Introduction to Semiconductor Devices

Prof. Dr. Naresh Kumar Emani

Indian Institute of Technology - Hyderabad

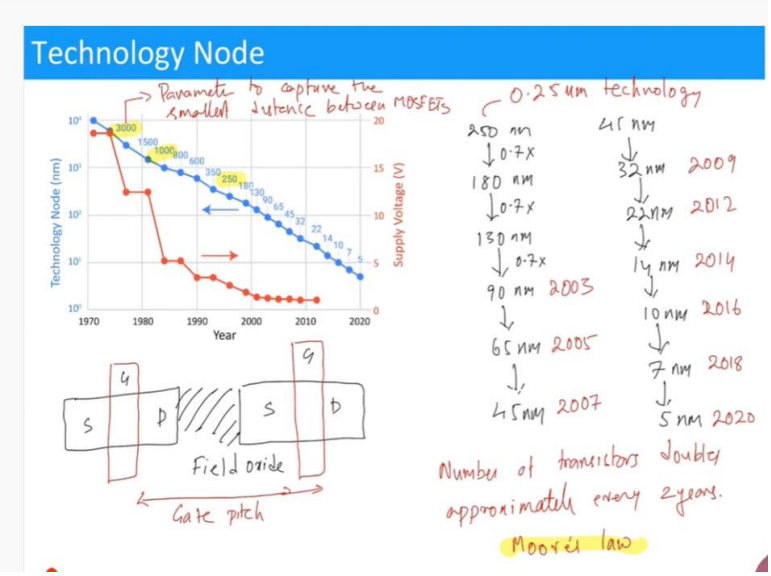
Module No # 13

Lecture No # 64

Limits of Scaling

Hello alright let us get started again in the previous video I introduce the concept of scaling and what we have shown is generally referred to as classical scaling scheme because of various limitations the classical scaling scheme cannot be implemented full. And because of that there is variation such as you know constant field scaling or generalized scaling or so on. So, there are many variation that are used in the industry but the idea of the course is not really go into details of it.

(Refer Slide Time: 00:58)



So, I will quickly tell you what are you know one of limitations of classical scaling and how that you know that comes from physical principles we can analyze that. So, I said that we have; to reduce the lateral and vertical dimension by a factor alpha and also increase the doping concentration and reduce the supply voltages. Well to reduce supply voltages for one of the problems this is the graph that we have seen already in the last video.

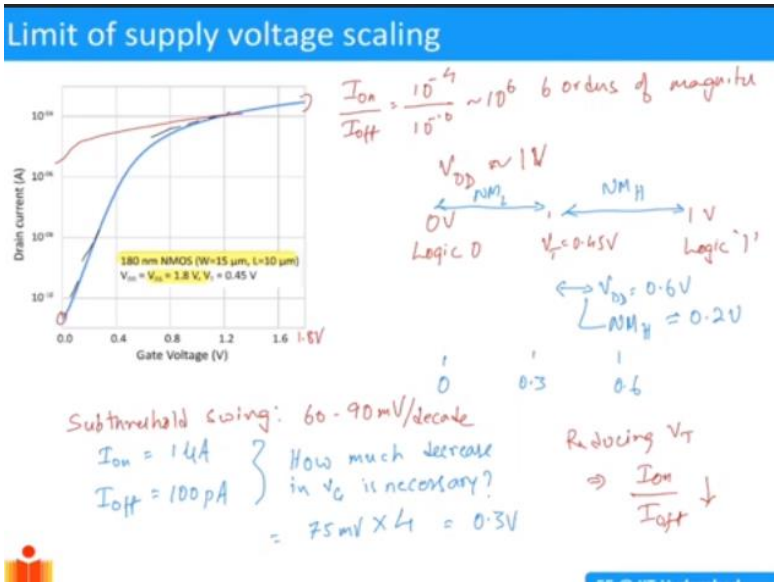
And in that we; have focused on the blue curve which is capturing on the technology node over the last 40, 50 years. But we have also plotted here the supply voltage so what you see is that let us say 1980's it was about 12 volts and then I then it came to 5 volts and then 3.3 volts and then after that it is continuously reducing. The reason why you do not see a gradual change in supply voltage is that our digital chip have to interact with the environment which is surrounding us it is not working in isolation right.

So let us say if our cell phone charger as some chips in it and the convertor will have certain output voltage let us say 3.3 volts it has to work at that particular voltage. Because if you change the voltages in every node then all the peripheral circuits that are used which are digital chips also have to change which is not a cost-effective thing to do that is why in eighties and nineties the voltages are not scaled along with the sizes.

But you know they were kept constant for some time and then scaled down so this gave us to what is known as constant field scaling rather voltage scaling. So, for 2 generations or 2 nodes the voltages were constant and then came down so this is sort of only the qualitative to picture and what you will also notice is that over the last 10 years there are not much change in voltage. Because that voltage was stuck at around from 1 to 1.1 volts or so; and I am not really given you the voltages in the latest generation because this data is not that easy to pass.

And from various sources I have put together these numbers so these are qualitatively the picture what happened over the last 40 years. So why this sort of trend on the voltage we can understand easily the argument that peripheral circuits will have to you know the digital chip have to interface with the peripheral circuits so the voltage as to remain constant that was fine. But the recent years the reasons why the voltage is not scaling is actually to do with the fundamental you know device.

(Refer Slide Time: 03:42)



So, you might have seen this graph already correct wherein we talked about the drain current and gate voltage, and we have plotted the drain current in the log scale and the gate voltage on the x

axis. One of the features of the digital circuits when we introduce the various device metrics, we said that I on to I off ratio in quadric metric so if you look at this picture here x axis in 1.8 volts, so this is 180 nanometer CMOS technology.

So, the channel length and width are quite large and the V_{DS} for this graph was 1.8 and V_{DD} is also 1.8 of course. So, you get this curve this is actually; real data that you know you will reproduce an experiment if you fabricate this using 180 nanometer technology. So, in this case the I on is basically here wherein let us say it is about let us say I on to I off here interested in the and I on is basically 10 power -4 let us take it approximately and I off is a gate voltage is 0 let us take it as 10 power -10.

$$\frac{I_{ON}}{I_{OFF}} = \frac{10^{-4}}{10^{-10}}$$

So, this ratio is about 6 orders of magnitude difference 6 orders of magnitude and we said when we are introducing the device metrics that digital circuits would you know prefer to have as I on to I off ratio. That means when the device is on, we want large current to flow and the device is off we want as small current as possible to flow. So that this ratio is very high then it is efficient digital circuit otherwise if the off current is large let us say off current 100 micro amps or you know 10 micro amps.

If the off current is somewhere here let us assume instead of this the device were performing like this something like this. This is second device let us just is it a good digital device well it is not because you see that I on to I off is 2 orders, so magnitude and it is no good. Because even when you switch off the MOSFET lot of current is flowing through it that means the batteries are getting discharged or there is lot of power dissipation which we do not like.

So we want to I on to I off ratio so now how is that related to the supply voltage. Well supply voltage s generally the largest voltage that you have in MOSFET. So, in this technology is 1.8

volts I mean you can always have some flexibility in adjusting that but not whole lot this is the nominal voltage you can call it. So now in this technology what is the threshold voltage here so you could say you could try to plot let us say you take this extrapolate this let us say you have this, and, in this direction, I will extrapolate, and I will say it is somewhere around this.

This is my V_T remember when I say on above this MOSFET is on and below this MOSFET is off. So this corresponds to logic one and this corresponds to logic 0. And we want you know sufficient margins so that the logic 1 and logic 0 has are not mixed up. Because we are building chips with billions of transistors and finally it is all voltages maybe 1 transistor will have 0.8 voltages. Other; transistor might have 0.85 other one might have 0.675 something like that.

So, when you have various voltages, we want to have good margin so that the device operation is not incorrect. Because if you have a any circuit that computes the values incorrectly then there is no use of having a computer. We need that reliability, so we make sure that there is sufficient margin so because of that let us say for this latest generation technologies. In the last 10 years we really had roughly about 1 volt approximately because some time it might be 1.9 each foundry will have a different number but roughly about 1 volt.

The reason for that is if say the threshold voltage is 0.4 or 0.45 so you can think of the higher side let us say if you are 0 is might ideal logic this is my 0 volt it is my logic 0 and 1 volt is logic 1. So now in between I have my V_T which is let us around 0.45 volts. So, I want what prevents me from reducing my supply voltage. After all I can build up battery with 0.6 volts so what prevents me from having that? The reason is a if reduce this if I let us say I put my supply voltage here 0.6 if I do this we did equal to let us say 0.6 volts.

If I do this, then the margin I have for the high logic hight ,right ,in this previous case it was 1 volt I could say this is my logic this margin or noise margin high NM_H and this is my noise

margin low. So my low side 0.4 noise margin and the high side I have 0.5 or so I have sufficiently large margin. So even if the 1 is not truly one you know instead of it might be 0.8 or 0.7 as well still it will be considered as logic high.

But the moment if I reduce my V_{DD} so let us say 0.6 then the noise margin high become only 0.2 sorry this will become noise margin high will be equal to 0.2 volts. That means instead of V_{DD} 0.6 what if the voltage for some reason you know because of some interference because of the sudden change in currents there is the spike. And then we saw the voltage drop down to 0.3 then it will be interpretated as logic low and we do not like that.

Because that leads to error in the computation right so for good device reliable operation, we need to have high noise much in this. So now you might ask well I mean why cannot I reduce my V_T I cannot completely use let us say I reduce my V_{DD} as the 0.6 and I will reduce my V_T to 0.3. So why I cannot have that let us say 0 0.6 and let us say V_T is 0.3 so I have 300 millivolts of noise margin high and 300 millivolts noise margin low why cannot we have this.

Well, the reason for this you know this is also difficult to achieve it is not that is impossible, but we pay a price and what is the price? The price is something to do is the sub-threshold screen that we talked about. You know we cannot arbitrarily reduce the threshold voltage because remember what we discussed about subthreshold swing. We said for a classical MOSFET you now where we have current flow because you know electrons jumping over barrier this is about 60 to 90 mill volts per decade.

That means if I have let us say 1 micro amp of current okay how much gate voltage should I have to reduce it to 1 nano amp. In a classical MOSFET let us say I have start with I am just taking an example let us say ion is 1 micro amp and my I off I will say I live with or rather you

know I want to have 4 orders of magnitude change in the current so I will leave with I off of 100 picoamps this is what I can have only 4 orders not 6 orders 4 orders I am asking.

If I want 4 orders of magnitude change in the current how much gate voltage should be preferred, you know how much reduction or decrease in V_G is necessary for 4 orders of magnitude right. What you need effectively is let us assume you know 0.75 or 75 millivolts as an average subthreshold slope for so which is realistic. You know generally devices nowadays have 75 to 90 millivolts per decade. So let us take it as 75 millivolts per decades but how many decades 4 decades 4 orders right.

So, I will need at least 0.3 volts if I do not have 0.3 volts I cannot switch off device. So that is why if you want to reduce let us say make V_T of you know 0.2 volts 200 millivolts V_T I can in principle think of V_T of you know 200 millivolts. But if I use that as CMOS circuit my off current is going to be large let us say if I want to plot this in graph let me erase these part of it you know and say if I want to have V_T of let us say 0.2 volts then it might have to fabricate a device which is like this I on to 1 off ratio not actually.

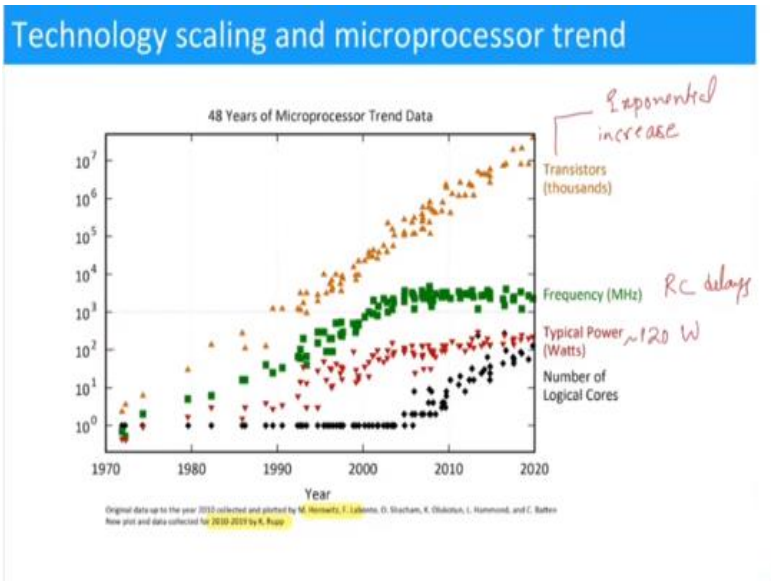
So I have this the current might flow something like this so there is no significant to I on to I off ratio that is why we cannot scale the supply voltages reducing to V_T implies I on to I off reduces which we do not like this is one of the limits. And because of that the threshold the supply voltages have struck in 1 volt so what you might have ask? Well we saw that the power dissipation in a chip it is going to be proportional to CV^2f capacitabce times voltage square times frequency.

Voltage square if I reduce supply voltage it will reduce my power dissipation and still increase my power frequency. So I can try to make my device operate faster but you know it turns out to that it is difficult challenge this is a very multi-dimensional problem that require some of the best

brains in the world to address and that is happen in last 40 years but there are lot of challenges this is what I wanted to tell you about.

So anyway so this is basically the fixed voltage scaling and it turned out that over the years there where many other challenges. And nowadays we change each parameter very different among so we call it generalized scaling and we can analyze what happens in each of these scenario's but it is not necessary for purpose of this course. What I wanted to show you now is?

(Refer Slide Time: 14:41)



Basically who convey the final impact of you understand the basic principle from classical scaling and we understand one of the limitations. Let us even doping density could not scale it there are so many other limitations please forget all the part. We understand at least at intellectual level what is requirement. So with all that this you know this is the result this is to be a very impressive slide. Because this data here shows you how the parameters of a microprocessor have behaved over a last 50 years.

So this was originally connected by you know horowitz others and fine recently in fact taken this data from github from the website of github from the website of karlupp group. So it as

collected this and then he keep updated it every now and then. So this is the data up to 2019 I think so what you are seeing in the y axis is logarithmic scale and the x axis you have the years. The first thing you notice is the number of transistors on a chip on microprocessor.

And it has been increasing exponentially over the last 50 years because it is logarithm so it is going to be a straight line so exponential increase in the number of transistors ,why? Because we said that you now every 2 years also we are doubling the number of transistors and because of that we get this sort of behavior great why large number of transistors? If you have large number of transistors you can implement more logic operations.

Now you might have studied in your basic electronic course how to build a counter or NAND gate or so on right. You might not have studied how to build a MOSFET but finally there are built out of MOSFET's. And let us say if you have a NAND gate you need 4 transistors 2 input NAND gate or in NOR gate might have replied again 4 to MOSFET's and so on. So if you are able to have more transistors you can built more logic circuits.

That means you have computers are more versatile microprocessor more versatile and both powerful. You can have a larger bit processor initial we had 16 bit then 32 bits then now 64 bit and so on right. So we are able to have that we are able to build in more transistors because of that the processing capability is increasing that is one of the features we have noticed. The other graph you see in green is the frequency of the processer.

You see that initially there is a processor increase and I mean it is logical because we saw that scaling is about reduced sizes but 0.7 or you know scale it as 0.7 we saw that the frequency is increasing about 40% each scaled on. So this is what happened over till 2000 or so but after 2000 it sort of saturated in a range of few gigahertz. The reason it happened is because finally frequency is said my how fast we are able to charge discharge capacitance in the circuit.

So it turned out that the RC delays became a problem it was not possible to discharge the capacitors as quickly as we want I really want to do in instantaneously. But if you do that I mean the current limitation of the technology does not permit so we are not able to go beyond this frequency. I mean we can do it in a single transistor level for sure but then once you integration let us say know 1000 transistor it is very difficult to achieve very high sparks high speeds especially with silicon.

There are other alternatives but silicon it is not that easy to increase the frequency more than this few gigahertz. So what is the solution? I mean this is the industry that never stops I mean there is a very famous saying of Gordon Moore that you know no exponential is forever but forever can be delayed. So now what happened was since the frequency as saturating what the industry did was they came up with a clever idea of introduction more number of cores.

So till 2000 till mid 2000 you never have seen a multi core processor but now a days everybody have a multi core processor. You cell phone have probably 8 core or you know 16 core in it right why? The reason is individual core is not able to do not able to scale the frequency of individual cores but I always parallelize it. I can distribute like computing job on you know more cores and I can make the overall system more efficient and that is why over the last 15 years you see the number of logical course are increased in a processor.

Fine and the last thing you know I mean I want you to know is basically the power in our classical scaling scheme. You saw that we are able to reduce the sizes by 30% and maintain the same power because the power dissipation was rather you know I say 50% deduction in the power so you can keep the same power in double the number of transistors. So I mean we wish that was true but it turns out that it practically there are many challenges in achieving it and so what was realistically possible was this.

You see that initially it is good number of you know power was increasing in the number of transistors but then after some time it you know we are not able to increase the capacity more. The reason is if you have more power dissipation you are heating up silicon once you heat up silicon you know all bad things will happen. Mobility can reduce and you know threshold voltage can shift whatever transistor or circuits we have designed so carefully they might not work as you want them.

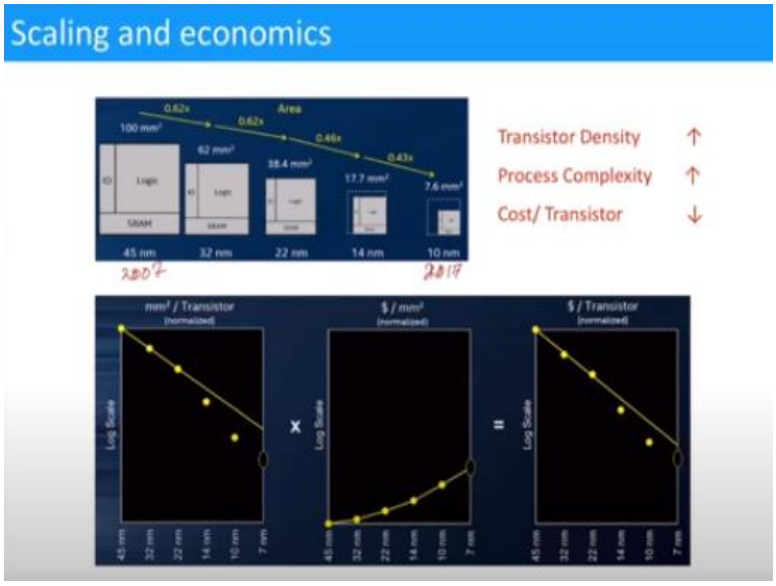
So we do not like them to be heated that is why you know you see them more certain computers are kept in especially servers are kept in air-condition rooms. Because if you do not have good temperature control then they might not work correctly they will have many glitches there life time will reduce. So that is why even a intrinsic even power dissipation CMOS strip is very important parameter is sort of we do not want it to be higher than about 100 watts.

I think in this case you know this is roughly about 100 watts 125, 150 watts maybe. So this is 122, 120 watts let us say approximately average we do not want to be much higher than that because if you want much higher than we have to cool it you know you have to already we have you know servers which have water cooling and things like that. We need much more aggressive cooling techniques to make sure that there are no problems in the technology.

But whatever it is you know overall definitely it is impressive technological achievement perhaps the greatest technological achievement in the last 100 years. The reason is I know many of our Indian kids are very interested in software you can do lot of software only when you have hardware. If you do not have the hardware you cannot to anything so this is the technology that has given you the sort of a hardware to run bottom program we want whatever we talk about artificial intelligence everything right.

There are been for a long time but people could not implement them because you know that computation capacities are limited but now you know we have this no huge systems which can actually enable artificial intelligence or making things like that. All of that are heart of it all semiconductor evolution that has happened over the last 5o years.

(Refer Slide Time: 22:15)



So I just want to give you few glimpses of you know what this is about this is not technical part of it I think this is important for us to understand and appreciate what the technology is all about. This is a slide which I have taken from the presentation by Mark Bhor at Intel in 2017 he gave presentation on continue More's law. And in that it shows how the processor area have reduced so you should this is 45 nanometers that this is about I would say about 2007 this is 10 nanometers would be about 2017 maybe 18, 17 or even 16 maybe.

So about 10 years the area has shrunk each time the area has shrunk like this so initially 100 out of 100 millimeters area of processor now is about 7.6 millimeters. Roughly I mean less than you know more than 10 times shrink which in terms of you know overall area. So the reason is been this has been possible is because of the scaling that we described we are able to increase the transistor density.

But simultaneously it turns out that the processor complexity on this is what I mean by process complexity is increased? The semiconductor chips have to go through a manufacturing process. And it is not easy to fabricate I mean this is done in a very high end fab and you have a thousands of step that has to be executed so that we have the semiconductor chips. So this is known as semiconductors fabrication process and that will be dealt in a technology related course VLSI technology related course.

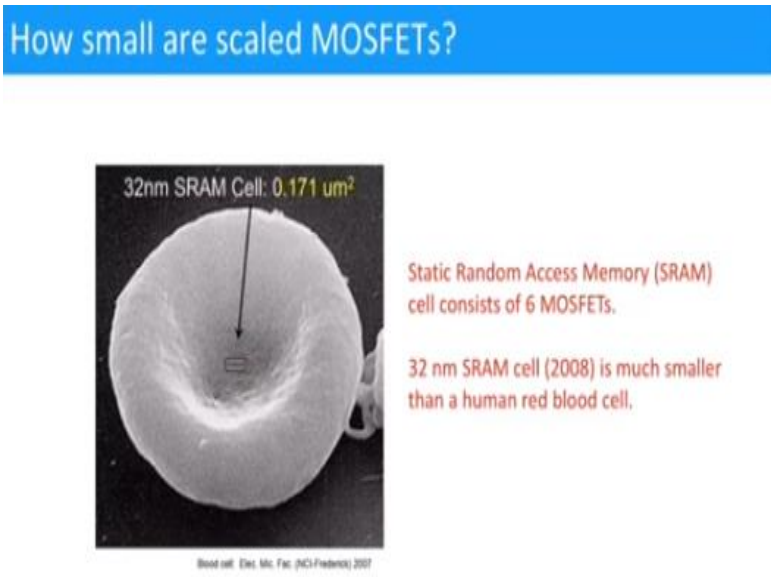
But essentially the message is that the process complexity has increased so that is what captured in the bottom you figures here. So this is evaluation from you last 10 years or so you see that the area per transistor is reduced this is a log scale. So exponentially reduced area per transistor and process cost also have increased exponentially. Because we are always you know if you want to go from let us say 45 to 32 nanometers you have invent lot of tools lot of material processing you know equipment has to be invented and developed.

So that we achieve this right so lot of cost is I mean the state of arts you know semiconductor fab's are you know upward of 5 billion dollars you know 8 billion dollars or so. If you want to establish a state of art it is too expensive. But the good thing is because you know finally the cost is going to be a product of this tool so what you get is? The cost for transistor which; has been exponentially reducing over the last 10 years.

The cost per transistor is reduced and we are also selling the devices you know millions and millions of cellphones and all. So some of the cost is getting of course semiconductors are industries always profitable because there is exponential growth in the demand. So the cost for transistor has reduced because enable us to have more and more powerful machines our phones right.

The super computer of the earlier days is now you know probably much worse than your cellphone ordinary cellphone which is probably 5000, 10000 rupees that is why probably much more powerful than the super computers of 60 years back. So this is how economics of the industry works and we have you know you can always have beautiful technologies but it is no economically reliable it is not going to become useful. So this is it is been economically viable technology that is why you have this revolution.

(Refer Slide Time: 25:37)



And I just wanted to close with a couple of images which I find them you know fascinating it is very easy to talk about the you know nanometers and modes and all that. But finally what is this? How small is the MOSFET the tradition MOSFETS are microns size how small is today's MOSFET. And I am not even showing right now the images of latest MOSFET it is still 10 years back you know in the lecture after next.

I will show you some images of latest MOSFETs but before that this is 32 nanometers technology which is about 10 years back. And in a particular technology the size is always want to drag about all the semiconductors industries drag about the size because you know there is something called as static random access memory this S RAM which is part of every computer.

And to build an S RAM cell you need 6 transistors so if you built 6 transistors together we make 1 S RAM cell and how much is the area of that cell that is a very important thing.

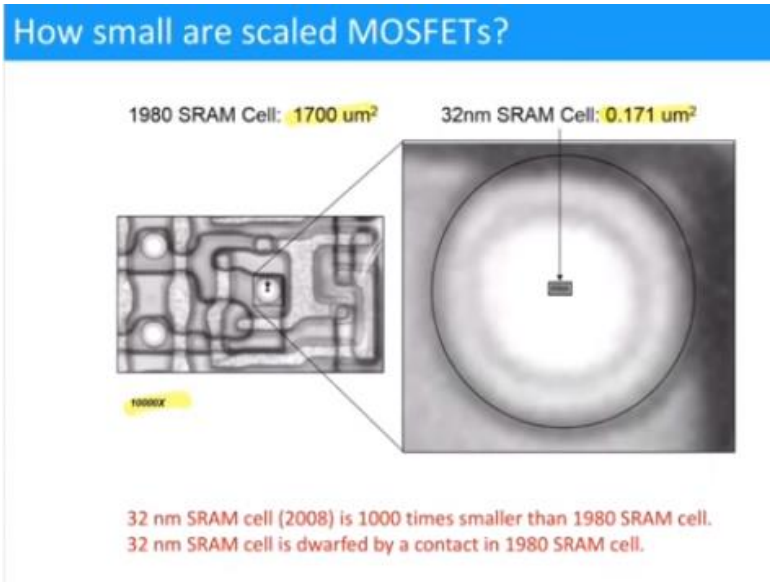
This is simplest circuit and computer always drag about this we even if we look at IEDM presentations of Intel you will see SRAM cell areas. The reason is very comparable you know that tells you how good the technology is? So this is static random access memory of 32 nanometer technology and you see that the area of this is only about 0.17 μm^2. It is not very illustrative when I say 0.17 μm^2 so what?

Well the size of these 0.17 μm^2 is this compared to this white area which is actually the red blood cell of a human body. So when you have a cut you see blood flowing out it has lot of red blood cells and you cannot even see them, You put it under microscope you will see a red blood cell this is going to be a size relative to the S RAM which is maybe 1000 times smaller than that.

And this as 6 transistors if you have a SRAM you know like a body human RBC of your body probably you could built 1000 SRAM cells in this you know 1 kilobyte of memory you can built probably or even 10 kilometers I do not know. You can pack all this stem cells together and built a memory array. So that will be the size of your memory that is why you are able to have you have SRAM's you know DRAM which are you know gigabytes now right DRAM sticks we are having 64 gigabytes and all that.

That is because we are able to pack more and more transistors this is how scale we are talking about. I sort of like this image that captures you know how small how advanced the technology is right now.

(Refer Slide Time: 28:14)



And other image is again I mean this is comparison the same SRAM but compared to the SRAM about 40 years back or 30 years back. In 80 the SRAM size are about 1700 micrometers square. By 2010 it become 0.171 or so in 30 years the size are reduced to 10000 times whatever the today's SRAM cell is much smaller than the contact you know this you know it is probably you will understand what all these things been here.

So these are all some connections and some contacts you see here this bright spots of the contacts so you SRAM cell are probably going to be very small compared to the contact of you know 30 years back that is the extent of technology that is where you know that advanced. So with that I would like to stop my discussion on scaling part of it. I hope you know understand what scaling means what technology is if you open a newspaper some time for you check on you know what Samsung is producing or TSMC is producing or Intel is producing.

We will talk about 5 nanometers technology 3 nanometers technology you research and all that right. So you might understand what they talking about. So I would ask you to just check out those things you know it is interesting this is one of the fore fronts of technology and always it is very exciting to keep up with that and with that I will stop this discussion. In the next lecture I will take I will describe scaled MOSFET's.

So we understand what scaling is what it looks like and all that what is the current characteristic of scaled MOSFET? I will discuss that in the next lecture. And finally you know I will discuss something call as short channel effects maybe the discussion is about half an hour to 45 minutes after that the last lecture in the MOSFET series will be the latest MOSFET's you know alright I will see you in the next videos thank you very much for you interest take care bye.