Lecture 1   24 febbraio

How were NPN devices build in the 1950's? → growth of N doped silicon, change the doping during the growth to P, back to N and then slice the wafer.

The planar process was one ~~two~~ of the processes that really changed the whole industry.

thin film deposition → photoresist application → exposure → development → etching → resist removal

deposited film, substrate, photoresist, light, etch mask

The IRDS (International Roadmap for Devices and Systems) has stipulated a roadmap for the different ages of scaling:

1. Geometrical scaling (1975 - 2002)
   reduction of horizontal and vertical physical dimensions in conjunction with improved performance of planar transistors

2. Equivalent scaling (2003 ~ 2024)
   reduction of only horizontal dimensions in conjunction with introduction of new materials and new physical effects. New vertical structures replace the planar transistors

3. 3D Power scaling (2025 ~ 2040)
   transition to complete vertical device structures. Heterogeneous integration in conjunction with reduced power consumption become the technology drivers

Semiconductors: let's take Si as a representative

- each atom is covalently bound to four neighbours
- thermal energy can break some of the bonds to generate free electrons (and holes)
- intrinsic carrier concentration varies as a function of T.

At room T Si is a very poor conductor.

(plot: intrinsic carrier concentration vs T, curves for GaAs and Si)

(diagram: E axis; conduction band, energy gap, valence band)

Atoms periodically arranged in the form of a crystal
↓
energy levels form a energy band
↓
Fermi - Dirac statistics fills the energy bands

Doping: carrier concentration (so also conductivity) of semiconductors can be easily altered by doping

(diagram: (Arsenic, Phosphorus), free electrons, E_c, E_D, N doping, E_A, P doping, E_v, bound electrons, mobile holes)

N doping: donors, 1 extra electron → ELECTRON

P doping: acceptors, 1 less electron → HOLE

In an undoped semiconductor we have $n = p = n_i$ ← intrinsic concentration

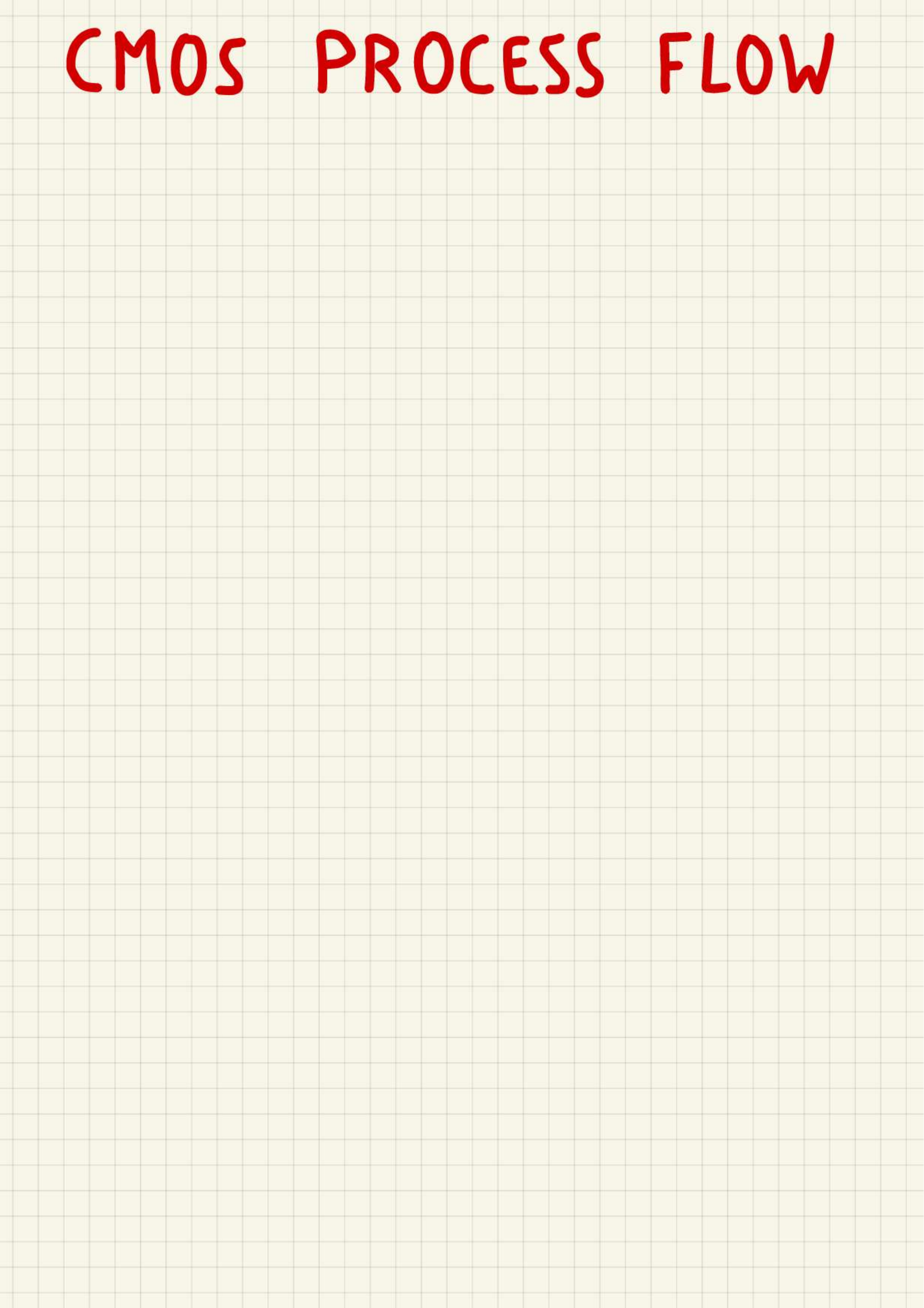
In a doped semiconductor this isn't true anymore, but $np = n_i^2$ (mass-action law)

Rule of thumb: if I dope with a certain type (for example Boron with N_A) we'll have roughly $n = N_A$

Freeze-out range: at very low T the number of carriers will be almost zero from the intrinsic material, and very low from the dopant.

A high T the doped semiconductor behaves like the undoped one, since all electrons can go into the conduction band, no matter the origin.

RESISTIVITY $\rho = \frac{1}{q(\mu_n n + \mu_p p)}$, $\mu_{n,p}$ are electron / hole mobility



CMOS PROCESS FLOW

Lecture 2 26 febbraio

CMOS process flow

Substrate choice = p-type (Boron during ingot growth), relatively high resistivity (25 ~ 50 Ω cm), (100) oriented substrates (less crystalline defects, better SiO_2 properties, same as diamond) are the common choice for CMOS manufacturing

STEP 0 - Water cleaning (removing contaminants and native oxidants): around 30 ~ 50 % of the process steps consist of cleaning (chemical, physical)

STEP 1 - we grow a thin layer of SiO_2 through O_2 or H_2O @ $T \sim 1000°C$
typical thickness of 4 ~ 40 nm. Thickness depends also on T.

STEP 2 - silicon nitride Si_3N_4 deposition: T 800°C, from 80 to 200 nm (CVD)
we ADD this material, like jam on a sandwich

CVD: chemical reaction between gases which leave behind, as a product, at a certain temperature, a solid-state product which deposits on the desired surface

Then we start our photolithographic process because we want to leave the Si_3N_4 only in the zones that we don't want to turn into isolation zones. The nitride would be covering the active portion of the wafer (active areas, which will host our MOS)

STEP 3 - photo-resist spinning

mask
photoresist
Si_3N_4
SiO_2
Si, P type, (100)

masks = 0

Photo-lithography using a mask is much faster (seconds) than the alternatives, like electron beam lithography (hours) which directly imprints the wanted patterns

Resolution of the patterning $R \propto \frac{\lambda}{NA}$ — λ: light wavelength; NA: numerical aperture of the lenses

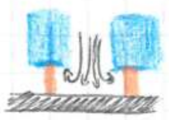
How to etch the photoresist?

H_2SO_4 wet etch

We also want to remove Si_3N_4, to expose the SiO_2, how? O_2 plasma dry etch

masks : 1

But to print fine lines we need an ANISOTROPIC etching



ISOTROPIC (WET) ETCHING

ANISOTROPIC (DRY) ETCHING

in plasmas we have radicals instead of molecules, so chemical reactions are accelerated (much lower T needed)

$O_2 \leftrightarrow 2O$

(NEUTRAL) MOLECULE NOT VERY REACTIVE

(NEUTRAL) RADICAL VERY REACTIVE

Wet etching is selective in the atoms (reaction), dry etching is selective in the direction (anisotropy).

To remove the remaining photoresist we can use H_2SO_4 because removes all the photoresist without touching Si_3N_4 and SiO_2

Active areas and isolation region formation :

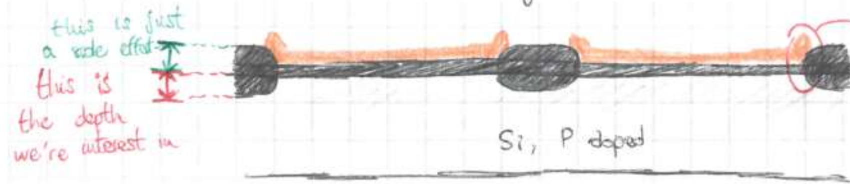
we will use a thick oxide to electrically insulate active regions

LOCOS (Local Oxidation Of Silicon) :

- thick SiO_2 is grown (water in a furnace with O_2 rich)
- Si_3N_4 (silicon nitride) is NOT oxidized and protects active regions
- Some oxidants penetrate at Si_3N_4 sidewalls → bird's beak
- Si_3N_4 removed by H_3PO_4 wet etch (NOT etching SiO_2)

oxygen must diffuse through Si to form SiO_2 because it needs Si. Oxygen diffusivity is VERY low in Si_3N_4 so it acts like a STOP!

NOTE : LOCOS isolation is no longer used in modern IC manufacturing



Why the first SiO_2 layer then? Is because of the temperature dilation. When we pass from the ~700°C of the process of Si_3N_4 deposition to room T, Si_3N_4 wants to shrink faster than Si, so tensile stress arises that would bend my wafer. SiO_2 has the opposite property: its thermal expansion coefficient is lower than Si, so would bend "less" than Si, causing an opposite tensile stress and balancing Si_3N_4.
(↑ compressive stress)

It's also easier to wet etch Si_3N_4 without touching the substrate SiO_2 than Si.

The bird's beak is an example of a problem with isotropic processes: we can't scale things down with isotropic processes because (for example) during LOCOS we create as much thickness as lateral growth

BIRD'S BEAK

FOCUS

$$NA = \sin\theta$$

We also lose the planarity of our surface due to LOCOS, which gets more and more important as we shrink down the dimensions, because of the resolution limit $R = \frac{\lambda}{NA}$ and DoF (depth of focus) $= \frac{\lambda}{(NA)^2} < 1$

So as my resolution increases, my depth of focus changes much faster, bringing easily items out of focus as they change their distance, which happens at the surface in LOCOS.

smaller chip → more resolution → lower DoF → higher planarity <u>needed</u>

WHY NO MORE LOCOS:
(why isn't used in modern VLSI)

- scalability problems
- no high resolution litography because of non-complanarity

Lecture 3   3 marzo

Ashing: dry etching using O_2 on carbon-based photoresist, which reduces it to ash.

What's a better alternative to LOCOS?

We would like something that doesn't consume our active area between the isolation regions AND keeps the surface planar, so to be able to shrink down the technology even further.

STI: Shallow Trench Isolation

STI is the state of the art isolation formation since it allows for lateral dimension shrinking and improves planarity.

photoresist (wet etching)
Si_3N_4
SiO_2
dry etching
Si(100)

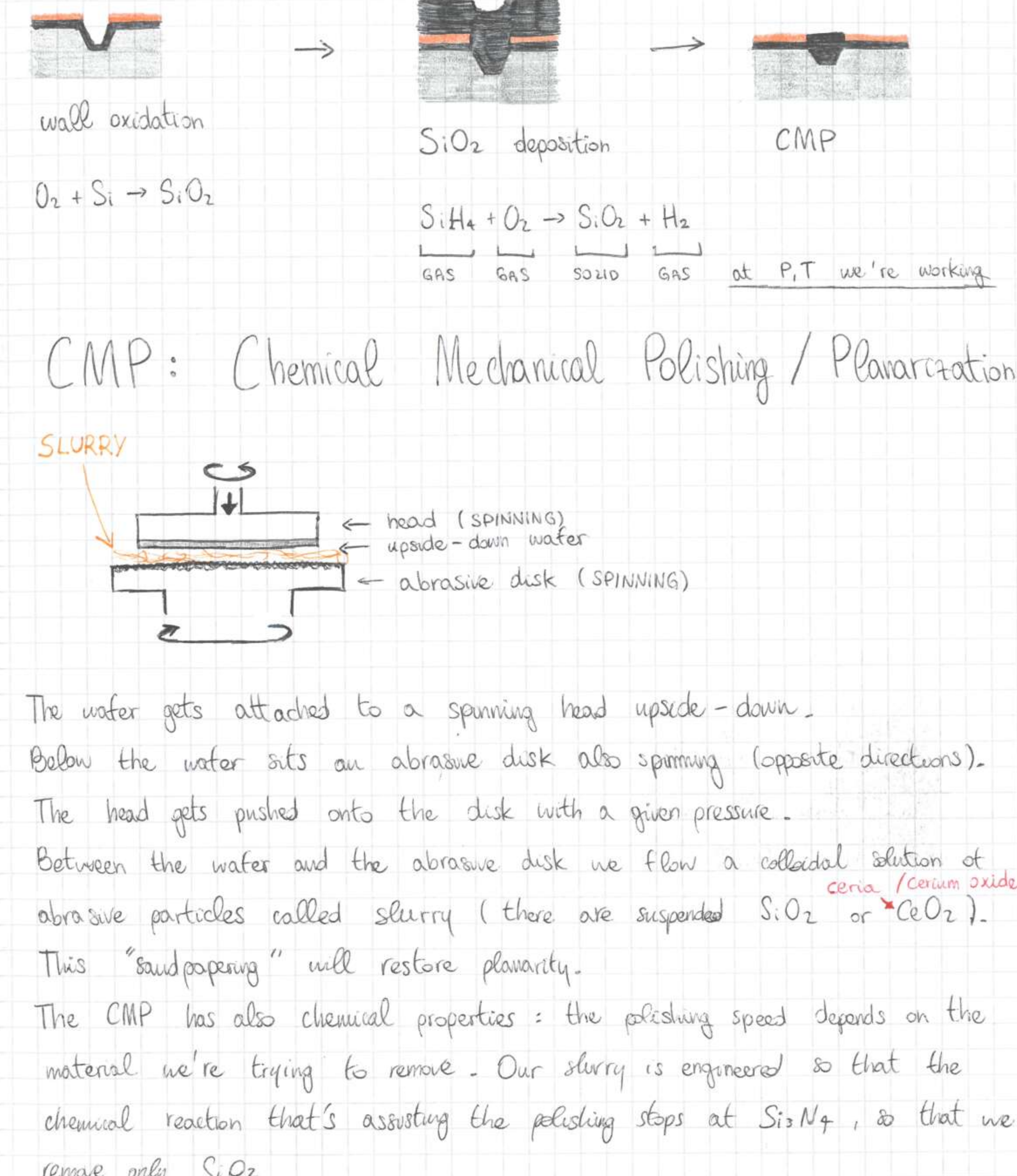
we can engineer the trench dry etching so that SiO_2, Si_3N_4 get removed vertically, while the trench gets excavated at an angle.

THEN we remove our photoresist

Now we do a wall **oxidation**: we'll grow a thin layer of (amorphous) SiO_2 on the walls of the trench. Why? Because dry etching damages the side walls of the trench, and that inner part of my Si is important. To get rid of all the defects induced by dry etching I grow a thin layer of SiO_2 before the SiO_2 deposition step (careful: first wall oxidation, then SiO_2 deposition)

GROW (CONSUMES THE SILICON)

DEPOSITED (CVD)



CMP: Chemical Mechanical Polishing / Planarization

The wafer gets attached to a spinning head upside-down.
Below the wafer sits an abrasive disk also spinning (opposite directions).
The head gets pushed onto the disk with a given pressure.
Between the wafer and the abrasive disk we flow a colloidal solution of abrasive particles called slurry (there are suspended SiO_2 or CeO_2 [ceria / cerium oxide]).
This "sandpapering" will restore planarity.
The CMP has also chemical properties: the polishing speed depends on the material we're trying to remove. Our slurry is engineered so that the chemical reaction that's assisting the polishing stops at Si_3N_4, so that we remove only SiO_2.

When do we stop "any" process on our wafer?

We do tests, so that we know the speed rate of every process:

- how to know when we stop etching?
  deposition?
  CMP?
  photolitography?
  STI?

do → measure → infer.

But we have variations between wafers and even different parts of the same wafer: that's why we implement additional techniques to control how my process goes.
For example I can shine a laser on my wafer as it's being polished and, from the interference pattern (reflected + refracted = interference) I can infer the thickness of the SiO_2 and know exactly when to stop.
Or I can measure in real time the torque needed to polish the wafer, which depends on the material: torque changes → I've exposed Si_3N_4 → I can stop.

NEXT we remove via wet etching (selective only to Si_3N_4) the filled oxide (name of the oxide put into isolation zones)

CMP → wet etch removal of filled oxide

removed by H_3PO_4

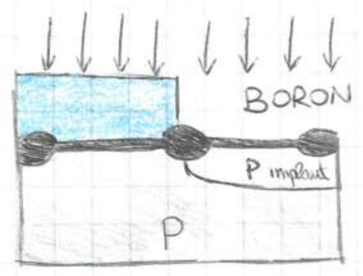
removed by wet etch

typically HF acid

NEXT we wet etch the Si_3N_4 using H_3PO_4

WELLS FORMATION

To build our active device, we need to properly dope silicon, which is done through ion implantation: for a CMOS transistor both p-type and n-type dopings are needed. Dose and energy of the implanted ions are chosen to match the wanted electrical properties for the CMOS.



MASKS = 2

- photo-resist spinning
- lithography on the photo-resist
- development of the photoresist

so now we have part of the wafer covered (protected) by the photoresist, so that if we do ion implantation (ALTERNATIVE which is less common: put the wafer in a furnace with Boron gas, which will diffuse into silicon) the masked portion won't be affected.

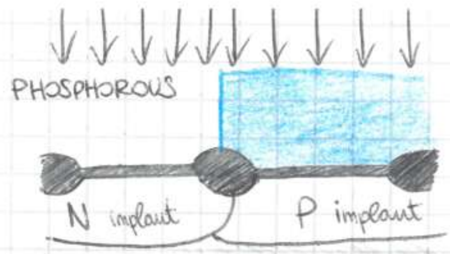
Ion implantation = basically a particle accelerator, so we take the source gas containing Boron (like BH_3), form a plasma (where B and H atoms will be separated and ionized). Divide B^{3+} from H^+ atoms and accelerate B^{3+} with an electric field into the wafer.

Since the beam of Boron ions is charged, we can know how many atoms we're implanting by measuring the current of the beam.
By tuning the electric field, we can tune the energy of B^{3+}, thus the depth at which they will be implanted.

NOTE: next page

typical B implantation: dose ~ 10^{13} atoms / cm^2

energy ~ 200 - 300 keV



MASKS : 3

we cover now the remaining part (the previous mask has been removed) and do N-type doping (phosphorous or arsenic commonly used)

NEXT we do an annealing step = thermal treatment with no chemical reaction happening

So we just heat the wafer at around 900° ~ 1100°C in an inert ambient (like N_2) for some hours.

<u>Why?</u>
- To be electrically active our implanted atoms will have to be in substitutional position to a Silicon atom in the crystal, so it must take the place of a Silicon atom.
  Usually after implantation the dopant atoms are in random positions in the lattice, NOT sitting in substitutional position. To obtain this we have to heat up the wafer. THIS PROCESS IS CALLED <u>ACTIVATION</u>

- We want the implanted species to diffuse to the right depth, and this happens when we heat up the wafer

- Reparation of crystalline damage = since we implant up to 10^{21} atoms/cm^3 (Silicon itself is 10^{22} atoms/cm^3, so <u>very</u> high implantation density) and the energy of each one of them is ~ 200 KeV (even 300 KeV), and the energy needed to dislodge a silicon atom from its crystalline lattice position is ~ 20 eV (from crystal to amorphous, the Si dislodged).

With high energy implantations we can even amorphize the silicon completely. But with the annealing process we can rebuild the crystalline structure

So after implantation, some kind of high-temperature treatment is needed.

NOTE: ion implantation affects also the SiO_2 layer, but that's not a problem since the only side effect would be that of changing the wet etch rate at which SiO_2 is removed, but that can be taken into account.

why did we keep this SiO_2 layer? → Metallic contaminant removal

During the ion implantation process, we know that there will be a small dose of metallic contaminants in the chamber (usually tungsten and molybdenum) which will be also accelerated and implanted, but since they're heavy metals, at the same energy as B or P their speed will be much lower, so their penetration depth will be lower → the thin SiO_2 layer will absorb most of them, and we will later remove this layer.

NOTE: are the implanted ions B^+, B^{2+}, B^{3+}? We can choose the ions to implant by putting a mass analyzer, which selects the ratio of (mass / charge) that we want.

NOTE: the wafer is electrically connected, so that the ions implanted regain electrons when they're inside the wafer.

NOTE: the Si wafer substrate is P doped because we generally want a body resistivity of $5 \sim 50\ \Omega \cdot cm$, so we achieve it this way

NEXT we remove the pad oxide (so also the contaminants present in the SiO_2 will be removed) via wet etching (HF is very selective to Si, so it won't be touched). Some isolation oxide will also be removed.

Now that we removed the contamined SiO_2 layer, we grow a new one (< 10 nm, thin) using again O_2 or H_2O, this will be our final oxide.

THERMAL BUDGET

When I heat my wafer to 900° ~ 1100°C during the annealing process, I have to keep in mind that the diffusivity of my implanted species grows exponentially with temperature, so I have a "thermal budget". It's important to know how much time the wafer will spend at which temperature during my whole process flow because the more I put it at high T, the more the dopants will diffuse deep into the substrate. Usually, as the process flow progresses, the operating temperatures become lower. At our first annealing process we can get to 1100°C because there's basically nothing on our wafer yet.

There are also alternative methods to the "1100°C for some hours" annealing we've seen. One of them is RTP (Rapid Thermal Process, which heats up the wafer up to ~1000°C more or less, in seconds or even microseconds, then are also cooled down quickly, so to use very little thermal budget), another way is to use lasers to heat up locally the wafer, nanosecond pulse lasers can be implemented, so even shorter timespans).

fine lezione: perché i side walls sono inclinati?

Lecture 4

10 marzo

Conformality : during the CVD process, conformality means that we have a uniform thickness of deposited film everywhere

CONFORMAL

same thickness everywhere

NON - CONFORMAL

thicker on horizontal surfaces

If my film deposition is non-conformal, I might run into the problem of a big hole in the middle

HOLE

BUT

we might run into the same "hole" problem with conformal CVD if we have negatively sloped walls (for some reason)

LAYER 4
LAYER 3
LAYER 2
LAYER 1

deposition is continuous, but I've shown layers for didactic purposes

HOLE

to solve this problems we combine CVD and plasma etching in a process called PECVD (Plasma Enhanced CVD)

Other acronyms also used are : HDP CVD (High Density Plasma CVD).

PE - CVD : Plasma Enhanced CVD

NOTE : in our plasma chamber we have :

- some neutral molecules (O_2, not important, non reactive)
- some radicals (O, neutral but highly reactive)
- ions + electrons (plasma, very reactive)

DRY ETCHING : radicals react with the surface atoms, form a gas and get pumped away

SPUTTERING : ions get accelerated and shoot away the atoms of the surface

CVD : radicals react with some pumped gases to form a thin film deposited.

non conformal = more on horizontal surface ↓ — vertical direction of accelerated ions ↙

If we combine CVD with SPUTTERING in the right amount, we can deposit and simultaneously remove selectively the horizontal plane, to have the same rate of deposition that we have on the vertical plane (lower rate due to non conformality, but also untouched by sputtering).

In this way we can completely fill our trenches without holes.

But this method works for aspect ratios of about 5:1 (depth : width), beyond that there's currently no way to completely fill up the trench without forming voids (holes), no matter the deposition technique used.

So we can fill the trenches in another way: SOD

SOD : Spin On Dielectric

sometimes referred as SOG : Spin On Glass

1. we take a polymer diluted in a solvent (just like the photoresist)

   Polysiloxazane in organic solvent (silicon, nitrogen, hydrogen)

2. spin coating : we spin it so it distributes evenly on the wafer, filling all of the trenches

3. pre-bake : we move it on a hot plate to evaporate the solvent, so we're left with the film in a solid state

4. curing with steam : we "oxidize" the polysiloxazane with water, this step removes the nitrogen and hydrogen and substitutes it with oxygen, giving us SiO_2

so <u>NOW</u> we could go back to wells formation / implantation.

NOTE : one of the reasons Silicon was chosen as the material for ICs is that its oxide is of excellent quality (and also isn't soluble in water ((SiO_2))). Si in the 100 direction has fewer atoms / cm^2, so the expansion given by SiO_2 isn't much of a trouble.

NOTE : the thin SiO_2 (< 10nm) we last grew is our final oxide, the one that will withstand the electric fields when we'll use the MOS. This oxide scales down with the MOS dimensions, but around 2nm we start having tunneling effect, so below this dimensions we need higher K oxides (higher dielectric constant = more insulating).

GATE STACK FORMATION

To form the gate electrode of our MOS we deposit by LPCVD (Low Pressure CVD) polycrystalline Silicon.

$$LPCVD: \quad SiH_4 \rightarrow Si + H_2 \qquad T: 450° \sim 600°C$$

"silane" (arrow to SiH_4); solid (Si), gas (H_2)

crystal = all atoms set in periodic structure

amorphous = no periodicity

polycrystal = areas where we have order (crystal), but no long-range periodicity

SiO_2 forms an amorphous structure when grown on top of silicon because the lattice constant of Si is too small for SiO_2, so it can't be taken as a template. The crystalline structure of SiO_2 is quartz.

Polycrystalline silicon (polysilicon)

GRAIN: here we have a crystalline structure

GRAIN with different orientation

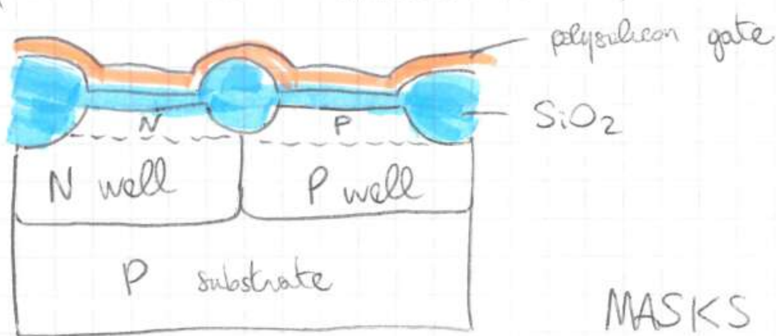
GRAIN with different orientation

GRAIN with different orientation

All grains have a crystalline structure, but they're "glued" together in different directions, so there's a disruption of periodicity at the boundaries between grains.

Why is this a polysilicon and not crystalline Si? Because we can't grow a crystalline structure without a template. Even worse: our template is SiO_2, which has a bigger lattice constant than monocrystalline Si.

Since polysilicon is in-between monosilicon and amorphous Si, its electrical properties are similar to monosilicon but slightly worse, since within a grain the properties are exactly like those of monosilicon.



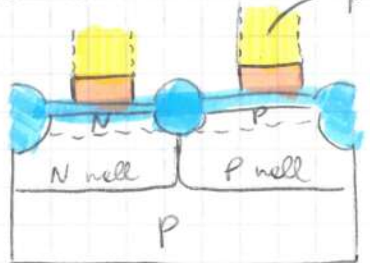
MASKS = 4

NEXT = Gate patterning

we put a mask on top of the polysilicon deposited by depositing the resist
then we dry etch the polysilicon:

- anisotropy of dry etching ensures vertical gate walls
- selectivity to the underlying SiO_2 must be very high

then we strip the photoresist



MASKS : 5

NOTE : the higher the T of deposition of my polysilicon, the more "crystalline" it will be. If I deposit at low T → amorphous Si instead of polysilicon.

NOTE : Why do we have additional N (P) doping inside the N (P) well? Because that additional doping is what ultimately decides the threshold voltage of my MOS device, so I can fine tune the doping of the "top layer" of the N (P) well to match the required criteria.

NOTE : the surface of the wafer will act as a sink for the interstitials (excessive atoms we have after ion implantation).

Problem!

Operating voltages don't scale as fast as transistor dimensions:

- 5V applied across a 2μm channel transistor → average E field ≃ $2,5 \cdot 10^4$ V/cm
- same voltage across 45nm transistor → average E field ≃ 1 MV/cm

But a high electric field in the channel means that we get hot holes and hot electrons (which are energetic enough to ionize the charges present in the drain zone and cause an avalanche effect). Thus hot e^-/holes might also inject charge through the gate (they go up from the channel to the gate = VERY BAD)

This problems get worse and worse as we shrink down the devices.
They're known as "SHORT CHANNEL EFFECTS".

To limit short channel effects we want shallow junctions (not doped to great depths), BUT junctions must also be sufficiently doped to guarantee ohmic contacts.

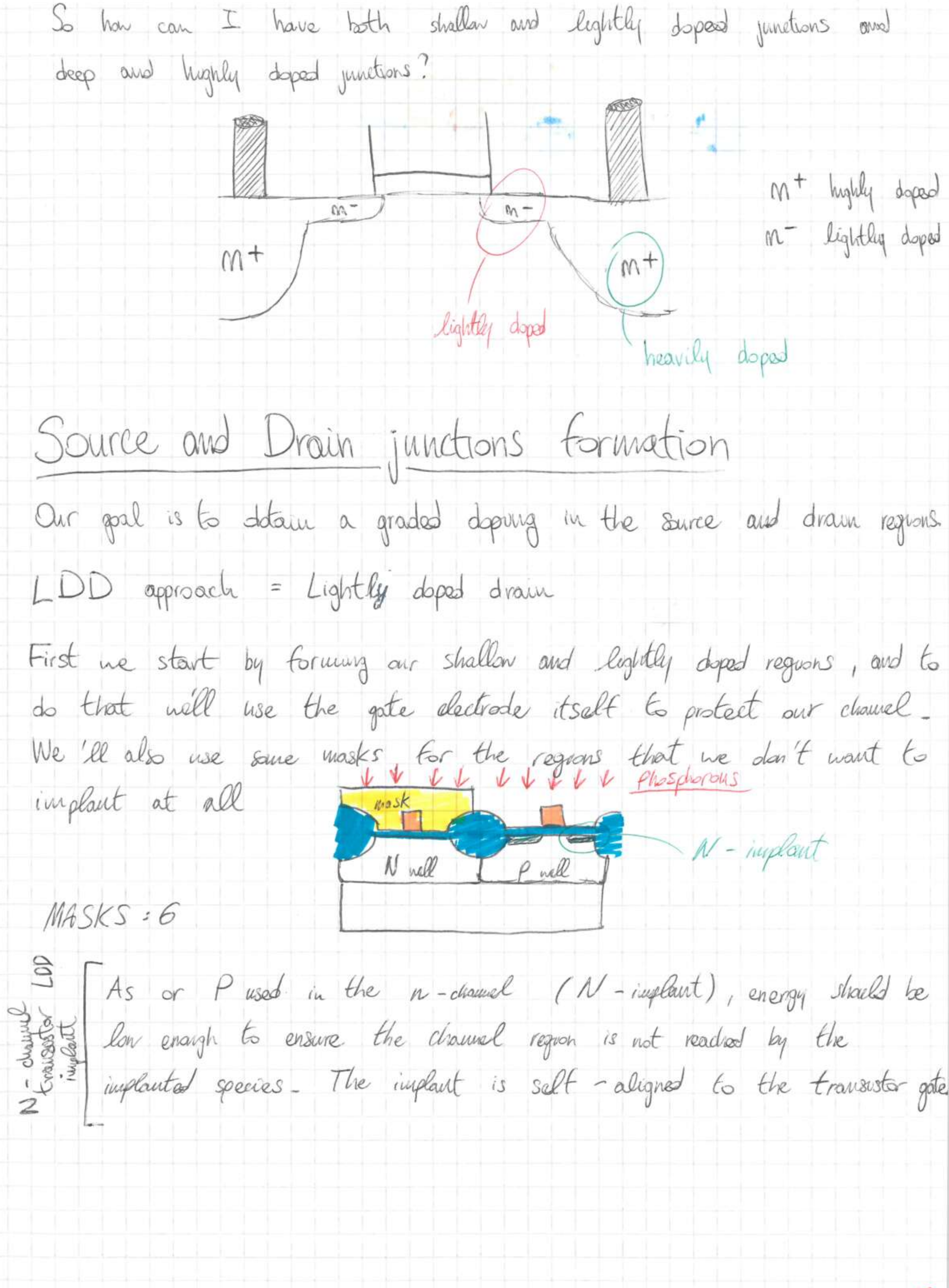
metal contact

n

p

ATTENTION!

In the contact region between metal and semiconductor, the higher the difference in resistivity the stronger it will resemble a Schottky diode! This is not an Ohmic contact (it might even draw current in the opposite direction)



So how can I have both shallow and lightly doped junctions and deep and highly doped junctions?

Source and Drain junctions formation

Our goal is to obtain a graded doping in the source and drain regions.

LDD approach = Lightly doped drain

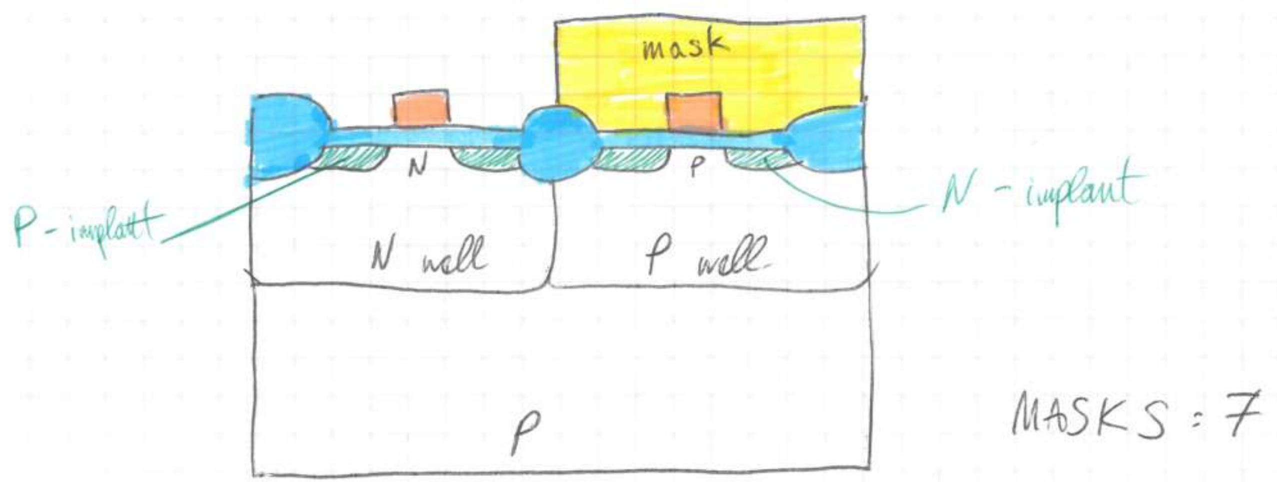
First we start by forming our shallow and lightly doped regions, and to do that we'll use the gate electrode itself to protect our channel. We'll also use some masks for the regions that we don't want to implant at all

MASKS : 6

N - channel transistor LDD implant

As or P used in the n-channel (N - implant), energy should be low enough to ensure the channel region is not reached by the implanted species. The implant is self - aligned to the transistor gate

Boron



P - channel transistor LDD⁻ implant :

B or In is used, similar doses as before ($\simeq 10^{12} - 10^{14} cm^{-2}$)

So the energy is tuned to make sure the ions don't go beyond the gate electrode — gate electrode.

NOTE : the fact that we're doping the gate is not a problem. It might even be beneficial since it changes the resistivity of the polysilicon.

NOTE : this method doesn't require a mask on top of the gate electrode since we tune the ions speed to not be able to go though its thickness. This is very good! Why?

1. One less mask means that the process is cheaper, especially because a mask on the gate (which is the smallest part of my chip) is the most expensive mask.
2. No mask → no alignment problems. Even the best alignment methods might miss the gate by a few nm (so around ~50 % error rate, too big!)

This alignment problem is called "overlay" or "registration" problem.
How can we know if we're correctly aligned? We measure the alignment between what we're printing and what was already there. This is done by using some special structures engraved in the wafer called scribe lines.

SCRIBE LINES

Scribe lines are just empty spaces in the wafer where later will pass the diamond saw and separate the chips. So scribe lines will be ultimately removed.
But until the packaging, we can use those spaces to print structures whose only purpose is alignment measures.
Examples of structures printed into the scribes:

ZOOM

printed before

printed after

- are the inside and outside lines aligned?
- is there any rotation of one square relative to the other?
- is there any rigid translation of one square relative to the other?

We can reach an alignment tolerance between subsequent masks of a few nanometers (2 ~ 3 nm), which is amazing but never zero!
And for gate nodes of 6 nm width, that's ~50% mismatch!

OR

So it's always better to find a way to self-align.

3. as we scale down our chip's dimensions, also the photoresist has to become thinner and thinner (since the λ I use for my lithography gets shorter and shorter AND the penetration of

light into matter becomes shorter and shorter, so if I want my resist to be developed fully, so in all of its thickness, I would need a thin photoresist if I use short wavelengths]. So even if I were to put a mask on top of the gate electrode, it would very likely be too thin to withstand the implantation itself.

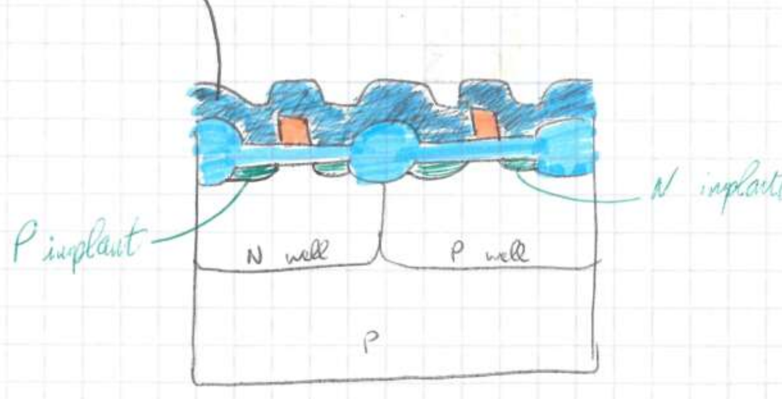
So this are the 3 main reasons why we don't use a mask for the gate electrode, plus doping the gate electrode might be beneficial in some cases.

So up until now we've formed the shallow and lightly doped junctions. How do we form the deep and highly doped outer junctions?
For the same reasons as before, we can't use a mask.
Can we find a self aligning process? Yes, through spacers formation.

Spacers formation

First we start with a conformal layer deposition (conformal = same thickness everywhere) then we do anisotropic dry etching

- LPCVD (low pressure CVD) oxide or nitride, i.e. $SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$ at ~400°C



- <u>very</u> anisotropic dry etching

  since etching only proceeds in the vertical direction, spacers are left on the gate sidewalls

<u>NOTE:</u> if nitride is used, selectivity to the underlying oxide is required

spacers

P implant

N implant

N well

P well

P

Now we repeat the steps used for the LDD approach: lay down a mask over the N well, implant Arsenic ($\sim 10^{15}\ cm^{-2}$), remove the mask, cover the n-channel transistors (P well), implant Boron, remove mask

Arsenic or Phosphorus

MASKS = 8

Boron or Indium

MASKS = 9

Let's see now just one channel :

P implant

N well

P+ implant

P

NEXT : we do an annealing for junction diffusion and crystalline damage. Usually is performed a high-T annealing, i.e. 1000°C for 1 min.

Diffusion of dopants must be strictly controlled (TED phenomena)

NEXT : oxide etching to expose areas of silicon to be "contacted".

And we're done, we made with just 9 masks our basic device.

This process is called FEOL.

Usually the integrated circuit microchip fabrication is divided in two categories: Front End and Back End.

- Front End : everything that happens <u>on</u> the wafer
- Back End : everything that happens after the wafer process is finished and you start separating the individual microchips and put them in packages, so they can be hosted on a circuit board

The Front End is subsequently divided into FEOL and BEOL
FEOL = Front End Of the Line
BEOL = Back End Of the Line

start ···→ FE ··· → BE ···
FE ↙ FEOL ↘ BEOL

- FEOL = everything you do to form your active devices sitting in the crystalline silicon

- BEOL = everything you do to form interconnections between those devices

Nowadays the FEOL and BEOL separation makes less sense, especially in memory device fabrication.

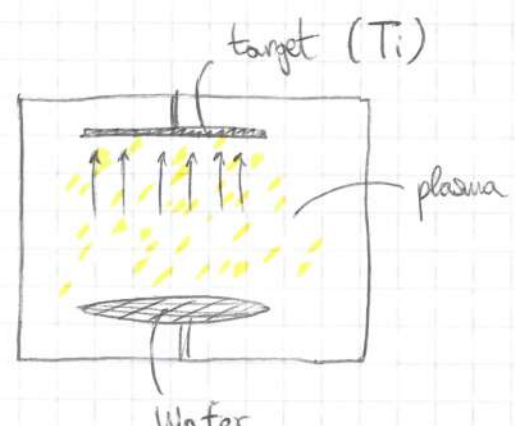
So we have terminated our FEOL process, now we need to create the interconnections between the devices (BEOL).

Local interconnect and active area metallization

To prepare our source and drain terminals for the metallic contacts, we **deposit** Titanium on the wafer by Sputter deposition (~ 50 nm).

Where the Ti meets the Si, it forms a silicide (compound between Silicon and a metal) $TiSi_2$, everywhere else TiN is formed because we react Ti in a N_2 ambient.

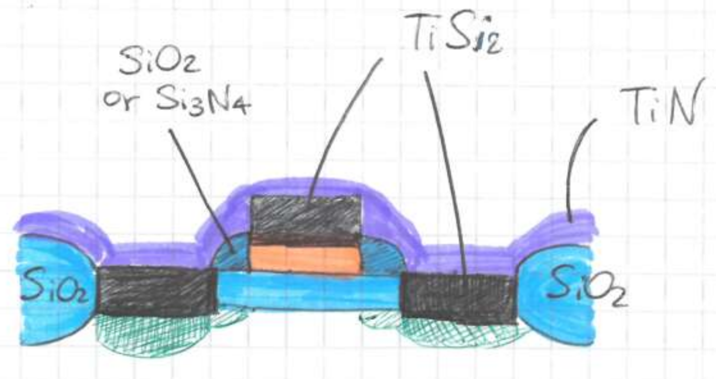
Sputter deposition: is a form of PVD (physical vapor deposition).



We put a piece of Ti (target) inside our chamber, use plasma to sputter ("polverizzare") the target, which then falls and recombines with the wafer

Sputter deposition is a room temperature process.

After the sputtering, we add N_2 and raise the temperature in order to promote a chemical reaction between Ti and the substrate.
Now we have $TiSi_2$ (only where Si is exposed) and TiN.



NOTE: $TiSi_2$ formation consumes part of the silicon substrate and part of source / drain junction too

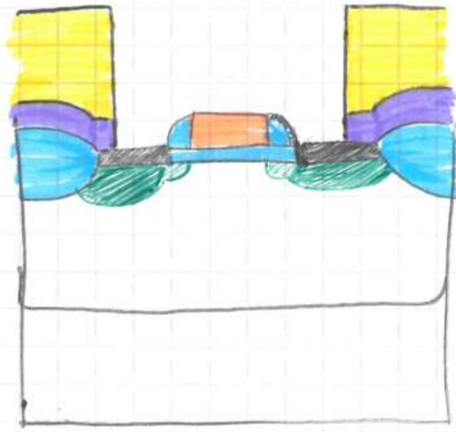
NOTE: spacers prevent $TiSi_2$ formation next to gate and active oxide

Then TiN is either patterned (not scalable, not used anymore) or simply stripped off and $TiSi_2$ is left on silicon regions. This process is usually called SALICIDATION (= Self - Aligned siLICIDATION).

Nowadays, instead of $TiSi_2$, $CoSi_2$ or Ni_2Si are used to adjust the resistivity, which changes when we scale down our dimensions.

OLD METHOD:

TiN etching by first masking, patterning TiN and wet etch in $NH_4OH : H_2O_2 : H_2O$ (1:1:5)



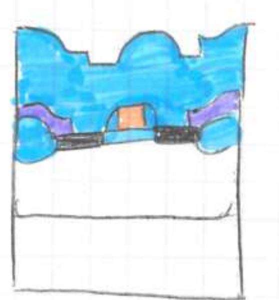
Wet etching stops at $TiSi_2$

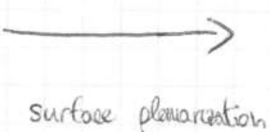
Lecture 5

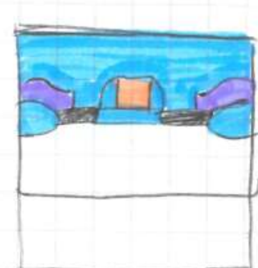
12 marzo

Contacts formation

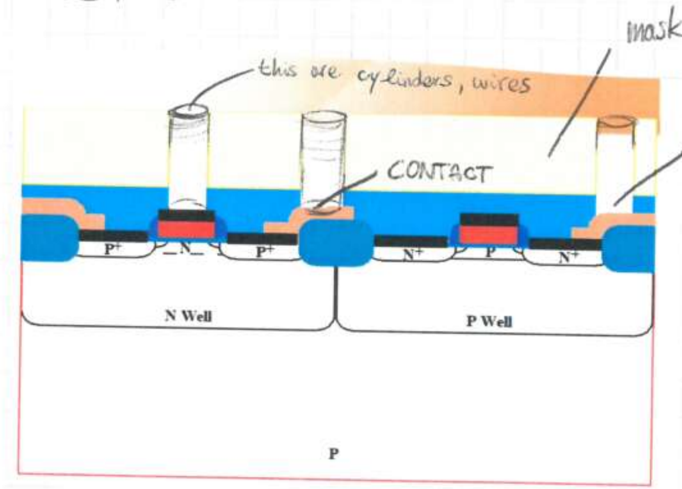
We start by depositing a dielectric (usually SiO_2, by CVD) and then restore planarization by CMP.





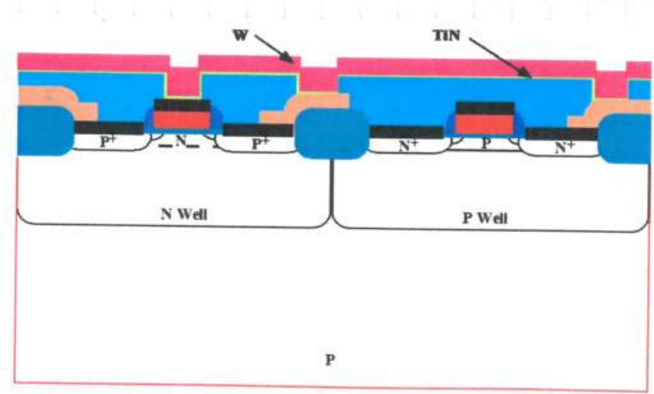


MASKS : 10



a mask is used to pattern the contact holes (they're cylinders, will become wires). Dry etching is used to remove oxide

MASKS : 11

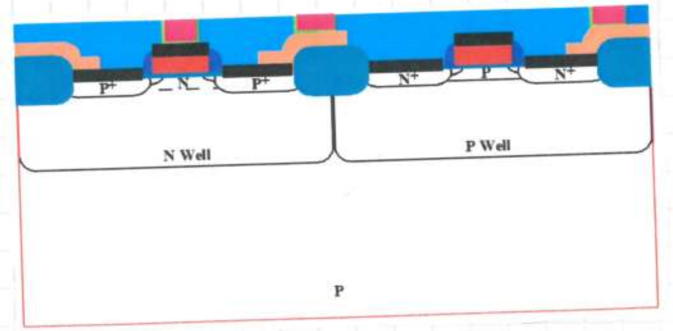


deposition of a few nm of TiN (or Ti) : it acts as an adhesive between substrate and W + protects underlying Si from W (since W is deposited by CVD, $WF_6 + 3H_2 \rightarrow W_{solid} + 6HF_{gas}$)

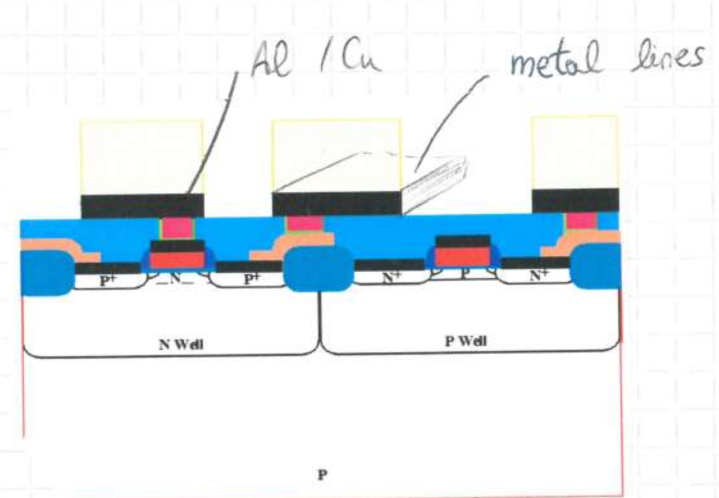
etches away silicon, not Ti/TiN

Planarization by CMP:

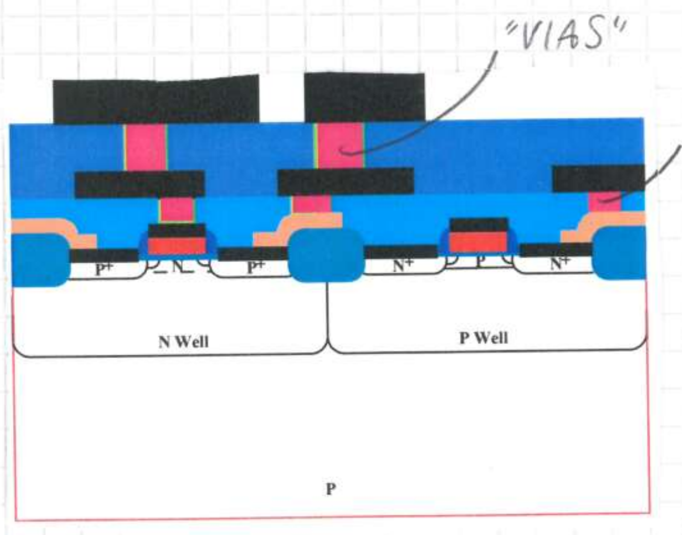
W and barrier are planarized and left only in contact holes



Al is deposited by sputtering alongside traces of Cu (to reduce electromigration effects)
Al / Cu is patterned using Mask 12 and dry etching



MASKS: 12



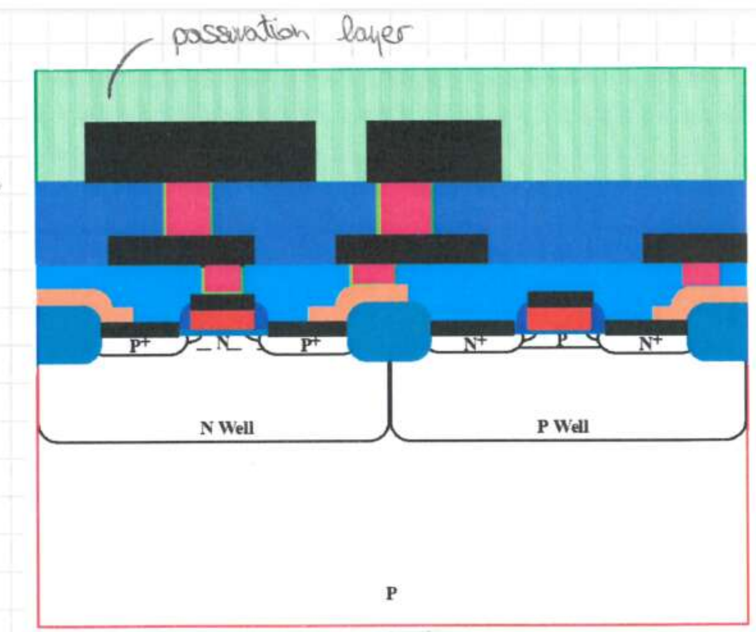
Multilayer metal: the process described previously can be repeated several times to form new plugs (called VIAS) and metal layers, needed to realize the desired interconnections

MASKS: 14

VIAS: when linking two metal lines
contacts: when directly attached to active device from a metal line

NOTE: due to its low resistivity, Cu is used instead of Al for advanced IC manufacturing. Cu is deposited by electroplating, and is not easy to etch. A damascene process similar to contact plug filling + planarization approach is used to realize metal level.

A passivation layer is deposited (SiO_2 or Si_3N_4) to protect the chip from humidity, contamination, ...

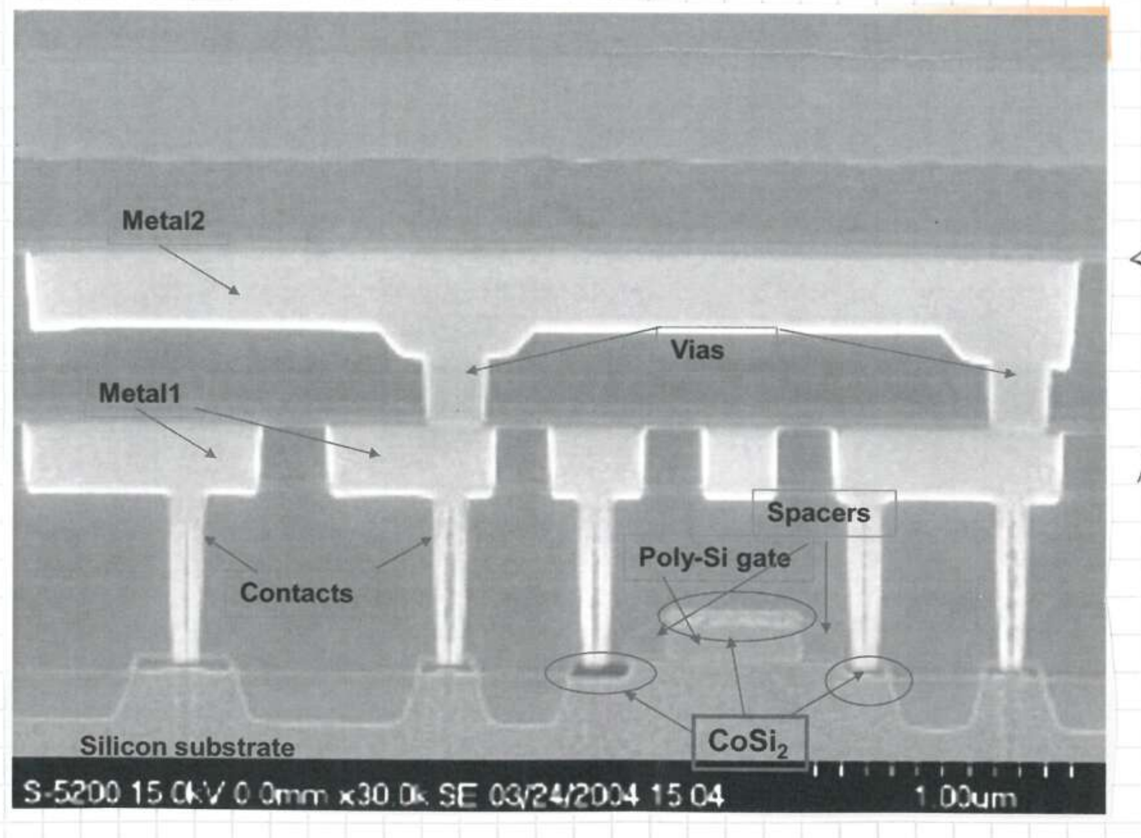


Passivation layer is patterned using MASK 15.

A final annealing at about 400°C in H_2 mixed with N_2 is used to alloy metals together and (mainly) to repair the active gate oxide bonds. This is the final step usually.

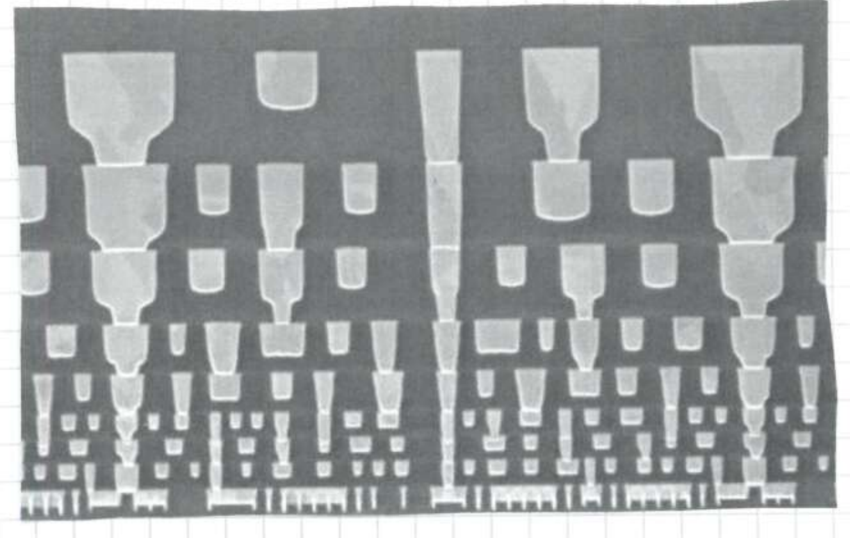
Why H_2? Because is a very fast diffuser, so it will penetrate and reach the gate oxide, repairing the charge defects.

NOTE: the patterning of the passivation layer is to reach the metal lines underneath.

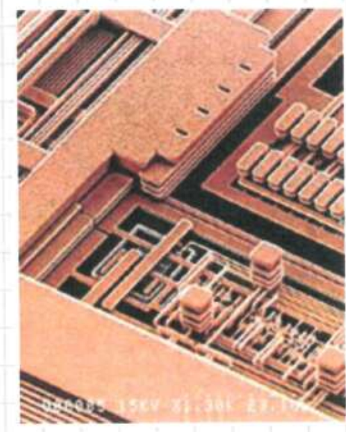


True transistor ←

(Numonyx 45nm NOR Flash HV circuitry)

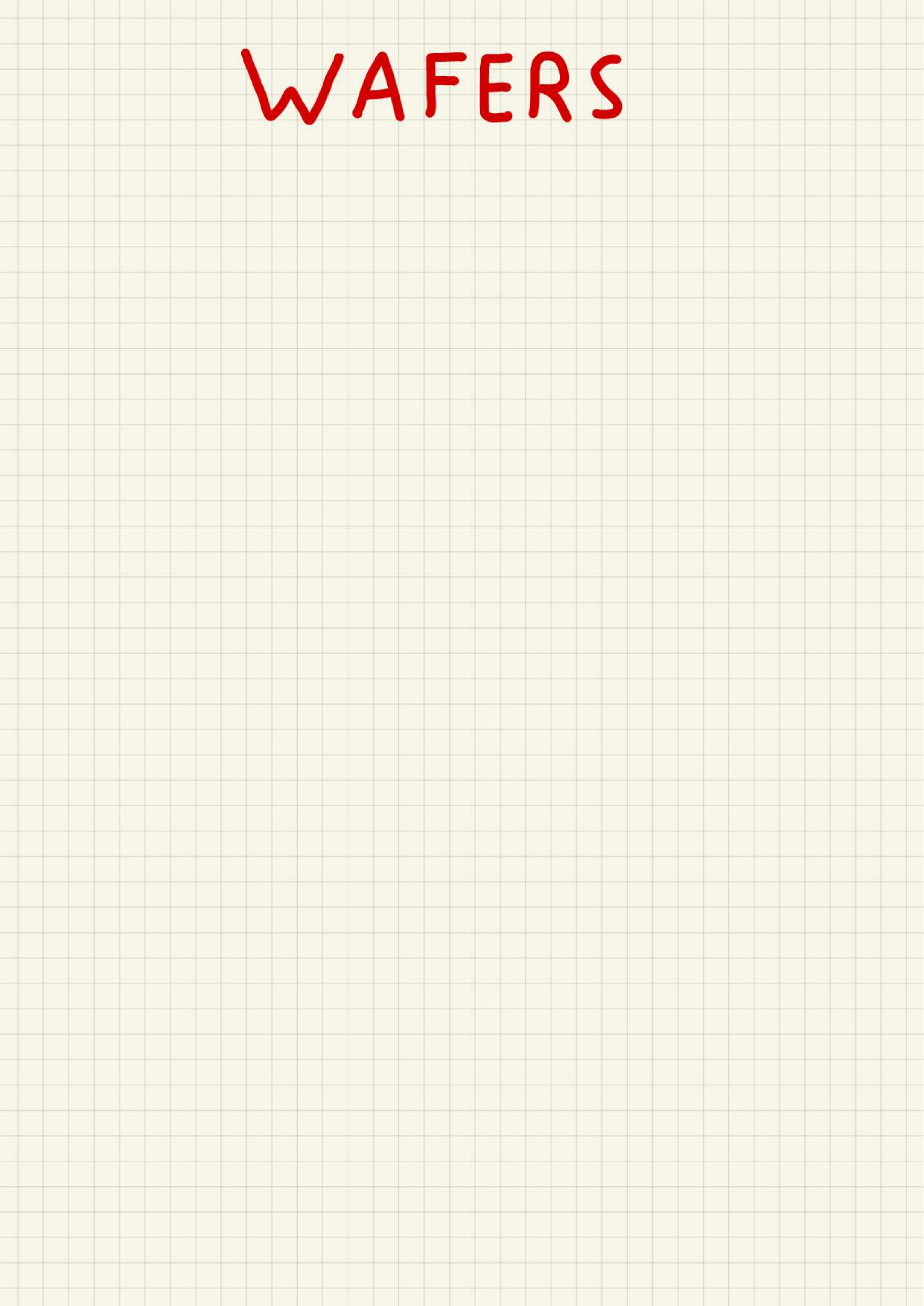


Intel 32 nm process



IBM 6 level metal Cu process

END OF CMOS PROCESS FLOW


WAFERS

SILICON WAFERS

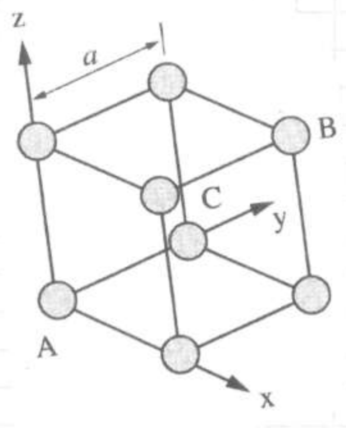
Bravais Lattice, two equivalent definitions

1. infinite array of discrete points with an arrangement and orientation that appears exactly the same, from whichever of the points the array is viewed

2. all points with position vectors $\vec{R}$ of the form $\vec{R} = n_1 \vec{a}_1 + n_2 \vec{a}_2 + n_3 \vec{a}_3$ where $\vec{a}_1, \vec{a}_2, \vec{a}_3$ are non-coplanar vectors. $n_1, n_2, n_3 \in \mathbb{N}$

If you can define a bravais lattice, you can simplify all of your calculations thanks to the periodicity of the crystal.

Examples of Bravais lattices and crystal structures:

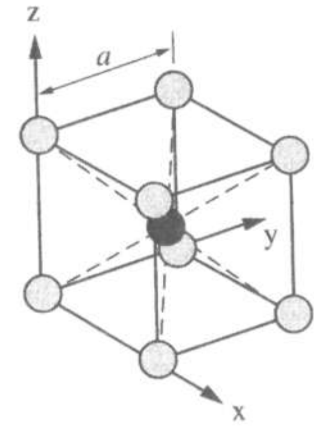


Simple Cubic Lattice

Nearest neighbours : 6

α - Polonium is the only element known to crystallize in a simple cubic form in normal conditions.

BCC – Body Centered Cubic

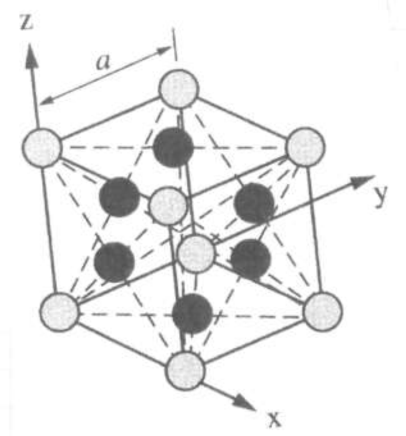


Nearest neighbours : 8

Element	a (Å)	Element	a (Å)
Ba	5.02	Na	4.23
Cr	2.88	Nb	3.3
Cs	6.05	Rb	5.59
Fe	2.87	Ta	3.31
K	5,23	Tl	3.88
Li	3.49	V	3.02
Mo	3.15	W	3.16

FCC – Face Centered Cubic

Nearest neighbours: 12

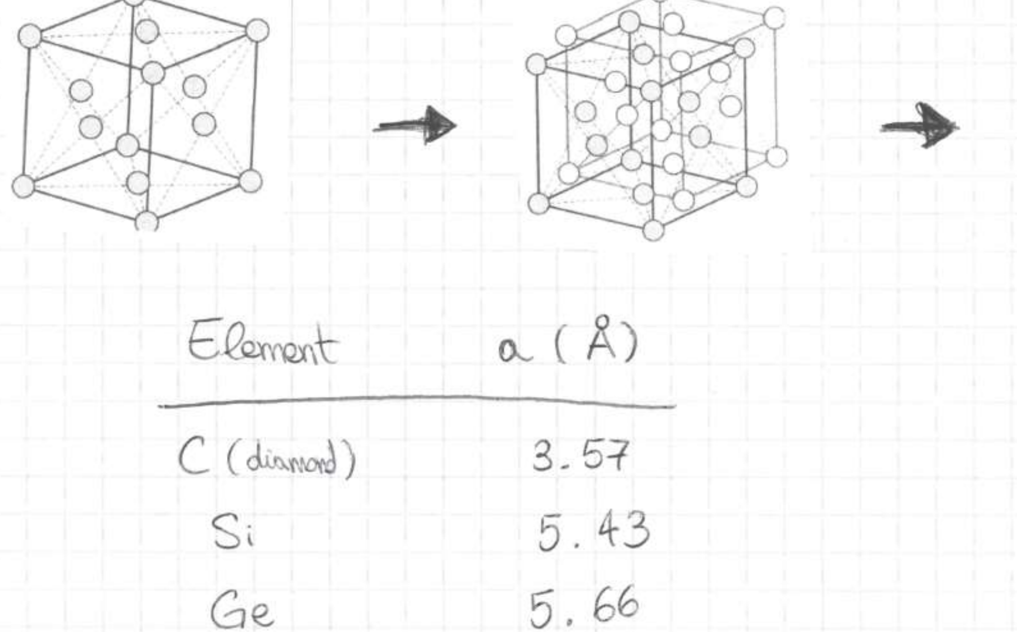


Element	a (Å)	Element	a (Å)
Ar	5,26	Ni	3.52
Ag	4.09	Pb	4.95
Al	4.05	Pd	3.89
Au	4.08	Pr	5.16
Ca	5.58	Pt	3.92
Ce	5.16	δ-Pu	4.64
β-Co	3.55	Rh	3.80
Cu	3.61	Sc	4.54
Ir	3.84	Sr	6.08
Kr	5.72	Th	5.08
La	5.30	Xe	6.20
Ne	4.43	Yb	5.49

Diamond structure - NOT a Bravais lattice

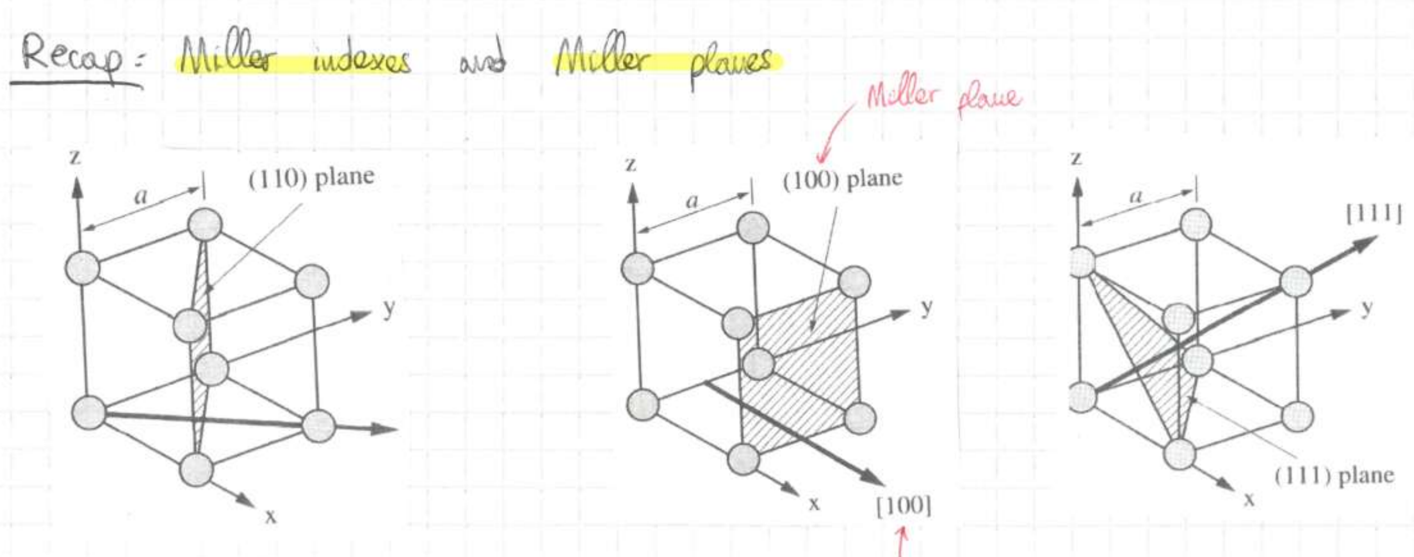
Silicon arranges itself following the diamond structure, which is not a Bravais lattice (but can be considered as an FCC Bravais lattice with a two-points basis).

Is obtained by two interpenetrating FCC lattices, with the origin of the second lattice offset of a/4 in every direction.



Element	a (Å)
C (diamond)	3.57
Si	5.43
Ge	5.66
α - Sn	6.49

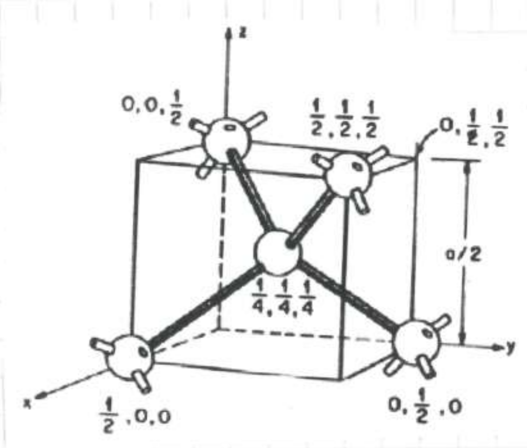
Recap: Miller indexes and Miller planes



[x y z] are directions, Miller indexes

(x y z) are planes ⊥ to [x y z], called Miller planes

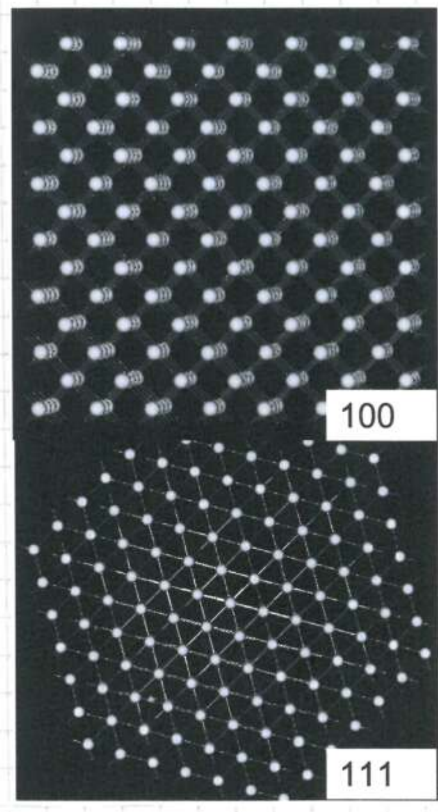
Silicon



a (= cubic cell side length) : 5,4 Å

closest neighbours distance : 2,36 Å

TOP VIEW:



(100) orientation is widely used for Si in microelectronics : lowest number of atoms per cm^2, lower oxidation rates, lower density of defects, ...

(111) orientation is adopted for older BJT technologies / power electronics.

NOTE: (100) is preferred for most MOS-based technologies also because of the less tensile stress it undergoes when SiO_2 is grown (SiO_2 = volume expansion compared to Si → mechanical stress → the less dense the Si wafer, the better)

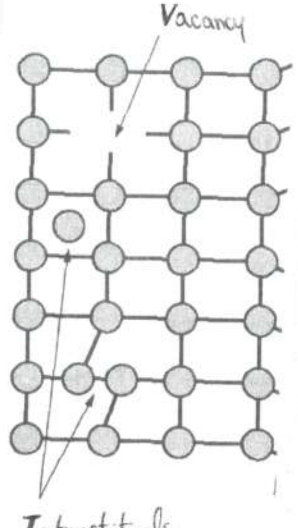
Defects

Generally classified in:

- point defects (interstitials, vacancies)
- line defects (dislocations)
- area defects (stacking faults)
- volume defects (precipitates)

Point defects

In a crystalline lattice you expect the points to be sitting in their Bravais lattice position, but $\forall \; T \neq 0°K$ there will always be a certain number of defects in my Silicon.



Intrinsic point defects:

- VACANCY (V): a lattice site without a Silicon atom

- INTERSTITIAL (I): an extra Silicon atom that resides NOT in a lattice site

NOTE: If the interstitial comes from a nearby vacancy we have a "Frenkel pair"

NOTE: dopants and impurities can be seen as point defects too (substitutional to Si atoms or interstitial)

The number of defects is inscribed in the thermodynamics of the crystal, so at a given temperature there will be a number of defects for my crystal to exist: it exists an equilibrium concentration of neutral point defects determined by a given entropy and enthalpy of formation.

$$C^*_{V_0}, C^*_{I_0} = N_s \, e^{\frac{S^f}{K}} \, e^{-\frac{H^f}{KT}}$$

N_s = density of lattice sites ($5 \cdot 10^{22}$ cm^{-3})

S^f = entropy of formation

H^f = enthalpy of formation

$C^*_{V_0}, C^*_{I_0}$ = concentration of neutral vacancies and interstitials

Exact values of S^f and H^f are still under debate, but a reasonable estimate of concentration at 1000°C gives:

$$C^*_{I_0} \simeq 10^{12} \, cm^{-3}$$

$$C^*_{V_0} \simeq 5 \cdot 10^{13} \, cm^{-3}$$

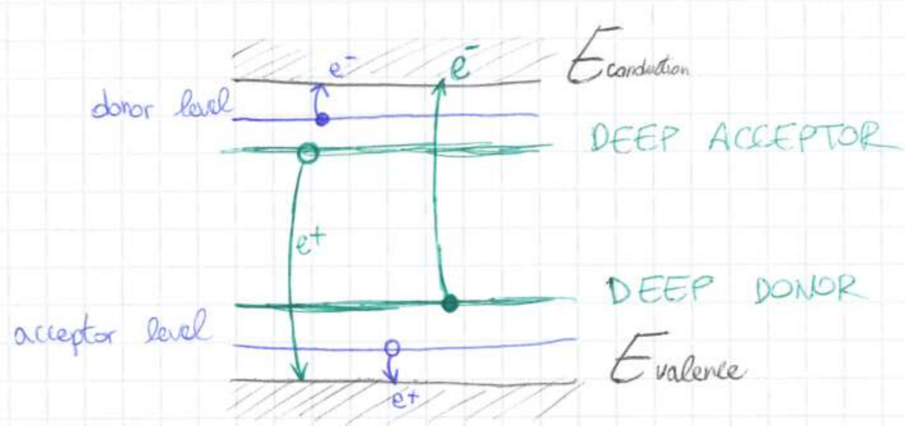
less than 1 in a billion defects!

NOTE: there's no thermodynamic requirement for the two concentrations to be equal!

Typical dopant concentrations $\simeq 10^{15} / 10^{20}$ cm^{-3}, so the defect concentration is very small.

NOTE: interstitials and vacancies can be charged!

Charged interstitials / vacancies show up as energy levels in the band gap and they usually act as deep donors / deep acceptors.



The concept is always the same for donor (or acceptor) and deep donor (or acceptor):

- donor level = a level that is ready to give away an electron to the conduction band, and when it's charged (so the e^- jumped to the $E_{conduction}$) this level is positively charged

- acceptor level = reverse of donor level (e^+, negatively charged, ...)

"Deep" = far away, compared to the non-deep level.

NOTE: the number of charged vacancies / interstitials depends on the position of the Fermi level, so by changing it (= doping) we change the total number of defects !

← OF CHARGED DEFECTS*

some examples (no need to remember):

$$C_{V+} = C^{*}_{V^0} \, e^{\frac{E_{V+} - E_F}{kT}}$$

$$C_{V++} = C^{*}_{V^0} \, e^{\frac{E_{V+} + E_{V++} - 2E_F}{kT}}$$

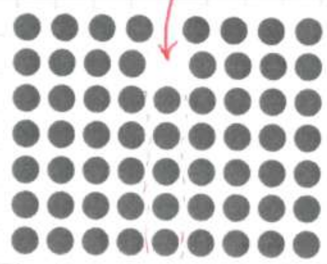
$$C_{V-} = C^{*}_{V^0} \, e^{\frac{E_F - E_{V-}}{kT}}$$

$$C_{V--} = C^{*}_{V^0} \, e^{\frac{2E_F - E_{V-} - E_{V--}}{kT}}$$

NOTE: doping doesn't "ionize" the neutral defects, those are determined by the temperature of my crystal. That's why doping changes the number of total point defects, because it's not converting neutral defects into charged ones but is creating charged defects.

Line defects / dislocations

A dislocation is a local deformation of silicon lattice:
a very easy example of dislocation is an "edge dislocation", which can be constructed by adding an extra plane terminating at the edge of the crystal.



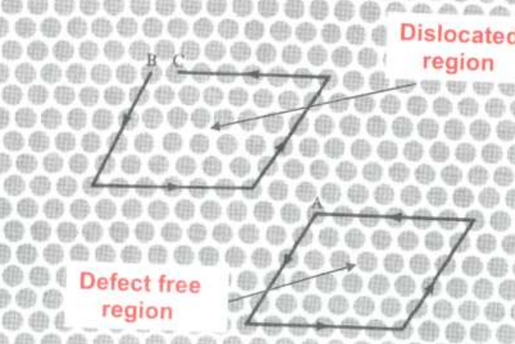
Dislocation formation can be induced by stress, point defects agglomeration, implantation damage, ...

BURGERS VECTOR

A way to find dislocation is through the Burgers vector:

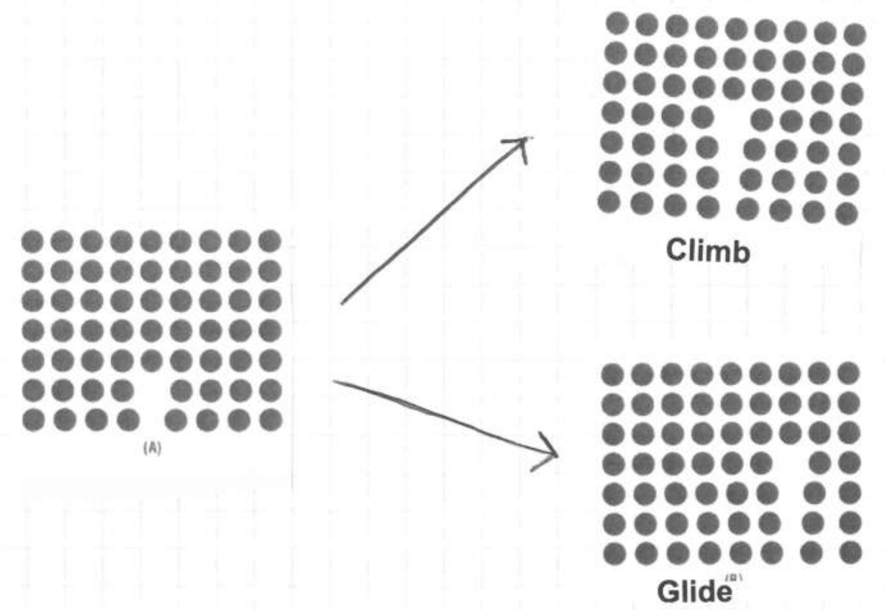
1. start from atom A, move n steps to the right
2. go up m steps (= lattice sites)
3. go left n steps
4. go down m steps

- if you go back to A → NO DISLOCATION
- if you don't, then the vector connecting you to A is called Burgers vector and identifies the dislocated region, enclosed by the path 1 → 2 → 3 → 4



NOTE: the enclosed region is defect-free only if the Burgers vector is vanishing.

NOTE: dislocations can move in the crystal, for example if there are too many point defects ("climb", vacancies or interstitials are absorbed by the dislocation) or under sheer stress ("glide")

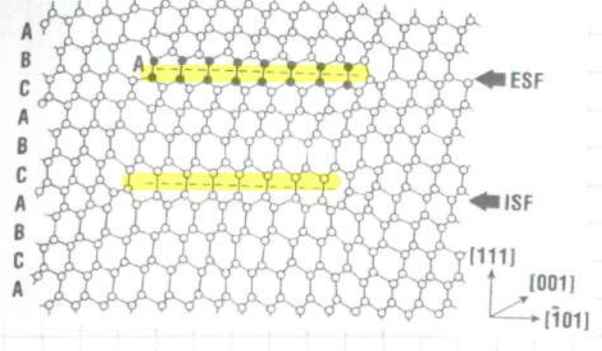


Dislocations make a structure resilient to stress and not brittle, since it's a way of absorbing and dispersing mechanical stress.
A perfect crystal with no dislocations at all would be very fragile.

Area defects / stacking faults

Area defects are divided into two categories:

- extrinsic stacking fault (ESF): an extra partial plane of atoms is inserted in the lattice
- intrinsic stacking fault (ISF): a partial plane of atoms is removed from the lattice



- Stacking faults in Si are usually {111} planes, don't know why.
- By definition, at the end of a stacking fault we have a dislocation (or the wafer ends).

"gettering" = process by which metal impurities bond with oxygen precipitates, instead of just laying around in my wafer.

Volume defects

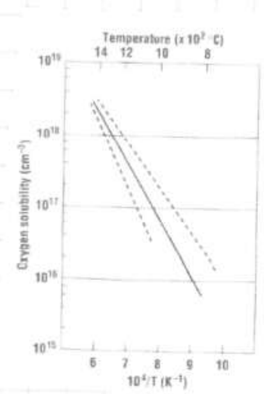
The most common form of 3D defects are **precipitates**.

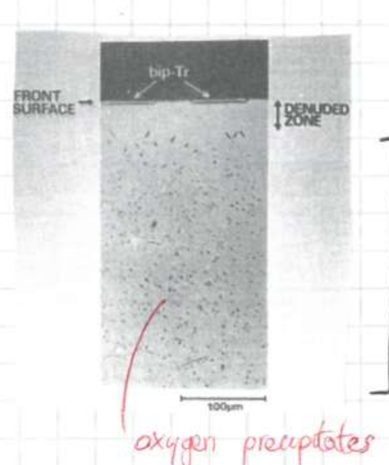
But how to form precipitates?
There's a thing called "**solid solubility**", which is the maximum concentration of an impurity that you can dilute into a solid, when you exceed solid solubility, that impurity will **precipitate in a 3D agglomerate**, like when we add too much sugar into coffee. Solid solubility depends on **temperature**, and that matters because Si growth happens at very high T (impurities dissolve), then the crystal cools down → impurities precipitate in the form of clumps of material

Usually this is a bad thing, but can also be exploited!

Gettering

O_2 can be incorporated in Si during crystal growth (very high T) so to **exceed solid** solubility for **low T**. During cooling precipitates are formed. A dedicated annealing process causes out diffusion of O_2 from the water surface (evaporation), creating a "**denuded zone**" in which active devices can be built. The "buried" damage zone is used to getter impurities during water processing.





oxygen precipitates

metallic impurities will bond to O_2 here, then I will remove this part and package the rest.

Lecture 6 17 marzo

For IC fabrication (usually) the semiconductor of choice is Silicon, which is very abundant : ≈ 27 % of Earth's crust.
Usually the starting material is quartzite (sand, SiO_2).
From sand (quartzite) then we can obtain Metallurgical grade Si (MGS, ≈ 98 % pure) by heating quartzite in a C source in a furnace at about 2000 °C.

$$2C_{solid} + SiO_{2\,solid} \rightarrow Si_{liquid} + 2CO_{gas}$$

To obtain Electronic grade Si (EGS, 4 nines = 99,99 % or 5 nines = 99,999 % purity) we react :

$$Si + HCl + catalyst \rightarrow SiHCl_3 \qquad \text{(trichlorosilane)}$$

⬇

$SiHCl_3$ purification by distillation (fractional distillation)

⬇

when the trichlorosilane is pure enough we can do a CVD :

@ 800°C $$2SiHCl_{3\,gas} + 2H_{2\,gas} \rightarrow 2Si_{solid} + 6HCl_{gas}$$

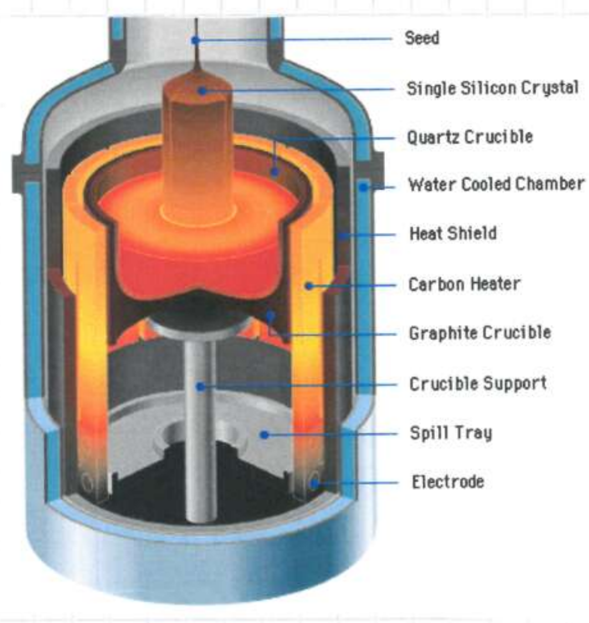
⬇

a polycrystalline silicon (poly-silicon) is obtained

REMEMBER: we can't get a monocrystal without a template!
The best we can do without it is polycrystalline.

How to get a monosilicon crystal to make wafers with?

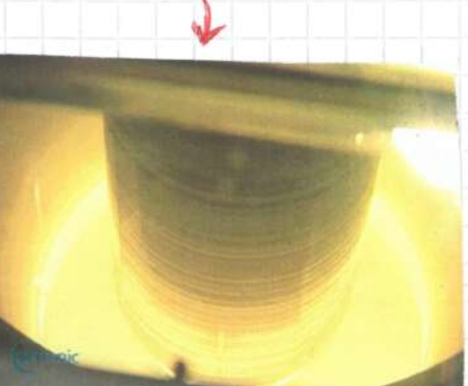
Czochralski technique (CZ)



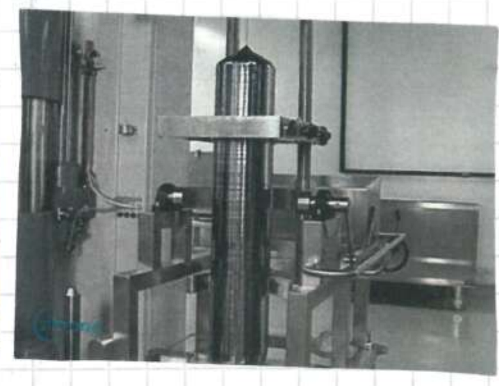
1. EGS (polysilicon) is melted in a SiO_2 crucible
2. a "seed" is lowered in contact with the melt. The seed is monocrystalline
3. the crystal is slowly pulled away: a single crystal ingot ("boule") will solidify while cooling down away from the melted surface
4. seed and crucible are rotated during pulling, to increase uniformity

With the CZ method modern boules diameter can reach 45 cm (even though 30 cm is the standard) and 1~2 m in length.

crucible (inside) ↓



boule ↓

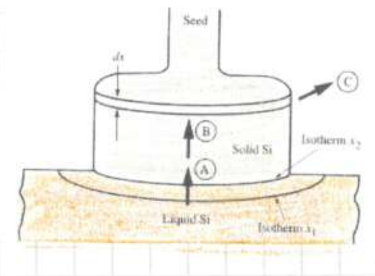


seed ↓

But if the CZ method is requiring a monocrystalline silicon, how do we get it? By chemical etching.
There are some chemical etching processes which etch faster along some directions, so if we etch a polysilicon we'll have only one direction which is etched away slowly, leaving at the end a monosilicon.

Pull rate

The bigger the diameter of the boule, the lower the pull rate, so the slower the process.



- liquid Si is above melting point
- solid Si is below melting point
- to make liquid Si solidify, we must dissipate at least the latent heat of fusion through the solid Si, so that the liquid Si can solidify onto it.

maximum pull rate $$V_{max} = \frac{k}{\rho L} \frac{dT}{dX_s} \simeq \frac{1}{\rho L} \sqrt{\frac{1}{r}}$$

k = solid Si thermal conductivity
ρ = Si density
L = latent heat of fusion
r = boule radius
dT/dX_s = T gradient at solid surface

↗ this will be derived next lecture

Doping of Si during CZ growth

How can we get in-situ-doped silicon? We could do this by injecting a gas (borane BH_3, phosphine PH_3, ...) into the chamber during the CZ growth. The problem is that the concentration of dopant won't be constant in my ingot but will increase as we go down.

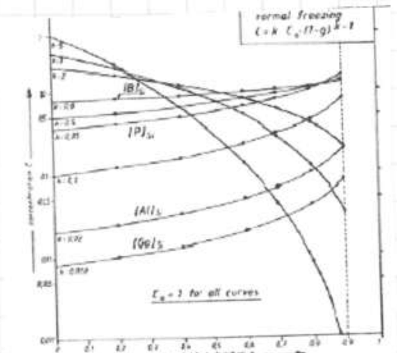
Why?

Because of different solubility of impurities in liquid compared to solid.

C_s solubility in solid (concentration)

C_e solubility in liquid (concentration)

$$k = \frac{C_s}{C_e}$$ segregation coefficient



so for $k \neq 1$ we will have a dopant concentration which is not uniform across the Si boule.

Element	Segregation coefficient
Al	0.002
As	0.3
B	0.8
P	0.35
Sb	0.023

$$C_s = k\, C_0\, (1 - X)^{k-1}$$

X = fraction of the melt that has solidified

C_0 = initial impurity concentration in the melt

So impurities "like" more to stay in the liquid phase of Si, so as the crystal grows, Si leaves more than the impurities, increasing the impurities' concentration.

There are also cases where $k > 1$, so where the impurities prefer to stay into the solid Si.

WAFER PREPARATION: next steps

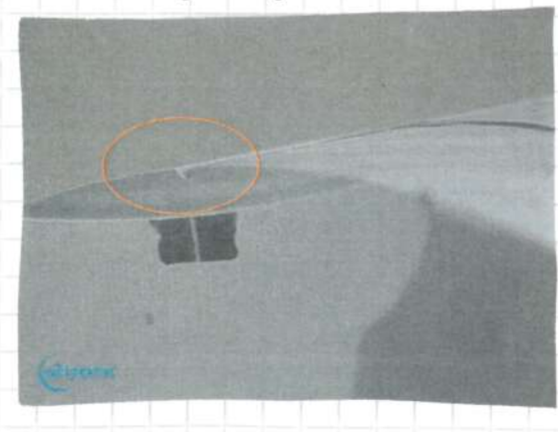
1. Ingot cutting

- top and bottom of the boule are removed
- portions of ingot are selected according to resistivity requirements

2. Diameter trimming

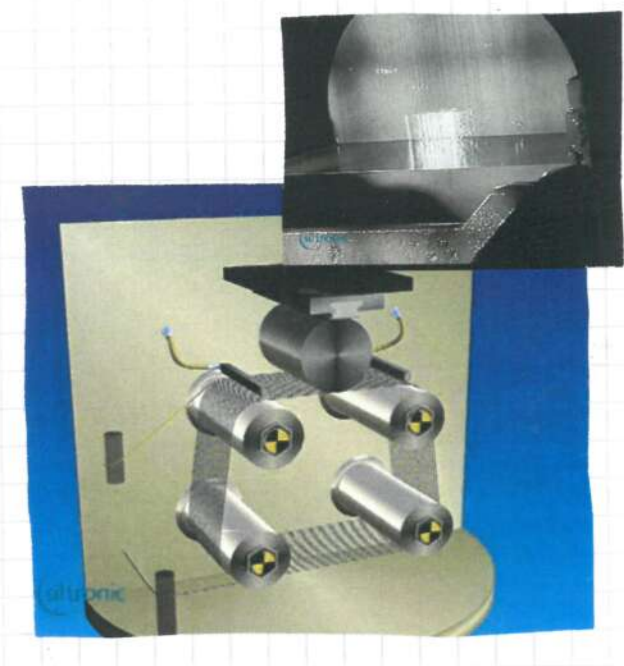
- the boule is reduced to the desired diameter

3. Notch grinding



- we form a little indentation (notch) on the surface, needed as a reference point (alignment). Usually directed along a crystalline direction, (100) usually.

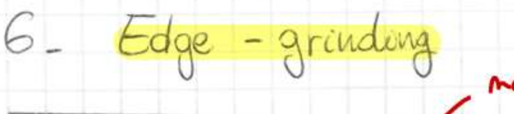
4. Wafer slicing



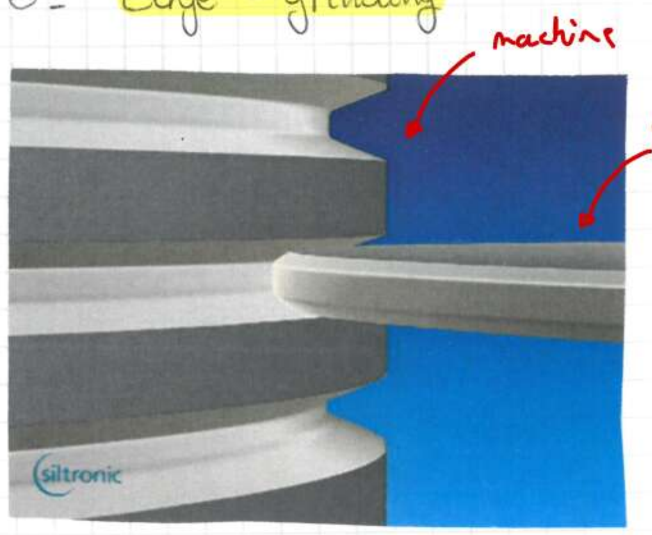
- wafers are sliced with a diamond wire saw

5. Mega-sonic cleaning

- wafers are cleaned in a chemical solutoon through mega-sowe cleaning

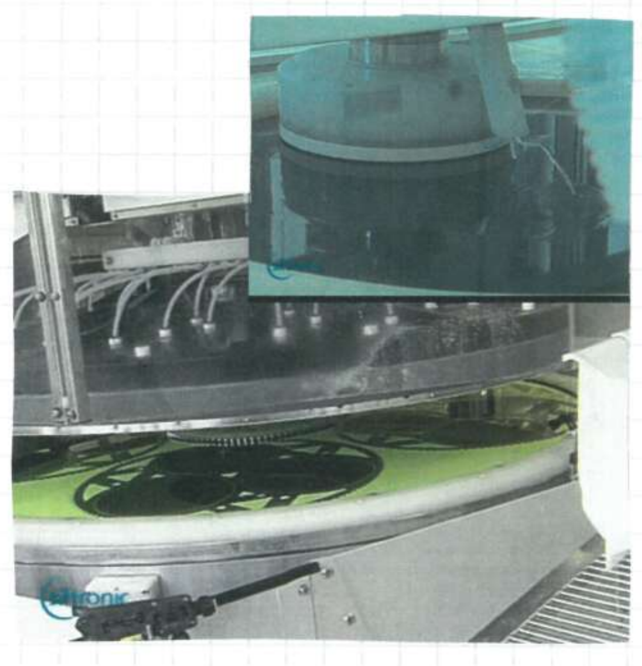


6. Edge-grinding



- the edges of the wafer are shaped based on the indications of the IC manufacturer requirements

7. Lapping and polishing



- wafers are lapped (= planarized) to the desired thickness
- wafer surfaces are polished

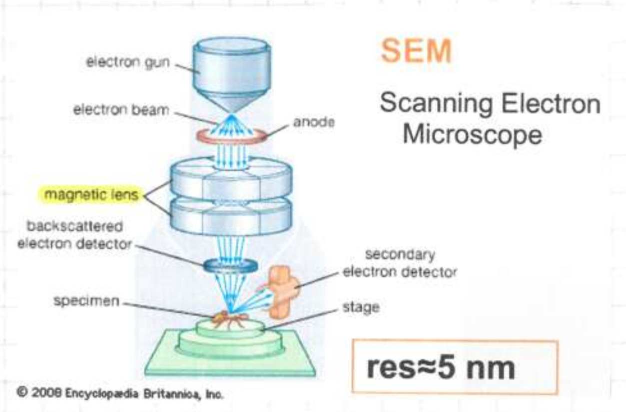
8. Annealing

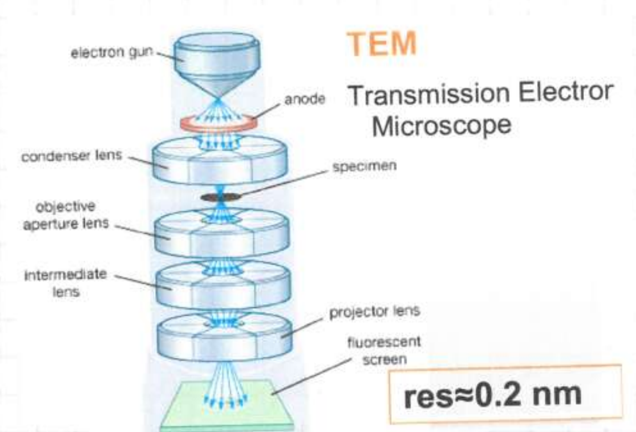
- thermal treatments are performed e.g. to form denuded zone
- if needed, epitaxial Si can be grown on top. This grown layer will also be monocrystalline and of higher quality than the underlying CZ monosilicon.

Measurement methods

For every property we want to check, there are measurement techniques to achieve that.

To observe our wafers we can't use normal microscopes, we need far more powerful ones: SEMs and TEMs. They use electrons instead of photons, so even if NA is in the order of 0.01 for magnetic lenses, the λ of electrons is $10^4 \sim 10^5$ times smaller compared to photons

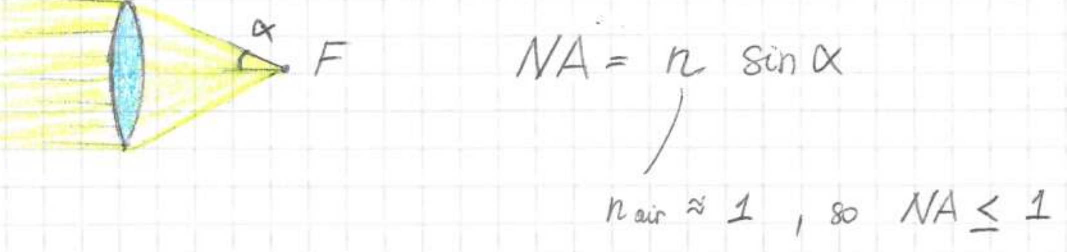




$$\text{resolution} = \frac{K\lambda}{NA}$$, but for electrons $$\lambda = \frac{h}{\sqrt{2qmV}}$$

small resolution = can resolve (see) smaller things

NA = numerical aperture



$$NA = n \sin\alpha$$

$n_{air} \approx 1$, so $NA \leq 1$

SINGLE SILICON ATOMS ! (TEM)

STI structures (TEM)

← in-situ spectroscopy (SEM)

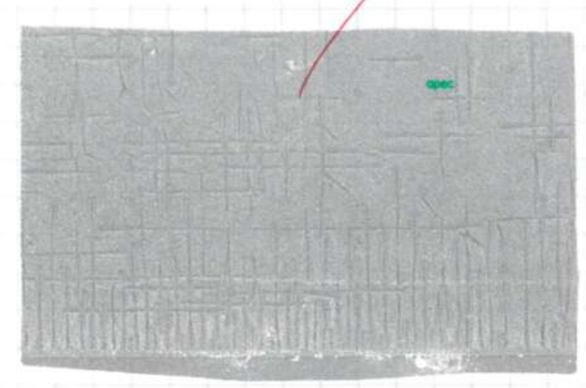
false colors

BLUE : oxygen containing materials

GREEN : cobalt containing materials

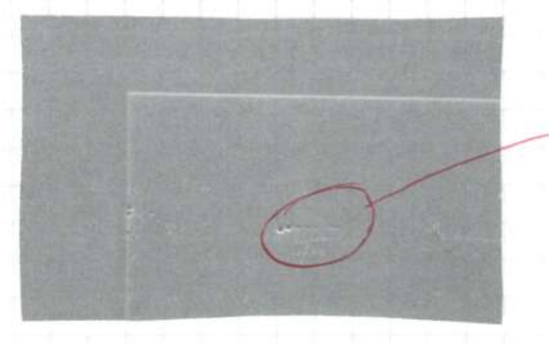
RED : titanium containing materials

extreme case of damaged silicon



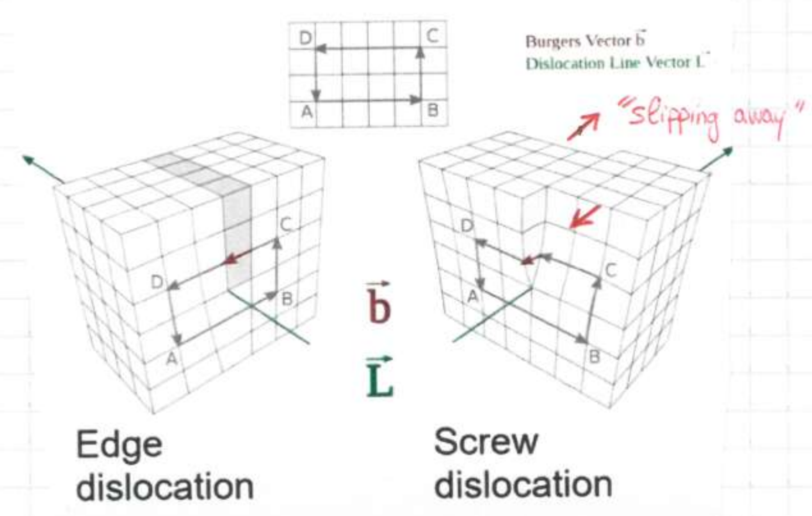
We can also spot dislocations !

Secco D'Aragona etching : $K_2 Cr_2 O_7 + HF$ has the property of etching away faster the damaged / deformed silicon, so we can see it with our SEM



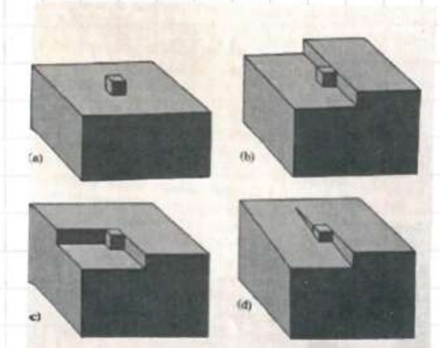
small damage

Additional notes on dislocations



We have already seen edge dislocations the previous lesson.
A screw dislocation is as if a portion of the plane "slipped away", so if you do a roundtrip you won't end up on the same plane as before but a little higher / lower. On an edge dislocation we ended up on the same plane as we started.

Dislocations are crucial in crystal growth, since atoms prefer to be surrounded by other atoms on all sides, favouring a spiral growth compared to a layer-by-layer growth.



(a) perfect crystal plane
(b) step between two planes
(c) corner
(d) screw dislocation

Atoms are relatively weakly attracted to perfect crystal planes (a), they're more strongly attracted to a step between two planes (b), and are most strongly attracted to a corner (c).
If the crystal contains a screw dislocation (d), then by adding atoms (as shown in the picture) the local planar structure can spiral endlessly around the dislocation.
Crystals can grow much more rapidly in this way, since the nucleation of new planes by the process shown in (d) is never required.

Calculation of point defects concentration

Let's focus on vacancy concentration (analogous for interstitials)

SINGLE VACANCY FORMATION ENERGY: $\left[G_F = H_F - T S_F \right]$

- G_F: free energy
- H_F: formation enthalpy
- T: temperature
- S_F: formation entropy (local, small scale)

Now we want to form n vacancies in my crystal, what would be the free energy associated with that configuration?

$$G = G_0 + n \cdot G_F - T S_{CONF}$$

- G_0: perfect crystal free energy
- S_{CONF}: configurational entropy (large scale entropy)

S_{CONF}: when I remove 3 (for example) vacancies, depending on where they are there will be a different entropy associated.

$\neq$ $\neq$

always 3 vacancies but different configuration

Since the crystal will tend to minimize the free energy G, by doing $\frac{\partial G}{\partial n} = 0$ we can find the number of vacancies associated with that minimum.

$$G = G_0 + n G_F - T S_{CONF}$$

$$\frac{\partial G}{\partial n} = 0 \rightarrow \frac{\partial}{\partial n} \left(G_0 + n G_F - T S_{CONF} \right) = 0$$

$\frac{\partial G_0}{\partial n} = 0$ since G_0 doesn't depend on n

$$\frac{\partial (n G_F)}{\partial n} = G_F$$

$$\frac{\partial (T S_{CONF})}{\partial n} = ?$$

we know (from statistical physics) that $S_{CONF} = K \ln \binom{N}{n} = K \ln \left(\frac{N!}{(N-n)!\, n!} \right)$

(K: Boltzmann constant; $\binom{N}{n}$: number of possible configurations of n vacancies over N atoms in my crystal)

(Stirling approximation $\ln x! \cong x \ln x$)

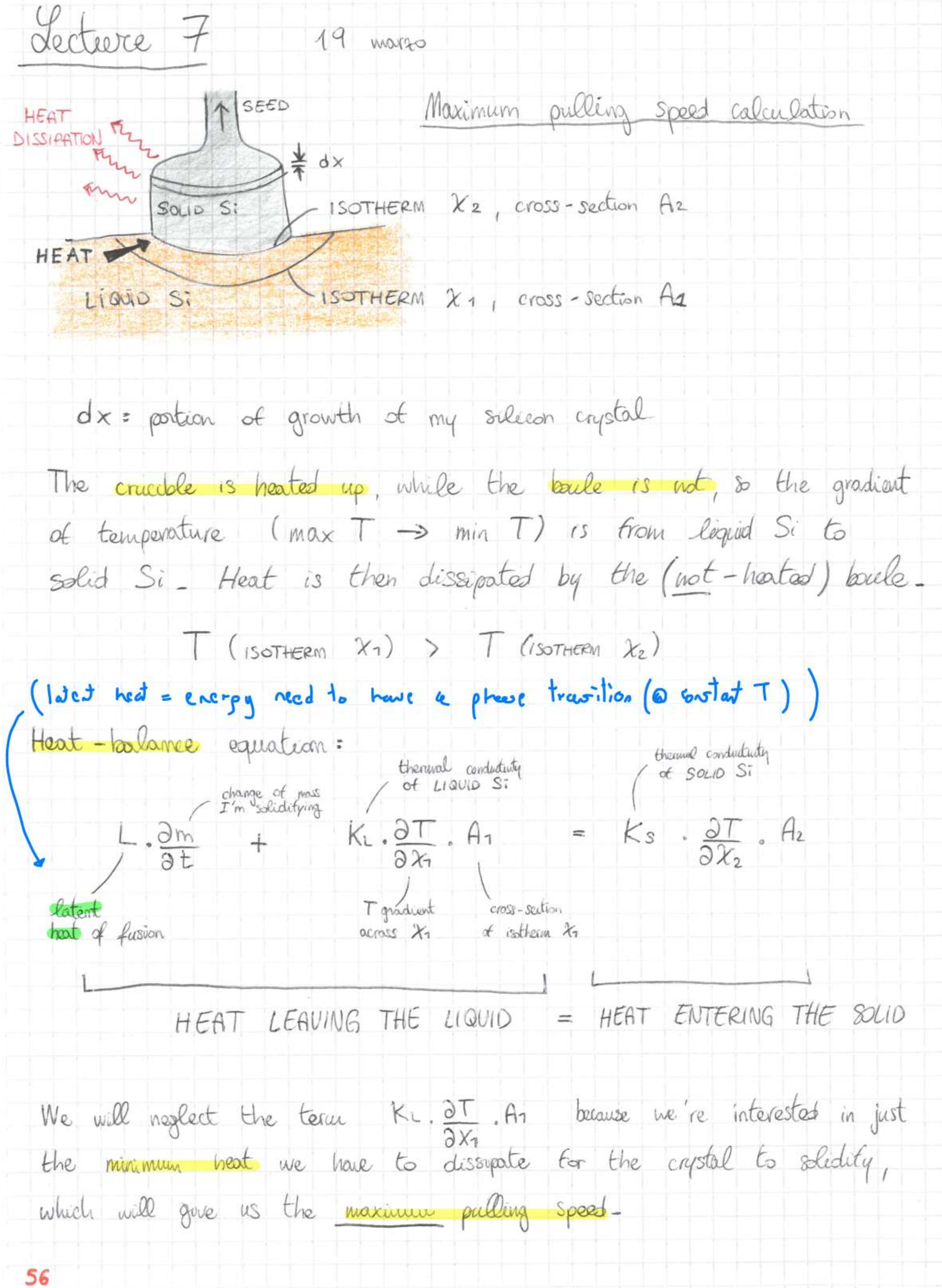
$$\frac{\partial}{\partial n} \left(K \ln \left(\frac{N!}{(N-n)!\, n!} \right) \right) \cong K \frac{\partial}{\partial n} \left[N \ln N - n \ln n - (N-n) \ln (N-n) \right] =$$

$$= K \left[-\ln n - \frac{n}{n} + \frac{N}{N-n} + \ln (N-n) - \frac{n}{N-n} \right] = -K \ln \left(\frac{n}{N-n} \right) \cong$$

assuming $n \ll N$ we get $\cong -K \ln \frac{n}{N} = -K \ln C_v$ ← vacancy concentration

$$\frac{\partial G}{\partial n} = 0 \rightarrow G_F + K T \ln C_v = 0 \rightarrow C_v = e^{-\frac{G_F}{KT}}$$

$$\boxed{C_v = e^{\frac{S_F}{K}} e^{-\frac{H_F}{KT}}}$$



Lecture 7

19 marzo

Maximum pulling speed calculation

dx = portion of growth of my silicon crystal

The crucible is heated up, while the boule is not, so the gradient of temperature (max T → min T) is from liquid Si to solid Si. Heat is then dissipated by the (not-heated) boule.

$$T(\text{isotherm } \chi_1) > T(\text{isotherm } \chi_2)$$

(latent heat = energy need to have a phase transition (@ constant T))

Heat-balance equation:

$$L \cdot \frac{\partial m}{\partial t} + K_L \cdot \frac{\partial T}{\partial \chi_1} \cdot A_1 = K_S \cdot \frac{\partial T}{\partial \chi_2} \cdot A_2$$

- L: latent heat of fusion
- $\frac{\partial m}{\partial t}$: change of mass I'm solidifying
- K_L: thermal conductivity of LIQUID Si
- $\frac{\partial T}{\partial \chi_1}$: T gradient across χ_1
- A_1: cross-section of isotherm χ_1
- K_S: thermal conductivity of SOLID Si

HEAT LEAVING THE LIQUID = HEAT ENTERING THE SOLID

We will neglect the term $K_L \cdot \frac{\partial T}{\partial \chi_1} \cdot A_1$ because we're interested in just the minimum heat we have to dissipate for the crystal to solidify, which will give us the maximum pulling speed.

we can also express $\frac{\partial m}{\partial t}$ in function of the pulling speed v_P

$$\frac{\partial m}{\partial t} = v_P \cdot A \cdot \rho_{Si}$$

we can assume $A_1 = A_2 = A$

ρ_{Si}: density of Silicon

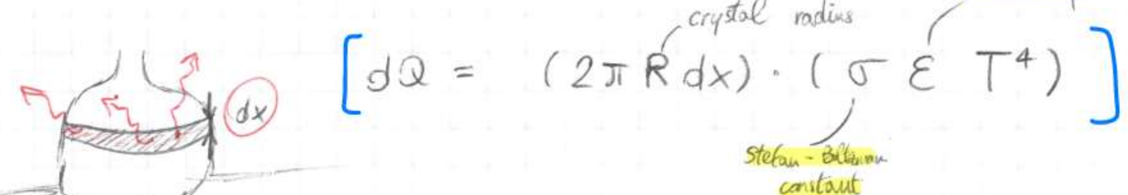
$$L \frac{\partial m}{\partial t} \cong K_S \frac{\partial T}{\partial x_2} A \longrightarrow \frac{\partial m}{\partial t} \cong \frac{K_S}{L} \frac{\partial T}{\partial x_2} A = v_P A \rho_{Si}$$

$$\left[v_{P_{MAX}} = \frac{K_S}{L \rho_{Si}} \frac{\partial T}{\partial x_2} \right]$$

$\frac{\partial T}{\partial x_2}$: ISOTHERM AT LIQUID - SOLID INTERFACE

we would like an expression that didn't contain $\partial T / \partial x_2$, but only geometrical factors and the melting temperature T_m of silicon.

Since all the heat we're dissipating is through radiation, we can write the Stefan - Boltzmann radiation loss equation:



$$\left[dQ = (2\pi R\, dx) \cdot (\sigma \, \varepsilon \, T^4) \right]$$

R: crystal radius; ε: emissivity of Silicon; σ: Stefan - Boltzmann constant

this new surface $(2\pi R\, dx)$ can now dissipate extra heat dQ !

We have seen that the heat conducted to the crystal is $K_S A \frac{\partial T}{\partial x}$ and we know that $A = \pi R^2$ so we get

$$Q = k_S (\pi R^2) \frac{\partial T}{\partial x}$$

and now we derivate

$$\frac{\partial Q}{\partial x} = K_S (\pi R^2) \frac{\partial^2 T}{\partial x^2} + \cancel{(\pi R^2) \frac{\partial T}{\partial x} \cdot \frac{\partial K_S}{\partial x}}$$

we neglect this, usually small

but also $dQ = (2\pi R \, dx) \cdot (\sigma \varepsilon T^4) \rightarrow \frac{dQ}{dx} = 2\pi R \sigma \varepsilon T^4$

$$\boxed{\frac{\partial^2 T}{\partial x^2} - \frac{2\sigma\varepsilon}{K_S R} T^4 = 0}$$

thermal conductivity at the melting temp. T_m

for $T \lessapprox 1000\,°C \rightarrow K_S \simeq K_M \frac{T_m}{T}$

$$\frac{\partial^2 T}{\partial x^2} - \frac{2\sigma\varepsilon}{K_m R T_m} T^5 = 0$$

from this we can get T, then $\frac{dT}{dx}$ and finally $v_{P_{MAX}}$

$$\left[v_{P_{MAX}} = \frac{1}{LN} \sqrt{\frac{2\sigma\varepsilon K_M T_m^5}{3R}} \right] \sim \sqrt{\frac{1}{R}}$$

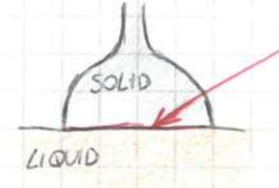
the larger the crystal, the slower I must pull

IMPORTANT: what we have to remember is the physical meaning behind this. We are transfering heat from the melt to the solid and the solid must dissipate that heat in order to freeze out. The minimum heat it must dissipate is the latent heat of fusion, which we assume moves to the crystal just by thermal conduction, NO convection. And is entirely dissipated by radiation (again NO convection). By equating heat IN = heat OUT and calculate how much heat is radiated by each incremental surface that I'm growing, we can get our pulling speed, which is a function of the material I'm growing (σ, ε, T_m, ...) and the geometry of the boule ($\propto \sqrt{\frac{1}{R}}$).

Dopant segregation in CZ

cross-section →

$$K_0 = \frac{C_S}{C_L} \quad \frac{\text{solid}}{\text{liquid}}$$



AT THE SOLID-LIQUID INTERFACE
the concentration of impurities in solid / concentration of impurities in liquid = segregation coefficient

The segregation coefficient K_0 is constant and is usually < 1, so dopants that we're interested in have higher solubility in liquid than in solid.

As we keep growing the crystal, more liquid Si solidifies than impurities, so the concentration increases from top to bottom.

$$C_S = C_0 K_0 (1 - f)^{K_0 - 1}$$

- C_S: concentration in the solid
- C_0: starting concentration in liquid Si
- K_0: segregation coefficient
- f: fraction of liquid Si that solidified

START:

initial volume V_0

initial impurity number I_0 (number of impurities)

initial concentration $C_0 = I_0 / V_0$

if now we freeze a volume dV, what will be the variation in the number of impurities that we'll have in the liquid?

$$dI = -C_S \, dV = -(K_0 C_L) \, dV$$

($C_S \, dV$: number of impurities we are subtracting to the liquid)

$$C_L = \frac{I_L}{V_L} = \frac{I_L}{V_0 - V_S} \longrightarrow dI = -K_0 \frac{I_L}{V_0 - V_S} dV$$

$$\int_{I_0}^{I_L} \frac{dI}{I_L} = -K_0 \int_0^{V_S} \frac{dV}{V_0 - V_S}$$

$$\ln\left(\frac{I_L}{I_0}\right) = \ln\left(1 - \frac{V_S}{V_0}\right)^{K_0}$$

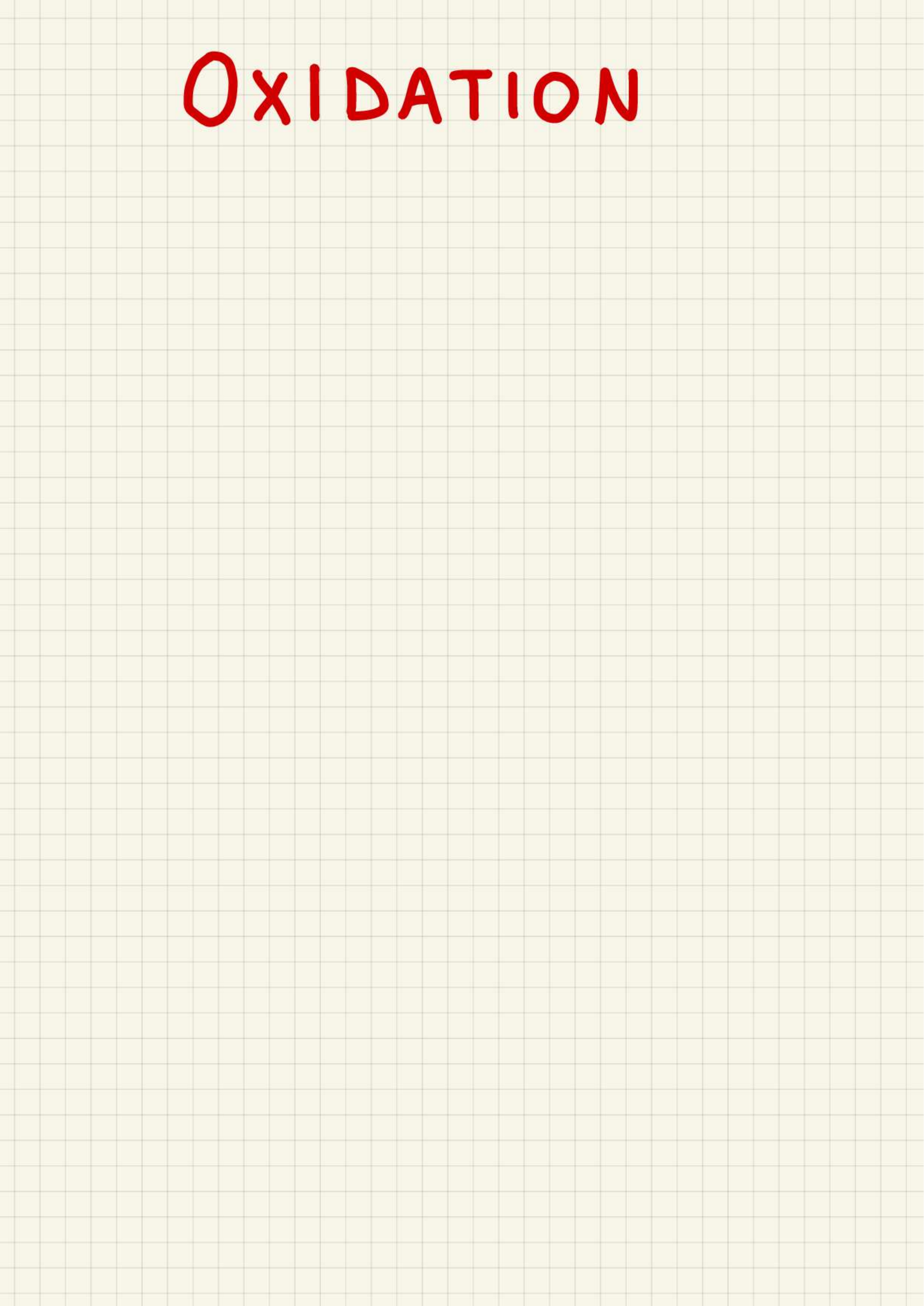
$$I_L = I_0 \left(1 - \frac{V_S}{V_0}\right)^{K_0}$$

number of impurities in the liquid as a function of crystal growth

to get the concentration in the solid : $C_S = -\frac{dI_L}{dV_S}$

$$\boxed{C_S = C_0 K_0 \left(1 - \frac{V_S}{V_0}\right)^{K_0 - 1}}$$

↑ fraction of solidified Si


OXIDATION

SILICON OXIDATION

Silicon oxidation is the formation of silicon dioxide (SiO_2) on silicon surface, in an oxidant ambient (O_2 or H_2O).

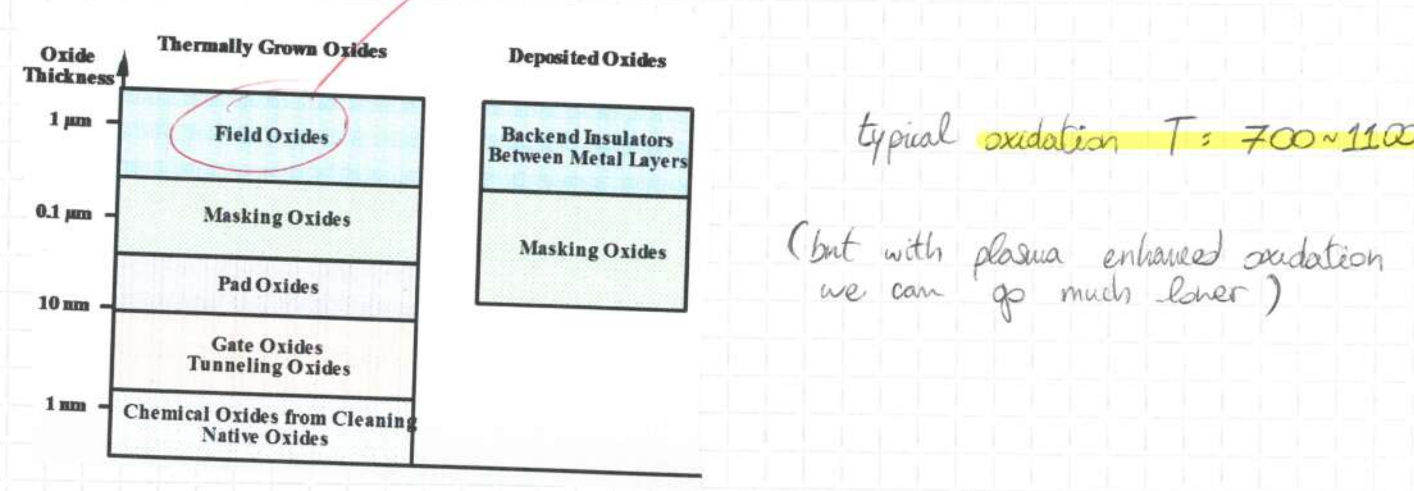
Remember that SiO_2 can also be deposited on silicon.

dry oxidation $Si + O_2 \rightarrow SiO_2$

wet oxidation $Si + H_2O \rightarrow SiO_2 + 2H_2$

SiO_2 can be found basically everywhere in an IC manufacturing facility. There are many kind of oxides, varying in thickness, in growth/deposition, ... For example silicon can also oxidize in air at room T, but O_2 won't have enough energy to diffuse through SiO_2 after a first thin oxide layer is formed, so it won't grow more than 2 nm ("native oxide").

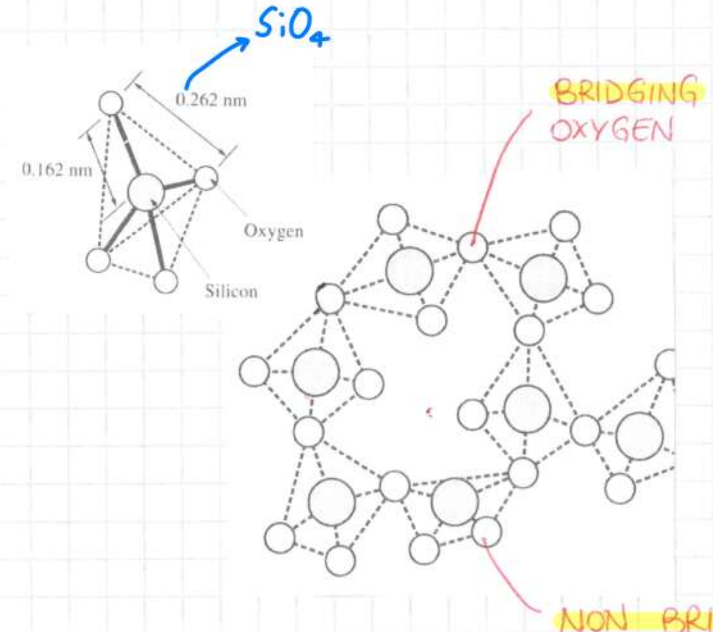
now we know that recent field oxides can be deposited (STI) instead of grown (LOCOS)



typical oxidation T: 700~1100°C

(but with plasma enhanced oxidation we can go much lower)

SiO_2 is grown as "fused silica" (amorphous SiO_2 with low range order). No long range order like we would have in a crystal.
Crystalline SiO_2 (= quartz) can't be grown on top of Si because of lattice mismatch.
The structure will be SiO_4 tetrahedra, with some of those oxygen atoms being bridging oxygen (shared between neighbouring Si atoms)

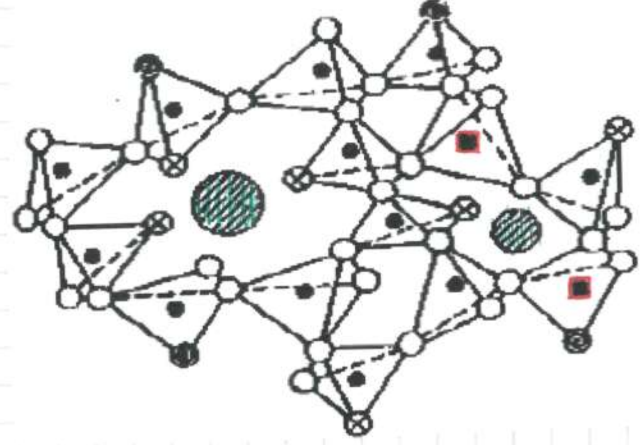


Higher fraction of bridging oxygen compared to non bridging means higher cohesion and structural strength of SiO_2

dry oxide > wet oxide (more bridging = more strength)

Impurities in SiO_2

- Bridging Oxygen
- Non-Bridging Oxygen
- Silicon
- Network Modifier — interstitial impurity
- Hydroxyl Group
- Network Former — substitutional impurity



SiO_2 structure can be modified by:

- **SUBSTITUTIONAL IMPURITIES**

  they replace silicon in the structure, so can eliminate or form bridging oxygens.

  they're also called network formers

- **INTERSTITIAL IMPURITIES**

  they increase the amount of NON bridging oxygen, making the structure weaker, more porous and they also increase the diffusivity of other species within the SiO_2

  also called network modifiers

We could also have water in SiO_2: $Si-O-Si \rightarrow Si-O-H + H-O-Si$

(bridging oxygen (strong) → bridging hydrogen (weak))

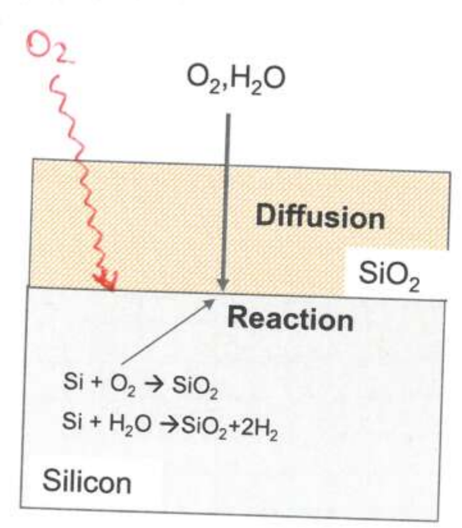
Hydroxyls $(-OH)$ are generated and substitute bridging oxygen.
The H bond is quite weak and can be easily removed, so charged states or traps can be generated.

For this reasons, water in SiO_2 is something that has to be taken care of.

SILICON OXIDATION

Now that we've seen the silicon oxide, let's see the oxidation process.

IMPORTANT : the oxidation process takes place at the Si/SiO_2 interface.

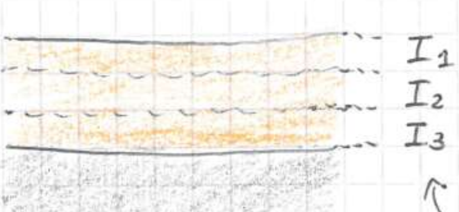


For the process to continue once the first layer of SiO_2 formed, my O_2 (or H_2O) molecules have to diffuse through the SiO_2 until they reach the surface of the silicon, where they can react.

This has been demonstrated (e.g. by oxygen isotopes profiling)

NOTE : this is not true for SiO_2 deposition, where the SiO_2 layers stack on top of each other

Oxygen isotopes profiling : during the growth we start with a given oxygen isotope (call it I_1), then switch (I_2), then I_3, ...



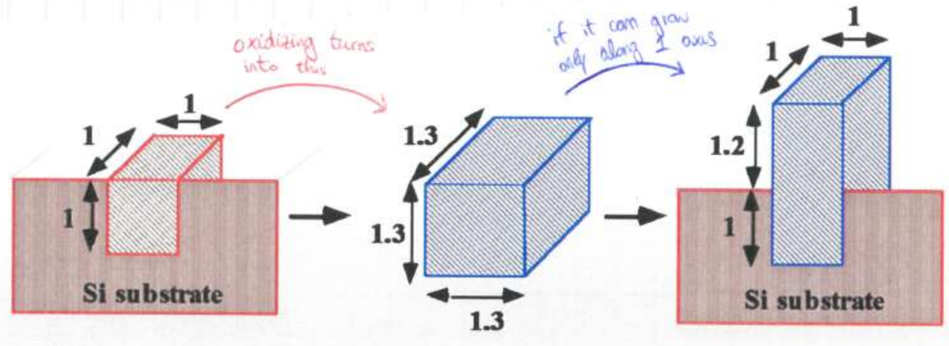
↖ the latest isotopes are found at the bottom, so they have diffused through SiO_2 to reach the surface of the silicon

NOTE: the diffusivity depends exponentially on T, so at room T oxygen will have very poor diffusivity → thin oxide.

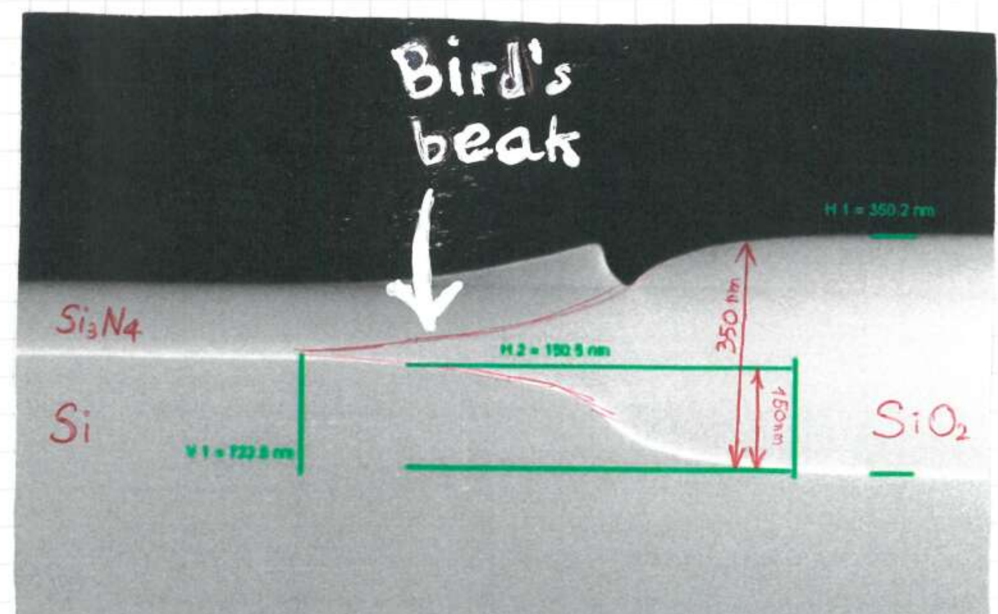
OXIDATION KINETICS

Silicon oxidation involves a volume expansion:

~46% of oxide thickness grows INTO the silicon, which means that the silicon surface is "pushed" away for a 46% of the oxide thickness. Mechanical stress can play a determinant role in oxidation kinetics.



LOCOS oxidation

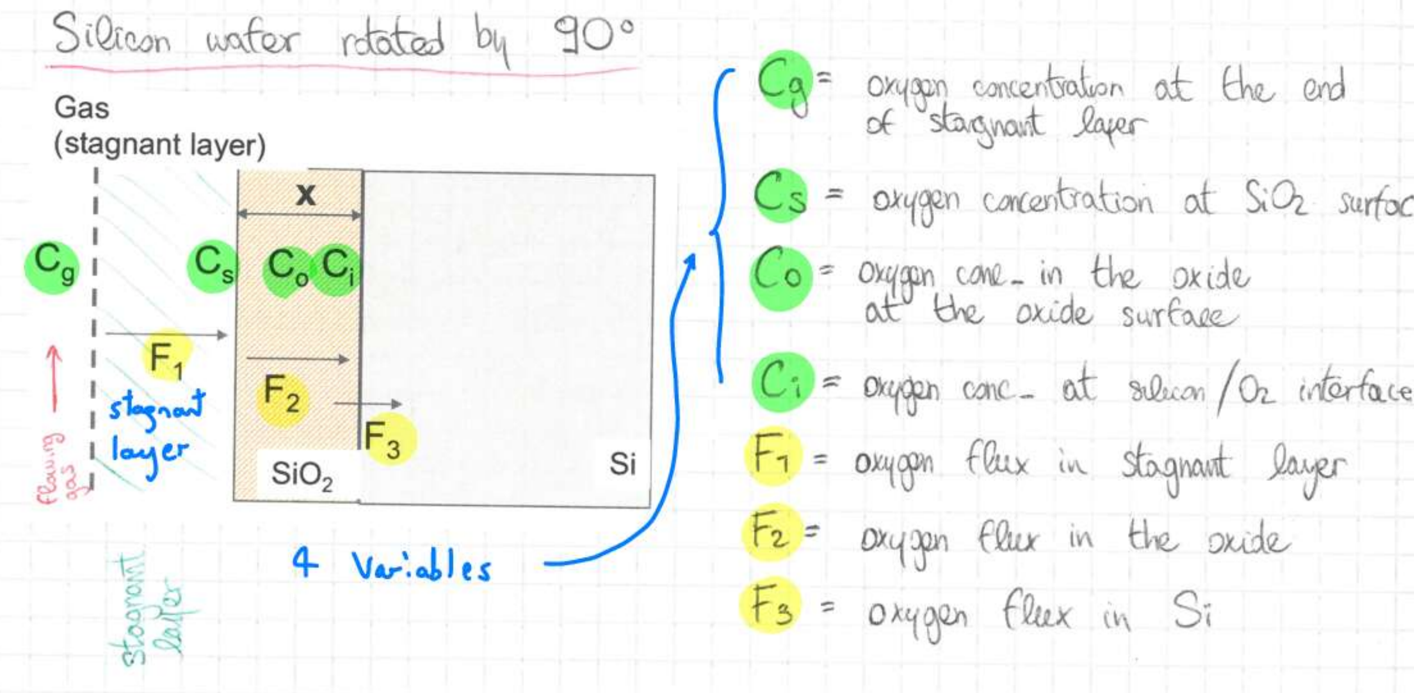


We can notice the bird's beak where the SiO_2 has infiltrated under the Si_3N_4 and lift it up.

Also, of the 350 nm of SiO_2 thickness, 150nm are inside the silicon.

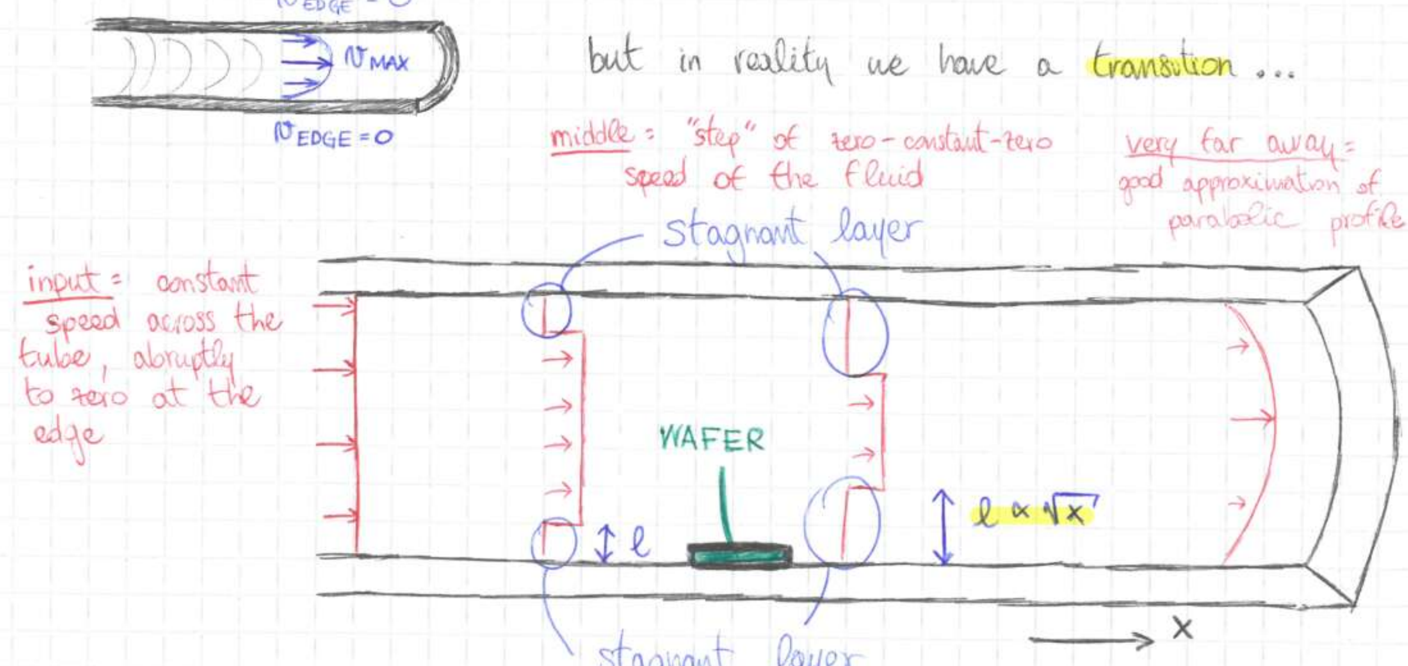
DEAL - GROVE MODEL

The Deal - Grove model is perhaps the most well - known model of silicon oxidation.



What is the "stagnant layer"?

Let's imagine a tube with a fluid moving inside of it. We know that the speed profile will be parabolic, with $v_{EDGE} = 0$.



NOTE: the height of the stagnant layer (ℓ) goes as $\sqrt{x}$

An oxygen molecule O_2 moving in the gas flow, to partecipate to the oxidation, must do 3 things:

1. move from the flowing gas, through the stagnant layer, to the surface of the silicon oxide SiO_2.
2. diffuse through the SiO_2 until it reaches the surface of silicon.
3. chemically react with silicon to form SiO_2

Gas (stagnant layer)
C_g C_s C_o C_i x
O_2 F_1 F_2 F_3
SiO_2 Si

the flux F_1 of oxygen molecules O_2 from the bulk of the gas, through the stagnant layer to the surface of SiO_2 can be written as mass transfer through the stagnant layer (in gas phase):

NOT exp. dependent on T

$$\left[F_1 = h_g \, (C_g - C_s) \right]$$ between gas flow and stagnant layer

h_g: mass transfer constant (property of the gas we're using, pressure, ...)

I believe it's Fick's

F_2: the flux through the solid (solid diffusion) can be described by Figg's law. So it depends on the gradient of concentration

movement is from more stuff → to less stuff

in SiO_2 $$\left[F_2 = -D \frac{\partial C}{\partial x} = D \left(\frac{C_o - C_i}{x} \right) \right]$$

D: "diffusivity" (constant)

at steady state we can write this (no time dependance)

$$\left[F_3 = K_s \, C_i \right] \text{ in } S_i$$

chemical reaction constant (K_s)

the third flux is simply the rate of chemical reaction, which depends on the concentration of O_2 available

NOTE: both D (diffusivity) and K_s have an exponential dependence on temperature, h_g NO (it's not a solid diffusion but a gas diffusion, completely different).

We have 4 variables (C_g, C_s, C_o, C_i) ← BUT ONLY 3 EQUATIONS. What we're missing is a way to describe $C_s \rightarrow C_o$, so how can O_2 enter the SiO_2 surface. We described $C_g \rightarrow C_s$, $C_o \rightarrow C_i$, $C_i \rightarrow SiO_2$ formation, but not how the O_2 molecules can go from outside SiO_2 to inside SiO_2.

SOLUTION: HENRY'S LAW

The concentration of gas in a solid (or a liquid) is proportional to the partial pressure of that same gas outside the surface of the solid (or liquid)

$$\left[C_o = H \, P_s \right]$$

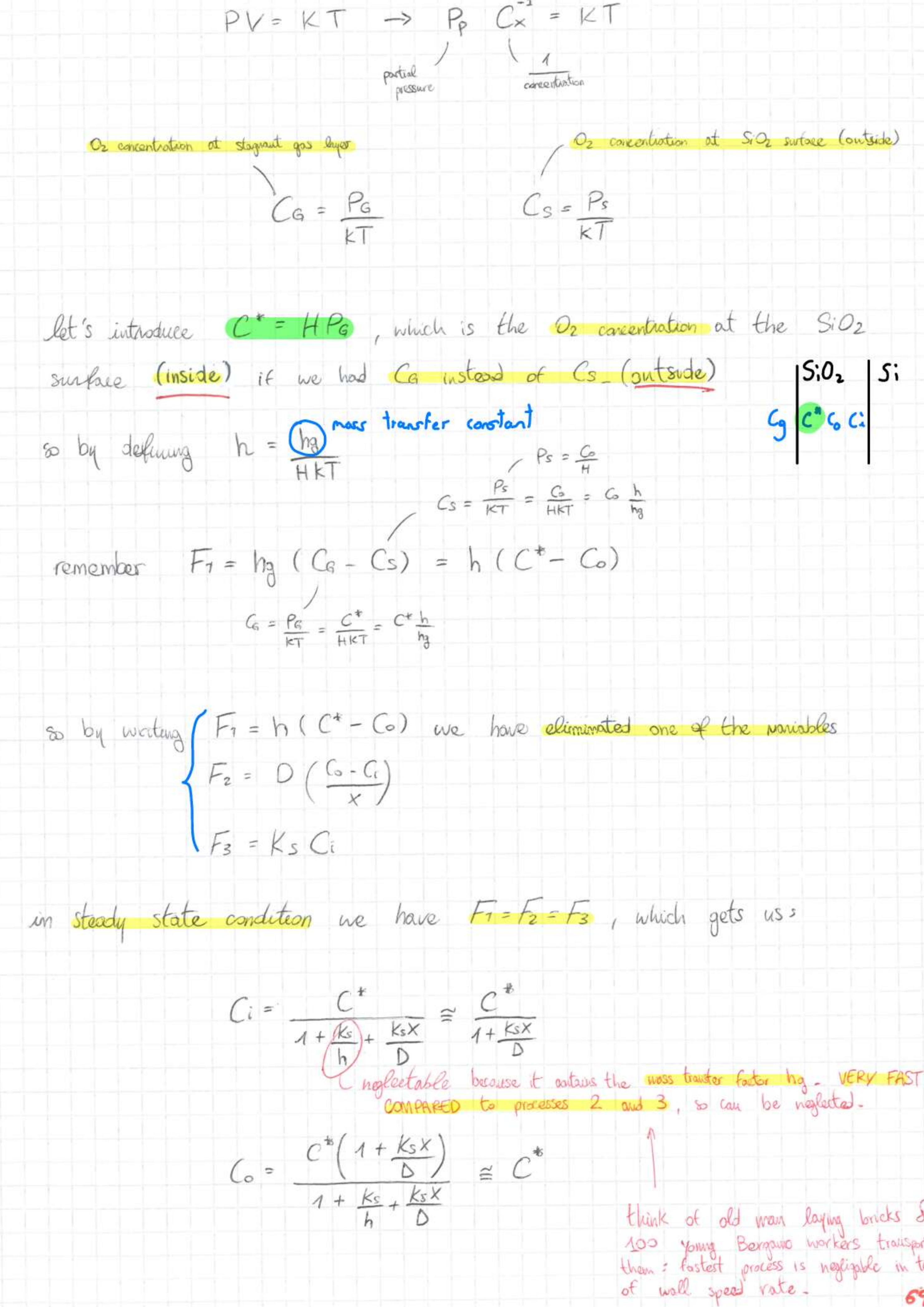
H: Henry's constant; P_s: partial pressure at oxide surface

(so that's why beer, champagne, sparkling water, ..., will stay carbonated only if I close the bottle)

Next: Perfect Gas Law

$PV = kT$ but instead of P we use P_p (partial pressure)

then we have to use C_x^{-1} (concentration)$^{-1}$ instead of V.



$$PV = KT \rightarrow P_p \, C_x^{-1} = KT$$

(P_p: partial pressure; C_x^{-1}: $\frac{1}{\text{concentration}}$)

O_2 concentration at stagnant gas layer:

$$C_G = \frac{P_G}{kT}$$

O_2 concentration at SiO_2 surface (outside):

$$C_S = \frac{P_s}{kT}$$

let's introduce $C^* = HP_G$, which is the O_2 concentration at the SiO_2 surface (inside) if we had C_G instead of C_S (outside)

SiO_2 | Si

C_g | C^* C_0 C_i |

so by defining $h = \frac{h_g}{HKT}$ (h_g: mass transfer constant)

$$P_s = \frac{C_0}{H}$$

$$C_S = \frac{P_s}{KT} = \frac{C_0}{HKT} = C_0 \frac{h}{h_g}$$

remember $F_1 = h_g \left(C_G - C_S \right) = h \left(C^* - C_0 \right)$

$$C_G = \frac{P_G}{KT} = \frac{C^*}{HKT} = C^* \frac{h}{h_g}$$

so by writing
$$\begin{cases} F_1 = h\left(C^* - C_0\right) \\ F_2 = D\left(\frac{C_0 - C_i}{x}\right) \\ F_3 = K_s C_i \end{cases}$$
we have eliminated one of the variables

in steady state condition we have $F_1 = F_2 = F_3$, which gets us:

$$C_i = \frac{C^*}{1 + \frac{K_s}{h} + \frac{K_s X}{D}} \cong \frac{C^*}{1 + \frac{K_s X}{D}}$$

($\frac{K_s}{h}$:) neglectable because it contains the mass transfer factor h_g - VERY FAST COMPARED to processes 2 and 3, so can be neglected.

$$C_0 = \frac{C^*\left(1 + \frac{K_s X}{D}\right)}{1 + \frac{K_s}{h} + \frac{K_s X}{D}} \cong C^*$$

think of old man laying bricks & 100 young Bergamo workers transporting them: fastest process is negligable in terms of wall speed rate.

$C_o \cong C^*$ means that the concentration at the SiO_2 surface isn't very different from that of O_2 in the stagnant layer, since the O_2 can't be depleted fast enough from the surface to get C_G different from C_S.

Let's now compare reaction rate ($K_S X$) and diffusion rate (D):

$$C_i \simeq \frac{C^*}{1 + \frac{K_S X}{D}}$$

LIMITING FACTOR — slow reaction rate; fast diffusivity

for $K_S X \ll D \rightarrow C_i \cong C^*$ REACTION CONTROLLED REGIME

for $K_S X \gg D \rightarrow C_i \cong 0$ DIFFUSION CONTROLLED REGIME

fast reaction rate; slow diffusivity LIMITING FACTOR

- $C_i \cong C^*$ means that the O_2 gets reacted ("removed") much more slowly than it can be replenished, so it will be $\cong C^*$ (SiO_2 | Si) almost constant
- $C_i \cong 0$ means that as soon as O_2 reaches the Si surface, it reacts and gets depleted very fast. (SiO_2 | Si) almost empty

The growth rate ($\partial x / \partial t$) is simply the flux F_3 divided by the number of oxidant molecules incorporated per unit volume of oxide grown (N)

$$\frac{\partial x}{\partial t} = \frac{F_3}{N} = \frac{K_S C^*}{N\left(1 + \frac{K_S}{h} + \frac{K_S X}{D}\right)} \xrightarrow{\text{separation of variables}} N\int_{x_i}^{X}\left(1 + \frac{K_S}{h} + \frac{K_S X}{D}\right)dx = K_S C^* \int_0^t dt$$

$\rightarrow \frac{\partial x}{\partial t} = \frac{B}{2X + A}$

$$\frac{X^2}{B} + \frac{X}{(B/A)} = t + \tau \quad \text{where:} \quad B = \frac{2DHP_G}{N} \text{ parabolic rate constant}$$

$$\frac{B}{A} = \frac{HP_G}{N\left(\frac{1}{K_S} + \frac{1}{h}\right)} \text{ linear rate constant}$$

if existing oxide is present: τ is the time it would've taken to grow that existing oxide

Lecture 8

24 marzo

$$\frac{x^2}{B} + \frac{x}{B/A} = t + \tau$$

with solution

$$x = \frac{A}{2}\left[\sqrt{1 + \frac{4B(t+\tau)}{A^2}} - 1\right]$$

so we can see that the thickness doesn't grow linearly with time!
It has a parabolic growth rate = it gets slower with time.

Why? Because as the SiO_2 grows thicker, it's not anymore the reaction rate that's slowing down the process but the concentration of O_2, since they can't diffuse easily through a thicker oxide.

$$B = \frac{2DHP_G}{N}$$

the parabolic term contains diffusivity, H, P_G, ..
so it's related to the solid state diffusion of oxygen into SiO_2

$$\frac{B}{A} = \frac{HP_G}{N\left(\frac{1}{K_s} + \frac{1}{h}\right)}$$

the linear term B/A is more related to the chemical reaction, contains also h ($\to h_g$), but is negligible (because it's very fast process)

How can we get B and B/A?

They have both a dependence on temperature T, so we can in principle write

$$B = C_1 e^{-\frac{E_1}{KT}}$$

$$B/A = C_2 e^{-\frac{E_2}{KT}}$$

Ambient	B	B/A
Dry	C_1=7.72E2 μm²hr⁻¹ E_1=1.23eV	C_2=6.23E6 μm²hr⁻¹ E_1=2.0eV
Wet	C_1=3.86E2 μm²hr⁻¹ E_1=0.78eV	C_2=1.63E8 μm²hr⁻¹ E_2=2.05eV

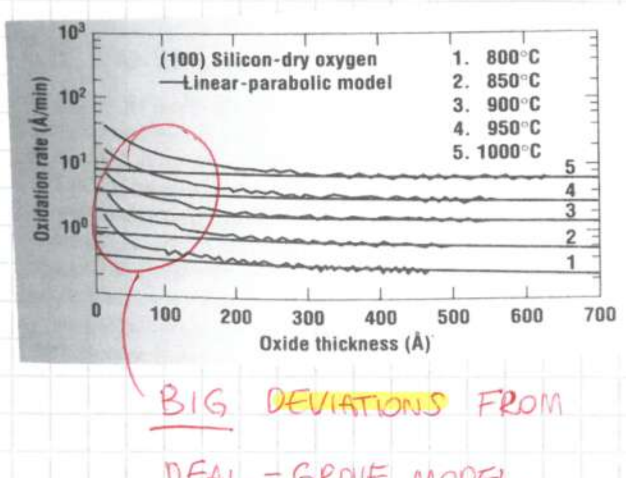
C_1, C_2, E_1, E_2 are tabulated, so we can look up B and B/A on a book.

NOTE: B and B/A are different for dry and wet oxidations

E_1 (wet) < E_1 (dry), E_2 (dry) ≅ E_2 (wet)

NOTE: this numbers are valid for (111) orientation, for (100) we must divide C_2 by 1,68 because chemical rate depends on atomic arrangements.

PROBLEM: the Deal-Grove model does NOT work for thin oxides



to model the growth of thick oxides (>30 ~ 40 nm) the Deal-Grove model works pretty well, below that thickness no (there's a substantial deviation even though we're seeing a LOG scale, so in LINEAR scale is much worse).

We yet don't know why thin oxides grow at faster rates, some ideas are:

- O_2^- ionic diffusion, which is boosted in very thin oxides
- thermoionic emission of electrons from Si into SiO_2, enhancing the reaction
- micro-pores in SiO_2 increasing oxidation rate
- more than one atomic layer of SiO_2 participating to oxidation

and many more, but no definitive answer.

There are also many models (empirical). The most used is the MASSOUD model

$$\left[\frac{\partial x}{\partial t} = \frac{B}{2x + A} + C e^{-\frac{x}{L}} \right] \quad \text{Massoud model}$$

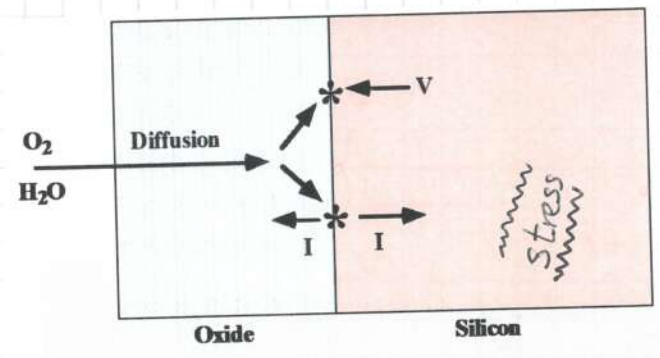
(Deal-Grove term: $\frac{B}{2x+A}$)

NOTE: the Massoud model is a correction to the Deal-Grove model, but it's empirical, there's no physical reasoning behind the correction, just there to take into account deviation from data.

NOTE: despite there not being a working model for the growth of thin SiO_2, IC manufacturers can grow very thin SiO_2 layers with extreme precision and repeatability across billions of wafers, and have been doing so for decades.

Other factors that influence oxidation rates

All kinds of substrate doping increase oxidation rate. Why?



- vacancies recombined
- interstitials pushed inside

When we oxidate silicon we "consume" vacancies and produce interstitials in my substrate. This is because silicon oxidation is associated with volume expansion, so if we have vacancies (= empty space) my oxidation will be "easier" because SiO_2 will have more space to accomodate as you're growing it. So the higher the vacancy level → higher oxidation rate.

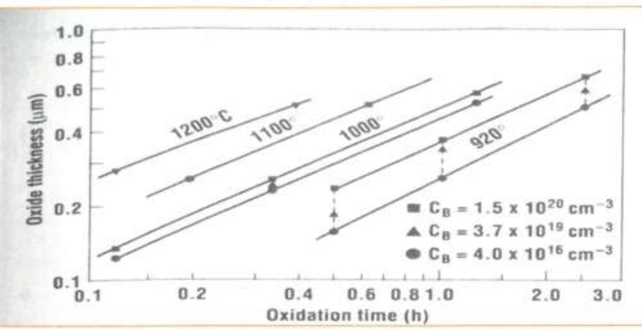
ANOTHER POV: to oxidize silicon we have to break Si-Si bonds. a Si atom with less bonds (= close to a vacancy) will be easier to oxidize.

Also, by SiO_2 formation we inject interstitials into the substrate. This happens because, due to volume expansion, Si atoms get "pushed away", so they increase the number of interstitials.

So we can write a modified microscopic oxidation reaction:

$$(1+2\gamma)Si + 2O + 2\beta V \leftrightarrow SiO_2 + 2\gamma I + \text{stress}$$

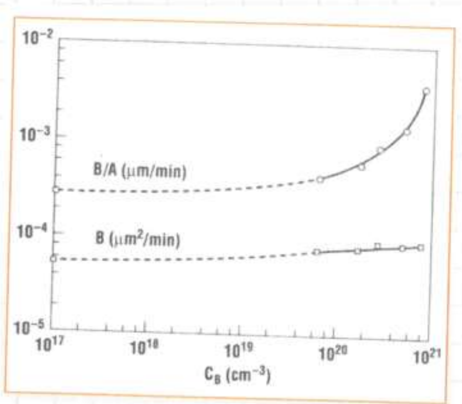
silicon is reacting with oxygen and vacancies to create SiO_2, interstitials and mechanical stress



→ Boron is incorporated in SiO_2 and weakens the oxide structure, enhancing oxidant diffusion

B segregates in SiO_2

Since B tends to segregate into SiO_2, it's incorporated in SiO_2. This weakens the structure of SiO_2, enhancing O_2 diffusion in SiO_2 which accelerates the oxidation rate of the substrate.

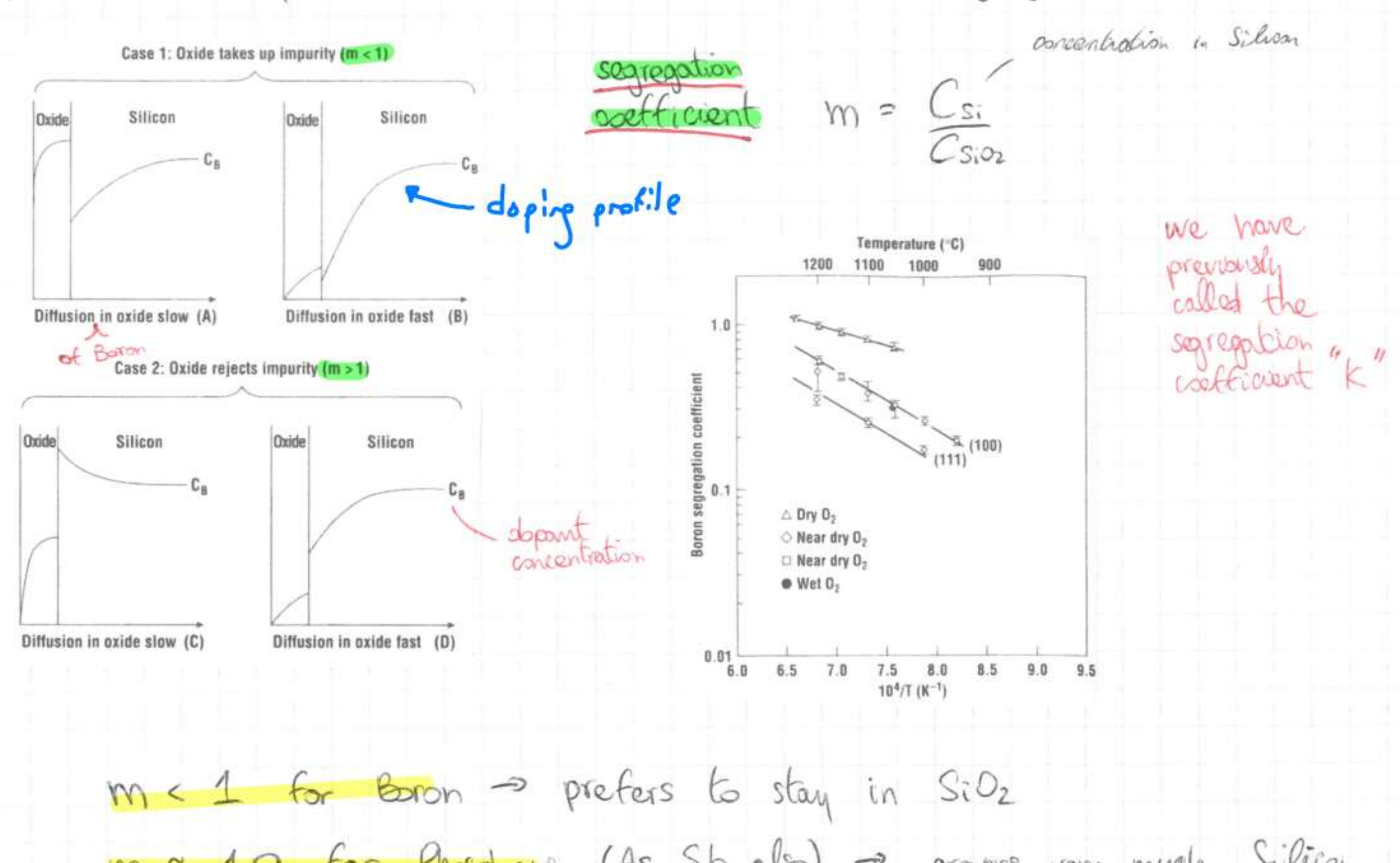


→ also Phosphorous increases the oxidation rate:
P doping shows a huge B/A increase!

P segregates in Si

P segregates into silicon (not SiO_2), and once they're there, they increase the concentration of point defects (vacancies), increasing the number of oxidation sites available → faster oxidation rate.
Why? Because Phosphorous doping shifts the fermi level E_F, increasing vacancy concentration.

The fact that we are oxidating the silicon will change the doping profile in my silicon substrate because of segregation (m)



$m < 1$ for Boron → prefers to stay in SiO_2

$m \simeq 10$ for Phosphorous (As, Sb also) → prevers very much Silicon

For $m < 1$ (case 1, picture on the left) we have a depletion of dopants from the substrate, at the Si / SiO_2 interface.

For $m > 2$ (case 2) we have a depletion of dopants from the SiO_2, and depending on the speed of diffusion of dopants in SiO_2 we can have an accumulation of dopants at the interface (slow diffusion), or a depletion of dopants at the interface (fast diffusion).
Nevertheless, for case 2 we'll always have $C_{Si} > C_{SiO_2}$ ($m > 1$).

BORON → DEPLETION OF DOPANTS AT Si / SiO_2 INTERFACE

PHOSPHOROUS → ACCUMULATION OF DOPANTS AT Si / SiO_2 INTERFACE

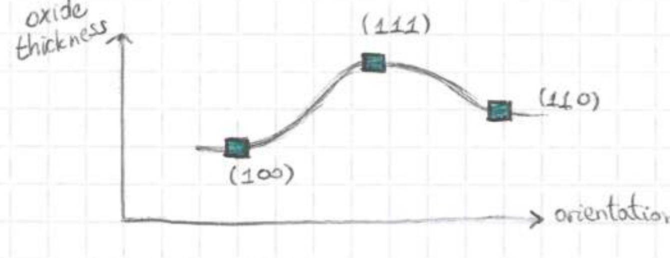
NOTE = in both cases we'll see an increase in resistivity at the interface.
For Boron this happens simply because we have less dopants in Si.
For Phosphorous this is because P atoms are NOT in substitutional position!, so aren't electrically active → we lose conductivity.

"PILE UP", a lot of P atoms piling up but not contributing to conductivity

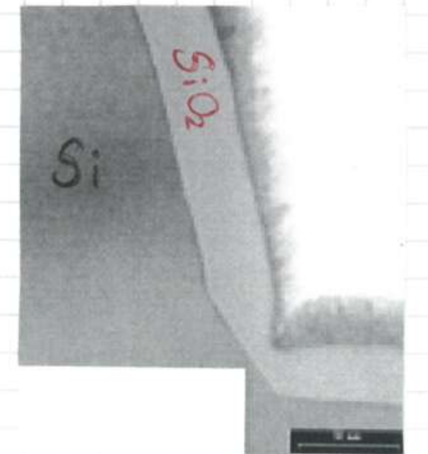
CRYSTAL ORIENTATION

Another thing that changes the SiO_2 growth rate is the crystal orientation.
B does not depend on crystal orientation but B/A does.

Why? Because change the number of Si bonds available for cm^{-2}, and also mechanical stress on Si depends on crystal orientation.



$$\left(\frac{B}{A}\right)_{111} = 1{,}68 \left(\frac{B}{A}\right)_{100}$$



STI picture at e^--microscope: see how the walls, which change orientation from vertical to horizontal, change their thickness.

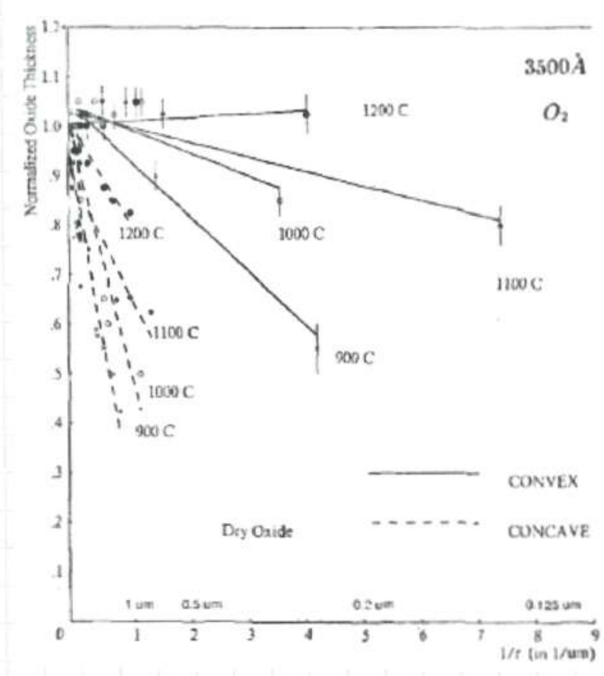
2D OXIDATION

The Deal - Grove model is a 1D model. When we go to 2D or 3D oxidation growth, we have to take into account many more parameters:

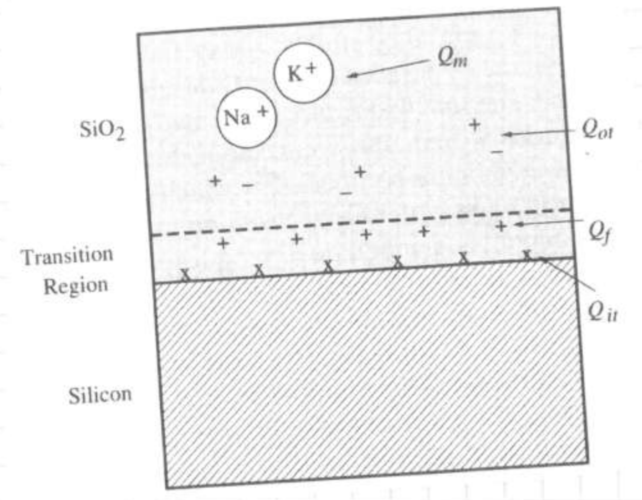
- effects of crystal orientation on oxidation rate
- dependence of oxidant diffusion on local surface curvature
- viscous flow
- stress effects

and many more

Example: Kao et al. studied oxidation rates as a function of radius of curvature of different structures



ELECTRICAL PROPERTIES OF SiO_2 : CHARGES IN SiO_2



usually at the Si/SiO_2 interface we will have around $\sim 10^9 - 10^{15}$ defects / cm^2 (which compared to the Si surface atoms of $\sim 10^{15}$ atoms / cm^2 is still very low).

The defects that we find in SiO_2 will have a charge associated with them.

There are 4 main kinds of "charge defects":

1. Q_f = fixed oxide charge

   is a sheet of positive charge close to the interface ($\simeq 2$ nm) and is (believed to be) formed of incompletely oxidized Si atoms. Being fixed, doesn't change during device operation.
   Can be reduced greatly by high T annealing after oxidation

2. Q_{it} = interface trapped charge

   are located at the interface, probably related to incompletely oxidized Si and/or dangling bonds.
   Can be positive, negative or neutral.
   They will have a level in the forbidden gap (close to valence/conduction) so during the operation of my device (biasing it), we move our quasi Fermi level (bending of bands) and those trap levels will be populated in a different way. To take care of Q_{it} we can run a 450°C H_2 anneal at the END of the process flow, otherwise they can re-form again.

3. Q_M = mobile ionic charges

we can have ionic contamination in our SiO_2, like alkali ions (Na^+, K^+, Li^+), negative ions, heavy metals, ...
Since they usually come from human beings, they can be reduced by keeping under control the clean room.
For the furnace tubes: we can clean them with HCl to control ionic contamination.

NOTE: ions in SiO_2 are usually mobile, while Q_f / Q_{it} NO.
So Q_m can move, and will move with an electric field.
This was what caused threshold voltage instability in early MOS devices. So the best way to keep Q_m low is an environmental control.

4. Q_{ot} = oxide trapped charges

usually located in the bulk of the oxide, can be positive or negative. Are usually broken Si - O bonds and will act as traps in the bulk oxide. Usually come from damage during the process flow itself rather than coming from the growth.
examples: ion implantation, plasma etching, ionizing radiation in general

Having trap charges and states both at the interface and the bulk of the oxide will change the electrical properties of the oxide, dielectric properties, leakage current level, ...

CV CHARACTERIZATION OF SiO_2

One of the ways of studying the properties of SiO_2 that we're growing for our MOS devices is measuring its electrical properties.

One of the key measurements we do for a MOS device is the **CV characteristics of the oxide SiO_2**.

The simplest MOS device we can form is a **MOS capacitor**, and from the CV characterization of SiO_2 **we can figure out where the fixed charges are, where the interface traps are allocated energetically in the forbidden gap**, ...

First: for a 2-metal-plates capacitor, there is **no dependance of C to V.**

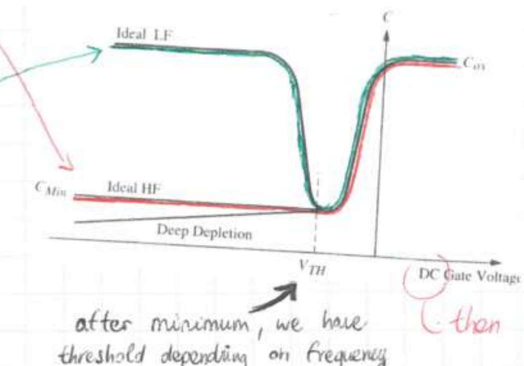
real "ideal" C : C — constant — V

metal / metal

For a **metal-semiconductor capacitor** we'll have a behaviour like:

HF ($\gtrsim$ 100 Hz) we have this

LF ($\lesssim$ 100 Hz) we have this



after minimum, we have threshold depending on frequency

NOT constant!

C depends on the frequency at which we measure

(then why DC? In reality is "DC + very small AC")

To measure our C(V), we apply a **DC Voltage** and then on top of it a **small AC signal** which gives $C(V) \neq 0$, because $I = C \frac{\partial V}{\partial t}$ and only when we apply an AC signal we see a current flowing.

ENERGY BANDS

$+V_G$

M O S

N Silicon Doping = N_D

C_{ox}

C, C_{ox}, V_G, −, +

(a)

E_C, E_F, E_V, $+V_G$

Metal, Oxide, Semiconductor

ACCUMULATION

the bands bend towards the electric field. Since we have N doped Si : E_F close to E_C. So we have (because of down bend) accumulation of e^- at the interface.

What happens when we apply a positive voltage ($+V_g$) at the metallic electrode?

$+V_g$

M O S

positive charges accumulate at postive end

e^- accumulate at the semiconductor, like a metal so until now we don't see any difference with a 2-metal-plate with a dielectric in the middle

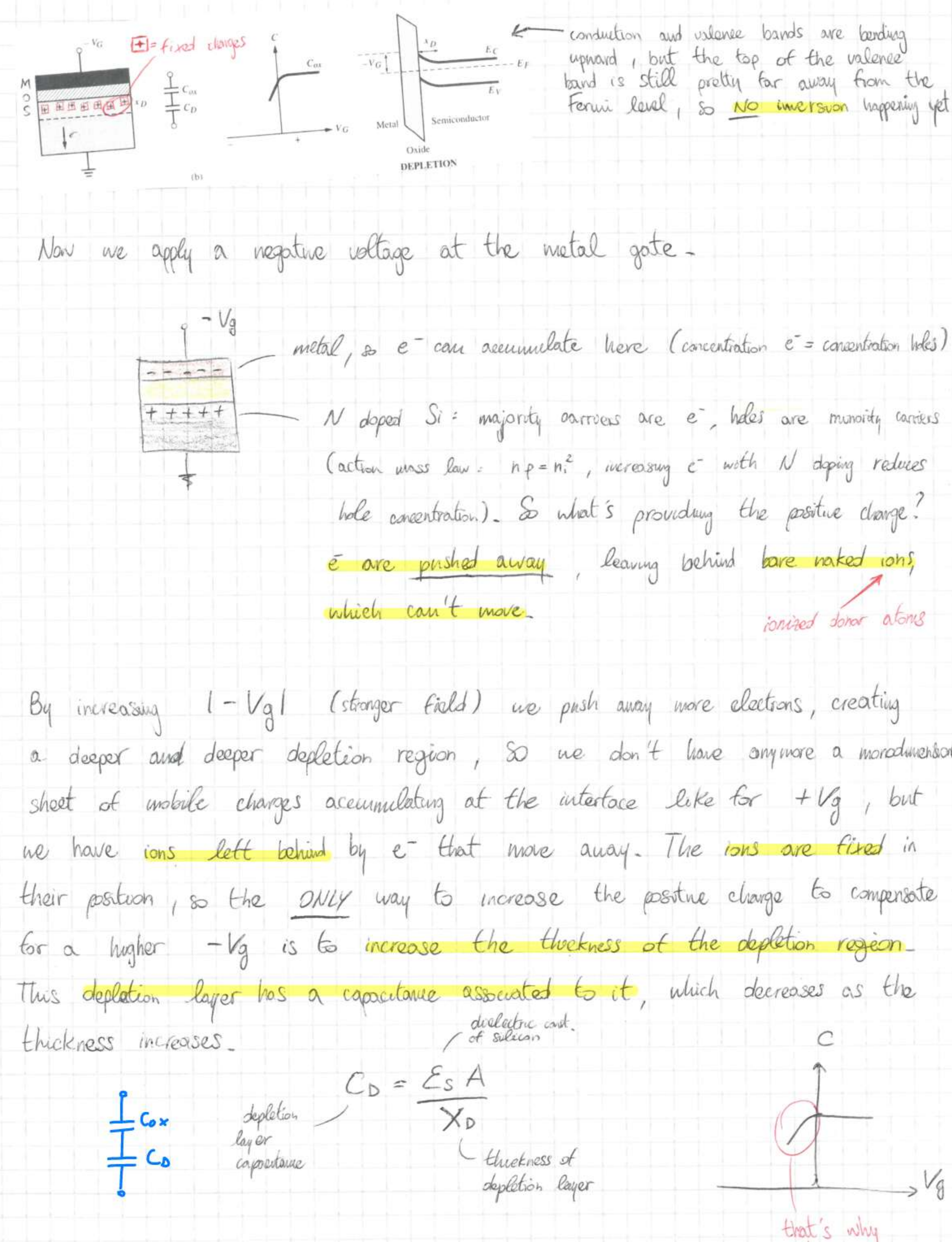
So if now we apply a small AC signal to measure the capacitance, we would measure just the capacitance associated with the oxide C_{ox}

$$C_{ox} = \frac{\varepsilon_{ox} A}{X_{ox}}$$

(ε_{ox}: dielectric const. of oxide; A: area of capacitor; X_{ox}: thickness of oxide)

At positive voltages our semiconductor just behaves like a metal because there's plenty of free moving e^- since our crystal is N doped.

And C stays constant $\forall\ V > 0$ that we apply.



← conduction and valence bands are bending upward, but the top of the valence band is still pretty far away from the Fermi level, so NO inversion happening yet

Now we apply a negative voltage at the metal gate.

$-V_g$

metal, so e^- can accumulate here (concentration e^- = concentration holes)

N doped Si: majority carriers are e^-, holes are minority carriers (action mass law: $np = n_i^2$, increasing e^- with N doping reduces hole concentration). So what's providing the positive charge?

e^- are pushed away, leaving behind bare naked ions, which can't move.

↑ ionized donor atoms

By increasing $|-V_g|$ (stronger field) we push away more electrons, creating a deeper and deeper depletion region, so we don't have anymore a monodimensional sheet of mobile charges accumulating at the interface like for $+V_g$, but we have ions left behind by e^- that move away. The ions are fixed in their position, so the ONLY way to increase the positive charge to compensate for a higher $-V_g$ is to increase the thickness of the depletion region. This depletion layer has a capacitance associated to it, which decreases as the thickness increases.

$$C_D = \frac{\varepsilon_S A}{X_D}$$

- C_D: depletion layer capacitance
- ε_S: dielectric const. of silicon
- X_D: thickness of depletion layer

C_{ox}

C_D

C

V_g

that's why decreases

INVERSION

holes accumulation

the bands bend upwards. Close to the interface we will have the valence band very close to the Fermi level, which results in holes accumulation (inversion)

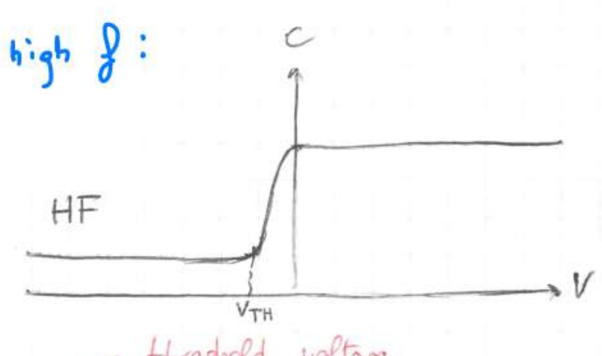
INVERSION

We keep decreasing V_g, so now we have $--V_g$ (very negative).

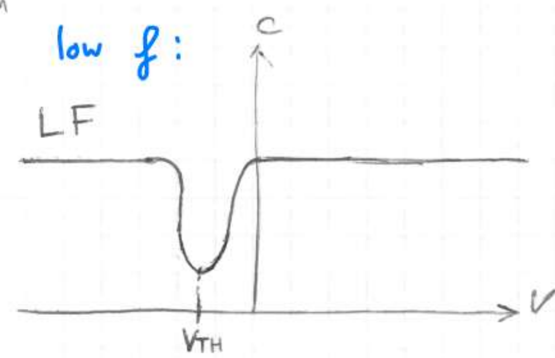
Two things can happen, depending on the frequence we use to probe our capacitance:

- LF (low frequency, $\simeq 10$ Hz) the curve will go up and go back to a metal-like behaviour
- HF (high frequency, $\simeq 100$ Hz) the curve will stay on the minimum

high f:

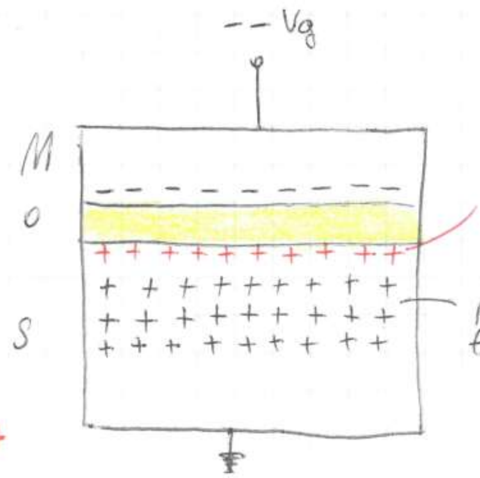


low f:



threshold voltage

V_{TH} = Voltage at which the concentration of holes overcomes that of e^- at the interface, so locally we can change the concentration of mobile charges and even invert it!



mobile positive charges from holes (minority carriers)

positive charges from fixed ionized donors

So after V_{TH} I start having an accumulation of positive charges coming from moving holes!

NOTE: the concentration of holes grows exponentially with V_g, while that of ionized donors goes sublinearly ($\propto \sqrt{V_g}$), so past V_{TH} we'll have a fast growing number of holes compared to the growth in thickness of the depletion zone (coming from ionized donors).

Now, when I try to increase V_g (more and more negative) I see a monodimensional sheet of holes appearing at the interface.

Let's now measure the capacitance, applying a small AC signal on top of DC signal already there.

NOTE: the electrons on the metallic electrode are always available, while the holes in the semiconductor are thermally generated, since they come from localized levels nearby the conduction bands.
The process that generates holes is slow, because I have to give the chance to an electron to jump from valence to conduction band to leave a hole behind.

So if my AC signal is very fast compared to the generation / ricombination rate ($\approx$ few 100s Hz), then what's reacting to the signal is not the holes but the depletion zone, since I'm not giving enough time to the holes to be generated, alternating too fast.
The capacitance I'd measure is around the same value we get thanks to the whole thickness of the depletion zone X_D (x at V_{TH}).

"around" = oscillation around X_D

If now I slow down the AC signal enough to allow generation and recombination of holes, what will respond to the signal won't be the ions but the holes, appearing and disappearing at the interface of the dielectric as a monodimensional sheet of charge.
This gives a capacitance like that of 2 metal plates.

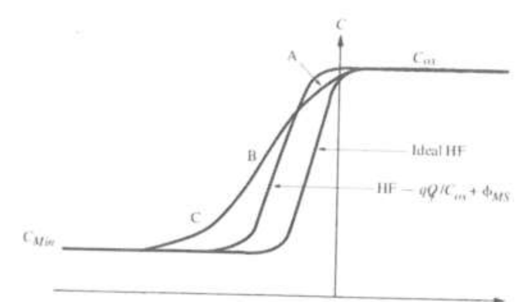
Lecture 9

26 marzo

NOTE: why, at HF, past V_{TH} we stay at a constant minimum?
Is it that the depletion zone doesn't have anymore space where to expand its thickness?
NO.
It is because after V_{TH} there are also the holes participating, and their response is exponential with voltage.
In fact: If we use a <u>too high</u> frequency, those holes can't contribute (not generated fast enough) and we have the depletion zone growing (DEEP DEPLETION)

Let's see how we can use this measurement (CV characterization of SiO_2) to detect some of the charge states (Q_f, Q_{it}, Q_m, Q_{ot})

-



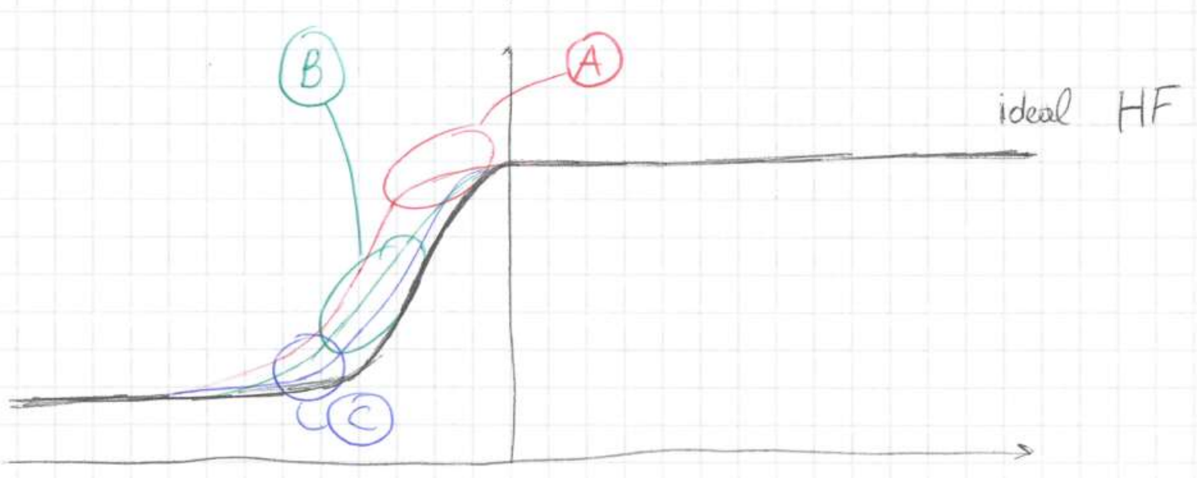
- Q_F = fixed charge

  Q_F are always positive, remember

  If we have a positive fixed charge in the oxide, we would have a negative image charge in the silicon, so the inversion would be harder to obtain.
  So what we would see if we had a fixed charge at the interface we would see our CV curve shifting rigidly to the left (that would also be true for P doped Si).
  So the effect of fixed charges is to lower the threshold voltage V_{TH}, no matter the substrate doping.

- Q_{it} = interface trapped charge

  Q_{it} has the effect of "stretching" the CV curve, and the stretching will happen in different zones of the curve depending on the position of the traps in the gap:

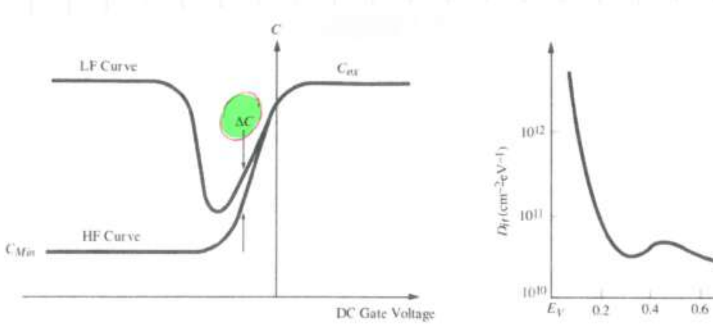


A zone : interface trapped states near the conduction band

B zone : interface trapped states near the mid band gap

C zone : interface trapped states near the valence band

and this is because the electric field is bending the bands (moving the quasi-Fermi level) so we're intercepting different regions of the gap.



Usually the stretching is much lower on the HF curve compared to the LF curve, and by measuring the change in capacitance (ΔC) between the two curves as a function of voltage, we can infer the density of interface states as a function of energy (picture to the right).
This also happens because the trap response time (trapping and detrapping) also takes time, so traps can't easily keep up with HF.

Other measurements that we can make:

- breakdown voltage (distructive measurement)
- TDDB (Time dependent dielectric breakdown) test, for reliability
- BTS (Bias temperature stress), for mobile charge estimation

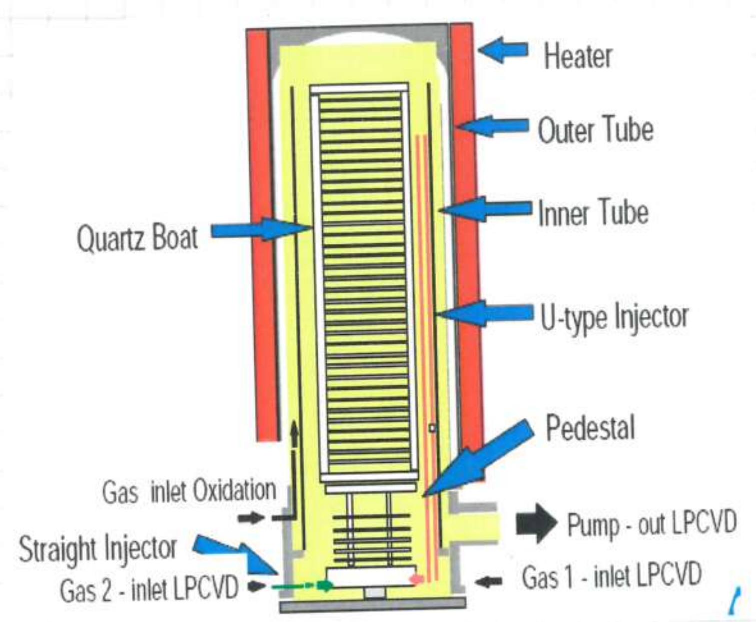
NITRIDATION

One thing we do (especially if we're using SiO_2 as the gate oxide material) to our gate oxide after forming the gate oxide is called nitridation, mainly to reduce Boron diffusion from Si to SiO_2 (because of segregation). We don't want our Boron to diffuse, reach the gate and dope the gate, or even just be incorporated into SiO_2 (which would weaken its structure).
Nitridation will hence improve oxide quality: by having a thin Si_3N_4 layer at the Si / SiO_2 interface will make the oxide more resistent and reliable, improve QBD (aka charge-to-breakdown, it's a destructive test method used to determine the quality of gate oxides in MOS devices), increase resistance to the trapping and generation of interface states, ...

The most common nitridation techniques consist in annealing treatments made after the thermal oxidation step, at high ~ medium T with NH_3, N_2O or NO by a furnace or RTP (rapid thermal processing). Nitridation by NO is becoming the leading technology.

The best results are obtained by the formation of a N rich layer at the Si / SiO_2 interface, avoiding N incorporation in the oxide bulk.

SILICON OXIDATION EQUIPMENTS

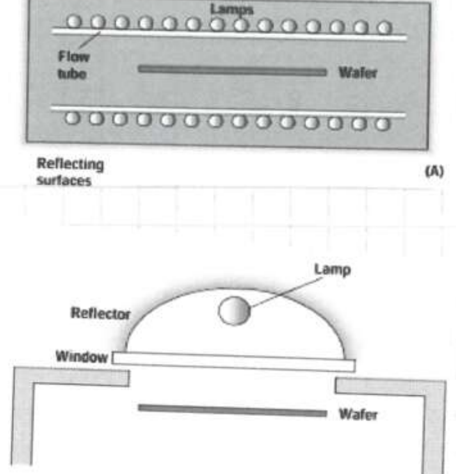


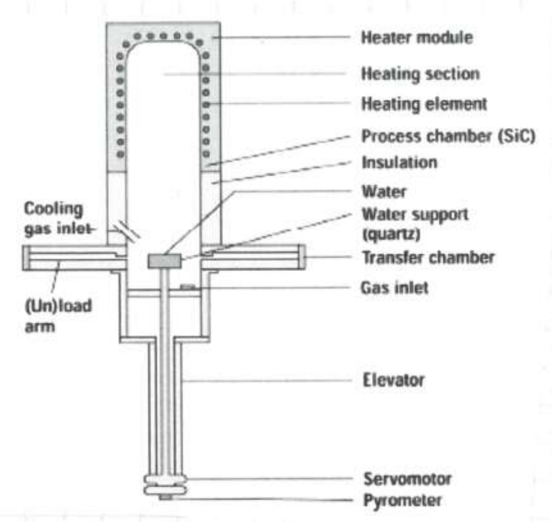
VERTICAL FURNACE

The mainstream tools used for silicon oxidation are vertical furnaces. We stack wafer batches (up to 100 wafers) in a quartz tube inside the furnace, then heat the furnace up by Joule effect, and in the meantime we flow reactant gases in the furnace (O_2 or H_2O)

NOTE: For bigger wafers the furnaces get smaller, because control of temperature uniformity across the wafer gets harder (12" wafers → up to ~25, for 8" wafers up to 100).

RTO (Rapid thermal oxidation)





This are single wafers tool in which we heat the wafers very rapidly (compared to batch furnaces tool) by using lamps. This lamps can heat the wafers to the desired T in a matter of seconds and once we're at desired T we flow in the gas and start the chemical reaction. ← takes hours for standard furnaces

There are various arrangements:

(A) array of lamps

(B) single lamp with a reflector

...

For a RTO the activation energies can be significantly different compared to batch furnace processes.

NOTE: being rapid, RTO reduces the thermal budget

RADICAL OXIDATION

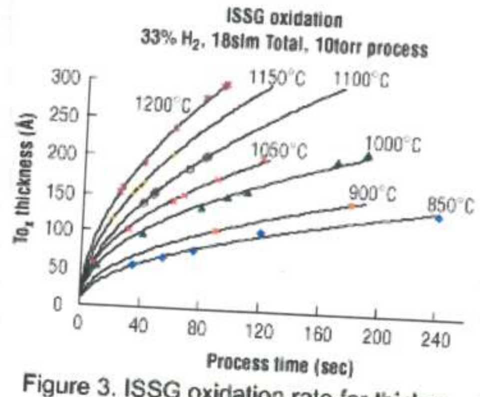
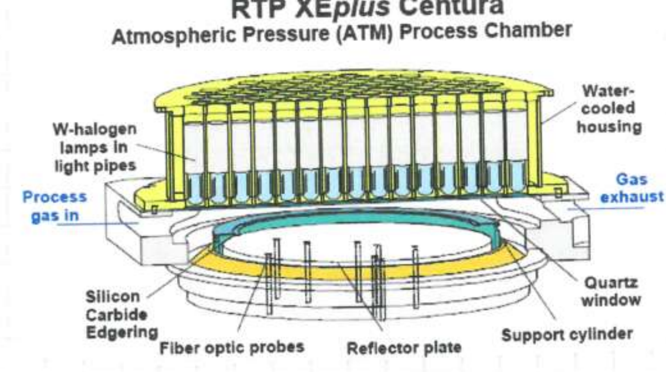


Figure 3. ISSG oxidation rate for thicker oxides (>50Å).

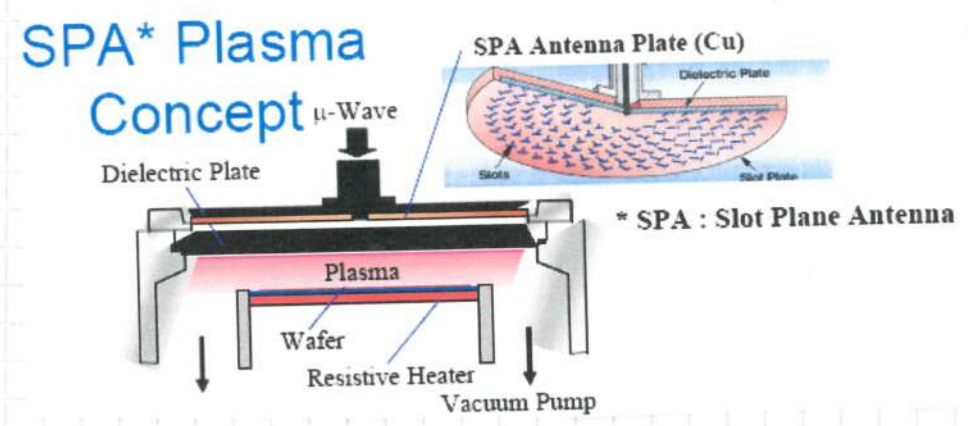


In a radical oxidation process we use RTP tools (lamps), and the Kinetics of the oxidation are changed in a way so that we directly inject H_2 and O_2 (NOT H_2O) into the RTO chamber, they react with the hot walls of the chamber and form water (H_2O) and radicals (O, OH).
This radicals inside the low pressure chambers guarantee high quality oxides at relatively low temperatures (~ 700 °C).

vs ~1000 °C for standard furnaces

NOTE: radical oxidation can have completely different oxidation Kinetics. For example, by fine-tuning the concentrations of H_2 and O_2 we can have an oxidation very insensitive to crystal orientation, very useful for when we want to grow a thermal oxide across a patterned surface without changing the thickness.

PLASMA OXIDATION



If your process requires a lower T for the oxide to grow (so we have a very low thermal budget), but you still want a high quality oxide, you can use plasma to form your oxide.

That's because when we want a chemical reaction to happen (whether is thin film deposition, oxidation, etching, ...) you can have it by increasing the temperature OR by forming a plasma, thus having the electric field providing the energy needed.

A very common tool for plasma oxidation is the SPA (slot plane antenna), which allows SiO_2 growth at ~300°C !

NOTE: what's "faster" or "slower" in this processes is the time required to reach the desired T on the wafer, NOT the oxidation rate, which is (in principle) exactly the same in all processes. That's not exactly true, but there's no substantial difference in the speed at which the thermal oxide grows.

ALTERNATIVE GATE OXIDES

A problem with scaling down the dimensions of our MOS devices is the k (dielectric constant) of SiO_2, since by scaling down the dimensions the area A decreases → C decreases. But to have a good control over our device we would have to also scale down the thickness x of the oxide (so C doesn't go down too much), but past a certain thickness (for SiO_2 $x \simeq 0{,}8$ nm) we have gate tunneling.

To avoid this while we keep scaling down, we have to use a dielectric with higher k.

$$C_{ox} = \frac{\varepsilon_{ox} A}{X_{ox}} = \frac{k\,\varepsilon_0 A}{X_{ox}}$$

This alternative gate dielectric must satisfy at least the following:

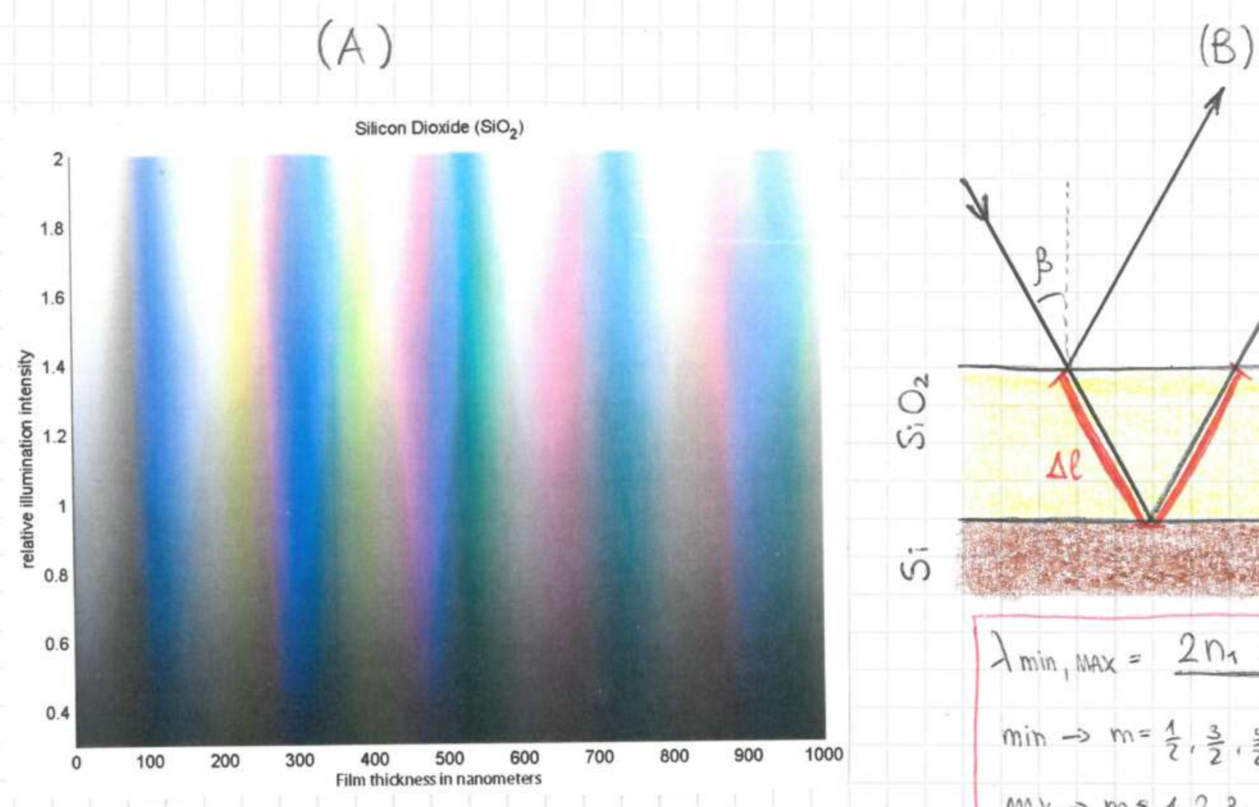
- high k
- band offset with respect to Si > 1 eV (at least)
- band gap > 4 ~ 5 eV (at least, to be sure is insulating)
- thermodynamically stable in contact with silicon
- compatible with IC process flow

SOME EXAMPLES

Dielectric	Dielectric constant	Bandgap (eV)	Conduction Band offset	Leakage reduction	Thermal stability
Silicon Dioxide (SiO_2)	3.9	9	3.5	1x	>1050°C
Silicon Nitride (Si_3N_4)	7	5.3	2.4	?	>1050°C
Aluminum Oxide (Al_2O_3)	~10	8.8	2.8	10^2-10^3x	~1000°C RTA
Tantulum Pentoxide (Ta_3O_5)	25	4.4	0.36	?	unstable with Si
Lanthanum Oxide (La_2O_3)	~21	6	2.3	?	?
Gadolinium Oxide (Gd_3O_3)	~12	?	?	?	?
Yttrium Oxide (Y_2O_3)	~15	6	2.3	10^4-10^5x	Form silicate
Hafnium Oxide (HfO_2)	~20	6	1.5	10^4-10^5x	~950°C

modern material of choice to replace SiO_2, usually → Hafnium Oxide (HfO_2)

HOW TO MEASURE SiO_2 THICKNESS



$$\lambda_{min, MAX} = \frac{2 n_1 x_0 \cos\beta}{m}$$

min → $m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$

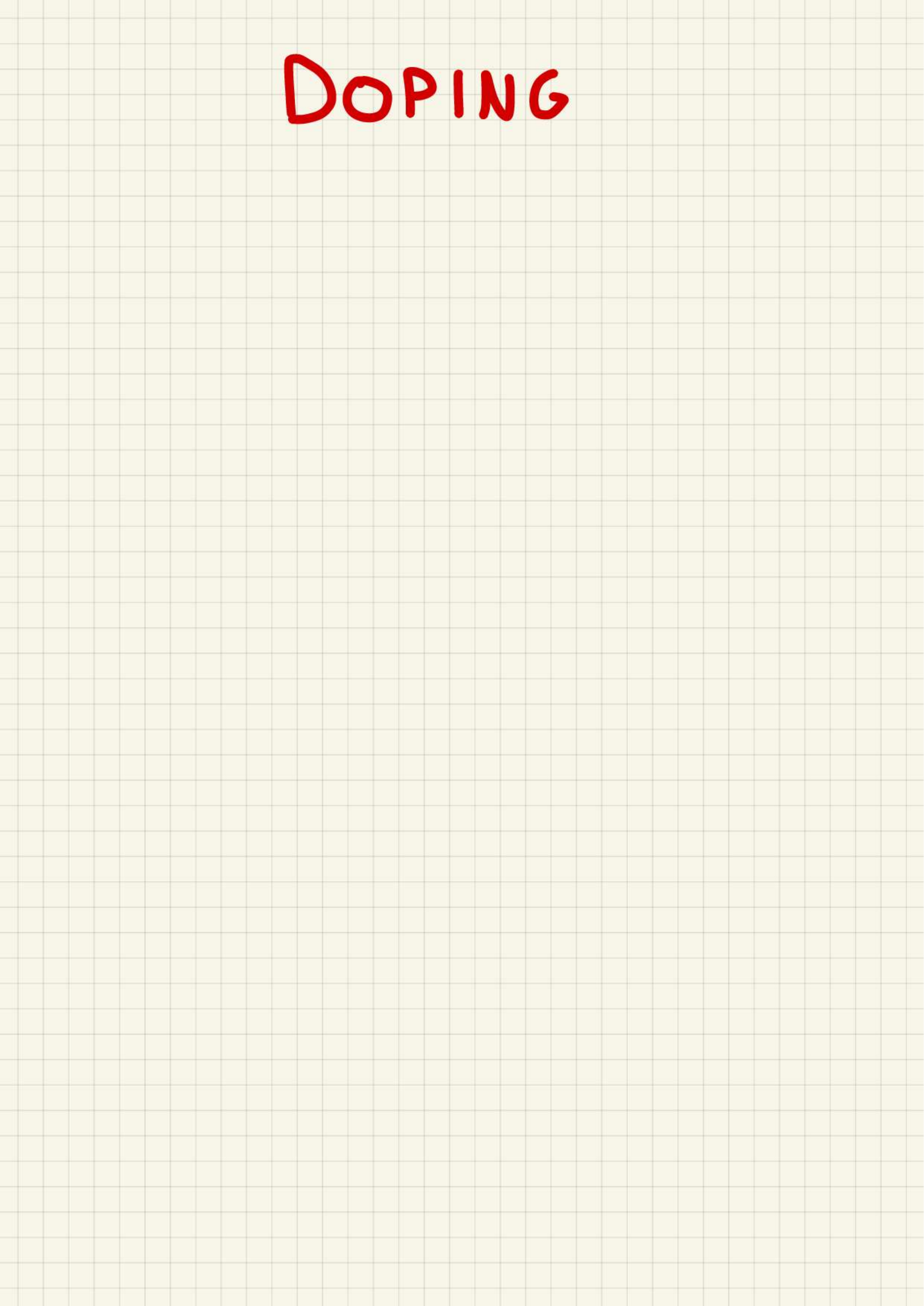
MAX → $m = 1, 2, 3, \ldots$

In the old days, the operators were given SiO_2 color charts (A) to make sure that the grown SiO_2 was of the right thickness: they would check it by looking the color of the reflected light at a certain angle by the wafer.

This happens because the light rays hitting the SiO_2 at an angle β will be partially reflected and partially refracted, with the refracted ray having traveled a surplus path of $\Delta\ell$. Depending on the wavelength of the impinging light, we could have constructive / destructive interference at the detector (or the eye, in the old days).

As we see in (B), we can have minima (or maxima) depending on the refractive index of SiO_2 (n_1), on its thickness (x_0), on the angle ($\cos\beta$).

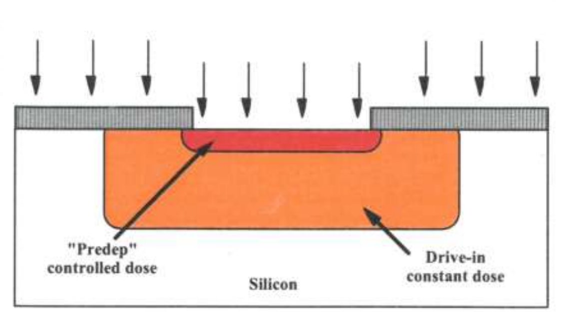
Modern techniques employ interferometers to analyze the interference patterns and derive the thickness of the oxide. We can use an ellipsometer, which derives oxide thickness, doping concentration, electrical conductivity, ... by measuring both intensity and polarization modulation.


DOPING

Lecture 10

31 marzo

DOPANT DIFFUSION



	MODERN Ion Implantation and Annealing	OLD DAYS Solid/Gas Phase Diffusion
Advantages	Room temperature mask	No damage created by doping
	Precise dose control	Batch fabrication
	10^{11} - 10^{16} atoms cm^{-2} doses	
	Accurate depth control	
Problems	Implant damage enhances diffusion	Usually limited to solid solubility
	Dislocations caused by damage may cause junction leakage	Low surface concentration hard to achieve without a long drive-in
	Implant channeling may affect profile	Low dose predeps very difficult

In the past, diffusion was the only known way to insert dopants into the substrate. Nowadays the pre-deposition is done by ion implantation.
The diffusion of dopants is separated in two phases:

1. pre-deposition: actual insertion of dopants into the substrate

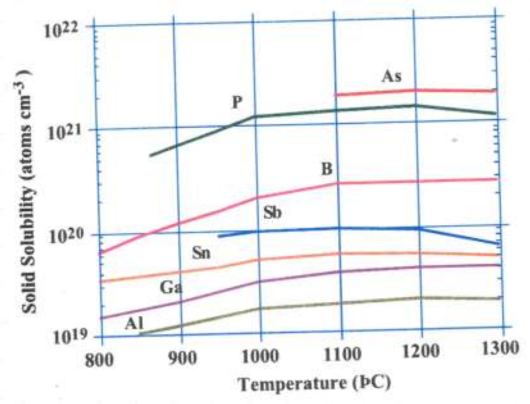
2. drive-in: movement of dopants with the subsequent thermal treatments

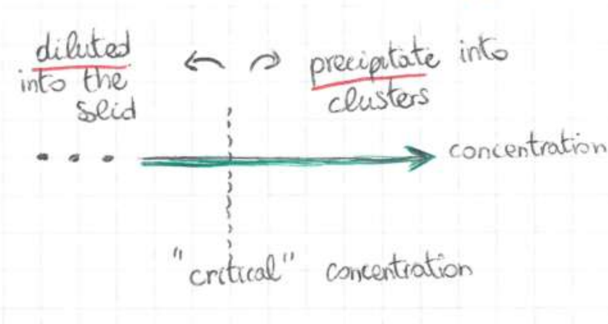
"Diffusion" is the redistribution of atoms from regions of high concentration of mobile species to regions of low concentration.
It occurs at all temperatures, but the diffusivity has an exponential dependence on T.

In the good ol' days, it was done by laying (or flowing a gas containing the dopant) a sled on top of the wafer for a certain time at a certain temperature, and let the dopants of the sled diffuse into the wafer.

SOLID SOLUBILITY

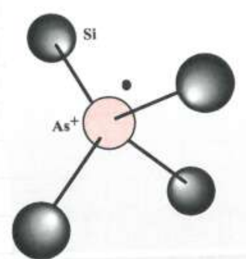
If you insert impurities into your substrate (no matter the impurities), after a certain concentration they precipitate into another phase. ⇒ like sugar in coffee



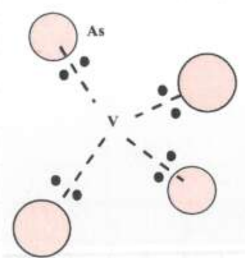


This is what concerns chemical solubility: the physical amount of dopant atoms you can actually put into a silicon crystal before forming precipitates.

We could also have "electrical" solubility: to have active dopants, we want them to sit in substitutional position with respect to Si atoms. If they're not, then they're chemically diluted but won't contribute to the doping profile of the wafer, because we can't use that extra electron (or hole)!



this Arsenic atom is sitting in substitutional position and is surrounded by 4 Silicon atoms → it is active



in this example (called As4V, because it's 4 Arsenic atoms surrounding one Vacancy) this complex of 4 Arsenic atoms is electrically INACTIVE, so they're chemically diluted, but not contributing electrically.

NOTE: so inserting dopants ≠ activating dopants
we have to make sure they sit in substitutional position, even though the limit of % of dopants we can activate is low.

FICK'S LAWS

The physical laws that describe how the dopants move in a solid are the Fick's Laws: (actually works for every impurity)

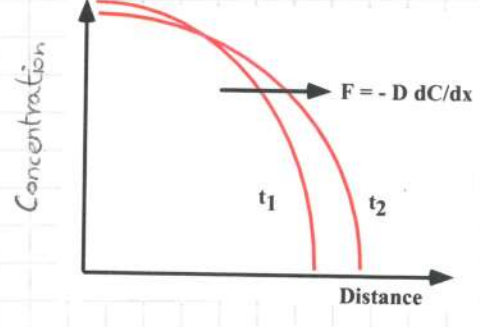
1. $F = -D \frac{\partial C}{\partial x}$

conservation of mass

2. $\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D \frac{\partial C}{\partial x} \right) = -\frac{\partial F}{\partial x}$

- F = dopant flux
- C = dopant concentration
- D = diffusivity

this laws are valid only if the concentration of the impurity is low compared to the density of the solid ⇒ NO interaction between impurity particles.



$$flux = \frac{\text{number of particles}}{\text{unit area} \cdot \text{unit time}}$$

A

1 second

the first Fick's law says that the flux is proportional to the gradient of concentration ($\partial C/\partial x$). This is very intuitive

D = diffusivity (strongly dependent on Temperature)

The **Second Fick's Law** is also very intuitive.

Consider a volume V (side Δx)

Consider the incoming flux F_{in} (particles in)

Consider the outgoing flux F_{out} (particles out)

F_{in} F_{out} Δx

$D \frac{\partial C}{\partial x}$ → variation of dopant concentration

$$\frac{\partial}{\partial x}\left(-F\right) = \frac{\partial C}{\partial t}$$

if $F_{in} > F_{out}$, the concentration inside the volume increases over time

if $F_{out} > F_{in}$, the concentration inside the volume decreases over time

$$-\frac{\partial F}{\partial x} = F_{IN} - F_{OUT} \longrightarrow \begin{cases} F_{IN} - F_{OUT} > 0 \Rightarrow \frac{\partial C}{\partial t} > 0 \\ F_{IN} - F_{OUT} < 0 \Rightarrow \frac{\partial C}{\partial t} < 0 \end{cases}$$

Assuming D is constant (e.g. at a fixed T) we can write

$$D = D_0 e^{-\frac{E_a}{KT}}$$

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2}$$

BOUNDARY CONDITION $\frac{\partial C}{\partial t} = 0$

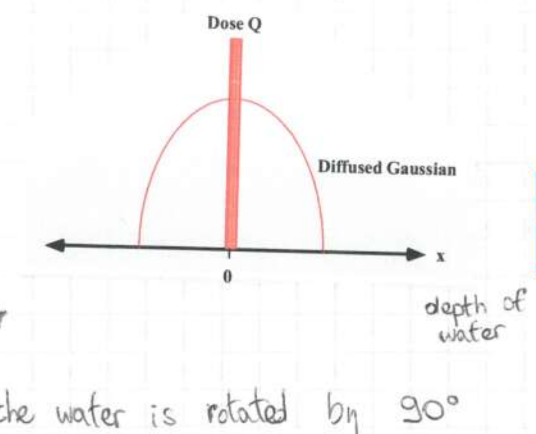
↓ steady state solution (concentration constant in time)

$$\frac{\partial C}{\partial t} = 0 \rightarrow C = a + bx$$

linear concentration

But this is NOT the only boundary condition that has an easy solution. Let's see some others.

δ - FUNCTION



let's consider a delta function of dopant in the middle of a lightly doped region

depth of water

the water is rotated by 90°

Boundary conditions

$C \to 0$ at $t=0$ for $x \neq 0$

$C \to \infty$ at $t=0$ for $x = 0$

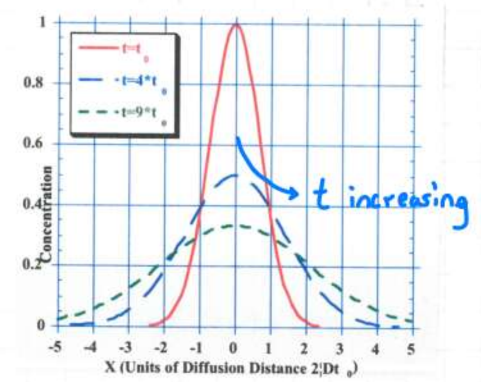
$$\int_{-\infty}^{+\infty} C(x,t)\, dx = Q$$

↑ dose Q

the solution to $\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2}$ with this boundary conditions is a gaussian profile

solution
$$C(x,t) = \frac{Q}{2\sqrt{\pi D t}} e^{-\frac{x^2}{4Dt}}$$

Dt → diffusivity · t

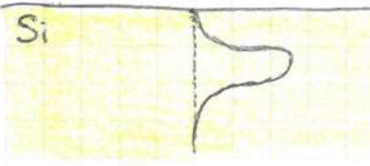


We see from the plotted solutions at consecutive times, that gaussian profiles always evolve into wider gaussian profile, so by this we get (for free!) also the solution to the boundary condition of a Gaussian instead of a δ-function.

δ - function $\xrightarrow{\text{solution}}$ Gaussian

Gaussian $\xrightarrow{\text{solution}}$ Gaussian

narrow distribution



time passes → at temperature T

wide distribution

This is very important because, for example, during ion implantation we reach gaussian profiles with a σ (sigma) so small that can be approximated to δ - functions. Nevertheless, after the thermal treatments, the evolution will give a wider gaussian profile.

δ - FUNCTION AT THE SURFACE

δ - function

before we had: δ - function

Let's suppose the δ - function is not at a certain depth but is right at the surface, now the dopants have no "up" to diffuse into (excluding evaporation), only "down".

The solution is half of a Gaussian: we just consider an imaginary gaussian outside the surface, then the solution is identical

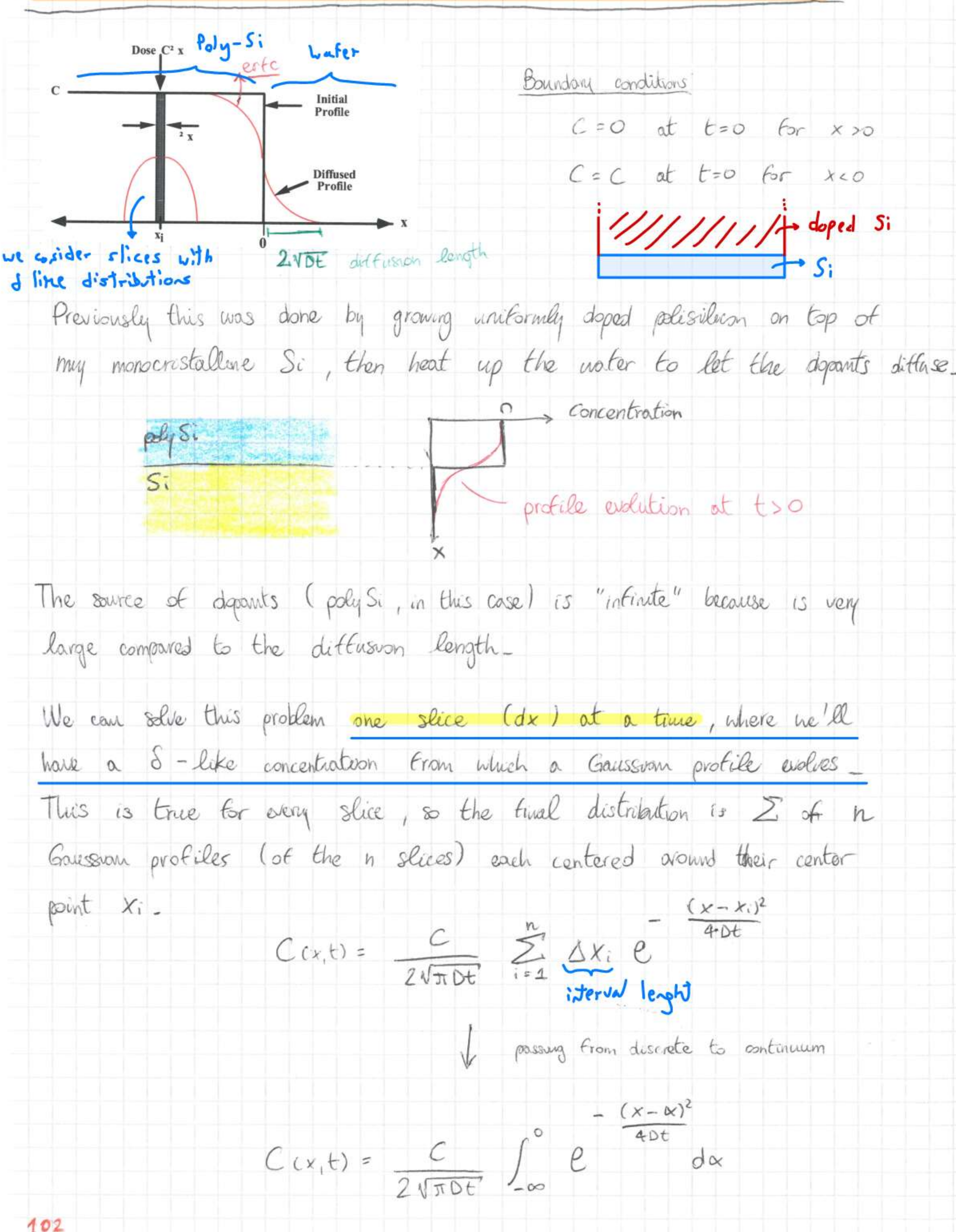
full Q in just half, before we had Q among both parts

solution $C(x,t) = \frac{Q}{\sqrt{\pi D t}} e^{-\frac{x^2}{4Dt}}$

NOTE: $\int_{-\infty}^{+\infty} C(x,t)\, dx = 2Q$

we would then take only the real atoms, from $0 \to +\infty$, discarding $-\infty \to 0$

This is useful because happens frequently to have the dopants very close to the surface (source and drain implants).



DIFFUSION FROM INFINITE SOURCE OF DOPANTS

Boundary conditions

$C = 0$ at $t = 0$ for $x > 0$

$C = C$ at $t = 0$ for $x < 0$

Previously this was done by growing uniformly doped polisilicon on top of my monocristalline Si, then heat up the water to let the dopants diffuse.

The source of dopants (poly Si, in this case) is "infinite" because is very large compared to the diffusion length.

We can solve this problem one slice (dx) at a time, where we'll have a δ-like concentration from which a Gaussian profile evolves.

This is true for every slice, so the final distribution is Σ of n Gaussian profiles (of the n slices) each centered around their center point x_i.

$$C(x,t) = \frac{C}{2\sqrt{\pi D t}} \sum_{i=1}^{n} \underbrace{\Delta x_i}_{\text{interval lenght}} \, e^{-\frac{(x-x_i)^2}{4Dt}}$$

↓ passing from discrete to continuum

$$C(x,t) = \frac{C}{2\sqrt{\pi D t}} \int_{-\infty}^{0} e^{-\frac{(x-\alpha)^2}{4Dt}} \, d\alpha$$

substituting $\eta = \frac{x-\alpha}{2\sqrt{Dt}}$

$$C(x,t) = \frac{C}{2\sqrt{\pi D t}} \int_{\frac{x}{2\sqrt{Dt}}}^{+\infty} e^{-\eta^2} \, d\eta$$

and if we remember the error function $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$

and the complementary error function $\operatorname{erfc}(z) = 1 - \operatorname{erf}(z)$

we can solve the integral above (which is similar but not identical)

$$C(x,t) = \frac{C}{2}\left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] = \frac{C}{2}\left[\operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)\right]$$

erfc (x)

0 x

0

$2\sqrt{Dt}$ diffusion length

$L = 2\sqrt{D \cdot t}$ = length $D = \left[\frac{m^2}{s}\right]$

- So we know that the wafer can be doped by ion implantation, by diffusion from a solid state source sitting on it.
- Another way to dope it is: you put your wafer in a (furnace) chamber containing a gas phase of your dopants. The concentration of dopants in the furnace is kept constant by continuously pumping new gas into it.

So now I'm interested to know what happens if I keep the concentration constant at the surface.

⇒ we already know, we get the error function tail

DIFFUSION FROM SURFACE CONSTANT CONCENTRATION OF DOPANTS

we note that as t passes, all erfc pass through the same point of constant concentration

$\frac{C}{2}$

From symmetry considerations we get to a concentration solution

$$C(x,t) = C\left[\operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)\right]$$

NOTE: at concentration $= \frac{C}{2}$ (mid-point, where all curves intercept) we have a constant concentration, so now I also have the solution to the "constant concentration at the surface" boundary condition. Just pay attention to the multiplication by 2!

From the concentration we can also get the dose Q:

$$Q = \int_0^{+\infty} C(x,t)\,dx = \frac{2C_s}{\sqrt{\pi}}\sqrt{Dt}$$

(C_s: concentration at the surface)

TO RECAP

1st Fick's Law $\xrightarrow{\text{conservation of mass}}$ 2nd Fick's Law

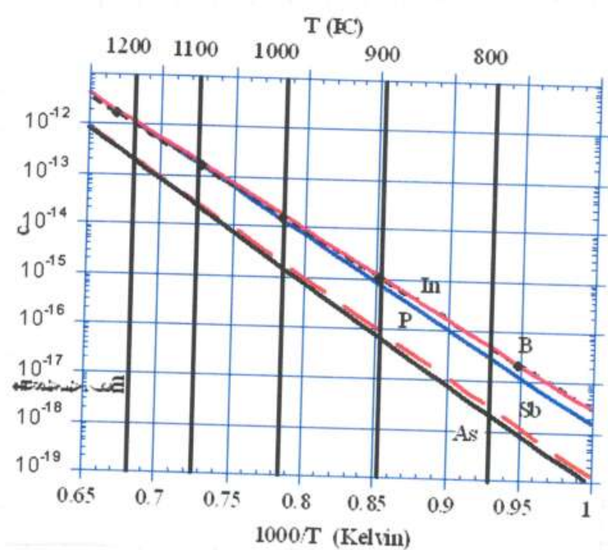
solutions:

- steady state solution
- δ - function
- Gaussian
- infinite source of dopants
- constant concentration at the surface

DIFFUSIVITY

For intrinsic doping, diffusion coefficients follow an Arrhenius law:

Arrhenius law
$$D = D_0 e^{-\frac{E_a}{KT}}$$
(E_a: activation energy; D_0: proportionality constant)



Element	D^0 (cm2sec-1)	E_a(eV)
Si	560	4.76
B	1	3.5
In	1.2	3.5
As	9.17	3.99
Sb	4.58	3.88
P	4.70	3.68

so the diffusivity has an exponential dependence on temperature and the diffusion length is $\ell = \sqrt{Dt}$

We could also account for different thermal treatments at different temperatures for different times, and get a total effective $D \cdot t$ product: (diffusion length squared)

(total diffusion length)^2 → $Dt_{effective} = \sum Dt = D_1 t_1 + D_2 t_2 + D_3 t_3 + \dots$

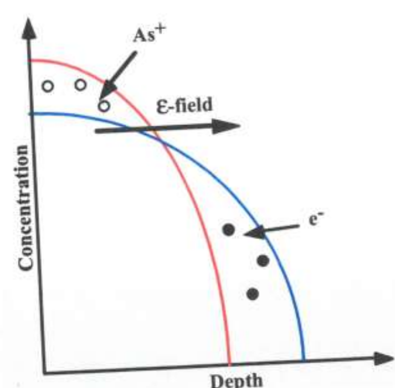
which can be rewritten as $Dt_{eff} = D_1 t_1 + D_1 t_2 \frac{D_2}{D_1} + \dots =$

$$= D_1 t_1 + D_1 t_2 \, e^{-\frac{E_a}{K T_1 T_2}(T_1 - T_2)} + \dots$$

so we see that if $T_1 \gg T_2$, it could be neglected (T_2) unless $t_2 \gg t_1$, but the dependence on T is exponential while only linear for t.

CORRECTIONS TO FICK'S LAWS

As always, being many more factors at play present in reality, Fick's Laws cannot predict accurately dopants diffusion in a wafer. Some corrections:



e^- and holes have higher mobility than ions so when we run a diffusion step, e^- and holes will diffuse faster, leaving behind charged donors (or acceptors).

This creates a polarization electric field which drags along donor/acceptor atoms too. So we'd have enhanced diffusivity

correction: $$\left[F = -h D \frac{\partial C}{\partial x} \right]$$

$$\text{with } h = 1 + \frac{C}{\sqrt{C^2 + 4 n_i^2}}$$

(n_i: intrinsic carrier concentration)

see after for full calculations!

Anther correction: in extrinsic conditions the concentration profiles don't match anymore neither Gaussian or erf / erfc, rather they resemble box - shaped profiles (for high concentration regions).

We had assumed that diffusivity didn't depend on concentration, but in the box-shaped profile we see that diffusion is faster where concentration is higher!

→ D depends on C

2nd Fick's law still holds, but in the form $\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D \frac{\partial C}{\partial x} \right)$

Element	D_0^0	E_a	D_0^+	E_a^+	D_0^-	E_a^-	D_0^{--}	E_a^{--}
Si	560	4.76						
B	0.05	3.5	0.95	3.5				
In	0.6	3.5	0.6	3.5				
As	0.01	3.44			31.0	4.15		
Sb	0.21	3.65			15.0	4.08		
P	3.85	3.66			4.44	4.0	44.2	4.37

Experimentally we can find:

$$D = D^0 + D^- \left(\frac{n}{n_i} \right) + D^{--} \left(\frac{n}{n_i} \right)^2 + \dots \qquad n \text{ type}$$

$$D = D^0 + D^+ \left(\frac{n}{n_i} \right) + D^{++} \left(\frac{n}{n_i} \right)^2 + \dots \qquad p \text{ type}$$

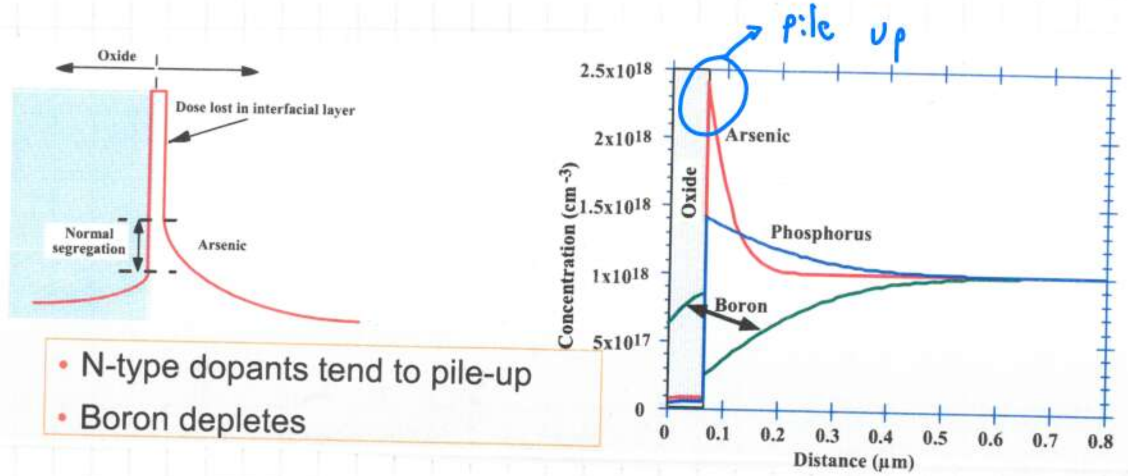
where every D (like $D^-, D^+, D^{--}, D^{++}, \dots$) can be written in the form:

$$D = D_0 e^{-\frac{E_a}{KT}}$$

with different activation energies and proportionality constants each

Lecture 11

7 aprile

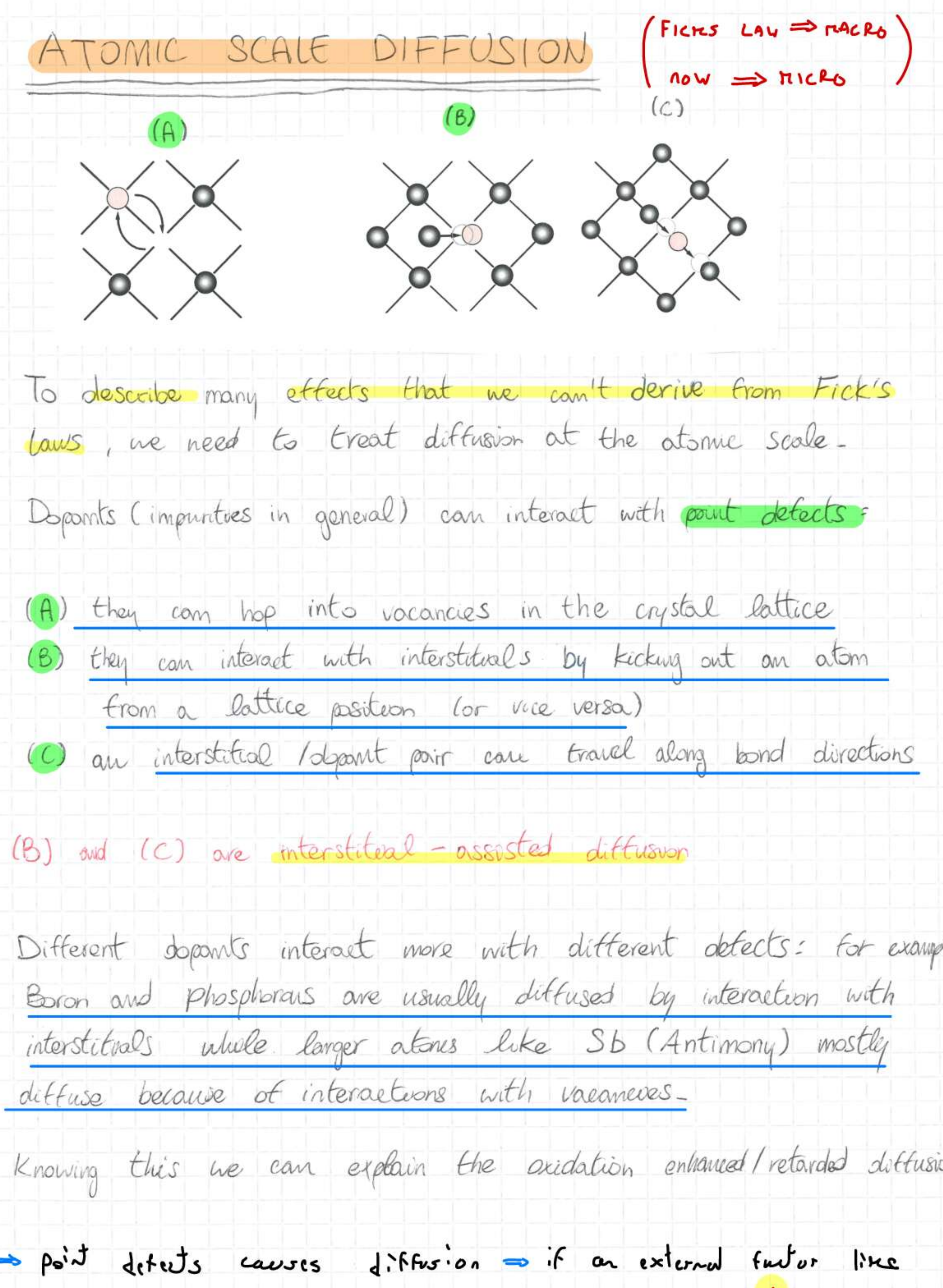


Other factors that change the diffusion profile are **segregation** and **pile-up** phenomena.

Segregation: different solid solubility in different materials

Interfacial dopant pile-up: n type dopants will tend to pile-up (since they prefer to stay in Si) at the Si/SiO_2 interface, on the Silicon side.

NOTE: usually piled-up dopants won't be electrically active.


ATOMIC SCALE DIFFUSION
(FICKS LAW ⇒ MACRO
NOW ⇒ MICRO)
(A)
(B)
(C)
To describe many effects that we can't derive from Fick's laws, we need to treat diffusion at the atomic scale.
Dopants (impurities in general) can interact with point defects:
(A) they can hop into vacancies in the crystal lattice
(B) they can interact with interstitials by kicking out an atom from a lattice position (or vice versa)
(C) an interstitial / dopant pair can travel along bond directions
(B) and (C) are interstitial-assisted diffusion
Different dopants interact more with different defects: for example Boron and Phosphorus are usually diffused by interaction with interstitials while larger atoms like Sb (Antimony) mostly diffuse because of interactions with vacancies.
Knowing this we can explain the oxidation enhanced / retarded diffusion
⇒ point defects causes diffusion ⇒ if an external factor like T changes the defects concentration, diffusion varies !
109

OXIDATION ENHANCED / RETARDED DIFFUSION

This picture shows the diffusion of Boron (and Antimony) in an oxidating ambient or in an inert ambient.

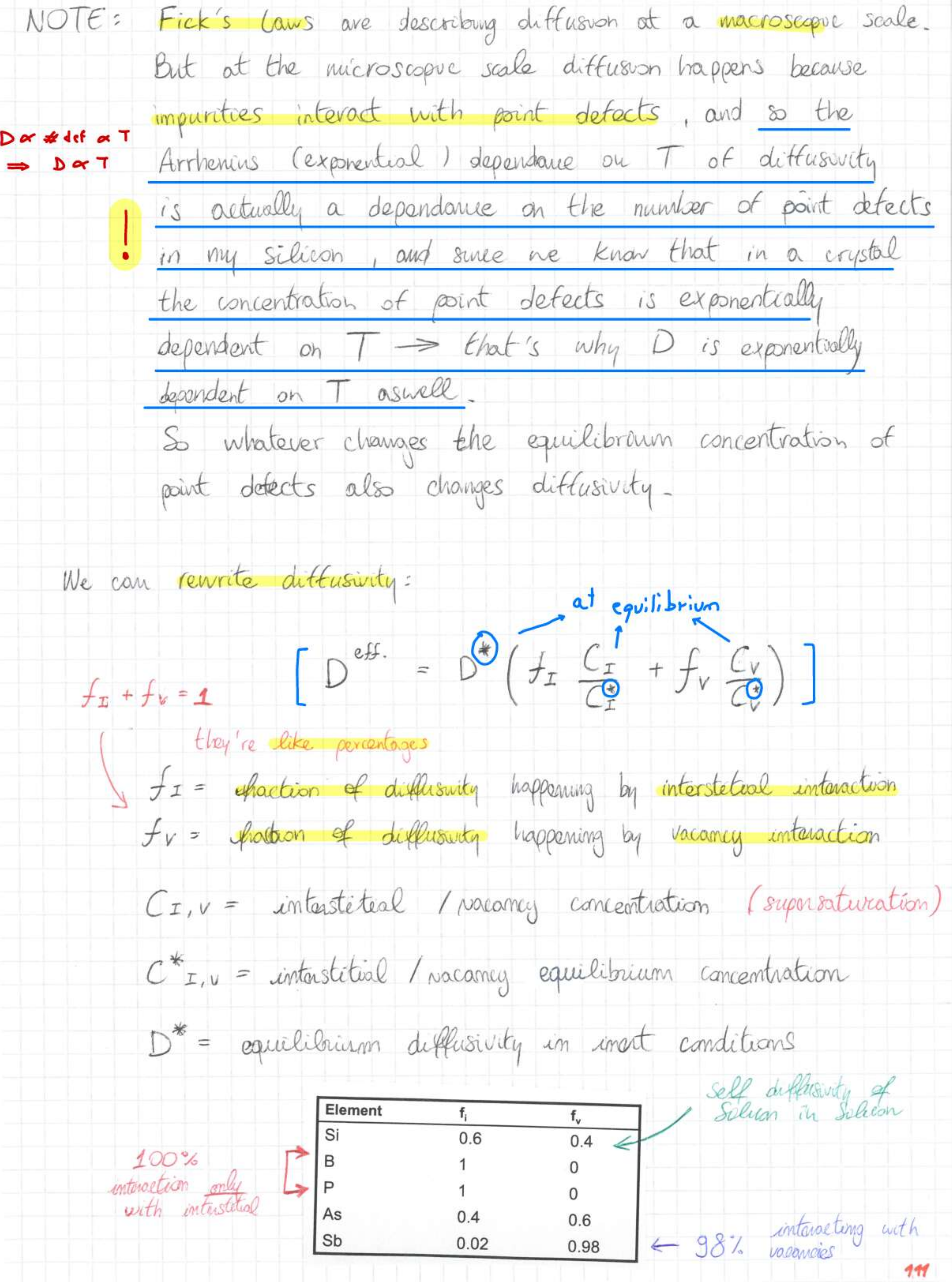
In an inert ambient (wafer in a chamber with an inert gas):

— Antimony (Sb)
— Boron (B)

With the same thermal budget (same t, T) but in an oxygen ambient:

— Antimony (Sb)
— Boron (B)

We can see that antimony is diffusing less, while boron much more. This is because oxidation (from a microscopic POV) injects interstitials into Silicon and absorbs vacancies (with respect to equilibrium concentrations at a given T).



D ∝ # def ∝ T
⇒ D ∝ T

NOTE: Fick's laws are describing diffusion at a macroscopic scale. But at the microscopic scale diffusion happens because impurities interact with point defects, and so the Arrhenius (exponential) dependance on T of diffusivity is actually a dependance on the number of point defects in my silicon, and since we know that in a crystal the concentration of point defects is exponentially dependent on T → that's why D is exponentially dependent on T aswell.

So whatever changes the equilibrium concentration of point defects also changes diffusivity.

We can rewrite diffusivity:

$$\left[D^{eff.} = D^{*}\left(f_I \frac{C_I}{C_I^{*}} + f_V \frac{C_V}{C_V^{*}} \right) \right]$$

(* = at equilibrium)

$f_I + f_V = 1$ — they're like percentages

f_I = fraction of diffusivity happening by interstitial interaction

f_V = fraction of diffusivity happening by vacancy interaction

$C_{I,V}$ = interstitial / vacancy concentration (supersaturation)

$C^{*}_{I,V}$ = interstitial / vacancy equilibrium concentration

D^{*} = equilibrium diffusivity in inert conditions

Element	f_i	f_v
Si	0.6	0.4
B	1	0
P	1	0
As	0.4	0.6
Sb	0.02	0.98

Si: Self diffusivity of Silicon in Silicon

B, P: 100% interaction only with interstitial

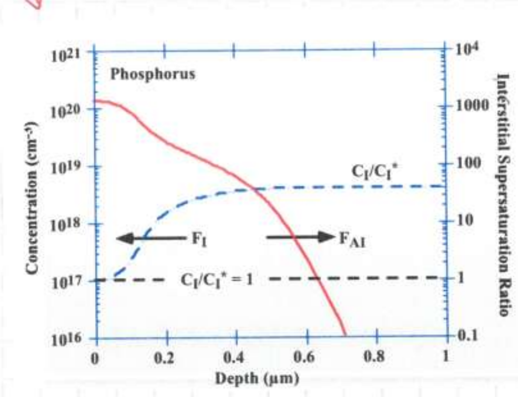
Sb: 98% interacting with vacancies

So we can think of diffusion not as dopants moving but as the couple dopant - point defect moving, because that's what's really happening! We can consider a dopant A interacting with an interstitial I as:

$$A + I \leftrightarrow AI$$

interstitial assisted mobile species

Interstitial assisted diffusion can generate a chemical pumping of interstitial defects into the substrate, explaining the "enhanced tail diffusion for P" and the "emitter push".



Emitter
Base
Emitter push
Collector

both diffuse due to interstitials
⇒ more interstitials where both Base and Emitter are ⇒ more diffusion

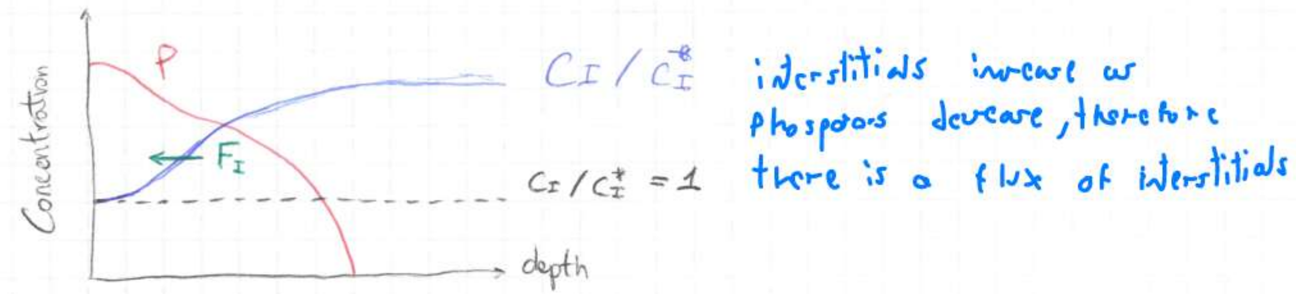
EMITTER PUSH

The Boron in the base is diffusing much more under the emitter than elsewhere (remember: this was done by gas phase diffusion, no implantation!). We would expect the Boron in the base, just because of thermal budget, in a uniform way.

Solution:

Phosphorous diffuses because of interstitial interaction, so at the emitter we'll have P - I couples moving. Boron also diffuses only due to interstitial interaction, so under the emitter there will be many more interstitials available → more diffusion for Boron there.
⇒ DUE TO PHOSPHOROUS FROM EMITTER

ENHANCED TAIL DIFFUSION FOR P



We see that the diffusion of phosphorous changes the supersaturation of interstitials (C_I) from $\frac{C_I}{C_I^*} = 1$ to $\frac{C_I}{C_I^*} \simeq 50$

This would obviously also lead to a flux of interstitials (F_I) following the gradient of concentration.

MODELING OF ATOMIC SCALE DIFFUSION

We can rewrite Fick's laws at the microscopic level by assuming that the dopant - interstitial (or vacancy) interaction can be treated as a chemical reaction:

could be vacancy

$$\left[A + I \leftrightarrow AI\right]$$

mobile species

which, if we suppose to be in chemical equilibrium, we can write the concentration of AI as

kind like the mass action law

chemical react. constant

$$\left[C_{AI} = K\, C_A C_I\right]$$

$$F_{AI} = -d_{AI} \frac{\partial C_{AI}}{\partial x} \quad (*)$$

- barely useful, we don't know d_{AI} and k

we can relate microscopic diffusion to macroscopic diffusion:

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} D_A^* C_A \frac{C_I}{C_I^*} \left[\frac{\partial}{\partial x} \ln\left(C_A \frac{C_I}{C_I^*} \frac{n}{n_i} \right)\right]$$

$(*)$ $F_{AI} = -d_{AI}\left(k C_I \frac{\partial C_A}{\partial x} + k C_A \frac{\partial C_I}{\partial x}\right)$

$$\frac{\partial C_A}{\partial t} = -\frac{\partial F_{AI}}{\partial x} = \frac{\partial}{\partial x} d_{AI}\left[k C_I \frac{\partial C_A}{\partial x} + k C_A \frac{\partial C_I}{\partial x}\right] \Rightarrow \frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} d_{AI}\left[\frac{C_{AI}}{C_A} \cdot \frac{\partial C_A}{\partial x} + \frac{C_{AI}}{C_I} \frac{\partial C_I}{\partial x}\right]$$

(see d50 after for same calculations!)

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} D_A^{eff} \frac{\partial C_A}{\partial x} \quad (7)$$

we consider an homogeneous distribution of points defects, we neglect the second term.

$$\Rightarrow \frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} d_{AI} \left[\frac{C_{AI}}{C_A} \cdot \frac{\partial C_A}{\partial x} + \frac{C_{AI}}{C_I} \frac{\partial C_I}{\partial x} \right]$$

(second term crossed out)

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} D_A^{eff} \frac{\partial C_A}{\partial x}$$

since D_A^{eff}, effective diffusivity:

$$\left[D_A^{eff} = d_{AI} \frac{C_{AI}}{C_A} \right]$$

$$\Rightarrow D_A^{eff} = d_{AI} \frac{C_{AI}}{C_A} = D_A^* \cdot \frac{C_I}{C_I^*}$$

D_A^* → equilibrium diffusivity

$\frac{C_I}{C_I^*}$ → supersaturation of interstitials

$$D_A^* \to D^* \left[f_I \frac{C_I}{C_I^*} + f_V \frac{C_V}{C_V^*} \right]$$

($f_V \frac{C_V}{C_V^*}$ crossed out, ≈ 0)

so

$$F_{AI} = -d_{AI} \left[\frac{C_{AI}}{C_A} \frac{\partial C_{AI}}{\partial x} + \frac{C_{AI}}{C_I} \frac{\partial C_I}{\partial x} \right]$$

$$F_{AI} = -D_A^* C_A \frac{C_I}{C_I^*} \left[\frac{1}{C_A} \frac{\partial C_A}{\partial x} + \frac{1}{C_I} \frac{\partial C_I}{\partial x} \right]$$

$$= -D_A^* C_A \frac{C_I}{C_I^*} \left[\frac{\partial}{\partial x} \ln \left[C_A \cdot \frac{C_I}{C_I^*} \right] \right]$$

from 2nd Ficks law: $\frac{\partial C_A}{\partial t} = -\frac{\partial F_{AI}}{\partial x}$

$$\left[F_{AI} = -D_A^* C_A \frac{C_I}{C_I^*} \left[\frac{\partial}{\partial x} \ln \left[C_A \frac{C_I}{C_I^*} \frac{n}{n_i} \right] \right] \right]$$

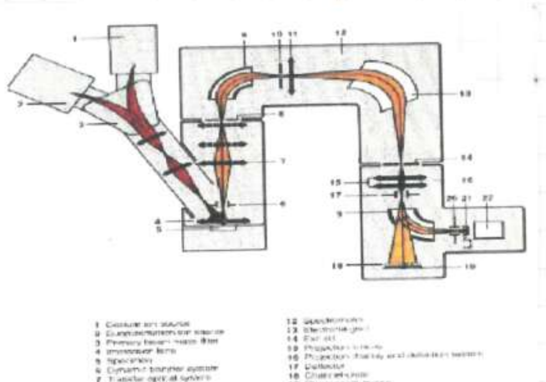
C_A: dopant concentration

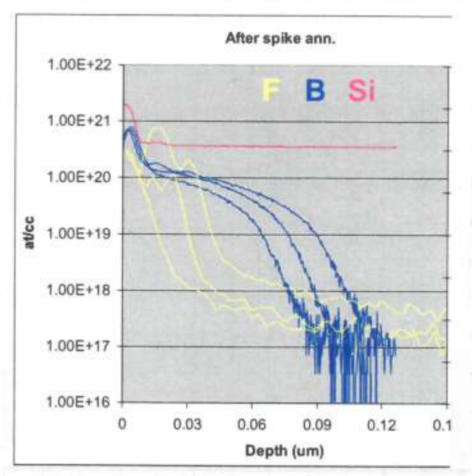
D_A^*: diffusivity of dopant @ equilibrium

SIMS (SECONDARY ION MASS SPECTROMETRY)

destructive, not precise, can't know if active dopants or not

the Sims





Very widely used to measure concentrations by bombarding my silicon wafer with an ion beam (made of Cs^+ or O^+). ← PRIMARY IONS

The impinging ions sputter sample (= wafer) atoms which then go through a mass analyzer (the sputtering ionizes the atoms of the wafer, ← SECONDARY IONS they get accelerated and go through a mass analyzer, where ∀ (mass / electric charge) ratio ∃! radius of curvature), which tells me what kind of species I'm producing and their concentration. Since as we sputter we keep going deeper (we remove more atomic layers), we can measure the concentration as a function of depth, provided we know the sputter rate of Cs^+ / O^+ on silicon.

NOTE: SIMS is a destructive technique and not very accurate, since the beam can be ~100s μm^2 (very big compared to transistor dimensions), so we get an average over that area.

We can see the dopant concentration profile as a function of depth in the picture in the right.

NOTE: with SIMS we can't know how many of the dopants were electrically active, since we can't know how many of them were sitting in substitutional position. To study that usually SSRM is used.

SSRM (Scanning Spreading Resistance Microscopy)

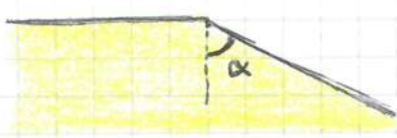
can find out if active

① water with a certain concentration profile

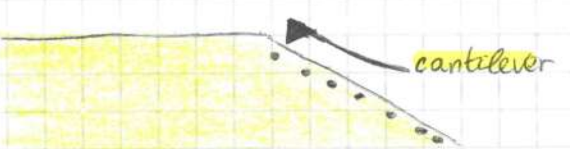
wafer

concentration

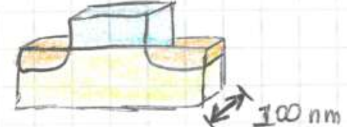
② cut the water at an angle α (usually gentle slope)



③ use an AFM (Atomic Force Microscope) and by measuring the current flowing through the tip I can measure the active dopant concentration as a function of depth



NOTE: more advanced SSRM techniques involve cutting a very thin (~100 nm) "lamella" of my device and running the AFM over it measuring the local resistivity / conductivity.



nanometer precision

ELECTRIC FIELD CORRECTION TO FICK'S LAW (cont.)

REMEMBER FOR EXAM

We assume doping >> intrinsic concentration

When we insert dopants into our silicon, the electrons (/holes) associated to them diffuse much faster, creating an extra electric field that accelerates diffusion (enhanced diffusivity).

So we must take into account the flux from 1st Fick's law plus a new flux F' (from $\vec{E}$ generated by electrons)

$$F_{TOT} = F + F' = -D\frac{\partial C}{\partial x} + C v$$

(flux can in general be written this way; v = speed of donor ions)

speed of donor ions due to $\vec{E}$ generated by e^-/holes $\Longrightarrow$

$$v = \mu \mathcal{E}$$

(μ = mobility, $\mathcal{E}$ = electric field) SPEED OF IONS

electric field

$$\mathcal{E} = -\frac{\partial V}{\partial x}$$

(V = potential in the crystal)

The higher the concentration (e^-/h^+) the more this phenomena are important, so if we assume a high concentration we can write the potential inside the crystal using the Boltzmann approximation

$$(\psi)\quad V = \frac{kT}{q}\ln\left(\frac{n}{n_i}\right)$$

(q = el. charge, n = concentration of e^-, n_i = intrinsic concentration of e^-)

$$V_{th}\cdot \ln\left(\frac{n}{n_i}\right)$$

We also know, from Einstein relashionship, that we can relate μ and D by:

$$\mu = \frac{D}{V_{th}} \qquad \mu = \frac{q}{KT} D$$

D → diffusivity

$$F_{TOT} = -D\frac{\partial C}{\partial x} + C v = -D\frac{\partial C}{\partial x} - DC\frac{\partial}{\partial x}\ln\left(\frac{n}{n_i}\right)$$

now we would like to find a way to replace the concentration of e^- with that of donors.

NEUTRALITY CONDITION
(assume all donors are ionized / acceptors)

$$N_D + p = N_A + n$$

- N_D: concentration of ionized donors
- p: concentration of holes
- N_A: concentration of ionized acceptors
- n: concentration of electrons
- $N_D + p$: total concentr. of positive charge
- $N_A + n$: total concentration of negative charge
- equal because crystal is neutral

MASS ACTION LAW

$$n \cdot p = n_i^2$$

n_i: intrinsic concentration

we can write

$$C = N_D - N_A = n - \frac{n_i^2}{n}$$

C: total concentration of dopants

From here it's just algebra, and the final result, as we have already seen, is:

$$\left[F = -h D\frac{\partial C}{\partial x} \quad \text{with } h = 1 + \frac{C}{\sqrt{C^2 + 4n_i^2}} \right]$$

see PDF on webeep

MODELING OF ATOMIC SCALE DIFFUSION (cont.)

As we have already seen, the diffusion happens because the atomic species (impurities) interact with point defects.

Let's consider an atomic species A that diffuses only through interactions with interstitials I (P or B, for example)

$$A + I \leftrightarrow AI \qquad (1)$$

mobile species

If we treat his as a chemical reaction, at "chemical equilibrium" the mass-action law holds true:

$$C_{AI} = K\, C_A\, C_I \qquad (2)$$

Mass-action law is true for every chemical reaction in chemical equilibrium

The assuption of treating (1) as a chemical reaction isn't just a working hypothesis, since we can experimetally see that whether we run a long or a short anneal, the mobile species AI tend to form very quickly and then reach a dynamic equilibrium, which is than kept through the whole anneal.

NOTE: the Fick's Law is true, the flux follows the gradient of concentration, but not the concentration of dopants but rather that of the dopant - interstitial complex.

or dopant - vacancy

diffusivity of AI complex

$$F_{AI} = -d_{AI} \frac{\partial C_{AI}}{\partial x} \qquad (3)$$

Fick's Law, considering AI, not only A.

But this equation isn't very useful because we can't directly observe C_{AI}, neither do we know d_{AI}.

Let's differentiate (2) and substitute into (3):

$$F_{AI} = -d_{AI}\left(K C_I \frac{\partial C_A}{\partial x} + K C_A \frac{\partial C_I}{\partial x}\right) \qquad (4)$$

At atomic scale, F_A = the flux of atomic species is zero, so the concentration of atomic species C_A doesn't change over time.
Why? F_A is zero because (for example) Boron isn't moving, it's the Boron-Interstitial couple that moves, so we have only a flux of Boron-Interstitial couple different from zero.
THIS flux is the only one that can change the concentration of Boron atoms over time.

2nd Fick's law

$$(5) \qquad \frac{\partial C_A}{\partial t} = -\frac{\partial F_{AI}}{\partial x} = \frac{\partial}{\partial x}\left[d_{AI}\left(K C_I \frac{\partial C_A}{\partial x} + K C_A \frac{\partial C_I}{\partial x}\right)\right]$$

NOTE: usually $C_{AI} \ll C_A$, since we could have $C_I \simeq 10^{13}$ atoms/cm^3, $C_A \simeq 10^{18}$ atoms/cm^3

so of those 10^{18} Boron atoms implanted, only 0,001% will have a chance of combining with an interstitial and then diffuse, with very high diffusivity.

From (2) we can rewrite K as: $K = \frac{C_{AI}}{C_A C_I}$
and substitute into (5)

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x}\left[d_{AI}\left(\frac{C_{AI}}{C_A}\frac{\partial C_A}{\partial x} + \frac{C_{AI}}{C_I}\frac{\partial C_I}{\partial x}\right)\right] \quad (6)$$

this is the microscopic Fick's Law, which we can compare to the MACROscopic one:

$D^{eff.} = D^*\left(\frac{C_I}{C_I^*} f_I + \frac{C_V}{C_V^*} f_V\right)$

↑ supersaturation

MACROSCOPIC FICK LAW

$$(7) \quad \frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} D_A^{eff.}\left(\frac{\partial C_A}{\partial x}\right)$$

assuming we can neglect $\partial C_I / \partial x$ (so we're away from Si surface, or away from the region we implanted, we can then assume the gradient of concentration of point defects to be negligable confronted with the gradient of concentration of dopant(s)), we can relate easily the microscopic and the MACROscopic Fick's Law:

microscopic | MACROscopic

$$\frac{\partial C_A}{\partial t} \simeq \frac{\partial}{\partial x} d_{AI}\left(\frac{C_{AI}}{C_A}\frac{\partial C_A}{\partial x}\right) = \frac{\partial}{\partial x} D_A^{eff.}\left(\frac{\partial C_A}{\partial x}\right)$$

$$\rightarrow \quad D_A^{eff.} = d_{AI}\frac{C_{AI}}{C_A} \quad (8)$$

but we also know that, in case of diffusion given only by interstitials ($f_I = 1$, $f_V = 0$):

$$D_A^{eff.} = D_A^* \frac{C_I}{C_I^*} \quad (9)$$

$$\rightarrow \quad d_{AI}\, C_{AI} = D_A^*\, C_A \frac{C_I}{C_I^*} \quad (10)$$

we can substitute (10) into (4):

$$F_{AI} = -d_{AI}\left(\frac{C_{AI}}{C_A}\frac{\partial C_A}{\partial x} + \frac{C_{AI}}{C_I}\frac{\partial C_I}{\partial x}\right) =$$

$$= -D_A^* C_A \frac{C_I}{C_I^*}\left(\frac{1}{C_A}\frac{\partial C_A}{\partial x} + \frac{1}{C_I}\frac{\partial C_I}{\partial x}\right)$$

and this flux expression contains only terms that we can measure, and also: IT WORKS!

We could correct it by also adding the electric field contribution

$$F_{AI} = -D_A^* C_A \frac{C_I}{C_I^*}\left[\frac{\partial}{\partial x} \ln\left(C_A \frac{C_I}{C_I^*}\frac{n}{n_i}\right)\right]$$

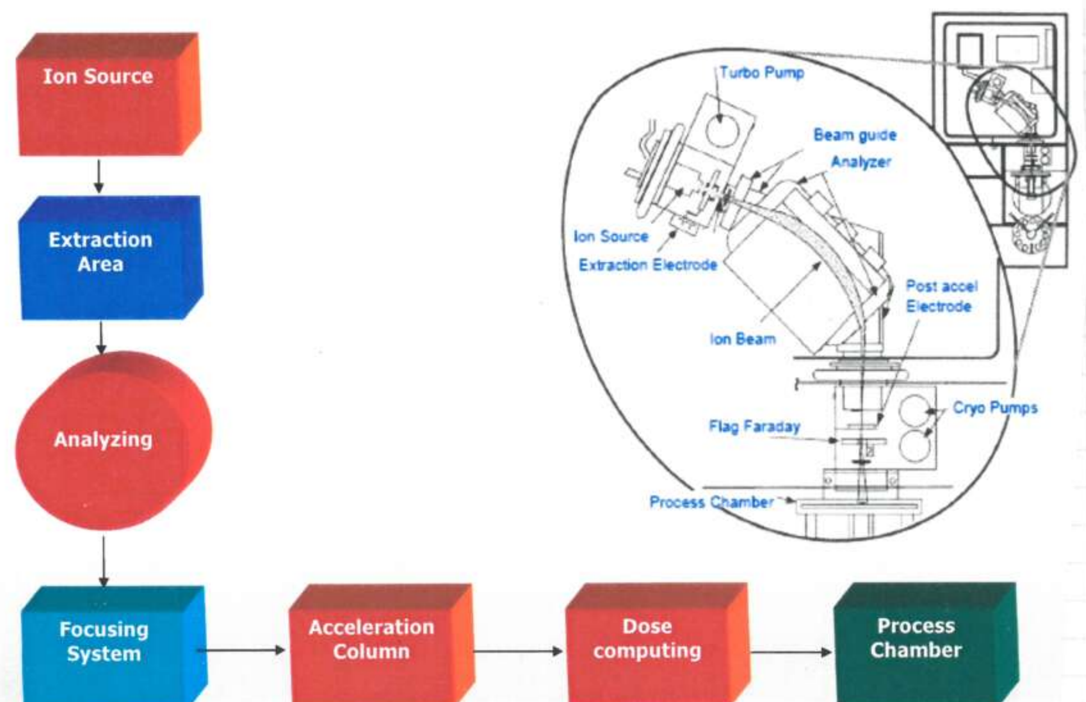
electric field contribution ($\frac{n}{n_i}$)

Lecture 12 9 aprile

ION IMPLANTATION

Diffusion is the study of how the dopants move inside the wafer, so naturally one thinks of also using it for doping the Si itself, like has been done for many years.
Ion implantation has been a huge step forward for device engineering since it allows for more precise dose control, easier masking, accurate depth control and also a large range of doses.

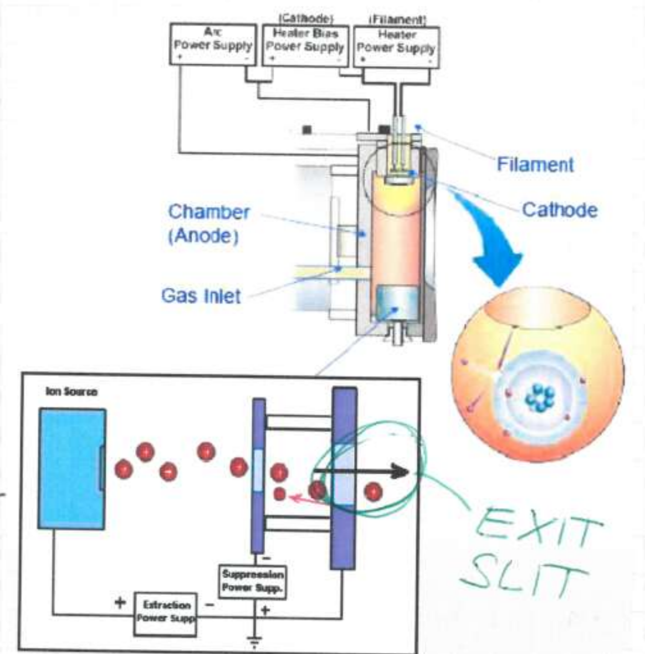
Let's see how implanters are made, what components do they have.



ION SOURCE and EXTRACTOR

The ion source is simply a chamber containing the gas form of the dopant as a plasma (borane BH_3, phosphine PH_3, arsine AsH_3, ...).

Gases are ionized by energetic electrons coming from an arc discharge.

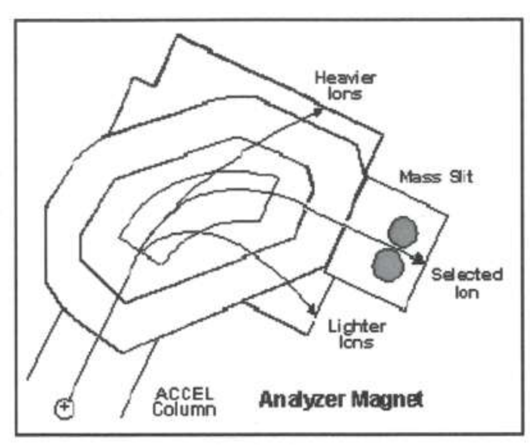


NOTE: a magnetic field is added so that electrons spiral (helicoidal trajectory), to increase ionization probability.

The filament is at a large negative potential with respect to the anode plate and the exit side (where the exit slit is) is biased at a large negative potential with respect to the filament.

ANALYZER

similar to the ones in SIMS! not the videogame



Once we get our ions out of the source chamber we have to select them:

we want only As^+ (for example), not H^+ or As^{++}, or As^{+++}.

How much ionized our dopant atom is matters because then the energy provided to $As^{+++} > As^{++} > As^+$, so the most ionized ones will accelerate more and reach greater depths, but we want perfect (or almost) control, so we use a mass analyzer (picture).

Ions are entering the mass analyzer with an energy given to them by the large voltage applied at the extractor (out of the source chamber). When they enter the chamber there's a magnetic field perpendicular to the plane of the paper (⊙ or ⊗). The ions will curve with a radius of curvature, which we can find:

$$V = \frac{qBR}{m}$$

$$\frac{mV^2}{R} = q\,|\vec{V} \times \vec{B}| = qVB$$

centripetal force (for $\frac{mV^2}{R}$); force acting on the ions (for $q\,|\vec{V} \times \vec{B}|$)

we can derive the speed just knowing the potential out of the extractor

$$V = \sqrt{\frac{2q V_{ext}}{m}}$$

V_{ext}: potential out of extraction chamber

$$\frac{q^2B^2R^2}{m^2} = \frac{2qV_{ext}}{m}$$

$$\sqrt{\frac{m}{q}} = \frac{1}{\sqrt{2V_{ext}}} RB$$

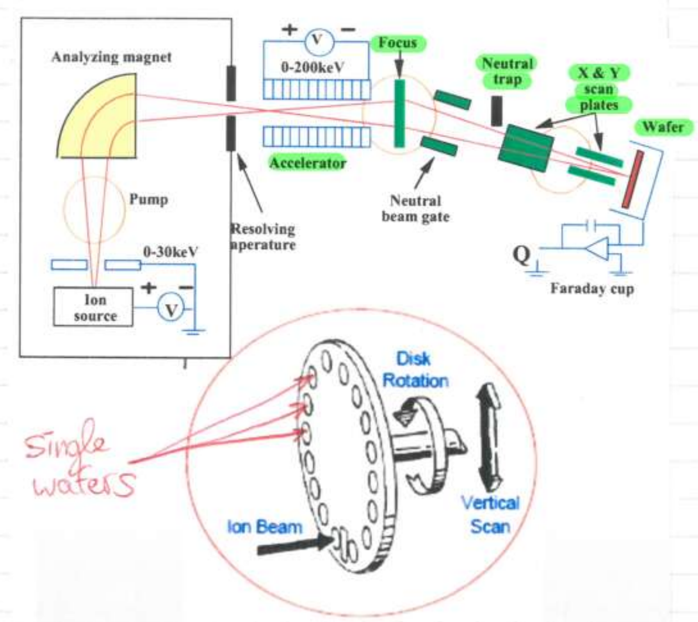
$\sqrt{\frac{m}{q}}$: ion mass/charge ratio

different ion mass/charge ratios can be selected by tuning $|\vec{B}|$

124

$$\frac{qBR}{m} = \sqrt{\frac{2qV_{ext}}{m}}$$

ACCELERATOR



If needed, the ions selected by the analyzer are further accelerated using a linear accelerator, where a linear electric field accelerates the ions (picture: "accelerator").
Then we have a magnetic lens focusing our ionic beam.
Then we have a neutral beam gate / neutral trap, which is just another electric field that just tilts my beam a little bit.

focus
neutral beam gate
neutral trap

That's because some ions can absorb thermal electrons and become neutral, so the neutral trap gets rid of them.

The problem with neutral dopant atoms is that we measure the implanted dose by measuring the current the ion beam produces, but neutral atoms won't produce a current → we will NOT know how many we have implanted. There's no problem associated with the fact that we're implanting ions into the wafer, since they'll be neutralized as soon as they get implanted.
After the neutral trap we have the X & Y scan plates, which are electric fields capable of moving the ion beam in the XY plane.
The wafers sit on top of a large disc which rotates, so that the ion beam can scan over a single wafer, then the disc rotates, doping of next wafer, ...

The wafers are in a Faraday Cup, which neutralizes the extra positive charge brought by the dopant ions. This neutralization is a current which can be measured and integrated to give us the dose. Again: if we can't measure the current, we can't know the implanted dose. ⟹ measure unused charge to find out how much have been implanted

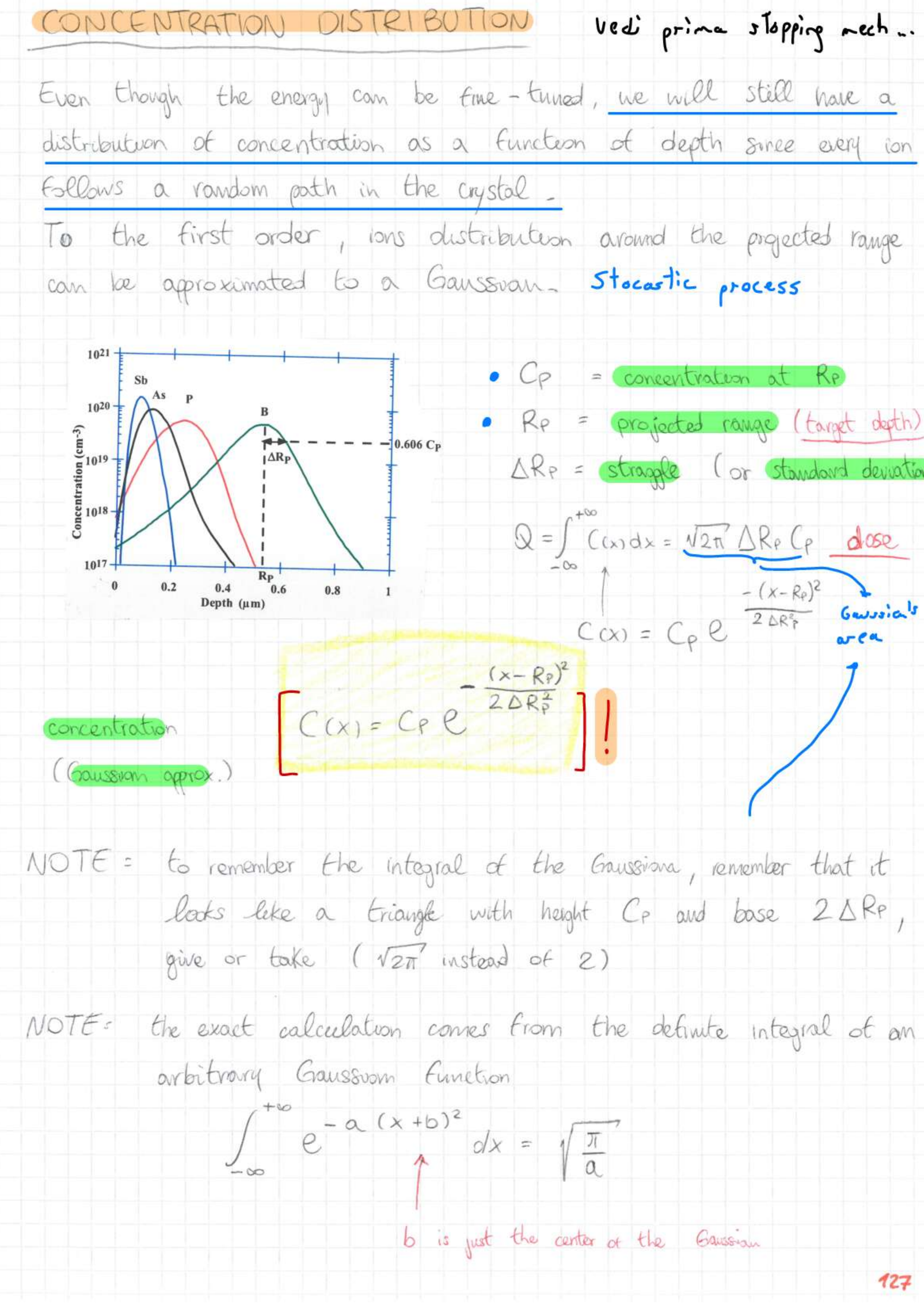
There's also a heating effect on the Si wafer brought by the ion beam, since they have high kinetic energy

$$\text{Energy} = \int \text{power } dt = \int IV \, dt = VQ$$

<u>example</u>:

for a dose of 10^{15} atoms/cm² at 60 keV → 10 kJ can be deposited on a 200mm wafer. (and that's considered low!)

The wafer can be cooled in the implanter itself, and the heating effect can even be exploited!



CONCENTRATION DISTRIBUTION

vedi prima stopping mech...

Even though the energy can be fine-tuned, we will still have a distribution of concentration as a function of depth since every ion follows a random path in the crystal.

To the first order, ions distribution around the projected range can be approximated to a Gaussian. Stocastic process

- C_P = concentration at R_P
- R_P = projected range (target depth)

ΔR_P = straggle (or standard deviation)

$$Q = \int_{-\infty}^{+\infty} C(x)\,dx = \sqrt{2\pi}\,\Delta R_P\, C_P \quad \text{dose}$$

$\sqrt{2\pi}\,\Delta R_P\, C_P$ → Gaussian's area

$$C(x) = C_P\, e^{-\frac{(x-R_P)^2}{2\Delta R_P^2}}$$

concentration (Gaussian approx.)

$$\left[C(x) = C_P\, e^{-\frac{(x-R_P)^2}{2\Delta R_P^2}} \right] \; !$$

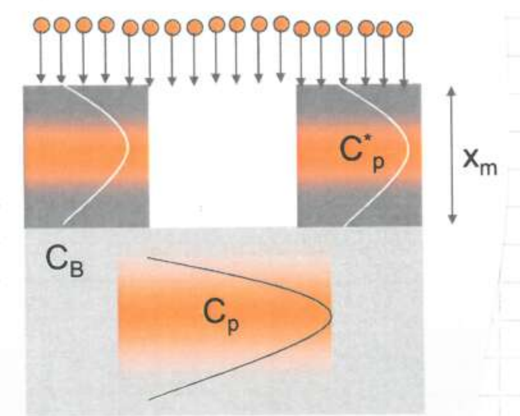
NOTE: to remember the integral of the Gaussian, remember that it looks like a triangle with height C_P and base $2\Delta R_P$, give or take ($\sqrt{2\pi}$ instead of 2)

NOTE: the exact calculation comes from the definite integral of an arbitrary Gaussian function

$$\int_{-\infty}^{+\infty} e^{-a(x+b)^2}\,dx = \sqrt{\frac{\pi}{a}}$$

b is just the center of the Gaussian

STOPPING POWER AND MASKING



One great advantage of ion implantation is that it's easy to mask, just by using photoresist. !
On the other hand, if we were to dope by gas diffusion we would need a material capable of withstanding the very high temperatures reached during gas diffusion (NOT photoresist).

For ion implantation we just run a lithography step and then use the photoresist as a mask, and after the implantation we can strip off the photoresist by using oxygen plasma or wet etching.

As also was for silicon, the concentration distribution in the photoresist can be approximated to a gaussian:

$$\left[\; C^*(x_m) = C_p^* \, e^{-\frac{(x_m - R_p^*)^2}{2\Delta R_p^{2*}}} \leq C_B \;\right]$$

Concentration in the photoresist

$C_p^* \neq C_p$, $R_p^* \neq R_p$, $\Delta R_p^{*2} \neq \Delta R_p^2$ since Si and photoresist are different materials they will have different stopping properties.

NOTE: we know that our photoresist has effectively "stopped" the ion implantation if the concentration at the bottom of the mask (near the Si interface) is lower than the bulk concentration of dopants C_B, which is the doping that gives the wafer its resistivity. So if C_B is due to Boron and we implant Phosphorous, we want the surface of the wafer to remain P type

we can calculate the thickness needed for the photoresist to have a concentration at the bottom which is at most the bulk concentration C_B:

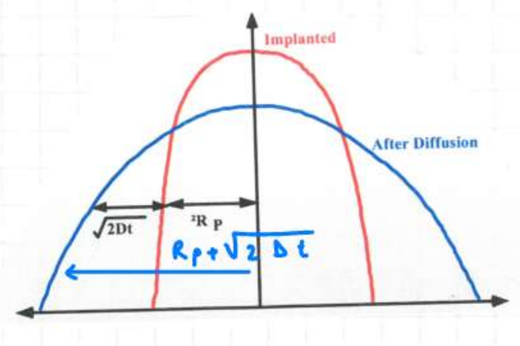
$$\left[X_m \geq R_P + \Delta R_P \sqrt{2 \ln\left(\frac{C_P^*}{C_B}\right)} \right]$$

! and since we have a Gaussian distribution, we will always have an amount of dopants penetrating through all the mask and ending on the wafer. That dose will be:

$$Q_P = \frac{Q}{\sqrt{2\pi}\,\Delta R_P^*} \int_{x_m}^{+\infty} e^{-\frac{(x-R_P^*)^2}{2\Delta R_P^{*2}}} dx = \frac{Q}{2} \operatorname{erfc}\left(\frac{x_m - R_P^*}{\sqrt{2}\,\Delta R_P^*}\right)$$

complementary error function

The fact that we have a Gaussian profile of dopants passing through the mask is convenient, since we have studied how concentrations with a Gaussian or δ profile diffuse as a function of t and T.

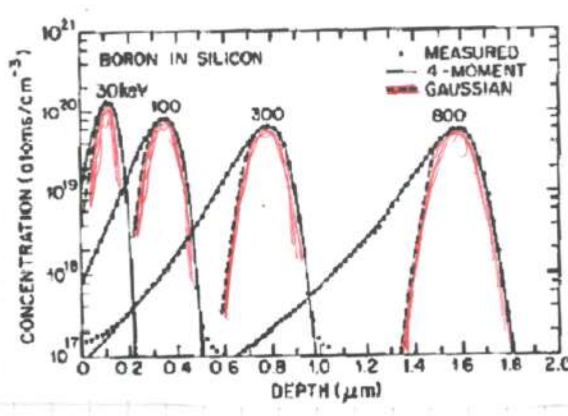


If we consider our Gaussian implanted distribution as the diffusion of a δ-function with $\Delta R_P = \sqrt{2Dt}$, then the effect of an annealing on our implanted distribution is:

$$C(x,t) = \frac{Q}{2\sqrt{\pi D t}} e^{-\frac{x^2}{4Dt}} \xrightarrow{\text{ANNEALING}} \frac{Q}{\sqrt{2\pi(\Delta R_P^2 + 2Dt)}} e^{-\frac{(x-R_P)^2}{2(\Delta R_P^2+2Dt)}}$$

still a Gaussian, just shorter and fatter, with a new sigma of $\sigma = R_P + \sqrt{2Dt}$

HIGHER ORDER MOMENTS



picture: real concentration profiles

Real concentration profiles can be approximated to a Gaussian only nearby the peak region.

An arbitrary function can be described by "moments":

moment of i-th order

$$m_i = \frac{1}{Q} \int_{-\infty}^{+\infty} (x - R_P)^i \, C(x) \, dx$$

center (R_P) — concentration distribution ($C(x)$)

(MOMENTI DI STATISTICA)

examples:

$$m_1 = \frac{1}{Q} \int_{-\infty}^{+\infty} x \, C(x) dx = R_P \quad \text{(average)}$$

$$m_2 = \Delta R_P$$ (the σ of the distribution)

$$m_3 = \gamma \Delta R_P^3$$ (γ = skewness, 0 for Gaussians)

$$m_4 = \beta \Delta R_P^4$$ (β = kurtosis, 3 for Gaussians)

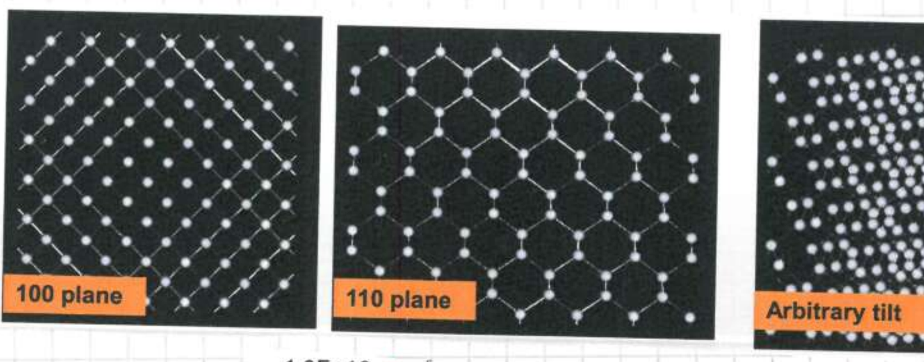
higher order moments can be used to generate more accurate distributions (es. Pearson's equation)

NOTE: the "zero" order moment is the Dose Q.

CHANNELING

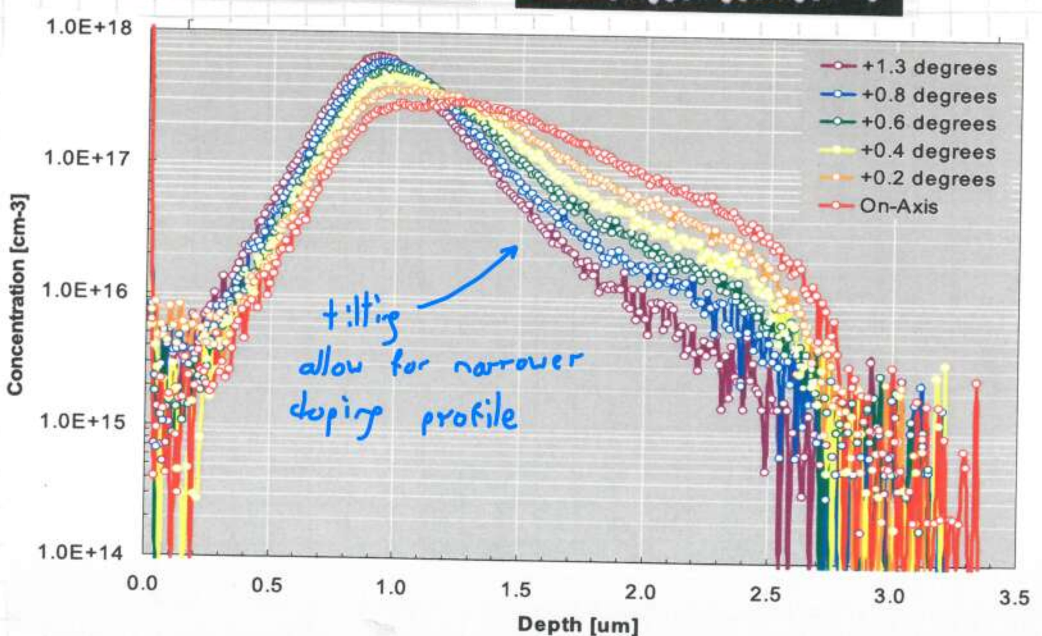
One of the reasons why the real distribution doesn't look like a Gaussian profile is a phenomenon called "channeling".

When we implant ions in a perfect crystal (or almost, but still a very high degree of simmetry), if the ion is moving along a symmetry direction in the crystal, the ion can channel in the lattice, resulting in a huge effect on doping profile. To avoid this effect we can do the implantation in amorphous silicon, use an amorphous screen oxide or just tilt the sample (picture).



effect of tilting by a small angle the wafer before implanting →

NB after implatation we need activation through heat



NOTE: to have a uniform doping, we do 2 implantations: $+\theta, -\theta$

NOTE: we don't want the channeling effect since it creates long tails of lucky ions who reach great depths (tails in the distribution).

STOPPING MECHANISMS

The trajectory of an ion in Silicon is deterministically known and it can experience two kind of interactions:

- ion - Si atom scattering (two body collision)
- electrons in the crystal slowing the ion down

we can write the energy lost as a function of depth as:

$$\frac{dE}{dx} = -N\left[S_n(E) + S_e(E)\right]$$

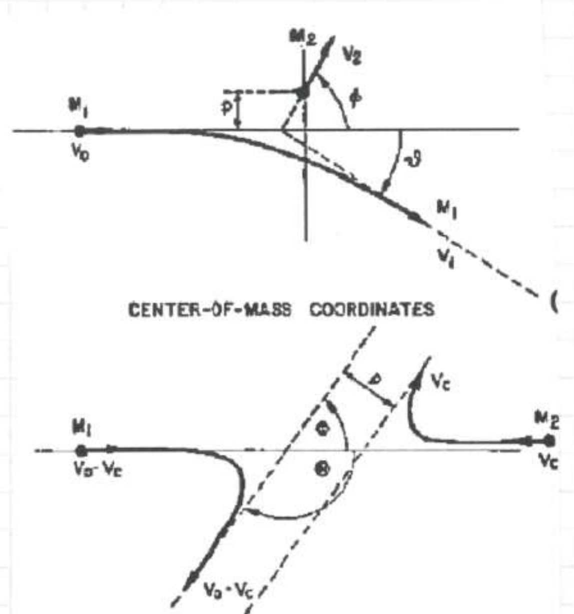
- N: target (Silicon) atom density
- $S_n(E)$: nuclear stopping power
- $S_e(E)$: electronic stopping power

from this we can derive the projected range R_p:

$$R_p = \int_0^{R_p} dx = \frac{1}{N}\int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)}$$

so all we need is a way to model $S_n(E)$ and $S_e(E)$

NUCLEAR STOPPING



Nuclear stopping power can be easily modeled as a Coulumb scattering, with a modified Potential in order to account for the shielding of electrons both from the Silicon atoms and the ion:

$$V(r) = \frac{q^2 Z_1 Z_2}{4\pi\varepsilon r} \, \Phi(r)$$

$\Phi(r)$ ← electron shielding correction factor

$S_n(E)$ is small at high energies → it dominates at the "end of the range" (where most of the damage is produced).

"End of range" = when the ion has sufficiently slowed down and is getting near the end of the projected range, its remaining energy is small & $S_n(E)$ dominates.

see picture in page 138

see comment in page 138

ELECTRONIC STOPPING

Electronic stopping can be divided into "local" and "non-local".

- **Non-local electronic stopping:**

  the positively charged ion travels through a sea of electrons, so a polarization field builds up, but it "lags" behind the moving ion, this creates a drag force similar to the motion in a viscous medium. This interaction is with the sea of electrons.

  OUTER ELECTRONS

- **Local electronic stopping:**

  Simply momentum exchange (= collision) with orbital electrons.

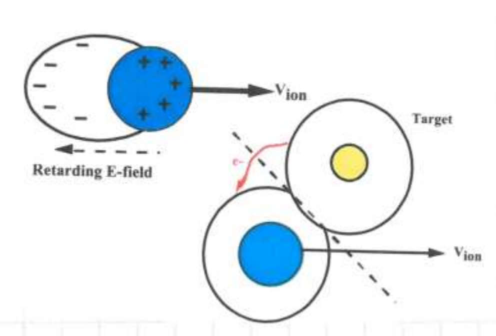
  INNER ELECTRONS

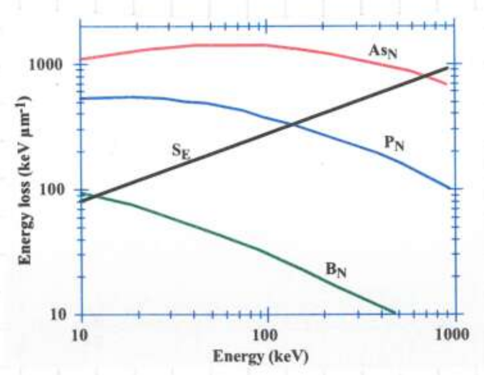
Both this mechanisms depend on ion velocity, so we can write

$$\left[S_e(E) = k\sqrt{E} \right]$$

k: proportionally constant; $\sqrt{E} \doteq$ velocity

NOTE: since $S_e(E)$ depends on speed ($\sqrt{E}$), it will dominate at the beginning of the trajectory, where the ion's speed is still high.





TRIM - ION IMPLANTATION

DAMAGE

Displacement energy in Si ≃ 15 eV

(energy needed to dislodge a Si atom from its place in the crystal → formation of interstitial and vacancy pair (**Frenkel pair**))

CASCADE

When we do ion implantation, **we can have 100 KeV or even more**, more than enough to have every ion able to dislocate thousands of silicon atoms, so **crystalline damage during ion implantation is unavoidable**.

While the ion stops some of the Frenkel pairs will ricombine, but still the damage is significant.

Every new ion creates an additional damage that can be written as:

$$\Delta n(x) = n \, f_{rec} \left(1 - \frac{N}{N_\alpha} \right)$$

- $\Delta n(x)$: damage accumulation
- n: number of defects generated in a single cascade
- f_{rec}: frequency recombination within a cascade
- N: pre-existing damage from previous cascades ($MAX = N_\alpha$)
- N_α: threshold defect density after which the crystal is considered amorphous

NOTE: for an amorphous crystal ($N = N_\alpha$) we have $\Delta n = 0$, so **no incremental disorder is introduced by additional cascades.**

NOTE: **as the crystal gets more damaged (N increases), the incremental damage brought by new ions decreases (Δn slows down).**

SPE = SOLID PHASE EPITAXY

method to fix damage

No matter what, with ion implantation we will end up with a significant crystalline damage. We could even amorphize the silicon! !
This is not a problem because every ion implantation will be followed by an annealing process, which is very important for 2 reasons:

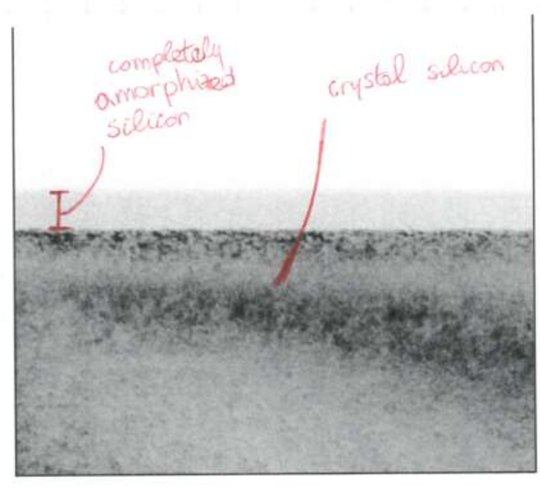
- electrically activate the dopants
- repair the crystalline damage

} annealing role

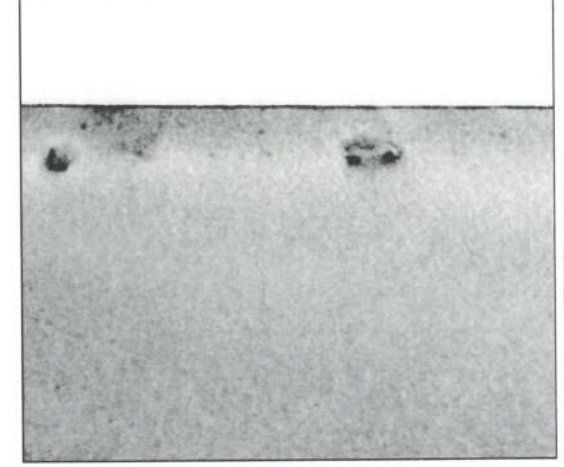
If the Si crystal has been amorphized during implantation, the way to repair it is through Solid Phase Epitaxy (SPE):
if you give the crystal enough energy, the amorphous part will re-crystallize using the underneath crystalline Silicon as a template (or as a "seed") and will crystallize again, at solid phase. → LOW TEMPERATURE
SPE is a quick process (50 nm / min at 600°C in <100>).
The activation energy is 2,3 eV (energy of Si - Si bond breaking / formation)

NOTE: if we didn't completely amorphize the crystal, in the early stages of annealing (or during the ion implantation itself) most of the Frenkel pair would recombine, leaving us with roughly 1 interstitial per dopant atom (this is called the +1 model). But if we go at high doses and / or energies we could amorphize the silicon.

+1 MODEL ⇒ 1 interstitial per dopant atom
all other interstitials recombined thanks to temperature, but one must remain (where the dopant is)

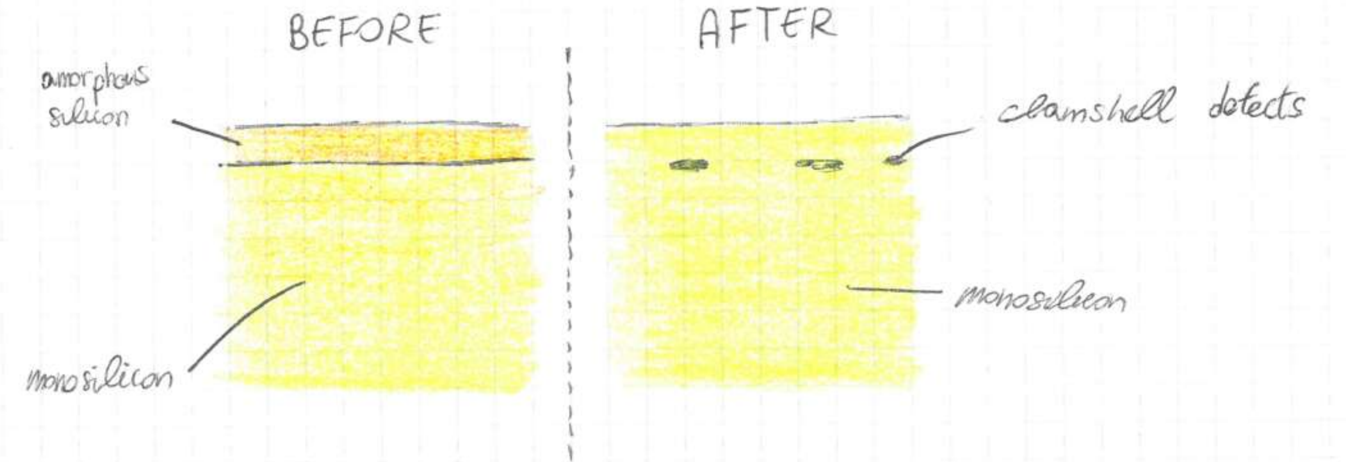


After implant



After annealing: end of range defects are visible

In this TEM pictures we can see (left) that the surface has been completely amorphized by the implantation process, and the crystal re-formed after an annealing (right) thanks to SPE phenomenon. On the right picture we don't have anymore the amorphous surface, but we can recognize "clam shell defects", which are the black spots. That happens because during SPE defects "agglomerate" as the crystal grows, and since during SPE the crystal re-forms from every direction, we end with a clamshell shape.
Clamshell defects tend to be located roughly at the end of where the amorphized region was before SPE, they're at the end of range*.



* "end of range" = at the end of the range of the implanted ions, which at low energies/speed get mostly slowed down by S_n.

Lecture 13

14 aprile

NOTE: if I was to amorphize a layer surrounded by monosilicon then I would get clamshell defects in the middle of the amorphized region, because during SPE both the top and the bottom monosilicon start growing at the expense of the amorphized region.

NOTE: EOR defects occurr because there is a large amount of damage (which is just below the threshold of amorphization) beyond the a/c interface. By definition, just beyond the a/c interface we have the maximum possible amount of damage that can exist in the crystal without it being amorphous (at the implant temperature). This damage is sufficiently high to be able to nucleate a mixture of {311} defects in a very narrow region just beneath the a/c interface, on the crystalline side of the interface.

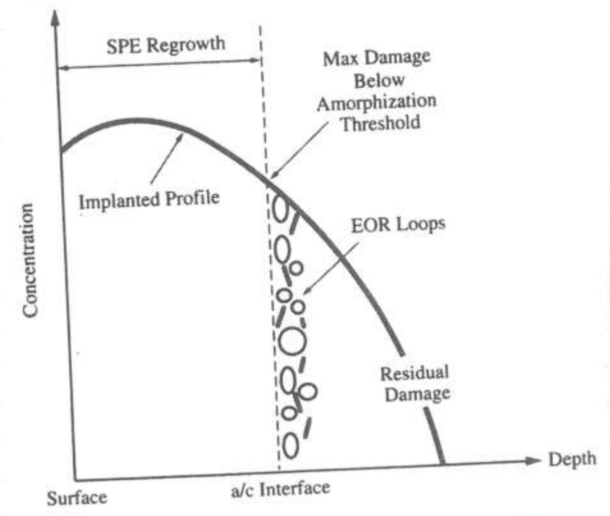


Figure 8-25 Schematic of the stable End-Of-Range (EOR) dislocation loops that form at the amorphous/crystalline (a/c) interface after solid-phase epitaxial regrowth.

+1 MODEL

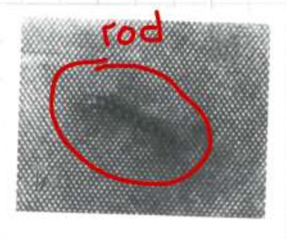
It's a simple but powerful model:

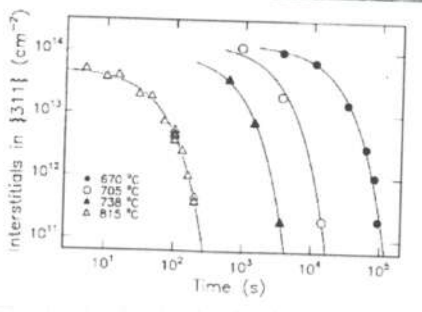
at relatively low T (~ 400°C) most of the interstitials and vacancies (Frenkel pair) recombine, but once the doping atoms have occupied their final lattice sites only the interstitials are left, hence we have +1 excess interstitials → "+1 model".

There's no experimental proof, but it gives a good estimate of excess interstitials, even though sometimes it might need corrections, becoming +2 or +3, but still simple and powerful.

What happens during the annealing process to this +1 excess interstitials?

Upon further annealing processes, the first thing they do is agglomerate in small rod-shaped clusters on {311} planes extending in the <110> direction

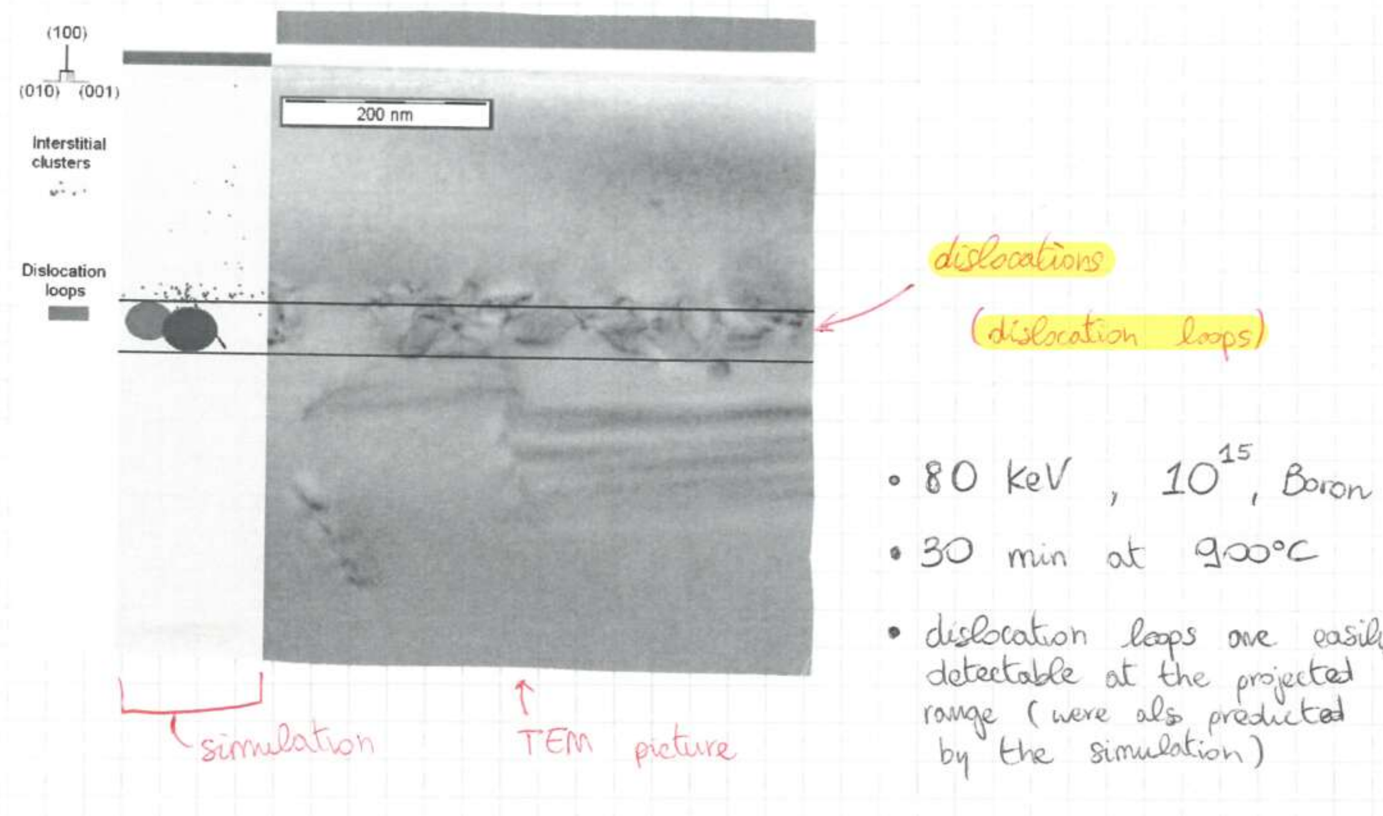




+1 interstitial agglomerate

Depending on the amount of damage (and distance from the surface) this clusters can:

- disappear by "evaporation" (emitting interstitials which then recombine at the surface)
- agglomerate even more by keeping absorbing interstitials and form dislocation loops, which can (1) agglomerate further to form huge dislocations or (2) dissolve at very high T (> 1100°C)



ACTIVATION

Once we have doped our water, we have to make sure that they sit in substitutional position (the dopants), so that they share their e^- / holes and become electrically active. This is done by a (relatively high T) anneal.

LOW DOSE
HIGH DOSE
1.0
0.1
0.01
Fraction Active
8×10^{12} cm^{-2}
2.5×10^{14}
2×10^{15}
Reverse Annealing
400 500 600 700 800 900 1000
T_A (°C)

As we can see (picture) we have peaks and valleys of activation as a function of T (some processes that compete with dopants activation - so lower the active fraction - "wake up" at a certain T, but still for high T we keep activating dopants).

example: Boron will form Interstitial - Boron inactive complexes that compete energetically with substitutionally positioned Boron.

"Reverse annealing" usually happens at very high doses (near solubility limit), instead of substitutional position the dopants can form precipitates.

Once you activated your dopants it's not taken for granted that they will stay there because maybe you activated a good fraction of dopants at very high T, but then later on in the process flow you can have a lower T step that will deactivate a portion of your dopants, because we might initiate a new phenomenon that competes with the activation process and creates inactive complexes.

ANNEALING VS DIFFUSION

Annealing after implantation is mandatory to dissolve defects, to re-grow crystalline silicon (if necessary) and to activate dopants

BUT

at high temperatures diffusion also occurs, and we know that diffusion can be enhanced by defects induced by implantation (Boron diffuses 100% because of interactions with interstitials), this phenomenon is called TED (Transient Enhanced Diffusion).

TRANSIENT ENHANCED DIFFUSION

diffusion caused by thermal annealing

more diffusion @ low T!

Left chart: Concentration (cm^{-3}) vs Depth (μm), 10^{16} to 10^{19}, 0 to 0.4; labels: 2 min 800°C, 10 sec 1000°C

Right chart: Concentration (cm^{-3}) vs Depth (microns), 10^{16} to 10^{21}, 0 to 0.4; labels: As profile, Boron profiles, Initial, Final

TED happens when interstitial damage from the implant enhances the dopant diffusion for a brief transient period.

It is the dominant effect today that determines junction depth in shallow junctions.

It is an anomalous diffusion because profiles can diffuse more at low T compared to high T for the same $D \cdot t$ (D: diffusivity, t: time)

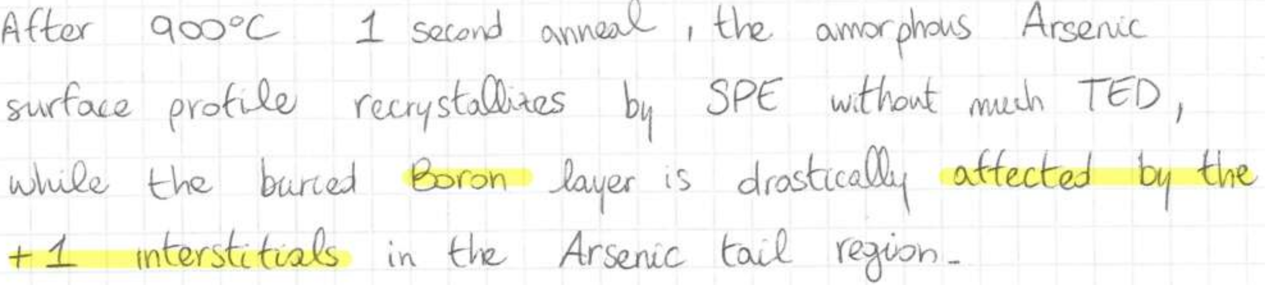
The basic model for TED assumes that all the implant damage recombines rapidly, leaving only 1 interstitial generated per dopant atom when the dopant atom occupies a substitutional site (+1 model). ⟹ we consider only +1 interstitials

picture (left):

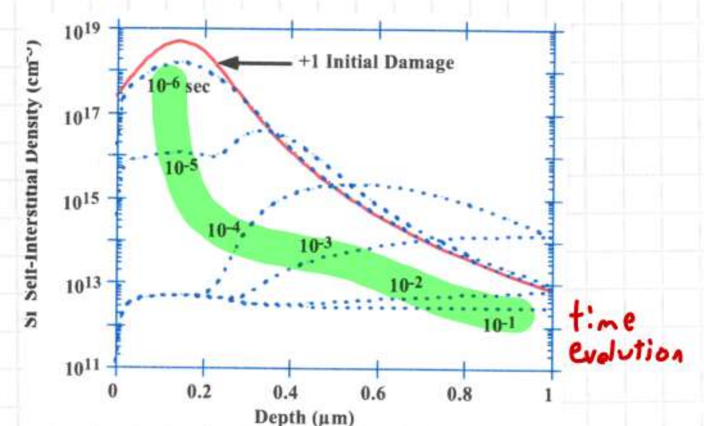
the Dt for the two profiles is roughly the same $\left(D(T) = D_0 e^{-\frac{E_a}{KT}}\right)$ but we see Boron diffusing much further away at low T compared to high T

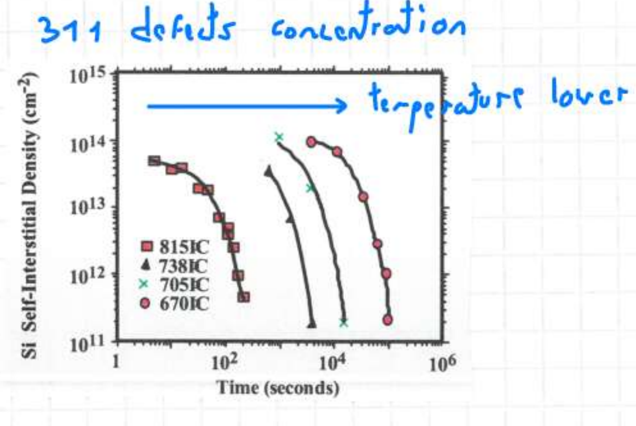
picture (right):

another anomalous diffusion happen when first we implant Arsenic then Boron (black profiles) then we anneal (green profiles). Arsenic doesn't move much while Boron diffuses a lot. Why?



After 900°C 1 second anneal, the amorphous Arsenic surface profile recrystallizes by SPE without much TED, while the buried Boron layer is drastically affected by the +1 interstitials in the Arsenic tail region.





picture (left):

Boron implants followed by 750°C anneals from 10^{-6} to 10^{-1} seconds, so we see the evolution in time.
Rapid clustering of interstitials into {311} rod defects reduces the enormous supersaturation of interstitials immediately after implant in a time scale of 10^{-2} seconds, and subsequently the supersaturation level is maintained approximately constant while the clusters are dissolving, until the clusters eventually disappear.
Most of the TED occurs during this period when the supersaturation of interstitials is at a high constant level, allowing the maximum doping motion to occur.

NOTE: {311} clusters are metastable, so they can be stable even for seconds ~ minutes! They form rapidly (and shrink during the anneal), driving the TED by emitting interstitials.
(at low T)
The higher the T of the anneal, the shorter the {311} lifetime, so the shorter the TED.

NOTE: by 0.1 seconds, at 750°C the {311} defects have formed and $C_I \approx 10^{13}$ cm^{-3} (concentration of interstitials), but $C_I^* \simeq 10^8$ cm^{-3} at 750°C ⟶ the enhancement is $> 10^5$!

picture (right, previous page):

we can see that on a much larger time scale, the {311} clusters decay. so TED can last hours at very low T, minutes at intermediate T and ms at very high T.

So: at 10^{-6} seconds we have our +1 initial damage, as time passes defects recombine but at $\sim 10^{-2}$ seconds we have a metastable cluster of {311} defects formed, keeping the interstitial supersaturation constant and driving TED. At much longer time scales (s, min, hours, ...) this clusters dissolve: at 670°C the concentration of {311} defects stays constant for $\sim 10^4$ seconds (~ 3 hours) before starting to dissolve (10^{11} concentration still after almost 30 hours!), while at 815°C they start dissolving after a couple of seconds!.

This kind of anomalous diffusion (TED) can explain the reverse short channel effect, where the damage induced by source and drain junctions implants "pushes" the channel species (Boron) towards the surface, shortening the channel and increasing the threshold voltage. The shorter the channel, the more pronounced the effect.

The way to solve this is to go at much higher T for very short periods of time, because increasing T reduces the lifetime of {311} clusters (thus reducing TED), but increases the amount of intrinsic interstitials (C_I^*), so we have again the problem of high diffusivity, but at a lower scale.

TO RECAP:

- Ion implantation is the mainstream doping method in IC industry nowadays because of a wide range of doses and energy (→ depth), precise dose control and easier masking.
- Implanted profiles can be thought as Gaussians (with some corrections)
- Silicon crystalline structure is important (remember the channeling effect)
- Crystalline damage is produced for every ion implantation (silicon displacement energy is much lower than any implant energy)
- Damage can be recovered by annealing but can also generate extended defects
- Damage (especially interstitials) plays a key role in implanted species diffusion (TED)

NOTE: pre-amorphizing (either by self-Si implantation or by implanting larger molecules, like BF_2) and then annealing (to re-crystallize via SPE) can also reduce significantly dopant diffusion, but increases EOR defects.

FILM DEPOSITION

Lecture 14 — 16 aprile

THIN FILM DEPOSITION

Many layers are needed in an IC process flow above the silicon substrates.

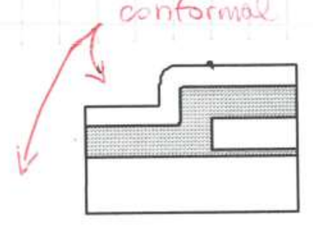
We are interested in:

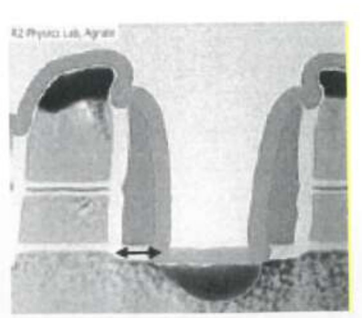
- film quality (composition, mechanical / electrical properties, defectivity / contamination levels, ...)
- film thickness control (from deposition to deposition, within the wafer, as a function of wafer topology, ...)

For example, we're interested in the step coverage (conformality) of the deposited layers, which is the property of showing a constant thickness on horizontal and vertical surfaces.

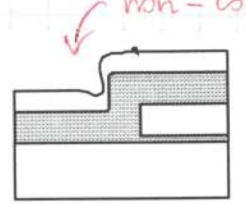
conformality

conformal

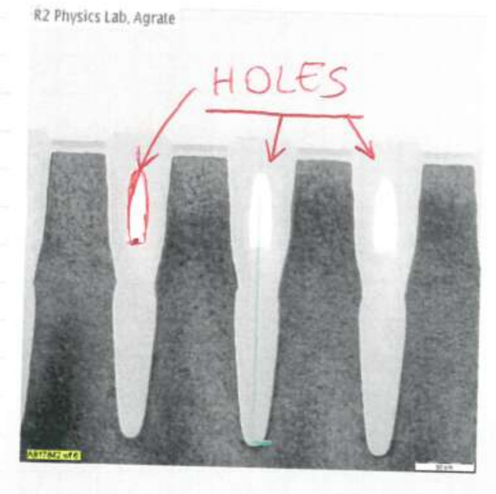


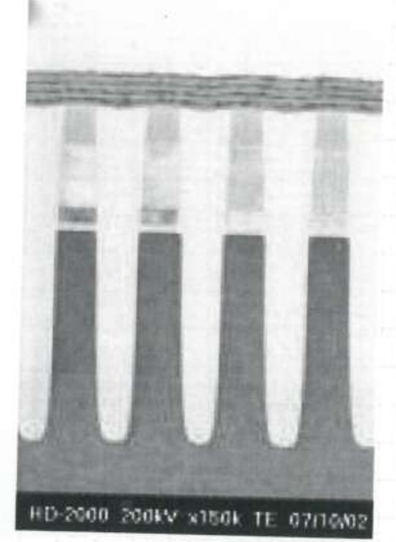


non-conformal



We might also want films that are good for gap-filling, so for filling deep holes (high aspect ratio) without forming holes





The main methods for film deposition are PVD and CVD, although also liquid coating + solidification (photoresist, Spin On Dielectric) ↑SOD and electrolitic deposition (Cu) are used.

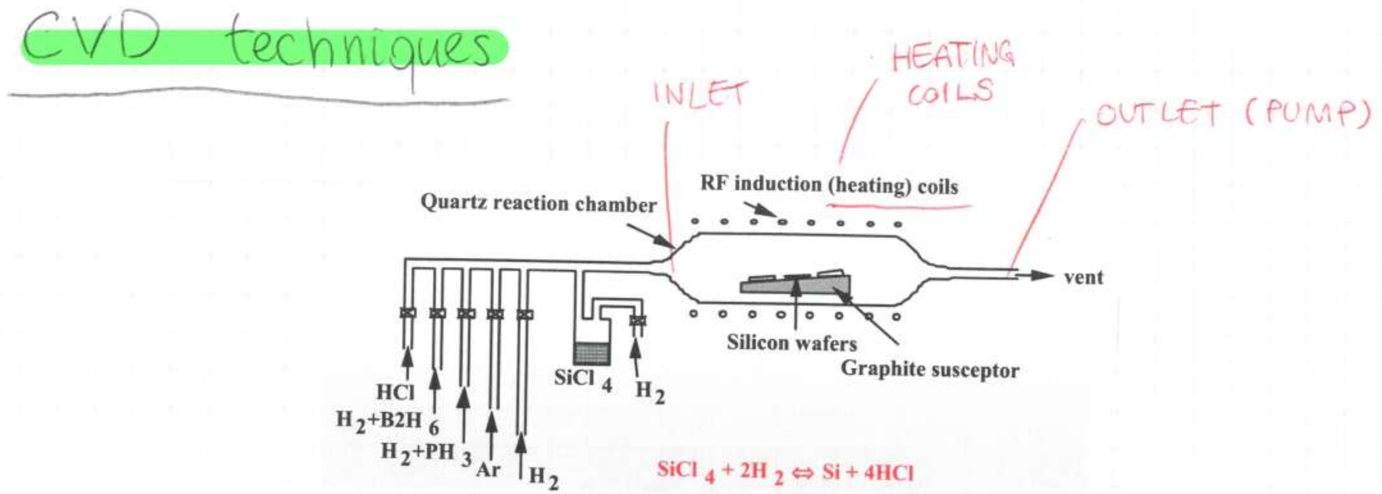
- CVD (Chemical Vapor Deposition) (reaction)

We introduce reactant gases entering the deposition chamber, they either react with one another or will react at the surface of the wafer, and the final compound will be at solid state at the deposition temperature (silane $SiH_{4(g)} + O_{2(g)} \rightarrow SiO_{2(s)} + 2H_{2(g)}$ @800°), this will be our thin film, we'll also have byproducts always in gas phase, which can be pumped away

- PVD (Physical Vapor Deposition) just physical deposition

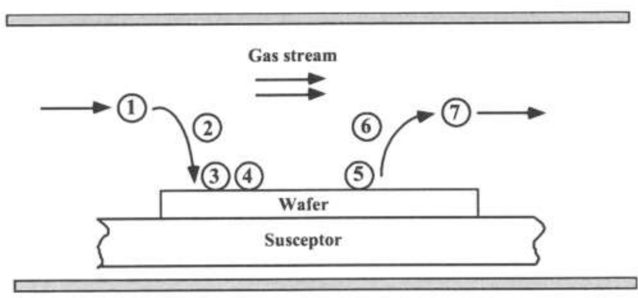
Most of the time we don't have any chemical reactions, just a solid piece of the material you wanna deposit, then you disgregate the material (by evaporation or sputtering) and let it re-condensate on the wafer surface

CVD techniques



the simplest CVD process is APCVD (Atmosferic Pressure CVD) which has been widely used, especially for epitaxial Si deposition. Wafers are held on heated graphite susceptors and reactant gases flow in the chamber at atmospheric pressure.

This process is almost abandoned in modern IC manufacturing.

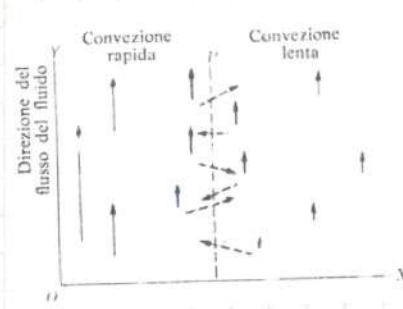


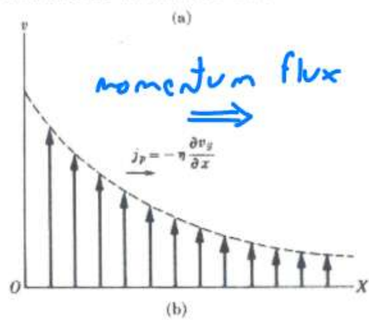
1. transport of reactants to the deposition region
2. transport of reactants from the main gas stream through the boundary layer to the wafer surface
3. adsorption of reactants on the wafer surface
4. surface reactions, including: chemical decomposition / reaction, surface migration to attachment sites, site incorporation and other surface reactions (e.g. emission and re-deposition)
5. desorption of byproducts
6. transport of byproducts through boundary layer
7. transport of byproducts away from the deposition region (pump away)

ONLY STEPS 2 - 5 WILL BE CONSIDERED TO MAKE OUR SIMPLE MODEL

We want to model the deposition rate in a CVD process, but before, let's talk about...

BASIC TRANSPORT MECHANISMS





Let's imagine we have a gas moving upward, and in the picture (left) we see the vectors representing the molecule's momentum. We have a gradient of speed along x (from left to right), which when we consider the molecules interacting with one another becomes a Fick-like law of momentum transfer:

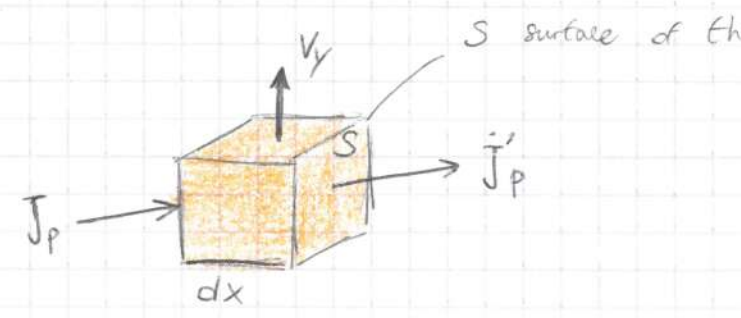
shear stress

$$\left[J_P = -\eta \frac{\partial V_y}{\partial x} \right]$$

- J_P: momentum flux
- η: viscosity (proportionality constant)
- $\frac{\partial V_y}{\partial x}$: speed gradient along x

$$\phi = -\rho V \eta \frac{\partial V}{\partial x}$$

so we have a flux of momentum (not mass like for Fick's law)
As we did for Fick's Law, we can consider:



cube with side dx, incoming momentum flux J_P, outgoing mom. flux J_P'

- if $J_P > J_P'$ the concentration of ENERGY inside the cube is increasing
- if $J_P < J_P'$ the opposite is happening

the total momentum gain for unit time in the volume $dV = S\,dx$ is:

(→ cube volume)

total momentum flux in the cube surface $\quad -\dfrac{\partial J_P}{\partial x} S\,dx$

$\dfrac{m \cdot v}{Volume} = \rho \cdot v$

and if p_y is the momentum per unit volume, then the total momentum gain for unit time is also

momentum gain in the cube per unit time $\quad \dfrac{\partial p_y}{\partial t} S\,dx$

Force, if external forces are present the momentum changes over time in the volume

$$\rightarrow \quad \frac{\partial p_y}{\partial t} = -\frac{\partial J_P}{\partial x} + F$$

$J_P = -\eta \dfrac{\partial V_x}{\partial x}$

if we divide everywhere by the density ρ of the gas:

$$\frac{\partial V_y}{\partial t} = \frac{\eta}{\rho} \frac{\partial^2 V_y}{\partial x^2} + \frac{F}{\rho}$$

if no external forces are present

$$\left[\frac{\partial V_y}{\partial t} = \frac{\eta}{\rho} \frac{\partial^2 V_y}{\partial x^2} \right]$$

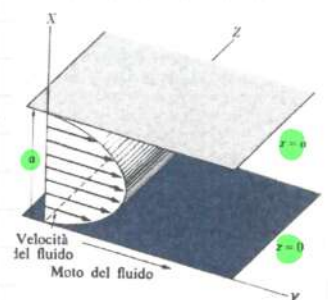
very similar to 2nd Fick's Law (remember Fick's law)

NOTE: the speed is along y, the gradient (derivative) is along x.

STEADY STATE CASE

Let's consider a simple case: $\frac{\partial v_y}{\partial t} = 0$

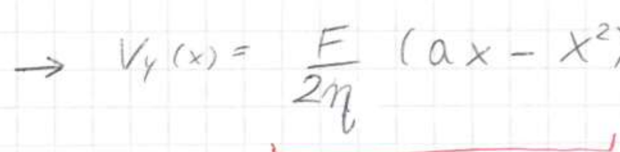
steady state



$$\frac{\partial^2 v_y}{\partial x^2} = -\frac{F}{\eta}$$

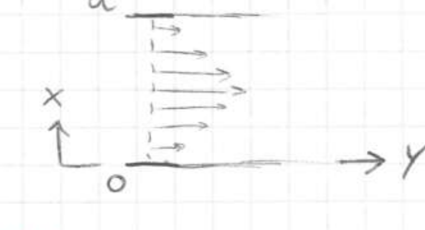
solution $V_y(x) = -\frac{F}{2\eta}x^2 + C_1 x + C_2$

we can find C_1, C_2 by the boundary conditions, so let's consider $V_y = 0$ for $x = 0$ and $x = a$ (water flowing in a river, in a tube, ...) and $F =$ const.



$$\rightarrow V_y(x) = \frac{F}{2\eta}(ax - x^2)$$

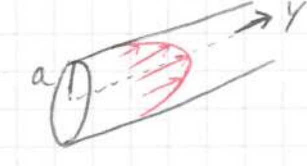
parabolic speed profile



Similarly, if we consider a cylinder of radius a, we get:

$$v_y(r) = -\frac{1}{4\eta}\frac{dp}{dy}(a^2 - r^2)$$

the Force is simply the pressure gradient along y



STAGNANT LAYER

In a real tube we won't always have the steady state condition since we also need to pump in the gas and pump away the byproducts.

When we pump in gases, we will insert a "plug" flow (flow with uniform velocity profile, not parabolic

initially we insert this →

$v=0$ | $v = v_{pump}$ | $v=0$ ... $v=0$ | v_{max} | $v=0$ → z

at the input the speed profile is uniform

after a while the transovent is done, the profile is parabolic

How much will it take for the fluid flow to reach the steady state?

It depends on the Reynolds number N_{re}

$$N_{re} = v_{pump} \frac{L \rho}{\eta}$$

(L: radius of the tube (a) — CHARACTERISTIC LENGTH)

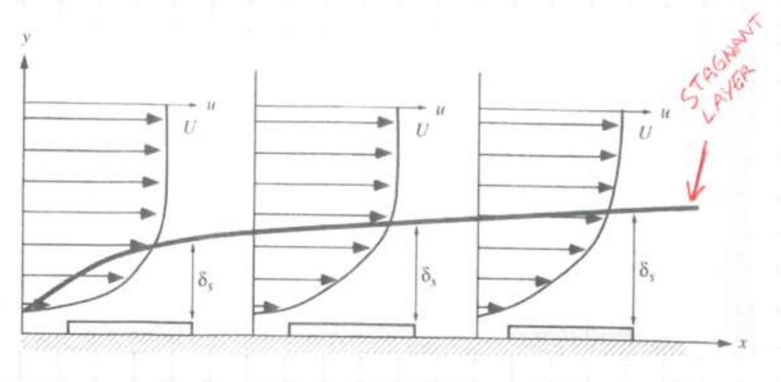
and the tube length needed will be $z_v \approx \frac{a}{25} N_{re}$ (a: radius of tube)

For large N_{re} the gas will mantain uniform velocity v_{pump} for a large portion of the tube.

So the velocity profile can be approximated as constant along r, apart from boundary / stagnant layer δ_s, where the flow speed is 0. It can be demonstrated that

stagnant layer thickness

$$\left[\delta_s(z) \approx \sqrt{\frac{\eta z}{\rho v_{pump}}} \right] \quad \left(\delta_s \propto \sqrt{z} \right)$$

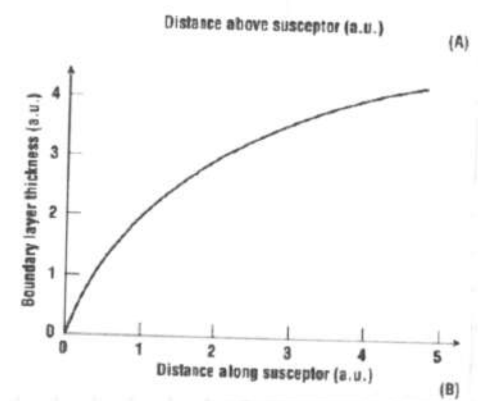


In a real tube usually the gas used have high N_{Re} and we can approximate the speed profiles as in the picture (left).

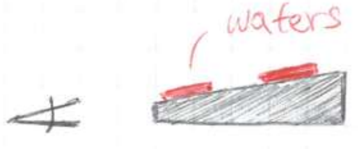
as we keep going →, the stagnant layer goes as $\sqrt{distance}$

stagnant layer profile :

not costant! bad, we want equal exposition



this is the reason why, in the picture at page , the susceptor has inclined surface, instead of being perfectly horizontal



susceptor holding the wafers

NOT like this:

The graphite susceptors have an inclined surface to take into account the fact that the stagnant layer isn't constant across the tube.

FLUX EQUATIONS FOR APCVD

stagnant layer

C_G

F_1

C_S

F_2

GAS

SILICON

apcvd = atmospheric pressure CVD

flux

$$F_1 = h_G (C_G - C_S)$$

$$F_2 = k_S C_S$$

flux

h_G = mass transport coefficient

k_S = chemical surface reaction rate

C_G = concentration of reactants in the gas flow

C_S = concentration of reactants at the surface

In the Deal-Grove model we also had F_3, but here we consider the chemical reactions always happening at the surface, so we have only F_1 = mass transport across the boundary / stagnant layer and F_2 = chemical reaction at the surface.

Assuming stationary conditions we can write $F_1 = F_2$

$$h_G (C_G - C_S) = k_S C_S$$

$$C_S = C_G \left(1 + \frac{k_S}{h_G}\right)^{-1} = \frac{h_G C_G}{h_G + k_S}$$

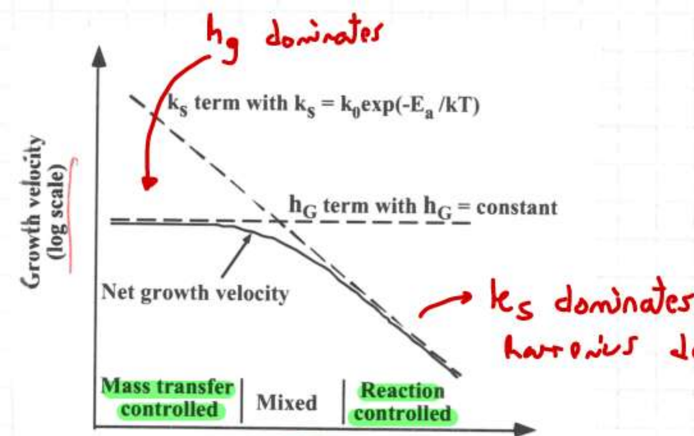
we can write the speed / rate of deposition as flux F divided by the density N (number of atoms incorporated per unit volume)

$$v = \frac{F}{N} = \frac{k_S C_S}{N} = \frac{k_S}{N}\left(\frac{h_G C_G}{k_S + h_G}\right) = \frac{k_S h_G}{k_S + h_G} \frac{C_T}{N} Y$$

where Y = molar fraction
C_T = total concentration of gaseous species
P_G, P_T = partial pressure / total pressure of incorporating species

$$\Upsilon \equiv \frac{C_G}{C_T} = \frac{P_G}{P_T}$$

so we get
$$\left[v = \frac{k_s h_G}{k_s + h_G} \frac{C_T}{N} \Upsilon \right]$$
(which is much simpler than the quadratic law of the Deal-Grove model of silicon oxidation)



Arrhenius plot (log scale / T^{-1})

because the T dependence of k_s and h_G are very different

this is a linear process, double the time? → double the thickness

that's because we don't have any solid state diffusion to go through like in silicon oxidation

As for Si oxidation, we can observe 2 distinct growth regimes:

① $k_s \ll h_G$ surface reaction controlled case

② $k_s \gg h_G$ mass transfer (or gas phase diffusion) controlled case

NOTE: they are both linear with time

NOTE: k_s has an Arrhenius dependence on T ($k_s = k_0 e^{-\frac{E_a}{kT}}$)

that's because is a chemical reaction, for higher T the reaction goes faster!

while h_G shows little dependency on T

← on a log scale, almost constant!

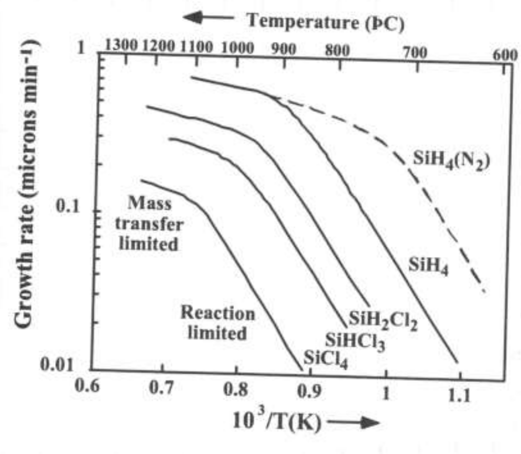
NOTE: the slower process dominates (determines the growth speed)

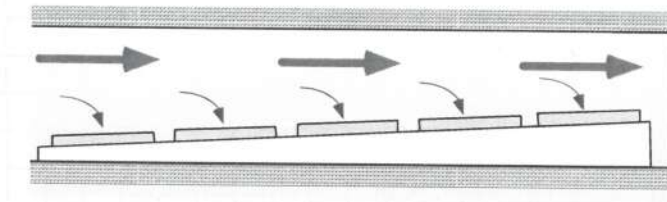
So for lower T: $v \propto K_S$

for higher T: $v \propto h_G$

But we have that δ (so also h_G) changes a lot along the chamber, heavily affecting my deposition speed. That's why we have the wafers sitting on an inclined susceptor, to balance the changes in h_G and gen an uniform layer (picture right)

If we instead go to the reaction controlled regime ($v \propto k_S$) we don't care anymore about the geometry of our chamber, but this happens at low T → low T usually means low thin film quality.





KEY POINTS:

- K_S limited deposition is VERY temperature sensitive
- h_G limited deposition is VERY geometry sensitive (stagnant / boundary layer)
- δ_S is not constant along the chamber length!

LPCVD (Low Pressure CVD)

A way to solve both problems is to act on the pressure, because:

$$\left[h_G = \frac{D_G}{\delta_S} \right]$$

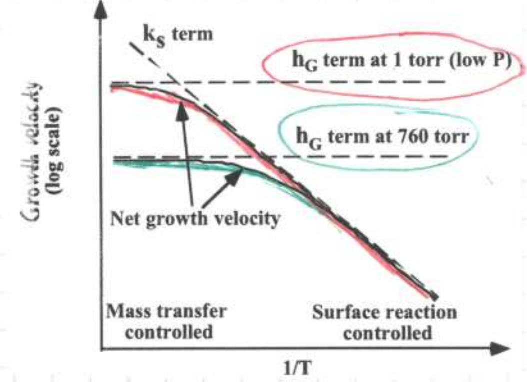
D_G – reactant gas diffusivity

δ_S – stagnant layer thickness

and it turns out that $D_G = \sqrt{T^3} \frac{P_G}{P_{tot}} \propto \frac{1}{P_{tot}}$

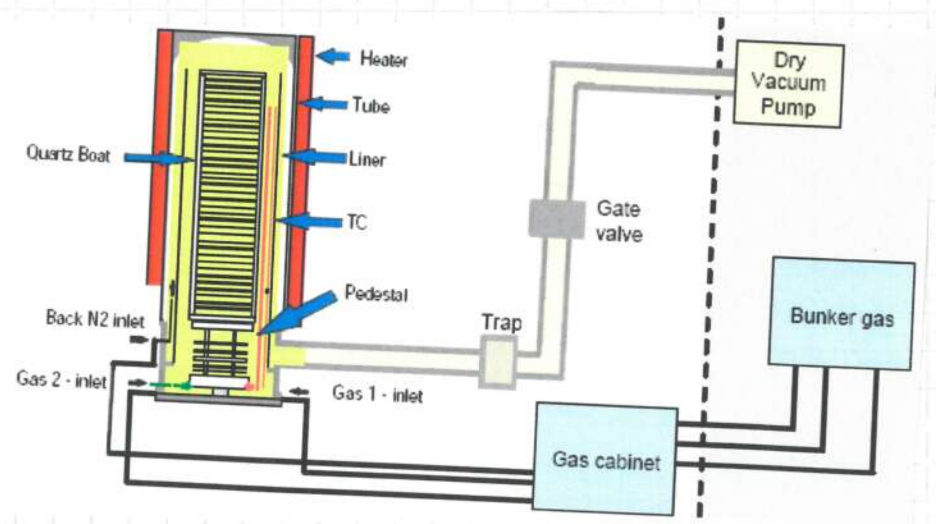
Diffusivity $\propto \frac{1}{P}$

so in this way we can move to surface reaction controlled regime at higher T, and that's why LPCVD is now vastly diffused in IC manufacturing



Advantages of LPCVD:

- wafers can be stacked
- less autodoping
- fewer gas phase reactions



PECVD (Plasma Enhanced CVD) ⟹ we don't want high T, plasma solves this

What if we're in a step of our process flow in which the T at which we would have to work at isn't compatible with the layers already present on our wafer?

e.g. metal layers that can't withstand high T
dopants that I don't want to diffuse further

Can we still get high quality films without going crazy with the geometry?
Yes, we just need another form of energy to promote the chemical reaction without using high T, for example a plasma.

What is a plasma? Ionized / partially ionized neutral gas (on average) ← avg neutral but locally Ionized

An easy way to get a plasma is heating the gas a lot ($T = 4000° \sim 20000°K$) but this isn't what we're interested in.

$T_{electrons} = T_{ions}$
thermal equilibrium

↑ that plasma (thermally generated) would also be completely ionized

We are interested in weakly ionized plasmas, which are electrically driven (ionization achieved throng a discharge across the plasma, that happens because since there are always some ionized atoms and free electrons in a gas at $T \neq 0°K$, if we use an electric field to accelerate those free charges they can scatter other atoms and ionize them, ..., thus ionizing more of the gas, while also some ions re-neutralize, giving us a partially / weakly ionized plasma), where the collisions between charged particles with neutral molecules are actually important (e.g. in the sun Temperature does everything), where the surface loss at boundaries is important (some electrons will be lost at the surface of the plasma chamber → compensated by a current of ions lost at the plasma chamber → this gives to our plasma their **SHEATH** REGIONS), and where electrons are NOT in thermal equilibrium with ions (for the same electric field applied, electrons have >> mobility compared to ions).

typical numbers

Pressure ≈ 1 mTorr − 1 Torr

$T_{electrons} \approx 1 - 10$ eV (electronic temperature) measured in Energy: $1eV \sim 10^4 K$

$T_{electrons} \gg T_{ions}$

$n \approx 10^8 - 10^{13} \frac{\text{ionic species}}{cm^3}$

REMEMBER

The purpose of using PECVD instead of LPCVD is because we might not want to work at higher T. Since the temperature inside the plasma chamber is around room T (can be controlled and increased to promote even more chemical reactions, up to me).

Lecture 15 21 aprile

SHEATHS / DARK REGIONS

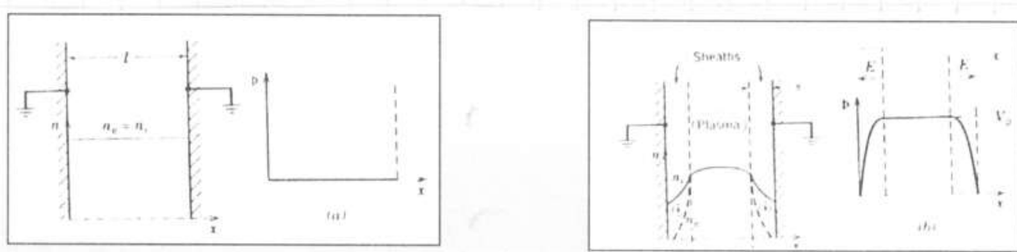
When we talk about plasmas it's very important to talk about the charge losses at the chamber walls.

We start with a neutral (on average plasma), but remember that the speed of the electron is much higher than that of the ions

$$v^{e^-}_{thermal} = \sqrt{\frac{e\,T_e}{m_e}} \gg v^{ions}_{thermal} = \sqrt{\frac{e\,T_i}{m_i}}$$

as soon as we ignite the plasma, the very fast electrons will be absorbed by the chamber walls, now an electric field is created (which compensates the loss of electrons) nearby the chamber walls → the center (bulk) of the plasma stays neutral while the regions near the walls have an electric field (sheaths)

e^- lost → E field → ion current → sheaths



This plasma dichotomy (bulk neutral, sheaths charged) is present even if we apply from the outside a bias potential to the walls, while the bulk always stays relatively at constant potential.

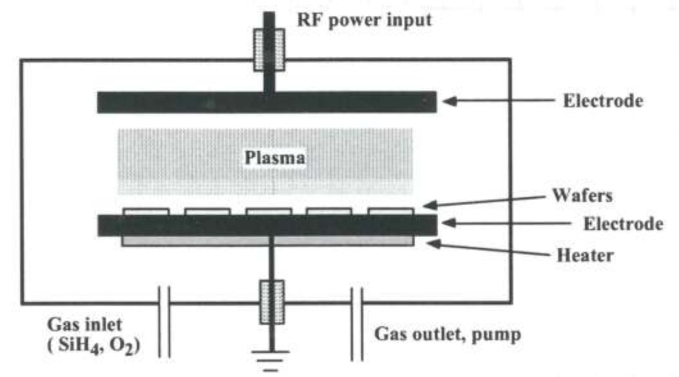
NOTE: plasma ON → fast electrons absorbed by walls → net electronic current builds up → ionic current balances = sheaths formed.

NOTE: sheaths are also called "dark regions" since there are far fewer ion-electron collisions (or atom-electron) from which light might come out.

Inside the chamber, we can have (other than neutral atoms):

- ions, sustaining the plasma (alongside electrons)
- excited atoms, providing the characteristic light of plasmas
- radicals (="incomplete molecules"), very very chemically reactive

atoms with one unpaired e^-, but not charged (like H)



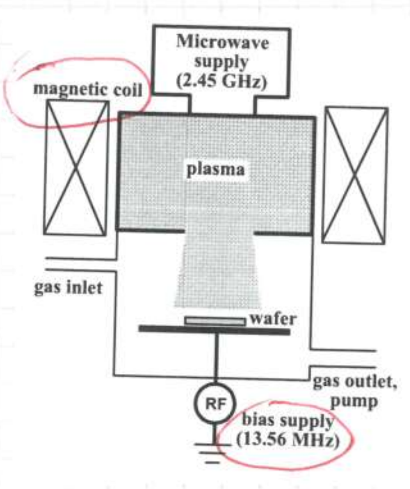
NOTE: modern plasma chambers only have 1 wafer inside, since they can better control the bigger wafers

We might also find neutral and ionized fragments of broken-up molecules, but what we're most interested in are the free radicals, which are electrically neutral but since they have incomplete bonding they're extremely reactive (e.g. $SiO_2 \rightarrow SiO + O$).
free radicals

This high reactivity is what allows us to work at much lower T [maintaining high film quality (as opposed to LPCVD)] combined with the presence of fragments of molecules and the ion bombardment of the surface, which is inevitable since we have our wafer immersed in a plasma

⇒ plasma, ions e^- and radicals ⇒ radicals are very reactive
there are also ions that help, ion bombardment

HDPCVD (High Density Plasma CVD)

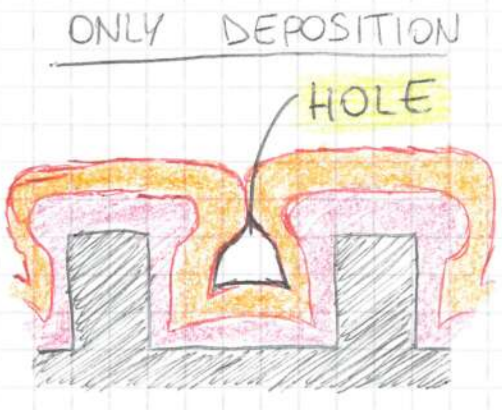


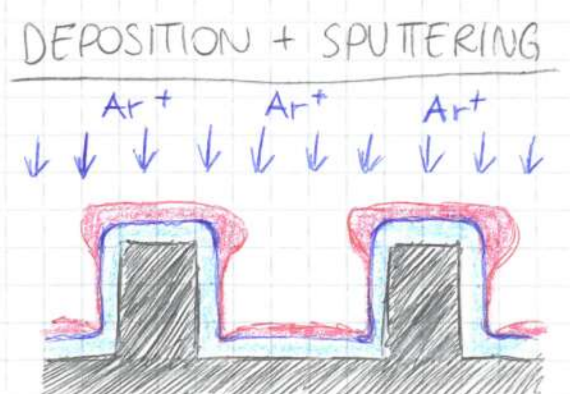
HDPCVD is a kind of PECVD where we have significantly higher ionic density in our plasma thanks to the presence of an added magnetic field which will make charged particles spiral, increasing the chances of ionization.

We can also have an additional electric field on the wafer so that we can drive the ion bombardment on the wafer.

Usually in our plasma chamber we will have some gases containing the deposited film you want to have (e.g. for $SiO_2 \rightarrow SiH_{4(g)} + O_{2(g)}$) but also some inert gases that will only be used for ion bombardment, like Argon.

What's the purpose of having deposition and sputtering / etching at the same time? To fill deep trenches (conformal growth), like STI.





So when we have only deposition, we will have an accumulation on the horizontal surfaces and top corners, because they will "see" more reactants that will then stick there, while all other places will have a more limited solid angle of interaction with reactants.

MANY

FEW

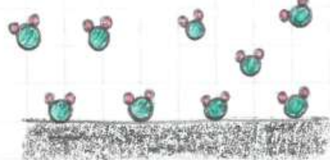
If we directionally bombard with ions the surface (vertically), we remove exactly the portion that accumulates the most, and with some tuning we can reach a conformal film growth.

This technique works up to aspect ratios ~ 5:1
beyond that we'll have to use other techniques to fill the trenches (like spin coating).

ALD (Atomic Layer Deposition) for very high aspect ratio

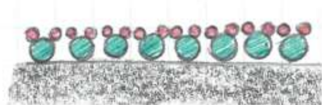
Composed of a sequence of self-limiting surface reactions, can be assisted by temperature, radicals or plasma. It has an atomic scale thickness control (high step coverage) and is typically used for DRAM capacitor high K oxides (aspect ratios of up to 100 : 1)

① FIRST PRECURSOR PULSE



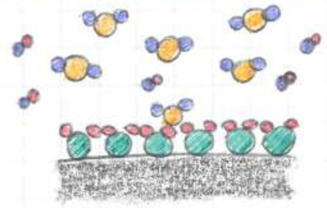
instead of having (for example) both SiH_4 and O_2 in the chamber and having them react to form $SiO_2 + H_2$, we insert only one precursor, for example SiH_4

② INERT GAS PURGE



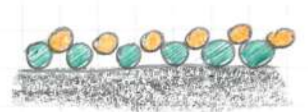
now we insert an inert gas in the chamber and then pump everything away, except the atoms of the 1st precursor gas that have been adsorbed (ADSORBED) by the surface

③ SECOND PRECURSOR PULSE



we insert the second precursor gas which will react with the first precursor molecules that are attached to the surface, making an extremely thin and conformal film, also is self-limiting because only the molecules attached to the surface can react

④ INERT GAS PURGE



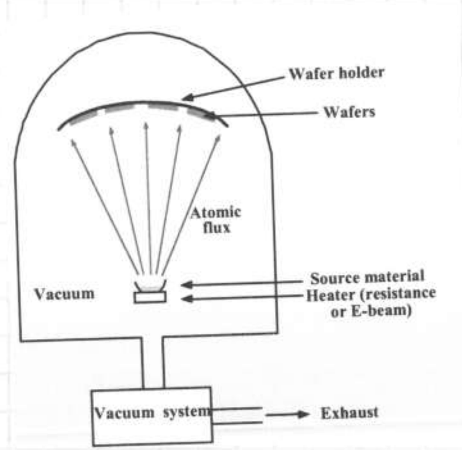
repeating step ② will now leave only our thin film on top of the substrate, ready to repeat the steps (more pulses → more thickness). We can deposit atomic layer by atomic layer.

NOTE: we can alternate different layers and have different compositions / dopings or even nano-laminates film structures

PVD techniques

In Physical Vapor Deposition you take your solid state piece of the material you want to deposit on your wafer, and then have it for example evaporating and then having it condensate on the wafer or sputtering it with ions using a plasma so the material recondensates on the wafer

EVAPORATION



The wafers are upside down stuck to the wafer holder inside a vacuum chamber. The source material is heated and it sublimates into gas form. The sublimated atoms run straight (no collisions) since they're in vacuum and land on the wafers, condensating.

NOTE: this is a room temperature process, despite having the crucible (source material holder, usually ceramic) heated up, the rest of the chamber isn't, so it remains at room T (also because there's vacuum)

NOTE: the deposition rate will depend on both source and reactor geometry

NOTE: evaporation is not commonly used in IC manufacturing today, but many concepts are useful to understand sputter-PVD.

DEPOSITION RATE : POINT SOURCE

We want to build a simple model of the deposition rate, and the first approximation we can do is think of the heated material evaporating as a point source.

Wafer holder
A_k
θ_k
ℓ
h
r
$A_k{}^p$
$\theta_i = \theta_k$
Point source

This point source has equal evaporation rate in all directions.

What is the emitted flux out of our point source?

flux = molecules evaporating / unit surface · unit time

flux

$$F_k^p = \frac{R_{evap}}{\Omega_0 r^2}$$

(flux; projection; angle θ_K; distance)

R_{evap} = evaporation rate

Ω_0 = solid angle of emission (4π all directions, 2π up only)

but if we do this we're calculating F_K, not F_K^p, so we're considering the flux if we had $\theta = 0$, which we don't have

F_K
F_K^p

ideal: isotropic

we have $\theta = \theta_K$, and to account for that $F_k = \frac{R_{evap}}{2\pi r^2} \cos\theta_K$

(we have considered $\Omega_0 = 2\pi$, only upward)

and if we have a source material with density N the deposition rate is:

$$v = \frac{F_K}{N} = \frac{R_{evap}}{2\pi N r^2} \cos\theta_K$$

NON UNIFORM DEPOSITION

$v \propto \cos\theta_K$

DEPOSITION RATE: SMALL PLANAR SOURCE

We can have a slightly more realistic model by considering a small planar surface source, taking into consideration that an extended object evaporating will usually emit the most atoms perpendicularly with respect to the surface (so at $\theta = 0$) and very little emission (ideally none) at $\theta = 90°$.

Wafer holder, A_k, θ_k, ℓ, A_k^p, h, r, θ_i, A_i, A_i^p, Small planar surface source

This can be done by adding a $2\cos\theta_i$ (the 2 is so we can have a normalized evaporation rate)

$$v = \frac{R_{evap}}{\pi N r^2} \cos\theta_k \cos\theta_i$$

Different (more realistic) angular distributions can be modelled as $\cos^n\theta_i$ functions, with $n > 1$

Some examples

uniform (isotropic) emission from a point source

ideal $\cos\theta$ emission from a small planar surface source ($n=1$)

non-ideal, more anisotropic emission from a small planar surface source ($n > 1$)

NOTE: no matter what, we can't get a uniform deposition on our surface through evaporation.

Usually many wafers sit in a spherical holder (called planetary) which rotates, to maximize deposition uniformity

Wafers

Location of source which behaves like an ideal point source

Location of source which behaves like an ideal small area surface source

EVAPORATION RATE (Langmuir - Knudsen theory)

The deposition (or evaporation) rate can be written as

$$R_{evap} = 5{,}83 \cdot 10^{-2} A_s \sqrt{\frac{m}{T}} P_e \quad \left[\frac{\text{grams}}{\text{second}}\right]$$

where: A_s = area of the source

m = molecular mass of the to-be evaporated material [grams]

T = temperature in K

P_e = vapour pressure (a partial pressure of 1-10 mTorr is required to obtain reasonable deposition rates)

NOTE: despite having explicitly a $T^{-1/2}$ dependence, there's an implicit Arrhenius (= exponential in T) dependence in P_e, so at higher T we increase the deposition rate.

EVAPORATION: PROs AND CONs

PROs

- almost every material can be evaporated
- no contamination or damage of the wafer (room T, low P, chill)

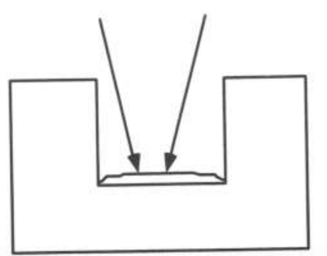
CONs

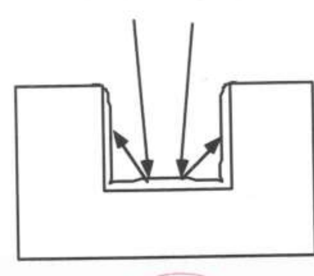
- some materials are very hard to evaporate (low P_e, like W)
- can't do in-situ cleaning of wafer before deposition
- deposition of alloys is difficult (different P_e, different T, no uniform)
- poor step coverage (due to no resputtering, no redistribution, only line of sight deposition, sticking coefficient ≈ 1)

LINE OF SIGHT ONLY →

STICKING COEFFICIENT

higher T, lower S_c





no conformality — more conformal

The sticking coefficient represents the probability of a molecule that lands on a given point of the surface to stay there and stick to the surface

$S_c \approx 1$ → no re-emission, where it first sticks is where it stays

$S_c < 1$ → there's a chance that a molecule that landed on the surface is re-emitted and lands somewhere else on the wafer

$$S_c = \frac{F_{reacted}}{F_{incident}}$$

$F_{reacted}$: flux of reacted molecules

$F_{incident}$: flux of total incident molecules

NOTE: the "reacted" molecules are the ones that permanently stick to the surface they land, so the ones that react with it.

NOTE: $S_c < 1$ gives more conformal films, and higher T lower S_c, so for evaporation $S_c \approx 1$ because the wafer is at room T.

Lecture 16

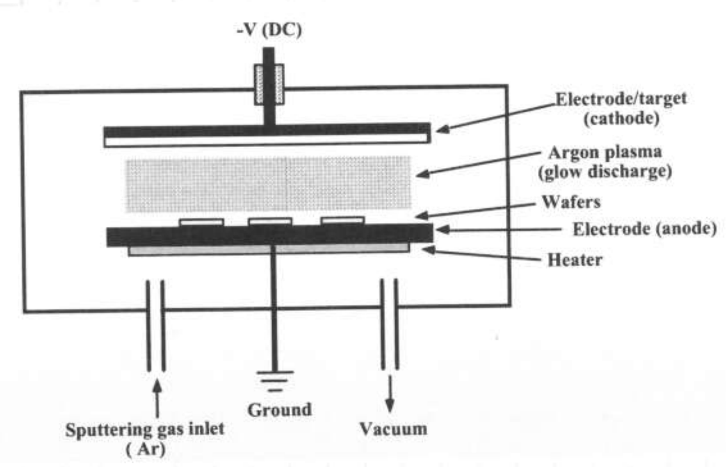
23 aprile

SPUTTER DEPOSITION

In sputter deposition the material that you want to deposit is in solid state inside the deposition chamber, and is usually called "target". Now we bombard the target with plasma ions (usually Ar) and dislodge the target's atoms, which are then deposited on the wafer.

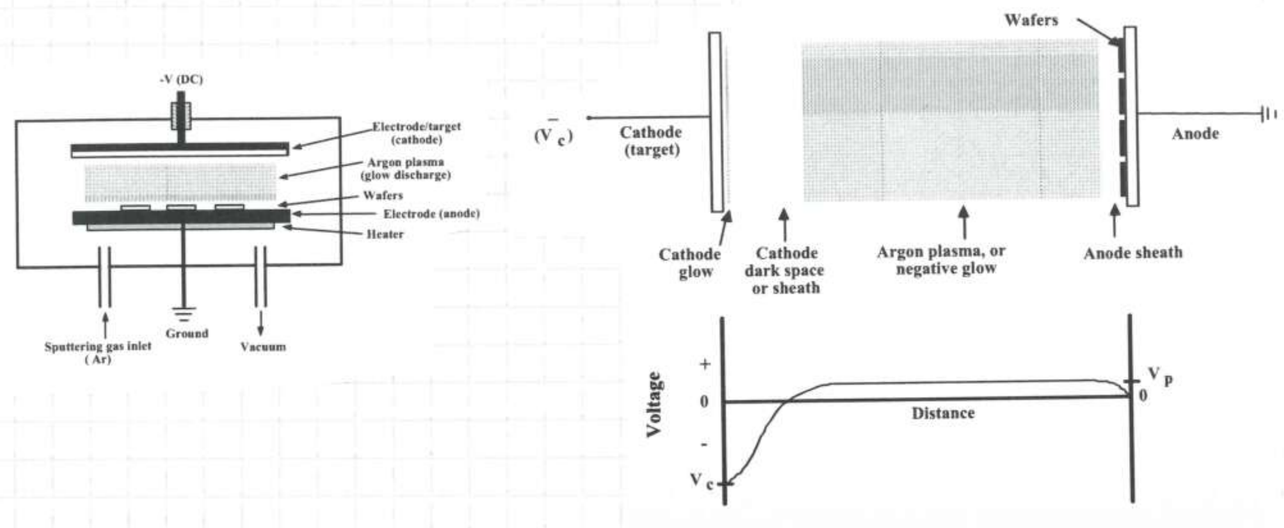
Higher pressure compared to evaporation, but still pretty low (1 - 100 mTorr)

Allows can be easily deposited



evaporation → no alloy
sputter → yes alloy

DC SPUTTER DEPOSITION for conductive target



In DC sputter deposition the wafers are sitting on the anode and the cathode has the target attached to it. The target is kept at a negative DC potential with respect to the wafers. We also have a plasma between them.

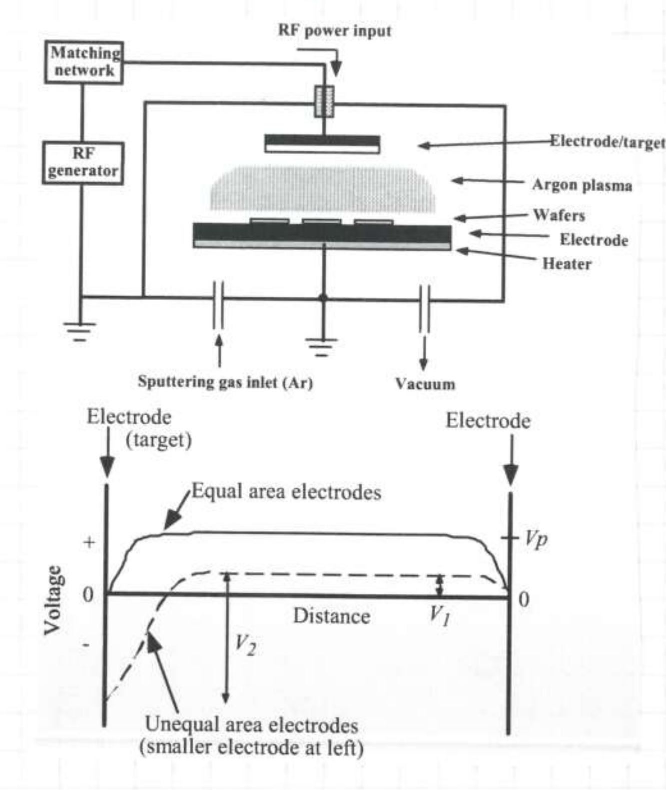
We know that when a plasma is present (not thermally generated) sheaths form (only place where theres average $\vec{E} \neq 0$).

NOTE: while the electric field in the bulk of the plasma is on average null (no voltage gradient), the bulk has a potential which is more positive than the sidewalls / sheath. This happens for every surface within the plasma: both ions and e^- will move and get into surfaces, but e^- much more often, creating always positive sheaths around the surface.

Positive ions (Ar^+) accelerate towards the cathode and dislodge target atoms which then (some of them) deposit on the wafer surface.
But this works only if the target is conductive (current DC must flow through it). What if it's not? → RF sputter deposition

RF (or AC) SPUTTER DEPOSITION

for not conductive target



What if, for example, we want to deposit a dielectric? Or any other non-conductive material?
We use an AC current.

As for the DC case, pretty quickly we reach a steady state situation (NOT equilibrium since $T_{e^-} \gg T_{ions}$) in which the bulk is at constant potential and positive sheaths are formed.

But since the electrons' response to the RF signal is almost istantaneous, the sputtering is continuous at both ends, so we're also sputtering the wafer! BAD (usually)

We can solve this by using electrodes with different areas, since the voltage drop between the bulk and the electrode goes as

$$\frac{V_1}{V_2} = \left(\frac{A_2}{A_1}\right)^m$$

$m = 4$ from theory
$m = 1 \sim 2$ experimentally

So by increasing the area of the electrode on which the wafers are sitting we can reach the same result of the DC sputter deposition: no sputtering of the wafer, only of the target. This is usually done by connecting the "wafer electrode" to the chamber walls, while the "target electrode" No.

More in detail...

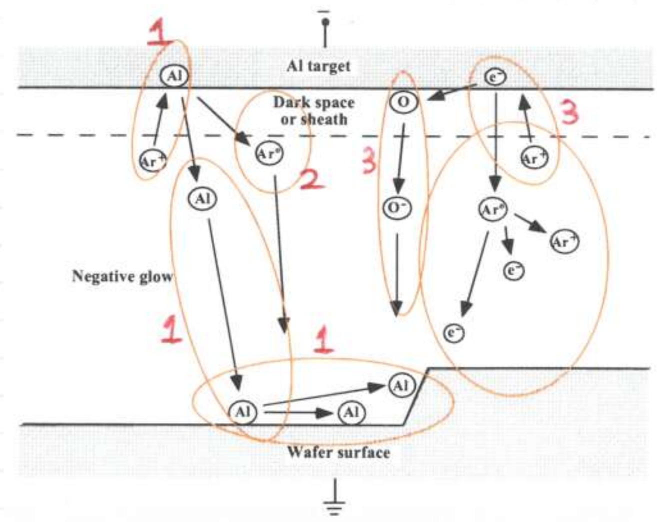
DC sputter deposition is not suitable for insulator deposition: with a negative DC voltage applied, positive Ar^+ ions from the plasma hit the negatively charged insulator and positive charge would accumulate. The negative surface voltage would become less than that required to sustain the glow discharge and the plasma would shut down. The time required for this to occur is $1 \sim 10\ \mu s$.

If we apply a high-frequency alternating voltage (RF, ~ 14 MHz), the positive charge buildup is neutralized by e^- bombardment over each cycle. The RF frequency is chosen to be high enough so that a continuous plasma discharge is maintained. The difference in mass (and so also mobility) between electrons and ions allows for continuous sputtering of the target throughout both half cycles of the RF voltage. The electrons have a high enough mobility to keep up with the changing electric field, but the heavier Ar^+ ions do not. During the first few complete cycles more e^- than Ar^+ are collected at each electrode, and this negative charge buildup is maintained during subsequent cycles.

NOTE: since we can easily choose which electrode will suffer the sputtering (just by connecting the one we want to the sidewalls) we can do an in-situ pre-cleaning of the wafer, before the deposition. For example if we want to clean any native oxides which may form during "air breaks" (= when the wafer is exposed to air during in-between steps). In the DC sputter deposition I just invert the voltages.

DC SPUTTER DEPOSITION

① Ar^+ ions accelerated in the sheath sputter target atoms (e.g. Al)



The Al atoms will migrate to the wafer surface where it can:

- stay adsorbed
- migrate to another surface site (by following surface potential)
- be re-emitted

② The Ar^+ ions can be neutralized during sputtering and can be incorporated in the growing film (we will always find traces of Argon in our Aluminum, while this is not the case for evaporation, which will be much more pure) sputtering less pure!

③ The Argon ions can emit secondary electrons (rip off electrons from the target) during sputtering, and they can:

- ionize an impurity (e.g. Oxygen) that can travel to the wafer
- ionize Argon and help sustaining the plasma

SPUTTERING YIELD

Sputtering yield $Y = \dfrac{\#\ \text{sputtered atoms}}{\#\ \text{incident ions}}$ USUALLY $(0{,}1 < Y < 3)$

The sputtering yield tells me how many target atoms every single incident ion dislodges, so how "effective" our sputtering is.
Y depends on the material being sputtered, on the ion energy (higher energy → higher Y) and on the ion arrival angle.

Ion arrival angle:

at 90° our ion can knock-on the target atoms, that is: the target atoms can recoil inside the material and not be expelled outside

TARGET

at a different angle, we can dislodge target atoms outside

NOTE: alloys can be easily deposited even if the source materials have very different Y from one another.

Since sputtering happens at the surface, if we have an alloy we will start sputtering first the atoms with higher sputtering yields, so we remove them first from the target surface, leaving now only low Y atoms, which don't have a choice and will also get sputtered. So is like a "dynamical equilibrium" where the elements with different Y will leave the surface at the same ratio. So the composition of the alloy will end up on the wafer the same as in the target.

This is why, when depositing new material, usually PVD sputtering is the method of choice, because is easy.

ARRIVAL ANGLE DISTRIBUTION

If I can describe, for each point of the wafer, how many particles are deposited per unit area / time (so the flux), then I can know the deposition rate.
But the uniformity of my film (= conformality) greatly depends on the arrival angle distribution, that is: the normal component of the incoming flux relative to the wafer → $\cos\Theta$.
For $\cos\Theta$ we get an isotropic flux, while for $\cos^n\Theta$, $n>1$ we get an anisotropic flux.

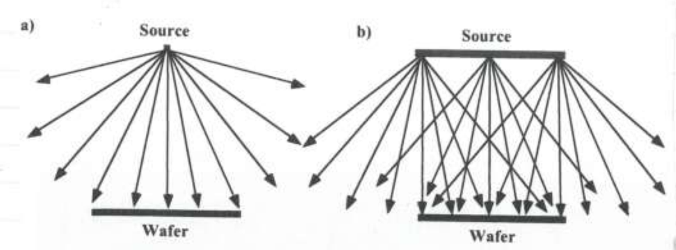
narrow distribution, coming mostly from **UP** so will cover primarily horizontal surfaces

a) Isotropic flux arrival
$n = 1$ in $\cos^n\theta$ arrival angle distribution

b) Anisotropic flux arrival
$n > 1$ in $\cos^n\theta$ arrival angle distribution

F^o θ
Small area at position i
Surface of wafer

How can we practically get a more uniform deposition?
If we use a point source we will get a narrow distribution of arrival angle for every point on the wafer, while a large target will act like many point sources, giving rise to a very uniform deposition since basically all the corners will be covered by some angle somehow.



OTHER SPUTTERING PROCESSES...

- **Reactive** sputter deposition

  A reactive gas is added to the plasma to help the layer deposition.
  e.g. TiN is deposited using a Ti target and an Ar / N_2 plasma

- **Bias sputtering**

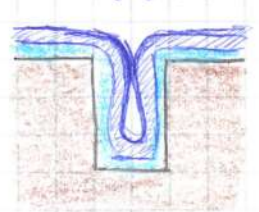
  The wafer can be opportunely biased so that we can perform sputter etch (for cleaning the surface), or we can do deposition and sputtering at the same time

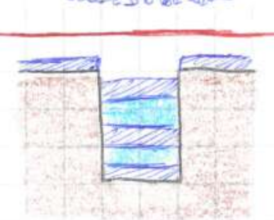
- **Magnetron sputtering**

  Magnets are used to increase the fraction of the plasma which is ionized (electrons will move in spirals → more collisions → more plasma)

- **Collimated sputtering**

  If we care more about gap filling than conformality we can collimate the arrival angle (→ VERY NARROW) by placing a plate full of circular / hexagonal holes between the wafer and the target.

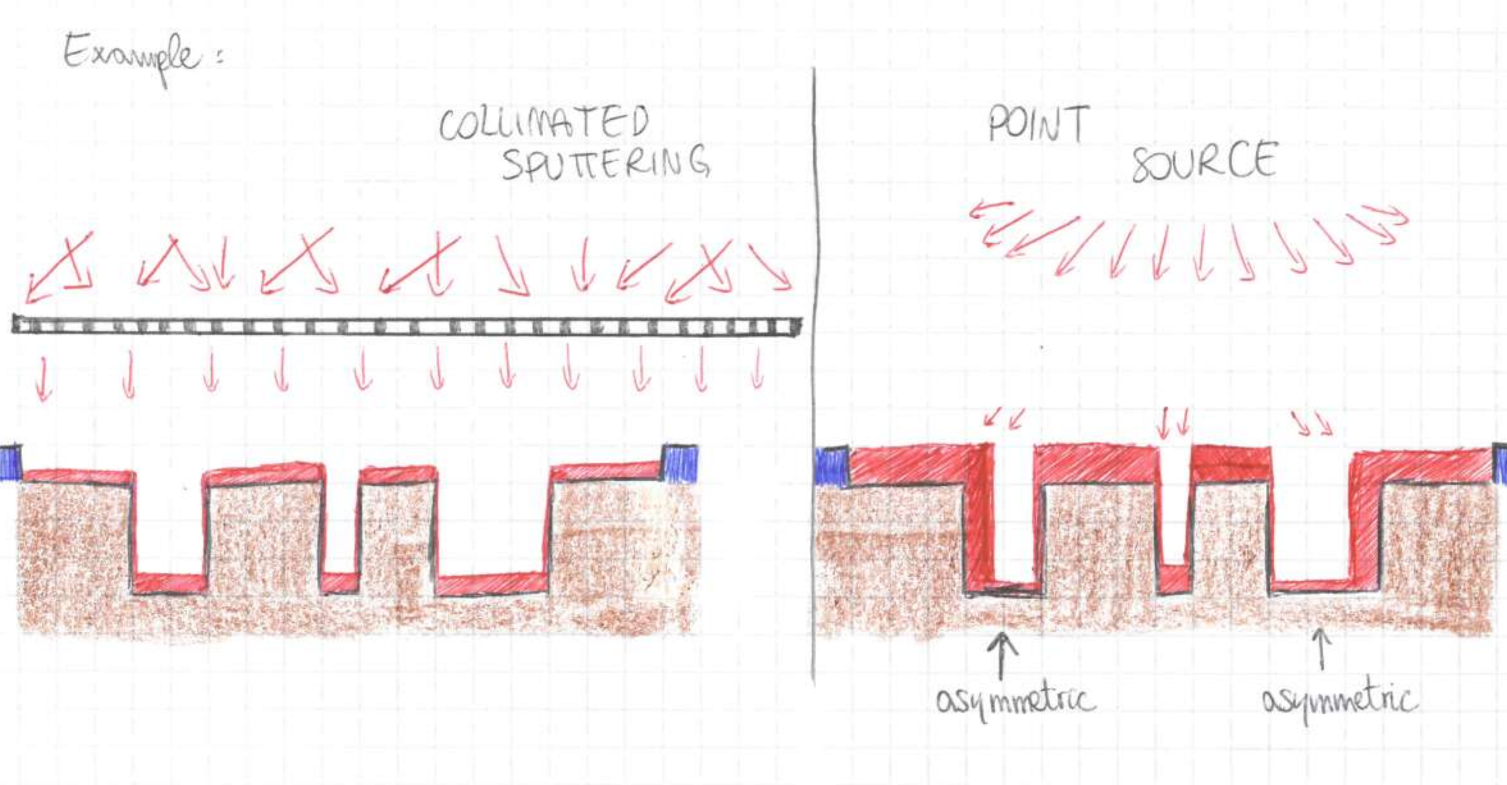




NOTE: collimated sputtering ≠ point source

While it's true that both have a narrow arrival angle, they are very different since collimated sputtering has atoms coming only from the top while the point source has atoms coming from all sides, but a narrow angle.

Example:



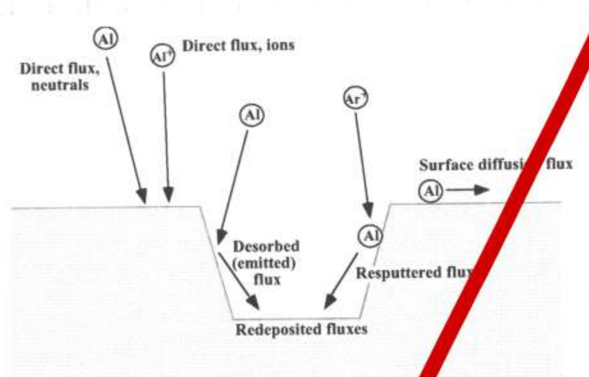
Why are the films deposited using a point source so asymmetric on the sides? Because of the narrow arrival angle!

So on the right side of the wafer: atoms are coming only from the left, none are coming directly from the top

In the end, both the arrival angle and the sticking coefficient will define the film conformality.

MODEL OF CVD AND PVD

To accurately model our deposition and predict the growth rate of the film, we have to be able to describe - for every point of the wafer - every reaction and flux that might contribute.



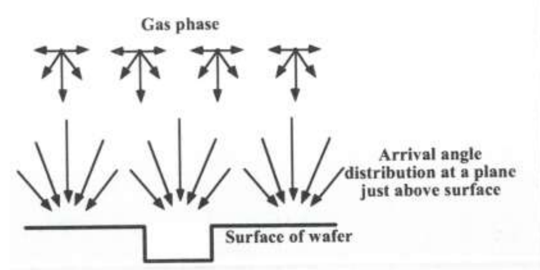
The total flux F^i_{net} $\forall$ point i $\in$ wafer will be comprised of:

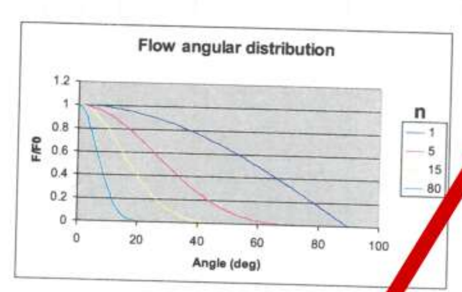
- → direct flux of neutral atoms $(+F^i_{direct(neutral)})$
- → direct flux of ions $(+F^i_{direct(ions)})$
- → but some of this molecules might not stick there and leave $(-F^i_{emitted})$, or can travel along the surface and be diffused $(-F^i_{diff.out})$
- → the incoming flux of ions/atoms might sputter some atoms of my wafer, so I'd get $-F^i_{sputtered}$ (they leave → negative flux)
- → some molecules that landed somewhere else on the wafer might not stick there and jump to the point i, giving us $+F^i_{redeposited}$, and also some neighbouring molecules will diffuse in $(+F^i_{diff.in})$.
  This is the opposite of $-F^i_{emitted}$ and $-F^i_{diff.out}$.

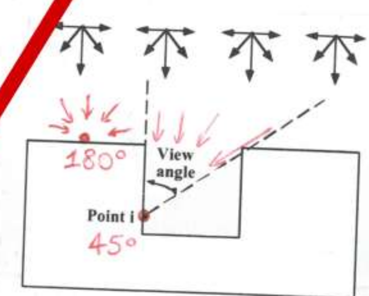
$$F^i_{net} = F^i_{direct(neutrals)} + F^i_{direct(ions)} + F^i_{redeposited} + F^i_{diff.in} - F^i_{emitted} - F^i_{sputtered} - F^i_{diff.out}$$

Depending on the type of deposition then I can "turn ON/OFF" every single parameter in my computer simulation, but an even simpler way to model the incoming flux is just to assume some distribution as a function of the angle formed with the surface.

→ $F_{direct}(\theta) = F_0 \cos^n \theta$ where $n > 1$ accounts for anisotropy.







Then I have to integrate the arrival angle θ over the view angle, so the more my point i can "see" the incoming particles from all angles, the greater the flux in that point. (picture right)

↑ THIS IS JUST TO MODEL F^i_{direct} ($F^i_{direct\,(neutrals)}$ and $F^i_{direct\,(ions)}$)

We can also model the diffusion along the surface and get the net flux:

$$F^i_{diff.\,net} = F^i_{diff.\,in} - F^i_{diff.\,out} = \frac{D_s}{kT} \gamma_s \Omega \nu \frac{\partial^2 K}{\partial s^2}$$

(k: Boltzmann constant)

D_s = surface diffusivity

γ_s = area surface energy

Ω = atomic (or molecular) volume

ν = atoms (or molecules) per unit area

K = curvature

S = unit length across surface

How to model $+F^{i}_{redep}$ and $-F^{i}_{emitted}$?

We use the sticking coefficient $S_c = \frac{F_{reacted}}{F_{incident}}$

High ($S_c = 1$)

Low ($S_c < 1$)

If I know the incident flux and S_c, the emitted flux is gonna be:

$$F^{i}_{emitted} = (1 - S_c)\, F_{incident}$$

Usually the flux is considered isotropic (perfectly diffuse angular distribution → $\cos\theta$ distribution).

And what about $F^{i}_{redeposited}$?

We can calculate the emission flux from ALL other points k on the surface, except i, taking also into account, for each ik couple of points, the distance and orientation between the points (so how likely is an atom from k to be able to jump to i?), and we put this geometrical considerations into g^{ik}.

flux emitted from point k

$$F^{ik}_{redeposited\,(emitted)} = g^{ik}\, F^{k}_{emitted} = g^{ik}\,(1 - S_c)\, F^{k}_{incident}$$

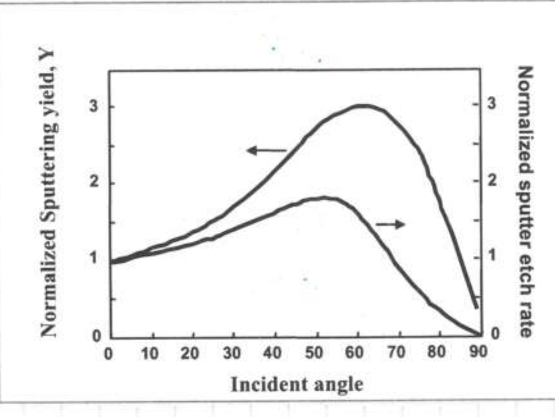
we then run k over all points ($k \neq i$) to get the total flux of atoms redeposited on i.

The last parameter to model is $-F^{i}_{sputtered}$

$$F^{i}_{sputtered} = Y\, F^{i}_{ions}$$

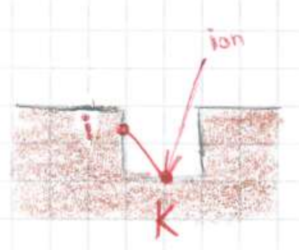
sputtering yield $Y = \frac{\#\ \text{sputtered atoms}}{\#\ \text{incident ions}}$

NOTE: Y has an angular dependence. The emission of sputtered ions is usually modelled as $\cos^n(\theta)$, with $n > 1$ for low energy sputtering and $n < 1$ for energetic ions (high energy sputtering).



Remember that also sputtered atoms can be redeposited!

as for $F^i_{redeposited}$, we can have atoms that move from their original location, but this time is due to sputtering, so due to ions dislodging atom K into i.



$$F^{iK}_{redep\,(sputtered)} = g^{iK} F^K_{sputtered} = g^{iK} Y F^K_{ions}$$

Impinging ions can also enhance the deposition rate by transfering energy to the species on the surface and thus helping the different surface processes and reactions. This can be done by adding a completely new term to the model:

$$F^i_{ion\ induced} = K_i F^i_{ions}$$

ion induced deposition enhancement factor

The competition between ion sputtering and ion-enhanced deposition can be very important in obtaining the final topography.

NOTE: F_{ions} includes all ions, both nonprecursor ions (like Ar^+) and ionized precursor species (like Al^+ or Ti^+).

LPCVD MODELLING

Now we can start "turning ON / OFF" the parameters we've seen in order to get the best possible model of the deposition technique we're using.

Standup wafers
Furnace - with resistance heaters
Direct flux, neutrals
Desorbed (emitted) flux
Redeposited fluxes

$F^i_{direct\,(neutrals)}$	YES	✓
$F^i_{direct\,(ions)}$	NO	✗
$F^i_{diff\,(net)}$	NO	✗
$F^i_{emitted}$	YES	✓
$F^i_{redeposited\,(emitted)}$	YES	✓
$F^i_{sputtered}$	NO	✗
$F^i_{redeposited\,(sputtered)}$	NO	✗
$F^i_{ion\ induced}$	NO	✗

So in the end our net flux will be $F^i_{net} = F_{direct(neutrals)} + F_{redeposited(emitted)} - F^i_{emitted}$

if we remember that $F^i_{emitted} = (1-S_c)\,F^i_{incident} = (1-S_c)\left(F^i_{direct(neutrals)} + F^i_{redepos(emitted)}\right)$

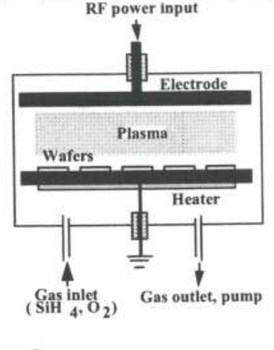
local deposition flux, $F_d = F^i_{direct(neutrals)} + F^i_{redeposited(emitted)}$

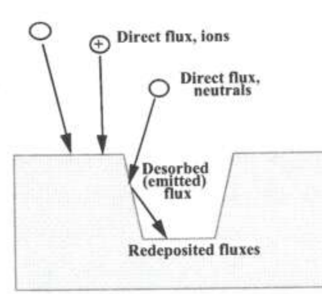
we get $F^i_{net} = S_c F_d$

if we divide by the film density N we get the rate of deposition:

$$rate = \frac{S_c F_d}{N}$$

PECVD MODELLING





$F^i_{direct\ (neutrals)}$	YES	✓
$F^i_{direct\ (ions)}$	NO	✗
$F^i_{diffu\ (net)}$	NO	✗
$F^i_{emitted}$	YES	✓
$F^i_{redeposited\ (emitted)}$	YES	✓
$F^i_{sputtered}$	NO	✗
$F^i_{redep\ (sputtered)}$	NO	✗
$F^i_{ion\ induced}$	YES	✓

similarly, we find a rate of deposition equal to

$$\text{rate} = \frac{K_d S_c F_d + K_i F_i}{N}$$

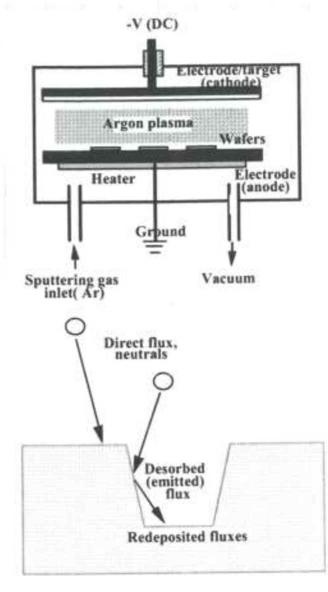
where K_d is a specific rate constant for non-ion enhanced component

K_i rate constant for the ion-enhanced deposition component

F_i is the net local ion flux

S_c is there to take into account the fact that the neutral species might jump around after their initial landing.

PVD MODELLING



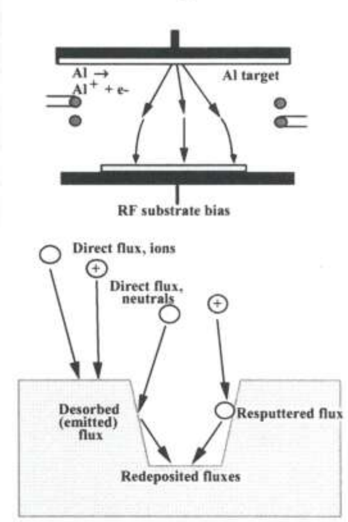
Ions don't play a significant role, so modelling is the same as in LPCVD systems (but with a different flux distribution and S_c):

$$rate = \frac{S_c F_d}{N}$$

but if we use high T, surface diffusion must be taken into account:

$$rate = \frac{S_c F_d + \frac{D_s}{kT} \gamma_s \Omega^2 \nu \frac{\partial^2 K}{\partial s^2}}{N}$$

IONIZED - PVD MODELLING



$F^i_{direct\,(neutrals)}$	YES	✓
$F^i_{direct\,(ions)}$	YES	✓
$F^i_{dir.\,(net)}$	NO	✗
$F^i_{emitted}$	YES	✓
$F^i_{redeposited\,(emitted)}$	YES	✓
$F^i_{sputtered}$	YES	✓
$F^i_{redeposited\,(sputtered)}$	YES	✓
$F^i_{ion\ induced}$	YES	✓

$$rate = \frac{S_c F_d + F_i - K_{sp} Y F_i + K_{rd} F_{rd}}{N}$$

In ionized PVD we ionize the (initially neutral) sputtered atoms as they travel in the plasma, then accelerate them towards the wafer at nearly 90° angle (perpendicular), achieving a very narrow arrival angle distribution (analogous to collimated sputtering).

FINAL REMARKS...

What's important to know / remember are <u>not</u> the rate formulae, but the physics behind all of this, so: which processes are at play? When we consider PECVD what changes? What process isn't present anymore? And so on...

very directional

	n (exponent in cosine arrival angle distribution)	S_C (sticking coefficient)
Sputter deposition		
-standard	~1 - 4	~1
-ionized or collimated	8 - 10	~1
Evaporation	3 - 80	~1
LPCVD silicon dioxide		
- silane	1	0.2 - 0.4
-TEOS	1	0.05 - 0.1
LPCVD tungsten	1	0.01 or less
LPCVD polysilicon	1	0.001 or less

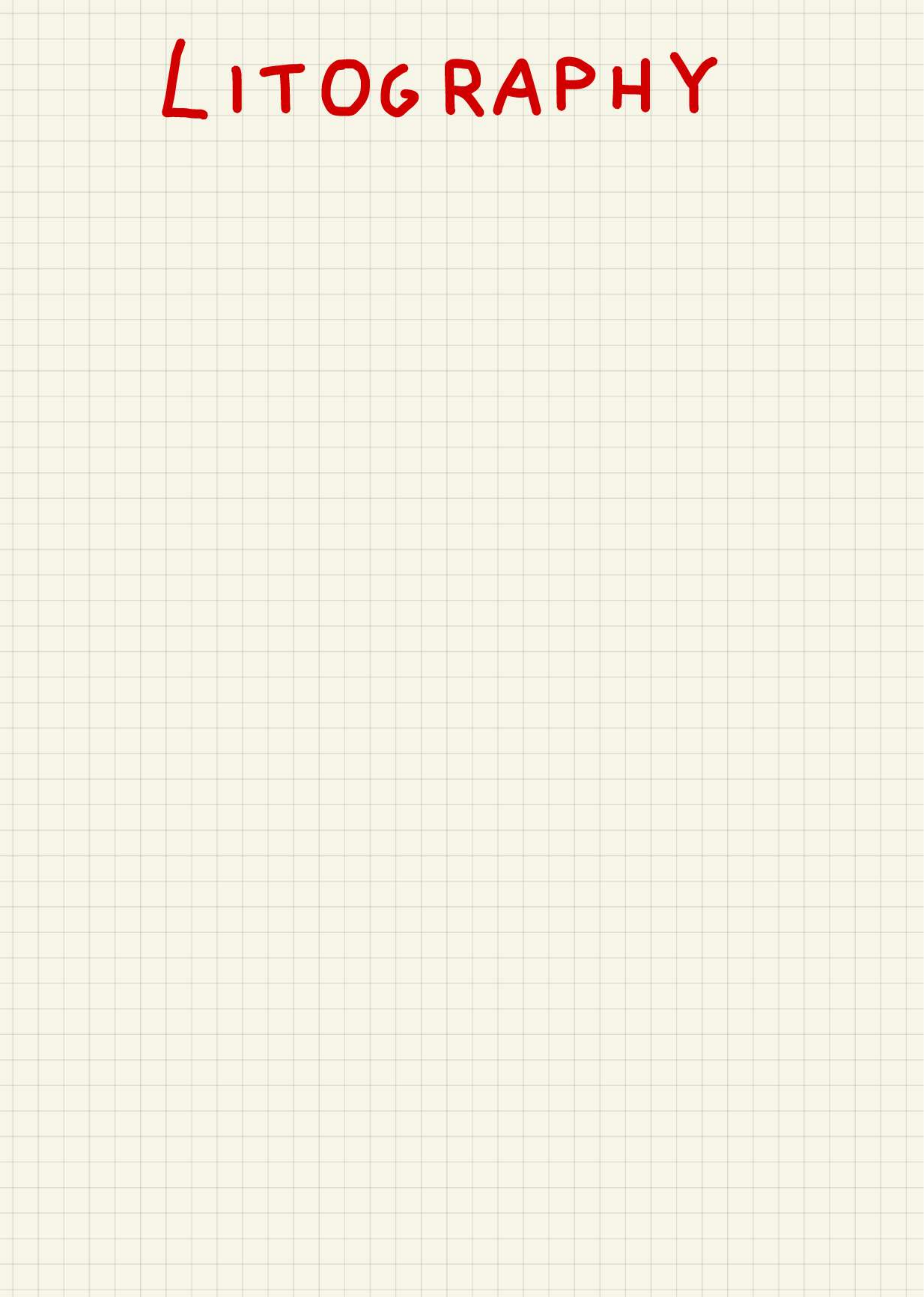
≈ 1 atoms won't move once they landed

high n + $S_C \approx 1$ → high <u>NON</u> - conformality

$n \simeq 1$ + low S_C → high conformality

Usually for CVD techniques $n \simeq 1$, while for PVD $n > 1$. This is because CVD is performed at a pressure considerably higher than PVD, so that the ions/atoms will collide many many times before hitting the wafer, giving them a wide angular distribution. HDPCVD systems are an exception, working at very low gas pressure.

ERROR 404 : page not found

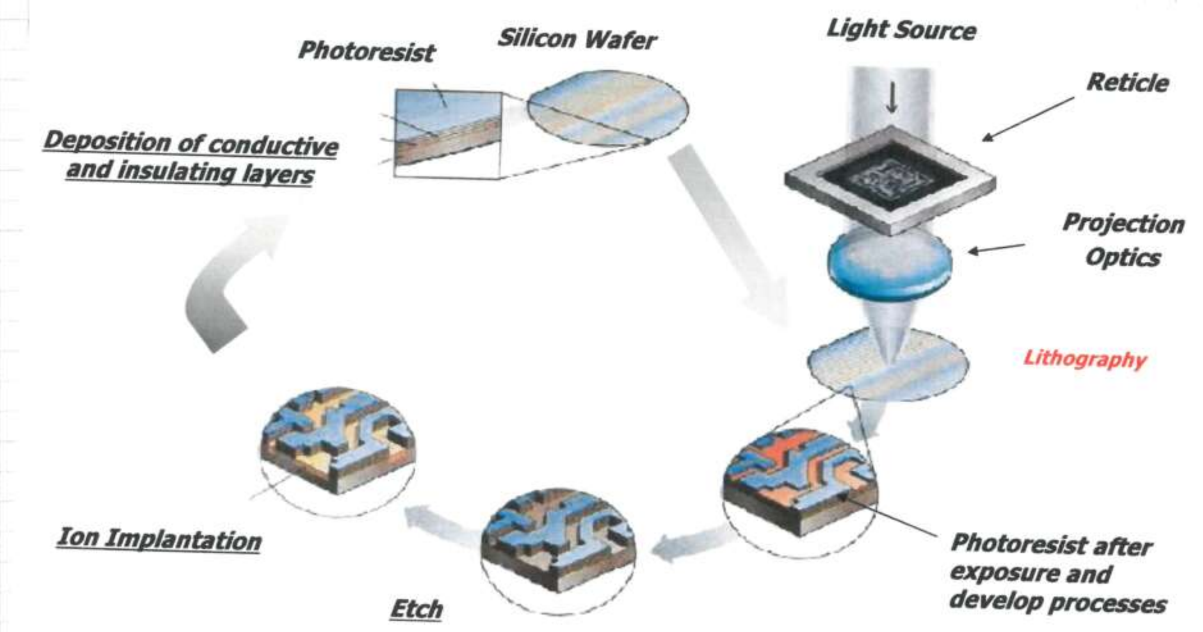

LITOGRAPHY

Lecture 17

28 aprile

LITHOGRAPHY

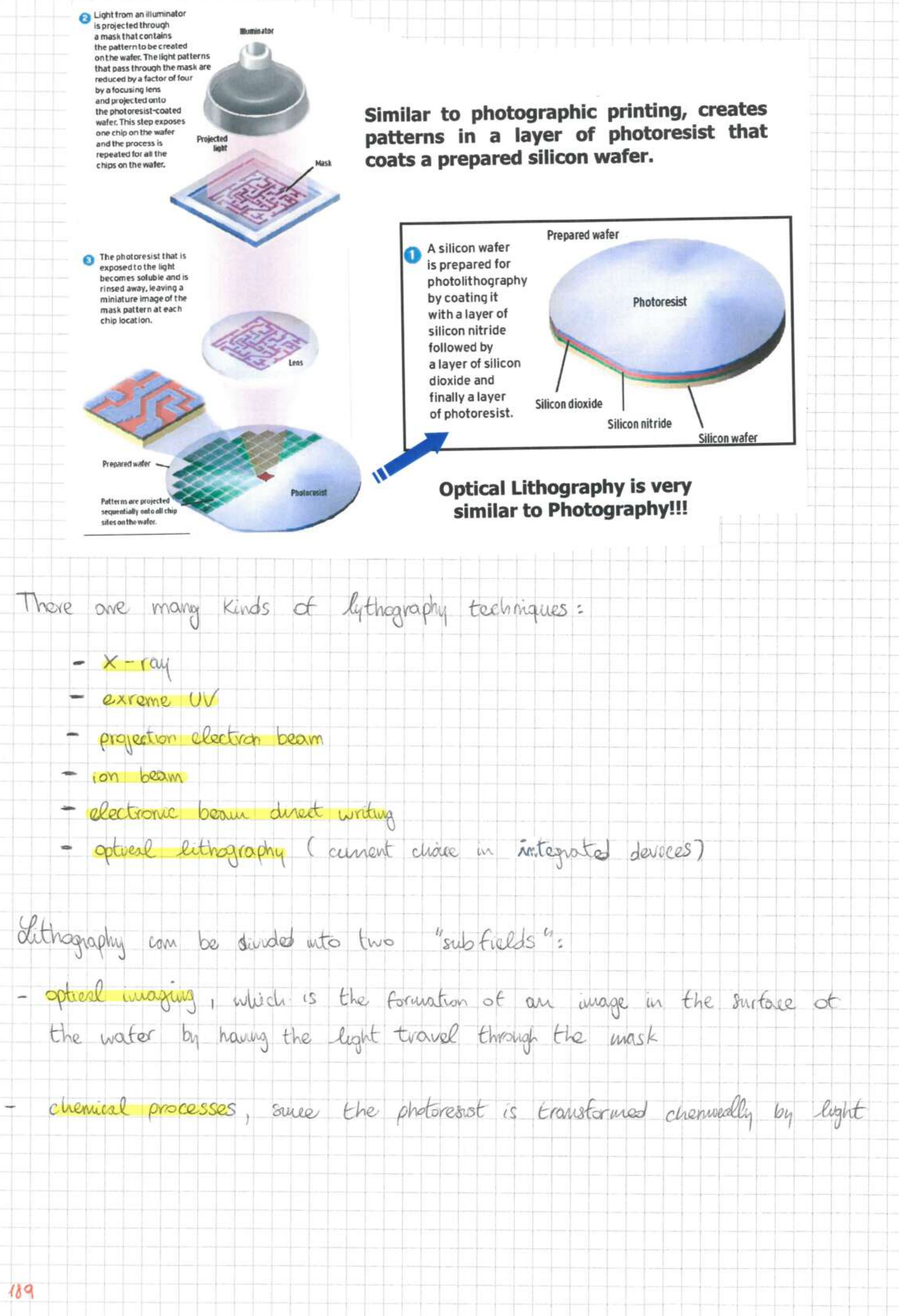
Lithography in the process flow of semiconductor devices is used to create patterns through masks which will later shield some portions of the wafer from ion implantation or etching.



The mask is simply a piece of quartz (glass) portions of which are covered in metal (usually chrome Cr). So if we illuminate the mask with a light source, it will go through only where there's glass.

The light hits (or not) the wafer, which is covered by a photosensitive material called photoresist which' solubility in another material (solvent) is modulated by the exposure to light. So when we expose the photoresist to light it will change its chemical structure and becomes soluble in a solvent (which is called developer).

After developing the photoresist, we have transferred the pattern on top of the wafer (to the photoresist, since we didn't etch the wafer yet.).



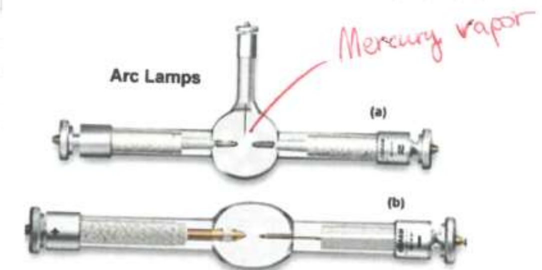
There are many kinds of lythography techniques:

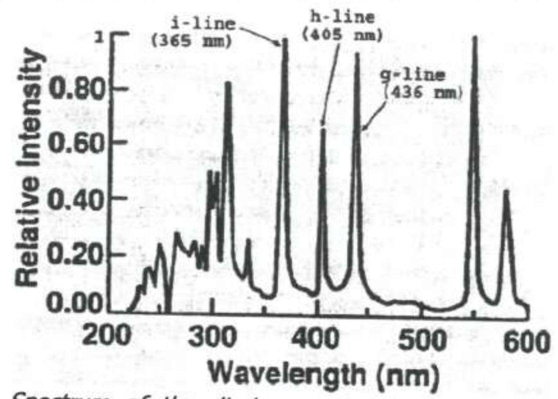
- X-ray
- exreme UV
- projection electron beam
- ion beam
- electronic beam direct writing
- optical lithography (current choice in integrated devices)

Lithography can be divided into two "subfields":

- optical imaging, which is the formation of an image in the surface of the wafer by having the light travel through the mask
- chemical processes, since the photoresist is transformed chemically by light

ILLUMINATION SOURCE





Spectrum of the discharge of a high pressure mercury arc lamp

The higher the resolution that I want to print (= smaller features) the shorter the wavelength of the photons.

Initially lithography was based on discharge lamps (rarified gas, high potential difference between electrodes → arc discharge creates a plasma → photon emission)

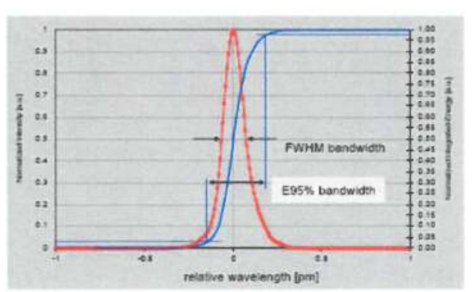
The specific emission lines have a name:

VIOLET	g-line	436 nm	emission lines
VIOLET/UV	h-line	405 nm	
UV	i-line	365 nm	

The need to print ever smaller features pushes us towards shorter wavelengths, but mercury (Hg) lamps don't provide the adequate intensity below 300nm. Lasers demonstrated to be the ideal candidates because of their high power and spectral purity. Current lithographic systems use excimer lasers in the DUV (Deep UV).

λ	Source	Name
436nm	Hg Lamp	G-line
365nm	Hg Lamp	I-line
248nm	Excimer Laser	KrF
193nm	Excimer Laser	ArF
157nm	Excimer Laser	F2

In Blue, current production systems



Laser Bandwidth ~ 0.2pm

"Excimer" = Excited Dimer

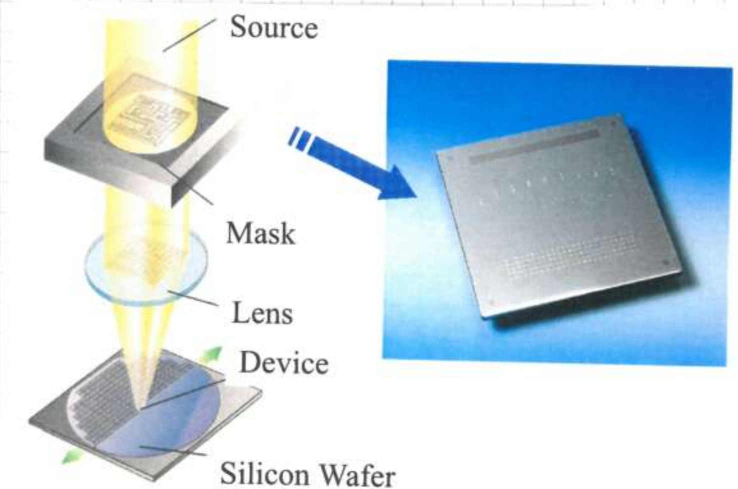
Take 2 elements which do not normally react in their unexcited state.
But when these elements are excited, a chemical reaction becomes possible, forming one compound.
Then when it returns to its ground state, a photon is emitted in the DUV and the molecule breaks up.
The pulsing can be ~ several kHz

NOTE: when changing the λ used, we also have to change the photoresist.

NOTE: 157nm (F^2 excimer laser) never made it to the market. The industry stopped at 193nm while still being able to shrink down the dimensions.

PHOTO MASKS

Photo-masks are high precision plates containing microscopic images of electronic circuits.
Current photo-masks are reduction reticles (5x for I-line systems, 4x for DUV systems).

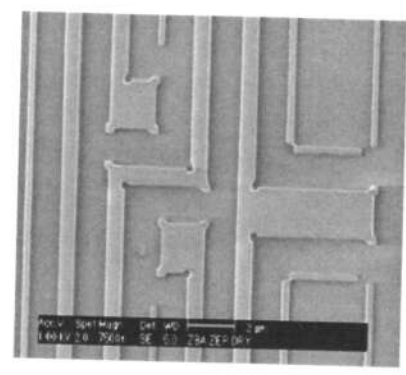


The process of fabricating a mask is very similar to the process flow for semiconductor devices: start from the glass substrate, deposit the metal, deposit a photoresist, pattern the photoresist through electron beam direct writing, but this process is very slow (~hours for 1 mask) so that's why we don't use electron beam direct writing directly on the wafers, which could have thousands of replicas of the mask pattern. With lithography we can process an entire wafer in ~10 seconds.

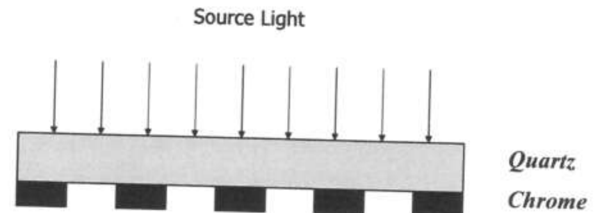
throughput = number of wafers processed / hour

The throughput of lithography is significantly higher than if we directly printed the pattern on the wafer using electron beams.
If you just have to do 1 ~ 2 patterns (e.g. for research purposes) you would use an SEM to pattern the photoresist.

NOTE: since current photo-masks are reduction reticles, they're 4 ~ 5 times bigger than the final pattern we want on the wafer, so we would need a lower resolution to create our photo-masks.



Top Down SEM picture of a photo mask (from Cr side)

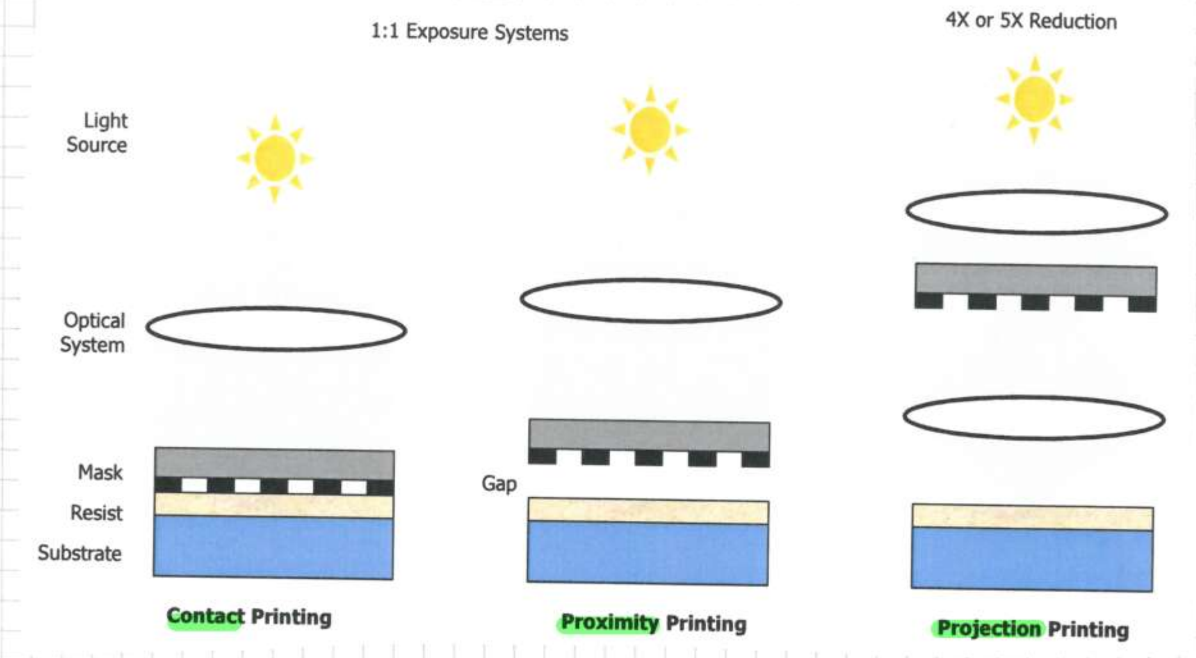


- **Quartz:** transmission 100% to DUV light
- **Chrome:** transmission 0% to DUV light

Cross Section sketch of a photo mask

Most of the time, the mask used is the positive of what we get on the wafer, since the light will pass only where there is NO chrome, and that portion of the photoresist will be removed

WAFER EXPOSURE SYSTEMS



If you run optical lithography there are 3 main systems we could use.

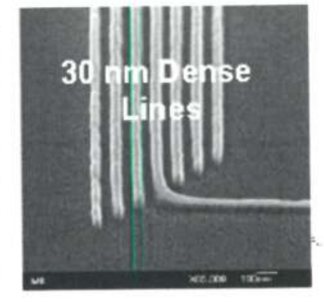
- Contact printing: the mask is in direct contact with the wafer (with the photoresist on top of it)
- Proximity printing: there's a small gap between the mask and the photoresist
- Projection printing: there's a lense in between the mask and the photoresist, so that we can reduce by 4 ~ 5 times the pattern.

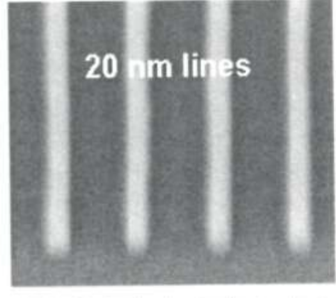
NOTE: contact and proximity printing are both 1 : 1 exposure systems, there's no "zoom" like in projection lithography.

NOTE: for masks with features $\approx \lambda$, diffraction becomes very important, and nowadays we have to account for them in the mask design.

For proximity printing we will use the near-field approximation, while for projection printing the far-field approximation.

CONTACT PRINTING





This is the oldest and simplest printing process, and since we have direct contact of mask and resist, diffraction effects are minimized, which makes this a high resolution technique (< 100 nm, well below that).

CON: hard contact results in damage to both mask and resist layer, resulting in high defect densities.

CON: 1× masks are required, which are ever more difficult to produce.

NOTE: the main drawback is the hard contact, since we'll have to reproduce it thousands of times across millions of wafers.

For a small laboratory / research environment, this is fine, but not for advanced IC manufacturing.

PROXIMITY PRINTING

When we move our mask away by a small gap (5 ~ 25 μm) we immediately get punished for our greed by diffraction effects.

CON: the gap between the mask and the wafer degrades the resolution of the printed patterns due to diffraction effects → near-field (or Fresnel) regime, which is when the image plane is close to the aperture

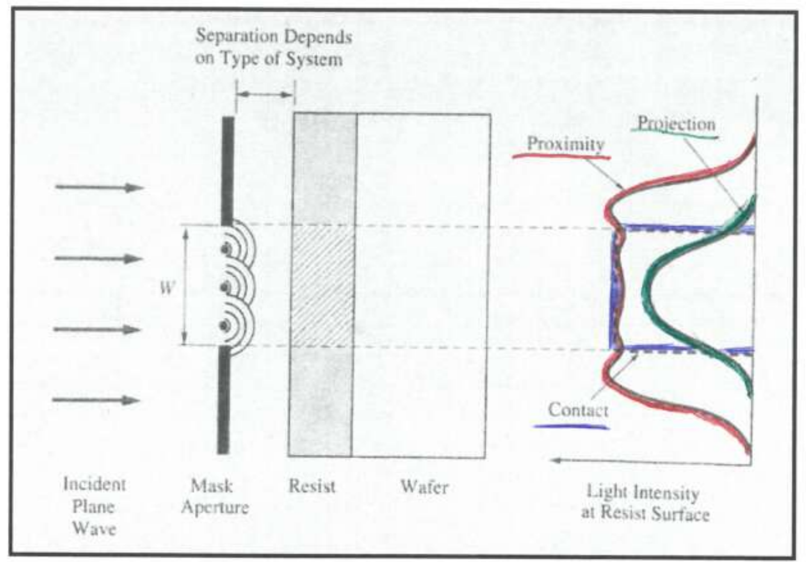
RESOLUTION LIMIT $R_{min} \approx \sqrt{K \lambda g}$

λ = wavelength of the source light
K = constant (depends from resist process) — usually $K \approx 1$
g = proximity gap

with $\lambda = 365$ nm (i-line), $g = 20\ \mu m$ → $R_{min} \approx 2{,}7\ \mu m$

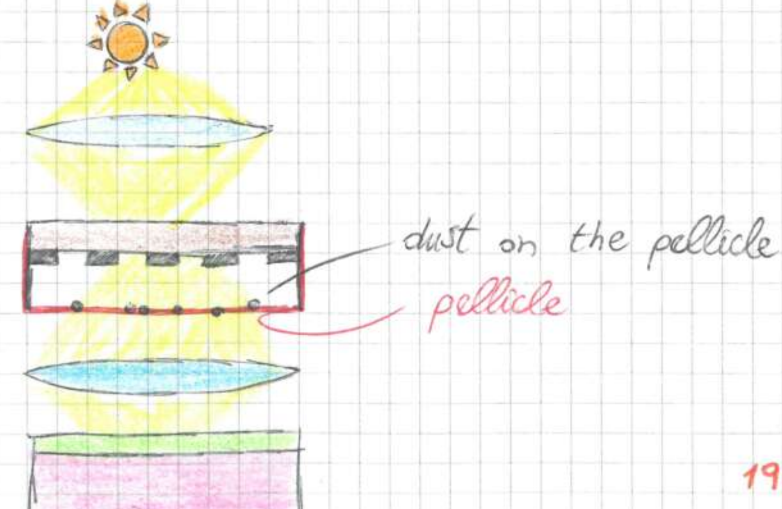
CON: 1× masks are required, which are ever more difficult to produce.

EXPOSURE SYSTEMS COMPARISON



For $W \sim \lambda$, we can consider every point of the aperture as a source of spherical waves (actually this can always be done, but only for $W \sim \lambda$ matters) and calculate how the resulting wavefront will look like for contact - proximity - projection.

NOTE: the mask actually has a frame around it with a pellicle suspended on top (just a thin layer of cellulose) to collect particles and dust so they don't land on the mask.
That's because on the mask they'd be on focus, while on the pellicle they will be out of focus and hence won't be projected.
What's degrading over time is the pellicle, not the mask itself.
The problem with soft x-rays / DUV is that few pellicles are compatible.



PROJECTION PRINTING

It's the dominant method of wafer exposure today because of its high resolution without the defect problem of contact printing.

The image reduction is currently 4 ~ 5 times, which is good for the control of reticle complexity and defectivity in the mask.

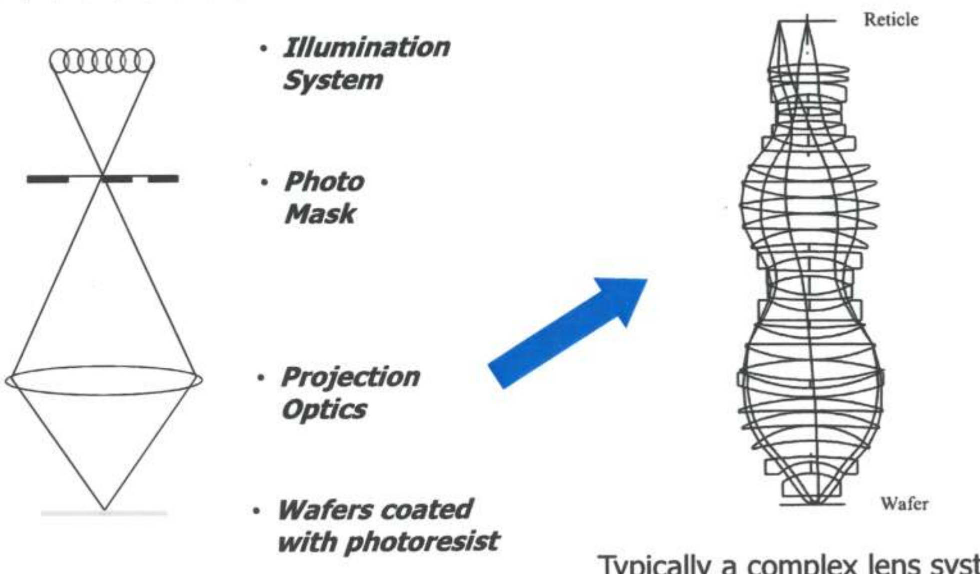
The resolution is limited by diffraction effects since the feature dimensions on the mask are comparable to the wavelength of the light.

We will be in the far-field (or Fraunhofer) approximation, where the image plane is far from the mask and a lens system is placed in between to capture and focus the image.

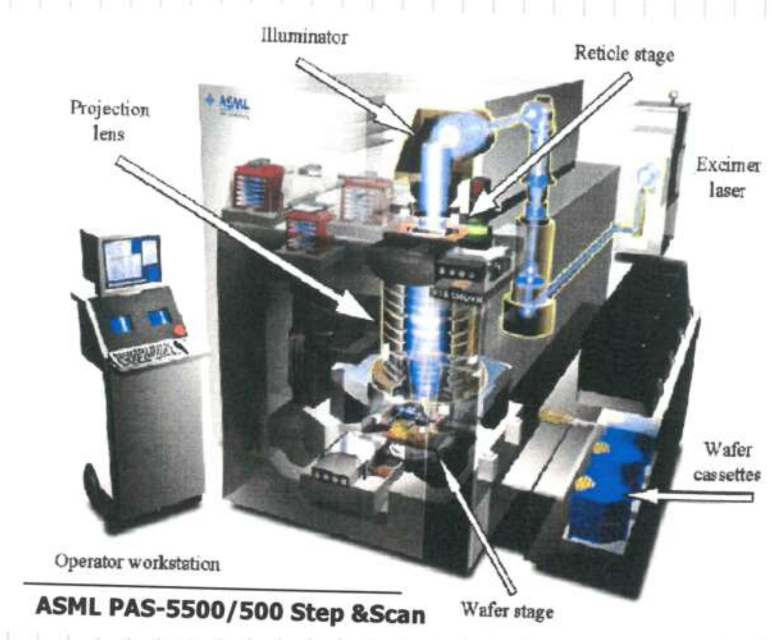
CON: resolution is limited by diffraction

CON: since the lens isn't infinite, it will collect and focus only part of the total diffraction pattern. Some information regarding the mask is lost while traveling to the wafer.

NOTE: the lens is the main drawback



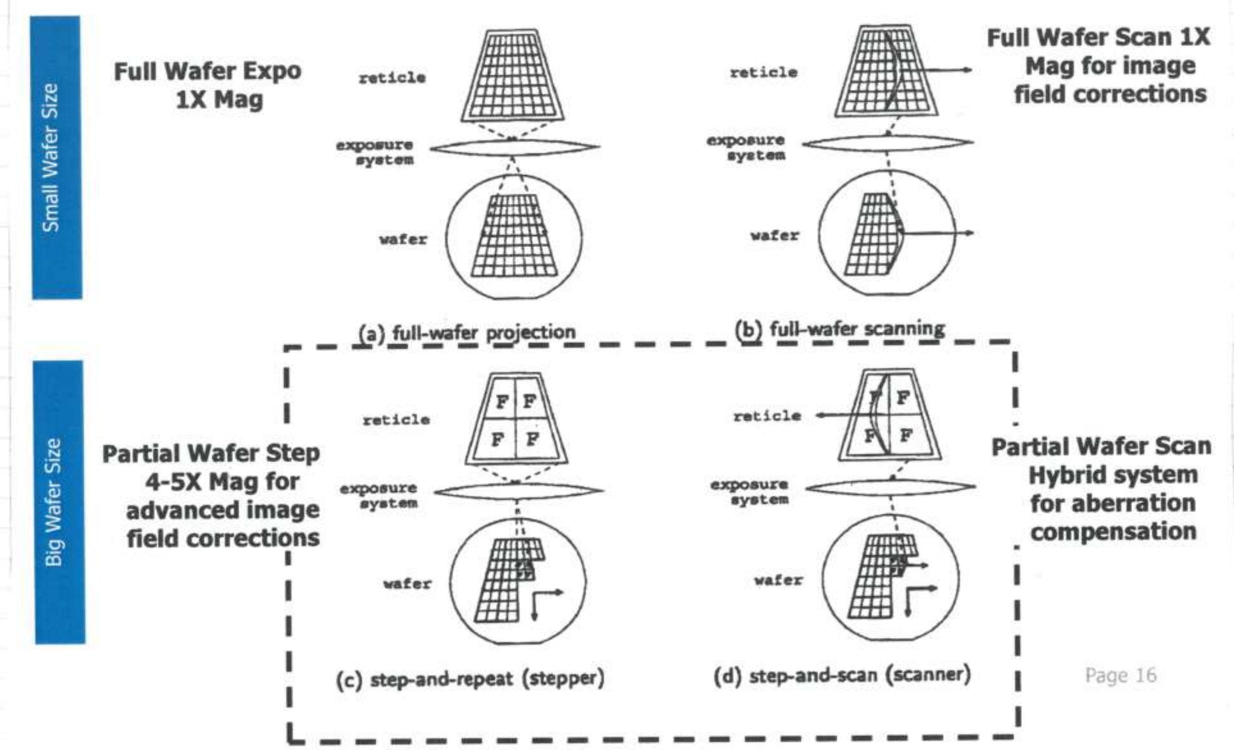
Typically a complex lens system with tens of elements



ASML PAS-5500/500 Step &Scan

In a real lithographic tool we don't have "a" lens, but a complex system of lenses (> 2 m tall, tens of lenses) and usually the exposure is separated from the rest of the lithographic process (physically separated: exposure is performed in a machine called Stepper, while the rest in the Track unit, where the wafer is processed before and after light exposure).

VARIOUS TYPES OF PROJECTION PRINTING



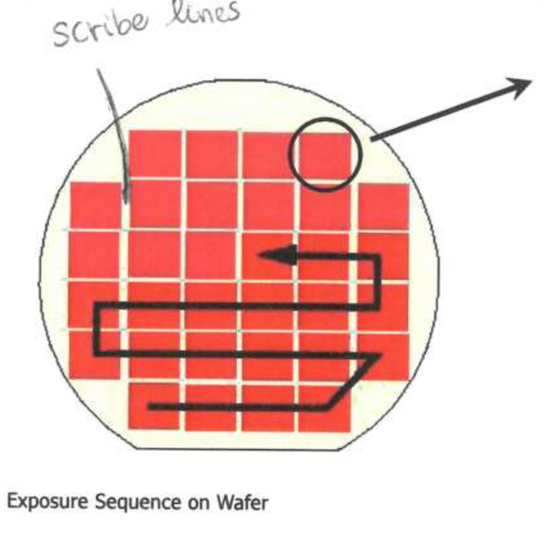
We can have several ways of printing what you want to print down on the wafer using projection printing.

We can either expose the whole reticle (TOP left), or scan it (TOP right). Scanning has the advantage of using a smaller portion of the lens by "selecting a slice" of the lens which is free of defects.

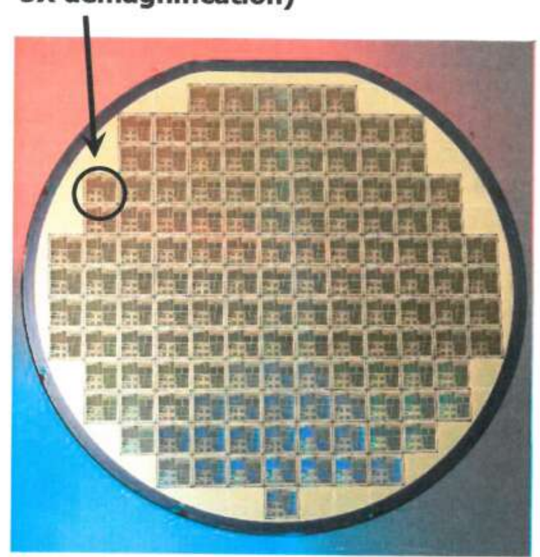
For modern wafers and masks, we won't ever have a full wafer exposure of a single reticle (the reticle is 4 ~ 5 times larger than what we will print), so what we do is print, move, repeat (STEPPER). But the most used method today is the SCANNER, where while the wafer moves the reticle is being scanned (picture bottom right), so we use a smaller portion of the lens.

STEPPER (step and repeat) and SCANNER (step and scan)

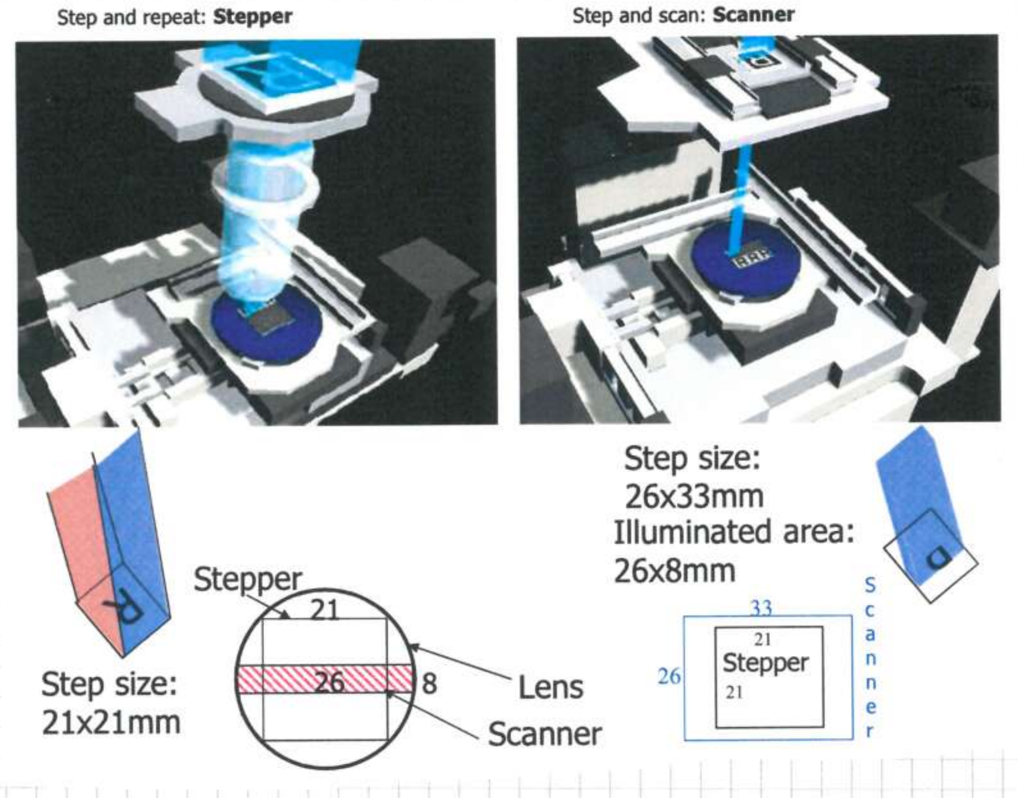
STEPPER



Exposure Sequence on Wafer



SCANNER

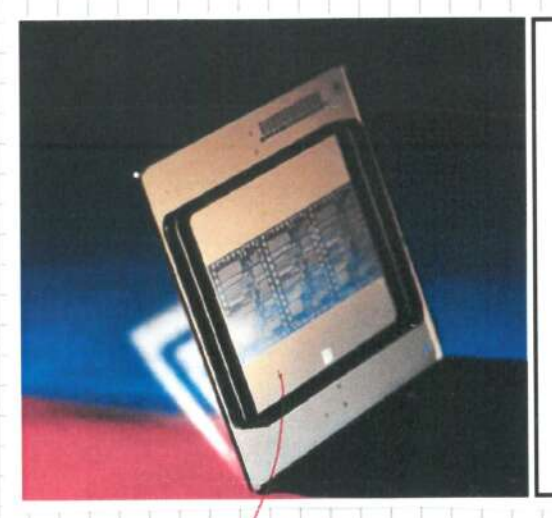


In the scan system we can reach:

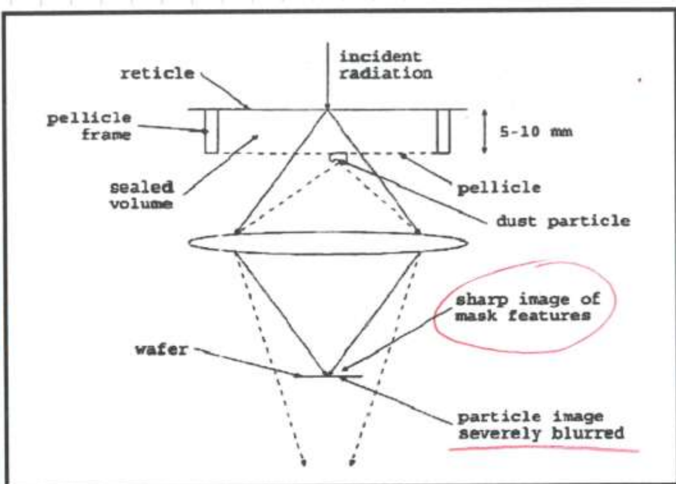
wafer stage speed	250 ~ 500 mm/s
acceleration	0.5 g (~ 4,9 m/s²)
reticle stage speed	1000 ~ 2000 mm/s

By using a smaller portion of the lens combined with synchronized movement of reticle and wafer, the scanner is able to achieve less geometrical aberration level (→ better dimension control) and wider exposure fields (→ higher throughput

PHOTO - MASK



pellicle is in here

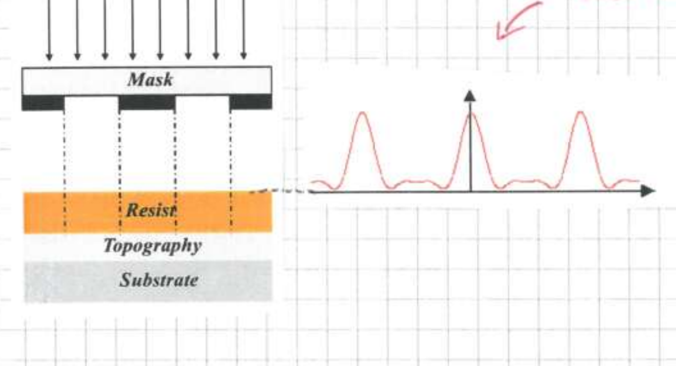


RESIST PROCESS (/ PHOTORESIST PROCESS)

① Aerial image formation at resist level

"aerial image" is the intensity modulation of light which is forming on top of the photoresist because of the optical projection system

aerial image



doesn't look like a step function because of diffraction

② Latent image formation inside the photoresist, in the exposed areas

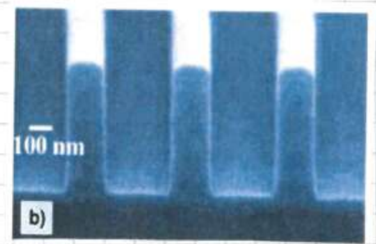
Where the photoresist was exposed to light forms a "latent image", latent because at this point of the process we don't see any pattern on the photoresist yet.

latent image

Resist
Topography
Substrate

③ Resist develop, where the areas exposed to radiation become soluble to developer and get removed.

positive resist



So photoresist materials are designed to respond to incident light by changing their properties when exposed to light.

Positive resist: exposed areas dissolve in resist developer

Negative resist: exposed areas become insoluble in develper, unexposed areas do

A good resist must have at least the following 3 characteristics:

- best possible pattern fidelity and resolution: the change in chemical properties must be as localized as possible, so good contrast
- highest possible sensitivity to incident energy (which is around mJ/cm^2) so we'd like to need few photons in order to promote chemical reaction.
- good resistance to subsequent processes (we want to be able to then keep the developed photoresist in order to use it during etching / ion implantation).

Lecture 18

30 aprile

G-LINE AND I-LINE PHOTORESISTS

→ remember that g-line is 436 nm and i-line is 365 nm

Those photoresists were made by 3 main components:

- inactive resin = polymer, is the base of the material
- solvent: it's used to adjust the viscosity of the resist, since we dilute the resist in the solvent so it's liquid at the moment of deposition, then the solvent will evaporate leaving behind the resin and the PAC
- PAC: stands for Photo Active Compound and upon exposure to light a chemical process occurs by which the compound becomes a "sensitizer" (= changes the dissolution rate of the resist in the solvent).

The most commonly used g-line and i-line resist today is DQN

N = Novolac (is the polymer, inactive resin)
DQ = Diazoquinone (is the PAC)

POSITIVE G-LINE / I-LINE CHEMICAL PROCESS

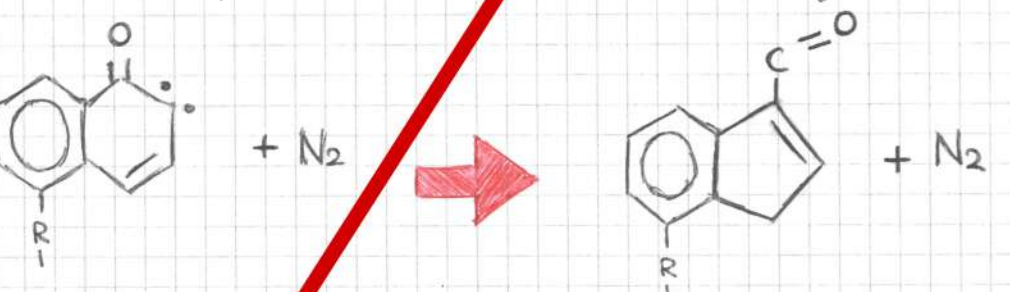
When the resist is exposed to light, the diazoquinone molecules chemically change: the N_2 molecule is weakly bonded in the PAC and the light breaks this bond

diazoquinone

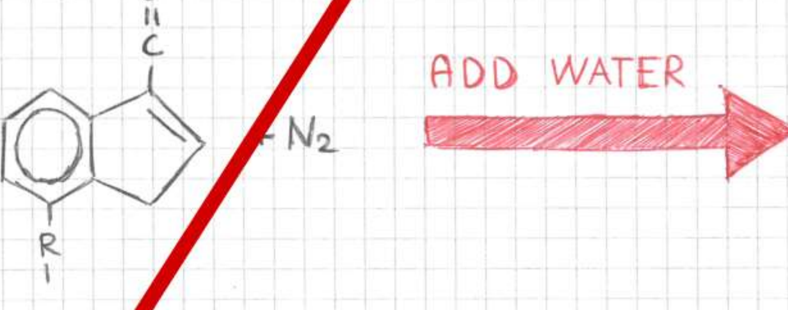
$+ N_2$

R = rest (NOVOLAC, for example)

But now the PAC structure is unstable, and to stabilize itself it moves a carbon atom outside the ring (so the hexagon becomes a pentagon) and the oxygen atom will be covalently bonded to it (Wolff rearrangement)



Now if we add water we obtain a carboxylic acid which is soluble in a basic developer, and the rest ("R", for example Novolac) is soluble in water and can easily dissolve



DEEP UV RESISTS

What changes when we move deeper in the UV spectrum to DQN?

DQN resist materials have 2 significant problems when shorter λ are used:

- resist absorption: below 365 nm incident photons are strongly absorbed, so the incident light can't penetrate the whole thickness of the resist

- resist sensitivity: the intensity of Hg arc lamps in DUV is very low. High sensitivity resist are needed for adequate exposure time, for this reasons PAC needed to be reviewed to work more effectively at 248 nm and below (like 193 nm).

NOTE: DUV resists today are NOT "modified DQN" resists, they are based on a new chemistry and make use of chemical amplification (CAR)

CHEMICALLY AMPLIFIED RESIST (CAR)

→ The incoming photons react with a photo-acid generator (PAG) molecule, creating an acid molecule.

→ This acid molecules act as catalysts during a subsequent resist bake to change the resist properties in the exposed regions

→ The reactions are catalytic, so the more I keep the wafer on a hot stove the more the reaction keeps on going because the acid molecule is regenerated after each chemical reaction and may participate in tens or hundreds of further reactions.

This catalytic feature of CAR allows us to work at relatively low doses of illumination:

exposure doses for I-line resist → 100 ~ 150 mJ/cm^2

exposure doses for DUV resist → 15 ~ 40 mJ/cm^2

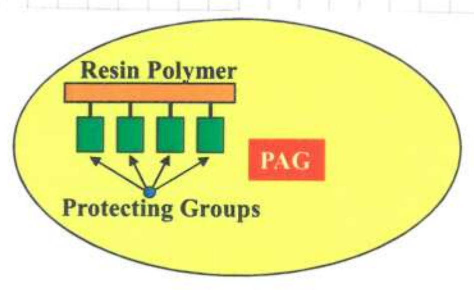
POSITIVE RESIST

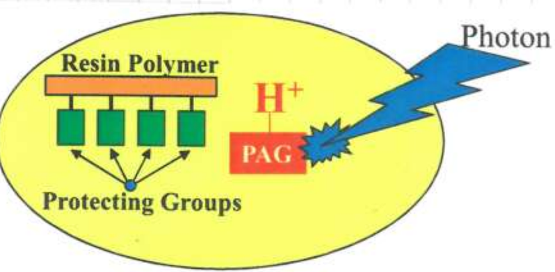
The resin polymer has attached to it protecting groups, which render the resin insoluble in the developer.
The incident DUV photons react with the PAG inside the resist to create an acid molecule.
After exposure to light, the wafer is baked ("Post Exposure Bake", PEB).
The heat provides the energy needed for the reaction between the acid molecules and the protective groups, removing the protecting groups and rendering the resin polymer soluble in the developer.

NOTE: the real reaction acts during the PEB

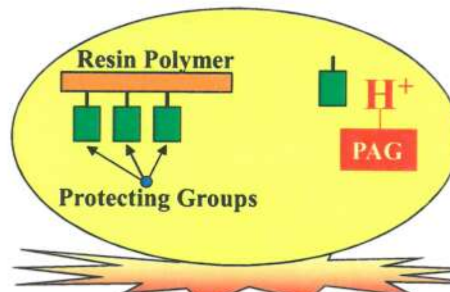
NOTE: PEB temperature and time are crucial for process control, since PEB is used to drive the chemical reaction that completes the resist exposure.

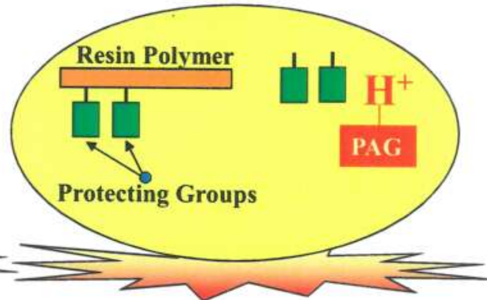
exposition → PAG (photo acid generator) → bake to remove protective groups → developer





1) Unexposed resist **2) Upon exposure acid production**

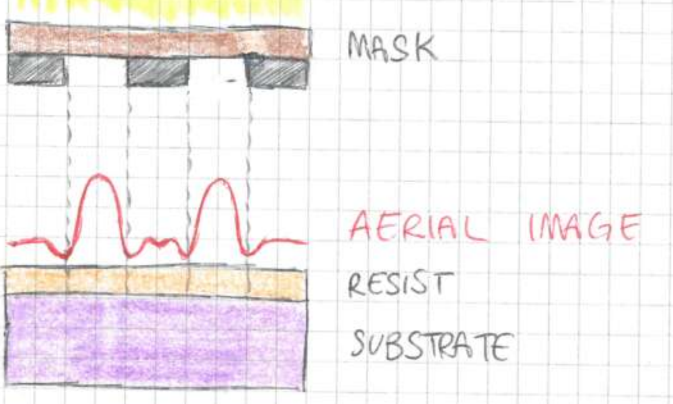


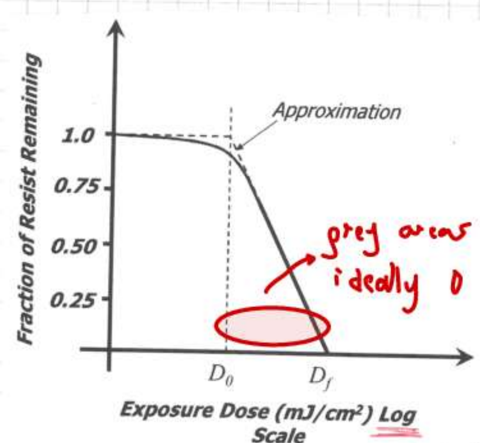


RESIST CONTRAST

CONTRAST

Contrast is a measure of the resist's ability to distinguish light from dark areas in the aerial image (PAGE 201) the exposure produces





In this graph (exposure dose → LOG scale) we can see how much resist is left as a function of the exposure dose

D_0 = dose at which the exposure first begins to have an effect

D_f = dose at which the exposure is complete (no more resist left)

What I would like to have in principle is a step function ($D_0 = D_f$).
The slope of the steep part of the curve is the contrast γ

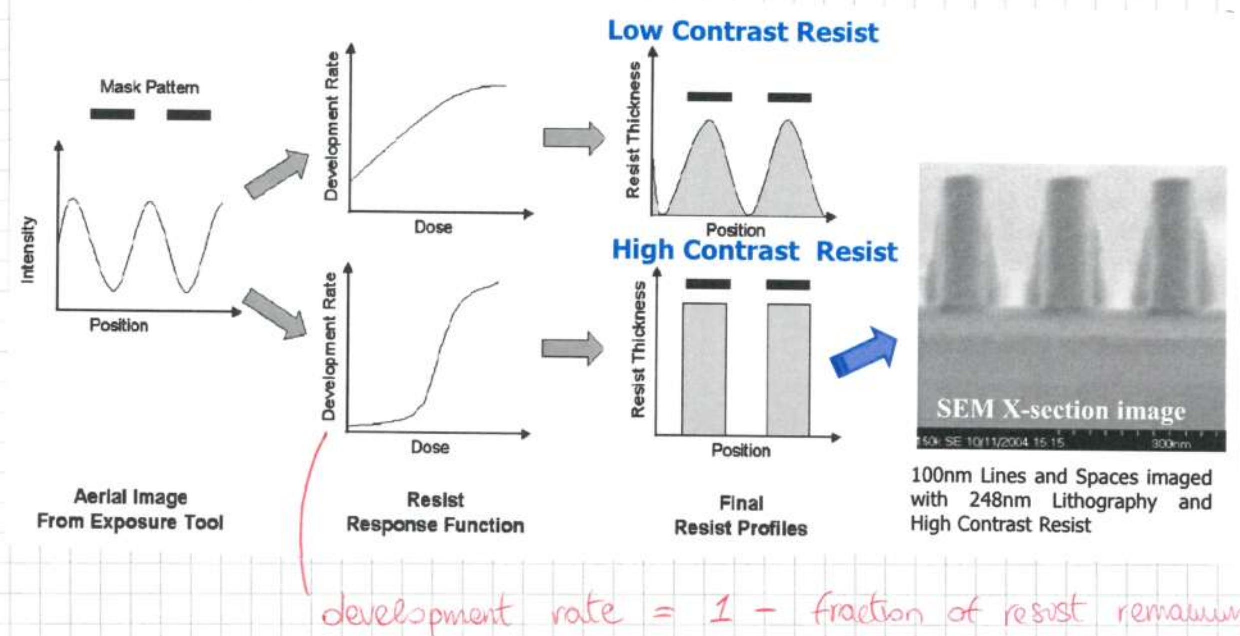
$$\gamma = \frac{1}{\log_{10}\left(\frac{D_f}{D_0}\right)}$$

We want the slope to be as steep as possible, since how the resist responds to "grey areas" ($D_0 \leq dose \leq D_f$) is fundamental

Contrast of the resist is determined experimentally:

for g-line / i-line is 2 ~ 3

for DUV contrast is 5 ~ 10



development rate = 1 − fraction of resist remaining

We see that the aerial image that we get from 2 apertures isn't comprised of square waves, and so the resist contrast plays a fundamental role in giving us the pattern we want developed.

PROCESS FLOW OF THE WAFER IN A LITHOGRAPHIC SYSTEM

Remember that the lithographic system is composed of 2 tools attached together: TRACK + STEPPER / SCANNER

TRACK
- Dehydration bake + HMDS application
- Resist application
- Pre exposure bake (SB: Soft Bake)

STEPPER or SCANNER
- Alignment + Exposure

TRACK
- Post exposure bake (PEB)
- Develop
- Final bake (HB: Hard bake)

TRACK

1. DEHYDRATION BAKE

This step is the first one inside the track, and is necessary to drive off any water vapor on the wafer surface.
Typical process is 150°C for 60 ~ 90 seconds

2. HMDS APPLICATION

HMDS = Hexamethyldisilane is an adhesion promoter, typically applied in vapor form. Without an adhesion promoter the photoresist wouldn't stick to the wafer, it would flow away (or lift off).

HMDS molecules have the possibility to form 2 bonds: one with the substrate and one with the resist.

NOTE: everything is deposited by a nozzle applying stuff while the wafer is rotating

3- RESIST APPLICATION

After the adhesion promoter we apply the photoresist (is liquid, contains a solvent).

The wafer spins at 1500 ~ 5000 RPM ~30 seconds

The film is very uniform regarding thickness (~ 2 - 10 nm).

The final thickness is defined by the speed of the spinning and viscosity of the resist.

The final thickness depends strongly on the application (etch, implant, ...), but typically it is around 0,1 μm ~ 5 μm

4- PRE EXPOSURE BAKE (or SOFT BAKE)

Wafer on a hot stove at ~ 100°C for 60 seconds, makes the solvent evaporate out of the photoresist and also improves adhesion since the heating strengthens the bonds between the resist, the HMDS and the substrate

STEPPER or SCANNER

5- ALIGNMENT + EXPOSURE

Now the wafer has entered the tool that does the real photolithographic part. It is properly aligned and is exposed to light, with doses varying wildly depending on resist sensitivity, the λ we're using, ...

I - line resist → 100 ~ 150 mJ/cm²
DUV resist → 15 - 50 mJ/cm²

The typical exposure time for modern systems is 60 ~ 90 seconds per wafer.

TRACK (again)

6_ POST EXPOSURE BAKE (PEB)

We are out of the Stepper / Scanner and back into the Track.

The PEB is a crucial step especially for CAR photoresist.

Wafer on a hot plate at 100 ~ 150 °C for 60 ~ 90 seconds.

The effects of PEB process depend on resist type (I-line or DUV), but still is a very important step for resist definition, since the chemical reaction rates and diffusion strongly depend on this process.

7_ DEVELOP

Before we apply the developer, only a latent image of the mask is present on the wafer.

We apply the developer on the spinning wafer, which removes the soluble parts of the resist.

The typical developer is TMAH (Tetra Methyl Ammonium Hydroxide), which is a basic solution.

NOTE: TMAH also etches silicon, so we have to be careful

Rinsing (= sciacquare) the wafer with H_2O stops the developing process

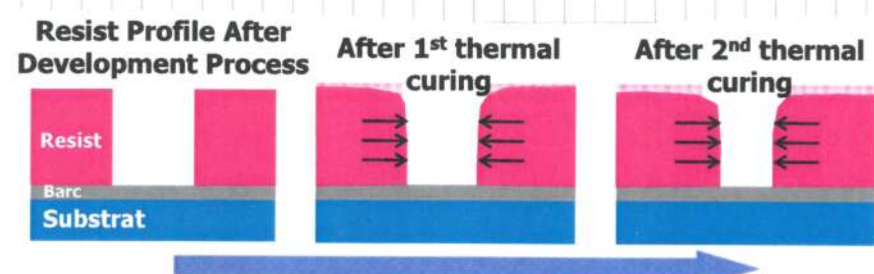
8_ FINAL BAKE (or HARD BAKE)

This step is done at ~ 150 °C and is designed to harden the resist and improve its resistance to subsequent etch or implantation processes.

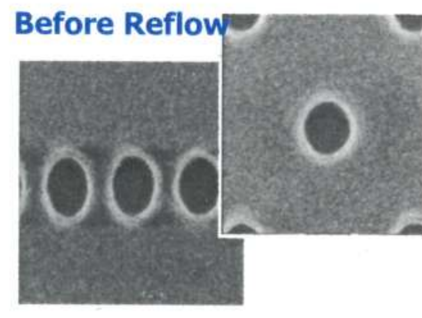
The final bake can be used to modify the final dimensions within the resist, since we can heat it up so much that it can start to reflow (like molten glass).

THERMAL REFLOW

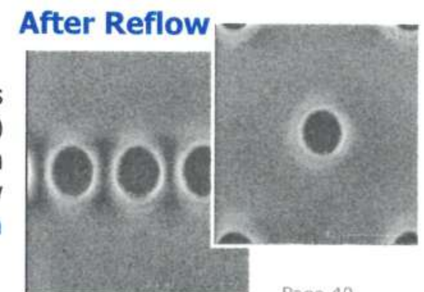
During the final bake, in order to ==reduce feature dimensions==, the photoresist can be processed with a controlled thermal process that "shrinks" dimensions



Progressive Dimension Reduction



120nm Contact Holes (Minimum pitch 200nm) imaged with 248nm lithography under reflow process. Contact dimension is reduced by 20nm

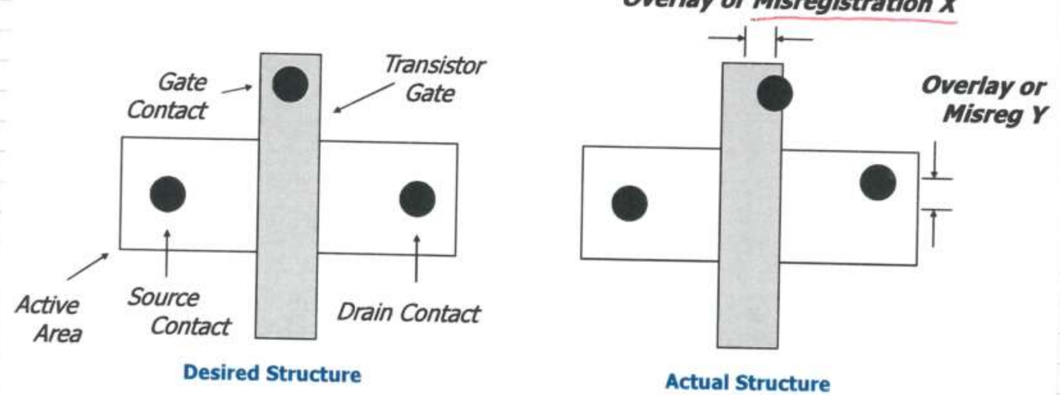


Page 40

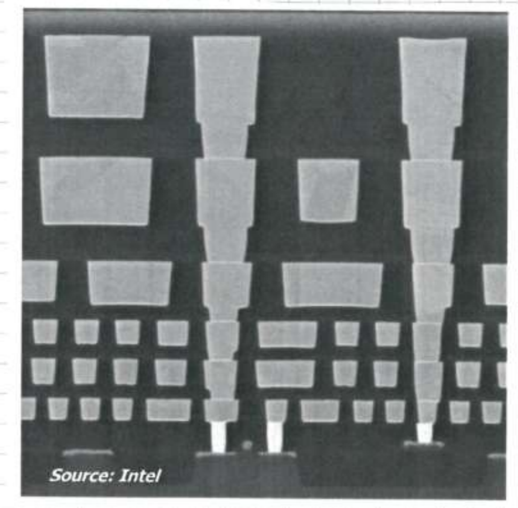
REGISTRATION

Photolithographic processes should be judjed with 3 metrics:

- resolution (depending on exposure system and λ)
- throughput (related to exposure speed)
- registration



Proper alignment from level to level is crucial for proper device functionality!

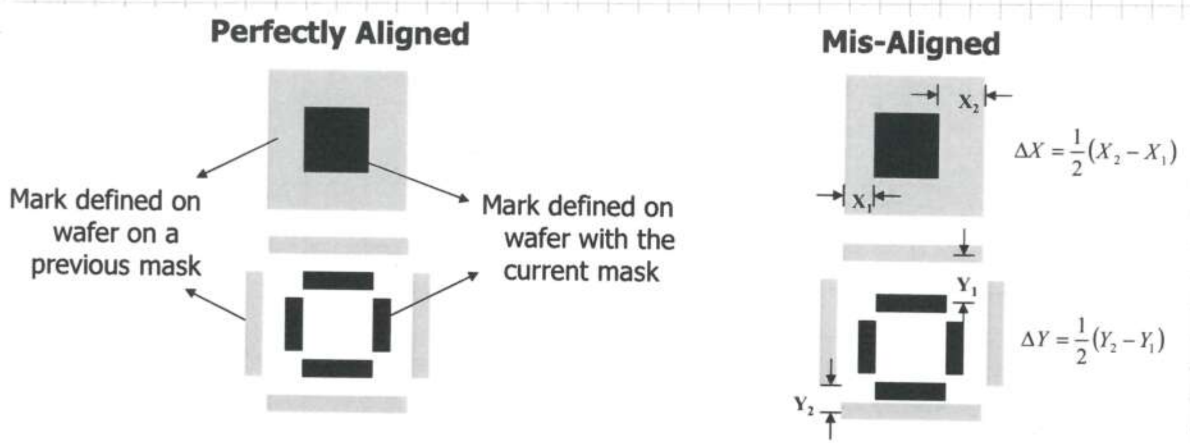


2008 ITRS Roadmap Predictions for Flash Memories

Year	2007	2008	2009	2010	2011	2012
Technology Node (nm)	54	45	40	36	32	28
Overlay 3σ (nm)	17.7	14.9	13.2	11.8	10.5	9.4

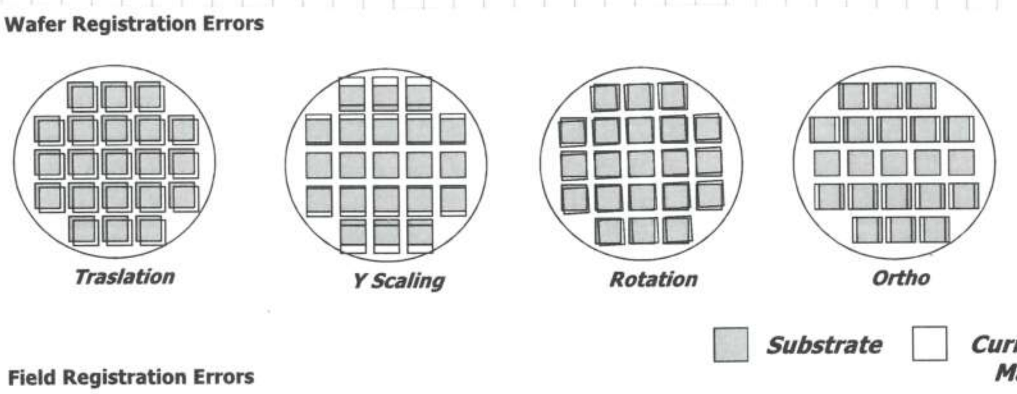
nowadays the overlay margin is a few nm (2 ~ 2,5 nm)

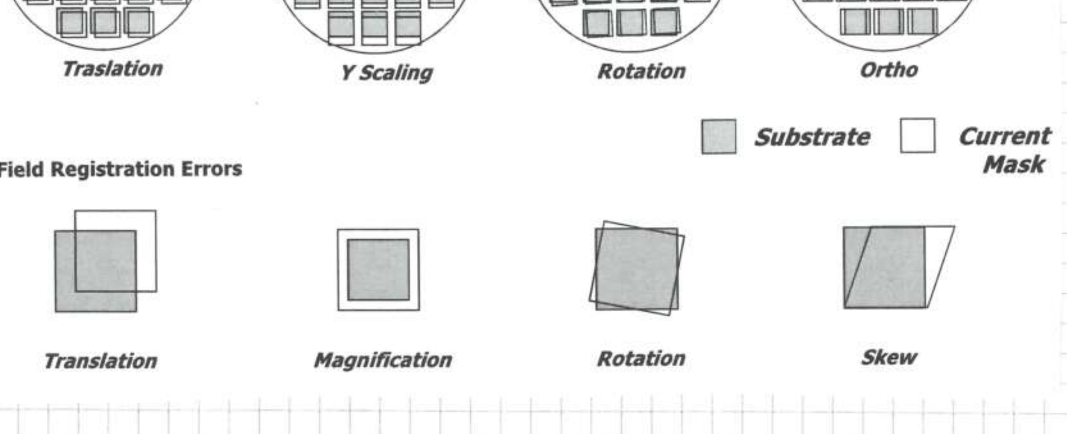
In the scribe lines, one of the test structures we can put is for registration measurements, either with Box in box or with Frame in Frame.



The dedicated features (Box in box or Frame in frame) are used to evaluate mutual alignment between 2 masks, and the structures are big enough to be measured with optical systems. If we're misaligned, we can go back into the lithographic tool and re-align.

NOTE: we can repeat this process almost as many times as we want, since we just have to strip away the photoresist and re-apply it, contrary to (for example) ion implantation: if you make a mistake you have to use another wafer.





ALIGNMENT STRATEGY

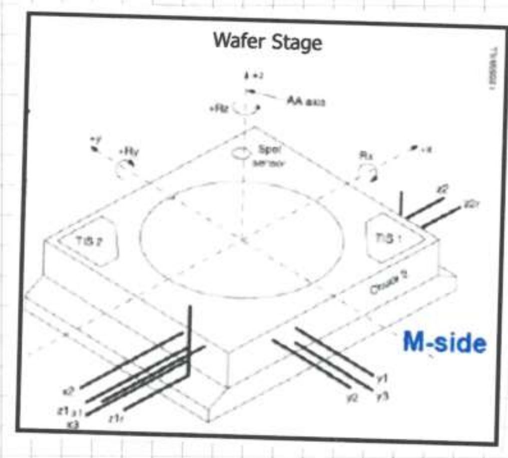
Special structures, called alignment marks, are generated on the wafer during the previous mask.

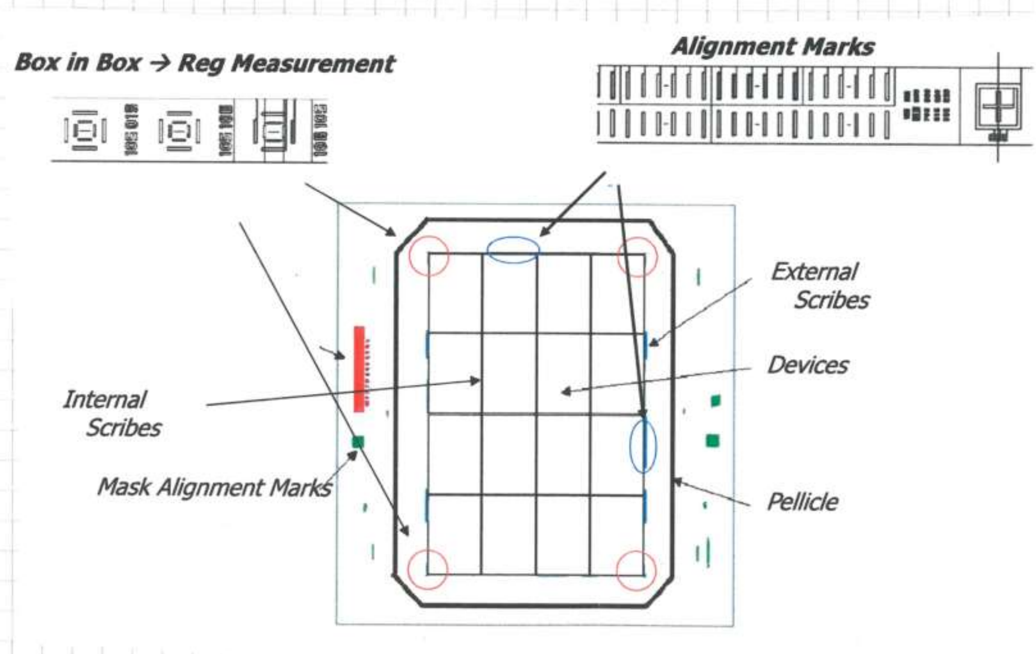
An alignment system is able to recognize these marks, calculating their coordinates with respect to an absolute reference system.

Fine alignment marks for a new level are positioned at the borders of the exposure mask.

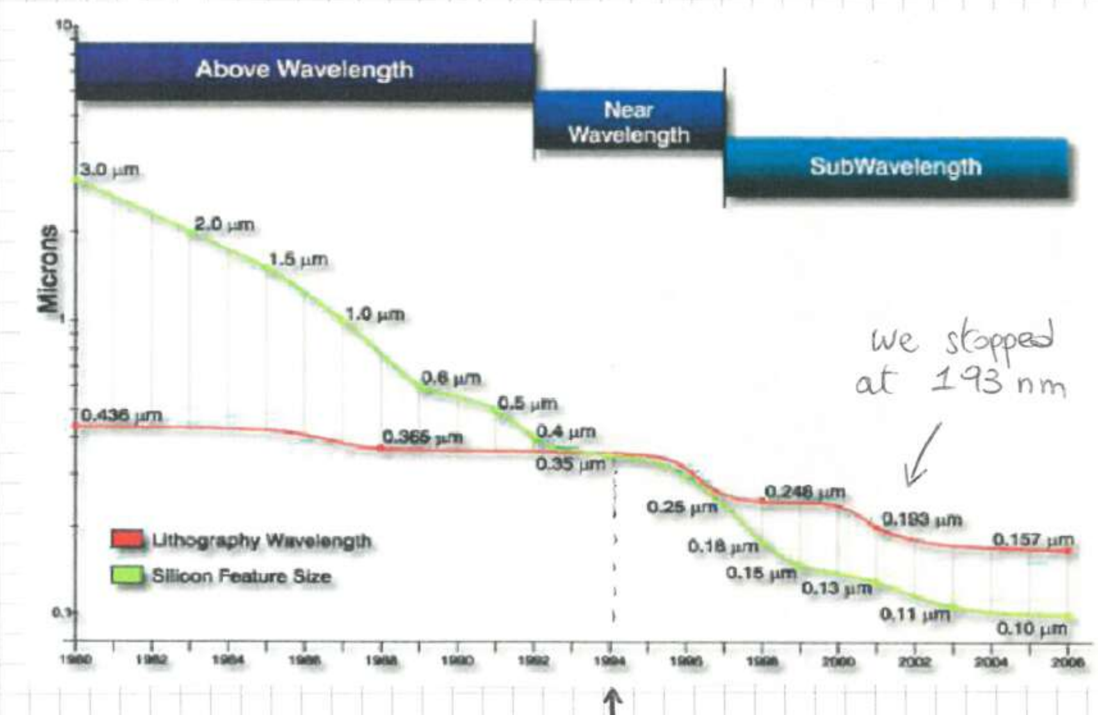
On the SCANNER, wafer and reticle can be moved with high precision using several interferometers that control X, Y, Z, R_x, R_y, R_z.

For state of the art tools the accuracy is of a few nm.





SUB-WAVELENGTH GAP



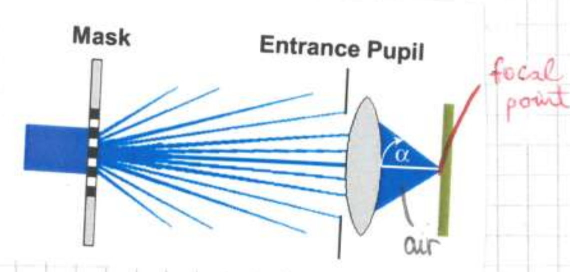
around here we started printing features smaller than the λ used

To keep pace with the demands of Moore's Law, a strong use of resolution enhancement techniques is necessary.

Let's first define some optical parameters:

NUMERICAL APERTURE

It determines the maximum number of diffraction orders that can be captured by projection lens, and thus the quality of the reconstructed image



$$NA = n \sin \alpha$$

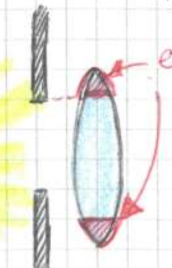
n: refraction index of medium (air = 1)

α: semi angle of the cone (vertex = focal point)

$0 \leq NA \leq n$ n if ∞ lens

NA gives an idea of how big the lens is

NOTE: usually we'll have a pupil in front of the lens covering part of it, because we don't want to use those parts since they cause aberrations.



excluded part of the lens because of aberrations

higher NA, more peaks captured, better

Another angle we need to take into account is the angle of illumination of the mask, and is usually referred as the coherence of the illumination

PARTIAL COHERENCE ($0 < \sigma < 1$)

max angle of light

mask

lens

α_0

α

focal point

illumination

entrance pupil

$$\left[\sigma = \frac{\sin \alpha_0}{NA} \right]$$

divided by NA only because its practical

- $\sigma = 0$, coherent illumination
- $\sigma = 1$, incohoerent illumination

if we have perfect parallel rays coming from infinite, we get $\sigma = 0$ (coherent light)

COHERENT LIGHT

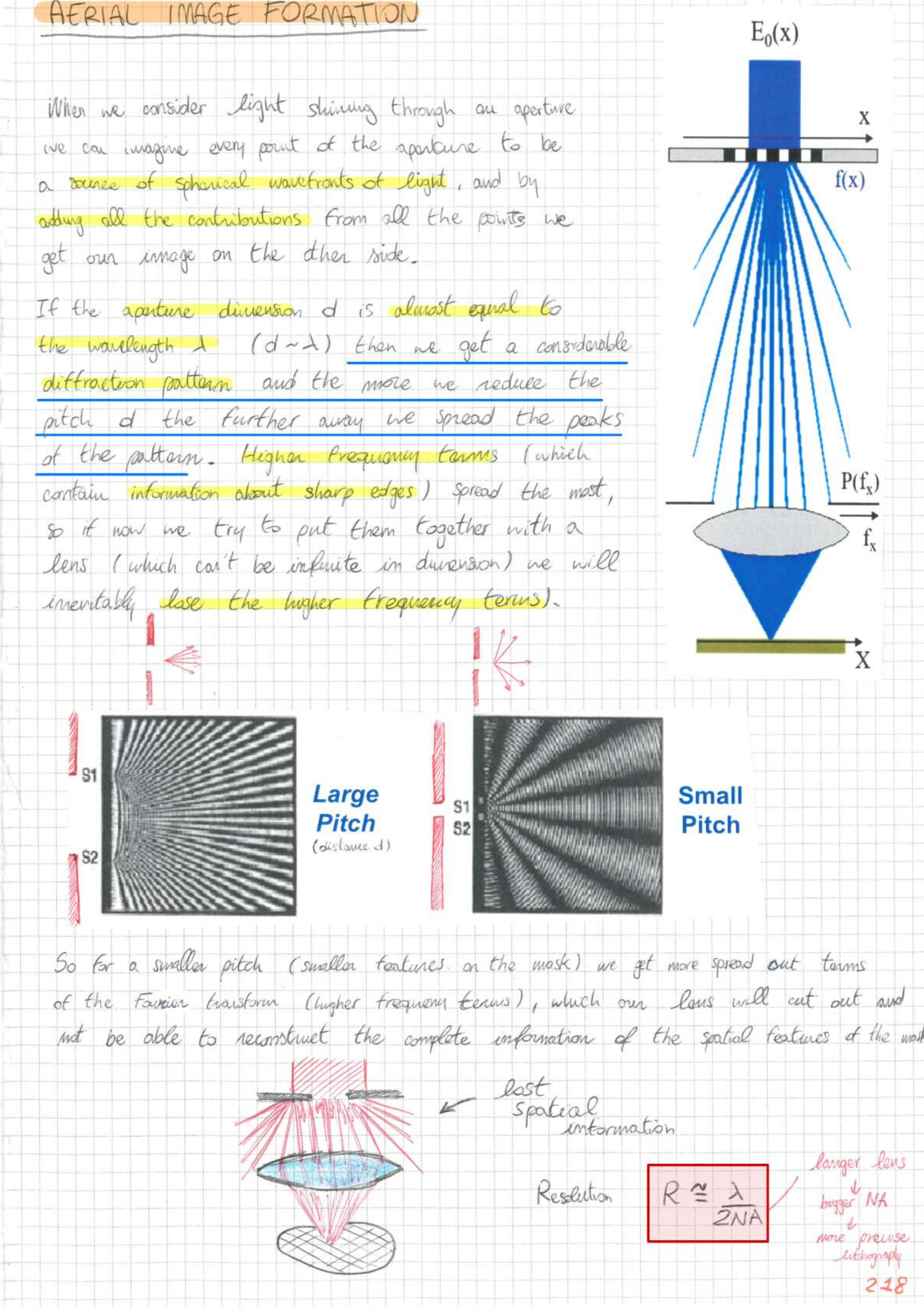
INCOHERENT LIGHT

if we have completely diffused light before the mask (so the maximum angle is $\alpha_0 = 90°$) we have completely incoherent light ($\sigma = 1$)

NOTE: this coherence σ has nothing to do with the coherence of the source

---

NB the interference pattern is the Fourier transform of the mask (in space) see after.



AERIAL IMAGE FORMATION

When we consider light shining through an aperture we can imagine every point of the aperture to be a source of spherical wavefronts of light, and by adding all the contributions from all the points we get our image on the other side.

If the aperture dimension d is almost equal to the wavelength λ ($d \sim \lambda$) then we get a considerable diffraction pattern and the more we reduce the pitch d the further away we spread the peaks of the pattern. Higher frequency terms (which contain information about sharp edges) spread the most, so if now we try to put them together with a lens (which can't be infinite in dimension) we will inevitably lose the higher frequency terms).

So for a smaller pitch (smaller features on the mask) we get more spread out terms of the Fourier transform (higher frequency terms), which our lens will cut out and not be able to reconstruct the complete information of the spatial features of the mask

Resolution $R \cong \frac{\lambda}{2NA}$

larger lens → bigger NA → more precise lithography

(1)

p

Lens Entrance Pupil

$E(f_x)$

-3/p -2/p -1/p 1/p 2/p 3/p

(2)

p

Lens Entrance Pupil

$E(f_x)$

-3/p -2/p -1/p 1/p 2/p 3/p

(3)

p

Lens Entrance Pupil

$E(f_x)$

-2/p -1/p 1/p 2/p

(4)

p

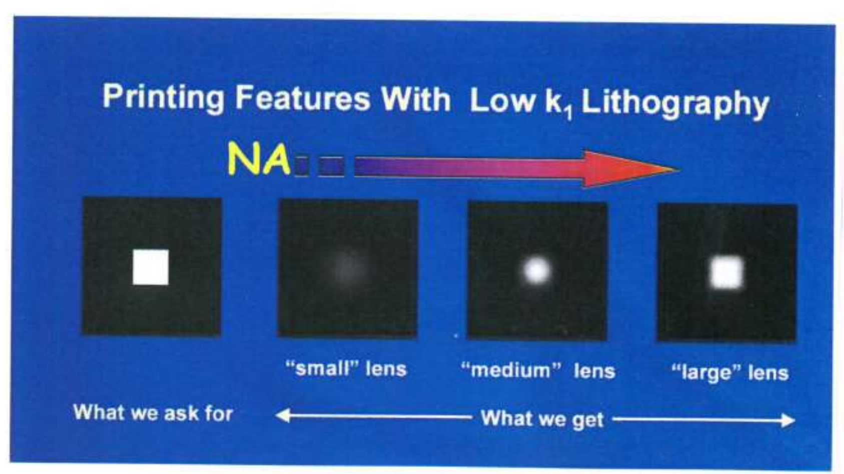
Lens Entrance Pupil

$E(f_x)$

-2/p -1/p 1/p

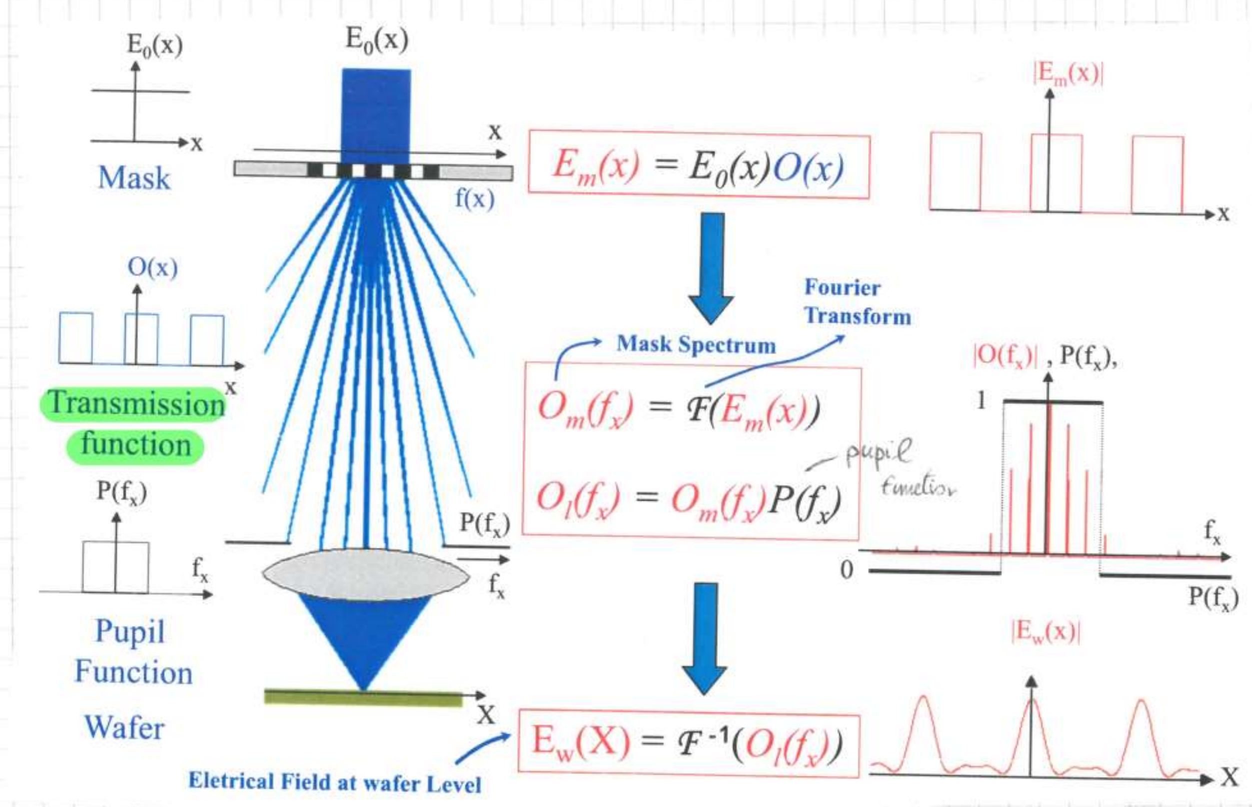
Just average intensity of the light

We can see that as the pitch (p) of the mask becomes smaller, the electric field spreads out more and more, until at (4) we collect only the central spectrum, which only contains the average intensity of the light → NO USEFUL INFORMATION



Lecture 19

5 maggio



Since we're using masks with features comparable (or even significantly smaller) to the wavelength λ of the light, diffraction phenomena will be important.
But after our light diffracts at angles from $-\infty$ to $+\infty$ the lens will be able to collect and reconstruct only part of the spectrum, losing some spatial features (high spatial frequency, like sharp corners) of the mask.

NOTE: remember that with closer features (smaller pitch) the diffraction spectrum (peaks and valleys) will spread further apart, losing more information.

NOTE: actually it's not about the dimensions of a single feature, but how many small things we want to cram very close all together, so it's easy to print a single 2nm square, <u>VERY</u> hard to print 2nm squares with 3nm space in between.

Let's try to formalize all of this.

MASK SPECTRUM

Mask — Quartz, Chrome ($p/2$, p)

Transmission Function — 1, 0, $O(x,y)$

The simplest model of a mask we can think of is a monodimensional alternation of quartz (100% transmitted light) and chrome (0% transmission)

The transmission function $O(x)$ will tell us the amount of light passing through.

The period p is equal to the pitch ($p/2$ is the aperture).

Since (compared to the λ used) we're relatively "far" from everything (mask far away from the wafer, lens far away from the mask) we can use the far-field / Fraunhofer approximation and can calculate the electric field using the Huygens-Fresnel principle.

NOTE: the Huygens-Fresnel principle tells us that we can treat every point of the wavefront in the slit as a source of spherical wavefronts which will then interfere

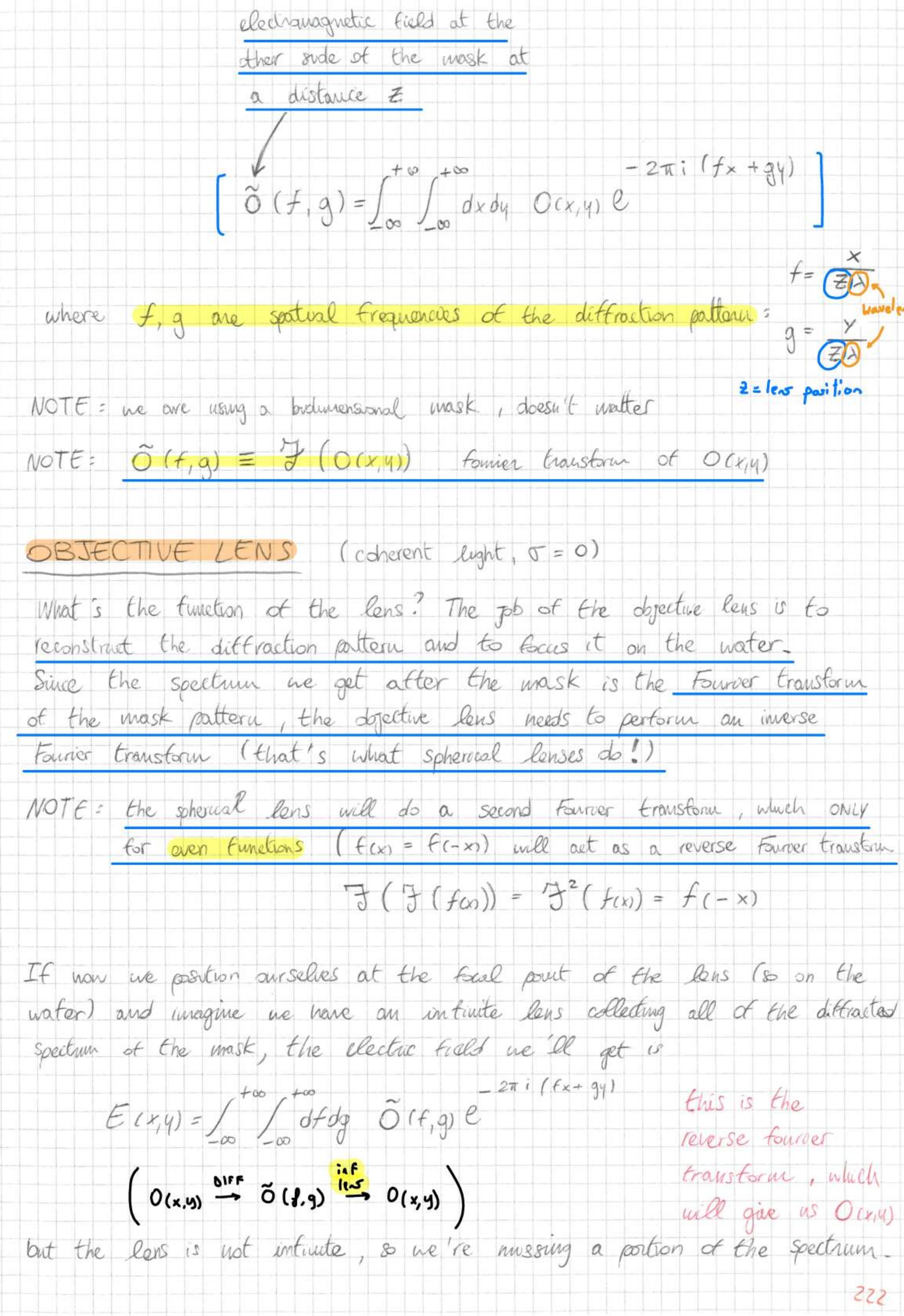
resulting wavefront

Fraunhofer → far field
Huygen-Fresnel → many spherical waves

!

NOTE: the far-field / Fraunhofer approximation can be used when we are looking at the diffracted pattern away from the source (= mask) of diffraction.

With the Fraunhofer approximation we can do "Fourier Optics", so we can calculate the electric field at a certain distance from the mask by just calculating the Fourier transform of the transmission function of the mask.



electromagnetic field at the other side of the mask at a distance z

$$\left[\ \tilde{O}(f,g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dx\,dy\ O(x,y)\, e^{-2\pi i (fx+gy)}\ \right]$$

where f, g are spatial frequencies of the diffraction pattern: $f = \frac{x}{z\lambda}$, $g = \frac{y}{z\lambda}$

(λ = wavelen., z = lens position)

NOTE: we are using a bidimensional mask, doesn't matter

NOTE: $\tilde{O}(f,g) \equiv \mathcal{F}(O(x,y))$ fourier transform of $O(x,y)$

OBJECTIVE LENS (coherent light, $\sigma = 0$)

What's the function of the lens? The job of the objective lens is to reconstruct the diffraction pattern and to focus it on the water. Since the spectrum we get after the mask is the Fourier transform of the mask pattern, the objective lens needs to perform an inverse Fourier transform (that's what spherical lenses do!)

NOTE: the spherical lens will do a second Fourier transform, which ONLY for even functions ($f(x) = f(-x)$) will act as a reverse Fourier transform

$$\mathcal{F}(\mathcal{F}(f(x))) = \mathcal{F}^2(f(x)) = f(-x)$$

If now we position ourselves at the focal point of the lens (so on the water) and imagine we have an infinite lens collecting all of the diffracted spectrum of the mask, the electric field we'll get is

$$E(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} df\,dg\ \tilde{O}(f,g)\, e^{-2\pi i (fx+gy)}$$

this is the reverse fourier transform, which will give us $O(x,y)$

$$\left(O(x,y) \xrightarrow{\text{DIFF}} \tilde{O}(f,g) \xrightarrow{\text{inf lens}} O(x,y) \right)$$

but the lens is not infinite, so we're missing a portion of the spectrum.

We will have to "cut away" some of the frequencies the lens is collecting, and we can do this with a function called **Pupil function**

Lens Pupil function (for circularly symmetric optical systems)

$$P(f,g) = \begin{cases} 1 & \text{if } \sqrt{f^2+g^2} = \nu \leq \frac{NA}{\lambda} = \nu_c \\ 0 & \text{otherwise} \end{cases}$$

pupil function, tells us where the lens ends

$\frac{NA}{\lambda}$: function of NA and λ

ν_c: critical cut-off frequency

so now our electric field at the water becomes

$$E(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} df\,dg \; P(f,g)\,\tilde{O}(f,g)\, e^{-2\pi i (fx+gy)}$$

NOTE: the lens is acting like a low-pass filter, cutting off frequencies higher than ν_c.

NOTE: we didn't once see time / frequency in time, always space and spatial frequency

so what the pupil function ultimately does is change the limits of the integrals:

$$E(x,y) = \iint_{\nu < \frac{NA}{\lambda}} df\,dg \; \tilde{O}(f,g)\, e^{-2\pi i (fx+gy)}$$

and from this we can get the intensity of the light at water level of the aerial image

$$I(x,y) = |E(x,y)|^2$$

Lens Performs Inverse Fourier Transform
$\mathcal{E}(x,y) = F^{-1}\{\mathcal{E}(f_x,f_y)P(f_x,f_y)\}$

Light Intensity
$I_i(x,y) = |\mathcal{E}(x,y)|^2$

Mask Transmittance
$t(x_1,y_1)$

Far Field Fraunhofer Diffraction Pattern
$\mathcal{E}(f_x,f_y) = F\{t(x_1,y_1)\}$

Pupil Function
$P(f_x,f_y)$

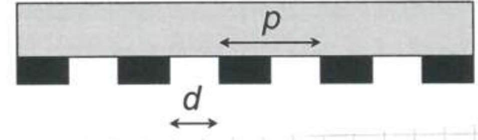
RESOLUTION LIMIT OF COHERENT IMAGING

We can rewrite the light intensity at the focal plane by normalizing our coordinates (let's consider 1D):

$$\hat{x} = x \frac{NA}{\lambda} \qquad \hat{f} = f \frac{\lambda}{NA}$$

$$\hat{I}(\hat{x}) = \left| \int_{-1}^{+1} \tilde{\hat{O}}(\hat{f})\, e^{-2\pi i \hat{f}\hat{x}}\, d\hat{f} \right|$$

components of the spectrum with frequency lower than $-1 \left(= -\frac{NA}{\lambda}\right)$ and greater than $+1 \left(= \frac{NA}{\lambda}\right)$ are cut off (low pass pupil filter) while the others are reconstructed to form the image.



Another way of defining our mask transmission function (as a function of the pitch, or period, $\hat{p}$ normalized; and $\hat{d}$, the aperture of the mask $\hat{d} = \hat{p}/2$) NOT TRUE IN GENERAL

$$\hat{d} = d \frac{NA}{\lambda} \qquad \hat{p} = p \frac{NA}{\lambda}$$

(sketch: d, p)

mask transmission function

$$\hat{O}(\hat{x}) = \begin{cases} 1 & \text{if } |\hat{x} - n\hat{p}| \le \frac{\hat{d}}{2}, \quad n \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases}$$

$\Downarrow \mathcal{F}$

mask spectrum (Fourier transform)

$$\tilde{\hat{O}}(\hat{f}) = \frac{\hat{d}}{\hat{p}} \frac{\sin(\pi \hat{f} \hat{d})}{\pi \hat{f} \hat{d}} \sum_{n=-\infty}^{+\infty} \delta\left(\hat{f} - \frac{n}{\hat{p}}\right)$$

(sinc) (sum of dirac deltas)

NOTE: mask spectrum is discrete rather than continuous. The frequencies are spaced with a $\frac{1}{\hat{p}}$ separation.

$|E|$ goes like a cardinal sine, intensity $I = |E|^2$

cardinal sine $\quad \text{Sinc}(x) = \dfrac{\sin(x)}{x}$

$$\tilde{\hat{O}}(\hat{f}) = \frac{\hat{d}}{\hat{p}} \frac{\sin(\pi \hat{f} \hat{d})}{\pi \hat{f} \hat{d}} \sum_{n=-\infty}^{+\infty} \delta\left(\hat{f} - \frac{n}{\hat{p}}\right) = \frac{\hat{d}}{\hat{p}} \, \text{sinc}(\pi \hat{f} \hat{d}) \sum_{-\infty}^{+\infty} \delta\left(\hat{f} - \frac{n}{\hat{p}}\right)$$

E field :

0th order of diffraction
1st order of diffraction
2nd order of diffraction
Sinc Envelope
Dirac Functions
O(f,d/p=1.25/3.2)
f(NA/lamba)

NOTE: the intensity is modulated by the sinc^2 function!

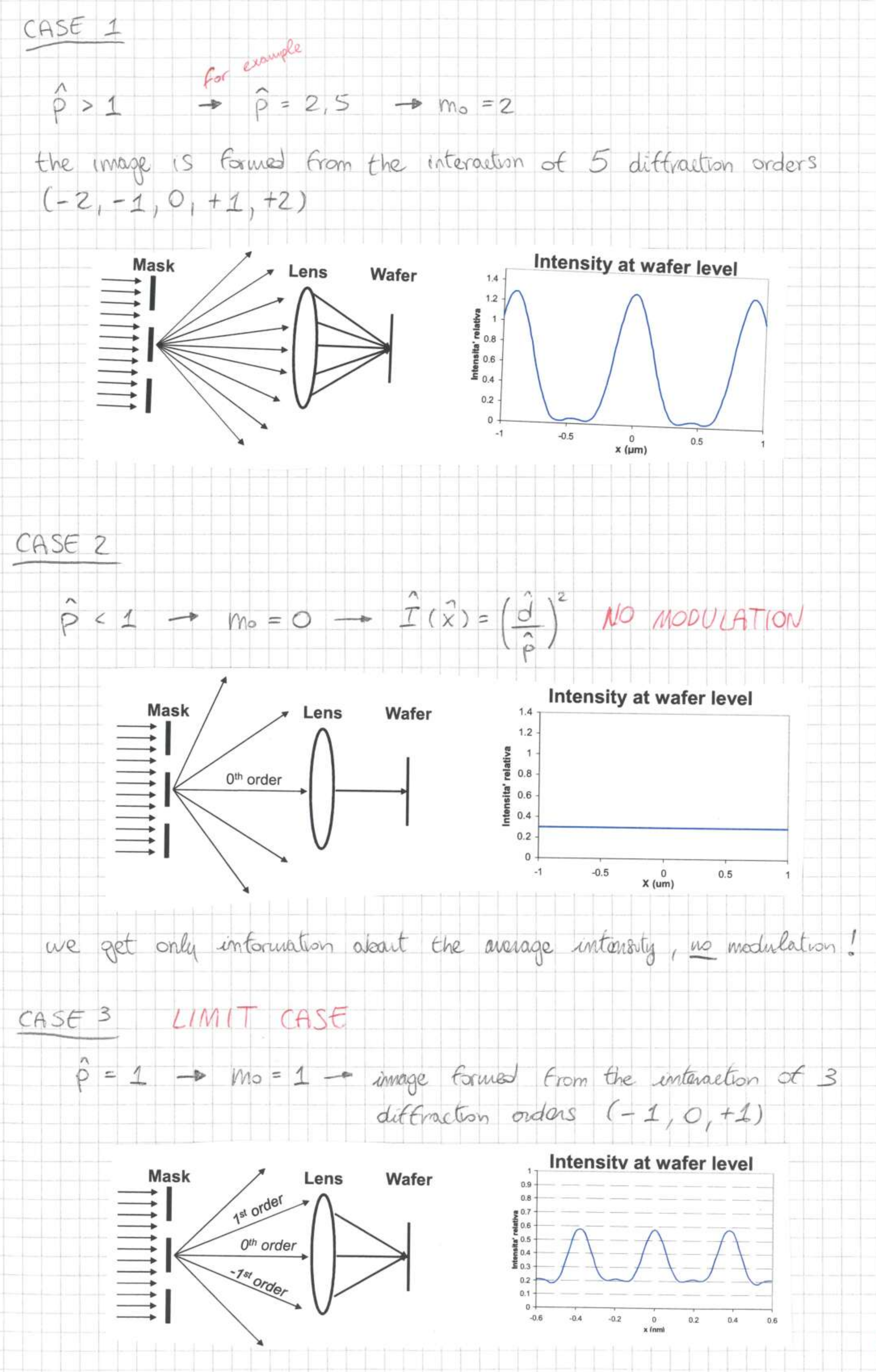
so inserting $\tilde{\hat{O}}(\hat{f})$ inside the formula of the intensity at wafer level, we get:

$$\hat{I}(\hat{x}) = \left(\frac{\hat{d}}{\hat{p}}\right)^2 \left| 1 - \sum_{m=1}^{m_0} 2 \, \text{sinc}\left(\frac{m\hat{d}}{\hat{p}}\right) \sin\left(\frac{2\pi m \hat{x}}{\hat{p}}\right) \right|^2$$

where $[m_0 \leq \hat{p} \leq m_0 + 1]$ $\quad$ m_0 tells us how many diffraction orders we have to consider

Let's see some examples.

NOTE: $\hat{I}(\hat{x})$ is now a continuous function of position $\hat{x}$


CASE 1
for example
$\hat{p} > 1 \rightarrow \hat{p} = 2,5 \rightarrow m_0 = 2$
the image is formed from the interaction of 5 diffraction orders
(-2, -1, 0, +1, +2)
Mask
Lens
Wafer
Intensity at wafer level
Intensita' relativa
x (µm)
CASE 2
$\hat{p} < 1 \rightarrow m_0 = 0 \rightarrow \hat{I}(\hat{x}) = \left(\frac{\hat{d}}{\hat{p}}\right)^2$
NO MODULATION
Mask
Lens
Wafer
0th order
Intensity at wafer level
Intensita' relativa
X (um)
we get only information about the average intensity, no modulation!
CASE 3
LIMIT CASE
$\hat{p} = 1 \rightarrow m_0 = 1 \rightarrow$ image formed from the interaction of 3
diffraction orders (-1, 0, +1)
Mask
Lens
Wafer
1st order
0th order
-1st order
Intensity at wafer level
Intensita' relativa
x (nm)
226

So we can see that the theoretical resolution limit is depending only on the period p and not on the opening width d.

$$\hat{p}_{min} = 1 \longrightarrow p_{min} = \frac{\lambda}{NA}$$

So $\hat{p}_{min} = 1$ is the minimum pitch at which we start having some modulation of intensity.

limit is on pitch

NOTE: since the resolution limit is on the pitch (how close together things are) and not on d, we can print single features with no resolution limit → a mask with 1 and only 1 opening has no resolution limit, since we will always have some modulation of light (then it's job of the photoresist to have the sensitivity / high contrast needed for it).

We can usually approximate the dimension of the features we want to print as half of the pitch, so the minimum resolution will be

minimum feature we can print

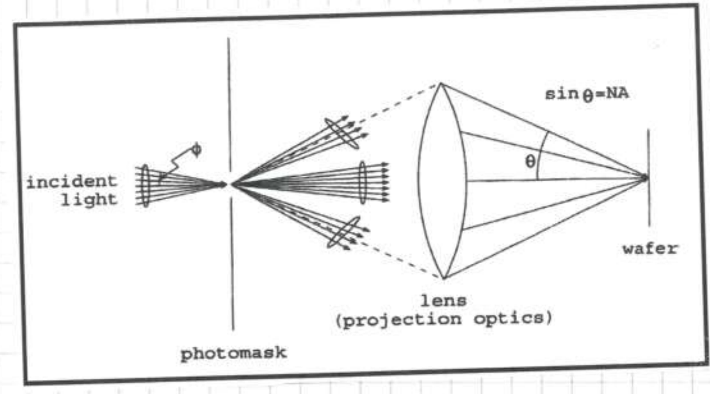
$$\left[R \cong \frac{p_{min}}{2} = \frac{\lambda}{2NA} \right] \quad !$$

(pitch)

So when we hear that a lithographic tool has a resolution limit of, say, 40 nm, it means that the minimum pitch it can print is 80 nm. So we can print even 3nm features BUT the period at which they appear can't be lower than 80 nm.

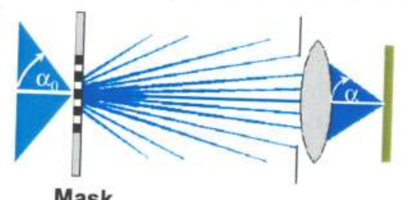
NOTE: now you can go back to page 219 and see how the pitch becoming smaller results in diffraction orders being more spread, so further apart, until in picture (4) the -1st, $+1^{st}$ diffraction orders are at the edges of the pupil and won't be collected, giving us no intensity modulation. Also check the picture of the pitch at page 218.

PARTIALLY COHERENT IMAGING



partial coherence

$0 < \sigma < 1$

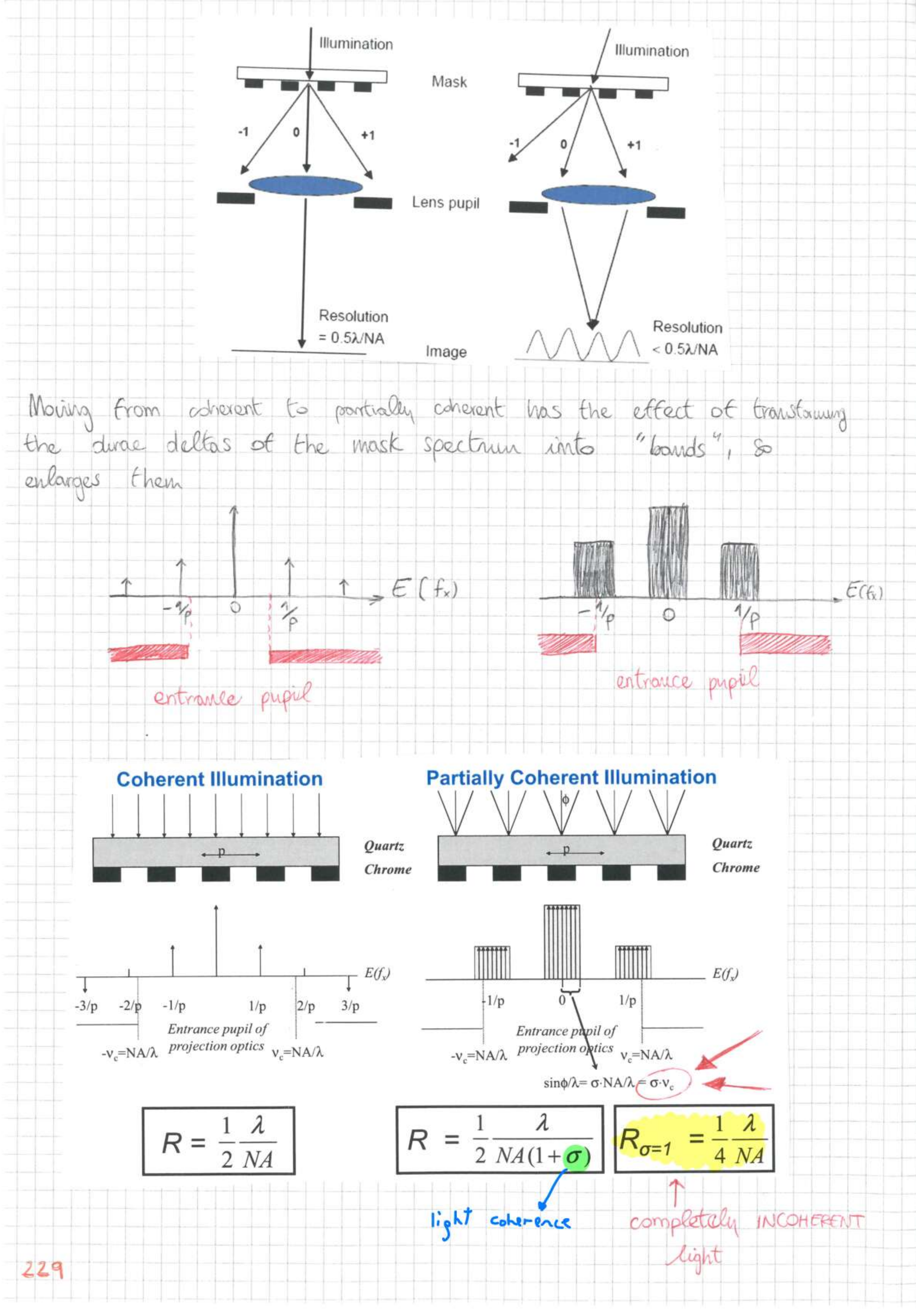


$$\sigma = \frac{\sin\alpha_0}{NA} = \frac{\sin\alpha_0}{n \cdot \sin\alpha}$$

We can have light coming from a variety of angles (for coherent light $\alpha_0 = 0$ the angle the light formed with the ⊥ of the surface).

What's the advantage of partially coherent light? (= off axis illumination)

Let's imagine we are just beyond our resolution limit with our COHERENT light, so the first orders of diffraction can't enter into the lens, so we don't have modulation.
Now if we tilt a little bit our illumination source (→ off axis illumination / partially coherent light) also the diffraction pattern will tilt, so the -1st order of diffraction will be deviated away from the lens, but we will collect the 0th and $+1^{st}$ diffraction orders, having some intensity modulation, effectively decreasing R and being able to print smaller things. (see picture next page)


Illumination
Mask
-1
0
+1
Lens pupil
Resolution
= 0.5λ/NA
Image
Illumination
-1
0
+1
Resolution
< 0.5λ/NA
Moving from coherent to partially coherent has the effect of transforming the dirac deltas of the mask spectrum into "bands", so enlarges them
E(fx)
-1/p
0
1/p
entrance pupil
E(fx)
-1/p
0
1/p
entrance pupil
Coherent Illumination
Partially Coherent Illumination
Quartz
Chrome
p
ϕ
Quartz
Chrome
p
E(fx)
-3/p
-2/p
-1/p
1/p
2/p
3/p
Entrance pupil of projection optics
-νc=NA/λ
νc=NA/λ
E(fx)
-1/p
0
1/p
Entrance pupil of projection optics
-νc=NA/λ
νc=NA/λ
sinϕ/λ= σ·NA/λ = σ·νc
R = 1/2 λ/NA
R = 1/2 λ/(NA(1+σ))
Rσ=1 = 1/4 λ/NA
light coherence
completely INCOHERENT light
229

Another very important factor to take into account is the intensity contrast of my image:

Aerial Intensity at wafer Level

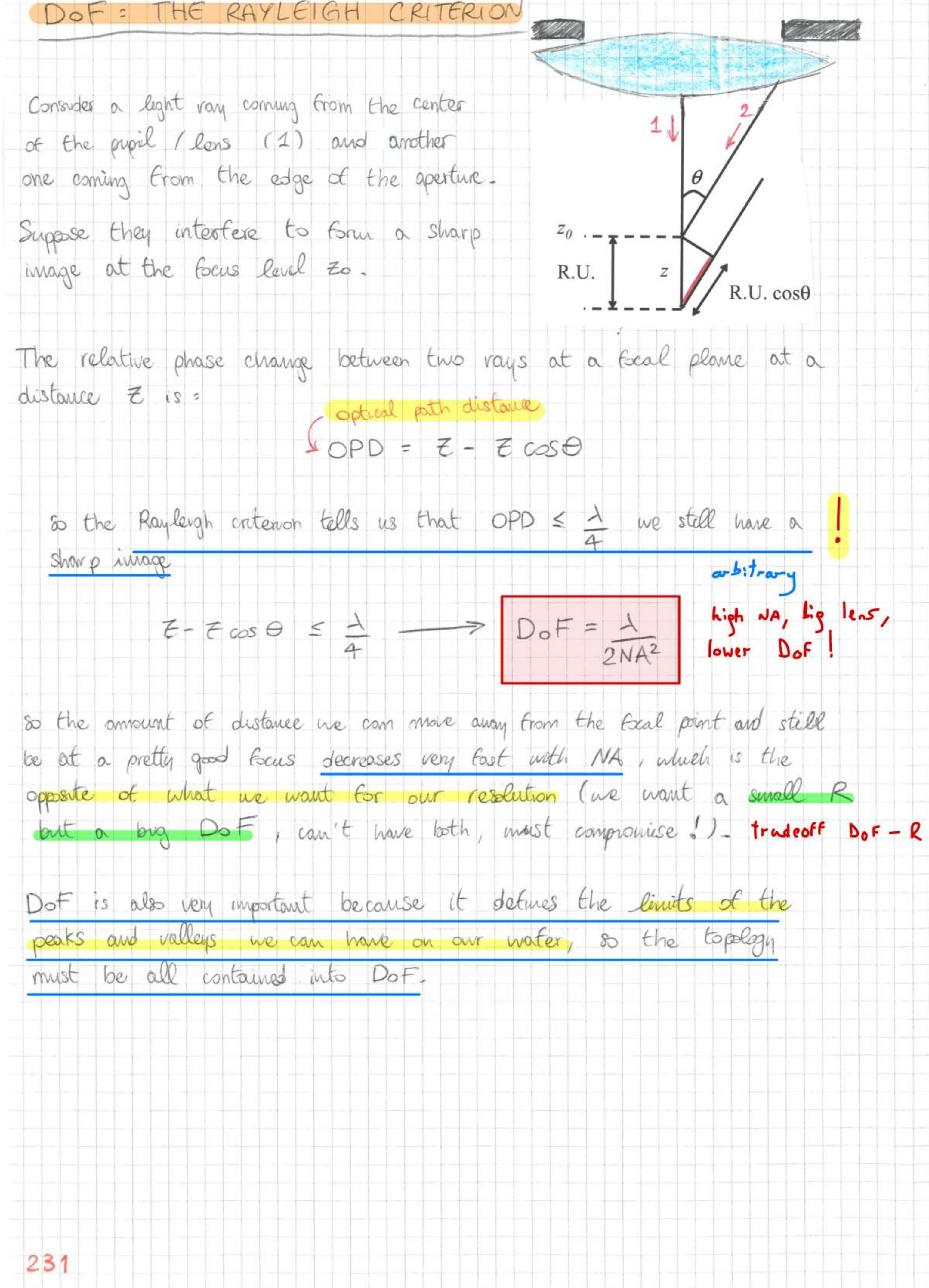
$$\underline{\text{Contrast}} = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \times 100\%$$

NOTE: at which intensity will the photoresist start reacting?

DEPTH OF FOCUS (DoF)

The definition of DoF is kinda qualitative: maximum amount of focus z that can be tolerated before the printed pattern size falls outside the specifications. !

In principle we want the wafer to sit exactly on the focal plane defined by the lens, but the best we can do is stay within some values, defined for example by the Rayleigh criterion.



DoF: THE RAYLEIGH CRITERION

Consider a light ray coming from the center of the pupil / lens (1) and another one coming from the edge of the aperture.

Suppose they interfere to form a sharp image at the focus level z_0.

The relative phase change between two rays at a focal plane at a distance z is:

optical path distance

$$OPD = z - z\cos\theta$$

so the Rayleigh criterion tells us that $OPD \leq \frac{\lambda}{4}$ we still have a sharp image !

arbitrary

$$z - z\cos\theta \leq \frac{\lambda}{4} \longrightarrow DoF = \frac{\lambda}{2NA^2}$$

high NA, big lens, lower DoF !

so the amount of distance we can move away from the focal point and still be at a pretty good focus decreases very fast with NA, which is the opposite of what we want for our resolution (we want a small R but a big DoF, can't have both, must compromise !). tradeoff DoF − R

DoF is also very important because it defines the limits of the peaks and valleys we can have on our wafer, so the topology must be all contained into DoF.

Ideal Optics | Real systems (+ resist response)

resolution

$$R = \frac{1}{2}\frac{\lambda}{NA(1+\sigma)} \qquad R = K_1 \frac{\lambda}{NA}$$

depth of focus

$$DoF = \frac{1}{2}\frac{\lambda}{NA^2} \qquad DoF = K_2 \frac{\lambda}{NA^2}$$

R ↓ then DoF ↑

For real systems, we cram every correction to $\frac{\lambda}{NA}$ into K_1 and every correction to $\frac{\lambda}{NA^2}$ into K_2.

- K_1 = "lithography aggressiveness". Process coefficient that depends on resist, lens, exposure tool. Theoretical limit is 0,25.
- K_2 = Process coefficient that depends on feature shape, coherence, aberrations, ...

NUMERICAL APERTURE AND WAVELENGTH EVOLUTION

1970s: g-line Hg, $\lambda = 436$ nm → 1984: i-line Hg, $\lambda = 365$ nm → 1989: KrF laser, $\lambda = 248$ nm → 1999: ArF laser, $\lambda = 193$ nm → 2020: EUVL, $\lambda = 13.4$ nm

$$R = K_1 \frac{\lambda \uparrow}{NA \downarrow}$$

1980s: NA = 0.5 → 1995: NA = 0.63 → 2001: NA = 0.75 → 2004: NA = 0.93 → 2006: NA > 1 (immersion)

RESOLUTION ENHANCEMENT TECHNIQUES (RET)

$$R = K_1 \frac{\lambda}{NA}$$

decreasing K_1 is a way to increase resolution

By applying a variety of manipulations to the optical wavefront, additional contrast in the high spatial frequency (corners) components has been obtained, with a significant enhancement of the practical resolution.

OAI • Off axis illumination (control of wavefront direction)

PSM • Phase shift mask (control of wavefront phase)

OPC • Optical proximity corrections (control of wavefront amplitude)

SRAF

OFF AXIS ILLUMINATION (OAI)

Based on the shapes of the critical features we want to print (CPU, Flash, DRAM, ...) [and their specific application] we customize the illumination shape in order to maximize contrast and minimize DC background light.

Annular

Cross-Pole

Dipole

Different illumination Schemes for critical Lithographic layers

Measured Intensity Distribution (a.u) of Dipole Illumination

Remember that the image contrast is the difference between I_{max} and I_{min} (normalized, percentage), while the photoresist contrast is the slope of the part between D_0 and D_f (D_0 = dose at which the exposure first begins to have an effect, D_f = dose at which the exposure is complete)

% of resist remaining

1

0

D_0

D_f

Exposure dose

$$\gamma = \frac{1}{\log_{10}\left(\frac{D_f}{D_0}\right)}$$

PHOTORESIST CONTRAST

$$\text{Light Contrast} = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \times 100\%$$

Intensity

I_{max}

I_{min}

position

Changing the illumination can improve the amount of diffraction orders that we can collect (PAGE 229) and there are also simulators that let us choose the best shape of illuminator based on shape and application (do we want to draw lines, holes, randomly oriented weird stuff?)

ATTENUATED PHASE SHIFT MASKS (Att - PSM)

Conventional Binary Mask

Quartz

Chrome

p

Transmission function $f(x)$

1

0

"Attenuated" Phase Shift Mask

Quartz

Shifter semi-transparent with 6% transmission

180° 0°

1

$-\sqrt{T} = -t$

Current T = 6%

Once we have fixed our NA and λ, to increase our resolution we can change our mask technology to get close to $K_1 \simeq 0,25$ (remember: $K_{1\,MAX} = 0,25$).

Instead of using binary masks (quartz 100% transmission, chrome 0% trans.) we can use an attenuated phase shift mask (quartz 100% transmission, semi-transparent material with ~6% transmission which also phase-shifts the light by 180°).

The net effect of using an Att - PSM is to change the relative amplitude of the peaks in the interference spectrum.

BINARY MASK

Intensity

-1   0   1

$-1/p$ 0 $1/p$ $\tilde{O}_x(f)$

ATTENUATED PHASE SHIFT MASK

-1   0   1

$-1/p$ 0 $1/p$ $\tilde{O}_x(f)$

The central peak carries no information about amplitude modulations, and so by having relatively higher ±1 orders we end up with more contrast.

BINARY MASK

Chrome

Quartz

e

Resist Threshold

i

Silicon

Binary

Bright (+)   Dark (0)   Bright (-)

$i \propto e^2$

Att - PSM

Phase Shifting Material

Quartz

e

Partially Transmitting

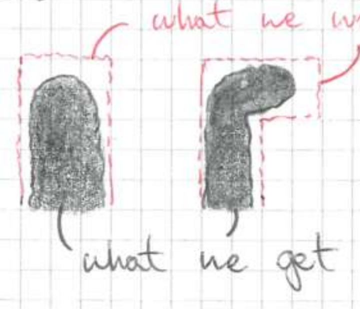
i

Silicon

EAPSM

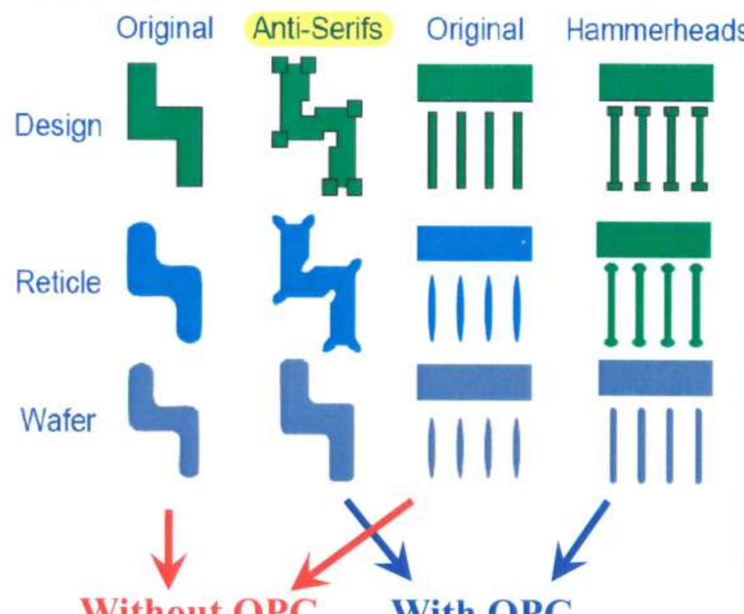
OPTICAL PROXIMITY CORRECTIONS (OPC)

Due to the current λ and feature sizes, the exposed pattern will result distorted (picture: left) with shorter lines and rounder corners.

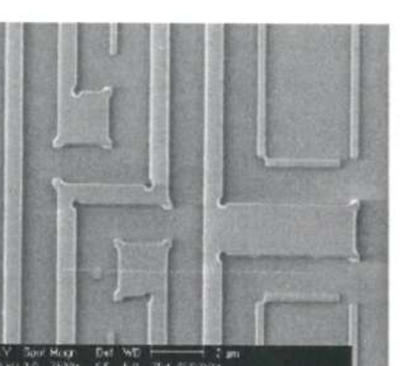


Mask (without optical proximity correction)
Light beam
Mask (with optical proximity correction)
Focusing lens
Wafer
Desired circuit pattern
Exposed circuit pattern

In order to solve this, we employ optical proximity corrections, that are sub resolution changes in the shape of the pattern on the mask, in order to counter the effects and get longer lines and sharper corners (picture: right)



Top-Down SEM Picture of a Lithographic reticle with OPC Correction to compensate optical proximity near the resolution liimit



We're reverse-engineering the shape of the mask to take into account the interference effects that will produce the shape that we want in the wafer. So by adding many more corners, those additional corners will be lost and we'll get the shape we were looking for.

SUB RESOLUTION ASSIST FEATURES (SRAF)

Let's suppose we want to print the following:

MASK WITHOUT SRAF

1 2 3 4 5 6 7 8

this would be very bad for our lithography since the pattern is terminating abruptly, some space is present, then the pattern starts again.
Especially the lines number 4 and 5 would suffer a lot (will be shorter, smaller, deformed, ...).

We can compensate that by adding to our mask some lines between 4 and 5 but we'll print them below our resolution limit, to make sure that they won't appear into the photoresist but can still contribute to the interference pattern, since now the pattern will be periodic with no interruptions.

MASK WITH SRAF

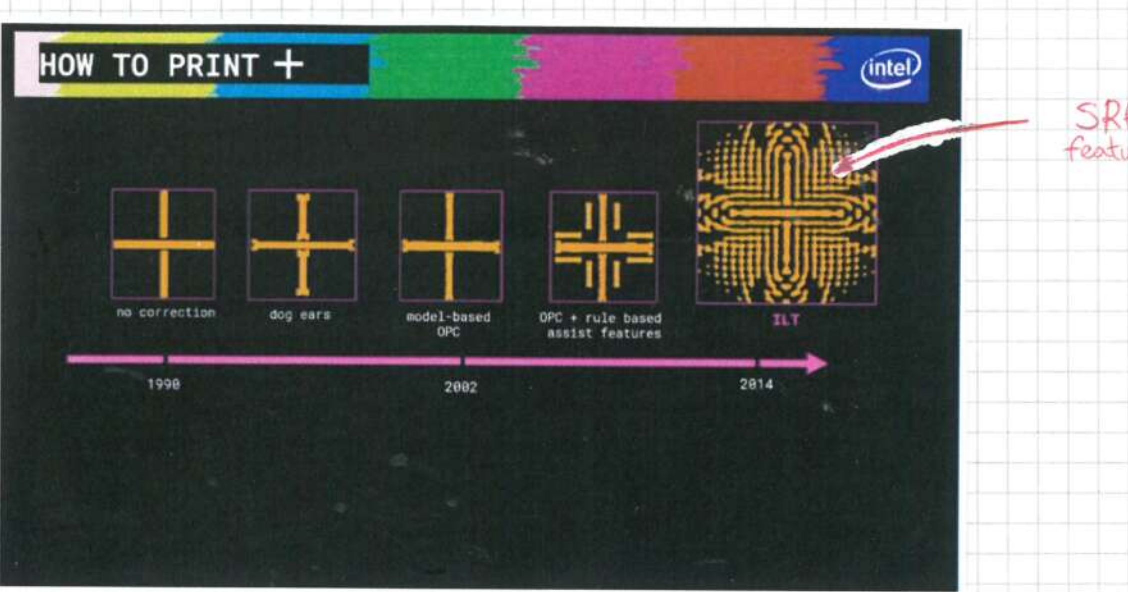
1 2 3 4 5 6 7 8

SRAFs

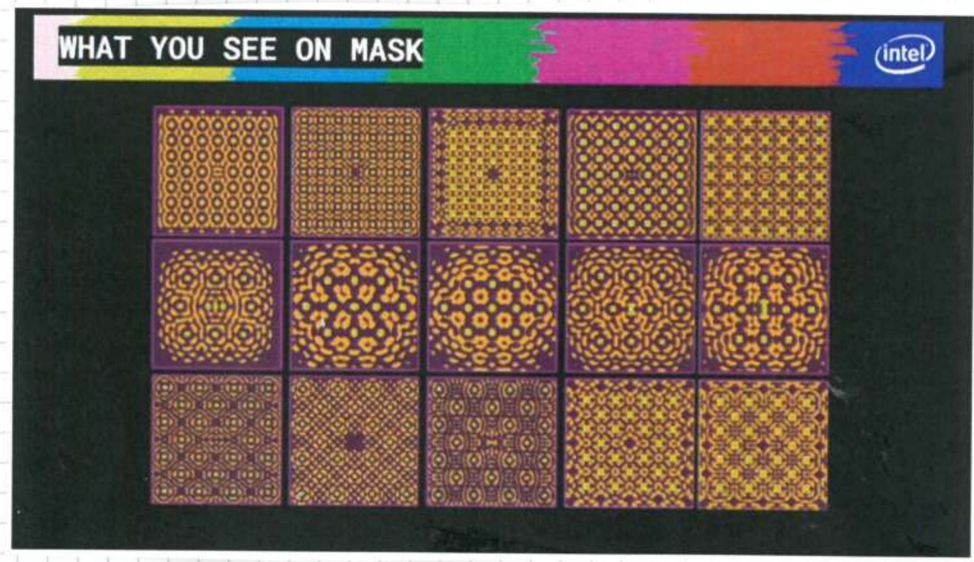
WHAT WE GET ON THE WAFER:

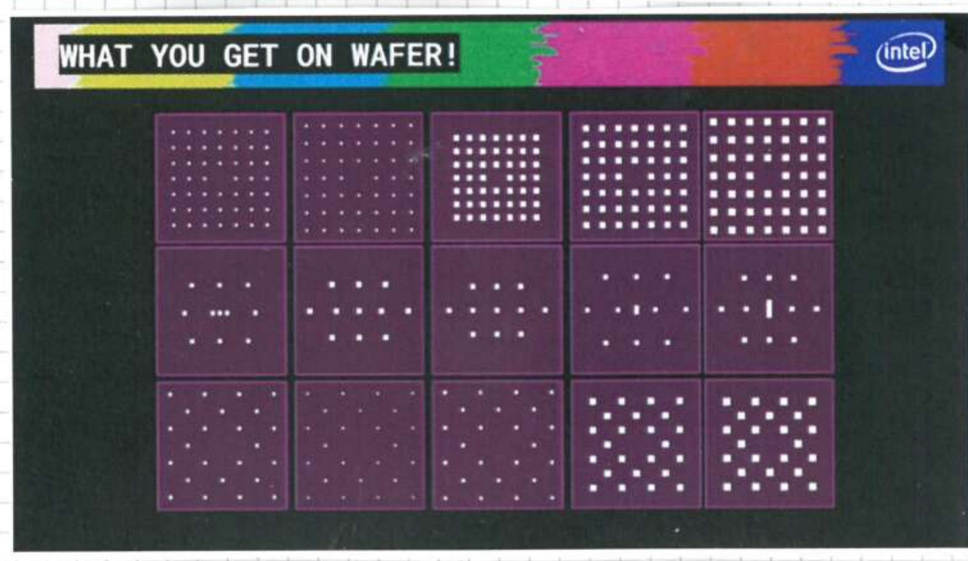
WITHOUT SRAF

WITH SRAF



Some more SRAFs, OPC and ILT (Inverse Lithography Technology, where you use a simulator to reverse-engineer the shape of the mask to get what you want on the wafer):





Lecture 20

7 maggio

IMMERSION LITHOGRAPHY

R ↓ since $\frac{\lambda}{NA} = \frac{\lambda}{n \sin\theta}$ ↓, also D.F ↑

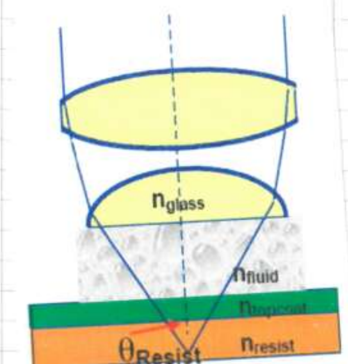
Around 1999 came out the ArF laser ($\lambda = 193$ nm) lithographic tools, and up until around 2020 (when EUVL $\lambda = 13,4$ nm came out) we have been able to keep up with Moore's law without needing to change wavelength.

One of the "tricks" used to get better resolution is immersion lithography, where we change the medium between the lens and the resist.

$$NA = n \sin\theta$$

$$R = K_1 \frac{\lambda}{NA} = K_1 \frac{\lambda}{n \sin\theta} = K_1 \frac{\lambda / n}{\sin\theta}$$

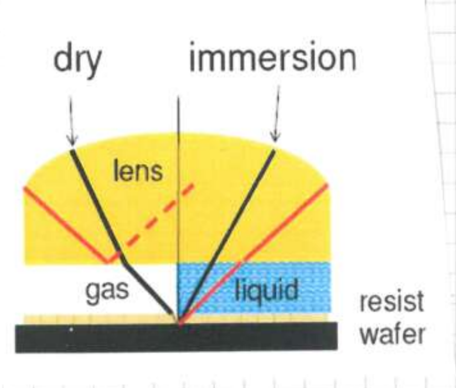
(λ / n: effective wavelength)



If instead of air ($n = 1$) we use a liquid (water $n = 1,34$), we effectively get a better resolution, without changing the laser or the photoresist.

State of the art lithographic tools are immersion scanners (not considering the EUVL tools, not very diffused yet).

Also, despite NA is increasing, immersion lithography has a better depth of focus (DoF) than dry lithography.
That's because by changing the refractive index n, we change the angle at which light rays diffract.

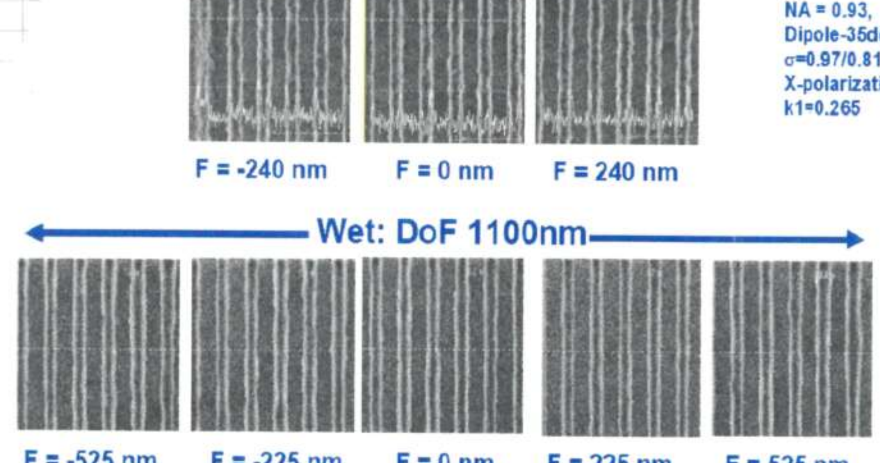


$$n_{lens} \sin\theta_{lens} = n_{medium} \sin\theta_{medium}$$

$$\sin\theta_{medium} = \sin\theta_{lens} \frac{n_{lens}}{n_{medium}}$$

$$\begin{cases} n_{lens} = 1,57 \text{ for quartz} \\ n_{medium} = 1 \text{ (air) or } 1,34 \text{ (water)} \end{cases}$$

So for a given diffraction order (and thus a given angle of the light inside the resist), the angle of the light inside an immersion fluid will be less than if air is used.
Smaller angles means smaller optical path difference, and so smaller degradation of the image for a given amount of defocus.



Top Down SEM Picture through focus of 55nm Lines and Spaces, with Dry and Wet Tool.

STATE OF THE ART LITHOGRAPHY

If we take the best tool in the immersion lithography market, we push every limit of resolution enhancement techniques (PSM, OPC, OAI, ILT, ...) the minimum pitch we can print is ~ 80 nm.

NOTE: we said minimum "pitch", not "feature" because there's no limit on a single feature, only on pitch (repeating patterns).

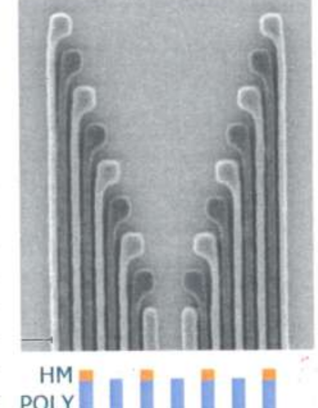
So 80 nm pitch will mean a 40 nm technology ($F = 40$ nm).

But if we take a look on the market, we can find a planar NAND by Micron of $F = 18$ nm ($\Rightarrow$ pitch $= 36$ nm).

How did they do it? Let's see some examples ...

DOUBLE PATTERNING

In double patterning you repeat the pattern twice, overcoming the limit of $k_1 = 0,25$



↑ we can use two masks
the first will print the grey lines
the second the black lines

Obviously the biggest limitation is the overlay, so how well aligned will the second mask be compared to the first one.

SELF-ALIGNED DOUBLE PATTERNING

40nm
20nm

For normal double patterning the problem of overlay is too big to ignore, since with state of the art alignment we can do 2nm, with a 36 nm (from double patterning) pitch we'll end up with ~18 nm space between the lines, which can vary in the range 16 ~ 20nm. NOT GOOD

So we need a self aligned technique.

POSITIVE APPROACH

(1) Print and Resist Trim (2) Etch Template (3) Form Spacers (4) Strip Template (5) Transfer Etch (6) STI Etch and ash

Oxide (or Nitride), APF, Nitride (or a-Si), APF, Silicon

Figure 5: Illustration of process flow for generating 22nm STI array.

CVD Spacer on APF template → Spacer Etch and APF Strip-Out → APF Hardmask Etch → APF Hardmask Top View → STI Etch and Ash

22nm Line and Space

photoresist

Template
Silicon

On top of our wafer we deposit a template (usually amorphous Carbon) and then on top of it we do standard lithography with a pitch double of what we want.

NOT photosensitive

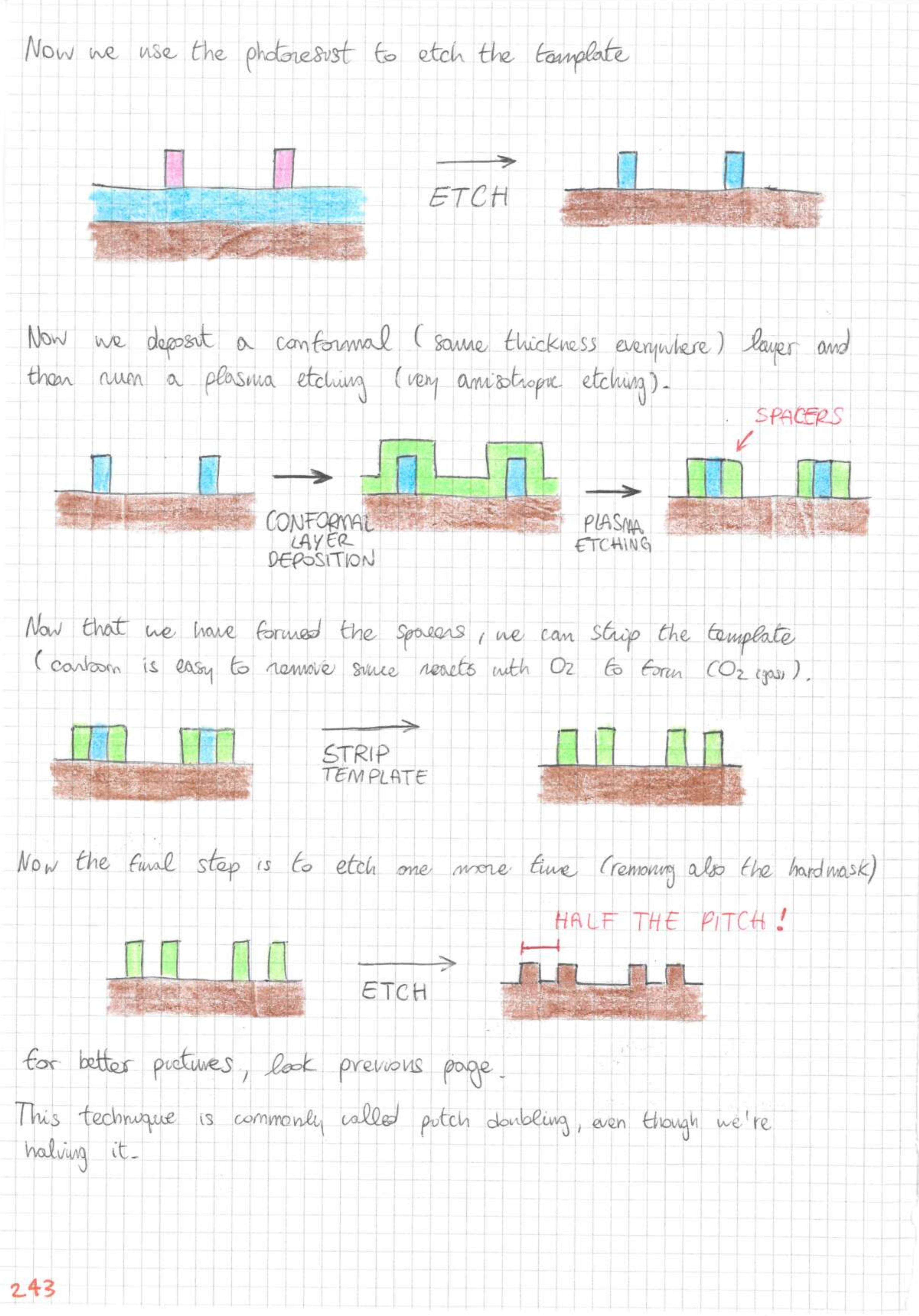
Then we shrink the photoresist lines to 1/3 their original thickness

the pitch doesn't change

TRIM PHOTORESIST

It's easy to change the thickness of the photoresist (e.g. increase PEB time if we're using CAR resist, we can use an oxygen plasma to trim the dimensions, ...)

post exposure bd


Now we use the photoresist to etch the template
ETCH
Now we deposit a conformal (same thickness everywhere) layer and
than run a plasma etching (very anisotropic etching).
SPACERS
CONFORMAL
LAYER
DEPOSITION
PLASMA
ETCHING
Now that we have formed the spacers, we can strip the template
(carbon is easy to remove since reacts with O2 to form CO2 (gas)).
STRIP
TEMPLATE
Now the final step is to etch one more time (removing also the hardmask)
HALF THE PITCH!
ETCH
for better pictures, look previous page.
This technique is commonly called pitch doubling, even though we're
halving it.
243

NOTE: Why did we put the template layer, and not print directly everything on the wafer using just the photoresist?

That's because the PR is very Temperature sensitive, so if we put a wafer with a PR on top in a tool at 200°C we'll contaminate our tool with evaporated PR.

The problem is that we want very conformal thin film, which usually needs high T (to promote re-emission → lower sticking coeff.). For that we use ALD (atomic layer deposition), which right now can even be done at ~30°C, but it's an extreme case.

In the picture at page 242 we can see that instead of a single template layer, we have two APF (Advanced Patterning Film, amorphous Carbon, which are two hardmasks) that are there because the photoresist alone isn't thick enough to withstand the whole etching process, so we need to add thickness, how? By adding a template + oxide (or nitride) layer.

Since usually both PR (photo resist) and template (or APF) are based on carbon, we can't etch one without also etching the other, so when we saw

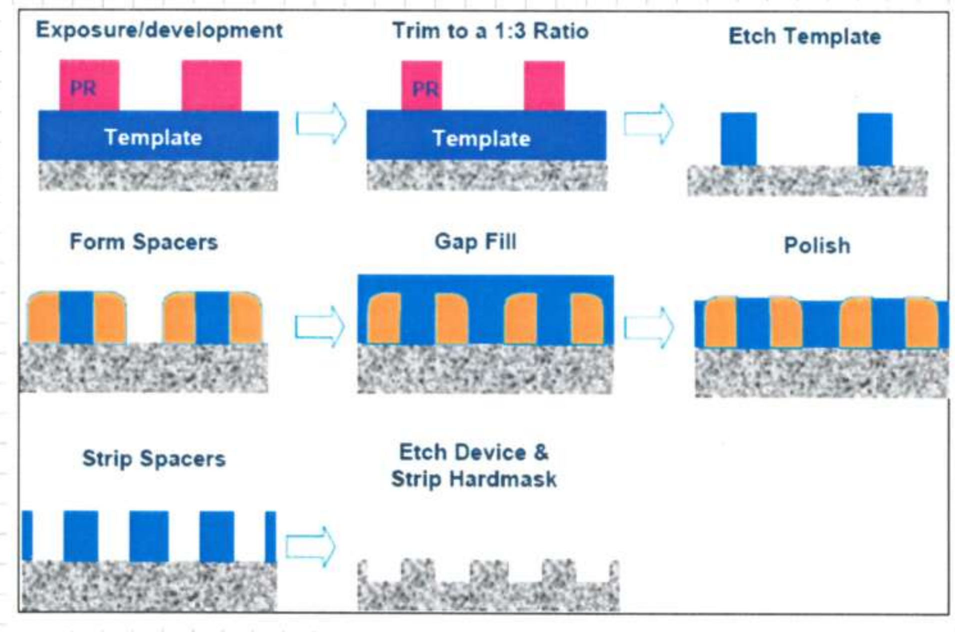
what we expect ← ETCH — what we get → ETCH

that's why we need a thin protective film (oxide or nitride) on top of APF:

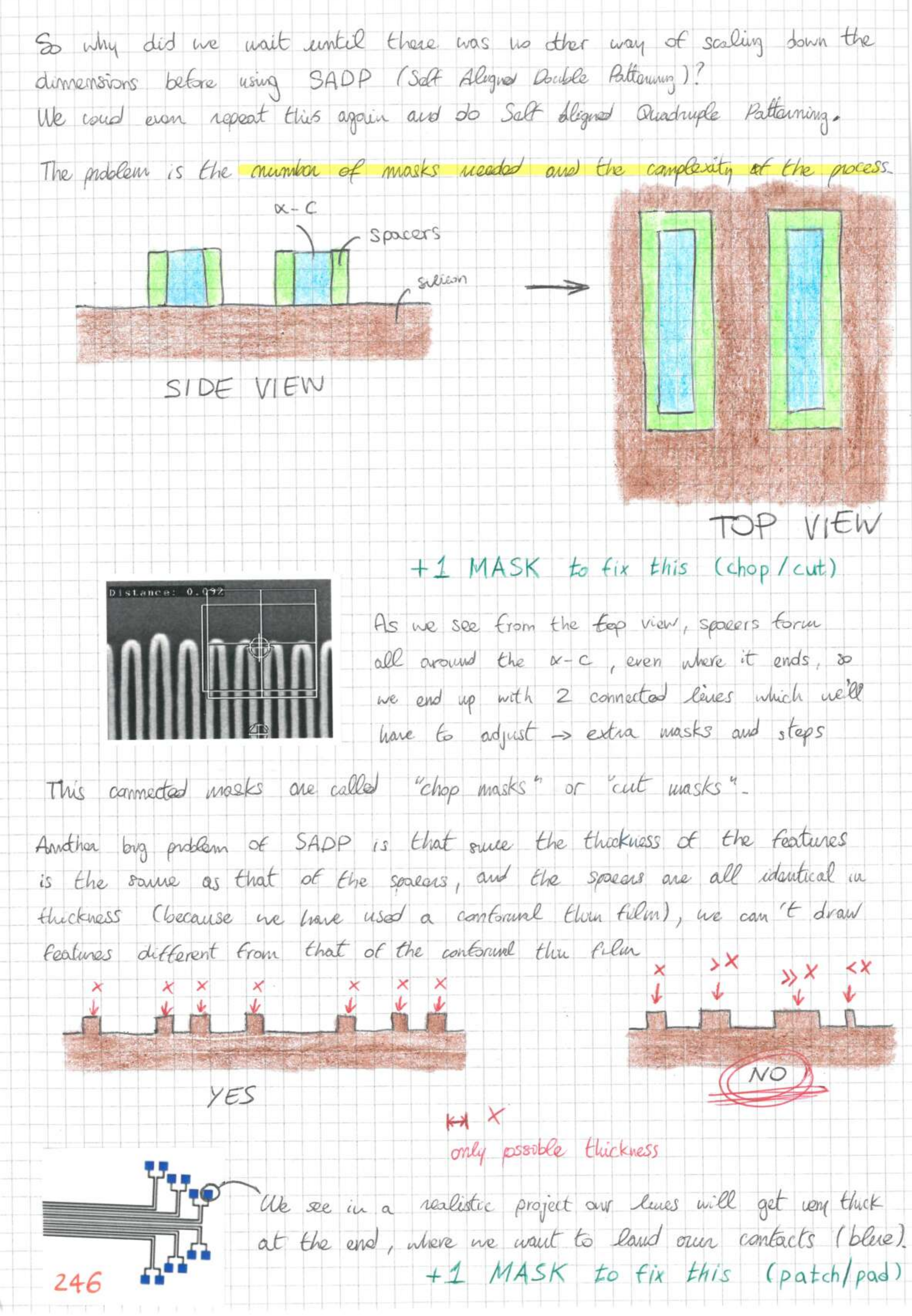
ETCH

NOTE: in the picture at page 242 we see 2 oxide - APF layers (APF - oxide + APF - oxide), why?
Because as before, the spacers also can't withstand the etching process until the end, since we can't be selective enough to them. So we would remove them before even indenting the wafer, so we put under the spacers another APF - oxide layer, so that during the etching process, when we remove the spacers, we etch the thin oxide layer, and THEN we can etch being very selective to the oxide, so we remove only the APF (α-C, very easy to remove with O_2 plasma).

NEGATIVE APPROACH



Very similar to the positive approach, but after the spacers formation we fill everything with the template material (α-C amorphous carbon) then planarize with CMP (Chemical Mechanical Polishing). Now instead of removing the α-C (positive approach) we remove the spacers (negative approach), then we etch the device, removing also the hardmask (α-C).



So why did we wait until there was no other way of scaling down the dimensions before using SADP (Self Aligned Double Patterning)?
We could even repeat this again and do Self Aligned Quadruple Patterning.

The problem is the number of masks needed and the complexity of the process.

+1 MASK to fix this (chop/cut)

As we see from the top view, spacers form all around the α-C, even where it ends, so we end up with 2 connected lines which we'll have to adjust → extra masks and steps

This connected masks are called "chop masks" or "cut masks".

Another big problem of SADP is that since the thickness of the features is the same as that of the spacers, and the spacers are all identical in thickness (because we have used a conformal thin film), we can't draw features different from that of the conformal thin film

↔ x
only possible thickness

We see in a realistic project our lines will get very thick at the end, where we want to land our contacts (blue).
+1 MASK to fix this (patch/pad)

So for our complete SADP process we would need (at least) 3 masks:

- SADP mask, the one that performes the pitch doubling
- chop / cut mask, to cut the connected wires
- patch / pad mask, to land the contacts (wider lines)

The SADP mask usually needs the most advanced lithographic tools since is very close to the resolution limit, while the other two can use n-1, n-2 gen lithographic tools (older tools, cheaper), but might still need very very good overlay precision, even if the masks are bigger.

So it's harder and less cost-effective than it might seem.

IMBALANCED PATTERN

The spacers determine the thickness of (in the picture) the odd spaces (I, III, ...), while the thickness of the SPACES between the spacers is a separate process with it's own error.

SPACES SPACES SPACERS → REMOVING TEMPLATE → SPACERS SPACES I II III IV V

ETCH → I 1 II 2 III 3 4

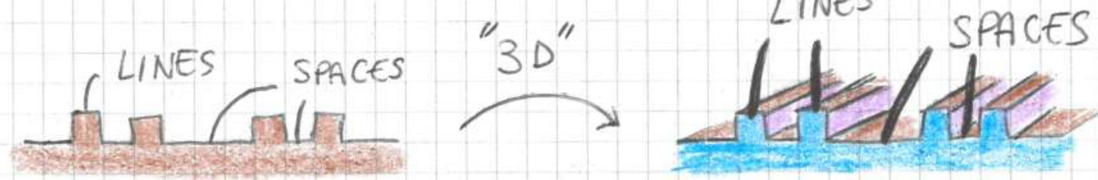
the distance (and so also the error) of spaces and spacers is different

If we have ±1 nm on spacers formation and ±1 nm on PR formation, the space between spacers 2 and 3 (picture) will have $\pm\sqrt{2}$ nm error (since we will add the error of photoresist + spacers, $\delta_{SPACES} = \sqrt{\delta^2_{PR} + \delta^2_{SPACERS}}$), while the spaces between spacers 1 and 2 will be just determined by the photoresist and its error.

NOTE: positive DP and negative DP have different advantages!

Positive double patterning has better control over the dimension of the lines (in the picture: 1, 2, 3, 4, ...) since they come all from the single process of conformal thin film deposition, but poor control over the dimensions of the spaces (I, II, III, ... some ± 1 nm, some $\pm\sqrt{2}$).

Negative DP is the opposite: you have great control over the spaces (since they all come from conformal thin film deposition) but poorer control over the lines.



NOTE: imbalances between odd and even spaces is one of the ways to reverse-engineer the technology used, if is double patterning or normal lithography.

ELECTRON BEAM DIRECT WRITING (EBDW)

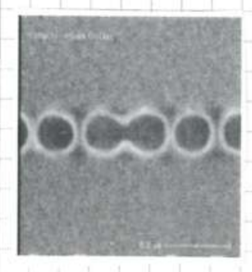
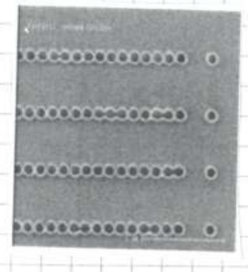
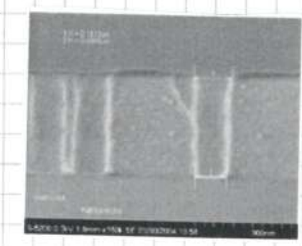
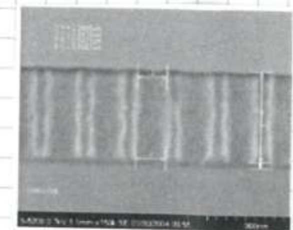
To to beyond the limits of optical lithography we could use EBDW.

ADVANTAGES:

- technology already well known, since it's used for mask manufacturing. Well enstablished
- extremely low wavelength $\lambda \sim 0,1$ nm (low diffraction)
- no need for a mask

DISADVANTAGES:

- coulumb repulsion lowers effective resolution (can't push two electrons too close together)
- high proximity effects: since the e^- hitting the resist can travel $> 1\ \mu m$ before stopping, they can partially expose dark regions
- VERY low throughput → 5 wafers / hour since the wafer is exposed pixel by pixel
  Optical systems ~ 60 wafers / hour

EXTREME UV LITHOGRAPHY (EUVL)

ADVANTAGES:

- employs step and scan printing
- supports RETs (OAI, PSM, OPC, ...)
- $\lambda \sim 13,4$ nm

DISADVANTAGES:

- needs high vacuum (while $\lambda \sim 193$nm was done in air/water)
- extremely expensive
- $\lambda \sim 13,4$nm is absorbed by every material

$\frac{\lambda}{4}$ Bragg reflector

Also we can't use anything other than mirrors (lens? → mirror, transmission masks? → reflective masks, ...).

λ/4 BRAGG REFLECTOR

If we alternate materials with high and low refractive index at the right thickness, we can obtain a reflected beam that interferes constructively with the beam reflected at each layer of the multiple reflections, maximizing reflection.

incident

5 constructive reflected

n_{air}

n_{HIGH}

n_{LOW}

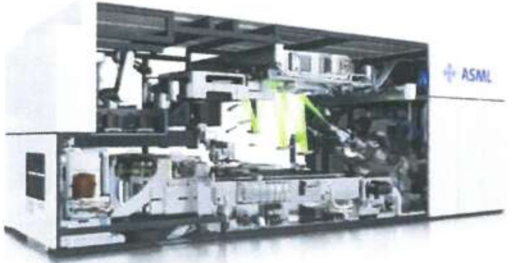
n_{HIGH}

n_{LOW}

$n_{substrate}$

transmitted

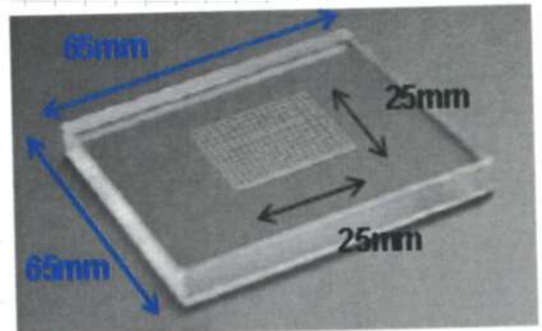
ArFi ($\lambda \sim 193$nm) EUV ($\lambda \sim 13$nm)



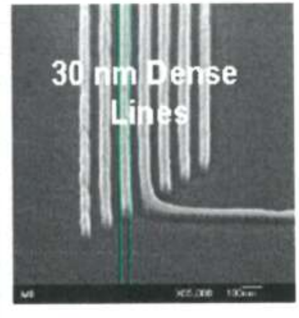
NANO IMPRINT LITHOGRAPHY (= CONTACT PRINTING)

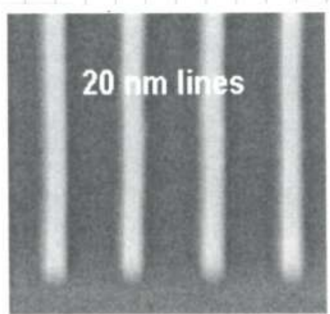
Nano imprint lithography is basically contact printing with a cooler name.

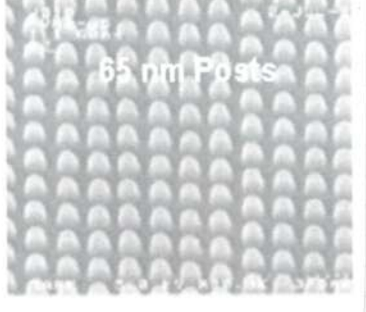
So it also needs a 1 : 1 mask, which is a very demanding requirement. We also have to consider the higher defectivity associated with it.

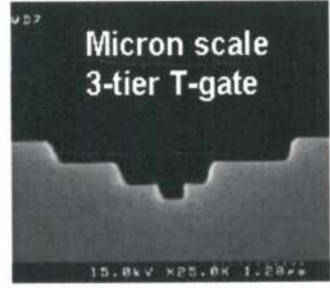


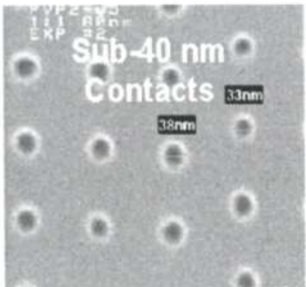
Imprint Templates

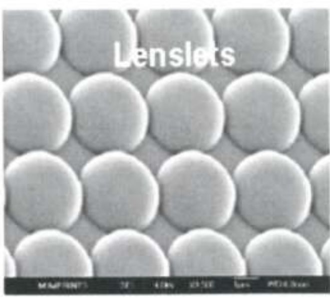












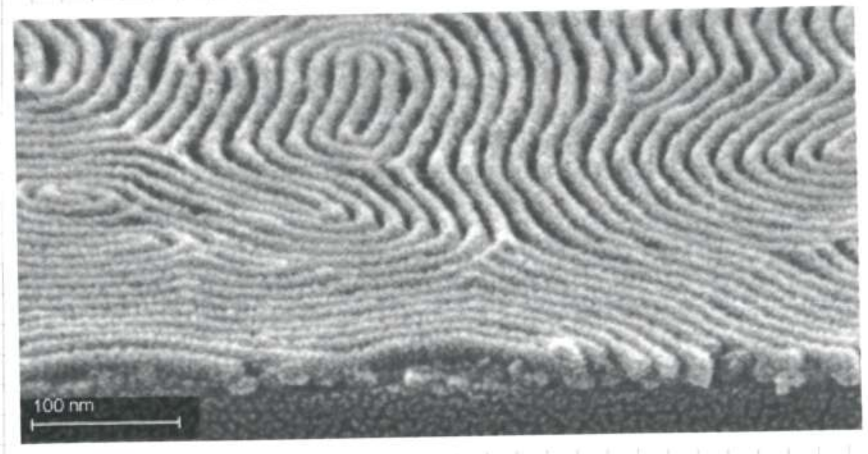
DIRECTED SELF ASSEMBLY (DSA)

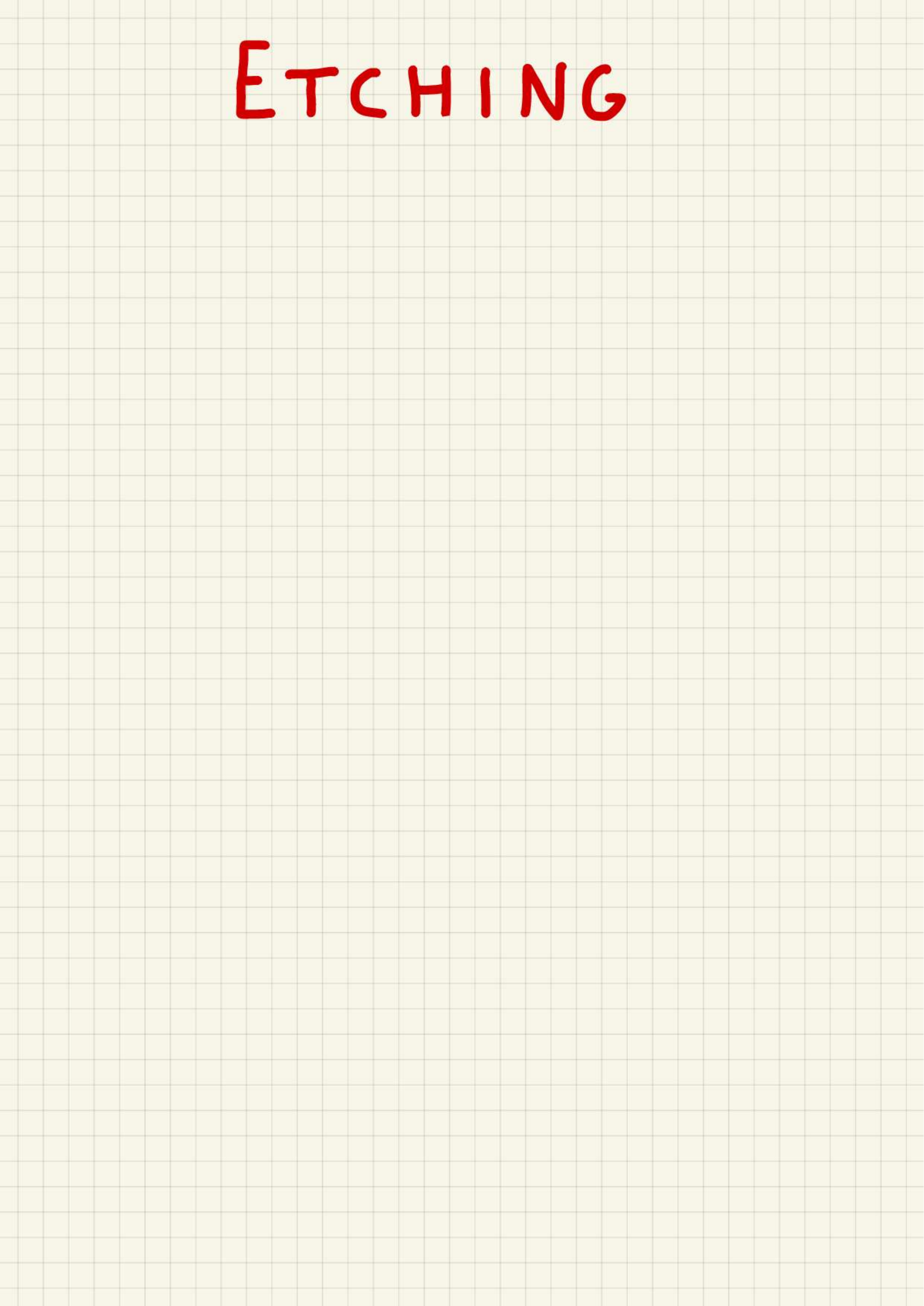
Random copolymer
Block copolymer
A-block
B-block

Block Copolymer Self Assembly
Increasing A-content
spheres
cylinders
lamella

BONUS PIC
DSA

If you take two Block copolymers (A-block and B-block), they can assemble into different morphological structures depending on the A/B ratio, self-assembling on the wafer.



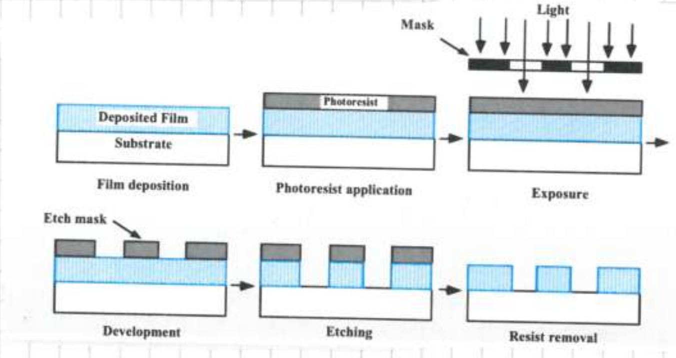

ETCHING

Lecture 21 — 12 maggio

ETCHING

After thin films are deposited, they are usually removed from the wafer surface to form the desired pattern (and sometimes the silicon substrate is patterned too).

After patterning, photoresist and/or hardmasks are etched away from the wafer, either by wet or dry (plasma) etching.



ETCH PARAMETERS

The parameters that we can tweak in our etching process are:

- **etch rate**: amount of material removed per unit time
- **selectivity**: proportion of etch rates for different materials (mask and materials underneath the film to be patterned MUST be preserved)
- **anisotropy**: ratio between vertical etch rate / horizontal etch rate (a certain degree of anisotropy is often needed, and actually anisotropy is crucial for downscaling)
- **uniformity**: how uniform the etch rate is across the wafer

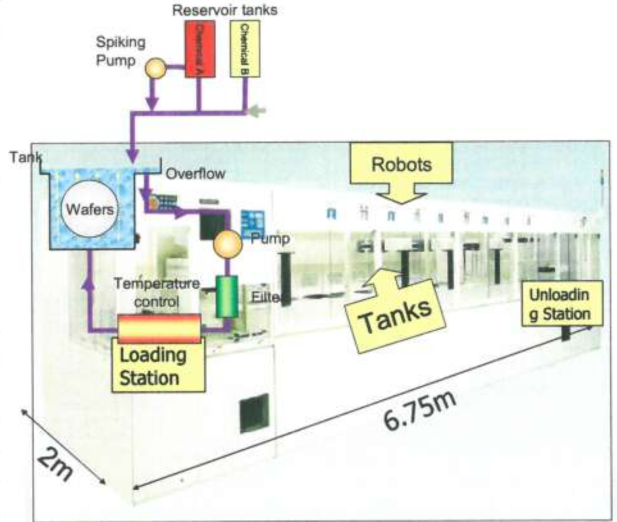
"SELECTIVE TO X" means that my process does NOT remove X. !

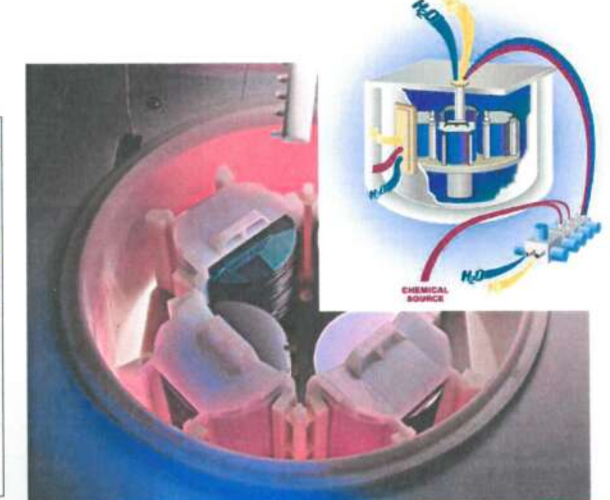
CAUTION: we can also have selective deposition (syntax: "selective to X" → deposit everywhere except X. "Selectively on X" → deposit only on X)

WET ETCHING

Wet chemical etchants are used: we immerse the wafers into baths or spray them with chemicals, and the chemical reaction will erode the material that it's NOT selective to.
It was the first method used in microelectronics, now widely used especially to remove hardmasks, photoresist and dry etching by-products.

PROS: very high selectivity, high uniformity and etch rate control
CONS: no anisotropy (except for crystalline orientation-dependent silicon etch) ⇒ scava in tutte le direzioni





	Designation	Chemical Constituent
SiO_2	Hydrofluoric acid	$HF:H_2O$
	BOE,BHF	$HF:NH_4F:H_2O$
	SC1	$NH_4OH:H_2O_2H_2O$
Si_3N_4	Hydrofluoric acid	$HF:H_2O$
	(orto-)phosphoric acid	H_3PO_4
Organics	Piranha,SPM	$H_2SO_4:H_2O_2$
	SOM	$H_2SO_4:O_3$
	Ozonized DIW	$DIW:O_3$
Si	Nitric:HF	$HF:HNO_3$
	Silicon etch	$HF:HNO_3: H_2SO_4: H_3PO_4$
	Tetramethylammonium-hydroxide	TMAH
	Ammonium hydroxide	NH_4OH
	Potassium hydroxide	KOH
	Hydrazine/water	$N_2H_4:H_2O$
W	Peroxide	H_2O_2
	SC1	$NH_4OH:H_2O_2H_2O$
	SC2	$HCl:H_2O_2:H_2O$
Al	SC2	$HCl:H_2O_2:H_2O$
	SC1	$NH_4OH:H_2O_2H_2O$

	Designation	Chemical Constituent
Cu	Ozonized water	$O_3:H_2O$
	SC2	$HCl:H_2O_2:H_2O$
Ti	SC1	$NH_4OH:H_2O_2H_2O$
	SC2	$HCl:H_2O_2:H_2O$
	Hydrofluoric acid	$HF:H_2O$
Co	SC2	$HCl:H_2O_2:H_2O$
	Piranha,SPM	$H_2SO_4:H_2O_2$
Ni	Piranha,SPM	$H_2SO_4:H_2O_2$
Ta	Hydrofluoric acid	$HF:H_2O$
	Nitric:HF	$HF:HNO_3$
TaN	Nitric:HF	$HF:HNO_3$
Al_2O_3	(orto-)phosphoric acid	H_3PO_4
	Hydrofluoric acid	$HF:H_2O$
HfO_2	Hydrofluoric acid	$HF:H_2O$
	Nitric:HF	$HF:HNO_3$
ZrO_2	Hydrofluoric acid	$HF:H_2O$
	(orto-)phosphoric acid	H_3PO_4
	Piranha,SPM	$H_2SO_4:H_2O_2$
	Nitric:HF	$HF:HNO_3$

DIW: de-ionized water (acqua distillata)

PLASMA ETCHING (DRY ETCHING)

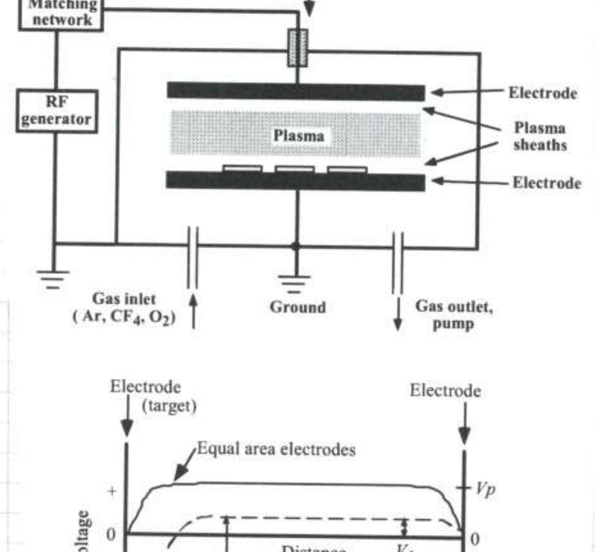
A plasma is used to etch the desired material.
Both chemical reactions and ionic bombardment contribute to film removal. (both radicals (neutral molecules) and ions in the plasma)

PROS: high anisotropy, profile shape controllability

CONS: contamination, damage (to the films, electrical damage by charge accumulation, ...), loading and memory effects, etch by-products to be removed.

NOTE: the selectivity achievable with dry etching isn't as good as wet etching.

NOTE: not all by-products are volatile and so can be pumped away. Some might be in solid state and will be removed by wet etching.



As we have already seen for PECVD or sputtering, we have a plasma generator (RF unit).
The plasma will have a mean electric field in the bulk = 0.
All the electric field will form across the sheaths.
We can modulate the electric field at one side by changing the electrodes area (or by just applying an additional $\vec{E}$ at the wafer side), so that we can not sputter the electrode opposite to the wafer.

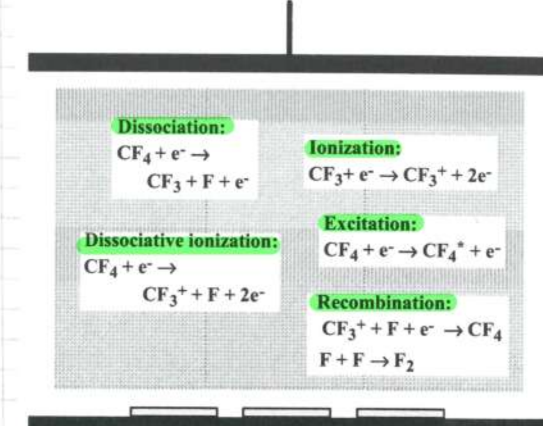
It's everything as it was for PVD sputtering, but instead of a target that gets sputtered we have a wafer that gets etched. !

NOTE: modern plasma tools are NOT multi-wafer, since it doesn't allow for a very good control over plasma uniformity.

PLASMA ETCHING BASIC MECHANISMS

(*) alogenuri

Etching gases contain halide species like CF_4, SiF_6, Cl_2, HBr, ... plus additives such as O_2, H_2 and Ar.
O_2 by itself is used to etch the photoresist.



Pressure : 1 mtorr ÷ 1 torr

Typically there are about 10^{15} neutral species / cm^3 (of these, around 1 ~ 10 % may be free radicals) and $10^8 \sim 10^{12}$ ions (and e^-) / cm^3

Usually (in standard systems with only one generator) the plasma density is coupled to the ion energy / speed since increasing the generator power increases both simultaneously.
Modern plasma tools have 2 generators, so they can control "indipendently" ion density and ion energy.

can't be really uncoupled at 100%

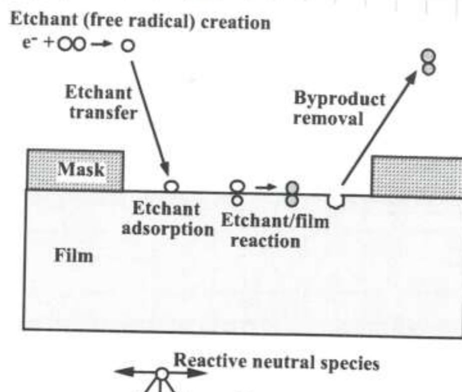
example (picture): CF_4 can go throug...

- dissociation $CF_4 + e^- \rightarrow CF_3 + F + e^-$
- ionization $CF_3 + e^- \rightarrow CF_3^+ + 2e^-$
- dissociative ionization $CF_4 + e^- \rightarrow CF_3^+ + F + 2e^-$
- excitation $CF_4 + e^- \rightarrow CF_4^* + e^-$
- recombination $CF_3 + F + e^- \rightarrow CF_4$ or $F + F \rightarrow F_2$

So now we have the reactive plasma. What will happen at the surface?

(*) "halide" = binary compound where one part is a halogen atom (group 17, so F, Cl, Br, I, At).

CHEMICAL ETCHING (in a plasma chamber)



We have a chemical reaction happening that will start from gas phase reactants that we inserted into the chamber, a solid state film (that we want to remove) and the byproduct of the chemical reaction will be volatile (and pumped away).

example:

$$CF_4 + e^- \rightarrow CF_3 + F + e^-$$

$$4F + Si_{(solid)} \rightarrow SiF_{4\,(gas)}$$

This chemical etching happening in a plasma chamber isn't very different from the one happening during wet etch.

We will have a very uniform arrival angle distribution ($\cos^n\theta$ with $n = 1$), low sticking coefficient S_c, high selectivity and (since it's purely chemical) a isotropic etch.

↑ instead of immersing in a liquid, we immerse in a gas

PHYSICAL ETCHING (SPUTTERING)

Basically the same as PVD sputtering, but instead of sputtering the target we sputter the wafer.

⊕ Ionic species

Mask

Film

The sputtering yield is usually low ($Y = 1$ atom / incident ion), but remember that depends on the angle and is not max at 90°. This leads to faceting.

Normalized Sputtering yield, Y

Incident angle

The sticking coefficient $S_c \approx 1$, so where the atoms arrive first they'll stay (anisotropic etching).

The degree of selectivity we can get is very poor (we'll have to make sure masks are thick enough).

physical etching →

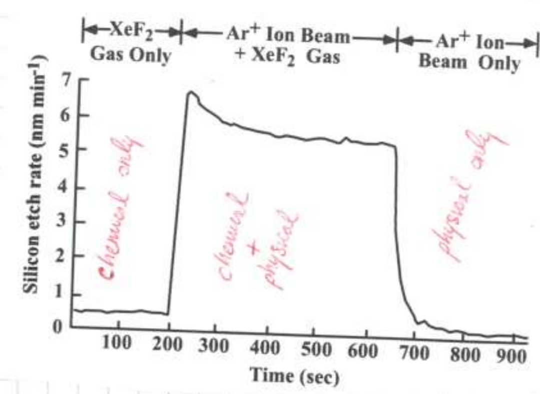
faceting

"Faceting" happens because Y is greater for $\theta > 90°$, so we might get an enhanced sputtering on the corners.

NOTE: if the species ionized by ion bombardment are reactive we could also have some chemical reaction

ION ENHANCED ETCHING

Ion + chemical



Ion bombardment and chemical etching seem to work in a synergetic way:

- the resulting etch rate is higher than the sum of single etch rates

- anisotropy is very high

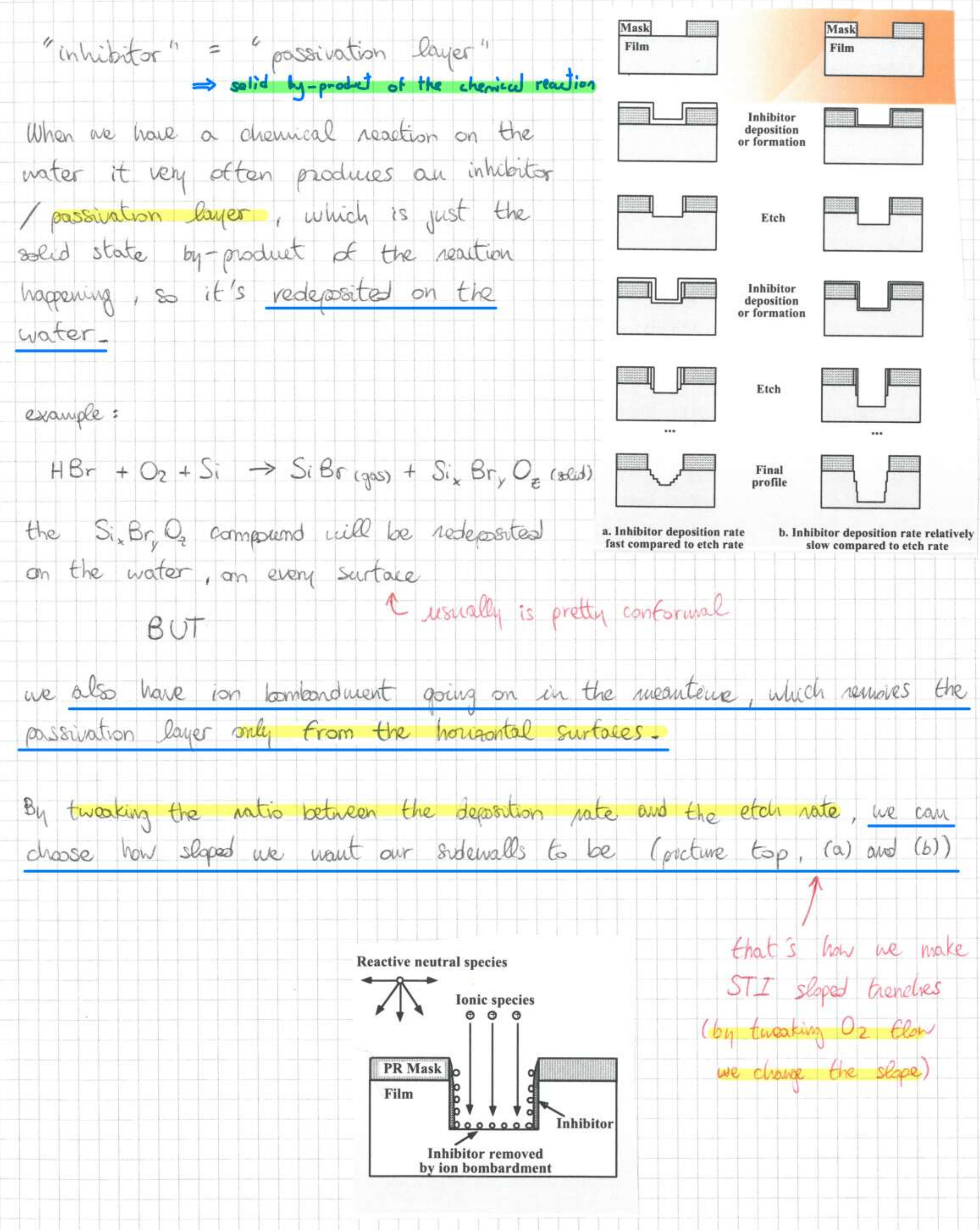
As we can see from the picture, we get that the ion enhanced etching is 6 times the sum of chemical only + physical only etch rates!

We don't yet know why this happens, although many hypothesis are present:

- ion energy might help breaking some bonds, enhancing chemical reactions
- the chemical reaction might produce some solid state products (called inhibitors) that then get removed by the ion bombardment, keeping the chemical reaction alive

possibly is a combination of the two phenomena.

ION ENHANCED INHIBITOR ETCHING



"inhibitor" = "passivation layer"

⇒ solid by-product of the chemical reaction

When we have a chemical reaction on the wafer it very often produces an inhibitor / passivation layer, which is just the solid state by-product of the reaction happening, so it's redeposited on the wafer.

example :

$$HBr + O_2 + Si \rightarrow SiBr_{(gas)} + Si_xBr_yO_z\ _{(solid)}$$

the $Si_xBr_yO_z$ compound will be redeposited on the wafer, on every surface

↑ usually is pretty conformal

BUT

we also have ion bombardment going on in the meantime, which removes the passivation layer only from the horizontal surfaces.

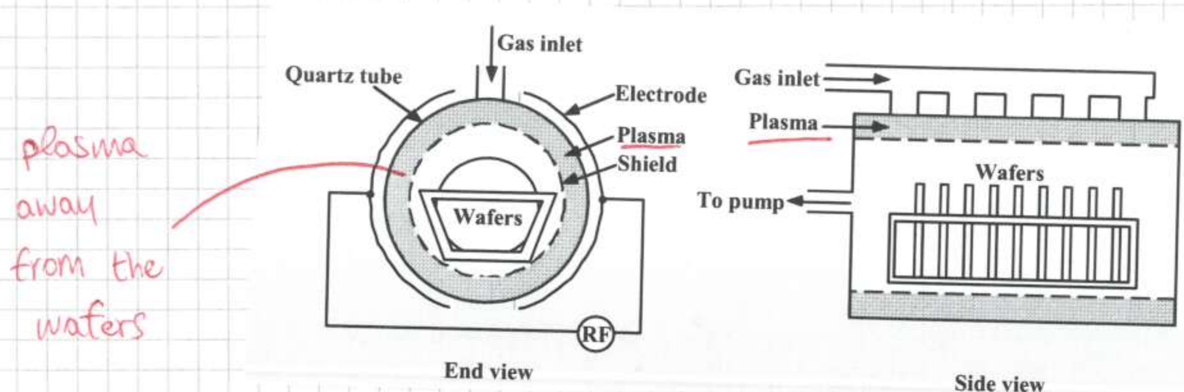
By tweaking the ratio between the deposition rate and the etch rate, we can choose how sloped we want our sidewalls to be (picture top, (a) and (b))

↑ that's how we make STI sloped trenches (by tweaking O_2 flow we change the slope)

ETCHING TOOLS (CHEMICAL)

Just plasma diffusing, hashing

plasma away from the wafers



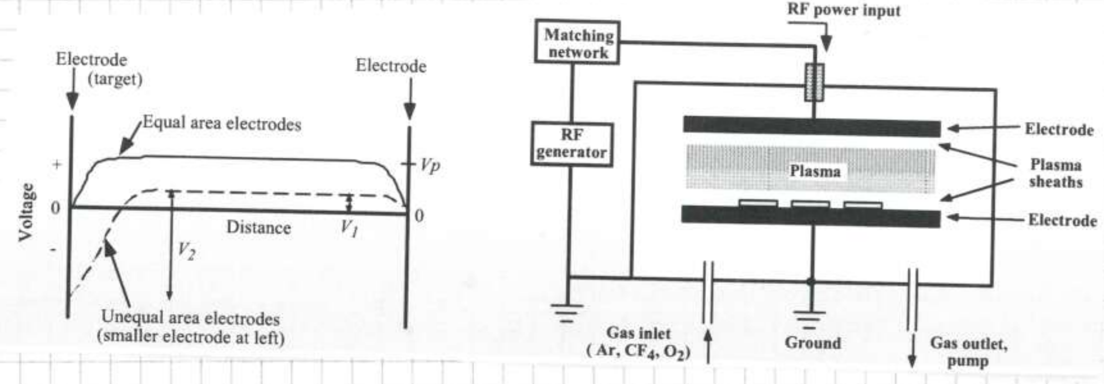
"Barrel etcher" is a downstream plasma etcher where the plasma is kept separate from the wafers, so we can't have ion bombardment.

So the plasma is formed away from the wafers, and all the fun is happening there (sheaths, electric fields, ...). In the bulk we will have just some plasma slowly diffusing from the outside.

Since this is basically purely chemical etching, it's main use is resist ashing (O_2 plasma used to burn carbon containing resists)

Nowadays isn't really used anymore in IC manufacturing for applications other than resist ashing. ?

ETCHING TOOLS (CHEMICAL + PHYSICAL)



"Parallel plate etcher", very similar to PE-CVD chambers (page 161)

Can be operated in 2 modes:

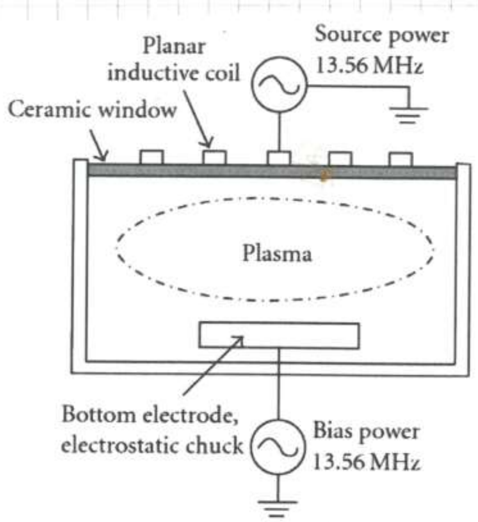
- plasma mode: symmetrical electrodes (same area) which leads to poor ion bombardment and small sheath voltage drop at wafer surface

- RIE (reactive ion etching) mode: asymmetric electrodes (different areas) guarantee enhanced ion bombardment

NOTE: all the etching systems in modern IC manufacturing are single wafer tools

ETCHING TOOLS (CHEMICAL + PHYSICAL)

"High density plasma systems" is a family of plasma tools where the plasma density and the ion energy are decoupled by using a RF source to control the bottom electrode (wafer holder), and the plasma density is controlled by ECR (Electron Cyclotron Resonance) or ICP (Inductively Coupled Plasma) systems.



Pressure = 1 ~ 10 mtorr

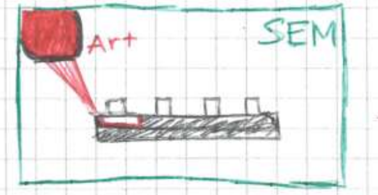
With this system we can achieve high ion density and relatively low ion bombardment, simultaneously.

ETCHING TOOLS (PHYSICAL)

"Sputter etching", which we have already seen

"Ion milling", we have a separate chamber which generates the plasma (usually Ar^+), which then is shaped into a beam which is then accelerated towards the target.

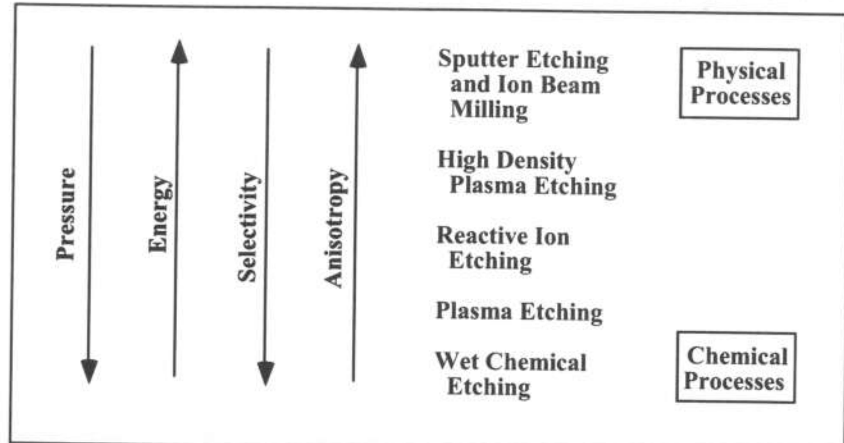
Ion milling can also be used for TEM sample preparation, by cutting a very very thin slice of the sample under a SEM and then putting it into a TEM.



⊏⊐ single transistor cut by ion milling, ready for TEM

SUMMARY

IMPORTANT

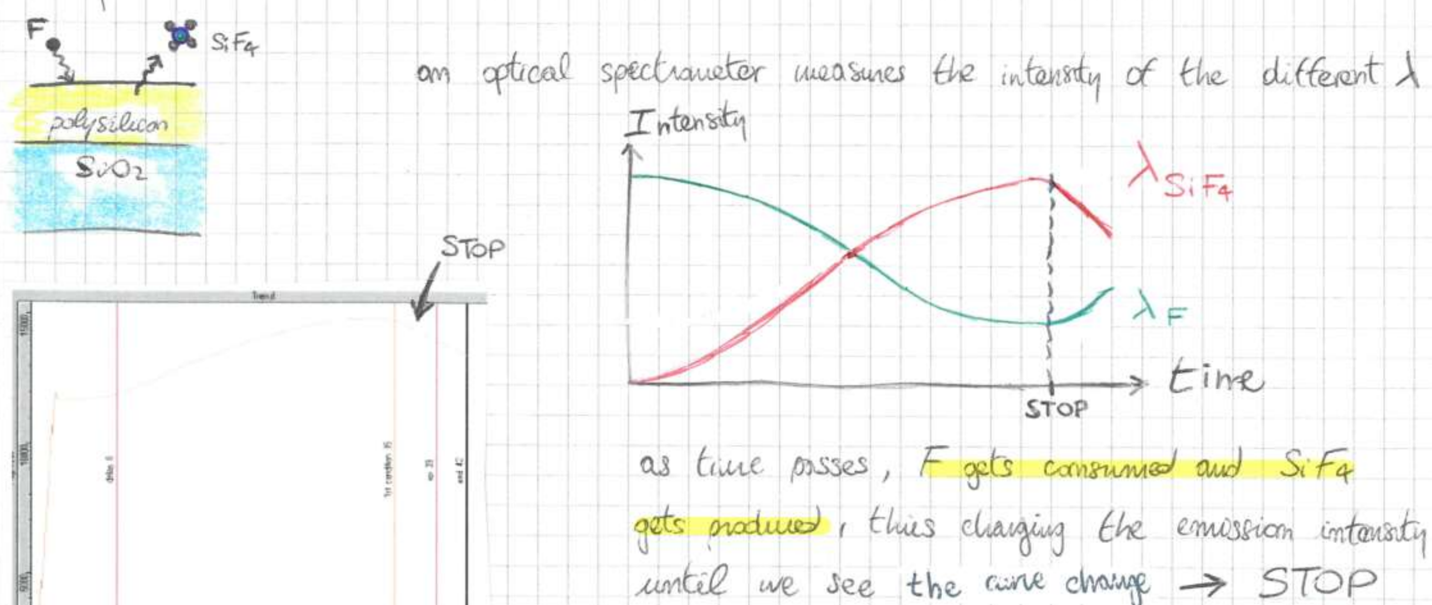


END POINT DETECTION

How do we know when to stop the etching? Besides taking the thickness we want to etch and divide it by the etching rate, to get the time we should keep on etching, is there a better way? The thickness may not be uniform from wafer to wafer, batch to batch,... and the etch rate can change over time depending on many factors.

The main end point detection (= when we have etched all the film, no overetch, no underetch) is based on Optical Emission Spectroscopy: we exploit the fact that every atom has its own emission lines, and many of the collisions in the plasma chamber don't ionize the molecules/atoms, just excite.

example: $F + Si \rightarrow SiF_4$ selective to SiO_2 (doesn't etch it)



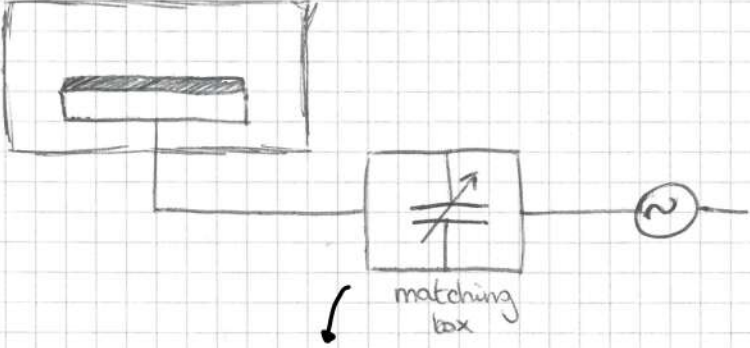
an optical spectrometer measures the intensity of the different λ

as time passes, F gets consumed and SiF_4 gets produced, thus changing the emission intensity until we see the curve change → STOP

NOTE: let's look at λ_{SiF_4} (for λ_F will be the opposite).
We start at zero emission intensity because we didn't produce any yet. Then the reaction starts → λ_{SiF_4} intensity goes up until it reaches steady state (SiF_4 produced = SiF_4 pumped away). When we see λ_{SiF_4} intensity go down, it means that we are pumping away and no more SiF_4 is being produced → we stop the process here.

Another end point detection technique is based on interferometry: we shine a laser at an angle on top of the surface and look at the interference pattern, which changes as the thickness of the film changes. This is rarely used in etching systems, while is used often in CMP tools.

- Yet another end point detection technique is using the information coming from the matching box (box containing variable capacitors that changes its impedance so that the maximum signal coming from the RF source reaches the wafer, without reflections. But since we're thinning down the wafer, the impedance is changing → matching box adaptes) to track the etch.



minimize reflected power, impedance matching

matching box is varying continuously, we can correlate the film thickness to the matching box.

PLASMA ETCHING PROCESS "KNOBS"

What are the variables we can change in our plasma etching tools?

- power

  RIE : 0,1 ~ 5 W/cm², ion energy 10 ~ 700 eV
  HDP = 0,1 ~ 3 W/cm², ion energy 10 ~ 500 eV

- pressure

  RIE = 10 ~ 100 mtorr
  HDP = 1 ~ 10 mtorr

- gas species
- flow rate
- flow distribution
- wafer temperature
- chamber temperature

when we have ultimated a process step with all the etching parameters right to remove the material we want, we have made a RECIPE (different etch recipes will etch different materials).

ETCH RATE DEPENDENCE ON "KNOBS"

Short answer: it's complicated.

Long answer:

- etch rate obviously depends on gas species
- increasing power increases etch rate (in HDP systems ion bombardment can also be independently controlled)
- increasing pressure decreases ion density but increases chemical flux (→ etch rate dependence on pressure is very hard to predict)
- flow rate has a minor impact on etch rate (second order, if it's increased more reactant species are supplied, but can reduce the residence time, so we might not give enough time to the reactant species to do their job).

ETCH PROFILE DEPENDENCE ON "KNOBS"

We can control the etch profile by:

- increasing / decreasing ion bombardment (power), since more ion bombardment means more anisotropy, but we have seen that we can reach an equilibrium between ion bombardment and passivation layer redeposition by which we can create particular profiles, so it's more of a balance than pure ion bombardment "knob".
- using different etchant gases we can control the by-product formation and passivation layers redeposition.

Lecture 22 — 14 maggio

PLASMA ETCHING ISSUES

If we take two different DRAM cells (for example 1GB vs 8GB, using different masks but exactly the same materials, structure, etch recipe, ...) we can see that they might have different etch rates, partially due to the different geometry → different local plasma density.

Loading effect is the etch rate dependence on local density

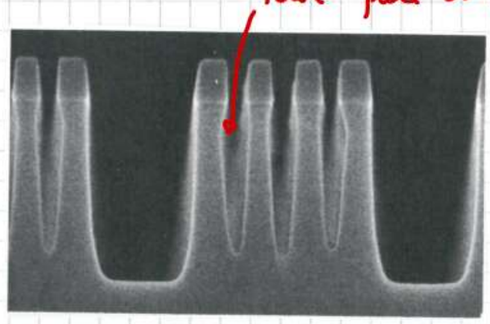
Micro-loading effect: structures with different mask openings or aspect ratios are etched with different etch rates

we didn't reach the bottom → look like we must stop, but we must wait

ox

Micro-loading effects can be due to:

lower pace of etching



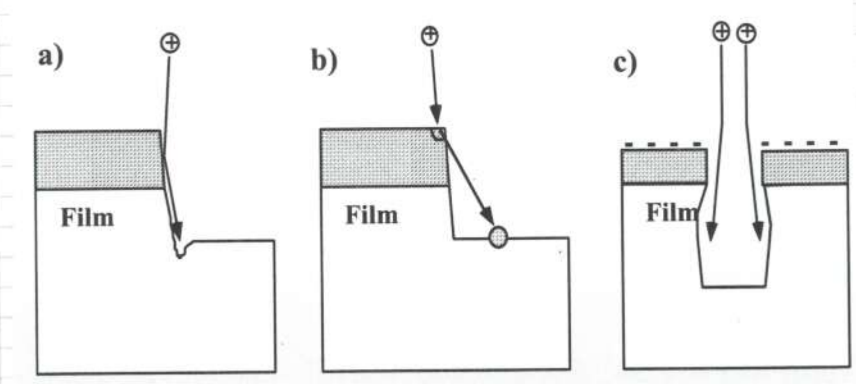
- depletion / trapping at the surface of reactant species
- ion paths distorted due to the surface getting charged and deviating the ions
- shadowing effects (the geometry shadows the flow of some reactants)

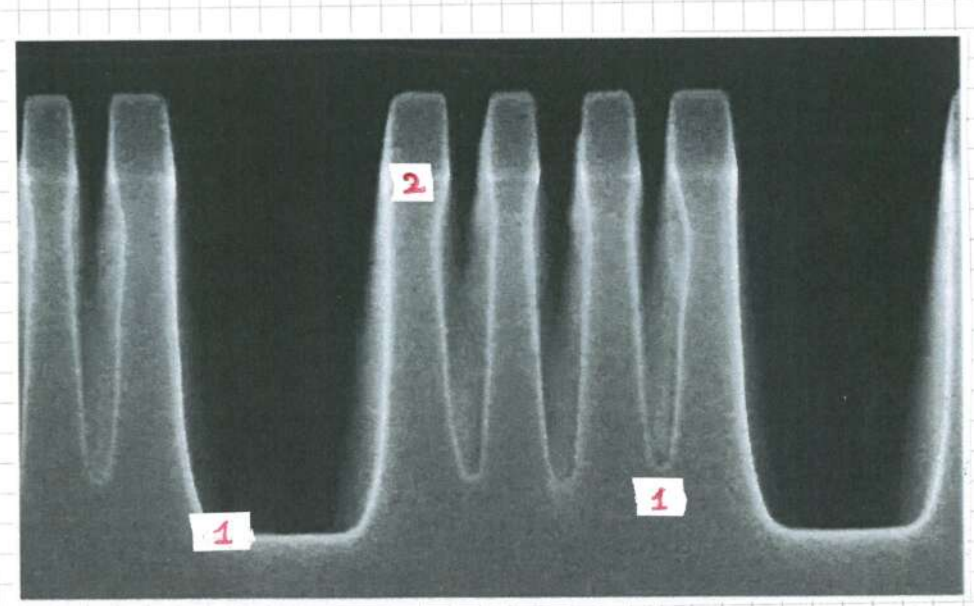
Usually what we see is that narrow openings are etched at a lower rate compared to large openings.

Sometimes the exact opposite happens (nobody knows really why) and it's called reverse micro-loading effect.

CD critical dimension smallest feature

The paths of the single ions become significant in determining the final profile of the etched structure. We could have ions recoiling and hitting spots that we might not expect (micro trenching, picture a), we could have the ions hitting the photoresist then resputtering it (resist resputtering, picture b), or we could have some of the free electrons charge up the surface and deviate the ions (ion trajectory distortion due to resist charging, picture c).





Set of STI trenches; 3 narrow and 1 wide (narrow is around ~50 nm).

We can see (1) how the wide trenches are much deeper than the narrow ones, because of the higher etch rate.
We can notice (2) the ion trajectory distortion due to the top surface charging.

Most of the times micro-loading effects must just be dealt with since we can't do much to mitigate them, while some other times we can fine tune the dry etching to minimize some effects.

Other problems arising from dry etching are radiation damage (you can implant Hydrogen and Carbon in the film since we're accelerating ions into the film, contaminating it to a certain depth) and plasma damage (or antenna effect).

ANTENNA EFFECT

Antenna Structure

b

Conductor

Field Oxide

Gate Oxide

Si

a

Antenna ratio : $\frac{b}{a}$

NB little to do with real antennas

The ratio between the area of the conductor exposed to the plasma divided by the area of the gate oxide is called "antenna ratio". The higher the ratio the easier will be to damage the device (by voltage breakdown of the gate oxide).

To avoid this we try to design the device in such a way as to avoid having large area discrepancies between the plates of the (wanted or unwanted) capacitors. !

But even if we design everything correctly, we might have plasma damage by accident:

mask

etching →

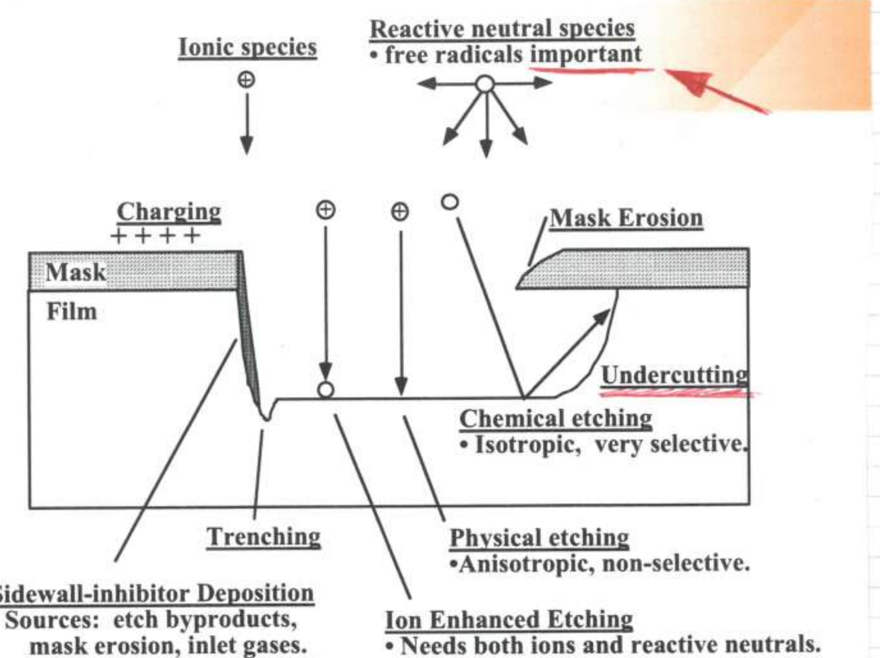
large area

Just before finishing our etching process, we have a very large surface that acts as a giant plate, accumulating charges

↓

possible plasma damage!

ALL IN ONE PICTURE



PLASMA ETCHING CHEMISTRIES ⇒ SILICON OXIDE

One very common way to etch Si and SiO_2 is through Flourine based chemistries, since F radicals are very reactive.

$$4F + SiO_2 \rightarrow SiF_{4\,(gas)} + O_2$$

For example, gases like NF_3, SF_6, CF_4 can perform both isotropic or anisotropic etching depending on wether or not we mandate re-deposition of solid state by-products. Selectivity to silicon is poor for this gases, so they etch Si as well as SiO_2.

Fluorine - poor chemistries can be used to reduce isotropy: CHF_3, C_3F_8, C_2F_6 are some examples. Increasing the amount of Carbon increases the polymer redeposition, increasing anisotropy (page)

SILICON and POLY-SILICON

Due to SiF_4 volatility, F-based chemistries can be used to etch Silicon as well.

PROBLEMS: SiO_2 selectivity is poor, relatively high anisotropy

Adding O_2 to CF_4 can increase oxide selectivity (Si etch is enhanced much more than SiO_2 etch).

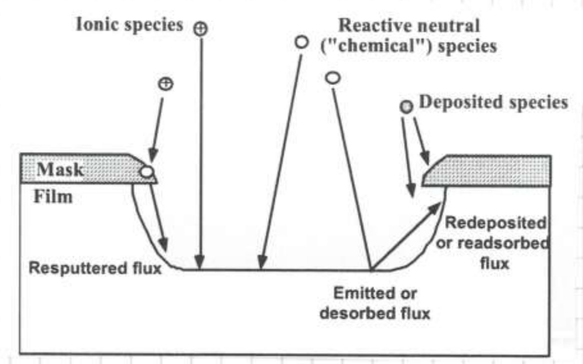
Halogen gases are used to increase both selectivity and anisotropy: Cl_2 and HBr are the most commonly used gases. Adding O_2 can activate inhibitor deposition, allowing more profile control.

ALUMINUM

For metals in general (except COPPER, which is very hard to dry etch) we usually have to run some sputtering to remove any native oxides on the surface (like Al_2O_3), then use clorine Cl to etch the metal.

CAUTION! Clorine can stick to sidewalls and once we take out our wafer, it can react with H_2O present in the air, form HCl and corrode the metal.

ETCHING MODELS



The model we're going to build will be pretty simple because many of the mechanisms at play are still not fully understood.

As always, we can sit on the surface, look up and model the arrival angle distribution as

$$F(\theta) = F^{\circ} \cos^{n}\theta$$

$n = 1$ for neutral chemical species (because they will come from everywhere) and $n = 10 \sim 80$ for ions (depending on ionization intensity, electric field, ...)

We can also use again the concept of sticking coefficient, where $S_c = 1$ for ionic species and very low for neutral chemicals (they "bounce" around a lot).

NEUTRAL SPECIES	IONS
S_c very low	$S_c = 1$
$n = 1$	n very high
↓	↓
isotropic etch	anisotropic etch

analogously as we did for the model of deposition, where S_c and n determined the isotropy of the process.

LINEAR MODEL

The simplest model will be a linear one, where we imagine that the sputtering and chemical reaction act independently:

$$\text{Etch rate} = \frac{(S_c K_f F_c + K_i F_i)}{N}$$

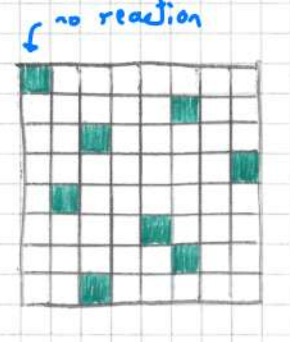
($F_c \to F_0 \cos^n(\theta)$, $F_i \to F_0' \cos^m(\theta)$)

- F_c (F_i) = flux of chemicals (ions) arriving at a given point of the surface
- K_f = reaction rate constant
- K_i = effective sputtering yield (depending both on incident angle and ion energy)
- N = density of the film to be etched

Micro-loading effects can be also simulated, since we can calculate the flux in every point in the surface, take into account the geometry, the bouncing around of the molecules, ...

ION ENHANCED ETCHING MODEL

In ion enhanced etching we have our neutral species reacting with the surface, forming some volatile byproducts (pumped away) and some solid-state, which will re-deposit, covering the surface and blocking the reaction (= chemical etching). The ion bombardment removes this passivation layer by sputtering, exposing the underlying surface, which can now react again with the neutral species.



take a piece of the surface, divide it into tiny pieces, then consider how many of them are covered (by the passivation layer formed by the reaction), they can't react anymore, unless ions bombard that piece and remove the passivation layer.

If we call the fraction of surface covered by the passivation layer θ, then the number of neutrals absorbed per unit area per unit time is $= S_c\,(1-\theta)\,F_c$

- S_c → sticking coefficient
- $(1-\theta)$ → fraction of available sites / little squares
- F_c → flux

The number of by-products removed per time per unit area $= K_i \theta F_i$
where F_i is the flux of ions impinging and K_i is the sputtering "efficiency".

At steady state condition the two process rates will be equal:

$$\left[S_c (1-\theta) F_c = K_i \theta F_i \right]$$

$$\downarrow$$

$$\theta = \frac{1}{1 + \frac{K_i F_i}{S_c F_c}}$$

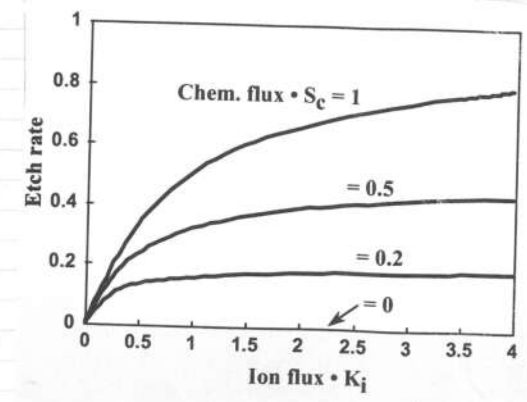
NOW: we know θ, we know that the rate at which the by-products are removed is $K_i \theta F_i$, the etch rate is $K_i \theta F_i$ divided by the film density N.

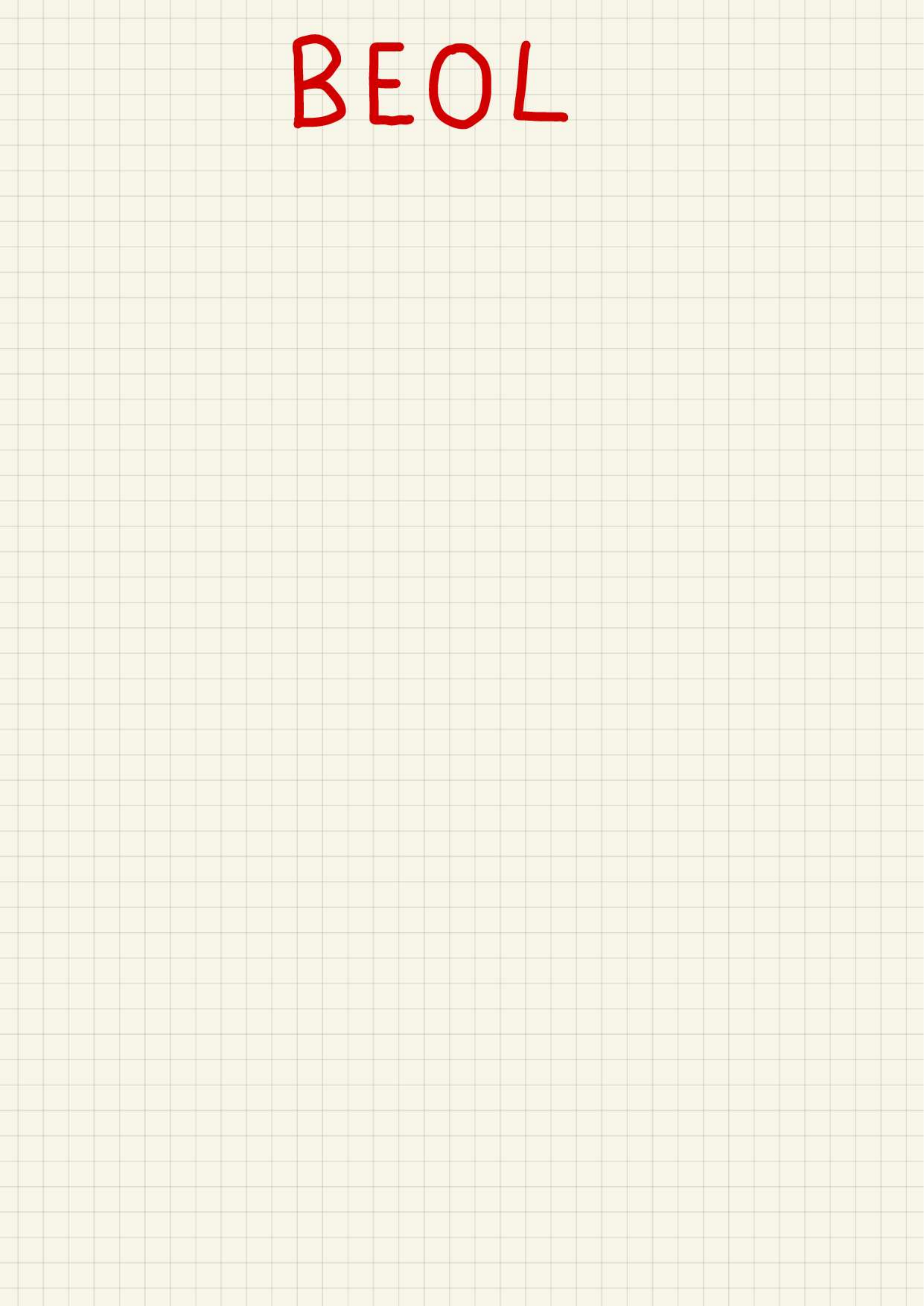
$$\text{Etch rate} = \frac{K_i \theta F_i}{N} = \frac{1}{N} \frac{1}{\left(\frac{1}{K_i F_i} + \frac{1}{S_c F_c}\right)}$$

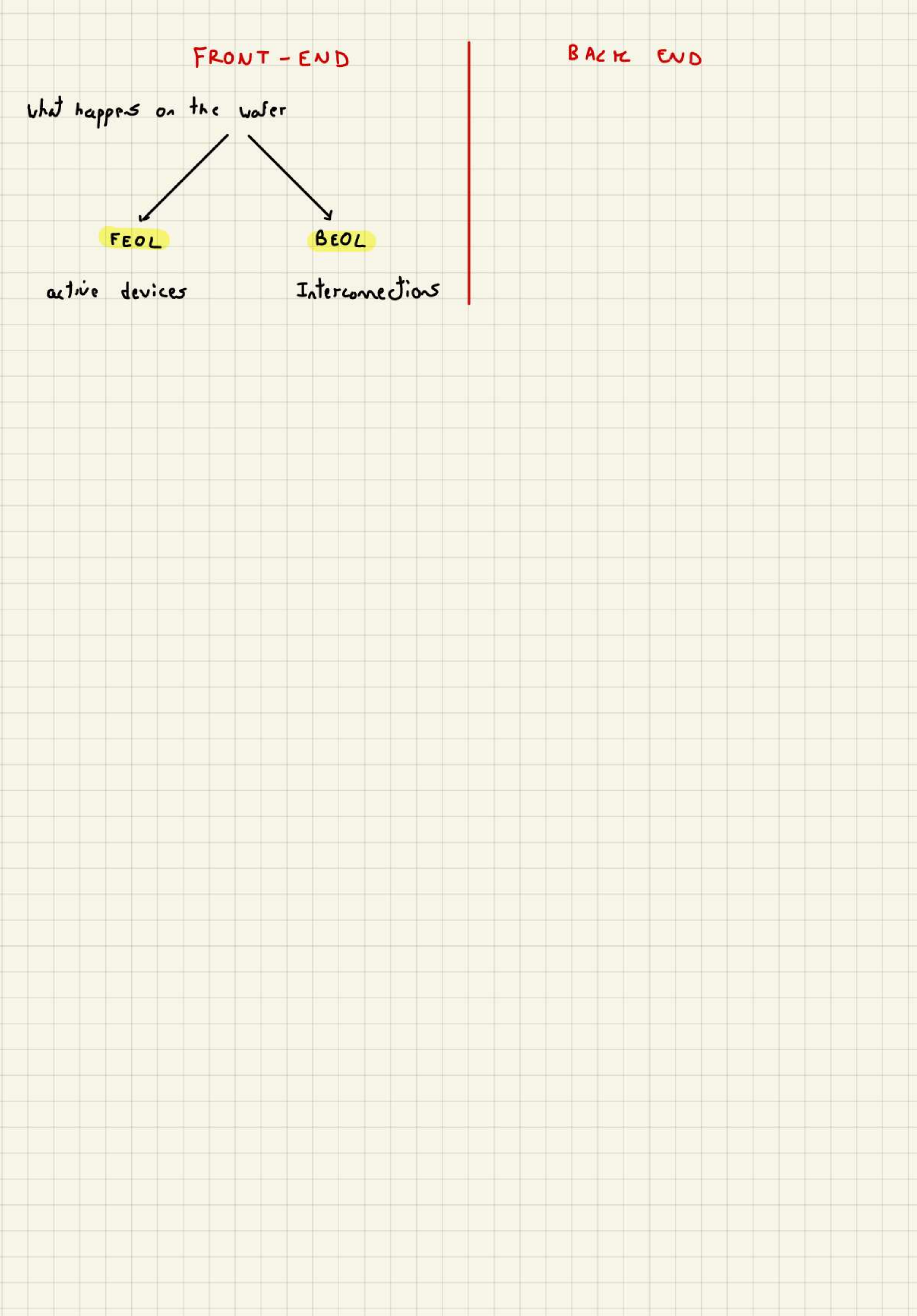
In this model if either flux is zero, the overall etch rate is zero since both are required to etch the material.

The etch rate saturates when one component gets too large relative to the other (as in deposition, the slower process dominates).

In this model we don't account for independently-formed inhibitor layer, or excess inhibitor layer formation.



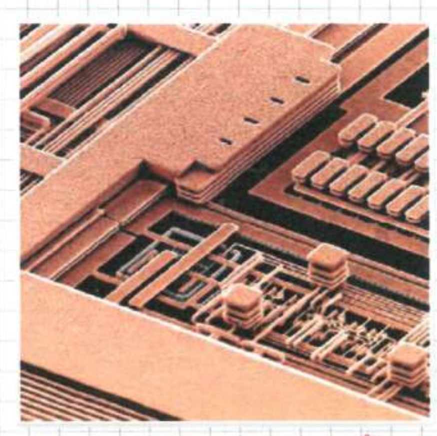

BEOL


FRONT - END
BACK END
what happens on the wafer
FEOL
BEOL
active devices
Interconnections

Lecture 23 — 19 maggio

BACK END OF THE LINE (BEOL)

In this process step we'll see all that is needed to wire the active devices into specific circuit configuration (interconnect layers, contacts, vias, insulating dielectric layers, ...)



Usually interconnections are divided into categories:

(1) • local interconnects (silicided transistor gates, silicided active areas, like source and drain regions)

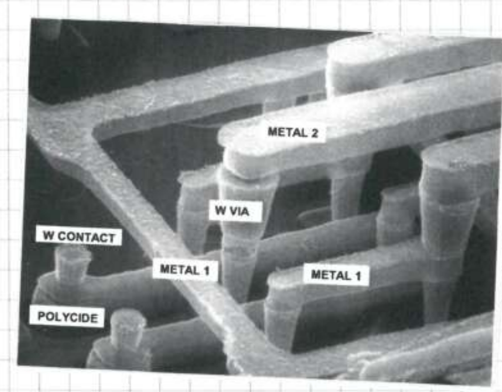
not really an interconnect, used when for example we have to "connect" 2 adjacent gates, VERY LOCAL

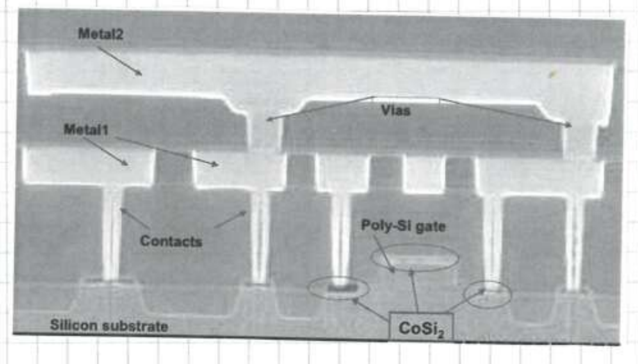
(2) • contacts (connecting local interconnects / active devices to first metal level)

(3) • intermediate / global interconnects (metal lines wiring active elements onto circuits)

(4) • dielectrics (insulating different metal levels and lines)

(5) • vias (connecting different metal lines)





ERROR 404 again = page not found

NOTE: as we go up, the pitch (= dimension) of the interconnects is usually designed to become larger, so that we have less resistance.

LOCAL INTERCONNECTS (1) Very local D-S

How do we turn the silicon surface into a conductor so we reduce local resistance. ⟹ we reduce Si resistance

For example: source and drain regions on top of which we'll want to put interconnects, and I want to also connect source and drain, so very local.

Local interconnects are below the lowest metal level.

How do we turn our silicon into a metal? → SILICIDE siliciuri

["silicide" = compound with silicon + metal]

Silicide properties:

Silicides properties:
- low resistivity (10 ~ 50 μΩ cm)
- high temperature stability
- good compatibility with silicon
- easy to plasma etch
- good compatibility / interface with other materials
- no electromigration

How do we form silicides? By deposition or chemical reaction.

- direct deposition: the silicide is directly deposited on the wafer by PVD or CVD (CVD used only for WSi_2) widely used

- Reaction by SALICIDE (self aligned silicide):

  - the metal is deposited (by sputtering) directly on exposed gates and source/drain regions
  - the wafer is annealed and the metal reacts with Silicon (no reaction occurs on unexposed silicon areas, so oxides, nitrides, STI, LOCOS, spacers, ...)
  - unreacted metal is stripped off (only unreacted)

EXAMPLE: SA-$TiSi_2$

- First we have to remove any SiO_2 that is present over the Si and poly-silicon areas that we want to turn into silicides. Very likely the silicon exposed areas grew a thin native oxide when exposed to air. We also have to make sure that no native oxide will be formed inbetween the cleaning of SiO_2 step and the metal deposition step. We can control the wet etch in a way as to passivate the surface with hydrogen, so we can retard the formation of native oxides.
- Then we'll deposit Ti by sputtering (50 ~ 100 nm)

  ↑ remember that we can control which electrode gets bombarded so we can do in-situ wafer cleaning before deposition.

- N_2 anneal at 600 ~ 700 °C for 15 ~ 60 seconds
  the Ti in contact with Si will react to form $TiSi_2$ (no reaction on SiO_2), while the Ti not in contact with Si will react with N_2 and form TiN.
  We keep the temperature relatively low because we don't want too much lateral $TiSi_2$ growth, which could short adjacent gates for example.

TiN

$TiSi_2$

SiO_2 Si SiO_2

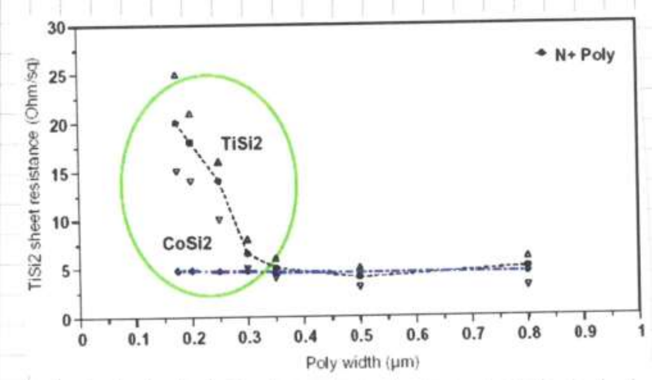
← volume expansion on all directions, could lead to damage!

reaction happening at expenses of Si! Si is being consumed

• TiN and unreacted Ti are stripped off (wet etch with ammonia NH_4OH diluted with water H_2O and hydrogen peroxide H_2O_2)

PROBLEM: the 600~700°C anneal gives us $TiSi_2$ in the so-called "C49 phase", which is not the low resistive crystalline phase that we want (that's the C54 phase)

• We heat the wafer at 800°C, so we convert the $TiSi_2$ into the C54 low resistivity crystalline phase (but temperature and time must be kept as low as possible to avoid agglomeration and dopant diffusion into the metal layer).



Over the years we have used many different silicides. Why?

As we can see from the picture, the resistivity of $TiSi_2$ becomes very large at very low thicknesses. How can the resistivity (not the resistance!) change with the dimension of the feature? That's because the silicides are polycrystalline materials, and we know that at grain boundaries the conduction decreases. But as we shrink down the dimension of the interconnects, grain boundaries become more and more relevant, degrading our resistivity.

ISSUE $TiSi_2$ is a polycrystalline material

as the device shrinks down, there are grains boundaries

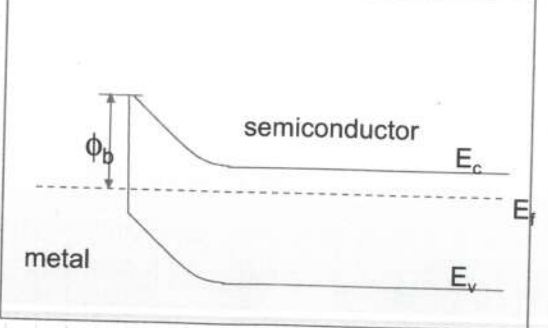
examples of silicides

Silicide	Resisitivity (μΩcm)	Sintering temperature (° C)	Stable on Si up to (° C)	nm of Si consumed per nm of metal	Nm of silicide per nm of metal	Barrier height to n-Si (eV)
PtSi	28-35	250-400	750	1.12	1.97	0.84
$TiSi_2$(C54)	13-16	700-900	900	2.27	2.51	0.58
$TiSi_2$(C49)	60-70	500-700		2.27	2.51	
WSi_2	30-70	1000	1000	2.53	2.58	0.67
CoSi	100-150	400-600		1.82	2.02	
$CoSi_2$	14-20	600-800	950	3.64	3.52	0.65
NiSi	14-20	400-600	650	1.83	2.34	
$NiSi_2$	40-50	600-800		3.65	3.63	0.66
$MoSi_2$	40-100	800-1000	1000	2.56	2.59	0.64
$TaSi_2$	35-55	800-1000	1000	2.21	2.41	0.59

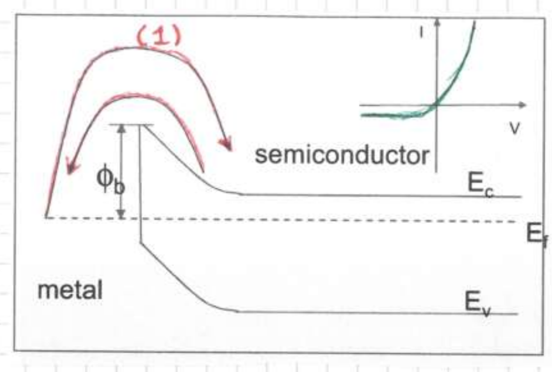
State-of-the-art transistors usually come with a metallic gate because you need high-k dielectrics, and to match the work functions we can't use polysilicon gate / silicide gate. Source and drain junctions are still being silicided in many processes, usually with NiSi / $NiSi_2$.

CONTACTS (2)

Contacts provide low resistance connections between active devices and the first metal layer: so it's a metal-semiconductor connection. That's a Schottky diode! We don't want a Schottky diode, we want an ohmic contact. We want to flow current IN and OUT both with low resistance.



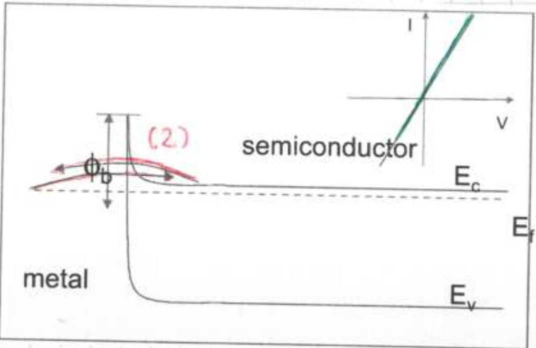
When we put a metal in contact with a semiconductor there will be a band off-set: the band of the semiconductor will bend up by Φ_b, the work function difference, from the Fermi level E_F of the semiconductor. So we'll have a barrier between the metal and the semiconductor, and for the conduction to happen the barrier must be overcome by the carriers (e^- and holes).



We can have 2 types of conduction:

(1) **thermoionic emission current**:

the carriers must go OVER the barrier, so they need a lot of energy, especially to cross in one direction → rectifying behaviour.

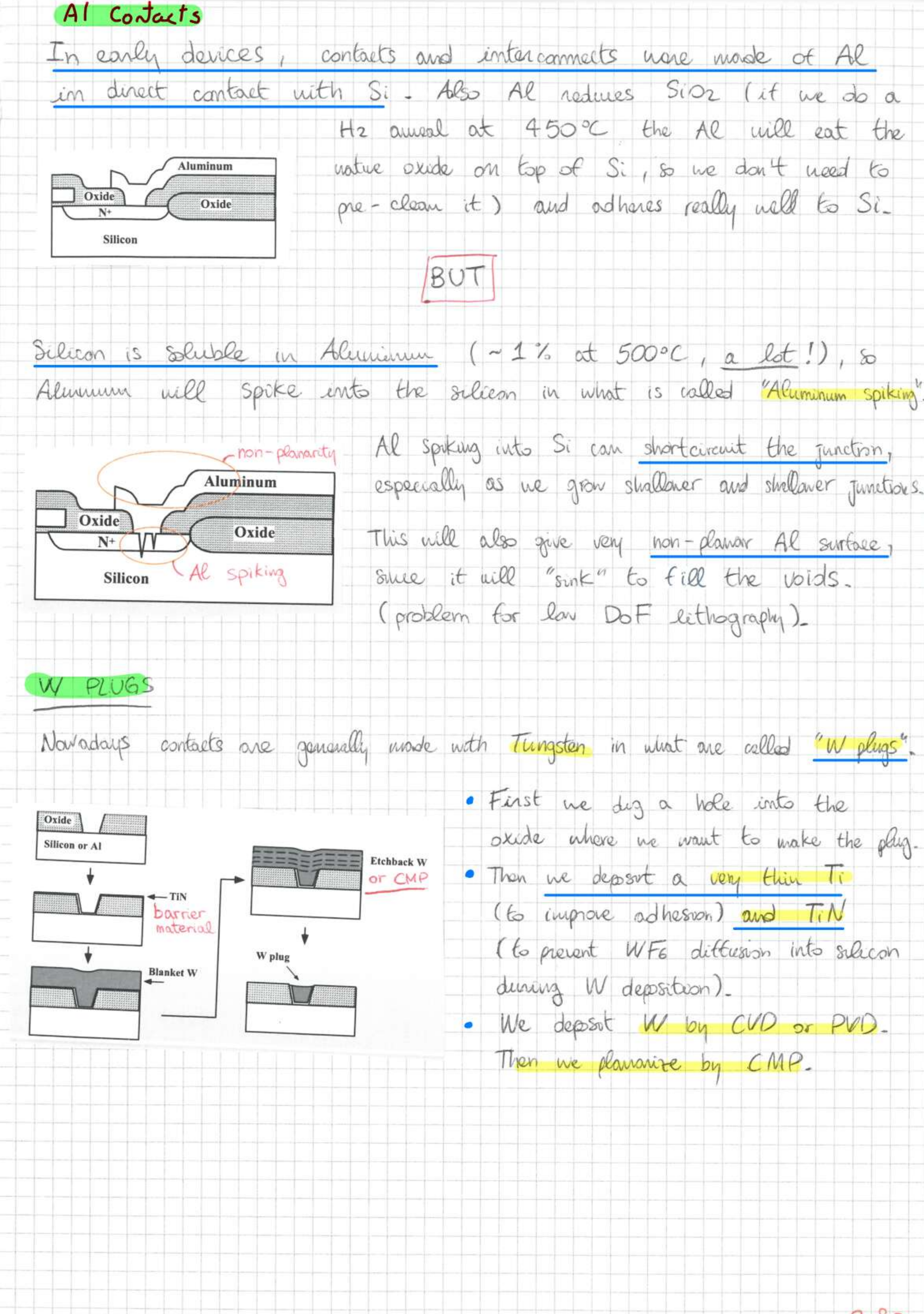


(2) **tunneling regime**:

if our Schottky junction is in tunneling regime (very thin barrier) then the carriers can go THROUGH the barrier, and we'll see an ohmic behaviour.
To shrink the barrier we can dope heavily the semiconductor, and also use a metal with a work function close to Si.

Contacts properties:

- ohmic electrical contact
- low resistivity
- high thermal stability during the process flow



Al Contacts

In early devices, contacts and interconnects were made of Al in direct contact with Si. Also Al reduces SiO_2 (if we do a H_2 anneal at 450°C the Al will eat the native oxide on top of Si, so we don't need to pre-clean it) and adheres really well to Si.

BUT

Silicon is soluble in Aluminum (~1% at 500°C, a lot!), so Aluminum will spike into the silicon in what is called "Aluminum spiking".

Al spiking into Si can shortcircuit the junction, especially as we grow shallower and shallower junctions.

This will also give very non-planar Al surface, since it will "sink" to fill the voids. (problem for low DoF lithography).

W PLUGS

Nowadays contacts are generally made with Tungsten in what are called "W plugs".

- First we dig a hole into the oxide where we want to make the plug.
- Then we deposit a very thin Ti (to improve adhesion) and TiN (to prevent WF_6 diffusion into silicon during W deposition).
- We deposit W by CVD or PVD. Then we planarize by CMP.

INTERCONNECTS (3) "Intermediate / global"

General properties of interconnects:

- low resistivity
- good adhesion to underlying films
- stability during processing and circuit operation (we want the metal lines to survive over time and use)
- can be deposited and etched

Material	Resistivity (μΩcm)	Melting point (° C)
Al	2.7/3.0	660
W	8/15	3410
Cu	1.7/2.0	1084
Ti	40/70	1670
TiN	50/150	~2950

Aluminum has been the mainstream material for also interconnects (for a very long time) because of its relatively low resistivity, good adhesion and relatively easy deposition / etching

BUT

thermal expansion for Aluminum = $23 \cdot 10^{-6}\ C^{-1}$

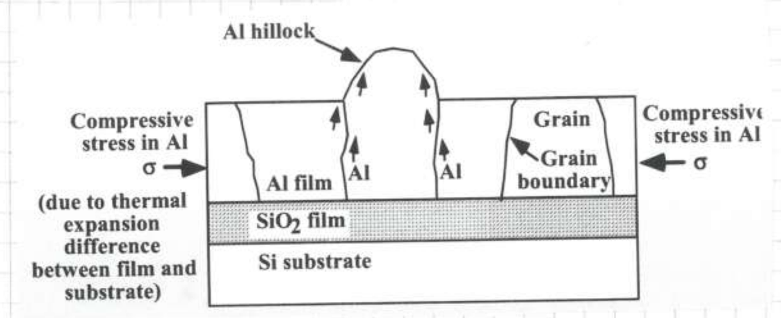
thermal expansion for Silicon = $2{,}6 \cdot 10^{-6}\ C^{-1}$

almost an order of magnitude !

NOTE: Al is still used for the very top metal layers, since there the dimensions are much bigger.

Al is still used

Al HILLOCKS AND VOIDS



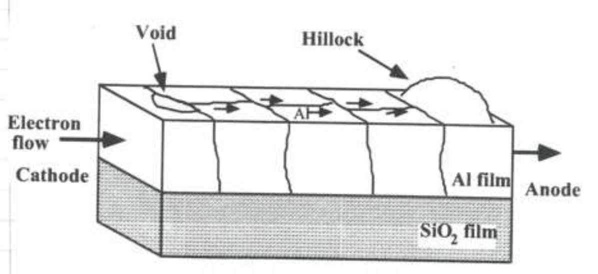
NOTE: everything we deposit on top of oxides / amorphous materials, will be polycrystalline (Al aswell).

When we heat up the wafer, the Al would like to expand a lot, while the underlying silicon substrate much less → COMPRESSIVE STRESS IN Al. Because of the compressive stress in Al, bumps form on the surface (called hillocks), and if 2 neighbouring hillocks touch, we can have a shortcircuit.

NOTE: the deformations usually happen at grain boundaries

When we cool down the wafer the opposite happens, and the tensile stress in Al will create voids (usually at grain boundaries).

Other than by differences in thermal expansion coefficients, we can form voids and hillocks also by electromigration.



Electromigration is the movement of Al atoms caused by high current densities. For 0.1 ~ 0.5 MA/cm² (or more) through a metal line, we'll have a lot of electrons flowing through the metal line. This high current of electrons can drag the Al atoms because we can imagine the Al atoms as partially positively ionized, since the outermost electrons are delocalized and form the "sea of electrons". This flow of Al atoms can form voids and hillocks (again, at the grain boundaries).

For lower current densities the deformation will be elastic, so no permanent damage will form. If we exceed the plasticity of the material voids and hillocks will form.

This problems can be partially solved by depositing instead of pure Al, an alloy of Al - Cu (Cu places itself at grain boundaries so mitigates this effects).

We can also deposit first Ti then Al : Al will try to mimic the grain shape of Ti, mitigating electromigration.

Ti / TiN can also shunt the metal line in case of voids or cracks formation

VOID
e^-
Al
TiN

"shunt" = low-resistance path for electric current.

Al metal lines are made simply by depositing and pattern it with a mask.

Once the metal layer has been completed, we might want to insulate one metal layer from the one above it, to avoid shortcircuit. To do this we use dielectrics.

DIELECTRICS (4)

General requirements for BEOL dielectrics:

- good electrical isolation
- low dielectric constant
- high breakdown electric field (> 5 MV / cm)
- good adhesion
- good stability
- good structural density
- no contaminants
- permeable to hydrogen !

Permeable to hydrogen : why?

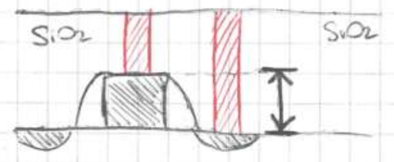
We have seen (page) that the way to repair SiO_2 from some of the traps / fixed charges at the Si / SiO_2 interface is to run an hydrogen anneal, because H will penetrate and saturate the dangling bonds between Si and SiO_2, improving the SiO_2 and Si/SiO_2 interface quality.

Usually the last step in our process flow is an Hydrogen anneal to repair crystalline damage (400 ~ 450 °C anneal, called also "Hydrogen alloying"), so everything we put on top of Si / SiO_2 must be permeable to hydrogen.

Some dielectric examples

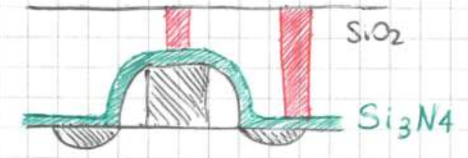
Material class	Material	K	Deposition technique
Inorganic	SiO_2 (including PSG, BPSG)	3.9/5.0	CVD, HDP
	FSG		CVD, HDP
	Spin on glass	3.9/5.0	SOD
	Si_3N_4 (only in multi-layer structures)	5.8/6.1	CVD/PE-CVD
Organic	Polymides	2.9/3.9	SOD/CVD
Hybrids	Si-O-C polymers	2.0/3.8	SOD/CVD
Aerogels	Porous SiO_2	1.2/1.8	SOD
Air bridge		1.0/1.2	

NOTE: silicon nitride Si_3N_4 is used to etch SiO_2 at different heights, so we can land our contacts.

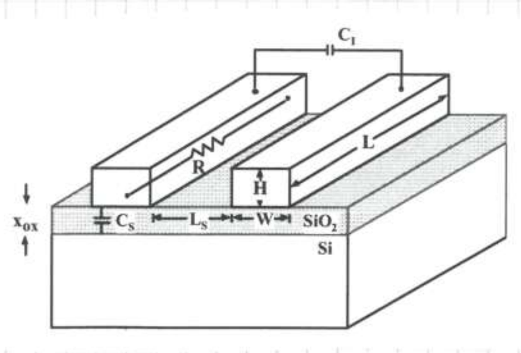


how do I etch this height difference without etching the gate?

We deposit a very thin layer of Si_3N_4 (thin because of it's very high k), then our SiO_2, planarize it, etch with a chemistry highly selective to Si_3N_4 and when we reach the bottom on both sides, switch chemistry to something that etches only Si_3N_4, then land the contacts.



Lecture 24   21 maggio



We know that the time delay of a signal transmitted over a metal line can be estimated as $\tau \simeq 0,9\,RC$.

The important thing is to remember that τ is proportional to RC.

So let's consider a simple example: 2 straight metal lines (picture)

They have width W, length L, resistance R, capacitance between them C_I since there's an insulator in between, there's also an oxide underneath of thickness X_{ox} which will give us an associated capacitance between the metal line and what's under the oxide (metal, substrate, ...) of C_S, the lines will have also height H and will be spaced by a distance L_S.

Resistance $R = \rho \dfrac{L}{WH}$ (ρ — resistivity)

Capacitance $C = C_S + C_I = \left(K_{ox}\,\varepsilon_0 \dfrac{WL}{X_{ox}}\right) + \left(K_{ox}\,\varepsilon_0 \dfrac{HL}{L_S}\right)$

(K_{ox} — oxide dielectric constant; ε_0 — dielectric permittivity of vacuum)

$$\rightarrow \text{time delay} \quad \tau = 0,89\,RC = 0,89\,K_i K_{ox}\,\varepsilon_0\,\rho\,L^2\left(\frac{1}{H X_{ox}} + \frac{1}{W L_s}\right)$$

($K_i \approx 2$)

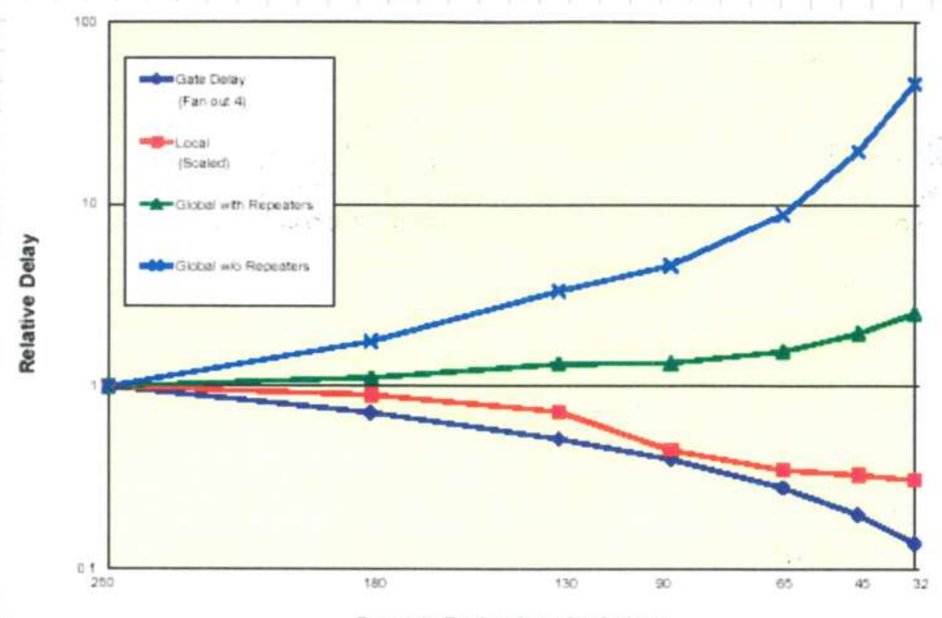
$K_i \approx 2$ takes into account the presence of other metal layers and electric field increasing the delay of the signal

As technology scales down, H, W, X_{ox} and L_s all scale down as ("proportionally to") F_{min}, which is the minimum feature dimension of the considered technology.

The maximum length of a line will go as $L_{max} \approx \sqrt{A}/2$, where A is the chip area (which increases as we scale down).

we end up with a time delay $\tau \approx 0,89\,K_{ox}\,\varepsilon_0\,\rho\,\dfrac{A}{(F_{min})^2}$

So as we shrink down the dimensions, our delay gets worse and worse, we'll need to act on material properties to keep τ under control.



As we can see from the picture, we are forced to insert repeaters inside our chip as the technology scales down. We have even some processors with 30% of their area covered with repeaters just for this problem!

Instead of repeaters, we can act on K_{ox} and ρ, so on the dielectrics and on the metals. That's why there has been a massive shift in the semiconductor industry from Al-based interconnects to Cu-based interconnects. Also low-k dielectrics have been introduced to minimize RC.

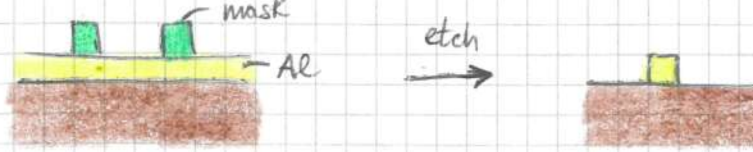
Copper issue

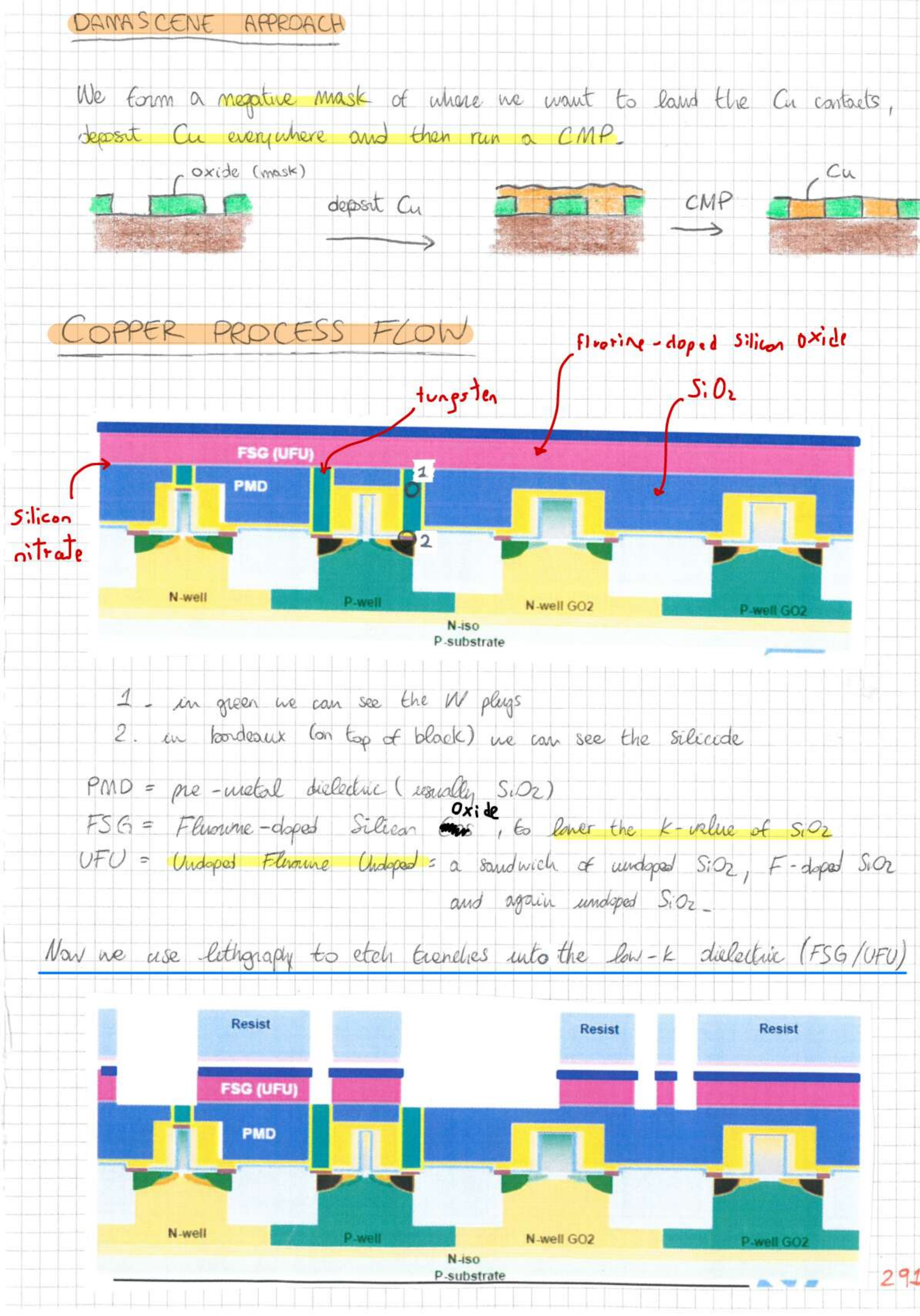
NOTE: Copper cannot be dry etched since its by-products are NOT volatile. It requires a damascene approach.

DAMASCATO

The damascene approach is the opposite of the subtractive approach we used to deposit Al interconnects.

Subtractive approach: you lay down Al, use a mask and etch the excess Al where we don't want it.





DAMASCENE APPROACH

We form a negative mask of where we want to land the Cu contacts, deposit Cu everywhere and then run a CMP.

COPPER PROCESS FLOW

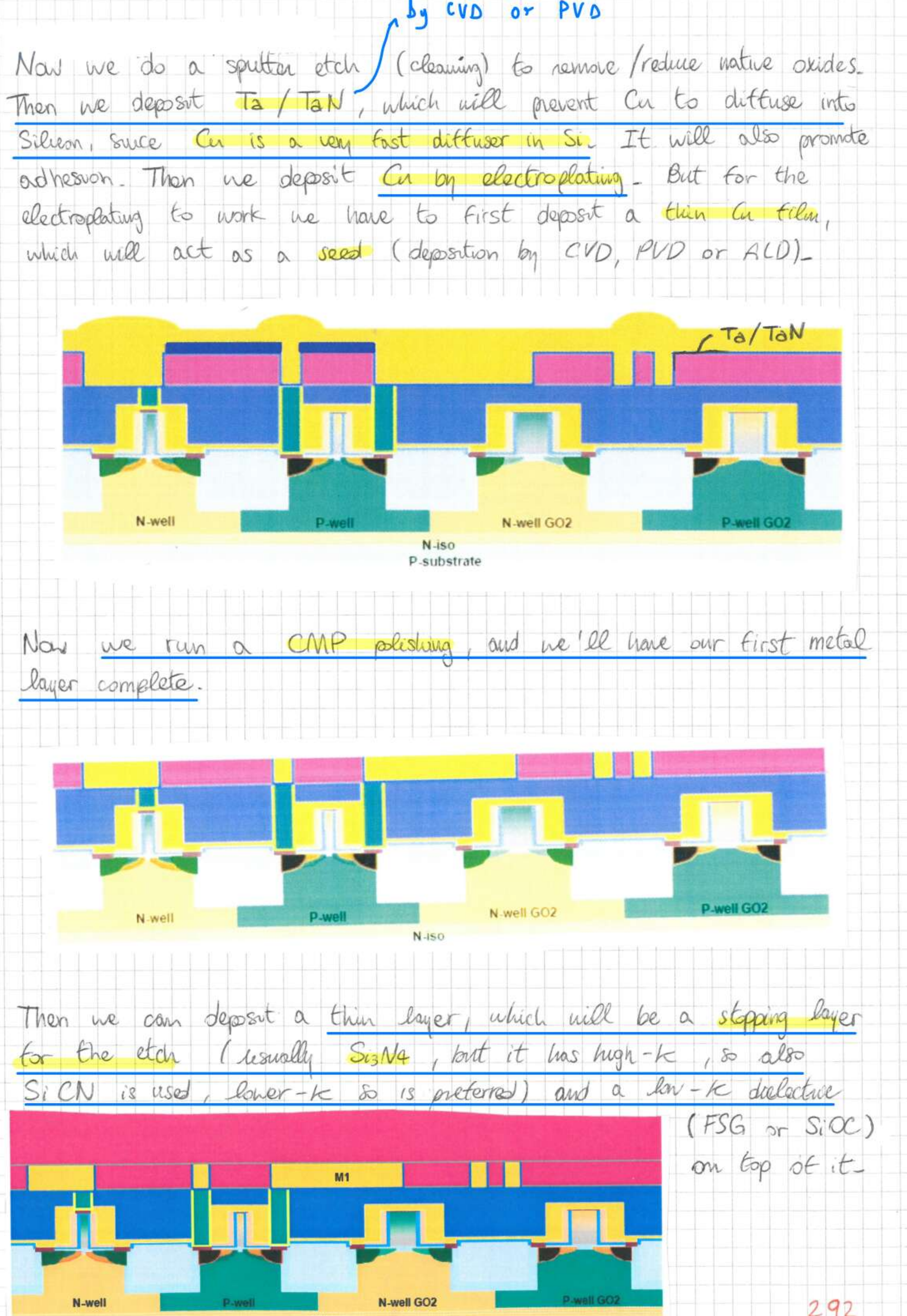
1. in green we can see the W plugs
2. in bordeaux (on top of black) we can see the silicide

PMD = pre-metal dielectric (usually SiO_2)

FSG = Fluorine-doped Silicon Oxide, to lower the k-value of SiO_2

UFU = Undoped Fluorine Undoped = a sandwich of undoped SiO_2, F-doped SiO_2 and again undoped SiO_2.

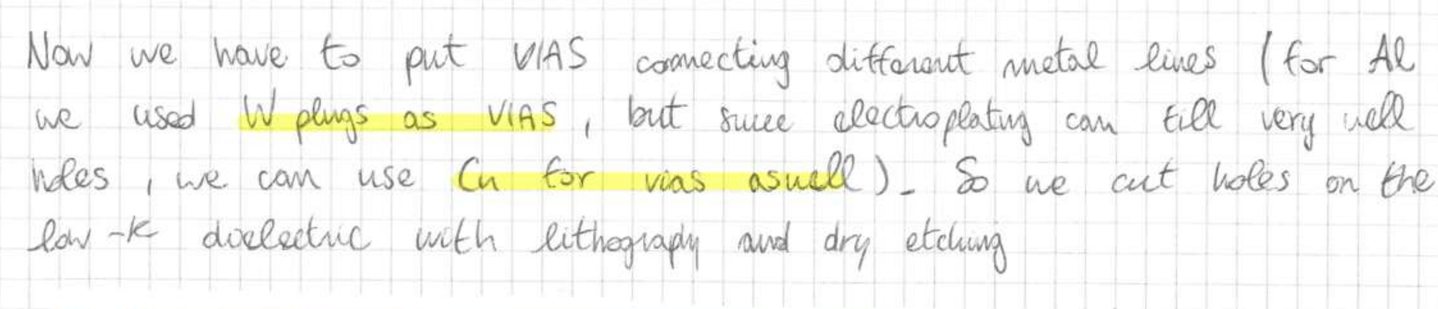
Now we use lithography to etch trenches into the low-k dielectric (FSG/UFU)



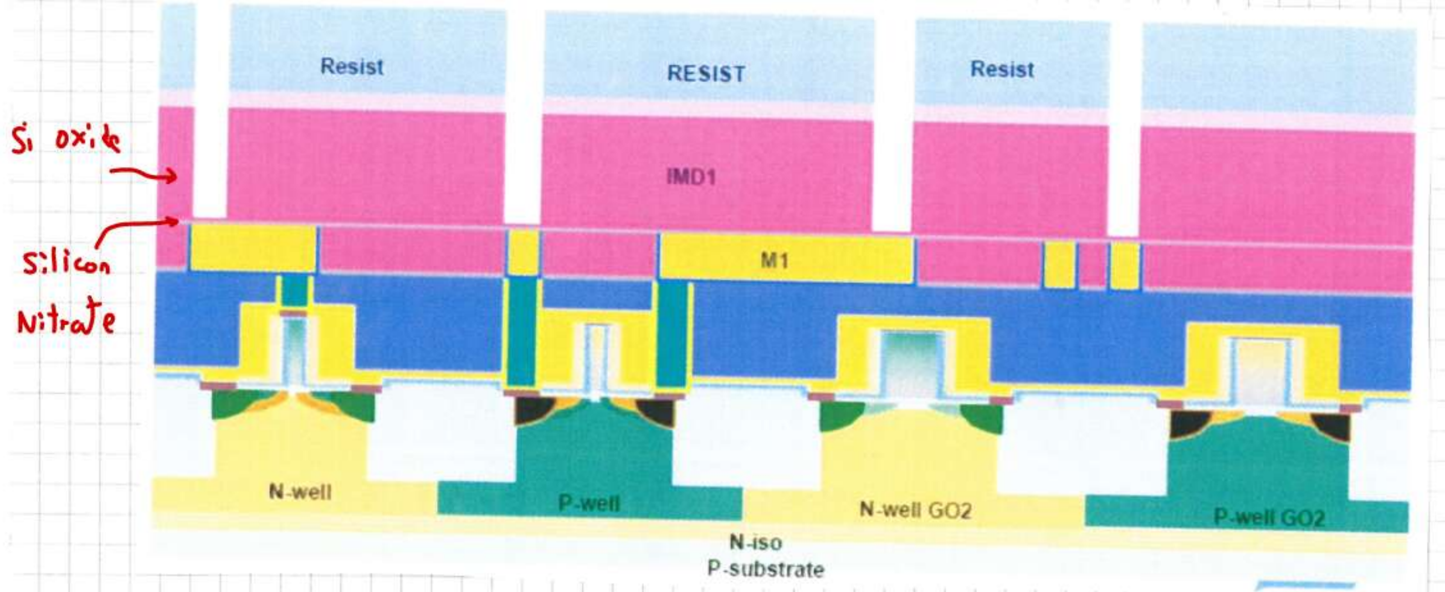
Now we do a sputter etch (cleaning) to remove/reduce native oxides. Then we deposit Ta / TaN (by CVD or PVD), which will prevent Cu to diffuse into Silicon, since Cu is a very fast diffuser in Si. It will also promote adhesion. Then we deposit Cu by electroplating. But for the electroplating to work we have to first deposit a thin Cu film, which will act as a seed (deposition by CVD, PVD or ALD).

Now we run a CMP polishing, and we'll have our first metal layer complete.

Then we can deposit a thin layer, which will be a stopping layer for the etch (usually Si_3N_4, but it has high-k, so also SiCN is used, lower-k so is preferred) and a low-k dielectric (FSG or SiOC) on top of it.

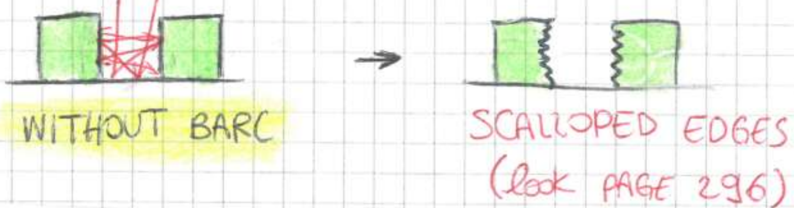


Now we have to put VIAS connecting different metal lines (for Al we used W plugs as VIAS, but since electroplating can fill very well holes, we can use Cu for vias aswell). So we cut holes on the low-k dielectric with lithography and dry etching

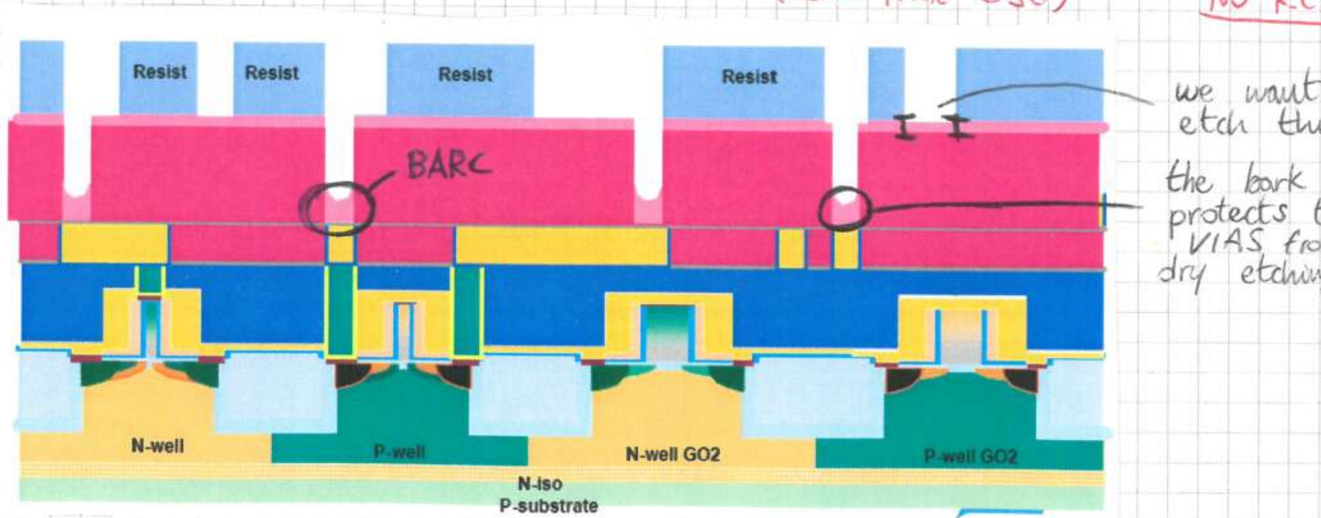


Then we strip away the portion of the resist where we will have our metal lines with another lithographic step (BARC will protect the VIAS from line etching).

BARC = Bottom anti-reflective coat, is an organic compound diluted in a solvent which is deposited before the photoresist to prevent the formation of standing waves, which will expose parts of the photoresist where we don't want.



WITH BARC
NO REFLECTION



we want to etch this

the bark here protects the VIAS from the dry etching

NOTE: obviously BARC isn't photosensitive, so we'll have to etch it away later.

We'll use the fact that we have BARC on top of the holes, so when we etch away part of the stopping layer to put the metal lines. We strip everything away (photoresist, BARC) =

N-well
P-well
N-well GO2
P-well GO2
N-iso
P-substrate

IN 3D

line (metal lines)

line

hole

cylinder (VIAS)

Then we just repeat the process as many times as we want =

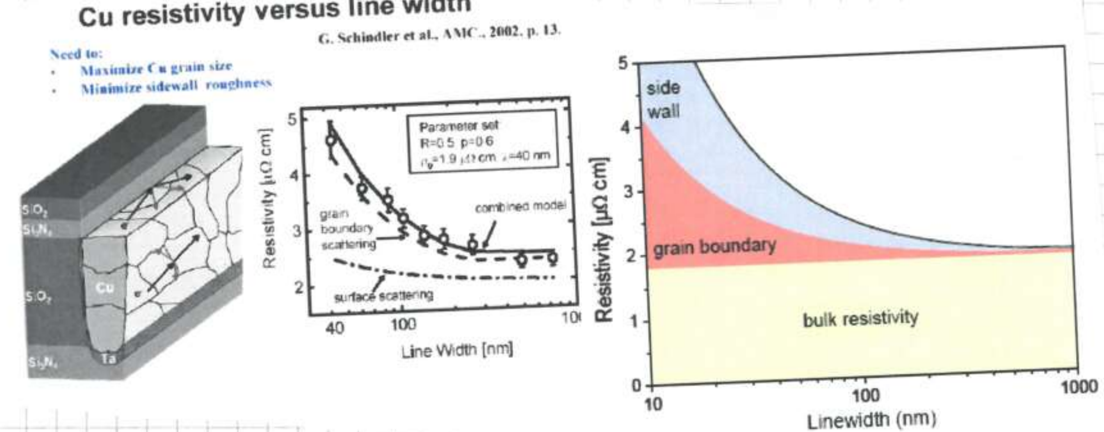
we deposit the Ta / TaN barrier, Cu seed deposition + electroplating, planarize with CMP, ...

Cu
W
N-well
P-well
N-well GO2
P-well GO2
N-iso
P-substrate

NOTE: the contacts (NOT the vias, so the first layer) are in W, not Cu, because we don't want any Cu near the active areas of the device. We have to also protect the back of the wafer for the same reason, because we don't want Cu to infiltrate from behind.

very last metal layer is Al, used also for bonding pads (less oxidation)

Cu MAIN ISSUES



Being copper a polycrystalline material, its conduction properties will also be an average between the bulk resistivity and grain boundary resistivity / side wall resistivity.

The more we shrink Cu, the more the boundary effects become important + the presence of the barrier (Ta / TaN) becomes relevant, since we can't reduce too much the thickness of the barrier otherwise Cu would diffuse into Si.

And also with thinner Cu lines comes greater Resistance R and greater resistivity ρ! (at such low dimensions, ρ starts becoming dependent on the geometry).

Then there are all the problems associated with the fact that our metal lines are made by CMP (and not by etching like Al), and being CMP also a mechanical process it can render the surface uneven, depending on the surface it started with. Let's see some examples.

PRE CMP POST CMP

Let's say the slurry (the white liquid used in the CMP) is selective to the gray material and not to the orange (metal), if we have large structures made of metal, the CMP can keep polishing more there, since the selective gray material is far away and the CMP disk is still pressing. This phenomenon is called "dishing". !

NOTE = the larger is the metal structure, the more pronounced the dishing.

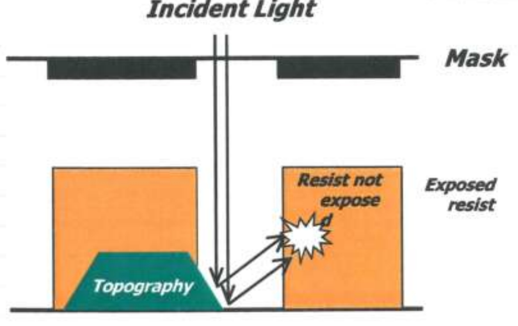
PRE CMP POST CMP

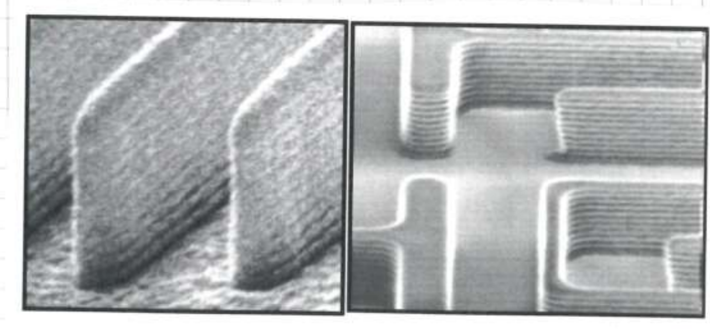
Also: the final profile could be a function of the initial profile. If we have Cu parts "popping up", they will be polished faster than the rest because there the abrasive disk is pressing more (more pressure = faster polishing rate).

This effects can be mitigated slightly, but not completely.

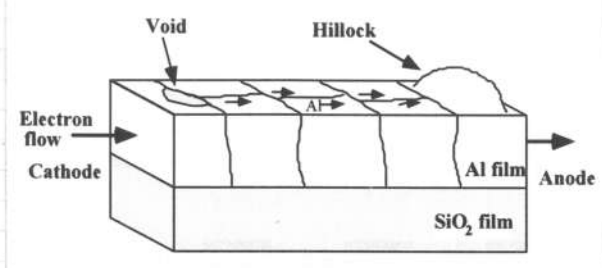
Interconnects are really important because the majority of the power in a chip is dissipated by them, rather than by all the rest.

Examples of scalloped edges due to the formation of standing waves, because of the absence of BARC layer at the bottom:





ELECTROMIGRATION



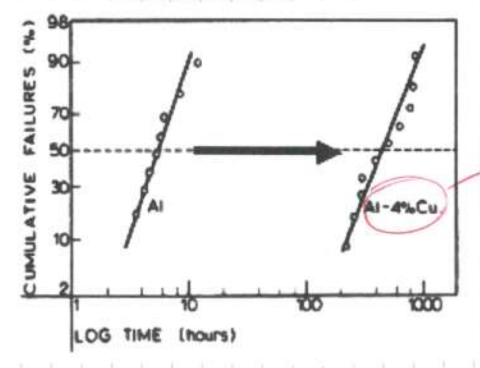
High current densities (0,1 ~ 0,5 MA/cm^2) can cause the movement of Al (and to a lesser degree, Cu too) atoms in the direction of electron flow, because the e^- flowing can "drag" the slightly positive Al atoms. Why slightly positive and not neutral? Because their outermost electrons are shared amongst all other Al atoms (metallic bond), so it's like the Al atoms are slightly positively ionized, making the high electronic current densities capable of attracting and dragging them, creating hillocks and voids (usually at grain boundaries).

Remember that voids and hillocks can also generate because of differences in thermal expansion coefficients between Al and Si.

LOOK PAGE 285, 286.

The failure of the device due to electromigration will inevitably happen sooner or later during the device's lifetime. Obviously this will happen because of it's usage, not because it's just sitting there.

The failure of the metal lines is measured statistically, and the cumulative probability of failures is usually plotted vs time, in a lognormal plot.



Al with 4% Cu
2 orders of magnitude better MTTF than pure Al

The medium time to failure (MTTF), so the time it takes for 50% of the structures to fail, depends on current and temperature as:

constant

$$\left[MTTF = \frac{A}{J^n} e^{\frac{E_a}{KT}} \right]$$

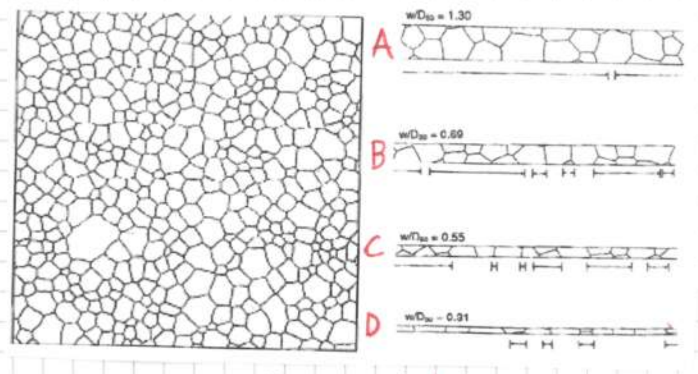
remember the exponent in the current density

So the MTTF has a polinomial dependence on J and exponential with T.

Once we find the exponent n and the activation energy E_a, we can run accelerated tests on our devices, so for example we can run tests at very high T and J, measure the MTTF and extrapolate the MTTF at lower T and J (which can be > 10 years, but we don't have to wait that long thanks to accelerated tests!)

NOTE: in the plot at page 297, just by using Al with 4% Cu we increased the MTTF by 2 orders of magnitude!

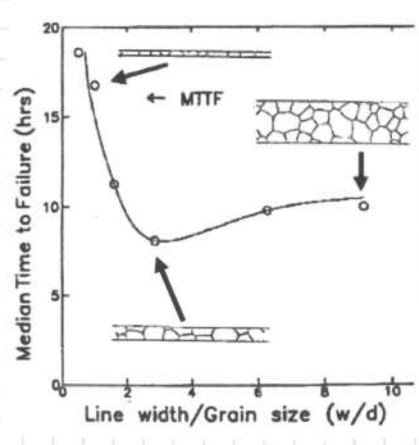
MTTF vs LINE WIDTH / GRAIN SIZE



Suppose we lay down a layer of Al, being polycrystalline it will have grains of different sizes and shapes.

Now take this layer and slice it into thin lines of different dimensions.

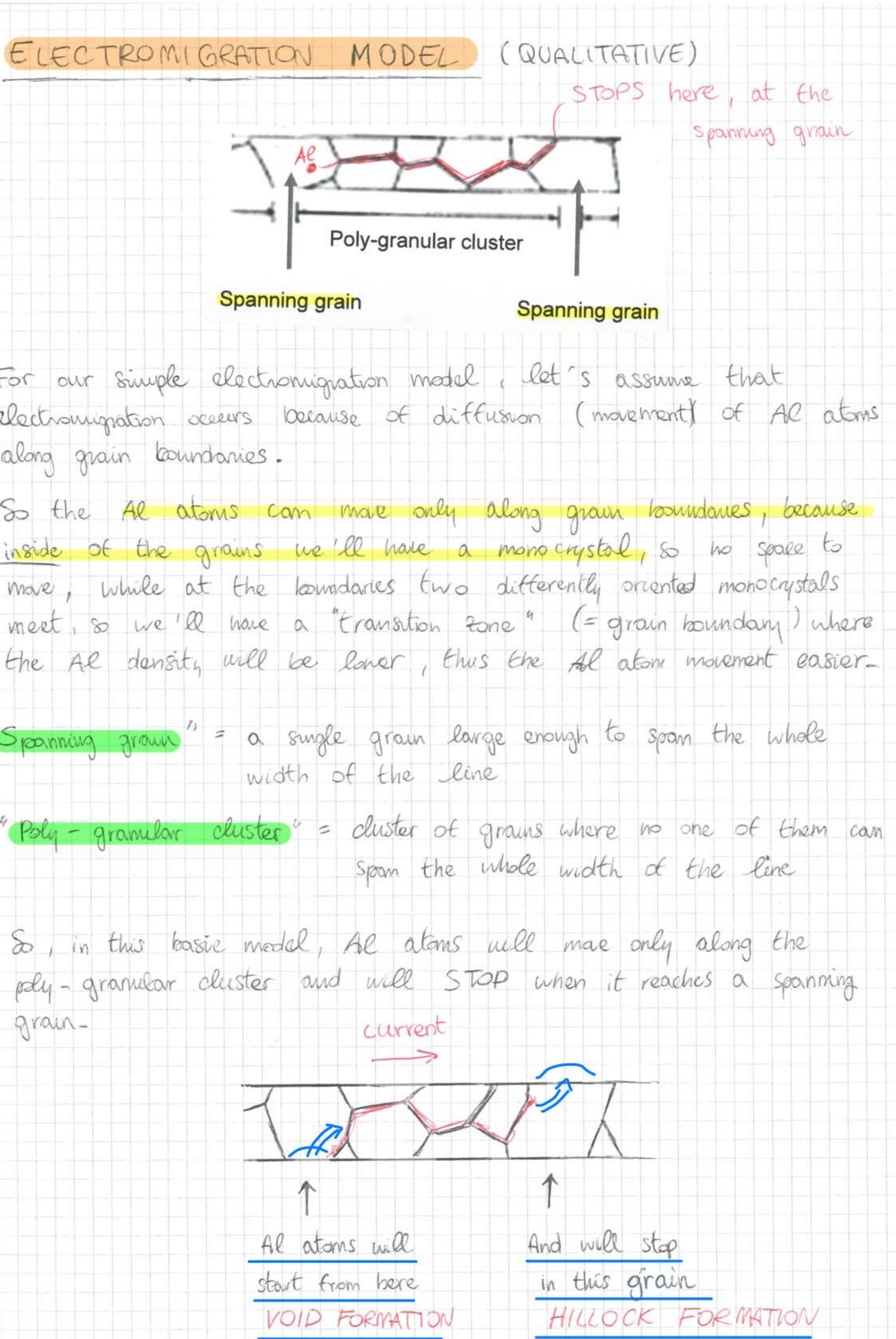
For thick lines (A), we will have multiple grains inside the line.
For thinner lines (B, C) we'll start having some grains spanning the whole width of the line, while still also having on some other spots multiple grains on a certain line width.
For very thin lines (D) we'll get a "bamboo structure", so a line where every grain will be as large as the line width.



If we now take the line width W measured in grain average sizes d, so their ratio w/d, and see the dependence of the MTTF, we can see that it's not a monotonic dependence!

Very small lines fail after a very long time.
Very large lines also fail after relatively long time.
Intermediate lines are the worst. Why?

We can try to explain that by building a simple electromigration model.



ELECTROMIGRATION MODEL (QUALITATIVE)

For our simple electromigration model, let's assume that electromigration occurs because of diffusion (movement) of Al atoms along grain boundaries.

So the Al atoms can move only along grain boundaries, because inside of the grains we'll have a monocrystal, so no space to move, while at the boundaries two differently oriented monocrystals meet, so we'll have a "transition zone" (= grain boundary) where the Al density will be lower, thus the Al atom movement easier.

"Spanning grain" = a single grain large enough to span the whole width of the line

"Poly-granular cluster" = cluster of grains where no one of them can span the whole width of the line

So, in this basic model, Al atoms will move only along the poly-granular cluster and will STOP when it reaches a spanning grain.

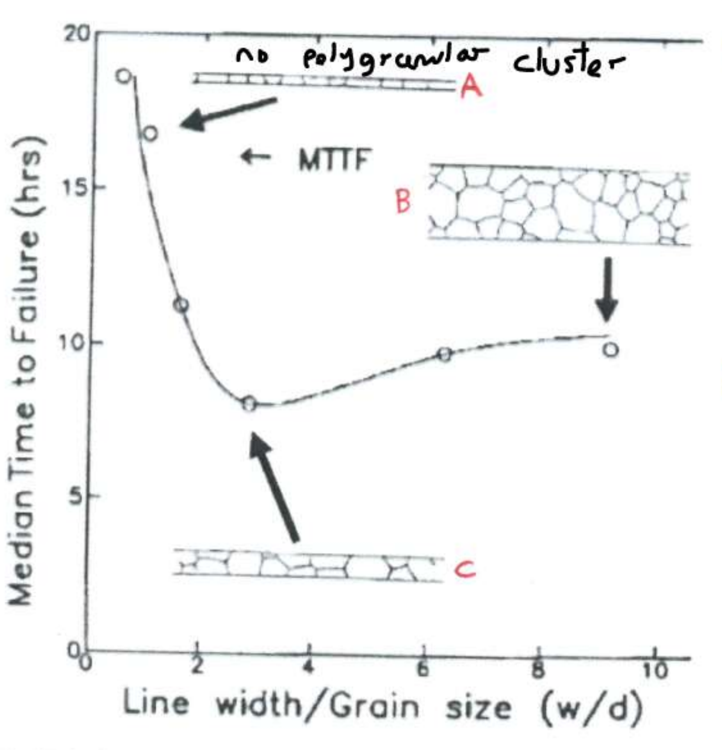
So the grain from which the atoms will start moving away will experience tensile stress, and will form voids. The grain on which they will stop will experience compressive stress, and will form hillocks.

This stress will build up, and so a counter-flux of Al atoms will form because of the gradient of concentration.

In steady state condition the two fluxes will be the same, and nothing will happen. But if we overcome the critical point of stress sustainable by the material, we undergo permanent deformation (voids, hillocks, ...).

NOTE: the longer the poly-granular cluster, the more stress builds up, so the easier is for failure to occurr.

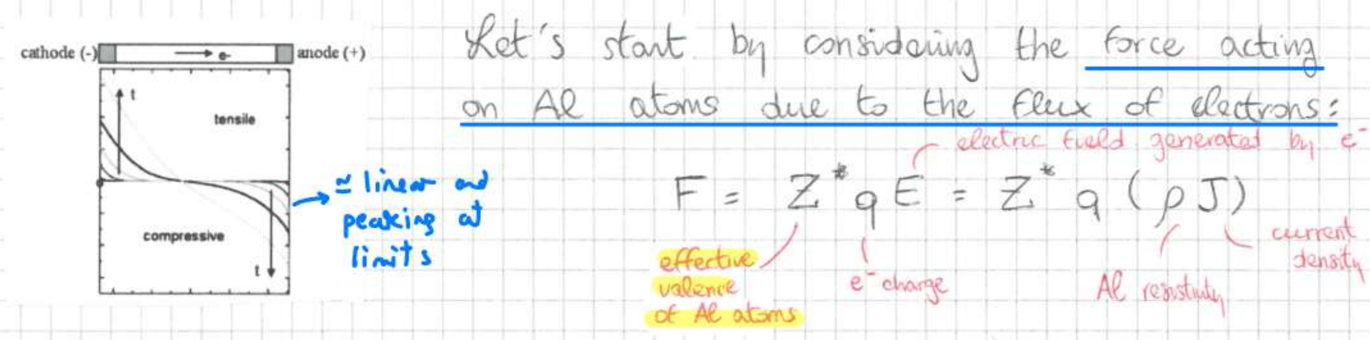
Now, back at the MTTF vs w/d plot:



- the thin lines (A) have very high MTTF because every grain is a spanning grain, so there's no boundaries across the current flow where the Al atoms can move. No electromigration.
- the thick lines (B) won't have any spanning grains, so the Al atoms will move but then won't stop and generate stress.

intermediate lines (C) will have both poly-granular clusters and spanning grains, so the MTTF induced by electromigration will be low.

ELECTROMIGRATION MODEL (ANALYTICAL)



Let's start by considering the force acting on Al atoms due to the flux of electrons:

$$F = Z^* q E = Z^* q (\rho J)$$

- Z^*: effective valence of Al atoms
- q: e^- charge
- E: electric field generated by e^-
- ρ: Al resistivity
- J: current density

Z^* is the effective valence of Al atoms, because the outermost electrons are delocalized, Al atoms have an effective ionic charge

We can write the flux of Al atoms as:

$$\varphi = \mu C F$$

- μ: mobility
- C: concentration
- F: force

but using Einstein's relation, as seen at page , we can write the mobility as

$$\mu = \frac{D}{KT}$$

- D: diffusivity

$$\rightarrow \left[\varphi = \frac{DC}{KT} F = \frac{DC}{KT} Z^* q \rho J \right]$$

this will be the flux of Al atoms along the grain boundaries due to a current flowing

but now also a counter flux will generate due to the presence of stress in the wire, and we can write that flux as

$$\varphi = \mu C \left(\omega \frac{\partial \sigma}{\partial x} \right)$$

- ω: atomic volume of Al atoms
- $\frac{\partial \sigma}{\partial x}$: gradient of stress along the current direction

In stationary conditions, the net flux will be zero:

$$\varphi = \frac{DC}{KT}\left(Z^* q \rho J + \omega \frac{\partial \sigma}{\partial x}\right) = 0$$

$$\rightarrow \left[\frac{\partial \sigma}{\partial x} = -\frac{Z^* q \rho J}{\omega} = -G\right]$$ gradient of stress

integrating: $\sigma = -Gx + C$

So we see a linear dependence on the position of stress, at $x=0$ (the beginning of the poly-granular cluster) we will have maximum tensile stress. At $x = L_P$ (end of the cluster) we will have maximum compressive stress. So we can derive that at $x = L_P/2$ the stress will be zero → boundary condition gives us $C = G L_P/2$ → see previous page

$$\left[\sigma = G\left(\frac{L_P}{2} - x\right)\right]$$

with $$\left[\sigma_{max} = \pm \frac{G L_P}{2} = \pm \frac{Z^* q \rho J}{\omega} \cdot \frac{L_P}{2}\right]$$

($\frac{Z^* q \rho J}{\omega}$: material properties; L_P: length of the poly-granular cluster)

So we have demonstrated that the longer the poly-granular cluster, the higher the maximum stress that builds up at the edges, the more likely σ_{max} will exceed the critical stress and fail.

⇒ Cu in Al tends to be along the grain boundaries, so grain boundaries are full and Al migration is reduced ⇒ higher MTTF