9.2.2 Laser Diodes

Stimulated Emission and Inversion

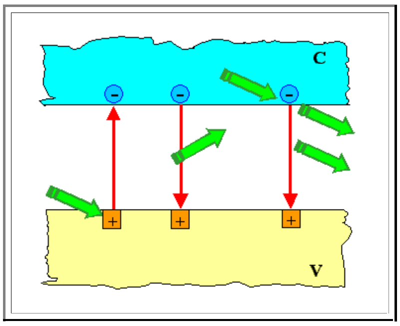
In principle, anything that emits electromagnetic radiation can be turned into a "**LASER**" - but what *is* a Laser?

- The word "*Laser*" was (and of course still is) an ***acronym***, it stands for "**L**ight **A**mplification by **S**timulated **E**mission of **R**adiation" By now, however, it is generally perceived as a standard word *in any language* meaning something that is more than the acronym suggests (and we will no longer write it with capital letters)!
- A Laser in the direct meaning of the acronym is a black box that emits (=outputs) more light of the same frequency than what you shine (=input) on it - that is the *amplifier* part. But the "*stimulated emission*" part, besides being the reason for amplification, has a second indirect meaning, too: The light emitted is exactly in phase or **phase coherent** (or simply *coherent*) to the light in the input. Coherent then also means that the waves are all traveling in exactly the same direction.
- Unfortunately, Lasers in this broad sense do not really exist. *Real*Lasers only amplify light with a very specific frequency i.e. **monochromatic** light - its like electronic amplifiers for *one frequency only*.

A Laser in the *general* meaning of the acronym thus produces intense monochromatic electromagnetic radiation in the wavelength region of light (including infrared and a little ultra violet; there are no sharp definitions) that it is coherent to the (monochromatic) input. If you "input" light containing all kinds of frequencies, only *one* frequency becomes amplified.

- A Laser in the *specific* meaning of everyday usage of the word, however, is more special. It is a device that produces a coherent beam (=only in direction) of monochromatic light and, at least for semiconductor Lasers, *without some input light* (but with a "battery" or power source hooked up to it). It is akin to an electronic oscillator that works by internally **feeding back** parts of the output of an amplifier to the input for a certain frequency only.
- Before the advent of hardware Lasers in the sixties, there were already "**Masers**" - just take the "**M**" for "microwave" and you know what it is.
- And even before that, there was the basic insight or idea behind Masers and Lasers and, as ever so often, it was **A. Einstein** who described the "**S**timulated **E**mission part in **1917/1924**. More to the history of Lasers can be found in the link.

Obviously, for understanding Lasers, we have to consider *stimulated emission* first, and then we must look at some *feedback* mechanism.



Understanding stimulated emission is relatively easy; all we have to do is to introduce one more process for the interaction between light and electrons. So far we considered two basic processes, to which we will now give their proper names:

- **1. Fundamental absorption**, i.e. the interaction of a photon with an *electron* in the *valence band* resulting in a electron - hole pair. This kind of absorption is fundamental because that is the way light is always fundamentally absorbed - not just in semicondcutors but in all materials: an electron is kicked to a higher energy levely a photon that then ceases to exist, i.e. gets absorbed.
- **2. Spontaneous emission** of a photon by the recombination of an electron-hole pair via direct band-band recombination. The word "spontaneous" refers to a certain stochastic component of this process. This is just the normal radiative recombination we have discussed before and, in a more general sense, the process that *always* generates light, e.g. in "excited" atoms or molecules.

This is clear enough and we have learned a lot of details about these processes without delving too deeply into quantum mechanics or quantum optics. This is no longer true for the third process concerning the interaction between light and electrons and holes - *stimulated emission*. So let's just describe it:

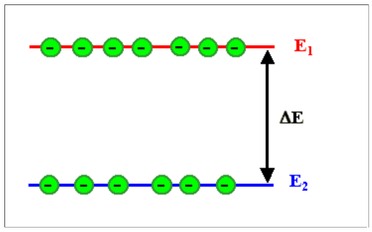
- **3. Stimulated emission**is simply a special kind of interaction of a photon with an electron on an "excited" energy level (=electron in the *conduction band* for semiconductors). A photon that happens to "fly by" then *stimulates* the recombination and thus the emission of a *second photon*. The way this happens is always such that the emitted photon is "identical" to the photon stimulating its generation - same energy, same direction, same phase.

- All three processes are schematically shown in the band diagram on the right. Stimulated emission, in a way of speaking, just takes the randomness out of the spontaneous emission.
- Looking at this picture, you should wonder why one obvious process is missing? How about an electron in the conduction band simply absorbing a photon and getting kicked up the energy scale in the conduction band?
  - In other words: An electron in the conduction band absorbs a photon, moves up the amount **h · ν** in the conduction band, and comes back to the band edge by tranferring its surplus energy to phonons, i.e. heating the lattice.
  - The answer is simple: this process does take place, but is not very strong if we do not have many electrons in the conduction band. It is also not necessary for "lasing", but rather detrimental and we will not consider it any more.
- Stimulated emission is a *resonant process*. It only works if the photons have exactly the right energy and momentum (=wave vector), corresponding to the energy that is released if the electron makes a transition to some allowed lower level. Of course, you have the usual quantum mechanical paradox that the incoming photon is in resonance with something it has not yet created; it is just, so to speak, probing possible outcomes of possible processes.
  - From the resonance argument it follows that two photons have the same wavelenth and are *exactly in phase* with each other. For semiconductors, the photon energy must be pretty much the energy of the band gap, because all conduction band electrons are sitting at the conduction band edge (with some small ΔE, of course), and the only available lower energy level are the free positions occupied by holes at the valence band edge, restricting photons that can be released (and be in resonance with incoming ones) to $E_{photon} = h\nu \approx E_g$.
  - Stimulated emission thus may be seen as a competing process to the fundamental band-band absorption process described above. But while *all* photons with an energy $h\nu > E_g$ may cause fundamental absorption because there are many unoccupied levels above E_g, *only* photons with $h\nu = E_g$ (give or take some small ΔE) may cause stimulated emission.
- Einstein showed that under "normal" conditions (meaning conditions not too far from thermal equilibrium), *fundamental absorption by far exceeds stimulated emission*. Of course, Einstein did not show that for semiconductors, but for systems with well defined energy levels - atoms, molecules, whatever.
  - However, for the special case that *more* electrons occupy an excited energy state than the related ground state - this condition we will call **inversion** - stimulated emission may dominate the electron-photon interaction processes. Then *two* photons of identical energy and being exactly in phase come out of the system for *one* photon going into the system.
  - The kind of *inversion* we are discussing here should not be mixed up with the *inversion* that turns **n**-type **Si** into **p**-type or vice versa that we encountered before. Same word, but different phenomena!
  - These two photons may cause more stimulated emission - yielding **4**, **8**, **16**, ... photons, i.e. an avalanche of photons will be produced until the excited electron states are sufficiently depopulated.
  - In other words: One photon **hν** impinging on a material that is in a state of *inversion* (with the right energy difference **hν** between the excited state and the ground state) may, by stimulated emission, cause a lot of photons to come out of the material. Moreover, these photons are all in phase (and thus travelling in the same direction), i.e. we have now a strong and coherent beam of light as output.
- We are now stuck with two basic questions:
  - **1.** What exactly do we mean with "inversion", particularly with respect to semiconductors?
  - **2.** How do we induce a state of "inversion" in semiconductors?
- Let's look at these questions separately

Obtaining Inversion in Semiconductors

- If you shine **10** input photons on a crystal, **6** of which disappear by fundamental absorption, leaving **4** for stimulated emission, you now have **8** output photons. In the next round you have **2 · (8 · 0,4)=6,4** and pretty soon you have none.
  - Now, if you reverse the fractions, you will get **12** photons in the first round, **2 · (12 · 0,6)=14,4** the next round - you get the idea.
- In other words, the *coherent* amplification of the input light only occurs for a *specific condition*:
  - There must be *more* stimulated emission processes than fundamental absorption processes if we shine light with $E = h\nu = E_g$ on a direct semiconductor - this condition defines "*inversion*" in the sense that we are going to use it.
  - We only look at *direct* semiconductors, because radiant recombination is always unlikely in indirect semiconductors, and while stimulated emission is generally possible, it also needs to be assisted by phonons and thus is unlikely, too.
- Lets first consider some basic situations for defining *inversion* in full generality.For the most simple system, we might have two energy levels E_1 and E_2, with the lower one (E_1) mostly occupied by electrons, the upper one (E_2) relatively empty. E_2 could be highest energy level still occupied in some atom or molecule, or whatever.

- *Inversion* then means that the number of electrons on the upper level, n_2, is *larger* or at least equal to n_1.
- In *equilibrium*, however, we would simply have



Two level system in inversion condition

$$\frac{n_1}{n_2} = \frac{D_1}{D_2} \cdot \exp - \frac{\Delta E}{kT}$$

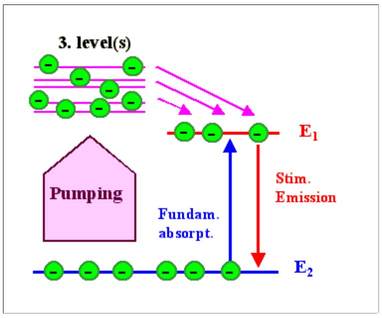
- With $\Delta E = E_1 - E_2$, and $D_{1,2}$= the maximum number of electrons allowed on $E_{1,2}$ (the "density of states").
- In words: In equilibrium we have far more electrons at E_2 than at E_1.

For inversion to occur, we must be *very far from equilibrium* if ΔE is on the order of **1 eV** as needed for visible light. In fact, the systems would have to have a *negative temperature* for such a distribution if we keep the concept of "temperature" that is only properly defined in equilibrium. This is something you should figure out by yourself.

- If we now shine some light on our two-levels system that we brought into inversion, stimulated emission would quickly depopulate the E_2 levels, while fundamental absorption would kick some electrons back. Nevertheless, after some (short) time we would be back to equilibrium having produced a short Laser light pulse at best.
- To keep stimulated emission going, we must move electrons from E_1 to E_2 *by some outside energy source* all the tine - we must "**pump**" the system. Doing this with some other light source providing photons of the only usable energy $\Delta E = E_1 - \Delta E = E_2$ would defeat the purpose of the game; after all that is just the light we want to generate. In semiconductors now, we could inject electrons from some other part of the device but a two-level system is not a semiconductor, so that is not possible here.
- *In short*: Two level systems are no good for practical uses of stimulated emission.

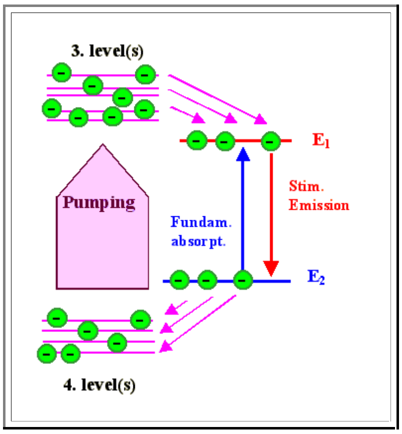
What we need is an *easy* way to move a lot of electrons to the energy E_2. This can be achieved in a **three level system** as shown below (and this was the way it was done with the first ruby Laser).

- The essential trick is to have a whole system of levels - ideally a band - *above* E_2, from which the electrons can descend efficiently to our single level E_1 but not easily back to E_2 where they came from. Schematically, this looks like in the figure on the right.
- The advantage is obvious. We now can take light with a whole range of energies - always larger than ΔE- to "*pump*" electrons up to E_2 via the reservoir provided by the third level(s).
- The only disadvantage is that we have to take the electrons from E_1. And no matter how hard we **pump** (this is the word used for this process), the probability that a quantum of the energy we pour into the system by pumping will find an electron to act upon, will always be proportional to the number (or density) of electrons available for kicking up to E_2. In the three level system this is still at most D_1. If we sustain the inversion, it is at most **0,5** · D_1, because by definition we have at least one-half of the available electrons already on E_2.

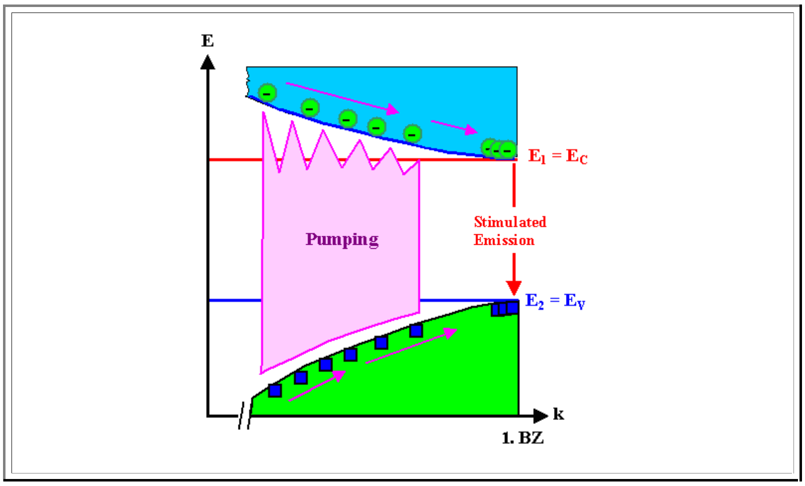


It is clear what we have to do: Provide a *fourth energy level* (even better: a band of energy levels) *below* E_1, where you have a lot of electrons that can be kicked up to E_2 via the third level(s). It is clear that we are talking semiconductors now, but lets first see the basic system:

- We simply introduce a system of energy states below E_1 in the picture from above. We now have a large reservoir to pump from, and a large reservoir to pump to.
- All we have to do is to make sure that pumping is a one-way road, i.e. that there are no (or very few) transitions from the levels **3** to levels **4**.
- This is not so easy to achieve with atoms or molecules, but, as you should have perceived by now, this is exactly the situation that we have in many direct band gap semiconductors. All we have to do to see this, is to redraw the **4**- level diagram at the right as a band diagram. To include additional information, we do this in *k-space*.



We have the following general situation for producing inversion in direct semiconductors:



Electrons may be pumped up from anywhere in the valence band to anywhere in the conduction band **-** always provided the transition goes vertically upwards in the reduced band diagram.

- The electrons in the conduction band as well as the holes in the valence band will quickly move to the extrema of the bands **-** corresponding to the levels E_2 and E_1 in the general four level system.
- "*Quickly*" means within a time scale of typicall **10^{-13} s**. This time scale is so small indeed that it introduces some uncertainties in the energies via the **uncertainty relation** but that need not bother us here.

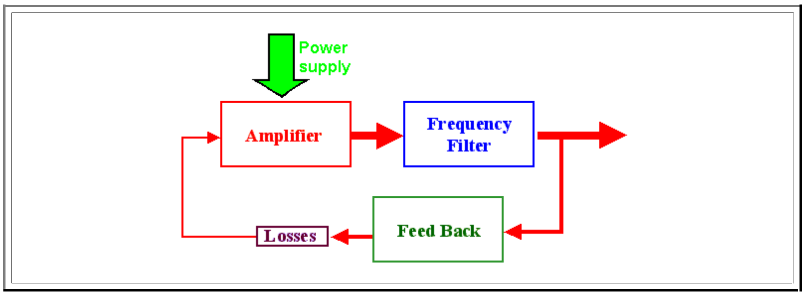
So all we have to do for a semiconductor Laser made from a fancy **LED** is to take the hetero junction configuration introduced before and run enough electrons and holes into the recombination zone defined by the small band gap semiconductor.

- This will raise the concentration of electrons at the conduction band edge of the small energy gap semiconductor and lower the concentration of electrons at the valence band edge (by rasing the concetrations of holes). If the current is large enough, we will always have more electrons at the conduction band edge than at the valence band edge, and this is what we called inversion above.
- In other words: run enough current into your **LED** (without heating it up too much!) and you achieve inversion and thus meet what is called the "**first Laser condition**".
- In orher words: We now can *amplify* light. All we need to do now for or making a Laser is to provide some feed back mechansims in the sense mentioned above.

A General Look at Feedback and Oscillations

So far we have the the amplification of light by stimulated emission. Making a Laser in the conventional sense of the word still requires to produce a light beam with a "battery" and withoutsome "input light".

- This is the same task as to produce an *oscillator* from an electronic amplifier, and the solution of this task for a Laser is achieved in the saem basic way.
- **Feed back** *one* frequency from the output of the amplifier to the input, and make sure it is in phase (or as we say for light, "coherent"). This frequency will be amplified, the feed back increases, it will become more amplified, ..., pretty soon your system is now an oscillator for the frequency chosen. You just need to feed back a large enough part of the output to account for losses that may be occurring in the feed back loop. The essential parts are shown in the drawing



- If you think about this, you will discover a problem. If there is enough signal at the input, the output will go up forever or until a fuse blows - there is no stability in the system
- We need some kind of servo mechanism that adjusts the amplification factor to a value where only the losses are recovered by amplification, so that a stable, preferably adjustable output amplitude is obtained.

This is clear enough for electrical signals, but how do we do this with light? Well, we do everything with **mirrors**:

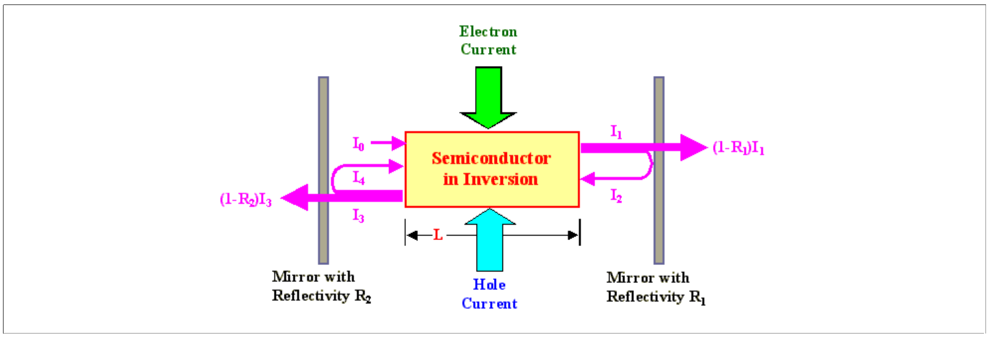
- **1**. The **feed back** part.
- **2.**The **coherency** requirement.
- **3.** The **selection of the frequencies**.
- **4.**The **guidance of the light** including the "*beam shaping*" of the output.

The **4th** point is new - after all, electrical signals go to wherever the wires go, but with light we have to make sure we get a *single* beam coming out in the right direction. We will only look at the general principle of light feedback without worrying much about the three other points other points..

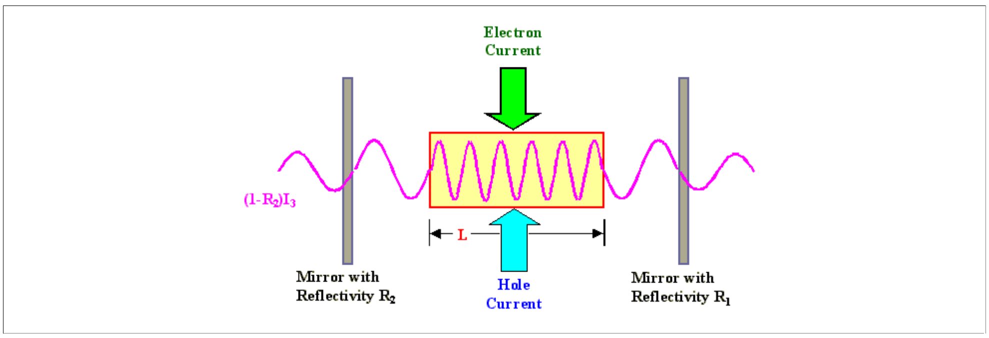
Feedback with Mirrors

All we have to do is to put the piece of semiconductor that is supposed to amplify the light by stimulated emission between two *partially transparent* mirrors

- The whole system then looks like this:



Principle of optical feed back



Formation of a standing wave as a result

Assuming some amplification, i.e. $I_1 = a \cdot {}_0$ with $a > 0$, and for reasons of symmetry, $I_3 = a \cdot {}_2$, we could now start some calculations.

- We won't, however, because all we could get at that level of simplicity is that the system is either quiet (too little feed back) or "explodes". What will really happen (if no meltdown occurs or a fuse blows first) is is that with increasing output intensity some losses will go up, too and the amplification factor ***a*** comes down until some balance is achieved.

- We have feed back now, but how do we achieve monochromatic light (i.e. what provides the filter) and coherency, not to mention the guidance of the light; i.e. the other points mentioned above?
  - Well, we have already started that we do that with mirrors, too - and we even use the same mirrors we used for the feed back.
- All we have to do is to chose the (optical) distance between the mirrors to be a *multiple of half the wave length* we want. Only in this case a **standing wave** in between the mirrors can be formed. Waves with wavelengths other than the "fitting "ones would simply cancel by interference.
  - This is best seen by looking at what would happen to light with a "wrong" wavelength. Every time it travels through the system, its phase is shifted to some extent, and pretty soon you have a wave with the phase $-\Phi$ for any wave with the phase Φ and destructive interference will cancel everything except waves with phases that fit.
  - This is the same old principle that governs diffraction of electrons or **X**-ray beams in crystals, all musical instruments, and, if you believe Richard **Feynman**, just about everything else, too.

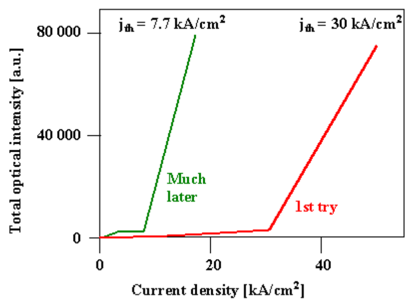
Making a Simple Laser Diode

- If we are not too particular about certain aspects of our semiconductor Laser, we do not even *need* external mirrors. We may now use the effect that annoyed as in the context of getting light out of a semiconductor: that the surfaces of the crystal already act as a partially transparent mirror (with a reflectivity of roughly **30 %** for perpendicular incidence)..
  - The surface mirrors will even be extremely plan-parallel (which is important for obvious reasons) if we obtain them by ***cleaving*** the crystals because with a bit of care they will fracture on the same lattice plane type (usually **{110}** or **{111}**); you can't do much better than this.
- The length crystal ***L***=distance between the "mirrors" then determines the allowed wavelengths as follows:

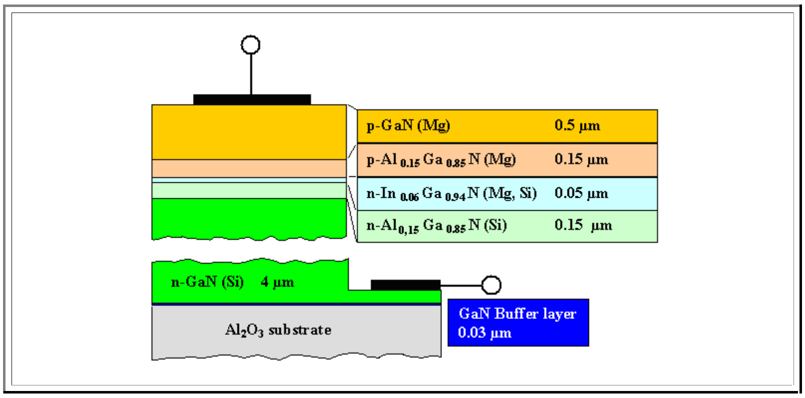
$$L = \frac{m \cdot \lambda_{air}}{2} = \frac{m \cdot \lambda}{2n_{ref}} \qquad m = 1, 2, 3, 4, \ldots,$$

$$\lambda = \frac{2 \cdot L \cdot n_{ref}}{m}$$

  - With λ_{air}= wave length in air, λ= wave length in the crystal, n_{ref}= refractive index of the crystal.
- However, with a typical Laser diode that is not extremely small, ***L*** is far larger than the the wavelength λ of visible light and we thus have many allowed wavelengths at this point. The question is how we make sure that we amplify just *one* of the many allowed ones.
  - Easy. Amplification depends on stimulated emission, and stimulated emission *only*works for $h \cdot \nu_{ref} = h \cdot c/\lambda = E_g$. From all the allowed waves that can live within our *Fabry-Perot resonator*, only the one with the proper energy will actually occur.
- If we have made a Laser and now turn it on we will typically encounter the following results:



  - We take the device and run current into it. For low current densities, we have a **LED** producing some light in all directions. Increasing the current produces more light and at some threshold current density j_{th} the losses in the optical feed back provided by our surface mirrors are overcome and lasing starts. We will notice that because now the total amount of light produced goes up sharply with thecurrent density and we also get the quitessential Laser beam
  - If we are lucky and get lasing at all, we probably will encounter a high $j_{th}$and our Laser will cease to operate after a few minutes or hours because all kinds of tricky processes (not always well understood) will kill it.
- Our "simple" Laser may have looked like that (after the basic inventor and investigator Nakamura)



- Now it's time for serious work. Later - and with far more complicated strucures than the one shown above, you may get the second curve in the diagram above with far lower j_{th} and hopefully a Laser life time of many years. Now you may start mass production, ebéventuyll selling your blue Laser diode for a few **€** a piece.
- All in all, we see that the claim from before is correct: the theory of (semiconductor) Lasers is rather complex, but the technology is not. However, that is only true as long as you don't care all that much about the quality of your Laser. In other words as long as you make mostly cheap low-power Lasers or exactly what is needed the mass market for **CD's**, **CVD's**, and watever comes after that.
  - If you think a minute about that: only the advent of **CD's** made progress in **PC's**and Laptops possible, i.e enabled microelectronics, the hard core of semicondutor technology to go on. **CD's** are unthinkable without cheap small Lasers with little power consumption, so semiconductor technolog*ies* are all intertwined by no and move on together, rapidly gaining momentum and complexity.