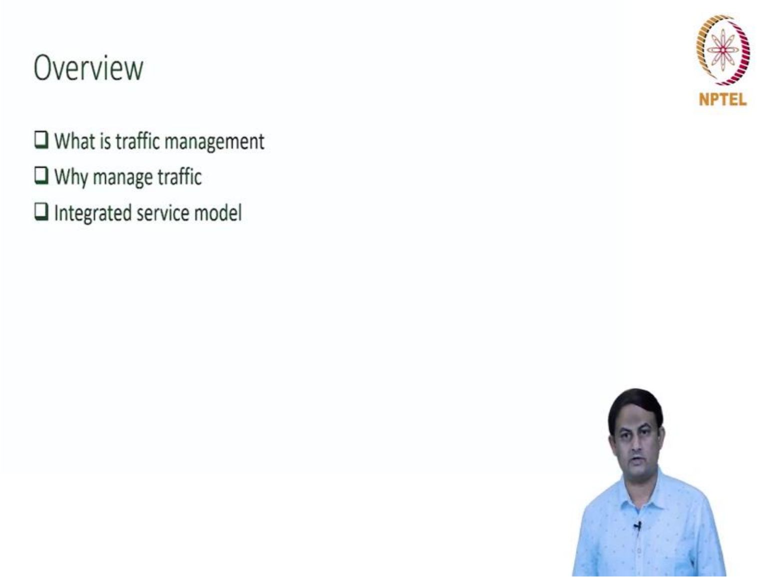
Advanced Computer Networks
Professor Doctor Neminath Hubbali
Department of Computer Science and Engineering
Indian Institute of Technology, Indore
Lecture 14
Traffic Management - Part 1

Welcome back. In today's lecture, we will talk about a topic called as the Traffic Management.

(Refer Slide Time: 00:24)

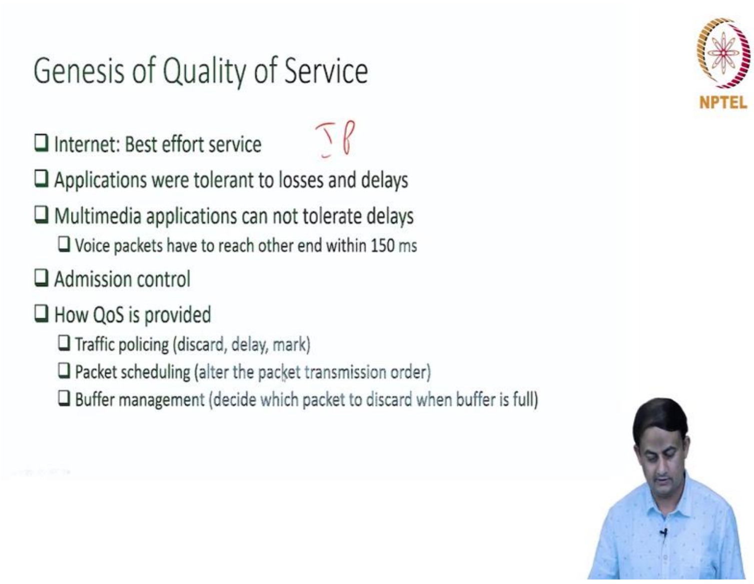


Overview

- ❑ What is traffic management
- ❑ Why manage traffic
- ❑ Integrated service model

And here is the agenda for today's lecture, we will try to understand what is traffic management, why we need to do the traffic management in the network and then finally, how do you actually manage the traffic with reference to one architecture called as the integrated service model, we see how traffic management is done in the network.

(Refer Slide Time: 00:51)



Genesis of Quality of Service

❑ Internet: Best effort service IP
❑ Applications were tolerant to losses and delays
❑ Multimedia applications can not tolerate delays
  ❑ Voice packets have to reach other end within 150 ms
❑ Admission control
❑ How QoS is provided
  ❑ Traffic policing (discard, delay, mark)
  ❑ Packet scheduling (alter the packet transmission order)
  ❑ Buffer management (decide which packet to discard when buffer is full)

So, before we actually look into the actual model, let us try to understand why traffic management is required in the network; where is the, what is the origin of the traffic management. So, the Internet was designed from the perspective of providing the best effort delivery service model.

So, when I say the internet, it is actually typically refers to the IP protocol inside the TCP IP model. So, the IP protocol in the TCP IP model basically tries to deliver the packet to the correct destination, but occasionally it do deliver the packets majority most of the time, but occasionally, it may not deliver it, the packets may be lost, it might be corrupted, and all things can happen.

So, this can be the service model was good enough in the earlier days, whereas the applications which were using these kinds of service delivery models were completely different. So, in the earlier days, the file transfer and the email were the two primary applications which were using, so they are not sensitive to these kinds of losses, delays and other things. So, at a higher layer, particularly at the TCP and the application layer probably you can handle such kind of the losses and the other errors that are introduced in the transmission.

So, but as days passed on, newer applications came into picture. So, we typically we had, we have the multimedia applications. So, these applications are sensitive to delay and also the losses, particularly if you take the case of the voice transmission, so it requires a very stringent delivery model. So, within the particular time frame, the packet has to reach the other end. So, in particular, in 150 milliseconds, the packet which originated at some source need to be delivered to the destination, if it takes beyond that time, then the audio will not be smooth.

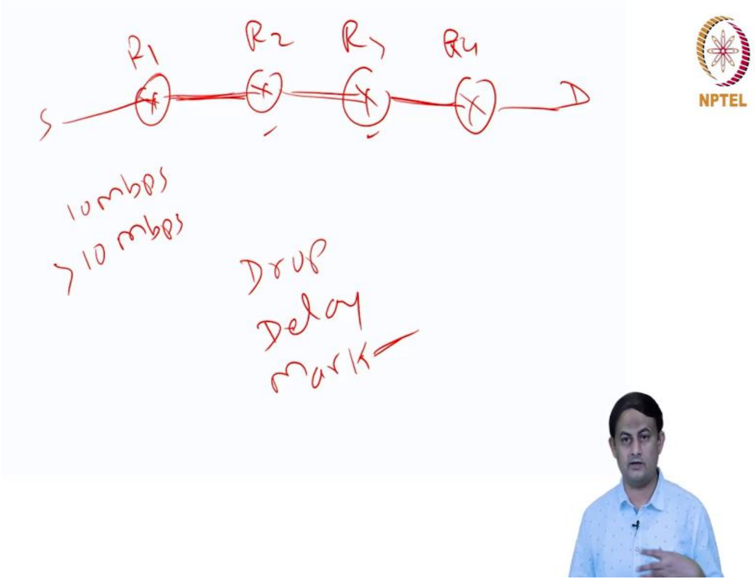
So now, the question is how do we actually run such kind of applications on top of these best effort delivery service models that the IP offers. In order to deliver such models, you require something called the quality of service. Quality of service here, in the sense, it means that the packet needs to be delivered to the right destination within the particular time frame or some other application might tell that I required so and so amount of the bandwidth and so forth, you can specify the quality of service in different parameters. We will see that in a minute’s time.

But that is what different applications can put different constraints on the delivery model. So, in order to facilitate or meet those requirements of the application, the routers in the network need to be aware of that. They need to make some preferred choices over the best effort to delivery model that the IP offers. A simple way to think of the providing the quality of service is to think of something called as the admission control. So, where the application comes and tells you or the network that I want so and so things from you, I want particular amount of the bandwidth, I want a particular throughput, I want a time delivery constraint of so and so.

So, under different, it can tell. So, if the network thinks that it can meet those requirements that the application tells, then it can actually admit that; otherwise, it may not admit that. That is a simple way to think of how the quality of service is achieved in the network. But it is not the only admission controls that work or at least working in practice.

There are a number of ways with which you can actually achieve the quality of service in the network. There are three things which are primarily used in today's network - one is called as the traffic policing and the second one is called as a packet scheduling and the third technical something called as the buffer management. So, let us try to understand each one of them in little more detail.

(Refer Slide Time: 05:10)



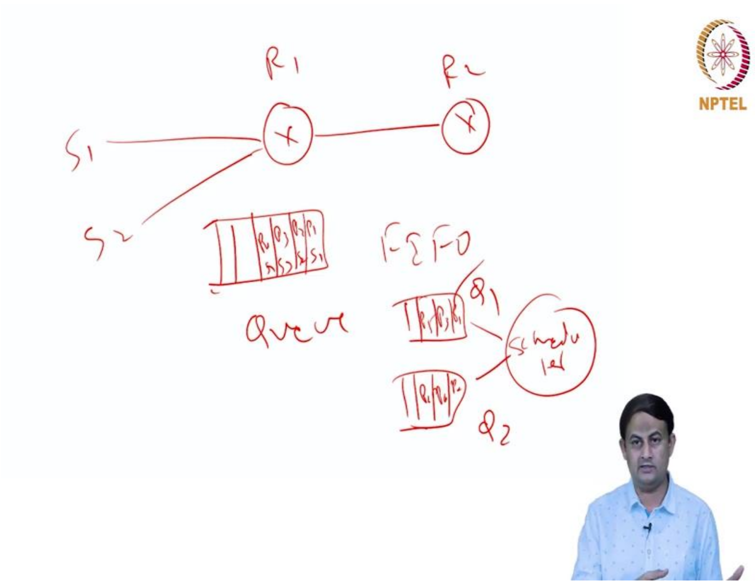
So, what traffic policing means is so let us say I have a source and destination and you need to deliver the packets originating from the source S to this destination D. And the source upfront it tells that I will go into transmit at a particular rate. So, let us say it says that the maximum rate that I am going to transmit is 10 Mbps and accordingly, some provision is made in the network. So, from this source to the distribution all around the path to carry 10 Mbps of the peak rate at which the source is generating the data, the network is provisioned, some resources are provisioned.

So, what if the source does not honor this thing? So, if it stops transmitting beyond 10 Mbps, what the network will do? The resources are provisioned only for 10 Mbps, but the source is transmitting beyond that rate. So, one simple way with which you can handle this is the router here, particularly the first router may drop the packet which are coming beyond the promised rate or it might delay the transmission of the packets originating from S or if there is a bandwidth available more than 10 Mbps it can still forward all those packets coming from S, but marking that a subsequent router, I can name these as R1, R2, R3 and R4.

So, if there is available bandwidth between R1 and R2, R1 can still transmit the packets in excess of 10 Mbps. But with the remark in the packet saying that if any of you, let us say the R2 or R3 or R4, any one of them are having issue in providing or transmitting more than 10 Mbps, they can probably drop the packets. So, either of these things. In a nutshell, policing is all about either somebody who is not honoring the what is promised transmission rate or the whatever the quality of service parameters we have negotiated, if it is violated then you in some form, you take some penal action.

So, either you drop or delay or you mark the packets and send it, where it might be subsequently subjected to other alterations in the network. So, this is called as the policing, using which you can actually achieve the or provide the quality of service. The second way up or doing or achieving the quality of service is something called as the packet scheduling.

(Refer Slide Time: 08:04)



So, let us take the example again. Here is the router, let me call that router as R1, here is a source which is actually transmitting or maybe two sources, S1 and S2 and the packets are coming to this router. So, every router in the network has got something called a buffer space, this you can also call as the queue and the packets coming from S1 and S2 are put into this queue. So, let us say packet P1 is coming from the S1 is put right here, packet P2 is coming from the S2, packet P3 is coming from S3, P4 from S1 and so forth, something like this.

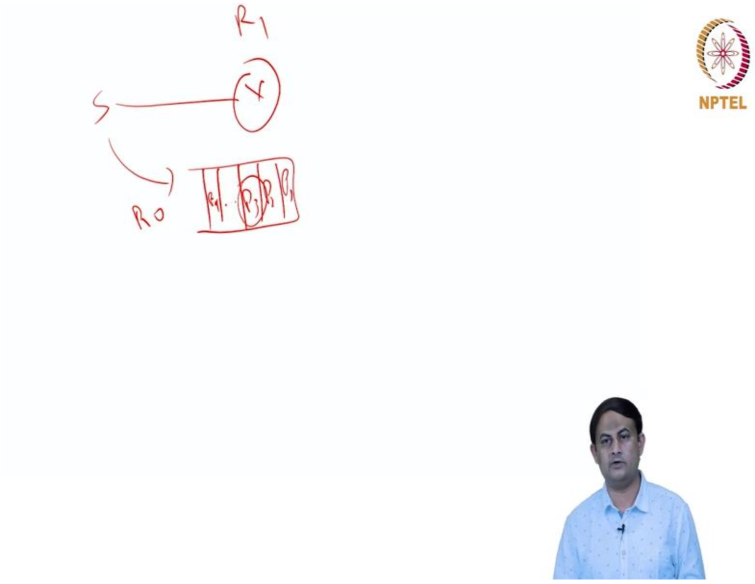
So, there is an outgoing link which is connecting this router to the next router and for the time being, let us assume that all of these packets are going to R2 in the network and the default strategy of transmitting these packets over the network is to pick up the packets in the order in which they have come to R1 and then transmit in the same order. So first, P1 gets transmitted, then second P2 gets transmitted and so forth, P1, P2, P3, P4 in order. So, this is the default strategy.

It is also called as the FIFO model or the FIFO scheduling first in first scheduling, this is the deep strategy. So, using, this is the best effort. So here the, if I want to provide a priority to the traffic originating from the S1 over the traffic originating from the S2 using the FIFO strategy, I cannot really do that. So, what you can think of is instead of having a single queue, I might have multiple queues, maybe the Q1 and Q2 and a simple way to visualize it all the packets coming from the source S1 with P1, P3 and P5 are put in the queue number 1 and all the packets coming from the source number 2, P, P4 and P6, something like this are put in the second queue.

And there is a component which is sitting ahead of this one, this is called the scheduler. What it does is it can pick up the packets from these two queues in a different order. So, maybe if I want to provide the priority for the transmission originating from S1 and I can tell my scheduler to pick up the packets if available from the Q1 always and if there are none in the first queue, then it go to second queue and then transmit the packets. So, by doing this, this is a kind of differentiated service, I am actually providing a kind of the priority to the S1's transmission over the S2, that is the quality of service.

So, using such it is not necessary that I should have only two queues, but I can have more than two queues as well. So, by maintaining multiple queues at a router, and by appropriately putting the packets in particular different queues, I can actually maintain or by picking the packets from the different queues, I can bring some kind of some notion of the quality of service to the transmission in the network, this is the second strategy with which you can actually achieve the quality of service and that also falls under the traffic management strategy.

(Refer Slide Time: 11:52)



And the third way of providing the quality of service is something called as the buffer management. So, let us again go back to the same example, you have a source and the source is transmitting to this router and there is a queue available. And in this queue, the packets are put - P1, P2, P3 and so on. So, this queue is basically a buffer where the packets are getting stored and the scheduler actually picks up the packets from this queue and then transmit it to the or schedule check to transmitted to the outgoing link.

This buffer is of finite space; given any router, you cannot have an infinite amount of memory the work space cannot be infinite. So, it has got, it can hold only a finite number of the packets. And at a particular time, it might happen that the buffer available for storage is actually full. So, for example, if it can accommodate maybe 9 packets up to P9, the things are filled inside this queue, P1 to P9 are already there. What if the source actually sends the 10 packet?

Now, the question arises, one simple way is to since there is no space in the queue, the new packet that is coming, I may not be able to put in the queue, so I drop the packet P10 itself. Or the alternative way to think of it is P10 is belonging to some communication, some application running in the source S, which is actually sensitive to delay and transmission delay, and I cannot afford to lose the packet P10. Then what I can do? I do not want to drop the packet P10.

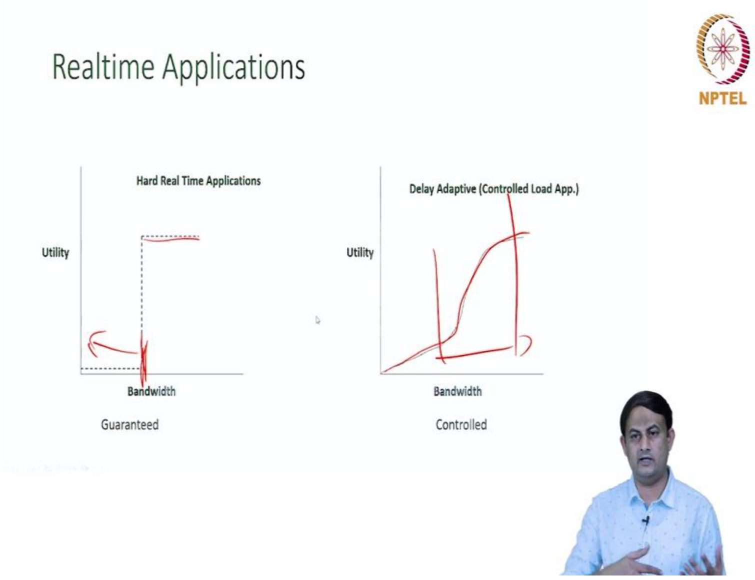
But then simple alternative could be I can, instead of dropping the packet P10, I might choose any other packet from the queue. So, maybe I can drop the packet P3 after knowing that P3 is not belonging to an application, which is sensitive to delay and bandwidth and other quality of service parameters. So, I can afford to lose P3. So, drop P3 and then make room for P10. Come and insert that P10 inside the queue so that the transmission experienced by the source and two endpoints is smooth enough.

So, when the resources is choked, when there is no buffer space available in router, I can decide which packet to draw from the queue, either it is P10 or P3 or some other packet. So, this is called as the buffer management. By dropping the packets belonging to applications which are not sensitive to losses and retransmission and delays, I can provide the quality of service for some other applications which are sensitive to such kind of the variations that happen in the network.

So, that is called as the buffer management. In simple words, buffer management deals with answering the question, if there is no space to accommodate the packets in the queue, then what strategy do I use to drop the packets from the queue to make room for the new packet that is coming? So, that is called as the buffer management.

By intelligently manipulating or choosing what packet to drop, I can still provide the quality of service to the end applications, which really requires that. So, these are the three strategies - either you do the polishing or you do the buffer management or you do the scheduler or you can do all of these three together to meet the quality of service requirements from the applications.

(Refer Slide Time: 15:59)



Realtime Applications

NPTEL
Hard Real Time Applications
Utility
Bandwidth
Guaranteed
Delay Adaptive (Controlled Load App.)
Utility
Bandwidth
Controlled

So, I said that some of the applications really cannot tolerate the delay and losses and all that, while some other applications can live with delay and losses and other manipulations that happen in the network. The applications which really cannot tolerate such kind of variations typically are of the type called as real time applications and these real time applications are again classified or divided into two types, one is called the hard-real time application and second one is called delay adaptive real time applications.

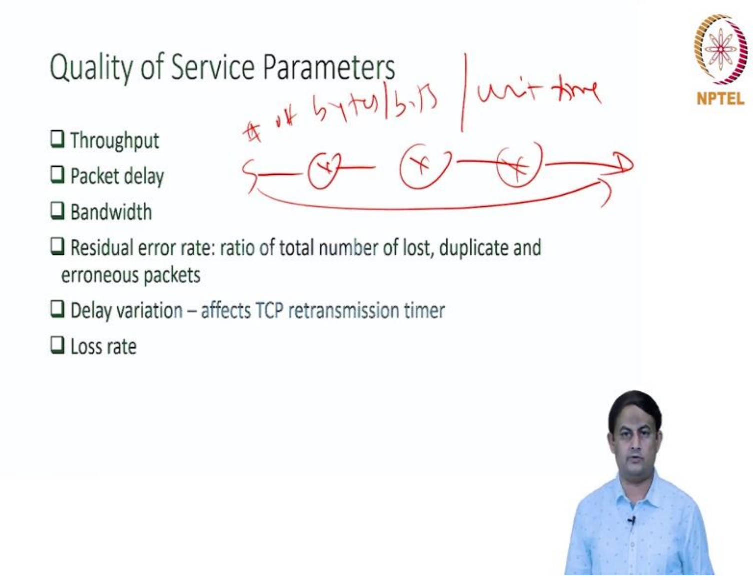
So, the difference is the following. In the hard-real time application, if I measure the utility quotient, when the quality of service parameter is below a certain threshold, here the quality of service parameter is bandwidth, when the bandwidth falls less than this threshold, the utility quotient of this application is 0, it is as good as not doing anything, but when the bandwidth or the quality of service parameter is beyond this one, the utility quotient will be 1. Either it would be 0 when you cannot meet the minimum requirement set by that application and it will be 1 when you are able to meet the requirements set by that application, it is either of these two.

So, such applications are called as the hard-real time applications. You need to provide such kind of the whatever the quality of service parameters that application requires, that has to be met. And in the second category, the utility quotient is not falling to 0. But instead depending upon the availability of the whatever the parameter that you want to meet here, in this case is the bandwidth, depending upon the availability of the bandwidth utility quotient keeps on increasing.

The more the bandwidth you have, the more the utility quotient is. So, only when the utility quotient will fall to 0 when you have 0 bandwidth. But as the bandwidth availability increases, then the utility will also increase. So, it can increase up to a level and beyond that point of time, then probably it will saturate. Here, in this case, you can still continue to utilize the application, but the experience, user experience may not be that great, but if you have a larger of bandwidth, then the user experience might be probably much better, that is the desired scenario.

So, I want to run depending upon the application type - whether it is hard real time or delay adaptive real time application, I can decide what kind of quality of service I want to provide for this particular application and I can make the routers in the network aware that so and so application is running, this is of this type and you need to provide this, this kind of service to that application.

(Refer Slide Time: 19:08)



Quality of Service Parameters

- ❑ Throughput
- ❑ Packet delay
- ❑ Bandwidth
- ❑ Residual error rate: ratio of total number of lost, duplicate and erroneous packets
- ❑ Delay variation – affects TCP retransmission timer
- ❑ Loss rate

So, it said that quality of service are different applications require different kind of the quality of service. Now, the question is how does the applications tell the routers or the network what kind of the quality of service it requires? There are a bunch of parameters which you can use to tell the network that the I want so and so kind of the quality of service. The first parameter is something called the throughput, which simply measures the number of the bytes or bits per unit time.

Maybe, the number of the packets per second, the number of the bytes per second, something like this. So, what is the throughput that you can achieve from this one? So, the application might come back and then say that I want you to transmit at least 100 packets per second or 1 million packets per second, below that, I cannot afford to do the useful transmission. So, that is one parameter you can use to specify the quality of service.

The second parameter is the packet delay. This packet delay is an end to end delay. From the source to the destination you are going or traversing through a bunch of intermediate nodes. From here to here, what is the maximum delay that the network can take to deliver the packet from that source to destination? That is called as the packet delay. Using that also you can specify the quality of service requirement.

And the third parameter with which you can specify the quality of service requirement is something called as the bandwidth, the application might come and tell that I require minimum x amount of the bandwidth, maybe I require 1 Mbps, I require 2 Mbps for useful utilization. So, below that, I cannot run smoothly. So, now, the application might tell that or the application might tell the quality of service requirement in terms of something called as the residual error rate.

So, error is basically the number of the packet that gets missed or lost in the network, some of the packets might get duplicated or some of the packets might arrive erroneously at the receiver or the destination. So, what fraction of the total transmissions that the sender does, the packet can arrive erroneously at the other end. So, it might be that 1 percent of the total packets transmitted can be erroneous or 0.1 percentage of the total packet transmitted can be erroneous or whatever is the requirement, the application can come and tell.

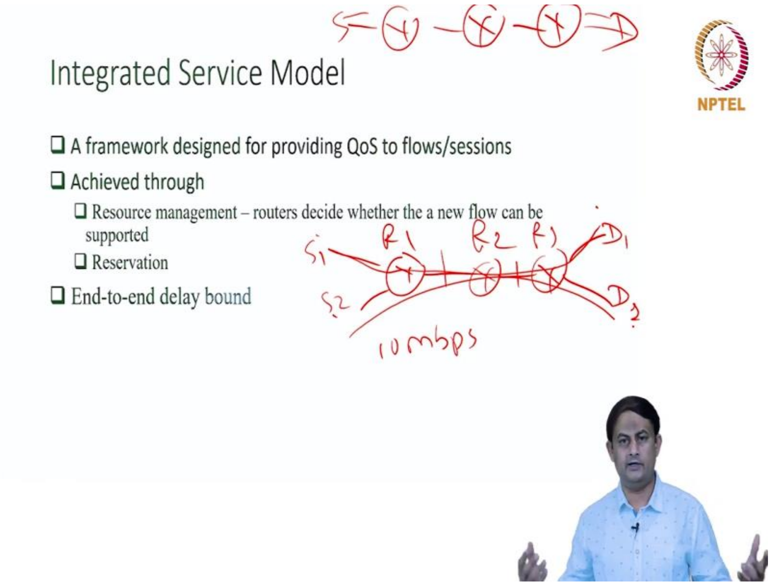
And the fifth parameter that the applications can use to tell the quality of service requirement is something called as the delay variation. A delay variation is the in between the successive transmissions. So, if you let us say I have P1 to P10 packets; from the P1 to P2, what is the delay that it took? And from P2 to P3 what is the delay that it took? And P3 to P4 and so forth. So, in between successive packet arrivals at the destination, what is the maximum or the peak delay that the application can tolerate?

So, the inter arrival time between the packets cannot go beyond the threshold, maybe 1 millisecond. So, within 1 millisecond, if P1 has been delivered, next 1 millisecond P2 has to come to the zone. You can tell the network to deliver the packets in this particular fashion. And the last parameter that the application can use to specify the quality of service requirement is something called as the loss rate.

Loss rate tells the fraction of the packet that gets lost in the network. So, it is not necessary that you use only one particular parameter to specify the quality of service requirement, a particular application can specify the quality of service requirement using more than one parameter. So, for example, some application might say that I want X amount of the throughput; at the same time, the bandwidth requirement is also Y.

So, using these two. Or some other application might come and tell that I want a throughput of X and the packet delay, end to end delay cannot exceed some Y milliseconds. And at the same time, the loss rate cannot be more than this. So, using a subset of all these parameters, a particular application can specify the quality of service requirements to the network and the routers need to be aware of all these parameters in order to meet those requirements set by the application. So, having studied these are the parameters that are using which you can specify the quality of service requirement.

(Refer Slide Time: 24:25)



Integrated Service Model

- ❑ A framework designed for providing QoS to flows/sessions
- ❑ Achieved through
  - ❑ Resource management – routers decide whether the a new flow can be supported
  - ❑ Reservation
- ❑ End-to-end delay bound

Now the question is, how do we actually meet these requirements put by the applications in reality? There are different service models that the network designers have come up with, in order to facilitate or meet the quality of service requirements set by different applications. And one of the simplest model of achieving the quality of service in the network is something called as the integrated service model. So, this service model, it is actually a bunch of things put together which actually provides this kind of service model.

What it does is, it applies or provides this quality of service to something called as the flows or the session. So, meaning, if a source is transmitting and the D is the destination going and again through a bunch of intermediate nodes, the packets are getting transmitted, the series of packets originating from one source going to one destination and the series of packet, this is called as the flow or the session. So, when the transmission begins till the transmission ends, that is called as one flow or one session.

It applies these quality of service requirements set by the application to D series of the packets, also called as the flows. So, integrated service model, whatever the specification that you do, that is applicable for one entire session or one particular flow. So, it might happen that on the same two endpoints, S and D, two different applications might run. They might have different requirements, but two endpoints, two applications together will define one flow to that flow the quality of service requirements are applied.

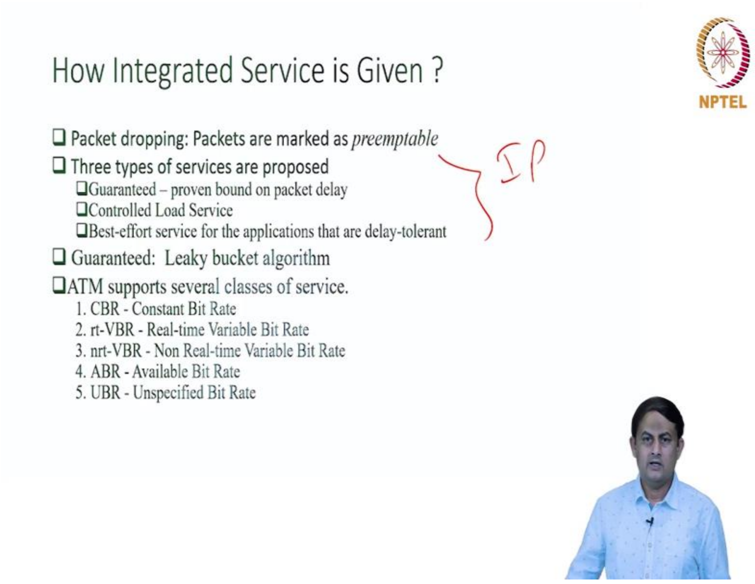
Now, how exactly this is done at least in the integrated service model. So, this is quality of service flow is provided with two mechanisms, one is called resource management with the admission control - you decide whether to admit a particular new flow or do not admit. So, for example, if we S1 and S2 are two transmitters and D1 and D2 are two destinations, and you have got three routers in between, S1 is connected to R1, R2 and then an R3, like this.

If S2 and D2 are already engaged in conversation, and S1 wants to talk to D1 and S1 says that I want a bandwidth of 10 Mbps and now, the question when this requirement comes to R1 S1 tells that I want 10 Mbps, R1 need to decide whether without disturbing the communication, ongoing conversation between S2 and D2, whether 10 Mbps of the bandwidth is available in the sling.

And similarly, R2 now also need to decide whether there is 10 Mbps bandwidth available in this link and so forth. So, without affecting the existing or ongoing conversation in the network, the routers need to decide whether to support the new flow coming into the network or not to support, this is called admission control. It does has a provision for the admission control and the second mechanism is once you decide to admit some flow, you will reserve the resources all around the path, meaning if I select say the network can indeed support the new flow between S1 and D1 and you reserve some resources or along the path, between S1 and D1.

So, at R1, you reserve some resource; at R2, you reserve some resource and at R3, you reserve some resource. So, resource can be anything - it might be a buffer space, it might be the order in which the packets are picked, the scheduler is altered or any of this, using which you bring the notion of this required quality of service. So, two things - one is the reservation and then the admission control together you actually bring the quality of service to the application, that is what the integrated service model does. So, if it incorporates these two, it gives some kind of the guarantee, a delay guarantee. So, I can deliver this packet to within this particular time bound.

(Refer Slide Time: 29:11)



How Integrated Service is Given ?

- ❑ Packet dropping: Packets are marked as *preemptable*
- ❑ Three types of services are proposed
  - ❑Guaranteed – proven bound on packet delay
  - ❑Controlled Load Service
  - ❑Best-effort service for the applications that are delay-tolerant
- ❑ Guaranteed: Leaky bucket algorithm
- ❑ATM supports several classes of service.
  1. CBR - Constant Bit Rate
  2. rt-VBR - Real-time Variable Bit Rate
  3. nrt-VBR - Non Real-time Variable Bit Rate
  4. ABR - Available Bit Rate
  5. UBR - Unspecified Bit Rate

So, now, how exactly this is achieved? I said that you can do the admission control, you can do the packet scheduling, you can also do the resource reservation, but do all the applications that are run between any two, any two applications require same kind of resources to be resolved all on the path, the answer is no. So, how do we actually provide different kinds of the resources to the different applications?

So, maybe the communication between S1 and D1 requires some amount of the resource and the communication between S2 and D2 requires different kinds of resources. So, it depends on the type of application, what it exactly wants from the network or the routers. The answer to this is with respect to the earlier discussion on the real time application being either the hard-real time application or the delay adaptive real time application. So, if it is hard real time application, it cannot tolerate the resources being short, is beyond below a certain threshold.

In order to meet that kind of the requirements, the integrated service model architecture gives you a kind of the service within its framework called as the guaranteed service. What it means is, the application tells you this is my requirement and the internet service model, actually, the guaranteed service model tells you that, I will guarantee you to meet that requirement that you have set. So, if let us say the application says that my bandwidth requirement is X, let us say 10 Mbps, so, I will guarantee you that and all the time I will provide you 10 Mbps.

How does it provide? I will reserve 10 Mbps of the bandwidth between the source sender and the receiver all along the path? If the requirement is on the jetter, at every hub, it takes a certain amount of the time for transmission. So, it calculates the end to end total time taken and it put constraints on the individual routers and individual hubs, thereby it guarantees that this is the total delay experienced by the packet from the source to the destination.

So, such kind of the hard requirements are met by this service called as the guaranteed service. And the second type of the service that it provides is something called as the controlled load service. So, this is typically for the second category of real time applications which are delay adaptive, where the requirement is not very stringent.

So, if you have more bandwidth available or the if you have the lesser end to end delay, the better it is, and the application can really adapt to the availability of the resources and what the control load service tells you is I will try my level best to weak or provide or guarantee a level of service and occasionally I might fall short of that, but I will try at maximum amount of the time I will be providing you the service which is at the best possible that is level, close to the to the peak that is available, that is what the second category of the model.

And the third category of the service model is really not for the real time application, which is the default service that the IP offers, it just utilizes that the best effort service and then transmit the data. So, there is no kind of differentiation or any kind of guarantee given to this kind of service model. So, given an application, you want to engage in a conversation and you can tell the application can tell them roughly map to one of these kinds of two service models.

So, either it is guaranteed or it is the control node service or it is the best effort service and based on that, the network can provision appropriate amount of resources and then provision the smooth transmission of the particular traffic to meet the quality of service requirements of that particular application. So, this kind of the service model, whatever I was talking about is with respect to the TCP IP model. In the TCP IP communication stack, particularly with the best effort delivery model of the IP, we have used integrated service model, these are the three differentiations that it can bring.

But it is not the only TCP IP model which actually guarantees such kind of the differentiated service, in other networking and also some particularly with the ATM networks also such kind of differentiated services will be provided. So, ATM also provides quite kind of the classes of services - one is called constant bitrate which says that I am going to give you between source and destination, the transmission rate is fixed. It will provide a second kind of service called as the real time variable bitrate.

Bitrate is varying, but I am going to ensure that the variation is not that much, still you can run a real time application using this transmission rate. It can also be non-real-time variable bitrate. Bitrate is variable, but this service model is not suitable for the real time applications. It might also give you service plus something called available bitrate. Here, this class at peak usually do not provide any kind of the guarantee on the transmission rate. Whatever is available I am running so and so applications, so and so flows are currently passing through me and they are consuming so much amount of the bandwidth.

And whatever is the residual bandwidth that is available using which I am going to give you some bitrate, that is called as the available bitrate; whatever is unutilized that I am going to promise you. The last category which is the, it is as good as providing the best effort delivery called as the unspecified bitrate, I am not going to give you any kind of guarantee, occasionally, I might also even drop the packets. So, that is what the service models in the ATM networks. So, the integrated service model that I was talking about can be run on the IP networks and also on the other net type of the networks like the ATM networks as well.