

Introduction to Semiconductor Devices

Prof. Dr. Naresh Kumar Emani

Indian Institute of Technology - Hyderabad

Module No # 13

Lecture No # 63

MOSFET Scaling and Technology Nodes

This document is intended to accompany the lecture videos of the course “Introduction to Semiconductor Devices” offered by Dr. Naresh Emani on the NPTEL platform. It has been our effort to remove ambiguities and make the document readable. However, there may be some inadvertent errors. The reader is advised to refer to the original lecture video if he/she needs any clarification.

Hello everyone welcome back to introduction in semiconductor devices in the last week we studied of basics of operation of a MOSFET. We looked at the saturation the linear and the sub-threshold regime of the operation. And then ,we also derived some basic expression of current as a function of gate and drain voltages. So what we have studied so far can be termed as classical MOSFET’s simple square law theory works there.

And these are the MOSFETs that were predominantly use in the sixties and seventies so after that there been a lot of developments. And today we will try to give you perspective of what happened in the last 40 years or so.

(Refer Slide Time: 00:57)

MOSFET Scaling

□ Proposed by Dennard *et. al.* in 1974

□ Reduce vertical and lateral dimensions by α

□ Reduce supply and threshold voltages by α

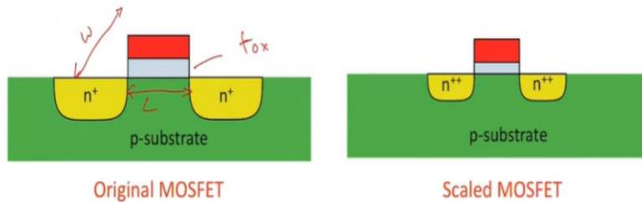
□ Increase doping concentrations by α

Scaling factor ($\alpha = 0.7$)

$$L \rightarrow \alpha L, w \rightarrow \alpha w, t_{ox} \rightarrow 0.7 t_{ox}$$

$$V_{DD} \rightarrow \alpha V_{DD}, V_t \rightarrow \alpha V_t$$

$$N_{sub} \rightarrow \frac{N_{sub}}{\alpha}$$



To start with, I would like to introduce the concept of scaling. The scaling was first proposed by Dennard and coworkers in the year 1974. So the essential idea is this so we have the MOSFET you know we have seen this we have the source drain and the oxide thickness and gate in all that. So we have the various parameters of the MOSFET ,such as ,you know ,the length here and then there is width which is going into the plane of the sheet.

And then there is oxide thickness and then the doping concentration right this is the most important parameters to calculate the threshold voltage and so on. So what Dennard proposed in 1974 was this, let us reduce the vertical and lateral dimensions right factor α so what you can do is, let us say we have the length L I will make it α time L. So I am introducing a scaling factor which I will call as α .

And most of the time this is so today we will assume that it is a 0.7, $\alpha = 0.7$,so we are essentially reducing the size by about 30% so we make α as 0.7 if L was initially 1 micron then after the scaling L become 0.7 microns. Similarly I would also reduce width by the same scaling factor and also the thickness and so on. So all the lateral and the vertical dimensions for example t_{ox} is vertical dimensions so I will make t_{ox} 0.7 times t_{ox} .

So I will do this sort of scaling along with that I will also scale my supply voltage for example the power supply VDD and the threshold. So VDD I will scale it has α times VDD and let us say V_T I will scale it as α time my V_T . So I will reduce my supply voltages and the threshold voltages and lastly we should also increase the substrate concentration or in all the doping concentrations. For example if I have drain concentration of let us say n^+ let us say scale geometry I will make it n^{++} .

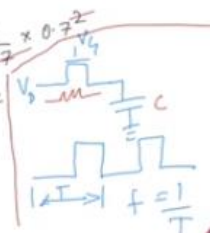
For example n^+ in the original MOSFET once I shrink the sizes I get a scale MOSFET like this all the sizes are smaller. But the doping concentration are higher in this case there is reason why this was proposed? We do not want to get into all that right now I just want to tell you know what is proposed and how it is benefices to us. So the doping concentration is also scaled some practical reasons alright.

So we are left with from the original MOSFET much you know smaller dimensional MOSFET with higher concentration so what? Why should this matters to us?

(Refer Slide Time: 04:02)

Scaling classical MOSFETS

Lateral and vertical dimensions reduce by 30% ($W = 0.7, L = 0.7$ and $T_{ox} = 0.7$)
 Supply and threshold voltages reduced by 30% ($V_{DD} = 0.7$ and $V_T = 0.7$)

Smaller	<p>Die area $w \times l = 0.7 \times 0.7 = 0.49$ \hookrightarrow Reduces by 50%.</p> <p>Total gate Capacitance $C_g = C_{ox} \cdot W \cdot L = \frac{\epsilon_{ox}}{t_{ox}} \cdot W \cdot L = \frac{1}{0.7} \times 0.7 \times 0.7$ \hookrightarrow Reduces by 30%.</p> <p>Gate Capacitance/unit area $C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{1}{0.7} = 1.43 \times$ \hookrightarrow Increases by 40%.</p> <p>Current $I_D = \mu \frac{C_{ox} \cdot W}{2L} (V_{GS} - V_T)^2 = \frac{1}{0.7} \times 0.7^2$ \hookrightarrow Reduces by 50%.</p>	
Smaller	<p>Delay in circuit $t_{delay} = \frac{C \cdot V}{I} = \frac{0.7 \times 0.7}{0.7} = 0.7$ \hookrightarrow Reduces by 30%.</p>	
Power efficient	<p>Power dissipation $P_{diss} = C \cdot V^2 \cdot f = \frac{0.7 \times 0.7^2}{0.7} = 0.49$ \hookrightarrow Reduces by 50%.</p>	

EE @ IIT Hyderabad

It matters to us because we will see that this scaling actually improve performance to show you that I will go step by step and compute some parameters here. So to start with I am assuming that the lateral dimensions and the vertical dimensions and reduce by 30% that means let us say initial W , as 1 I will make 0.7 L was 1 I will make it 0.7 t_{ox} was 1 I will make it 0.7.

It is just shrunk by 30% similarly voltages are scaled by 30% so V_{DD} of 0.7, V_T of 0.7. So now we want to what happens to these parameters after scaling. To start with a die area maybe I should not call it die area let me call it MOSFET area die is basically a chip you fabricate various devices so you make blocks you call them as dies. But anyway for our purpose it does not really make any difference we can just call it MOSFET area and we are only looking at the active region you know under the gate.

So this area for the MOSFET would be $W \times L$ right you have a length and width that is the area under which inversion layer found. So this is what happens to $W \times L$ once you scale it it becomes 0.7×0.7 that should be equal to 0.49. So after scaling the MOSFET area reduces by 50% nothing is surprising so far right that is what we will expect I mean it reducing the sizes. What happen to the total gate capacitance?

Well remember this total gate capacitance is in farads so the expression for this is let me call it C_G will be C_{ox} which is the capacitance per unit area we have defined for oxide times the width into length by area right. So this will be equal to ϵ_{oxide} by t_{oxide} into $W \times L$ so if I scale it ϵ_{oxide} is a material parameter it does not change. So I will still retain that but thickness is scaled so it is divided by $0.7 \times 0.7 \times 0.7$.

So what I end up getting is 0.7 times reductions in the gate capacitance so this reduces by 30% what about gate capacitance per unit area? Gate capacitance per unit area we have already seen this is simply C_{ox} and that will be ϵ_{oxide} by t_{oxide} . So this reduces by or rather you know

it is divided by 0.7 so it effectively it turns out to be equal to 1.43 into whatever those capacitance. So, it increases by 43% or you know let us round it of to 40%.

So this increases by 40% why are we doing this? Well we are just trying to evaluate this so that we can draw some conclusion of what happens in a scaled MOSFET or you know if you shrink the size of a MOSFET what happens. So now what happens to current we know the expression of current I_D is basically $\mu C_{ox} \frac{W}{L}$ I will take saturation current so I will simply say $(V_{GS} - V_T)^2$ there is no harm in this by $2L$ so this is my I_D expression.

$$I_D = \mu C_{ox} \frac{W}{2L} (V_{GS} - V_T)^2$$

And remember in this I_D expression this is per unit area C_{ox} is capacitance per unit area. So let us substitute and see so μ anyway it is material parameter I do not care C_{ox} is 1 by 0.7 W and L both will scale by 0.7 so they cancel out. So I will be left with 0.7 square so what you see is the current in a MOSFET reduces by 30%. And what happens to the delay? So delay is a important parameter in any MOSFET circuit the reason is if you think about any digital chip we will deal with logic 0 and logic 1.

So essentially the transistor is switching between 0 and 1 and this simplest way to model this is form of a RC circuit. So what happens is let us say if I have a MOSFET I will small let me take a MOSFET and then I will basically MOSFET are drive going to drive some loads typically the output of the MOSFET will be given the input of another MOSFET which is essentially input means the gate.

So, gate will present a capacitance so I can model in simplest form a MOSFET circuit as you know simply a RC network. So the capacitor the load is capacitance C and MOSFET I can model

as resistance. So when I have RC network I have studied in basic circuits what happens right we have we analyzed it you know you would have analyzed it great detail. So what is the time it takes to charge a capacitor in RC network? Well essentially it is going to depend on the time constant RC.

In this case let us say I want to charge it to certain voltage V and I know that the MOSFET is contributing a current I . So the time to charge the capacitor will simply be I will call it T delay that is going to be the capacitance is C and you want to charge it to voltage V . So into V is the charge and if I divided by current it is flowing through MOSFET then I will get my time it takes to charge that much of charge on the capacitor.

So time delay and clearly you know this capacitance is going to be total capacitance oxide so this will be 0.7 into voltage will be scale by 0.7 divided by current scale by 0.7 . So roughly this will be equal to 0.7 so that, time delay as reduced by 0.7 . Why is this important? Well, the reason it is important is the digital circuit ,in a digital circuit you will have switching behaviour like you have a input it switches and so on right will have cycles like this.

So that time delay this is going to be a t and the frequency of operation it is going to be determined by the time delay. The simple frequency is going to be inversely proportional to 1 over t . So if a time delay is small I can switch to a circuit faster that mean I can charge and discharge so essentially I can perform lot of operations. So if my time delay is reduced the frequency of operation is increasing if it is reduced by you know 0.7 that means it is reduced here my notation is 30% .

Then the frequency increase by about 43% and finally the power dissipation is another important metrics. So we said that RC network is essentially a simple model of a MOS digital circuit and if you trying to charge capacitor to voltage V then the amount of energy is stored in the capacitor

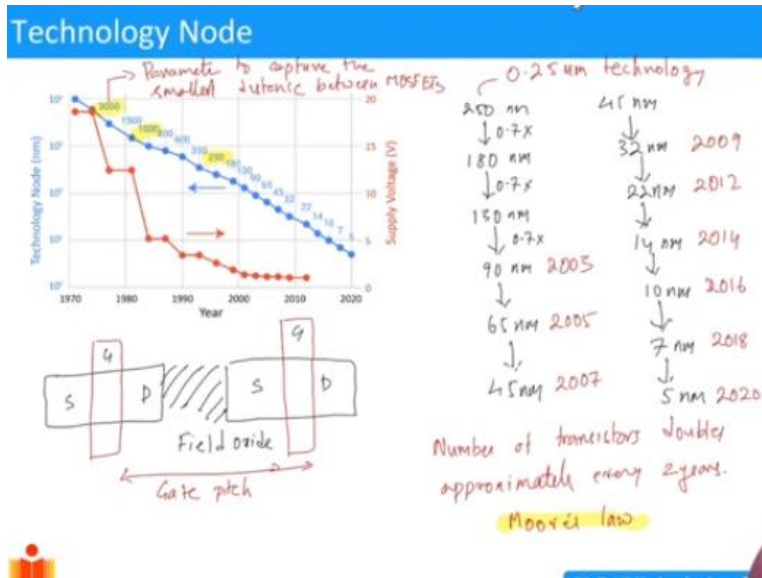
after charging is going to be $\frac{1}{2}CV^2$. So you turn on the MOSFET you charge it to $\frac{1}{2}CV^2$, you turn it off, that voltage will be the energy will be dissipated into the resistor or in the MOSFET.

And the total amount of energy will be in the process you will lose CV^2 , of energy so that dynamic power dissipation of a MOSFET in dynamic is going to be $CV^2 \times \text{frequency}$. Well I mean if you study any digital circuits you will actually going to get detail of what is dynamic or dissipation static power dissipation and so on. But for now I want to accept that if you a C as capacitance and V as supply voltage.

But total amount of power dissipated when you are switching up 0 at 1 and 0 is going to be CV^2f where f is a frequency. So what happens to this parameter well C is reduced for 0.7 voltage is 0.7 square F is basically 1 over t so this will be divided by 0.7. So what we end of getting of this is going to be scaled by 0.49. So power dissipation reduces by 50% so what happens because of scaling when you shrink sizes of MOSFET you are first of all making it MOSFETs smaller right you are making it faster.

And you also making it power efficient so these are very good properties to have so digital circuit has become much more efficient and this is the basic reason process why scaling process is adopted.

(Refer Slide Time: 13:49)



So this was as simple as that so if you look at the impact of scaling in the nineteen seventies and eighties in the early nineties that is captured in this graph. So what you are seeing is basically the frequency of a microprocessor over the years from seventies to 2000. And many you may not even know or remember or be aware that in the eighties or even seventies frequency of the processors was basically few megahertz eighties it was hundreds of megahertz.

So the 286 processor having may be you know 200 megahertz or 100 megahertz or so that is it. So here 286 processors which is one of the popular this is 10 megahertz even smaller. By the time we come to 386 it is still you know 25, 50 megahertz and finally when you come to Pentium it is hundreds of the first Pentiums are 100 megahertz or so. And then finally they kept on increasing the reason this happened was process of scaling that was described.

So this increase because of scaling of course the lot of technically much developed to achieve this but over time in the 30 years we have tremendously improved the processing capability of the MOSFET. So if you take let us say you know when you are actually designing processor one of the most when you are actually designing a processor of the most important features is power dissipation.

Let us say power dissipation we can accept about let us say 100 milli watts not 100 milli watts it is 100 watts of power dissipation in a chip. So if you scale a particular technology and make it smaller what happen? We are able to more transistors you know double number the transistors for same power dissipation. Because we saw that the power dissipation reduces by 20% so effectively you can increase the number of MOSFETS by same amount you know the same amount.

You can make them number of MOSFETs double we still have the power dissipation and then what else the time delay has reduced. So the frequency of operations has increased by about 42% or 43% we saw. So overall this scaling process is enable technology to evolve over the last 30 to 40 years. If it seems very simple well the idea is quite simple but implemented took lot of challenges , we have to overcome lot of challenges.

You might have heard about nanotechnology you know is a popular term but the entire field of nanotechnology I could say was developed so that the tools and techniques were developed so that we can achieve scale in the MOSFETs. So in that way the MOSFET is the driving force of nanotechnology. Even today the foremost forefront of nanotechnology is actually what we use in IC industry.

So this is the background and so this leads me to next step which you know what we call as Moore's law you might have heard about Moore's law in nineteen sixties Gordon E. Moore was the chairman of Intel he predicted that the number of transistors in chip actually doubles every 2 to 3 years. So we want to understand what that it is and we are trying to understand that I would like to introduce a term called as technology node.

So this is very important because even today you know follow the technology news you will see Samsung as you know come out of 5 nanometer technology where Intel as demonstration 5

nanometer technology and 3 nanometer technology in the research stage. So what does it mean? That is what we have to introduce so in the previous lectures we have shown you how the MOSFET looks like we have shown you the top view side view to the MOSFET.

So if you look at it from the top a MOSFET will have the source drain region separated by gate region. So this is the view we said source drain and the gate and on a chip you can have many such MOSFETs. For example I can have another MOSFET which will be adjacent to that so let us say it has same width and same gate length this is another MOSFET. Now one of the prominent or important feature of the technologies the spacing between these 2 MOSFET's.

So of course in between them they have this what will call as the isolation region you know field oxide or now a days it is called shallow trench isolation different technologies. Essentially you have to you do not want this MOSFET's to communicate with each other so will have what is called as field oxide in all the times now it is STR. So now the important parameter to characterize a particular technology is this distance which we will call as gate pitch which essentially how close can your transistor be in a particular technology.

So of course we will like to be in 50 nanometer apart or 10 nanometer apart but that is not possible especially if you do not have the tools to achieve that sort of a precision. So and you know this is what was called as a technology node parameter of course I am being very you know I am only giving a qualitative description there is a more precise description of this which will call us you know the distance between the half pitch of the DRAM cell.

So; DRAM memory is basically always the more aggressive form of scaling so you have the memory switches which was scale fast so they have the smallest devices we make. And if you look at the distance between 2 gates and look at that as pitch and take half of that , that is called as

technology node parameter. So if you look at nineteen seventies this was about 3 microns and by eighties it became 1 micron and by quick nineties let us say this parameter became 250 nanometers.

So what this number is telling is essentially it is simply a parameter to capture minimum feature sizes or capture the smallest distance between MOSFETs. The smallest distance between MOSFETs this was the past but now a you know we actually the meaning is lost actually now days it does not physically represent anything it is just that you know this notation is being continuously used even today.

So what happens is let us take 1 number let us start say 250 nanometers right so I can say 250 nanometers I will write it here. I will take that distance that particular technology, this was I think demonstrated in 95 let us say approximately. What I will do is, I will shrink it by a factor 0.7x what happens what do you get? The features size now will be close to 180 nanometers and then if you shrink it again by 0.7 times this will be 130 nanometers and if you shrink it again we get 90 nanometers shrink it by 0.7 and so on.

So if you want to write out more it will be after this it was 65 nanometers and then that lead to 45 nanometers and that lead to 45 nanometers will lead to 32 nanometers, that will lead to 22 I believe 32 into 0.7 will be 22 nanometers and then 14 nanometers then it will come to 10 nanometers and then seven nanometers and then 5 nanometers. So essentially what we do is every, you know we are physically packing more transistors and we saw that when we scale it 0.7 x the areas reduces by 50%.

So essentially we can increase by number of transistors I mean you, can double the number of transistors. If you take a certain chip in let us say 250 nanometer technology okay when I say 250 nanometers or you know sometimes we also call it as 0.25 micron technology. So what this means is the smallest MOSFETs that you can make roughly at the order of 25 microns. This is a very now

a days this term does not have any double precise meaning it is essentially simply the capture the number of transistor is doubling.

Sometimes you know nowadays it means that the effective performance of the processor is or your chip is effectively as though it is scaled by this distance. So this is called as technology nodes this numbers are known as technology nodes and we see that over the last 40 years there is a continuous decrease in the number technology you know the size of the technology nodes and this is the log scale to remind you this is on the log.

So you are seeing almost exponential reduction of the sizes and let us say if you take 250 nanometers. This was somewhere 95 the once I remember I have encountered personally was basically is 90 which was introduced I think by year 2003 65 which is 2005 and then 45 which was in 2007 and then 32 in 2009 22 I think in 2012 it took a little bit of time. Because Intel introduced FinFETs there and then 2014, 2016, 2018 and 2020 so what you are seeing is?

Every 2 years the number of transistors on chip is essentially doubled so this is what was you know the famous prediction of Gordon E. Moore was. And we sort of the semiconductor industry price to keep up with sort of a trend. Even though effectively now the transistors are definitely be not you know you do not have 2 transistor spaced by about 5 nanometers it is not true it is much more complicated than that. But we still use this notation you know in a kind of a homage to Gordon Moore prediction.

So essentially the number of transistors doubles approximately every 2 years or effective performance I could say of the transistors approximately doubles every 2 years. So this is what is known as technology scaling and Moore's law which is Moore's law. So while we do this process the performance improves significantly. It turns out that this is not that simple there are various challenges of course if you try to fabricate a PCB you would have seen, that you know the PCB

that you see in the computers or in an electronic components they have this fine copper lines which are connecting electronic components.

And if you try to fabricate them at home you will see that it is quite challenging to make very fine features. Now we are talking of nanometer scale features how do we do that it is a whole lot of a difficult process but thankfully semiconductor technology or IC industry have developed these processes. So that these can be realized so I will in the next couple of I will stop this video now in the next video I will talk little bit about what is the various challenges in scaling?

And what is the limits some of the limits? And then we will take it forward from that alright thank you very much for you attention.