6.012 Final Project

Aaron Wubshet and Eswar Anandapadmanaban

1a. Transistor or Simmons Hall? An Exercise Left to the Reader.

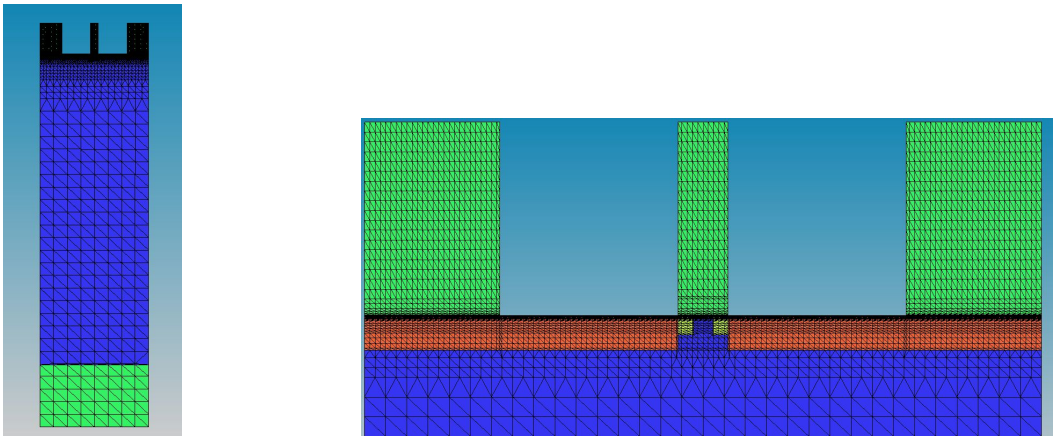
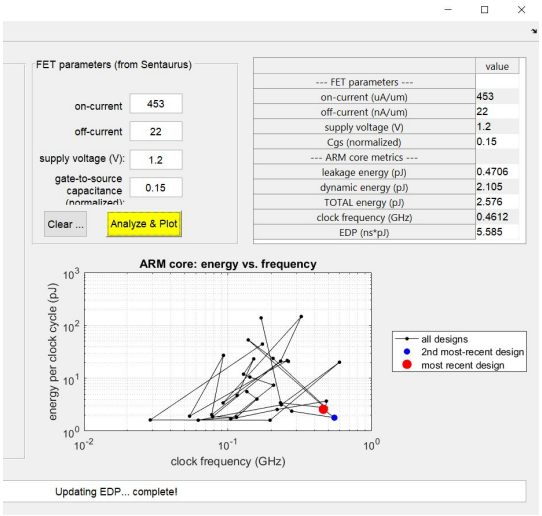


Figure 1: Transistor geometry. Images from Sentaurus

Our transistor seen above features the following improvements over the the initial high power, high performance transistor of the ARM microprocessor:

1. 13nm channel
2. High k-metal gate with hafnium oxide and silver electrodes
3. Vacuum oxide spacer between gate and source as well as gate and drain
4. Reduced operating voltage from 1.8 V to 1.2 V
5. Increased channel doping by a factor of 2.2 from 3e+18 to 6.5e+18
6. Shallow junctions on the source and drain
7. Increased source and drain doping by a factor of 1.5 from 5e+19 to 7.6e+19
8. Lightly doped drain (LDD) pockets added with doping of 3 orders of magnitude smaller than the doping of the source and drain



Results

Our resulting transistor from these changes has an on current of 452 uA/um and off current of 22 nA/um at a supply voltage of 1.2 V. This leads to an EDP of 5.585 ns*pJ and clock frequency of .4612 GHz

compared to the initial transistor with 811.3 ns * pJ and .1706 Ghz.
Our normalized Cgs was 0.15. Our leakage energy and dynamic energy were 0.4706 pJ and 2.104 pJ respectively leading to a total energy of 2.576 pJ.

2. Half Design, Half Divergence: The Mythical Program - Sentaurus

The overarching design philosophy used to arrive at the above transistor actually changed over the course of the design process, from pure Dennard scaling to aggressively small channels and halo doping to fany shallow junction and LDD. Ultimately, it was an approach of balance that won out. The results are summarized below:

2a. ".000000000003nm Node Reached (it's actually a lot Moore)"

Our first approach to designing the transistor was to follow Dennard constant field scaling and reduce the size of our transistor without changing the electric field. Because of the fabrication and design limits preventing us from reducing the oxide thickness or total physical dimensions we chose to reduce only the channel width. To increase the capacitance of gate we changed the material as opposed to reducing the thickness of the oxide since the capacitance is given by $\boldsymbol{\varepsilon}_{ox}/t_{ox}$. We chose hafnium oxide due to its high-k characteristics giving the gate much more control over the channel.

However, the dielectric constant of hafnium oxide is 4-6 times greater than that of silicon oxide. Thus, this led to a capacitance increase of ~5 fold, and we had to account for this change. Looking to follow constant field scaling, we reduced the channel length and decreased the supply voltage proportionally while increasing the doping concentrations. This meant a 7nm channel, 1.5e+19 channel doping, and a supply voltage at the lower bound allowed by the specifications of 400mV. With this setup, the off current was far too high and the on current was far too low. Dennard scaling behaved as expected, in good ways and bad. The plummeting threshold voltage made our off current skyrocket. We tried to back off on the scaling but leave the supply voltage and doping values where they were, but this didn't work either. Half the time, the simulation wouldn't even converge with the unscaled supply voltages and doping concentrations, and the other half of the time the EDP benefits were eliminated.. We were left with a transistor with far too high leakage energy. Reducing the channel any further lead to short channel effects that could not be countered with constant scaling of the electric field.

At this point we tried using shallow junctions which theoretically should have reduced the short channel effects. However, our perception of scaling was way off base. We started with 35nm source and drains and after using the shallow junctions ended up with a 5 nm source and drain heights. This drastic difference seemed to help reduce the off current, but at the same time reduce the on current and we couldn't figure out what was going on, because

we didn't realize the competing side effect of shallow junctions was reducing our current by increasing our contact resistance. We soon abandoned this design methodology and moved on to what we thought was our saving grace.

2b. "My Saving Grace"[1] (Halo)

To address these short channel effects we decided to take aggressive measures and add Lowly Doped Drain (LDD) patches as well as halo patches. After adding these patches, we continued to shrink the channel to about 7 nm. Remembering that fabrication limits prevent us from fabricating anything smaller than 4 nm, we realized that 7nm is far too small to be able to create LDD patches because they would need to be at least 4 nm wide themselves, leading to 8nm of just the LDD patches, making that larger than the channel itself! This put us in a dilemma. To be able to reasonably use LDD and halo we would need to have at least a 20 nm channel. This would be a jump several steps backward leading to a worse off transistor in the end. Realizing this fabrication limit, we scaled the channel back up to 20nm, but still needed to properly size the LDD and halo patch. Starting with just the LDD, we made them 4nm by 4nm square patches. But through experimentation, we found that a taller LDD region was actually more beneficial since this shrinks the depletion region even further away from the gate, facilitating the gate to have more control over the channel deeper into the substrate as well. However, the taller/deeper LDD region also led to lower on-current which was problematic.

With a large channel (>15nm) and LDD pockets added, the EDP benefits were still not being realized, so we assumed (without actually checking) that the problem was short channel effects. Thus, we started to add halo and super halo doping to mitigate our fabricated short channel effects. If we actually had a channel experiencing those effects, these techniques could have helped, but instead all they did was decrease our on current and reduce our overall performance. Our solution to that was adjust the channel and source/drain doping as well as the supply voltage, but this brought us back to our former problem of the simulator not converging. After back and forth on altering the doping values of the halo pockets and LDD and channel and source/drain, we eventually decided to start adjusting the size of the pockets. Even going as far as having the super halo doping extend across the bottom off the channel. After running into the same issues as above, we abandoned the use of halo/super halo as well as LDD and re-increased our channel size and went back to drawing board.

2c. Episode III: Dennard Strikes Back

Our final, and most successful approach features a more intelligent and balanced version of Dennard scaling. The convergence issues with certain combinations of supply

[1] Beyonce et. all - "Halo"

voltage and doping concentrations meant we had to be clever with how we scaled down the channel length otherwise our off current and on current would suffer drastically.

At this point we decided to ignore Dennard scaling and continue to decrease the channel length anyway. The resultant was as expected: short channel effects. Off current started to creep up, on current took a hit due to pinch off, we even saw the threshold voltage drop to the point of the device not ever turning off. The doping values of the source, drain, and channel were at the maximum allowed by the mesh's convergence pattern and the supply voltage lever had already been pulled as far as we could without suffering from EDP drawbacks. The voltage had been reduced to about 1.2 V and the doping values were 2-4 times their original values. However, we were still using silicon dioxide as our insulator between the gate and the channel at this point and thus, we had the option of replacing it with hafnium oxide and thereby increasing the control the gate had on the channel (lower threshold voltage, etc). The higher CSF meant we could continue to shorten the channel, but soon enough the short channel effects crept back up.

The next lever we decided to pull in this approach was not halo doping, but rather shallow junctions. We dropped the height of the drain and source from 35 nm to 20nm and the short channel effects started to disappear. With that we began our process of reducing the channel width (currently 20nm also) bit by bit (approximately 1 nm at a time) until short channel effects showed up and then reducing the height of the drain and source even more. However, we soon ran into issues with on current dropping too low. This was because the shallow junctions have the side effect of increasing contact resistance significantly. This could be mitigated by increasing the doping in the source and drain which increases the number of free carriers and thus decreases the resistivity and resistance. Another way is to use strained silicon which increases the mobility of the source and drain and thus the decreases the resistivity and resistance. Yet another way was to use a more highly conductive electrode, such as silver or gold instead of aluminum. However, after the shallow junctions reached about 13 nm and the channel reached approximately 15nm, the doping and strained silicon and silver electrodes were not enough to keep the contact resistance low. With no way to counter the shallow junctions side effects, we couldn't keep reducing the channel length and using the shallow junctions as a mitigator for short channel effects. We needed another way to fight short channel effects, so we used the LDD technique.

The lightly doped drain didn't seem very effective at first. We started with the difference in doping of the source and drain vs the LDD pockets being 2-5 times. However, we soon realized that the greater the difference doping (to a certain point) the more effective the LDD pockets. Thus, we reducing the doping in the LDD pocket by orders of magnitude relative to the source and drain doping, and the effects became more obvious. The depletion region was pushed back into the drain and source and the channel could be safely reduced again. Each time we reduced the channel length, we went back to the doping values and supply voltage to see if a better combination was achievable to reduce the EDP. When

creating the LDD pockets, we had to decide on a height and width of the pockets. After experimenting with very short and very tall pockets (as in our second approach) we saw that while the taller the pocket, the better it was at mitigating short channel effects, the on current also dropped with such a lightly doped pocket in the channel. To strike a balance between these competing factors, we made the height of the pocket half that of the channel while making it as narrow as fabrication limits would allow (approximately 4nm). With the LDD pockets in place we were able to reduce the channel width even more to 13nm. The figure below shows a comparison of the initial transistor and our short channel transistor. The top figure is our final transistor and the bottom figure is the initial transistor provided.

With each change we made following the above philosophy, I mentioned that we went back and checked for better combinations of supply voltage and channel/source/drain doping values. This is because with each change, the mesh shifted, and thus the convergence

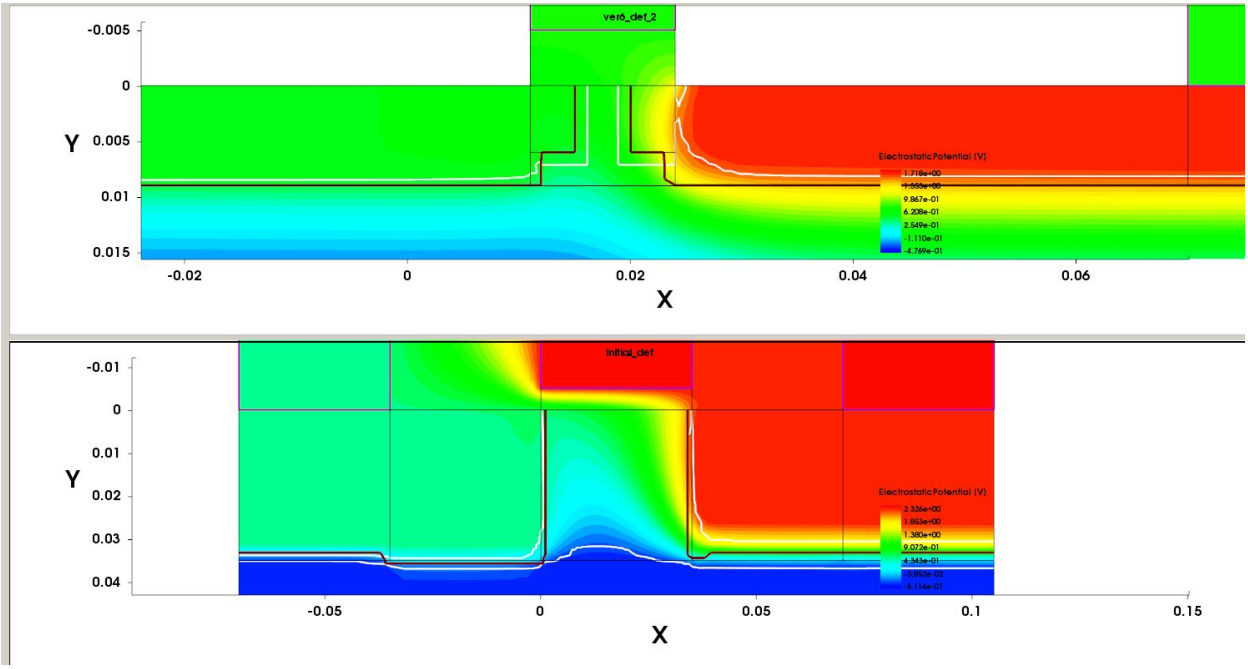


Figure 3: Depletion region and electro-static potential visualizations from Sentaurus Visual.

possibility shifted. This meant we could continue to appropriately Dennard scale the doping and voltages. It was with this methodology that we ended up at the supply voltage of 1.02V, however, the specifications provided demanded a speed of at least 200 MHz, and the supply voltage was too low to reach that speed, so we increased the voltage to 1.2 which bumped our speed up to above 400 MHz.

3. Lethal Transistor: ARMed with Math

There were quite a few calculations that went into optimizing the transistor. Modeling the performance of the transistor requires knowing the on and off current as well as the gate

to source capacitance as a proportion of the original capacitance. Useful metrics for understanding the performance before running the MatLab simulation include the threshold voltage as well as the charge sharing factor. The more repeated calculation we did was scaling the channel width, supply voltage, and doping as we did Dennard scaling. We also attempted to scale LDD and halo pockets using similar scaling principles, though with much less success. Below is a sample of the calculations we did for each of the above mentioned metrics:

Threshold Voltage:

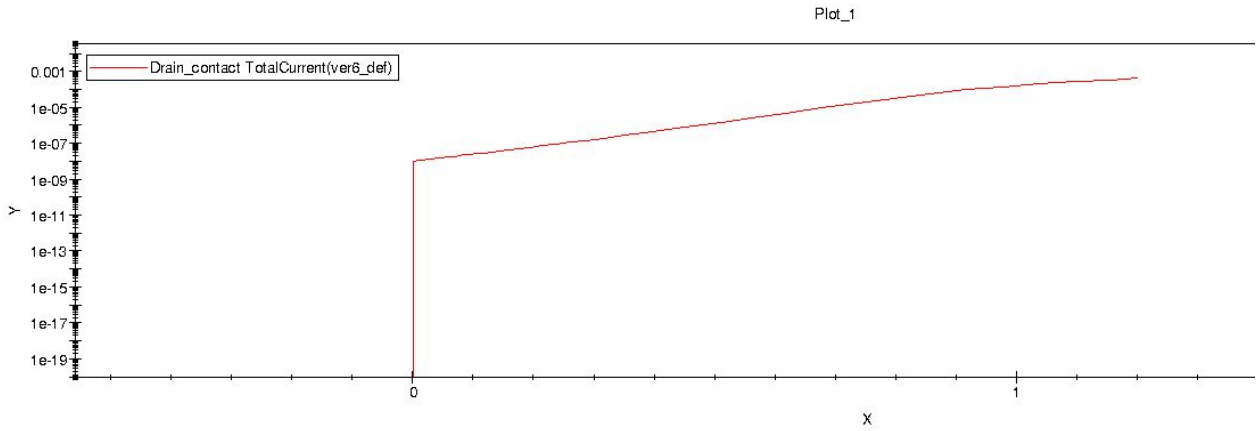


Figure 4: Log scale plot of total current and voltage on gate

The presence of the LDD pockets disrupts the uniform doping assumed when deriving the threshold voltage equation, thus we decided to estimate the threshold voltage. Based of the IV curve above and below we can estimate the threshold voltage at approximately 0.6V.

Charge Sharing Factor (CSF)

Because of the presence of the LDD pockets, fully calculating the charge sharing factor proved difficult. Thus, we chose to take a hybrid approach using the final formula for calculating CSF, but measuring/estimating values for the individual components.

$$CSF = 1 - \sqrt{\frac{W_s + W_D}{2 * L}}$$

Plotting the geometry of the transistor and visualizing the depletion region, we measured the widths which were approximately 5 nm on either side although the source side depletion region does look to be slightly smaller. We approximated both to be 5 nm.. The length of the channel (13nm) was used for L giving us a final CSF of 63%.

Capacitance:

Within this transistor we have 2 capacitances of importance. The C_{gs} and C_{gb}. To calculate C_{gs} we used the following equation:

$$C_{gs} = \frac{\varepsilon_{ox}}{t_{ox}}$$

Using this we calculated the following :

The initial Cgs was 0.11nF with $\boldsymbol{\varepsilon}_{ox}$ of 3.9 and t_{ox} of 35nm. The initial C_{GB} was 0.78 nF with the same $\boldsymbol{\varepsilon}_{ox}$ and a thickness of 5 nm. The final C_{GB} was 5nF using halfnium with an $\boldsymbol{\varepsilon}_{ox}$ of approximately 25 and and the same thickness. The final C_{GS} was 0.0175 nm with an $\boldsymbol{\varepsilon}_{ox}$ of 1 due to a lack of oxide (vacuum) and thickness of 57 nm. Thus the relative C_{GS} of our final transistor is 0.15.

Using a combination of the design philosophy discussed in section 2, the calculations above, and a lot of testing that we got the performance shown in the IV curve below.

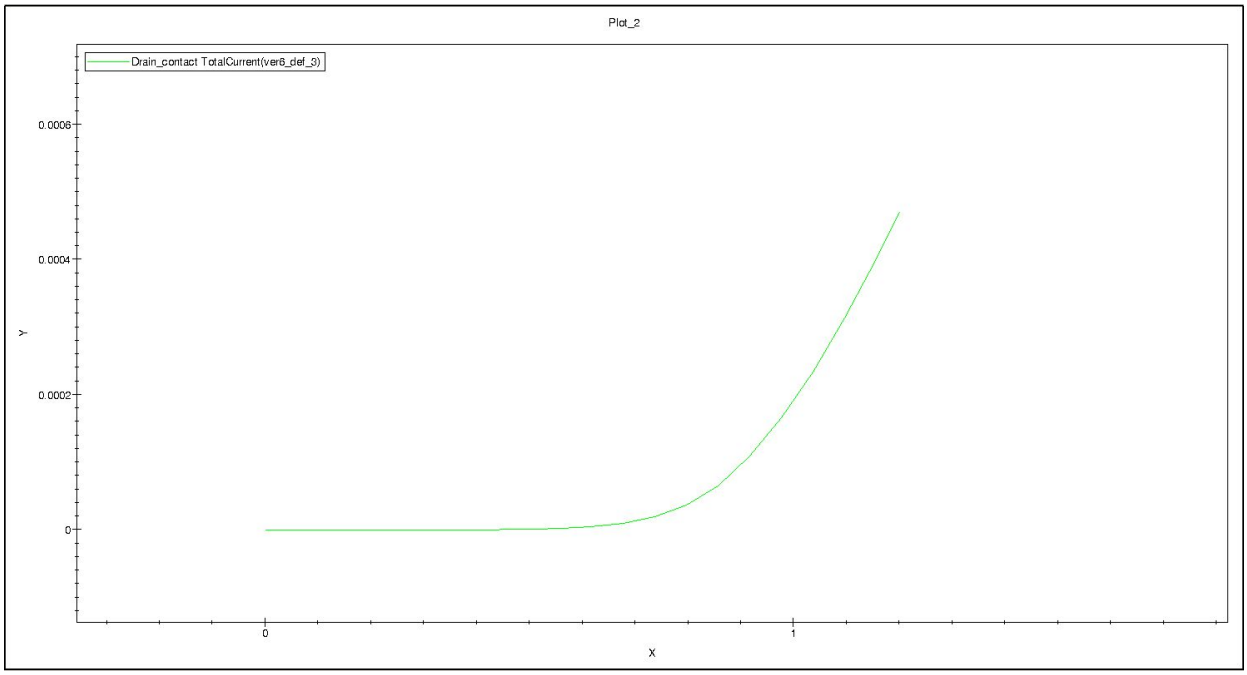


Figure 5: Total current vs. gatevoltage

4. More Surprising than Spectre

Through the process of designing this transistor we learned several things about the Sentaurus software, design pipeline, limitations and surprising findings about the techniques learned. Despite being imprecise, the 2d Drawing tools within sentaurus were quite useful to visualize the corners of different regions and patches, allowing us to then identify the precise corners and create patches via scripting.

Proper convergence was an issue that we ran into on multiple occasions. Non-symmetrical designs led to convergence failures as did certain combinations of relatively

high doping values. In the later stages of our designs, we identified a technique to improve convergence. Changing the supply voltage very slightly to have more significant digits led to proper convergence. For example, changing a supply voltage of 1.2 to 1.201 would help the simulation converge.

From our findings we noticed that some benefits were useless in isolation and others were not as beneficial as expected. Strained silicon for example, seemed to make little to no difference in our on-current although we expected the change in material to increase our on-current. Similarly the addition of LDD pockets, shallow junctions, and high K- metal gate can all have detrimental effects if overdone or used when unnecessary. This is why our initial use of some of these techniques made our transistor worse off, but used properly in our third iteration we realized the expected benefits

.

5. The Secret to intel igent Design

Ultimately, we found that the approach that yielded the most EDP benefits was only achieved after understanding the relevant calculations that needed to be made in addition to being more familiar with the software. The latter point gave us the ability to quickly test out ideas and develop intuition as to how the EDP changes as we change everything from channel length to LDD height to supply voltage. After understanding the limits of doping and supply voltages on convergence, we were able to optimize other areas and achieve the EDP benefits. Also prioritizing energy efficiency over speed (to a certain extent) provided better performance within the specifications of the device.