

Nanoelectronics: Devices and Materials
Prof. Navakanta Bhat
Centre for Nano Science and Engineering
Indian Institute of Science, Bangalore

Lecture - 02
CMOS Scaling Theory

In the last class you know we had just talked with the slide on the periodic table of materials, and I had mentioned that if I will to ask somebody may be 10 years ago you know what are the ingredients of the silicon chip you would probably see just a handful of materials; as it is indicated here.

(Refer Slide Time: 00:33)

New materials are essential ...

The periodic table of the elements

1A	2A	3A	4A	5A	6A	7A	8	1B	2B	3B	4B	5B	6B	7B	0		
1 H	2 Li	3 Be	4 B	5 C	6 N	7 O	8 F	9 Ne	10 Na	11 Mg	12 Al	13 Si	14 P	15 S	16 Cl	17 Ar	
18 K	19 Ca	20 Sc	21 Ti	22 V	23 Cr	24 Mn	25 Fe	26 Co	27 Ni	28 Cu	29 Zn	30 Ga	31 Ge	32 As	33 Se	34 Br	35 Kr
36 Rb	37 Sr	38 Y	39 Zr	40 Nb	41 Mo	42 Tc	43 Ru	44 Rh	45 Pd	46 Ag	47 Cd	48 In	49 Sn	50 Sb	51 Te	52 I	53 Xe
54 Cs	55 Ba	56 L	57 Hf	58 Ta	59 W	60 Re	61 Os	62 Ir	63 Pt	64 Au	65 Hg	66 Tl	67 Pb	68 Bi	69 Po	70 At	71 Rn
72 Fr	73 Ra	74 A															
		75 L	76 La	77 Ce	78 Pr	79 Nd	80 Pm	81 Sm	82 Eu	83 Gd	84 Tb	85 Dy	86 Ho	87 Er	88 Tm	89 Yb	90 Lu
		91 A	92 Ac	93 Th	94 Pa	95 U	96 Np	97 Pu	98 Am	99 Cm	100 Bk	101 Cf	102 Es	103 Fm	104 Md	105 No	106 Lr

NPTEL

Metals
Metalloids
Non-metals
Transition Metals
Gases

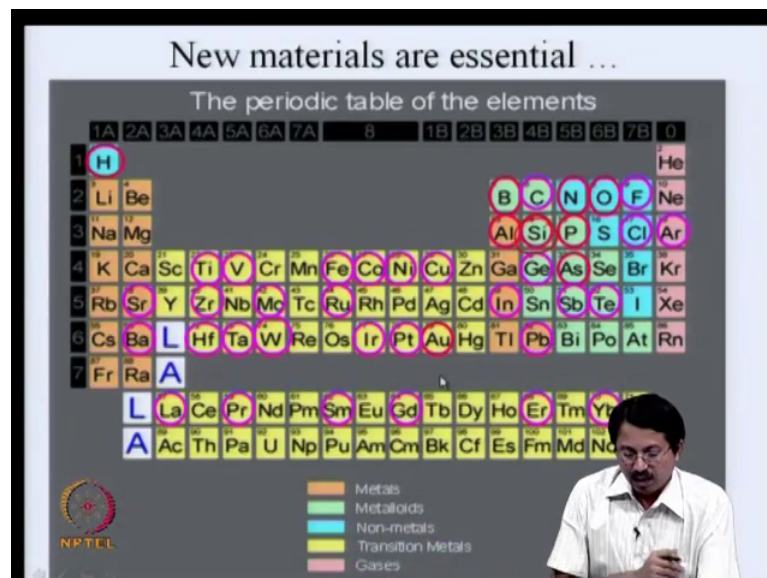
Silicon of course is a group 4 semiconductor and in addition you would need arsenic and phosphorus to dope it n-type and boron to dope it p-type right. And of course you would need oxide because we have been talking about the FET structure you need silicon oxide as a gate insulator and also you would need oxide for isolation and so on and so forth right. And once you complete your transistors you need to connect them so you would certainly have aluminum as an inter connect layer right. And you know hydrogen is a lightest element no matter what you do you will always have hydrogen.

And it turns out we also intentionally introduce hydrogen in our silicon technology to do what is called surface passivation of silicon. In other words when you put oxide on silicon you see it is a system of 2 different materials when you go from one material to

the other material at the interface you will get a lot of defects right. By having hydrogen out there hydrogen can passivate these defects which are essentially dangling bonds of silicon which are un-terminated bonds. And hydrogen would minimize the defects which is extremely important in a field effect transistor, because interface is the key in a FET device right.

And you know once you complete your chip you passivate the chip with what is called a silicon nitride and hence you will have also nitrogen in it. And you when you put it in package you will probably have gold wires connecting the silicon pads on to the package pins right. So this was very simple just handful of materials where ingredients of your silicon chip.

(Refer Slide Time: 02:13)

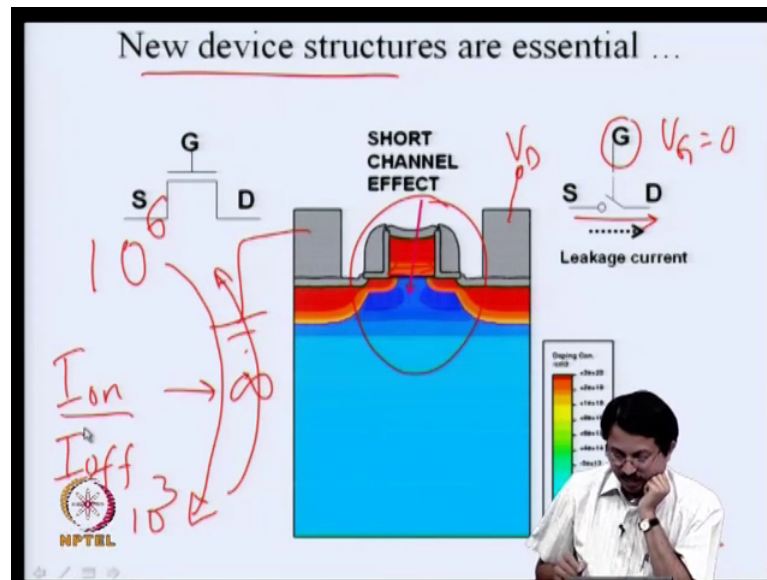


However, the world has changed completely now you know if you were to ask the same question what is the ingredient of a silicon chip today you will see you know a plethora of materials in the state of that silicon chip today, right. As an example for example, you know in addition to the doping materials which were phosphorus and arsenic they are also using antimony as a doping material right, we are also using indium as a doping material right. And you know for materialization we are using a variety of materials as well for example, copper is replacing aluminum right.

So, similarly there are reasons why we introduced a variety of materials. In fact it is fair to say that we are sort of digging through the periodic table it is sort of an exploration in

periodic table to really build nanoelectronics chips. In fact, as you know the course is titled nano electronics devices and materials. In fact as we go on in future lectures we will also understand why some of these materials are being introduced right. So, this is very crucial right. So, an appreciation for materials technology is very important.

(Refer Slide Time: 03:18)



New device structures are also very essential going forward right. This is our conventional MOS structure, metal oxide semiconductor field effect transistor right. And as I have already mentioned what we have done over the years when we have been scaling the technology is really to you know scaled this gate length; you know the length between the distance between the source terminal and the drain terminal that is your gate length. In other words, typically in a n channel transistor the source terminal is at ground potential and you would be applying your drain voltage which is positive voltage to the drain terminal.

Now, what is happening with scaling is that this drain terminal is coming closer and closer to the source, this is what we call a proximity effect right. I mean we need to consider the 2 dimensional distribution of electric field. In other words the drain electric field has started influencing the behavior of the transistor out here. Even without the gate being there to turn on the transistor meaning: I really when the gate voltage is 0, right your V_g equal to 0 should corresponds to very very ideally 0 current, but you never get the 0 current none the less a very small current should be out there but because of this

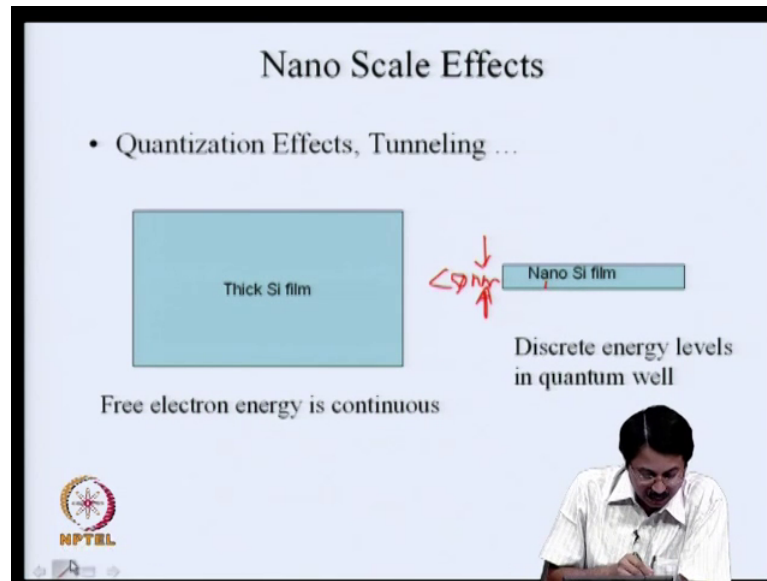
proximity effect you have very significant leakage current. You know it is becoming harder and harder to turn off a transistor to turn off an FET as we are scaling the dimensions of the FET.

One of the important abstraction if we recall they said transistor is a switch right I mean for it to act as a switch ideally it should have a large on current and very small off current. In other words one of the very important metric of transistor is what we call I_{on} over I_{off} ratio. Ideally we wanted to be infinity right meaning I_{off} should be 0, I_{on} typically for a given on current. But of course you know you cannot get infinite on to off current ratio, but traditionally in the first we have been easily getting on to off current ratio which is of the order of 10^6 ok million times on current is a million times more than off current which is a fairly a good abstraction of a switch right because when it is on it is conducting I mean 10^6 times current when it is off.

But now you see this has started coming down, as we are starting to move this down you know it could come down as low as 10^3 or even less than that. Now imagine if your on current is only 100 times greater than your off current you know that is not a very good abstraction of a switch you see. So, in other words we need to do something out here something in this transistor structure so that we can bring this on to off current ratio back to reasonable number you know 10^4 ; 10^5 ; that kind of a thing right. We can do that only by using new device structures.

And hence, as we will go on in the course we will also see how do we engineer these transistor, so that we can really get better transistor right in spite of the fact that we are scaling, we are bringing this grain voltage closer to the source terminal right. So that is very important challenge that we need to deal with.

(Refer Slide Time: 06:47)

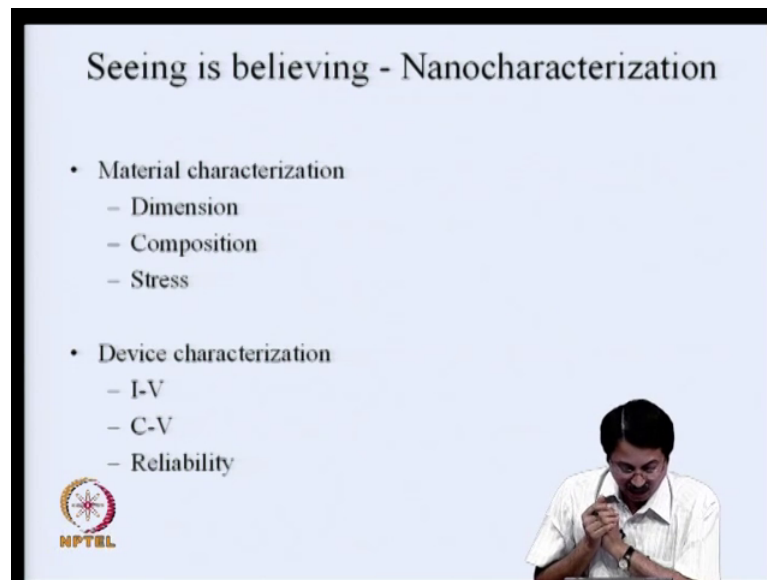


So, in addition to although is you know there is something interesting that also comes about, in other words when we start scaling these dimensions to Nano scale, you know some physical properties themselves start changing you see for example, we have been building chips on silicon as I said you know we call it bulk silicon, bulk silicon essentially means that you know you have fairly thick silicon film.

Now, today we are talking of building chips in silicon film which could be as thin as of the order of you know less than 50 nanometer, you know that is ultra thin film. Now as we start thinning down it further right the properties of this silicon film which is a nanosilicon film are no longer the same as the properties of a bulk silicon film right, the physics changes at this nanoscale. For example, the band gap is no longer the same as the thick silicon film, if the band gap changes. Obviously you would expect all the semi conducting properties would change like carrier concentration so on and so forth.


So, that is something that we also need to understand and you know accordingly design of course you cannot get rid of quantum effects right, the size effect is essentially coming because of quantization but we will exploit those effects and intelligently device and design the transistors right so that we would be able to scale the technology further.


(Refer Slide Time: 08:27)



Seeing is believing - Nanocharacterization

- Material characterization
 - Dimension
 - Composition
 - Stress
- Device characterization
 - I-V
 - C-V
 - Reliability

 MPTCL



So then characterization is also very important right you have made something and you need to really see if you have tried to make a transistor which is 40 nanometer, after doing all the processes in your nano Farad, have you actually got 40 nanometer right you need to do lot of characterization, material characterization, device characterization and so on and so forth.

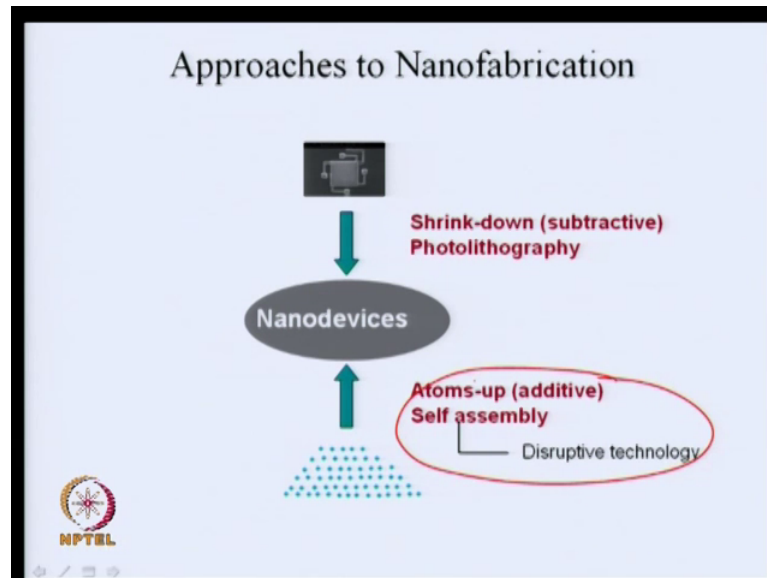
We will also have quite a bit of discussion in this course on material characterization which includes measuring dimensions right whether it is 40 nanometers, 35 nanometers, and so on and so forth. Composition, if you are making a new material you know what is the composition of that material right we will have techniques to really investigate that right. Stresses become extremely important right you have multiple materials kept together there you know thermal expansion coefficients are not going to be identical.

So, hence there will be stresses that will be developed in these materials and that they could have significant impact on the device sometimes detrimental. On the other hand we can also exploit this stresses for our benefit and design devices intelligently as we will see later right we actually do strain engineering and we know get better devices.

And of course you know in terms of I-V characterization, eventually your electrical property is what is important when you use it as circuits. So such as current voltage and capacitance voltage characteristic and reliability of the device will it operate only today

or it will sustain all the harsh environment for next 5 to 10 years you know that becomes extremely important right. So, we will have a discussion on all these aspects as well.

(Refer Slide Time: 10:08)



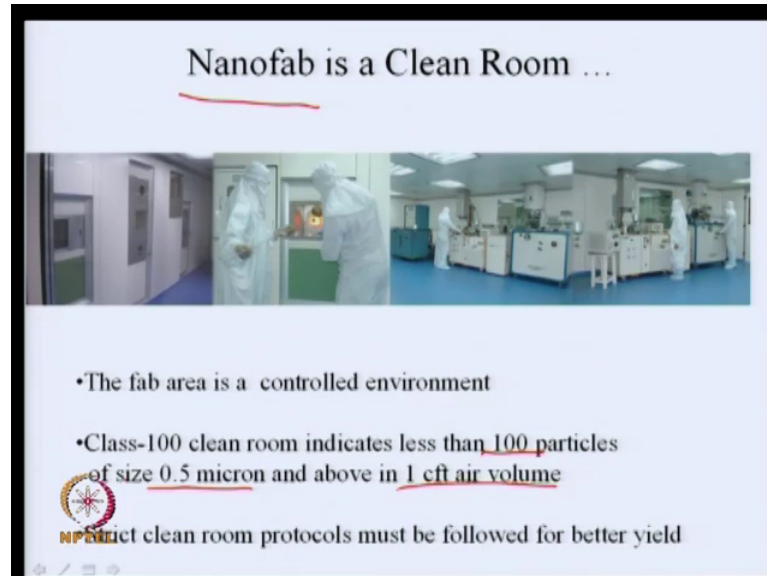
And also when we talk of various processing right you know there are 2 ways you can approach nano dimension right, one is what is called a top down approach that is you start from large dimensions and start etching materials that you do not want and bring it down to the nanoscale right so that is what we call a top down or shrink down approach, which is also sometimes called subtractive approach. Because you are trying to get rid of films that you do not want and the key enabler here is photolithography right using that we create these devices.

There is another approach that chemists have been working on this kind of approach you bring atoms together and you get a new product right that is assembling the atoms right from the atomic scale you can reach the nano dimension right. So, these are the 2 ways you can arrive at nanodevices.

But as of today you know we have no clue yet although there is lot of research going on in figuring out how to really create technologies the so called self assembly technology to create nanodevices. If we do that, that would really be disruptive technology, biology builds the network big systems like that right but all the technology that we have been building is essentially shrink down technology right. So, all the processes what we are going to discuss in this course is only going to be restricted on this there maybe a very

brief mention on you know self assembly kind of technology is right so that you know where we are having.

(Refer Slide Time: 11:50)

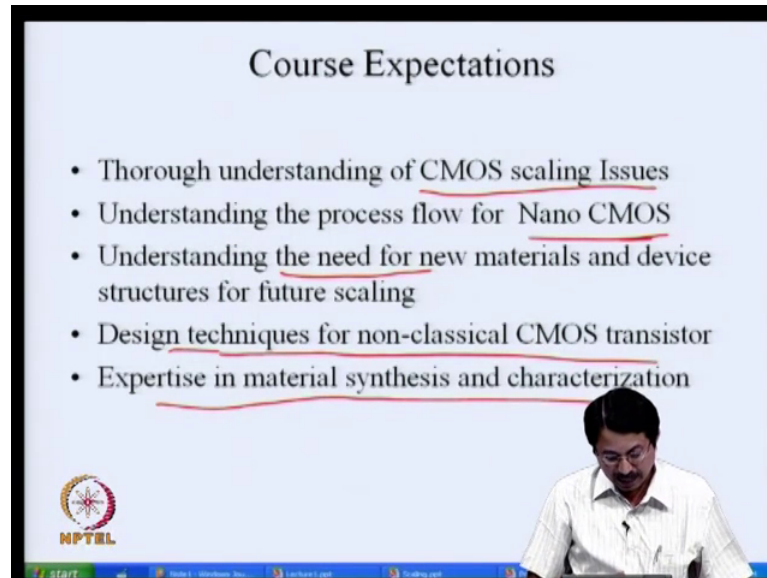


And you know these devices, these chips are essentially made in facility is that are called nanofabrication facility right which is essentially a clean room right. Clean room means that you know this is an area where we have extremely fine control on environmental parameters right. The environmental parameters include the dust particle in that room for example, we say that a clean room typically is a class 100 or class 10th clean room, what it means is that in the if it is a class 100 clean room, if I take 1 cubic foot of air volume in that room. And if I start counting particles of size 0.5 micron and above I should not count more than 100 particles, then it qualifies as class 100 clean room.

Just to give you a proper context the typical air conditioned office room may be class one million room meaning you will be easily count a million particles right. So, we will have to go through a very elaborate process to filter out all that and create a clean environment. And in addition there will be very strict temperature and humidity consideration. This is important because only when you do that you will get very high yield, what we mean by yield here is if I make 1000 chips out of 1000 how many are working chips right that is what we mean by yield 100 percent yield is what we always air for.

If you have defects for example, if you have dust particle if they sit on your silicon wafers. Obviously, you will not get that chip in that location as a working chip right so that is as simple as that ok. So thus ends to just set the expectations for this course.

(Refer Slide Time: 13:31)



Course Expectations

- Thorough understanding of CMOS scaling Issues
- Understanding the process flow for Nano CMOS
- Understanding the need for new materials and device structures for future scaling
- Design techniques for non-classical CMOS transistor
- Expertise in material synthesis and characterization

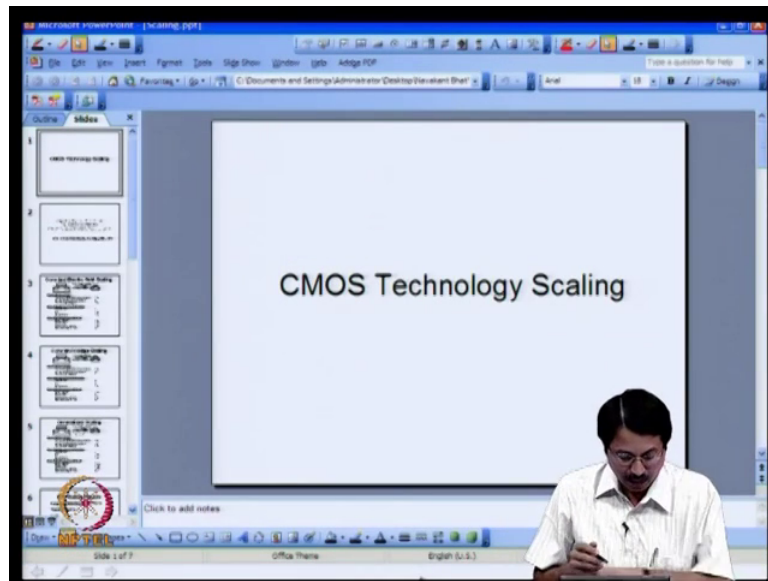
NPTEL

You will gain a very thorough understanding of CMOS scaling issues. In fact, we will get started on CMOS scaling immediately after this slide and you will also understand the state of the art process flow, how does one put together different semiconductor processes. In other words what is called process integration to be able to create a state of the art CMOS technology, we will also understand that through this course.

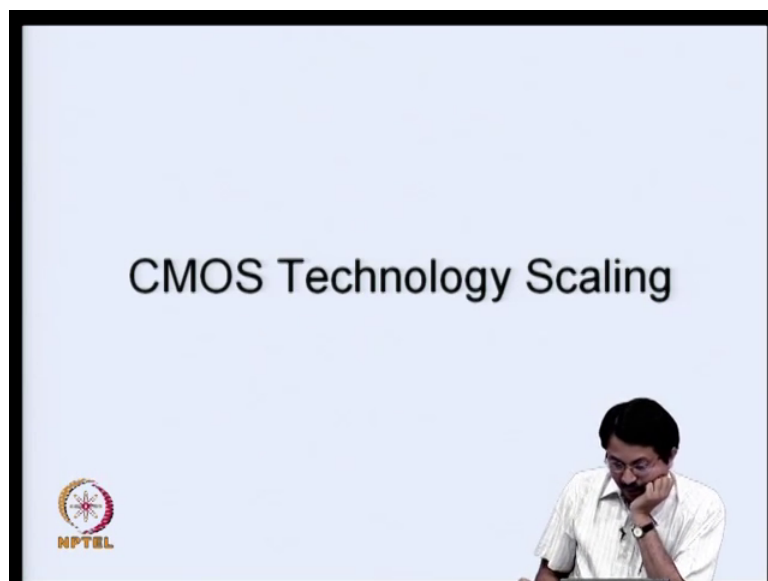
And you will also understand why do we need new materials and device structure I have already mentioned that very briefly but we will take several cases and explain why do you need this right and you will also have a fairly good knowledge on design techniques to be able to do what is called a non classical CMOS transistor. Meaning, remember the device structure I showed you earlier a simple a source drain and gate structure that is what we call a classical transistor which has really work for last of 4 decades also. Now we are talking of non classical CMOS transistors right.

So you will also have a fairly good idea on you know understanding the design techniques for that and hopefully you will also gain fairly good expertise in material synthesis and characterization through this course right, so this is what we aim for through this course. So, then let us gets started with the scaling ok.

(Refer Slide Time: 15:08)

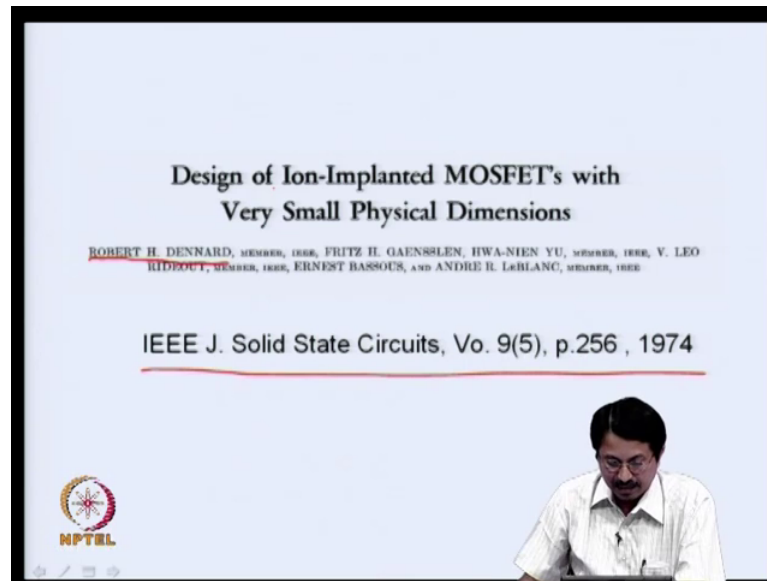


(Refer Slide Time: 15:11)



And again here we are looking at CMOS Technology Scaling.

(Refer Slide Time: 15:20)



So, you know you should look at this paper sometime you know this is of course available IEEE archives which was published in journal of solid state circuits in 1974, this is a very classic paper right which is being referred whenever you talk of scaling.

This was you know published by a group headed by Dennard at IBM way back in 1974. And this was essentially titled as design of iron implanted MOSFETs with very small physical dimensions right this was as you know published in journal of solid state circuits. And essentially through this paper the authors present certain guidelines for scaling and also sort of explain why do you want to scale the CMOS technology. So, let us try to understand that ok.

Constant Electric Field Scaling

Technology scaling
Scaling factor $K > 1$

Primary scaling factors:

T_{ox} , L, W, X_j (all linear dimensions)	$1/K$
N_a , N_d (doping concentration)	K
V _{dd} (supply voltage)	$1/K$

Derived scaling behavior of transistor:

Electric field	1
I_{ds}	$1/K$
Capacitance	$1/K$

Derived scaling behavior of circuit:

Delay (CV/I)	$1/K$
Power (VI)	$1/K^2$
Power-delay product	$1/K^3$
Circuit density ($\propto 1/A$)	K^2

The diagram shows a cross-section of a transistor with a channel length L. The scaled version has a channel length of L/K. The diagram is annotated with red circles and arrows.

The MPTEL logo is visible in the bottom left corner.

Now, there are different kinds of scaling but when you go through that paper you come across this term called constant electric field scale, this is the keyword here constant electric field scaling. So, what this scaling theory this is what we call ideal scaling theory for constant electric field scaling of CMOS technology ok.

So, what it says as that you have a transistor at any given time, let us say this is a state of the our technology right does not matter what dimensions are right it has certain gate length, it has certain source range injunction there and so on and so forth. And now you want to come up with a new generation of technology and that new generation of technology will have at same MOSFET, except that it is a miniaturized version of this transistor a smaller transistor. So, this is what we call a technology scaling starting from a bigger transistor to a smaller transistor and we define a scaling factor called K which is greater than 1 ok.

Now, this constant electric field scaling theory provided the 3 fundamental guidelines right which is what we call primary scaling factors or primary scaling guidelines. It says that given a transistor that you want to scale in future you scale all linear dimensions all linear dimensions no matter which area of the transistors that you are looking at by a factor $1/K$. In other words if your transistor as a channel length l , here this transistor will have a channel length which is l divided by K that is what we mean by you know scaling it by $1/K$. Since K is greater than unity the length has come down right in a new transistor by that scaling factor K right which turns out that you know traditionally we have used a scaling factor which is about 1.4 from one technology generation to other technology generation.

In other words $1/K$ is approximately point seven now it also becomes clear in a minute why we use them, you see oxide thickness essentially this is your transistor and this is your gate insulator right that is oxide thickness T_{ox} that should also come down here by the same factor, length we already talked about and the transistor has a width right in other direction right so the width should also scale down by the same amount right so that is the another important parameter to think about.

(Refer Slide Time: 19:22)

Constant Electric Field Scaling

Technology scaling
Scaling factor $K > 1$

Primary scaling factors:

T_{ox}, L, W, X_j (all linear dimensions)	$1/K$
N_A, N_D (doping concentration)	K
Vdd (supply voltage)	$1/K$

Derived scaling behavior of transistor:

Electric field	1
I_{ds}	$1/K$
Capacitance	$1/K$

Derived scaling behavior of circuit:

Delay (CV/I)	$1/K$
Power (VI)	$1/K^2$
Power-delay product	$1/K^3$
Circuit density ($\propto 1/A$)	K^2

You see here.

(Refer Slide Time: 19:40)

Gate
 T_{ox} SiO_2
 n^+ (W) n^+ L
p-sub (N_A) X_j

① Linear dimensions $1/K$ T_{ox}, L, W
② Supply voltage $1/K$ $E = V$
③ Doping increases K

So, let us consider this transistor here right let us say we restrict our discussion for the time being to n channel transistor right, so this is Si O 2 and this is your gate electrode, which typically is polycrystalline and silicon although it has change now we are talking of metal gate transistors today.

And this transistor you know we will extend in this direction right because this is what we call a width of the transistor correct, so whereas this is the length of your transistor.

So, the length will scale down by that factor and this is what we already said is T_{ox} right oxide thickness will also scale down. So, T_{ox} goes as T_{ox} over K and the same thing happens to length width and also the junction depth this is what we call x_j , which is your junction depth. So, this is junction depth will also scale down in a new transistors, so all linear dimensions we will scale down by that factor. So, that is the first point of the scaling theory.

The second point of the scaling theory is that the supply voltage also scales down by the same factor just as your linear dimensions all your linear dimensions have gone down right as 1 over k ; the supply voltage will also go down 1 over k . Now you understand why is it called constant electric fields scaling theory right because if your voltages are going down by a same factor distances are going down by the same factor right as a result of that your electric field which is essentially voltage divided by any distance is invariant in your original transistor and a new transistor right so that is the idea behind constant electric field scaling theory.

Now, there is a third scaling guideline which is very interesting because in order to satisfy constant electric field scaling theory it appears that these 2 guidelines are more than sufficient right. But the third guidelines are your doping concentrations should increase by a factor K that is all linear dimensions and supply voltages coming down. However, your doping should go up in the new transistor.

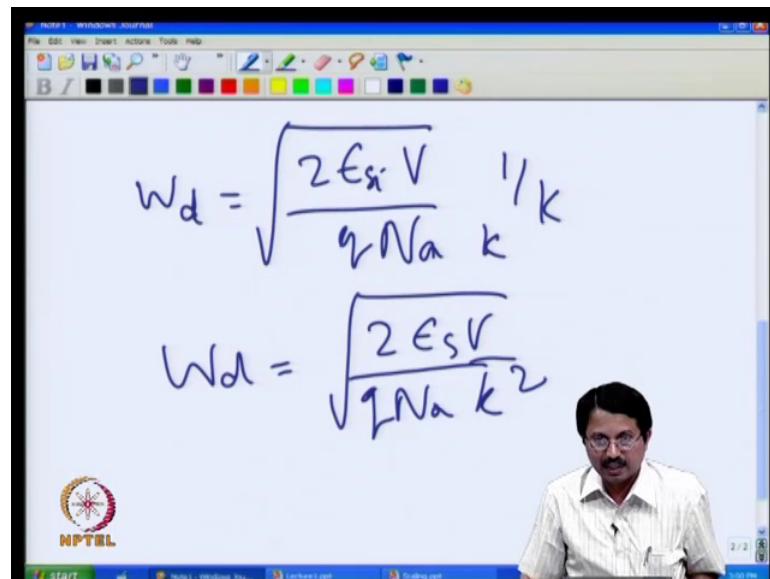
Now for example, what doping are we talking about this is anyway n plus meaning it is degenerately doped very close to the atomic density of silicon. So, there is not much we can do there we are really talking about what is called a substrate doping this is p type substrate, in which we have made a n channel transistor. The doping concentration here which is designated as n_a which is acceptor impurities that we have in the transistor that should go up in the new transistor as K times N_a ok.

Now, if you think about it this really comes about because of the fact that this linear dimension you see all linear dimension but for one linear dimensions are defined by as using photolithography process like the length of the transistor you print the length of the transistor, width of the transistor you are printing the width of the transistor and so on and so forth right.

The junction there you control by how long you do the diffusion right, but there is one very important parameter which is what we call a depletion width right there is a p n junction here, p n plus p junction which is what we call a one sided junction right which is n plus and this is likely doped p region and there is always going to be certain depletion width here you see this depletion width, let us denote as W_d suffix d depletion width that is a width of the depletion region.

You see this is also a linear dimension right if the scaling guideline the first guideline is to scale all linear dimension we should also enable the scaling of the depletion width right. So, how do you enable the scaling of the depletion width?

(Refer Slide Time: 23:54)

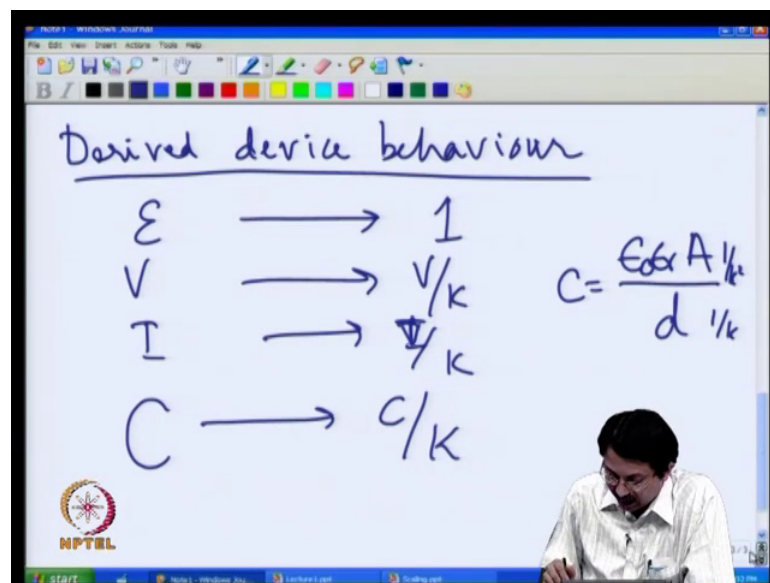


You see for the one sided depletion junction, the depletion width is essentially given by this expression right, this is the expression for depletion width for one sided junction epsilon silicon encompasses the permittivity as well as the free space permittivity, so it is epsilon r times epsilon naught right that is the permittivity of silicon and V is the voltage across that depletion region and N a is the doping concentration. When we talk of one sided junction we are talking of doping concentration in the lightly doped junction region of the junction ok.

Now, you see this V is scaling as 1 over K. Remember that the scaling guideline as already told us to scale V as 1 over K. Now if I scaled this N a as K times N a right you see then what happens in the new transistor is that you get K square under root here right

$q Na K$ square because, you are increasing $N a$ and you are decreasing V by a same factor which is K . In other words your new transistor will now have a depletion width which is K times smaller than the previous transistor right and hence you have been able to scale or linear dimensions consistently as per the requirement of a constant electric field scaling theory right. So, this is essentially the basis for the constant electric field scaling theory.

(Refer Slide Time: 25:20)

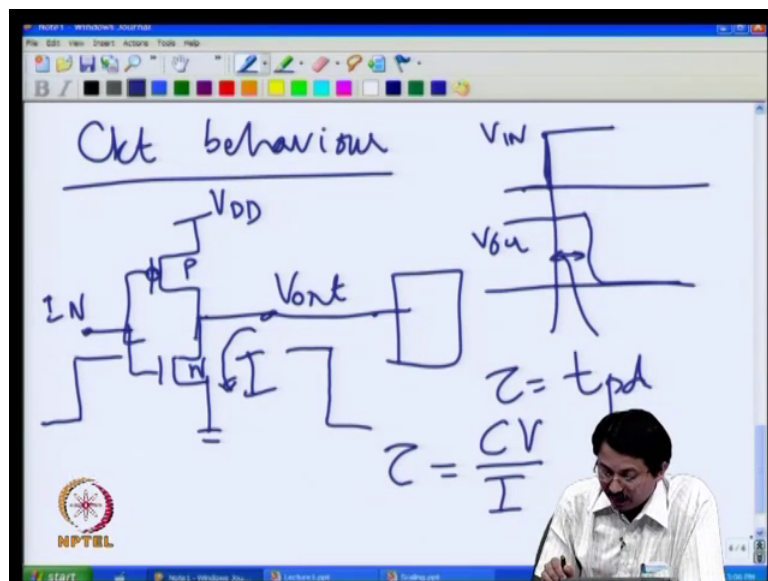


Now, something very interesting comes out if you follow these 3 primary guidelines right. Now let us first do what is called Derived Device behavior, if you follow that guideline what happens to the device behavior we are now looking at 4 important device behaviors here right, one is electric field we already said electric field is invariant, it does not scale it remains same right previous electric field multiplied by 1 gives the new electric field correct. Voltage is important matrix for a device we know voltage is scaling as V over K because of that your current will also scale over as sorry I over K .

Now, one other very important parameter for device is the capacitance. In fact, that is the most important parameter when we talk of CMOS circuits. What is capacitance? You see capacitance is again epsilon naught epsilon r times A divided by distance between you know whether it is depletion capacitance or a parallel plate capacitance it does not matter.

Now let us think about what happens to the device capacitance what is it is scaling behavior right A is area which has 2 linear dimensions in the numerator, and d is also one linear dimension in the denominator, in another words A goes down as 1 over K square and d goes down as 1 over K effectively your C will go down as 1 over K . So, your C_e goes down as C over K . In a new device capacitance decreased you see capacitance is like inertia for these devices if your inertia goes down these transistors will start switching faster and faster you know that is the key we will find out that in a minute in what is called derived circuit behavior.

(Refer Slide Time: 27:20)



Now, with this background let us look at the circuit behavior right if I follow this what happens to your circuit when we talk of CMOS circuit we should envision the CMOS circuit as capacitive circuits you see for example, let me show you a very simple the simplest CMOS circuit which is an inverter which has a p channel transistor and an n channel transistor connected in series this is V_{DD} this is your P MOS and this is your n channel transistor and this is your output and this is your input you see this is an inverter right that is when your input is high output is low and you know vice versa ok

But the figure of merit for us is not the steady state how long does it take to switch from one state to the other state that is your switching speed that intern will determine your circuit speed right any complex circuit whether it is microprocessor or memory is all you know large number of such smaller circuits aggregated together you see. So, the

switching speed is very important now when we talk of switching speed remember this circuit is going to drive a next stage of a CMOS circuit right does not matter what that is it could be a similar inverter or it could be a NAND gate or you know so on and so forth.

In other words when we look at the input of any CMOS circuit we need we will essentially see a gate capacitance right we see a capacitive input right because it is going to the gate insulator here $f \cdot v \cdot t$ right. So you know you have capacitance here this capacitance really has to be switch between 0 and supply voltage depending on when I am switching it from 0 to V_{DD} or V_{DD} to 0, in other words if my input goes from 0 to V_{DD} , then my output should go from V_{DD} to 0, this is what I want to be able to do in CMOS circuit right.

But you know there is going to be if I were to do it in a time axis, if this is my input and this is my output if this input goes up at time t equal to 0, you know my output will have certain latency right it will not in switch instantaneously right and this is a very crucial parameter which is typically called propagation delay, we say that in CMOS circuit the propagation delay which we also sometimes write as τ , τ is given by a metric called $C \cdot V$ over I . Whereas c is the capacitance that you are switching at any node in a CMOS circuit, V is the voltage that capacitance needs to be switched you know as I said between 0 to V_{DD} or V_{DD} to you know ground as I have mentioned already.

And then you know I is the current that is available for you to charge and discharges capacitor. In other words for this to go from output to go from high to low right output was high you see output was high how will it go to 0, it will go to 0 only by discharging that node through this transistor which is sitting down otherwise you cannot bring it down right.

And that happens because you have made this input high earlier when it was low p channel was all and hence your output was pulled up to V_{DD} right, now I have switched of the p channel transistor but capacitance still has that charge you see that needs to be discharged unless you do that you will not go to 0. So, this transistor will draw out that charge. And hence the current in this transistor decides; what is the time it takes to discharge that right, larger the current you have quicker is it to discharge and enhance the discharge time is inversely proportional to current, larger the capacitance

more is a charge to r. And hence it takes longer. Similarly larger the voltage more is the charge to over it takes longer to discharge the capacitor.

So, you know this is in fact you know q times I has a dimension of time right you know it is dimensionally consistent also. Now this is what is the key right.

(Refer Slide Time: 31:44)

$$Z = \frac{CV}{I} \rightarrow \frac{C/K \cdot V/K}{I/K} \rightarrow \frac{Z}{K}$$

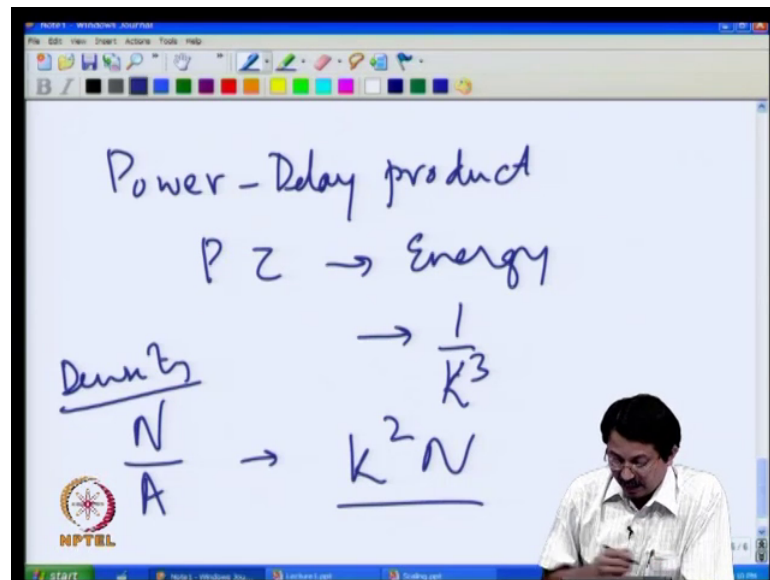
$$P = VI \rightarrow \frac{V}{K} \cdot \frac{I}{K} \rightarrow \frac{P}{K^2}$$

Now given this back ground let us now see tau is C V over I this is my dealing I know that C is scaling as C over K, V is scaling over V over K and I is scaling over I over K. So, what does it mean your delay goes down as K? This is amazing. What it tells you is that you do not even have to redesign any circuit you take the circuit today it could be as I said a simple inverter or a more complex micro processor which is on a 90 nanometer technology without even redesigning you scale it down to 65 nanometer technology immediately your circuit starts operating faster, ok.

Of course when you go to a new generation of technology you do lot of circuit engineering, redesign and also lot of system level architecting and hence your speed benefit is much more than what technology it gives but this is one of the very important reason why you want to scale the technology circuit starts operating faster. But that is not the end of the story you see in circuits speed is one important metric and power is another important metric which is V times I. So, what will happen to power V is scaling as V over K, I is scaling as I over K, power is scaling as P over K square right.

This is even better right this circuit which was operating earlier at certain frequency, now operates at much higher frequency but consuming significantly lower amount of power. So, why would you not want to do a scaling of a technology? So that is why there has been a phenomenal push towards scaling technology.

(Refer Slide Time: 33:31)



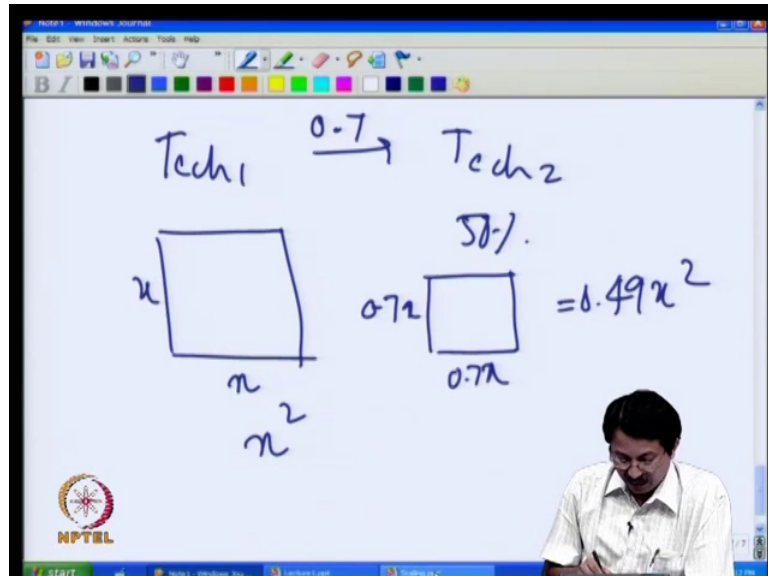
So in fact we sort of capture this whole thing based on what is called a power delay product which is essentially $p \tau$, you know power times delay has you know the dimension of energy you see.

Now, what will happen to power delay product power is going over $1/K^2$, delay is going over $1/K$. In other words power delay product goes as $1/K^3$ right K is greater than 1. So, what it tells you is the following right energy consumed to perform any given operation is going down as $1/K^3$ you just by scaling the device dimension. So, this is you know very good right this is great actually and that is why we have been scaling the technology.

And of course what happens to the circuit density right what is circuit density, the circuit density is essentially your number of transistor that you can pack per unit area correct this is your circuit density right density of the circuit. So, what happens to that remember A goes as $1/K^2$ and that K^2 comes to the numerator right so your circuit density goes up as K^2 . In other words you are able to pack more number of

devices in a given area compared to what you are doing earlier right. So, this is what is very important to understand.

(Refer Slide Time: 35:11)



And I also mentioned some time some ago that typically when we are scaling the technology right technology 1 to the technology new technology right we have used a scaling factor which is 0.7 which is $1/K$, K is $1/0.7$ I mention that you know a while ago right. So, you may wonder why do we use this factor 0.7 right, the ideas of using this factor 0.7 is the following right you know eventually you are building a chip right you are going to scale the dimensions by a factor 0.7.

In other words this dimension length and width both will come down as 0.7, in other words the area here is 50 percent lower than the area here right because this is 0.7, if this is x and this is x if it is squared this is $0.7x$, and this is $0.7x$, so what you got here in terms of area is $0.49x^2$ as oppose to x^2 here right.

So, that is a very good metric for scaling you have a chip you have reduce that area by 50 percent right if you want to reduce it very significantly you may ask the question why not 0.3 you know 0.3 is like a with the jumping a big step you see we are at you know 100 nanometer technology and we want to directly go to 30 nanometer technology right you know skipping all the intermediary steps you know that is very daunting task you know it is almost impossible to do that if you want to that we may want to wait for 10 years right it does not make sense in this first phase of the technology right.

On the other hand why not just 0.95 because that gives a very incremental improvement right you have not really scale the technology is very significantly. If you take x and scale it by $0.95x$ right, whereas you know $0.7x$ historically you know we have must along that path that has been a fairly good scaling number that we have discovered right. So, and this is what we do right so hence the constant electric fields scaling theory gives you all these benefits and that is why we would like to scale the technology.

So, you know if were to go by that and as I mentioned already you know your delay goes down, your power goes down, your power delay product goes down, your circuit density goes up and so on and so forth right. So, this is a very important consideration but it turns out if we look at the scaling we have not necessarily done constant electric field scaling right for various reasons right. Let me illustrate that to you through this typical scaling scenario right.

(Refer Slide Time: 37:54)

Non Scaling Factors ↑ $\frac{1}{L^2}$

Bandgap of Silicon $E_g = 1.12\text{eV}$


Thermal voltage $\frac{kT}{q}$ -40°C - 125°C

Mobility degradation ↳ Subthreshold slope
Increasing doping and electric field

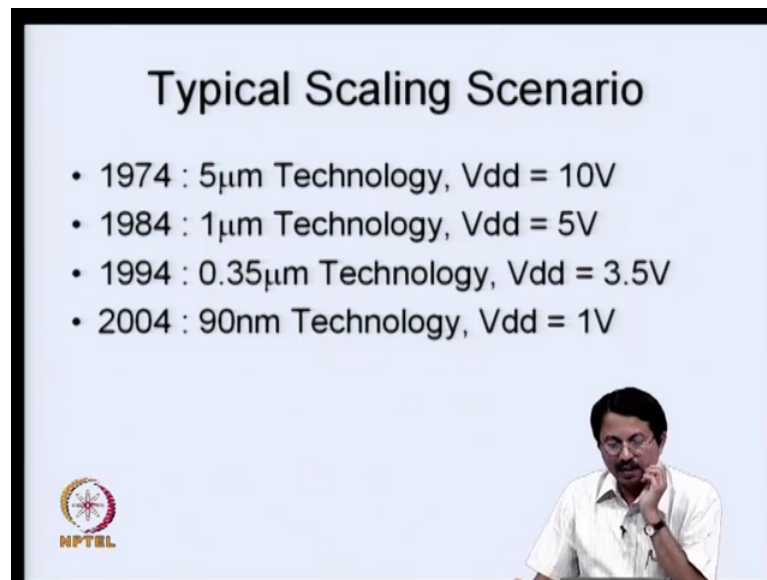
Velocity saturation $\frac{V}{k}$ $\frac{I}{C}$

Parasitic s/d resistance $I_{DS} = \mu_n C_{ox} \frac{W}{L} \frac{(V_{GS} - V_T)^2}{2}$

Process tolerance $-\frac{k}{k^2}$ $\frac{C_{ox} \epsilon_0}{10x}$



(Refer Slide Time: 37:55)



Typical Scaling Scenario

- 1974 : 5 μ m Technology, Vdd = 10V
- 1984 : 1 μ m Technology, Vdd = 5V
- 1994 : 0.35 μ m Technology, Vdd = 3.5V
- 2004 : 90nm Technology, Vdd = 1V

HPTEL

As I mentioned this paper was published way back in 1974. During that period we had 5 micron technology which was operating at a supply voltage of 10 volt and you know a decade later in 1984, we had something like 1 micron technology which operated at a supply voltage of 5 volt. You can obviously see that we scale the dimension by 5 x, but the voltage is we are not scaled by 5 x the voltages is we are scale by only 2 x right. So, then of course from 1 micrometer we came down to 0.35 whereas, the supply voltage came down from 5 volt to 3.5 volt again you know it is not the same scaling factor.

However, here it is very close from 0.35 micro meter which is 350 nanometer we came down to 90 nanometer and the supply voltage came down from 3.5 volt to 1 volt right I mean it is very close to a 0.9 volt right at least here we are very close to constant electric field scaling theory right whereas, here in these scenarios we have actually let the electric field go up in the device right. And now you know we are the first of all why did it happen right.

There was always a resistance to scale voltage because you see eventually you are going to use this chips and build systems right a system designer would have already design the system at working at 10 volt right and 2 years down the line if you come back to the system designer and tell the designer that I have a better chip but you will have to redesign your entire system to operate at 8 volt you know there will be lot of resistance from system designer right.

Now, that is by historically the voltage scaling has been very slow, there has always been a resistance to scale voltage as far as possible keep the voltage you know only if it comes to such a bad condition that the device will breakdown it will not operate, so first of all why do you scale voltages.


Because if you do not scale voltages your electric fields in the device will be so large that you will have the device breakdown taking place right of course you cannot let the voltages be stay at the same right at the same value all along. Voltages have scaled as long as they satisfy the reliability consideration they have not scaled beyond that right they have just take the voltage that is adequate right, accordingly 3.5, 1 volt and so on and so forth right.

(Refer Slide Time: 40:33)

Generalized Scaling

Technology scaling
Scaling factor $K > 1$
 $1 < \alpha < K$

<u>Primary scaling factors:</u>	
Tox, L, W, Xj (all linear dimensions)	$1/K$
Na, Nd (doping concentration)	αK
Vdd (supply voltage)	α/K
<u>Derived scaling behavior of transistor:</u>	
Electric field	α
Ids	α^2/K
Capacitance	$1/K$
<u>Derived scaling behavior of circuit:</u>	
Delay (CV/I)	$1/\alpha K$
Power (VI)	α^3/K^2
Power-delay product	α^2/K^3
Circuit density ($\propto 1/A$)	K^2



And that is why we define a few other scaling scenarios that are called Constant Voltage Scaling, in which case this is an another extreme right you do not scale the voltage from the current generation of technology to the new generation of the technology right which means all linear dimensions are scaled voltage is not scaled. If you do not scale the voltage you want to scale all linear dimension you are doping concentration has to go up by K square correct this is done whatever we worked out earlier right. If you do that when we are divides the electric field will go up by this factor K, drain current will go up capacitance of course will come down because our linear dimensions have been scaled and capacitance is only a function of linear dimension, ok.

Your delay of course goes down much faster 1 over K square as opposed to constant electric field scaling theory where, the delay was decreasing only as 1 over K , your circuit is little more faster you know that is also another good thing if your circuit becomes fast and withstands the reliability constrained then you are you do not necessarily you have to scale the voltage right.

However, the flip side is that your power you know will go up your power delay product will not scale as efficiently as it used to earlier right so these are other issues right this is circuit density of course will improve no matter what right because your scaling the linear dimension and circuit density will go up ok.

So but as I mentioned we have neither done constant electric field scaling theory nor done constant voltage scaling, but instance what we have actually done is something called generalized scaling here what we say is that we introduce one more parameter called alpha we have a linear dimension scaling factor which is K , all linear dimensions are scaled by the same factor K , your supplied voltage is scaled as alpha divided by K . Now when alpha is equal to K , you know it is a constant voltage scaling theory right otherwise you know it is a constant you know for different values of alpha between the extreme you go between electric field and voltage constant voltage scaling and in between we call it is a generalized scaling right voltage is as scaled but not as aggressively as the scaling in linear dimensions ok.

So, again you can go through the math that we did earlier you see that the electric fields are going up by a factor alpha, but not as much as constant voltage scaling theory in which case it would have gone up by K where here alpha is less than K idea is goes scales as this capacitance will scale as this and your delay will go down as this as I mentioned already right when alpha is equal to 1 , it is a constant electric field scaling when alpha is equal to K it is constant voltage scaling theory right. So, you know that is what we have done we have really done what is called more generalized scaling.

Now, this is all fine when we are taking of what is called idealized scaling right. But you see there are certain parameters that we call are non scaling. Let us start with band gap of silicon right the band gap of silicon is 1.12 electron volt, whether you know your transistor as a gate length of 1 micron or 0.5 micron your band gap of silicon does not change.

And you know that is a very important parameter right you know why is it important it turns out band gap is important, because this will determine what is going to be your bulk potential your dope silicon you introduce certain number of impurities your carrier level in silicon will change and that is what we call a bulk potential and intrinsic silicon as a 0 bulk potential, you convert it to p type or n type and Fermi level will go down if it is p type below mid gap if it is n type it will go above the mid gap right and that is the bulk potential right and that bulk potential also depends on what is your band gap right and your intrinsic carrier concentration depends on band gap right. So, there are a lot of device parameters which are governed by band gap right so they will remain invariant they do not really scale.

And thermal voltage you see this is also another important parameter what is thermal voltage essentially it is kT/q right k is Boltzmann constant, q is the electron charge and T is the absolute temperature. You see if your chip commercial chips most of the commercial chips aspect to operate between minus 40 degrees centigrade to plus 125 degree centigrade right that is your typical operating temperature of your chip right whether you do a chip in a you know 350 nanometer technology or 65 nanometer technology it is going to the same application and hence this operating range is not going to change.

As a result of that the temperature remains same you know the temperature of operation is not changing so kT/q does not change right again this has very significant implication as we will see little later, because this kT/q among other things it determines one very important metric of a device and that is called sub threshold slope, this sub threshold slope we will determine what is the leakage current of a transistor ok. So this is a very important parameter.

In fact, we will see that one of the reason why the on current off current ratio is degrading, as I mentioned earlier is because you are leakage increasing and that is because your sub threshold slope is invariant you are decreasing the threshold voltage. But the sub threshold slope is not changing and hence you have very large leakage current right we will we will talk about that later.

Mobility degradation we made very simplistic assumptions earlier we said that voltage scales as $1/V$ over K current will scale as V over K . But in reality what is the current that

we are talking about the current that we are talking about is the MOSFET current right you know a MOSFET current. For example, if it is in the saturation region then it is essentially given by you know expression which would look something like this right where this is what we call the mobility ok.

You know if you were to see this you know what we mentioned awhile ago right this mobility we have assumed is not going to change, the mobility will remain constant. First of all now let us verify, if mobility were to be constant whether this is indeed the case right remember this C_{ox} here is a per unit area capacitors in other words this C_{ox} here is ϵ_{ox} , ϵ_{ox} naught divided by T_{ox} that is the C_{ox} , C_{ox} is not the total capacitance here.

Now, you see T_{ox} scales as 1 over k . So, 1 over K , K will come to the numerator all voltages we have said scale as 1 over K , right there is a square term here voltage square term right so hence you will have a 1 over K square term here in the denominator right. So, in other words your idea scaling rule will be K coming from C_{ox} in the numerator, K square coming in the denominator. Because of V_g minus V_t square term you see and that is how we said V scales over by this dimension and I will also scale as I over K that assumes that mobility is not scaling you see but that is not longer true mobility is also getting affected and why is that because the scaling theory told us to increase the doping concentration.

You see we had this transistor right this is the source and this is the drain and this is my p type region which is doped with certain impurity concentration. When the electrons are travelling in this channel to reach the drain terminal and contribute to your drain current they get started because of the presence of impurities in this channel and that is what is called the impurity scattering.

And impurities scattering has a significant impact on the mobility of a transistor right on in general carrier mobility as I started increasing the doping concentration here as dictated by the scaling guideline I had to do that you see my mobility will degrade in the channel, because these carriers will starts seeing more impurity atoms and as a result of that your mobility is no longer constant right we are implicitly made this assumption that mobility is constant. And hence V scales as V over I will scale as I over K right so these

are the secondary effects that we need to worry about when we actually look at a practical device.

(Refer Slide Time: 50:37)

Non Scaling Factors

Bandgap of Silicon $E_g = 1.12\text{eV}$

Thermal voltage kT/q $-40^\circ\text{C} - 125^\circ\text{C}$

Mobility degradation \rightarrow Subthreshold slope

Increasing doping and electric field

Velocity saturation $\frac{V}{K} \quad \frac{I}{C}$

Parasitic s/d resistance $I_{DS} = \mu_n C_{ox} \frac{W}{L} \frac{(V_{GS} - V_t)^2}{2}$

Process tolerance $\frac{k}{x^2} \quad \frac{C_{ox}}{bx}$

(Refer Slide Time: 50:39)

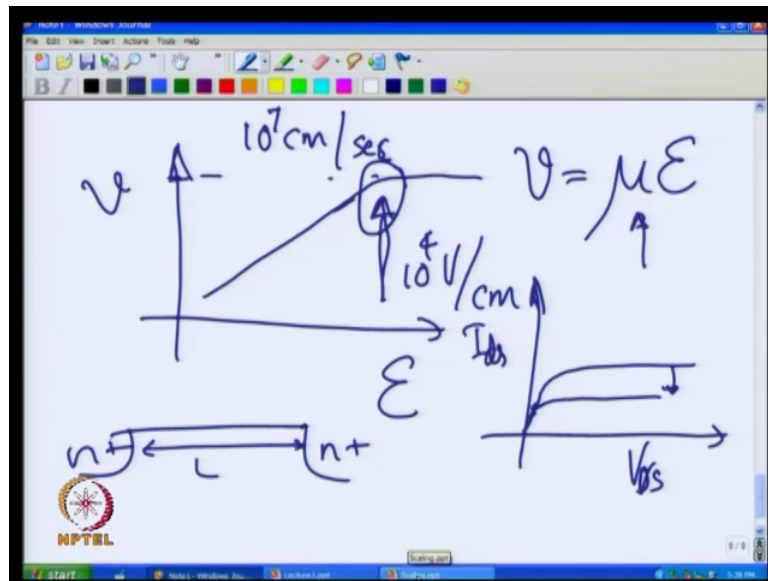
Tech₁ $\xrightarrow{0.7}$ Tech₂

50%

x x x^2

$0.7x$ $0.7x$ $= 0.49x^2$

(Refer Slide Time: 50:42)



There is also another issue of velocity saturation you see and this is also you know something very important issue that we need to address and that essentially ok If you look at electric field versus velocity curve in any semiconductor you know it will look like this remember what is mobility your velocity is your mobility times electric field and this is the proportionality constant mobility right so hence V versus E is expected to be linear. But every material has a what is called the saturation velocity the velocity the maximum velocity is a carriers can attain in that given material.

For example, in silicon the saturation velocity is about 10^7 meter per second and typically this saturation velocity we attain this saturation velocity at electric field such as 10^4 to the power 4 volt per centimeter 10^5 volt per centimeter kind of electric field you start departing from this linear relationship.

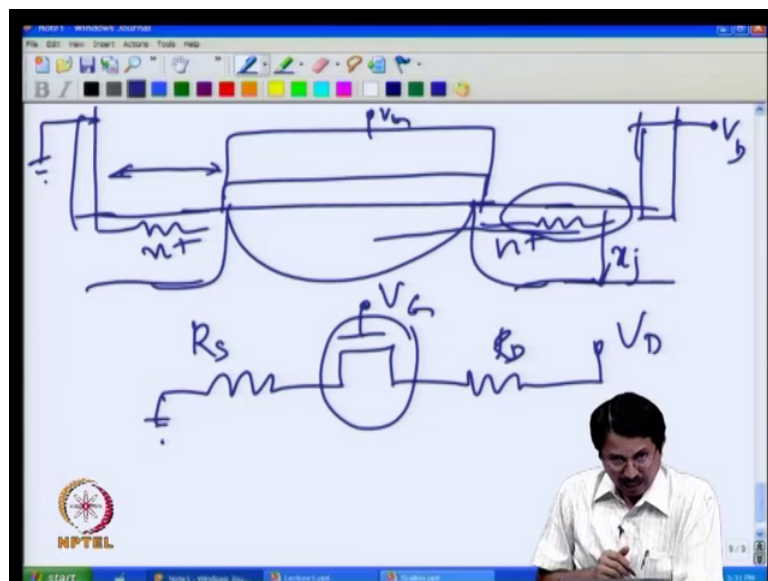
Now, why is this important because remember this transistor that we said $n+$ plus $n+$ and this is my length, the electric field in this lateral direction where the carriers are conducting right and this is the electric field which is accelerating the carriers and they are attaining certain velocity this electric field is increasing as I said because we have not been following constant electric field scaling theory. If we have this electric field approaching this region then you know your current really does not respond to the voltage beyond certain point you are increasing the voltage but you are already in this

regime carries cannot attend any higher velocity right. So, your current will not really increase.

In other words you will have current saturation, even before the pinch of condition typically in a MOSFET, you get a current saturation only when you reach it pinch of condition in other words if you where to look at I_{ds} versus you know V_{ds} right so you have this kind of a behavior right this is the linear region and this is the saturation region right and you transition from a linear region to saturation region when you have a pinch of condition correct.

Now, if you have a velocity saturation which takes place even before pinch off has occurred, you are increasing the voltage current is increasing. But now you have saturated the velocity of the carriers current can no longer increase beyond this point so you have an early saturation of the velocity and hence your current gets impacted. So, this is another important effect that we have to address we have to look at the device and ask the question are we operating the device in velocity saturation region if so, we will have to use different set of equations to describe the device behaviour. So, it will no longer be a quadratic behaviour between drain current to the gate to source voltage right, so that is also another important parameter to consider.

(Refer Slide Time: 54:08)



So, the other thing of course is the parasitic source and drain resistance if a when you see when we talked about the transistor, right this is the source and drain region and this is

your gate by applying gate voltage you are only controlling this channel eventually we will have a metal contact which will be coming in the source and drain region somewhere out here with a proper distance maintained between the gate edge and this contact if you do not maintain this proper edge you will may have a danger of this contact shorting gate and this drain or gate and the source right you need to have an appropriate distance right.

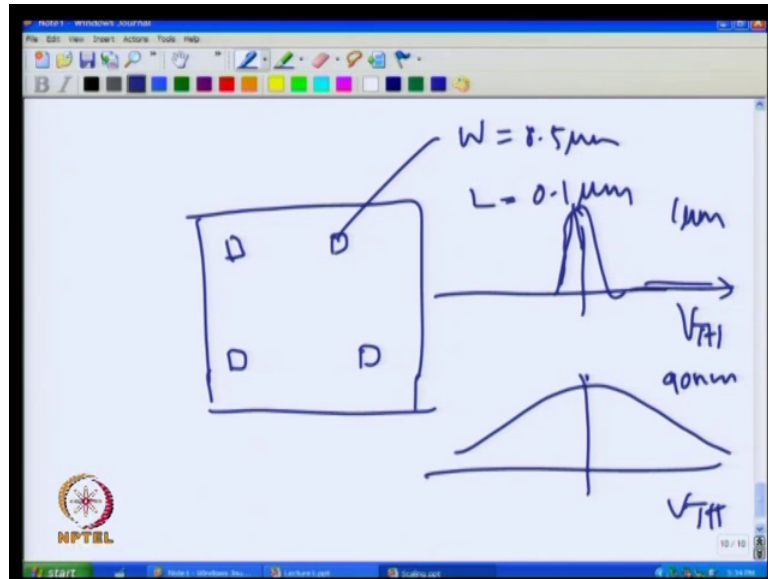
Now, what is means is the following right I am applying a ground potential to the source terminal which means at the metal contact I am applying a drain voltage out here again at the metal contact right because this transistor sitting you remember sitting all the way down in your hierarchy of the inter connects and silicon that we looked at in the last class right. So, the supply is applied out there if it is metal 5, that is where I have applying the drain voltage and the ground potential and that signal will have to come down all the way down to the silicon although in that case fortunately we can ignore all the metal resistances right for all practical purpose. But you see this resistance out here, it is no longer a metal it is highly doped silicon and this is what we call parasitic resistance.

In other words, I have a gate control only in the channel region if you were to light an equivalent model for this transistor then you are equivalent model for your transistor would look something like this, this is your drain voltage, this is your gate voltage, and this is your ideal as transistor. But then you have a parasitic source resistance and a parasitic drain resistance which is which should not be there ideally, because eventually if this parasitic resistance is large then it can impact the entire transistor behaviour.

Now, is it becoming large? Yes right, why is it becoming large you see we have decrease a junction depths, the junction that is coming down and you know the current that is coming out here in a channel has to flow through this region with decreasing junction then that resistance has also increased phenomenally right, so that is also another important consideration. Now and of course we do something very intelligent to overcome all this right now this is what you can understand why do we need new materials you know how do we bring this resistance back to where it should be and so on and so forth right so that is something that you need to keep in mind ok.

Then the last one is what we call a process tolerance you know we will not really discussed quite a bit about this in this course but what it tells you is the following right.

(Refer Slide Time: 57:17)



If you are talking about a circuit right you are making printing, manufacturing large number of transistors, you have this chip as a mentioned there is a transistor sitting here a transistor sitting here and so on and so forth. Let suppose that you have designed this transistors to be identical transistor with W is equal to let us say 0.5 micro meter and L is equal to let us say 100 nanometer or 0.1 micrometer. But when you actually fabricate this transistor it turns out these transistors if you were to measure electrical characteristics they do not come out identical to each other then is what is called a process variation. When you are trying to print here, it may not print exactly 0.5 here, it may print as 0.52, and here it may print as 0.49 and here the length may print as 0.11 and here it may be point you know 0.99 or whatever it is right.

So, there is going to be variation in the processing that we have there is the process tolerating. Now what is happening is that it is becoming harder and harder for us to make tiny transistor exactly identical to each other. In other words, if we were looking at a 1 micron technology and look at any device parameters such a threshold voltage which is a very important parameter for a transistor you will still now a distribution. But their distribution would have looked something like this very tired distribution with very very low standard deviation. On the other and this is what you would see in 1 micron transistor, on the other hand in the 90 nanometer technology if you were to look at a V_{TH} distribution you may see a distribution which would look like this. You have made a million transistor which were supposed to be exactly identical to each other, but they are

no longer identical, they have actually huge spread how do you manage this spread. This is also a very important consideration that one has to keep in mind.

So, this is what we call as some of the parameters which are non-scaling and we will have to deal with these parameters appropriately right. But just to sort of summarize and conclude we required scaling, because scaling will immediately enhance the circuit performance, make the circuit more efficient in terms of speed and consumption of power and so on and so forth.

However, there are a lot of issues that will come about because of scaling, because the world is not ideal. As we looked at some of the non ideal factors we need to be able to design the transistors, so that we overcome all these on idea that it is and still be able to derive all the benefits that were indicated in the beginning in the ideal scaling theory right.

We will look at some of these things in the next class.