Microelectronic Technologies

Lorenzo Pasquetto, Politecnico di Milano

A.A. 2022-2023

1 CMOS process flow

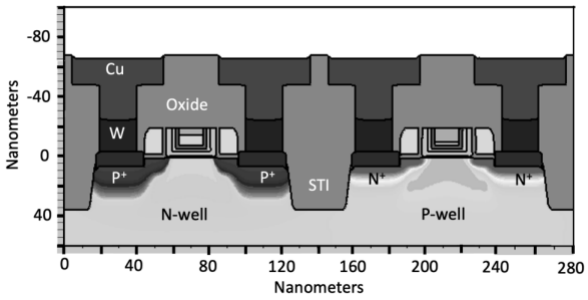

Figure 1: CMOS cross section, image from [1].

1.1 Substrate

Before we begin actual wafer fabrication, we must choose the starting wafer. We have to decide the type (N or P), resistivity (doping level), crystal orientation (ex. Si(100)), wafer size (ex. 300mm) and other parameters. The major choices are **type**, **resistivity**, **orientation**. In most CMOS integrated circuit, the substrate has moderate high resistivity (25-50 Ωcm), therefore a doping level of the order of $10^{15} cm^3$. The active regions on the wafer have doping levels on the order of $10^{17} cm^3$. The other important parameter to specify is crystal orientation. All modern silicon integrated circuits are manufactured today from wafers with (100) surface orientation. The principal reason for that is that the properties of Si/SiO_2 interface are significantly better when the (100) crystal is used.

1.2 Active region formation

Now we have our specific wafer. Before and after each step the surface is cleaned using reagents. Modern CMOS chips integrated millions of active devices (NMOS and PMOS) side by side in a common silicon substrate. This individual active regions are designed to not interact directly with each other except through their circuit interconnection. To isolate the active regions we use the SiO_2. The process of local oxidation is also known as **LOCOS**. A thermal SiO_2 is grown on Si surface by placing wafer in a high temperature furnace. A typical furnace cycle might be 15 minutes at $900C$ in H_2O atmosphere. The oxide layer of SiO_2 is about $40nm$ thick. The wafers are then transferred to a second furnace, which is used to deposit a thin layer of Si_3N_4 (typically $80nm$). This deposition occurs when reactants like NH_3 and SiH_4 are introduced into the furnace at $800C$. The nitride layers deposited are usually highly stressed. This produces a large compressive stress in the underlying Si substrate which can lead to generate defects. In fact, the major purpose of SiO_2 layer under the nitride is to help to relieve this stress. By choosing the proper thickness, the stresses in the two layers compensate each others. The final step is to deposit the photoresist layer in preparation for masking. Since photoresists are liquid at room temperature they are spun onto the wafer. The typical thickness is $1\mu m$. After the spun the wafers are baked around $100C$ in order to drive off the solvent from the layer. The resist is then exposed using a mask, defining LOCOS pattern. Now to remove the Si_3N_4 following the LOCOS pattern we can use to different etching. We can use wet etching (ex. using H_2SO_4 in solution), in this case we have an isotropic etching. To ensure the anisotropy etching is performed in RF plasma chamber. The resist is chemical removed in sulfuric acid. We remain with the following structure.

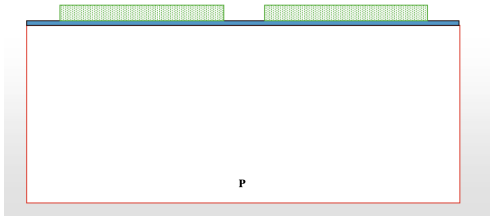

Figure 2: After plasma etching

Following cleaning, the wafer is placed into the furnace in oxidation ambient. This **grows** a thick layer of SiO_2 only between the islands. In fact, the Si_3N_4 layer on the surface prevents oxidation because it is a very dense material and it limits the diffusion of H_2O and O_2. The final result is SiO_2 structures between the islands with thickness of $500nm$. Note that the diffusion of H_2O or O_2 is isotropic, therefore we obtain the following structures. An alternative way for forming isolation

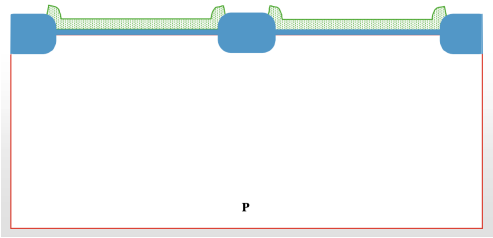

Figure 3: LOCOS

regions is the STI (Shallow Trench Isolation). This is the method actually used. STI actually etches trenches between the active regions and fills it with SiO_2. Such a process eliminates the bird's beak shape of LOCO isolation. The stress related issues, which constrained the thickness of layer in the LOCO case, are relaxed. We can create SiO_2 thickness of $10-20nm$ and Si_3N_4 thickness of $50-100nm$. Starting from the same wafer of photoresist onto $Si_3N_4/SiO_2/Si$ we define the pattern using a mask and then the nitride and oxide layers are etched using fluoride-based plasma chemistry for both materials. The next step is to etch the trenches in the substrate using bromide-based plasma chemistry. In this step is important that the trench walls be vertical to avoid undercutting in the active regions. Now we reach this situation:

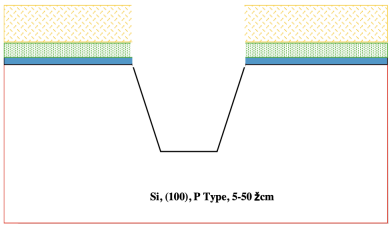

Figure 4: STI trenches formation

The next step is to thermally grow a thin liner oxide on the trenches sidewalls. This thermal growing this layer produces a better Si/SiO_2 interface. The next step is the deposition of SiO_2 all around followed by CMP (Chemical Mechanical Polishing) where the nitride layer serves as a polishing stop. Once the CMP operation is completed Si_3N_4 can be chemically removed. At this stage the wafers are ready for device fabrication. Now we reach this situation:

1.3 N and P well formation

We now return to our CMOS process flow using LOCUS isolation technology. At this stage, the wafers are ready for device fabrication in each isolated region. We want to put in a P type well creating a local substrate for the NMOS transistors and and an N type well to creating a local substrate for PMOS transistors. We use photoresist to pattern where the doping goes. Now that we have define the region for doping, we want to replace some fraction of silicon atoms in P well with

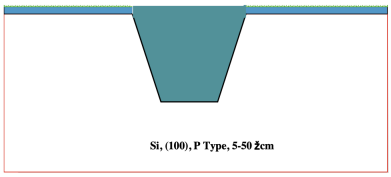

Figure 5: STI

boron atoms. The process is called **ion implantation**. A small version of ion accelerator is used to accelerate the boron ions, shooting them at the silicon wafer. Note that this process is precise, controllable and reproducible. But ion implantation comes at a cost. Boron atoms can easily strike a silicon atom in the lattice creating a damage in the overall lattice. In fact the ion are accelerated at 10-100 keV while the binding energy of Si-Si is $12eV$. This damage must be repaired somehow since the device we want to build require virtually perfect crystalline substrates. In this case the damage can be repaired simply increasing the temperature. We have used the boron and phosphorous atoms

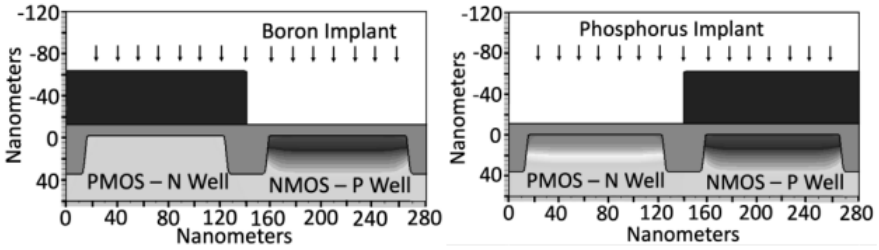

Figure 6: Ion implantation, image from [1].

since we want the wells to be symmetrical and they have roughly the same diffusive coefficient in silicon. Now to adjust the threshold barrier we put N and P regions near to the surface. The purpose of these implants in the channel region is to fine-tune the doping concentration under the gate oxide of the transistor to set the threshold voltage. The original thin oxide on the wafer surface has been exposed to several implants at this stage in the process, which create damage in the SiO_2, so stripping and regrowing a new oxide produces a higher quality gate oxide.

1.4 Gate Oxide and Polysilicon Gate

After regrowing the gate oxide we deposit polysilicon immediately to avoid contamination. Since we don't have deposition techniques that are selective, the usual material is to deposit the material everywhere, do lithography and etch it away in the region where we don't want it. The polysilicon needs to be an N-type conductor for the NMOS and P-type conductor for the PMOS, therefore we need another implantation of the dopants. This polysilicon layers can be used to provide wiring between active devices.

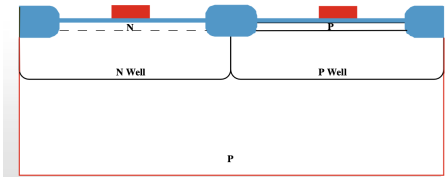

Figure 7:

1.5 Source and drain junction formation

Thirty years ago, MOS devices used in ICs were built with minimum dimension well above one micron and were operated with supply voltage of 5V. From the beginning of 2000 device dimensions have been reduced to $0.25\mu m$. However, supply voltage in circuits have not been proportional reduced. Typical supply voltage were between 2.5 and 3 V. There is great benefit at the system level

to maintaining a standard power supply level because then new ICs are compatible with older parts and do not have to be redesigned (also noise margins remain adequate). However if we reduce the dimension of the channel and we do not scale the voltage, the electric field increase. For example, five volts applied to 2 micron channel length implies an average electric field of $2.5 \times 10^4 V/cm$ while the same voltage applied to 0.25 micron channel length has an average electric field of $10^5 V/cm$. This high electric fields causes hot electron problems. High energy electrons can cause impact ionization which create additional electron-hole pair breaking Si-Si bonds. Such carriers can gain sufficient energy to surmount the energy gap of SiO_2. These carriers injected into the gate dielectrics may be trapped and cause device reliability problems. To overcome this problem is used the Light Doped Drain (LDD). The idea is to produce N^+N^-P profile between the drain and the channel in the NMOS devices and P^+P^-N in the PMOS devices. This allows the drain voltage to be dropped over a larger distance that would be the case if an abrupt N^+P junction were formed. Since many of the deleterious effect of high electric fields in modern MOS devices depend exponentially on the electric field, modest reductions in the field strength obtained through the LDD structures can make a significant difference in device reliability. Another problem regarding the shrinking in size of devices is linked to short channel effects. These effect results when the drain electric field penetrates into the channel region. This results in drain currents that are not controlled by the gate. An important strategy to minimize this effects is the use of shallow junction. Such junction are less susceptible to short channel effects essentially because their geometry minimizes the junction areas adjacent to the channel.

Photoresist is spun on the wafer and a mask is used to protect all the devices except from the NMOS transistors. A phosphorous implantation is done to form the N^- region. Similar process for the PMOS. We now proceed to the second step in forming the LDD regions. Spacers are formed by putting down a thin layer of conformal oxide that is uniform everywhere in the wafer. Because the deposition is conformal, it is much thicker from the top to the bottom along the side walls of the gate. So, if now we do a highly anisotropic etch of the film we just deposited, it will strip the film in the flat regions and leave a narrow spacer at the edges of the polysilicon gates. This clever trick makes the sidewall spacers without using lithography (the process is cheaper). Now the final step is to put in the deep source/drain regions that are self aligned to the edges of the spacer regions. Note that, during the formation of spacer region, the dry etching of flat region would etch the thin oxide region by varying amount. For this reason, the remaining oxide would be wet etch in HF down to the silicon and a thin layer of oxide would be regrown on the surface. A mask exposes the drain/source regions of NMOS device and arsenic implant is performed. Analogue for the PMOS. The last step is to repair the damages caused by ion implantation. A short high temperature anneal (RTA) or flash anneal is used to minimize the amount of dopant diffusion.

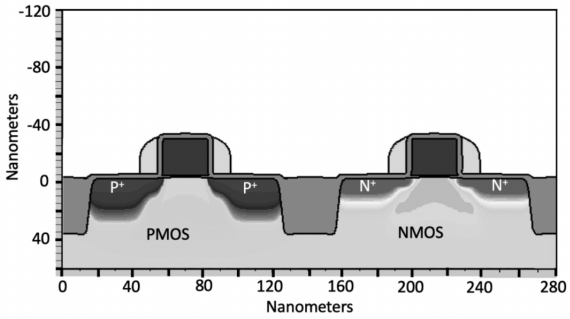

Figure 8: Image from [1]

1.6 Contact Formation

The first step in forming the contacts is to lower the sheet resistance of the source/drain regions by a process called silicidation. By depositing a reactive metal such as Ti using sputtering and performing a short anneal at $350 - 450C$, a metal silicide layer is formed. This thin layer of metal silicide forms in the source/drain regions and gate region and lowers the sheet resistance that connect the contacts to the channel. Since the anneal is done in N_2 atmosphere, if Ti is used, the top surface reacts to form a TiN layer. The two reaction (with N and Si) form a TiN/TiS_2 composite layer. Usually the TiN layer is etched away, leaving only the TiS_2 layers. It is also possible to use TiN as local

interconnection since it is a fairly good conductor.

The final step in our CMOS process involve the deposition and patterning of the two layers of metal interconnect. At this stage the the surface is highly non planar and the depositions of metal in such structures can cause several problems. To planarize the device a fairly thick SiO_2 layer is deposited using CVD. Then, CMP is used to planarize the surface. Photoresist is spun into the surface and a mask is used to pattern the contact holes. The SiO_2 layer is etched in plasma and the photoresist is stripped off the wafer. A very thin layer of TiN is sputtered using CVD. This layer provides good adhesion to the SiO_2. It also acts as an effective barrier layer between the upper metal layers and the lower local interconnection layers. The next step is the deposition of a blanket W layer by CVD and a planarization using CMP. Metal 1 is then deposited usually by sputtering, and defined using resist and mask. This metal is commonly Al with a small percentage of Si and Cu. The Si is used because Si is soluble in Al up to a few percent and if it is not already present in aluminium, it may be absorbed by the Al layer from the underlying silicon level. Cu is added because it helps to prevent a reliability problem known as electromigration in Al film. The process described above is repeated several times to form new plugs and metal layers, needed to realize the interconnections.

The very last step is the deposition of passivation layer (SiO_2 or Si_3N_4) to protect the chip during subsequent handling. The last mask (16) is used to open holes holes in this layer over the bonding pads.

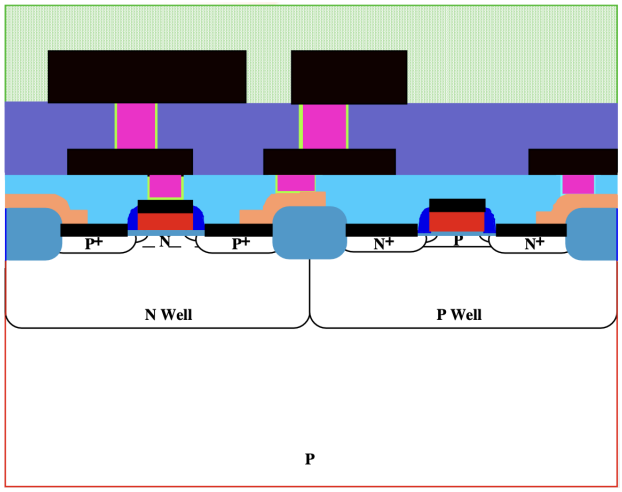

Figure 9: CMOS

2 Silicon Substrate

2.1 Introduction

Almost from the beginning, it was clear that the silicon was the best choice in the production of integrated circuits. The abundance of silicon, the availability of simple techniques for refining it and growing single crystal, the essential properties of Si/SiO_2 interface and the invention of manufacturing techniques based on planar process, all led to the dominance of silicon based device by the early 1960s. The most important concept in solid state physics is the definition of Bravais lattice. It is defined as a discrete set of points with an arrangement and orientation that appear exactly the same, from whichever of the points the array is viewed. This definition is equivalent to say that the Bravais lattice consists of all points with position vector $\mathbf{R}$ of the form:

$$\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$$

where $\mathbf{a}_i$, $i = 1, 2, 3$ are the linear independent vectors that span all the lattice points and n_i, $i = 1, 2, 3$ integer numbers. Silicon has diamond cube crystal lattice, i.e. a face centered cubic lattice with a basis of two Si lattice with a lattice parameter of $a = 5.43$ Å. In crystallography planes are identified by Miller indices (hkl). The choice of a particular plane is very important. In fact, it defines mechanical, electronic and optical properties of the interface. Nowadays the most most used plane is $Si(100)$ but in the past was used also $Si(111)$. The main properties of $Si(100)$ are:

- Low density per cm^2.
- Low oxidation rate.
- Low density of defects.

2.2 Points Defects

Many different types of crystal defects can exist. The first class the one of point defects. We can demonstrate from fundamental principles that the concentration of vacancies (V) and interstitial (I) increase with temperature and:

$$C^*_{Vo}, C^*_{Io} = N_S e^{\frac{S_f}{k_b}} e^{-\frac{H_f}{kT}}$$

where N_S is the number of lattice sites, S_f is formation entropy and H_f is the formation enthalpy. This is valid for I and V in neutral charge state. Also charged point defects can exist. Charged point defects have energy levels within the silicon forbidden region therefore they act as donors/acceptors.

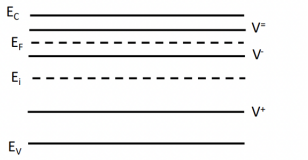

Figure 10: Charged point defects

The V and I concentrations can be written as:

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

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

2.3 Line, Area and Volume Defects

The second type of defects are the so called line defects. A possible line defect is the dislocation. It is a local deformation of silicon lattice. Dislocation are usually defined by nonvanishing Burger vectors. Note that, dislocation are active defects in crystal, that is they can move when subjected to stresses or when excess of point defects are present. The process of climb occurs when excess point defects are absorbed by the dislocation. Under shear stress the dislocation can glide.
The third type of defects are area defects like stacking fault. It is an extra partial plane of atom inserted/removed in the lattice. The last type of defects are volume defects. Usually we call volume defects agglomerate of point defects such as vacancies, precipitate of dopants or impurities. For example, when doping concentrations are pushed too high (beyond the solubility of the dopant in the crystal), precipitation of dopant atoms in amorphous or crystalline cluster is observed. Another example occurs in Czochralski silicon because of the relatively high oxygen impurity level. The oxygen usually incorporated during the process is about $10^{18} cm^{-3}$, far higher than oxygen solubility at normal temperatures. During the cooling process, precipitates are formed. This event can be useful since these damage zones are kept far from the active regions (with annealing processes) and can collect impurities like heavy metals.

3 Lithography

There exist different types of lithography. The actual choice is the so called Optical Lithography. Optical Lithography is similar to photographic printing. The image (in our case the pattern) is transfered to a photoresist solution which coats the wafer. Light is then projected from a light source (with a particular directionality, intensity and wavelength) through a mask that contains the pattern. The light pattern is then reduced by a factor of four or five using optical lens. The illumination is repeated for each one of the chip of the wafer. The photoresist that is exposed to light becomes soluble and is rinsed away, leaving the mask pattern on the photoresist.

High resolution lithography need small wavelength photons. Discharge of high pressure mercury lamp provides a suitable choice of light for lithography, for wavelength between 300nm and 450nm.

Usually filters are used to choose a particular emission line (ex. i-line at 365nm, g-line at 436nm). The need to print even small features drive to short wavelength. Lasers showed to be the best candidates because of high power and spectral purity. Current lithographic systems use Deep Ultra Violet lasers in particular excimer lasers. An excimer is an excited dimer. Two elements that generally don't react if excited can react forming a compound of two elements. When they returns to ground state, they emit a photon in the DUV and molecule breaks up. Nowadays ArF (193 nm) and KrF (248 nm) excimer lasers are used.

Photomasks are high precision plate containing microscopic image of electric circuit/features. As we have already said, the projected light through the mask is then reduced (4x in DUV). To create a Litho-mask we have to use the same techniques used in semiconductor process flows. Masks are made of thin layer of glass on top of chrome features which create the negatives transmitted to the photoresist.

There exist several type of **Exposure systems**. The first one is the contact printing. Here we have direct contact of mask and resist. Therefore we can reduce the diffraction effect having an high resolution (which depends on the mask). The cons are related to the hard contact of mask and resist and therefore damage in both sides and higher density of defects. The other drawback is related to the 1x image transfer. To reduce these cons, proximity printing has been used. A gap of 5-25μm between the mask and photoresist is used. Obviously we have degradation of image due to diffraction effects in near field regime. The resolution power is then reduced:

$$R_{min} = \sqrt{k\lambda g}$$

where λ is the wavelength, g is the proximity gap and k depends on the photoresist.

Nowadays the most used printing method is **Projection Printing**. We can obtain high resolution without having problems related to contact defects. Image reduction is used (5x and 4x) but resolution is limited by diffraction effects. The Fresnel Integral is approximated in Far-field regime (Fraunhofer approximation). The cons regard the finite dimension of focusing lens and therefore the collection of only a subset of diffraction order points.

We have two types of projection printing, step-and-repeat and step-and-scan.

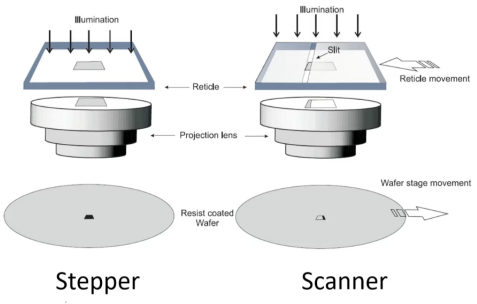

Figure 11: Stepper and Scanner

For step-and-repeat (stepper) the full pattern is transfered in one go to the wafer portion. In the case of step-and-scan the image is transfered through a small slit and the mask is scanned (the wafer and recticle are in movement). Using scanner we can obtain less geometrical aberrations and wider exposure field. The masks for projection printing have a thin transparent membrane on the chrome side to put out of focus any possible dust particles present in projection system.

The image transfered to the photoresist is called aerial image. It will interact with the resist forming the latent image. This image becomes soluble in developer. The transfer pattern, therefore, depends on the photoresist properties:

- Pattern Fidelity and Distribution
- Sensitivity to incident energy
- Resistance to subsequent processes

What we have described until now is the so called positive photoresist in which the exposed areas become soluble in developer. Negative photoresist become insoluble in developer. The three main components of photoresists are:

- Inactive resin

- Solvent, to regulate the viscosity
- PAC, photo active compound

The most common photoresist for i-line and g-line is DQN, Diazoquinone (PAC) and Novolac (a polymer). After light exposure, DQ loses N_2 and to minimize the energy the system rearranges forming the Ketene which is soluble in basic developer. DQN for DUV lithography has two main problems. The first concerns the high resist absorption of light which leads to a non complete penetration throughout the resist. The second is linked to small sensitivity at DUV regime. Photoresist today are based on CAR (Chemical Amplificator).

The incoming photons react with Photo-Acid Generator, creating acid molecules. These acid molecules act like catalyst during subsequent resist bake to change resist properties in the exposed regions. Acid molecules are regenerated after each reaction and may participate to hundred of other reactions. The typical exposure dose for i-line photoresist is in the order of $100mJ/cm^2$ while for DUV 20 mJ/cm^2.

The photoresist is composed of PAG and blocked polymer (not soluble). After light exposure, the wafer is baked providing the activation energy for acid-protective group reactions. The blocked polymer becomes soluble.

The generation of acid molecules can cause contamination. The acid molecules can react with airborne contaminant like amines blocking the reaction at the surface.

An important parameter in PR is the contrast. It is the ability for the PR to distinguish light to dark areas in aerial images. The ideal case (of resist remaining) is a step function for a threshold dose ($20mJ/cm^2$). Before the threshold the resist does not react while after becomes soluble. In the

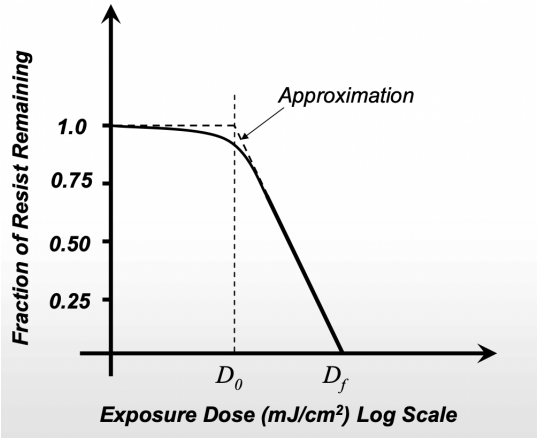

Figure 12: Contrast is the slope of the steep part

realistic case D_o is the dose at which the exposure begins to have effect and D_f the dose at which the exposure is complete.

$$\gamma = \frac{1}{\log_{10} \frac{D_f}{D_o}}$$

For DUV contrast is 5-10.

3.1 PhotoResist Process Flow

- Dehydration Bake to remove any water contamination and HDMS (hexamethyldisilane) application to increase the adhesion of PR.
- Photoresist is applied and then the wafer is spun. The final thickness depends on the rotatory velocity and resist viiscosity.
- Pre-Expo Bake performed at 100°C to remove the solvent within the PR and improve adhesion.
- Alignment and Exposure to i-line, g-line or DUV.
- Post-Expo Bake particularly important for PAG for DUV photoresist.
- Development Step where the soluble parts are removed using TMAH. Rinsing the wafer with water stops the delevopment.

- Final Bake to harden the resit. It is also done to modify the final dimension of resist features.

Due surface non planarity, oblique reflection can impact non exposed region. An other problem is linked to surface reflectivity causing scalloped edges, this problem is called standing waves. In order to avoid this problem, a thin layer of anti-reflective material is deposited before the PR application.

One of the most important metric to judge a photo-lithography process is **registration**. In order to have good alignment, alignment marks are generated ion the wafer at previously mask. Litho systems are able to recognize these marks an align the exposure mask. Typical registration errors are divided into two categories:

- Intra-field: errors within exposure field
- Inter-field: errors from field to field over the wafer.

4 Projection Lithography

Why do we have resolution limits? Mask features are comparable with light source wavelength. A diffraction pattern is form at the lens plane. The point is that, the lens is able to collect only a portion of diffraction pattern to reconstruct the image at the wafer plane. High frequency components are placed far away from the center of the lens.

Under the same light condition, the most important parameter is the pitch length. Smaller is the pitch higher is the spatial frequency and the diffraction angle becomes larger:

$$sin\phi = n\frac{\lambda}{p}$$

In order to resolve small pitch we need to collect many diffraction orders. This parameter is the so called Numerical Aperture. It is the maximum number of diffraction orders that can be capture by the projection lens. The image is then focused on the wafer. The semi-angle between the projection lens and wafer is α. The Numerical Aperture (NA) is:

$$NA = n * sin(\alpha)$$

where n is the refractive index of the medium. Another important parameter is the **partial coherence** σ. It defines the type of illumination at the mask. For perfect on-axis illumination the semi-angle at the source α_o is 0 and σ is zero. The partial coherence is then defined as:

$$\sigma = \frac{sin(\alpha_o)}{NA}$$

Let us consider the transmission function of the mask $O(x, y)$. The electric field after the mask, using Fresnel integral in far-field approximation, can be calculated with the Fourier transform of the transmission function:

$$\hat{O}(f, g) = F(O(x, y))$$

The transmission function is then reconstructed by spherical lens which perform an anti-Fourier transform:

$$E(x, y) = F^{-1}(P(f, g)\hat{O}(x, y))$$

where $P(x, y)$ is the pupil function:

$$P = 1 \quad \text{if} \quad \sqrt{f^2 + g^2} < \nu_c, \qquad \text{zero otherwise}$$

The ν_c is called cut-off frequency and is linked to the numerical aperture:

$$\nu_c = \frac{NA}{\lambda}$$

The reconstruction of electric field is limited to low frequency (the pupil lens can be considered as low pass filter). Therefore we loose information about the transmission function due to finite lens. Let us consider a periodicity of the pitches p with width d. The mask spectrum will be delta functions modulated by a sinc function:

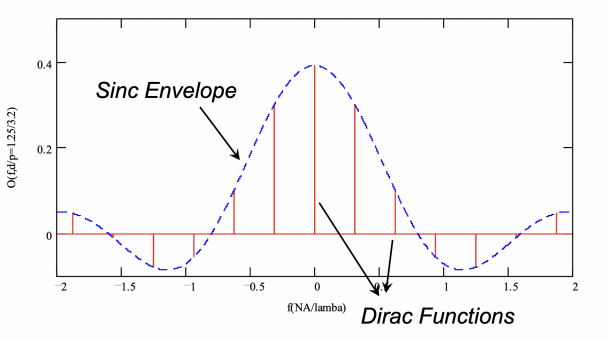

Figure 13: Mask spectrum after the mask. Note that we can collect only the peaks from -1 to +1.

$$\hat{O}(\hat{f}) = \frac{\hat{d}}{\hat{p}} sinc(\pi \hat{f} \hat{d}) \sum_{n=-\infty}^{+\infty} \delta(\hat{f} - \frac{n}{\hat{p}})$$

where $\hat{f}$, $\hat{d}$ and $\hat{p}$ are the normalized frequency, normalized feature dimension and normalized periodicity. Taking the square modulus of the anti-Fourier transform of the mask spectrum, we can obtain the intensity at the wafer. We must restrict our domain between -1 and +1.

$$I(\hat{x}) = (\frac{\hat{p}}{\hat{d}})^2 |1 - \sum_{m=1}^{m=m_o} 2sinc(\frac{m\hat{d}}{\hat{p}}) sin(\frac{2\pi m \hat{x}}{\hat{p}})|^2$$

where:

$$m_o < \hat{p} < m_o + 1$$

Example 1

$$\hat{p} = p \cdot \frac{NA}{\lambda} > 1$$
$$\hat{p} = 2.5, m_o = 2$$

The image is formed from 5 diffraction orders. We can have the following intensity image at the surface:

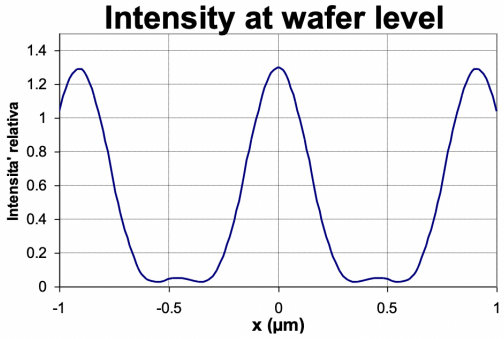

Figure 14: Intensity at the surface with 5 diffraction orders -2, -1, 0, 1 and 2.

Example 2 Let us consider a normalized pitch less than 1.

$$\hat{p} < 1 \quad \text{therefore} \quad \frac{p}{\lambda} \cdot NA < 1$$

This means that $m_o = 0$ and the intensity is not modulated:

$$I(\hat{x}) = (\frac{d}{p})^2$$

Example 3 The limit case is that in which $\hat{p} = 1$ and therefore $m_o = 1$, In the same condition of λ and NA this correspond to take the minimum value of p without losing modulation:

$$\hat{p} = p \cdot \frac{NA}{\lambda} = 1$$

$$p_{min} = \frac{\lambda}{NA}$$

The Resolution is nothing else but:

$$R = \frac{p_{min}}{2}$$

therefore

$$R = \frac{1}{2} \cdot \frac{\lambda}{NA}$$

The resolution depends only on the period p (which defines the spatial frequency) and not on the slit d. The limit resolution is also affected by partial coherence of light. Since light arrives at the mask with a semi-angle α_o, we have to take in account all the images arising from different angle within α_o.

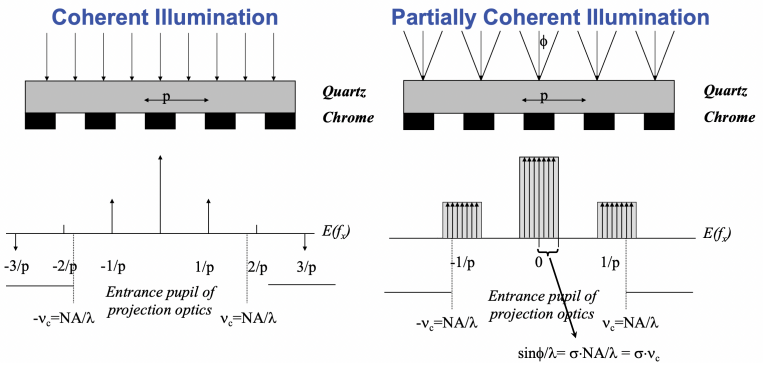

Figure 15: Coherent Illumination vs Partially Coherent Illumination

The Resolution becomes:

$$R = \frac{1}{2} \frac{\lambda}{NA \cdot (1 + \sigma)}$$

The resolution can be enhanced using partially coherent light up to $R = \lambda/4NA$.

4.1 Quantification of Image Quality

The first parameter is the **Modulation Transfer Function** MTF. It is used to determine the resolution limit of optical imaging. It is use-full to calculate the fraction of energy transmitted by the optical system at a particular frequency. Another important parameter is the **contrast**, the difference between the maximum and minimum intensity divided by they sum. The **Depth of Focus** depends on the specification of a particular optical system. It is the maximum variation on vertical direction that the system can tolerate before going out of specification. The DoF can be approximated using Rayleigh criterion in which the maximum difference in the optical path between the on-focus and off-focus ray is a quarter of the wavelength:

$$DoF = \frac{\lambda}{2NA^2}$$

To recap for ideal optics:

$$R = \frac{1}{2} \frac{1}{(1 + \sigma)} \frac{\lambda}{NA}$$

$$DoF = \frac{1}{2} \frac{\lambda}{NA^2}$$

while in real systems:

$$R = k_1 \cdot \frac{\lambda}{NA}$$

$$DoF = k_2 \cdot \frac{\lambda}{NA^2}$$

where k_1 (theoretical limit 0.25) is called the lithography aggressiveness and k_2 is a process coefficient dependent on feature shape, coherence and aberrations.

4.2 Resolution Enhancement Techniques

$$R = k_1 \frac{\lambda}{NA}$$

- Off-Axis Illumination: using non perpendicular photon source, therefore partially coherent illumination, we can collect non zero diffraction orders. The diffraction orders are rotated and we can collect -1 and 0 or +1 and 0.

- Destructive interference between the light traveling in the quartz and in the shifter material. This produces nodes in the intensity figure and a more well defined image. The spectrum after the mask:

$$\hat{O}(f) = [(1+t)\frac{\hat{d}}{\hat{p}} \cdot sinc()\hat{f}\hat{d} \cdot \sum_{n=-\infty}^{\infty} \delta(f - \frac{n}{\hat{p}})] - t\delta(\hat{f})$$

  The non zero orders are enhanced (in intensity) by a factor $(1+t)$.

- Optical Proximity Correction: rounded corners and short lines are usually affected by diffraction distortion. Note that in order to construct an orthogonal edge we need several diffraction orders (high spatial frequency). If we now that we can collect only a certain amount of diffraction orders we can modify the mask pattern in order to have edges and short lines.

- As we have seen the numerical aperture depends on the diffraction angle and on the refractive index of the medium (air, liquid or vacuum):

$$NA = n * sin(\phi)$$

  The resolution can be written:

$$R = k_1 \cdot \frac{\lambda}{n \cdot sin(\phi)}$$

$$R = k_1 \cdot \frac{\lambda/n}{sin(\phi)}$$

  We can increase the refractive index to increase the resolution power. This technique is called Immersion Lithography. This technique has also advantages in terms of Depth of Focus. Smaller the angle difference in lens and medium, the smaller is the optical path difference. From the Snell's law:

$$n_{lens} \cdot sin(\theta_{lens}) = n_{medium} \cdot sin(\theta_{medium})$$

  Hence:

$$sin(\theta_{medium}) = \frac{n_{lens}}{n_{medium}} \cdot sin(\theta_{lens})$$

  The refractive index for the glass is $n = 1.34$. Using water as medium, instead of air, the difference of the angle decreases and the depth of focus increases.

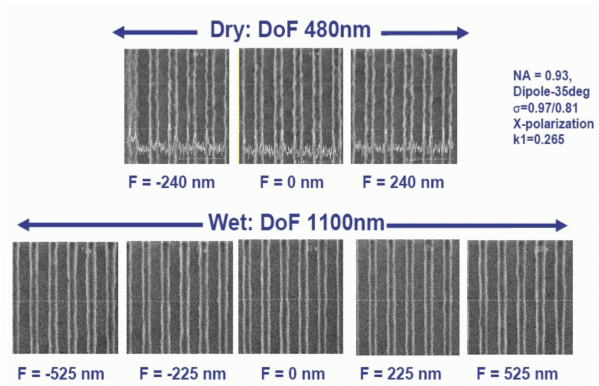

Figure 16: Dry vs Wet Lithoraphy

5 Diffusion

Diffusion is the redistribution of atoms from region of high concentration of mobile species to low concentration region. It happens at all temperature but it depends exponentially on temperature. In IC manufactures we have two phase. Pre-deposition in which doping proceeds by an initial pre-deposition step to introduce the require dose of dopant into the substrate. A subsequent drive-in anneal redistributes the dopant giving the required sheet and surface concentration.
Dopants are soluble in bulk silicon up to a temperature where they precipitate. The two important equations to describes diffusion are the Fick's laws:

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

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}(D\frac{\partial C}{\partial x})$$

Now suppose that the diffusivity D doesn't depend on x, the steady state general solution is:

$$C = a + bx$$

Now suppose a delta function dopant in the middle of lightly doped region. The boundary condition becomes:

$$C \xrightarrow{x=0} \infty \quad \text{and} \quad C \xrightarrow{x<0} 0$$

while:

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

In this case the solution of the second Fick's Law becomes:

$$C(x,y) = \frac{Q}{2\sqrt{\pi Dt}} \cdot exp(-\frac{x^2}{4Dt})$$

Let us now consider $x = 0$ as the surface and, using low energy ion implantation, a delta function of concentration at the surface. We have to multiply the solution by a factor of two:

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

In addition to gaussian profile, the other solution that is useful in silicon processing is the case in which we consider diffusion from an infinite source of dopant. This might corresponds to put an heavily doped epitaxial layer on lightly doped wafer.

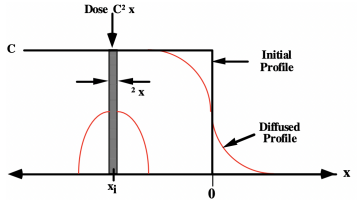

Figure 17: Infinite source at $x < 0$.

The boundary conditions becomes:

$$C(x,t) = 0 \quad \text{for} \quad x > 0$$

$$C(x,t) = C \quad \text{for} \quad x < 0$$

The solution can be seen as the sum of multiple slice Δx_i having gaussian solution. The solution becomes:

$$C(x,t) = \frac{C}{\sqrt{2\pi Dt}} \sum_{i=1}^{n} \Delta x_i exp(-\frac{(x-x_i)^2}{4Dt})$$

After some manipulation, the concentration becomes:

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

Remember that the error function is defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x exp(-\eta^2) d\eta$$

Therefore, the concentration becomes:

$$C(x,t) = \frac{C}{2}(1 - \text{erf}(\frac{x}{2\sqrt{Dt}}))$$

or better:

$$C(x,t) = \frac{C}{2} \cdot \text{erfc}(\frac{x}{2\sqrt{Dt}})$$

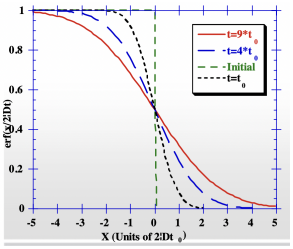

Figure 18: Diffusion from infinite source within the substrate.

Few consequences about the solution. Symmetries about mid-point allows constant surface concentration to be derived. Moreover dose beyond $x = 0$ continue to increase with the time.

Note that the diffusivity in intrinsic doping condition depends on temperature following an Arrhenius law:

$$D = D_o exp(-\frac{E_a}{kT})$$

If we have multiple diffusion step, the total effective $D \cdot t$ is:

$$Dt_{eff} = \sum Dt = D_1 t_1 + D_2 t_2 + ...$$

Since D depends exponentially on the temperature, only high temperature diffusion steps are important.

5.1 Correction to Fick's law

The higher mobility of electrons and holes compared to ions leaves behind charged donors and acceptors. An electric field is settled up. This effective field drag donors and acceptors to. We can derive the corrected Fick's law:

$$F = -hD \cdot \frac{\partial C}{\partial x}$$

where:

$$h = 1 + \frac{C}{\sqrt{C^2 + (n_i)^2}}$$

Another correction comes from diffusivity dependence on concentration. Second Fick's law stands in the form:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x} D(T,C) \cdot \frac{\partial C}{\partial x}$$

Dopants have different solubilities in different materials and so they redistribute at the interface until the chemical potential is the same in both sides of the interface. The ratio of equilibrium dopant concentration on each side of the interface is defined as the segregation coefficient.

The difference in dopant's solubility in each phase drives a diffusion flux until the chemical potential is equalized.

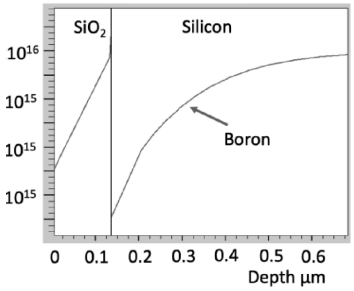

Figure 19: Boron is depleted from the bulk into the oxide due to segregation.

5.2 Diffusion mechanism

Dopants diffusivity is strongly related to the presence of point defects. It seems clear that a vacancy near a dopant atoms provides a mechanism for the dopant atom to hop the adjacent site. At the same time an interstitial can kicked out an atom from its lattice position. This interstitial/dopant pair can travel along bonds direction. The last mechanism is called **interstitial assisted diffusion**. Let us consider an oxidation, injection of O_2 atoms which consumes vacancies and creates interstitials. Experimentally is observed an enhancement of Boron and Phosphorous diffusion while Antimony diffusion is quenched. This suggest a B, P interstitial assisted diffusion and Sb vacancy assisted diffusion.

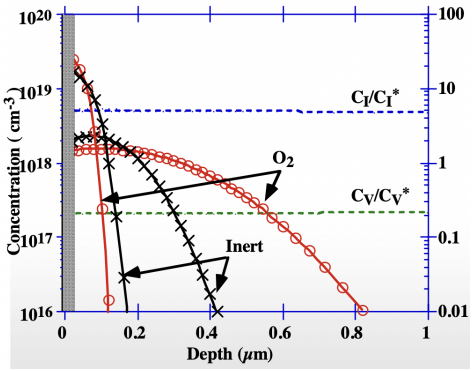

Figure 20: Oxidation can enhance or retard the diffusion.

After this consideration, we can write the effective diffusivity as:

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

where f_I and f_V is mechanism fraction (weight of interstitial/vacancy assisted) while C_I and C_V are the interstitial/vacancy concentration. For example in the case of Antimony $f_I = 0.02$ and $f_V = 0.98$.

Let us consider interstitial assisted reaction:

$$I + A \longleftrightarrow AI$$

The concentration of AI will depend on the concentration of atoms and interstitial with a reaction constant k:

$$C_{AI} = kC_A \cdot C_I$$

The Fick's law for the AI specie is:

$$F_{AI} = -d_{AI} \frac{\partial C_{AI}}{\partial x}$$

$$F_{AI} = -d_{AI}(kC_A \frac{\partial C_I}{\partial x} + kC_I \frac{\partial C_A}{\partial x})$$

We know, from the second Fick's law that:

$$\frac{\partial C_A}{\partial t} = -\frac{\partial F_{AI}}{\partial x}$$

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x}(d_{AI}(\frac{C_{AI}}{C_A}\frac{\partial C_A}{\partial x} + \frac{C_{AI}}{C_I}\frac{\partial C_I}{\partial x}))$$

Assume to neglect C_I :

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

where the effective atomic diffusivity is:

$$D_A^{\text{eff}} = d_{AI} \cdot \frac{C_{AI}}{C_A}$$

The last diffusivity is the same we have calculated previously n term of mechanism fractions:

$$D^{\text{eff}} = D^* f_I \frac{C_I}{C_I^*}$$

The diffusion is:

$$F_{AI} = -d_{AI}\frac{\partial C_{AI}}{\partial x} = (\frac{C_{AI}}{C_I}\frac{\partial C_I}{\partial x} + \frac{C_{AI}}{C_A}\frac{\partial C_A}{\partial x}) = -d_{AI}C_{AI}(\frac{1}{C_I}\frac{\partial C_I}{\partial x} + \frac{1}{C_A}\frac{\partial C_A}{\partial x})$$

$$F_{AI} = -d_{AI}\frac{C_{AI}}{C_A} \cdot C_A \cdot (\frac{1}{C_I}\frac{\partial C_I}{\partial x} + \frac{1}{C_A}\frac{\partial C_A}{\partial x})$$

$$F_{AI} = -D_A^{\text{eff}} \cdot C_A(\frac{1}{C_I}\frac{\partial C_I}{\partial x} + \frac{1}{C_A}\frac{\partial C_A}{\partial x})$$

Finally we can write that:

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x} D_A^{\text{eff}} \cdot C_A \cdot \frac{C_I}{C_I^*}\frac{\partial}{\partial x} ln(\frac{C_I}{C_I^*} \cdot C_A \cdot \frac{n}{n_i})$$

The last expression links microscopic parameters to macroscopic measurable parameters.

In order to measure the diffusion we use secondary ions mass spectroscopy. Using ion bombardment, atom are dislodged and analyzed using a mass spectrometer. Another method to measure the doping is Scanning Spreading Resistance Measure. AFM is used to measure the conductivity of a cut crystal. All these method are destructive.

6 Ion Implantation

Ion implantation has been the most dominant doping technique for 50 years. In this process the ions are accelerated towards 1-100keV and are smashed into a crystalline semiconductor substrate. At first glance, it seems strange that such a energetic process it is still used nowadays for very precise doping. As we have already said, ion implantation offers a very precise mean to introduce specific dose of dopants in semiconductor substrate. The basic requirement is an ion source from which the charge particle are accelerated. Ion source can be composed of gas source (ex. BF_2 atom in gas phase) which can be ionized with energetic electrons. The ions are extracted by a voltage bias and filtered using a mass analyzer. Tuning the magnetic field and the slit position it is possible to select only a particular ratio between mass and charge:

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

we can control the extracting bias, the curvature by positioning the slit and the magnetic field. The ions are then accelerated through a linear accelerator and neutral atoms are filtered using a deviator. The ions undergo an x-y electrostatic deflection which scans the beam into the wafer. Charged effects can become an issue, particularly for insulator layer such as SiO_2. To avoid charged surface, which can deplete other ions, a low energetic electron beam is used.

Ion implantation is a random process. Once ions arrive to the substrate (or masks) every ion follows a random path. Concentration distribution can be modelled, in first approximation, on a gaussian centered in R_p called projected range with a straggle ΔR_p (standard deviation):

$$C(x) = C_p \cdot exp(-\frac{(x - R_p)^2}{2\Delta R_p})$$

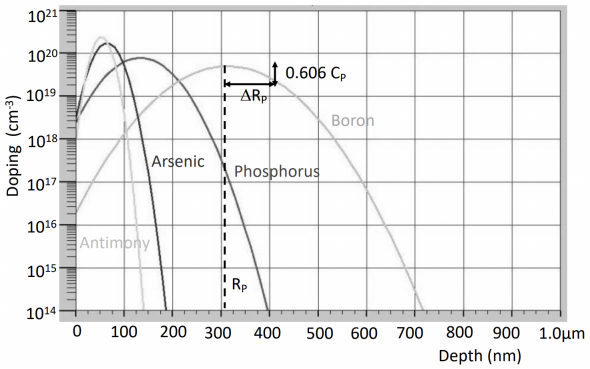

Figure 21:

Actual concentration can be approximated as gaussian only near the peak. For a better description we also must include higher order momentum (Kurtosis and Skewness). Ions can be implanted in some zones masked by a masking material. In order to avoid ion implantation under these zones, masking material must be characterized by a critical thickness and a straggle. The dose penetrating the mask will be the integral from x_m (the thickness of masking layer) to infinity of the concentration:

$$Q_p = \frac{Q}{R_p\sqrt{2\pi}} \int_{x_m}^{\infty} exp(-\frac{(x-R_p)^2}{2\Delta R_p^2})dx$$

After ion implantation, it is common to anneal the sample to obtain the final implant profile. If the implant is shallow enough the implanted profile can be considered as delta function. As we have seen, the concentration profile will be:

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

in one side gaussian approximation. If the implant is deep enough, we must consider two side gaussian, the solution will be the convolution of two gaussian:

$$C(x,t) = \frac{Q}{\sqrt{2\pi(\Delta R_p^2 + 2Dt)}} \cdot exp(-\frac{(x-R_p)^2}{2(\Delta R_p^2 + 2Dt)})$$

The initial $G(R_p, \Delta R_p)$ becomes $G(R_p, \Delta R_p + 2Dt)$

Semiconductor structures are crystalline structure with a regular array of atoms. This leads to the existence of planar and axial channels that have dramatic effects on ion implantation. The implanted region is much more deep than expected. In order to avoid channel effects we can simply tilt the sample, deposit an amorphous layer on the surface or implant directly on amorphous substrate.

6.1 Atomic description

From atomistic point of view, the trajectory of ions in silicon substrate is completelly deterministic. The are two stopping forces: nuclear and electronic.

$$\frac{dE}{dx} = -N(S_e(E) + S_n(E))$$

$$R_p = \int_{E_o}^{0} -\frac{1}{N(S_e(E) + S_n(E))} dE$$

The nuclear stopping force can be modelled as a coulomb interaction between positive ions and positive nucleus. It can be described as:

$$V(r) = \frac{q^2 Z_1 Z_2}{4\pi\epsilon_o r} \phi_{corr}$$

The electronic stopping force is more complicated. It can be modelled as a viscous stopping force (the ions excited substrate electrons of the conduction band creating drag forces since the substrate

is supposed to be a dielectric) and a local electronic stopping (from the interaction of orbitals). Both depends on velocity and we can write:

$$S_e(E) = k \cdot \sqrt{E}$$

Note that, the typical energy to dislodge silicon atoms is $15eV$. It can create an interstitial-vacancy pair. The energy used in ion implantation is few keV, therefore damage is produced in the silicon crystal during the implantation. For every ion, n silicon atoms are dislodge from their sites and some of those can recombine. Every new ion creates an additional damage that can be written as:

$$\Delta n(x) = n f_{rec}(1 - \frac{N}{N_\alpha})$$

where N is the pre-existing damage and N_α is the amorphizing damage. If silicon has been amorphized during ion implantation, crystal structure can be reconstructed using Solid Phase Epitaxy (SPE). Crystal can reconstruct using undamaged portion as seed at 600°C for few minutes. The SPE can be quicker if the silicon is doped.
The interstitials and vacancies generated by implanted ions are generated in equal numbers so if they recombine we have a perfect crystal. But we must take in account also the implanted species. They must sit on a lattice site (substitutional position) to be activated. All this leads to a silicon interstitial after the substitutional implantation. This is called "+1 model".

Therefore in good approximation, all damage recombine leaving an interstitial atom behind. Upon further annealing excess interstitials tends to agglomerate into a small rod clusters on the (311) surface along the [110] direction. Depending on the amount of damage, these rods can disappear by evaporation or can form dislocation loops.
Until now we have seen that annealing after implantation is needed to activate dopants, to reconstruct the crystal and to dissolve defects. But at high temperature diffusion occurs. Diffusion of particular species can be enhanced by interstitial from implantation. This phenomena is called Transient Enhanced Diffusion. We have seen that the diffusivity at atomic scale depends strongly on the type of dopant:

$$D^{eff} = D_o(f_I \frac{C_I}{C_I^*} f_V \frac{C_V}{C_V^*})$$

In IC manufactures the 311 cluster must be removed first because they are crystallographic defects but also because they can inject interstitials near the surface. This enhances the diffusivity of interstitial driven species, such as boron. The 311 rods can last minutes driving the TED for long time. In the following image we can see the interstitial density at different instants. After 0.1 seconds the rods are formed and the $C_I = 10^{13} cm^{-3}$. Actually the equilibrium interstitial concentration is:

$$C_I^* = N e^{\frac{S_f}{k_b}} e^{-\frac{H_f}{k_b T}}$$

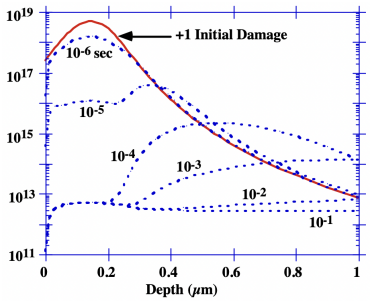

Figure 22:

The presence of injection enhance the interstitial concentration by a factor of 10^5!

7 Backend

With backend we mean the wiring between active regions. Most of the innovation in backend material and processes has been aimed at three issues:

- Reducing the resitance R of the wires.
- Reducing the parasitic capacitance C that exists between wires in complex backend structure.
- Increasing the number of interconnect levels.

Let us consider the following figure. The line resistance R is:

$$R = \rho \frac{L}{WH}$$

Treating the capacitors simply as parallel plate capacitor, the total capacitance derives from the dielectric layer and the two metal wires:

$$C = K_{ox}\epsilon_o \frac{WL}{x_D} + K_{ox}\epsilon_o \frac{HL}{L_s}$$

The time delay of the signal can be expressed as:

$$\tau = 0.89 K_i K_{ox} \epsilon \rho L^2 (\frac{1}{HL_s} + \frac{1}{Hx_D})$$

As technology scales down also H, W, L_s and x_D scale down. Let us take all these parameters as minimum features F_{min}:

$$\tau = 0.89 K_{ox}\epsilon_o \rho \frac{L^2}{F_{min}^2}$$

Therefore to minimize the delay we need low K dielectrics and low resistivity.

Local interconnects: silicides are used as or with local interconnects to reduce the sheet resistance of poly-gates and allowing low resistance contacts with source/drain. Silicides are composed of metals and silicon (ex. $TiSi_2$). These material are characterized by low resistivity, high temperature stability (expansion follows silicon), easy to plasma etch and good compatibility with silicon and other materials. Silicide layer formations are generally from direct deposition or reaction on surface. In reaction layer formation metal and amorphous silicon are deposited by sputtering. Silicon is deposited and patterned before metal deposition. After an annealing process metal and silicon can reacts forming the silicide. Another possibility is the formation of self alignment silicide. The metal is directly deposited by sputtering on the exposed region (after stripping any oxide from selected region). The wafer is annealed and silicide is formed of selected region. Unreacted material are stripped off.

An example of Self-Alignment Silicide is $TiSi_2$. The native screen of silicon dioxide is removed by wet etching using HF. Titanium is sputtered and the wafer is annealed at 700°C in N_2 environment. Ti reacts with silicon at the bottom forming $TiSi_2$ and with nitrogen at the surface forming a titanium nitride TiN layer. Metallic titanium and titanium nitrogen are stripped off using selective etching. Performing the annealing at 700°C the silicide is in C49 phase. To decrease the resistivity an 800°C is performed to obtain C51 phase.

Contacts: provide low resistance connection between active device and the first metal layer. Metal semiconductor junction can vary from rectify to Ohmic behaviour. The rectify behaviour comes from the Schottky barrier. In this regime the conduction comes from thermoionic emission. Ohmic behaviour can be obtained if operating in tunneling regime. Therefore we need heavily doped semiconductor. In the early stages of IC manufactures, contacts and interconnects were made of Al in direct contact with silicon. The choice of Al arise from its good adhesion and low resistivity. The problems in using aluminum come from its solubility in Si (1% at 500°C). This forms spickes in the silicon substrate.
Nowadays contacts and interconnections are made of tungsten. In a typical process flow, pre metal dielectric is deposited and planarized. Contact plug is etched and Ti and TiN are deposited to improve adhesion and acts like barrier to prevent WF_6 diffusion. Tungsten is then deposited by CVD or PVD. The system is planarized by wet etching or CMP (pollishing).

Another problem in the using of Al is its thermal expansion, ten times greater than that of silicon. Heating places Al under compressive stress causing hillocks. Cooling back places Al under tension causing voids. Adding copper can stabilize grain boundaries and minimize hillocks. High current density can cause electromigration, movement of metal atoms in the direction of current. This can cause hillocks and voids leading to shorts or opens of the circuit. Adding Cu in few percentage can inhibit electromigration. Multilayer stacks like Ti+TiN+Al(Cu)+TiN is used. Ti helps controlling Al grains, reducing electromigration. TiN blocks the diffusion of WF_6. Ti+TiN can shunt metal lines in the case of voids and hillocks. All this structure can contains stresses.
Dielectric: between different metal lines dielectrics are needed to avoid shorts. We need good electrical isolation, low dielectric constant and high breakdown electric field.

8 NOR Flash Memory process flow

The most important structure in Flash memories are the floating gate transistors. They are composed as usual by two active areas (Drain and Source) and two gate, the usual control gate and the floating gate. The following image schematizes a flash cell.

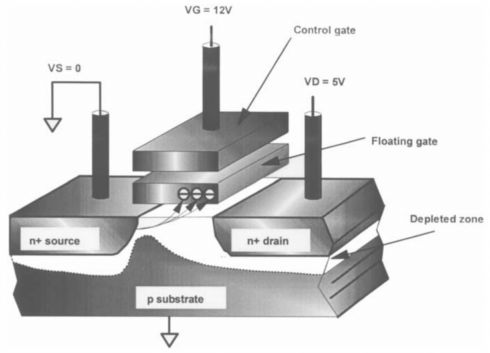

Figure 23: Flash cell cross section in writing mode.

During the write operation of the memory, the gate and drain are kept at high voltage. With a given oxide thickness of 20nm between the floating and control gates and 12nm between floating gate and the channel, there are a strong electric field which let the electrons pass from the channel to the floating gate. This is known as Hot Electrons Injection. To erase the NOR flash memory, we can apply an high voltage between gate and source. This mechanism is called Fowler-Nordheim tunneling.

8.1 Process Flow

As usual, we start with a silicon wafer with a pad oxide (7-20nm) and a LPCVD-SiN which protects the wafer from sub sequential oxidation and acts like stopping layer for Chemical Mechanical Polishing. We then pattern the memory array with lithography. This step is very crucial since define the node technology of the system. Therefore, we use double patterning to archive half pitch resolution. After Dry etching, we have formed the shallow trench isolation. The same is done for LV-HV CMOS in the periphery.
Dry etching used for STI formation damage the silicon. Trench sidewalls are oxidized to repair etch

damages and corners are rounded. The trenches are now ready for trench filling. In order to control the thin film oxidation, advanced technique are used such as ISSG (In Situ Steam Generation). Trenches are filled with silicon dioxide, using High Density Plasma, SA-CVD or other techniques. CMP is then used to planarize the system using silicon nitride as stopper. Now the field oxide is lowered using HF wet etching and nitride layer is removed using H_3PO_4.

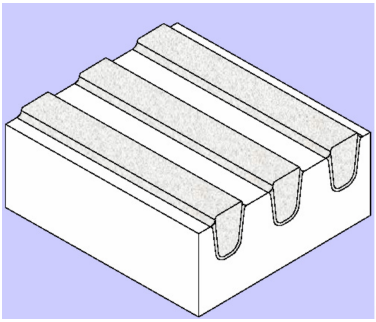

Figure 24: Open field region after field oxide lowering and nitride removals.

Residual pad oxide is used for N-buried (P-buried) implant (used for channel erasing) and it is used for P-Well (N-Well) formation. Then the oxide pad is removed using HF wet etching and tunnel oxide is grown using advance techniques. Note that this is a critical step in the process flow since the quality of the tunnel oxide is essential for the memory performance and reliability. For this reason, nitridation is used to avoid trapping charge effects.

After the formation of tunnel oxide, a P-doped polysilicon layer is deposited (LPCVD).
Now we can pattern our floating gate made of poly1. We use the same mask used for STI formation. Etch must guarantees effective removal of poly1, high anisotropy and selectivity. From this step can arise overlay control issues. To avoid these issues self-alignment floating gate techniques are used. The first one is poly-CMP. Implantations are performed before STI formation. After STI definition, trenches are fulfilled in silicon dioxide and polished using CMP. Then nitride layer is removed with wet etching and tunnel oxide is grown. The next step is the deposition of poly1 followed by CMP.

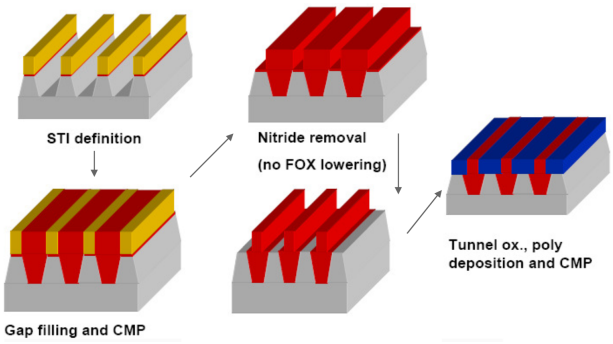

Figure 25: poly-CMP.

The great advantage of poly-CMP is that we have ridden of one exposure step (lowering the costs and no overlay prolems). During nitride wet etching we can control the lateral dimension of floating gate (increase the capacitance of the floating gate).

The second technique is called Advanced Self Alignment STI. The Well and threshold implant is done before STI formation. We grow the tunnel oxide and deposit poly1 and nitride before STI formation. The, using lithography we create STIs. Why don't we use this technique before? The problem is strictly linked to the stability of tunnel oxide during dry and wet etching followed by oxidation of trench sidewalls.
We know that, after trenches formation the sidewalls are plenty of by-products. To remove this products we have to use wet etching, that is usually selective to SiO_2. The subsequent oxidation consumes poly1 enlarges the tunnel oxide corners.

We now can define the control gate. To separate the floating from the control gate, we lay down an interpoly dielectric. The interpoly layer must ensure control/floating gate coupling (note that control gate "controls" the channel through capacitive coupling with the floating gate) but must

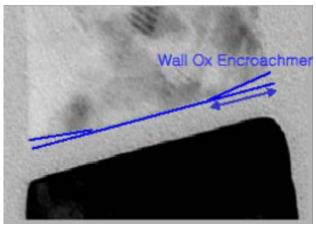

Figure 26: ASASTI oxide enlargement issue.

avoid charge transfer from floating gate to control gate. This must be an high quality dielectric (high T oxide). The best choice is a three layers structure of nitride-oxide-nitride.

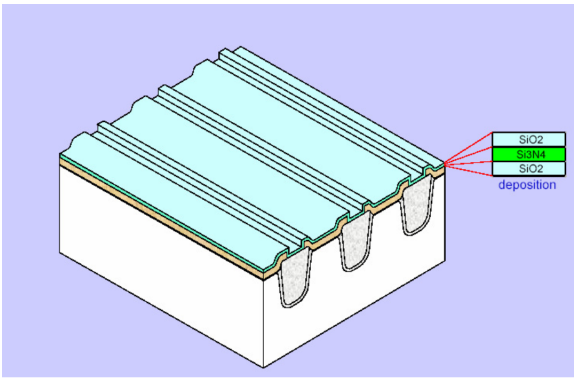

Figure 27: ONO conformal deposition.

We can now focus our attention in the peripheral. Let's develop the Low and High Voltage N-P-MOS. Until now, we have the following structure:

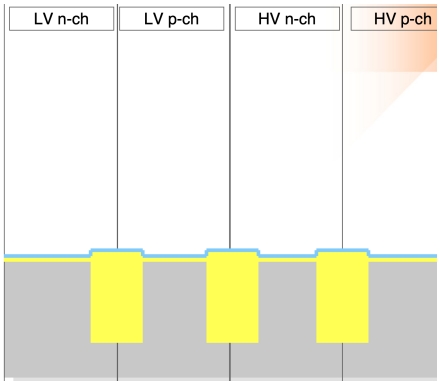

Figure 28: HV and LV.

To form HV Well, we first remove the ONO structure covering the memory array with a mask. Tunnel oxide becomes as sacrificial oxide. Using different masks HV p-well and n-well are formed. Tunnel oxide is then removed using a Buffered Oxide Etch (avoiding to peel photoresist form memory array). HV oxide tunnel is then grown. The same approach is used for LV wells formation. In this case the LV oxide must be very thin (2-3nm). After the wells formation an undoped polysilicon is deposited. This will constitutes the control gate material for the transistors. (HV and LV) and memory cells.

8.1.1 STI corners

STI corners and moat shape of polysilicon can lead to oxide thinning and anomalous conduction. This corner changes the I-V characteristic of the MOSFET in what is called Hump effect:

Different etch methods have been developed in order to improve leakage, rounding the corners. We know that trench morphology can changes using dry or wet etching. After STI formation in the cleaning process under cuts are formed, exposing trench corners.

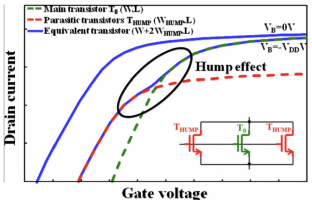

Figure 29: Hump effect.

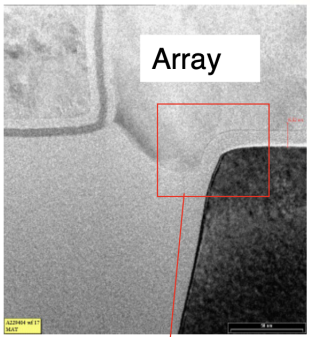

Figure 30: An example of STI corner and moat shape in memory array. The moat is formed in poly1 while the corner is in the active region.

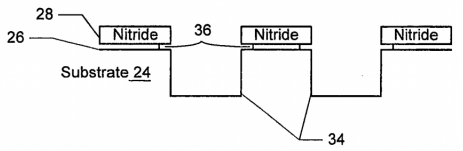

Figure 31: Formation of undercut due HF selectivity to silicon dioxide.

After multiple sidewalls oxidation, these corners are rounded. Note that stresses can affect oxidation kinetics lead to current leakages and mobility variation

References

[1] J.D. Plummer. *Silicon VLSI Technology: Fundamentals, Practice and Modeling.* Dorling Kindersley, 2009.