Abstract

This document is part of the specification of the free silicon manufacturing standard for manufacturing the LibreSilicon standard logic cells[1] and related free technology nodes from the LibreSilicon project.

For this initial revision 0.1 a gate-first approach has been chosen which led to the choice of polysilicon as the gate electrode material because of the simplicity of the gate alignment. For better isolation properties of the transistors and gates in overall a box-isolation approach has been chosen. All of these choices have been made with the future scale down from the recent $1\mu m$ to smaller structure sizes. **This process is for manufacturing $1\mu m$ only!** But further releases which will have been tested with smaller structure sizes can be expected.

[1] `https://github.com/chipforge/StdCellLib`

Contents

Libre Silicon process design

David Lanzendörfer

June 3, 2018

We need to optimize our process to be TTL compatible (5V logic levels) and at the same time being as fast and power efficient as possible. In order to have a good propagation delay with a technology node of around $1\mu m$ we will have to have gates with up to four stacked MOS transistors.
Acceptable input signal voltages range from 0 volts to 0.8 volts for a low logic state, and 2 volts to 5 volts for a high logic state. Acceptable output signal voltages shall range from 0 volts to 0.5 volts for a low logic state, and 2.7 volts to 5 volts for a high logic state[1]

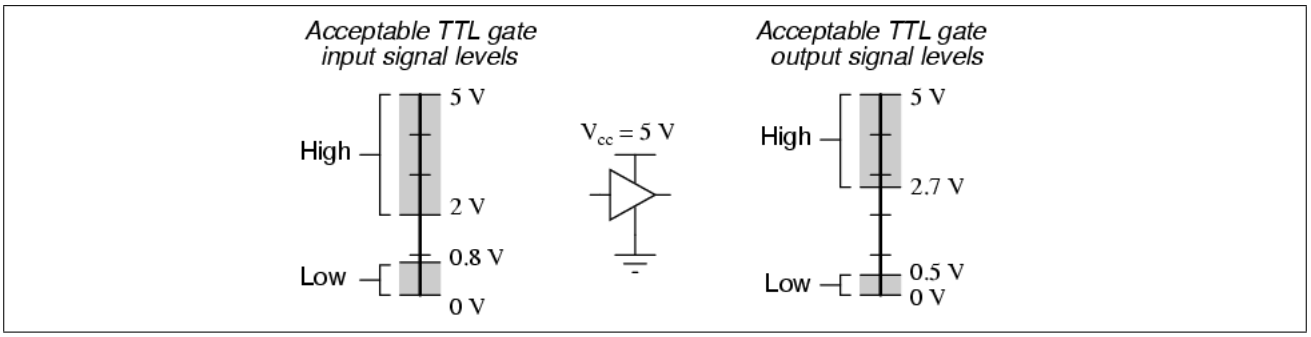

Figure 1: TTL logic levels

As shown in Figure 1 we have some margin to make our PMOS and NMOS transistors work with each other in order to form a CMOS circuit which is actually working without getting warm.
Or more clearly defined

$$V_{off} \leq 0.8V \tag{1}$$

and

$$V_{on} \geq 2V \tag{2}$$

which are limits, elementary to our design.

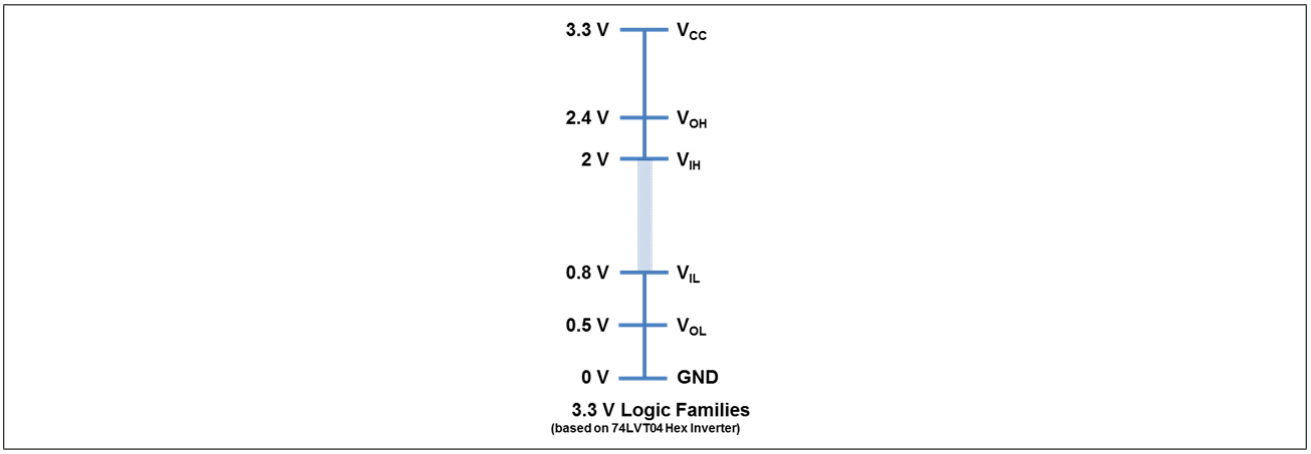

Figure 2: CMOS 3.3V logic levels

This means that we also will be compatible to CMOS logic level output pins since their ON/OFF levels are within our tolerance range[2] as it is shown in Figure 2.
We target threshold voltages of $V_{Tn} \approx 0.8V$ and $V_{Tp} \approx -0.8V$ which should be enough. We can internally always shift the voltage supply levels to compensate for threshold variations.[3]

[1] `https://www.allaboutcircuits.com/textbook/digital/chpt-3/logic-signal-voltage-levels`
[2] `https://learn.sparkfun.com/tutorials/logic-levels/33-v-cmos-logic-levels`
[3] Hagen! Please explain this part here

1 Substrate

The Hong University of science and technology (short HKUST) provides us with two types of wafers.

- Prime Grade Silicon Wafer, [100] N-type
  - Front-side polished, backside etched
  - Dopant: Phosphorus
  - Thickness: $525\mu m \pm 25\mu m$
  - Resistivity:4 to 7ohm-cm
  - Growth Method: CZ
  - Diameter: 100mm +/- 0.5 mm
  - Primary & secondary flat locations: (In compliance with the SEMI)
    - Carbon concentration<2.5 x e16 atm/cc
    - Oxygen concentration $< 9.0 \cdot 10^{17} \frac{atm}{cc}$
    - $TTV < 10\mu m$
    - $TIR < 6\mu m$
    - $Bow/Warp < 40\mu m$
- Prime Grade Silicon Wafer, [100] P-type
  - Front-side polished, backside etched
  - Dopant: Boron
  - Thickness: $525\mu m \pm 25\mu m$
  - Resistivity:15 to 25 ohm-cm
  - Growth Method: CZ
  - Diameter: 100mm +/- 0.5 mm
  - Primary & secondary flat locations: (In compliance with the SEMI)
    - Carbon concentration<2.5 x e16 atm/cc
    - Oxygen concentration $< 9.0 \cdot 10^{17} \frac{atm}{cc}$
    - $TTV < 10\mu m$
    - $TIR < 6\mu m$
    - $Bow/Warp < 40\mu m$

For this process the p-doped mono crystalline silicon substrate is being used, but forks and modifications will be very well possible based on a Graphene substrate or alike, still under the LSPL. The starting material is a p-doped <100> oriented mono crystalline silicon wafer

Reasons for using p-doped substrate:

- We can't use two different substrates for our design because in the design both PMOS and NMOS is present. We have to choose which is more beneficial from fabrication point of view. In general or say it's true that NMOS devices are always more in the Semiconductor Industry in comparison to PMOS devices. For your reference-SRAM requires 6 transistors (4 NMOS, 2 PMOS).
- Another reason for more number of NMOS is because of difference of mobility of electron and holes. Electron mobility is almost twice of holes mobility and because of this ON-RESISTANCE of n-channel device is half of p-channel device with the same geometry and under the same operating conditions. That means to achieve same impedance size of n-channel transistors is almost half of p-channel devices. Same thing I can say in the different way that for same size of wafer, we can have more number of NMOS (means can perform more logical operation) in comparison to PMOS.
- Since we only have the choice between P and N doped substrate, we use P doped substrate, because of the carrier mobility

Using the web app[4] we get a doping concentration between $8.76 \cdot 10^{14} \frac{1}{cm^3}$ and $5.23 \cdot 10^{14} \frac{1}{cm^3}$. The average of this range is $N_B = \frac{8.76+5.23}{2} \cdot 10^{14} \frac{1}{cm^3} \approx 7 \cdot 10^{14} \frac{1}{cm^3}$

[4] `http://www.solecon.com/sra/rho2ccal.htm`

2 Isolation

For the isolation in this design the STI-LOCOS hybrid approach is being chosen. Shallow trench isolation (STI), also known as box isolation technique, is an integrated circuit feature which prevents electric current leakage between adjacent semiconductor device components.[5] STI is generally used on CMOS process technology nodes of 250 nanometers and smaller.

Reasons for using box isolation:

- We want to be forward compatible to future LibreSilicon nodes with a size of 100nm or smaller
- It simplifies the construction of the gate and allows us to use Aluminum instead of Polysilicon for the gate contact

Issues we have to keep in mind is that the depth is not uniform and can variate strongly within a $2\mu m$ range! This means we have to make the well at least "deep enough" at the shallowest place, so that it provides adequate isolation between the transistors everywhere on the die.
One way to reduce the variation in depth is to have a uniform width of the isolation.
Also the non-uniform thickness of the oxide is a problem.

CMP (Chemical Mechanical Planarization) for evening the oxide out is being chosen, because the hard mask can be removed in the very same process step.

3 Interconnect

The interconnects and the gate electrode are being made using Aluminum which is a very commonly used material to do interconnects in low-frequency and low-resolution applications

Reasons for using Aluminum:

- It's a well explored material for interconnect with a lot of literature on how to process it
- Aluminum is easy to etch compared to copper
- Aluminum isn't contaminating everything like copper does and doesn't require special separated setup for handling
- There are many technology nodes even down to 180nm which still are using Aluminum and it's not having a big impact on the clock frequencies, so it should be good enough for us as well.

As soon as we've got CMOS all figured out, we will tackle copper interconnect in release 2.0

[5] `https://www.google.com/patents/US7985656`

4 MOS gate material

We decided to use the gate-first approach because realizing the gate self-alignment is much easier this way. We further on decided to use the best-practice material polysilicon for the gate electrode, because it is easy to deposit and etch and virtually every manufacturer out there has at least one machine in their lab to deposit it. Because of its high resistance however, we had to throw in another layer of silicide in order to reduce the gate resistivity from these $100\Omega - m$ or so to a a few Ohms per meter. A nice side effect is the better etch-stop properties of this low-resistance film adding to the gate, source and drain contacts.
A down side is that we will have to get "a hang on" the reaction times of silicides because a lot of the details of the reactivity between silicon and titanium to form titanium-silicide seem to be under NDA and secrets of the diverse factories running their own CMOS processes.

5 MOS gate thickness

As the continuous down-scaling of the device size has lead to very thin gate oxides, the leakage current that can flow from the channel to the gate comes into the order of the subthreshold leakage current and the gate cannot be considered as an ideally insulated electrode anymore. This affects the circuit functionality and increases the standby power consumption due to the static gate current. For dynamic logic concepts the gate leakage drastically reduces the maximum clock cycle time[6]. Two tunneling mechanisms are responsible for the gate leakage, Fowler-Nordheim tunneling and direct tunneling[7]. The gate leakage increases exponentially as the oxide thickness is reduced. This limits the down-scaling of the oxide thickness to about 1.5-2 nm when looking at the total standby power consumption of a chip[8]. To further decrease the effective oxide thickness alternative high dielectric constant materials can be used[9]. On the other hand, a thin gate oxide reduces the short-channel effect and improves the driving capabilities of a MOS transistor. However, a tradeoff between this benefit and the gate leakage is necessary.

With $1\mu m$ we don't have to worry about this leakage yet because our gate oxide thickness is too high for these effects to actually become a problem, but we want to do our home work already in preparation of scale-down and also for curiosity.
We for now just use 40 nm. That's still doable with a precision high enough when using dry oxidation and a temperature of 1000°Celsius.

5.1 Subthreshold leakage

The sub-threshold leakage current can be calculated with[10]

$$I_{sub} = I_0 \cdot \left(1 - \exp\left(-\frac{V_{ds}}{V_{th}}\right)\right) \cdot \exp\left(\frac{V_{gs} - V_T}{n \cdot V_{th}}\right) \tag{3}$$

where

$$I_0 = \frac{W}{L}\mu_0 V_{th}^2 \sqrt{\frac{N_A \cdot q \cdot \epsilon_{Si}}{2 \cdot \phi_{sub}}} \tag{4}$$

$V_{th} = 26mV$ is the thermal voltage, V_T is the threshold voltage, V_{ds} and V_{gs} are the drain-to-source and gate-to-source voltages respectively. W and L are the effective transistor width and length, respectively. C_{ox} is the gate oxide capacitance, μ_0 is the carrier mobility and $n = 1 + \frac{C_{dep}}{C_{ox}}$ is the subthreshold swing coefficient.
First of all, lets say $W = L$ which leads to a square:

$$I_0 = \mu_0 V_{th}^2 \sqrt{\frac{N_A \cdot q \cdot \epsilon_{Si}}{2 \cdot \phi_{sub}}} \tag{5}$$

With

- $\epsilon_0 = 8.85 \cdot 10^{-14} \frac{F}{cm}$. is the electric permittivity in vacuum
- $\epsilon_{ox} = 3.9 \cdot \epsilon_0$ is the relative permittivity of silicon dioxide

[6] N. Wang, Digital MOS Integrated Circuits, Prentice-Hall, Englewood Cliffs, NJ, 1989
[7] A. Schenk and G. Heiser, "Modeling and Simulation of Tunneling through Ultra-Thin Gate Dielectrics"J.Appl.Phys., vol. 81, no. 12, pp. 7900, 1997
[8] Y. Taur, "The Incredible Shrinking Transistor," IEEE Spectrum, pp. 25-29, July 1999.
[9] S. Thompson, P. Packan, and M. Bohr, "MOS Scaling: Transistor Challenges for the 21st Century," Intel Technology Journal, vol. Q3, 1998
[10] `http://ecee.colorado.edu/~bart/book/book/chapter3/ch3_4.htm#3_4_2`

- $\epsilon_{Si} = 11.68 \cdot \epsilon_0$ is the relative permittivity of silicon

The carrier mobility μ_0 can be calculated with[11]

$$\mu(N) = \mu_{min} + \frac{\mu_{max} - \mu_{min}}{1 + \left(\frac{N}{N_r}\right)^{\alpha}} \tag{6}$$

using the fitting parameters from Table 1

	Arsenic	**Phosphorus**	**Boron**
$\mu_{min}[\frac{cm^2}{Vs}]$	52.2	68.5	44.9
$\mu_{max}[\frac{cm^2}{Vs}]$	1417	1414	470.5
$N_r[\frac{1}{cm^3}]$	$9.68 \cdot 10^{16}$	$9.20 \cdot 10^{16}$	$2.23 \cdot 10^{17}$
α	0.68	0.711	0.719

Table 1: Parameters for calculation of the mobility as a function of the doping density

5.2 Gate tunneling current

The tunneling of electrons (or holes) from the bulk and source/drain overlap region through the gate oxide potential barrier into the gate (or vice-versa) is referred as gate oxide tunneling current. This phenomenon is related with the MOS capacitance concept. There are three major gate leakage mechanisms in a MOS structure. The first one is the electron conduction-band tunneling (ECB), where electrons tunneling from conduction band of the substrate to the conduction band of the gate (or vice versa). The second one is the electron valence-band tunneling (EVB). In this case, electrons tunneling from the valence band of the substrate to the conduct band of the gate. The last one is known as hole valence-band (HVB) tunneling, where holes tunneling from the valence band of the substrate to the valence band of the gate (or vice- versa)
Each mechanism is dominant or important in different regions of operation for NMOS and PMOS transistors. For each mechanism, gate leakage current can be modeled by

$$I = W \cdot L \cdot A \cdot \left(\frac{V_{ox}}{t_{ox}}\right)^2 \exp\left(\frac{-B\left(1 - \left(1 - \frac{V_{ox}}{\phi_{ox}}\right)^{\frac{3}{2}}\right)}{\frac{T_{ox}}{t_{ox}}}\right) \tag{7}$$

[11] https://ecee.colorado.edu/~bart/book/book/chapter2/ch2_7.htm#2_7_2

6 Thresholds

6.1 NMOS threshold

First we take a look at the pull down network, which is being formed by the NMOS transistors. We have to make sure that the pull down network will effectively tie our output signal to ground, latest when reaching the voltage defined as high signal (inverter!)
As shown in Figure 1 our acceptable voltages for our CMOS "ON" state range from 0V to 2V

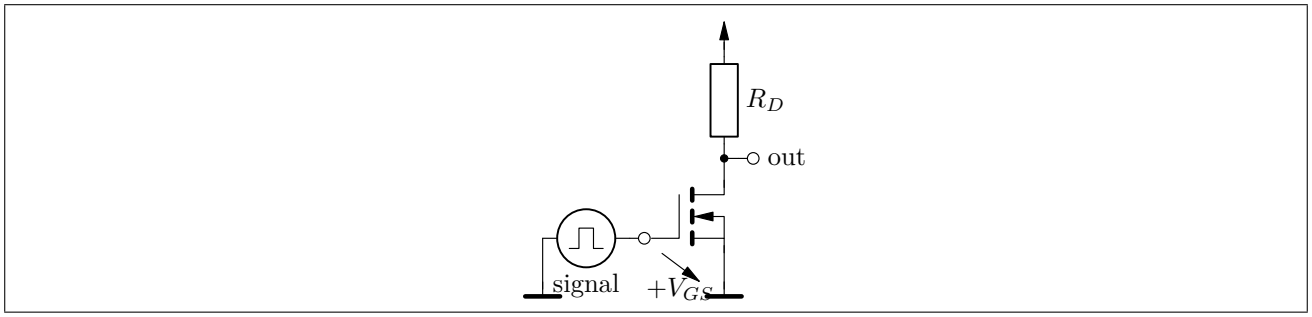

Figure 3: enhancement-mode NMOS transistor use-case

$$V_{off} \leq 0.8V \tag{8}$$

With the values derived from section 5 which gives us the thickness of the gate ($\approx 40nm$) we target a threshold voltage of $0.8V$, so that we switch the pull down circuit as soon as the low-signal is on.
We target a concentration of $N_p = 10^{16}\frac{1}{cm^3} = 10^{22}\frac{1}{m^3}$.
The depletion zone thickness at its peak will be $W_{dmax} \approx 2.73 \cdot 10^{-7}m = 273nm$
With an implantation (or constant source diffusion step), we can now set a range/energy and dosage in order to cover the depletion zone area.
For getting the energy and dose we look at **??** or use the web tool linked in the implant chapter.
The depth of the p-well $\approx 2\mu m$ comes mainly from the need to fulfill the condition from **??**

$$x_e = 2 \cdot \sqrt{D_e \cdot t_e} \gg 2 \cdot \sqrt{D_v \cdot t_v} = x_v \tag{9}$$

We already got the background ($N_B \approx 7 \cdot 10^{14} \frac{1}{cm^3} = 7 \cdot 10^{20} \frac{1}{m^3}$) concentration from the specs of the basis substrate.

$$N_p - N_B = 10^{22} \frac{1}{m^3} - 7 \cdot 10^{20} \frac{1}{m^3} = 9.3 \cdot 10^{21} \frac{1}{m^3} \tag{10}$$

We use a drive in temperature of $1150°C$ which is $T = 1423.15°K$ in Kelvin which gives us the diffusion coefficient $D = 9.1 \cdot 10^{-17} \frac{m^2}{s}$
Now using

$$N(x,t) = \frac{Q}{\sqrt{\pi \cdot D \cdot t}} \cdot \exp\left(\frac{-x^2}{4 \cdot D \cdot t}\right) \tag{11}$$

We set the conditions and get the required diffusion time as well as the initial dosage in one shot:

$$N(0,t) = \frac{Q}{\sqrt{\pi \cdot D \cdot t}} = N_p - N_B = 7 \cdot 10^{20} \frac{1}{m^3} \tag{12}$$

$$x = 2 \cdot \sqrt{D \cdot t \cdot \ln\left(\frac{N_T}{N_B}\right)} = 2\mu m = 2 \cdot 10^{-6} m \tag{13}$$

$$\Rightarrow t \approx 4259s \approx \underline{70min} \tag{14}$$

$$\Rightarrow Q = 7 \cdot 10^{20} \frac{1}{m^3} \cdot \sqrt{\pi \cdot D \cdot t} \approx \underline{1.02 \cdot 10^{16} \frac{1}{m^2}} \tag{15}$$

If we plot the functions from our calculation we can yield the below graphics[12]

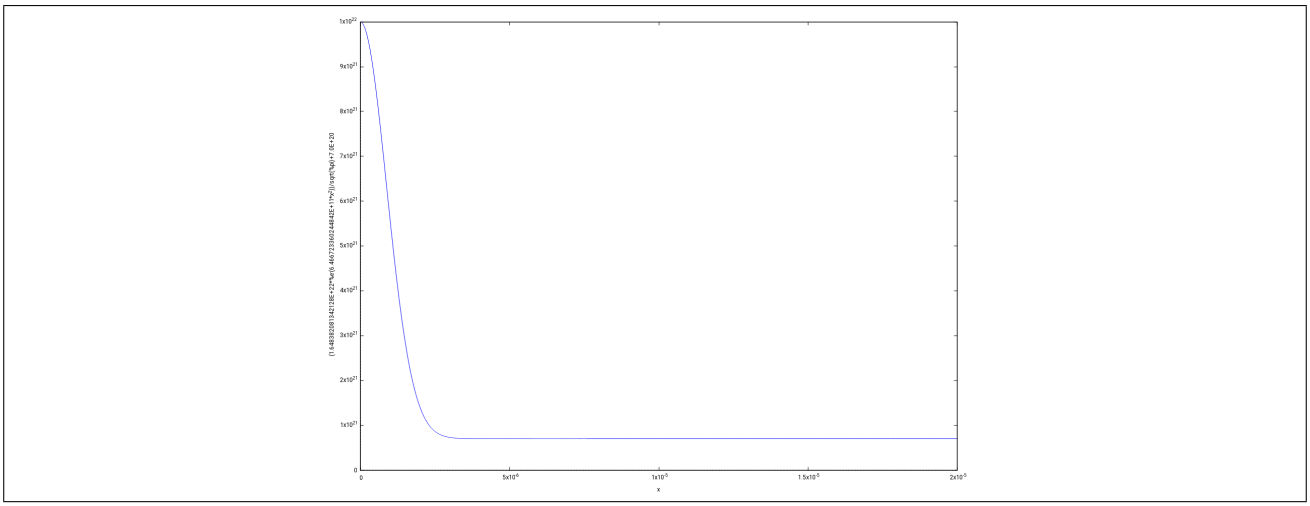

Figure 4: Dopant concentration after around 70 minutes

In Figure 4 we can see that after roughly an hour we already have the desired even gradient and deep penetration of dopants, which will give us a low R_{DS}.

[12] see simulation/diffusion_pwell.wxmx

6.2 PMOS threshold

Now we take a look at the worst case of 4 stacked PMOS transistors, which is the highest stacking amount which will be possible in technologies relying on this process.

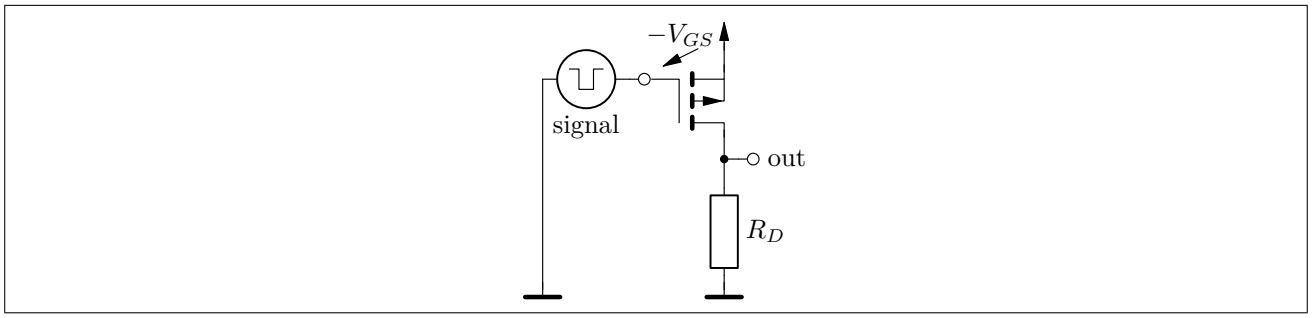

Figure 5: enhancement-mode PMOS transistor use-case

$\approx 4\mu m$ come mainly from the need to fulfill the condition from **??**

$$x_e = 2 \cdot \sqrt{D_e \cdot t_e} \gg 2 \cdot \sqrt{D_v \cdot t_v} = x_v \tag{16}$$

We already got the background ($N_B \approx 7 \cdot 10^{14} \frac{1}{cm^3} = 7 \cdot 10^{20} \frac{1}{m^3}$) concentration from the specs of the basis substrate.
We use a drive in temperature of $1150°C$ which is $T = 1423.15°K$ in Kelvin which gives us the diffusion coefficient $D = 9.1 \cdot 10^{-17} \frac{m^2}{s}$
Now using

$$N(x,t) = \frac{Q}{\sqrt{\pi \cdot D \cdot t}} \cdot \exp\left(\frac{-x^2}{4 \cdot D \cdot t}\right) \tag{17}$$

We set the conditions and get the required diffusion time as well as the initial dosage in one shot:

$$N(0,t) = \frac{Q}{\sqrt{\pi \cdot D \cdot t}} = N_p - N_B = 7 \cdot 10^{20} \frac{1}{m^3} \tag{18}$$

$$x = 2 \cdot \sqrt{D \cdot t \cdot \ln\left(\frac{N_T}{N_B}\right)} = 4\mu m = 4 \cdot 10^{-6} m \tag{19}$$

$$\Rightarrow t \approx 16162s \approx 269min \approx \underline{4h30min} \tag{20}$$

$$\Rightarrow Q = 7 \cdot 10^{20} \frac{1}{m^3} \cdot \sqrt{\pi \cdot D \cdot t} = 7 \cdot 10^{20} \frac{1}{m^3} \cdot \sqrt{\pi} \cdot 2 \cdot 10^{-6} m \approx \underline{2.48 \cdot 10^{15} \frac{1}{m^2}} \tag{21}$$

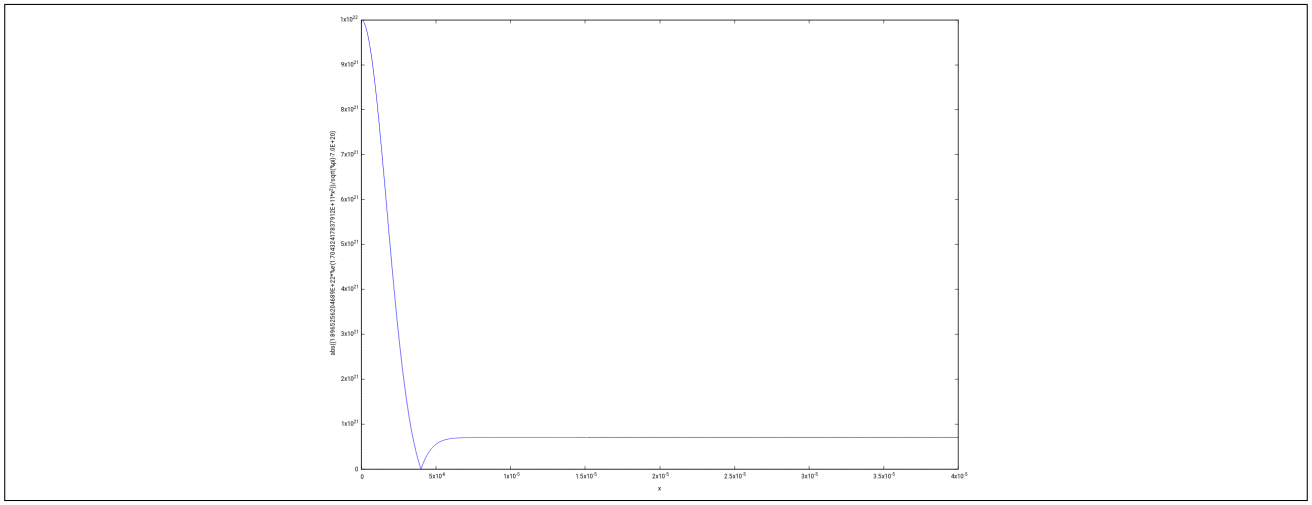

Figure 6: Dopant concentration after 4 hours 30 minutes

7 MOS gate verification

Now we have to verify that we don have relevant leakage on the gate with the thresholds and doping values we've set.
For that we use the formulas from subsection 5.1
We can now plot multiple leakages for N- and P-channel transistors with a gate oxide thickness[13] and with a surface concentration of $1e16\frac{1}{cm^3} = 1e22\frac{1}{m^3}$ each

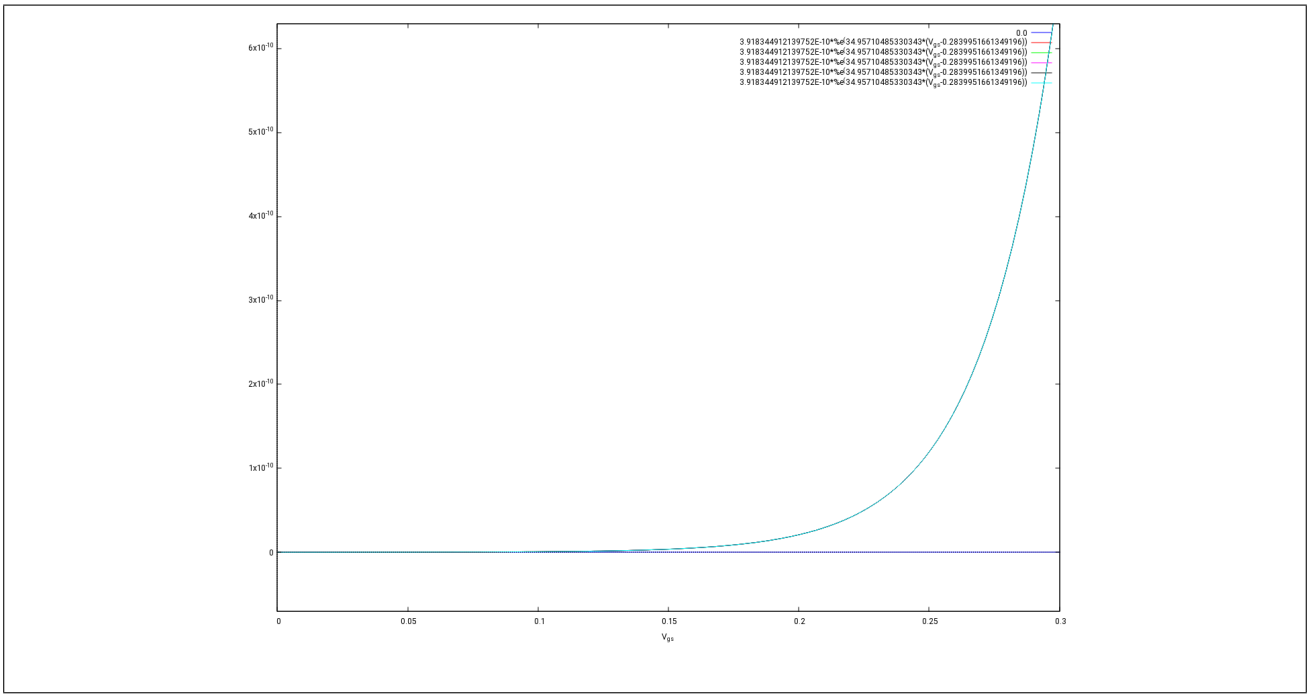

Figure 7: Subthreshold leakage plot(in Ampere)

In Figure 7 we see that with our gate oxide thickness this is really no problem, as we had expected. From 0V up to 5V and further there is basically no leakage on the gate from the sub threshold current with $V_{Tn} \approx 0.39V$ and $V_{Tp} \approx -0.30V$. That's good enough, as we will see in subsection 6.1 and subsection 6.2. There is actually a reduction of current when reaching the threshold because of the inversion of the capacity in the depletion zone[14], but I didn't include this into the calculation, because "TL;DR". It's a TODO for release 2.1 of this process which will go sub $1\mu m$

[13] See simulation/gate.wxmx
[14] `https://people.eecs.berkeley.edu/~hu/Chenming-Hu_ch5.pdf`

8 Source/Drain junctions

As junction thickness decreases, the series resistance of the junction increases. This cannot be neglected for conventional shallow junction technologies.

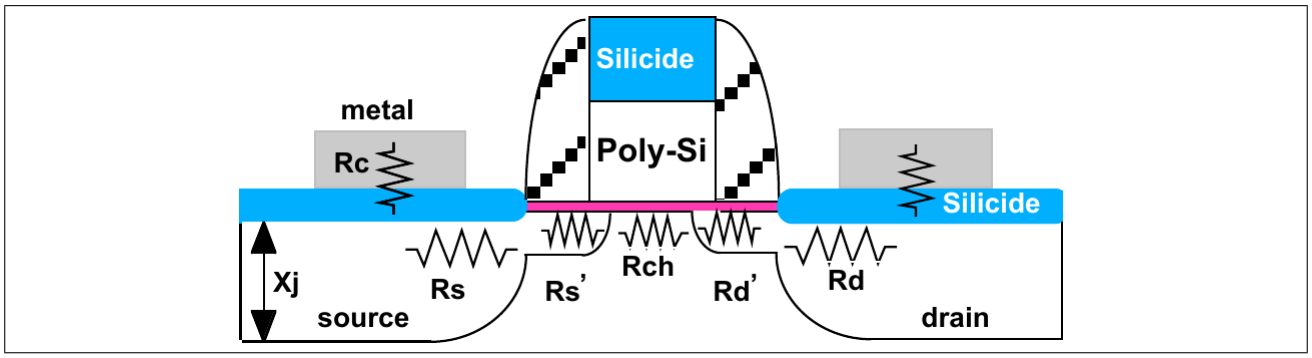

Figure 8: Cross section with resistivities

Sheet resistance is given by

$$R_{sh} = \frac{\rho_s S}{W} \tag{22}$$

Where the sheet resistivity is

$$\rho_s = \frac{\rho}{x_j} \propto \frac{1}{N_{sd} x_j} \tag{23}$$

The channel resistance can be approximated by

$$R_{ch} \propto \frac{L_{ch} t_{ox}}{V_{gs} - V_{th}} \tag{24}$$

As L_g scales down

- R_{ch} scales down
- R_{sd} does not scales as maximum doping is limited by solid solubility
- R_{sd} becomes comparable to R_{ch}
- An increase in R_{sd} becomes an important factor for device current
- Parasitic portion of the device is now playing important role in device performance and CMOS scaling

9 Alignment strategy

When having multiple layers exposed after each other, there is the problem on how to make sure that for instance vias are actually making contact with the below wire and the wire above. For this purpose we have to align the masks, in order to avoid issues like shown in Figure 9

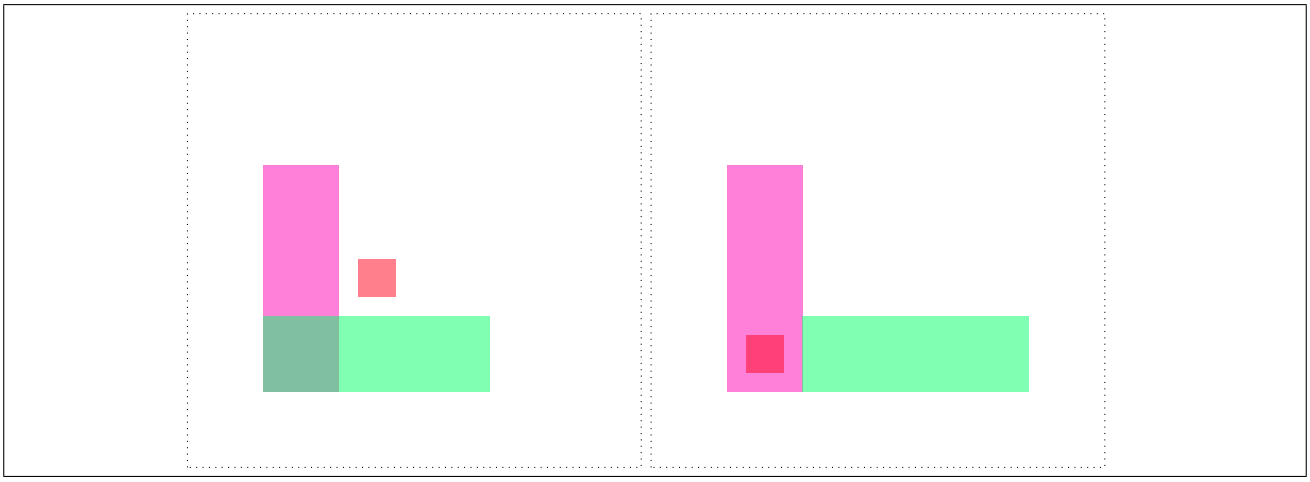

Figure 9: What can go wrong with alignment

In Figure 9 we can see how the wires and vias are missing each other because of an exposure offset and the via is going nowhere in the best case and creates an (of course) undesired short circuit in the worst case. This has to be avoided by using alignment.
We have decided to use backside alignment because shining through the wafer from behind with infrared and finding an orientation marker isn't a problem with our simple CMOS process.[15]
The stepper machine at HKUST[16] does that for us.
We need to add orientation cross hairs onto the layout edge in order to identify them during alignment

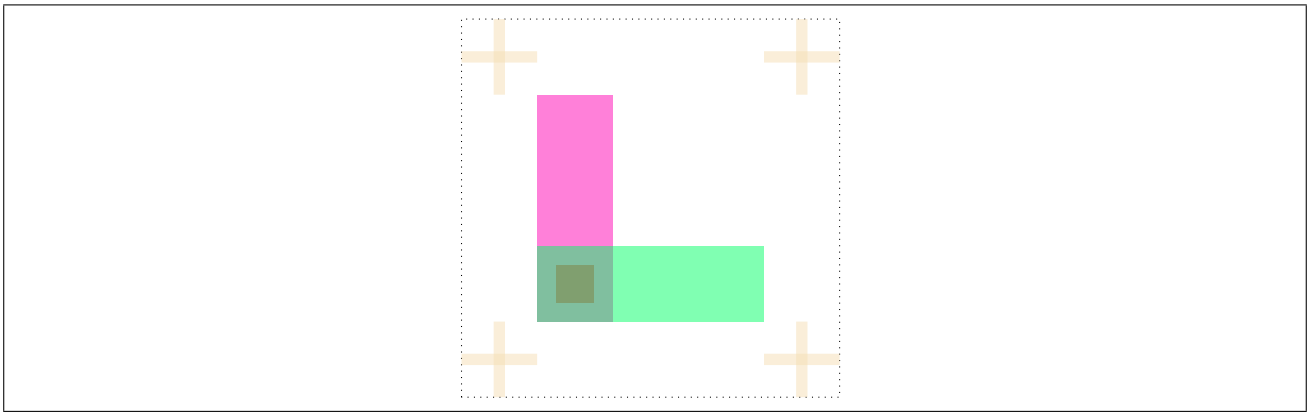

Figure 10: Mask alignment with markers

Using the STI (active) mask for the cross hair structure is a good choice, because it's the lowest layer, will never be covered by any additional material and gives us a good contrast because it's silicon next to silicon dioxide.

[15] https://patents.google.com/patent/US6376329
[16] http://www.nff.ust.hk/en/equipment-and-process/equipment-list/photolithography-module.html

The "DRIE Etcher" can't etch more precise than $0.5\mu m$ (Minimum Line/Space: $0.5\mu m$) which means we will have to give the cross hair structure with the dimensions as shown in Figure 11

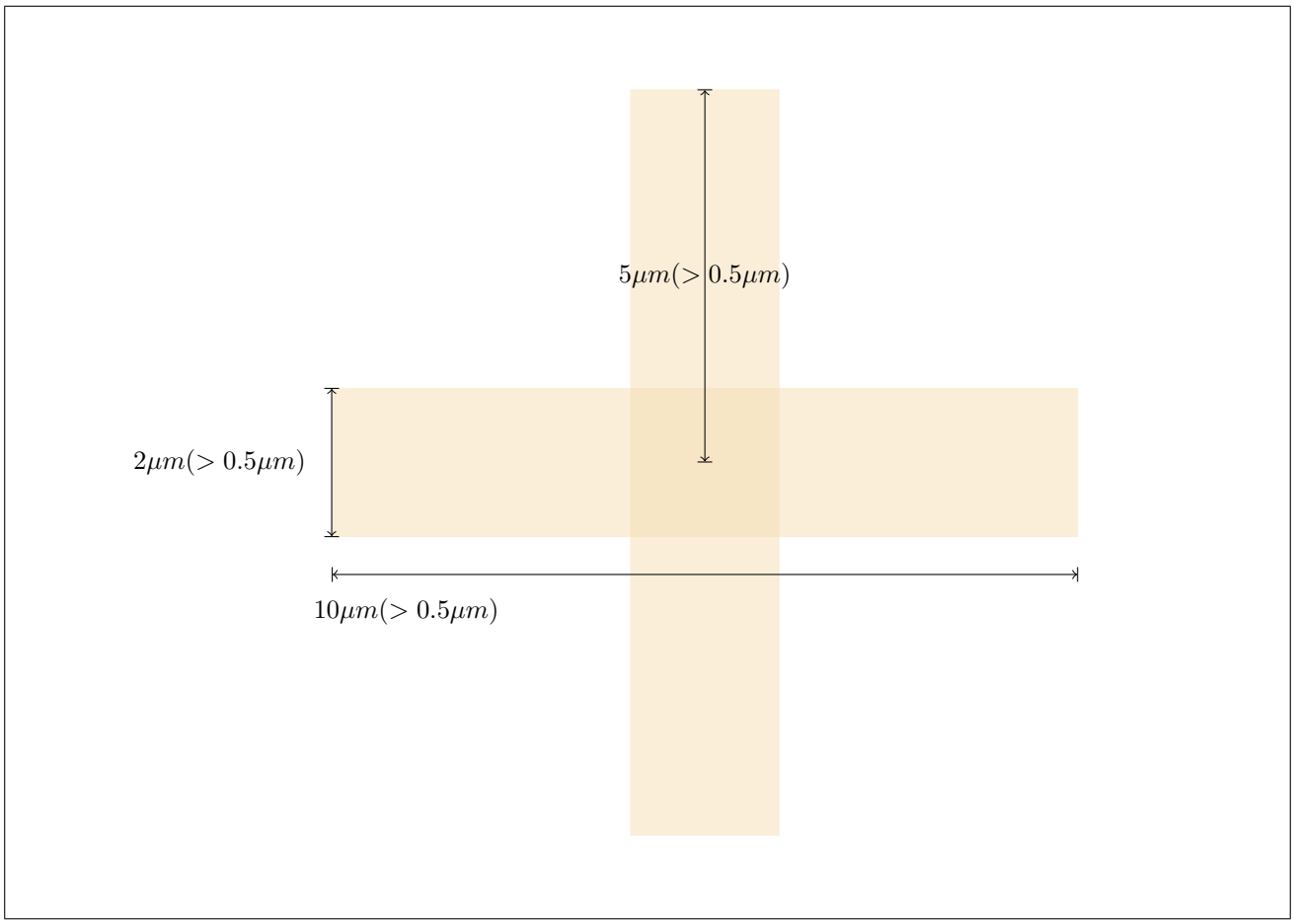

Figure 11: Cross hair dimension

We can now manufacture masks as shown in Figure 12 and align them using the infrared light from the stepper mask alignment and the provided electronic microscope.

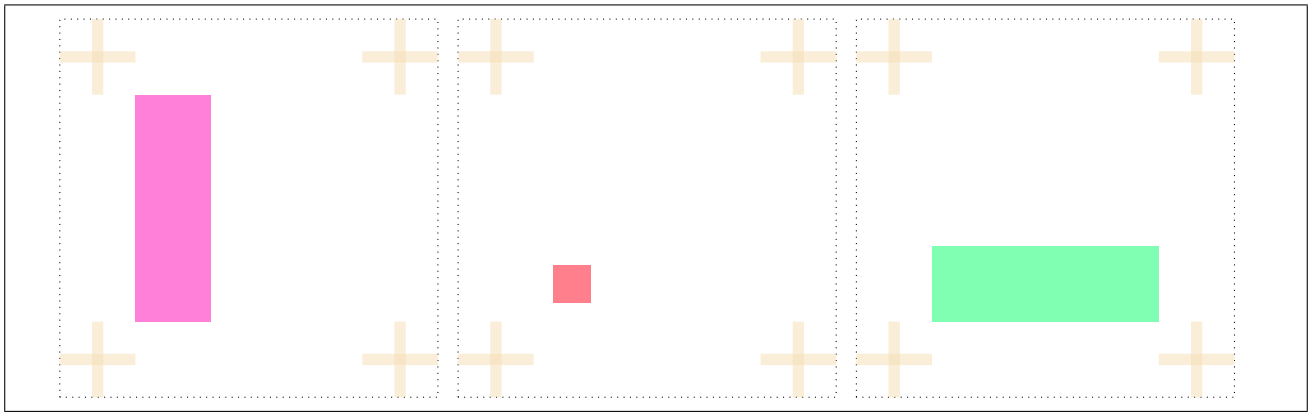

Figure 12: Layout masks with alignment markers

10 Simulation with parameters