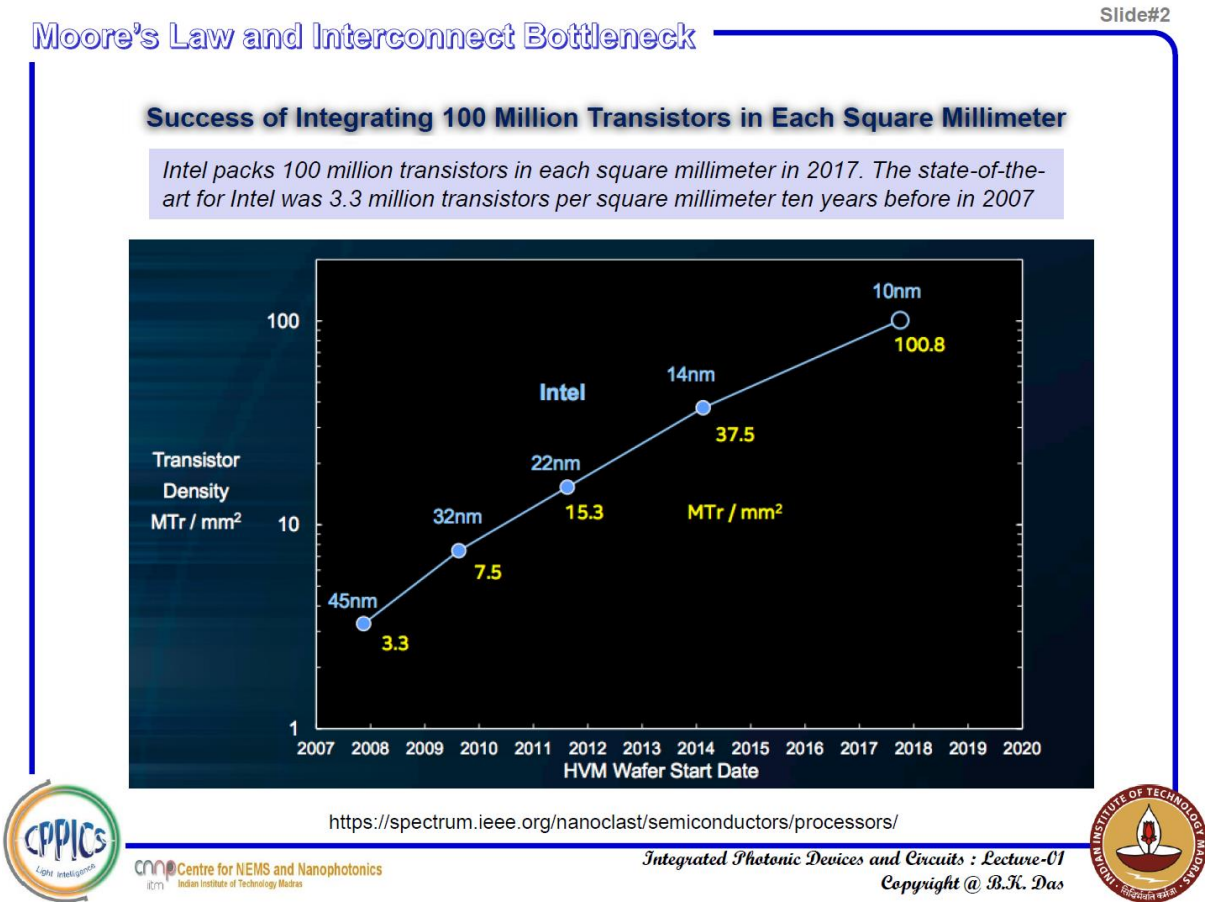
Integrated Photonics Devices and Circuits
Prof. Bijoy Krishna Das
Department of Electrical Engineering
Indian Institute of Technology – Madras

Lecture –02
Moore's Law and Interconnect Bottleneck

Hello everyone, in the last lecture, we have already given an overview of the course structure. So, today I will be discussing Moore's law and interconnect bottleneck.

(Refer Slide Time: 00:34)



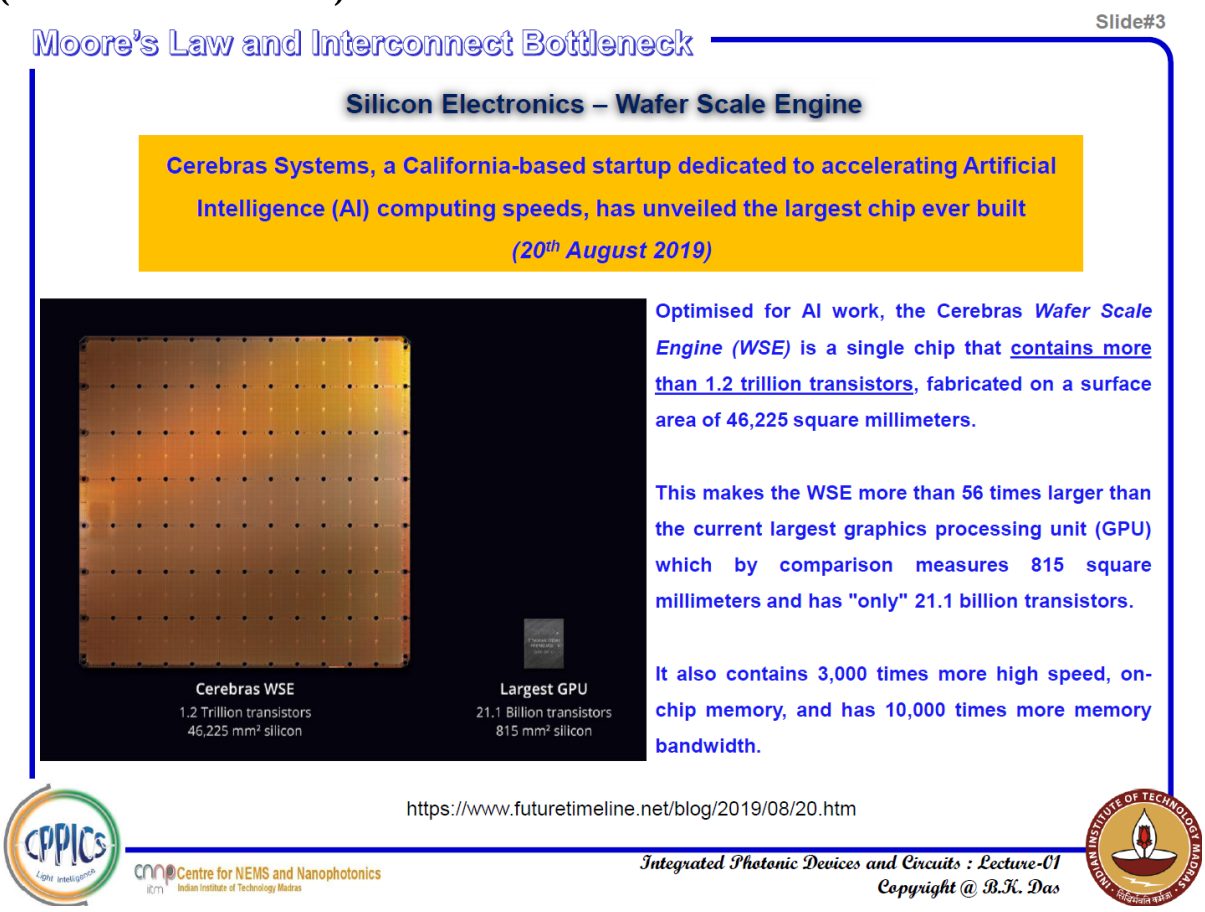
So, what is that? Suppose you see the progress track of integrated electronic circuits. In that case, we find that for the first time in 2017, one of the semiconductor giants, Intel, packed 100 million transistors (MTR) in each square millimetre area in Silicon. In 2007, the state-of-the-art for Intel was 3.3 million transistors per square millimetre.

In 10 years, Intel's transistor density per square millimetre has improved from 3.3 million to 100 million transistors. The above figure shows the transistor density achieved from 2007 to 2018 and the technology node. In 2007, the technology node was 45nm technology, which is

the source to the gate length of the MOSFET transistor. A technology node is defined based on the channel length of the MOSFET transistor to fabricate. So, with 45nm technology used in 2007, Intel could demonstrate 3.3 million transistors per millimetre square. In further years, Intel increased the transistor density to 7.5, 15.3 and 100 million transistors with 32nm, 22nm and 10nm technology nodes, respectively. So, the advancement of technology has enabled the scaling of the transistor size and wafer-level integration of trillions of transistors, also known as wafer-scale engines.

For more information on the picture, you can refer to the link (https://spectrum.ieee.org/nanoclast/semiconductors/processors/) mentioned in the bottom part of the slide.

(Refer Slide Time: 03:46)

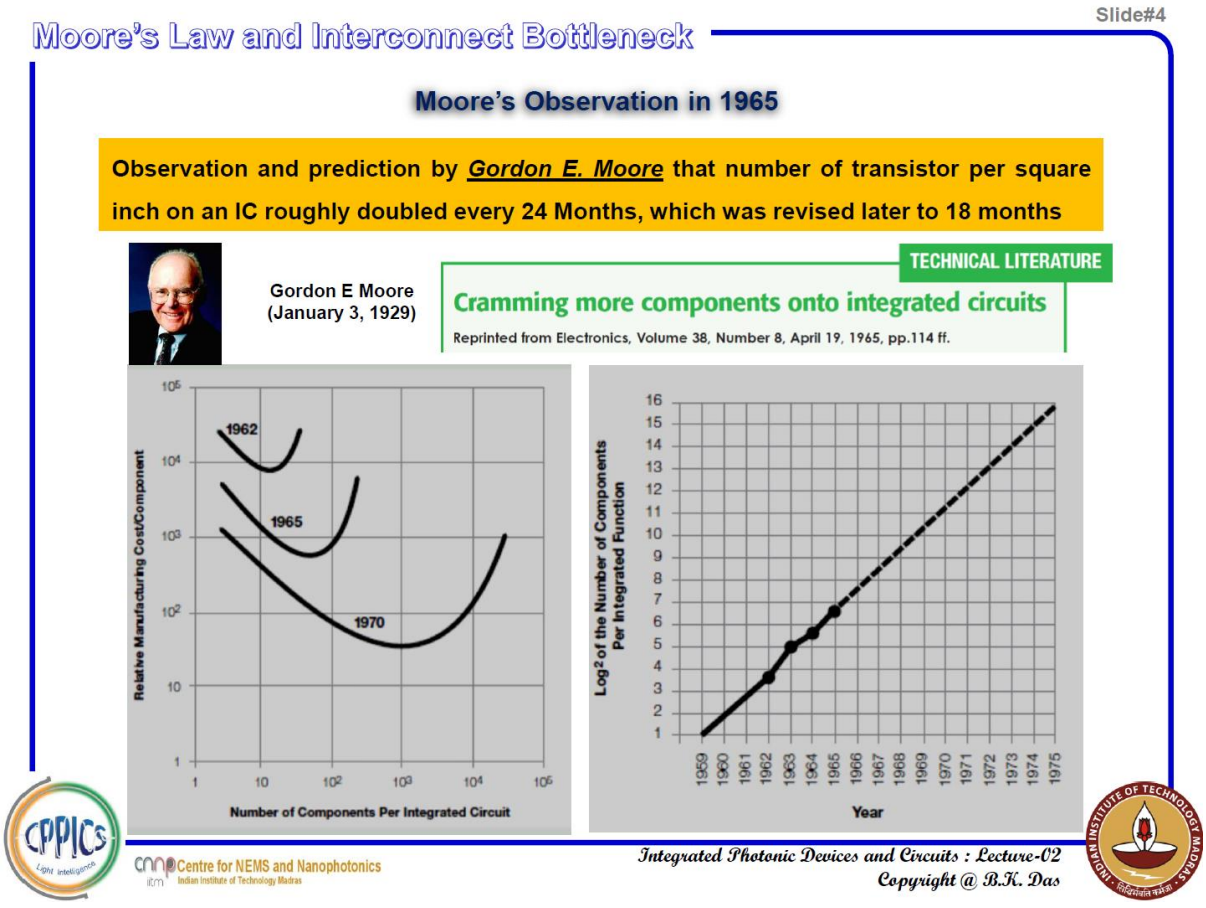


As another example of improved integration density due to the advanced technology node, I want to show the wafer-scale electronics engine demonstrated by California based startup, Cerebras systems. Optimised for artificial intelligence and machine learning work, the Cerebras

WSE is a single chip that contains more than 1.2 trillion transistors fabricated on a surface area of 46,225 square millimetres using 10nm technology.

As mentioned earlier, the advancement in the technology node has enabled the scaling of the transistor size. But how small can you reduce the transistor size? That is one limitation faced by the integrated electronic industry, but we will focus on another restriction in the background for the next few minutes.

(Refer Slide Time: 05:05)



I mentioned that John Bardeen, Walter Brattain and William Shockley, the three stalwarts, first demonstrated the transistor 1947-1948 and discussed the progress of integrated circuits from a single transistor to wafer-scale engines.

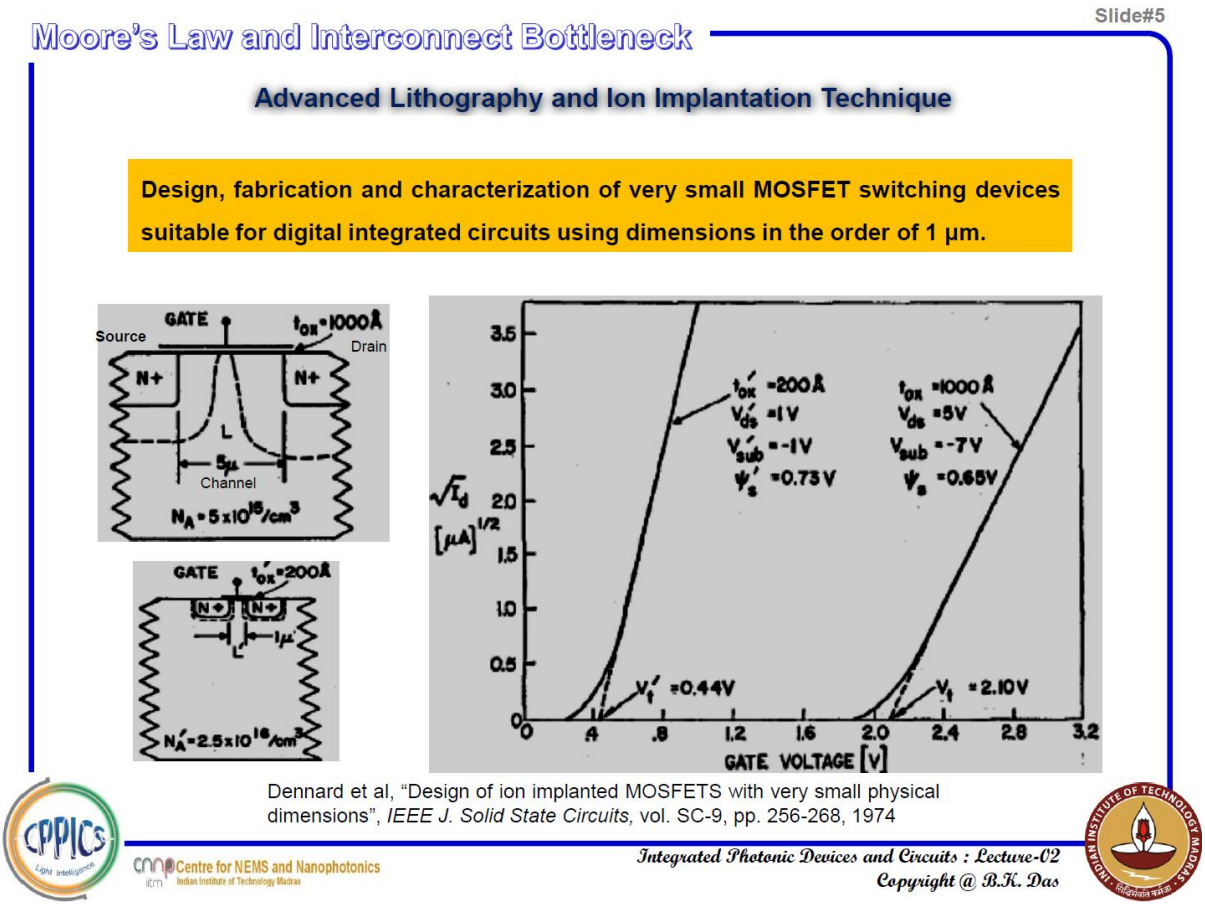
Gordon Moore, while he was working for Intel, he observed a fascinating thing, and he published a paper in 1965 with a title called "Cramming more components onto integrated circuits". What is that? He presented a graph with the x-axis as the number of components per integrated circuit and the y-axis as the relative manufacturing cost per component (left side

diagram in slide four). In 1962, as you increase the number of components per integrated circuit, the cost reduces up to say about 10-12 components because of the manufacturing technology. But if we want to integrate further, the price increases, thus taking the shape of a parabola. Similarly, in 1965, with the improved technology, the economic integration of up to 75-80 components on an integrated circuit was possible. Based on these two results, Gordon Moore predicted the technology advancement for 1970. He stated that in 1970, the economic integration of up to 1000 components on a single integrated chip would be possible. So, that is the plot.

Further, Gordon Moore considered the number of components integrated into the circuit economically (minima in the parabolic curve). He extrapolated the data for future years (right side diagram in slide four). Here Moore plotted the minima of the parabolic curve in log2 scale in the y-axis, and the x-axis is the year which is almost a linear curve. Looking into the technology and cost analysis trend, he commented that after looking into that, he concluded that the number of transistors per square inch on an IC integrated circuit roughly doubled every 24 months. This prediction is well known even today as Moore's law. However, in the future years, the technology advancement was much faster and instead of 24 months, **the community** has revised Moore's law to 18 months. So that means for every 18 months, the number of components you can integrate economically is double, and it is continuing.

But over time, engineers and scientists had to solve a lot of techniques and many problems to continue maintaining Moore's law. Okay, how that happened, I will discuss with you, that will help us understand the integrated photonics or photonic integrated circuits technology, research progress and the investment.

(Refer Slide Time: 10:37)



Here, I will take the help of the paper by Dennard et al. in 1974 in the IEEE Solid State Circuit journal titled "Design of ion-implanted MOSFETs with very small physical dimensions." The keyword in the title is ion-implanted MOSFETs. In this paper, the authors claim that the transistor can be reduced dramatically with the help of ion implantation. Here we will not go through the basics of MOSFETs; I will assume you all know about the MOSFETs working principle and operation.

You know MOS transistor or any transistor you need p-type doping and n-type doping. Before 1974 or so, diffusion doping was the preferred doping process in the CMOS industry. The process required the furnace to operate at high temperatures. The top-left diagram in slide five shows the sketch of an older version of MOSFET with diffusion doping. The MOSFET has a gate region between two n-plus contact regions which are called the source and drain. There is an oxide layer of 1000 Angstrom (100 nm) thickness on top of the gate layer. The channel length of the MOSFET is 5 micrometres. So, before 1974 the channel length was 5 micrometres, and this is because the depletion layer was this depletion typically was very large.

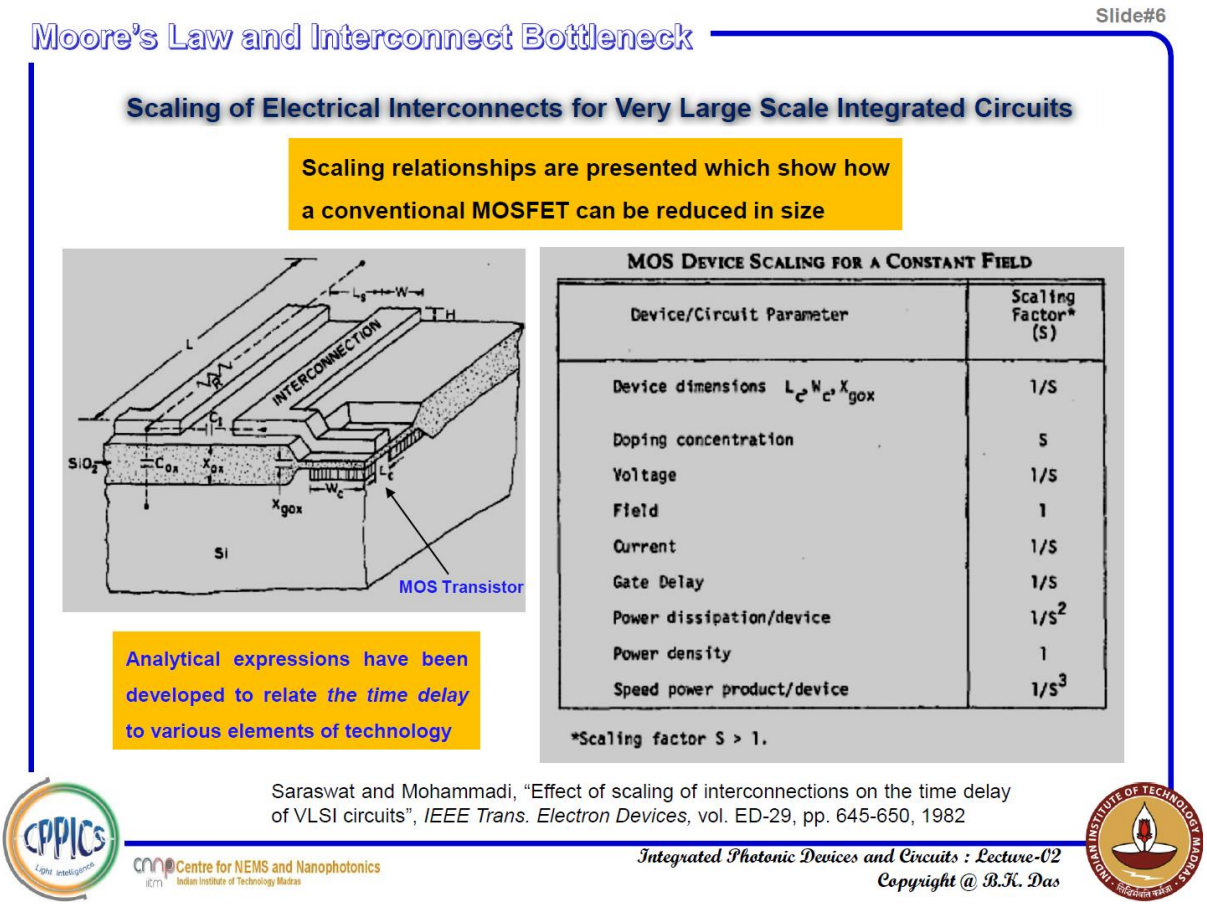
To achieve higher doping with diffusion doping was complicated. Therefore, the depletion region width was more, resulting in the reduction of the effective gate length as shown in the sketch.

But, in this paper, an accelerated ion was implanted directly on the silicon substrate to yield high doping. For example, if you want to do n-type doping, you must use Phosphorus atoms in Silicon. If you wish p-type doping, then you have to do Boron implantation. You can photo-lithographically define a small region, and shallow doping is possible depending on the energy of the implanted atoms.

Therefore, you can achieve a high doping concentration in a smaller window, which is why you can control the width and depth of the source/drain region. In the bottom left diagram of slide 5, you can see the reduction in the size of the n-plus doped source and drain region. The broadening of the depletion region is not possible because of the heavy doping, and the oxide thickness and the channel length could be reduced to 200 Angstrom and 1 micrometre, respectively. Because of the reduction in the device dimensions, you get much better performance from the MOSFET, as shown in the right-side diagram of slide 5.

The graph shows the square root of the measured source to drain current (the y-axis) to the applied gate voltage (the x-axis) for diffusion doped and ion-implanted MOSFET. The channel length was 5 microns for the device with diffusion doping and 1 micron for the ion-implanted device. The threshold voltage, defined as the smallest voltage required to enable the current flow from source to drain, improved to 0.44 V from 2.1 V due to smaller dimensions. The slope of the source to drain current with gate voltage is also high for the ion-implanted device. So, the improvement in transistors' performance and size was a dramatic breakthrough observed in 1974.

(Refer Slide Time: 15:59)



In the last slide, we have seen a technique based on advanced lithography and ion implantation to reduce the size of the transistor. I will continue the discussion with the help of another celebrated paper published in IEEE transaction of Electron Devices in 1982 by Saraswat et al.

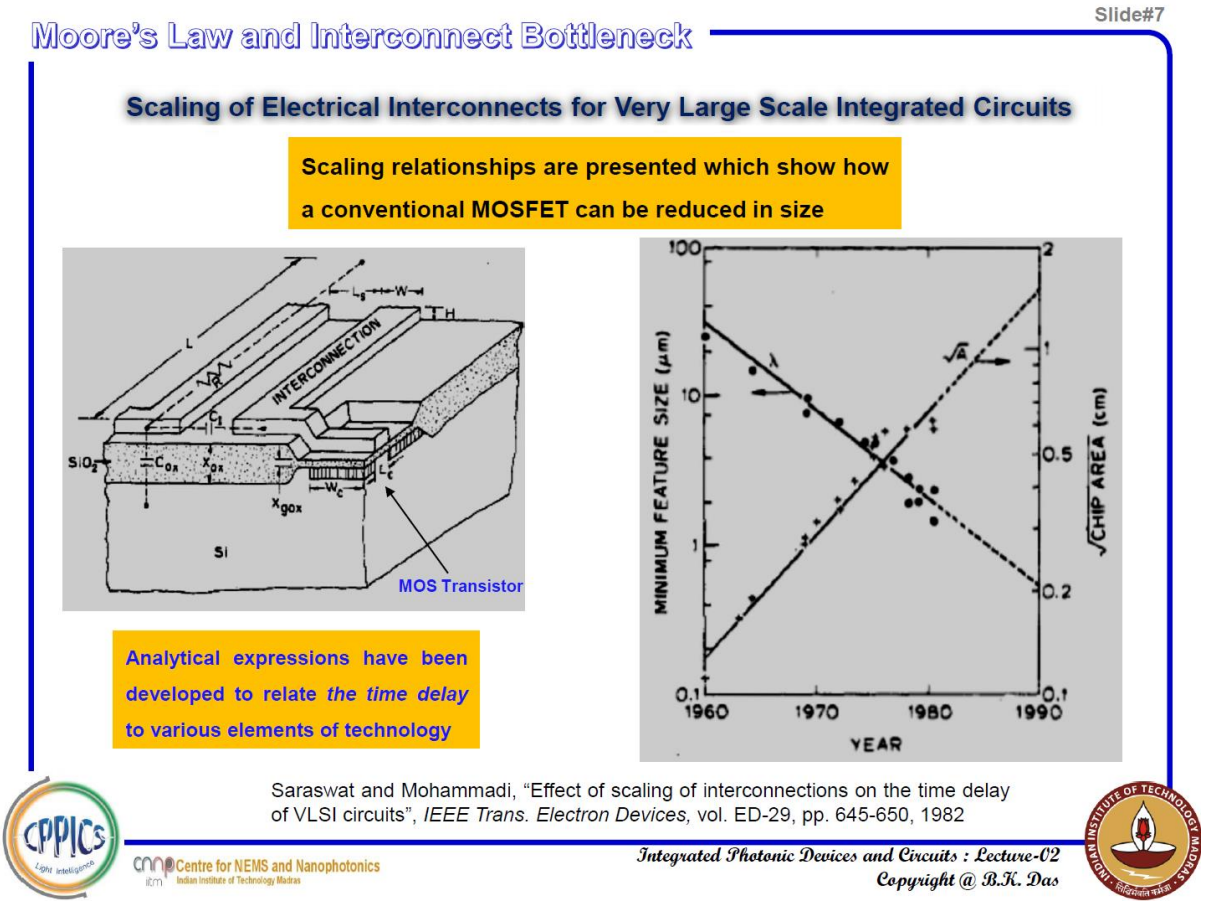
The left side of slide 6 shows a 3D view graph of a MOS transistor circuit with a source, drain and gate oxide. An interconnect connected to the top gate is also shown in the view graph. L_c defines the channel length, the channel's W_c width, and X_{gox} as the gate oxide thickness of the MOS transistor.

If these three parameters, width and length of the channel and thickness of the oxide, are scaled by S, S is the scaling parameter. To maintain the same gate field, you need to improve the doping concentration by S times. If you scale down device size by S times ($S > 1$), you have to increase the doping concentration by the S factor. The scaling will result in the reduction of both threshold voltage and drain current by $1/S$ times. Gate delay, the output switching time with the input signal change, is essential for switching applications. The scaling of transistor

size results in the improvement of delay by the S factor. With further calculation, we can show that the power dissipation per device will reduce by $(1/S^2)$ times, a dramatic improvement. Power density, power consumed per unit area in a circuit, will remain the same. Speed and power product per device is a crucial parameter as a figure of merit that will be scaled by $(1/S^3)$ times.

Therefore, we can conclude from this slide that scaling down the transistor size helps in the integration of 100 million transistors per millimetre square; as reported by Intel, it also helps enhance device performance.

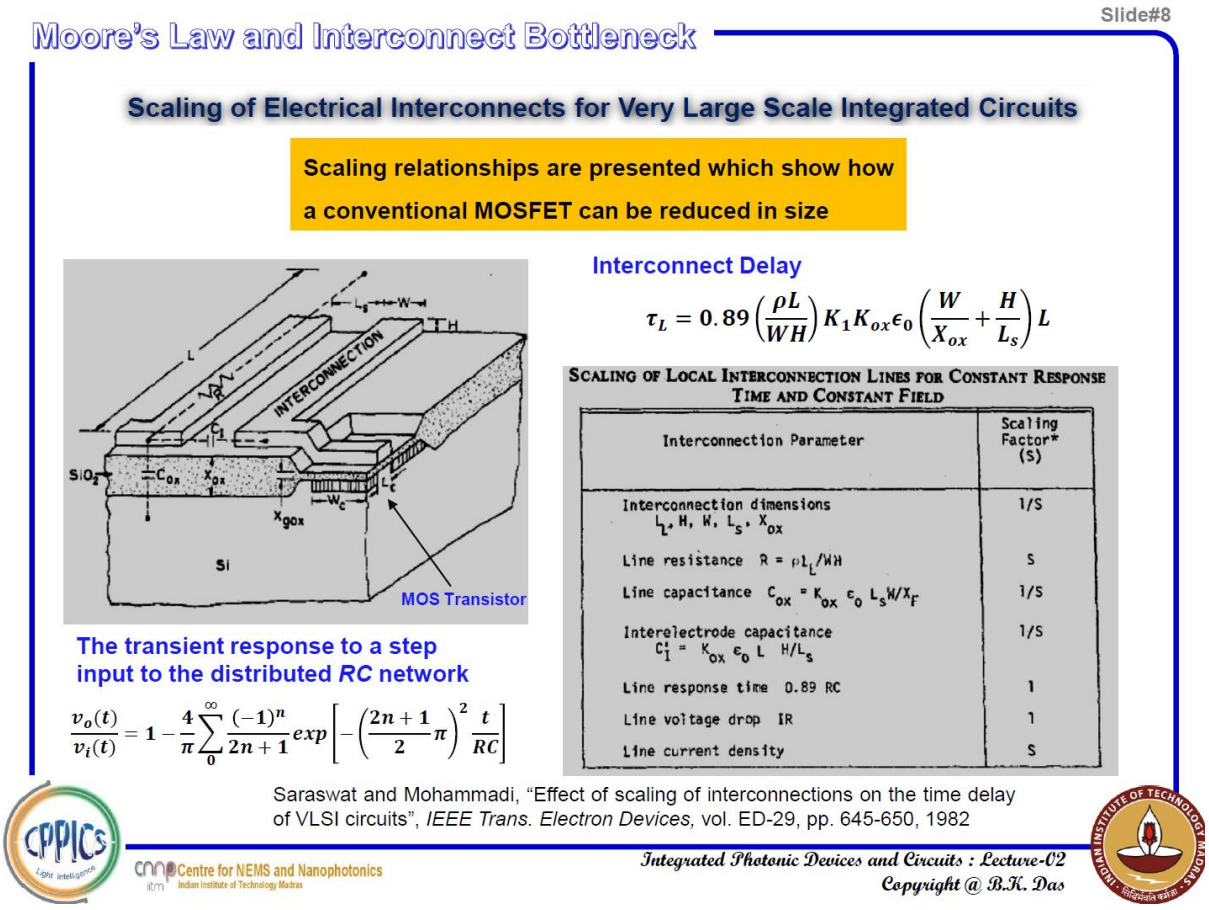
(Refer Slide Time: 20:21)



From this 1982 article, I have taken the plot (right-side diagram of slide-7) to show the evolution of electronic integrated circuits. The graph shows the technology advancement in terms of minimum feature size possible and the increase in chip area (as a consequence of reduced transistor size) from 1960 to 1990. This technology advancement continued beyond the mentioned years of the graph, and it reached the wafer-scale engine today.

In this paper, the authors have also shown the analytical expressions to relate the time delay with various elements such as technology and topology, which I will discuss in the next slide.

(Refer Slide Time: 21:45)



The next problem in a MOS transistor circuit is how to provide the input electrical signal to the on-chip integrated transistor. An interconnect is a metal electrode used to apply external voltage to the transistor, as shown in the 3D view graph of a MOS transistor circuit (left-side diagram in slide 8). The next question we need to ask ourselves is how important are the interconnect characteristics in the performance of an integrated electronic circuit? We will discuss the same in this and the coming slides.

Let us consider an interconnect with width W and height H, as shown in the 3D view graph of a MOS transistor circuit (left-side diagram in slide 8). We know that an interconnect is a metal line connecting two different terminals of a circuit. So, your interconnect size also has to be scaled down; otherwise, you cannot pack more transistors. So, the lower the interconnect cross-section (W times H), the resistance will be higher.

Next, let us consider L_s as the separation between two adjacent interconnect lines, as shown in the 3D view graph of a MOS transistor circuit (left-side diagram in slide 8). The increased integration density results in closely spaced interconnect (reduction in L_s), resulting in the capacitance between the two interconnect lines. When two closely spaced metal plates are at a potential difference, there will be some capacitance between the two plates. Therefore, if you scale down the interconnect dimensions, the resistance will increase, and due to the close spacing of the interconnect, the capacitance effect between the metal lines will increase. These two effects are undesirable for a MOS transistor circuit, and we can model it by considering the interconnect as a transmission line.

In this paper, the authors have shown that the interconnect response depends on the RC constant. The MOS transistor response along with the interconnects, i.e. the change in output voltage $v_o(t)$ of a time-varying input $v_i(t)$, is given by the following expression:

$$\frac{v_o(t)}{v_i(t)} = 1 - \frac{4}{\pi}\sum_{0}^{\infty}\frac{(-1)^n}{2n+1}\exp\left[-\left(\frac{2n+1}{2}\,\pi\right)^2\frac{t}{RC}\right]$$

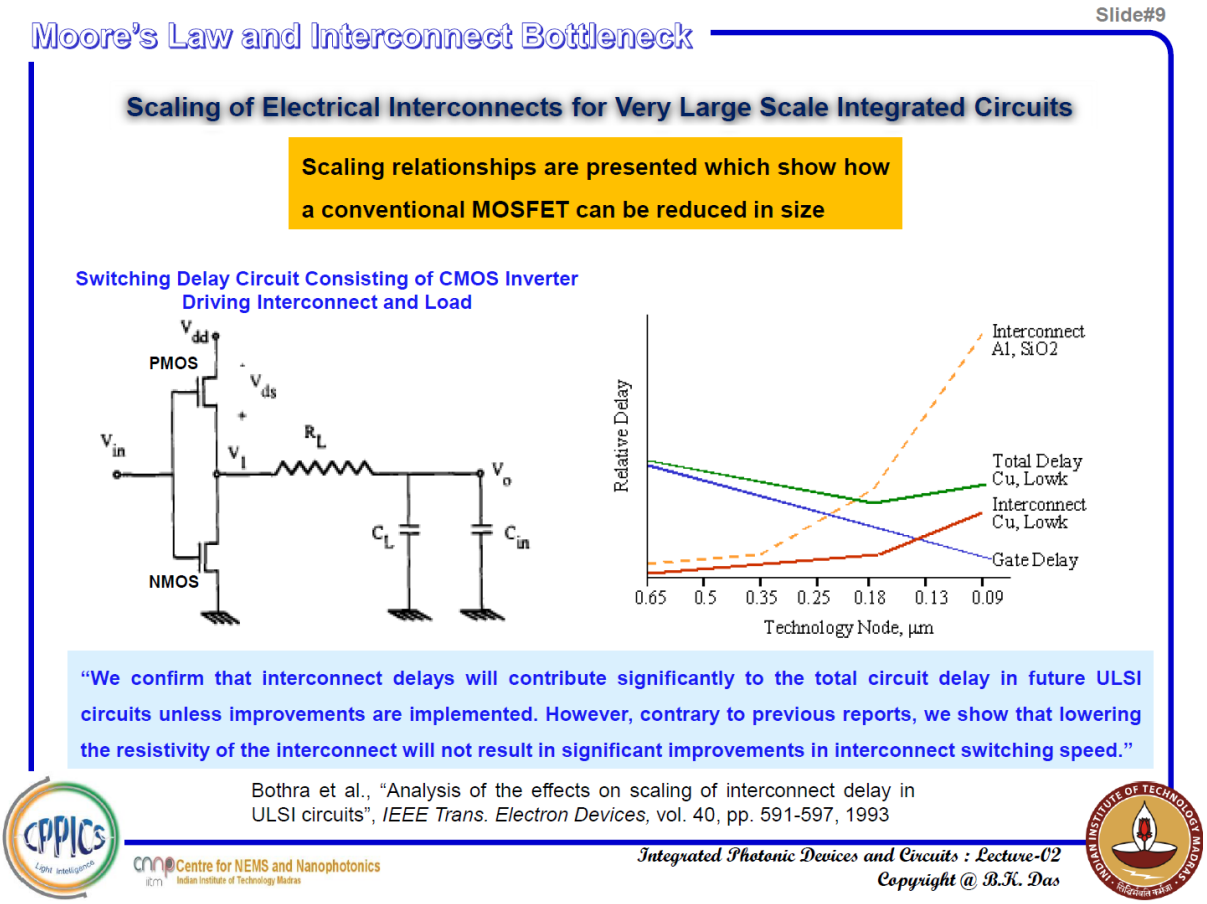
The voltage response is an infinite series. Further, the authors have simplified the term t/RC in the exponential. They derived an expression for the interconnect delay in terms of the transistor and the interconnect dimensions. The interconnect delay means how fast your applied signal can induce a response in the transistor located at a distance L, as shown in the 3D view graph of a MOS transistor circuit (left-side diagram in slide 8). The interconnect delay is given by:

$$\tau_L = 0.89\left(\frac{\rho L}{WH}\right)K_1 K_{ox}\epsilon_0\left(\frac{W}{X_{ox}} + \frac{H}{L_s}\right)L,$$

where, the term $\left(\frac{\rho}{WH}\right)$ is the distributed resistance per unit length, $K_{ox}\epsilon_0\left(\frac{W}{X_{ox}} + \frac{H}{L_s}\right)$ is the distributed capacitance per unit length, K_1 is a constant value corresponding to the overlap integral for the capacitance calculations. Both the distributed capacitance and resistance are multiplied with the transmission length L. The number 0.89 corresponds to the 90% of the total switching delay.

So, if you have control over W, H value and reduce the transmission length L that means you can reduce the delay time, and you can switch the transistor faster. To summarize the discussion, transistors overall switching delay depends not only on the gate delay but also on interconnect delay, which depends on the interconnect dimensions.

(Refer Slide Time: 28:05)



To understand the significance of the interconnect delay, we will consider a simple switching delay circuit consisting of a CMOS inverter driving a interconnect and a load. A CMOS inverter scheme consists of PMOS and NMOS transistors, as shown in the left-side diagram of slide 9. You have only one input common to both the PMOS and NMOS. If V_{in} is low, then the PMOS will be on so you can get the output exactly as V_{dd} (high). When the input is high, in that case, PMOS will be off, and NMOS will be on, and the voltage at the output terminal can discharge to the ground through NMOS and result in a low, so it is an inverter.

However, when joining the output to load (or the next transistor), we do it through a interconnect. Here, we consider R_L as the connector resistance, C_L as the connector capacitance and also C_{in} as the input capacitor of the following transistor. To understand the interconnect delay with the transistor scaling in terms of increase in resistance, the effect of fringing capacitance etc., in this slide, I wanted to explain why that is the case.

If you consider the capacitor, its reactance is given by:

$$X_c = \frac{1}{2\pi f C}$$

where f is the frequency of operation (input voltage signal, $V_{in} = V_{oi}\sin(2\pi f t)$). As the frequency of operation increases, X_C reduces, i.e. the resistance of the capacitance is reduced. So, whenever you send an input at a very high speed, in that case, your signal will not be able to travel longer distances through the interconnect. So, that is why the interconnect will be bandwidth limited.

The right-side diagram of slide 9 shows the delay contributions with advancement in technology node, starting from 650 nm to 90 nm. The scaling of the gate length was possible with the technology advancement, and it helped reduce the gate delay (blue line in the graph). However, the interconnect delay has seen an increment with the scaling and increased resistance and capacitance.

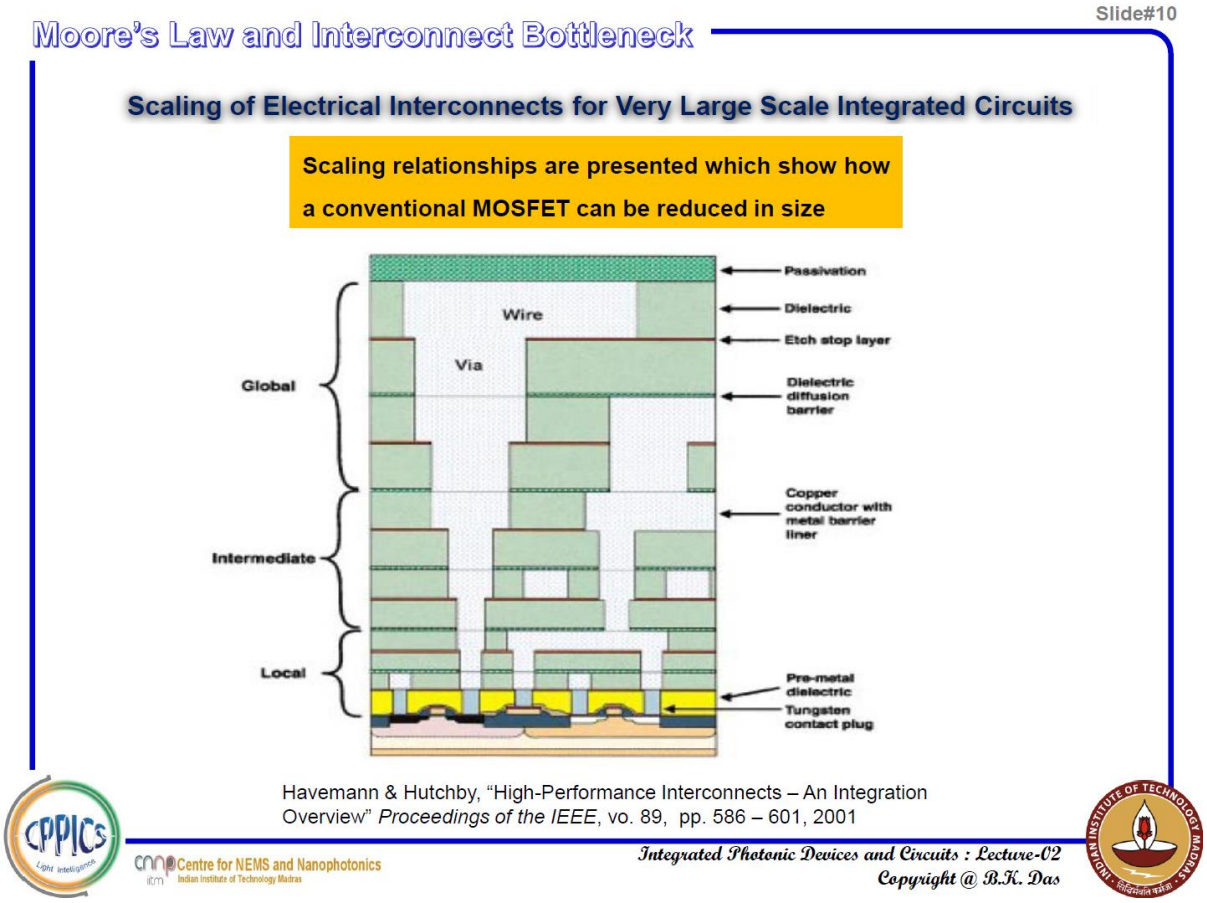
In the earlier days, Aluminum surrounded with silicon dioxide was the common interconnect used in the CMOS industry. The Aluminum interconnects delay rise faster with the technology evolution, as shown by the yellow dotted line in the graph. This considerable delay has resulted in the growth of copper interconnect. The relative delay is in constant rise for copper interconnect surrounded with low K dielectric material (red line in the graph). The delay slope further gets steeper post 180nm technology. Consequently, the interconnect delay is the major contributor to the total delay (interconnect delay + gate delay) for the post 180nm technology node. Therefore, we see no scope of improvement in terms of switching speed despite technological advancement.

This phenomenon of an increase in interconnect delay with advanced technology is known as the interconnect bottleneck. Even though you can increase the transistor packing density, you cannot gain in terms of the transistor's switching speed.

So, here in this paper, Bothra et al. in 1993 comment that "we confirm that interconnect delays will contribute significantly to the total circuit delay in future ULSI (ultra large scale integrated circuit) circuits unless improvements are implemented." ULSI means whenever you are talking about trillions of transistor integration. If you do not have any alternative solution, interconnect delay you will hamper, and you will be stuck with your so-called celebrated Moore's law. The doubling of transistors in a single every 18 months will not be possible. So, that is a huge

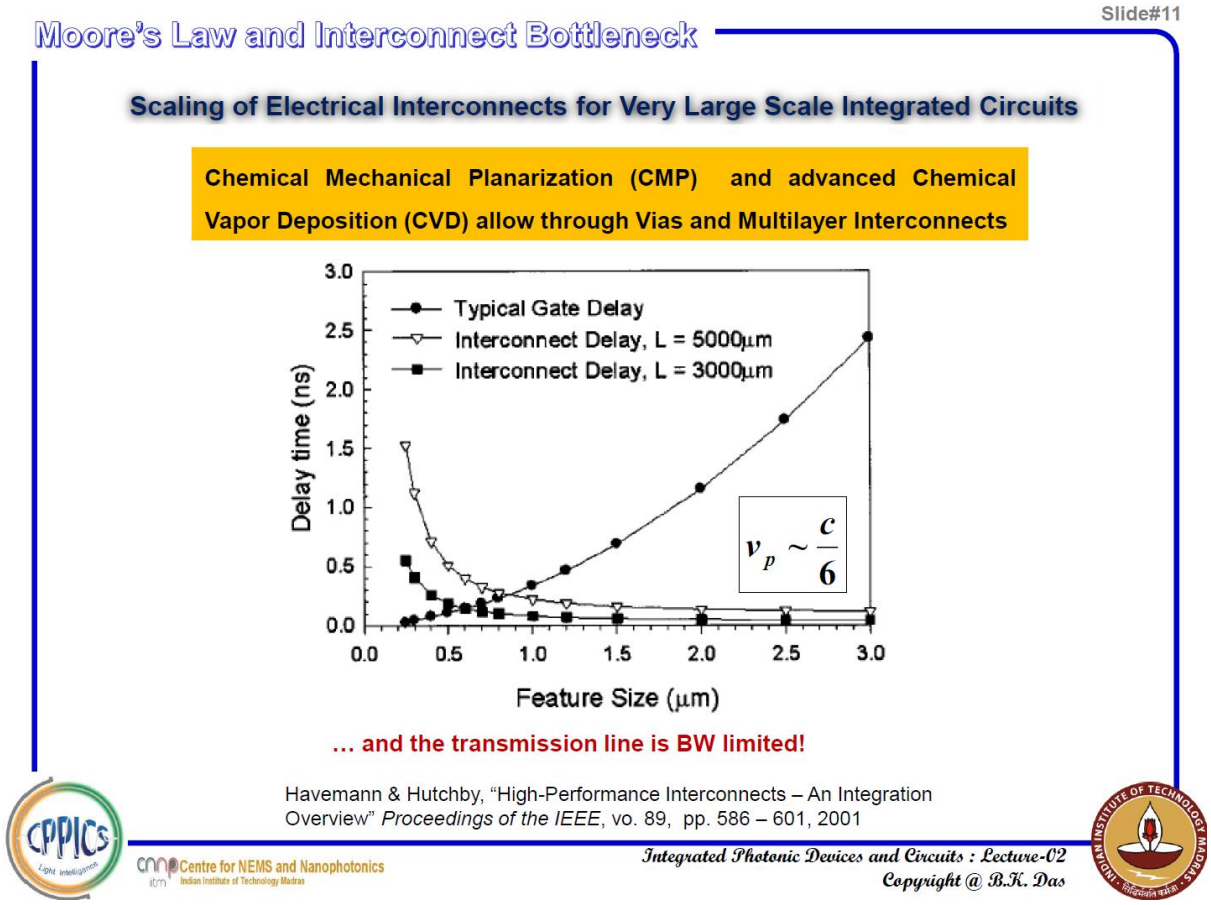
problem. However, people tried to solve this situation technologically, and I will discuss it in the next slide.

(Refer slide Time: 35:25)



Havemann and Hutchby, in 2001, published an excellent article in IEEE proceedings titled "High-Performance Interconnects - An Integration Overview." This article proposes the vertical stacking of interconnects (of different width's) on top of each other through vias. The interconnects can be stacked up to local, intermediate and global levels based on their destination. You can go up to the local layer in case of transistor-to-transistor connection. Whenever you want to connect two logic circuits, which are 10s of a micrometre away from each other, you can increase the interconnect cross-section through a via and go up to the intermediate layer. The increase in the cross-section enables to reduce the interconnect resistance and capacitance and, finally, its delay. In another example, if you want to transmit a clock signal from one end of the chip to the other, you can use a global interconnect with a higher cross-section. In this way, people have demonstrated the interconnect stacking to 9 metallic layers to mitigate the interconnect delay.

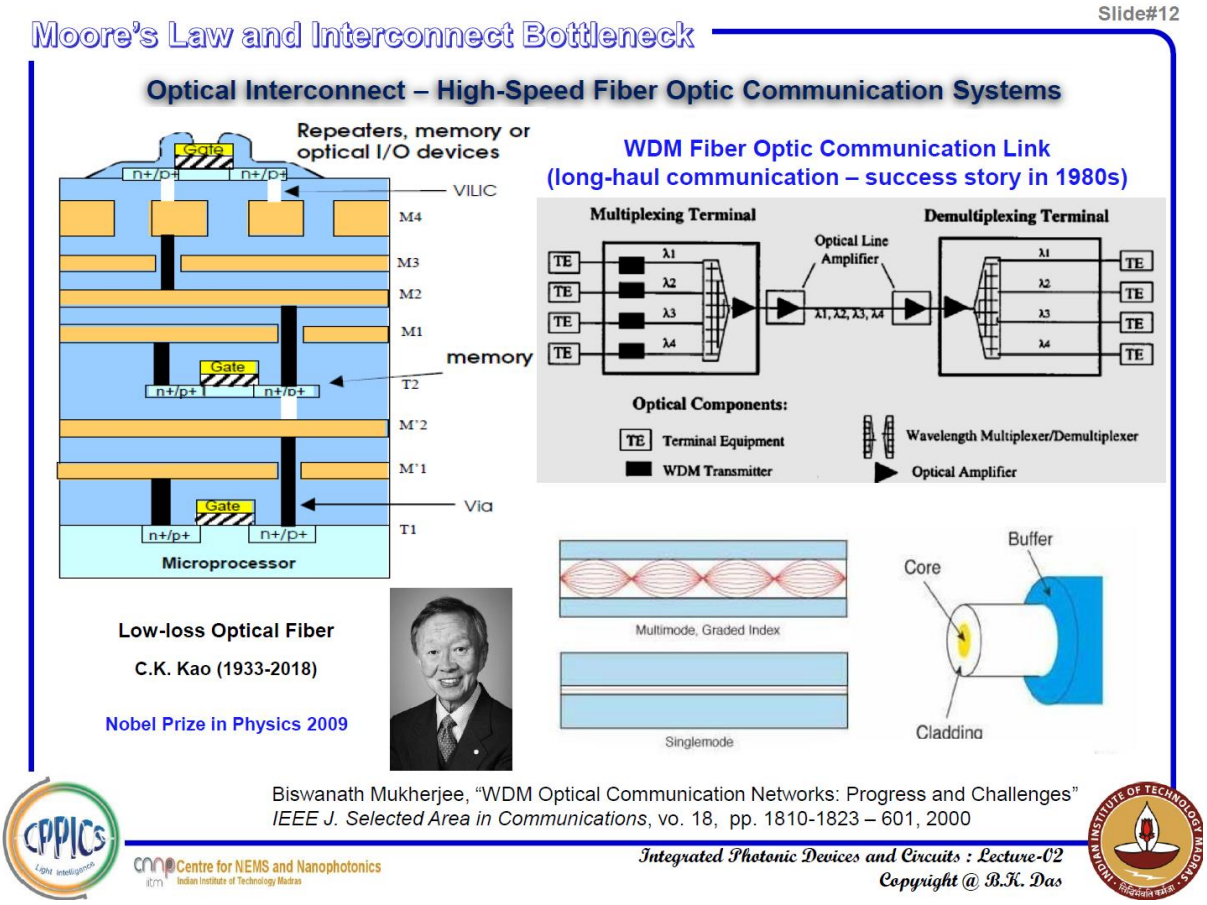
(Refer Slide Time: 38:33)



The demand for switching speed is increasing day by day in our current world. Even though stacking the metals into layers technique has helped mitigate the interconnect delay, there is a limit in terms of the number of metal layers grown on the surface and the interconnect width allowed. The interconnect delay for the delay line of 5000 micrometres and 3000 micrometres and the gate delay for different feature sizes is shown in the plot in slide 11.

If you increase the interconnect cross-section with a via, then the delay is in control. We don't see any significant improvement in the delay time for less than 1-micron feature size. The delay time saturates to the value c/6, where c ($=3 \times 10^8$) is the velocity of light. Therefore, we can conclude that the transmission line is bandwidth limited for the feature size bigger than 1-micron. You cannot send data very fast so, what is the solution?

(Refer Slide Time: 40:53)



The 3D integration of the transistor was one of the solutions proposed to overcome the interconnect bottleneck (left-side diagram of slide 12). Similar to the interconnect layers, the proposed scheme suggests multiple transistor layers based on the circuit operation. The bottom layer for computations and logic operations, the second layer for memory purposes and so on. Whenever needed, the via can the information between the layers can be shared with the help of a via.

Despite the proposed new architecture, long-distance on-chip communication is still a problem. So if we can have some device that can convert the electrical signals into optical signals and propagate the data in a waveguide. In this case, we can transmit the data with the speed of light with low loss.

The proposed idea of data communication using light already existed in long-haul fibre-optic communications, where the optical frequency is around 193 THz (1550 nm wavelength). The success of broadband fibre-optic communication paved the way for transatlantic and

transpacific optical fibre cables. These fibres allow high-speed, secure data transmission from one country to another in no time.

Charles Kao demonstrated the fabrication of long-length, low loss fibre optical fibre. It confines light due to total internal reflection and constitutes a thin circular core region and a thick cladding region around the core (bottom ride-side pictures in slide 12). In a practical fibre-optic cable, a buffer jacket surrounds the core and cladding for protection. We can adjust the fibre dimensions to allow single-mode guidance.

The WDM (wavelength division multiplexing) fibre-optic communication link is shown in slide 12 (top right-side diagram). The system has a multiplexing (transmitter) and demultiplexing (receiver) terminal. The terminal equipment (TE) at the transmitter converts the electrical data into optical data with a particular wavelength (λ). Therefore, the transmitter TE consists of a laser source and a modulator. Here this specific example shows four wavelengths ($\lambda_1 - \lambda_4$) to encode the data. Next, we can use a multiplexer to transmit the multi-channel encoded data to a low-loss fibre-optic link for long-distance transmission. We can also have optical amplifiers between the transmitter and receiver to boost up the signal.

We can demultiplex the wavelengths at the receiver end and send them to terminal equipment to decode the optical signal into an electrical data signal. The terminal equipment (TE) at the receiver end is the photodetector. As mentioned earlier, this example shows the four-channel WDM. However, people have already demonstrated a dense wavelength division multiplexing (DWDM) system with 160 wavelength channels to encode and transmit the data at higher rates. Due to the success of DWDM systems, today, we can get optical fibres to our doorstep (FTTH, JioFiber) and enjoy high-speed internet.

So, in today's lecture, I was trying to discuss the electrical interconnect bottleneck the possible solution available to convert electrical data into optical data and use optical fibres. This way, you have an enormous bandwidth; your speed can be close to the speed of light with the help of an optical interconnect. I also mentioned the success of optical interconnects in long-haul communications. In the following lecture, I will discuss the progress of on-chip optical interconnect for short-haul communication instead of copper wires.