In short, it should look like this



  - We can define what we want in terms of coercivity and remance. Ideally, the hystereses curve is very "square.
  - At some minimum field, the magnetization is rather large and does not change much anymore.
  - If we reverse the field direction, not much happens for a while, but as soon as we move above slightly above the coercivity value, the magnetization switches direction completely.
  - Ferromagnetic losses are unavoidable, we simply must live with them
- Pretty much all possible applications - consult the list in the next section - either calls for *soft* or for *hard* magnets; there isn't much in between.
  - So we now must turn to the second and third question:

Tailoring Hystereses Curves

- The question was: What is available in terms of *hystereses curves*? Good question; it immedately provokes another questions:
  - What is available in terms of ferromagnetic *materials*? The kind of hystereses behavior you get is first of all a property of the specific material you are looking at.
  - For arbitrary chemical compounds, there is little predictive power if they are ferromagnetic or not. In fact, the rather safe bet is that some compound *not* containing **Fe**, **Ni**, or **Co** is *not* ferromagnetic.
  - Even if we restrict ourselves to some compound or alloy containing at least one of the ferromagnetic elements **Fe**, **Ni** or **Co**, it is hard to predict if the result will be ferromagnetic and even harder to predict the kind of hystereses curve it will have. Pure **Fe** in its (high temperature) **fcc** lattice variant is not magnetic, neither are most variants of stainless steel, for example.
- But progress has been made - triggered by an increasing theoretical understanding (there are theories, after all), lots of experience and semi-theoretical guide lines - and just plain old trying out in the lab.
  - This is best demonstrated by looking at the "strength" of permanent magnets as it went up over the years:

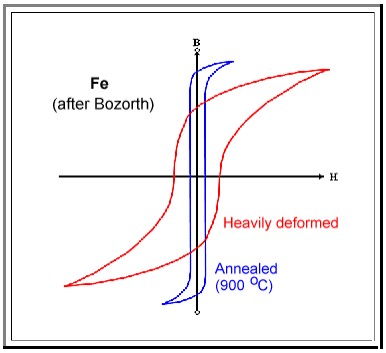


- Not bad. And pretty exotic materials emerged. Who thinks of Cobalt - Samarium compounds, or Neodymium - iron - boron?
  - What will the future bring. Well, I don't know and you shall see!
  - But we can do a little exercise to get some idea of what might be possible

Exercise 6.3.1

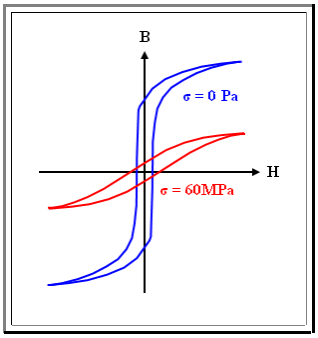
Maximum Magnetization

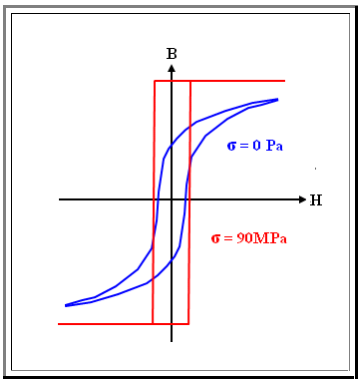
- The final question was: Can I change the hystereses curve of a given material in a defined direction?
  - The answer is: Yes, you can - within limits, of course.
  - The hystereses curve results from the relative ease or difficulty of moving domain walls in a given material. And since domain walls interact with stress and strain in a material, their movement depends on the internal structure of the material; on the kind and density of crystal lattice defects.
- This is best illustrated by looking at hystereses curves of *one and the same material* with different internal structures.



- There is a big difference for annealed, i.e. relatively defect free iron and heavily deformed iron, i.e. iron full of dislocations, as the figure on the left nicely illustrates
- We will find similar behavior for most ferromagnetic materials (not for all, however, because some are amorphous).
- Instead of manipulating the defects in the materials to see what kind of effect we get, we can simply put it under mechanical stress, e.g. by pulling at it. This also may change the hystereses curve very much:

Here we have the hystereses curves of pure **Ni** samples with and without mechanical tension. The effects are quite remarkable





- In this case the tension force was parallel to the external field ***H***
- There is a big change in the remanence, but not so much difference in the coercivity.

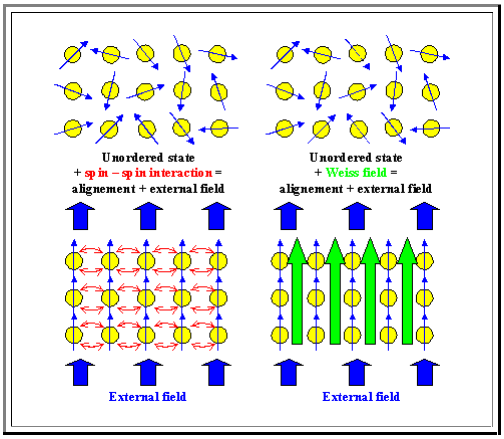
- In this case the tension force was at *right angles* to the external field ***H***
- Big changes in the remanence, not so much effect in the coercivity. We have an almost box-like shape, coming close to the ideal hard magnet from above.

The final word thus is:

- There is a ***plethora*** of ways to design ferromagnetic properties out there. The trouble is, we are just learning now how to do it a little bit better than by pure trial and error.
- The future of magnetism looks bright. With an increased level of understanding, new materials with better properties will result for almost sure. Time will tell.

4.3.7 Summary to: Ferromagnetism

- In ferromagnetic materials the magnetic moments of the atoms are "correlated" or lined-up, i.e. they are all pointing in the same direction
  - The physical reason for this is a quantum-mechanical spin-spin interaction that has no simple classical analogue.
  - However, exactly the same result - complete line-up - could be obtained, if the magnetic moments would feel a strong magnetic field.
  - In the "mean field" approach or the "Weiss" approach to ferromagnetism, we simply assume such a magnetic field H_{Weiss} to be the cause for the line-up of the magnetic moments. This allows to treat ferromagnetism as a "special" case of paramagnetism, or more generally, "orientation polarization".

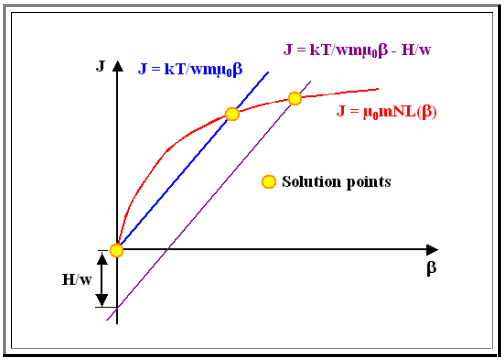


- For the magnetization we obtain ⇒
  - The term $w \cdot J$ describes the Weiss field via $H_{loc} = H_{ext} + w \cdot J$; the Weiss factor w is the decisive (and unknown) parameter of this approach.
  - Unfortunately the resulting equation for J, the quantity we are after, cannot be analytically solved, i.e. written down in a closed way.

$$J = N \cdot m \cdot \mu_0 \cdot L(\beta) = N \cdot m \cdot \mu_0 \cdot L\left(\frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}\right)$$

- Graphical solutions are easy, however ⇒
  - From this, and with the usual approximation for the Langevin function for small arguments, we get all the major ferromagnetic properties, e.g.
    - Saturation field strength.
    - Curie temperature T_C.

$$T_C = \frac{N \cdot m^2 \cdot \mu_0^2 \cdot w}{3k}$$

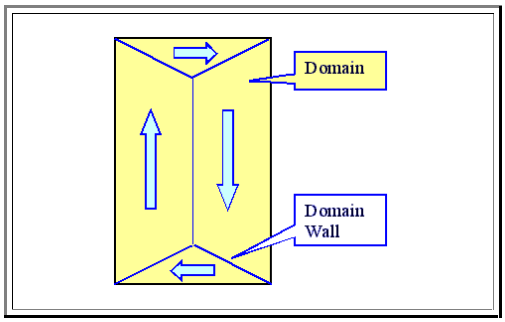


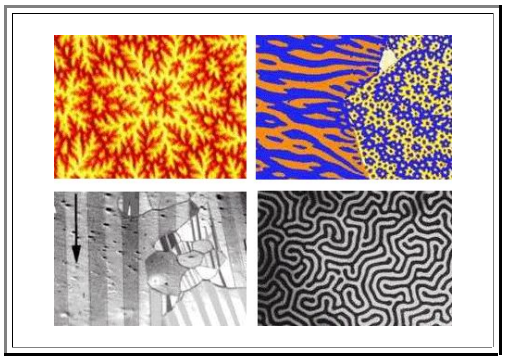
    - Paramagnetic behavior above the Curie temperature.
    - Strength of spin-spin interaction via determining w from T_C.
  - As it turns out, the Weiss field would have to be far stronger than what is technically achievable - in other words, the spin-spin interaction can be exceedingly strong!

- In single crystals it must be expected that the alignments of the magnetic moments of the atom has some preferred crystallographic direction, the "easy" direction.

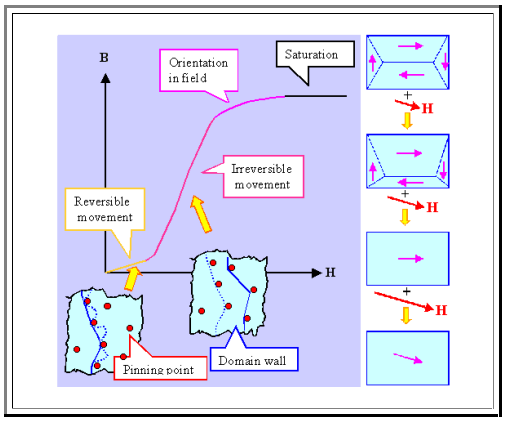
Easy directions:
Fe (bcc) <100>
Ni (fcc) <111>
Co (hcp) <001> (c-direction)

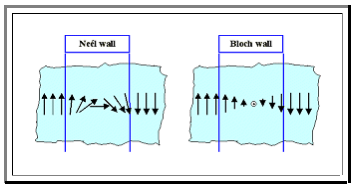
- A single crystal of a ferromagnetic material with all magnetic moments aligned in its easy direction would carry a high energy because:
  - It would have a large external magnetic field, carrying field energy.
- In order to reduce this field energy (and other energy terms not important here), magnetic domains are formed ⇒. But the energy gained has to be "payed for" by:
  - Energy of the domain walls = planar "defects" in the magnetization structure. It follows: Many small domains —> optimal field reduction —> large domain wall energy "price".
  - In polycrystals the easy direction changes from grain to grain, the domain structure has to account for this.
  - In all ferromagnetic materials the effect of magnetostriction (elastic deformation tied to direction of magnetization) induces elastic energy, which has to be minimized by producing a optimal domain structure.
- The domain structures observed thus follows simple principles but can be fantastically complicated in reality ⇒.



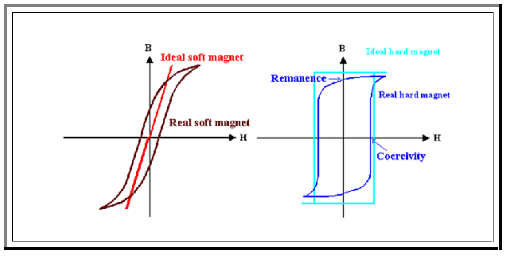


- For ferromagnetic materials in an external magnetic field, energy can be gained by increasing the total volume of domains with magnetization as parallel as possible to the external field - at the expense of unfavorably oriented domains.
  - Domain walls must move for this, but domain wall movement is hindered by defects because of the elastic interaction of magnetostriction with the strain field of defects.
  - Magnetization curves and hystereses curves result ⇒, the shape of which can be tailored by "defect engineering".
- Domain walls (mostly) come in two varieties:
  - Bloch walls, usually found in bulk materials.
  - Neél walls, usually found in thin films.





- Depending on the shape of the hystereses curve (and described by the values of the remanence M_R and the coercivity H_C, we distinguish hard and soft magnets ⇒.
- Tailoring the properties of the hystereses curve is important because magnetic losses and the frequency behavior is also tied to the hystereses and the mechanisms behind it.
  - Magnetic losses contain the (trivial) eddy current losses (proportional to the conductivity and the square of the frequency) and the (not-so-trivial) losses proportional to the area contained in the hystereses loop times the frequency.



- The latter loss mechanism simply occurs because it needs work to move domain walls.

It also needs time to move domain walls, the frequency response of ferromagnetic materials is therefore always rather bad - most materials will not respond anymore at frequencies far below **GHz**.

Questionaire
Multiple Choice questions to all of 4.3

4.4 Applications of Magnetic Materials

4.4.1 Everything Except Data Storage

General Overview

What are typical applications for magnetic materials? A somewhat stupid question - after all we already touched on several applications in the preceding subchapters.

- But there are most likely more applications than you (and I) are able to name. In addition, the material requirements within a specific field of application might be quite different, depending on details.
- So lets try a systematic approach and list all relevant applications together with some key requirements. We use the abbreviation **M_S**, **M_R**, and **H_C** for saturation, remanence, and coercivity, resp., and **low** ω, **medium** ω, and **high** ω with respect to the required frequency range.

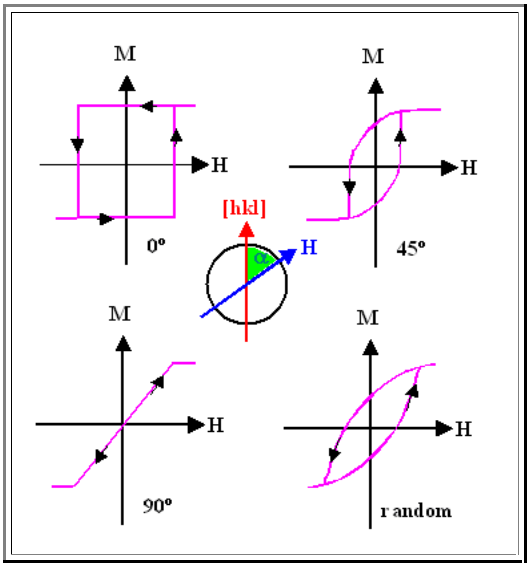
Field of application	Products	Requirements	Materials
Soft Magnets			
Power conversion electrical - mechanical	Motors Generators Electromagnets	Large **M_R** Small **H_C** Low losses **=** small conductivity low ω	**Fe** based materials, e.g. **Fe +** ≈ **(0,7 - 5)% Si** **Fe +** ≈ **(35 - 50)% Co**
Power adaption	(Power) Transformers		
Signal transfer	Transformer ("Überträger")	Linear **M - H** curve	
	LF ("low" frequency; up to ≈ **100 kHz**)	Small conductivity medium ω	**Fe +** ≈ **36 % Fe/Ni/Co** ≈ **20/40/40**
	HF ("high" frequency up to ≈ **100 kHz**)	Very small conductivity high ω	**Ni - Zn** ferrites
Magnetic field screening	"**Mu-metal**"	Large **dM/dH** for **H** ≈ **0** ideally **μ_r = 0**	**Ni/Fe/Cu/Cr** ≈ **77/16/5/2**
Hard Magnets			
Permanent magnets	Loudspeaker Small generators Small motors Sensors	Large **H_C** (and **M_R**)	**Fe/Co/Ni/Al/Cu** ≈**50/24/14/9/3** **$SmCo_5$** **Sm_2Co_{17}** **"NdFeB" (= $Nd_2Fe_{14}B$)**
Data storage analog	Video tape Audio tape	Medium **H_C**(and **M_R**), hystereses loop as rectangular as possible	**NiCo**, **CuNiFe**, **CrO_2** **Fe_2O_3**
Data storage digital	Ferrite core memory Drum		
	Hard disc, Floppy disc		
	Bubble memory	Special domain structure	Magnetic garnets (**AB_2O_4**, or **$A_3B_5O_{12}$**), e.g. with **A =** Yttrium (or mixtures of rare earth), and **B =** mixtures of **Sc**, **Ga**, **Al** Most common: **$Gd_3Ga_5O_{12}$**
Specialities			

Quantum devices	***GMR*** reading head	Special spin structures in multilayered materials	
	MRAM		

- As far as materials are concerned, we are only scratching the surface here. Some more materials are listed in the link
  - Data storage is covered in a separate module, here we just look at the other applications a bit more closely.

Soft Ferromagnets

- The general range of applications for soft magnets is clear from the table above. It is also clear that we want the hystereses loop as "flat" as possible, and as steeply inclined as possible. Moreover, quite generally we would like the material to have a high resistivity.
  - The requirements concerning the maximum frequency with which one can run through the hystereses loop are more specialized: Most power applications do not need high frequencies, but the microwave community would love to have more magnetic materials still "working" at **100 Ghz** or so.
- Besides trial and error, what are the guiding principles for designing soft magnetic materials? There are simple basic answers, but it is not so simple to turn these insights into products:
  - Essentially, *remanence* is directly related to the ease of movement of domain walls. If they can move easily in response to magnetic fields, remanence (and coercivity) will be low and the hystereses loop is flat.
- The essential quantities to control, partially mentioned before, therefore are:
  - The *density of domain walls*. The fewer domain walls you have to move around, the easier it is going to be.
  - The *density of defects* able to "pin" domain walls. These are not just the classical lattice defects encountered in neat single- or polycrystalline material, but also the cavities, inclusion of second phases, scratches, microcracks or whatever in *real* sintered or hot-pressed material mixtures.
  - The general *anisotropy* of the magnetic properties; including the anisotropy of the magnetization ("*easy*" and "hard" direction, of the magnetostriction, or even induced the shape of magnetic *particles* embedded in a non-magnetic matrix (we must expect, e.g. that elongated particles behave differently if their major axis is in the direction of the field or perpendicular to it). Large anisotropies generally tend to induce large obstacles to domain movement.
- A few general recipes are obvious:
  - Use well-annealed material with few grain boundaries and dislocations. For **Fe** this works, as already shown before.
  - Align the grains of e.g. polycrystalline **Fe**-based material into a favorable direction, i.e. use materials with a **texture**.
  - Doing this by a rather involved process engineered by **Goss** for **Fe** and **Fe-Si** alloys was a major break-through around **1934**. The specific power loss due to hystereses could be reduced to about **2.0 W/kg** for regular textured **Fe** and to **0.2 W/kg** for (very difficult to produce) textured **Fe** with **6% Si** (at **50 Hz** and **B** ≈ **1 T**)
  - Use isotropic materials, in particular **amorphous metals** also called **metallic glasses**, produced by extremely fast cooling from the melt. Stuff like $Fe_{78}B_{13}Si_{19}$ is made (in very thin very long ribbons) and used.
  - Total losses of present day transformer core materials (including eddy current losses) are around **0,6 W/kg** at **50 Hz** which, on the one hand, translates into an efficiency of **99,25 %** for the transformer, and a financial loss of roughly **1 $/kg** and year - which is not to be neglected, considering that big transformer weigh many tons.



- Reduce the number of domains. One solution would be to make very small magnetic particles that can only contain one domain embedded in some matrix. This would work well if the easy direction of the particles would always be in field direction, i.e. if all particles have the same crystallographic orientation pointing in the desired direction as shown below.
- This picture, by the way, was calculated and is an example of what can be done with theory. It also shows that single domain magnets can have ideal soft or ideal hard behavior, depending on the angle between an easy direction and the magnetic field.
- Unfortunately, for randomly oriented particles, you only get a mix - neither here nor there.

Well, you get the drift. And while you start thinking about some materials of your own invention, do not forget: We have not dealt with eddy current losses yet, or with the resistivity of the material.

- The old solution was to put **Si** into **Fe**. It increases the resistivity substantially, without degrading the magnetic properties too much. However it tends to make the material brittle and very hard to process and texture.
- The old-fashioned way of stacking thin insulated sheets is still used a lot for big transformers, but has clear limits and is not very practical for smaller devices.
- Since eddy current losses increase with the square of the frequency, metallic magnetic materials are simply not possible at higher frequencies; i.e. as soon as you deal with signal transfer and processing in the **kHz**, **MHz** or even **GHz** region. We now need *ferrites*.

Questionaire

Multiple Choice questions to 4.4.1

4.4.2 Magnetic Data Storage

This topic was regularly handled in the Seminar and was therefore not included here. The Seminar has been abandoned, but this page nevertheless has not been written.

- You know anyway that Magnetic Data Storage is a top key technology of our times. If you don't know this or are unsure why, I strongly advice you to quit studying engineering and go for something simple.

4.4.3 Summary to: Technical Materials and Applications

- Uses of ferromagnetic materials may be sorted according to:
  - Soft magnets; e.g. **Fe** - alloys

    - **Everything profiting from an "iron core": Transformers, Motors, Inductances, ...**
    - **Shielding magnetic fields.**

  - Hard magnets; e.g. metal oxides or "strange" compounds.

    - **Permanent magnets for loudspeakers, sensors, ...**
    - **Data storage (Magnetic tape, Magnetic disc drives, ...**

- Even so we have essentially only **Fe**, **Ni** and **Co** (+ **Cr**, **O** and **Mn** in compounds) to work with, innumerable magnetic materials with optimized properties have been developed.
  - New complex materials (including "nano"materials) are needed and developed all the time.

Strongest permanent magnets:
Sm_2Co_{17}
$Nd_2Fe_{14}B$

- Data storage provides a large ***impetus*** to magnetic material development and to employing new effects like "**GMR**"; giant magneto resistance; a purely quantum mechanical effect.

4.5. Summary: Magnetic Materials

The **relative permeability μ_r** of a material "somehow" describes the interaction of magnetic (i.e. more or less all) materials and magnetic fields ***H***, e.g. vial the equations ⇒

- ***B*** is the **magnetic flux density** or **magnetic induction**, sort of replacing ***H*** in the Maxwell equations whenever materials are encountered.
- ***L*** is the inductivity of a linear solenoid, or)coil or inductor) with length ***l***, cross-sectional area ***A***, and number of turns ***t***, that is "filled" with a magnetic material with μ_r.
- ***n*** is *still* the index of refraction; a quantity that "somehow" describes how electromagnetic fields with extremely high frequency interact with matter.
  For all practical purposes, however, $\mu_r = 1$ for optical frequencies

$$B = \mu_0 \cdot \mu_r \cdot H$$

$$L = \frac{\mu_0 \cdot \mu_r \cdot A \cdot w^2}{l}$$

$$n = (\epsilon_r \cdot \mu_r)^{1/2}$$

Magnetic fields inside magnetic materials polarize the material, meaning that the vector sum of magnetic dipoles inside the material is no longer zero.

- The decisive quantities are the *magnetic* dipole moment ***m***, a vector, and the *magnetic* Polarization ***J***, a vector, too.
- Note: In contrast to dielectrics, we define an additional quantity, the **magnetization *M*** by simply including dividing ***J*** by μ_0.
- The magnetic dipoles to be polarized are either already present in the material (e.g. in **Fe, Ni or Co**, or more generally, in all **paramagnetic** materials, or are induced by the magnetic fields (e.g. in **diamagnetic** materials).
- The dimension of the magnetization ***M*** is **[A/m]**; i.e. the same as that of the magnetic field.

$$B = \mu_0 \cdot H + J$$

$$\underline{J} = \mu_0 \cdot \frac{\Sigma \underline{m}}{V}$$

$$M = \frac{J}{\mu_0}$$

The magnetic polarization ***J*** or the magnetization ***M*** are *not* given by some magnetic surface charge, because ⇒.

There is no such thing as a *magnetic monopole*, the (conceivable) counterpart of a negative or positive electric charge

The equivalent of "Ohm's law", linking current density to field strength in conductors is the *magnetic* Polarization law:

- The decisive material parameter is $\chi_{mag} = (\mu_r - 1) =$ **magnetic susceptibility**.
- The "classical" induction ***B*** and the magnetization are linked as shown. In essence, ***M*** only considers what happens in the material, while ***B*** looks at the total effect: material plus the field that induces the polarization.

$$M = (\mu_r - 1) \cdot H$$

$$M := \chi_{mag} \cdot H$$

$$B = \mu_0 \cdot (H + M)$$

Magnetic polarization mechanisms are formally similar to dielectric polarization mechanisms, but the physics can be entirely different.

Atomic mechanisms of magnetization are not directly analogous to the dielectric case

Magnetic moments originate from:

- The intrinsic magnetic dipole moments ***m*** of elementary particles with spin is measured in units of the Bohr magneton $\mathbf{m_{Bohr}}$.

- The magentic moment $\boldsymbol{m^e}$ of the electron is ⇒

- Electrons "orbiting" in an atom can be described as a current running in a circle thus causing a magnetic dipole moment; too

$$m_{Bohr} = \frac{h \cdot e}{4\pi \cdot m^*_e} = 9.27 \cdot 10^{-24}\ Am^2$$

$$m^e = \frac{2 \cdot h \cdot e \cdot s}{4\pi \cdot m^*_e} = 2 \cdot s \cdot m_{Bohr} = \pm m_{Bohr}$$

- The total magentic moment of an atom in a crystal (or just solid) is a (tricky to obtain) sum of all contributions from the electrons, and their orbits (including bonding orbitals etc.), it is either:
  - **Zero** - we then have a **diamagmetic material**.

**Magnetic field induces dipoles, somewhat analogous to elctronic polarization in dielectrics.
Always very weak effect (except for superconductors)
Unimportant for technical purposes**

  - In the order of a few Bohr magnetons - we have a essentially a **paramagnetic material**.

**Magnetic field induces some order to dipoles; strictly analogous to "orientation polarization" of dielectrics.
Always very weak effect
Unimportant for technical purposes**

- In some **ferromagnetic** materials spontaneous ordering of magenetic moments occurs below the Curie (or Neél) temperature. The important familiess are
  - Ferromagnetic materials ⇑⇑⇑⇑⇑⇑⇑ large μ_r, **extremely important**.
  - Ferrimagnetic materials ⇑⇓⇑⇓⇑⇓⇑ still large μ_r, **very important**.
  - Antiferromagnetic materials ⇑⇓⇑⇓⇑⇓⇑ $\mu_r \approx 1$, unimportant

**Ferromagnetic materials:
Fe, Ni, Co, their alloys
"AlNiCo", Co_5Sm, $Co_{17}Sm_2$,
"NdFeB"**

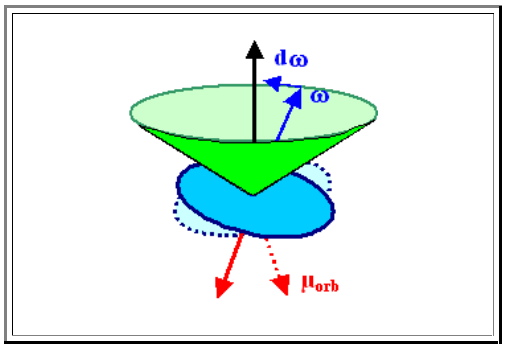
- There is characteristic temperatuer dependence of μ_r for all cases

- Dia- and Paramagentic propertis of materials are of no consequence whatsoever for products of electrical engineering (or anything else!)
  - Only their common denominator of being essentially "non-magnetic" is of interest (for a submarine, e.g., you want a non-magnetic steel)
  - For research tools, however, these forms of magnitc behavious can be highly interesting ("paramagentic resonance")

Normal diamagnetic materials: $\chi_{dia} \approx -(10^{-5} - 10^{-7})$
Superconductors (= ideal diamagnets): $\chi_{SC} = -1$
Paramagnetic materials: $\chi_{para} \approx +10^{-3}$

Diamagnetism can be understood in a semiclassical (Bohr) model of the atoms as the response of the current ascribed to "circling" electrons to a changing magnetic field via classical induction ($\propto$ **dH/dt**).

- The net effect is a precession of the circling electron, i.e. the normal vector of its orbit plane circles around on the green cone. ⇒
- The "Lenz rule" ascertains that inductive effects oppose their source; diamagnetism thus weakens the magnetic field, $\chi_{dia} < 0$ must apply.



Running through the equations gives a result that predicts a very small effect. ⇒
A proper quantum mechanical treatment does not change this very much.

$$\chi_{dia} = -\frac{e^2 \cdot z \cdot \langle r \rangle^2}{6\, m^*_e} \cdot \rho_{atom} \approx -(10^{-5} \cdot 10^{-7})$$

The formal treatment of paramagnetic materuials is mathematically completely identical to the case of orientation polarization

- The range of realistc β values (given by largest ***H*** technically possible) is even smaller than in the case of orientation polarization. This allows tp approximate **L(β)** by **β/3**; we obtain:

$$\chi_{para} = \frac{N \cdot m^2 \cdot \mu_0}{3kT}$$

- Insertig numbers we find that χ_{para} is indeed a number just slightly larger than **0**.

$$W(\varphi) = -\mu_0 \cdot \underline{m} \cdot \underline{H} = -\mu_0 \cdot m \cdot H \cdot \cos\varphi$$

Energy of magetic dipole in magnetic field

$$N[W(\varphi)] = c \cdot \exp -(W/kT) = c \cdot \exp \frac{m \cdot \mu_0 \cdot H \cdot \cos\varphi}{kT} = N(\varphi)$$

(Boltzmann) Distribution of dipoles on energy states

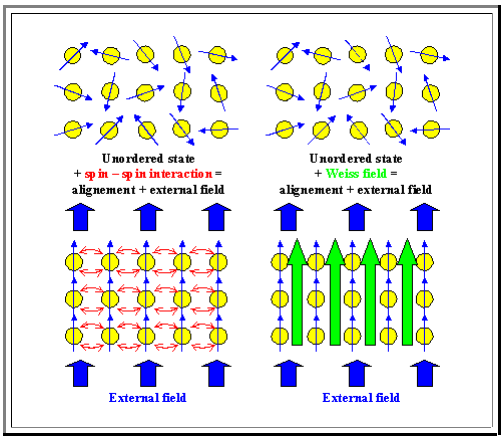
$$M = N \cdot m \cdot L(\beta)$$

$$\beta = \frac{\mu_0 \cdot m \cdot H}{kT}$$

Resulitn Magnetization with Langevin function **L(β)** and argument β

In ferromagnetic materials the magnetic moments of the atoms are "correlated" or lined-up, i.e. they are all pointing in the same direction

- The physical reason for this is a quantum-mechanical spin-spin interaction that has no simple classical analogue.
- However, exactly the same result - complete line-up - could be obtained, if the magnetic moments would feel a strong magnetic field.
- In the "mean field" approach or the "Weiss" approach to ferromagnetism, we simply assume such a magnetic field H_{Weiss} to be the cause for the line-up of the magnetic moments. This allows to treat ferromagnetism as a "special" case of paramagnetism, or more generally, "orientation polarization".



For the magnetization we obtain ⇒

- The term $w \cdot J$ describes the Weiss field via $H_{loc} = H_{ext} + w \cdot J$; the Weiss factor ***w*** is the decisive (and unknown) parameter of this approach.

$$J = N \cdot m \cdot \mu_0 \cdot L(\beta) = N \cdot m \cdot \mu_0 \cdot L\left(\frac{m \cdot \mu_0 \cdot (H + w \cdot J)}{kT}\right)$$

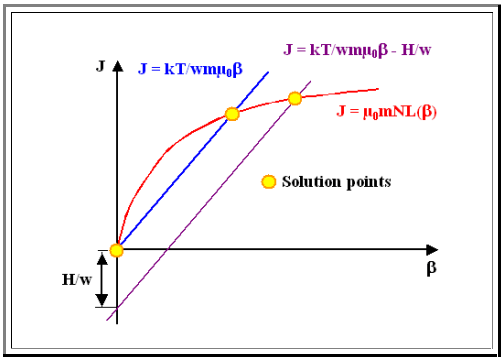
- Unfortunately the resulting equation for ***J***, the quantity we are after, cannot be analytically solved, i.e. written down in a closed way.

Graphical solutions are easy, however ⇒

- From this, and with the usual approximation for the Langevin function for small arguments, we get all the major ferromagnetic properties, e.g.
  - Saturation field strength.
  - Curie temperature T_C.

$$T_C = \frac{N \cdot m^2 \cdot \mu_0^2 \cdot w}{3k}$$

  - Paramagnetic behavior above the Curie temperature.
  - Strength of spin-spin interaction via determining w from T_C.
- As it turns out, the Weiss field would have to be far stronger than what is technically achievable - in other words, the spin-spin interaction can be exceedingly strong!



In single crystals it must be expected that the alignments of the magnetic moments of the atom has some preferred crystallographic direction, the "easy" direction.

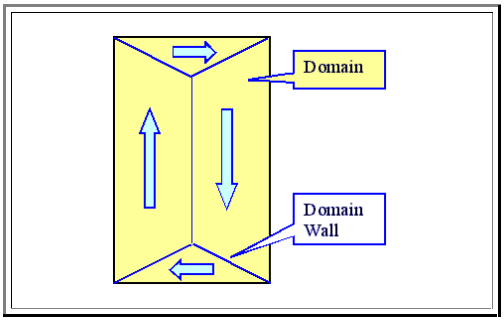
Easy directions:
Fe (bcc) <100>
Ni (fcc) <111>
Co (hcp) <001> (c-direction)

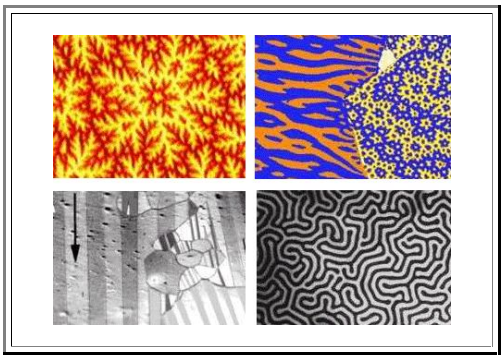
A single crystal of a ferromagnetic material with all magnetic moments aligned in its easy direction would carry a high energy because:

- It would have a large external magnetic field, carrying field energy.

In order to reduce this field energy (and other energy terms not important here), magnetic domains are formed ⇒. But the energy gained has to be "payed for" by:

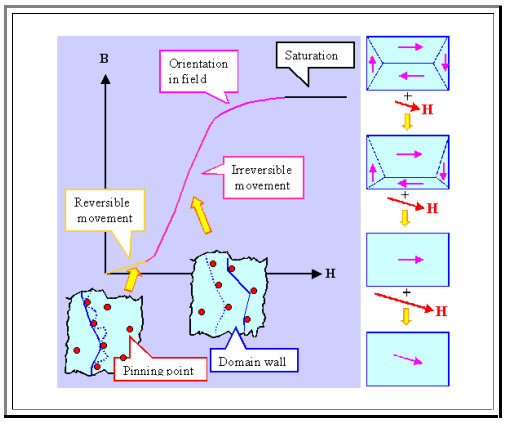
- Energy of the domain walls = planar "defects" in the magnetization structure. It follows: Many small domains —> optimal field reduction —> large domain wall energy "price".
- In polycrystals the easy direction changes from grain to grain, the domain structure has to account for this.
- In all ferromagnetic materials the effect of magnetostriction (elastic deformation tied to direction of magnetization) induces elastic energy, which has to be minimized by producing a optimal domain structure.



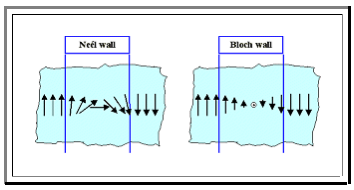


The domain structures observed thus follows simple principles but can be fantastically complicated in reality ⇒.

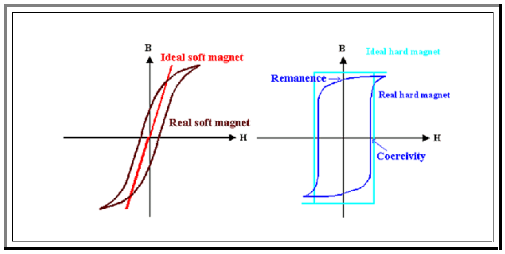
- For ferromagnetic materials in an external magnetic field, energy can be gained by increasing the total volume of domains with magnetization as parallel as possible to the external field - at the expense of unfavorably oriented domains.
  - Domain walls must move for this, but domain wall movement is hindered by defects because of the elastic interaction of magnetostriction with the strain field of defects.
  - Magnetization curves and hystereses curves result ⇒, the shape of which can be tailored by "defect engineering".



- Domain walls (mostly) come in two varieties:
  - Bloch walls, usually found in bulk materials.
  - Neél walls, usually found in thin films.



- Depending on the shape of the hystereses curve (and described by the values of the remanence M_R and the coercivity H_C, we distinguish hard and soft magnets ⇒.



- Tailoring the properties of the hystereses curve is important because magnetic losses and the frequency behavior is also tied to the hystereses and the mechanisms behind it.
  - Magnetic losses contain the (trivial) eddy current losses (proportional to the conductivity and the square of the frequency) and the (not-so-trivial) losses proportional to the area contained in the hystereses loop times the frequency.
  - The latter loss mechanism simply occurs because it needs work to move domain walls.
- It also needs time to move domain walls, the frequency response of ferromagnetic materials is therefore always rather bad - most materials will not respond anymore at frequencies far below **GHz**.

- Uses of ferromagnetic materials may be sorted according to:
  - Soft magnets; e.g. **Fe** - alloys

    - **Everything profiting from an "iron core": Transformers, Motors, Inductances, ...**
    - **Shielding magnetic fields.**

  - Hard magnets; e.g. metal oxides or "strange" compounds.

    - **Permanent magnets for loudspeakers, sensors, ...**
    - **Data storage (Magnetic tape, Magnetic disc drives, ...**

- Even so we have essentially only **Fe**, **Ni** and **Co** (+ **Cr**, **O** and **Mn** in compounds) to work with, innumerable magnetic materials with optimized properties have been developed.
  - New complex materials (including "nano"materials) are needed and developed all the time.

Strongest permanent magnets:
Sm_2Co_{17}
$Nd_2Fe_{14}B$

- Data storage provides a large ***impetus*** to magnetic material development and to employing new effects like "**GMR**"; giant magneto resistance; a purely quantum mechanical effect.

Questionaire

Multiple Choice questions to all of 4

5. General Aspects of Silicon Technology

5.0 Required Reading

5.0.1 Basic Bipolar Transistor

5.0.2 Basic MOS Transistor

5.0.3 Summary to: Required Reading to Chapter 5

5.1 Basic Considerations for Process Integration

5.1.1 What is Integration?

5.1.2 Basic Concepts of Integrating Bipolar Transistors

5.1.3 Basic Concepts of Connecting Transistors

5.1.4 Integrated MOS Transistors

5.1.5 Integrated CMOS Technology

5.1.6 Summary to: 5.1 Basic Considerations for Process Integration

5.2 Process Integration

5.2.1 Chips on Wafers

5.2.2 Packaging and Testing

5.2.3 Summary to: Chips on Wafers

5.3 Cleanrooms, Particles and Contamination

5.3.1 Cleanrooms and Defects

5.3.2 Summary to: 5.3 Cleanrooms, Particles and Contamination

5.4 Development and Production of a New Chip Generation

5.4.4 Summary to: 5.4 Development and Production of a New Chip Generation

5.5.1 Summary: General Aspects of Silicon Technology

5. General Aspects of Silicon Technology

5.0 Required Reading

5.0.1 Basic Bipolar Transistor

- For the purpose of this basic module, we simply take the contents of the "Bipolar Transistor" module from the *Semiconductor Hyperscript*.
  - There you will always find the newest version; the module is reproduced below.
  - It is about as basic as it can be - just assumming that you know the *basics about **pn**-junctions*.
  - If you remember **pn**-junctions diodes only vaguely (or not at all), turn to the diode parts of the Semiconductor Hyperscripts and check the links from there.
- If you understand German; this link will bring you to the relevant parts of the Hyperscript "Einführung in die Materialwissenschaft **II**"

Bipolar Transistors: Basic Concept and Operation

- We are not very 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 d_{base} is much smaller than the diffusion length 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:
- For the *hole* forward current, we have in the simplest approximation (ideal diode, no reverse current; no **SCR** contribution):

$$j_{hole}(U) = \frac{e \cdot L \cdot n_i^2}{\tau \cdot N_{Acc}} \cdot \exp - \frac{e \cdot U}{kT}$$

  - and the relevant quantities refer to the *hole* properties in the ***n** - doped base* and the doping level N_{Acc} in the ***p** - doped emitter*. For the electron forward current we have accordingly:

$$j_{electron}(U) = \frac{e \cdot L \cdot n_i^2}{\tau \cdot N_{Don}} \cdot \exp - \frac{e \cdot U}{kT}$$

  - and the relevant quantities refer to the *electron* properties in the ***p** - doped emitter* and the doping level N_{Don} in the ***n** - doped base*.
  - The relation between these currents, i.e. $j_{hole}/j_{electron}$, which we call the **injection ratio** κ, then is given by

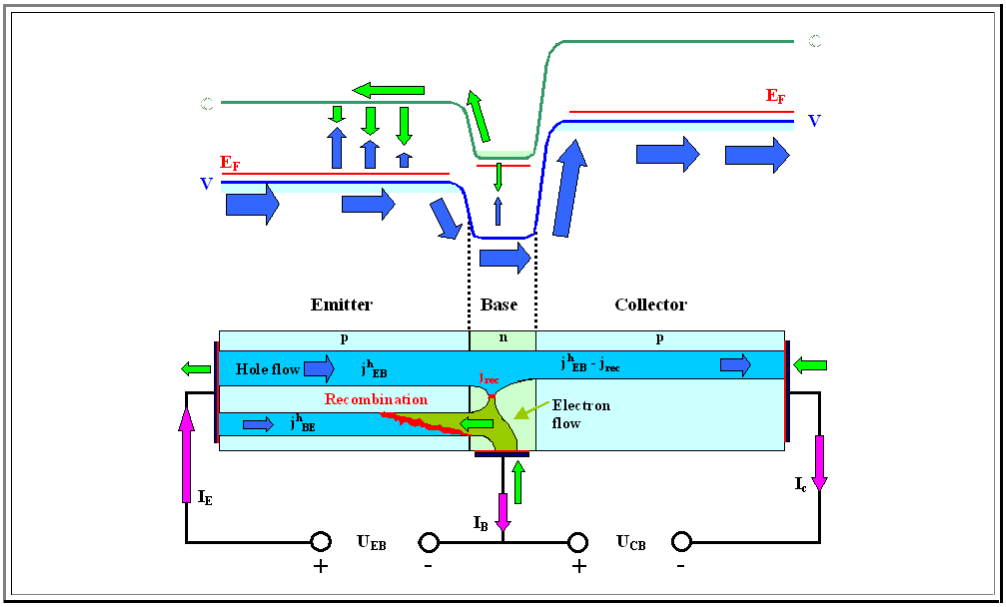
$$\kappa = \frac{\frac{L_h}{\tau_h \cdot N_{Ac}}}{\frac{L_e}{\tau_e \cdot N_{Don}}} = \frac{N_{Ac}}{N_{Don}}$$

- 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 optoelectronic devices! (in a separate lecture course).

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 they 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.



Let's 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 from 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 Υ of

$$\Upsilon = \frac{I_C}{I_B} = \frac{I_E}{I_B} - 1$$

  - 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 Υ:

$$\frac{I_E}{I_B} = \frac{j_{BE}^{h} + j_B^{e}}{j_B^{e}} = \kappa + 1$$

$$\Upsilon = \frac{I_E}{I_B} - 1 = \kappa + 1 - 1 = \kappa = \frac{N_{Ac}}{N_{Don}}$$

- *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

$$\Upsilon = \frac{\frac{L_h}{\tau_h \cdot N_{Ac}}}{\frac{L_e}{\tau_e \cdot N_{Don}}} \cdot \left(1 - \frac{d_{base}}{L}\right) \approx \frac{N_{Don}}{N_{Ac}} \cdot \left(1 - \frac{d_{base}}{L}\right)$$

  - 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 **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*.

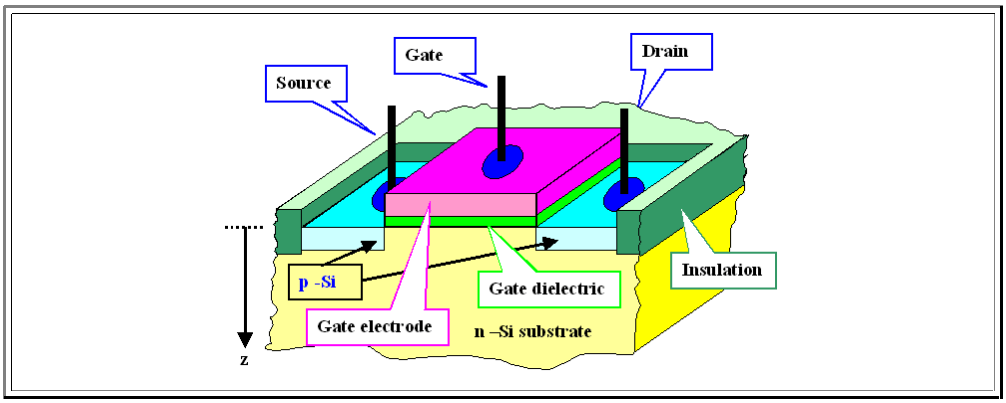
Questionaire

Multiple Choice questions to 5.0.1

5.0.2 Basic MOS Transistor

Qualitative Description

- The basic concept of a ***MOS* Transistor** transistor is simple and best understood by looking at its structure:



  - It is always an *integrated* structure, there are practically no single individual **MOS** transistors.
- A **MOS** transistor is primarily a switch for digital devices. Ideally, it works as follows:
  - If the voltage at the **gate electrode** is "*on*" , the transistor is "*on*", too, and current flow between the **source** and **drain** electrodes is possible (almost) without losses.
  - If the voltage at the gate electrode is "*off* ", the transistor is "*off*", too, and no current flows between the source and drain electrode.
- In reality, this only works for a given *polarity* of the gate voltage (in the picture above, e.g., only for negative gate voltages), and if the supply voltage (always called U_{DD}) is not *too small* (it used to be **5 V** in ancient times around **1985**; since then it is declining and will soon hit an ultimate limit around **1 V**).
  - Moreover, a **MOS** transistor needs *very thin* **gate dielectrics** (around, or better below **10 nm**), and *extreme control* of materials and technologies if real **MOS** transistors are to behave as they are expected to in "ideal" theory.
- What is the working principle of an "ideal" **MOS** transistor?
  - In order to understand it, we look at the behavior of carriers in the **Si** under the influence of an external electrical field under the gate region.
- Understanding **MOS** transistor *qualitatively* is easy. We look at the example from above and apply some source-drain voltage U_{SD} in either polarity, but *no gate voltage yet*. What we have under these conditions is
  - A **n**-type **Si** substrate with a certain equilibrium density of electrons $n^e(U_G = 0)$, or $n^e(0)$ for short. Its value is entirely determined by doping (and the temperature, which we will neglect at the present, however) and is the same everywhere. We also have a much smaller concentration $n^h(0)$ of holes.
  - Some **p**-doped regions with an equilibrium concentration of holes. The value of the hole concentration in the source and drain defined in this way also is determined by the doping, but the value is of no particular importance in this simple consideration.
  - Two **pn**-junctions, one of which is polarized in forward direction (the one with the positive voltage pole), and the other one in reverse. This is true for any polarity; in particular one junction will *always* be ***biased*** in reverse. Therefore *no source-drain current I_{SD} will flow* (or only some small reverse current which we will neglect at present).
  - There will also be no current in the forwardly biased diode, because the **n-Si** of the substrate in the figure is not electrically connected to anything (in reality, we might simply ground the positive U_{SD} pole and the substrate).
- In summary, for a gate voltage $U_G = 0$ **V**, there are no currents and everything is in equilibrium. But now apply a *negative voltage* at the gate and see what happens.
  - The electrons in the substrate below the gate will be electrostatically repelled and driven into the substrate. Their concentration directly below the gate will go down, $n^e(U)$ will be a function of the depth coordinate z .

$$n^e = n^e(z) = f(n^e(0), U)$$

  - Since we still have equilibrium, the mass action law for carriers holds anywhere in the Si, i.e. .

$$n^e(z) \cdot n^h(z) = n_i^2$$

  - With n_i **= intrinsic carrier density in Si = const.(U,z)**

- This gives us

$$n^h(z) = \frac{n_i^2}{n^e(z)}$$

- In other words: If the electron concentration below the gate goes down, the hole concentration goes up.

If we sufficiently decrease the electron concentration under the gate by cranking up the gate voltage, we will eventually achieve the condition $n^h(z = 0) = n^e(z = 0)$ right under the gate, i.e. at $z = 0$

- If we increase the gate voltage even more, we will encounter the condition $n^h(z) > n^e(z)$ for small values of z, ie. for $z_c > z > 0$.

In other words: Right under the gate we now have *more holes than electrons*; this is called a state of **inversion** for obvious reasons. **Si** having more holes than electrons is also called *p-type Si*. What we have now is a **p**-conducting **channel** (with width z_c) connecting the **p**-conducting source and drain.

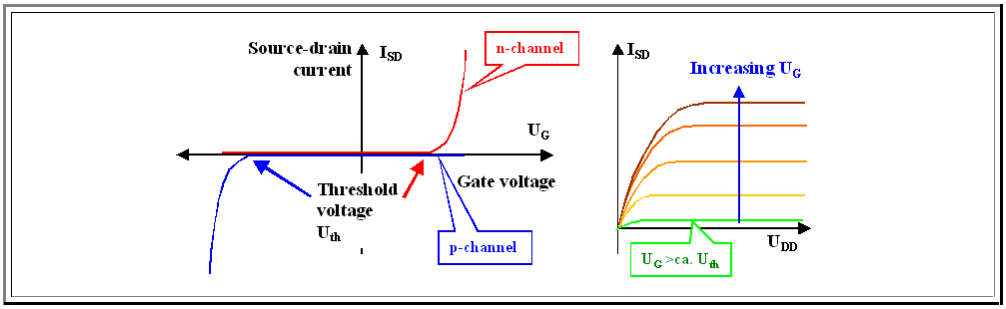
- There are no more **pn**-junctions preventing current flow under the gate - current can flow freely; only limited by the ohmic resistance of contacts, source/drain and channel.

Obviously, while cranking up the gate voltage *with the right polarity*, sooner or later we will encounter inversion and form a conducting channel between our terminals which becomes more prominent and thus better conducting with increasing gate voltage

- The resistivity of this channel will be determined by the amount of **Si** we have inverted; it will rapidly come down with the voltage as soon as the **threshold voltage** necessary for inversion is reached.

If we reverse the voltage at the gate, we attract electrons and their concentration under the gate increases. This is called a state of **accumulation**. The **pn** junctions at source and drain stay intact, and no source - drain current will flow.

- Obviously, if we want to switch a **MOS** transistor "on" with a *positive* gate voltage, we must now reverse the doping and use a **p**-doped substrates and **n**-doped source/drain regions.
- The two basic types we call "**n-channel MOS**" and "**p-channel MOS**" according to the kind of doping in the channel upon inversion (or the source/drain contacts).
- Looking at the electrical characteristics, we expect curves like this:



- The dependence of the source-drain current I_{SD} on the gate voltage U_G is clear from what was described above, the dependence of I_{SD} on the source-drain voltage U_{SD} with U_G as parameter is maybe not obvious immediately, but if you think about it a minute. you just can't draw currents without some U_{SD} and curves as shown must be expected qualitatively.

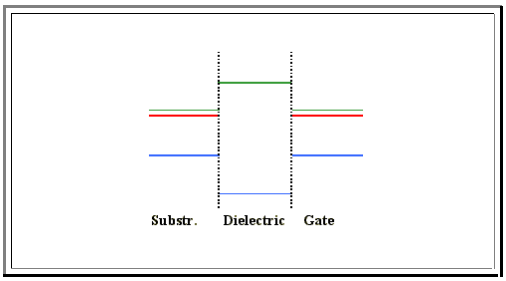
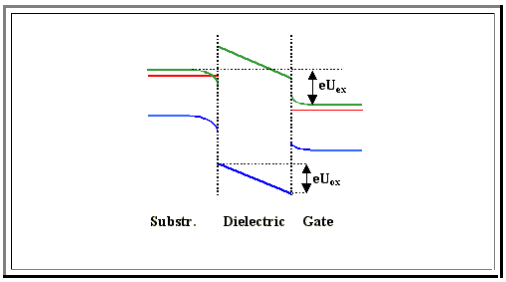
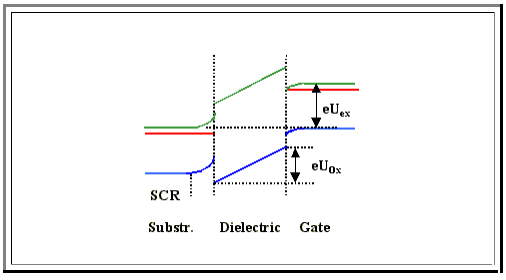
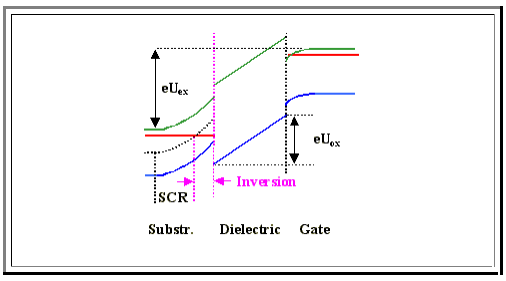
What can we say *quantitatively* about the working of an **MOS** transistor?

- What determines the threshold voltage U_{th}, or the precise shape of the $I_{SD}(U_{th})$ curves? Exactly how does the source - drain voltage U_{SD} influence the characteristics? How do the prime quantities depend on material and technology parameters, e.g. the thickness of the gate dielectric and its dielectric constant ϵ_r or the doping level of substrate and source/drain?
- Plenty of questions that are, as a rule, not easily answered. We may, however, go a few steps beyond the qualitative picture given above.

Some Quantitative Considerations

The decisive part is achieving *inversion*. Lets see how that looks like in a band diagram. To make life easier, we make the gate electrode from the same kind of **n-Si** as the substrate, just highly doped so it is as metallic as possible - we have the same kind of band diagram then to the left and right of the gate dielectric

- Lets look schematically what that will give us for some basic cases:

Voltage at the gate	Conditions in the Si	Voltage drop	Charge distribution
Zero gate voltage. "**Flat band**" condition			
	Nothing happens. The band in the substrate is perfectly flat (and so is the band in the contact electrode, but that is of no interest).	We only would have a voltage (or better potential) drop, if the Fermi energies of substrate and gate electrode were different	There are no *net* charges
Positive gate voltage. **Accumulation**			
	With a positive voltage at the gate we attract the electrons in the substrate. The bands must bend down somewhat, and we increase the number of electrons in the conduction band accordingly. (There is a bit of a space charge region (**SCR**) in the contact, but that is of no interest).	The voltage drops mostly in the oxide	There is some *positive charge* at the gate electrode interface (with our **Si** electrode from the **SCR**), and *negative charge* from the many electrons in the (thin) accumulation layer on the other side of the gate dielectric.
Small negative gate voltage. **Depletion**			
	With a (small) negative voltage at the gate, we repel the electrons in the substrate. Their concentration decreases, the hole concentration is still low - we have a layer depleted of mobile carriers and therefore a **SCR**.	The voltage drops mostly in the oxide, but also to some extent in the **SCR**.	There is some *negative charge* at the gate electrode interface (accumulated electrons with our **Si** electrode), and *positive charge* smeared out in the the (extended) **SCR** layer on the other side of the gate dielectric.
Large negagive gate voltage. **Inversion**			
	With a (large) negative voltage at the gate, we repel the electrons in the substrate very much. The bands bend so much, that the Fermi energy (red line) is in the lower half of the band close to the interface. In this region holes are the majority carriers, we gave	The voltage drops mostly in the oxide, but also to some extent in the **SCR** and the inversion layer.	There is more *negative charge* at the gate electrode interface (accumulated electrons with our **Si** electrode), some *positive charge* smeared out in the the (extended) **SCR** layer on the other side of the gate dielectric, and a lot of *positive charge* from the

	inversion. We still have a **SCR**, too.		holes in thin inversion layer.

- Qualitatively, this is clear. What happens if we replace the (highly **n**-doped) **Si** of the gate electrode with some metal (or **p**-doped **Si**)?
  - Then we have *different Fermi energies* to the left and right of the contact, leading to a *built-in potential* as in a **pn**-junction. We will than have some band bending at zero external voltage, flat band conditions for a non-zero external voltage, and concomitant adjustments in the charges on both sides.
  - But while this complicates the situation, as do unavoidable fixed immobile charges in the dielectric or in the **Si**-dielectric interface, nothing new is added.
- Now, the decisive part is achieving inversion. It is clear that this needs some minimum threshold voltage U_{th}, and from the pictures above, it is also clear that this request translates into a request for some *minimum charge* on the capacitor formed by the gate electrode, the dielectric and the **Si** substrate.
  - What determines the amount of charge we have in this system? Well, since the whole assembly for any distribution of the charge can always be treated as a simple capacitor C_G, we have for the charge of this capacitor.

$$Q_G = C_G \cdot U_G$$

  - Since we want U_{th} to be small, we want a *large gate capacitance* for a large charge Q_G, and now we must ask: What determines C_G?
- If all charges would be concentrated right at the interfaces, the capacitance *per area unit* would be given simply by the geometry of the resultant plate capacitor to

$$C_G = \frac{\epsilon\epsilon_0}{d_{Ox}}$$

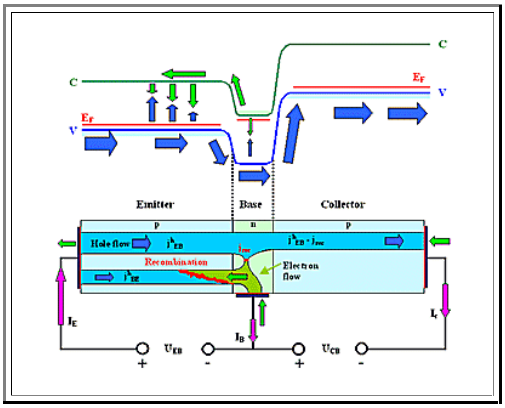
  - With d_{Ox} = thickness of the gate dielectric, (so far) always silicon dioxide SiO_2.
- Since our charges are somewhat spread out in the substrate (we may neglect this in the gate electrode if we use metals or very highly doped **Si**), we must take this into account.
  - In electrical terms, we simply have a second capacitor C_{Si} describing the effects of spread charges in the **Si**, switched in series to the geometric capacitor which we now call **oxide capacitance** C_{Ox} . It will be rather large for concentrated charges, i.e. for accumulation and inversion and small for depletion.
  - The total capacitance C_G then is given by

$$\frac{1}{C_G} = \frac{1}{C_{Ox}} + \frac{1}{C_{Si}}$$

- For inversion and accumulation, when the most of the charge is close to the interface, the total capacitance will be dominated by C_{Ox}. It is relatively large, because the thickness of the capacitor is small.
  - In the depletion range, C_{Si} will be largest and the total capacitance reaches a minimum.
  - In total, C_G as a function of the voltage, i.e. $C_G(U)$ runs from a constant value at large positive voltages through a minimum back to about the same constant value at large positive voltages. The resulting curve contains all relevant information about the system. Measuring $C_G(U)$ is thus the first thing you do when working with **MOS** contacts.
  - While it is not extremely easy to calculate the capacitance values and everything else that goes with it, it can be done - just solve the Poisson equation for the problem.
- All things considered, we want C_{Ox} to be *large* , and that means we want the dielectric to be *thin* and to have a *large* dielectric constant - as stated above without justification.
  - We also want the dielectric to have a large breakdown field strength, no fixed charges in the volume, no interface charges, a very small **tg δ**; it also should be very stable, compatible with **Si** technology, and cheap.
  - In other words, we wanted SiO_2 - even so its dielectric constant is just a mediocre **3.9** - for all those years of microelectronic wonders. But now (**2001**), we want something better with respect to dielectric constants. Much work is done, investigating, e.g., CeO_2, Gd_2O_3, ZrO_2, Y_2O_3, $BaTiO_3$, **BaO/SrO**, and so on. And nobody knows today (**2002**) which material will make the race!

5.0.3 Summary to: Required Reading to Chapter 5

Essentials of the bipolar transistor:

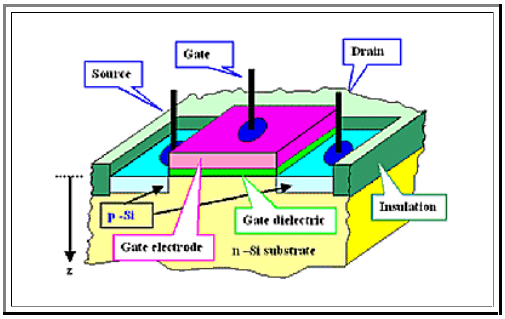


- High emitter doping (N_{Don} for **npn** transistor here) in comparison to base doping N_{Ac} for large current amplification factor $\gamma = I_C/I_B$.
- $N_{Don}/N_{Ac} \approx \kappa$ = injection ratio.

$$\gamma \approx \frac{N_{Don}}{N_{Ac}} \cdot \left(1 - \frac{d_{base}}{L} \right)$$

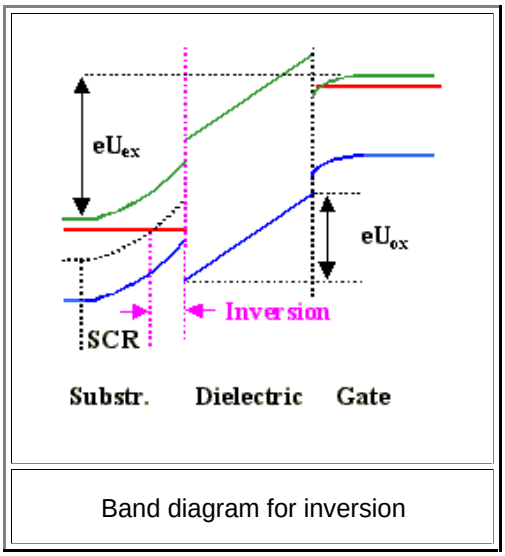
- Small base width d_{base} (relative to diffusion length L) for large current amplification.
- Not as easy to make as the band-diagram suggests!

Essentials of the **MOS** transistor:



- Gate voltage enables Source-Drain current
- Essential process. Inversion of majority carrier type in channel below gate by:
  - Drive intrinsic majority carriers into bulk by gate voltage with same sign as majority carriers.
  - Reduced majority concentration n_{maj} below gate increases minority carrier concentration n_{min} via mass action law

$$n_{maj} \cdot n_{min} = n_i^2$$



Band diagram for inversion

  - An inversion channel with $n_{min} > n_{maj}$ develops below the gate as soon as threshold voltage U_{Th} is reached.
  - Current now can flow because the reversely biased **pn**-junction between either source or drain and the region below the gate has disappeared.

The decisive material is the gate dielectric (usually SiO_2). Basic requirement is:

- High capacity C_G of the gate electrode - gate dielectric - **Si** capacitor = high charge Q_G on electrodes = strong band bending = low threshold voltages U_G
- It follows:
  - Gate dielectric thickness $d_{Di} \Rightarrow$ High breakdown field strength U_{Bd}
  - Large dielectric constant ϵ_r
  - No interface states.
  - Good adhesion, easy to make / deposit, easy to structure, small leakage currents, ...

$$Q_G = C_G \cdot U_G$$

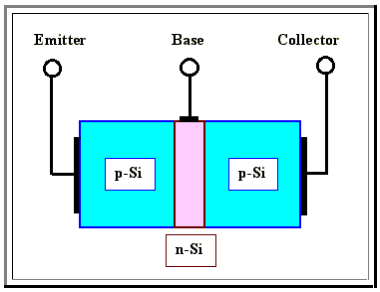
Example:
$U = 5$ V, $d_{Di} = 5$ nm $\Rightarrow$ $E = U/d_{Di} = 10^7$ V/cm !!

$\epsilon_r(SiO_2) = 3.9$

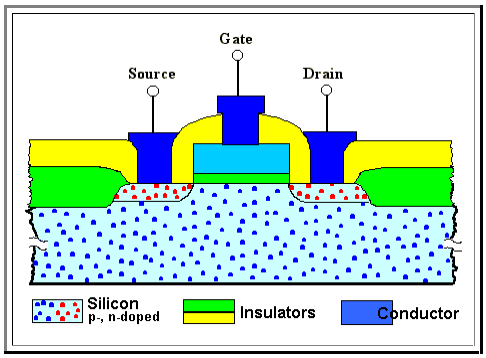
5.1 Basic Considerations for Process Integration

5.1.1 What is Integration?

- The key element of electric engineering, computer engineering, or pretty much everything else that is remotely "technical" in the last thirty years of the **2nd** millennium, is the **integrated transistor** in a **Silicon crystal** - everything else comes in second - at best.
  - *Integrated* means that there is more than one transistor on the same piece of **Si** crystal and thus in the same package. And "*more than one*" means at the present stage of technology some 10^7 *transistors per* cm^2 of Silicon.
  - *Silicon crystal* means that we use huge, extremely perfect single crystals of **Si** to do the job. Why **Si** and not. for example**Ge**, **GaAs** or **SiC**? Because if you look at the sum total of the most important properties you are asking for (crystal size and perfection, bandgap, extremely good and process compatible dielectric, ...) **Si** and its oxide, SiO_2 are so vastly superior to any possible contender that there is simply no other semiconductor that could be used for complex integrated circuitry.
- The lowly **integrated circuit** (***IC***), mostly selling for a few Dollars, is the most marvelous achievement of Materials Science in the second half of the **20th** century. Few people have an idea of the tremendous amount of science and engineering that was (and still is) needed to produce a state-of-the art **chip**, the little piece of **Si** crystal with some other materials in precise arrangements, that already starts to rival the complexity of the brains of lower animals and might at some day in the not so distant future even rival ours.
- If we want to make a *circuit* out of many transistors (some of which we use as resistors) and maybe some capacitors, we need *three basic ingredients* - no matter if we do this in an integrated fashion or by soldering the components together - and on occasion some "spices", some special additions:
- **1. Ingredient:** *Transistors.* "Big" and "small" ones (with respect to the current they can switch), for low or high voltage, fast or not so fast - the whole lot. We have two basic types to chose from:
  - **Bipolar transistors** (the hopefully familiar **pnp**- or **npn**-structures) *are usually drawn as follows*



  - Note right here that no real transistor looks even remotely like this structure! It's only and purely a schematic drawing to show essentials and and has nothing whatsoever to do with a real transistor.
  - The name *bipolar* comes from the fact that *two* kinds of carriers, the negatively charged electrons and the positively charged holes, are necessary for its function.
  - ***MOS*** **Transistors** or **unipolar Transistors**, more or less only exist in integrated form and usually are drawn as follows:



- **2. Ingredient:** *Insulation.* Always needed between transistors and the other electrically active parts of the circuit.
  - In contrast to circuits soldered together where you simply use air for insulation, it does not come "for free" in **ICs** but has to be made in an increasingly complex way.
- **3. Ingredient:** *Interconnections* between the various transistors or other electronic elements - the wires in the discrete circuit.
  - The way you connect the transistors will determine the function of the device. With a large bunch of transistors you can make everything - a microprocessor, a memory, anything - only the interconnections must change!

- Then we may have *special elements*
  - These might be capacitors, resistors or diodes on your chip. Technically, those elements are more or less subgroups of transistors (i.e. if you can make a transistor, you can also make these (simpler) elements), so we will not consider them by themselves.
- If you did the required reading, you should be familiar with the basic physics of the two transistor types; *otherwise do it now!!!*
  - **Basic bipolar transistor**
  - **Basic MOS transistor**
- The list of necessary ingredients given above automatically implies that we have to use several different materials. At the very minimum we need a *semiconductor* (which is practically always Silicon; only **GaAs** has a tiny share of the **IC** market, too), an *insulator* and a *conductor*. As we will see, we need many more materials than just those three basic types, because one kind of material cannot meet all the requirements emerging from advanced **Si** technology.
  - Since this lecture course is about *electronic materials* , it may appear that all we need now is a kind of list of suitable materials for making integrated circuits. But that would be far too short sighted. In **IC** technology, *materials and processes* must be seen as a unit - one cannot exist without the other.
  - We therefore have to look at both, materials with their specific properties and their integration into a **process flow**.
- Todays integrated circuits contain mostly **MOS** transistors, but we will start with considering the integration of bipolar transistors first. That is not only because historically bipolar transistors were the first ones to be integrated, but because the basic concepts are easier to understand.

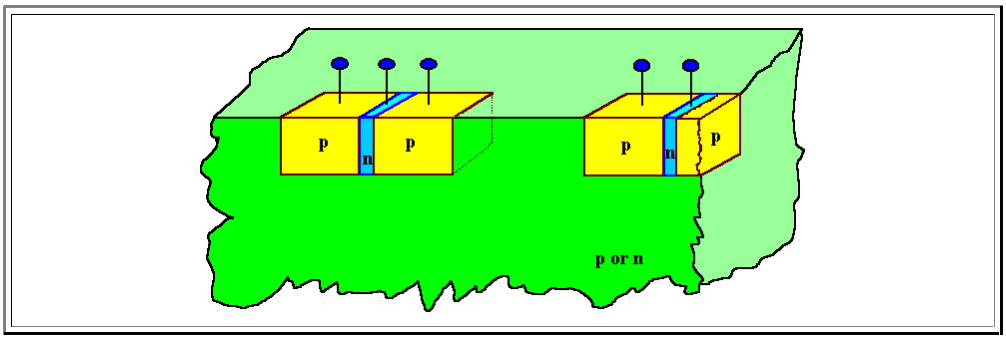
Questionaire

Multiple Choice questions to 5.1.1

5.1.2 Basic Concepts of Integrating Bipolar Transistors

How *Not* to Make an Integrated Bipolar Transistor

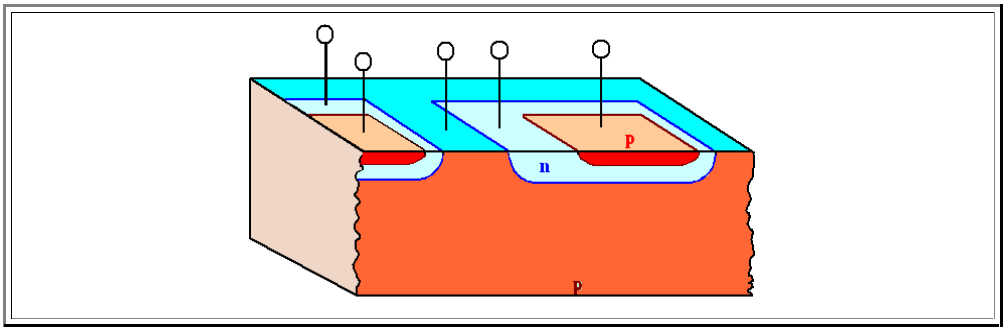
Obviously, embedding the three slices of Si that form a bipolar transistor into a **Si** crystal will not do you any good - we just look at it here to see just how ludicrous this idea would be:



What is the problem with this approach? Many points

- The transistors would not be insulated. The **Si** substrate with a certain kind of doping (either **n**- or **p**-type) would simple short-circuit all transistor parts with the same kind of doping.
- There is not enough place to "put a wire down", i.e. attach the **leads**. After all, the base width should be very small, far less than **1 µm** if possible. How do you attach a wire to that?
- How would you put the sequence of **npn** or **pnp** in a piece of **Si** crystal? After all, you have to get the right amount of , e.g. **B**- and **P**-atoms at the right places.

So we have to work with the really small dimensions in ***z***-direction, into the **Si**. How about the following approach?



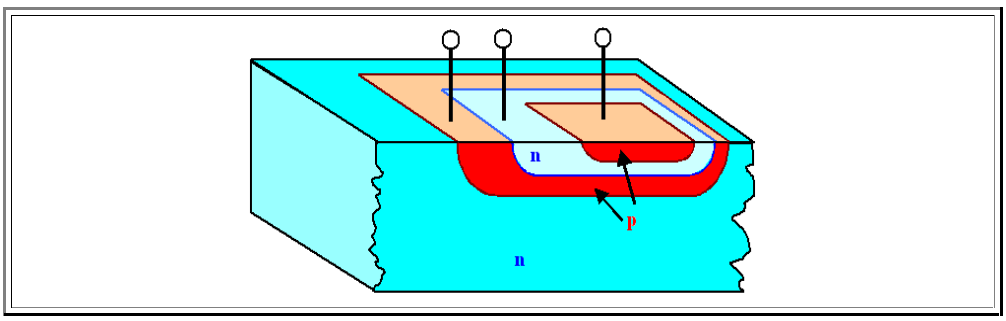
This is much better, but still not too convincing. The *pro* arguments are:

- Enough space for *leads*, because the lateral dimensions can be as large as you want them to be.
- It is relatively easy to produce the doping: Start with **p**-type **Si**, diffuse some **P** into the **Si** where you want the Base to be. As soon as you overcompensate the **B**, you will get **n**-type behavior. For making the emitter, diffuse lots of **B** into the crystal and you will convert it back to **p**-type.
- The base width can be very small (we see about this later).

But there is a major shortcoming:

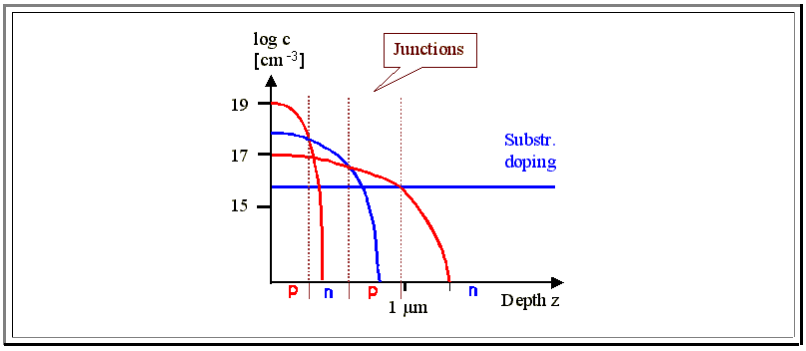
- Still *no insulation* between the collectors - in fact the **Si** crystal is the collector of all transistors and that is not going to be very useful.

Easy you say, lets add another layer of doped **Si**:



This would be fine in terms of insulation, because now there is always a **pn**-junction between two terminals of different transistors which is always blocked in one direction for all possible polarities.

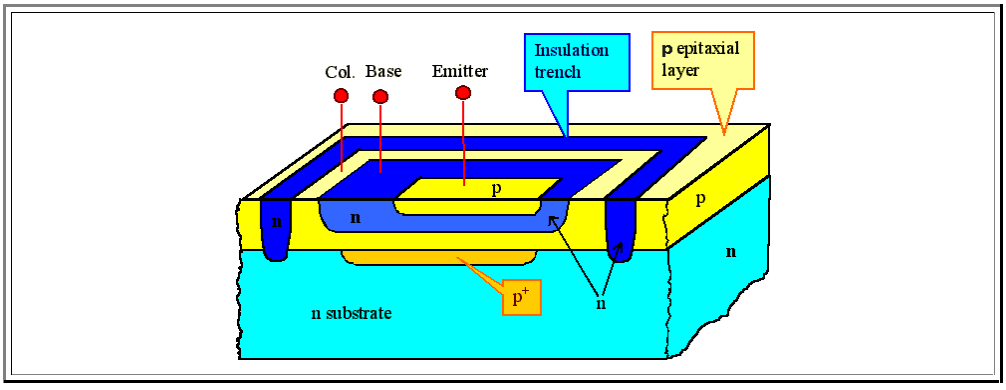
- However, you now have to change an **n**-doped substrate to **p**-doping by over-compensating with, e.g. **B**, then back to **n**-type again, and once more back to **p**-type. Lets see, how that would look in a **diffusion profile** diagram:



- The **lg** of the concentration of some doping element as shown in the illustration above is roughly what you must have - except that the depth scale in modern **ICs** would be somewhat smaller.
  - It is obvious that it will be rather difficult to produce junctions with precisely determined depths. Control of the base width will not be easy.
  - In addition, it will not be easy to achieve the required doping by over-compensating the doping already present three times. As you can see from the diagram, your only way in the resistivity is *down*. If the substrate, e.g., has a doping of **10** Ω**cm**, the collector can only have a lower resistivity because the doping concentration must be larger than that of the substrate, so lets have **5** Ω**cm**. That brings the base to perhaps **1** Ω**cm** and the emitter to **0,1** Ω **cm**. These are reasonable values, but your freedom in designing transistors is severely limited
  - And don't forget: It is the relation between the doping level of the emitter and the base that determines the amplification factor γ
- There must be *smarter* way to produce integrated bipolar transistors. There is, of course, but this little exercise served to make clear that integration is far from obvious and far from being easy. It needs *new ideas*, *new processes*, and *new materials* - and that has not changed from the first generation of integrated circuits with a few **100** transistors to the present state of the art with some **100** million transistors on one chip.
  - And don't be deceived by the *low costs* of integrated circuits: Behind each new generation stands a huge effort of the "best and the brightest" - large scale integration is still the most ambitious technical undertaking of mankind today.

How to Make an Integrated Bipolar Transistor

- So how is it done? *By inventing special processes*, first of all: **Epitaxy**, i.e. the deposition of thin layers of some material on a substrate of (usually, but not necessarily) the same kind, so that the lattice continues undisturbed.
  - Lets look at a cross-section and see what *epitaxy* does and why it makes the production of **IC**s easier.



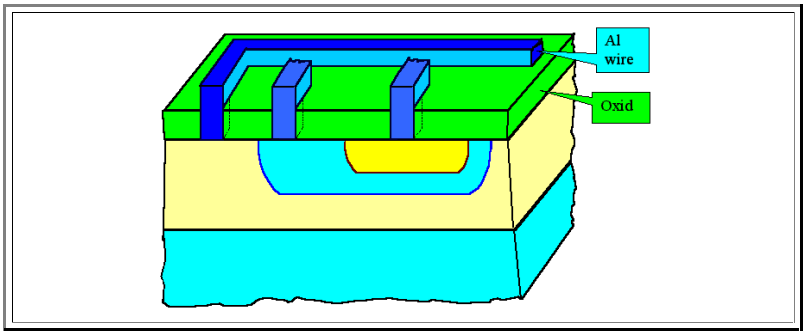
- We start with an **n**-doped wafer (of course you can start with a **p**-doped wafer, too; than everything is reversed) and diffuse the $\mathbf{p^+}$ layer into it. We will see what this is good for right away.
  - On top of this wafer we put an *epitaxial layer of **p**-doped Silicon* , an *epi-layer* as it is called for short. Epitaxial means that the crystal is just continued without change in orientation. The epitaxial layer will always be the collector of the transistor.
  - Next, we diffuse a closed ring of **n**-material around the area which defines the transistor deeply into the **Si**. It will insulate the transistors towards its neighbours because no matter what voltage polarity is used between the collectors of neighbouring transistors, one of the two **pn**-junctions is always in reverse; only a very small leakage current will flow.
  - Then we diffuse the **n**-base- and **p**-emitter region in the epi-layer.
- Looks complicated because it is complicated. But there are many advantages to this approach:
  - We only have *two "critical" diffusions* , where the precise doping concentration matters.
  - The transistor is in the epitaxial layer which, especially in the stone age of integration technology (about from **1970 - 1980**) had a much better quality in terms of crystal defects, level and homogeneity of doping, minority carrier lifetime τ, ...) than the **Si** substrate.
  - We get one level of wiring *for almost free*, the $\mathbf{p^+}$ layer below the transistor which can extend to somewhere else, contacting the collector of another transistor!
- This leads us to the next big problem in integration: The "wiring", or how do we connect transistors in the right way?

Questionaire

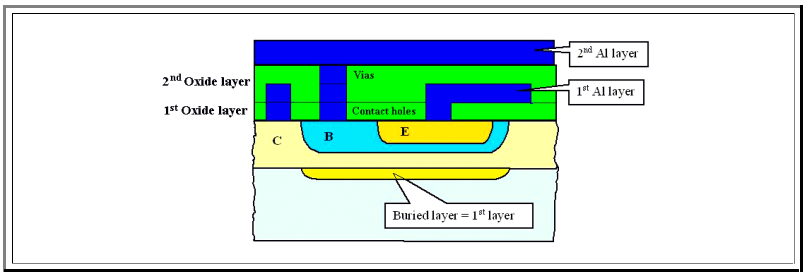
Multiple Choice questions to 5.1.2

5.1.3 Basic Concepts of Connecting Transistors

- How do we connect a few million transistors - i.e. run signal wires from transistor **x** to transistor **y** (**x** and **y** being arbitrary integers between **1** and about **50 000 000**?) and connect *all* transistors to some voltage and current supply - and all that *without wires crossing*?
  - For state-of-the-art **ICs** this is one of the bigger challenges. Obviously you must have wiring on several planes because you cannot avoid that connections must cross each other.
  - The first level is simple enough - in principle! Lets see this in a schematic drawing.

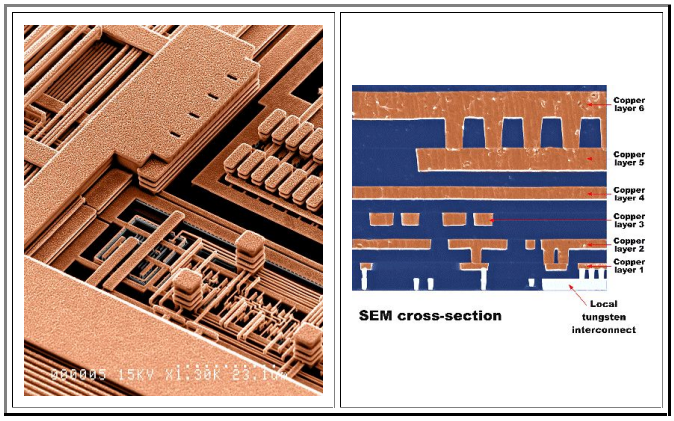


- So all you do is to cover everything with an insulator. For that you are going to use **SiO_2**, which is not only one of the best insulators there is, but is easily produced and fully compatible with **Si**.
  - On top of this oxide you now run your "wires" from here to there, and wherever you want to make a contact to a transistor, you make a **contact hole** in the **SiO_2** layer.
  - Every transistor needs three contact holes - and as you can see in the drawing, you rather quickly run into the problem of crossing connections.
- What we need is a **multi-level metallization**, and how to do this is one of the bigger challenges in integration technology.
  - Fortunately, we already have a second level in the **Si** - it is the "**buried layer**" that we put down before adding the epitaxial layer. It can be structured to connect the collectors of all transistors where this makes sense. And since the collectors are often simply connected to the power supply, this makes sense for most of the transistors.
  - But this is not good enough. We still need more metallization layers on top. So we repeat the "putting oxide down, making contact holes, ..etc". procedure and produce an second metallization layer:



- If you get the idea that this is becoming a trifle complicated, you get the right idea. And you haven't seen anything yet!
  - State-of-the-art **ICs** may contain **7** or more connection (or metallization) layers. For tricky reasons explained later, besides Aluminium (**Al**), Tungsten (**W**) is employed, too, and lately **Al** is being replaced by Copper (**Cu**).
  - Between the metal layers we obviously need an "**intermetal dielectric**". We could (and do) use **SiO_2**; but for modern chips we would rather use something better. In particular, a material with a smaller dielectric constant (**SiO_2** has a value of about **3.7**). Polymers would be fine, in particular polyimides, a polymer class that can "take the heat", i.e. survives at relatively high temperatures. Why we do not have polyimides in use just now is an interesting story that can serve as a prime example of what it means to introduce a new material into an existing product.
- Why are we doing this - replacing trusty old **Al** by tricky new **Cu** - at considerable costs running in the billion **$** range?
  - Because the total resistance ***R*** of an **Al** line is determined by the specific resistivity **ρ = 2,7 $\mu\Omega$cm** of **Al** and the geometry of the line. Since the dimensions are always as small as you can make it, you are stuck with ρ.
  - Between neighbouring lines, you have a parasitic capacitance ***C***, which again is determined by the geometry and the dielectric constant ϵ of the insulator between the lines. Together, a **time constant $R \cdot C$** results, which is directly proportional to $\rho \cdot \epsilon$. This time constant of the wiring - found to be in the **ps** region - gives an absolute upper limit for signal propagation. If you don't see the probem right away, turn to this basic module.
  - In other words: Signal delay in **Al** metallization layers insulated by **SiO_2** restricts the operating frequency of an **IC** to about **1 GHz** or so.

- This was no problem before **1998** or so, because the transistors were far slower anyway. But it is a problem *now* (**2000 +**)!
  - Obviously, we must use materials with lower ρ and ϵ values. Choices are limited, however - **Cu (ρ = 1,7 μΩcm**) is one option that has been chosen; the last word about a suitable replacement for **SiO_2** (having ϵ **= 3,7**) is not yet in.
- Here are famous pictures of an advanced **IBM** chip with **7** metallization layers, completely done in **W** and **Cu**. In the picture on the left, the dielectric between the metals has been etched off, so only the metal layers remain.



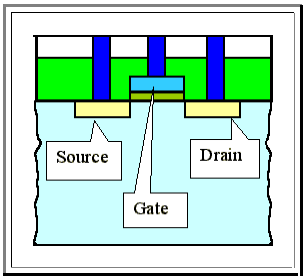
- The transistors are not visible at this magnification - they are too small. You would find them right below the small "local tungsten interconnects" in the cross sectional view.
- Before we go into how one actually does the processes mentioned (putting down layers, making little contact holes, ...), we have to look at how you make *MOS transistors* as opposed to *bipolar* transistors. We will do that in the next sub-chapter.

Questionaire

Multiple Choice questions to 5.1.3

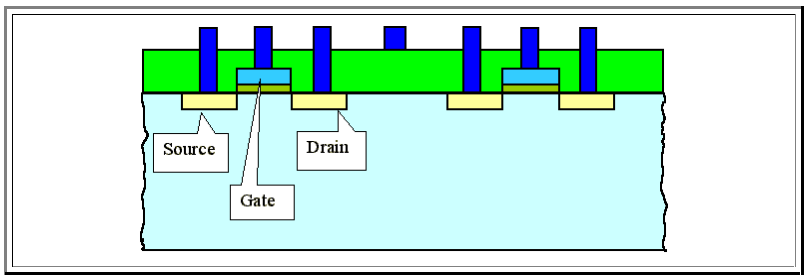
5.1.4 Integrated MOS Transistors

MOS transistors are quite different from bipolar transistors - not only in their basic function, but also in the way they are integrated into a **Si** substrate. Lets first look at the basic structure

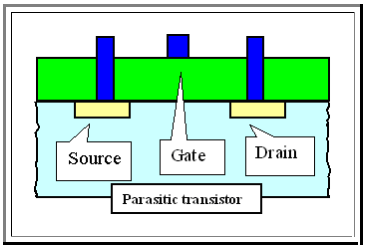


- We have a **source** and **drain** region in the **Si** (doped differently with respect to the substrate) with some connections to the outside world symbolically shown with the blue rectangles. Between source and drain is a thin **gate dielectric** - often called **gate oxide** - on top of which we have the **gate electrode** made from some conducting material that is also connected to the outside world.
- To give you some idea of the real numbers: The thickness of the gate dielectric is below **10 nm**, the lateral dimensioon of the source, gate and drain region is well below **1 µm**.
- You know, of course, what a **MOS** transistor is and how it works - at least in principle. If not: Use the link **Basic MOS transistor**.

If we integrate **MOS** transistors now, it first appears (wrongly!) that we can put them into the same **Si** substrate as shown below:



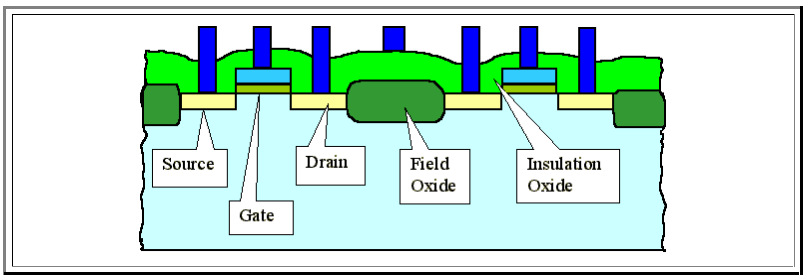
There seems to be no problem. The transistors are insulated from each other because one of the **pn**-junctions between them will always be blocking. *However:* We must also consider "**parasitic transistors**" not intentionally included in our design!



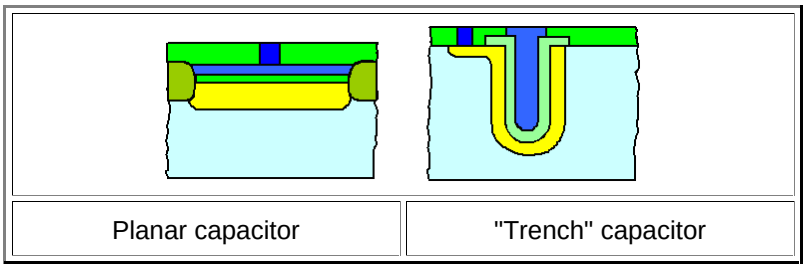
- If in the space between transistors a wire is crossing on top of the insulating layer as shown in the illustration, it will, on occasion be at high potential. The drain of the left transistor together with the source of the right transistor will now form a *parasitic transistor* with the insulating layer as the gate dielectric, and the overhead wire as the gate electrode.
- Everything being small, the threshold voltage may be reached and we have a current path where there should be none.
- This is *not* an academic problem, but a typical effect in *integrated* circuit technology, which is not found in *discrete* circuits: Besides the element you want to make, you may produce all kinds of unwanted elements, too: parasitic transistors, capacitors, diodes, and even thyristors.

The solution is to make the threshold voltage larger than any voltage that may occur in the system. The way to do this is to increase the *local thickness* of the insulating dielectric.

- This gives us the structure in the next illustration



- How we produce the additional insulator called **field oxide** between the transistors will concern us later; here it serves to illustrate two points:
  - Insulation is not just tricky for bipolar transistors, it is a complicated business with **MOS** technology, too
  - There is now some "topology" - *the interfaces and surfaces are no longer flat*. Looks trivial, but constitutes one of the major problems of large-scale integration!
- Note that the gate - substrate part of a **MOS** transistor is, in principle, a **capacitor**. So we can now make capacitors, too.
  - However, if we need a *large* capacitance - say some **50 fF** (femto Farad) - we need a *large* area (several **μm^2**) because we cannot make the dielectric arbitrarily thin - we would encounter tunneling effects, early breakdown, or other problems. So we have to have at least a thickness of around **5 nm** of **SiO_2** . If the capacitor area than gets too large, the escape road is the *third dimension*: You fold the capacitor! Either into the substrate, or up into the layers on top of the **Si**.
  - The "simple" way of folding integrated capacitors into the substrate is shown in the right hand side of the next illustration



Planar capacitor	"Trench" capacitor

- The planar capacitor (on the left) and the "**trench**" capacitor (on the right) have a doped region in the **Si** for the second electrode, which must have its own connection - in the drawing it is only shown for the trench capacitor. We learn two things from that:
  - **1.** Large scale integration has long since become three-dimensional - it is no longer a "planar technology" as it was called for some time. This is not only true for truly three-dimensional elements like the trench capacitor, but also because the processes tend to make the interfaces rough as we have seen already in the case of the field oxide.
  - **2.** The names for certain features generally accepted in the field, are on occasion simply **wrong**! The capacitor shown above is not folded into a *trench* (which is something deep and long in one lateral direction, and small in the other direction), but into a *hole* (deep and small in both lateral directions). Still, everybody calls it a **trench capacitor**.
- The key processes for **ICs** more complex than, say, a **64 Mbit** memory, are indeed the processes that make the surface of a chip halfway flat again after some process has been carried out.
- Again, there is a special message in this subchapter: Integrating **MOS** transistors, although supposedly simpler than bipolar transistors (you don't need all those **pn**-junctions), is far from being simple or obvious. It is again ***intricately*** linked to specific combinations of materials and processes and needs lots of ingenuity, too.
  - But we are still not done in trying to just get a very coarse overview of what integration means. If you take an arbitrary chip of a recent electronic product, changes are that you are looking at a ***CMOS*** chip, a chip made with the "*Complementary Metal Oxide Semiconductor* " technology.
  - So lets see what that implies.

Questionaire

Multiple Choice questions to 5.1.4

5.1.5 Integrated CMOS Technology

Power Consumption Problem

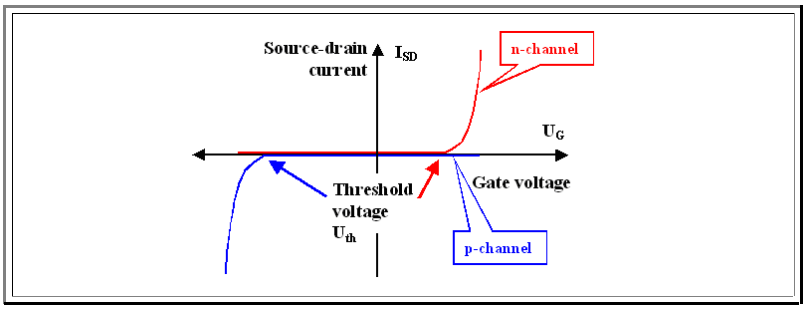
- The first integrated circuits hitting the markets in the seventies had a few **100** transistors integrated in bipolar technology. **MOS** circuits came several years later, even though their principle was known and they would have been easier to make.
  - However, there were insurmountable problems with the *stability* of the transistor, i.e. their threshold voltage. It changed during operation, and this was due to problems with the gate dielectric (it contained minute amounts of alkali elements which are some of many "**IC** killers", as we learned the hard way in the meantime).
  - But **MOS** technology eventually made it, mainly because bipolar circuits need a lot of power for operation. Even for all transistors being "off", the sum of the leakage current in bipolar transistors can be too large for many application.
- **MOS** is principally better in that respect, because you could, in principle, live with only switching voltages; current per se is not needed for the operation. **MOS** circuits do have lower power consumption; but they are also slower than their bipolar colleagues. Still, as integration density increased by an average **60%** per year, power consumption again became a problem.
  - If you look at the data sheet for some state of the art **IC**, you will encounter power dissipations values of up to **1 - 2 Watts** (before **2000**)! Now (**2004**) its about **10** times more. If this doesn't look like a lot, think again!
  - A chip has an area of roughly **1 cm^2** . A power dissipation of **1 Watt/cm^2** is a typical value for the hot plates of an electrical range! The only difference is that we usually do not want to produce french fries with a chip, but keep it cool, i.e. below about **80 ^{o}C**.
- So power consumption is a big issue in chip design. And present day chips would not exist if the ***CMOS*** technique would not have been implemented around the late eighties. Let's look at some figures for some more famous chips:
  - Early Intel microprocessors had the following power rating:

Type	Architecture	Year	No. transistors	Type	Power
4004	4bit	1971	2300	PMOS	
8086	16bit	1978	29000	NMOS	1,5W/8MHz
80C86	16bit	1980	?50000?	**CMOS**	250mW/30Mhz(?)
80386	16bit	1985	275000	**CMOS**	
Pentium 4		2004		**CMOS**	80 W/3 GHz

  - **CMOS** seems to carry the day - so what is **CMOS** technology?

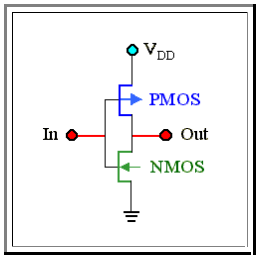
CMOS - the Solution

- Lets first see what "**NMOS**" and ***PMOS***" means. The first letter simply refers to the kind of carrier that carries current flow between source an drain as soon as the threshold voltage is surpassed:
  - *PMOS* stands for transistors where *positively* charged carriers flow, i.e. *holes*. This implies that source and drain must be **p**-doped areas in an **n**-doped substrate because current flow begins as soon as inversion sets in, i.e. the **n**-type **Si** between source and drain is inverted to **Si** with holes as the majority carriers
  - *NMOS* then stands for transistors where negatively charged carriers flow. i.e. electrons. We have **n**-doped source and drain regions in a **p**-doped substrate.
- The characteristics, i.e. the source-drain-current vs. the gate voltage, are roughly symmetrical with respect to the sign of the voltage:



- The red curve may stand for a **NMOS** or **n**-channel transistor, the blue one then would be the symmetrical **PMOS** or **p**-channel transistors. The threshold voltages are not fully symmetric if the same gate electrode is used because it depends on the difference of the Fermi energies of the gate electrode materials and the doped **Si**, which is different in the two cases.
- Anyway, for a given gate voltage which is larger than either threshold voltage applied to the transistor, one transistor would be surely "on", the other one "off".

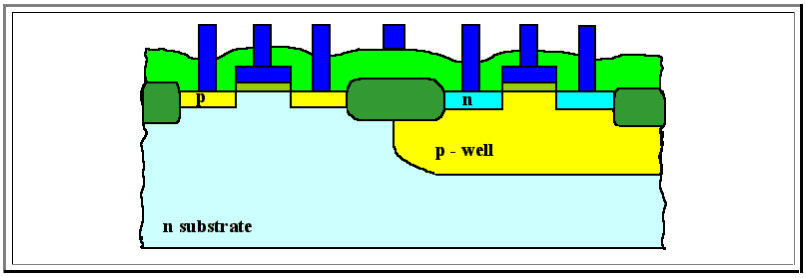
So if you always have a **NMOS** and a **PMOS** transistor in series, there will *never* be any static current flow; we have a small dynamic current component only while switching takes place.



- Can you make the necessary logical circuits this way?
- Yes you can - at least to a large extent. The illustration shows an **inverter** - and with inverters you can create almost anything!
- Depending on the right polarities, the blue **PMOS** transistor will be closed if there is a gate voltage - the output then is zero. For gate voltage zero, the green **NMOS** transistor will be closed, the **PMOS** transistor is open - the output will be $\mathbf{V_{DD}}$ (the universal abbreviation for the *supply voltage*).

So now we have to make *two* kinds of transistors- **NMOS** and **PMOS** - which needs substrates with different kind of doping - in *one* integrated circuit. But such substrates do not exist; a Silicon wafer, being cut out of an homogeneous crystal, has always *one* doping kind and level.

- How do we produce differently doped areas in an uniform substrate? We remember what we did in the bipolar case and "simply" add another diffusion that converts part of the substrate into the different doping kind. We will have to diffuse the right amount of the compensating atom rather deep into the wafer, the resulting structure is called a **p**- or **n**-**well**, depending on what kind of doping you get.
- If we have a **n**-type substrate, we will have to make a **p**-well. The **p**-well then will contain the **NMOS** transistors, the original substrate the **PMOS** transistors. The whole thing looks something like this:



By now, even the "simple" **MOS** technology starts to look complicated. But it will get even more complicated as soon as you try to put a metallization on top. The gate structure already produced some "roughness", and this roughness will increase as you pile other layers on top.

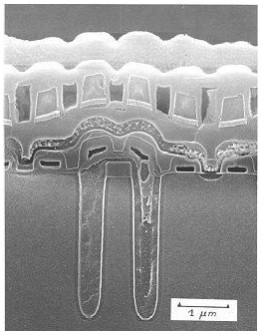
- Let's look at some specific metallization problems (they are also occurring in bipolar technology, but since you start with a more even surface, it is somewhat easier to make connections).
- A cross-section through an early ***16 Mbit DRAM*** (**DRAM =** *Dynamic Random Access Memory*; the work horse memory in your computer) from around **1991** shown below illustrates the problem: The surface becomes exceedingly wavy. (For enlarged views and some explanation of what you see, click on the image or the link)

Adding more metallization layers becomes nearly impossible. Some examples of the difficulties encountered are:

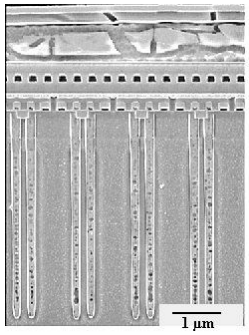
- **1.** With wavy interfaces, the thickness between two layers varies considerably, and, since making connection between layers need so-called "**via**" holes, the depths of those *vias* must vary, too. This is not easily done! And if you make all vias the same (maximum) depth, you will etch deeply into the lower layer at places where the interlayer distances happens to be small.
- **2.** It is very difficult to deposit a layer of anything *with constant thickness* on a wavy surface.

- **3.** It is exceedingly difficult to fill in the space between **Al** lines with some dielectric without generating even more waviness. The problem then gets worse with an increasing number of metallization layers.

The **64 Mbit DRAM**, in contrast, is very flat. A big break-through in wafer processing around **1990** called "**Chemical mechanical Polishing**" or ***CMP*** allowed to planarize wavy surfaces.



Cross section 16 Mbit DRAM (Siemens)



Cross section 64 Mbit DRAM (Siemens)

State of the Art

Lets get some idea about the state of the art in (**CMOS**) chip making in the beginning of the year **2000**. Above you can look at cross-sectional pictures of a **16 Mbit** and a **64 Mbit** memory; the cheap chip and the present work horse in memory chips. The following data which come from my own experience are not extremely precise but give a good impression of what you can buy for a few Dollars.

Property	**Number**
Feature size	**0,2 µm**
No. metallization levels	**4 - 7**
No. components	**> 6 · 10^8 (Memory)**
Power	**several W/cm^2**
Speed	**600 MHz**
Lifetime	**> 10 a**
Price	**$2 (memory) up to $ 300 (microprocessor)**
Complexity	**> 500 Process steps**
Cost (development and **1** factory)	**ca. $ 6 · 10^9**

How will it go on? Who knows - but there is always the official semiconductor roadmap from the **Semiconductor Industry Associaton** (***SIA***)

- That's it. Those are holy numbers which must not be doubted. Since they are from **1993**, the predictive power can be checked.

Semiconductor Industry Association Roadmap (1993)							
		1992	**1995**	**1998**	**2001**	**2004**	**2007**
Feature size (µm)		0.5	0.35	0.25	0.18	0.12	0.1
Bits/Chip	DRAM	16M	64M	256M	1G	4G	16G
	SRAM	4M	16M	64M	256M	1G	4G
Chip size (mm^2)	Logic / microprocessor	250	400	600	800	1000	1250
	DRAM	132	200	320	500	700	1000
Performance (MHz)	on chip	120	200	350	500	700	1000
	off chip	60	100	175	250	350	500

Maximum power (W/chip)	high performance	10	15	30	40	40-120	40-200
	portable	3	4	4	4	4	4
Power supply voltage (V)	desktop	5	3.3	2.2	2.2	1.5	1.5
	portable	3.3	2.2	2.2	1.5	1.5	1.5
No. of interconnect levels - logic		3	4-5	5	5-6	6	6-7
Number of I/Os		500	750	1500	2000	3500	5000
Wafer processing cost ($/cm^2)		$4.00	$3.90	$3.80	$3.70	$3.60	$3.50
Wafer diameter (mm)		200	200	200-400	200-400	200-400	200-400
Defect density (defects/cm^2)		0.1	0.05	0.03	0.01	0.004	0.002

Questionaire

Multiple Choice questions to 5.1.5

5.1.6 Summary to: 5.1 Basic Considerations for Process Integration

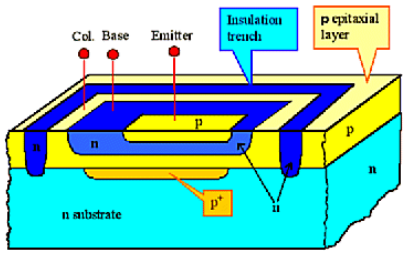
Integration means:

- **1.** Produce a large number (up to **1.000.000.000**) of transistors (bipolar or **MOS**) and other electronic elements on a **cm^2** of **Si**
- **2.** Keep thoses elements electrically insulated from each other.
- **3.** Connect those elements in a meaningful way to produce a system / product.

It ain't easy!

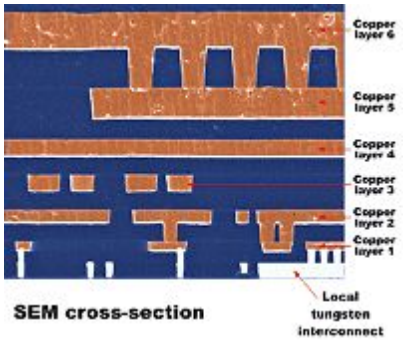
An integrated bipolar transistor does not resemble the textbook picture at all, but looks far more complicated ⇒.

- This is due to the insulation requirements, the process requirements, and the need to interconnect as efficiently as possible.
- The epitaxial layer cuts down on the number of critical diffusions, makes insulation easier, and allows a "buried contact" structure.



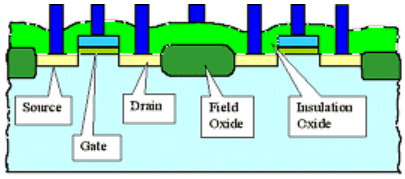
Connecting transistor / elements is complicated; it has to be done on several levels

- Materials used are **Al** ("old"), **Cu** ("new"), **W**, (highly doped) poly-**Si** as well as various silicides.
- Essential properties are the conductivity σ of the conductor, the dielectric constant ϵ_r of the intermetal dielectric, and the resulting time constant $\tau = \sigma \cdot \epsilon_r$ that defines the maximum signal transmision frequency through the conducting line.



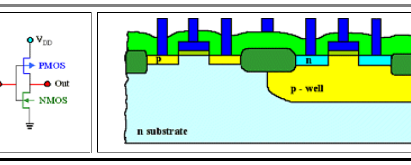
Integrating **MOS** transistors requires special measures for insulation (e.g. a field oxide) and for gate oxide production

- Since a **MOS** transistor contains intrinsically a capacitor (the gate "stack"), the technology can be used to produce capacitors, too.



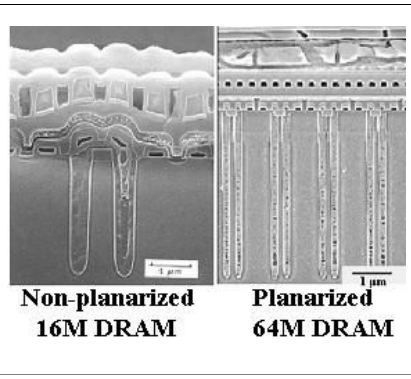
CMOS allows to reduce power consumption dramatically.

- The process, however, is more complex: Wells with different doping type need to be made.



Using the third dimension (depth / height) might become necessary for integrating "large" structures into a small projected are (example: trench capacitor in **DRAMs** ⇒).

- Unwanted "topology", however, makes integration more difficult.
- Planarized technologies are a must since about 1995! ⇒



It ain't neither easy nor cheap!

Property	Number
Feature size	**0,2 µm**
No. metallization levels	**4 - 7**
No. components	**> 6 · 10^8 (Memory)**
Complexity	**> 500 Process steps**
Cost (development and **1** factory)	**ca. $ 6 · 10^9**

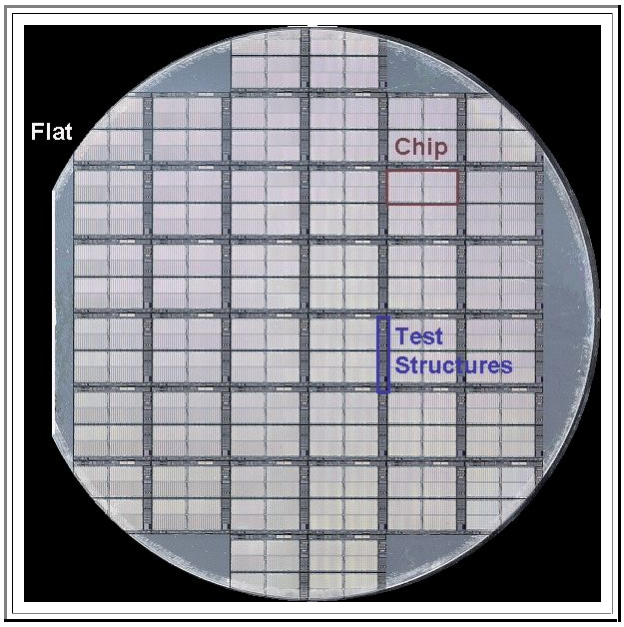
Questionaire

Multiple Choice questions to all of 5.1

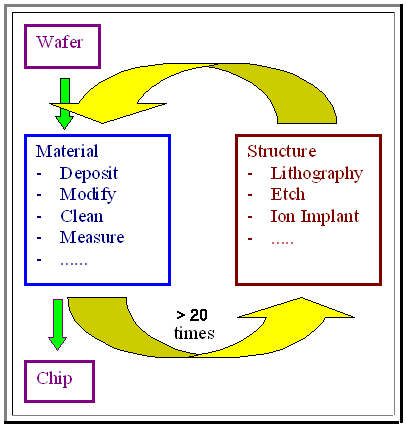
5.2 Process Integration

5.2.1 Chips on Wafers

- We now have a crude idea of *what* we want to make. The question now is *how* we are going to do it.
  - We start with a suitable piece of a **Si** crystal, a **Si wafer**. A wafer is a thin (about **650 µm)** round piece of rather perfect **Si** single crystal with a typical diameter (in the year **2000**) of **200 mm**. Nowadays (**2007**) you would build you factory for **300 mm**.
  - On this wafer we place our **chips**, square or rectangular areas that contain the complete integrated circuit with dimensions of roughly **1 cm^2**.
  - The picture below shows a **150 mm** wafer with (rather large **1st** generation) **16 Mbit DRAM** chips and gives an idea about the whole structure.



- The chips will be cut with a diamond saw from the wafer and mounted in their casings.
  - Between the chips - in the area that will be destroyed by cutting - are test structures that allow to measure certain technology parameters.
  - 'The (major) **flat**" of the wafer is aligned along a **<110>** direction and allows to produce the structures on the wafer in perfect alignment with crystallographic directions. It also served to indicate the crystallography and doping type of the wafer; consult the link for details
  - Don't forget: **Si** is *brittle like glass*. Handling a wafer is like handling a thin glass plate - if you are not careful, it breaks.
- How to get the chips on the wafer? In order to produce a **CMOS** structure as shown before, we essentially have to go back and forth between two two basic process modules:
- *Material module*
  - Deposit some material on the surface of the wafer (e.g. **SiO_2**), or
  - Modify material already there (e.g. by introducing the desired doping), or
  - Clean the material present, or
  - Measure something relative to the material (e.g. its thickness), or
  - - well, there are a few more points like this but which are not important at this stage.
- *Structuring module*
  - Transfer the desired structure for the relevant material into some light sensitive layer called a photo-resist or simply **resist** (which is a very special *electronic material*!) by **lithography**, i.e. by projecting a slide (called a **mask** or more generally **reticle**) of the structure onto the light sensitive layer, followed by developing the resist akin to a conventional photographic process, and then:
  - Transfer the structure from the resist to the material by **structure etching** or other techniques.
- Repeat the cycle *more than **20** times* - and you have a wafer with fully processed chips.
  - This is shown schematically in the drawing:



- For the most primitive transistor imaginable, a minimum of **5** lithographic steps are required. Each process module consists of many individual **process steps** and it is the art of **process integration** to find the optimal combination and sequence of process steps to achieve the desired result in the most economic way.

It needs a lot of process steps - most of them difficult and complex - to make a chip.

- Even the most simple **5** mask process requires about **100** process steps.
- A **16 Mbit DRAM** needs about **19** masks and **400** process steps.

To give an idea what this contains, here is a list of the ingredients for a **16 Mbit DRAM** at the time of its introduction to the market (with time it tends to become somewhat simpler):

- **57** layers are deposited (such as **SiO_2** (**14** times), **Si_3N_4**, **Al**, ...).
- **73** etching steps are necessary (**54** with "plasma etching", **19** with wet chemistry).
- **19** lithography steps are required (including deposition of the resist, exposure, and development).
- **12** high temperature processes (including several oxidations) are needed.
- **37** dedicated cleaning steps are built in; wet chemistry occurs **150** times altogether.
- **158** measurements take place to assure that everything happened as designed.

A more detailed rendering can be found in the link.

Two questions come to mind:

- *How long does it take to do all this?* The answer is: *weeks* if everything always works and you never have to wait, and *months* considering that there is no such thing as an uninterrupted process flow all the time.
- *How large is the success rate?* Well, lets do a back-of-the-envelope calculation and assume that each process has a success rate of ***x* %**. The overall **yield *Y*** of working devices is then
  $Y = (x/100)^N$ % with N = number of process steps. With $N = 450$ or **200** we have

x	*Y* for *N* = 450	*Y* for *N* = 200
95%	$9{,}45 \cdot 10^{-9}$ %	$3{,}51 \cdot 10^{-3}$ %
99%	1,09 %	13,4 %
99,9%	63,7 %	81,9 %

- **N = 200** might be more realistic, because many steps (especially controls) do not influence the yield very much.

But whichever way we look at these numbers, there is an *unavoidable conclusion* : Total perfection at each process step is absolutely necessary!

Questionaire

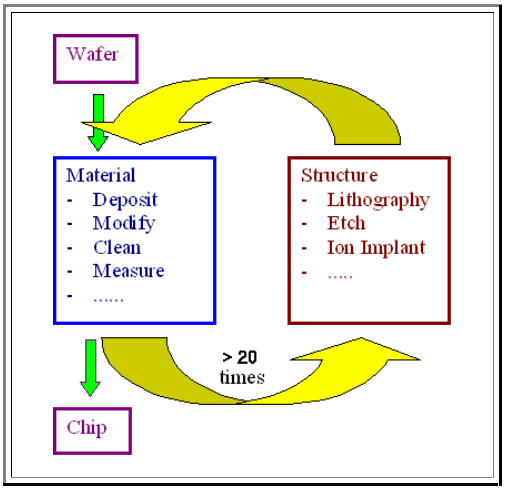
Multiple Choice Questions to 5.2.1

5.2.2 Packaging and Testing

- Very complex, very important - and not covered here. Sorry

5.2.3 Summary to: Chips on Wafers

- Typical wafer size for new factories (**2007**) : **300 mm** diameter, **775 µm** thickness, flatness in lower **µm** region
  - Chip size a few **cm^2**, much smaller if possible
  - Yield ***Y*** = most important parameter in chip production = % of chips on a wafer that function (= can be sold).
  - ***Y* = 29 %** is a good value for starting production
- Chip making = running about **20** times (roughly!!) through "materials" - "structuring" loop.
  - About **400 - 600** individual processing steps (= in / out of special "machine") before chip is finished on wafer
  - More than **30** processing steps for packaging (after separation of chips by cutting)
  - Simple estimate: **99.9%** perfection for each processing step means***Y* < 70 %.**

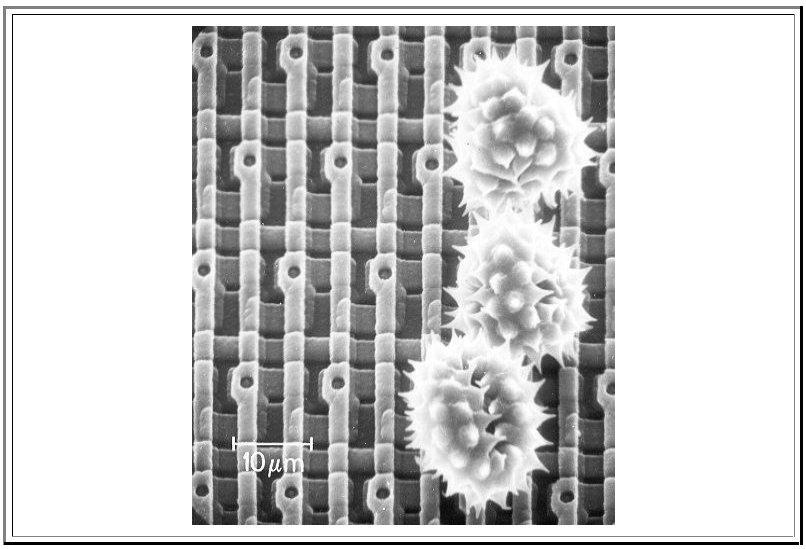


5.3 Cleanrooms, Particles and Contamination

5.3.1 Cleanrooms and Defects

Particles

- Normal air is full of "dirt" usually called **particles**. The fact that *we* cannot see them (except the bigger ones in a bright beam of light) does not mean that the air is clean. What happens when a particle (e.g. pollen, scrapings from whatever, unknown things) falls on a chip is shown in the picture below.
  - Anything that can "fall" on a chip is called a *particle*; independent of its size and of what it is. Particles smaller than some **10 µm** usually do not "feel" gravity anymore (other forces dominate); so they do not "fall" on a chip. However, they may be attracted electrostatically and that makes it quite difficult to remove them.
  - Often anything that disturbs the structure of a chip by lying *on* the layers of the integrated circuit is called a **"defect"**. Defects may not only be particles, but all kinds of other mishaps too, e.g. small holes in some coating.
  - However, we will *not* use that terminology here, but restrict the name "*defect*" to **crystal lattice defects** *in* the **Si**, i.e. *in*, not *on*, the integrated circuit.
  - A pretty old chip (a **256k** memory as sold around **1985**) was chosen for the following illustration because its structures are clearly visible. It has a few pollen grains (from "Gänseblümchen") on its surface (which essentially shows the wiring matrix of a memory array). What would have happened if a pollen grain would have fallen on the chip while it was made needs no long discussion: The chip would be dead!

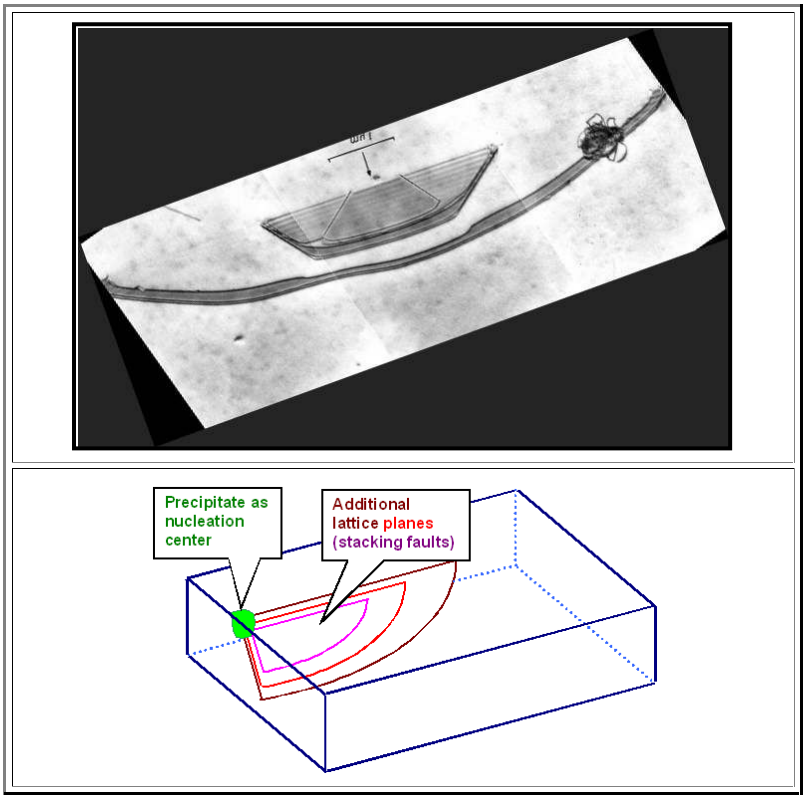


- At feature sizes **< 0,2 µm**, everything that falls on a chip with sizes **> 0,1 µm** or so will be deadly. All those defects- the **particles** - must be avoided at all costs. There are three major sources of particles:
- **1.** The *air* in general. Even "clean" mountain air contains very roughly 10^6 particles **> 1 µm** per *cubic foot* (approximately **30** liters). We need a "**cleanroom**" serving two functions:
  - It provides absolutely clean air (usually through filters in the ceiling), and
  - It immediately removes particles generated somewhere in the cleanroom by pumping large amounts of clean air from the ceiling through the (perforated) floor.
  - Avoiding and removing particles while processing **Si** wafers has grown into a science and industry of its own. The link provides some information about cleanrooms and cleanroom technology.
- **2.**The *humans* working in the cleanroom.
  - Wiping your (freshly washed) hair just once, will produce some **10 000** particles. If you smoke, you will exhale thousands of particles (size about **0,2 µm**) with every breath you take. A ***TI*** add once said: Work for us and we will turn you into a non-smoker.
  - The solution is to pack you into **cleanroom garments**; what this looks like can be seen in the link. It is not as uncomfortable as it looks; but it is not pure fun either. Graphic examples of humans as a source of particles can be found in the link.
- **3.** The *machines* (called "**equipment**") that do something to the chip, may also produce particles.
  - As a rule, *whenever something slides on something else* (and this covers most mechanical movements), particles are produced. Layers deposited on chips are also deposited on the inside of the equipment; they might flake off. There is no easy fix, but two rules:
  - - Use special engineering and construction to avoid or at least minimize all possible particle sources, and
    - Keep your equipment clean - frequent "special" cleaning is required!

- But even with state-of-the art cleanrooms, completely covered humans, and optimized equipment, particles can not be avoided - look at the picture gallery to get an idea of what we are up to. The most frequent process in chip manufacture therefore is "cleaning" the wafers.
  - Essentially, the wafers are immersed in special chemicals (usually acids or caustics or in combination with various special agents), agitated, heated, rinsed, spin-dried, ... , its not unlike a washing machine cycle.
  - This cleaning process in all kinds of modifications is used not only for removing particles, but also for removing unwanted atoms or layers of atoms which may be on the surface of the wafers. This brings us to the next point:

Contamination and Crystal Lattice Defects

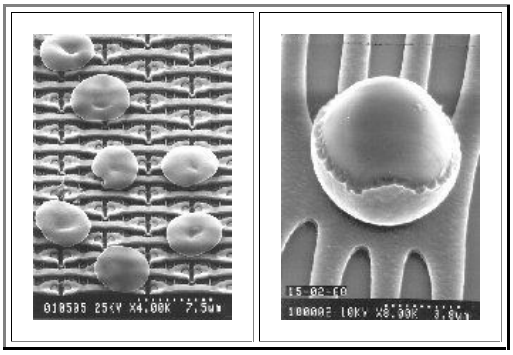
- The **Si** single crystals used for making integrated circuits are the (thermodynamically) most perfect objects in existence - at least on this side of Pluto. They are in particular completely free of dislocations and coarser defects as, e.g., grain boundaries or precipitates of impurities, and have impurity concentrations typically in the **ppt** (parts per trillion) or **ppqt** (parts per quadrillion) range - many orders of magnitude below of what is normally considered "high-purity".
- *Defects* (always now in the meaning of "*crystal lattice defects*") will without fail kill the device or change properties!
  - *Dislocations* or *precipitates* in the electronically active region of the device; e.g. in or across **pn**-junctions, simply "kill" it - the junction will be more or less short-circuited.
  - *Point defects* in solid solution (e.g. **Cu**, **Au**, **Fe**, **Cr**, ...most metals) in the **Si** crystal reduce the *minority carrier lifetime* and thus influence device characteristics directly - usually it will be degraded. Alkali atoms like **Na** and **K** in the gate oxides kill **MOS** transistors, because they move under the applied electrical field and thus change the charge distribution and therefore the transistor characteristics.
  - But point defects do more: If they precipitate (and they all have a tendency to do this because their solubility at low temperatures is low) close to a critical device part (e.g. the interface of **Si** and SiO_2 in the **MOS** transistor channel), they simply kill that transistor. Possibly worse: Even very small precipitates of impurities may act as the nuclei for large defects, e.g. dislocations or stacking faults, that without help at the nucleation stage would not have formed.
- This simply means that we have to keep the **Si**-crystal free of so-called **process-induced defects** during processing, something not easily achieved. Cleaning helps in this case, too.
  - Below a picture of what a process-induced defect may look like. It was taken by a *transmission electron microscope* (***TEM***) and shows the projection of a systems of stacking faults (i.e. additional lattice planes bounded by dislocations) extending from the surface of the wafer into the interior. The schematic picture outlines the three-dimensional geometry



- The central ***precipitate*** that nucleated the stacking fault system is visible as a black dot. The many surplus **Si** atoms needed to form the excessive lattice planes were generated during an oxidation process.
  - Oxidation liberates **Si** interstitials which, since in supersaturation, tend to agglomerate as stacking faults.
  - However, without "help", the nucleation barrier for forming an extended defects can not be overcome, the interstitials then diffuse into the bulk of the crystal were they eventually become immobile and are harmless.
- Defects like the one above are known as "**oxidation induced stacking faults**" or ***OSF***. They form in large densities if even trace amounts of several metals are present which may form precipitates. In order to provide enough metal atoms, it is sufficient to hold the wafer just once with a metal tweezer and subject it to a high temperature process afterwards.
- There are many more ways to generate lattice defects, but there are two golden rules to avoid them:
  - *1. Keep the crystal clean!*
    Even **ppt** of **Fe**, **Ni**, **Cr** (i.e. stainless steel) **Cu**, **Au**, **Pt** or other notorious metals will, via a process that may develop through many heat cycles, eventually produce large defects and kill the device.
  - *2. Keep temperature gradients low!*
    Otherwise mechanical stress is introduced which, if exceeding the yield strength of **Si** (which decreases considerably if impurity precipitates are present), will cause plastic deformation and thus the introduction of large amounts of **dislocations**, which kill your device.
- Via the link a gallery of **process-induced defects** can be accessed together with short comments to their nature and how they were generated.

- There is a simple lecture that can be learned from this: *Electronic Materials* in the context of microelectronics comprise not only the semiconductors, but
  - Anything that can be found in the finished product - the casing (plastics, polymers. metal leads, ..), the materials on and in the **Si**, the **Si** or **GaAs** or... .
  - Anything directly used in making the chip - materials that are "sacrificial", e.g. layers deposited for a particular purpose after which they are removed again, the wet chemicals used for cleaning and etching, the gases, etc.
  - Anything used for handling the chip - the mechanisms that hold the **Si** in the apparatus or transport it, tweezers, etc.
  - Anything in contact with these media - tubing for getting gases and liquids moved, the insides of the processing equipment in contact with the liquid or gaseous media, e.g. furnace tubes.
  - Anything in possible contact with these parts - and so on! It never seems to end - make one mistake (the wrong kind of writing implement that people use in the cleanroom to make notes (not on (dusty) paper, of course, but on clean plastic sheets) - and you may end up with non-functioning chips.
- The link provides a particularly graphic example of how far this has to go!

5.3.2 Summary to: 5.3 Cleanrooms, Particles and Contamination



- Dirt in any form - as "particles" on the surface of wafer, or as "contamination" inside the wafer is almost always deadly
  - Particles with sized not much smaller than minimum feature sizes (i.e. < **10 nm** in **2007**) will invariably cover structures and lead to local dysfunction of a transistor or whatever.
  - Point defects like metal atoms in the **Si** lattice may precipitate and cause local short circuits etc. from the "inside", killing transistors
  - One dysfunctional transistor out of **1.000.000.000** or so is enough to kill a chip!

- Being extremely clean is absolutely mandatory for high Yields ***Y***!
  - Use cleanrooms and hyper-clean materials!
  - It won't be cheap!

5.4 Development and Production of a New Chip Generation

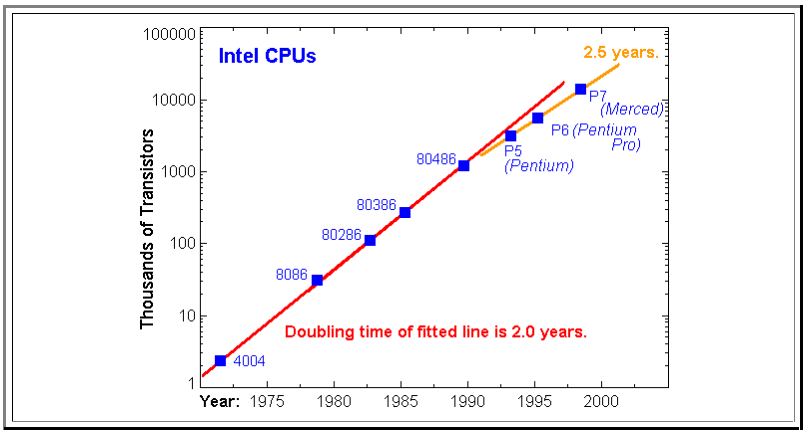
5.4.1 Money, Time, and Management

You may be the foremost *materials* expert in the world, but if you try to leave your mark on chip development without regard to some boundary conditions of a more *economical* nature, you will not achieve much. And if you are the *manager* (which you should be with the kind of education you get here), you better be aware of the following points that are special to research, development and manufacture of (memory) chips.

- There is no other product with quite such brutal requirements, even considering that *all* technical product development must follow similar (but usually much more relaxed) rules.

***1.** A new generation with four-fold capacity will appear on the market every three years.*

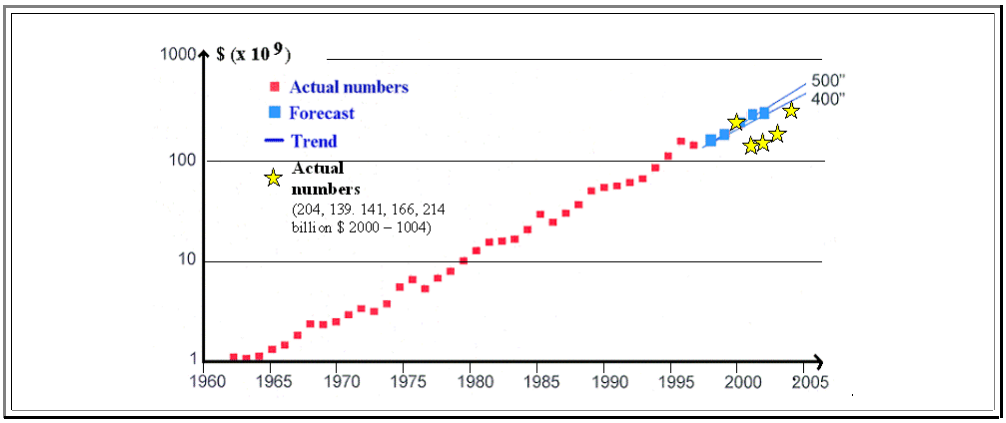
- That is an expression of "**Moore's law**". It is, of course, not a "law" but an extrapolation from observation and bound to break down in the not so distant future (with possible disastrous consequences to the economy of the developed countries).
- The original observation made in **1965** by Gordon **Moore**, co-founder of **Intel**, was that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future. In subsequent years, the pace slowed down a bit, but data density has doubled approximately every **18** months, and this is the current definition of Moore's Law, which Moore himself has blessed. Most experts, including Moore himself, expect Moore's Law to hold for at least another two decades.
- Here is a graphic representation for microprocessors:



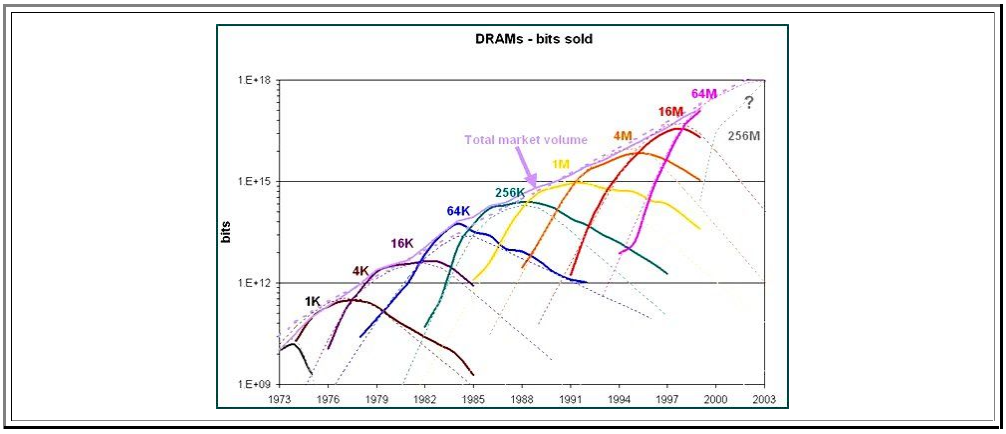
- Not bad - but of course Moore's law must break down sometime (at the very latest when the feature size is akin to the size of an atom (when would that be assuming that the feature size of the **P7** is **0,2 µm**?). This is illustrated in a separate module.
- Still, as long as it is true, it means that you either have your new chip generation ready for production at a well-known time in the future, or you are going to loose **large amounts of money**. There are some immediate and unavoidable consequences:
- You must spent *large amounts of money* to develop the chip and to built the new factory **2 - 3** years before the chip is to appear on the market, i.e. at a time were you do not know if chip development will be finished on time. And *large* means several billion **$**.
- The time allowed for developing the new chip generation is a constant: You can't start early, because everything you need (better lithography, new materials,) does not exist then. But since chip complexity is ever increasing, you must do more work in the same time. The unavoidable conclusion is more people and *shift work*, even in research and development.
- It follows that you need ever increasing amounts of money for research and development of a new chip generation (there is a kind of Moore's law for the costs of a new generation, too). Look at it in another way in a separate module.

***2.** The market for chips grows exponentially*

- That is an expression of the insatiable demand for chips - as long as they provide more power for less money! That this statement was true is shown below. Note *that the scale is logarithmic*!

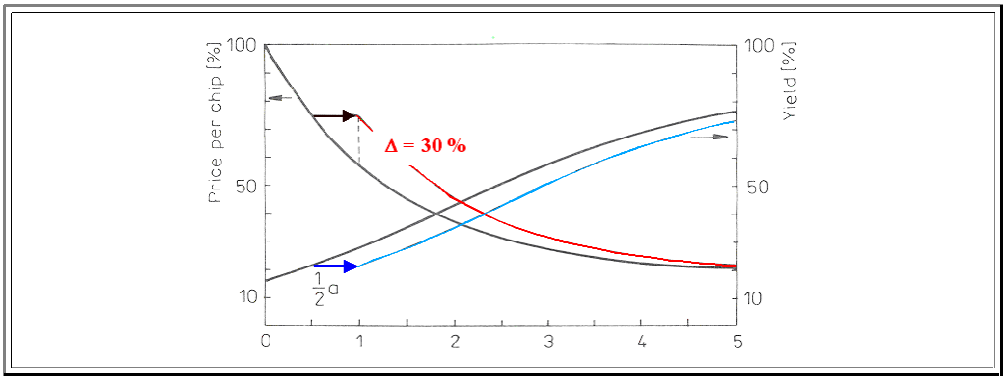


- Shown is the total amount of money grossed in the semiconductor market from **1960 - 2000** (Source: Siemens / Infineon).
  - The essentially straight line indicates exponential growth - including the foreseeable future. Extrapolations, however, are still difficult. The two extrapolated (in **2000)** blue lines indicating a difference of **100.000.000.000 $** of market volume in **2005** are rather close together. The error margins in the forecast thus correspond to the existence or non-existence of about **10** large semiconductor companies (or roughly **150.000** jobs).
  - We also see the worst downturn ever in **2001**. Sales dropped from **204** billion $ in **2000** to **139** billion $ in **2001**, causing major problems throughout the industry.
- More specific, we can see the exponential growth of the market by looking at sales of **DRAMs**. Shown is the total number of memory *bits* sold. Note that just a fourfold increase every three years would keep the number of *chips* sold about constant because of the fourfold increase of storage capacity in one chip about every three years.



  - The unavoidable consequence is: Your production capacity must grow exponentially, too, if you just want to keep your share of the market. You must pour an *exponentially growing amount of money* into investments for building factories and hiring people, while the returns on these investments are delayed for at least **2 - 3** years. In other words: the difference of what you must spent and what you earn increases, very roughly, exponentially. This is not a healthy prospect for very long, and you must make *large amounts of money* every now and then (e.g. by being the first one on the market with a new chip or by having a quasi monopoly).
  - You must make (and sell at a profit) an *exponentially increasing number of chips* (since the prize per chip is roughly constant) to recover your ever increasing costs. Since chip sizes increase and prices must stay halfway constant, you must use larger wafers to obtain more chips per processing. This puts a lot of pressure on developing larger **Si** crystals and equipment to handle them.
  - You must produce as many chips as you can during the product life time (typically **5** years). Continuous shift work in the factory (**7** days a week, **24** hours a day) are absolutely mandatory!
- ***3**. Chip prices for memories decay exponentially from the time of their market introduction (roughly **$60**) by two orders of magnitude within about **5** years (i.e. at the end you get **$1**).*
  - The prize development up to the **16 Mbit DRAM** can be seen in an illustration via the link. Microprocessors may behave very differently (as long as Intel has a quasi monopoly). The rapid decay in prizes is an expression of fierce competition and mostly caused by:
  - **1.** The "**learning curve**", i.e. the increase of the percentage of good chips on a wafer (the **yield**) from roughly **15%** at the beginning of the production to roughly **90%** at the end of the product life time (because you keep working like crazy to improve the yield).
  - **2.** Using a "**shrink strategy**" . This means you use the results of the development efforts for the next two generations to make your present chip smaller. Smaller chips mean more chips per wafer and therefore cheaper chips (the costs of making chips are mostly the costs of processing wafers).

- An immediate consequence is that if you fall behind the mainstream for **6** months or so - you are dead! This can be easily seen from a simple graph:



- The descending black curve shows the expected trend in prizes (it is the average exponential decay from the illustration). The ascending curve is the "learning curve" needed just to stay even - the cost of producing one chip then comes down exactly as the expected prize.
- Now assume you fall behind **6** month - your learning curve, i.e. your yield of functioning chips does not go up. You then move on the modified blue learning curve. The prize you would have to ask for your chip is the modified red prize curve - it is **30 %** above the expected world market prize in the beginning (where prizes are still moderately high).
- Since nobody will pay more for your chips, you are now **30 %** behind the competition (much more than the usual profit margins) - you are going to loose *large amounts of money*!
- In other word: You *must* meet the learning curve goals! But that is easier said then done. Look at real yield curves to appreciate the point

To sum it up: If you like a quiet life with just a little bit of occasional adrenaline, you are better off trying to make a living *playing poker in Las Vegas*.

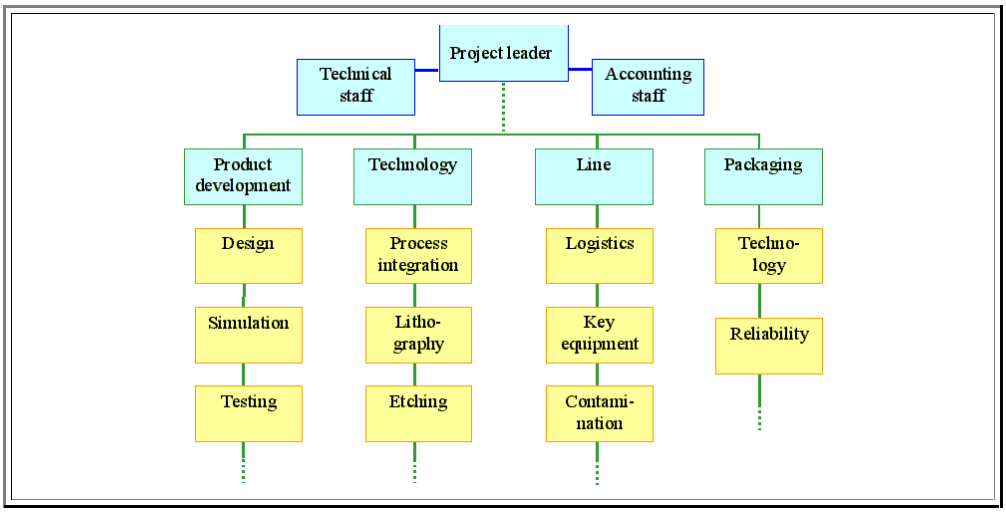
- Developing and producing new chip generation in our times is a quite involved gamble with billions at stake! There are many totally incalculable ingredients and you must make a lot of million **$** decisions by feeling and not by knowledge!
- So try at least to know what can be known. And don't forget: *It's fun*!

To give a few more data, here is a table with many numbers:

Type	**4 kb**	**16 kb**	**64 kb**	**256 kb**	**1 Mb**	**4 Mb**	**16 Mb**	**64 Mb**	**256 Mb**	**1 Gb**	**4 Gb**
Begin of production	1974	1976	1979	1982	1985	1988	1991	1994	1997	2001	2004
Equivalent of type written pages	0,23	1	4	16	64	250	1000	4000	16000	64000	250000
	Growth per year about **+ 60 %**										
Prize for 1 Mbit memory (DM)	150000.-	50000.-	10000.-	800.-	240.-	60.-	10.-	1.-	0.25.-	0.11.-	0.05.-
	Growth about **– 40%** per year										
Chip size (mm^2)	24	16	25	45	54	91	140	190	250	400	?
Structure size (μm^2)	6	4	2	1.5	1.2	0.8	0.6	0.4	0.3	0.2	0.15
Number of process steps	70	80	8	120	280	400	450	500	600	?	?
Size of "killer" particles (>μm^2)	1.5	1.3	0.8	0.6	0.4	0.2	0.15	0.1	0.07	0.05	0.03
Total development costs (M$)	(90)	(140)	200	450	650	1000	2000	3500	5000	7000	?

5.4.2 Working in Chip Development and Production

- Most material scientists and engineers in the **Si** semiconductor industry will be involved in chip development and production.
  - They will be part of a *large* team that also includes colleagues from electrical engineering (design, testing), computer engineering (on-chip software, functionality, testing routines) physicists and chemists and, not to forget, "money" people.
- Three major tasks can be distinguished:
  - **1.** Development of the next chip generation up to the point where the factory takes over.
  - **2.** Improving yield and throughput in the factory for the respective technology (making money!)
  - **3.** Introducing new products based on the existing technology.
- However, this three fields started to grow together in the late eighties:
  - Development of new technologies takes place in a factory because pure research and developments "**lines** " - a cleanroom with the complete infrastructure to process (and characterize) chips - are far too expensive and must produce some sellable product at least "at the side". More important, without a "base load" produced at a constant output and high quality, it is never clear if everything works at the required level of perfection!
  - Improving the yield (and cutting down the costs) is easily the most demanding job in the field. It is hard work, requires lots of experience and intimate knowledge of the chip and its processes. The experts that developed the chip therefore often are involved in this task, too.
  - There are not only new products based on the new technology that just vary the design (e.g. different memory types), but constant additions to the technology as well. Most important the "shrink" designs (making the chips smaller) that rely on input from the ongoing development of the next generation and specific processes (e.g. another metallization layer) that need development on their own.
- A large degree of interaction therefore is absolutely necessary, demanding flexibility on the part of the engineers involved.
- Lets look briefly on the structure and evolution of a big chip project; The development of the **16 Mbit DRAM** at the end of the eighties. The project structure may look like this:

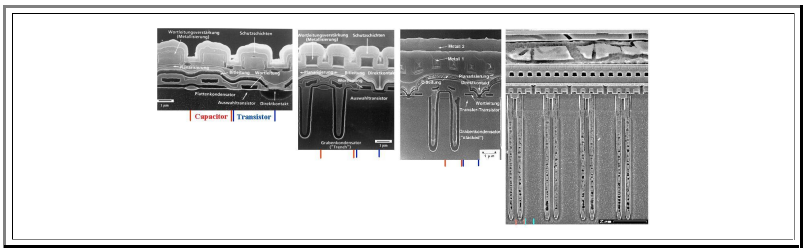


  - The number of experts working in a project like this may be **100 - 200**; they rely on an infrastructure (e.g. clean room personnel) that counts in the thousands (but these people only spend part of their time for the project).
- While there are many tasks that just need to be done on a very high level of sophistication, some tasks involve topics never done before: New technologies (e.g. trench- or stacked capacitor process modules, metallization with chemical-mechanical polishing (***CMP*** , one of the key processes of the nineties), new materials (e.g. silicides in the eighties or **Cu** in the nineties), new processes (always lithography, or, e.g., plasma etching in the eighties, or electrodeposition in the nineties).
  - The problem is that nobody knows if these new ingredients will work at all (in a mass production environment) and if they will run at acceptable costs. The only way of finding out is to try it - with very high risks involved.
  - It is here were you - a top graduate in materials science of a major university - will work after a brief training period of **1 - 2** years.
- One big moment in the life of the development team is the so-called "**First Silicon**".
  - This means the first chips ever to come out of the line. Will they work - at least a little bit? Or do we have to wait for the next batch, which will be many weeks behind and possibly suffer from the same problems that prevents success with the first one?

- Waiting for first **Si** can be just as nerve racking as waiting for the answer to your job applications, your research proposal or the result of presidential elections (this was written on Nov. **17th** in **2000**, when **10** days after the election nobody knows if Bush or Gore will be the next president of the **USA**).
- In the link, the results of first Silicon for the**16 Mbit DRAM** at Siemens are shown, together with how it went on from there.

5.4.3 Generation Sequences

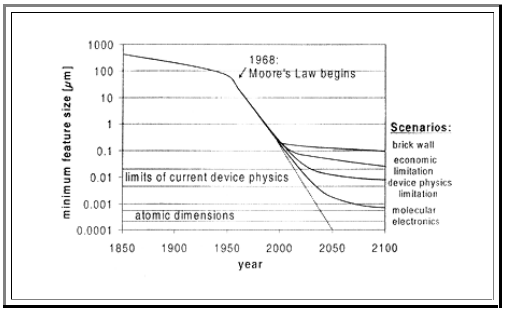
- It is quite instructive, if difficult to arrange, to look at several **generations** of **DRAMs** in direct comparison.
  - The picture below shows cross sections though the transistor - capacitor region necessary to store **1** bit - from the **1 Mbit DRAM** to the **64 Mbit DRAM** (all of **Siemens** design)
  - The pictures have been scaled to about the same magnification; the assembly is necessarily quite large. It starts with the **1 Mbit DRAM** on the left, followed by the **4 Mbit**, **16 Mbit** and **64 Mbit** memory.



- Decrease in **feature size** and some new key technologies are easily perceived. Most prominent are:
  - Planar capacitor ("Plattenkondensator") for the **1 Mbit DRAM**; **LOCOS** isolation and **1** level of metal ("Wortleitungsverstärkung") parallel to the **poly-Si** "Bitleitung" (bitline) and at right angles to the **poly-Si/Mo-silicide** "Wortleitung" (wordline).
  - Trench capacitor for the **4 Mbit DRAM**, "FOBIC" contact, and **TiN** diffusion barrier.
  - Two metal levels for the **16 Mbit DRAM**, *poly-ONO-poly* in trench; improved planarization between bitline - metal **1**, and metal **1** - metal **2**.
  - Box isolation instead of **LOCOS** for the **64 Mbit DRAM**, very deep trenches, **W**-plugs, and especially complete planarization with **chemical mechanical polishing** (***CMP***), the key process of supreme importance for the **64 Mbit** generation and beyond.
- If some of the technical expressions ***eluded*** you - don`t worry, be happy! We will get to them quickly enough.
- Parallel to a reduction in feature size is always an increase in **chip size**; this is illustrated the link.
  - You may ask yourself: Why do we not just make the chip bigger - instead of **200 4 Mbit DRAMs** on a wafer we just as well produce **50 16 Mbit Drams**?
  - Well. Le's say you have a very high yield of **75 %** in your **4 Mbit** production. This gives you **150** good chips out of your **200** - but it would give you a yield close to zero if you now make **16 Mbit DRAMs** with that technology.
  - What's more: Even if you solve the yield problem: Your **16 Mbit** chips would be exactly **4** times more expensive then your **4 Mbit** chip - after all your costs have not changed and you now produce only a quarter of what you had before. Your customer would have no reason to buy these chips, because they are not only not cheaper per bit, but also not faster or less energy consuming.
  - Progress performance only can come from reducing the feature size.
- The cost per bit problem you also can address to some extent by using larger wafers, making more chips per process run.
  - This has been done and is being done: Wafer sizes increased from **< 2** inch in the beginning of the seventies to **300 mm** now (**2002**) - we also went metric on the way.

5.4.4 Summary to: 5.4 Development and Production of a New Chip Generation

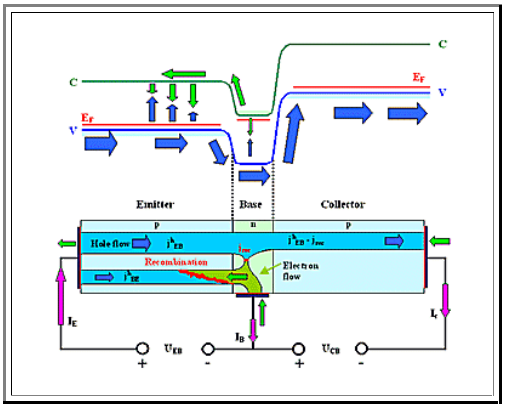
- Moore's law predicts exponentially growth of "chip complexity" with a high growth rate - how far will it reach?
  - Problems and costs are growing exponentially with every new generation.
  - It follows: The market must grow exponentially too, if you want to make a profit.
  - It follows: Large amounts of money can be easily made - or lost.
- Falling behind thecompetition in your technology and yields means certain death for companies without a monopoly in some product.



5.5.1 Summary: General Aspects of Silicon Technology

- Essentials of the bipolar transistor:
  - High emitter doping (N_{Don} for **npn** transistor here) in comparison to base doping N_{Ac} for large current amplification factor $\gamma = I_C/I_B$.
  - $N_{Don}/N_{Ac} \approx \kappa$ = injection ratio.

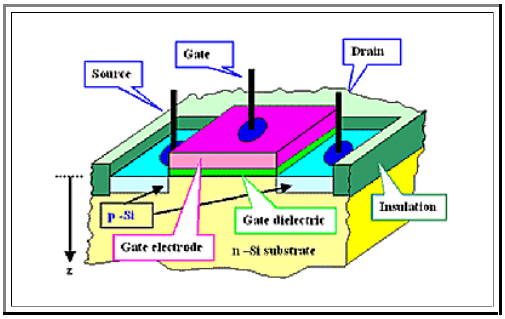
$$\gamma \approx \frac{N_{Don}}{N_{Ac}} \cdot \left(1 - \frac{d_{base}}{L} \right)$$

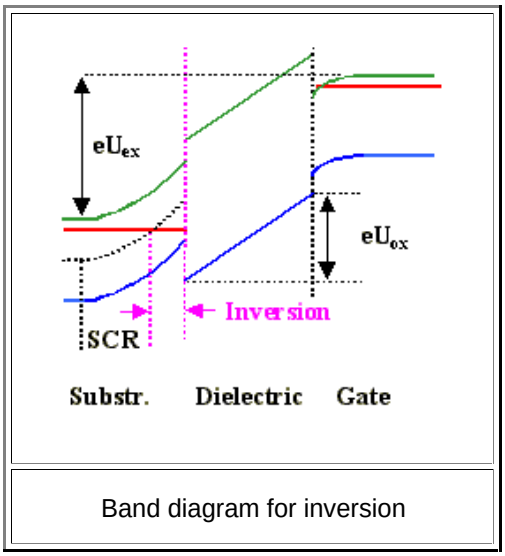


  - Small base width d_{base} (relative to diffusion length L) for large current amplification.
  - Not as easy to make as the band-diagram suggests!

- Essentials of the **MOS** transistor:
  - Gate voltage enables Source-Drain current
  - Essential process. Inversion of majority carrier type in channel below gate by:
    - Drive intrinsic majority carriers into bulk by gate voltage with same sign as majority carriers.
    - Reduced majority concentration n_{maj} below gate increases minority carrier concentration n_{min} via mass action law

$$n_{maj} \cdot n_{min} = n_i^2$$





Band diagram for inversion

    - An inversion channel with $n_{min} > n_{maj}$ develops below the gate as soon as threshold voltage U_{Th} is reached.
    - Current now can flow because the reversely biased **pn**-junction between either source or drain and the region below the gate has disappeared.

- The decisive material is the gate dielectric (usually SiO_2). Basic requirement is:
  - High capacity C_G of the gate electrode - gate dielectric - **Si** capacitor = high charge Q_G on electrodes = strong band bending = low threshold voltages U_G
  - It follows:
    - Gate dielectric thickness d_{Di} ⇒ High breakdown field strength U_{Bd}
    - Large dielectric constant ϵ_r
    - No interface states.
    - Good adhesion, easy to make / deposit, easy to structure, small leakage currents, ...

$$Q_G = C_G \cdot U_G$$

Example:
$U = 5$ V, $d_{Di} = 5$ nm ⇒ $E = U/d_{Di} = 10^7$ V/cm !!

$\epsilon_r(SiO_2) = 3.9$

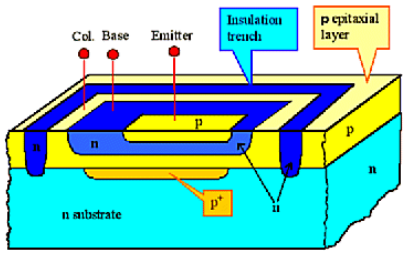
Integration means:

- **1.** Produce a large number (up to **1.000.000.000**) of transistors (bipolar or **MOS**) and other electronic elements on a **cm^2** of **Si**
- **2.** Keep thoses elements electrically insulated from each other.
- **3.** Connect those elements in a meaningful way to produce a system / product.

It ain't easy!

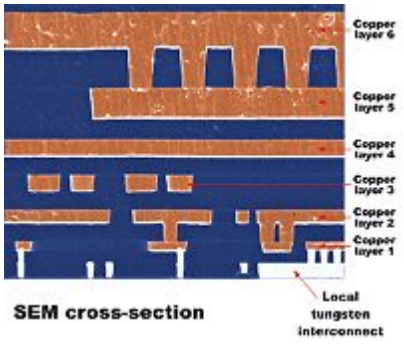
An integrated bipolar transistor does not resemble the textbook picture at all, but looks far more complicated ⇒.

- This is due to the insulation requirements, the process requirements, and the need to interconnect as efficiently as possible.
- The epitaxial layer cuts down on the number of critical diffusions, makes insulation easier, and allows a "buried contact" structure.



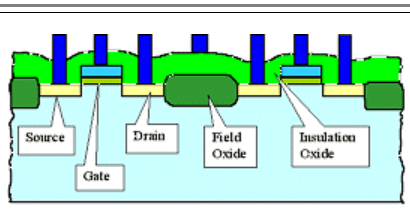
Connecting transistor / elements is complicated; it has to be done on several levels

- Materials used are **Al** ("old"), **Cu** ("new"), **W**, (highly doped) poly-**Si** as well as various silicides.
- Essential properties are the conductivity σ of the conductor, the dielectric constant ϵ_r of the intermetal dielectric, and the resulting time constant $\tau = \sigma \cdot \epsilon_r$ that defines the maximum signal transmision frequency through the conducting line.



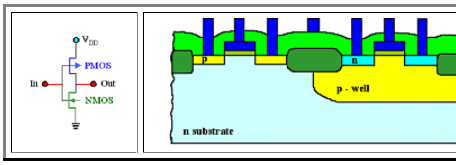
Integrating **MOS** transistors requires special measures for insulation (e.g. a field oxide) and for gate oxide production

- Since a **MOS** transistor contains intrinsically a capacitor (the gate "stack"), the technology can be used to produce capacitors, too.



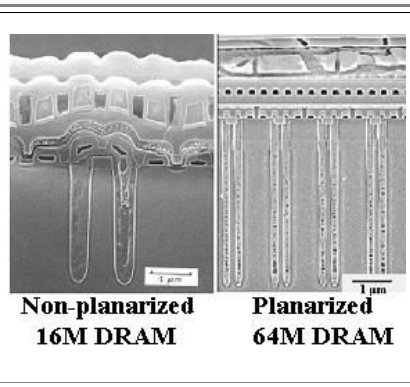
CMOS allows to reduce power consumption dramatically.

- The process, however, is more complex: Wells with different doping type need to be made.



Using the third dimension (depth / height) might become necessary for integrating "large" structures into a small projected are (example: trench capacitor in **DRAMs** ⇒).

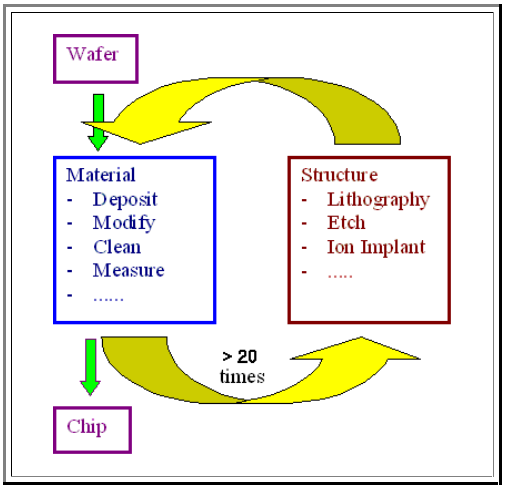
- Unwanted "topology", however, makes integration more difficult.
- Planarized technologies are a must since about 1995! ⇒



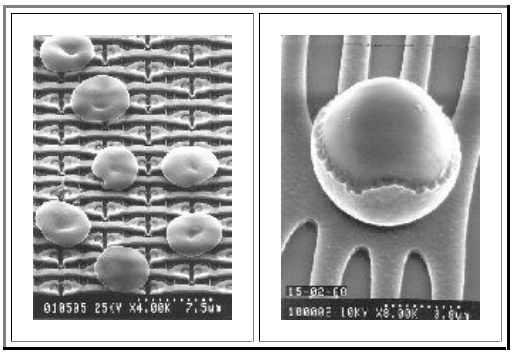
- It ain't neither easy nor cheap!

Property	Number
Feature size	**0,2 µm**
No. metallization levels	**4 - 7**
No. components	**> 6 · 10^8 (Memory)**
Complexity	**> 500 Process steps**
Cost (development and **1** factory)	**ca. $ 6 · 10^9**

- Typical wafer size for new factories (**2007**) : **300 mm** diameter, **775 µm** thickness, flatness in lower **µm** region
  - Chip size a few **cm^2**, much smaller if possible
  - Yield ***Y*** = most important parameter in chip production = % of chips on a wafer that function (= can be sold).
  - ***Y* = 29 %** is a good value for starting production



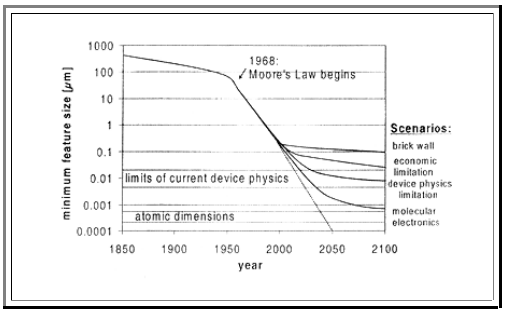
- Chip making = running about **20** times (roughly!!) through "materials" - "structuring" loop.
  - About **400 - 600** individual processing steps (= in / out of special "machine") before chip is finished on wafer
  - More than **30** processing steps for packaging (after separation of chips by cutting)
  - Simple estimate: **99.9%** perfection for each processing step means***Y* < 70 %.**



- Dirt in any form - as "particles" on the surface of wafer, or as "contamination" inside the wafer is almost always deadly
  - Particles with sized not much smaller than minimum feature sizes (i.e. < **10 nm** in **2007**) will invariably cover structures and lead to local dysfunction of a transistor or whatever.
  - Point defects like metal atoms in the **Si** lattice may precipitate and cause local short circuits etc. from the "inside", killing transistors
  - One dysfunctional transistor out of **1.000.000.000** or so is enough to kill a chip!

- Being extremely clean is absolutely mandatory for high Yields ***Y***!
  - Use cleanrooms and hyper-clean materials!
  - It won't be cheap!

- Moore's law predicts exponentially growth of "chip complexity" with a high growth rate - how far will it reach?
  - Problems and costs are growing exponentially with every new generation.
  - It follows: The market must grow exponentially too, if you want to make a profit.
  - It follows: Large amounts of money can be easily made - or lost.
- Falling behind thecompetition in your technology and yields means certain death for companies without a monopoly in some product.



Questionaire

Multiple Choice questions to all of 5

6. Materials and Processes for Silicon Technology

6.1 Silicon

6.1.1 Producing Semiconductor-Grade Silicon

6.1.2 Silicon Crystal Growth and Wafer Production

6.1.3 Uses of Silicon Outside of Microelectronics

6.1.4 Summary to: 6.1 Materials and Processes for Silicon Technology

6.2 Si Oxides and LOCOS Process

6.2.1 Si Oxide

6.2.2 LOCOS Process

6.2.3 Summary to: 6.2 Si Oxide and LOCOS Process

6.3 Chemical Vapor Deposition

6.3.1 Silicon Epitaxy

6.3.2 Oxide CVD

6.3.3 CVD for Poly-Silicon, Silicon Nitride and Miscellaneous Materials

6.3.4 Summary to: 6.3 Chemical Vapor Deposition

6.4. Physical Processes for Layer Deposition

6.4.1 Sputter Deposition and Contact Hole Filling

6.4.2 Ion Implantation

6.4.3 Miscellaneous Techniques and Comparison

6.3.4 Summary to: 6.4 Physical Processes for Layer Deposition

6.5 Etching Techniques

6.5.1 General Remarks

6.5.2 Chemical Etching

6.5.3 Plasma Etching

6.5.3 Summary to: Etching Techniques

6.6 Lithography

6.6.1 Basic Lithography Techniques

6.6.2 Resist and Steppers

6. Materials and Processes for Silicon Technology

6.1 Silicon

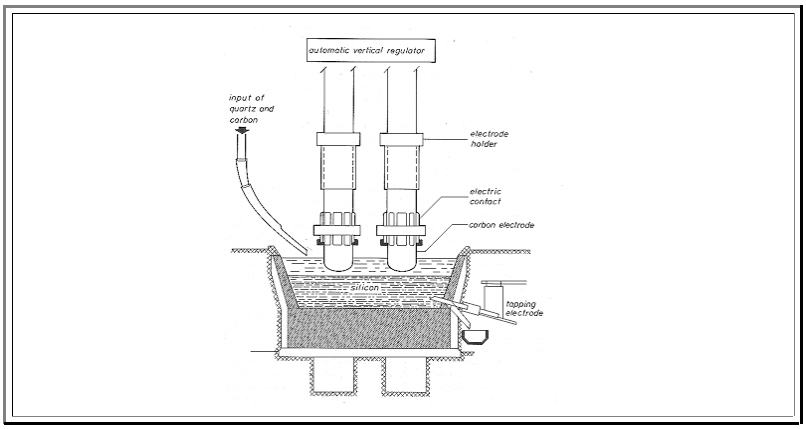
6.1.1 Producing Semiconductor-Grade Silicon

Introductory Remarks

- It is written somewhere that in the beginning God created heaven and the earth. It is not written from what.
  - We do not know for sure what the heaven is made of but we do know what the the earth is made of, at least as far as the upper crust is concerned. Interestingly enough, he (or she) created mostly **Silicon** and Oxygen with some dirt (in the form of the other **90** elements) thrown in for added value.
  - Indeed, the outer crust of this planet (lets say the first **100 km** or so) consists of all kinds of silicates - **Si** + **O** + something else - so there is no lack of **Si** as a raw material. **Si**, in fact, accounts for about **26 %** of the crust, while **O** weighs in at about **49 %**.
- However, it took a while to discover the element **Si**. **Berzellius** came up with some form of it in **1824** (probably amorphous), but it was **Deville** in **1854** who first obtained regular crystalline **Si**.
  - This is simply due to the very high chemical reactivity of **Si**. Pure **Si** (not protected by a thin layer of very stable SiO_2 as all **Si** crystals and wafers are) will react with *anything*, and that creates one of the problems in making it and keeping it clean.
  - *Liquid* **Si** indeed does react with all substances known to man - it is an universal solvent. This makes crystal growth from liquid **Si** somewhat tricky, because how do you contain your liquid **Si**? Fortunately, some materials - especially SiO_2 - dissolve only very slowly, so if you don't take too long in growing a crystal, they will do as a vessel for the liquid **Si**.
  - But there will always be some dissolved SiO_2 and therefore oxygen in your liquid **Si**, and that makes it hard to produce **Si** crystals with very low oxygen concentrations.
- What we *need*, of course, are **Si crystals** - in the form of **wafers** - with extreme degrees of perfection.
  - What we *have* are ***inexhaustible*** resources of **Silicondioxide**, SiO_2 , fairly clean, if obtained from the right source. Since there is no other material with properties so precisely matched to the needs of the semiconductor industry, and therefore of the utmost importance for our modern society, the production process of **Si** wafers shall be covered in a cursory way.

Producing "Raw" Silicon

- Fortunately, the steel industry needs **Si**, too. And **Si** was already used as a crucial alloying component of steel before it started its career as the paradigmatic material of our times.
  - Most of the world production of **raw Si** still goes to the steel industry and only a small part is ***diverted*** for the semiconductor trade. This is why this stuff is commonly called "**metallurgical grade**" **Si** or ***MG-Si*** for short. The world production in **2006** was around **4** Mio tons per year.
  - How is **MG- Si** (meaning poly crystalline material with a purity of about **99%**) made? More or less like most of the other metals: Reduce the oxide of the material in a furnace by providing some reducing agent and sufficient energy to achieve the necessary high temperatures..
- Like for most metals, the reducing agent is *carbon* (in the form of coal or coke (**=** very clean coal)). The necessary energy is supplied electrically.
  - Essentially, you have a huge furnace (lined with **C** which will turn into very hard and inert **SiC** anyway) with three big graphite electrodes inside (carrying a few **10.000 A** of current) that is continuously filled with SiO_2 (**=** quartz sand) and carbon (**=** coal) in the right weight relation plus a few added secret ingredients to avoid producing **SiC**. This looks like this



- The chemical reaction that you want to take place at about **2000 ^{o}C** is

$$SiO_2 + 2C \Rightarrow Si + 2CO$$

- But there are plenty of other reactions that may occur simultaneously, e.g. **Si + C** ⇒ **SiC**. This will not only reduce your yield of **Si**, but ***clog up*** your furnace because **SiC** is not liquid at the reaction temperature and extremely hard - your reactor ends up as a piece of junk if you make **SiC**.

Still, we do not have to worry about **MG-Si** - a little bit of what is made for the steel industry will suffice for all of **Si** electronics applications.

- What we do have to do is to *purify* the **MG-Si** - about **10^9** fold!

This is essentially done in three steps:

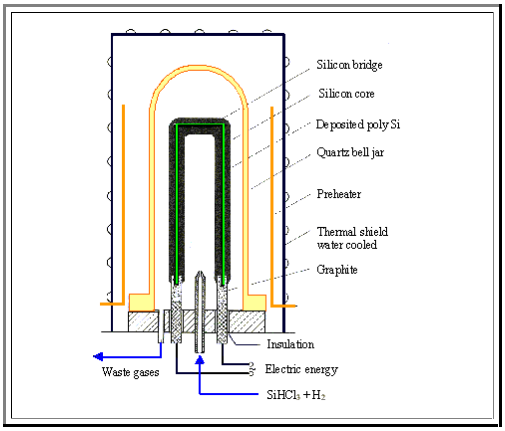
- *First*, **Si** is converted to **$SiHCl_3$** in a "fluid bed" reactor via the reaction

$$Si + 3HCl \Rightarrow SiHCl_3 + H_2$$

- This reaction (helped by a catalyst) takes place at around **300 ^{o}C**. The resulting **Trichlorosilane** is already much purer than the raw **Si**; it is a liquid with a boiling point of **31.8 ^{o}C**.
- *Second*, the **$SiHCl_3$** is distilled (like wodka), resulting in extremely pure Trichlorosilane.
- *Third*, high-purity **Si** is produced by the **Siemens process** or, to use its modern name, by a "**Chemical Vapor Deposition**" (***CVD***) process - a process which we will encounter more often in the following chapters.

Producing Doped Poly-Silicon

The doped **poly-Si** (not to be confused with the poly-**Si** layers on chips) used for the growth of single **Si** crystals is made in a principally simple way which we will discuss by looking at a poly-**Si CVD** reactor



- In principle, we have a vessel which can be evacuated and that contains an "U" shaped arrangements of slim **Si** rods which can be heated from an outside heating source and, as soon as the temperature is high enough (roughly **1000 ^{o}C**) to provide sufficient conductivity, by passing an electrical current through it.
- After the vessel has been evacuated and the **Si** rods are at the reaction temperature, an optimized mix of ***$SiHCl_3$*** (**Trichlorosilane**), **H_2** and doping gases like **AsH_3** or **PH_3** are admitted into the reactor. In order to keep the pressure constant (at a typical value of some mbar), the reaction products (and unreacted gases) are pumped out at a suitable place.

- On *hot* surfaces - if everything is right this will only be the **Si** - a chemical reaction takes place, reducing the **$SiHCl_3$** to **Si** and forming **HCl** (hydrochloric acid) as a new compound:

$$SiHCl_3 + H_2 \Rightarrow Si + 3\,HCl$$

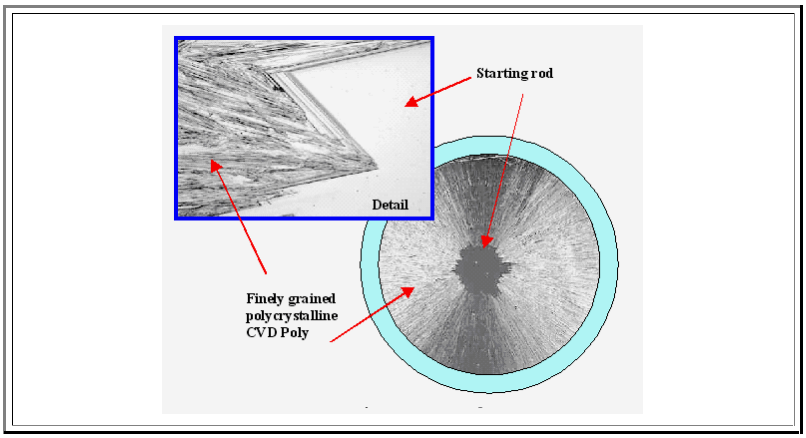
- Similar reactions provide very small but precisely measured amounts of **As**, **P** or **B** that will be incorporated into the growing polysilicon
- The **Si** formed will adhere to the **Si** already present - the thin rods will grow as fresh **Si** is produced. The incorporation of the dopants will produce *doped polysilicon*.

In principle this is a simple process, like all **CVD** processes - but not in reality. Consider the complications:

- You have to keep the **Si** ultrapure - all materials (including the gases) must be specially selected.
- The chemistry is extremely dangerous: **AsH_3** and **PH_3** are among the most poisonous substances known to mankind; **PH_3** was actually used as a toxic gas in world war **II** with disastrous effects. **H_2** and **$SiHCl_3$** are easily combustible if not outright explosive, and **HCl** (in gaseous form) is even more dangerous than the liquid acid and extremely corrosive. Handling these chemicals, including the safe and environmentally sound disposal, is neither easy nor cheap.
- Precise control is not easy either. While the flux of **H_2** may be in the **100 liter/min** range, the dopant gases only require **ml/min**. All flow values must be precisely controlled and, moreover, the mix must be homogeneous at the **Si** where the reaction takes place.
- The process is slow (about **1 kg/hr)** and therefore *expensive*. You want to make sure that your hyperpure (and therefore expensive) gases are completely consumed in the reaction and not wasted in the exhaust - but you also want high throughput and good homogeneity; essentially conflicting requirements. There is a large amount of optimization required!
- And from somewhere you need the slim rods - already with the right doping.

Still, it works and abut **10.000 tons** of poly-**Si** are produced at present (**2000**) with this technology, which was pioneered by **Siemens AG** in the sixties for the microelectronic industry. (in **2007** it is more like **21.000 to** plus another **30.000** tons for the solar industry).

- **Electronic grade Si** is not cheap, however, and has no obvious potential to become very cheap either. The link provides todays specifications and some more information for the product. Here is an example for the poly-crystalline rods produced in the Siemens process:



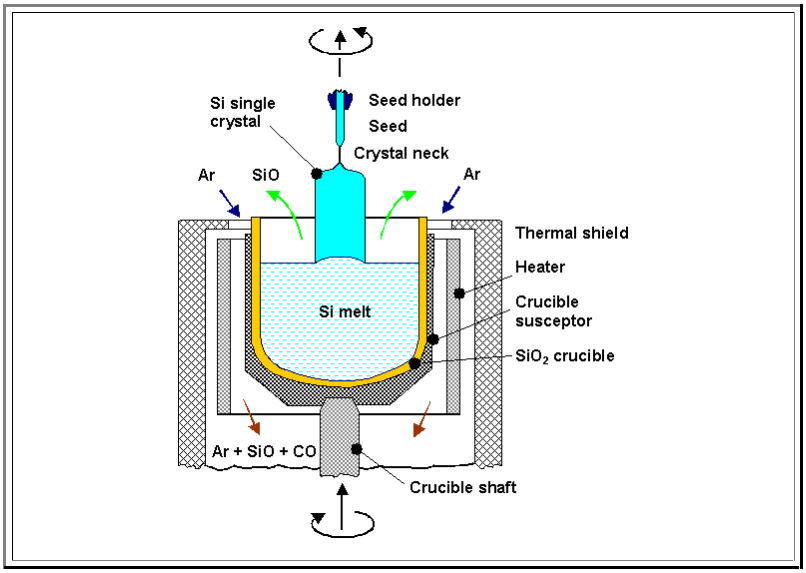
- While this is not extremely important for the microelectronics industry (where the added value of the chip by far surpasses the costs of the **Si)**, it prevents other **Si** products, especially *cheap solar cells* (in connection with all the other expensive processes before and after the poly-**Si** process). Starting with the first oil crisis in **1976**, many projects in the USA and Europe tried to come up with a cheaper source of high purity poly-**Si**, so far without much success.
- By now. i.e. in **2007**, demand for electronic grade **Si** si surging because of a booming solar cell industry. A short ovberveiw of the current Si crisis can be found in the link.

This Module is available in Romanian

6.1.2 Silicon Crystal Growth and Wafer Production

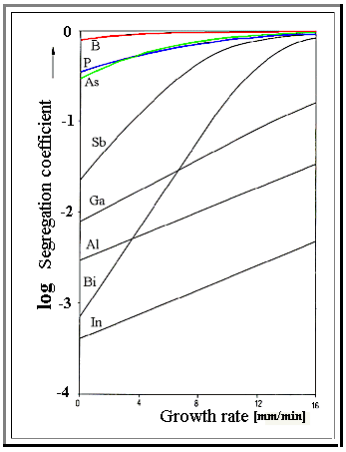
Single Crystal Growth

- We now have hyperpure poly-**Si**, already doped to the desired level, and the next step must be to convert it to a *single crystal*. There are essentially two methods for crystal growth used in this case:
  - **Czochralski** or ***crucible* grown** crystals (***CZ*** crystals).
  - **Float zone** or ***FZ*** crystals.
- The latter method produces crystals with the highest purity, but is not easily used at large diameters. **150 mm** crystals are already quite difficult to make and nobody so far has made a **300 mm** crystal this way. Float zone crystal growth, while the main method at the beginning of the **Si** age, is now only used for some specialities and therefore will not be discussed here; some details can be found in the link.
  - The Czochralski method, invented by the Polish scientist **J. Czochralski** in **1916**, is the method of choice for high volume production of **Si** single crystals of exceptional quality and shall be discussed briefly. Below is a schematic drawing of a crystal growth apparatus employing the Czochralski method. More details can be found in the link.



  - Essentially, a crystal is "pulled" out of a vessel containing liquid **Si** by dipping a **seed crystal** into the liquid which is subsequently slowly withdrawn at a surface temperature of the melt just above the melting point.
  - The *pulling rate* (usually a few mm/min) and the *temperature profile* determines the crystal diameter (the problem is to get rid of the heat of crystallization).
  - Everything else determines the quality and homogeneity - crystal growing is still as much an art as a science! Some interesting points are contained in the link.
- Here we only look at one major point, the **segregation coefficient** k_{seg} of impurity atoms.
  - The segregation coefficient in thermodynamic equilibrium gives the relation between the concentration of impurity atoms in the growing crystal and that of the melt. It is usually much lower than **1** because impurity atoms "prefer" to stay in the melt. This can be seen from the liquidus and solidus lines in the respective phase diagrams.
  - In other words, the *solubility* of impurity atoms in the melt is larger than in the solid.
  - "Equilibrium" refers to a growth speed of **0 mm/min** or, more practically, very low growth rates. For finite growth rates, k_{seg} becomes a function of the growth rate (called k_{seff}) and approximates **1** for high growth rates (whatever comes to the rapidly moving interface gets incorporated).

- This has a positive and a negative side to it:
  - On the positive side, the crystal will be *cleaner* than the liquid, crystal growing is simultaneously a purification method. Always provided that we discard the last part of the crystal where all the impurities are now concentrated. After all, what was in the melt must be in the solid after solidification - only the distribution may now be different.
  - This defines the negative side: The *distribution of impurities* - and that includes the doping elements and oxygen - *will change along the length of a crystal* - a homogeneous doping etc. is difficult to achieve.
  - That segregation can be a large effect with a sensitive dependence on the growth rate is shown below for the possible doping elements; the segregation coefficients of the unwanted impurities is given in a table.



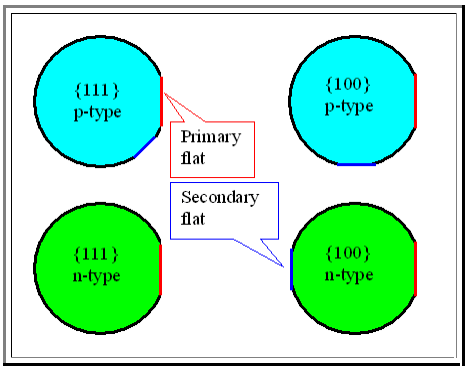
Atom	Cu	Ag	Au	C	Ge	Sn	
k_{seg}	$4 \cdot 10^{-4}$	$1 \cdot 10^{-6}$	$2{,}5 \cdot 10^{-5}$	$6 \cdot 10^{-2}$	$3{,}3 \cdot 10^{-1}$	$1{,}6 \cdot 10^{-2}$	
Atom	O	S	Mn	Fe	Co	Ni	Ta
k_{seg}	1,25	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$8 \cdot 10^{-6}$	$8 \cdot 10^{-6}$	$4 \cdot 10^{-4}$	$1 \cdot 10^{-7}$

  - We recognize *one* reason why practically only **As**, **P**, and **B** is used for doping! Their segregation coefficient is close to **1** which assures half-way homogeneous distribution during crystal growth. Achieving homogeneous doping with **Bi**, on the other hand, would be exceedingly difficult or just impossible.
- Present day single crystals of silicon are the most perfect objects on this side of Pluto - remember that perfection can be measured by using the second law of thermodynamics; this is not an empty statement! A very interesting and readable article dealing with the history and the development of **Si** crystal growth from W. **Zulehner** (Wacker Siltronic), who was working on this subject from the very beginning of commercial **Si** crystal growth until today, can be found in the link.
  - What the finished crystal looks like can be seen in the link. What we cannot see is that there is no other crystal of a different material that even comes close in size and perfection.
  - Our crystal does not contain dislocations - a unique feature that only could be matched by Germanium crystals at appreciable sizes (which nobody grows or needs)[1]. It also does not contain many other lattice defects. With the exception of the doping atoms (and possible interstitial oxygen, which often is wanted in a concentration of about **30 ppm**), substitutional and interstitial impurities are well below a **ppb** if not **ppt** level (except for relatively harmless carbon at about **1 ppm**) - unmatched by most other "high purity" materials.
  - Our crystal is homogeneous. The concentration of the doping atoms (and possibly interstitial oxygen) is radially and laterally rather constant, a feat not easily achieved.
- The crystal is now ready for cutting into wafers.

Wafer Technology

- It may appear rather trivial now to cut the crystal into slices which, after some polishing, result in the **wafers** used as the starting material for chip production.
  - However, it is *not* trivial. While a wafer does not look like much, its not easy to manufacture. Again, making wafers is a closely guarded secret and it is possibly even more difficult to see a wafer production than a single **Si** crystal production.
  - First, wafers must all be made to exceedingly tight geometric specifications. Not only must the diameter and the thickness be precisely what they ought to be, but the flatness is constrained to about **1 µm**.
  - This means that the polished surface deviates at most about **1 µm** from an ideally flat reference plane - for surface areas of more than **1000 cm^2** for a **300 mm** wafer!
  - And this is not just true for one wafer, but for all **10.000** or so produced daily in *one* factory. The number of Si wafers sold in **2001** is about **100.000.000** or roughly **300.000** a day! Only tightly controlled processes with plenty of know-how and expensive equipment will assure these specifications. The following picture gives an impression of the first step of a many-step polishing procedure.

- In contrast to e.g. polished metals, polished **Si** wafers have a *perfect* surface - the crystal just ends followed by less than two **nm** of "**native oxide**" which forms rather quickly in air and protects the wafer from chemical attacks.
  - Polishing **Si** is not easy (but fairly well understood) and so is keeping the surface clean of particles. The final polishing and cleaning steps are done in a *cleanroom* where the wafers are packed for shipping.
  - Since chip structures are always aligned along crystallographic directions, it is important to indicate the crystallography of a wafer. This is done by grinding **flats** (or, for very large wafer - **200 mm** and beyond **notches**) at precisely defined positions.
  - The flats also encode the doping types - mix ups are very expensive! The convention for flats is as follows:



  - The main flat is always along a **<110>** direction. However, many companies have special agreements with wafer producers and have "customized" flats (most commonly no secondary flat on **{100} p**-type material).
  - More about flats in the link

- Typical wafer specifications may contain more than **30** topics, the most important ones are:
  - *Doping type*: **n** or **p**-type (**p**-type is by far the most common type) and *dopant* used (**P**, **As** or **B**). *Resistivity* (commonly between **100** Ω **cm** to **0,001** Ω**cm** with **(5 - 1)** Ω**cm** defining the bulk of the business. All numbers with error margins and homogeneity requirements
  - *Impurity concentrations* for metals and other "life time killers" (typically below $\mathbf{10^{12}\ cm^{-3}}$), together with the *life time* or diffusion length (which should be several **100 µm**).
  - *Oxygen* and *carbon* concentration (typically around $\mathbf{6 \cdot 10^{17}\ cm^{-3}}$ or $\mathbf{1 \cdot 10^{16}\ cm^{-3}}$, respectively. While the carbon concentration just has to be low, the oxygen concentration often is specified within narrow limits because the customer may use "**internal gettering**", a process where oxygen precipitates are formed intensionally in the bulk of the wafer with beneficial effects on the chips in the surface near regions.
  - *Microdefect densities* (after all, the point defects generated in thermal equilibrium during crystal growth must still be there in the form of small agglomerates). The specification here may simple be: ***BMD*** ("bulk micro defect") density **=** $\mathbf{0\ cm^{-3}}$. Which simply translates into: Below the detection limit of the best analytical tools.
  - *Geometry*, especially several parameters relating to flatness. Typical tolerances are always in the **1 µm** regime.
  - Surface *cleanliness*: No particles and no atomic or molecular impurities on the surface!
- This link provides a grapical overview of the complete production process - from sand to **Si** wafers - and includes a few steps not covered here.
  - Appreciate that the production of wafers - at last several thousands per day - with specifications that are always at the cutting edge of what is possible - is an extremely involved and difficult process.
  - At present (Jan. **2004**), there are only a handfull of companies world wide that can do it. In fact, **4** companies control about **80%** of the market.

- This link leads to a recent (**1999**) article covering new developments in **Si CZ** crystal growth and wafer technology (from A.P. **Mozer**; Wacker Siltronic) and gives an impression of the richness of complex issues behind the production of the humble **Si** wafer.
- This link shows commercial wafer specifications.
  - To give an idea of the size iof the industry: In **2004** a grand totoal of about **4.000.000 m^2** of polished **Si** wafers was produced, equivalent to about **1.25 · 10^8 200 mm** wafers.

Questionaire

Multiple Choice questions to 6.1.2

---

[1] No longer true in **2004**! Germanium wafers may (or may not) make a come-back; but they are certainly produced again.

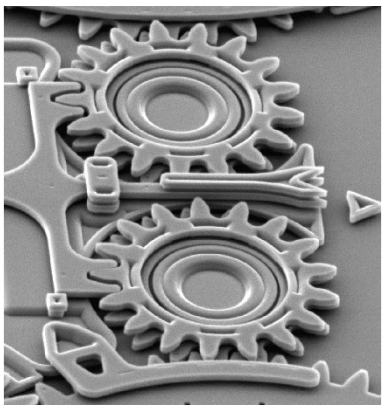
6.1.3 Uses of Silicon Outside of Microelectronics

Solar Cells

- Besides integrated circuits, electronic grade **Si** is used in rather large quantities for the production of **solar cells**. While there are solar cells made from other semiconductors, too, the overwhelming majority of solar cells really producing power out there is made from (thick) **Si**. We may distinguish three basically different types.
  1. Solar cells made from "thick" (**< 300 µm**) slices of *single-crystalline* **Si**. The substrates are essentially made in the same way as wafers for microelectronics, except that quality standards are somewhat relaxed and they are therefore cheaper.
  2. Solar cells made from "thick" (**< 300 µm**) slices of *poly-crystalline* **Si** with preferably large grains. This material is therfore usually referred to as "**multicrystalline Si**".
  3. Solar cells made from thin (some **µm**) layers of fine-grained poly-crystalline **Si** deposited on a (cheap) glas substrate. This type of solar cell is at present (**2005**) in the research and development stage.
- We will not discuss solar cells here in any detail, but refer the matter to the Hyperscript "*Semiconductors*" where some background on Si solar cells is provided.

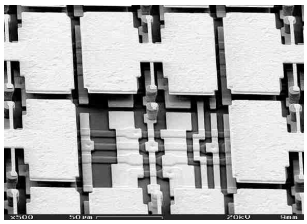
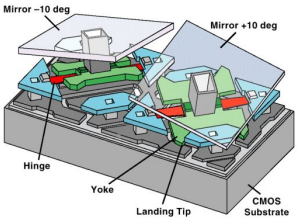
MEMS - Micro Electronic and Mechanical Systems

- Micromechanical devices made from **Si** are rapidly gaining in importance. Their production process utilizes most everything used in microelectronics, plus a few special processes.
- Again, we will not discuss **MEMS** in this Hyperscript, but show only a few pictures of what can be made.
  - Let's look at *mechanical* **MEMS** first. On top, a microscopic gear wheel systems from *Sandia Labs*. It could be used for mechanically "locking" your computer; which might be more secure than just software protection.
  - On the bottom, more or less the same thing with a dust mite on it. This is the little animal that lives in your rug, bed and upholstery and gives a fair share of us the infamous dust ("Hausstaub") allergy



Pictures: Courtesy of Sandia National Laboratories, SUMMiTTM Technologies, www.mems.sandia.gov"

  - While gear wheels look very good, the real use of **MEMS** so far is in sensors, in particular for acceleration. The sensor exploding your air bag when you wrap your car around a tree is the paradigmatic **MEMS** product.
- If we look at *optical* **MEMS**, we are mostly also looking on a mechanical microstructure, in this case at arrays of little mirrors which can be addresses individually and thus "process" a light beam pixel by pixel.

	
Courtesy of "ISiT" (Fraunhofer Institut Silizium Technologioe; Itzehoe, Germany)	Courtesy of Texas Instruments

- On the left we have an array of microscopic mirrors that can be moved up and down electrically (from the **ISiT** in Itzehoe). The central mirror is removed to show the underlying structure
- On the right a schematic drawing of the "mechanical" part of Texas Instruments ("***DLP***" = digital light processing) chip, the heart of many beamers.
- But many other things are possible with **MEMS**, suffice it to mention "bio-chips", micro-fluidics, sensors and actuators for many uses, microlenses and lens arrays, and tunable capacitors and resonators, and, not to forget, very down-to-earth products like the micro-nozzles for ink jet printers.

Miscellaneous

- There are many more applications, most in the development phase, that exploit the exceptional quality of large **Si** crystals, the unsurpassed technology base for processing, or simple emerging new features that might be useful. Here are few examples:
- While there are no conventional lenses for **X**-rays or neutron beams, some optics is still possible by either using reflection (i.e. imaging with mirrors) or diffraction.
  - An good **X**-ray mirror, like any mirror, must have a roughness far smaller than the wavelength. For useful applications (like "**EUV**" = Extreme Ultraviolet) lithography (it is really X-ray lithography, but this term has been "burned" in the **80**ties and is now a dirty word in microelectronics), this quickly transfers into the condition that the mirrors must be more or less atomically flat over large areas. This can be only done with large perfect single crystals, so your choice of materials is no choice at all: You use **Si**.
  - If you want to "process" a neutron beam, e.g. to make it monochromatic, you use Bragg diffraction at a "good" crystal. Again, mostly only large and perfect single crystals of **Si** meet the requirements
- Si is fully transparent for **IR** light and is thus a great material for making **IR** optics. In this field, however, there is plenty of competition from other materials. But **Si** is the material of choice for mirrors and prisms needed for **IR** spectroscopy.
- Since about **1990** "porous **Si**" is emerging as a totally new kind of material. It is electrochemically made form single-crystalline **Si** and comes in many variants with many, partially astonishing properties (optically activity, highly explosive, ...)
  - A review about this stuff can be found in the link . Here we simply note that a number of projects explores possible uses as for example electrodes for fuel cell, very special optical and **X**-ray filters, biochips, fuses for airbags, "normal" and biosensors, or special actuators.

6.1.4 Summary to: 6.1 Materials and Processes for Silicon Technology

- Making "metallurgical" (= "dirty") **Si_{met}** is easy: ⇒
  - A large scale **Si_{met}** production (**> 1 Mio tons/a**) exists for metallurgical ("alloying") and chemical ("silicones") uses
- A small amount of **Si_{met}** (some **20.000 to/a**) is purified (factor **10^9** or so) to "semiconductor grade **Si**" ⇒
  - Produce high-purity trichlorosilane (**$SiHCl_3$**) gas in a reactor and distill.
  - Use **$SiHCl_3$** and **H_2** to deposit **Si** on some **Si** core by a **CVD** process
- The final result is ultra-high purity (and expensive) **poly Si** (already doped if so desired)

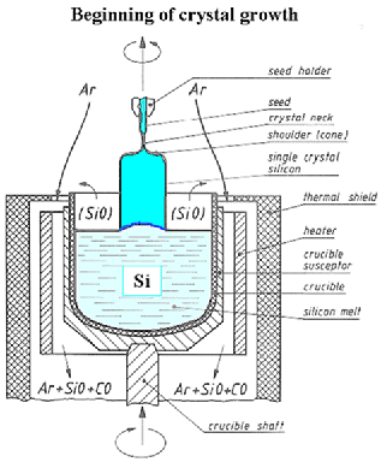
$$SiO_2 + 2C \Rightarrow Si_{met} + 2CO$$

$$Si + 3HCl \Rightarrow SiHCl_3 + H_2$$

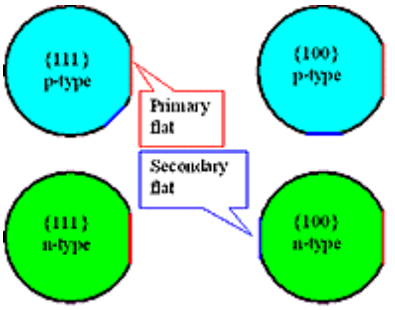
$$SiHCl_3 + H_2 \Rightarrow Si + 3\,HCl$$

- Growing a "perfect" single crystal from this **poly-Si** is not easy - but possible.
  - The major crystal growth method is the **CZ** (= Czrochalski) method: "Pull" the crystal from a crucible full of molten **Si**. ⇒
  - Some (ususally **< 300 mm** diameter) crystals are grown by the **FZ** (= float zone) method. Somewhat better perfection, but more expensive than **CZ**.
- Major problem: Impurity segregation = general tendency for most impurities (including doping atoms) to remain (= enrich) in the melt.
  - Segregation coeffcient = c_{cryst}/c_{melt} at interface, often **<< 1** and dependent on parameters like growth speed (usually a few **mm/min**).
  - **+** Crystal is purer than melt.

    – It is practically impossible to grow a crystal with a uniform impuritiy (including dopant!) concentration along its length.



- Produce wafers by cutting, grinding and polishing
  - Extreme precision for a mass product is needed.
  - "Flats" or "notches" (for wafers **> 200 mm**) identify the crystallographic orientation and the doping type.
  - Beware! Flats are often custome specific and different from the norm. ⇒



Questionaire

Multiple Choice questions to all of 6.1

6.2 Si Oxides and LOCOS Process

6.2.1 Si Oxide

The Importance of Silicon Dioxide

- Silicon would not be the "miracle material" without its oxide, ***SiO_2***, also known as ordinary **quartz** in its bulk form, or as **rock crystal** if you find it in single crystal form (and relatively pure) somewhere out there in the mountains.
  - Not only the properties of **Si** - especially the achievable crystalline perfection in combination with the bandgap and easy doping - but also the properties of **SiO_2** are pretty much as one would have wished them to be for making integrated circuits and other devices.
  - From the beginning of integrated circuit technology - say **1970** - to the end of the millennium, **SiO_2** was a key material that was used for many different purposes. Only now (around **2004**), industry has started to replace it for some applications with other, highly "specialized" dielectrics.
- What is so special about **SiO_2**?
  - First of all, it comes in many forms: There are several *allotropes* (meaning different crystal types) of **SiO_2** crystals; the most common (stable at room temperature and ambient pressure) is "low quartz" or α-quartz, the quartz crystals found everywhere. **But SiO_2** is also rather stable and easy to make in amorphous form. It is amorphous **SiO_2** - homogeneous and isotropic - that is used in integrated circuit technology. The link provides the phase diagram of **SiO_2** and lists some of its allotropes.
  - **SiO_2** has excellent *dielectric properties*. Its dielectric constant ϵ_r is about **3.7 - 3.9** (depending somewhat on its structure). This *was* a value *large* enough to allow decent capacitances if **SiO_2** is used as capacitor dielectric, but *small* enough so that the time constant $R \cdot C$ (which describes the time delay in wires having a resistance R and being insulated by **SiO_2**, and thus endowed with a parasitic capacitance C that scales with ϵ_r) does not limit the maximum frequency of the devices. It is here that successors for **SiO_2** with larger or smaller ϵ_r values are needed to make the most advanced devices (which will hit the market around **2002**).
  - It is among the *best insulators* known and has one of the highest *break-through field strengths* of all materials investigated so far (it can be as high as **15 MV/cm** for very thin layers; compare that with the values given for normal "bulk" dielectrics).
  - The *electrical properties of the **Si - SiO_2** interface are excellent*. This means that the interface has a very low density of energy states (akin to surface states) in the bandgap and thus does neither provide recombination centers nor introduce fixed charges.
  - **SiO_2** is relatively *easy to make* with several quite different methods, thus allowing a large degree of process leeway.
  - It is also relatively *easy to structure*, i.e. unwanted **SiO_2** can be removed selectively to **Si** (and some other materials) without many problems.
  - It is very *stable and chemically inert*. It essentially protects the **Si** or the whole integrated circuit from rapid deterioration in a chemically hostile environment (provided simply by humid air which will attack most "unprotected" materials).
- What are the uses of **SiO_2**? Above, some of them were already mentioned, here we just list them a bit more systematically. If you do not quite understand some of the uses - do not worry, we will come back to it.
  - **Gate oxides**: As we have seen before, we need a thin dielectric material to insulate the gate from the channel area. We want the channel to open at low threshold voltages and this requires large dielectric constants and especially no charges in the dielectric or at the two interfaces. Of course, we always need high break through field strength, too. No dielectric so far can match the properties of **SiO_2** in total.
  - *Dielectrics in integrated capacitors*. Capacitors with high capacitance values at small dimensions are needed for so-called **dynamic random access memories** (***DRAM***), one of the most important integrated circuits (in terms of volume production). You want something like **30 fF** (femtofarad) on an area of **0.25 µm^2**. The same issues as above are crucial, except that a large dielectric constant is even more important. While **SiO_2** was the material of choice for many **DRAM** generations (from the **16 kbit DRAM** to the **1 Mbit DRAM**), starting with the **4 Mbit** generation in about **1990**, it was replaced by a triple layer

of **SiO_2 - Si_3N_4 - SiO_2**, universally known as "***ONO***" (short for oxide - nitride - oxide); a compromise that takes not only advantage of the relatively large dielectric constant of silicon nitride (around **7.5**) while still keeping the superior quality of the **Si - SiO_2** interface, but has a few added benefits - at added costs, of course.

- *Insulation*: Some insulating material is needed between the transistor in the **Si** as well as between the many layers of wiring on the chip; cf. the many pictures in chapter four, starting with the one accessible via the link. **SiO_2** was (and still is) the material of choice. However, here we would like to have a material with a *small* dielectric constant, ideally **1**, minimizing the parasitic capacitance between wiring and **SiO_2** may have to be replaced with a different kind of dielectric around **2003**.
- *Stress relieve* layer: **SiO_2** becomes "*viscous*" at high temperatures - it is a glass, after all. While it is a small effect, it is large enough to absorb the stress that would develop between unwielding materials, e.g. **Si_3N_4** on **Si**, if it is used as a "**buffer oxide**", i.e. as a thin intermediary layer.
- **Masking**: Areas on the **Si** which are not be exposed to *dopant diffusion* or *ion implantation* must be protected by something impenetrable and that also can be removed easily after "use". It's **SiO_2**, of course, in many cases,
- **"Screen oxides"** provide one example of so-called **sacrificial layers** which have no direct function and are disposed off after use. A screen oxide is a thin layer of **SiO_2** which stops the low energy ***debris*** that comes along with the high-energy ion beam - consisting, e.g., of metal ions that some stray ions from the main beam banged off the walls of the machine. All these (highly detrimental) metal and carbon ions get stuck in the screen oxide (which will be removed after the implantation) and never enter the **Si**. In addition, the screen oxide scatters the main ion beam a little and thus prevents "*channeling*", i.e. deep penetration of the ions if the beam happens to be aligned with a major crystallographic direction.
- **Passivation**: After the chip is finished, it has to be protected from the environment and all bare surfaces need to be electrically passivates - its done with **SiO_2** (or a mixture of oxide and nitride).

- **Gate oxide for Transistors**
- **Dielectric in Capacitors**
- **Insulation**
- **Stress relieve layer**
- **Masking layer**
- **Screen oxide during Implantation**
- **Passivation**

Enough reasons for looking at the oxide generation process a little more closely? If you think not - well there are more uses, just consult the list of processes for a **16 Mbit DRAM**: You need **SiO_2** about **20** times!

How is **SiO_2** made - in thin layers, of course? There are essentially *three* quite distinct processes with many variations:

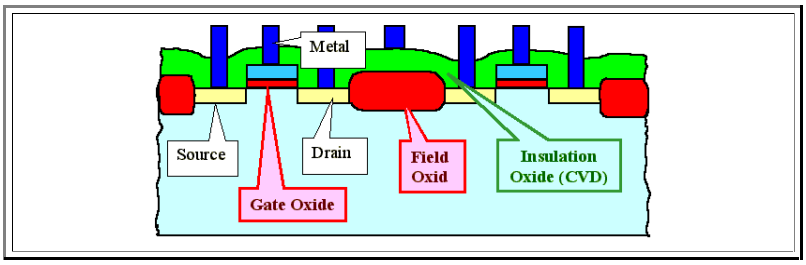
- **Thermal oxidation**. This means that a *solid state reaction* (**Si + O_2 ⇒ SiO_2**) is used: Just expose **Si** to **O_2** at sufficiently high temperatures and an oxide will grow to a thickness determined by the temperature and the oxidation time.
- **CVD oxide deposition**. In complete analogy to the production of poly-**Si** by a **CVD** (= chemical vapor depositions) process, we can also produce **SiO_2** by taking the right gases and deposition conditions.
- **Spin-on glass** (***SOG***). Here a kind of polymeric suspension of **SiO_2** dissolved in a suitable solvent is dropped on a rapidly spinning wafer. The centrifugal forces spread a thin viscous layer of the stuff on the wafer surface which upon heating solidifies into (not extremely good) **SiO_2**. Its not unlike the stuff that was called "***water* glass**" or "liquid glass" and that your grandma used to conserve eggs in it.

There is one more method that is, however, rarely used - and never for mass production: **Anodic oxidation**.

- Anodic oxidation uses a current impressed on a **Si** - electrolyte junction that leads to an oxidation reaction. While easy to understand in principle, it is not very well understood in practice and an object of growing basic research interest.

Thermal Oxidation

- In this paragraph we will restrict ourselves to **thermal oxidation**. It was (and to a large extent still is) one of the key processes for making integrated circuits. While it may be used for "secondary" purposes like protecting the bare **Si** surface during some critical process (remember the "screen oxide" from above?), its major use is in three areas:
  - **Gate oxide** (often known as "***GOX***")
  - **Capacitor dielectric**, either as "simple" oxide or as the "bread" in an "**ONO**" (**=** oxide-nitride-oxide sandwich)
  - **Field oxide** (***FOX***) - the lateral insulation between transistors.
- We can use a picture from chapter 4.1.4 to illustrate **GOX** and **FOX**; the capacitor dielectric can also be found in this chapter.



  - We must realize, however, that those drawing are *never to scale*. The gate oxide is only around **10 nm** thick (actually, it "just" (**2007**) petered out at **1.2 nm** accoding to Intel and is now replaced by a thicked **HfO_2**), whereas the field oxide (and the insulating oxide) is in the order of **500 nm**. What it looks like at atomic resolution in an electron microscope is shown in this link.
- There are essentially *two* ways to do a thermal oxidation.
- "**Dry oxidation**", using the reaction
.

$$2\,Si + O_2 \Rightarrow 2\,SiO_2 \quad (800\,^oC - 1100\,^oC)$$

  - This is the standard reaction for thin oxides. Oxide growth is rather slow and easily controllable.
  - To give an example: Growing **700 nm** oxide at **1000 °C** would take about **60 hr** - far too long for efficient processing. But **7nm** take only about **15 min** - and that is now too short for precise control; you would want to lower the temperature.
- "**Wet oxidation**", using the reaction

$$Si + 2\,H_2O \Rightarrow SiO_2 + 2\,H_2 \quad (800\,^oC - 1100\,^oC)$$

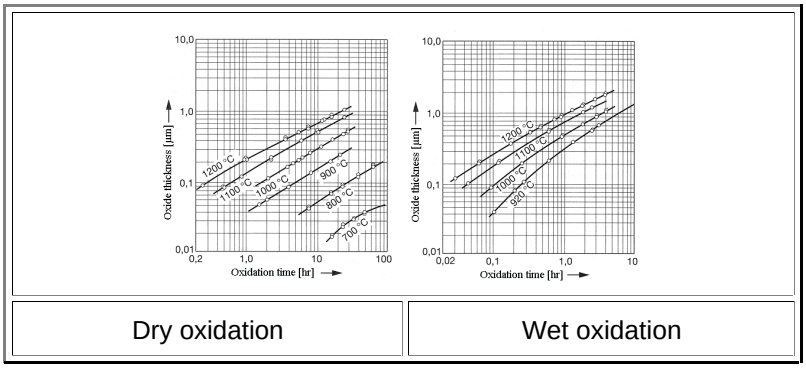
  - The growth kinetics are about **10x** faster than for dry oxidations; this is the process used for the thick *field oxides*.
  - Growing **700 nm** oxide at **1000 °C** now takes about **1.5 hr** - still pretty long but tolerable. Thin oxides are never made this way.
- In both cases the oxygen (in the form of **O**, **O_2**, **OH^-**, whatever,...) has to *diffuse through the oxide already formed* to reach the **Si - SiO_2** interface where the actual reaction takes place.
  - This becomes more difficult for thick oxides, the reaction after *some time of oxide formation* is always **diffusion limited**. The thickness d_{ox} of the growing oxide in this case follow a general "*square root*" law, i.e. it is proportional to the diffusion length $L = (Dt)^{1/2}$ (D = diffusion coefficient of the oxygen carrying species in **SiO_2**; t = time).
  - We thus have a general relation of the form .

$$d_{thick\text{-}ox} = const. \cdot (D \cdot t)^{1/2}$$

  - For *short times or thin oxide thicknesses* (about **< 30 nm**), a *linear* law is found

$$d_{thin\text{-}ox} = const. \cdot t$$

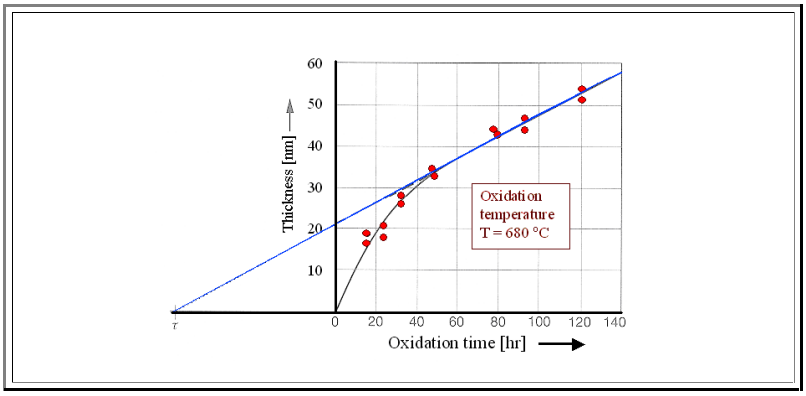
  - In this case the limiting factor is the rate at which **O** can be incorporated into the **Si - SiO_2** interface.
- This kind of behavior - linear growth switching to square root growth - can be modelled quite nicely by a not too complicated theory known as the **Deal-Grove model**. Some results of experiments and modeling are shown below.



Dry oxidation	Wet oxidation

- The left diagram shows dry, the right hand one wet oxidation. The solid curves were calculated with the Deal-Grove model after parameter adjustment, the circles indicate experimental results. *The theory seems to be pretty good !*.

However, for *very thin oxides* - and those are the ones needed or **GOX** or capacitors - things are even more complicated as shown below.

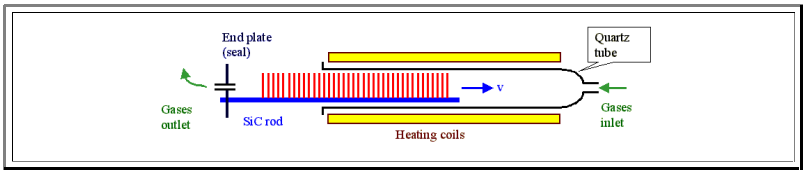


- The red points are data points from an experiment (at an unusually low temperature); the blue curve is the best Deal-Grove fit under the (not justified) assumption that at $t = 0$ an oxide with a thickness of about **20 nm** was already present.
- The Deal-Grove model is clearly inadequate for the technologically important case of very thin oxides and experimentally an *exponential law* is found for the dependence of the oxide thickness on time for very short times
- Moreover, the detailed kinetics of oxide growth are influenced by many other factors, e.g. the *crystallographic orientation* of the **Si** surface, the *mechanical stress* in the oxide (which in turn depends on many process variables), the substrate *doping*, and the *initial condition* of the **Si** surface.

And this is only the thickness! If we consider the *properties* of the oxide, e.g. the amount of fixed charge or interface charge, its etching rate, or - most difficult to ***assess*** - how long it will last when the device is used, things become most complicated. An oxide with a nominal thickness d_{ox} can be produced in many ways: dry or wet oxidation, high temperatures and short oxidation times or the other way around - its properties, however, can be very different.

- We won't look into details here but only use the issue to illustrate an important point when discussing processes for microelectronics:
- Learning about microelectronic processes involves very little Math; and "theory" is needed only at an elementary to medium level! But this does *not* make the issue trivial - quite the contrary. If you would have a theory - however complicated - that predicts all oxide properties as a function of all variables, process development would be easy. But presently, even involved theories are mostly far too simple to come even close to what is needed. On an advanced level of real process development, it is the interplay of a solid foundation in materials science, lots of experience, usage of mathematical models as far as they exist, and finally some luck or "feeling" for the job, that will carry the day.
- So do not consider microelectronic processes "simple" because you do not find lots of differential equations here. There are few enterprises more challenging for a materials scientist then to develop key processes for the next chip generation!

- How is a thermal oxidation done in real life? Always inside an oxidation furnace in a **batch process**; i.e. many wafers (usually **100**) are processed at the same time.
  - Oxidation furnaces are complicated affairs, the sketch below does not do justice to the intricacies involved (nowadays they are usually no longer horizontal as shown below, but vertical). For some pictures of real furnaces use the link.



  - First of all, temperature, gas flow etc., needs not only to be very constant but precisely adjustable to your needs. Generally, you do not load the furnace at the process temperature but at some lower temperature to avoid thermal shock of the wafers (inducing temperature gradients, leading to mechanical stress gradients, leading to plastic deformation, leading to the generation of dislocations, leading to wafers you have to throw away). After loading, you "ramp up" the temperature with a precisely defined rate, e.g. **15 ^{o}C/min**, to the process temperature selected.
  - During loading and ramping up, you may not want to start the oxidation, so you run **N_2** or **Ar** through the furnace. After the process temperature has been reached, you switch to **O_2** for dry oxidation or the right mixture of **H_2** and **O_2** which will immediately burn to **H_2O** if you want to do a wet oxidation.
  - After the oxidation time is over, you ramp down the temperature and move the wafers out of the furnace.
- Moving wafers in and out of the furnace, incidentally, is not easy.
  - First you have to transfer the wafers to the rigid rod system (usually **SiC** clad with high purity quartz) that moves in and out, and than you have to make sure that the whole contraption - easily **2 m** long - moves in and out at a predetermined speed **v** without ever touching anything - because that would produce particles.
  - There is quite a weight on the rods and they have to be absolutely unbendable even at the highest process temperature around **1150 ^{o}C**.
  - Of course, the rods, the quartz furnace tube and anything else getting hot or coming into contact with the **Si**, must be ultrapure - hot **Si** would just lap up even smallest traces of deadly fast-diffusing metals.
- And now to the *difficult* part: After the process "works", you now must make sure that it works exactly the same way (with tolerances of **< 5%**) on all **100** wafers in a run in, from run to run, and independently of which one of the **10** or so furnaces is used.
  - So take note: Assuring *stable process specifications* for a *production environment* may be a more demanding and difficult job than to develop the process in the first place.
- You see: At the final count, a "simple" process like thermal oxidation consists of a process recipe that easily calls for more than **20** precisely specified parameters, and changing anyone just a little bit may change the properties of your oxide in major ways.

All these points were emphasized to demonstrate that even seemingly simple processes in the production of integrated circuits are rather complex.

  - The processes to be discussed in what follows are no less complex, rather more so. But we will not go into the finer points at great depth anymore.
- There are two more techniques to obtain **SiO_2** which are so important that we have to consider them in independent modules:
  - *Local oxidation* - this will be contained in the following module, and
  - ***CVD** deposition* of oxide - this will be part of the **CVD** module.

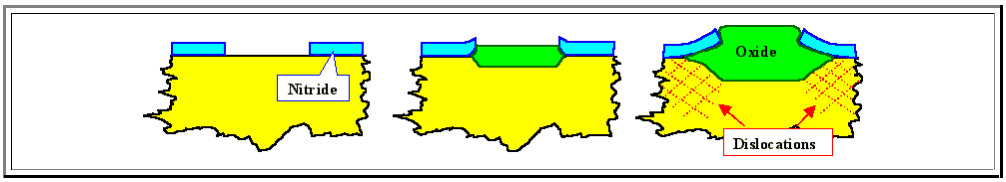
Questionaire

Multiple Choice questions to 6.2.1

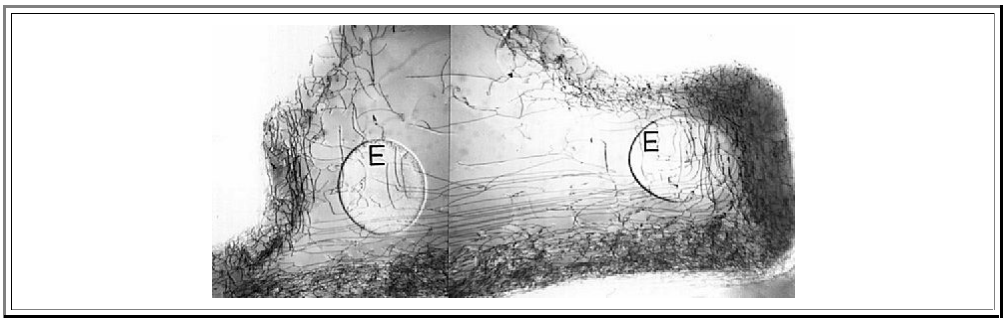
6.2.2 LOCOS Process

Basic Concept of Local Oxidation

- The abbreviation "***LOCOS***" stands for "**Local Oxidation of Silicon**" and was almost a synonym for **MOS** devices, or more precisely, for the insulation between single transistors. **LOCOS** makes the isolation between **MOS** transistors considerably easier then between bipolar transistors, cf. the drawings discussed before:
  - For bipolar transistors, you have to separate the collectors. This involves an epitaxial layer and some deep diffusion around every transistor.
  - For MOS transistors, no isolation would be needed weren't it for the possible parasitic transistors. And this problem can be solved by making the "gate oxide" of the parasitic transistors - which then is called **field oxide** - sufficiently thick.
- The thick field oxide has been made by the **LOCOS** process from the beginning of **MOS** technology until presently, when **LOCOS** was supplanted by the "box isolation technique", also known as "***STI***" for "**Shallow trench isolation**".
  - Since the **LOCOS** technique is still used, and gives a good example of how processes are first conceived, are optimized with every generation, become very complex, and are finally supplanted with something different, we will treat it here in some detail
- As the name implies, the goal is to oxidize **Si** only *locally*, wherever a field oxide is needed. This is necessary for the following reason:
  - *Local* (thermal) oxide penetrates into the **Si** (oxidation is using up **Si**!), so the **Si - SiO_2** interface is *lower* than the source - drain regions to be made later. This could not be achieved with oxidizing all of the **Si** and then etching off unwanted oxide.
  - For device performance reasons, this is highly beneficial, if not absolutely necessary.
- For a *local* oxidation, the areas of the **Si** that are not to be oxidized must be protected by some *material* that does not allow oxygen diffusion at the typical oxidation temperatures of **(1000 - 1100) ^{0}C**. We are talking electronic materials again!
  - The *only material* that is "easily" usable is **Silicon nitride**, ***Si_3N_4***. It can be deposited and structured without too much problems and it is compatible with **Si**.
  - However, **Si_3N_4** introduces a major new problem of its own, which can only be solved by making the process more complicated by involving yet another materials. This gives a ***succinct*** example of the statement made before: That materials and processes have to be seen as a unit.
  - Lets see what would happen with just a **Si_3N_4** layer protecting parts of the **Si** from thermal oxidation.



  - Oxygen diffusion through the oxide already formed would also oxidize the Si under the **Si_3N_4**; i.e. there would be some amount of lateral oxidation. Since a given volume of **Si** expands by almost a factor of **2** upon oxidation (in other words: Oxidizing **1cm^3** of **Si** produces almost **2 cm^3** of **SiO_2**), the nitride mask is pressed upwards at the edges as illustrated.
  - With increasing oxidation time and oxide thickness, pressure under the nitride mask increases, and at some point the *critical yield strength* of **Si** at the oxidation temperature is exceed. *Plastic deformation* will start and dislocations are generated and move into the **Si**. Below the edges of the local oxide is now a high density of dislocations which kill the device and render the **Si** useless - throw it out.
  - This is not "theory", but eminently practical as shown in the ***TEM*** picture from the early days of integrated circuit technology:



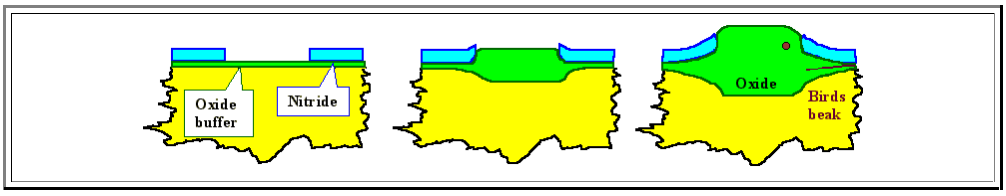
- We are looking through a piece of **Si**. The dark lines are the projections of single dislocations, the "dislocations tangles" corresponds to oxide edges; "**E**" shows contact areas (emitters) to the **Si**. Another picture can be found in the link.
- Actually, it doesn't even need the oxidation to produce dislocations. Si_3N_4 layers are always under large stresses at room temperature and would exert great shear stresses on the **Si**; something that can not be tolerated as soon as the nitride films are more than a few **nm** thick.

We arrive at a simple rule: You *cannot* use Si_3N_4 directly on **Si** - never ever. What are we to do now, to save the concept of local oxidation?

Buffer Oxide

We need something *between* the Si_3N_4 mask and the **Si**; a thin layer of a material that is compatible with the other two and that can *relieve the stress* building up during oxidation. Something like the oil in you motor, a kind of ***grease***.

- This "grease" material is SiO_2, as you might have guessed - it was already mentioned before under its proper name of "**buffer oxide**". The hard Si_3N_4 (which is a ceramic that is very hard not yielding at a "low" temperature of just about **1000 °C**), is now pressing down on something "soft", and the stress felt by the **Si** will not reach the yield stress - if everything is done right.
- The situation now looks like this



- No more dislocations, but a comparatively large lateral oxidation instead, leading to a configuration known as "**birds beak**" for the obvious reason shown in the picture to the right (the inserts just are there to help you see the bird).

So we got rid of one problem, but now we have another one: The lateral extension of the field oxide via the birds beak is comparable to its thickness and *limits the minimum feature size*.

- While this was not a serious problems in the early days of **IC** technology, it could not be tolerated anymore around the middle of the eighties.
- One way out was the use of a poly-**Si** layer as a sacrificial layer. It was situated on top of the buffer oxide below the nitride mask and was structured with the mask. It provided some sacrificial **Si** for the "birds beak" and the total dimension of the field oxide could be reduced somewhat.
- This process is shown in comparison with the standard process in the link.

But even this was not good enough anymore for feature sizes around and below **1 µm**. The **LOCOS** process eventually became a very complicated process complex in its own right; for the Siemens **16 Mbit DRAM** it consisted of more than **12** process steps including:

- **2** oxidations, **2** poly-**Si** deposition, **1** lithography, **4** etchings and **2** cleaning steps.
- It was one of the decisive "secrets" for success, and we can learn a simple truth from this:

Before new materials and processes are introduced, the existing materials and processes are driven to extremes! And that is not only true for the **LOCOS** process, but for all other processes.

- Still, with feature sizes shrinking ever more, **LOCOS** reached the end of its useful life-span in the nineties and had to be replaced by "Box isolations", a simple concept in theory, but hellishly difficult in reality.
- The idea is clear: Etch a hole (with vertical sidewalls) in the **Si** wherever you want an oxide, and simple "fill" it with oxide next. More about this process can be found in the link above.

6.2.3 Summary to: 6.2 Si Oxide and LOCOS Process

- Silicondioxide (SiO_2) has been the "ideal" dielectric with many uses in chip manufacture
  - Only recently (**2007**) is it replaced by "low **k"** and "high **k**" dielectrics, i.e. dielectrics with a dielectric constant either lower or larger than that of SiO_2
  - "Low **k**" dielectrics (polymers, porous SiO_2, ..; the ideal material has not yet been found) are used for intermetal insulation; low **k** is important here to keep the **RC** time constants small
  - "High **k**" dielectrics (the present front runner is HfO_2) will replace the gate oxides. They can be somewhat thicker than SiO_2 without sacrificing capacity, while strongly reducing tunneling currents.

- **Gate oxide for Transistors**
- **Dielectric in Capacitors**
- **Insulation**
- **Stress relieve layer**
- **Masking layer**
- **Screen oxide during Implantation**
- **Passivation**

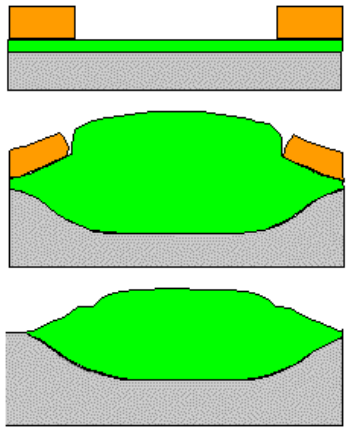
- SiO_2 can be made in several ways:
  - Dry oxidation is relatively slow but gives best oxide qualities as defined by:
    - Uniformity
    - thickness control
    - Break down field strangt
    - Interface quality
    - Reliability

    Typical use: Highest quality gate oxid.
  - Wet oxidation is about 10 times faster; it is used whenever relatively thick oxides are needed.
    Typical use: Field oxide.
  - The other methods are needed whenever there is no **Si** available for oxidation (e.g. intermetal dielectrics).

- Dry thermal oxidation:
  $2\,Si + O_2 \Rightarrow 2\,SiO_2$
- Wet thermal oxidation:
  $Si + 2\,H_2O \Rightarrow SiO_2 + 2\,H_2$
- "Chemical Vapor Deposition" (next sub-chapter)
- "Spin-on techniques (next sub-chapter)
- "Anodic oxidation (presently not used in technology)

- As long as the process is diffusion controlled (i.e. the time it takes oxygen to diffuse through the already formed oxide determines rates, the thickness increases protorional to $t^{1/2}$
  - For thin oxides the growth rate is reaction controlled and the thickness - time dependence becomes complicated.

- Growing oxide only locally ("LOCOS") was a key process for field oxides.
  - Without a "buffer" oxide below the masking nitride, large mechanical strain develops, producing plastic deformation and thus dislocations around the oxide edges.
  - These "Oxide edge dislocations" kill the transistor.
  - Buffer oxides solve the problem, but create new problems: A "birds beak" develops, increasing lateral dimensions beyond the mask dimension.
- "LOCOS" is a good example for a universal feature of **Si** technology: Solutions to "old" problems create new problems. Solutions to the new problems... and so on. It follows:
  - Process complexity increases all the time.
  - New materials are needed all the time.



Questionaire

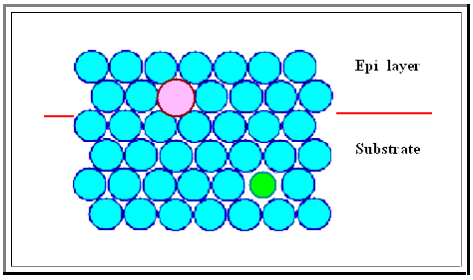
Multiple Choice questions to all of 6.2

6.3 Chemical Vapor Deposition

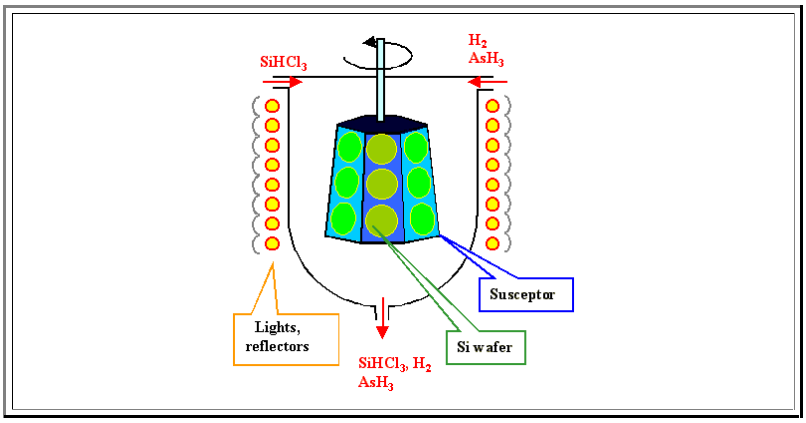
6.3.1 Silicon Epitaxy

- We have encountered the need for epitaxial layers before, and we also have seen a **Si CVD** process for making poly-crystalline material good enough for growing crystals. All we have to do now is to put both things together.
  - We can use essentially the same **CVD** process as before, but instead of thin rods of poly-**Si** which we want to grow in diameter as fast as possible, we now want to make a thin, but absolutely perfect **Si** layer on top of a wafer.

- We now must have tremendous process control. We require:
  - A *precise continuation* of the substrate lattice. There should be no way whatsoever to identify the interface after the epitaxial layer has been deposited. This means that *no lattice defects whatsoever* should be generated.
  - *Doping* of the epitaxial layer with high precision (e.g. **5 Ωcm ± 5%**), and the doping is usually very different from that of the substrate. The picture on the right symbolizes that by the two differently colored doping atoms.
  - *Precise thickness control*, e.g. ***d* = 1.2 µm ± 10%** over the entire wafer, from wafer to wafer and from day to day. Now there is a challenge: If you met the first point and thus can't tell where the interface is - how do you measure the thickness? (The answer: Only electronically, e.g. by finding the position of the **pn**-junction produced).
  - *Cleanliness*: No contaminants diffusing into the substrate and the epitaxial layer are allowed.



- This looks tough and it is indeed fairly difficult to make good epitaxial layers. It is also quite expensive and is therefore avoided whenever possible (e.g. in mainstream **CMOS** technology). It is, however, a must in bipolar and some other technologies and also a good example for a very demanding process with technical solutions that are far from obvious.
- Lets look at a typical epitaxial reactor from around **1990** (newer ones tend to be single wafer systems). It can process several wafers simultaneously and meets the above conditions. Here is a muchly simplified drawing:



- The chemical reaction that produces the **Si** is fairly simple:

$$SiCl_4 + 2\ H_2 \Rightarrow Si + 4\ HCl$$

(1000 °C - 1200 °C)

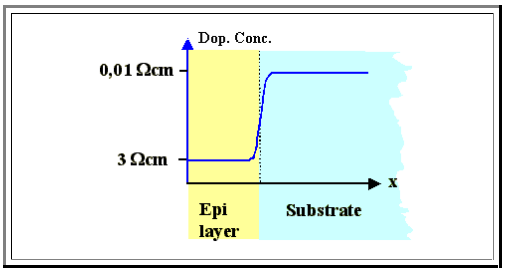
  - The dopant gases just decompose or react in similar ways. However, instead of **$SiCl_4$** you may want to use **SiH_xCl_{4-x}** .
  - The essential point is that the process needs high temperatures and the **Si** wafer will be at high temperature! In an Epi reactor as shown above, the **Si** wafer surfaces (and whatever shows of the susceptor) are the only hot surfaces of the system!
- How is the process actually run? You need to meet some tight criteria for the layer specifications, as outlined above, and that transfers to tight criteria for process control.

- **1.** *Perfectly clean* **Si** surface before you start. This is not possible by just putting clean **Si** wafers inside the Epi-reactor (they always would be covered with SiO_2), but requires an in-situ cleaning step. This is done by first admitting only H_2 and Cl_2 into the chamber, at a very high temperature of about **1150 °C**. **Si** is etched by the gas mixture - every trace of SiO_2 and especially foreign atoms at the surface will be removed.
- **2**. Temperature gradients of at most **(1 - 2) °C**. This is (better: was) achieved by **heating with light** as shown in the drawing. The high intensity light bulbs (actually rods) consume about **150 kW** electrical power (which necessitates a **30 kW** motor running the fan for air-cooling the machinery).
- **3.** Extremely tightly controlled gas flows within a range of about **200 l/min** H_2, **5 l/min** $SiCl_4$ (or **Si** HCl_3), and fractions of **ml/min** of the doping gases.

Not to forget: Epi-reactors are potentially very dangerous machines with a lot of "dirty" output that needs to be cleaned. All things taken together make Epi-reactors very expensive - you should be prepared to spend several million **$** if you want to enter this technology.

- **Si** epitaxy thus is a process that is avoided if possible - it costs roughly **$5** per wafer, which is quite a lot. So when do we use epitaxy?

Epitaxy is definitely needed if a *doping profile* is required where the *resistivity in surface near regions is larger than in the bulk*. In other words, a profile like this:



- By diffusion, you can always lower the resistivity and even change the doping type, but *increasing the resistivity by diffusion* is not realistically possible.
- Consider a substrate doping of 10^{16} **cm**^{3}. Whatever resistivity it has (around **5 - 10** Ω**cm**), if you diffuse **2 ·** 10^{16} **cm**^{3} of a dopant into the substrate, you *lowered* the resistivity of the doped layer by a factor of **2**.
- To *increase* the resistivity you have to compensate half of the substrate doping by diffusing a dopant for the reverse doping type with a concentration of **5 ·** 10^{15} **cm**^{3}. Not only does that call for much better precision in controlling diffusion, but you will only get that value at a particular distance from the surface because you always have a diffusion profile. So all you can do by diffusion is to increase the resistivity somewhat near the surface regions; but you cannot make a sizeable layer this way.

You also may use epitaxial layers if you simply need a *degree of freedom in doping* that is not achievable otherwise.

- While **DRAMs** were made without epitaxy up to the **16 Mbit** generation (often to the amazement of everybody, because in the beginning of the development work epitaxy seemed to be definitely necessary), epitaxial **Si** layers are now included from the **64 Mbit DRAM** upwards.

6.3.2 Oxide CVD

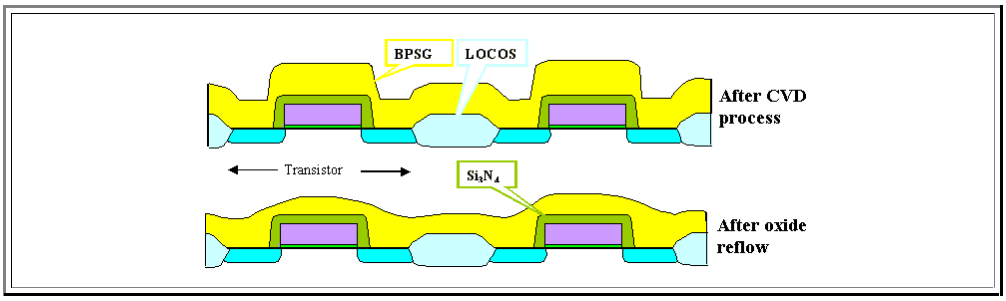
- Whenever we need **SiO_2** layers, but can not oxidize **Si**, we turn to **oxide CVD** and deposit the oxide on top of the substrate - whatever it will be
  - Again, we have to find a suitable chemical reaction between gases that only occurs at high temperature and produces **SiO_2**. There are several possibilities, one is

$$SiH_2Cl_2 + 2NO_2 \Rightarrow SiO_2 + 2HCl + \tfrac{1}{2}N_2 \quad (900\ ^oC)$$

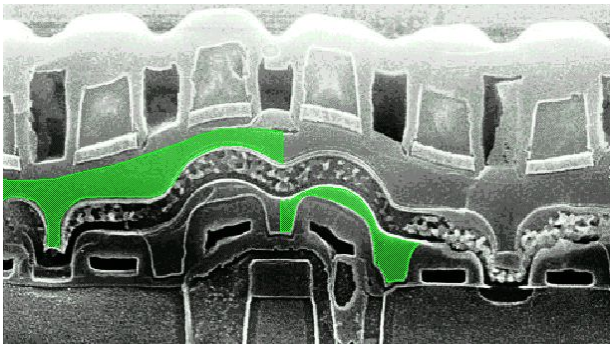
  - While this reaction was used until about **1985**, a better reaction is offered by the "***TEOS***" process:

$$Si(C_2H_5O)_4 \Rightarrow SiO_2 + 2H_2O + 4C_2H_4 \quad (720\ ^oC)$$

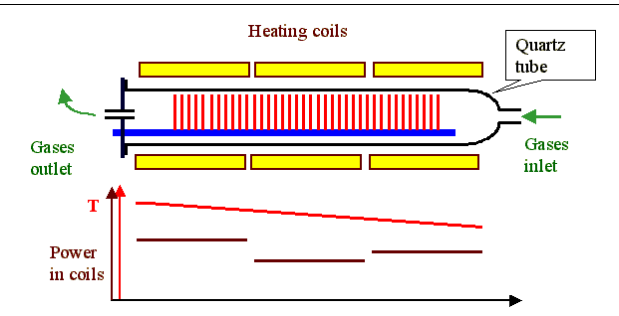
  - **$Si(C_2H_5O)_4$** has the chemical name **Tetraethylorthosilicate**; abbreviated **TEOS**. It consists of a **Si** atom with the four organic molecules bonded in the four tetrahedral directions. The biggest advantage of this process is that it can be run at lower temperatures, but it is also less dangerous (no **HCl**), and it produces high quality oxides.
- Low temperature processes are important after the transistors and everything else in the **Si** has been made. Every time the temperature must be raised for one of the processes needed for metallization, the dopant atoms will move by diffusion and the doping profiles change.
  - Controlling the "**temperature budget**" is becoming ever more important as junction depths are getting smaller and smaller.
- **CVD** techniques allow to tailor some properties of the layers deposited by modifying their chemistry. Often, an oxide that "flows" at medium temperature, i.e. evens out the topography somewhat, is needed. Why is shown below.



  - After the transistor has been made, there is a veritable mountain range. Here it is even worse than before, because the whole "gate stack" has been encapsulated in **Si_3N_4** - for reasons we will not discuss here. (Can you figure it out? The process is called "**FOBIC**", short for "Fully Overlapping Bitline Contact").
  - It is important for the next processes to flatten the terrain as much as possible. While this is now done by one of the major key process complexes introduced around **1995** (in production) called "***CMP***" for "**Chemical-mechanical polishing**", before this time the key was to make a "**flow glass**" by doping the **SiO_2** with **P** and/or **B**. Conventional glass, of course is nothing like **SiO_2** containing ions like **Na** (which is a no-no in chip making), but **P** and **B** are also turning quartz into glass.
  - The major difference between *glass* and *quartz* is that glass becomes a kind of viscous liquid above the *glass temperature* which depends on the kind and concentration of ions incorporated.
- So all you have to do during the **SiO_2** deposition, is to allow some incorporation of **B** and/or **P** by adding appropriate gases.
  - As before phosphine (**PH_3**) is used for **P**, and "**TMB**" (= **$B(OCH_3)_3$ =** trimethylborate) for **B**. Concentrations of both elements may be in the **%** range (**4% P** and **2% B** are about typical), the resulting glass is called "***BPSG***" (= Bor-Phosphorous Silicate Glass). It "flows" around **850 ^oC**, i.e the viscosity of **BPSG** is then low enough to allow the surface tension to reduce the surface areas by evening out peaks and valleys.
  - How much it "flows" can be further influenced by the atmosphere during the annealing: **O_2** or even better, **H_2O** like in wet oxidation, enhances the viscosity and helps to keep the thermal budget down
  - The **BPSG** process was a key process to ***VLSI*** (= **Very Large Scale Integration**), this can be seen in any cross section of a real device. Lets look at the cross section of the **16 Mbit DRAM** again that was shown before:



- Two layers of **BPSG** are partially indicated in green
  - The lower layer has been etched back to some extent; it only fills some deep crevices in some places.
  - Both layers smoothed the topography considerably; but there can never be complete planarization with **BPSG** glasses, of course.
- How do we carry out an oxide **CVD** process? Of course, we could use a "*tool*" like an epi-reactor, but that would be an overkill (especially in terms of money).
- For "simple" oxide **CVD**, we simply use a furnace as in the thermal oxidation process and admit the process gases instead of oxygen. However, there are certain (costly) adjustments to make:
  - **CVD** processes mostly need to be run at *low pressure* (then often abbreviated **LPCVD**) - some **mbar** will be fine - to ensure that the layers grow smoothly and that the gas molecules are able to penetrate into every nook and cranny (the mean free path length must be large). The furnace tube thus must be vacuum tight and big pumps are needed to keep the pressure low.
  - We want to *minimize the waste* (dangerous gases not used up) which at the same time maximizes the conversion of the (costly) gases to SiO_2. But this means that at the end of the tube the partial pressure of the process gases is lower than at the beginning (most of it has been used up by then). To ensure the same layer thickness for the last wafer than for the first one, requires a higher temperature at the end of the furnace tube because that leads to a higher reaction rate countering the lower gas concentration.
  - The first wafers to be exposed to the gas flow are "air-cooled" by the process gas to some extent. We therefore need to raise the temperature a bit at the front end of the furnace.
- Essentially, we must be able to run a *defined temperature gradient* along the **CVD** furnace tube! This calls for at least three sets of heating coils which must be independently controlled.
  - The whole thing looks like this



- Again, we see that there are many "buttons" to adjust for a "simple" **CVD** oxide deposition.
  - Base pressure and temperature, flow rates of the gases, temperature profile of the furnace with the necessary power profile (which changes if a gas flow is changed), ramping up and ramping down the temperature, etc., - all must be right to ensure constant thickness of the deposited layer for every wafer with minimum waste of gases.
  - Changing any parameter may not only change the oxide thickness, but also its properties (most important, maybe, its etch rate in some etching process).
  - Developing a "new" oxide **CVD** process thus is a lengthy undertaking, demanding much time and ingenuity. But since this is true for every process in microelectronics, we will from now on no longer emphasize this point.
- **CVD** furnaces have a major disadvantage: Everything inside (including the quartz tube) is hot and will become covered with oxide (including the wafer back sides). This is not quite so bad, because the quartz tube will simply grow in thickness. Nevertheless, in regular intervals everything has to be **cleaned** - in special equipment inside the cleanroom! Very annoying, troublesome and costly!
- A "conventional" **CVD** furnace is, however, not the only way to make **CVD** oxides. Several dedicated machines have been developed just for **BPSG** or other variants of oxides.

- One kind, adding also something new to the process, merits a short paragraph: ***PECVD*** or "**Plasma Enhanced CVD**"

Plasma Enhanced CVD

- As the thermal budget gets more and more constrained while more and more layers need to be added for multi-layer metallization, we want to come down with the temperature for the oxide (or other) **CVD** processes.
  - One way for doing this is to supply the *necessary energy for the chemical reaction* not by heating everything evenly, but just the gas. The way to do this is to pump electrical energy into the gas by exposing it to a suitable electrical field at high frequencies. This could induce dielectric losses, but more important is the *direct energy transfer* by collisions as soon as the *plasma* stage is reached.
  - In a gas plasma, the atoms are ionized and both free electrons and ions are accelerated in the electrical field, and thus gain energy which equilibrates by collisions. However, while the average kinetic energy and thus the temperature of the heavy ions is hardly affected, it is quite different for the electrons: Their temperature as a measure of their kinetic energy may attain **20.000 K**.
  - (If you have problems with the concept of *two* distinctly different temperatures for *one* material - you're doing fine. Temperature is an *equilibrium* property, which we do not have in the kind of plasma produced here. Still, in an approximation, one can consider the electrons and the ions being in equilibrium with the other electrons and ions, respectively, but not among the different species, and assign a temperature to each subgroup separately.)
  - The chemical reactions thus may occur at low nominal temperatures of just a few **100 ^{o}C**.
- There are many kinds of **PECVD** reactors, with **HF** frequencies from **50 kHz to >10 MHz** and electrical power of several **100 W** (not to be sneered at in the **MHz** range!).
  - Since after the first **Al** deposition, the temperature has to be kept below about **400 ^{o}C**, (otherwise a **Si - Al** eutectic will form), **PECVD** oxide is the material of choice from now on, rivaled to some extent by spin-on glass.
  - However, its properties are not quite as good as those of regular **CVD** oxide (which in turn is inferior to thermal oxide).

Footnote:	There are certain names used for the "hardware" needed to make chips that are not immediately obvious to the uninitiated:
	Simple people - e.g. you and me or operators - may talk of "*machinery*" or "*machines*" - which is what those big boxes really are.
	More sophisticated folks - process engineers or scientists - talk about their "*equipment*"
	Really sophisticated people - managers and **CEOs** - will contemplate the "*tools*" needed to make chips.

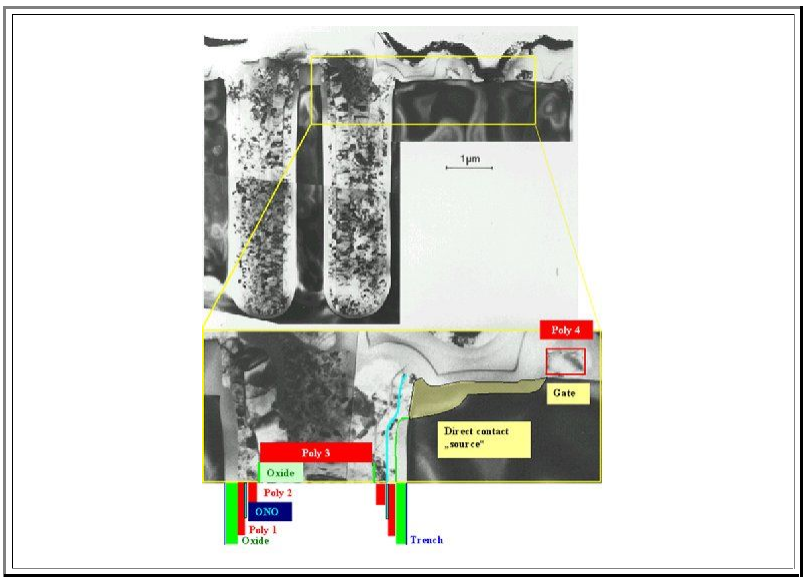
6.3.3 CVD for Poly-Silicon, Silicon Nitride and Miscellaneous Materials

Poly Silicon CVD

- If we were to use an epitaxial reactor for wafers covered with oxide, a layer of **Si** would still be deposited on the hot surface - but now it would have no "guidance" for its orientation, and poly-crystalline **Si** layers (often just called "**poly**" or "polysilicon") would result.
  - **Poly-Si** is one of the key materials in microelectronics, and we know already how to make it: Use a **CVD** reactor and run a process similar to epitaxy.
  - If doping is required (it often is), admit the proper amounts of dopant gases.
- However, we also want to do it cheap, and since it we want a polycrystalline layer, we don't have to pull all the strings to avoid crystal lattice defects like for epitaxial **Si** layers.
  - We use a more simple **CVD** reactor of the furnace type shown for oxide **CVD**, and we employ smaller temperatures (and low pressure, e.g. **60 Pa** since we only need thin layers and can afford lower deposition rates). This allows to use SiH_4 instead of $SiCl_4$; our process may look like this:

$$SiH_4 \xrightarrow[630^oC]{60\ Pa} Si + 2\ H_2$$

  - Much cheaper! The only (ha ha) problem now is: Cleaning the furnace. Now you have poly-**Si** all over the place; a little bit nastier than SiO_2, but this is something you can live with.
- What is poly-**Si** used for and why it is a key material?
  - Lets look at a **TEM** (= transmission electron microscope) picture of a memory cell (transistor and capacitor) of a **16 Mbit DRAM**. For a larger size picture and additional pictures click here.



- All the speckled looking stuff is poly-**Si**. If you want to know exactly what you are looking at, turn to the drawing of this cross section. We may distinguish **4** layers of poly **Si**:
  - "*Poly **1***" coats the inside of the trench (after its surface has been oxidized for insulation) needed for the capacitor. It is thus one of the "plates" of the capacitor. In the **4 Mbit DRAM** the substrate **Si** was used for this function, but the space charge layer extending into the **Si** if the capacitor is charged became too large for the **16 Mbit DRAM**.
  - The "*Poly **2***" layer is the other "plate" of the capacitor. The **ONO** dielectric in between is so thin that it is practically invisible. You need a ***HRTEM*** - a high resolution transmissin electron microscope - to really see it.
  - Now we have a capacitor folded into the trench, but the trench still needs to be filled. Poly-**Si** is the material of choice. In order to insulate it from the poly capacitor plate, we oxidize it to some extent before the "*poly **3***" plug is applied.
- One plate of the capacitor needs to be connected to the source region of the transistor. This is "simply" done by removing the insulating oxide from the inside of the trench in the right place (as indicated).
  - Then we have a *fourth poly layer*, forming the gates of the transistors.
  - And don't forget: there were two sacrificial poly-**Si** layers for the LOCOS process!
- That makes **6** poly-**Si** deposition (that we know off). Why do we like poly-**Si** so much?

- Easy! It is perfectly compatible with single crystalline **Si**. Imagine using something else but poly-**Si** for the plug that fills the trench. If the thermal expansion coefficient of "something else" is not quite close to **Si**, we will have a problem upon cooling down from the deposition temperature.
- No problem with poly. Moreover, we can oxidize it, etch it, dope it, etc. (almost) like single crystalline **Si**. It only has *one* major drawback: Its conductivity is not nearly as good as we would want it to be. That is the reason why you often find the poly-gates (automatically forming one level of wiring) "re-enforced" with a **silicide** layer on top.
- A silicide is a metal silicon compound, e.g. **Mo_2Si**, **PtSi**, or **Ti_2Si**, with an almost metallic conductivity that stays relatively inert at high temperatures (in contrast to pure metals which react with **Si** to form a silicide). The resulting double layer is in the somewhat careless slang of microelectronics often called a "*polycide*" (its precise grammatical meaning would be the killing of the poly - as in fratricide or infanticide).
- Why don't we use a silicide right away, but only in conjunction with poly-**Si**? Because you would loose the all-important high quality interface of (poly)-**Si** and **SiO_2**!

Si_3N_4 Deposition

- We have seen several uses for silicon nitride layers - we had **LOCOS**, **FOBIC** (and there are more), so we need a process to deposit **Si_3N_4** .
  - Why don't we just "nitride" the **Si**, analogous to oxidations, by heating the **Si** in a **N_2** environment? Actually we do - on occasion. But **Si_3N_4** is so impenetrable to almost everything - including nitrogen - that the reaction stops after a few **nm**. There is simply no way to grow a "thick" nitride layer thermally.
  - Also, don't forget: **Si_3N_4** is always producing tremendous stress, and you don't want to have it directly on the **Si** without a buffer oxide in between. In other words: We need a **CVD** process for nitride.
- Well, it becomes boring now:
  - Take your **CVD** furnace from before, and use a suitable reaction, e.g.
    .

$$3\ SiH_2Cl_2 + 4NH_3 \Rightarrow Si_3N_4 + 2HCl + 1{,}5\ H_2 \quad (\approx 700\,^{o}C))$$

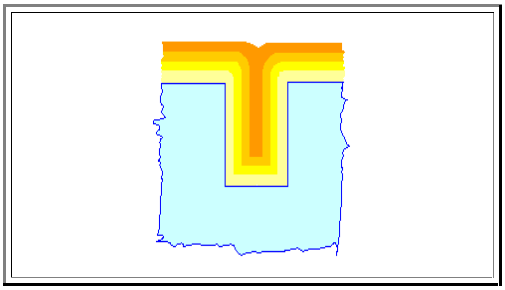
  - Nothing to it - except the cleaning bit. And the mix of hot ammonia (**NH_3**) and **HCl** occurring simultaneously if you don't watch out. And the waste disposal. And the problem that the layers, being under internal stresses, might crack upon cooling down. And, - well, you get it!

Tungsten CVD

- For reasons that we will explain later, it became necessary at the end of the eighties, to deposit a metal layer by **CVD** methods. Everybody would have loved to do this with **Al** - but there is no good **CVD** process for **Al**; nor for most other metals. The candidate of choice - mostly by default - is **tungsten** (chemical symbol ***W*** for "Wolfram").
  - Ironically, **W-CVD** comes straight form nuclear power technology. High purity **Uranium** (chemical symbol ***U***) is made by a **CVD** process not unlike the **Si** Siemens process using **UF_6** as the gas that decomposes at high temperature.
  - W is chemically very similar to **U**, so we use **WF_6** for **W-CVD**.
- A **CVD** furnace, however, is not good enough anymore. **W-CVD** needed its own equipment, painfully (and expensively) developed a decade ago.
  - We will not go into details, however. **CVD** methods, although quite universally summarily described here, are all rather specialized and the furnace type reactor referred to here, is more an exception than the rule.

Advantages and Limits of CVD Processes

- **CVD** processes are ideally suited for depositing thin layers of materials on some substrate. In contrast to some other deposition processes which we will encounter later, **CVD** layers always follow the contours of the substrate: They are conformal to the substrate as shown below.



- Of course, conformal deposition depends on many parameters. Particularly important is which process dominates the reaction:
  - *Transport controlled process* (in the gas phase). This means that the rate at which gas molecules arrive at the surface controls how fast things happen. This implies that molecules react immediately wherever they happen to reach the hot surface. This condition is always favored if the *pressure is low enough*.
  - *Reaction controlled kinetics*. Here a molecule may hit and leave the surface many times before it finally reacts. This reaction is dominating at high pressures.
- Controlling the partial pressure of the reactants therefore is a main process variable which can be used to adjust layer properties.
  - It is therefore common to distinguish between ***APCVD*** (**=** atmospheric pressure **CVD**) and ***LPCVD*** (**=** low pressure **CVD**).
  - **LPCVD**, very generally speaking, produces "better" layers. The deposition rates, however, are naturally much lower than with **APCVD**.
- **CVD** deposition techniques, though quite universal and absolutely essential, have certain *disadvantages*, too. The two most important ones (and the only ones we will address here) are
  - They are not possible for some materials; there simply is no suitable chemical reaction.
  - They are generally not suitable for *mixtures* of materials.
- To give just one example: The metallization layers for many years were (and mostly still are) made from **Al** - with precise additions of **Cu** and **Si** in the **0,3% - 1%** range
  - There is no suitable **Al**-compound that decomposes easily at (relatively low) temperatures. This is not to say that there is none, but all **Al**-organic chemicals known are too dangerous to use, to expensive, or for other reasons never made it to production (people tried, though).
  - And even if there would be some **Al CVD** process, there is simply no way at all to incorporate **Si** and **Cu** in the exact quantities needed into an **Al CVD** layer (at least nobody has demonstrated it so far).
- Many other materials, most notably perhaps the silicides, suffer from similar problems with respect to **CVD**. We thus need alternative layer deposition techniques; this will be the subject of the next subchapter.

Questionaire

Multiple Choice questions to 6.3.3

Footnote:	The name "*Poly Silicon*" is used for at least three qualitatively very different kinds of materials:
	1. The *"raw material" for crystal growth*, coming from the "Siemens" **CVD** process. It comes - after breaking up the rods - in large chunks suitable for filling the crucible of a crystal grower.
	2. Large ingot of cast **Si** and the *thin sheets made from them*; exclusively used for solar cells. Since the grains are very large in this case (in the cm range), this material is often referred to as "**multi crystalline Si**".
	3. The *thin layers of poly **Si*** addressed in this sub-chapter, used for micro electronics and micro mechanical technologies. Grain sizes then are µm or less.
	In addition, the term poly **Si** might be used (but rarely is) for the *dirty stuff coming out of the **Si** smelters*, since this **MG-Si** is certainly poly-crystalline

6.3.4 Summary to: 6.3 Chemical Vapor Deposition

- Chemical Vapor Deposition (CVD) is simple in principle
  - Find to gases that react ot the desired material at elevated temperatures
  - Put your wafer(s) into some machine, evacuate, heat to the desired temperature (preferably only the wafers) and admit the gases (and remove undesired reaction products).
  - There are many quute different technical ways (all of them expensive) to realize a **CVD** apparatus

- Major **CVD** process are
- Deposition of epitaxial **Si** layers - obviously always on (atomically clean) **Si** substrates. By admitteing some gases carrying doping atoms (e.g. **AsH_3, AsH_3**) the layer can be doped in-situ.
- Depositon of poly crystalline **Si** layers.
  - Chemically similar to epitaxial layers, in reality quite different because the CVD reactos can be simpler.
  - Poly-**Si** is needed for many uses: Gate electrode, interconnect, filling of holes, sacrifical layer.
  - Its great advantage is its ull compatibility with **Si** and **SiO_2**; its great disadvantage is its mediocre conductivity (for heavy doping).
- Deposition of **Si_3N_4**
  - Very important. Always prone to produce mechanical stress (**Si_3N_4** is an unyielding ceramic!).
- **W** (and Silicides, and ...)
  - Not "good" processes, but sometimes unavoidable!

Epitaxial Si layer

$$SiCl_4 + 2\,H_2 \Rightarrow Si + 4\,HCl \quad (1000\,^oC - 1200\,^oC)$$

Polycrystalline Si layer

$$SiH_4 \xrightarrow[630^oC]{60\ Pa} Si + 2\,H_2$$

SiO_2 layer ("TEOS process)

$$Si(C_2H_5O)_4 \Rightarrow SiO_2 + 2H_2O + C_2H_4 \quad (720\,^oC)$$

Si_3N_4 layer

$$3\,SiH_2Cl_2 + 4NH_3 \Rightarrow Si_3N_4 + 2HCl + 1{,}5\,H_2 \quad (\approx 700\,^oC))$$

W layer

$$WF_6 + H_2 \xrightarrow[470^oC]{10^4\ Pa} W + \text{(nasty) gases}$$

Questionaire

Multiple Choice questions to all of 6.3

6.4. Physical Processes for Layer Deposition

- The technologies discussed in this subchapter essentially cover *physical processes* for layer deposition as opposed to the more *chemical* methods introduced in the preceding subchapters. In other words, we deal with some more techniques for the material module in the process cycle for **ICs**.
  - While still intimately tied to the electronic materials to be processed, these technologies are a bit more of a side issue in the context of electronic materials, and will therefore be covered in *far less detail* then the preceding, more material oriented deposition techniques.
  - On occasion, however, a particular tough problem in **IC** processing is elucidated in the context of the particular deposition method associated with it. *So don't skip these modules completely!*
- Essentially, what you should know are the basic technologies employed for layer deposition, and some of the major problems, advantages and disadvantages encountered with these techniques in the context of chip manufacture.

6.4.1 Sputter Deposition and Contact Hole Filling

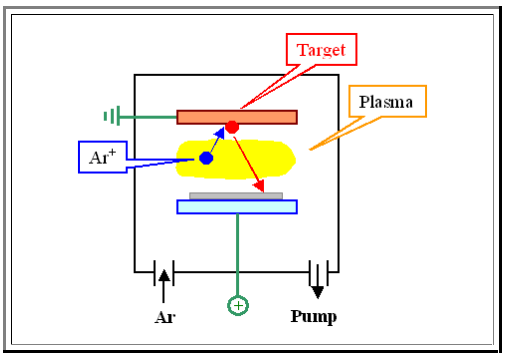
General Remarks

- It should be clear by now that the *deposition of thin layers* is the key to all microelectronic structures (not to mention the emerging micro electronic and mechanical systems (**MEMS**), or **nano technology**).
  - *Chemical* vapor deposition, while very prominent and useful, has severe limitations and the more alternative methods exist, the better.
  - *Physical* methods for controlled deposition of thin layers do exist too; and in the remainder of his subchapter we will discuss the major ones.
- What are *physical* methods as opposed to *chemical* methods? While there is no ironclad distinction, we may simply use the following rules
  - If the material of the layer is produced by a *chemical reaction* between some primary substances in-situ, we have a chemical deposition process. This does not just include the **CVD** processes covered before, but also e.g. galvanic layer deposition.
  - If the material forming the layer is sort of transferred from some substrate or source to the material to be coated, we have a *physical process*. The most important physical processes for layer deposition which shall be treated here are

    - **Sputtering techniques**
    - **Ion implantation**
    - **Spin on coating**

Basic Sputter Process

- "**Sputtering**" or "**sputter deposition**" is a conceptually simple technique:
  - A "**target**" made of the material to be deposited is bombarded by energetic ions which will dislodge atomes of the target, i.e., "***sputter*** them off".
  - The dislodged atoms will have substantial kinetic energies, and some will fly to the substrate to be coated and stick there.
- In practice, this is a lot easier than it appears. The basic set-up is shown below:



- The ions necessary for the bombardment of the target are simply extracted from an **Ar plasma** burning between the target and the substrate.

- Both target and substrate are planar plates arranged as shown above. They also serve as the cathode and anode for the gas discharge that produces an **Ar** plasma, i.e. ionized **Ar** and free electrons, quite similar to what is going on in a fluorescent light tube.
- Since the target electrode is always the cathode, i.e. negatively charged, it will attract the $\mathbf{Ar^+}$ ions and thus is bombarded by a (hopefully) constant flux of relatively energetic **Ar** ions.
- This ion bombardment will liberate atoms from the target which issue forth from it in all directions.

Reality is much more complex, of course. There are many ways of generating the plasma and tricks to increase the deposition rate. Time is money, after all, in semiconductor processing.
Some target atoms will make it to the substrate to be coated, others will miss it, and some will become ionized and return to the target. The important points for the atoms that make it to the substrate (if everything is working right) are:

- **1.** The target atoms hit the substrate with an energy large enough so they "*get stuck*", but not so large as to liberate substrate atoms. Sputtered layers therefore usually stick well to the substrate (in contrast to other techniques, most notably evaporation).
- **2.** *All* atoms of the target will become deposited, in pretty much the same composition as in the target. It is thus possible, e.g., to deposit a silicide slightly off the stoichiometric composition (advantageous for all kinds of reason). In other words, if you need to deposit e.g. $\mathbf{TaSi_{2-x}}$ with $\mathbf{x \approx 0.01 - 0.1}$, sputtering is the way to do it because it is comparatively easy to change the target composition.
- **3.** The target atoms hit the substrate coming from *all directions*. In a good approximation, the flux of atoms leaving the target at an angle Φ relative to the normal on the target is proportional to **cos** Φ. This has profound implications for the coverage of topographic structures.
- **4.** Homogeneous coverage of the substrate is relatively easy to achieve- just make the substrate holder and the target big enough. The process is also relatively easily scaled to larger size substrates - simply make everything bigger.

Of course, there are problems, too.

- Sputtered layers usually have a very bad crystallinity - very small grains full of defects or even amorphous layers result. Usually some kind of annealing of the layers is necessary to restore acceptable crystal quality.
- Sputtering works well for metals or other somewhat conducting materials. It is not easy or simply impossible for insulators. Sputtering $\mathbf{SiO_2}$ layers, e.g., has been tried often, but never made it to production.
- While the **cos** Φ relation for the directions of the sputtered atoms is great for over-all homogeneity of the layers, it will prevent the filling of holes with large **aspect ratios** (aspect ratio **=** depth/width of a hole). Since contact holes and vias in modern **ICs** always have large aspect ratios, a serious problem with sputtering A**l(Si Cu)** for contacts came up in the nineties of the last century. This is elaborated in more detail below.

More or less by default, sputtering is the layer deposition process of choice for **Al**, the prime material for metallization.

- How else would you do it? Think about it. We already ruled out **CVD** methods. What is left?
- The deposition of a metallization layer on a substrate with "heavy" topography - look at some of the drawings and pictures to understand this problem - is one of the big challenges **IC** technology and a particularly useful topic to illustrate the differences between the various deposition technologies; it will be given some special attention below.

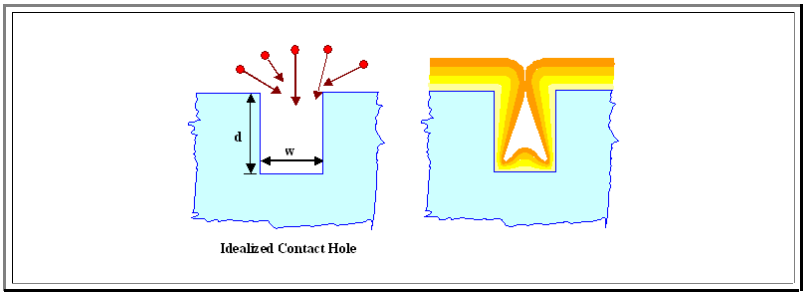
The Contact Hole Problem

The metallization of chips for some **30** years was done with Aluminum as we know by now - cf. all the drawings in chapter 5 and the link.

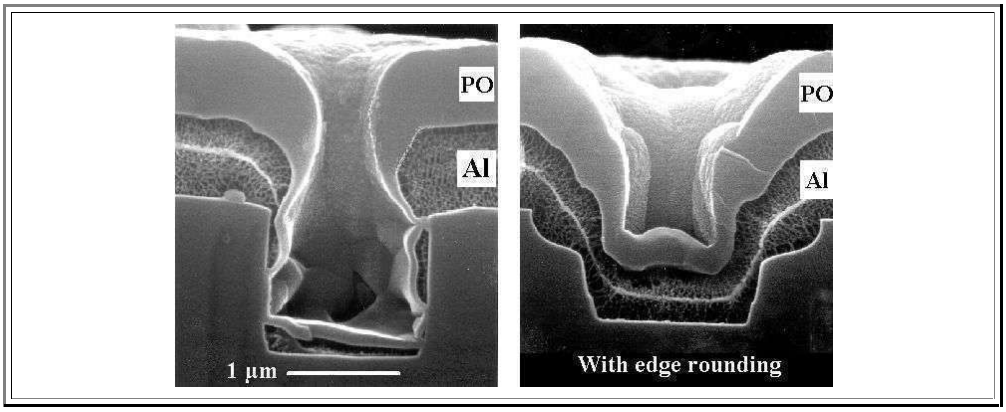
- **Al**, while far from being optimal, had the best over-all properties (or the best "figure of merit") including less than about **0,5 % Si** and often a little bit (roughly **1 %**) of **Cu**, **V** or **Ti**. These elements are added in order to avoid deadly "spikes", to decrease the contact resistance by avoiding epitaxial Si precipitates and to make the metallization more resistant to electromigration.
- While you do not have to know what that means (you might, however, look it up via the links), you should be aware that there are even more requirements for a metallization material than listed before.

Sputtering is the only process that can deposit an **Al** layer with a precisely determined addition of some other elements on a large **Si** substrate. There is an unavoidable problem however, that becomes more severe as features get smaller, related to the so-called "edge coverage" of deposition processes.

- Many **Al** atoms hitting the substrate under an oblique angle, will not be able to reach the bottom of a contact hole which thus will have less Al deposited as the substrate surface.
- To make it even worse, the layer at the edge of the contact hole tends to be especially thick, reducing the opening of the hole disproportionately and thus allowing even less **Al** atoms to bottom of the hole. What happens is illustrated below.



- Fort aspect ratios ***A = w/d*** smaller than approximately **1**, the layer at the edge of the contact hole will become unacceptably thin or will even be non-existing - the contact to the underlying structure is not established.
  - The real thing together with one way to overcome the problem is shown below (the example is from the **4Mbit DRAM** generation, around **1988**).



  - ***PO*** denotes "**Plasma oxide**", referring to a PECVD deposition technique for an oxide. This layer is needed for preparation purposes (the upper **Al** edge would otherwise not be clearly visible)
- Clearly, the **Al** layer is interrupted in the contact hole on the left. **Al** sputter deposition cannot be used anymore without "tricks".
  - One possible trick is shown on the right: The edges of the contact hole were "rounded". "**Edge rounding**", while not easy to do and consuming some valuable "real estate" on the chip, saved the day for the at or just below the **1µm** design rules.
  - But eventually, the end of sputtering for the **1st** metallization layer was unavoidable - despite valiant efforts of sputter equipment companies and **IC** manufacturers to come up with some modified and better processes - and a totally new technology was necessary
- This was **Tungsten CVD** just for filling the contact hole with a metal.
  - In some added process modules, the wafer was first covered with tungsten until the contact holes were filled (cf. the drawing in the **CVD** module). After that, the tungsten on the substrate was polished off so that only filled contact holes remained.
  - After that, **Al** could be deposited as before.
- However, depositing **W** directly on **Si** produced some new problems; related to interdiffusion of **Si** and **W**.
  - The solution was to have an intermediary **diffusion barrier** layer (which was, for different reasons, already employed in some cases with a traditional **Al** metallization).
  - Often, this diffusion barrier layer consisted of a thin **$TiSi_2$/Ti /TiN** layer sequence. The **$TiSi_2$** formed directly as soon as **Ti** was deposited (by sputtering, which was still good enough for a very thin coating), the **Titanium Nitride** was formed by a **reactive sputtering process**.
  - Reactive sputtering in this case simply means that some **N_2** was admitted into the sputter chamber which reacts immediately with freshly formed (and extremely reactive) **Ti** to **TiN**.
- A typical contact to lets say **p**-type **Si** now consisted of a **p-Si/p^+-Si/$TiSi_2$/Ti/TiN/W/Al** stack, which opened a new can of worms with regard to contact reliability. Just imagine the many possibilities of forming all kinds of compounds by interdiffusion of whatever over the years.
  - But here we stop. Simply because meanwhile (i.e. **2001**), contacts are even more complicated, employing **Cu** (deposited galvanically after a thin **Cu** layer necessary for electrical contact has been sputter deposited), various barrier layers, possibly still **W**, and whatnot.
  - So: Do look at a modern chip with some awe *and remember*: We are talking electronic materials here!

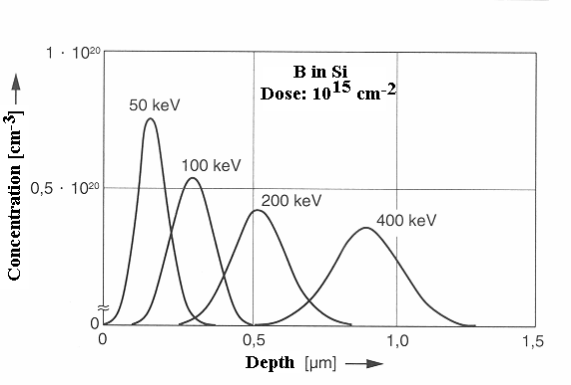
Questionaire

Multiple Choice questions to 6.4.1

6.4.2 Ion Implantation

Ion Implantation Basics

- What is **ion implantation**, often abbreviated ***I²*** ? The name tells it succinctly: *Ions* of some material - almost always the dopants **As**, **B**, **P** - are *implanted*, i.e. shot into the substrate.
- Ion implantation may be counted among layer deposition processes because you definitely produce a layer of something different from the substrate even so you do not deposit something in the strict meaning of the term.
- How is it done? Obviously you need an ion beam, characterized by three basic parameters:
  - **1.** The *kind of the ions*. Almost everything from the periodic table could be implanted, but in practice you will find that only **Sb** (as dopant) and occasionally **Ge** and **O** are being used besides the common dopants **As**, **B**, and **P**.
  - **2**. The *energy of the ions* in **eV**. This is directly given by the accelerating voltage employed and is somewhere in the range of **(2 - 200) kV**, always allowing for extremes in both directions for special applications. The energy of the ion together with its mass determine how far it will be shot into a **Si** substrate. The following graph gives an idea of the distribution of **B** atoms after implantation with various energies. The curves for **As** or **P** would be similar, but with peaks at smaller depth owing to the larger mass of these atoms.



  - There are several interesting points to this graph: Obviously everything happens at dimensions ≤ **1 µm**, and a **dose** ***D*** of 10^{15} **cm**$^{-2}$ gives a pretty high doping density in the peak value. Moreover, by changing the implantation energy, all kinds of concentration profiles could be produced (within limits). That is completely impossible by just diffusing the dopant into the **Si** from the outside.
  - **3.** The *flux* (number of ions per **cm²** and second), i.e. the current (in **µA** or **mA**) carried by the ion beam. In order to obtain some rough idea about the current range, we assume a certain beam cross section ***A*** and a constant current density ***j*** in this cross section. A total current ***I*** then corresponds to a current density $j = I/A$ and the implanted dose is

$$D = \frac{j \cdot t}{e} = \frac{I \cdot t}{e \cdot A}$$

  - ***t*** is the implantation time, **e** the elementary charge **= 1,6 · 10^{-19}C**.
- For an implantation time of **1 s** for an area ***A* = 1 cm²** and a dose ***D* = 10^{15} cm^{-2}**, we obtain for the beam current

$$I = \frac{D \cdot e \cdot A}{t} = 10^{15} \cdot 1{,}6 \cdot 10^{-19}\ \mathrm{C \cdot s^{-1}} = 1{,}6^{-4}\ \mathrm{A}$$

  - Total implantation time for *one* **200 mm** wafer than would be about **300 s**, far too long. In other words, we need implanters capable to deliver beam currents of **10 mA** and more for high doses, and just a few precisely controlled **µA** for low doses.
  - Think a minute to consider what that means: **10 mA** in **1 cm^{-2}** at **200 kV** gives a deposited power of **2 kW** on **1 cm^{-2}**. That is three orders of magnitude larger than what your electric range at home has to offer - how do you keep the **Si** cool during implantation? If you do nothing, it will melt practically instantaneously.
- Since the beam diameter is usually much smaller than the **Si** wafer, the ion beam must be scanned over the wafer surface by some means. For simplicities sake we will dismiss the scanning procedure, even so it is quite difficult and expensive to achieve with the required homogeneity. The same is true for all the other components - ion beam formation, accelaration, beamshaping, and so on.

- If we take everything together, we start to see why ion implanters are very large, very complicated, and very expensive (several million **$)** machines. Their technology, while ingenious and fascinating, shall not concern us here, however.

- Why do we employ costly and complex ion implantation?
  - Because there simply is no other way to dope selected areas of a **Si** substrate with a precisely determined amount of some dopant atoms and a controlled concentration profile.
  - In our simple drawing of a **CMOS** structure we already have three doped regions (in reality there are much more). Just ask yourself: How do you get the dopants to their places?
- With ion implantation it is "easy":
  - Mask with some layer (usually **SiO_2** or photo resist) that absorbs the ions, and shoot whatever you need into the open areas.
- The only alternative (used in the stone age of **IC** semiconductor technology) is to use **diffusion** from some outside source.
  - Again, mask every area where you do not want the dopants with some layer that is impenetrable for the atoms supposed to diffuse, and expose the substrate to some gas containing the desired atoms at high temperature. After some time, some atoms will have diffused into the **Si** - you have your doping.
  - But there are so many problems that direct diffusion is not used anymore for complex **ICs**: Accuracy is not very good, profiles are limited, the necessary high temperatures change the profiles already established before, and so on. Simply forget it. Ion implantation is what you do.
- But like everything (and everybody), implantation has it limits.
  - For example: How do you dope around the trench shown before in the context of integrated capacitors? Obviously, you can't just shoot ions into the side wall. Or can you? Think about how you would do it and then turn to the advanced module.

Defects and Annealing

- After implanting the ions of your choice with the proper dose and depth distribution, you are not yet done.
  - Implantation is a violent process. The high energy ion transfers its energy bit by bit to lattice atoms and thus produces a large number of defects, e.g. vacancies and interstitials. Often the lattice is simply killed and the implanted layer is **amorphous**. This is shown in an illustration module.
  - You must restore order again. Not only are **Si** crystal lattice defects generally not so good for your device, but only dopant atoms, which have become neatly incorporated as substitutional impurities, will be electrically active.
- Implantation, in short, must always be followed by an annealing process which hopefully will restore a perfect crystal lattice and "activate" the implanted atoms.
  - How long you have to anneal at what temperature is a function of what and how you implanted. It is a necessary evil, because during the annealing the dopants will always diffuse and your neat implanted profiles are changing.
  - Much research has been directed to optimal annealing procedures. It might even be advantageous to anneal for very short times (about **1 s**) at very high temperatures, say **(1100 - 1200) ^{o}C**. Obviously this cannot be done in a regular furnace like the one illustrated for oxidation, and a whole new industry has developed around "**rapid thermal processing**" (***RTP***) equipment.

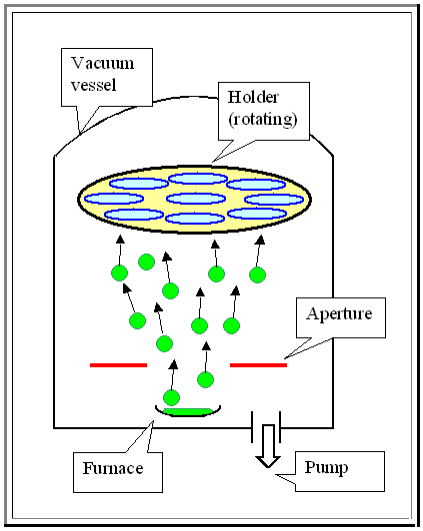
Questionaire

Multiple Choice questions to 6.4.2

6.4.3 Miscellaneous Techniques and Comparison

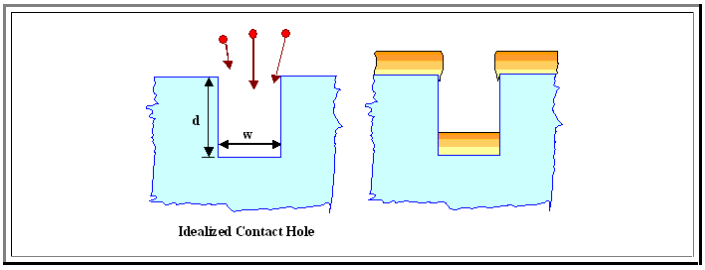
Evaporation

- By now you may have wondered why the time-honored and widely used technique of evaporation has not been mentioned in context with **Si** technology.
  - The answer is simple: *It is practically not used*. This is in contrast to other technologies, notably optics, where evaporation techniques played a major role.
  - In consequence, this paragraph shall be kept extremely short. It mainly serves to teach you that there are more deposition techniques than meets the eye (while looking at a chip).
- What is the evaporation technique? If your eye glasses or your windshield ever fogged, you have seen it: Vapor condenses on a cold substrate.



- What works with water vapor also works with all other vapors, especially metal vapor.
- All you have to do is to create the vapor of your choice, always inside a vacuum vessel kept at a good vacuum. The (usually) metal atoms will leave the crucible or "boat" with an kinetic energy of a few **eV** and sooner or later will condense on the (cooled) substrate (and everywhere else if you don't take special precautions).
- Your substrate holder tends to be big, so you can accommodate several wafers at once (Opening up and loading vacuum vessels takes expensive time!)

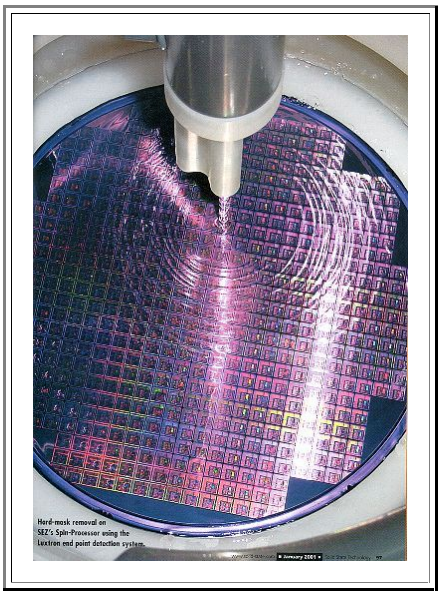
- The technique is relatively simple (even taking into account that the heating nowadays is routinely done with high power electron beams hitting the material to be evaporated), but has major problems with respect to **IC** production:
  - The atoms are coming from a "point source", i.e. their incidence on the substrate is nearly perpendicular. Our typical contact hole filling problem thus looks like this:



  - In other words: Forget it!
  - It is also clear that it is very difficult to outright impossible to produce layers with arbitrary composition, e.g. **Al** with **0,3% Si** and **0,5% Cu**. You would need three independently operated furnaces to produce the right mix.
- All things considered, sputtering is usually better and evaporation is rarely used nowadays for microelectronics.

Spin-on Techniques

- **Spin-on techniques**, a special variant of so-called *sol-gel techniques*, start with a liquid (and usually rather viscous) source material, that is "painted" on the substrate and subsequently solidified to the material you want.
  - The "*painting*" is not done with a brush (although this would be possible), but by spinning the wafer with some specified **rpm** value (typically **5000 rpm**) and dripping some of the liquid on the center of the wafer. Centrifugal forces will distribute the liquid evenly on the wafer and a thin layer (typically around **0,5 µm**) is formed.
  - Solidification may occur as with regular paint: the solvent simply evaporates with time. This process might be accelerated by some heating. Alternatively, some chemical reaction might be induced in air, again helped along by some "baking" as it is called.
- As a result you obtain a thin layer that is rather smooth - nooks and crannies of the substrate are now planarized to some extent. The film thickness can be precisely controlled by the angular velocity of the spin process (as a function of the temperature dependent viscosity of the liquid).
  - Spin-on coating is the technique of choice for producing the light-sensitive **photo resist** necessary for lithography. The liquid resist rather resembles some viscous paint, and the process works very well. It is illustrated on the right.
  - Most other materials do not have suitable liquid precursors, the spin-on technique thus can not be used.



- A noteworthy exception, however, is **spin-on glass**, a form of SiO_2 mentioned before.
  - The liquid consists basically of Silicon-tetra-acetate ($Si(CH_2COOH)_4$) (and some secret additions) dissolved in a solvent. It will solidify to an electronically not-so-good SiO_2 layer around **200 °C**.
  - Using spin-on glass is about the only way to fill the interstices between the **Al** lines with a dielectric at low temperatures. The technique thus has been developed to an art, but is rather problematic. The layers tend to crack (due to shrinkage during solidifications), do not adhere very well, and may interact detrimentally with subsequent layers.
- A noteworthy example of a material that can be "spun on", but nevertheless did not make it so far are **Polyimides**, i.e. polymers that can "take" relatively high temperatures
  - They look like they could be great materials for the intermetal dielectric - low ϵ_r, easy deposition, some planarizing intrinsic to spin-on, etc. They are great materials - but still not in use. If you want to find out why, and how new materials are developed in the real world out there, use this link.

Other Methods

- Deposition techniques for thin layers is a rapidly evolving field; new methods are introduced all the time. In the following a couple of other techniques are listed with a few remarks
- **Molecular Beam Epitaxy** (***MBE***). Not unlike evaporation, except that only a few atoms (or molecules) are released from a tricky source (an "effusion cell") at a time.
  - **MBE** needs ultra-high vacuum conditions - i.e. it is very expensive and not used in **Si-IC** manufacture. **MBE** can be used to deposit single layers of atoms or molecules, and it is relatively easy to produce multi layer structures in the **1 nm** region. An example of a **Si-Ge** multilayer structure is shown in the link
  - **MBE** is the method of choice for producing complicated epitaxial layer systems with different materials as needed, e.g., in advanced optoelectronics or for superconducting devices. An example of what you can produce with **MBE** is shown in the link
- **Laser Ablation**. Not unlike sputtering, except that the atoms of the target are released by hitting it with an intense Laser beam instead of **Ar** ions extracted from a Plasma.
  - Used for "sputtering" ceramics or other non conducting materials which cannot be easily sputtered in the conventional way.
- **Bonding techniques**. If you bring two ultraflat **Si** wafers into intimate contact without any particles in between, they will just stick together. With a bit of annealing, they fuse completely and become bonded.
  - Glass blowers have done it in a crude way all the time. And of course, in air you do not bond **Si** to **Si**, but SiO_2 to SiO_2. One way to use this for applications is to produce a defined SiO_2 layer first, bond the oxidized wafer to a **Si** wafer, then polish off almost all of the **Si** except for a layer about **1 µm** thick

- Now you have a regular wafer coated with a thin oxide and a perfect single crystalline **Si** layer - a so-called "**silicon on insulator**" (***SOI***) structure. The **Si** industry in principle would love **SOI** wafers - all you have to do to become rich immediately, is to make the process cheap. But that will not be easy. You may want to check why **SOI** is a hot topic, and how a major company is using wafer bonding plus some more neat tricks, including mystifying electrochemistry, to make **SOI** affordable.
- Bonding techniques are rather new; it remains to be seen if they will conquer a niche in the layer deposition market.

Galvanic techniques, i.e. electrochemical deposition of mostly metals. Galvanizing materials is an old technique (think of chromium plated metal, anodized aluminium, etc.) normally used for relatively thick layers.

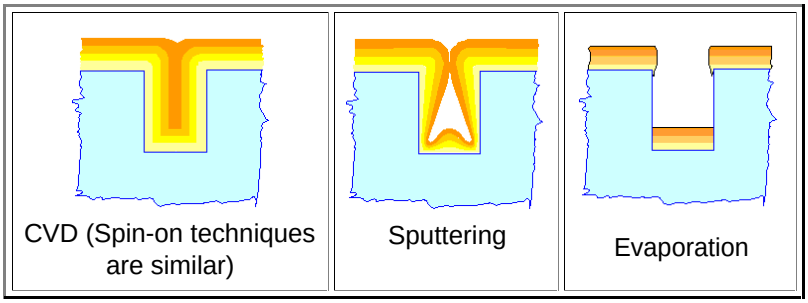
- It is a "dirty" process, hard to control, and still counted among the black arts in materials science. No self-respecting **Si** process engineer would even dream of using galvanic techniques - except that with the advent of **Cu** metallization he was not given a choice.
- Using **Cu** instead of **Al** for chip metallization was unavoidable for chips hitting the market around **1998** and later - the resistivity of the **Al** was too high.
- As it turned out, established techniques are no good for **Cu** deposition - galvanic deposition is the method of choice. **Cu** metallization calls for techniques completely different from Al metallization - the catchword is "damascene technology". The link takes you there - you may also enjoy this module from the *"Defects" Hyperscript* because it contains some other interesting stuff in the context of (old) materials science.

And not to forget: Galvanic techniques are also used in the packaging of chips

Comparison of Various Layer Deposition processes

First, lets look at **edge coverage**, i.e. the dependence on layer thickness on the topography of the substrate. This is best compared by looking at the ability to fill a small contact hole with the layer to be deposited.

- We have the following schematic behavior of the major methods as shown before.



Second, lets look at what you can deposit.

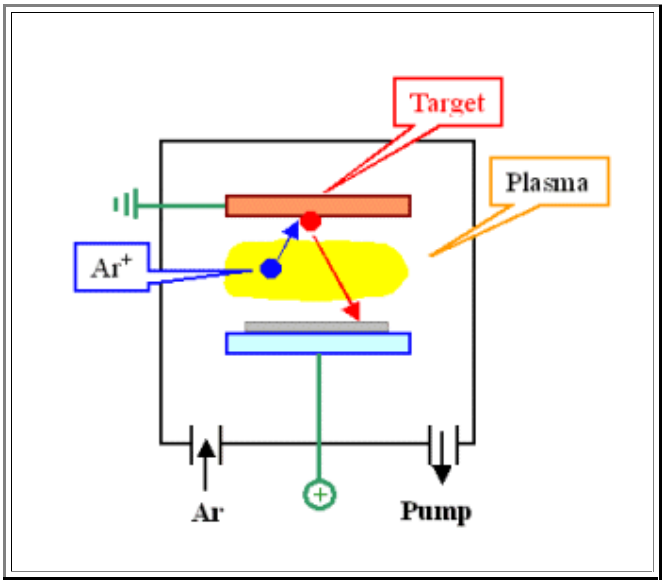
- **CVD** methods are limited to materials with suitable gaseous precursors. While it is not impossible to deposit mixtures of materials (as done, e.g. with doped poly **Si** or flow glass), it will not generally work for arbitrary compositions.
- Sputter methods in practice are limited to conducting materials - metals, semiconductors, and the like. Arbitrary mixtures can be deposited; all you have to do is make a suitable target. The target does not even have to be homogeneous; you may simply assemble it by arranging pie-shaped wedges of the necessary materials in the required composition into a "cake" target.
- Evaporation needs materials that can be melted and vaporized. Many compounds would decompose, and some materials simply do not melt (try it with **C**, e.g.). If you start with a mixture, you get some kind of distillation - you are only going to deposit the material with the highest vapor pressure. Mixtures thus are difficult and can only be produced by co-evaporation from different sources.

Questionaire

Multiple Choice questions to 6.4.3

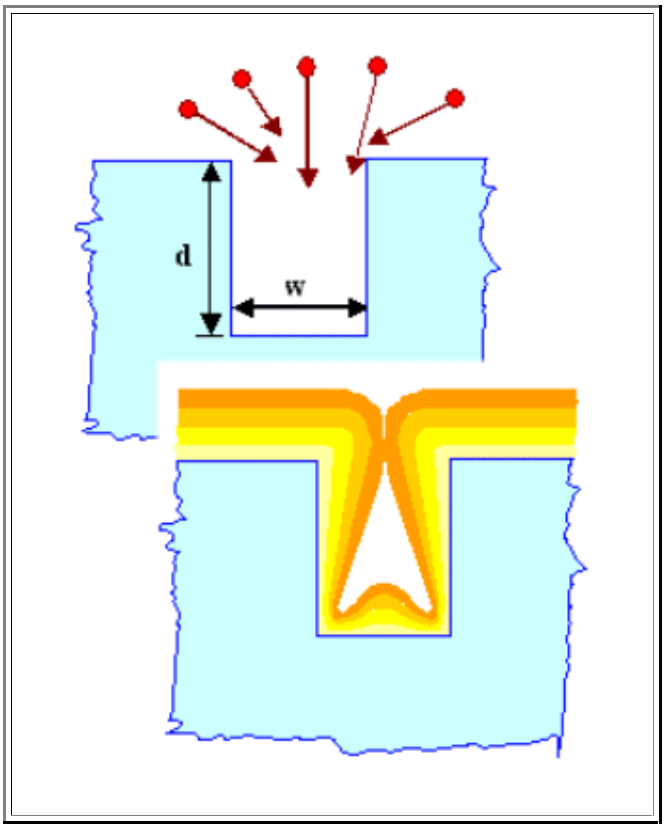
6.3.4 Summary to: 6.4 Physical Processes for Layer Deposition

- Sputter Deposition



  - Simple in principle: Shoot off any mixture of atom contained in a target be an ion beam produced by applying high (RF) voltage between target and Si wafer; see picture.
  - Great advantage is easy deposition of mixtures of elements, e.g. Al plus traces of Cu etc.
  - Disadvantage:
    1. target should be conducting; no (easy) deposition of insulators like SiO_2.
    Target atoms are emitted in all directions, leading to:

- "Contact hole filling problem"



  - The picture tells it all. At some point you need to go back to CVD processes.

- Ion implantation. Might be considered to be a "deposition" technique but actually shoots ions into the Si target.
  - The technique of choice for doping selected areas with typically B, As or P.. Depth distribution and concentrations finely adjustable in a wide range.
  - Major problems:
    1. The Si crystals gets more or less destroyed; in the extreme it turns amorphous. Implantation thus always need a follow-up annealing process that changes dopant distribution by diffusion and may not be able to restore perfectness of the lattice
    2. Implanters are huge, complex and very expensive machines.

- Other physical deposition techniques.
  There are plenty, often quite specialized. Of special important for chip making are.
  - Evaporations. Easy but very limited. Rarely used in chip production
  - Spin-on techniques (for deposition the light sensitive "resist" needed for lithography
  - Molecular Beam Epitaxy (MBE);: hugely important for III-V technology.
    Galvanic techniques. Hated but used
    Many other.

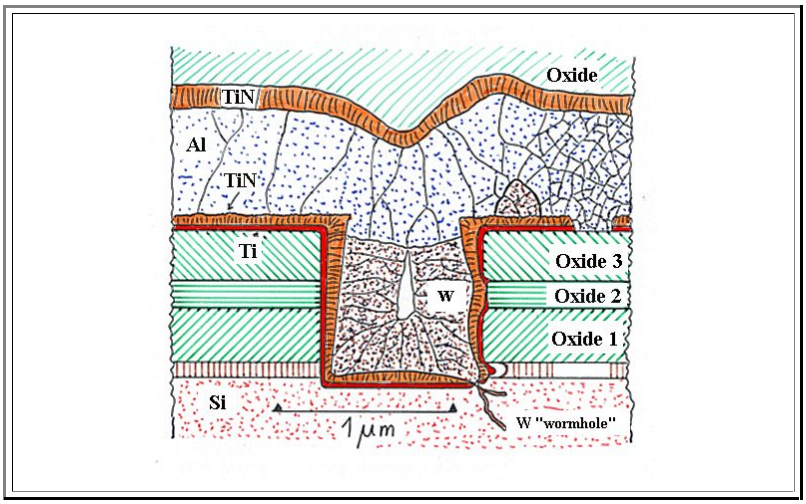
Questionaire

Multiple Choice questions to all of 6.4

6.5 Etching Techniques

6.5.1 General Remarks

- After we have produced all kinds of layer, we must now proceed to the structure module of our basic process cycle. First we discuss **etching techniques**.
  - Lets see what it means to produce a structure by etching. Lets make, e.g., a contact hole in a somewhat advanced process (and do some of the follow-up processes for clarity).
  - What the stucture contains may look like this:

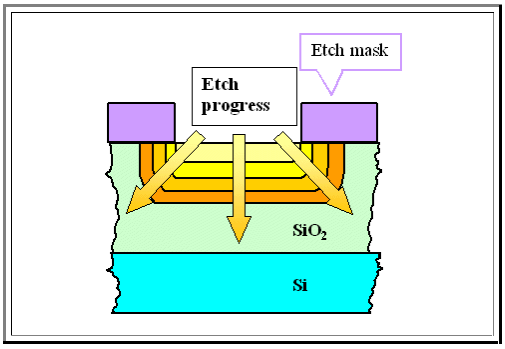


- Obviously, before you deposit the **Ti/TiN** diffusion barrier layer (and then the **W**, and so on), you must etch a hole through **3** oxide layers and an **Si_3N_4** - and here we don't care why we have all those layers. (The right hand side of the picture shows a few things that can go wrong in the contact making process, but that shall not concern us at present).
- There are some obvious requirements for the etching of this contact hole that also come up for most other etching processes.
  - **1.** You only want to etch *straight down* - not in the lateral direction. In other words, you want strongly **anisotropic etching** that only affects the bottom of the contact hole to be formed, but not the sidewalls (which are, after all, of the same material).
  - **2.** You want to *stop* as soon as you reach the **Si** substrate. Ideally, whatever you do for etching will not affect **Si** (or whatever material you want not to be affected). In other words, you want a large **selectivity** (**=** ratio of etch rates).
  - **3.** You also want reasonable *etching rates* (time is money), the ability to etch through several *different* layers in *one* process (as above), no *damage* of any kind (including rough surfaces) to the layer where you stop, and sometimes *extreme* geometries (e.g. when you etch a trench for a capacitor: **0,8 µm** in diameter and **8 µm** deep) - and you want perfect *homogeneity and reproducibility* all the time (e.g. all the about **200.000.000.000** trenches on *one* **300 mm** wafer containg **256 Mbit DRAMs** must be identical to the ones on the other ***500 - 1000*** wafers you etch today, *and* to the *thousands* you etched before, or are going to etch in the future).
- Lets face it: *This is tough*! There is no single technique that meets all the requirements for all situations.
  - Structure etching thus is almost always a search for the best compromise, and new etching techniques are introduced all the time.
  - Here we can only scratch at the surface and look at the two basic technologies in use: *Chemical or wet etching* and *plasma or dry etching*.

6.5.2 Chemical Etching

- Chemical etching is simple: Find some (liquid) chemical that dissolves the layer to be etched, but that does not react with everything else.
  - Sometimes this works, sometimes it doesn`t. Hydrofluoric acid (**HF**), for example will dissolve **SiO_2**, but not **Si** - so there is an etching solution for etching **SiO_2** with extreme selectivity to **Si**.
  - The other way around does not work: Whatever dissolves **Si**, will always dissolve **SiO_2**, too. At best you may come up with an etchant that shows somewhat different etching rates, i.e. some (poor) selectivity.

- Anyway, the thing to remember is: Chemical etchants, if existing, can provide extremely good selectivity and thus meet our *second* request from above.
  - How about the *first* request, anisotropy? Well, as you guessed: It is rotten, practically non-existent. A chemical etchant always dissolves the material it is in contact with, the forming of a contact hole would look like this:



- There is a simple and painful consequence: As soon as your feature size is about **2 µm** or smaller, *forget chemical structure etching*.
  - Really? How about making the opening in the mask smaller, accounting for the increase in lateral dimensions?
  - You could do that - it would work. But it would be foolish: If you *can* make the opening smaller, you also want your features smaller. In real life, you put up a tremendous effort to make the contact hole opening as small as you can, and you sure like hell don't want to increase it by the structure etching!
- Does that mean that there is no chemical etching in **Si** microelectronics? Far from it. There just is no chemical *structure* etching any more. But there are plenty of opportunities to use chemical etches (cf. the statistics to the **16 Mbit DRAM** process). Lets list a few:
  - *Etching off whole layers*. Be it some sacrificial layer after it fulfilled its purpose, the photo resist, or simply all the **CVD** layers or thermal oxides which are automatically deposited on the wafer backside, too - they all must come off eventually and this is best done by wet chemistry.
  - Etching *coarse structures*, e.g. the opening in some protective layers to the large **Al** pads which are necessary for attaching a wire to the outside world.
  - Etching off unwanted *native oxide* on all **Si** or poly-**Si** layers that were exposed to air for more than about **10 min**.
  - All *cleaning steps* may be considered to be an extreme form of chemical etching. Etching off about **1,8 nm** of native oxide might be considered cleaning, and a cleaning step where nothing is changed at the surface, simply has no effect.
- While these are not the exciting process modules, experienced process engineers know that this is where trouble lurks. Many factories have suffered large losses because the yield was down - due to some problem with wet chemistry.
- A totally new field, just making it into production for some special applications, is **electrochemical etching**. A few amazing (and not yet well understood) things can be done that way; the link provides some samples.

6.5.3 Plasma Etching

- **Plasma etching**, also known as **dry etching** (in contrast to *wet* etching) is the universal tool for structure etching since about **1985**. In contrast to all other techniques around chip manufacture, which existed in some form or other *before* the advent of microelectronics, plasma etching was practically unknown before **1980** and outside the microelectronic community.
  - What is Plasma etching? In the most simple way of looking at it, you just replace a liquid etchant by a plasma. The basic set-up is not unlike sputtering, where you not only deposit a layer, but *etch the target* at the same time.
  - So what you have to do is to somehow produce a plasma of the right kind between some electrode and the wafer to be etched. If all parameters are right, your wafer might get etched the way you want it to happen.
- If we naively compare chemical etching and plasma etching for the same materials to be etched - lets take **SiO_2** - we note major differences:

Chemical etching of **SiO_2**	*Plasma etching* of **SiO_2**
Etchant: **HF + H_2O** (for etching **SiO_2**.	*Gases:* **CF_4 + H_2** (or almost any other gas containing **F**).
Species in solution:: **F^-**, **HF^-**, **$H^+SiO_4^{2-}$**, **SiF_4**, **O_2** - whatever chemical reactions and dissociation produces.	*Species in plasma and on wafer:* **CF_x^+ ($x \leq 3$)**, and all kinds of unstable species not existent in wet chemistry. Carbon based polymers, produced in the plasma which may be deposited on parts of the wafer.
Basic processes: **SiO_2** dissolves	Etching of **SiO_2**, formation of polymers, deposition of polymers (and other stuff) and etching of the deposited stuff, occurs simultaneously
Driving force for reactions: Only "chemistry", i.e. reaction enthalpies or chemical potentials of the possible reactions; essentially equilibrium thermodynamics	*Driving force for reactions:* "Chemistry", i.e. reaction enthalpies or chemical potentials of the possible reactions, including the ones never observed for wet chemistry, near equilibrium, *and* non-equilibrium physical processes", i.e. mechanical ablation of atoms by ions with high energies.
Energy for kinetics: Thermal energy only, i.e. in the **1 eV** range	*Energy for kinetics:* Thermal energy, but also kinetic energy of ions obtained in an electrical field. High energies (several **eV** to hundreds of **eV**) are possible.
Anisotropy: None; except some possible **{hkl}** dependence of the etch rate in crystals.	*Anisotropy:* Two major mechanisms **1.** Ions may have a preferred direction of incidence on the wafer. **2.** Sidewalls may become protected through preferred deposition of e.g. polymers Completely isotropic etching is also possible
Selectivity: Often extremely good	*Selectivity:* Good for the chemical component, rather bad for the physical component of the etching mechanism. Total effect is open to optimization.

- If that looks complicated, if not utterly confusing - that's because it is (and you thought just chemistry by itself is bad enough).
  - Plasma etching still has a strong black art component, even so a lot of sound knowledge has been accumulated during the last **20** years.
  - It exists in countless variants, even for just *one* material.
  - The many degrees of freedom (all kind of gases, pressure and gas flux, plasma production, energy spread of the ions, ...), or more prosaically, the many buttons that you can turn, make process development tedious on the one hand, but allow optimization on the other hand.
- The two perhaps most essential parameters are: **1.** the relative strength of chemical to physical etching, and **2.** the deposition of polymers or other layers on the wafer, preferably on the sidewalls for protection against lateral etching.

- The physical part provides the absolutely necessary anisotropy, but lacks selectivity
- The chemical part provides selectivity.

Polymer deposition, while tricky, is often the key to optimized processes. In our example of **SiO_2** etching, a general finding is:

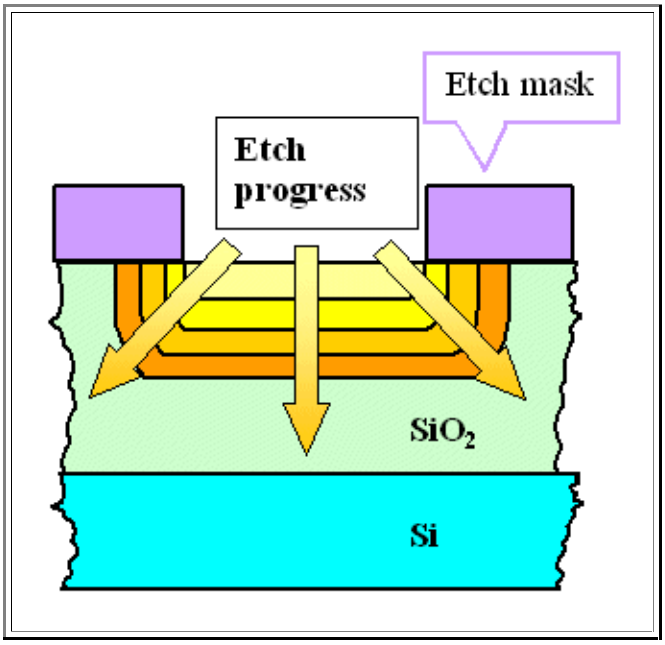
- **Si** and **SiO_2** is etched in this process, but with different etch rates that can be optimized
- The (chemical) etching reaction is always triggered by an energetic ion hitting the substrate (this provides for good anisotropy).
- The tendency to polymer formation scales with the ratio of **F/H** in the plasma. The etching rate increases with increasing **F** concentration; the polymerization rate with increasing **H** concentration.
- Best selectivity is obtained in the border region between etching and polymer formation. This will lead to polymer formation (and then protecting the surface) with **Si**, while **SiO_2** is still etched. The weaker tendency to polymer formation while etching **SiO_2** is due to the oxygen being liberated during **SiO_2** etching which oxidizes carbon to **CO_2** and thus partially removes the necessary atoms for polymerization

Enough about plasma etching. You get the idea.

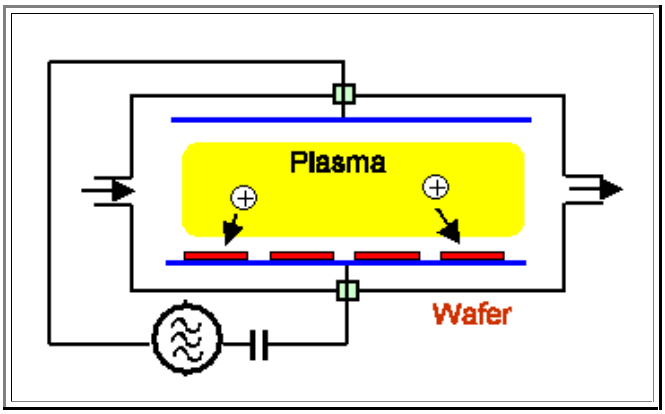
- A taste treat of what it really implies can be found in an advanced module.

6.5.3 Summary to: Etching Techniques

- Chemical etching:
  Relatively simple, with luck extremely selective, relatively fast, but:
  - Material removal always isotropic (see figure) and thus enlarging features and thus not applicable to "micro" or better "nano" structuring
  - Might be extremely dangerous (corrosive, poisonous, inflammable, explosive, ...) and often hard to recycle / dispose off.



- Plasma Etching:
  Simple in principle (see figure) quite complex and expensive in praxis.
  - Given the right plasma (not easy to find; plasma chemistry counts among the "black arts"), the etching can be completely anisotropic and sufficiently selective.
  - Innumerable variants exist and new methods are coming up all the time. Definitely on of the key technologies for micro / nano technology.

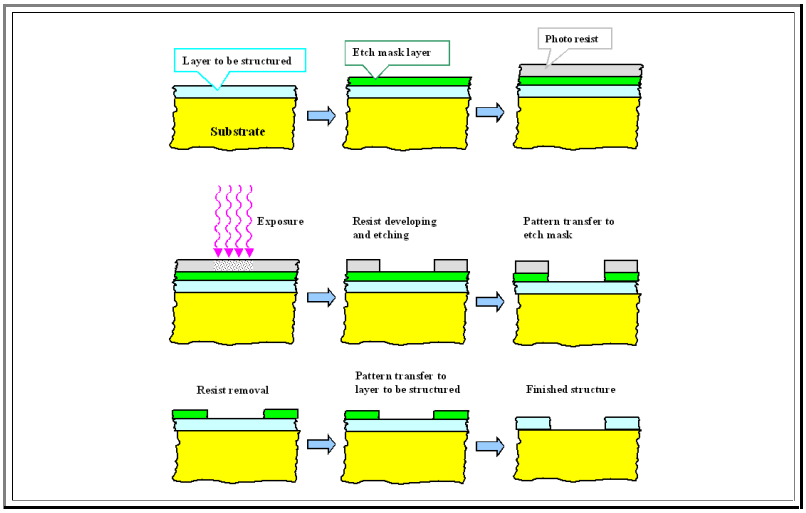


6.6 Lithography

6.6.1 Basic Lithography Techniques

Process flow of Lithography (and Pattern Transfer)

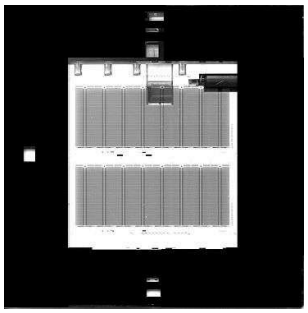
- Lets start by considering the basic processes for the complete structuring module.
  - Shown is a more complex process flow with a special etch mask layer (usually **SiO_2**).
  - Often, however, you just use the *photo resist as masking layer for etching*, omitting deposition, structuring, and removal of the green layer. A photo resist mask generally is good enough for ion implantation (provided you keep the wafer cool) and many plasma etching processes.



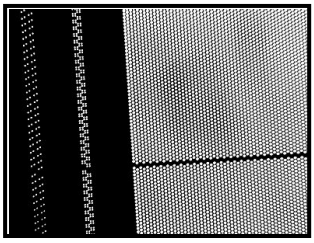
- As far as lithography is concerned, it is evident that we need the following key ingredients:
  - A **photo resist** [1)], i.e. some light sensitive material, not unlike the coating on photographic film.
  - A **mask** (better known as **reticle** [2)]) that contains the structure you want to transfer - not unlike a slide.
  - A **lithography unit** that allows to project the pattern on the mask to the resist on the wafer. Pattern **No. x** must be *perfectly aligned* to pattern **No. x - 1**, of course. Since about **1990** one (or just a few) chips are exposed at one time, and than the wafer is moved and the next chip is exposed. This step-by-step exposure is done in machines universally known as **steppers**.
  - Means to **develop** and structure the resist. This is usually done in such a way that the *exposed* areas can be removed by some etching process (using **positive resist**). For some special purpose, you may also use **negative resists**, i.e. you remove the *unexposed* areas.
- In principle, it is like projecting a slide on some photosensitive paper with some special development.
  - However, we have some very special requirements. And those requirements make the whole process very complex!
  - And with *very complex* I mean *really* complex, super-mega-complex - even in your wildest dreams you won't even get close to imagining what it needs to do the lithography part of a modern chip with structures size around **0,13 µm**.
- But relax. We are not going to delve very deep into the intricacies of lithography, even though there are some advanced material issues involved, but only give it a cursory glance.

Reticles

- For any layer that needs to be structured, you need a ***reticle***. Since the projection on the chip usually reduces everything on the reticle fivefold, the reticle size can be about **5** times the chip size
  - A reticle then is a glass plate with the desired structure etched into a **Cr** layer. Below, a direct scan of an old reticle is shown, together with a microscope through-light image of some part.
  - "Obviously", the regular lattice of small opening in the non-transparent **Cr** layer is the array for the trenches in a memory chip. The smallest structures on this reticle are about **5 µm**.

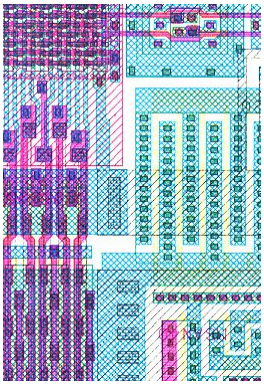


Typical reticle, about original size



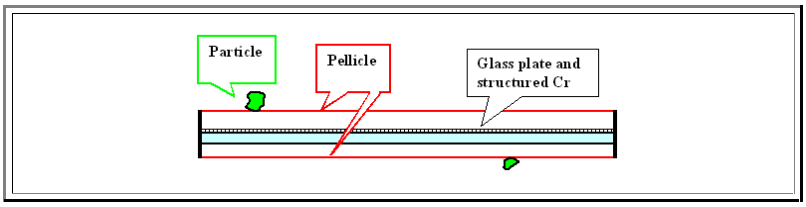
Enlargement (**x 100**)

- Before we look at the requirements of reticles and their manufacture, lets pause for a moment and consider how the structure on the reticle comes into being.



- First, lets look at these structure, or the **lay-out** of the chip. Shown on the left is a tiny portion of a **4 Mbit DRAM**.
- Every color expresses *one* structured layer (and not all layers of the chip are shown).
- A print-out of the complete layout at this scale would easily cover a soccer field.
- The thing to note is: it is *not* good enough to transfer the structure on the reticle to the chip with a resolution *somewhat better* than the smallest structures on the chip, it is also necessary to superimpose the various levels with an **alignment accuracy** *much better* than the smallest structure on the chip!
- And remember: We have about **20** structuring cycles and thus reticles for one chip.

- The *lay-out* contains the *function* of the chip. It established where you have transistors and capacitors, how they are connected, how much current they can carry, and so on.
  - This is determined and done by the product people - electrical engineers, computer scientists - no materials scientists are involved.
  - The *technology*, the making of the chip, determines the *performance* - speed, power consumption, and so on. This is where material scientists come into their own, together with semiconductor physicists and specialized electrical engineers (who e.g., can simulate the behavior of an actual transistor and thus can tell the process engineers parameters like optimal doping levels etc.).
  - In other words, the reticles are the primary input of the product engineers to chip manufacturing. But they only may contain structures that can actually be made. This is expressed in **design rules** which come from the production line and must be strictly adhered to. Only if *all* engineers involved have some understanding of *all* issues relevant to chip production, will you be able to come up with aggressive and thus competitive design rules!
- What are the requirements that reticles have to meet (besides that their structures must not contain mistakes from the layout. e.g. a forgotten connection or whatever).
  - Simple: They must be *absolutely* free of defects *and* must remain so while used in production! Any defect on the reticle will become transferred to *every* chip and more likely than not will simply kill it.
  - In other words: Not a single particle is ever allowed on a reticle!
- This sounds like an impossible request. Consider that a given reticle during its useful production life will be put into a stepper and taken out again a few thousand times, and that every mechanical movement tends to generate particles.
  - Lithography is full of "impossible" demands like this. Sometimes there is a simple solution, sometimes there isn't. In this case there is:
  - First. make sure that the freshly produced reticle is defect free (you must actually check it pixel by pixel *and* repair unavoidable production defects).
  - Then encase it in **pellicles** [3] (**=** fully transparent thin films) with a distance of some mm between reticle and pellicle as shown below.



- One of the bigger problems with steppers - their very small (about **1 µm**) *depth of focus* - now turns to our advantage: Unavoidable particles fall on the pellicles and will only be imaged as harmless faint blurs.

How do we make reticles?

- By writing them pixel by pixel with a finely focussed electron beam into a suitable sensitive layer , i.e. by direct writing **electron-beam lithography**.
- Next, this layer is developed and the structure transferred to the **Cr** layer.
- Checking for defects, repairing these defects (using the electron beam to burn off unwanted **Cr**, or to deposit some in a kind of e-beam triggered **CVD** process where it is missing), and encasing the reticle in pellicles, finishes the process.

Given the very large pixel size of a reticle (roughly $\mathbf{10^{10}}$), *this takes time* - several hours just for the electron beam writing!

- This explains immediately why we don't use electron beam writing for directly creating structures on the chip: You have at most a few seconds to "do" one chip in the factory, and e-beam writing just can`t deliver this kind of throughput.
- It also gives you a vague idea why reticles don't come cheap. You have to pay some **5000 $ - 10 000 $** for *making* one reticle (making the lay-out is not included!). And you need a set of about 20 reticles for one chip. And you need lots of reticle sets during the development phase, because you constantly want to improve the design. You simply need large amounts of money.

---

1) Something as a protective coating that resists or prevents a particular action (Webster, second meaning)

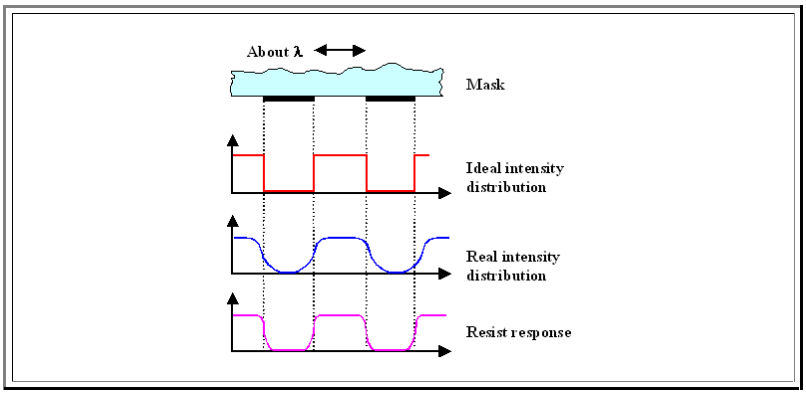
2) A system of lines, dots, cross hairs, or wires in the focus of the eyepiece of an optical instrument (Webster)

3) A thin skin or film, especially for optical uses

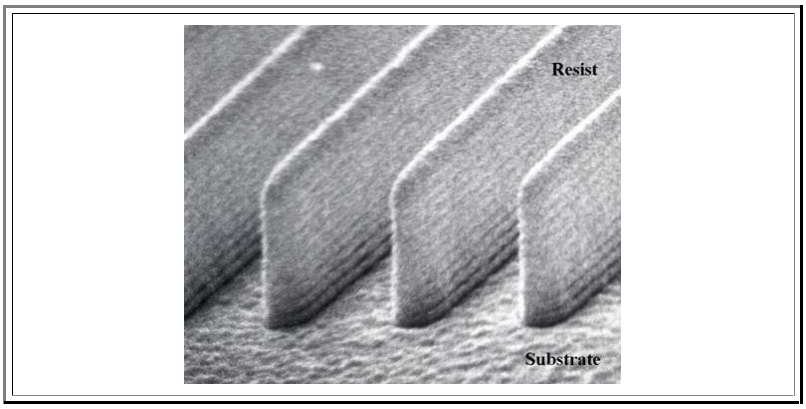
6.6.2 Resist and Steppers

Photo Resists

- Lets just look at a list of requirements for resists. We need to have:
  - *High sensitivity* to the wavelength used for imaging, but not for all optical wave lengths (you neither want to work in the dark, nor expose the resist during optical alignment of the reticle which might be done with light at some other wave length). Not easy to achieve for the short wave lengths employed today.
  - *High contrast*, i.e. little response (= "blackening") to intensities below some level, and strong response to large intensities. This is needed to sharpen edges since diffraction effects do not allow sharp intensity variations at dimensions around the wavelength of the light as illustrated below.



  - *Compatibility* with general semiconductor requirements (easy to deposit, to structure, to etch off; no elements involved with the potential to contaminate **Si** as e.g. heavy metals or alkali metals (this includes the developer), no particle production, and so on).
  - *Homogeneous "blackening" with depth* - this means little absorption. Simply imagine that the resist is strongly absorbing, which would mean only its top part becomes exposed. Removal of the "blackened" and developed resist than would not even open a complete hole to the layer below.
  - No *reflection of light*, especially at the interface resist - substrate. Otherwise we encounter all kinds of interference effects between the light going down and the one coming up (known as "**Newton fringes**"). Given the highly monochromatic and coherent nature of the light used for lithography, it is fairly easy to even produce *standing* light waves in the resist layer as shown below. While the ripple structure clearly visible in the resist is not so detrimental in this example, very bad things can happen if the substrate below the resist is not perfectly flat.



  - This would call for a *strongly absorbing* resist - in direct contradiction to the requirement stated above. Alternatively, an **anti-reflection coating** (***ARC***) might be used between substrate and resist, adding process complexity and cost.
  - Suitablity of the resist as *direct mask* for ion-implantation or for plasma etching.
  - Easy *stripping* of the resist, even after it was turned into a tough polymer or carbonized by high-energy ion bombardment. Try to remove the polymer that formed in your oven from some harmless organic stuff like plum cake after it was carbonized by some mild heat treatment without damaging the substrate, and you know what this means.
- Enough requirements to occupy large numbers of highly qualified people in resist development!
  - Simply accept that resist technology will account for the last **0,2 µm** or so in minimum structure size. And if you do not have the state-of-the-art in resist technology, you will be a year or two behind the competition - which means you are loosing *large amounts of money*!

Stepper

- A stepper is a kind of fancy slide projector. It projects the "picture" on the reticle onto the resist-coated wafer. But in contrast to a normal slide projector, it does not *enlarge* the picture, but *demagnifies* it - exactly fivefold in most cases. Simple in principle, ***however:***
- **1.** We need the *ultimate in optical resolution*!
  - As everybody knows, the **resolution limit** of optical instruments is equal to about the wave-length λ. More precisely and quantitatively we have

$$d_{min} \approx \frac{\lambda}{2NA}$$

  - With $\mathbf{d_{min}}$ **= minimal distinguishable feature size**; i.e the distance between two **Al** lines, and ***NA*** = **numerical aperture** of the optical system (the ***NA*** for a single lens is roughly the quotient of diameter / focal length; i.e. a crude measure of the size of the lens).
  - Blue light has a wave length of about **0.4 µm**, and the numerical apertures ***NA*** of very good lenses are principally **< 1**; a value of **0.6** ~~**is**~~ was about the best you can do (consider that all distortions and aberrations troubling optical lenses become more severe with increasing ***NA***). This would give us a minimum feature size of

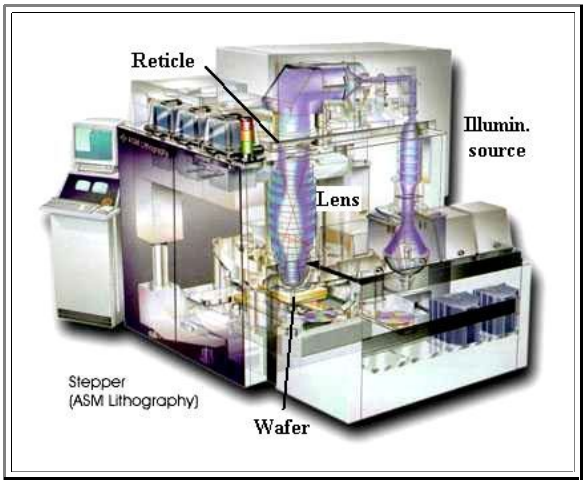
$$d_{min} \approx \frac{0.4}{1.2} = 0.33\ \mu m$$

  - Since nowadays you can buy chips with minimum features of **0.18 µm** or even **0.13 µm**; we obviously must do better than to use just the visible part of the spectrum.
- **2.** Resolution is not everything, we need some *depth of focus*, too. Our substrate is not perfectly flat; there is some topography (not to mention that the **Si** wafer is also not perfectly flat).
  - As anyone familiar with a camera knows, your depth of focus Δf *decreases*, if you *increase* the aperture diameter, i.e. if you *increase* the ***NA*** of the lens. In formulas we have

$$\Delta f \approx \frac{\lambda}{(NA)^2} = \frac{0.4}{0.6^2} = 1.11\ \mu m$$

  - Tough! What you gain in resolution with larger numerical apertures, you loose (quadratically) in focus depth. And if you decrease the wavelength to gain resolution, you loose focus depth, too!
- **3.** We need to *align* one exposure exactly on top of the preceding one. In other words, we need a wafer stage that can move the wafer around with a precision of lets say **1/5 of** $\boldsymbol{d_{min}}$ - corresponding to **0.18/5 µm = 0.036 µm =** ***36 nm***.
  - And somehow you have to control the stage movement; i.e. you must measure where you are with respect to some alignment marks on the chip *with the same kind of precision*. We need some **alignment module** in the stepper.
  - Alignment is done optically, too, as an integral (and supremely important) part of stepper technology. We will, however, not delve into details.
- **4.** We need to do it *fast*, *reliable* and *reproducible* - **10 000** and more exposures a day in one stepper.
  - Time is money! You can't afford more than a few seconds exposure time per chip.
  - And you also can not afford that the machine breaks down frequently, or needs frequent alignments. Therefore you will put your stepper into separate temperature and humidity controlled enclosures, because the constancy oft these parameters in the cleanroom (Δ**T** ≈ **1 °C**) is not good enough. You also would need to keep the atmospheric pressure constant, but ingenious engineers provided mechanism in the lens which compensates for pressure variations of a few **mbar**; still, when you order a stepper, you specify your altitude and average atmospheric pressure).

- How do we built a stepper? By combining elements from the very edge of technology in a big machine that costs around **10.000.000.000 $** and that can only be produced by a few specialized companies.
  - The picture below gives an impression. The basic imaging lens of the stepper is a huge assembly of many lenses; about **1 m** in length and **300 kg** in weight.



  - We need intense monochromatic light with a short wave length. If you use colored light, there is no way to overcome the chromatic aberrations inherent in all lenses and your resolution will suffer.
  - The wave lengths employed started with the so-called **g**-line (**436 nm**) of **Hg**, fairly intense in a **Hg** high pressure arc lamp and in the deep blue of the spectrum. It was good down to about **0.4 µm** as shown in the example above.
  - Next (around **1990**) came the **365 nm i**-line in the near ultra violet (***UV***). This took us down to about **0.3 µm**.
  - Next came *a problem*. There simply is no "light bulb" that emits enough intensity at wavelengths considerably smaller than **365 nm**. The (very expensive) solution were so-called *excimer Lasers*, first at **248 nm** (called deep **UV** lithography), and eventually (sort of around right now (**2001**)), at **194 nm** and **157 nm**.
  - Next comes *the end*. At least of "conventional" stepper technology employing lenses: There simply is no material with a sizeable index of refraction at wavelengths considerably below **157 nm** that can be turned into a high-quality lens. Presently, lots of people worry about using single crystals of **CaF_2** for making lenses for the **157 nm** stepper generation.
- What do you do then? First you raise obscenely *large amounts of money*, and than you work on alternatives, most notably
  - **Electron-beam lithography**. We encountered it before; the only problem is to make it much, much faster. As it appears today (Aug. **2001**), this is not possible.
  - **Ion beam lithography**. Whatever it is, nobody now would bet much money on it.
  - **X-ray lithography**. Large-scale efforts to use **X**-rays for lithography were already started in the eighties of the **20** century (involving huge electron synchrotons as a kind of light bulb for intense **X**-rays), but it appears that it is pretty dead by now.
  - **Extreme UV lithograpy** at a wave length around **10 nm**. This is actually soft **X**-ray technology, but the word "**X**-ray lithography" is loaded with negative emotions by now and thus avoided. Since we have no lenses, we use mirrors. Sounds simple - but have you ever heard of mirrors for **X**-rays? Wonder why not? This is what the US and the major US companies favor at present.
- Well, lets stop here. Some more advanced information can be found in the link.
  - *But note*: There are quite involved materials issues encountered in lithography in general, and in making advanced steppers in particular. **CaF_2** is an electronic material! And the success - or failure - of the global enterprise to push minimum feature size of chips beyond the **100 nm** level, will most likely influence *your* professional life in a profound matter.
  - This is so because the eventual break down of Moores law will influence in a major way *everything* that is even remotely tied to technology. And what will happen is quite simply a question if we (including you) succeed in moving lithography across the **100 nm** barrier.