

Real-Time Systems
Prof. Dr. Rajib Mall
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module No. # 01

Lecture No. # 37

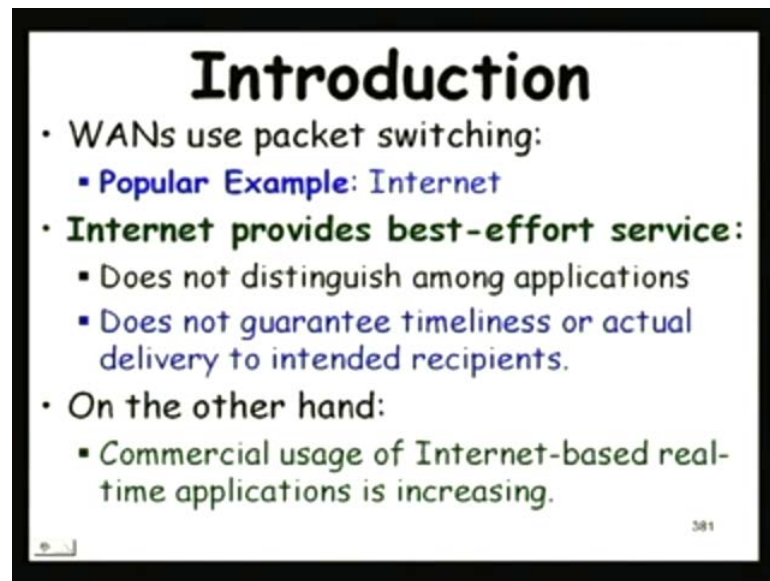
Real-Time Communication Over Packet Switched Networks

Good morning. So, let us get started from where we left last time and we were just starting to discuss about real-time communication over a packet switched network. So, we had seen over a LAN and now we will see how real-time communication can be achieved in a wide area network and (()).

(Refer Slide Time: 00:33)

The slide features a black border. Inside, the title "Real-Time Communication over Packet Switched Networks" is written in blue, with "(Lecture 37)" below it in a smaller blue font. A yellow rectangular box in the center contains the text "Dr. RAJIB MALL", "Professor", "Department Of Computer Science & Engineering", and "IIT Kharagpur." in black. In the bottom left corner, the date "03/08/10" is displayed, and in the bottom right corner, the number "300" is shown.

(Refer Slide Time: 00:41)



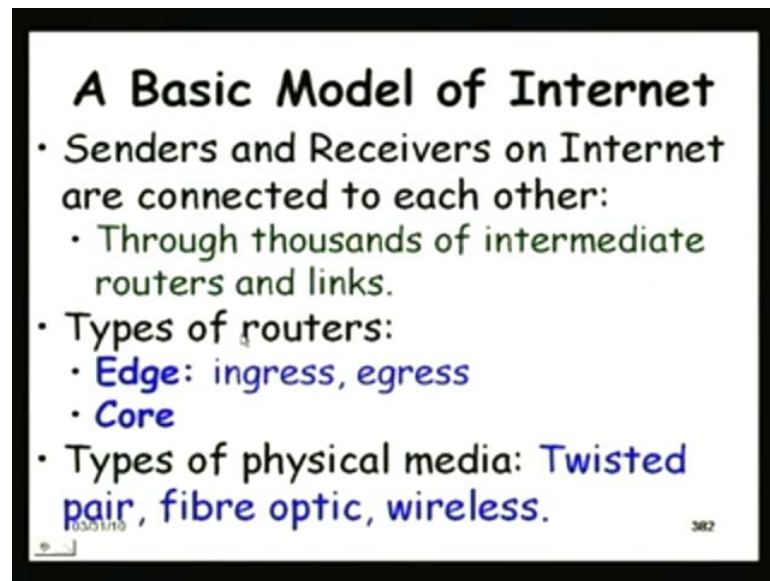
Introduction

- WANs use packet switching:
 - Popular Example: Internet
- Internet provides best-effort service:
 - Does not distinguish among applications
 - Does not guarantee timeliness or actual delivery to intended recipients.
- On the other hand:
 - Commercial usage of Internet-based real-time applications is increasing.

381

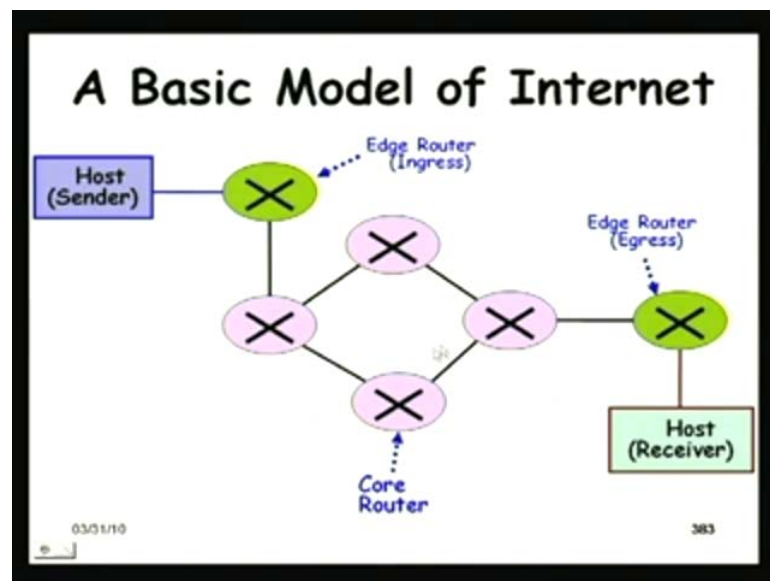
So, let us continue from there. So, you know that the wide area networks are based on packet switching and popular example is the internet where you widely used and we also know that internet provides best-effort service; does not distinguish between the sources and for every source it has to utilize the resources and provide the best service. It does not distinguish among applications and as a result it also does not guarantee timeliness or actual delivery to intended recipients and it just does the best that is possible. But, on the other hand, we see that the commercial usage of internet based real-time applications is increasing every day. We are doing online banking transaction video conferencing VOIP and so on. So, in that context let us see how a real-time communication can be achieved over the internet or any wide area network.

(Refer Slide Time: 01:44)



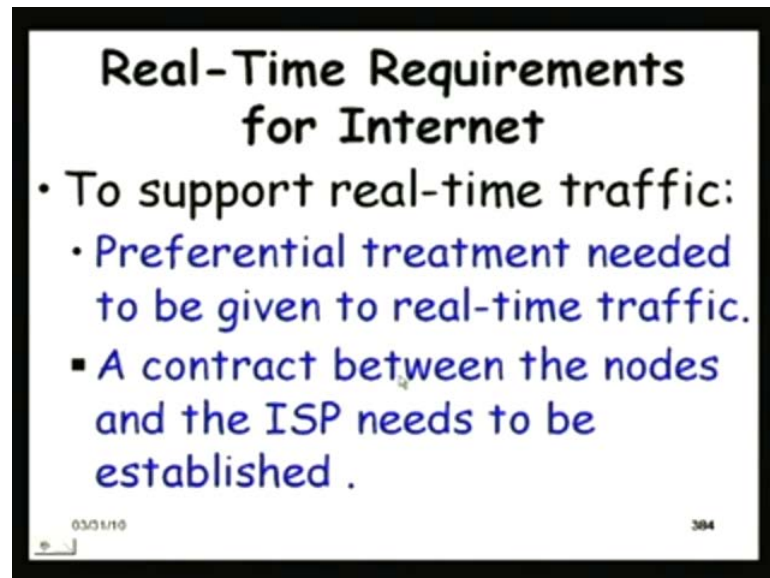
So, first let us understand the basic model of the internet here. The sender and receiver are connected to each other through various links and routers that are in between and if we look at the internet there are 3 types of routers: the edge routers which handle the requests from the sender and also connect to the receiver. So, depending on whether they are handling the sender or the receiver it is ingress or egress type and then the core routers which are in between and various types of physical media is used over different links for example, some may be twisted pair fiber optic wireless and so on.

(Refer Slide Time: 02:48)



If we look at this basic model we will see that the host is connected to the ingress router and then that the core routers and then the egress router connects to the receiver.

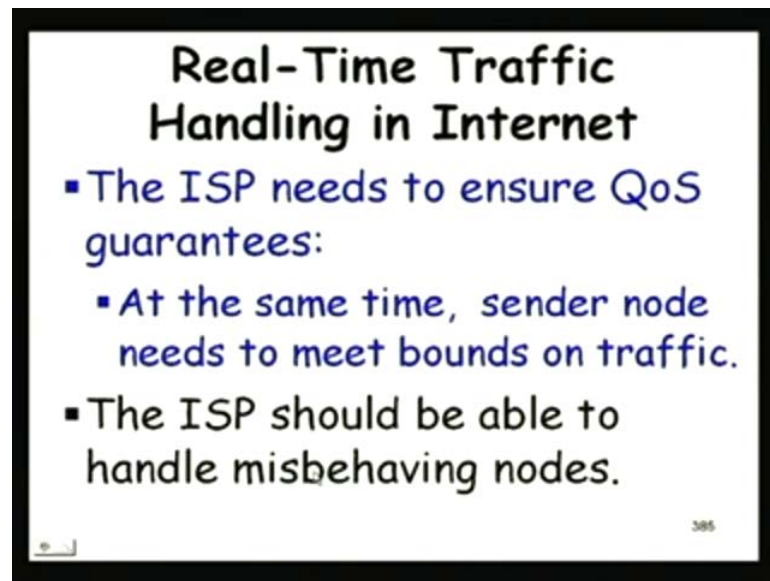
(Refer Slide Time: 03:05)



Now, let us see how the real-time requirements of applications can be met in the internet environment to support real-time traffic. Of course, there has to be some preferential treatment given to the real-time traffic because if it does not distinguish between the traffic sources; then it cannot give any guarantee real-time guarantee.

So, what it finally turns out that the sender needs to communicate or needs to differ hand need to identify its traffic needs, kind of traffic, the characteristics and the quality of service requirement and the network on the other hand needs to agree to this. The service provider needs to agree to this and once this contract between the sender and the ISP is agreed upon then, the ISP would ensure that the required quality of service for the given traffic source will be met. So, that is the basic way in which it will work.

(Refer Slide Time: 04:33)



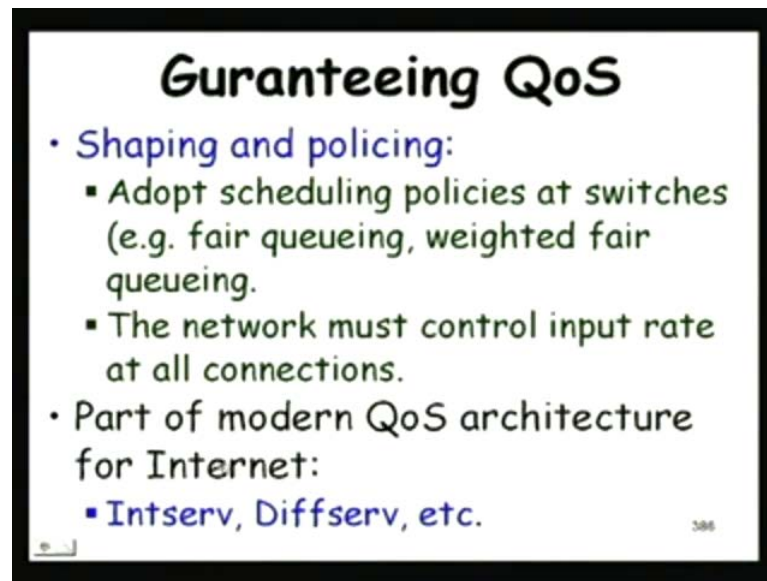
Real-Time Traffic Handling in Internet

- The ISP needs to ensure QoS guarantees:
 - At the same time, sender node needs to meet bounds on traffic.
- The ISP should be able to handle misbehaving nodes.

So, the ISP would somehow ensure the quality of service guarantee. We will look at that; how it can do that? But, at the same time we must remember that the sender node needs to meet the bounds of the traffic for which the quality of service was agreed upon. If it comes in different traffic or a traffic having different characteristics, they will what was agreed upon. Then of course, it would be difficult to meet the required quality of service not only for this source, but even for the other sources, other connections.

So, it becomes very important before the real-time communication how it is handled we discuss. Let us try to address how a traffic source can be characterized so that, for a given traffic source the agreement between the ISP and the sender can be obtained and if in case a sender has agreed upon some traffic characteristics and then it misbehaves, comes in more traffic or traffic having a different characteristic it had agreed for a CBR traffic and then it just generates a burst traffic, then it will inconvenience or it will make it difficult to meet the QOS requirements for the other applications also.

(Refer Slide Time: 06:20)



Guaranteeing QoS

- Shaping and policing:
 - Adopt scheduling policies at switches (e.g. fair queueing, weighted fair queueing.)
 - The network must control input rate at all connections.
- Part of modern QoS architecture for Internet:
 - Intserv, Diffserv, etc.

388

So, the ISP should be able to handle the misbehaving nodes and possibly it can restrict their traffic or even may reject a source to guarantee the quality of service requirements. We will see that one main technique that is used is traffic shaping and policing. One is shaping traffic - it can even police the traffic. So, let us see what traffic is shaping. So, in traffic shaping the network will change the characteristics of the traffic. If it is a burst traffic too much of bursts are there, it can smoothen out; for example, the way it might do it is that by adopting suitable scheduling policies at the switches. So, even though the packets are generated too rapidly by the source it might queue them and only generate a CBR traffic or may be a traffic characteristic that was agreed upon, right? So, that is the traffic shaping the shape of the traffic that was given by the source was not acceptable and therefore, the network might shape the traffic reshape. And even this can occur at the different routers because, the traffic from different sources arises there and it may be that the combined effect of different sources distort the shape of the traffic that was originally generated. So, it might reshape the traffic **right** using certain scheduling techniques.

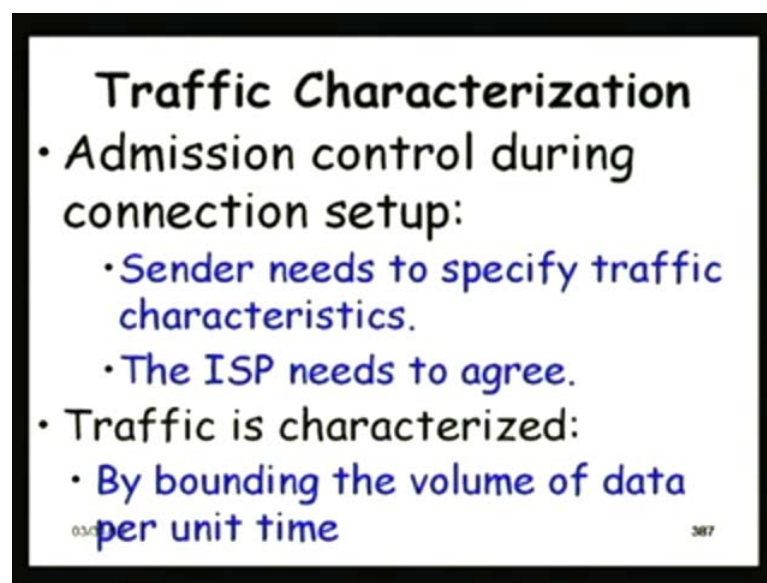
Sir is it all possible that (())

Yes; see, one is that traffic shaping right, a shaping does not involve rejecting the source. It is trying to make the best that it can do to make the traffic into the desired characteristics; not only at the sender end, but also in the intermediate routers **right**? Here, rejection is not considered, but in policing where it checks whether the traffic that

was given by the source is not of the required character it can be also rejected. So, in policing the network controls the input rate at the connections, we will see that this is part of the modern quality of service architecture for the internet where real-time transactions can be done.

Now, let us look at the way it is done in the modern internet we will discuss about the intserv and the diffserv which are used for providing quality of service to the receivers and the sender the connections basically.

(Refer Slide Time: 09:28)



Traffic Characterization

- Admission control during connection setup:
 - Sender needs to specify traffic characteristics.
 - The ISP needs to agree.
- Traffic is characterized:
 - By bounding the volume of data per unit time

03/07 387

Now, the first thing we need to discuss is about how do we characterize the traffic because, based on this the sender and the (()) right. So, let us look at the traffic characterizations. So, during the call initiation, the sender needs to specify the traffic characteristics and the network in turn will do an admission control that is it will check whether the given traffic characteristic can be given the required quality of service that is being demanded by the source. So, the admission control will be carried out at the call initiation based on the traffic characteristics.

So, once it finds that it can support the specified traffic for the given quality of service requirement the ISP would agree to the application that the required quality of service for the specified traffic can be given and we will see that the main way in which the traffic is characterized is by bounding the volume of data per unit time. What is the maximum

data that can be transmitted per unit time? But, if we say that just bound the maximum data rate, it is not really a correct statement; because, for burst traffic if we just bound the data rate it will lead to unnecessary resource reservation which are wasted most of the time because the bursts occur once in a while, is not it? So, need a more accurate characterization of the traffic.

(Refer Slide Time: 11:29)

Traffic Characterization

- An accurate bound on traffic:
 - Can help reduce network resource that needs to be reserved .
 - Helps to reduce resource idling.
- For traffic characterization:
 - Many models have been proposed.


388

If we are able to accurately specify the traffic that is bound on the data that is being generated by the source then, it will reduce the resource that needs to be reserved. Otherwise, conservatively too much resource might have to be reserved for burst traffic.

So, by accurately specifying the traffic we can reduce the amount of resource that needs to be reserved for a specific connection and it will increase the, improve the network utilization and also help to reduce resource idling. So, characterization - traffic characterization is a very important point in specifying the traffic **to is p**. Many models suggest for traffic characterization which are being used and these different models are suitable for different types of traffic.

(Refer Slide Time: 12:38)

(X_{\min}, S_{\max}) Model

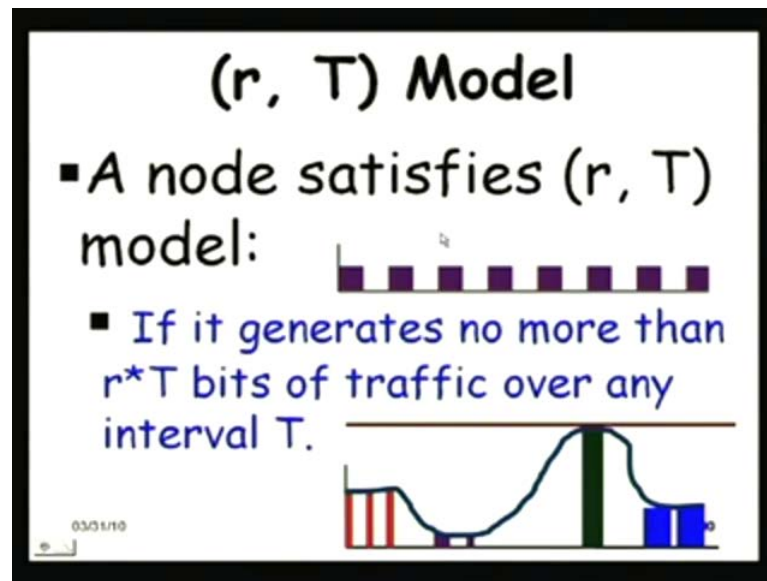
- A traffic satisfies this model:
 - If the inter-arrival times between two packets is always less than X_{\min} , and 
 - Size of the largest packet does not exceed S_{\max} .
- Provides tight bound for CBR traffic.

389

So, let us look at the simplest model: the $X_{\min} S_{\max}$ model. A traffic would satisfy this model if the inter arrival times between two packets is always less than X_{\min} and size of the largest packet does not exceed S_{\max} . It would provide a tight bound for CBR traffic. I am sorry there is a mistake here. Actually, the inter arrival time between two packets is always greater than X_{\min} . X_{\min} is the minimum difference between the arrival of two packets. This is the minimum separation it would not arrive any sooner than X_{\min} . ,If it arrived one packet the next one there must be a X_{\min} separation. In other words if there are bursts where too much of data gets generated then of course, X_{\min} would may be less than the one that is specified and this would not be appropriate model.

Similarly, the S_{\max} is the size of the largest packet would be S_{\max} then, the packet size will always be less than S_{\max} . Here, the inter arrival time will be more than X_{\min} this is the... and the size will be less than S_{\max} , right.

(Refer Slide Time: 14:29)

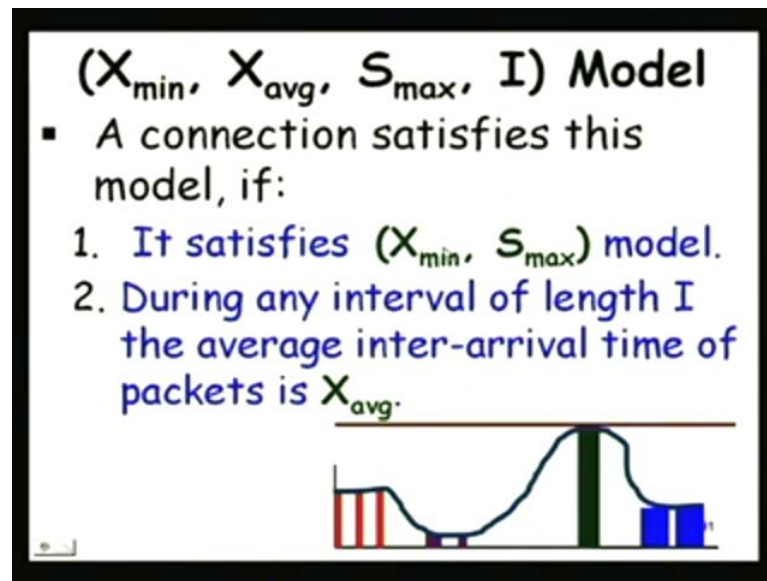


So, if you look at it then it is it can actually model a constant bit rate traffic. It can provide a tight bound to the CBR traffic because here these two parameters are enough because S_{\max} is always a constant and X_{\min} is also a constant. So, by the $X_{\min} S_{\max}$ specification, we can for any interval, we can tell what is the maximum traffic that is arising at that and accordingly resource reservation can be made

Now, let us look at the r comma T model in this model. The traffic would satisfy the model if it generates no more than r into T bits over any interval T ; just look at this any interval T . So, let me just emphasis this over any interval T . So, that means, T can be large or T can be small, but still r into T bits will be generated. In other words, r is the maximum rate that is data will be generated in any instant, right? Is that clear?

So, this is again a good model for CBR traffic, but for busted traffic it will make a very conservative - it will, we will have to specify a very conservative value of r **right** that is, the maximum data rate or the maximum burst that can occur. Data burst and then the resource has to be reserved; according to that it does not tell about the average that have it; is not it? So, for a CBR traffic it would work well and resource reservation can be made, but for burst traffic like this where there can be burst, it will lead to reserving resources at the maximum rate. So, it is not a good model for burst traffic, but for CBR traffic it is all **right**.

(Refer Slide Time: 17:08)



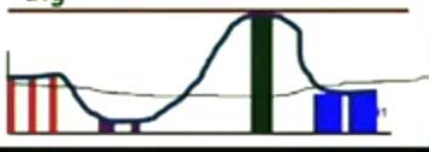
Now, let us look at another model the X_{\min} X_{avg} S_{\max} and I model we will say that a connection would satisfy this model if it satisfies the X_{\min} S_{\max} model. So, this specifies what the X_{\min} S_{\max} specifies. What, yes, can anybody answer? What does X_{\min} S_{\max} specify? It is the maximum rate at which the data will be generated, right? The X_{\min} S_{\max} model specifies the maximum rate at which data is generated and we had said that if the maximum rate is equal to the average rate then this is a good model. But for burst traffic the maximum rate will be much higher than the average rate. So, you need to also characterize the average rate of data generation. Now, let us see how this model specifies the average rate of data generation.

So, a connection would satisfy this model if we specify its maximum data rate through the X_{\min} S_{\max} . X_{\min} and S_{\max} parameters and also the X_{avg} and I parameters we will specify over any interval I . what is the average inter arrival time? see X_{\min} is the minimum inter arrival time **right**? X_{avg} is the average inter arrival time. So, this would specify the average rate of data, is not it? So, in the other model we could only specify the maximum rate. Here, even these two parameters X_{\min} S_{\max} they specify the maximum rate at which data is being generated.

(Refer Slide Time: 19:19)

$(X_{\min}, X_{\text{avg}}, S_{\max}, I)$ Model

- A connection satisfies this model, if:
 1. It satisfies (X_{\min}, S_{\max}) model.
 2. During any interval of length I the average inter-arrival time of packets is X_{avg} .



We have the second parameter the X average and I which will specify the average rate; may be this one may be somewhere here, **right**? So, you have, we are bounding it by the maximum rate and the average rate is that. So, this would be a better characterization of a burst traffic even though it is not really an accurate approximation or an accurate characterization of burst traffic. Why is that? why it is not an accurate characterization of burst traffic? Yes...

(())

No, we are specifying both the maximum and the average, but still it is not a good characterization of burst traffic.

(()) average

Yes

For that method burst may not be the same as we would have expected burst is.

Ok.

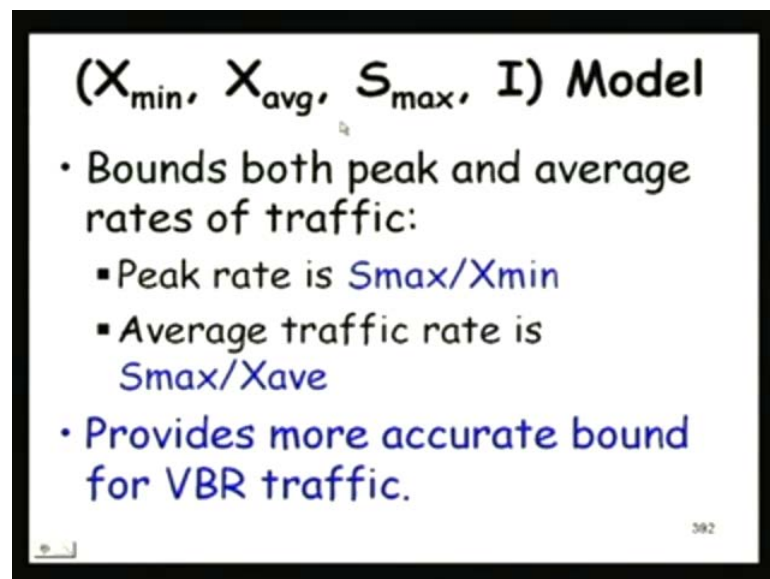
(())

Yeah.

May not (()) So it basically...

So, what as he says that it basically, gives two information: what is the average rate at which data will be generated and what is the peak rate at which data will be generated. But, how frequently the peaks will occur or what will be the sizes of the peaks in different instances? It does not tell anything regarding that. So, let us see the different models, another model.

(Refer Slide time: 20:52)



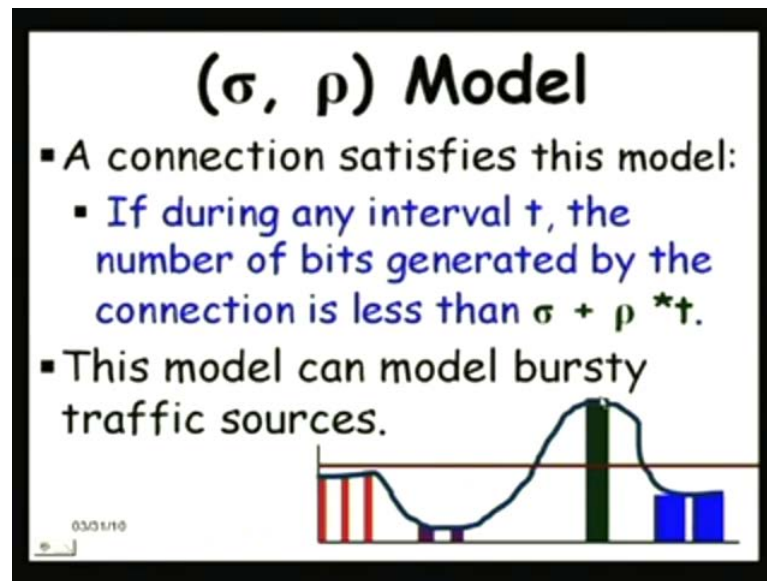
$(X_{\min}, X_{\text{avg}}, S_{\max}, I)$ Model

- Bounds both peak and average rates of traffic:
 - Peak rate is S_{\max}/X_{\min}
 - Average traffic rate is S_{\max}/X_{ave}
- Provides more accurate bound for VBR traffic.

392

So, whatever we discussed is again here. So, it bounds the peak and the average rate of the traffic. The peak rate is S_{\max} by X_{\min} as we are seeing and the average rate is S_{\max} by X_{average} . So, where an average interval of X_{average} S_{\max} is getting generated right? X_{average} is the average separation between two packets. It definitely provides a more accurate bound for VBR traffic than the $X_{\min} S_{\max}$ model. But still it is not really a good characterization of burst traffic where bursts of different sizes can occur.

(Refer Slide Time: 21:46)



Now, let us look at the sigma cum a rho model. A connection would satisfy this model if during any interval t , so, again mark this one any interval t . So, that is up to us whatever we consider the number of bits generated by the connection is less than sigma plus rho into t . So, what does this give us if we say that over any interval it should be less than this, what will it give us? (())

That is, but what does this total expression give?

(())

This is the maximum number of bits rate is generated because any interval no. So, it will, you can include the largest burst that exists. So, it is sigma plus rho into t is the maximum rate at which data will be generated right?

Sir (())

We will see. So, this model can, this can model burst traffic because sigma is the average rate and rho into t will specify the maximum burst that can occur over the average over the averages not the bursts size, but that the increase in data rate over the average rate.

They agree to that because we are saying that this is the maximum rate and rho is the average rate and sorry, sigma is the average rate and rho into t will give this one right, is

not it? The increase over the average rate in a burst, are you agreeing to that, yes or no?
 (()) The sigma is the average value. So, what is the average that which it will be
 generated and the rho into t is the maximum data that can occur over the average rate; is
 that ok?

(Refer Slide Time: 24:29)

Multiple Rate Bounding

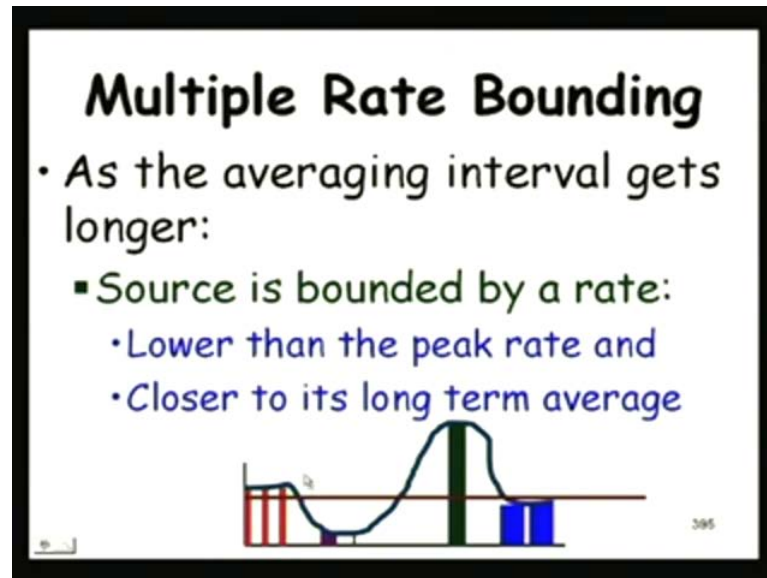
- Bursty traffic sources can be more accurately characterized:
 - By bounding the traffic over multiple averaging intervals.
- A traffic would satisfy:
 - $\{(r_1, T_1), (r_2, T_2), \dots\}$ if $T_1 < T_2 < T_3 \dots$,
 - Over any interval I the number of bits generated is bounded by $r_i * T_i$, if $T_{i-1} < I < T_i$.

Now, let us look at, if the traffic is really I mean, the resource reservation is really crucial to give an exact characterization of the traffic - burst traffic. We can bound the rate of a burst traffic over multiple bounding intervals. So, you can bound the traffic yeah, over multiple averaging intervals and traffic would satisfy the multiple rate bounding. If we specify several r_1, T_1 we had discussed about the r, T model right?

So, we specify several bounding intervals where T_1 is less than T_2, T_3 etcetera. So, over any interval I , the number of bits generated will be less than $r_i * T_i$ where i is between T_{i-1} and T_i . So, you just give me any interval; whether you tell 5 seconds or 3 seconds or some milliseconds, I can tell you what is the maximum rate at which data will be generated over that. I will just look at the corresponding T_{i-1} and T_i and then I will take T_i and I will say that see the data rate will be between $r_i * T_{i-1}$ and $r_i * T_i$, sorry, $r_{i-1} * T_{i-1}$ is not it?

So, here the different sized bursts can be modeled, is not it? But, of course needs many parameters and more complicated. So, unnecessarily we would not use it unless it is really required.

(Refer Slide Time: 26:35)



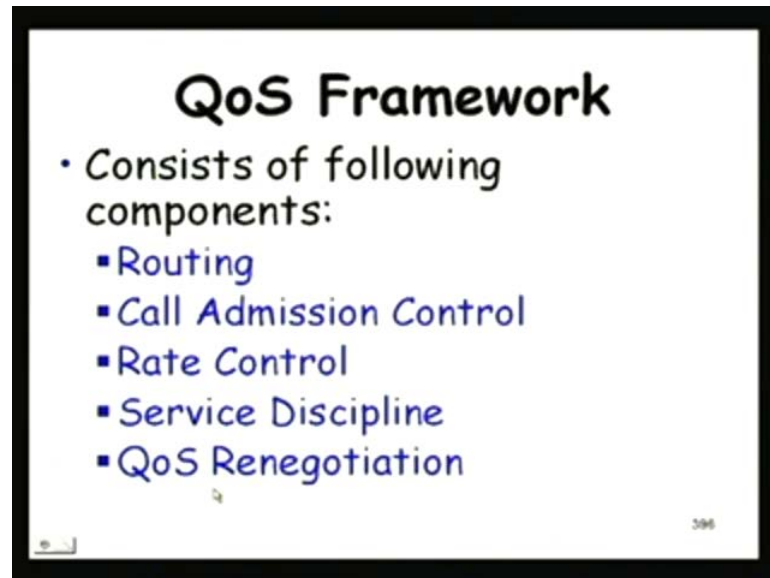
(()) That depends yeah. So, just look at the question. The question is that for how many bounds we have to give how many T_i ? So, the answer to that is that it depends on the type of traffic that we are trying to model the different bursts that might exist.

So, if the burst is unpredictable, bursts can be of any size. Then we will have to use many T_i 's; we cannot really say how many T_i 's **right**? Large number of T_i 's, but if we know that the bursts can be two or three types, if the bursts are up to bursts are up two or three types, sometimes a burst up three packets per 10 milliseconds can occur. Sometime, it can be 10 packets per 2 milliseconds etcetera. So, then we can specify those bounding intervals **right**, but if nothing can be said about the traffic it would become extremely difficult to model the traffic.

So, one thing to notice here is that as the averaging interval gets longer that is $T_1 T_2 T_i$ etcetera, as T_i becomes longer and longer the source is bounded by a rate which is lower than the peak rate, but closer to the long term average. The peak rate will occur when we are considering the smallest T_i , is not it? The smallest interval over which you are specifying the traffic, that will tell us about the peak rate and as the bounding interval

increases, it will come closer and closer to the long term average. So, for this kind of traffic let us say we have few bursts sizes then we can specify for those bounding intervals and that will give us an exact characterization of this kind of traffic.

(Refer Slide Time: 28:46)

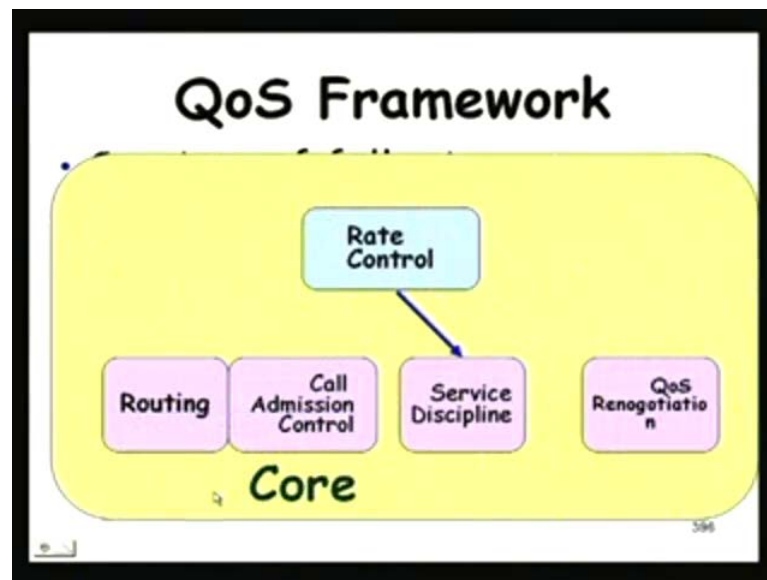


Now, we have discussed about how traffic can be modeled **right**. Now, let us see once the source specifies the traffic model and its quality of service requirement. What does the network do to check whether it can accept the request and how it will ensure that the quality of service demand will be made? So, let us proceed with that. That we will call as the quality of service framework. So, this is what the network will do to honor or not to honor the request of the source and then having honored, what it will do to meet the demanded quality of service? The quality of service framework would consist of the following component: one is routing. Routing is an important component in providing the required quality of service. Call admission control: it will check whether at all for the specified traffic the requested quality of service can be delivered. Rate control: what if the traffic after having requested rate, it misbehaves? So, you need to control the rate.

The service discipline: the service discipline is as the rate has been agreed upon through the intermediate routers, at what rate it will be serviced. The service discipline will check how to the queue traffic, how it will get transmitted, **right** at the intermediate nodes the intermediate routers. And then, the quality of service renegotiation because sometimes the traffic the source might generate large amount of data then what was planned and it

will try to renegotiate the quality of service. Or maybe it does not generate traffic at that rate anymore; at least for some time. Then it will try to renegotiate the quality of service and it might change its traffic characteristics; it might change its quality of service requirement. So, that is the quality of service renegotiation.

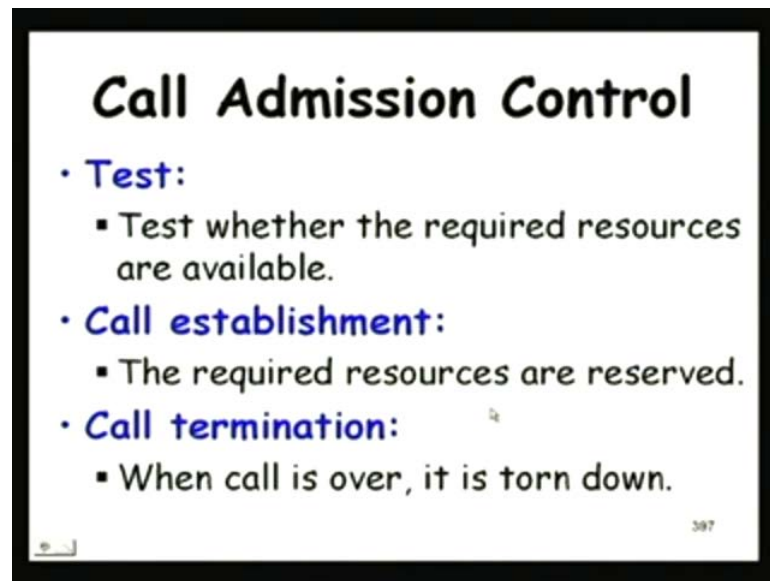
(Refer Slide Time: 31:20)



So, if we look at the different routers. So, this is a core router and even at the edge routers we will have these components there which handle the quality of service requirement. So, one is routing the call admission control, this is not really required at the core.

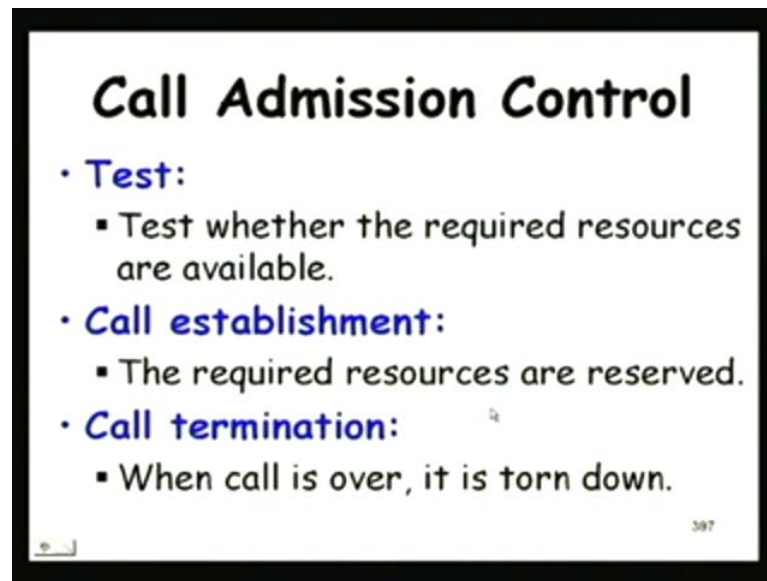
The service discipline, see the rate control it generates the traffic that needs to be transmitted and the service discipline as I was saying you, it will schedule the traffic to be transmitted on the link on the outgoing link and then the quality of service renegotiation component which we will handle any requests for renegotiating the a great quality of service and the traffic.

(Refer Slide Time: 32:09)



So, first let us look at the call admission control. We will discuss this slightly more detail as we proceed, but let us have the overall idea in call admission control. You need to test whether the required resources are available to meet the demand and if the required resource is available to meet the quality of service demand for the specified traffic characteristic. Then, the call establishment will be carried out. But what we mean by call establishment? In call establishment the required resources will be reserved at the various routers and then finally, the call termination when the call is over it is torn down. What it really means is that the resources are being reserved. So, these are the important task that will be carried out by the call admission control module.

(Refer Slide Time: 33:23)



Let us look at the rate control and the service modules the rate control module will ensure that the source does not violate the agreed upon traffic rate and we were saying that there are different ways it can be handled. One is by traffic shaping where even if the traffic that is being generated is not of the agreed upon shape, it will try to reshape and then the policing control. And one thing that we will as we proceed we will observe is that as the traffic proceeds through the network it becomes burstier and burstier. So, the different routers a traffic which was burst to a very small extent as it proceeds through the network it will become burstier and burstier why is that anybody can.

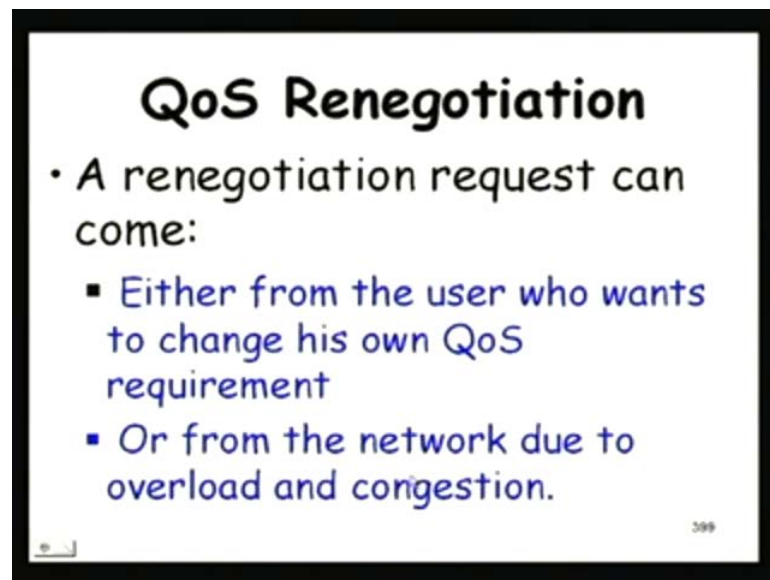
(())

Exactly, so, he has the answer that it will get stored at different routers and then it will be transmitted at the peak rate and then at sometime there will be no traffic. So, as it precedes through the network a traffic which was only to a very small extent burst it will become burstier and burstier and therefore, traffic shaping will be needed at the intermediate routers.

The service discipline: it will check the reservations that were made by the, during channel establishment. The resource reservations that were made for a specific connection and then it will take up the queue traffic from different sources and then it will transmit at that rate. So, it will schedule the packets that were queued up at the

agreed upon rates **right**? These are the reservations made for example, it is a bandwidth may be reserved. So, it will, the service discipline will ensure that the intermediate routers where reservations were made the traffic. The queued up traffic will be transmitted based on the bandwidth reserved.

(Refer Slide Time: 36:00)



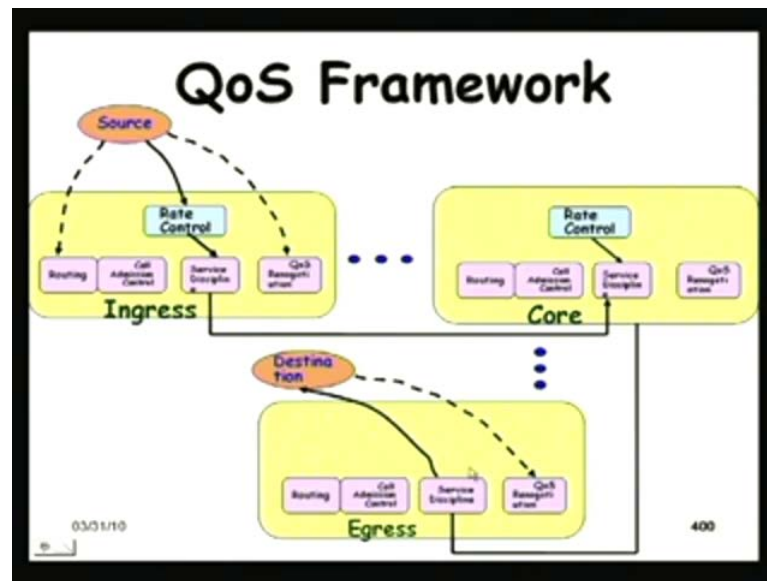
QoS Renegotiation

- A renegotiation request can come:
 - Either from the user who wants to change his own QoS requirement
 - Or from the network due to overload and congestion.

399

Now, let us look at the quality of service renegotiation. A renegotiation request can come from a connection which would like to change its requirement regarding the traffic or the quality of service requirement or it can be from the network due to overload and congestion; that suddenly there is too much of load in the network or may be some links failed it might try to renegotiate the quality of service with the application.

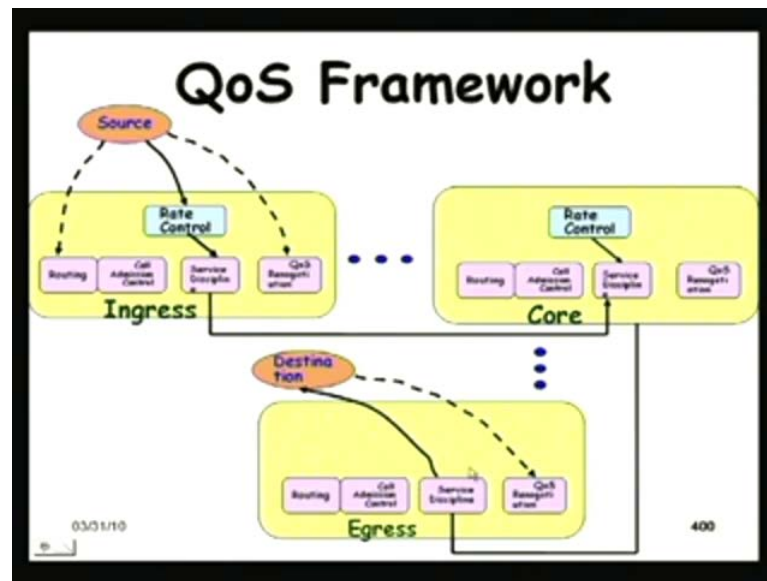
(Refer Slide Time: 36:38)



So, if we look at this whole picture we will have different core routers here, 100 or 1000 of them the internet and then there is the edge routers the incoming traffic handled by the ingress router.

The source: it first requests the connection routing and call admission control takes place and then it gives the traffic. It will subject to rate control and then the service discipline and as it gets forwarded to different core routers it again undergoes rate control and the service discipline and finally, it reaches the destination and the destination can even make quality of service renegotiation. We will just look at this the context of multicast routing.

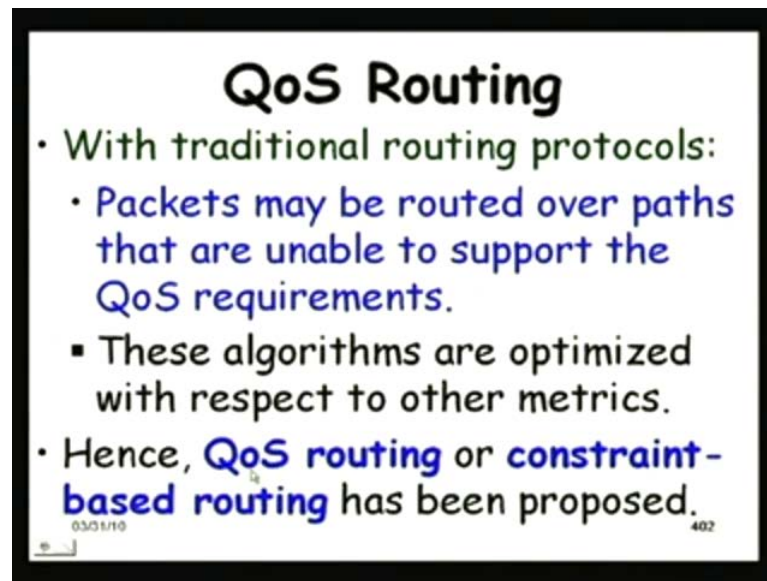
(Refer Slide Time: 37:46)



Now, let us look at routing. We had said that packet routing is a very important means by which quality of service guarantees can be given to applications. During routing we know that the route selection takes place and the route selection takes place during the connection establishment for both unicast and multicast routing that we will consider now. But, one thing is that the routing algorithms that are used in the traditional protocols they consider the shortest path routing they optimize with respect to a number of hops in the route selection.

But, we will see that this not really appropriate for a real-time request because here, shortest path is not really the only criteria. Here there are several quality of service criteria and it needs to meet all those not just the shortest path.

(Refer Slide Time: 38:57)



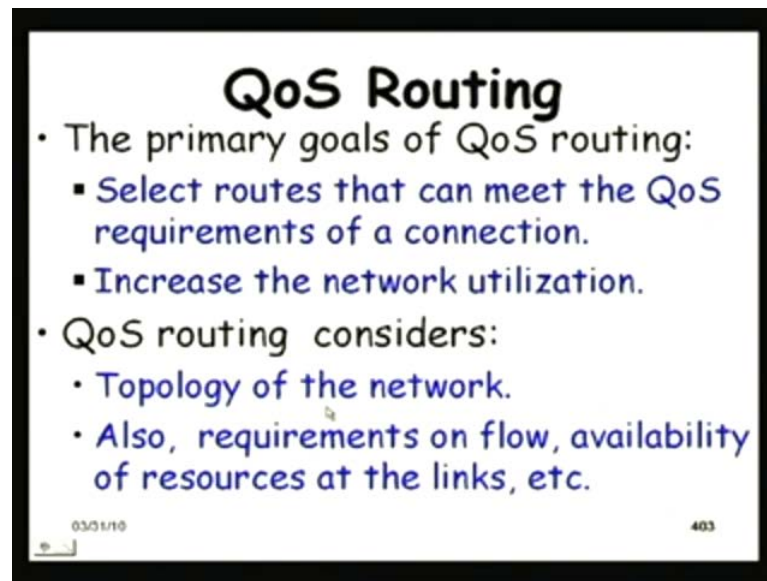
QoS Routing

- With traditional routing protocols:
 - Packets may be routed over paths that are unable to support the QoS requirements.
 - These algorithms are optimized with respect to other metrics.
- Hence, QoS routing or constraint-based routing has been proposed.

So, if we use a traditional protocol we will observe that it would route packets over paths that are unable to meet the quality of service requirements even though there might exist some path where it might have been possible to provide the quality of service requirement.

So, if we think to consider quality of service during routing which will call as QoS routing, the traditional routing we will just refer as routing and when we need to consider the various quality of service parameters we will call it as the QoS routing. In QoS routing, the algorithms need to be optimized with respect to the other metrics which will help insure the required quality of service. So, we will sometimes call it as constraint-based routing or quality of service routing.

(Refer Slide Time: 40:05)



QoS Routing

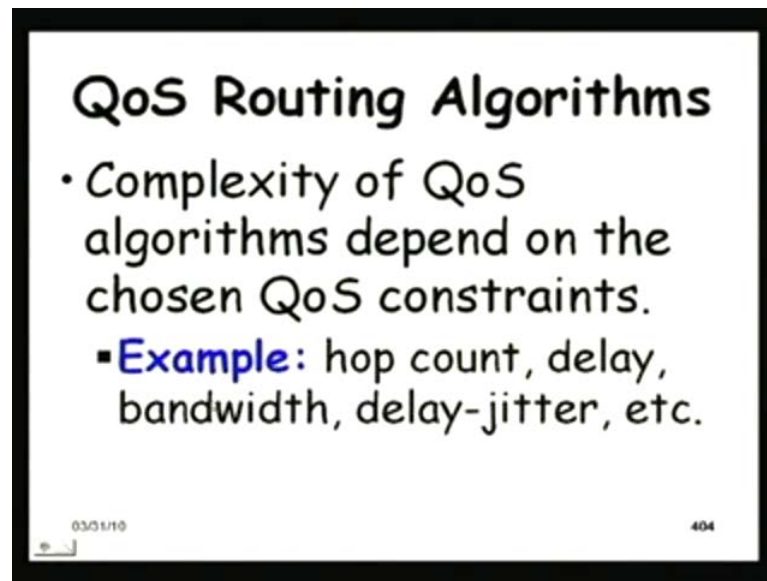
- The primary goals of QoS routing:
 - Select routes that can meet the QoS requirements of a connection.
 - Increase the network utilization.
- QoS routing considers:
 - Topology of the network.
 - Also, requirements on flow, availability of resources at the links, etc.

03/01/10 403

Let us look at the quality of service routing here. The primary goal is to select a route that will meet the quality of service requirement of a connection. So, it needs to consider what quality of service requests were made during the connection establishment and then it will select a path that will meet all the quality of service requirement for the specified traffic.

Subject to increasing network utilization it is not just finding any path that will meet the quality of service requirement, but also it has to ensure that the load in the network is balanced and it will increase the network utilization. Of course, the quality of service routing just like the traditional routing needs to consider the topology of the network in addition to the quality of service requirement and then it has to consider the various requirement and the availability resources at the links and so on.

(Refer Slide Time: 41:18)



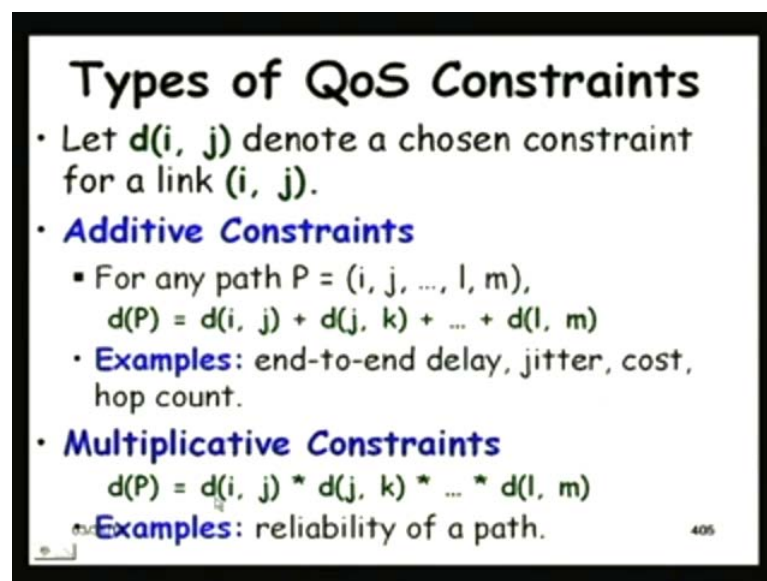
QoS Routing Algorithms

- Complexity of QoS algorithms depend on the chosen QoS constraints.
 - **Example:** hop count, delay, bandwidth, delay-jitter, etc.

03/01/19 404

The complexity, there are several algorithms actually exist for QoS routing, but the complexity of a quality of service routing algorithm would depend on what kind of quality of service the application requires. The specific constraints: the different constraint quality of service requirements may be hop count delay bandwidth delay-jitter, etcetera.

(Refer Slide Time: 41:51)



Types of QoS Constraints

- Let $d(i, j)$ denote a chosen constraint for a link (i, j) .
- **Additive Constraints**
 - For any path $P = (i, j, \dots, l, m)$,
$$d(P) = d(i, j) + d(j, k) + \dots + d(l, m)$$
 - **Examples:** end-to-end delay, jitter, cost, hop count.
- **Multiplicative Constraints**
 - $$d(P) = d(i, j) * d(j, k) * \dots * d(l, m)$$
 - **Examples:** reliability of a path.

405

Now, we will try to classify the different constraints quality of service constraints into three main types, whatever quality of service constraints we had discussed. So far let us

try to classify them depending on whether it is easy to or difficult to meet the constraint and what are the general characteristics of the constraint. Now, let us say d_{ij} is a constraint for a link i, j . So, that is the case the d_{ij} is a constraint between a link i, j then we can have one class of constraint is a additive constraint. In a additive constraint if there is a path consisting of many links i, j, j, j plus l, m , etcetera then the quality of constraint for the entire path will be a summation of the quality of constraint for the individual paths, is that **ok**?

So, we have some quality of service requirement for the connection i, j whole. We can split that into quality of service requirements on the individual links. Now, what are examples of this end to end delay? So, if we need some end to end delay, we can split the end to end delay between various links. Jitter, what is the maximum jitter that will be permitted? You can split the jitter also cost, hop count, etcetera, these can be split. So, these are all examples of additive constraints. There can be multiplicative constraints. So, here the constraint in the entire path will be expressed as a multiplication of the constraints and the individual links. What is an example of this reliability? So, if we have reliability on the entire path then we need to multiply the reliability of the individual links in the path.

(Refer Slide Time: 44:05)

Types of QoS Constraints

- **Concave Constraints**
 $d(P) = \min\{d(i, j), d(j, k), \dots, d(l, m)\}$
• **Example:** bandwidth
- **The problem of finding a path:**
 - **Subject to two or more additive and/or multiplicative constraints in any possible combination is NP-Complete (Wang and Crowcroft).**

03/01/10 406

Now, we have some constraints which are concave type concave constraints in the concave constraint the constraint on the entire path is actually a minimum of the constraints on the individual paths.

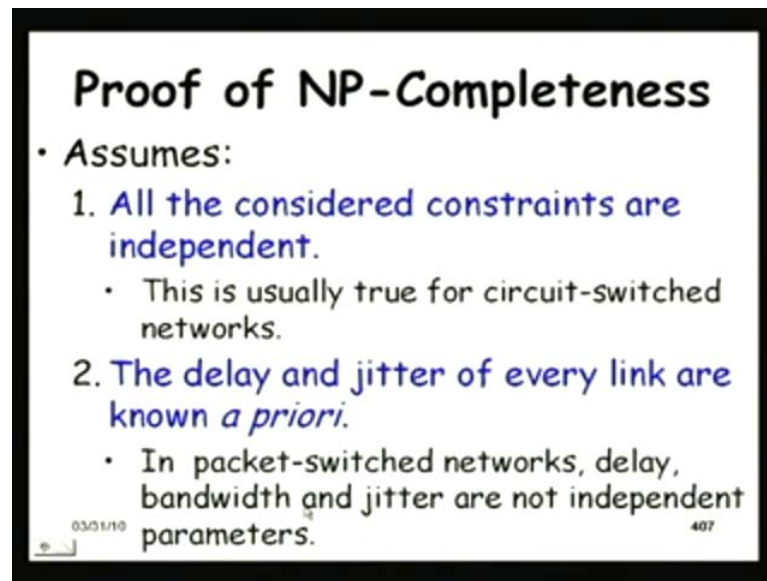
So, we do not bother what is the maximum that can be provided by a link. Our constraint will be guided by the minimum that is available in any of the links what an example bandwidth is. So, even if in some of the intermediate links too much of bandwidth is available, but as far as the constraint on the path is concerned the minimum bandwidth is what we will consider. (()) Concave because we are considering the minimum. The minimum that is available if it was a maximum that we are considering that will be a convex constraint. So, concave we can think of that if the bandwidth on different links that is available is let us say is like this. So, we need to consider the minimum here, but on a convex we will consider the maximum that is available in any of the links and we will say that the path constraint is in the maximum that is available we do not have a example of a convex constraint; none of the quality of constraints that we discussed. So, we did not discuss about a convex constraint.

(())

No convex constraint is hard to think of. Anything in the current quality of service that we discuss the different QoS parameters these will fall into any of these three. But, I do not know in future there might be new quality of service requirements which might satisfy convex, but at the moment there are no convex constraints a delay is additive constraint and it is an established result reported by Wang and Crowcroft that the problem of finding a path or a route subject to two or more additive or multiplicative constraints is NP complete.

So, this is an important result; that it is a difficult problem. If we have two or more additive or multiplicative constraints where you have to satisfy this because concave constraint we just take the minimum, that is all.

(Refer Slide Time: 47:17)



Proof of NP-Completeness

- Assumes:
 1. All the considered constraints are independent.
 - This is usually true for circuit-switched networks.
 2. The delay and jitter of every link are known *a priori*.
 - In packet-switched networks, delay, bandwidth and jitter are not independent parameters.

03/01/10 407

But one thing is that if you look at that result the proof of the NP completeness assumes that the constraints that we are considering are independent. That is one assumption that is made in the NP completeness proof. But, in our situation that is not the case. See, if it, we were trying to establish this result in the context of a circuit switched network this would be true because the different paths are independent here, **right**? But, we are considering a packet switched path - a packet switched link and it is not true actually or a packet switched link and the other assumption is that the delay and jitter link are known *a priori*. But here we know that these are related constraint. The delay and jitter are not really independent parameters; the delay bandwidth jitter these are related. So, these are not independent really, but these are related. So, possibly I should have written this point under this one.

So, the constraints are not really independent because, if you delay and bandwidth for example, they are interrelated, is not it? If you have more bandwidth delay will be less; similarly, the jitter is also affected by these parameters.

(Refer Slide Time: 48:58)

QoS Routing

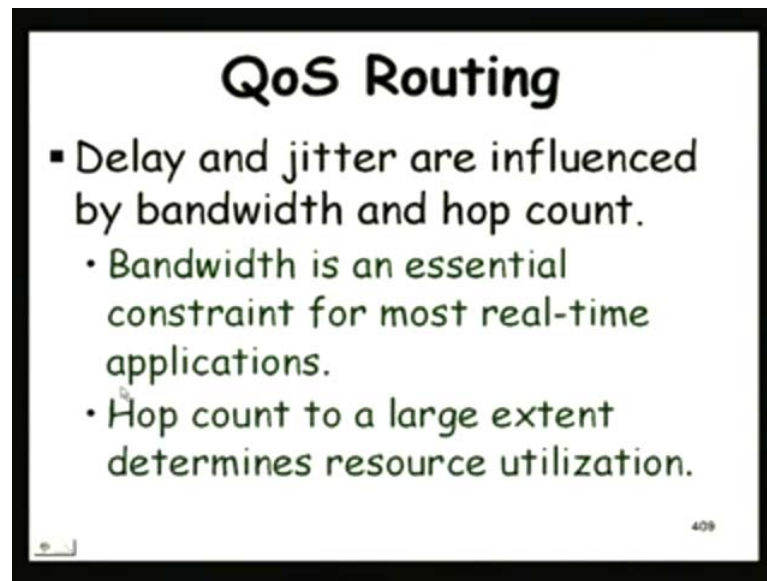
- Polynomial algorithms like Bellman-Ford and Dijkstra's algorithm:
 - Compute routes using hop count, delay and jitter constraints.
- Bandwidth and hop count are more important constraints.
- As compared to delay or jitter
- Why?

03/01/19 408

But even then, even though the result that was established is not strictly true for the case that we are considering for a packet switched network, but still it indicates that it would be hard to develop a optimal algorithm efficient optimal algorithm for quality of service routing we know that in a traditional network we have polynomial algorithms like Bellman-Ford, Dijkstra's, etcetera which are based on hop count delay, jitter, etcetera.

We cannot really make straightaway use, these need different algorithms, but one thing that we must keep in mind is that out of the different constraints, the bandwidth and hop count are important. The others are dependent on this because if we know bandwidth we can also tell about delay **right** and this is also easier to handle in a network during the quality of service establishment and so on.

(Refer Slide Time: 50:25)



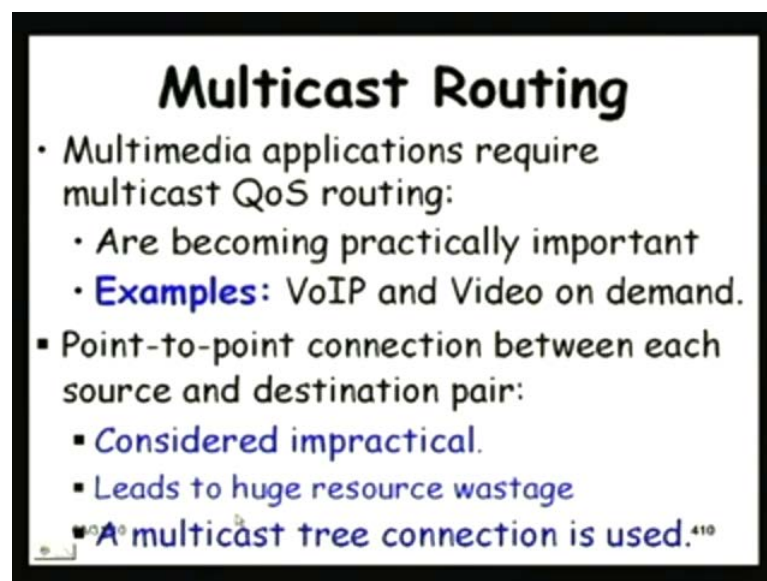
QoS Routing

- Delay and jitter are influenced by bandwidth and hop count.
 - Bandwidth is an essential constraint for most real-time applications.
 - Hop count to a large extent determines resource utilization.

409

So, that is what it says that delay and jitter are influenced by bandwidth and hop count and bandwidth is a parameter that is easily controlled. Even for the service discipline, if bandwidth we specify it can be easily unforced and it is considered as an essential constraint for all real-time application and also the hop count determines resource utilization.

(Refer Slide Time: 50:50)



Multicast Routing

- Multimedia applications require multicast QoS routing:
 - Are becoming practically important
 - **Examples:** VoIP and Video on demand.
- Point-to-point connection between each source and destination pair:
 - **Considered impractical.**
 - **Leads to huge resource wastage**
 - **A multicast tree connection is used.**⁴¹⁰

One thing that we must consider is the multicast routing because whenever we have multimedia applications, multicast routing comes naturally. These are becoming

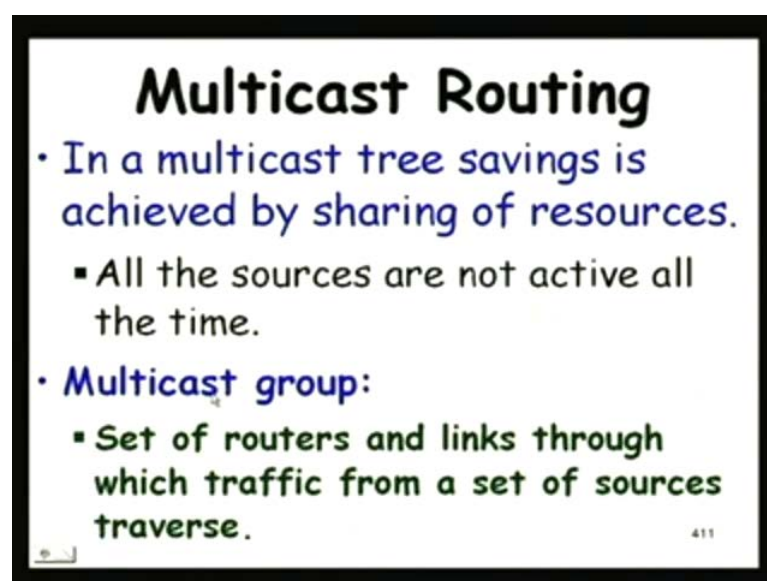
practically important; for example, VOIP where you have multiple persons participating in a VOIP session. They are doing a conference call using the VOIP or may be video on demand where there is a video server on which different persons are getting the streamed video, right?

But they might be using the same video stream and therefore, this would require a multicast routing because the same source is being shared by different receivers and in this situation the one that we discussed about a conference call about the same source video source being saved by multiple receivers, if we think of a point to point connection between the source and destination pair.

For example, here if a we establish a point-to-point connection between a source and all the receivers or here between the different sources and receivers it will be a extremely inefficient solution in practical. Actually, because too much of bandwidth resources etcetera you wasted because they share the links in between all these different connections they share the resource here you must consider that and that is the (()) of the multicast routing.

So, unless we use a multicast routing for this kind of application it will lead to huge resource wastage and we have to use a multicast tree connection

(Refer Slide Time: 53:02)



Multicast Routing

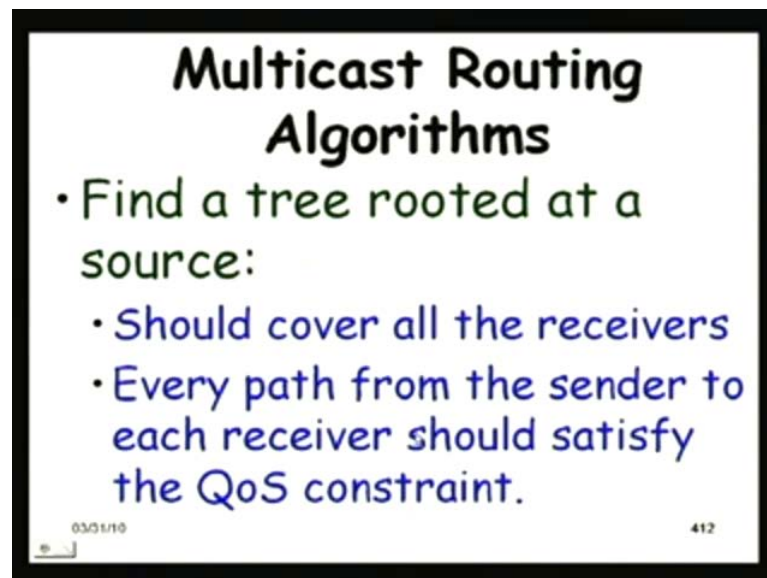
- In a multicast tree savings is achieved by sharing of resources.
 - All the sources are not active all the time.
- Multicast group:
 - Set of routers and links through which traffic from a set of sources traverse.

411

The essential idea behind the multicast routing is that savings in resources is achieved by sharing because after all the same data with same quality of service requirements etcetera is getting transmitted. So, it is unnecessary to duplicate them. They should be able to share the same set of resources; we should not handle them as independent sources and receivers. We should handle them together that is the idea and even if we consider a VOIP not all resources sources are active at the same time and we must make use of that.

So, in that context we need to define a multicast group. A multicast group is a set of routers and links through which the traffic from a set of sources traverse to some common receivers **right**. May be a single source, the set may be a single source to which it traverse to some receivers. Or it may be that the traffic traverses from a few sources to some of the receivers. So, basically they are sharing the data when this source sends one source sends it reaches to some of the destinations and another source also reaches these destinations.

(Refer Slide Time: 54:33)

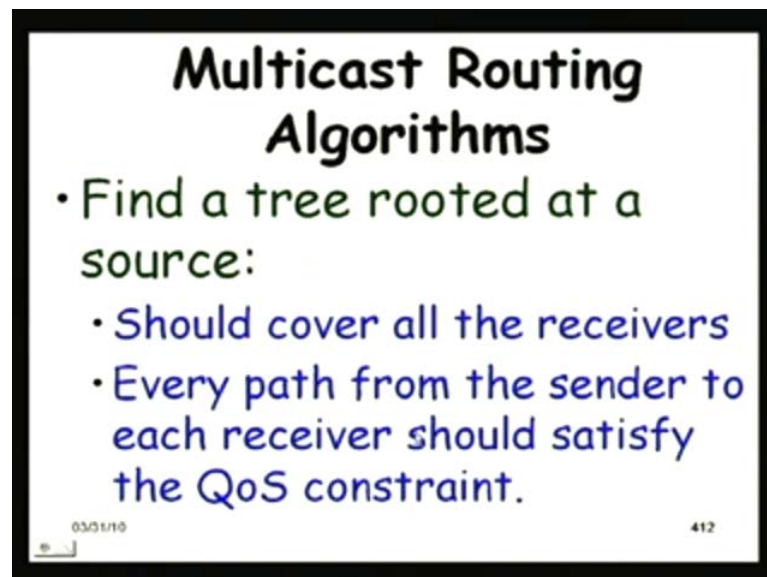


The multicast routing algorithms if we look at it basically find a tree that is a rooted at the source. So, from multiple sources we need to have all these trees that are originated at the sources and connect the destinations or the receivers.

So, the tree should be rooted at the sources and should cover all the receivers and for simplicity, we can consider a single source where the tree would be rooted at the source

and it should cover all the receivers. And, every path from the sender to each receiver should satisfy the quality of service constraint, **right**. We have to consider all the receivers and the paths that are existing between the sources and then find that along all the paths the quality of service constraint should be able to satisfy.

(Refer Slide Time: 55:34)



Multicast Routing Algorithms

- Find a tree rooted at a source:
 - Should cover all the receivers
 - Every path from the sender to each receiver should satisfy the QoS constraint.

03/01/10 412

So, one important **thing** here is the tree construction the multicast tree construction and essentially, if we look at the various tree construction algorithms that exist, two types. One is the source based tree construction where these algorithms initiate construction from each source and then reach the destination and there can be a core based algorithm. Also, several core based algorithm exist where the core is first formed in a network which satisfies the required quality of service and then the receivers join the core.

So, we will not discuss a specific algorithm because there large number of algorithms exist which are being used. We will, if you are interested you can look at specific algorithms, but they will fall into these categories. So, we will stop here and we will continue on our discussion about how the quality of service is being provided in a wide area network situation and especially in the internet; thank you.