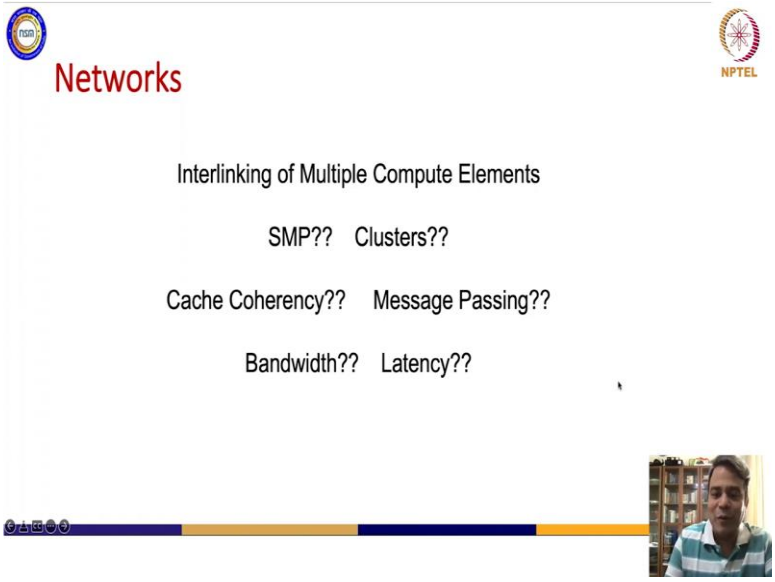
Applied Accelerated Artificial Intelligence
Prof. Ashrut Ambastha
School of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 14
Design Principles for Building High Performance Clusters
Networking Fundamentals Part - 1

Let us start today's session. My name is Ashrut Ambastha. I am Principal Architect at NVIDIA, and my domain of expertise is in Networking. And we will get into the fundamentals of networking as well as how we build large clusters in today's HPC data centers.

(Refer Slide Time: 00:39)



Networks

Interlinking of Multiple Compute Elements

SMP?? Clusters??

Cache Coherency?? Message Passing??

Bandwidth?? Latency??

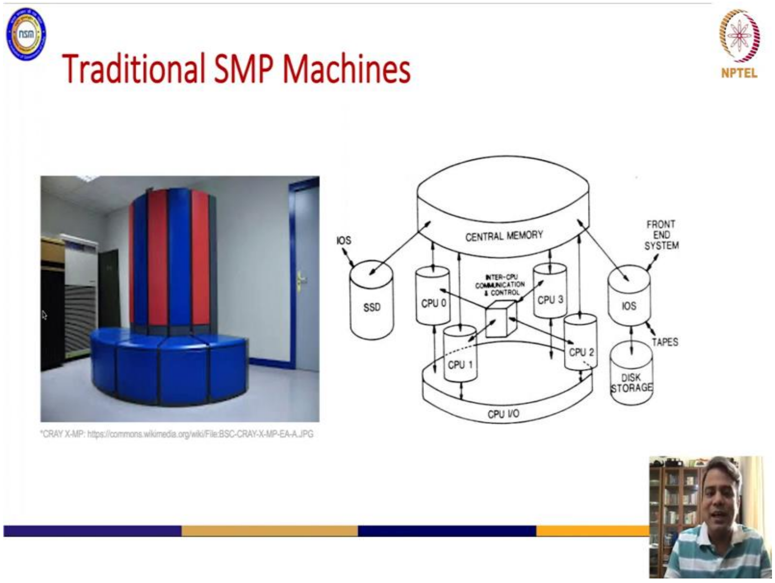
So, when you look at networks, what does first come to our mind? Now, why do you need or in a large high performance computing cluster, what is a cluster? And why is there a cluster? So, let us go back a little bit and see why networking came into being, ok.

So, main thing was that when you a processor was created, right, now you wanted more and more power, people wanted more and more compute, and therefore, the only way of increasing more and more compute was to make the processor faster or was to actually have more and more such processors, ok.

When you have more and more processors, you actually need to interlink all of these, you need to interlink multiple compute elements and then you can share your load across multiple compute elements. So, many things came into being. You must have heard the term SMP, symmetric multiprocessing systems. Then, clusters came into being you look at things like cache coherency, you look at concepts of message passing between various compute elements, and when all these elements come into picture, you look at bandwidth and latencies.

So, today, we are going to cover all these fundamental concepts and kind of build-up on what you already know as standard networks to how you end up making a high performance cluster using high performance network. Or what is the difference between the standard networks that you see in an office environment or in your home with traditional gigabit LAN and Ethernet or the Wi-Fi routers or even a telco, large telco network versus what you do in a large HPC and AI kind of cluster. So, let us go forward.

(Refer Slide Time: 02:39)



Traditional SMP Machines

*CRAY X-MP: https://commons.wikimedia.org/wiki/File:BSC-CRAY-X-MP-EA-A.JPG

So, traditionally, I mean this is a very popular picture, right. Everybody kind of recognizes this. This was one of the first Cray machines which actually enabled connecting 4 CPUs together, and all the 4 CPUs shared a certain central memory. And they were also connected to each other for running the parallel workload, ok.

And to connect all these elements together, there was some inter CPU communication and control. This is where networks for HPC comes into picture, right. You need to have some sort of mechanism to distribute the workload between various CPUs. You need to have this mechanism to coordinate between the various compute elements. You need to have this mechanism to do some sort of memory locking, so that if one element is writing on a certain section of the memory, then the second compute should not be over writing that and so on and so forth, right.

So, now this is traditionally how it started. But you know humankind has always been not satiated with simple stuff, so we wanted more and more. So, I said ok, now I do not be keep on adding more and more CPUs.

And now you can see that over a period of time or if you keep adding more and more CPUs, you cannot the complexity; that is required to connect them together starts increasing manifolds. The central block that you see on this picture, the inter CPU control and communication part, it becomes more and more complex. And there is only a certain limit to which it can handle the number of compute elements.

(Refer Slide Time: 04:33)

Modern Supercomputing Clusters

And this is where industry started moving into something called clusters. And starting from that you know standard typical SMP machines to these large modern supercomputing clusters, is a journey of various technologies. We are going to be covering the networking portion of this.

So, in a modern supercomputing cluster, you do have 100s and 1000s and even 10s of 1000s of these compute elements, but now they cannot all connect into a central location. Because imagine, what is this central location? It is after all some sort of logic, it is a silicon device, right which will coordinate between all this.

Now, imagine if you have to connect 1000 processes to a single silicon trying to coordinate and do data passing between the compute elements, it is impossible to construct. And that is where standard high performance networks comes into picture, where now we are able to link tons and tons of CPUs together to create a single large supercomputer.

The programming paradigms, the use cases have changed quite a lot, because if you look at a standard symmetric multiprocessor machine where whole set of RAM was completely shared and exposed to all the compute elements, you cannot have it in a clustered environment.

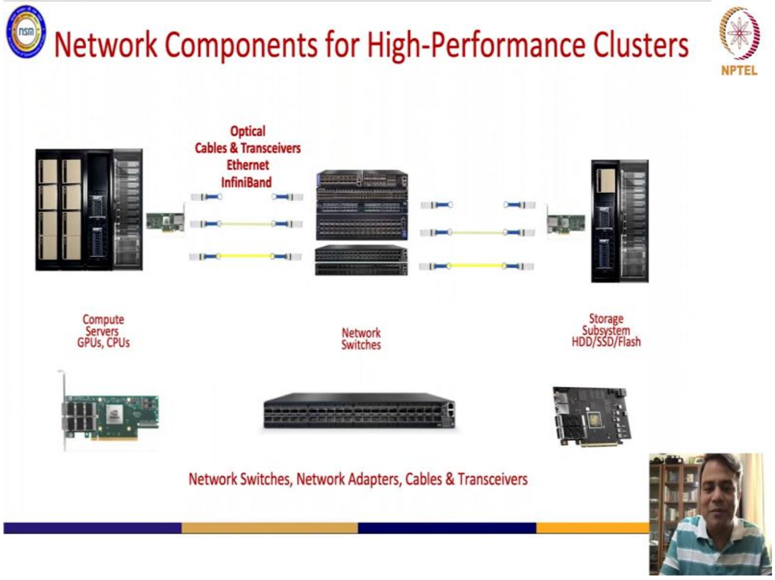
So, a new paradigm of exchanging data came into play which is called message passing with which most of the people who are into HPC domain or AI domain are aware of, right. You need to have some sort of a communication library that packetizes the data from one CPU element, and then sends it across to another element and it does it using some sort of messages which is what we call as an MPI, Message Passing Interface, right.

So, to do all that, you need to have the basis of connecting all these elements. You know when you say message passing, it is a very higher layer of abstraction. It is in the domain of coders. But before that we need to create a cluster. And that is what we will talk about from the fundamentals, what are the design principles, what are the limitations, and what limits us in going from a to b, right.

If let us say today I am saying that we are able to construct a cluster of 20,000 elements, why cannot I construct more than that. If I am saying that today I have a certain amount of bandwidth between the elements, why cannot I increase, or why, what is the limiting factor, what technology defines the limits. So, we will look at those limits.

And hopefully people who are interested in this domain will be very interested to continue research and you know see it as a challenge, ok; why is this limit there, what is the state of the art today, and how it can be improved, ok. So, with that introduction let us get on to today's session.

(Refer Slide Time: 07:43)



Network Components for High-Performance Clusters

NPTEL

Optical
Cables & Transceivers
Ethernet
InfiniBand

Compute
Servers
GPUs, CPUs

Network
Switches

Storage
Subsystem
HDD/SSD/Flash

Network Switches, Network Adapters, Cables & Transceivers

So, when you look at the networking components for a high performance cluster, you actually break it down into 3 4 elements, ok. You have, why do you need this networking? You need this networking because you have a large enough job which cannot fit into a single compute element.

You need to distribute it amongst multiple compute elements. And therefore, you have multiple compute elements that you see on the left. You have some sort of interconnect media that you will use to connect this to another element.

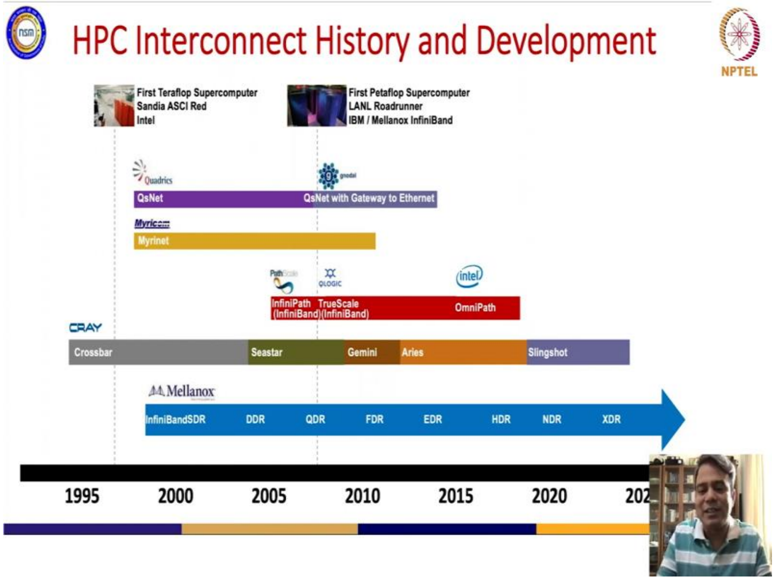
But, now since your cluster is not only a point to point connection between two elements, you need to have 100s or maybe 1000s of compute element, you need to have that central you know interconnect unit which is where we bring in a switch a network switch or a combination of multiple network switches. And therefore, you have a end interface device which is called a network interface card or a host channel adapter in various terminologies.

It is nothing but an interface which kind of talks to the processing element is able to address the memory of the processing element, and its able to send data from its port externally to a particular switch which will receive the data try to make sense out of it and then send it across to its destination, ok.

So, there is element inside your compute which is the channel adapter or a network adapter. There is an interconnect media which are nothing, but your cables. There are switches, and the same thing repeated again, right for the destination side.

So, we will look at all of these elements. And we will mostly look at the fundamentals of the fabric which is from the compute to another compute, what are the elements involved.

(Refer Slide Time: 09:54)



HPC Interconnect History and Development

So, before I go there let us do a little bit of history, you know. So, if you look at the HPC interconnect history and development, it starts right down from I think the early 90s, where we had Cray as one of the leading supercomputing companies. They started off with a crossbar interconnect to link multiple of there of the compute elements, ok.

Then, there was an explosion of multiple standards and technologies by the end of 1990s, ok. So, towards 1999 and 2000, there were companies like Quadrics, Myricom and Mellanox, which standardized high performance interconnect for HPC and HPC cluster. Actually, there was no AI back then.

Now, now AI comes into picture, and we will talk about how the fundamental requirements of AI are completely equivalent to what HPC systems need, ok. So, the systems that we design for HPC now are actually being called HPC/AI systems.

So, yes, many companies started off and there were lot of standards, ok. The whole aim of all these standards was to have the fattest possible pipe between the compute elements. So, that we can transfer the highest amount of data passed in the least amount of time which corresponds to your bandwidth.

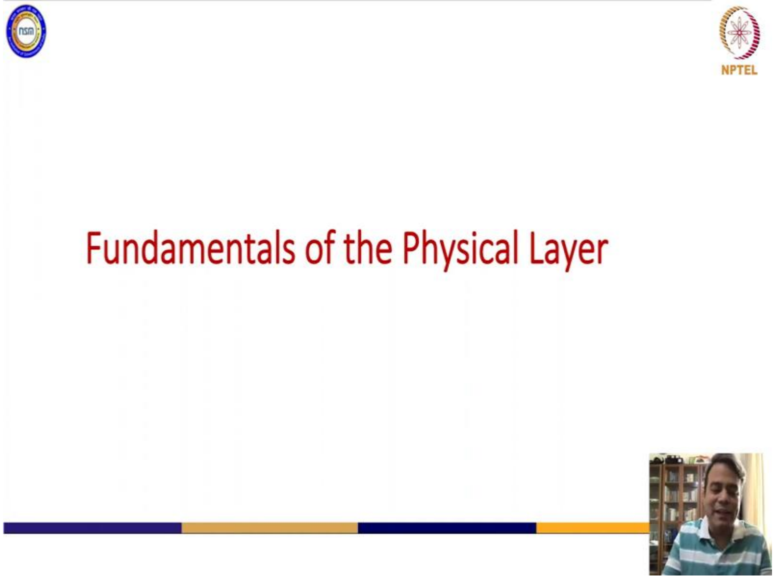
It also aimed at reducing the latency of these transfers, because every time when you were doing parallel compute loads, people would not do very large data set transfers. There might be simple communication wanting to like a mutex kind of communication, right or simple flags or simple you know single element data exchange.

So, it was imperative to have these network be very low latency because whatever time that is spent by the data in flight is the time no compute can happen, right. If you are looking at synchronous compute and communication, the time spent on moving data from one point to the other is time wasted. So, you have to have very small as small as latency as possible.

So, the two fundamentals for high performance networks for building clusters was high bandwidth and low latency. And over the years many companies you know did lot of things. And right now if you look at today’s scenario two of the companies are surviving, one is Mellanox which was actually purchased by NVIDIA and it is now which was the proponent of InfiniBand as an interconnect technology.

The other is Cray which got actually acquired by Hewlett Packard, HPE, and they propagate their own proprietary interconnect for making supercomputers. Some other companies came into being, but then they have stopped manufacturing HPC networks. Again, this is more of history and technology and companies where things stand today. But, what we will discuss is the fundamentals, ok. And then will map it into the current generation.

(Refer Slide Time: 13:22)

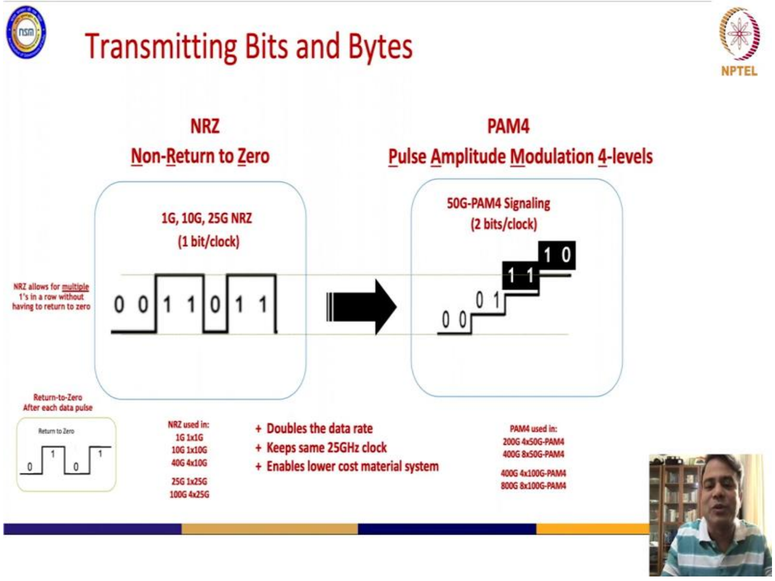


Fundamentals of the Physical Layer

So, let us go into the fundamentals. I thought of having this session with top downs, top down approach where we will talk about today’s cluster and then will drill down into the underlying technologies. But, if since this is more of you know knowledge sharing session, so I thought let us let me do a bottoms up approach.

So, we start with the fundamentals. The fundamental is the physical layer, ok. If you have to connect two elements together, you need to have some method of connecting it; and what are we aware of, ok.

(Refer Slide Time: 13:59)



Transmitting Bits and Bytes

NRZ
Non-Return to Zero

1G, 10G, 25G NRZ
(1 bit/clock)

NRZ allows for multiple 1's in a row without having to return to zero

PAM4
Pulse Amplitude Modulation 4-levels

50G-PAM4 Signaling
(2 bits/clock)

Return-to-Zero
After each data pulse

NRZ used in:
1G 1x1G
10G 1x10G
40G 4x10G

25G 1x25G
100G 4x25G

+ Doubles the data rate
+ Keeps same 25GHz clock
+ Enables lower cost material system

PAM4 used in:
200G 4x50G-PAM4
400G 8x50G-PAM4

400G 4x100G-PAM4
800G 8x100G-PAM4

So, first of all when you look at the physical layer, it is nothing, but a wire. It can be a wire, right. You take you take two elements and you connect a communication cable between them. And what does this cable do? Well, actually it just transmits 0s and 1s, right. It sends out sequences of 0s and 1.

And there are various you know people who are from communication engineering background they know that there are various methods of sending these 0s and 1. You can have different kind of coding, you can have higher frequency, and so on. And that is what we do today, ok.

And the thing that we are most commonly aware of is a standard LAN cable, right. A LAN cable, RJ-45 cable contains copper conductors inside. And there is a certain speed at which you do send the bits, you sends send the 0s and 1s. In today’s world, we are looking at sending the 0s and 1 over a single pair of copper conductor as fast as that 25 gigabits per second or even 50 gigabits per second, ok.

And depending on what kind of signalling are you doing, if you are doing non return to 0 where 0 corresponds to 0 volts and 1 corresponds to a high voltage, or you can divide the levels of analog signal by in multiple levels represent every level to a particular bit pattern. So, you do you know amplitude modulation, right. So, you do different kind of signaling. But it is the media that transmits the data between these two compute elements.

(Refer Slide Time: 15:36)

What We Don’t Want to Achieve

NPTEL
HOW STANDARDS PROLIFERATE:
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

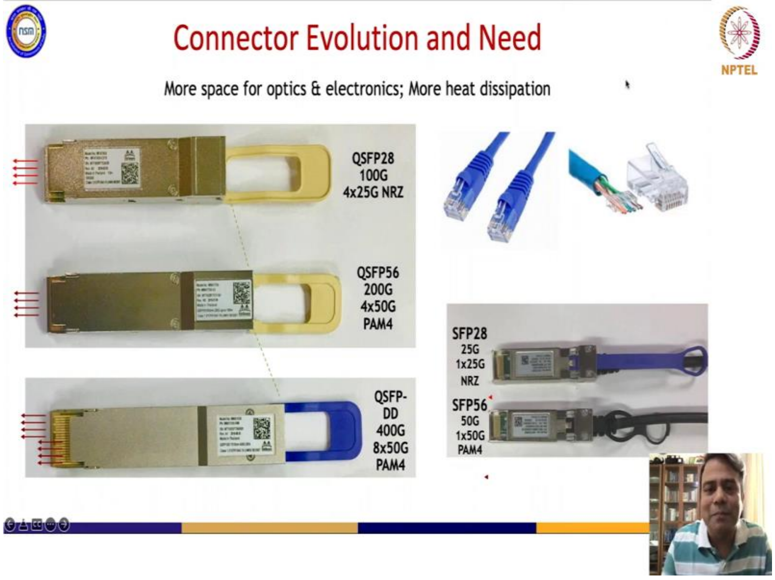
Now, one nice thing we have seen you know over the years, there are so many ways of doing this thing, ok. And every time what happens is you know one company propagates a particular standard, another company propagates another standard, ok. One says, ok my analog designers are good to send bits very fast. So, they can design their ICs, so that the switches are very fast, ok. So, I can do 25 gigabits per second.

But another company says, hey I have got engineers which cannot design the on off control, so fast enough, but they can do it very accurately. So, I can do multiple levels of analog voltages, ok. So, I will propagate a standard which is you know non-return to 0 and it does a modulation, different kind of modulation. So, there are so many things.

One company says, ok I will send it across one pair of copper conductor, other company says I will send it across 4 pairs because I can do parallel, ok. So, all this resulted in multiple competing standards. And then you know there is always a consortium of companies, they come together and say hey it is ridiculous we need to develop one universal standard that will cover everything and it will cover everybody’s use case.

And eventually what happens, it results in another standard, a 15th standard, right. So, now there are 15 competing standards. Again, there is no I would say there is no way of addressing this. This is how it has been. And therefore, we try to achieve as much standardization as possible. But because HPC and AI or high performance networking for HPC and AI cluster, is such a niche I would say field that whoever does things the best and the fastest gets standard, ok or puts things in a standard.

(Refer Slide Time: 17:42)



Connector Evolution and Need

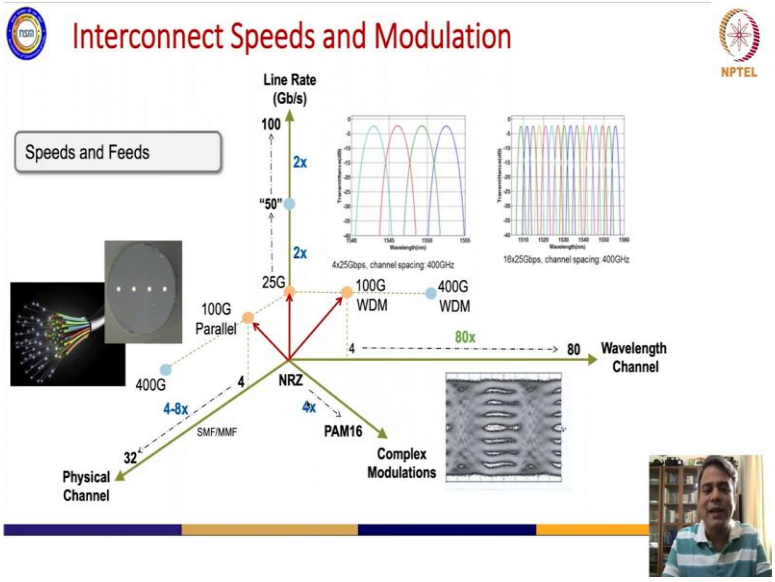
More space for optics & electronics; More heat dissipation

So, let us look at we will talk about standards; let us look at some simple things. I am talking about connecting servers together, right. When I am talking about connecting servers together, I am looking at these physical media connectors, ok. You all are aware of these standard LAN cables. It contains multiple copper conductors inside and it connects to your simple RJ-45 port and it basically sends out data.

But this as you all know is not very suitable for very high speed networking. What is the maximum that you see? You see a maximum of 1 gigabit per second or you see a maximum of nowadays 10 gigabit per second, right. In HPC or AI clusters we are looking at 400 gigabits per second. Obviously, these standard links are not capable of carrying that much. And we will touch a bit of why they are not capable of carrying that much, ok.

Therefore, you have standards which are based on SFPs, standards which are based on QSFPs. These are various connectors which are used to link two systems together or link a system to your switches, ok. And they have different different capabilities. But the fundamental capability comes from the fact that you need to send data as fast as possible.

(Refer Slide Time: 19:16)



Interconnect Speeds and Modulation

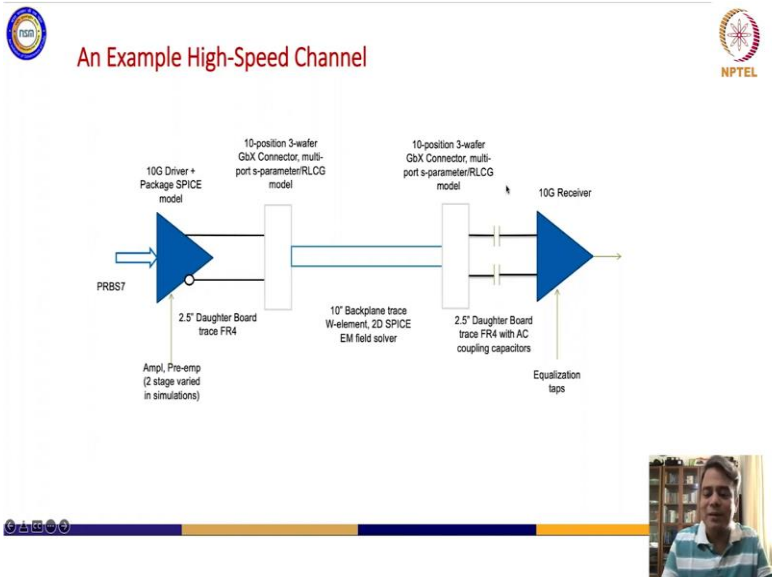
Speeds and Feeds

What are the norms that you can tweak to send this data fast enough? What you can do. We talked about simply encoding, right. So, you can encode more rather than having voltage level represent 0 and 1, you can actually do encoding, so that you can carry two bits in every clock cycle. You can increase the clock cycle itself rather than clocking data at 25 gigabits per second, you can do it at 50 gigabits or 100 gigabits.

You can use multiple physical channels, so that when you connect two systems together, if you go in your normal RJ-45 there are 4 twisted pairs, right. Why is it called UTP? Twisted pair. So, so unshielded twisted pair. There are 4 pairs of copper. So, you can have more parallel lanes to increase the bandwidth, ok.

All of or when you are going optics you can start multiplexing multiple wavelengths into an optical fiber. So, all of these parameters have got certain limit and limits are very much related to current technology physics. So, I will just concentrate 5 minutes on one of the limits or how this limit is reached, right.

(Refer Slide Time: 20:29)



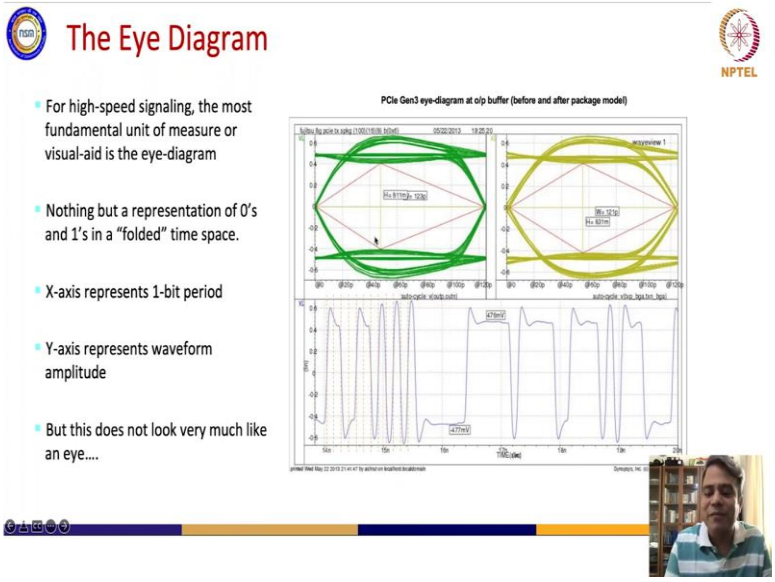
An Example High-Speed Channel

10G Driver + Package SPICE model
10-position 3-wafer GbX Connector, multi-port s-parameter/RLCG model
10-position 3-wafer GbX Connector, multi-port s-parameter/RLCG model
10G Receiver
PRBS7
2.5" Daughter Board trace FR4
10" Backplane trace W-element, 2D SPICE EM field solver
2.5" Daughter Board trace FR4 with AC coupling capacitors
Ampl, Pre-emp (2 stage varied in simulations)
Equalization taps
NPTEL

Let us look at a standard high speed channel between transmitting element and a receiving element; that means, between server 1 and server 2. Whatever is there inside the server, it is a black box.

So, this triangle, it is the transmitter, it represents that black box. It goes into a connector. The connector goes into a channel, the channel can be a cable, it can be a backplane of a large supercomputer where you plug in multiple servers as blades, ok that channel carries signals. There is a receiver, connector, and then there is a receiver amplifier and therefore, this entire thing forms a high speed channel.

(Refer Slide Time: 21:15)



The Eye Diagram

- For high-speed signaling, the most fundamental unit of measure or visual-aid is the eye-diagram
- Nothing but a representation of 0's and 1's in a "folded" time space.
- X-axis represents 1-bit period
- Y-axis represents waveform amplitude
- But this does not look very much like an eye....

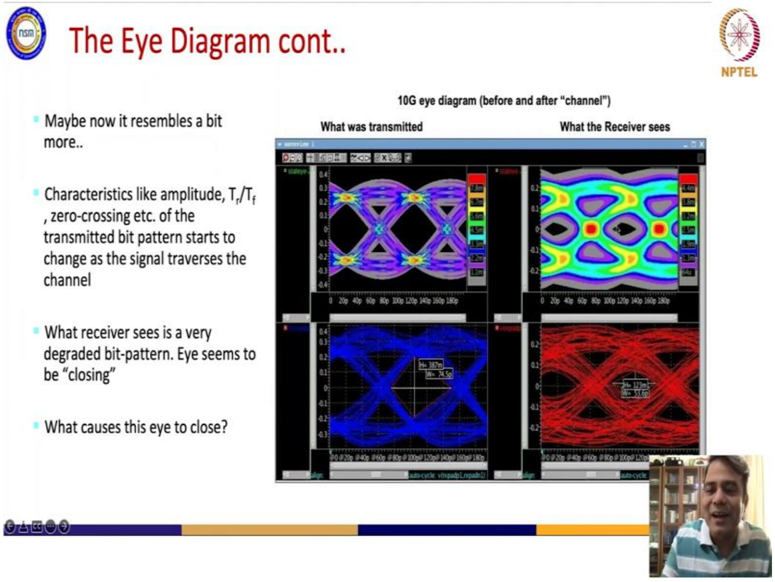
PCIe Gen3 eye-diagram at o/p buffer (before and after package model)

What; now. When we are sending data on this high speed channel you send it in 0s and 1s. You must have seen in many you know interconnect related blog post or papers that you study, you always see something called an eye diagram.

Whenever there is a high speed network shown, it shows a very nice you know diagram which looks like an eye. This one is not looking like a eye, but I just want to show you why I have it here. What is an eye? What is this diagram? It is nothing but a representation of the bits that are being sent from one server to the other. The bits are these 0s and 1. Every 0 and 1 has got a bit period. The period of this bit is nothing, but the reciprocal of the frequency.

So, if you are doing transmitting data at 25 gigahertz, the bit period is certain amount of nanoseconds. So, if you plot all the bits in that window and then overlap all the other bits, you will see something that looks like an eye, ok where x axis represents the period of your bit, y axis represents the amplitude of your bit.

(Refer Slide Time: 22:31)



The Eye Diagram cont..

- Maybe now it resembles a bit more..
- Characteristics like amplitude, T_r/T_f , zero-crossing etc. of the transmitted bit pattern starts to change as the signal traverses the channel
- What receiver sees is a very degraded bit-pattern. Eye seems to be "closing"
- What causes this eye to close?

NPTEL
10G eye diagram (before and after "channel")
What was transmitted
What the Receiver sees
H= 387m
W= 74.5p
H= 123m
W= 53.6p
0 20p 40p 60p 80p 100p 120p 140p 160p 180p

If now let me go here. This looks more like it, right. This looks like an eye. This is nothing, but the bits that are actually travelling through your interconnect channel. And this is what people who are electrical engineers who are working on circuits with high speed networking would see.

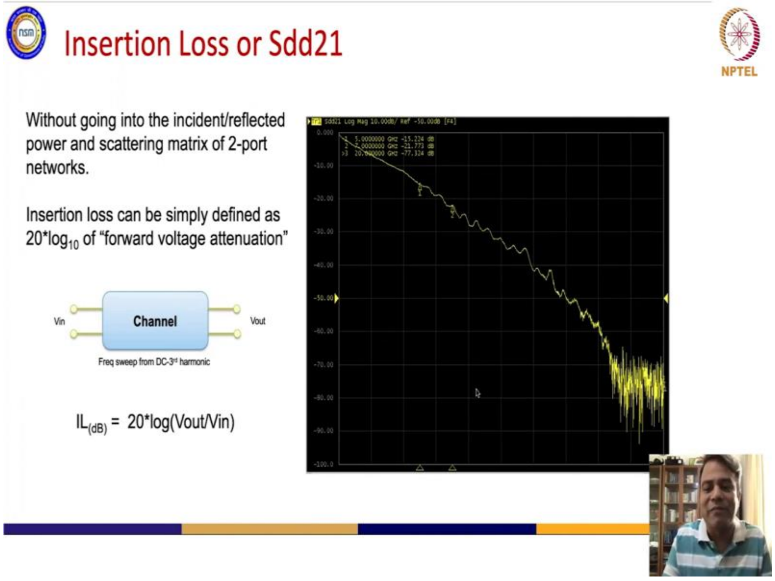
This diagram just represents that if there is a 0 or if there is a you know which is mapped to the negative voltage; if there is a 1, it is mapped to the positive voltage. And the margin in between where I have no bit present or no sample present is my margin. That means, my receiver can distinguish between a 0 and a 1.

Why it can distinguish between 0 and 1? Because the voltage levels are distinct over here. So, if I have a comparator, I can make what is make out what is a 0 and what is a 1 by comparing against a level crossing, ok. When we are transmitting this data through a channel the bit period or the eye diagram starts looking somewhat like this.

The eye; this is what we say that the eye is closing. What it means is now at the receiver end, I am not able to make out or you know there are times when the channel is very long, the eye completely closes. So, there is no more black portion in the center over here. Therefore, my receiver cannot make out whether it has received a 0 or a 1.

So, this defines one of the fundamental limits, ok. Why? Unless and until if you are able to make out between 0 and 1, you have not actually transmitted data from the source to destination. What defines this fundamental limit? This is where we come into.

(Refer Slide Time: 24:17)



Insertion Loss or Sdd21

Without going into the incident/reflected power and scattering matrix of 2-port networks.

Insertion loss can be simply defined as 20*log$_{10}$ of “forward voltage attenuation”

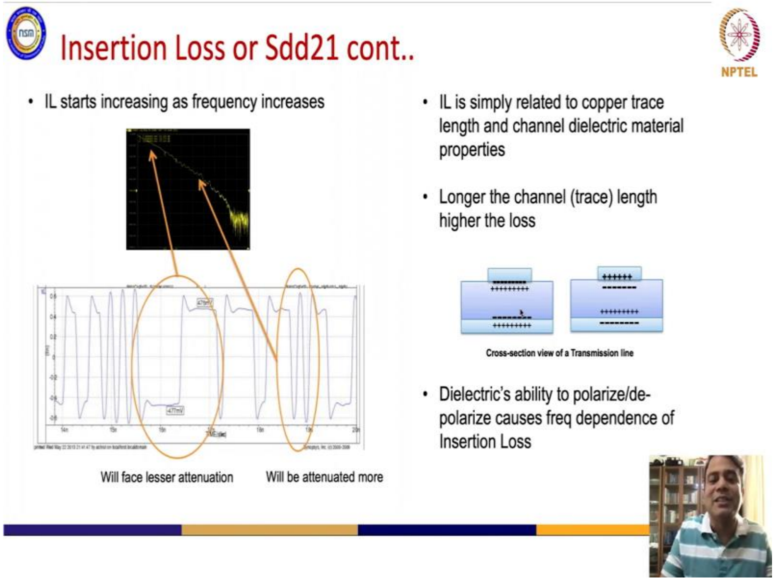
$IL_{(dB)}$ = 20*log(Vout/Vin)

NPTEL
Vin
Channel
Vout
Freq sweep from DC-3rd harmonic
Tr1 Sdd21 Log Mag 10.00dB/ Ref -50.00dB [F4]
1 5.0000000 GHz -15.224 dB
2 7.0000000 GHz -21.773 dB
>3 20.000000 GHz -77.324 dB
0.000
-10.00
-20.00
-30.00
-40.00
-50.00
-60.00
-70.00
-80.00
-90.00
-100.0

So, I am getting into why you know what are the limits right now or why we cannot you know why do we have only 10 gbps, why cannot I make 1 tbps today, ok. The reason is that I have a certain media, ok and it has got certain characteristics, it has got certain kind of loss characteristics. Every media has got frequency dependent, I would say conductance, right.

So, when you start increasing the frequency of the electromagnetic wave that you are transferring, the amount of resistance that is offers it increases. There are more losses, ok. And channels are not capable of keeping up to a certain speed limit. Why they are not able to keep up with a certain speed limit? Again, you go into the physics of it. There are copper losses, there are reflections. I will talk a little bit about that.

(Refer Slide Time: 25:13)



Insertion Loss or Sdd21 cont..

- IL starts increasing as frequency increases

Will face lesser attenuation

Will be attenuated more

- IL is simply related to copper trace length and channel dielectric material properties
- Longer the channel (trace) length higher the loss

Cross-section view of a Transmission line

- Dielectric's ability to polarize/de-polarize causes freq dependence of Insertion Loss

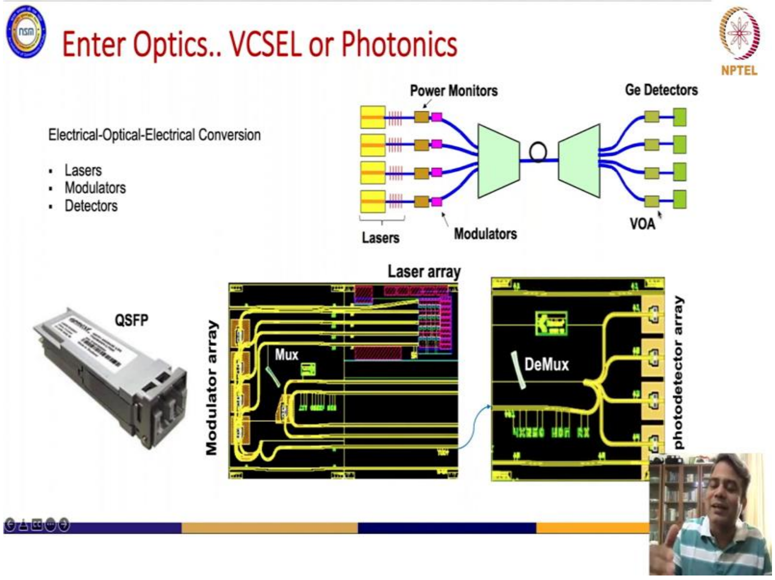
But what happens is; why do you see that closing eye? You see that closing eye because when you are transmitting 0s and 1, it can be any random pattern. It is not 0 1 0 1 0 1. It can be multiple 0s followed by multiple 1s or some 0, some 1s or it can be even 0 1 0 1, depending on what signal is going.

When you look at that bit period or that bit diagram, multiple 0s followed by multiple 1s, it is a lower frequency, right because the time period is high. When the bit pattern is a continuous sequence of 0s and 1s, the frequency is higher. So, you are actually transmitting a wide range of frequencies, it is a wide band.

And obviously, when you transmit the lower frequencies, the attenuation of the media to these frequencies correspond to this side of the spectrum, the left side of the spectrum. When you go on higher frequencies you have more attenuation, that is standard property of copper conductors.

When you are transmitting electromagnetic waves, it is a transmission line. There is obviously, a dielectric. There is a ground reference, there is a plus, and a minus, and there is a dielectric in between. When the electromagnetic wave is going through the media, the dielectric also needs to polarize and depolarize at the same frequency at which you are transmitting the waves. Obviously, this dielectrics ability to polarize and depolarize causes frequency dependence of insertion loss and because of which higher frequencies get attenuated more, lower frequencies get attenuated less, and you see a closing eye.

(Refer Slide Time: 26:58)



Enter Optics.. VCSEL or Photonics

Electrical-Optical-Electrical Conversion

- Lasers
- Modulators
- Detectors

So, these are some of the limits. And therefore, you know industry obviously, moved to optics. There are tons of optics out there now. Why? Because well insertion loss in optics we know is very less, ok. When you are sending an optical wave, it goes through optical fibers and you do not get so much insertion loss. So, now, I can carry longer distances, I can carry more speeds. But when you go into optics you require more power, you require more energy, you require more complicated electro optical converters.

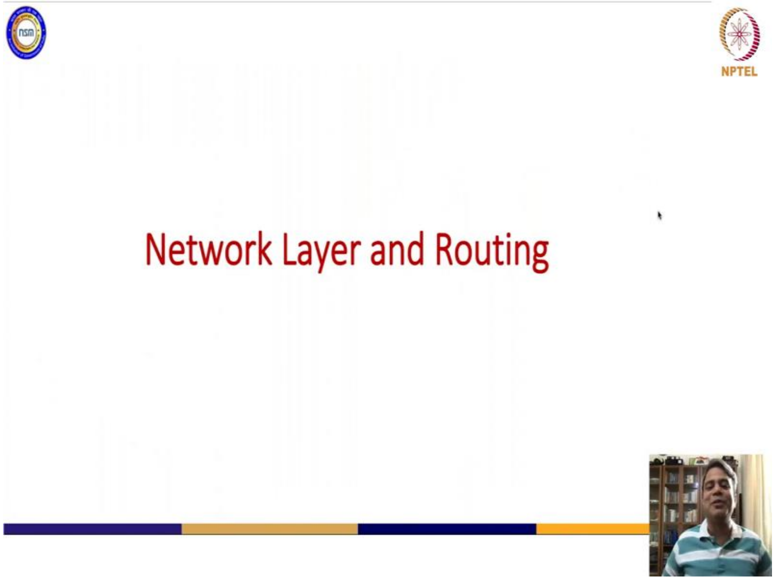
How do you launch data into the optical fiber? Basically, it is nothing, but a laser which is being turned on and off at the same frequency as your data rate, right. And the laser when it turns on it sends a pulse of light and the photo detector detects it and then converts it back into electrical signal, makes it a 1. If the laser is off, it is a 0.

There are other technologies that people work on now which is related to photonics, where you also try to modulate; instead of turning the laser on and off, you try to modulate the laser. So, that you know you do not you operate in a region where you have lesser delays between turning on and turning off and therefore, you can transmit data faster and so on and so forth.

So, starting from standard LAN cables or if people you know people from the 90s would also remember standard modems, right, cable modems, telephone modems, in which we used to get kbps of speeds, right because it was you know a certain kind of media.

Standard UTP connectors would give you that time 100 mbps. So, starting from that today the industry is able to do with single ports link speeds of around you know 400 gigabits per second. So, this is where we solve the bandwidth problem. And of course, there are developments going on in which we would actually increase it from 400 to maybe a terabit per second and beyond, ok.

(Refer Slide Time: 29:11)



Network Layer and Routing

So, this is on the physical side. Definitely, we do not want to spend all our time on the physical side. So, let me come into the next side.