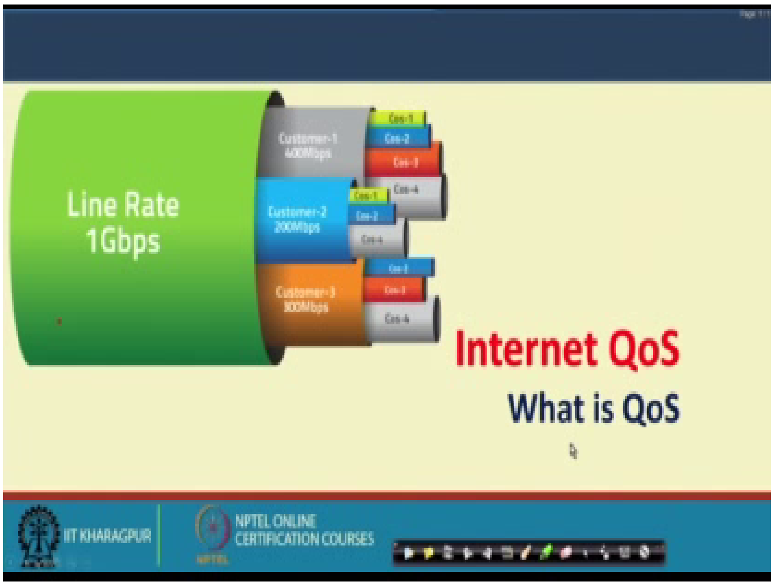
Computer Networks and Internet Protocol
Prof. Sandip Chakraborty
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 31
Internet Qos – I (What is Qos)

Welcome back to the course on Computer Networks and Internet Protocols. So, today we will are we are going to start a new topic which we call a Extra Network Quality of Service.

(Refer Slide Time: 00:28)



So, in this introductory lecture of the internet quality of service we will first look into what is quality of service, and what do we require quality of service. Indeed in today's network quality of service is an very important topic because, nowadays we are going to use different kind of multimedia data and multimedia traffic like we use mobile phones, and in mobile phones we have thousands of minuet different apps or applications which are running there. And we do different kind of multimedia based applications like this YouTube app like Facebook life then this Hotstar, Netflix all these different type of applications which use multimedia streaming of data.

So, this streaming data the nature of streaming data is much different compared to the nature of normal file transfer. Why? because in case of normal file transfer you just need to transfer the bytes at the bits at the other end and in the other end you can just combine all the bits all together whenever you are receiving, or whether you have received all the bits and then you can reconstruct the file.

But in case of multimedia streaming say just think of the example, of YouTube streaming say for example, when you are watching this video during that time the data is coming over the internet. So, and you are playing it simultaneously, so do not think about the offline video download that link that is a different ballgame that is similar to that file transfer, you are just transferring the video file and playing it offline, but whenever you are playing it online over say YouTube.

During that time from the YouTube server to your client which is running at your browser the video data is getting transmitted continuously, and by the time this video data is getting transmitted during the same time you are playing the video. And, many of the time I think you have observed that whenever your network quality is poor you do not have sufficient bandwidth. You may observe for degradation of the video quality suddenly you will see that, the video quality is getting dropped or sometime you may observe some events what we call as the re-buffering. So, what happens that the video got stuck and you keep on seeing the circular thing that is rounding about for trying to download the video data.

Now, for this kind of video transmission we need to maintain certain level of quality of service, quality of service in the sense that what the other end or the client side is expecting from the internet. So, here in case of this YouTube video streaming the client side the YouTube client it is expecting a continuous stream of videos such that it can render the videos directly on the player and play the video without having this kind of quality drop or re-buffering. So, that is why or to ensure such kind of application quality we need to provide certain special services at the internet level.

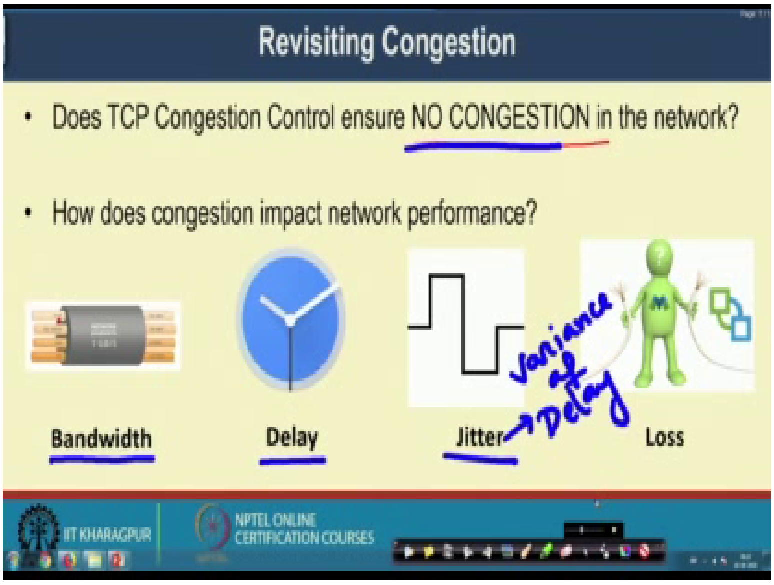
Now, remember that providing this kind of special service over the internet requires something like a dedicated line or a dedicated resource that is given to you. So, do not get confused over the fact that while you are observing the videos over YouTube why you are not able to see properly whenever your video is getting bandwidth, whenever your network bandwidth is getting dropped or whenever you see some quality degradation. That is because your network bandwidth on which you are currently subscribed that does not provide that kind of service level agreement, that you require to get the network quality of service.

Now, today we will look into briefly that what type of quality of service parameters are required for this kind of smooth video streaming, or to improve the quality of service or in other words sometimes we call it as quality of experience. The, what term quality of experience actually indicates the users perceived quality of particular applications. For example, how good you are thinking about the video streaming service or the quality of the video that you are observing in YouTube.

So, on that the network need to provide certain services. So, that you can look into those videos more smoothly and; obviously, for that you have to pay more to your network service provider, and you have to go for an agreement with your network service provider that you require these particular types of video quality of service. And the network service provider should provide you that level of quality of service. So, this gives you a kind of broad overview of what quality of service actually means. Now let us go into the details of different parameters, which we need to play with while providing quality of service to the end users from the network perspective and how we can insure it over a large internet.

Indeed in the beginning of the lecture let me give you an information that, in today's internet providing quality of services indeed a very challenging task considering the scale of the network. So, that is why we actually apply certain kind of approximation there and you will never get a perfect quality of service unless you have a dedicated least line connection, where you are getting the entire bandwidth. But, for a general internet for a general data transfer over the internet we try to provide certain level of quality of service. So, let us look into the details of this particular topic.

(Refer Slide Time: 06:29)



So, to start with our journey on quality of services, we start with this specific question that, does this recipe congestion control ensures no congestion in the network? So, our objective is to see that whether this TCP congestion control algorithm ensures that there will be no congestion in the internet unfortunately this if you look into the TCP congestion control algorithm in details what TCP does TCP tries to avoid congestion.

So, in other words what TCP does whenever TCP detects that there is certain congestion in the internet during that time, TCP simply finds out the reason for congestion; that means, whether you have you have exceeded the slow start phase and you are observing certain kind of packet loss due to congestion. And TCP simply reduces its rate and because of that what TCP does or the TCP congestion control behavior. It is like that once the congestion happens in the network TCP detects it and then only TCP responses to the congestion by reducing the sending rate of the sender.

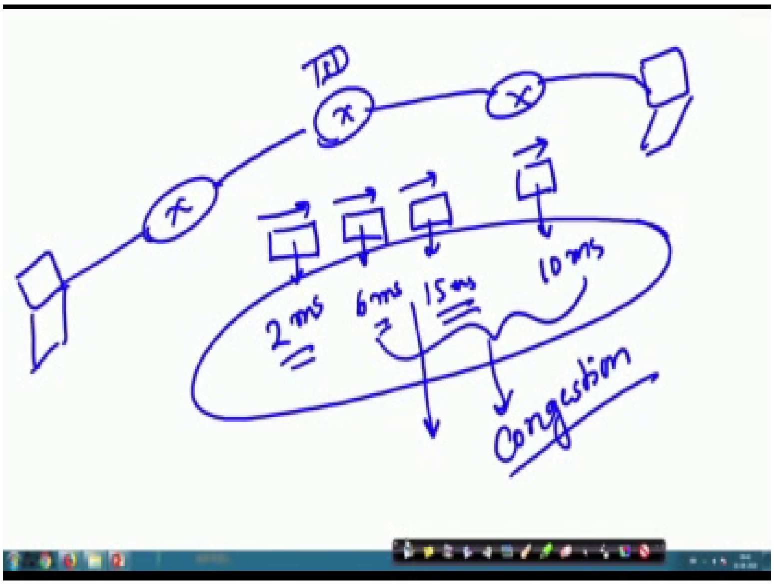
So, that way you are actually never avoiding the congestion in the network, rather what you are trying to do you are trying to have congestion in the network and then once congestion happens in the network, you are trying to coming out of that congestion scenario. So, TCP congestion control algorithm does not ensure that congestion will never happen in the network rather it works in a different way, like it first finds out whether there is certain congestion in the network by observing the packet loss. And if there is congestion in the network then only it acts or it updates its sending rate, to avoid the congestion from the network.

Now, the question comes that if there is still a possibility of congestion in the network how this congestion impact the network performance. So, we look into four specific parameters on top that impact the network performance, when there are congestion in the network the first parameter is the network Bandwidth. Obviously, whenever the congestion is there you are expecting less Bandwidth from the network, because the same (Refer Time: 08:58) Bandwidth is getting shared by multiple applications.

The second parameter is the Delay. We will look into the different components of Delaying details and we will see that whenever, the congestion is there in the network during that time that delay gets impacted like if there is a congestion; that means, their packet need to wait for more amount of time in the packet buffer. Because the packet needs to wait for more amount of time in the packet buffer you will experience that the packet will be transmitted, with a higher delay because the queuing delay the time to wait inside the queues at intermediate routers that get gets increased.

The third parameter which we are going to talk about we call it Jitter. So, possibly you have heard about this term bandwidth and delay earlier, but jitter may be a new term for you to so, but in respective of quality of service jitter is a very important parameter. So, what is Jitter? So, jitter is basically the variance of delay. So, we call the variance of delay as the jitter. So, what do you mean by variance of delay? So, variance of delay means assume there are two packets; which are coming from the or which are going through the network. So, whenever there are two packets which are going to the network.

(Refer Slide Time: 10:24)



So, assume that these are the sequences of packets, which are being transferred over the network. Individual packets will have individual delay. So, say this first packet has a delay of 10 millisecond; that means, you have a source here you have a destination here and from this source to destination there are multiple routers, and over that you are transferring this packets. Now, whenever you are transferring this packet during that time these different packets may have different delay. So, said this packet has 10 millisecond delay this packet has say 15 millisecond delay then this packet has say 6 millisecond delay and this packet expect some say 2 millisecond delay.

Now, why this there is a variance here? So, what you can see from here that the variance of delay among those four packets is significant like it the maximum delay is something like 15 millisecond and the minimum delay is something like 2 millisecond and, why there can be variance because different packets may experience different level of congestion at the intermediate routers.

So, it may happen that whenever this first two or three packets who are getting transmitted, during that time you have the congestion in the network and as a result due to this congestion the packets have experienced more waiting time inside the packet queues at this intermediate routers. But this last packet it came out of the congestion and because it was able to came out of the congestion true the TCP congestion control algorithm which is running at the transport layer.

So, maybe at this point of time TCP has experienced our TCP has found out a packet loss, and whenever TCP has found out a packet loss TCP has reduced it rate and as the TCP has reduced it rate the congestion will slowly get out of the network, and you will see less delay, but this delay variation among multiple packets that actually impact the performance of the quality of video streaming services.

Now, the question comes that how it impacts the quality of video streaming services. Just think of the scenario when you are observing a video in a YouTube player and, during that time every data packet is coming one after another that is getting buffered inside the client buffer that you have at the YouTube player and then the YouTube player is rendering that video.

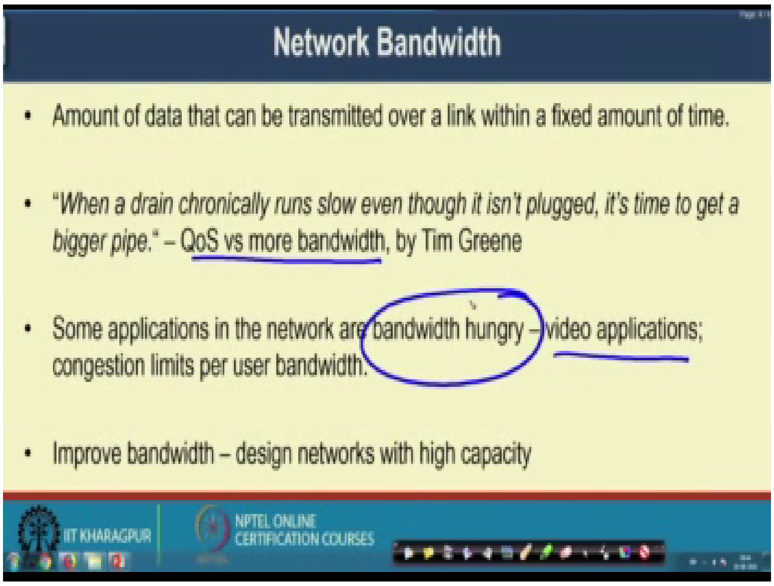
Now, if different packets has different delay then, the problem is comes for the live streaming. So, for the live streaming what happens whatever I am doing here or whatever I am say recording here that is getting streamed immediately. So, currently you are observing something called a buffer streaming because, the video has been recorded and you are getting the video from the video server, but if there is a live streaming session just think of an NPTEL live streaming session which is going on.

Now, in that live streaming session whenever some video is getting recorded immediately that video is getting transferred. And that video is getting played in the YouTube player where you are being observed. Now if different data packets has different amount of delay what will happen some packets will reach there at the client site and the client will play it then the next packet comes. So, the next packet has see a higher delay. So, the client wait for some amount of time and place it again that third packet say it comes very fast. So, the client plays it immediately the fourth packet gets a higher delay, and there is a delay in playing that particular video and as a result what happens that at the client site you will see lots of this kind of jerkiness in quality.

So, the quality some time it good bit becomes good some time there is a it is waiting for the next video frame, but it is not getting that after that it is again immediately getting the frame, then again there is a delay. So, you will see a lots of jerkiness in the quality level because the data which is coming from the server it is not coming at a constant bitrate it is coming in a variable bitrate and that is too way at a high variation. So, that is why jitter is indeed important for ensuring quality of service. So, that we do not see lots of up and downs in video quality.

So, in that particular context or that particular per perspective jitter is an important parameter for consideration and the final parameter is the packet loss. So, it triggers that how much packet loss the a particular video streaming service or a particular quality of service associated application can sustain. So, certain application they can sustain level of quality of service, but for some other application say for the voice application loss system very important parameters. So, if you have a significant amount of packet loss you will not be able to hear the voice correctly if you are transferring the voice over the normal IP network data. Now, we will look into all these individual parameters in little details so.

(Refer Slide Time: 15:55)

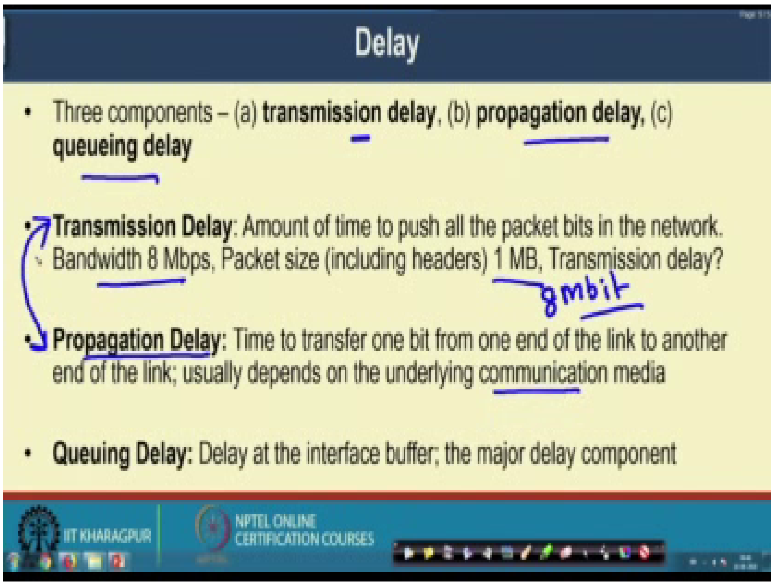


So first let us look into the network bandwidth. So, by definition as you know that the network bandwidth is the amount of data that can be transmitted over a link within a fixed amount of time. Now network bandwidth is something like we do not have much control over that. So, I actually liked very much this line by Tim Greene from a book called QoS versus more bandwidth an article there he mentions that when drain chronically runs low even though it is not plugged it is time to get a bigger pipe. So, what does it mean that if your connection does not have sufficient bandwidth, but you require more bandwidth then you will never be able to manage with the existing pipe or existing in line that you have data to go for a subscription of a higher bandwidth line.

So, that is regarding bandwidth if you have a lower capacity pipe or lower capacity network channel you will always experience bandwidth if you are going to going to say deliver or if you are going to use more amount of bandwidth. Say for example, you have a think of you have a 1 Mbps list line and with that 1 Mbps list line you are always trying to observe high definition video.

Now, high definition video if you are trying to observe it every time 1 Mbps may not be sufficient for you. And because of that you require more bandwidth you need to go from 1 Mbps to 8 Mbps or even more bandwidth. Now some applications in the networker bandwidth hungry such as, these video applications and congestion limits per user bandwidth and that is why we need to design networks with high capacity to improve the bandwidth. So, if you think that your application or your network is going to run applications like video applications which are kind of bandwidth hungry applications, then you purchase network lines with more bandwidth ok.

(Refer Slide Time: 18:04)



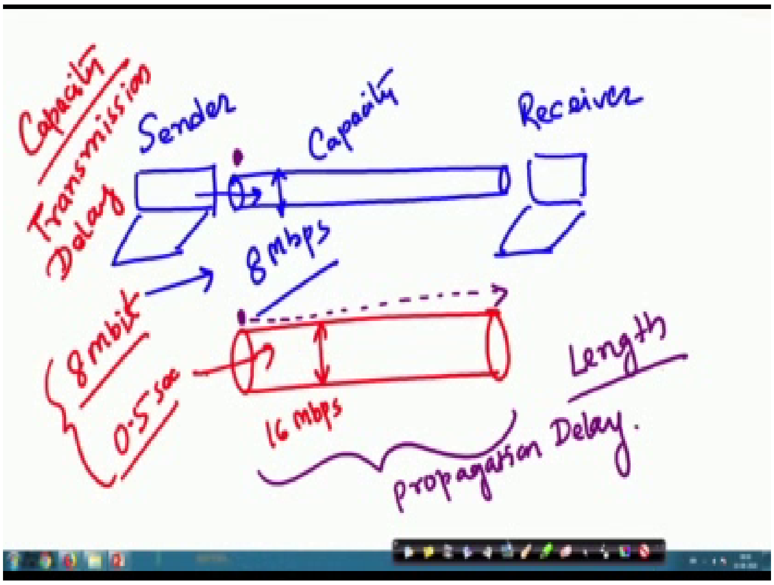
Now, coming to delay in network there are three different types of delay or you say that; there are three components of delay, the first one is the Transmission delay then we have the Propagation delay and the third one is the Queuing delay.

Now, what is a transmission delay? The transmission delay is the amount of time to push all the packet bits in the network. So, it is like that it actually depends on the capacity of your network. So, for example, if you network bandwidth is 8 Mbps and your packet size including the packet headers is 1 megabyte then what is the transmission delay. So, 1 megabyte packet means it is 8 megabit packet and you are transferring data at a rate of 8 Mbps. So, your network capacity is 8 Mbps.

Now if you are the 8 megabit bit of packet then to transfer that 8 megabit data over 8 megabit per second line you require exactly 1 second. So, that is the transmission delay that we have over the network. So, it depends on the amount of bandwidth that is given to you and based on the amount of bandwidth that is given to you need to transfer the data over that capacity. And the example, that I have given if your capacity of 8 megabit per second and you are going to transfer 8 megabit of packet; that means, you can transfer that 8 megabit data for 1 second. So, it will take 1 second to transfer the packet from the sender to the receiver.

So, this particular delay component we call it the Transmission Delay component. So, this transmission delay component depends on the capacity of the channel now the second component of the delay we call it as the Propagation Delay. So, the propagation delay is time to transfer one bit from one end of the link to another end of the link. So, usually depends on the underlying communication video. Now what is the difference between this transmission delay and the propagation delay. So, let us see it with the help of an example.

(Refer Slide Time: 20:05)



Now, think of a Sender and a Receiver. So, this is my sender and this is my receiver now the capacity. So, you can think of that we have a pipe between the sender and the receiver through which you are sending data. Now, the capacity of the pipe will depend on what is the width of this time. So, how much data it can pump to that pipe. So, if you think of that the capacity of data is say 8 Mbps the example, that I was giving; that means, from the sender side you can pump data to this capacity channel at a rate of 8 megabit per second. So, you can push 8 megabit data to this particular channel.

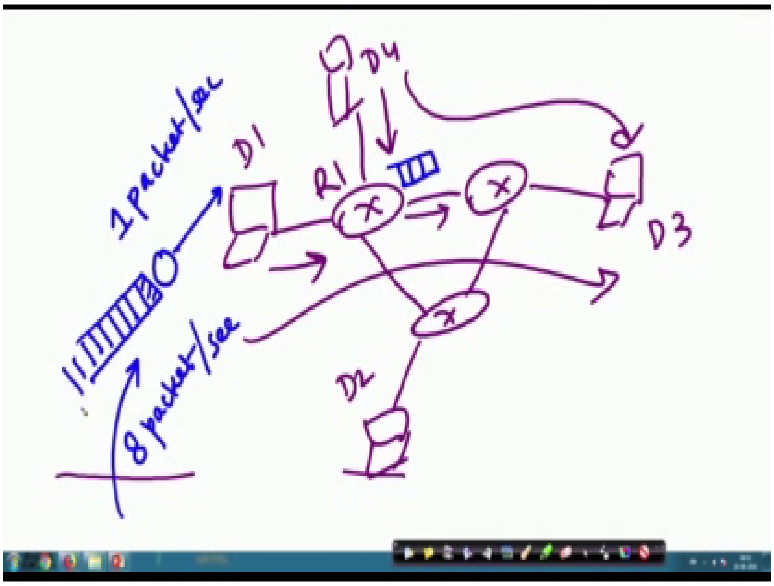
So, it is if it is depending on this capacity or the width of the pipe. Now if you have a wider pipe say; if you have a pipe of something like this between the same sender and the receiver where the capacity the width is more say it is 16 Mbps. If that is the case; that means, you can push data at despite at a rate of 16 megabit per second. So, if your packet size is still 8 megabit then it will take 0. it will take 0.5 second to push this entire data in this pipe.

So, because you are increasing the capacity of the pipe; so, this particular delay to transfer or to push the data in the pipe that we call as the Transmission Delay. Now what is the propagation delay now think of that you have pushed the first bit of data in the say the at the first bit of data has been pushed in the pipe. Now that particular data that particular bit need to reach at the other end.

Now, the time to reach for this bit to the other end that particular delay we call it as the Propagation Delay. So, whenever you are transferring a single bit you can think of it in the form of a signal that, you are transferring now the signal will get propagated over this particular channel, and if the length of the channel is more, it will take more time to get propagated. So, that is why this transmission delay it depends the transmission delay component it depends on the capacity of the pipe or the capacity of the channel and whereas, the propagation delay it depends on the length of the pipe or the length of the channel of the distance between the sender and the receiver.

So, these are the two different delay components that we have the Propagation Delay and the Transmission Delay and there is a third delay component which we call as the Queuing Delay. Now the queuing delay is an interesting component. So, the queuing delay is the delay at the interface buffer. So, whenever you are transferring certain data.

(Refer Slide Time: 23:37)

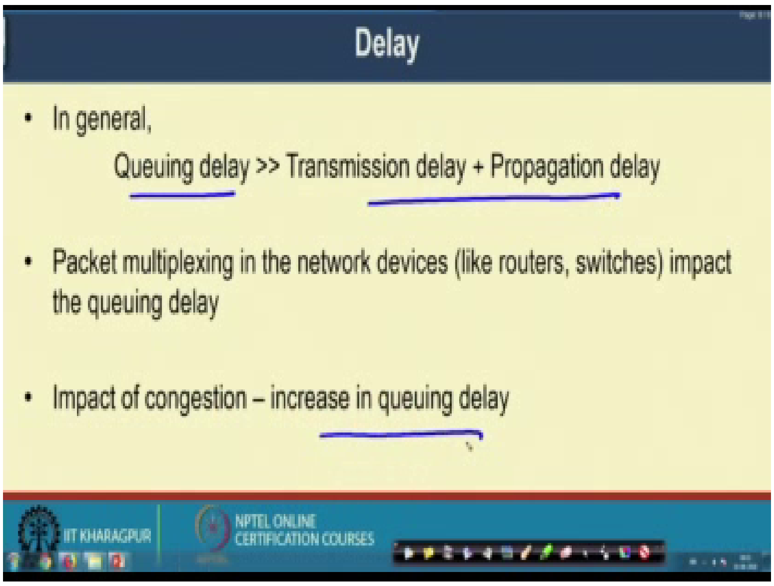


During that time you have from the sender to the receiver you have multiple intermediate routers. And these routers form a network and you have say multiple senders if you are sending data. So, you have device 1 device 2 and device 3 which are sending data simultaneously. So, assume that so another device a device 4 which is again sending data. So, device 4 is sending some data device D 1 is sending some data, but you have certain fixed capacity of this out going link say assume D 4 is sending data to D 3 as well as D is sending data to D 3.

But at this router say R1 you have some fixed capacity of this particular router, because you have some fixed capacity of this particular router the packet gets include in the intermediate buffer queue, or packet buffer queue that is your that is associated with this particular router. Now, as you have more data which are coming to this router your delay for transferring the data from this queue will increase. So, that packet has to wait for more amount of time in the queue. So, you can think of it in this way say the packet this particular queue it can process say 1 packet per second. So, if it can process 1 packet per second; that means, at the outgoing link you can send 1 packet per second and assume that you are receiving data at a rate of 8 packets per second.

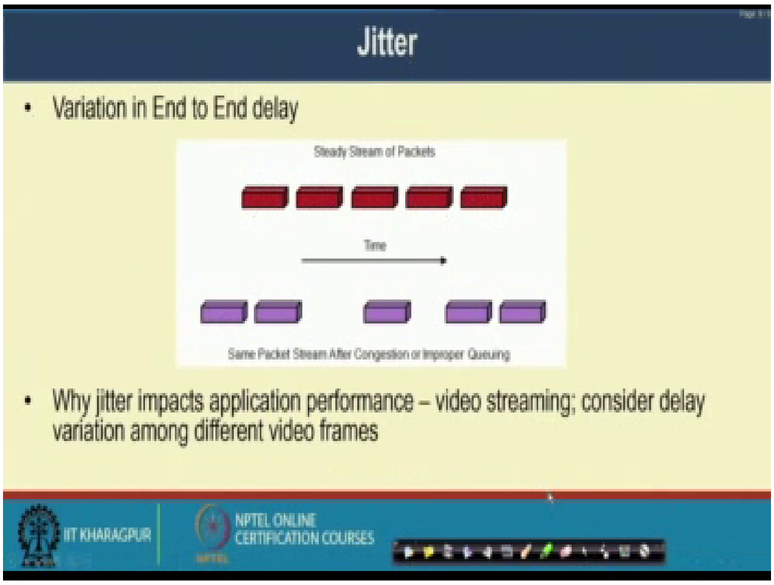
So; that means, at every second 8 packets are getting n queued and that is, but the router is able to process only 1 packet per second so; that means, by the time the routers will send this 1 packet in the channel another 8 packet will get n queued. So, you can just think of a line a queuing line in the in the gate of movie hall, ticket counters say; at INOX ticket counter and the people are coming at a faster rate compared to the service rate of that ticket counter. So, you have to wait for more amount of time in that particular line. So, the same thing happens here and this particular delay component we call it as the Queuing Delay. So, the queuing delay is the major delay component which impacts the network.

(Refer Slide Time: 26:10)



Now, in general what we have that these Queuing delay it is significantly much more than the Transmission delay and the Propagation delay. So, that is why the queuing delay it dominates the network. Now packet multiplexing in the network devices loud like the routers at the switches it impact the queuing delay, the example, that I have just shown you. Now if you have congestion in the network congestion in the network means; more packets are coming to the queue and it was not able to serve it. So, it increased the queuing delay ok.

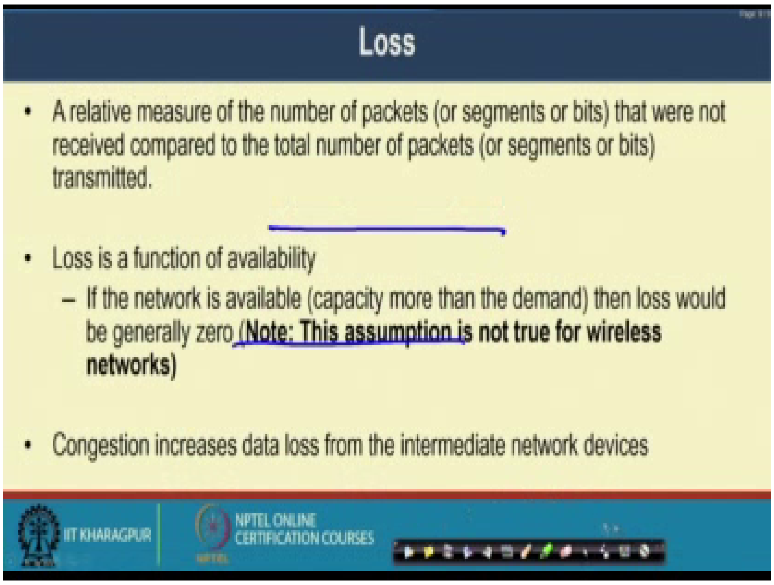
(Refer Slide Time: 26:44)



Now, coming to the third component which we call as the Jitter; so, as we have mentioned jitter is the Variation in End to End delay. So, you do not have a study stream of packets like this one rather what you have you have a stream of packets which are irregulated and there are different packets are transmitted different instants of time, and you do not have a steady stream of packets which impact the jitter.

So, in case of video streaming as the example that we have given earlier this kind of jitter it impacts significantly the application performance. So, the example, that we have given earlier that you are watching a live streaming and then packets are coming we differ in delay. So, it is not coming at a constant bitrate. So, the video player will not be able to play the service at a constant rate. So, it will see lots of up and downs in a video quality ok.

(Refer Slide Time: 27:43)

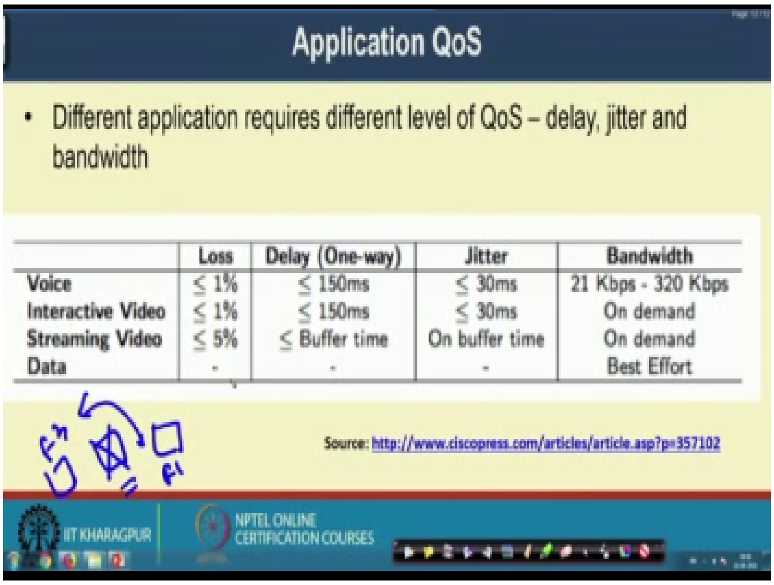


Now, coming to the fourth component which is the Loss; so, this loss is a relative measure of the number of packets or sometime we represent it in the form of segments or bits, that were not received compared to the total number of packets or segments or bits. So, this loss is a function of availability in generally if the network is available; that means, if your capacity is more than the demand then the loss will generally be zero you do not see any last there, but if your capacities list then you will see a significant loss from the network.

So, that is why whenever there is a congestion you will see there is a significant loss, and that is why TCP takes that loss as an indication of congestion, but this particular principle does not hold the exactly in case of wireless network, because in case of wireless network there can be loss from the channel as well like; you are transferring data in a open media and that is why you can just think of that lots of people are talking together in a single room. If a lots of people are talking together in a single room over that open environment then it will create a noise and non will be able to hear other voice.

So, the same thing happens in case of wireless media what we call as the interference during the discussion of wireless physical layer will talk about this interference in more details, but this interference is also results in packet loss. So, in wireless there are multiple results the multiple reasons for packet loss are indeed that is the reason you will experience more amount of packet loss in a wireless environment.

(Refer Slide Time: 29:26)



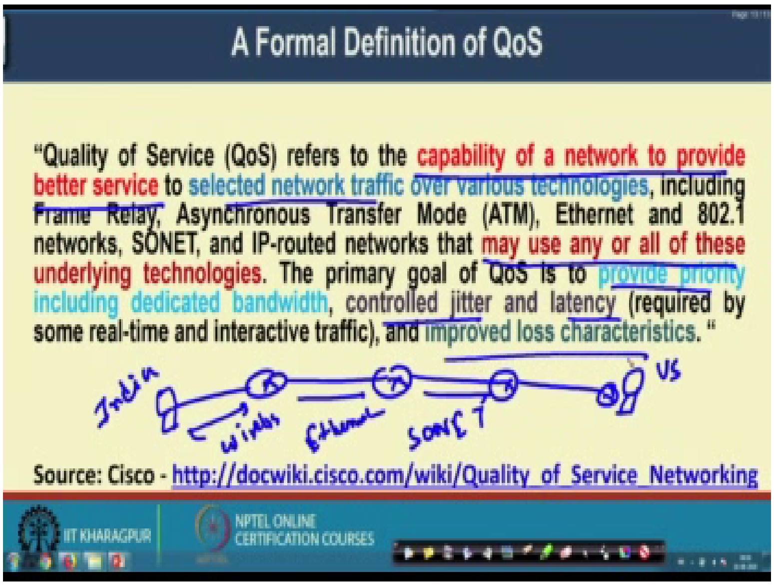
So, if we look into the Application level Quality of Service. So, different application requires different level of quality of service in terms of delay jitter and bandwidth. And here is an example chart for that. So, you can see that the for the voice application so this data is taken from c square typical networking company. So, the voice data has a loss tolerance of less than equal to 1 percent you need to it can tolerate very few losses the deal delay should be less than equal to 150 millisecond the jitter need to be less than equal to 30 millisecond and you require some 21 Kbps to 320 Kbps of dedicated bandwidth.

For interactive video like the live streaming the lost need to be less than 1 percent the delay can be around 150 milliseconds the jitter need to be 30 millisecond and the bandwidth is on demand like if you are watching a high quality video you will require high bandwidth. If you are watching standard quality video we may require little less bandwidth.

For the streaming video like the video that you are watching right now in YouTube, where the video has been pre recorded and now getting streamed. It can sustain more amount of loss why it can sustain more amount of loss because the video is already recorded. So, that is why say for example, you have received frame 1 frame 2 has lost and you have then received frame 3 then by doing averaging over frame 1 and frame 3 you can recover certain part of the frame the lost frame, because it is just like images frames means; images one after another. So, that is why it can sustain for more amount of loss the delay is equals to the buffer time at the client side jitter is again depends on the buffer time and the video is on demand

If you are going for high definition video you will require more amount of bandwidth for normal data transferred we do not have any bound on lost delay or jitter and the bandwidth is the best for the best bandwidth that you can support for that application ok.

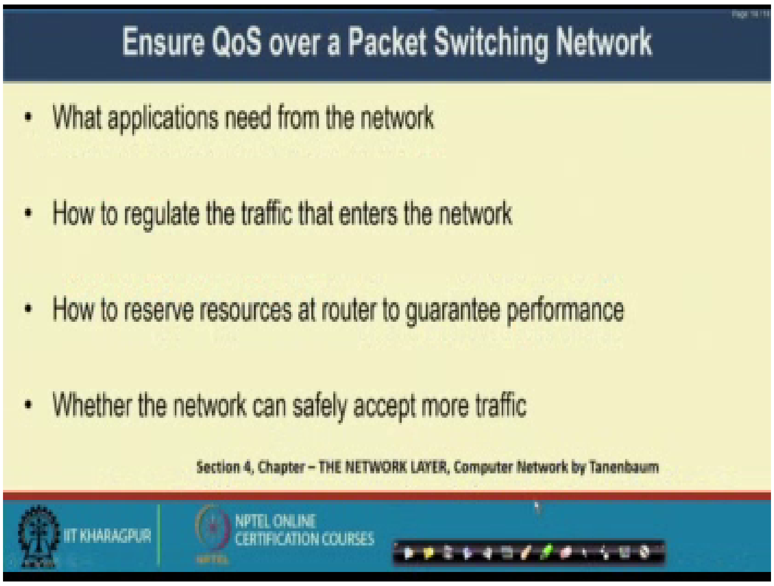
(Refer Slide Time: 31:24)



So, here is the formal definition of quality of service which is coming from Cisco. So, I have highlighted the different part of the definition in different color the important key words. So, quality of service refers to the capability of a network to provide better service to selected network traffic why it is selected network traffic, because we are looking into network traffic which are quality of service associated like the voice traffic or the video traffic, over various technologies including firmly layers asynchronous transfer mode ATM Ethernet and 802 that 1 networks, SONET, and IP based networks. So, you need to remember that, whenever you are trying to transfer data from one indorse to another indorse, say this host is India and this host is in US in between you have multiple such links and different links may be of different types.

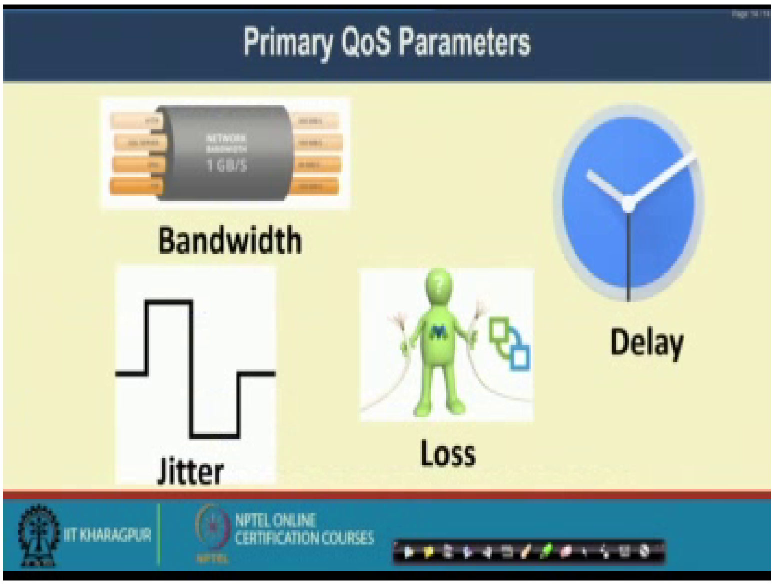
Here the first part may be Well is then the second top maybe Ethernet the top of may be optical network like SONET. So, you can have multiple networking technologies in between. So, you need to provide quality of service over multiple technologies; IP route network that may use any or all of these underlying technologies. So, the primary goal of quality of service is to provide priority including dedicated bandwidth, for certain classes of traffics controlled jitter and latency means one way delay required by some real time and interactive traffic and improve loss characteristics. So, this is the formal definition of quality of service to ensure quality of service.

(Refer Slide Time: 33:01)



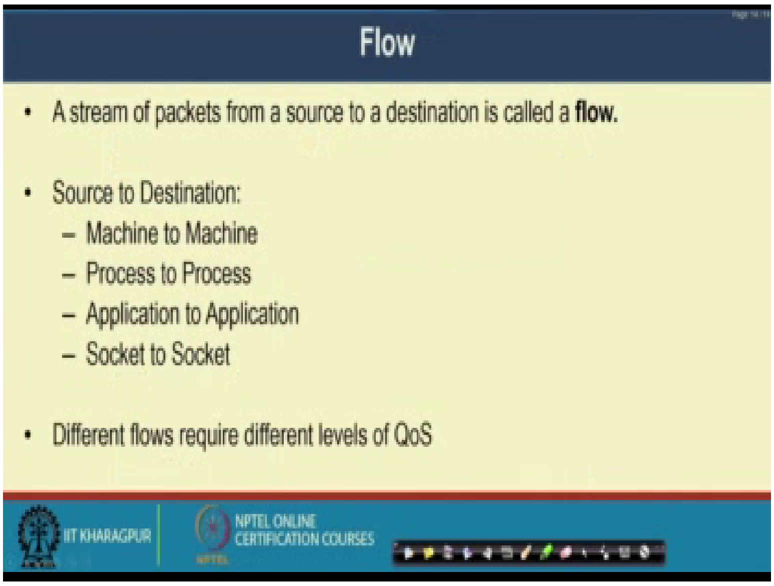
Over a Packet Switched Network we need to find out these four different things. First of all What the application it from the network, What type of quality of service it is expecting, How to regulate the traffic that enters the network to provide quality of service we will look into all these things in details in the subsequent lectures. Then how to reserve the resources at a router to guarantee performance because you need to have n to n dedicated resources to ensure certain classes of quality of service, and whether the network can safely accept more traffic while not violating the quality of service of the existing traffics.

(Refer Slide Time: 33:43)



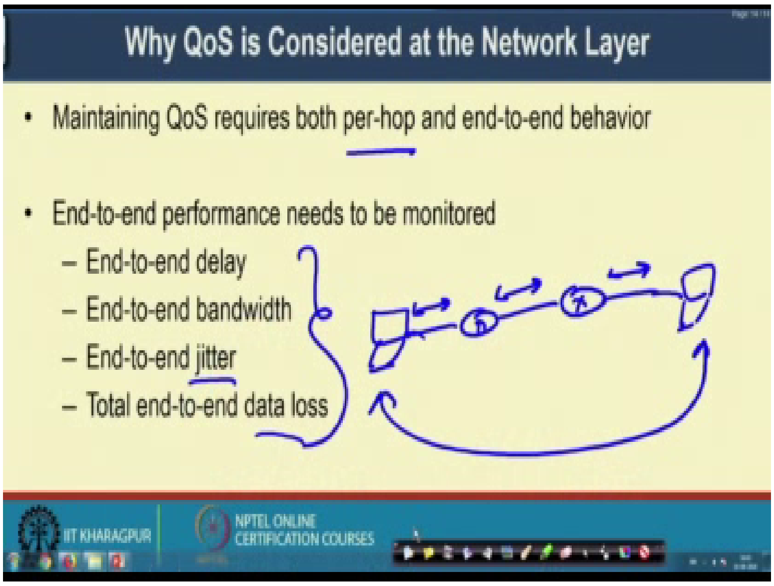
So, the Primary Quality of Service Parameters that you have already seen there are four such parameters the Bandwidth, Delay, Jitter and Loss.

(Refer Slide Time: 33:53)



So, in terms of quality of service many of the time will use this term Flow. So, this flow is a stream of packets from a source to a destination now when we call as a source to destination there can be multiple set definitions. It can be a machine to machine communication, so we want to ensure quality of service between two machines, it can be processed to process communication like we want to provide quality of service between two different processes. When we are communicating with each other we may want to ensure quality of service from application to application or from socket to socket different ways we can define the two end of a flow. Now, different flows require different levels of quality of service and accordingly we need to provide quality of service to those flows.

(Refer Slide Time: 34:40)

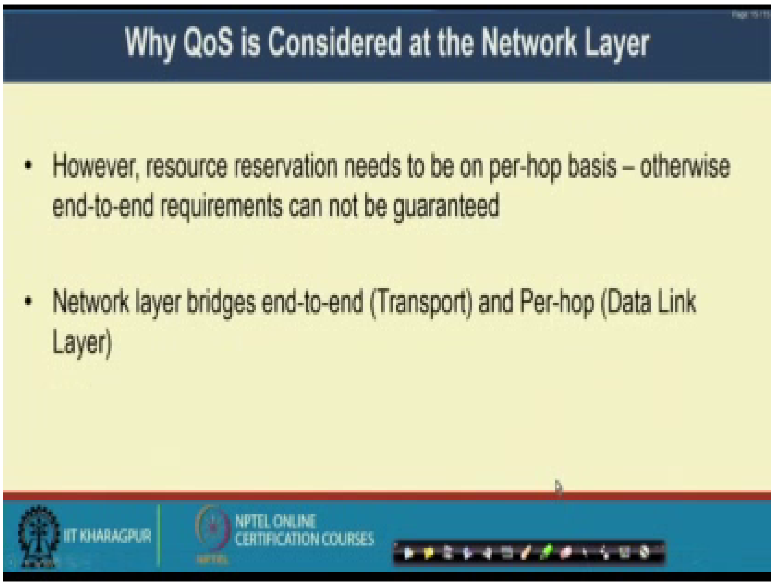


Now, another important question is that; why quality of service is considered at the network layer, because maintaining quality of service requires both per happen n to n behavior. So, whenever you want to ensure quality of service over the network, say you have these two different n host and between these two different windows you have multiple intermediate routers or switches.

Now, what you have to do you have to see that what is this End to end requirement to ensure quality of service, what End to end bandwidth you want to ensure, what End to end delay you want to ensure what End to end jitter you want to ensure, or what end to end data loss you want to ensure, but to ensure this you have to this reserve the resources at every individual hop of the network. So, to guarantee quality of service you require this per hop behaviors are you need to reserve the resources at every individual hops in the end to end path.

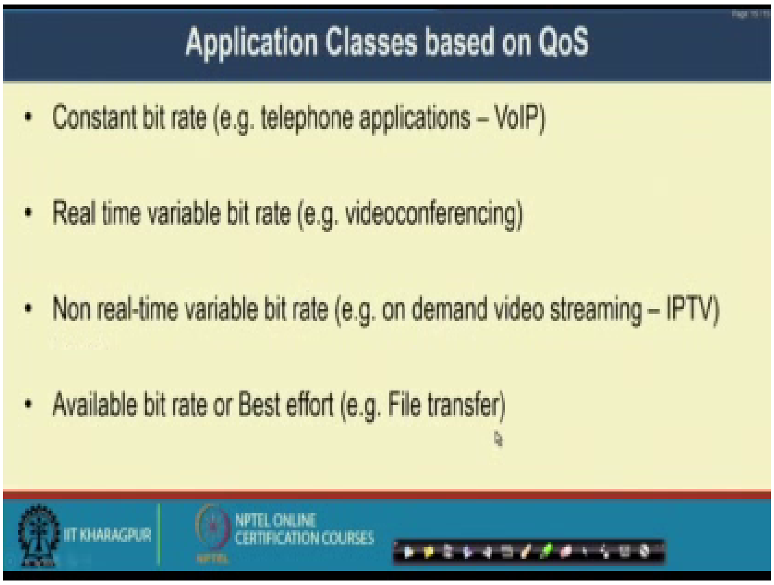
So, that is why we need to consider quality of service by considering the information from end to end perspective as well as per hop perspective. Now, if you look carefully this network clear it sits in between the transpose layer and the data link layer. So, the data link layer considers per hop behaviors and the network layer provides you this routing over that individual hop whereas, the transport layer it gives you the end to end information. So, you can get a feedback from the transport layer and apply the things to the data link layer. So, that is why you implement quality of service and then network later.

(Refer Slide Time: 36:20)



So, that is the point I have this mentioned resource reservation need to be on per hop basis otherwise, end to end requirements cannot be guaranteed and because the network layer bridges the end to end like the transport layer and per hop the data link layer you implement quality of service at the network layer.

(Refer Slide Time: 36:37)



Now a based on the quality of service we define multiple application classes like the constant bit rate classes so, example of constant bit rate is telephone applications like Voice over IP. So, you require constant streaming of data bits they in the second to on a studial time variable bit rates. So, in case of real time variable bit rate example, is some video conferencing service where the bit rate can be variable depending on the frames that you are going to transfer, but it need to be a real lifetime.

The third one is Non real time variable bit rate like on demand video streaming. So, for example, television service over IP network. So, if you want to observe a TV over IP network, it is a variable bit rate at the same time it did not to be in real time and in available bit rate service at the best effort service like the file transfer services. So, we look into all these classes of traffic in more details. And, in the next class we will look into that considering this different level of quality of service or different classes of quality of service. How the network should design itself so, that it can provide the desired level of quality of service.

So, hopefully you have got a basic idea about what quality of service means and in the next class we will go to the details about how the network actually provides quality of service in the internet.

So, thank you all for attending this class.