

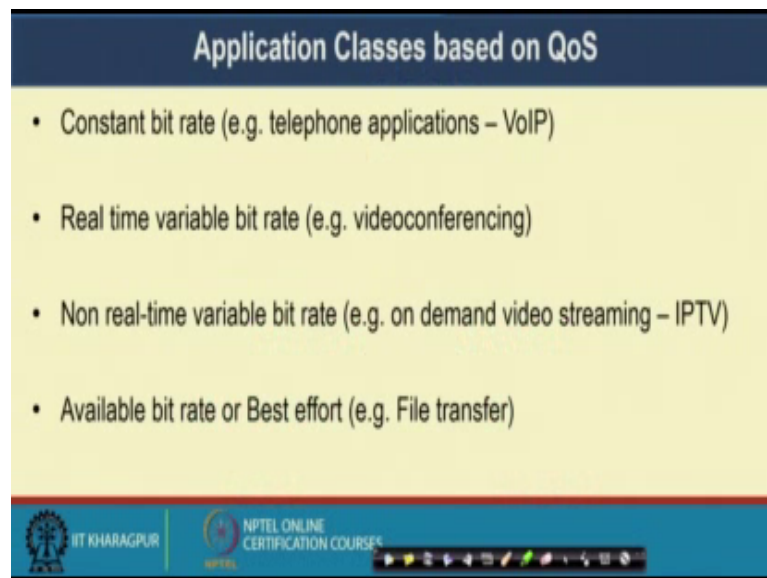
Computer Networks and Internet Protocol
Prof. Sandip Chakraborty
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 32
Internet QoS – II (Basic QoS Architecture)

Welcome back to the course on Computer Network and Internet Protocols and in the last 2 class we were discussing about quality of service in the internet and we have looked into the basic definition of quality of service and what is mean by quality of service and what are the parameters that impact quality of service and we have seen that 4 parameters a bandwidth delay jitter and loss they have an significant impact on the network quality of service. And whenever we are trying to ensure quality of service over the internet we need to have a control over these 4 parameters.



Now, we look into that what support do we require from the TCP IP network to ensure the quality of service in the network. So, we look into the architecture in details in today's lecture.

(Refer Slide Time: 01:17)



Application Classes based on QoS

- Constant bit rate (e.g. telephone applications – VoIP)
- Real time variable bit rate (e.g. videoconferencing)
- Non real-time variable bit rate (e.g. on demand video streaming – IPTV)
- Available bit rate or Best effort (e.g. File transfer)

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, as we are discussing in the last class that based on the quality of service we can have multiple classes of applications, we can have constant bit rate application, constant bit rate application means in us so, in short form we call it as a kind of severe application. So, in case of severe application you expect the data at a constant bit rate. So, the

receiver expects that the data will be coming at a constant bit rate and then the receiver will process that data and we will be able to render the data further.

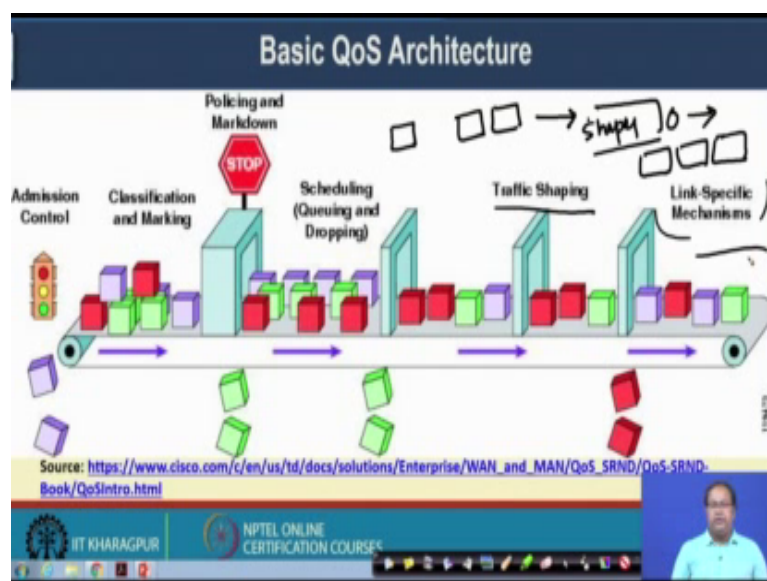
So, the example of constant bit rate requirement QoS requirement is telephone application or these voice over IP application say whenever you want to make a VoIP call to make a VoIP call you have to transfer the data or the voice data over the IP network. So, whenever you are transferring the voice data over the IP network the voice is actually converted to a digital signal, the voice signal is digitized and after digitizing the voice signal that data need to be transferred at a constant bit rate.

Then the second class is real time variable bit rate. So, here the bit rate can be variable, but you need to transfer the data in real time. So, by real time we mean that there is a maximum delay by which you need to send the data. If you fail to send the data by that time then your transfer your data transfer we consider to be failure. So, the example of real time variable bit rate is the video conferencing or live streaming.

Then the third class is non real-time variable bit rate. So, here you do not need to send the data in real time you do not have a such strict delay requirement that by within this time you have to transfer the data buffering is supported, but you require certain QoS still you have a kind of loose delay on the loose bound on the delay or loose bound on the amount of packet loss that can be tolerated and that is example can be on demand video streaming or IPTV kind of services.

And then finally, comes to the fourth class of service, which we call as the available bit rate or the best effort service. So, best effort means whatever is the available bandwidth now in the network you utilize that bandwidth for transferring the data for say file transfer. For file transfer as such we do not need any strict requirement on the quality of service parameters you can transfer it with whatever available bandwidth is there in the network. So, that kind of transfer we call it as available bit rate transfer or Best Effort transfer.

(Refer Slide Time: 04:06)



So, this is the basic quality of service architecture in the internet this figure I have taken from a Cisco website and the link is given here. So, you can see that there are multiple stages in this end to end pipeline. So, let us explain this figure and then we will go to the internals of every individual step. So, the packets are moving from one end to another and at every individual step we are applying certain filters, filters means we are actually looking into the property of this packet and taking certain kind of corrective measurement to ensure quality of service in the end to end application perspective.

So, the first thing is kind of admission control. So, if you remember whenever we discussed about that what is our expectation from the network to ensure quality of service? One expectation was to find out whether the network can take or whether the network can accept more flows without violating the quality of service of the existing flows. Now to ensure that we have the first module which is called the admission control module.

So, what this admission control module does?

It basically admit a new flow in the network by ensuring that even if you are admitting this new flow in the network and if you know that what is the quality of service requirement for that particular kind of flow. So, this requirement we call it as service level agreement or SLA. So, we will discuss all these terminologies in details later on just giving you a broad overview kind of backside view of the entire system.

So, based on the service level agreement you know that, what is the expected QoS level for this particular flow? Now whenever you are entering this flow in the network during that time you try to estimate that, if you enter this flow in the network, then whether you will be able to satisfy the quality of service for all the existing flows plus this new flow in the same network or not. If you are able to do that then you admit or allow that new flow to transfer the packet otherwise you block that flow.

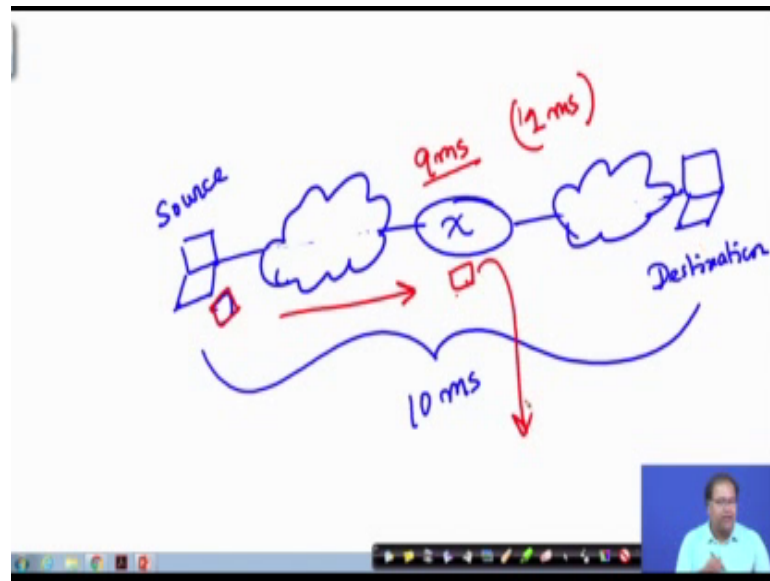
So, that particular module is termed as admission of control after the admission control is done then the next step is classification and marking. So, what is classification module does, the classification module basically identifies the classes of packets which are there. So, as you have discussed that there are 4 different classes of application, they are kind of broad classes in different network context you can have much more granular definition of quality of service classes.

So, considering those 4 classes of service the guaranteed bit rate or the constant bit rate, the real time variable bit rate, the non real time variable bit rate and a best effort are available bit rate. You classify and mark the packet that which packet require say constant bit rate, which packet require non-real time variable bit rate, which packet required real time variable bit rate and which packet requires this available bit rate. So, basically does not require any particular bit rate.

So, based on that you mark the packet so, just like a blue packet, a red packet, a green packet, or a yellow packet, like that and then go to the next filter next level of filters. Now all these filters are actually implemented that all the layer 3 devices; that means, inside the router. So, inside the router so, we implement all these individual filters that we are currently talking about.

And the third filter is traffic policies and markdown. So, what traffic policy in looks into, traffic policy in looks into that whether certain kind of flows or certain kind of packets in the flows is significantly violating the quality of service requirement or not. If it is violating the quality of service requirement then you simply drop those packets. So, the idea is there that if you know that well your end to end delay requirement is say 10 millisecond, let me explain it with the help of an example figure say your end to end delay requirement is 10 milliseconds.

(Refer Slide Time: 08:56)



So, you have a host a source host then coming to an intermediate router and then there is this destination host. And in between you have an intermediate network through 2 different networks they are say connected ok. So, in this network now say that end to end delay requirement is 10 milliseconds so, my end to end delay requirement is 10 millisecond.

Now one up once so, you are sending a particular packet say you are sending a particular packet from the source and once this packet reaches to this router then this router finds out that the packet has already experienced 9 millisecond of delay. If the packet has already experienced 9 millisecond of delay and this router is very sure that it is impossible to transfer this packet to the destination within 1 millisecond.

So; that means, if the packet has already experienced a 9 millisecond delay, then within 10 milliseconds you have to send it to the destination; that means, within one millisecond you have this remaining one millisecond you have to send the packet from this router to the final destination. And the router knows that it is totally impossible if that is the case then what the router does, the router simply drop this packet.

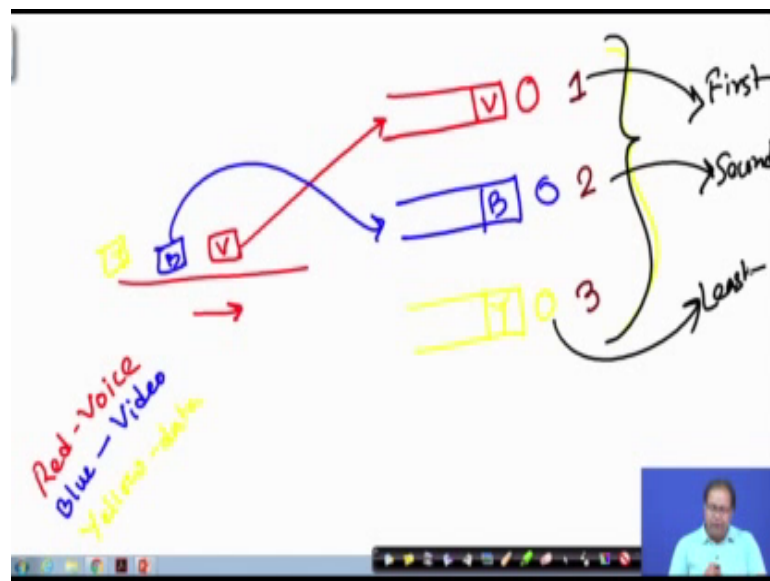
Why the router drops this packet? The reason is that you know that you will not be able to satisfy the quality of service for this particular packet. If you are not able to satisfy the quality of service then there is no meaning to send that packet unnecessary to the link, because it will again clog the link it will take certain bandwidth, it will reach at the final

destination, the application again need to process that packet. So, there is a fail lot of over rate which are associated with this entire process.

So, if you are sure at a intermediate step that you will not be able to satisfy the quality of service requirement for some specific packets of some specific traffic classes then you immediately drop those packets. Coming back to the stages so, that is called the policies and policing and markdown.

Then the next step is scheduling scheduling where we applied the queuing and the dropping policies. So, scheduling means based on individual requirement of traffic classes we basically prioritize individual packets, prioritizing individual packet means say again let me give you an example.

(Refer Slide Time: 11:53)



Say your packets are coming to an input link and you have multiple service queues ok. So, say this is a red queue, I have a blue queue and I have a say yellow queue. Now these 3 queues has 3 different priority levels. So, say the red queue has a priority of 1, the blue queue has a priority of 2, and the yellow queue has a priority of 3. So, priority 1 means it is the highest priority and priority 3 means it is the lowest priority.

Now, the scheduling works in this way say, if you are receiving a red packet then this red packet you put it in the red queue ok. So, red packet means say that this is a voice packet. So, red means voice say voice has the highest priority because we have seen that it has a

very strict requirement on delay jitter and bandwidth. So, say red means voice. So, whenever you are getting a voice packet you are putting it in the say red queue. Whenever you are getting a blue packet say this blue packet is assume it is a video packet the second class of traffic you are putting it in the blue queue.

And whenever you are getting a yellow packet, say yellow packet means data it is the lowest priority as we have seen. So, whenever you are getting a yellow packet you are putting it in the yellow queue. Now here you have the server that we will finally, serve this individual 3 queues in a round robin fashion and there you apply this priority queuing mechanism. So, the priority queuing mechanism says that if you have some packets in the red queue you first transfer those packets.

So, this is the first transfer if the red queue is empty then only you go to the blue queue and transfer the packets besides the second priority and if both the red queue and the blue queues are empty then you transfer the packets from the yellow queue. So, this is the least priority. So, that way you are ensuring that the voice traffics which are there in the link they are getting highest priority and they are transmitted first. So, that they are strict queue is delay requirement gets satisfied.

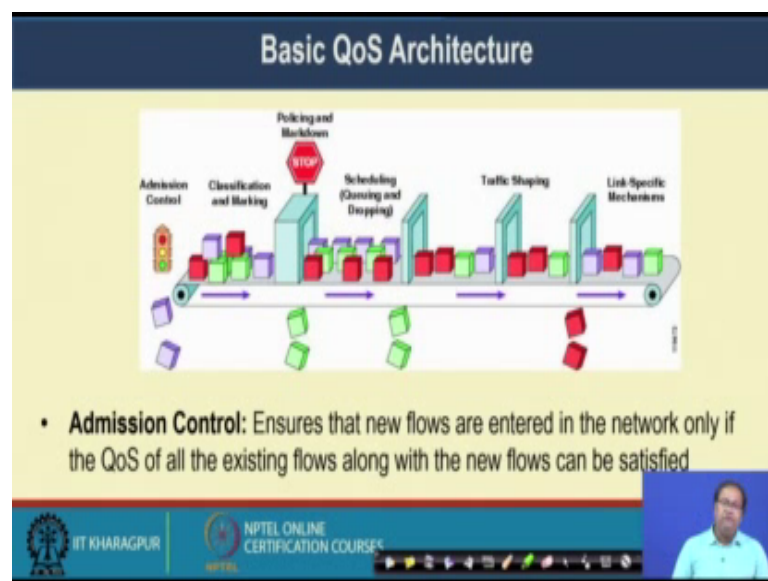
So, this is just an example of different kind of scheduling mechanism this is not the only scheduling mechanism which is a used or which is implemented in the network. So, this is one type of scheduling mechanism which is called priority queuing. There are other kind of queuing mechanism like custom queuing, weighted fair queuing, fair weighted fair queuing and so on. So, we will discuss those in details later on. So, the scheduling means scheduling the packets in different application queues based on our different flow cubes based on their QoS requirements ok.

Then the step which we call as the traffic shaping. So, traffic shaping basically tells that say in a link you are getting the packets. So, traffic shaping actually ensures that smooth jitter in the network, it controls the jitter in the network say you are getting the packets at random delay, now you send it to a shaper the shaper will output the packet at a constant rate. So, you are actually getting the packet at with the jitter introduced. So, the shaper will remove the jitter and send it to the outgoing queue um. So, that is the objective of the traffic shaper so, that you can regulate the flow of traffic over the outgoing link.

So, it is a kind of traffic regulator and finally, we apply certain link specific mechanism like if it is a wi-fi links, then wi-fi has their own QoS service QoS provisioning service like a prioritizing traffic for link layer, channel access and so on. So, these link specific mechanisms are finally, applied. So, this is the broad mechanism brought QoS architecture through which we try to guarantee end to end quality of service in the internet.

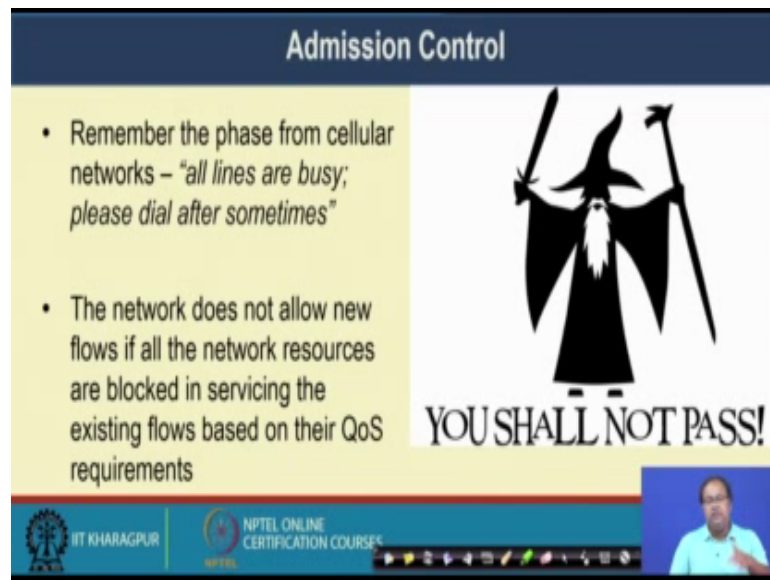
So, we will now look into all these individual steps in little more details.

(Refer Slide Time: 16:40)



So, let us start with admission control as we are mentioning that admission control ensures that new flows are entered in the network only if the quality of service of all the existing flows including the new flow can be satisfied.

(Refer Slide Time: 16:59)



The slide is titled "Admission Control" in a dark blue header. The main content area has a yellow background on the left and a white background on the right. On the left, there are two bullet points: "Remember the phase from cellular networks – 'all lines are busy; please dial after sometimes'" and "The network does not allow new flows if all the network resources are blocked in servicing the existing flows based on their QoS requirements". On the right, there is a black silhouette of a wizard with a long white beard, holding a sword and a staff. Below the wizard, the text "YOU SHALL NOT PASS!" is written in a bold, serif font. At the bottom of the slide, there is a blue footer bar containing the IIT Kharagpur logo, the text "IIT KHARAGPUR", the NPTEL logo, and the text "NPTEL ONLINE CERTIFICATION COURSES". A small video inset in the bottom right corner shows a man in a light blue shirt speaking.

Admission Control

- Remember the phase from cellular networks – “all lines are busy; please dial after sometimes”
- The network does not allow new flows if all the network resources are blocked in servicing the existing flows based on their QoS requirements

YOU SHALL NOT PASS!

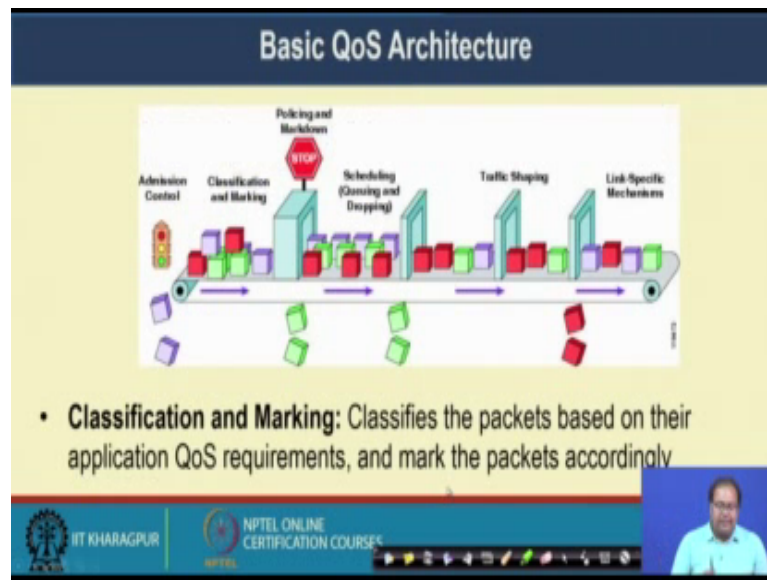
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, you have actually experienced this context of admission control, many of the time whenever you are dialing over say the cellular network you have hard disk space like all lines are busy please dial after sometime.

So, whenever you are hearing this voice that is actually blocking your call because the cellular service provider it does not have sufficient amount of resource to ensure the minimum quality of service for your call. So, that is why it is blocking your call and for long distance call it is pretty the common that it will say that, all lines are busy please dial after some time. So, it is like that the network does not allow new flows if all the network resources are blocked in servicing the existing flows based on their quality of service requirement.

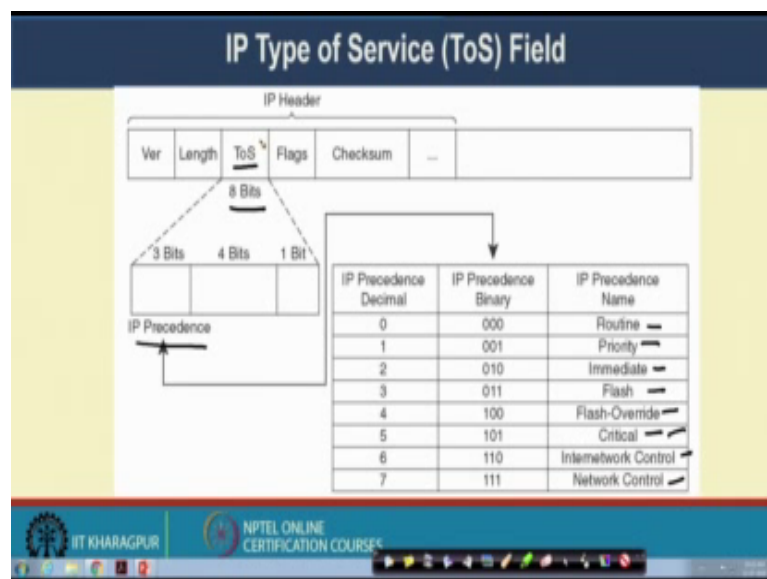
So, that is the first phase of maintaining quality of service. So, you do not admit a new flow if you cannot support quality of service for all the flows including the new flows.

(Refer Slide Time: 18:01)



The second mechanism was classification and marking as I mentioned that classifies the packet based on their application QoS requirements and then mark the packets accordingly.

(Refer Slide Time: 18:13)



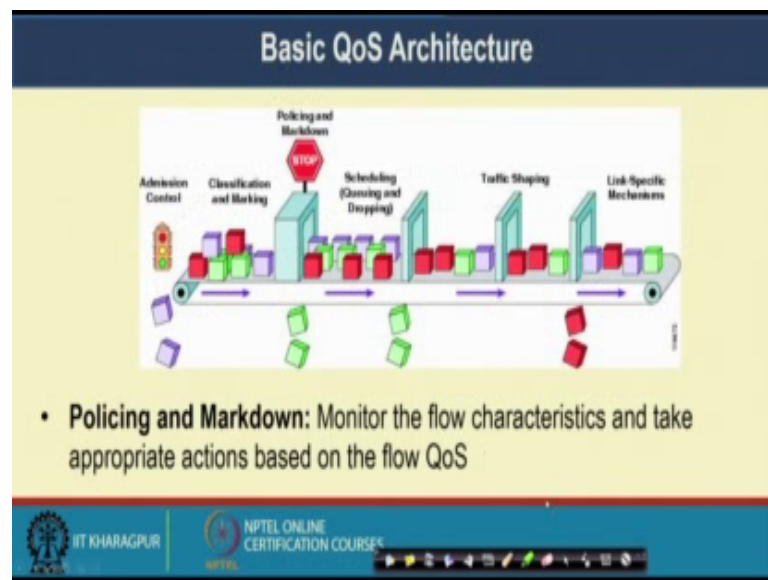
So, to classify it in case of IP we use a header in the IP header field it is called the IP type of service. So, we have a 8 bit field, 8 bit type of service field in the IP header in that type of service field in the IP header you have 3 bits for IP precedence.

So, this IP precedence value it define different kind of traffic say if the precedence value is 0 then it is a kind of routine packets. Then if the precedence value is 1 it is the highest priority packet, if the precedence value is 2 it is the immediate priority packet, if it is a priority 3 it is a kind of flash packet, if it is priority 4 flash override packet, for 5 the kind of critical packet, for 6 it is internetwork control packets say when it is network control packets.

So, that way we define these 8 different classes of traffic based on the IP precedence. Then the next 4 bits it actually defines the priority inside the classes for example, if you want to send voice video voice and streaming video simultaneously, you can use it under this critical class and under the critical class you can again relatively prioritize a voice give more priority to voice over the video traffic.

So, that way this 8 bit IP type of service header can be type of service field in the IP header can be utilized to identify or to mark the packet to a specific quality of service class.

(Refer Slide Time: 19:57)



Now, the third filter which we apply for quality of service it is policing and markdown. So, policing and markdown means monitor the flow characteristics and take appropriate action based on the flow QoS.

(Refer Slide Time: 20:13)

Traffic Policing

- **Service Level Agreements (SLA):** An agreement or a contract between the customer and the service provider to maintain QoS of an application

```
1 ip sla 11
2 icmp-echo 10.0.10.11 source-ip 68.68.1.2
3 frequency 5
4 !
5 track 10 ip sla 11 reachability
6   delay down 15 up 15
7 !
8 ip sla schedule 11 life forever start-time now
9 ip sla enable reaction-alerts
```

- **Traffic policing** monitors the flow of traffic and mark them to take appropriate actions (reduce priority, drop etc.)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, for traffic policy as we are mentioning initially that we have a terminology called service level agreement. So, this service level agreement says that you have an agreement or a contract between the customer and the service provider to maintain the quality of service of an application. Say for example, if you want certain quality of service for the VoIP data, you have to go for a service level agreement with your service providers say if you are purchasing network from Airtel or from say Vodafone with the service provider you have to make an agreement that, well I want to transfer the VoIP traffic, for that VoIP traffic I require this class of service. So, for that what is the money that I have to pay?

So, you have to pay that much of money and in that case you have a service level agreement or SLA with your service provider. Now this SLA will determine that how your packets will be treated when the packets are going over the network. Say one interesting example of service level agreement is whenever you are subscribing for say broadband connection. So, whenever you are subscribing for broadband connection there are you will see that there are multiple packages say you can have 1 Mbps Mbps of leased line for 1 month, you have can have 256 Mbps of leased line for 15 days and so on.

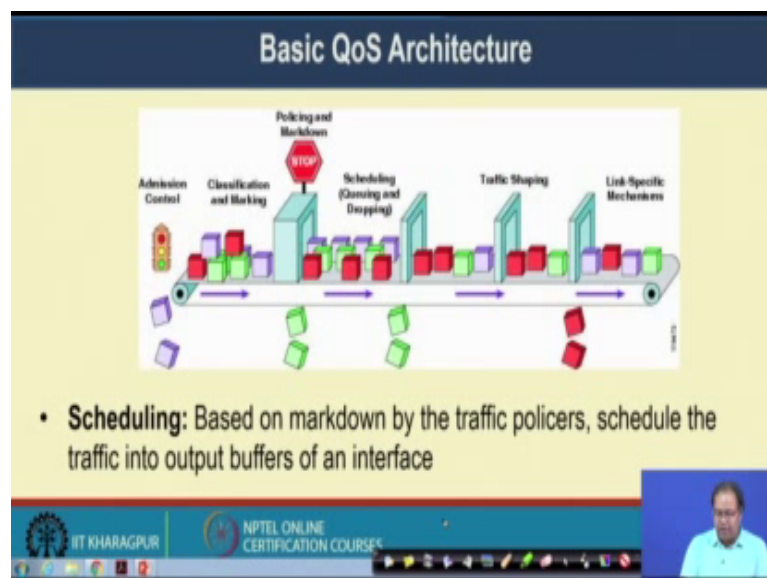
So, these packages are kind of service level agreements. So, they are saying that whatever data you will send for that data we will try to give you 1 Mbps of peak

bandwidth or 256 Mbps of peak bandwidth. Many of the time you will see that there is a differentiation between the uplink traffic rate and a downlink traffic rate. So, it will say that you can have 256 Mbps of uplink rate and say 1 Mbps of downlink rate. All those things are kind of service level agreement that you have with the service provider whenever you are subscribing for a specific package for getting your internet connection.

Now, this kind of service level agreements are actually also embedded in the IP packets and this is one example of configuring service level agreement in a particular router that track 10 ip sla 11 reachability and agreement is delay down 15 up 15; that means, in the downlink it can tolerate up to 15 millisecond delay in uplink it can also tolerate into up to 50 milliseconds of delay.

So, this is one service level agreement which has been configured in a particular router. So, I have just taken the trace of a router to show you that this way you can configure the service level agreement in the each router or the gateway routers of a specific network service provider. So, the network service provider so, whenever you are going for a service level agreement the network service provider actually writes down all those service level agreement in the network based on this their policies based on their architecture and so on.

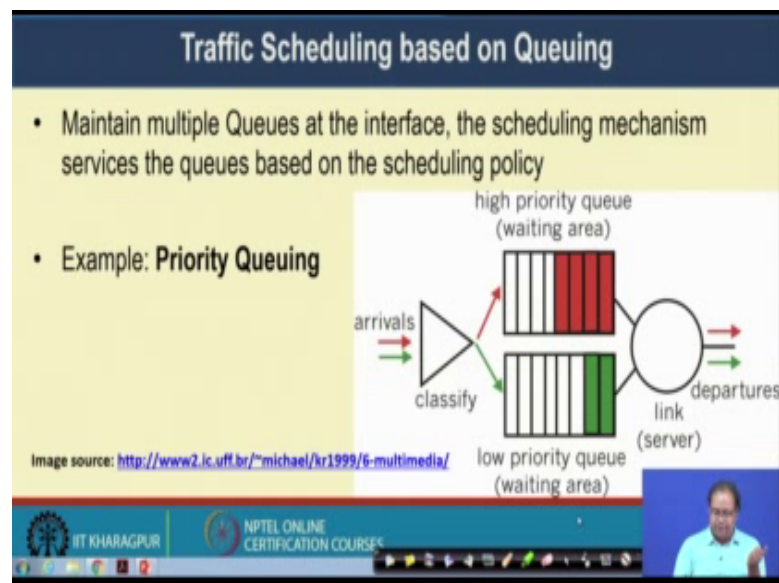
(Refer Slide Time: 23:40)



Now, this traffic policy it monitors the flow of traffic and marked them to take appropriate action, like whether you want to reduce the priority or whether you want to drop the packets and so on ok.

Next the fourth step was traffic scheduling as we have mentioned that based on the marked down by different traffic policers, the scheduler schedule the traffic into output buffers of an interface.

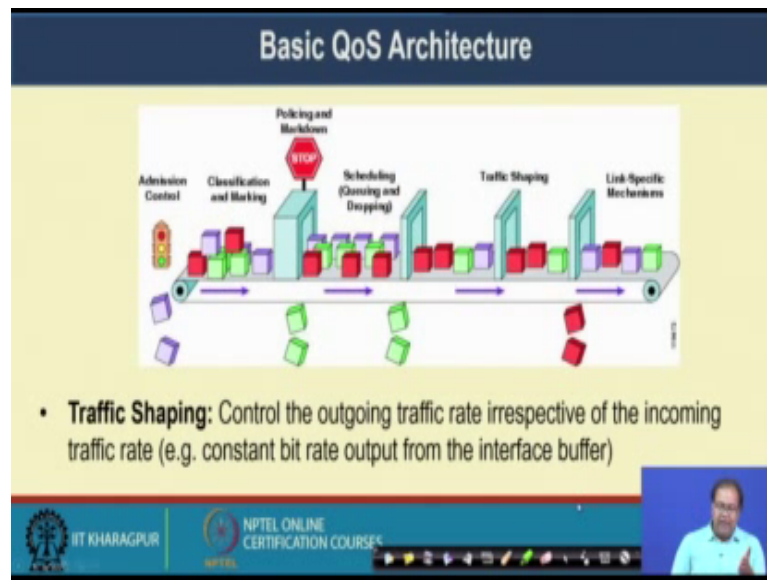
(Refer Slide Time: 23:58)



And the example that I was discussing in the terms of priority queue that you maintain multiple queues at the interface, the scheduling mechanism service the scheduling mechanism services the queues based on the scheduling policy. So, one example is the priority queue that I was mentioning. So, on arrival of a packet you classify the packet and then put it either in the high priority queue or in the low priority queue.

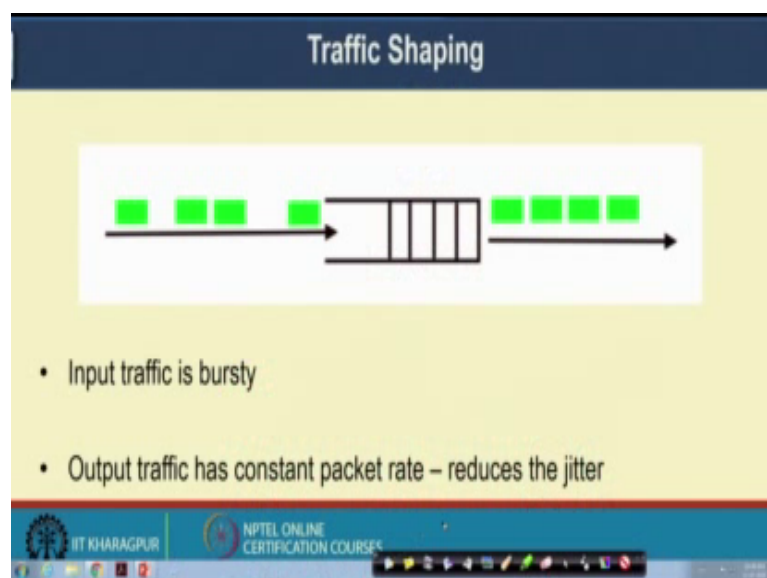
If you are in the high priority queue you will be service first compared to a customer and the low priority queue. So, the link will sub it one by one and send a packet in the outgoing link. So, that way you are giving priority to certain classes of traffic which will experience better quality of service compared to others.

(Refer Slide Time: 24:47)



Finally, the traffic shaping as you are mentioning the so, the traffic shaping control the outgoing traffic rate, irrespective of the incoming traffic rates. So, you are always regulating that what is going to be your outgoing traffic rate. So, either it has constant bit rate output from the interface buffer or you also one certain delay or certain jitter based on the application requirement.

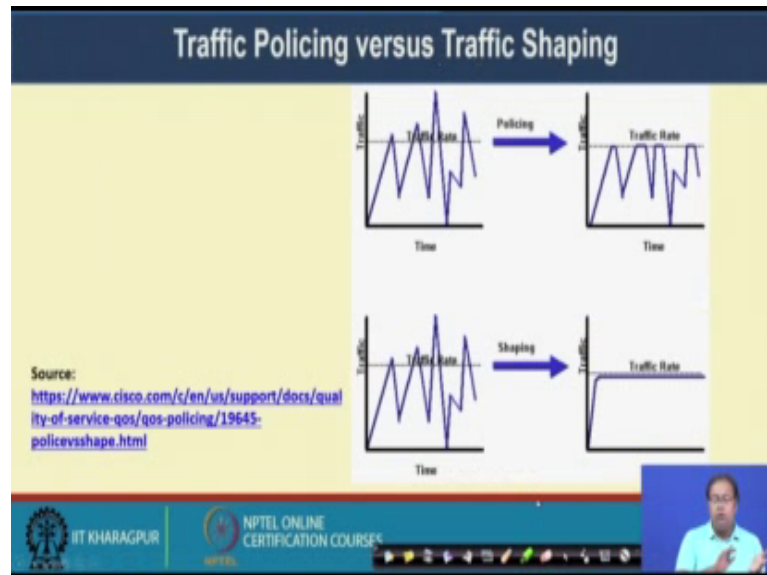
(Refer Slide Time: 25:12)



So, here is the example of traffic shaping. So, you are having irregular traffic at the input and at the output you are getting the regulated traffic, which by the from the figure

you can directly see that it is minimizing the jitter in the network. So, output traffic has a constant packet rate so it reduces the jitter in the network.

(Refer Slide Time: 25:35)



Now, these there are 2; this 2 different terms the traffic policing and traffic shaping. So, let us look into the difference between traffic policing and traffic shaping.

So, traffic policy what it tries to do it just looks that whether certain flows or certain packets are violating the QoS service requirements or not. So, you are getting the traffic at this rate and your expected traffic rate is this dotted line, in that case whatever are the peaks you just drop those packets. So, this is your traffic policy.

So, you see whenever whether something is violating the requirement if someone is violating the requirement you drop the packet. In case of traffic shaper it does something the traffic shaper it actually have this irregular traffic and it tries to regulate the traffic further.

So, you are regulating the traffic at the expected traffic rate. So, this is the difference between traffic policy and traffic shaping. So, we require both in the network because see traffic shaping can may not always be able to give you a smoother rate. If you are average rate is more than this expected rate then traffic shaping will not work directly in that case, you have to apply the traffic polisher to drop the packets, which are violating the quality of service requirements, which are violating the service level agreement then

from the remaining packets where your average rate or your mean traffic rate is less than your expected bound, then you can apply traffic shaper to regulate the output rate.

So, remember this difference between traffic shaping and traffic policy. So, this gives you a kind of broad overview of the basic QoS architecture and in the next class we will go to more deeper inside this, 3 components which are very important from the QoS perspective, the traffic shaping, traffic policy and traffic scheduling. So, we look into the details of these individual filters that we apply for quality of service so.

Thank you all for attending this class.