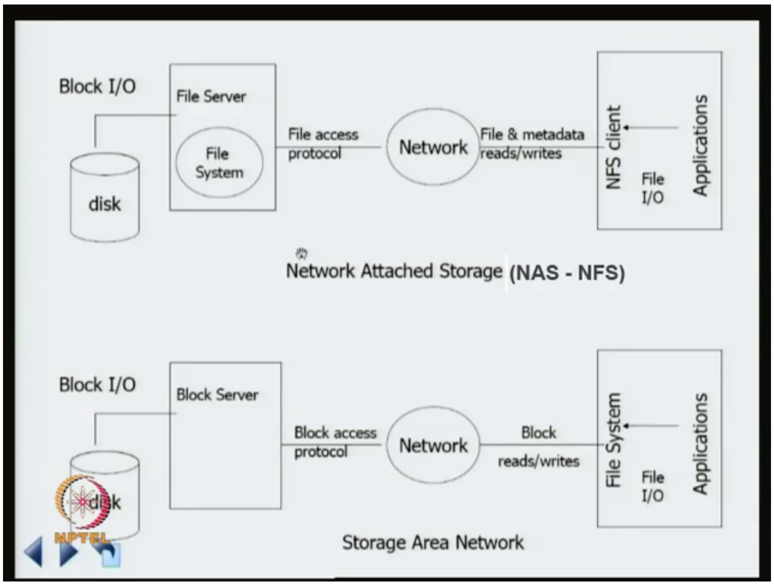
Storage Systems
Dr. K. Gopinath
Department of Computer Science and Engineering
Indian Institute of Science, Bangalore

Communication Protocol for Networked Storage Systems
Lecture - 07
Protocol for Networked Storage Systems, Modern Network Storage Protocol Stack, Flow Control in SAN/NAS, SCSI Protocol Vs. Fibre Channel Protocol, Fibre Channel Protocol - layers/working/classes-of-service/exception handling/flow

Welcome again to the NPTL course on Storage Systems.

(Refer Slide Time: 00:24)



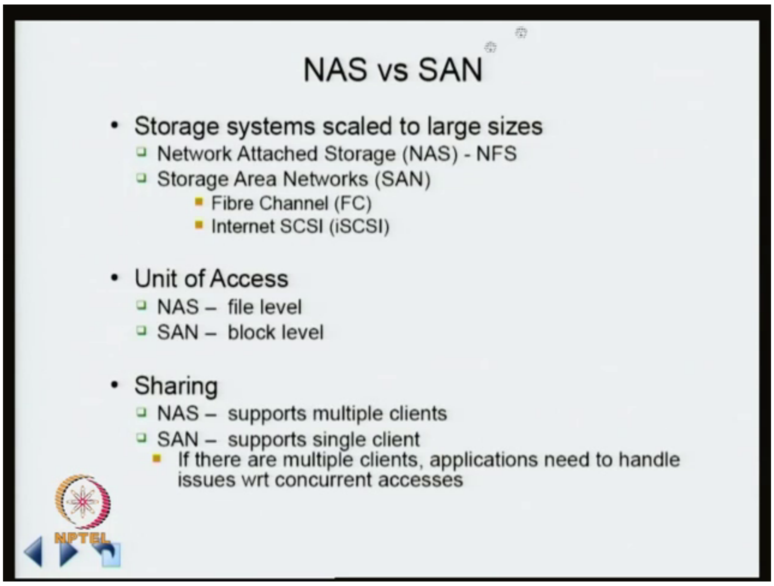
In the previous classes we look at the disk as a device. We also looked at how a computer system can interact with the disk, plus protocol called SCSI, that is what we looked at in the previous class.

Now, there are various space in which the same protocol SCSI protocol can be used to interconnect multiple devices and with applications. So, we look at there are 2 possibilities something called storage area network, something call a network attached storage, and there are also other models like distributed storage. So, we will look today at some aspect little to the storage area network, and then we look at network attached storage and later we look at how distribute the storage models work.

So, first storage area networks we can use this SCSI protocol. As I mentioned before, this SCSI protocol usually in the beginning when it was initially designed, it was meant to be running on an a electrical. Let us say, transport might call it, and the idea here is if I am going to make multiple devices, multiple disks or multiple I O units communicate using SCSI protocol. You might need electrical connections might be problematic for longer distances. So, the idea is to see how we can do it better. So, normally the problem with the electrical connections is that for example, if you have 32-bit transfers, you need 32-bit electrical lines, 32 electrical lines, and these there are this problem calls block skew which creates problems. Therefore, what people like to do is to go for serial protocols.

Now what I have ideally what we want to do is to take SCSI protocol encapsulated into a serial communication medium, and then interconnect things. So, that is one way to do it. The other way to do it is to think of not in the SCSI level. But in the file level you think of the application talking to this storage system at the file level, not at the block level. So, if you are using the SCSI, we are essentially going to begin with the block level, whereas if you go to slightly higher-level protocol. The file protocol, then basically that is going to be using these lower level possibly SCSI kind of protocols to access there. So, there are 2 different types of base of accessing it. And you would like to look into each of this in some detail. So, what I will do right now is to briefly summarize the difference between these 2, and then proceed with discussing how fibre channel protocol enables you to encapsulate SCSI commands.

(Refer Slide Time: 04:00)



NAS vs SAN

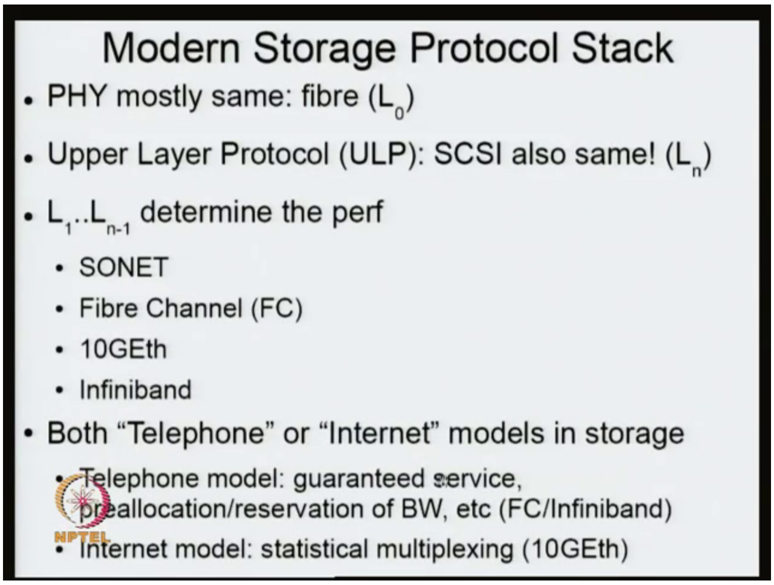
- Storage systems scaled to large sizes
  - Network Attached Storage (NAS) - NFS
  - Storage Area Networks (SAN)
    - Fibre Channel (FC)
    - Internet SCSI (iSCSI)
- Unit of Access
  - NAS – file level
  - SAN – block level
- Sharing
  - NAS – supports multiple clients
  - SAN – supports single client
    - If there are multiple clients, applications need to handle issues wrt concurrent accesses

So, essentially again to summarize, if you want to make storage systems scaled to larger sizes. Then you can have what is called network attached storage, or storage area networks; so the unit of access for network attached over this file level, for whereas the storage networks at the block level.

And it has got some implications otherwise also. For example, in sharing it turns out that network attached storage typically can support multiple clients, because it is a file protocol. So, usually there is a file system involved, because the file system is involved the file system can take care of concurrent accesses. For a storage area network typically, it supports a single client, if you really want to make it handle multiple clients, then there has to be something elsewhere either the application level somewhere else. That has to handle concurrent accesses, it can not it is basically has a simple idea that; I talked block things. And therefore, there is no other party in between.

So, the machine directly talks to this storage. So, this is are 2 differences, we will look at network attached to this. Later one thing I should mention is that both this network data storage and storage area networks start around the same time. For example, network attached to this started about the early 1980s and store the network started close to later 1980s. IBM cannot with their first model, and later to eliminate let us say there is some I am sorry, to standardize basically the command channel kinds of protocols were device. I can just to give some background.

(Refer Slide Time: 05:45)



Modern Storage Protocol Stack

- PHY mostly same: fibre (L_0)
- Upper Layer Protocol (ULP): SCSI also same! (L_n)
- $L_1..L_{n-1}$ determine the perf
  - SONET
  - Fibre Channel (FC)
  - 10GEth
  - Infiniband
- Both “Telephone” or “Internet” models in storage
  - Telephone model: guaranteed service, preallocation/reservation of BW, etc (FC/Infiniband)
  - Internet model: statistical multiplexing (10GEth)

It turns out that if you look at any protocols stack nowadays, inefficient high-speed storage protocol stack, the physical layer is almost always the same it is still fibre. Either it is fibre or copper, you can not use copper for long distances. So, if you have thinking about longer distance has still fibre. And this basically means that you are able to transmit right optical speeds or with basically at the speed of light. So, it is extremely fast, but that means that since you are exchanging photons. It means that when it comes to an electrical circuit, because finally your computer system is built of electrical circuits. That photon has to be converted into some kind of particles or whatever that interact with electrical systems like electrons. There has to be some converters between photons and electrons.

That is something to be done here. So, at the lower level it is always fibre, and then there are some converters from photons to the electrical signals. And the upper level protocol is also typical always SCSI, this SCSI has become standard; that means, that the lowest level and the topmost level both are standardized. So, in in between things are finally, going to determine exactly what is going to happen, because both are identical the to is identical the lower level also identical.

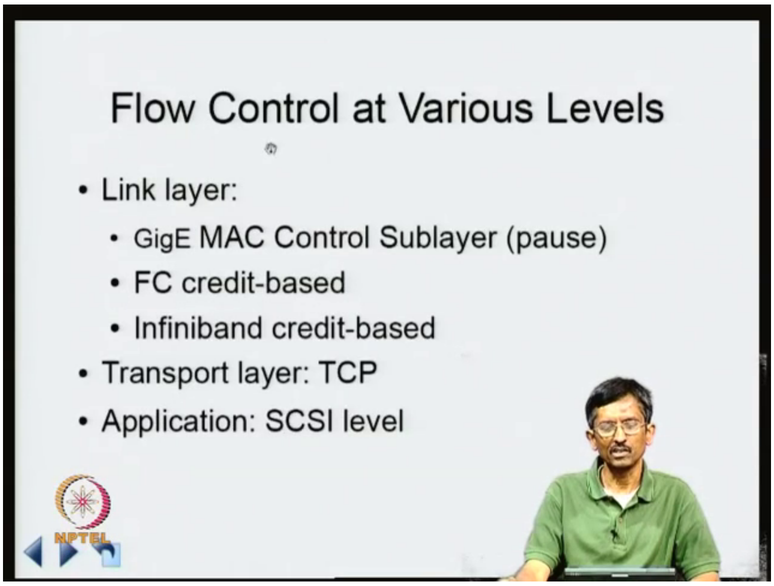
So, what are the in between things that actually finally, determine the price and let us say, the performance of the system. So, various protocols have been have been attempted in the past. For example, I will not go into it, but this has been attempt to it something called SONET now. Further channel has been quite heavily used in enterprise systems. You will see that with the speed so for gigabit ethernet etcetera and higher speeds like 10 gigabit ethernet, this also has become very attractive.

And there is also another model called infiniband which was devised as what is called a system interconnect; that means, connecting CPUs not necessarily storage. So, this has been also been attempted to be used to interconnect storage and multiple storage units along with to a the computing engines; so various possibilities. So, what it is interesting to also notice is that broadly there are 2 types of interconnect models when you might call the telephone model, mathematically called internet model. The telephone model typically you come up with what are called virtual circuits, and they usually typically give him guaranteed service. There is preallocation and reservation of bandwidth, and this is typically followed in models like fibre channel infinity band; this also an internet model, which basically attempts statistical multiplexing. For example: gigabyte ethernet or 10 gigabit ethernet.

So, the idea here is there is no motion of guaranteed service or preallocation resolution of bandwidth. Now the storage it turns out that these things are quite important, guaranteed service and a possible bandwidth etcetera, because if we are trying to do for example, backup and those kinds of things the heavy duty and you cannot really depend on statistical multiple into all the time. So, it is very important for you to do this; that is why a long time fibre channel and those kinds of protocols have been reasonably predominant. I think with the if 10 gigabit ethernet becomes low cost or you have something called metro ethernets, where you have ethernet at speeds of multiple gigabits within a city those things become possible widely possible.

Then again models based on gigabit ethernet can also become extremely popular. Now these they use the kinds of things that we do whether go for the telephone model, or internet model, you have to actually do lots of other things that go with these particular characteristics. For example, I will just briefly mention one thing the flow control.

(Refer Slide Time: 10:19)



Flow Control at Various Levels

- Link layer:
  - GigE MAC Control Sublayer (pause)
  - FC credit-based
  - Infiniband credit-based
- Transport layer: TCP
- Application: SCSI level

It turns out to the flow control of each of these kinds of models is different. And your performance or your controls that you need in your system depends on the kinds of local delays there. For example, you can do flow control for these network systems, network storage systems are the link layer itself, or you can do it is wit the higher levels or even in the application level. For example, it link layer.

If you look at gigabit ethernet; it does it had something called mac control sub layer. So, if it comes out that you can actually control the react with somebody sending information, by saying that the recipient part you can say, I cannot handle your speed it is pause. And it tells in terms some nanoseconds, it tells me it tells you for example, please pause for so many nanoseconds you sending it earlier. There is also something called credit based where you can send only so many credits of data.

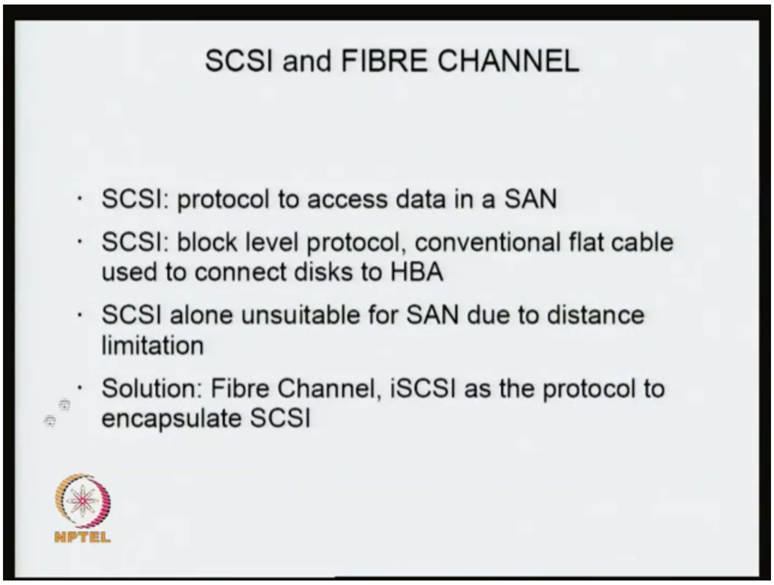
For example, you might decide beforehand, the initiator and the target can decide that are going to be 64 credits or 256 credits, and then every time an add comes, then your create increases, every time we send data you decrease your credit; that means, that you have determine the throughput is determined by the round trip. Because you will have some credits of beginning you send it, and other party has to act it. Then only we can then only we get one more credit. So, that we start with 250 it is true that you can send 256 packets in beginning, but after that you have to wait till we keep getting acts. Once you get each act then we can again you are allow to send them more packet.

Now if you have is credit based; that means, that you really have some reasonable control over how the network is going to behave, but the same time it creates some problems with respect to the throughput. We will discuss it later, why that is the case. We already mentioned that in the previous class, that in SCSI you have an ability to control, the flow basically you have this motion of ready to transfer and by which you can do it.

Sometimes it is also the case that, you encapsulate any of these protocols through TCP or some such thing because, you really want to access internet scale systems. And but across internet typically TCP is only thing that is available right now. So, means that you need to encapsulate whatever protocol you have either fiber channel or whatever through TCP. And it also does it is own flow control. Again, we will discuss this more extensively in the next class, then we talk about or TCP does it, and how storage protocols based on TCP IP for example, how they perform.

So, what I wanted to mention here is that your you can decide your physical layer, but still there will be some other layers that need to also be are in the picture, and they will also do certain different kinds of that is say management of the transfers. And these also are usually quite closely tied to the kind of physical model that is there.

(Refer Slide Time: 13:47)



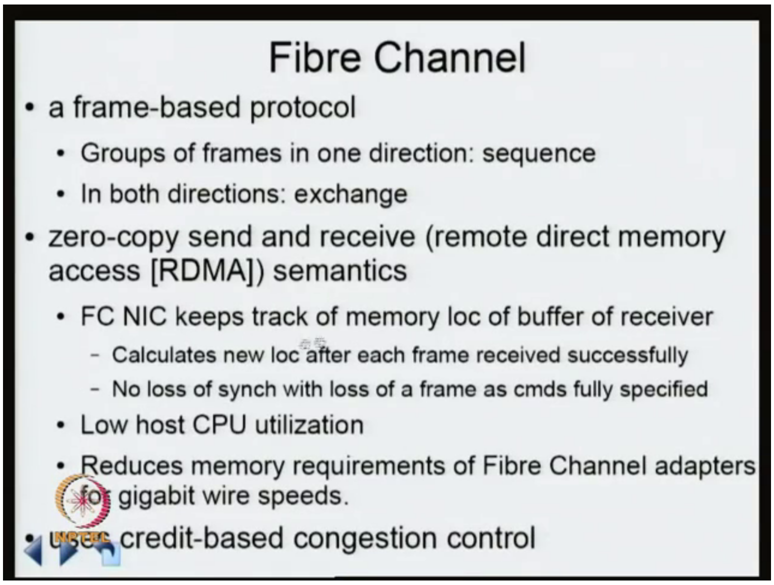
SCSI and FIBRE CHANNEL

- SCSI: protocol to access data in a SAN
- SCSI: block level protocol, conventional flat cable used to connect disks to HBA
- SCSI alone unsuitable for SAN due to distance limitation
- Solution: Fibre Channel, iSCSI as the protocol to encapsulate SCSI

So, let us take a look at how SCSI and fibre channel work together. So, in a storage area network, as I mentioned it is a block protocol; that means, it uses this SCSI protocol to access data. So, if you are in a single desktop kind of system which is a SCSI, basically that some kind of flat cable that is used to connect disc to the host bus adapter, the agent for the computing system, the storage agent.

But if you want to do it across longer distances, then SCSI is not suitable. Because it does not have it has some in the original even SCSI came around, they did not really have the notion of being able to interconnect longer distances. So, the solution that was adopted in later versions of the SCSI for example, SCSI 3 for example, as SCSI to was to use further channel or some other protocols as iSCSI which again we will discuss later as the protocol is encapsulates SCSI.

(Refer Slide Time: 14:54)



Fibre Channel

- a frame-based protocol
  - Groups of frames in one direction: sequence
  - In both directions: exchange
- zero-copy send and receive (remote direct memory access [RDMA]) semantics
  - FC NIC keeps track of memory loc of buffer of receiver
    - Calculates new loc after each frame received successfully
    - No loss of synch with loss of a frame as cmds fully specified
  - Low host CPU utilization
  - Reduces memory requirements of Fibre Channel adapters for gigabit wire speeds.
- uses credit-based congestion control

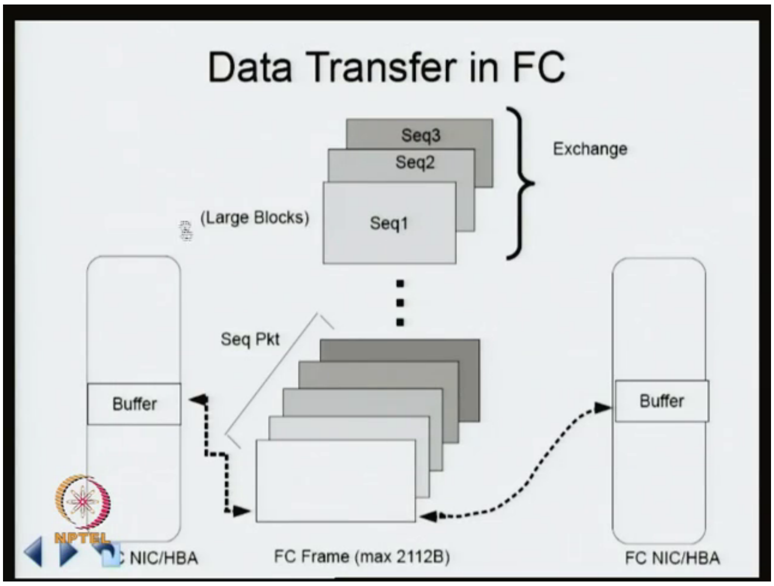
So, what is fibre channel it is a frame based protocol and groups of frames in one direction is called a sequence, in both directions it is called an exchange. What is interesting on fibre channel is it; it uses what is called 0 copy send and receive semantics sometimes called remote DMA transfers. What this means is that, the network interface cards from the fibre channel, after some negotiation they essentially keep track of the memory location of the buffer the receiver.

So, because they know the new location buffer the receiver, the DMA transfer that takes place across the network can directly deposit the transferred data directly into the memory of the recipient. So, you can essentially avoid or that extra copies. And after one frame is transferred, again the network interface card for the fibre channel it calculates a new location for the next frame, and it does it every single every time. So, therefore, it always can keep track on the communication. And even if there is a loss of synchronization because of loss of made a mistake. Even if the loss of frame it turns out there is no loss of synchronization.

Basically, because the comments that they exchanged, they essentially specify all the details beforehand hand fully. And because of there is it turns out there is no host CPU utilization. This is in contrast with for example, gigabit ethernet, which went through a lot of difficulties in the beginning. Because it turned out that there is a lot of copies going on, and you have to do a lot of re-engineering of the network stack so that the copy would not take place, and it was before the copies could be re-engineered away. The CPU was always involved in the copy and therefore, you should have a very high CPU utilization.

And for the same reason, because you are not doing any extra buffering it reduces memory requirements for fibre channel in what these speeds, it is a little earlier it uses rate based condition control we will look at it soon.

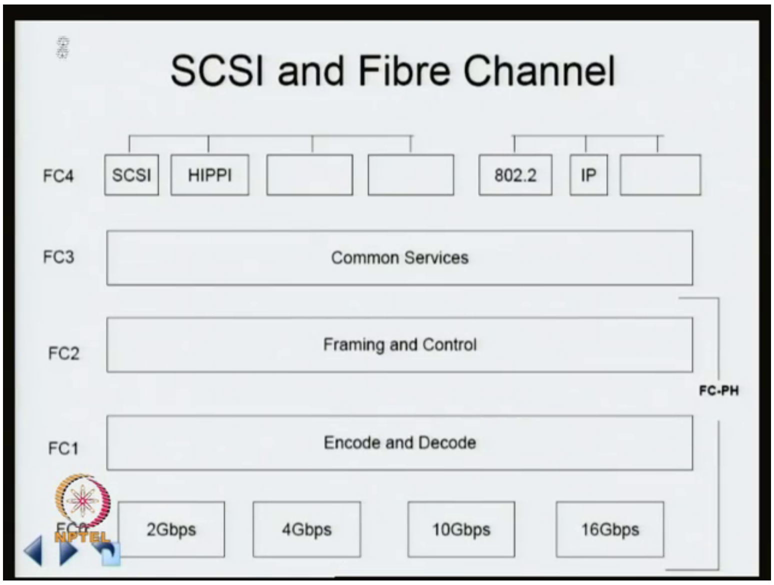
(Refer Slide Time: 17:17)



Data Transfer in FC

So, dramatically you can see the following there is a buffer here and then NIC, there is another before in the NIC, this is each of the frame, it turns out in fibre channel the maximum size of a frame is 2012 bytes and multiple frames is what is called a sequence, and multiple sequences is called an extreme. This is always in a single direction, this is bi directional. This is an important thing is that these particular NIC knows exactly the address of this buffer, it has to be deposited in the application therefore, it can actually directly deposited.

(Refer Slide Time: 17:59)



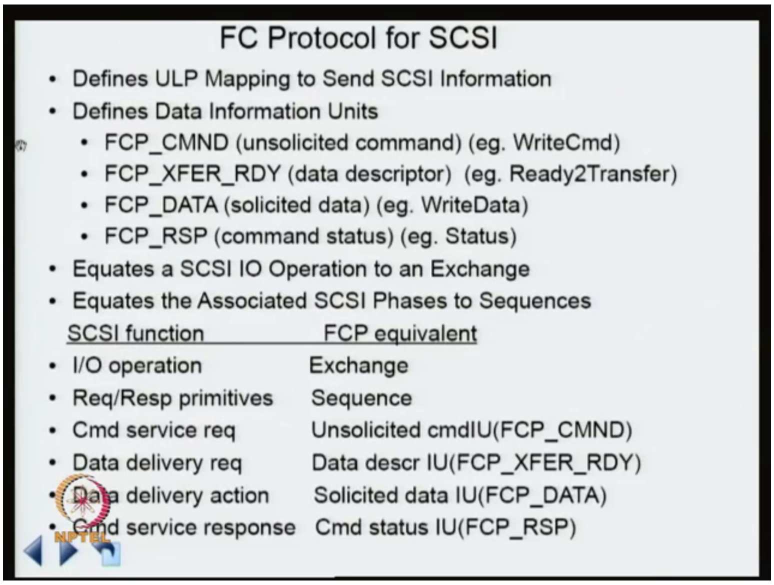
SCSI and Fibre Channel

Diagrammatically you will see the layering as follows. At the lowest level is basically the optical communication. In the beginning is started with is a 1 gigabit per second (Refer Time: 18:16) lower, but nowadays 10 gigabit per second and 16 gigabit per second fibre channels are is available. Because you are transmitting at very high speeds and reduce possibilities of errors, they do a lot of encoding and decoding. There is something called 8-bit 10 bit; that means, 8 bits of data is encoded into 10 bits.

Or you can also do 64-bit 66 bit. So, there is a lot of encoding it happens because can write UCE the chance of error because you transfers it very high speeds. And we will look at this framing and control the control flow as part of it especially. So, there is also common services. For example, it is possible that you can the encryption etcetera of course, this FC 3 typically no less is empty nobody got uses this typically. And top of it you have SCSI fibre channels started in the beginning with supporting some protocol hippi now other protocols, I am not going to go into them it can also support other protocols like IP. For example, so basically for us what is interesting is SCSI control flow, and the speeds on such rest of it I think are certainly important.

But I am not going to discuss them. These 3 things are basically what is called the physical layer starting from this to this let us see how we encoding and decode.

(Refer Slide Time: 19:44)



FC Protocol for SCSI

- Defines ULP Mapping to Send SCSI Information
- Defines Data Information Units
  - FCP_CMND (unsolicited command) (eg. WriteCmd)
  - FCP_XFER_RDY (data descriptor) (eg. Ready2Transfer)
  - FCP_DATA (solicited data) (eg. WriteData)
  - FCP_RSP (command status) (eg. Status)
- Equates a SCSI IO Operation to an Exchange
- Equates the Associated SCSI Phases to Sequences

SCSI function	FCP equivalent
• I/O operation	Exchange
• Req/Resp primitives	Sequence
• Cmd service req	Unsolicited cmdIU(FCP_CMND)
• Data delivery req	Data descr IU(FCP_XFER_RDY)
• Data delivery action	Solicited data IU(FCP_DATA)
• Cmd service response	Cmd status IU(FCP_RSP)

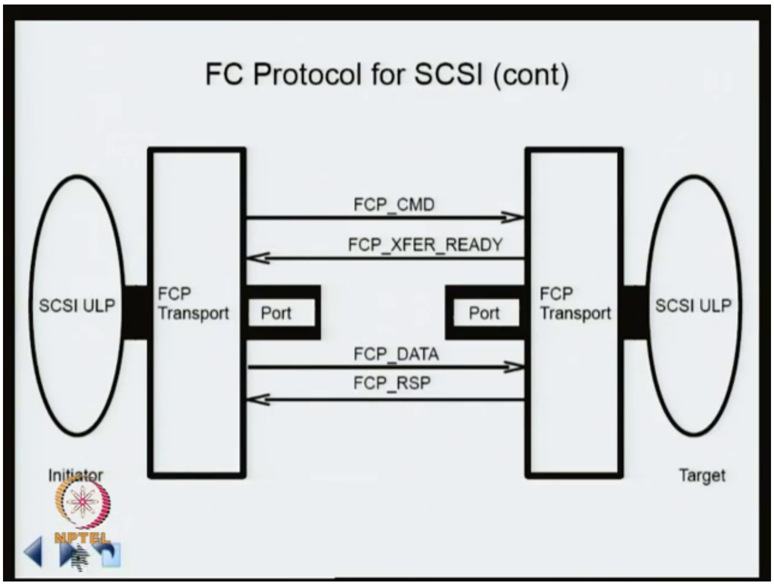
Let us how the mapping happens. We saw in the SCSI protocol for example, here is write commands that is what the initiator tries to do, then the target after all after having allocated buffers, it says I am ready to receive your packets. So, it sends a ready to transfer, then the initiator sensor data, and then once the data transfer is done, then the target sends back a status.

So, these things have to be mapped to the FC protocol data units. That basically what I am talking calling as a mapping. So, for example, there is a one to one correspondence. Write command becomes FCP command, it is unsolicited command, because the target it has to come it is the beginning part. So, it is unsolicited it comes out of the blue some cells. Then there is some people strictly transfer ID which basically means it is the target now we say that I am ready to accept data.

And again, the one that happens then the target again sends the data by encapsulating in FCP data protocol unit. And then again, the target sends back using FCP RSP other aspects of it is that; this is per protocol unit, you can essentially take a SCSI I O operation to be an exchange, I need associated SCSI phase is to be sequence right. So, this again that in a tabular form administered again. So, I O operation in this can exchange that means it is bi directional.

This is the request response primitives that is there could be multiple, let us say sequence units or the frames that are sent to with respect to 1, let us say and I O operation for example, the write and in turn the command is basically the SCSI command is mapped to this the data delivery is mapped this, but. So, IH the being ability to sorry, I made a mistake the data by if delivery, basically the target being ready to accept the data being sent is this transfer may be protocol unit. And the FCP data is what is the corresponding to the write data of the SCSI command, and the response FCP response is the corresponding response for this SCSI.

(Refer Slide Time: 22:39)



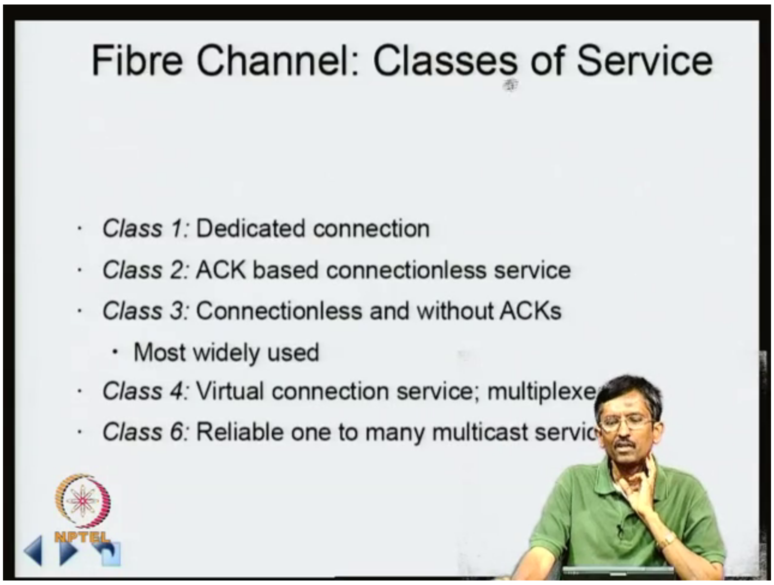
FC Protocol for SCSI (cont)

So, this is an example of a mapping for SCSI commands on 2 FCP protocol units. Again, this diagrammatic one way this command goes this way transfer it is comes back, the data goes this way response comes back. And these are the 2 initiative in targets which are at the SCSI upper level protocol it is a transport. And this transport can be let us say, can be arbitrarily complicated, because there can be arbitrary network sitting out here, and there is something called a port just like your IP address, you can have a port number, and there could be another port number and some writing also we start.

So, you are not getting too much detail all these things, but there is a whole amount of infrastructure here, which takes care of routing it takes kind of discovery for example, how do you know that there is some like a target out here. So, there is also issues about concur issues about security issues about can this initiated actually access this particular network. So, there is some motion on what is called login; that means, that this particular initiate we can actually access this thing. As I mentioned earlier science, are for a single client typically, unless you do something else application level to make sure concurrence is taken care of.

So, there could be multiple single clients which their applications are somewhat ensuring that they do not interfere, but that means that each client should be somehow be able to get access to this infrastructure; that means that there has to be some protocol in the beginning itself. And, there are some specific things called login procedures have to be done. Again, I am not going to go into too much in detail about these things.

(Refer Slide Time: 24:18)



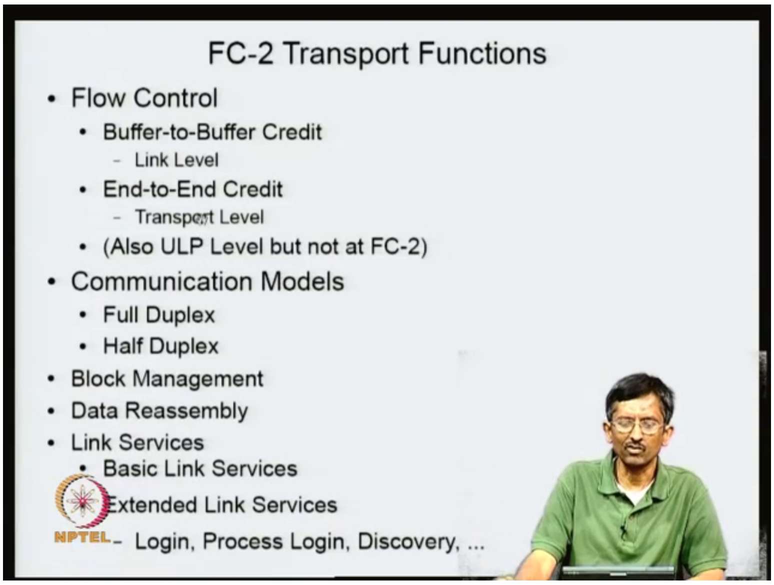
Fibre Channel: Classes of Service

- *Class 1:* Dedicated connection
- *Class 2:* ACK based connectionless service
- *Class 3:* Connectionless and without ACKs
  - Most widely used
- *Class 4:* Virtual connection service; multiplexe
- *Class 6:* Reliable one to many multicast servic

Now, let us take a look at why fibre channel has been used extensions to which, first one thing to notice about fibre channel is that they have quite a few characteristics of class of service.

It can have a what is called a dedicated connection, ACK based connectionless service, connection less without acks, and this class thing is the one which is might be used. So, you might be surprised to know why connectionless, and without ACKs is the one that is used it is used because how the channel is extremely good medium it is error rates are extremely low. Therefore, anybody trying to do is ACKs at one they are going to be slower than if you are use without ACKs. That is why the expectation is that since the error rate is so low if anything bad happens it will be handled at higher layers. So, there are few other classes also which I would not go into. Let us look at; as I mentioned if you look at the previous one we are talking about mostly connectionless user ACKs.

(Refer Slide Time: 25:39)



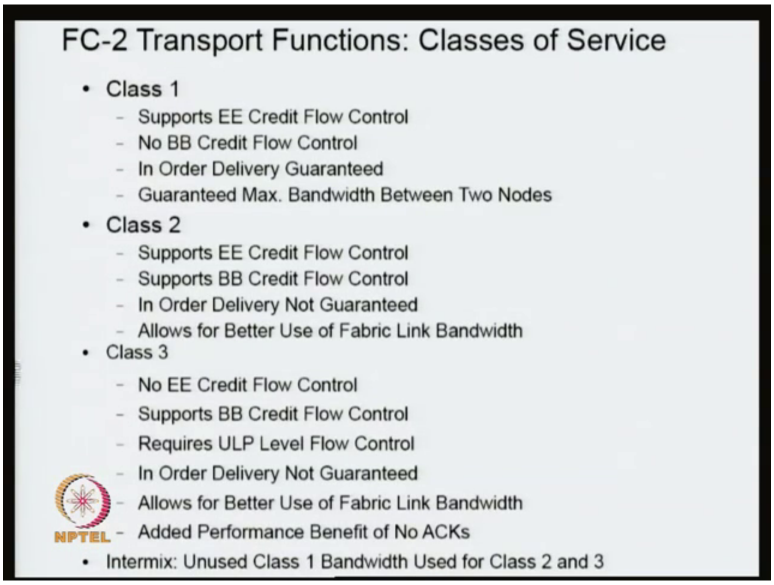
FC-2 Transport Functions

- Flow Control
  - Buffer-to-Buffer Credit
    - Link Level
  - End-to-End Credit
    - Transport Level
  - (Also ULP Level but not at FC-2)
- Communication Models
  - Full Duplex
  - Half Duplex
- Block Management
- Data Reassembly
- Link Services
  - Basic Link Services
  - Extended Link Services
    - Login, Process Login, Discovery, ...

So, because there were AKSs; that means that there is no inherent way in which you can avoid buffer overruns. So, that means, that there has to be some flow control. Now there are 2 possibilities. You can do flow control of the link to link level or you can do the transport level, end to end. What I am talking about is; if you have a big fibre channel network, you can have what is called link level flow control; that means, that you only for every jump from one link layer device to the next link layer device, you have flow control. That is, it is happening in small segments, one segment at a time, but it is end to end it is across the whole set of devices, or end between routers whatever rest there. So, of course, they also defined other models like full duplex and half duplex.

And as I mentioned earlier you need things like login discovery of this kind of capabilities, and this this particular layer of fibre channel FC 2 and solve these issues.

(Refer Slide Time: 26:58)



FC-2 Transport Functions: Classes of Service

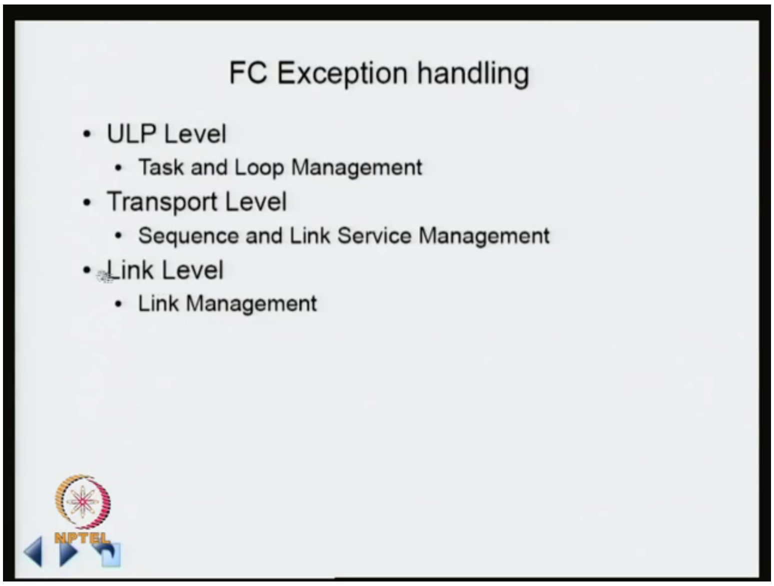
- Class 1
  - Supports EE Credit Flow Control
  - No BB Credit Flow Control
  - In Order Delivery Guaranteed
  - Guaranteed Max. Bandwidth Between Two Nodes
- Class 2
  - Supports EE Credit Flow Control
  - Supports BB Credit Flow Control
  - In Order Delivery Not Guaranteed
  - Allows for Better Use of Fabric Link Bandwidth
- Class 3
  - No EE Credit Flow Control
  - Supports BB Credit Flow Control
  - Requires ULP Level Flow Control
  - In Order Delivery Not Guaranteed
  - Allows for Better Use of Fabric Link Bandwidth
  - Added Performance Benefit of No ACKs
- Intermix: Unused Class 1 Bandwidth Used for Class 2 and 3

So, we will go more or bit into these details. So, in class 1 as I mentioned earlier; it supports end to end credit flow control. There is no buffer to buffer credit flow control in order the deliver is guaranteed guarantied max bandwidth approval. So, this is the one which it takes a lot of trouble to make sure that, your data that is sent is in the same order that was keep to it send to it. And also, is making sure that across the whole fibre channel network. There is credit flow control, the end to end credit flow control. And so, this is going to be the most difficult one to engineer, for one point of view of the fibre channel network it requires a lot of infrastructure lot of it is have here whereas, the one which is common.

As I mentioned is this one there is no end to end credit flow control, there is only between one there is a device fibre channel device on get work; that is, it and it is upper level protocols have a take care of the actual flow control across end to end that is what they have to do in addition no in order to get a delivery is guaranteed so that something else not to worry about. Basically, reason why people have the storage systems and fibre channel typically cluster is because it allows better use of the fabric link bandwidth, because you do not have any ACKs etcetera. There is also another one called class 2, it basically is has end to end credit flow. But does not support in order delivery; so, but they the most widely used is this one. So, we will not discuss all the ones.

There are statistical model also available, if you want to use it. If there is some class one bandwidth is not used you can use it for class 2 and class 3, this is the some kind of statistical multiplexing model.

(Refer Slide Time: 29:05)



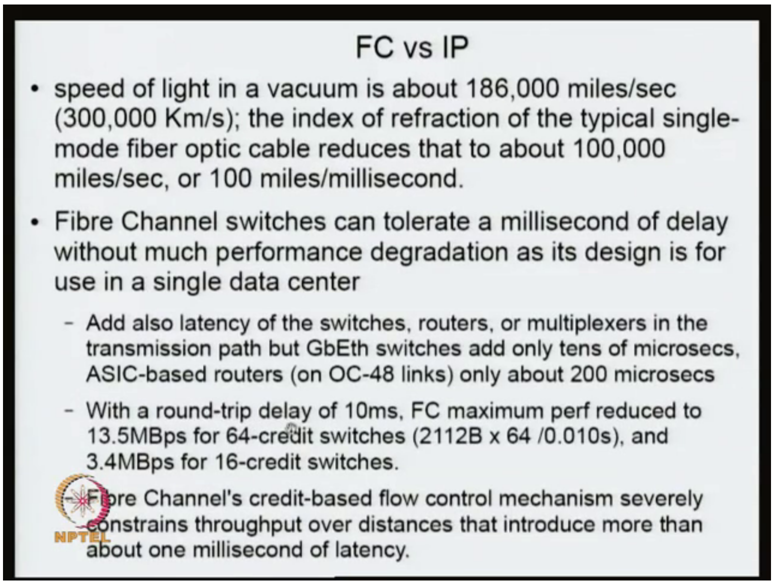
FC Exception handling

- ULP Level
  - Task and Loop Management
- Transport Level
  - Sequence and Link Service Management
- Link Level
  - Link Management

Again, the exception handling can happen at multiple levels. So, for example, if there is a frame that is dropped, it might be done at a link level, it might be or it can be done at slightly higher levels. And there is also some other aspects, basically when the network comes online first, the very first time the discovery aspects have to be also taken into account. And some of these issues also have to be done at multiple levels; so here when we are talking about what happens when some packet that is transmitted gets lost.

As I mentioned earlier this is very rare, but it does happen. And so, they have to be handled which are using of course, the class one kind of models. Then the each other proper channel devices that are there in the in the fabric, they have to handle it in the hardware itself. And whereas, we use class 3, then that some of the exception handling is done by some of other at the protocol level.

(Refer Slide Time: 30:23)



FC vs IP

- speed of light in a vacuum is about 186,000 miles/sec (300,000 Km/s); the index of refraction of the typical single-mode fiber optic cable reduces that to about 100,000 miles/sec, or 100 miles/millisecond.
- Fibre Channel switches can tolerate a millisecond of delay without much performance degradation as its design is for use in a single data center
  - Add also latency of the switches, routers, or multiplexers in the transmission path but GbEth switches add only tens of microsecs, ASIC-based routers (on OC-48 links) only about 200 microsecs
  - With a round-trip delay of 10ms, FC maximum perf reduced to 13.5MBps for 64-credit switches (2112B x 64 /0.010s), and 3.4MBps for 16-credit switches.
  - Fibre Channel's credit-based flow control mechanism severely constrains throughput over distances that introduce more than about one millisecond of latency.

Now, let us see what is the issue with respect to credit, credit based flow control. Now as I mentioned earlier the fibre channel the medium of transmitting the data is fibre; which is basically optical fibre. And as you know the speed of light is about 186000 miles per second, because of other issues like index of refraction etcetera.

The actual speed is about 100 thousand miles per second; that is about 100 miles per millisecond. So, if you have 100 miles per millisecond, now we can start thinking how it can actually impact your design if you are expecting some kind of round trip delays because of acknowledgments. So, essentially the design of fibre channel has been to make the assumption that, you can tolerate about a millisecond of delay. Because typically has been designed for a single large data center. So, you are going to be within 100 miles or so definitely. So, about a millisecond it is because we can do the round trip within a millisecond. So, even if you add other I things latency of the switches, routers etcetera, right instead of it at the most is about a millisecond it is not ever going to tens of microseconds. It is not too much.

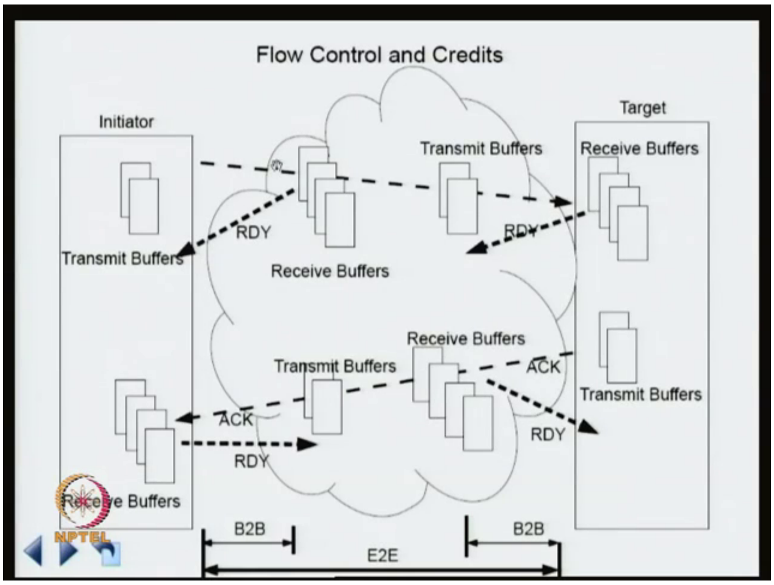
So, if you assume that it is a millisecond or at the maximum 10 milliseconds, you will notice that a performance of fibre channel maxes out. You can see this for example, suppose you have what is called 64 credit switches. So, you have how much channels switches, routers inert call it at the link level link air level, so much of 64 credit switches. We are able to send one packet at a time, but you have to get add back right. So, you are able to get, you can send about 2112 bytes, but if you are assume, because of additional delays, let us say it goes this can be as much as 10 milliseconds, essentially what it can base you can send 2112 bytes into 64 by 10 milliseconds.

So, that is like it 13.5, if you make it to one millisecond, you can be 134 megabyte per second. So, if you are of course, lower than that you can be slightly better, but you can have the most be about 200 megabyte per second, because of this which was something smaller 16 credit switches in this example which your performance also decreases that much.

So, in a sense your round trip delay has a very big important impact on the performance. It is true for TCP also, but it turns out that because of the way the flow control is going TCP, you can by using bigger buffers, you can essentially do much higher throughput stands. For example, people have done multiple gigabit per second transmission TCP ethernet (Refer Time: 33:43) across wide area network, whereas in fibre channel it is not possible, because the number of credits determines how much you can send.

And then you have to keep on waiting for the ACKs to come back when the ACKs comes back, then you get a chance to send one more packet. So, this is one important issue that it is there in fibre channel. And so, the fibre channel was designed in the mid 90's when the kind of speeds were talking about was rather high that time. Therefore, it was not an issue, but nowadays with the 10 gigabit ethernet, and those kind of speeds also becoming common and fibre being extensively used the design aspect with respect to credit based flow control has been some slightly problematic.

(Refer Slide Time: 34:38)

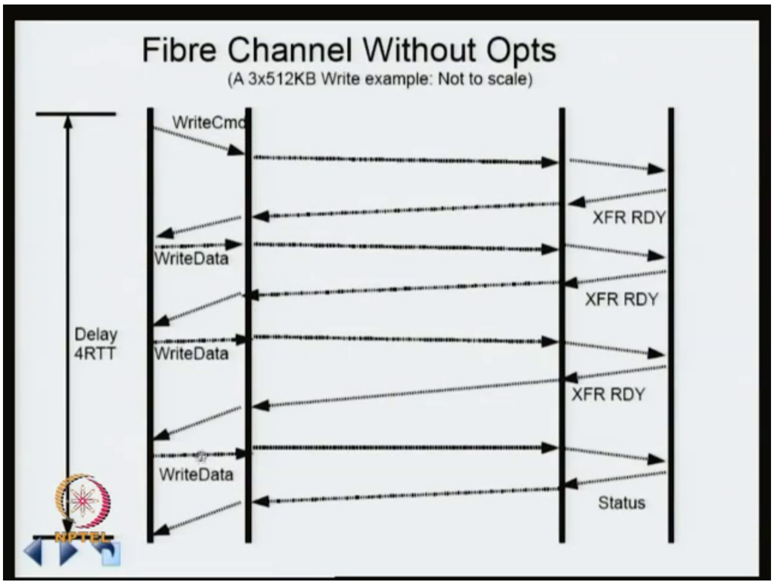


So, again to the diagrammatically you can see the following, this is a initiated this is a target, this is the transmit buffer.

Basically, the application has written some data into here, and then you want to send that you come back to this target. As I mentioned earlier the target has to send a really signal back so that it can do it and this signal has to percolate backwards. So, suppose there is someone are holding multiple fibre channel switches are in between lot of switches are in between; that means, that there is everybody has to because they are all although all of these entities are talking fibre channel protocol SCSI based fibre channel right. So, that means, that every time they receive that (Refer Time: 35:29) then only this black and send something.

So, everybody also has this institution; that means that if I want to send something to here, one packet then it is a sum of all the delays; so in a sense the number of switches that are there in between and their delays; that are there for transmitting across this whole section of the switches. And if there could be some other a network that is a deal is also a if you are really trying to encapsulate fibre channel through some other protocols. Then all these delays are going to be involved before I can get one, packet back saying I used you can now send one more packet, right. And essentially is this what I was a problem, and I can have been illustrated here by saying that for example, the buffer to buffer is between these 2, and end to end is between this end and this end. So, all the delays essentially matter in fibre channel.

(Refer Slide Time: 36:35)



Fibre Channel Without Opts

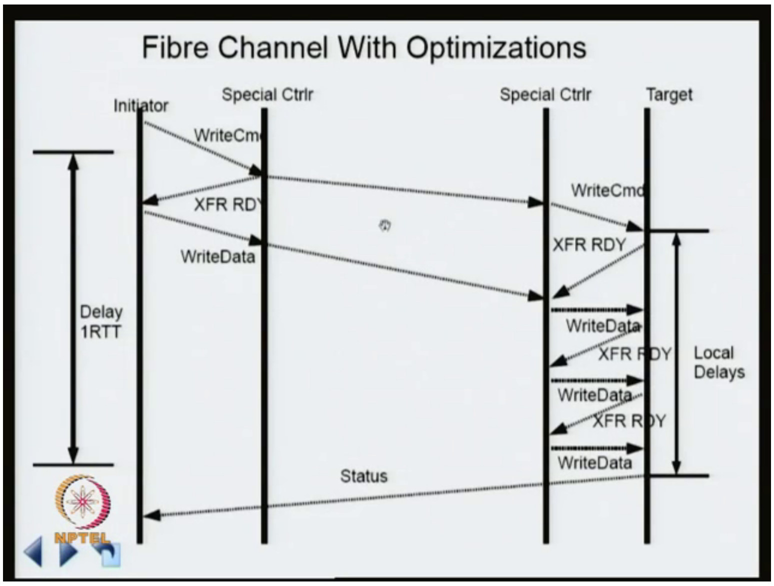
(A 3x512KB Write example: Not to scale)

Suppose you want to do something better than this. Suppose I want to write 512 kilobytes, and this is the initiator this is the target. And basically, I am going to this is going to be my first level, fibre channel switch out here. This is my on the other side is going to be my first level, fibre channel switch. So, this is you can call it the initiator, as you can call the entry into the fibre channel network. This is the exit from the fibre channel network, and this is the target. So, this can be a disc for example, and this can be an application HP for example. So, what we are talking about is I do a right, then it has to traverse this fibre channel network. It goes all the way up to this point is read command, then the command once it received by the target. It will allocates at some buffers and saying I am ready to transfer.

Again, has to calculate back all the way up to this point. Again, once the write command this the transfer ID commands commit, then if you actually can send the data suppose I am doing a 3 into 512 kilobyte transfer. So, I sent 512 kilobytes it has to come back all the way. So, these have one 512 kilobyte, and this was second kilobyte is for the third one. And then the status will come back. So, you can notice that there is going to be these delays, 1 2 3 essentially some kind of 4 round trips are involved. Now if you are going to incorporate some wide area network in between. And someone is able to encapsulate the fibre channel protocol inside it, then all these delays anymore.

So, the throughput is going to be quite small it is going to be basically, 512 with a 1-point megabyte divided by that 4 times round trip require, it is going to be very speed. And this basically because fibre channel has this notion of credit based flow control you actually have to get the credits, before you can see anything. In the beginning you can send something, but again once you have sent it you have to get back the ACKs before we can we you have to be the (Refer Time: 39:02) are the replenished, before you can send it back. So, if this is the case are you limited to this? Luckily as I mentioned earlier, in it SCSI protocol, you can introduce intermediate agents, right. You can also introduce intimate agents here also.

(Refer Slide Time: 39:22)


Fibre Channel With Optimizations

And recon to the forum I have the initiator, and then I basically have a special controller. I have a special controller here also which actually talks to the fibre channel network. So, it is a initiated is a target, I have a special controller which is sitting in front of me, and it takes care of the transfer from this to this through the some other protocol.

Let us say some highly optimized protocol that can do wider network transfers efficiently the receiver. So, what I am going to do is; where the initiator sends a write command, this pressure controller traps it. And it pretends also it is the target. And so, it sends back transferring and since initiator and special event and nearby right. So, this command can come back immediately, and the data can be blasted with the higher speeds possible within these two entities, and there is some special protocol from going from here to here. That can essentially take all the data and blast it back here. These are very some, somehow what happens is that this command goes out here, and then the transfer ID happens, and once this data has come here right.

This pressure controller has already got the transfer ID from here. So, what is happening out here? Basically, what we are talking about is; once this write command comes here because it is going to both the places. This write command is sent to the target, and they tried the target essentially says now I am ready. And this is going to be local. Let us assume like this pressure controller in the target are collocated somewhere. Because they are co located, the minute the special controller gets it is write command from this distance through wide area network of whatever, this write command essentially is sent is received quite quickly. And this target can respond immediately as quickly as it can.

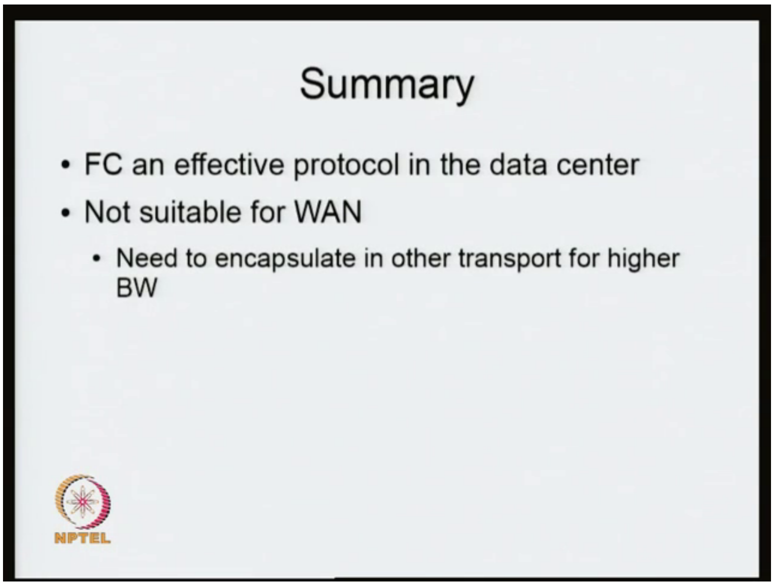
And once this transfer it is done, then this data that is sent in one in a bulk fashion as fast as you can not imagine is possible across whatever special protocols are using, this is data can come here. And then this data is parcelled out segmented and parcelled out in this target as ended. And because this particular special control and target are co located; therefore, we do not have any serious problems of trying to navigate this wide area network round tripped that they will not be coming in a picture. So, for example, what will happen is that; this will take place very quickly, and then this status will send there. I am I should only that this particular picture has not been drawn to scale. So, this time units are quite small.

So, then basically what is happening is that there is one can going to be only 1 round trip delay, we send it out here. And then you get back this there is a status comes back. And there the throughput can be that much aware. So, basically what we are saying is that, because you can virtualize the SCSI endpoints, and therefore, you can in incorporate multiple devices in between. Even if sometimes some problems to the protocol sometimes you can work around it, but it requires some special engineering to make this happen.

And again, we will next class what we will do is we will study how TCP IP does this inflate different way it is not going to, it is not a create base flow control model. And so, that requires some other types of engineering, and it was thought that because TCP IP, it can suffer losses. Because it can suffer losses it requires a is different kind of flow control and that creates on the complications of 0 copy. So, it turns out that I can do 0 copy equal 12, because the protocol actually keeps track of the memory it does it consisted a remote DNA, but TCP had time doing 0 copy kinds of transfers it is possible to create, but it is not so easy. And so, there are both plus and minus points in these cases. So, in TCP essentially you need higher performance devices so that you can do all the TCP protocol processing whereas, in the fibre channel, you may not need that high-end system.

So, what are the protocols that you choose for trace typically there are lot of implications across the layers and that is something had to worry about. And then we are not talked about other aspect was security another respects all these things also have very big impact on the performance of system.

(Refer Slide Time: 44:51)



Summary

- FC an effective protocol in the data center
- Not suitable for WAN
  - Need to encapsulate in other transport for higher BW

So, in summary basically it turns out fibre channel is a ineffective protocol in the data center, but as I discussed before it is not that suitable for other network, you need to do something about encapsulating this particular protocol per have bandwidth. And so, it turns out that fibre channel nowadays is also going through some other transformations, basically the idea is to see how you can encapsulate fibre channel in 10 gig ethernet. And it is called fibre channel over ethernet fcoe; that is also something that is taking place in a wide scale.

I think this I am going to stop todays class. In the next class I will look at suppose I use internet as a way to deliver SCSI, all that can be done.