

Real Time Systems
Professor Durga Prasad Mohapatra
Department of Computer Science and Engineering
National Institute of Technology Rourkela
Lecture 55
Rate Control

Good afternoon to all of you. Now we will take up in this class, one the important topic that is Rate Control or Rate Control in Real Time Communication.

(Refer Slide Time: 00:31)

CONCEPTS COVERED

- Rate Control
- Service Discipline
- Traffic Distortion and its Control
- Types of Service Disciplines
- Conserving Disciplines

The slide features a video inset of Professor Durga Prasad Mohapatra in the bottom right corner. At the bottom, there are logos for NITRR (National Institute of Technology Rourkela) and NPTEL (National Programme on Technology Enhanced Learning).

KEYWORDS

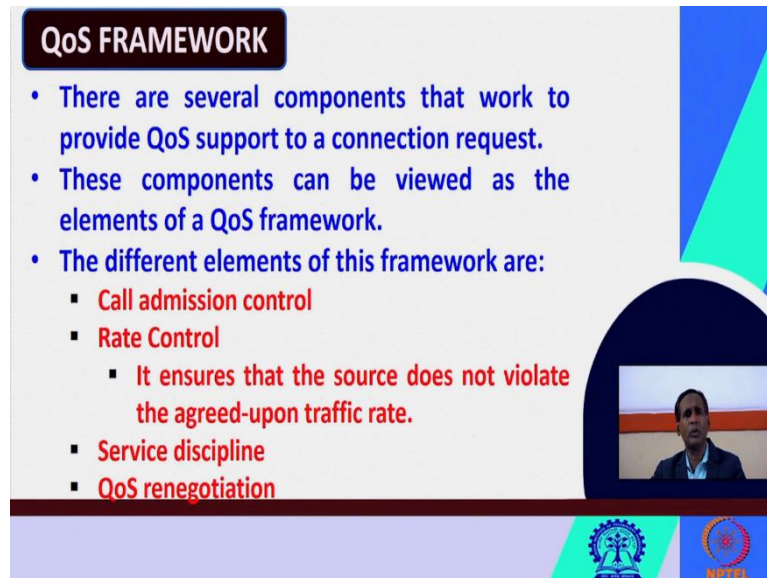
- Token Bucket
- Scheduler
- Rate Controller
- Work Conserving Discipline
- Non-Work Conserving Discipline

The slide features a video inset of Professor Durga Prasad Mohapatra in the bottom right corner. At the bottom, there are logos for NITRR (National Institute of Technology Rourkela) and NPTEL (National Programme on Technology Enhanced Learning).

We will discuss the concept of rate control, what is service discipline, the traffic distortion and its control, what are the different types of service disciplines, and we will see about these

conserving disciplines. We will discuss these, what is, what do you mean by token bucket, scheduler, rate controller, work conserving discipline, and non-work conserving discipline.

(Refer Slide Time: 00:54)



QoS FRAMEWORK

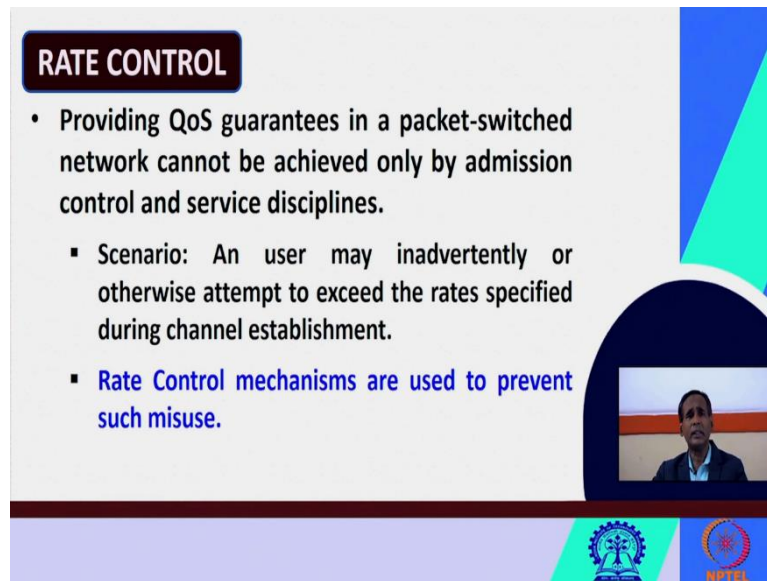
- There are several components that work to provide QoS support to a connection request.
- These components can be viewed as the elements of a QoS framework.
- The different elements of this framework are:
 - **Call admission control**
 - **Rate Control**
 - It ensures that the source does not violate the agreed-upon traffic rate.
 - **Service discipline**
 - **QoS renegotiation**

The slide features a video inset of a man speaking, set against a background with blue and green geometric shapes. Logos for IIT Bombay and NPTEL are visible at the bottom right.

In one of the previous classes, I have told you something QoS framework. We have discussed that there are four important elements in this QoS framework. Those are, call admission control, rate control, service discipline and QoS renegotiation. Please recall, in some of the earlier classes I have already taught this. Today we will discuss, will emphasize on this second important element that is rate control in real time communication.

What does this rate control do? It ensures that the source does not violate the agreed upon traffic rate. So, what this rate control ensures? It ensures that the source node, it does not violate the traffic rate which was agreed upon. It ensures that the source does not violate the agreed upon traffic rate.

(Refer Slide Time: 01:41)



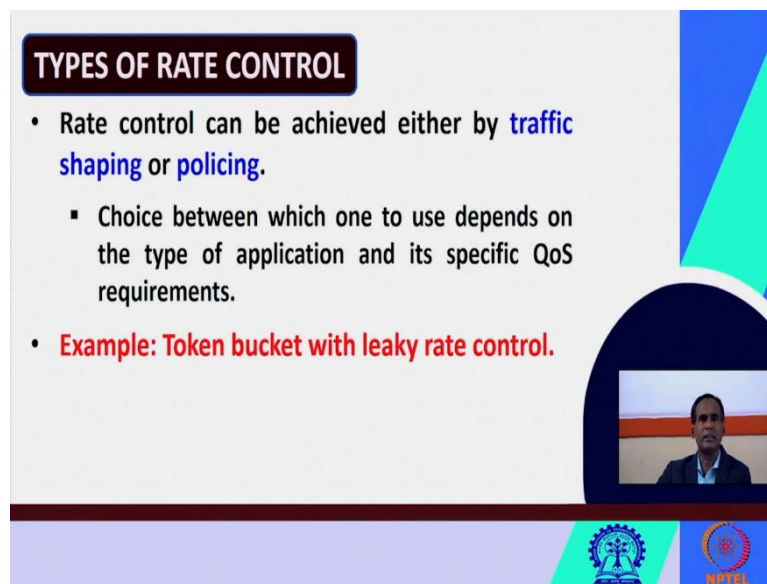
RATE CONTROL

- Providing QoS guarantees in a packet-switched network cannot be achieved only by admission control and service disciplines.
 - Scenario: An user may inadvertently or otherwise attempt to exceed the rates specified during channel establishment.
 - Rate Control mechanisms are used to prevent such misuse.

We will see something more details on this rate control. So providing quality of service guarantees in a packet-switched network, it cannot be achieved just by admission control and service disciplines. If you are using just a simple admission control and service disciplines, you cannot give guarantee that QoS guarantees will be provided in a packet-switched network. Something more is required. Let us take, let us see, why just by using admission control and service disciplines, this quality of service guarantees cannot be provided? Let us take a small scenario or a small example.

A user may inadvertently or otherwise, attempt to exceed the rates specified during the channel establishment. Suppose during channel establishment, some rate is specified. And now, a user, he may deliberately, he may inadvertently or otherwise, he may attempt to exceed that rate then what you can do? You cannot provide the quality of service guarantees. So, rate control mechanisms, they have to be used to prevent such misuse. So in order to avoid these misuses, a mechanism called as rate control mechanism, should be used for preventing these misuses.

(Refer Slide Time: 02:59)



TYPES OF RATE CONTROL

- Rate control can be achieved either by **traffic shaping** or **policing**.
 - Choice between which one to use depends on the type of application and its specific QoS requirements.
- **Example: Token bucket with leaky rate control.**

The slide features a video inset of a man speaking, and logos for IIT Bombay and NITEL at the bottom.

So let us see, what are the different types of rate control. So rate control can be achieved either by traffic shaping or by policing. The choice between, which one want to use, whether you should use traffic shaping or policing, it depends on the type of the application that you are using and its specific quality of service requirements. So whether to use traffic shaping or policing, you require two important things, say, what is the type of application you are using, what is your specific quality of service requirement, those things have to be considered, while making a choice between these two alternatives, traffic shaping or policing.

One of the important examples of or one of the popular example of traffic shaping and policing is the token bucket with leaky rate control. I will discuss this. I hope in computer networking paper, you must have seen about leaky buckets. Those things you must have seen earlier. So this example token bucket with leaky buckets rate control I will discuss now. But before going to this, let us first see, what are the differences between traffic shaping and policing.

(Refer Slide Time: 04:07)

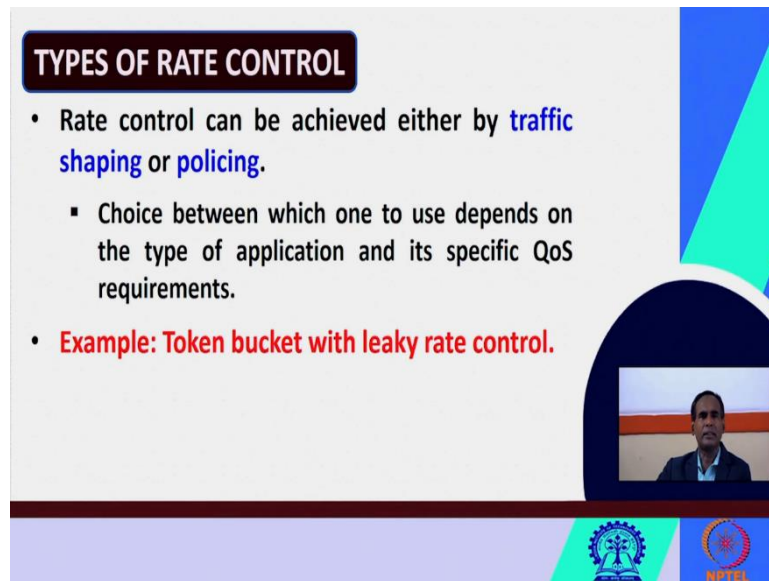
	Traffic Shaping	Policing
Objective	Buffers the packets that are above the committed rates.	Drops the excess packets over the committed rates. Does not buffer.
Handling bursts	Uses a leaky bucket to delay traffic, achieving a smoothing effect.	Propagates bursts.
Advantage	Avoids retransmission due to dropped packets.	Avoids delay due to queuing.
Disadvantages	Can introduce delays due to queuing.	Can reduce throughput of affected streams.

So these are the differences between traffic shaping and policing. The objective of traffic shaping is the followings that it buffers the packets which are above the committed rates but in case of policing, it drops the excess packets over the committed rates. If something is excess than the committed rates, then it will drop for the excess packets. It does not buffer, but here, in traffic shaping, the buffer concept is there. It buffers the packets, which are above the committed rates, but in policing, simply they are dropped.

In context to the handling bursts, so this traffic shaping, it uses a leaky bucket algorithm, to delay the traffic, by thus achieving a smoothing effect. But in case of policing, it propagates the bursts. And the advantage in case of traffic shaping is that it avoids retransmission due to the dropped packets. But policing, it avoids the delay due to the queuing of the packets.

And the disadvantage of traffic shaping is that it can introduce delays due to the queuing of the packets, but in policing it can reduce the throughput. What is the total throughput, it can reduce or it may reduce the throughput of the affected stream? So these are some of the differences between traffic shaping and polishing.

(Refer Slide Time: 05:23)



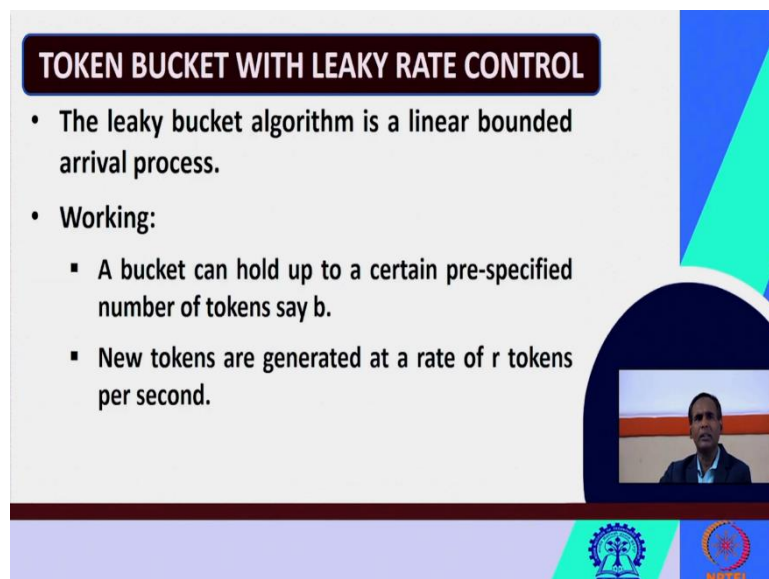
TYPES OF RATE CONTROL

- Rate control can be achieved either by **traffic shaping** or **policing**.
 - Choice between which one to use depends on the type of application and its specific QoS requirements.
- **Example: Token bucket with leaky rate control.**

The slide features a video inset of a man in a suit and tie, and logos for IIT Bombay and NPTEL at the bottom.

So let us first, what I have already told you? One of the examples of this traffic shaping and policing is that token bucket with leaky rate control.

(Refer Slide Time: 05:30)



TOKEN BUCKET WITH LEAKY RATE CONTROL

- The leaky bucket algorithm is a linear bounded arrival process.
- **Working:**
 - A bucket can hold up to a certain pre-specified number of tokens say b .
 - New tokens are generated at a rate of r tokens per second.

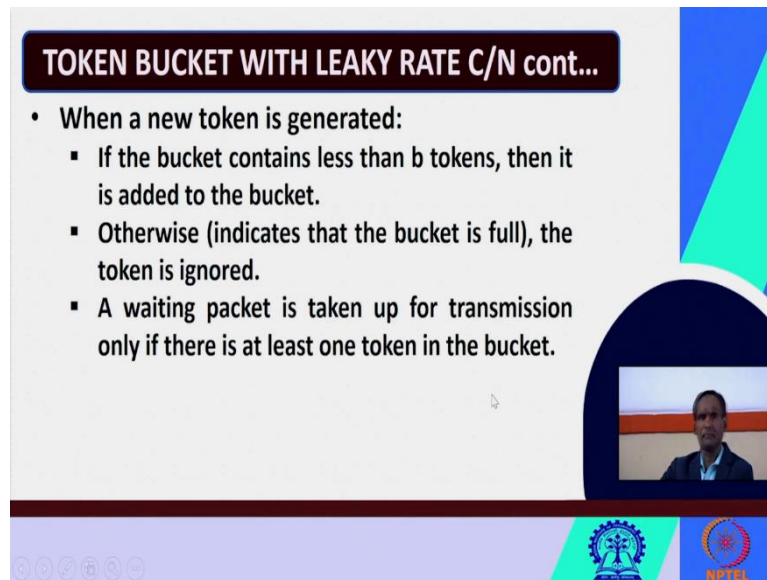
The slide features a video inset of a man in a suit and tie, and logos for IIT Bombay and NPTEL at the bottom.

Let us quickly look at this token bucket with leaky rate control. The leaky bucket algorithm is a linear bounded arrival process. So I hope leaky bucket algorithm you must have studied in the computer network paper earlier. Those who have forgotten, please, look at that algorithm again. The leaky bucket algorithm, it is a linear bounded arrival process.

Let us see, how does it work? A bucket can hold up to a certain predefined number of tokens. A bucket, it can contain maximum some pre-specified or predetermined number of tokens, say

it is b . Suppose a bucket can hold maximum b number of tokens. Now, when new tokens are generated, new tokens are generated at a rate of r tokens per second. So new tokens, they can be generated at a rate of r number of tokens per second.

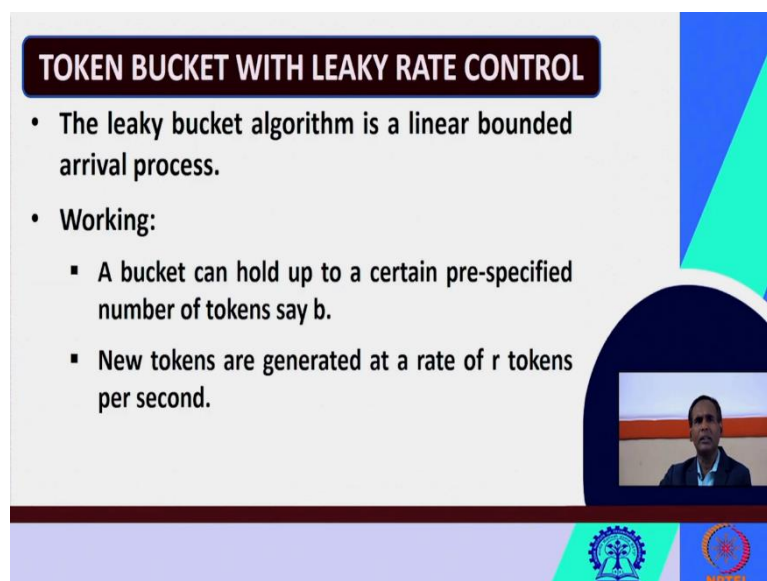
(Refer Slide Time: 06:20)



TOKEN BUCKET WITH LEAKY RATE C/N cont...

- When a new token is generated:
 - If the bucket contains less than b tokens, then it is added to the bucket.
 - Otherwise (indicates that the bucket is full), the token is ignored.
 - A waiting packet is taken up for transmission only if there is at least one token in the bucket.

The slide features a dark blue header with the title in white. The main content is a list of bullet points. A small video inset of a speaker is visible in the bottom right corner. The footer contains navigation icons and the NPTEL logo.



TOKEN BUCKET WITH LEAKY RATE CONTROL

- The leaky bucket algorithm is a linear bounded arrival process.
- Working:
 - A bucket can hold up to a certain pre-specified number of tokens say b .
 - New tokens are generated at a rate of r tokens per second.

The slide features a dark blue header with the title in white. The main content is a list of bullet points. A small video inset of a speaker is visible in the bottom right corner. The footer contains navigation icons and the NPTEL logo.

Now when a new token is generated, if the bucket contains less than the maximum number of tokens. That is you have to, what is the maximum number of tokens I have told here? ' b '. So when a new token is generated, if the bucket contains less than the maximum number of tokens that it can hold that is the b number of tokens, then that token can be added to the bucket. Because some space is there.

Otherwise, what does it mean? Already the bucket, it contains the maximum number of tokens that means the bucket is full, because already it has contained b tokens. Otherwise, what happens? The token is ignored. Why the token is ignored? Because, in the bucket no more space. It has already contained the maximum b number of tokens. It indicates that the bucket is full and hence the token which is generated, it is ignored.

A waiting packet is taken up for transmission, only if there is at least one token in the bucket. So if a packet is there in the queue, it is waiting in the queue, then it can be considered for transmission. It can be taken up for transmission, only if there is an at least one token in the bucket. So if there is no token in the bucket, the waiting packet cannot be transmitted.

(Refer Slide Time: 07:30)

TOKEN BUCKET WITH LEAKY RATE C/N cont...

- If sufficient number of tokens are not present, then
 - In case of a *shaper*, the packet waits until the bucket has enough tokens.
 - In case of a *policer*, the packet is discarded.
- The token is removed after the transmission of a packet.
- The maximum burst-size possible is b .

The slide features a dark blue header with the title in white. The main content is on a light grey background. A small video inset of a man in a suit is in the bottom right. The footer contains the IIT Bombay and NPTEL logos.


TOKEN BUCKET WITH LEAKY RATE C/N cont...



- When a new token is generated:
 - If the bucket contains less than b tokens, then it is added to the bucket.
 - Otherwise (indicates that the bucket is full), the token is ignored.
 - A waiting packet is taken up for transmission only if there is at least one token in the bucket.

The slide features a dark blue header with the title in white. The main content is on a light grey background. A small video inset of a man in a suit is in the bottom right. The footer contains the IIT Bombay and NPTEL logos, along with navigation icons on the left.

TRAFFIC SHAPING VS POLICING


	Traffic Shaping	Policing
Objective	Buffers the packets that are above the committed rates.	Drops the excess packets over the committed rates. Does not buffer.
Handling bursts	Uses a leaky bucket to delay traffic, achieving a smoothing effect.	Propagates bursts.
Advantage	Avoids retransmission due to dropped packets.	Avoids delay due to queuing.
Disadvantages	Can introduce delays due to queuing.	Can reduce throughput of affected streams.





TOKEN BUCKET WITH LEAKY RATE CONTROL

- The leaky bucket algorithm is a linear bounded arrival process.
- Working:
 - A bucket can hold up to a certain pre-specified number of tokens say b .
 - New tokens are generated at a rate of r tokens per second.



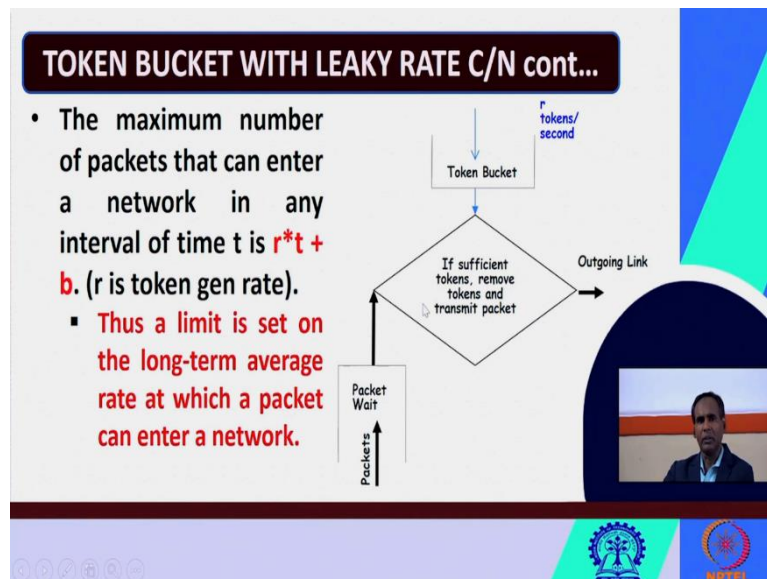



So that is what I have told here, if sufficient number of tokens are not present, then what will happen? I have already told you, a packet will be transmitted only when at least one token is there in the bucket. I have already told you, there are two policies that shaping and policing. So, if sufficient number of tokens are not present then in case of a shaper, what will happen?

The packet, it waits until the bucket has enough tokens. So in case of these two things, I have already told you, traffic shaping and policing that means shaper and policer. So if sufficient number of tokens are not present then in case of a traffic shaper, the packet will wait until the bucket has enough number of tokens, but in case of policer, simply the packet is discarded, it will be not stored. It will not be saved or it will not be stored. Simply the packet will be discarded. The token is removed after the transmission of the packet.

So when the packet is transmitted then that token is removed from the bucket. The token is removed after the transmission of a packet. So, what is the maximum burst-size possible? I have already told you here, that we have already told that a bucket can hold up to a certain predefined number of tokens, say it is b here.

(Refer Slide Time: 08:49)



So the maximum burst-size possibly is, it is b . So this working of the token bucket with leaky rate, what I have explained just now, it has just explained in the form of a flowchart. Like r number of tokens per second are entering into the token bucket. Now, if sufficient tokens are there in the bucket then what will happen?

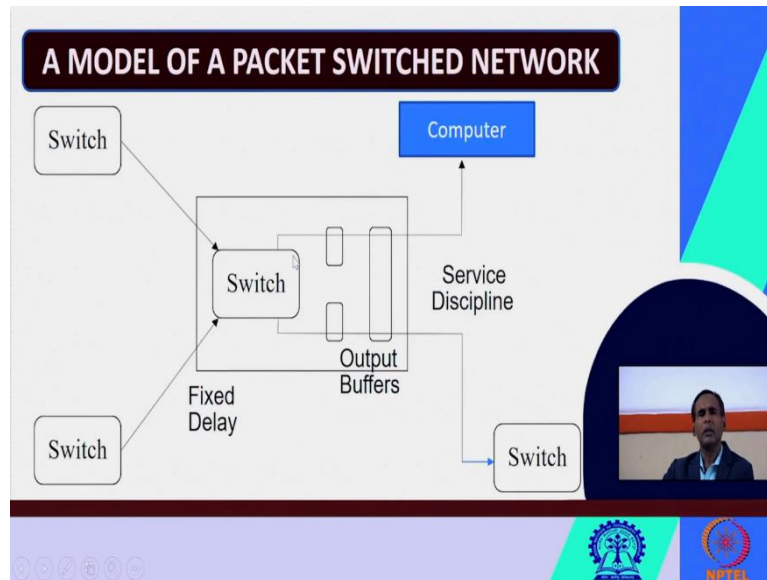
You remove one token and transmit the packet. And you can see, packets are just, they are waiting in a queue and if at least one token is there, if sufficient tokens are there, then you remove one token and transmit the packet. So this is the outgoing link. This is, how the token bucket with leaky rate, it works.

You can see now, one important derivation. One important formula you can derive. That is as follows. The maximum number of packets, which can enter a network in any interval of time t , is given by r into t plus b . I am repeating again, what is the maximum number of packets that can enter a network in any interval of time t , it can be given as r into t plus b . What is r ? r number of tokens are passing per second. So r is also known as the token generation rate.

So the maximum number of packets will be equal to the maximum number of packets that can enter a network, is given by, in any interval of time t , is given by r into t plus b . Thus, a limit

is set on the long term average rate at which a packet can enter a network. It cannot be just indefinite. It cannot be unbounded. So thus a limit is set, a limit is put on the long term average rate with which a packet can enter into a network. So some constraint is put, some limit is put and the limit is r into t plus b .

(Refer Slide Time: 10:41)



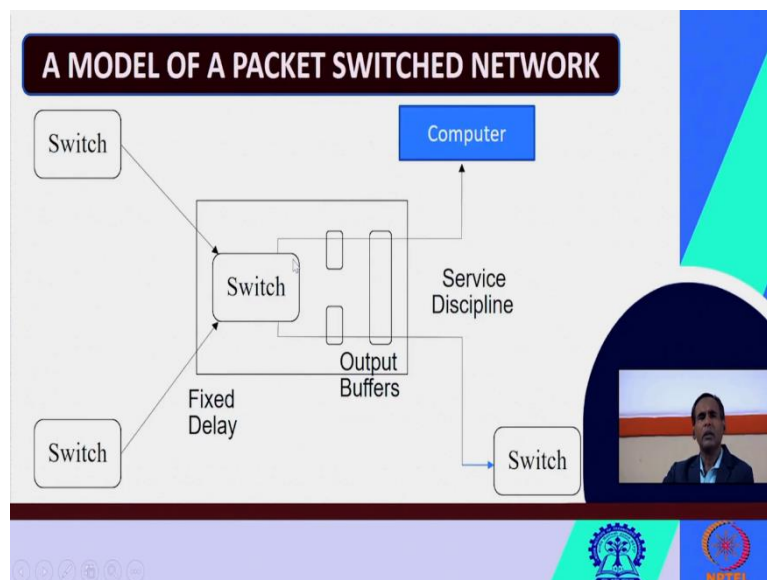


So now, let us look at this packet switched network. We have already discussed packet switched network perhaps earlier, you have also read in computer network paper. In a packet switched network, you see, there are several switches are connected, so there are several incoming and outgoing links are there, from each switch. Also a computer is connected.

So some output buffers are there, where the data packets may be stored. The delay is fixed here. Something delay is there. So here, in packet switched network, we use one important concept called as a service discipline. So now, we will discuss, what is service discipline, what are the different types of service disciplines.

(Refer Slide Time: 11:20)

SERVICE DISCIPLINE

- Depending upon the destination of a packet, it is queued in the corresponding outgoing link.
- A scheduler selects packets to be transmitted from the buffer at the output link based on some scheduling policy.
- Service discipline is the mechanism used to schedule incoming packets for transmission.

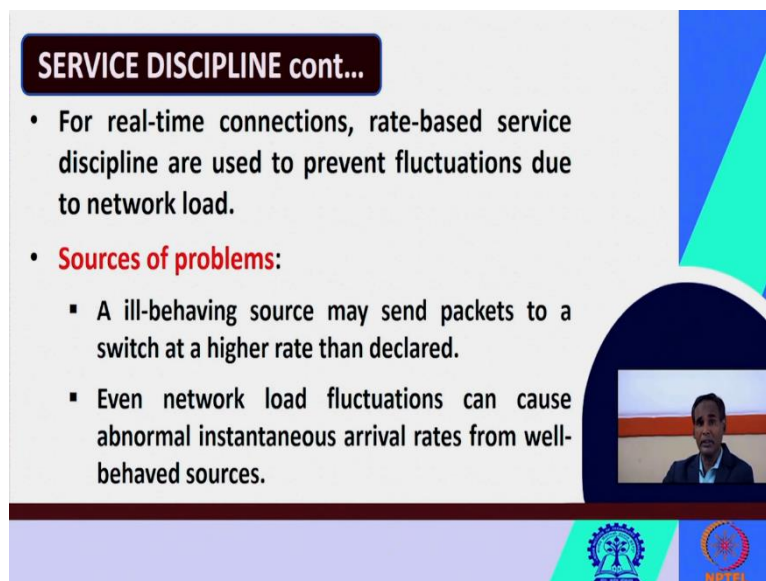


So we will discuss the basic concepts of service discipline first. Depending upon the destination of a packet, the packet is queued in the corresponding outgoing link. So there are different outgoing links, you can see, these two are incoming links for this switch. These two are outgoing links. So depending upon the destination of a packet, the packet is queued in the corresponding outgoing link.

Then a scheduler selects the packets, which are to be transmitted from the buffer at the output link based on some scheduling policy. I am repeating again, there is a scheduler, which selects, which packets will be transmitted from the buffer, are the output links, based on some scheduling policy. Here, you can see that these are the outgoing links. Now you see, some of the output buffers are there.

So now, what you have to see? A scheduler will select the packets, which are to be transmitted. From where? From the buffers at the output link, based on some scheduling policy. Service discipline is the mechanism, which is used to schedule the incoming packets for transmission. So, how to schedule the incoming packets for transmission? Which packet will be taken up? So that is determined by a mechanism called as service discipline. So service discipline is a mechanism, which is used to schedule the incoming packets for transmission.

(Refer Slide Time: 12:55)



SERVICE DISCIPLINE cont...

- For real-time connections, rate-based service discipline are used to prevent fluctuations due to network load.
- **Sources of problems:**
 - A ill-behaving source may send packets to a switch at a higher rate than declared.
 - Even network load fluctuations can cause abnormal instantaneous arrival rates from well-behaved sources.

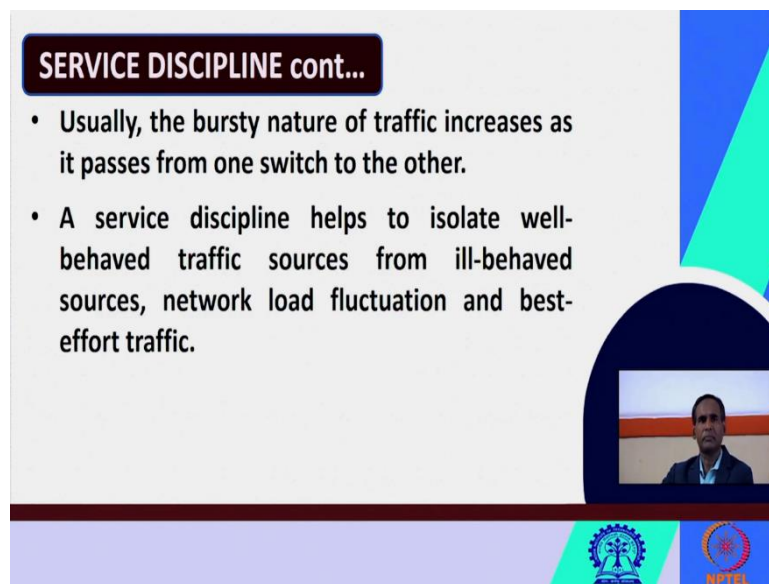
The slide features a video inset of a man in a suit on the right side. At the bottom, there are logos for IIT Bombay and NPTEL.

For real time connections, rate-based service discipline are used, to prevent fluctuations due to network load. So in case of real time systems, we use a discipline, we use a service discipline called as the rate based service discipline. So these rate based service disciplines are used to prevent the fluctuations due to the network load.

So, now let us see, what are the sources of problems? Why there might be some fluctuations due to the network load? Some of the sources of the problems I have listed below, like, an ill-behaving source may send packets to a switch at a higher rate than declared. So for every source, some rate is fixed, with which the packets can be, they can be sent.

But there are some normal sources and some ill-behaving sources. An ill-behaving source, it may send some packets to a switch, at some higher rate, at some higher rate, which was declared, which was prescribed, it sends the data packets with much higher rate then, of course, fluctuations may occur. Even network load fluctuations can cause abnormal instantaneous arrival rates from well-behaved sources. Even if in some cases the network load fluctuations, they can also cause some abnormal instantaneous arrival rates from the well-behaved sources.

(Refer Slide Time: 14:17)



SERVICE DISCIPLINE cont...

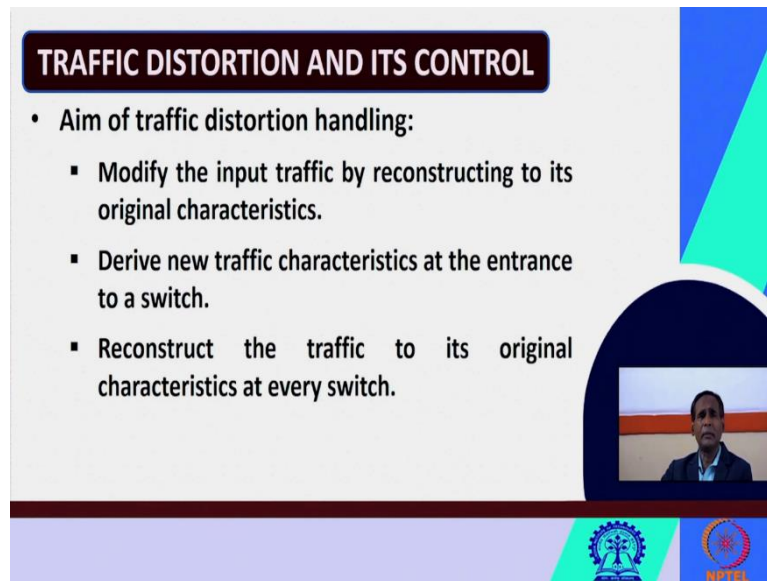
- Usually, the bursty nature of traffic increases as it passes from one switch to the other.
- A service discipline helps to isolate well-behaved traffic sources from ill-behaved sources, network load fluctuation and best-effort traffic.

The slide features a video inset of a man speaking, set against a background with blue and green geometric shapes. At the bottom, there are logos for IIT Bombay and NITEL.

Usually, the bursty nature of the traffic, it increases as it passes from one switch to other. So normally, the bursty nature of the traffic, it increases, when it passes or as it passes from one switch to another switch. A service discipline will help to isolate the well-behaved traffic sources from ill-behaved sources, from network load fluctuation and best-effort traffic.

So, how does this service discipline help you? A service discipline will help you, to isolate the well-behaved traffic sources, the perfect traffic sources or the well behaved traffic sources, from, from the ill-behaved sources, from the network load fluctuation and from the best-effort traffic.

(Refer Slide Time: 15:04)



TRAFFIC DISTORTION AND ITS CONTROL

- Aim of traffic distortion handling:
 - Modify the input traffic by reconstructing to its original characteristics.
 - Derive new traffic characteristics at the entrance to a switch.
 - Reconstruct the traffic to its original characteristics at every switch.

The slide features a dark blue header with the title in white. The main content is on a light grey background. A video inset in the bottom right shows a man speaking. The footer contains logos for IIT Bombay and NITEL.

So, what is the aim of traffic distortion and its control? What is the aim of traffic distortion handling? How, why you should handle the traffic distortion? What is its objective? So, the traffic distortion handling mechanism, it modifies the input traffic, by reconstructing to its original characteristics. So any traffic distortion handling scheme, it will modify the input traffic, by reconstructing to its original characteristics.

Another objective of this traffic distortion handling mechanism is to derive the new traffic characteristics at the entrance to the switch and the still one more objective of traffic distortion handling is that it is used to reconstruct the traffic. This traffic distortion handling mechanism is used to reconstruct the traffic to its original characteristics at every switch. So, this traffic distortion handling mechanism, it can be used to reconstruct the traffic to its original characteristics at every switch.

(Refer Slide Time: 16:05)

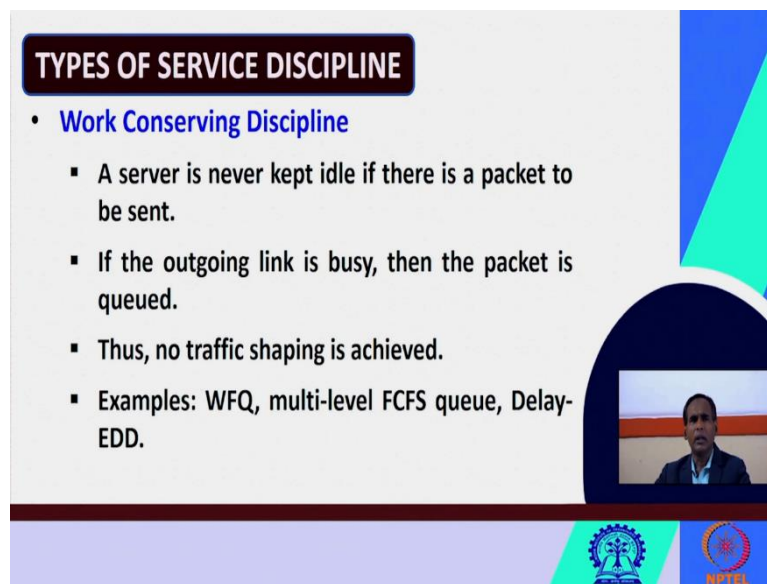
TWO COMPONENTS OF SERVICE DISCIPLINE

- **Rate Controller**
 - It controls the distortion on each connection using regulators and allocates bandwidth to them.
- **Scheduler**
 - It orders transmission of packets from different connections, and provides per-connection delay bounds.

There are two important components of service discipline. One is rate control, another is scheduler. This rate control component, it controls the distortion on each connection. This rate controller, it controls the distortion on each connection using some regulators and it allocates bandwidth to them. So the rate controller, it controls the distortion on each connection, using some regulators and it allocates the bandwidth to them.

What does the scheduler component do? The scheduler component, it orders the transmission of packets. It sequences the transmission packets, how the packets will be transmitted, what will be sequence, what is the order, it prepares that order. A scheduler, it orders or sequences the transmission of packets from different connections and also it provides per-connection delay bounds. It also provides the per-connection delay bounds.

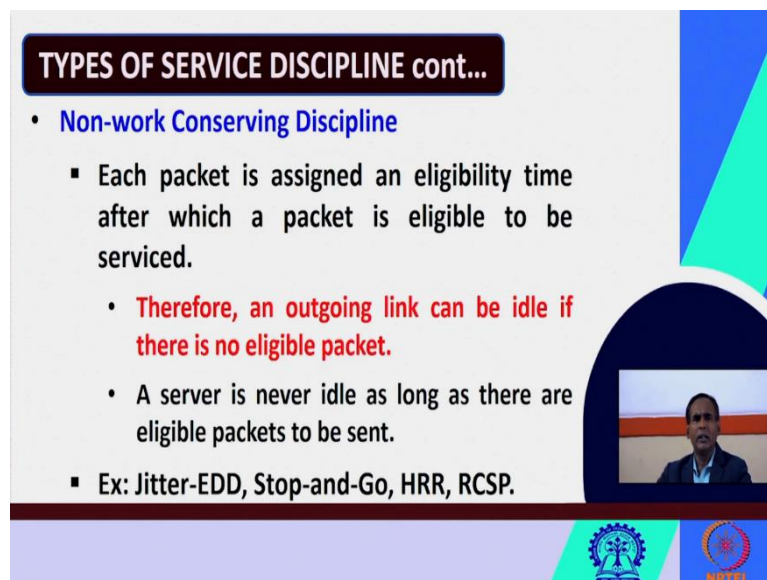
(Refer Slide Time: 17:02)



TYPES OF SERVICE DISCIPLINE

- **Work Conserving Discipline**
 - A server is never kept idle if there is a packet to be sent.
 - If the outgoing link is busy, then the packet is queued.
 - Thus, no traffic shaping is achieved.
 - Examples: WFQ, multi-level FCFS queue, Delay-EDD.

The slide features a video inset of a man in a suit speaking, and logos for IIT Bombay and NPTEL at the bottom.



TYPES OF SERVICE DISCIPLINE cont...

- **Non-work Conserving Discipline**
 - Each packet is assigned an eligibility time after which a packet is eligible to be serviced.
 - **Therefore, an outgoing link can be idle if there is no eligible packet.**
 - A server is never idle as long as there are eligible packets to be sent.
 - Ex: Jitter-EDD, Stop-and-Go, HRR, RCSP.

The slide features a video inset of a man in a suit speaking, and logos for IIT Bombay and NPTEL at the bottom.

Let us see the types of service discipline. There are two important types of service discipline. One is Working Conserving Discipline; another is Non-Work Conserving Discipline. Let us first this work conserving discipline, let us first see this one. In a work conserving discipline, a server is never kept idle, if there is a packet to be sent. Please remember. So whenever a packet has to be sent, in a work conserving discipline, a server is never kept idle.

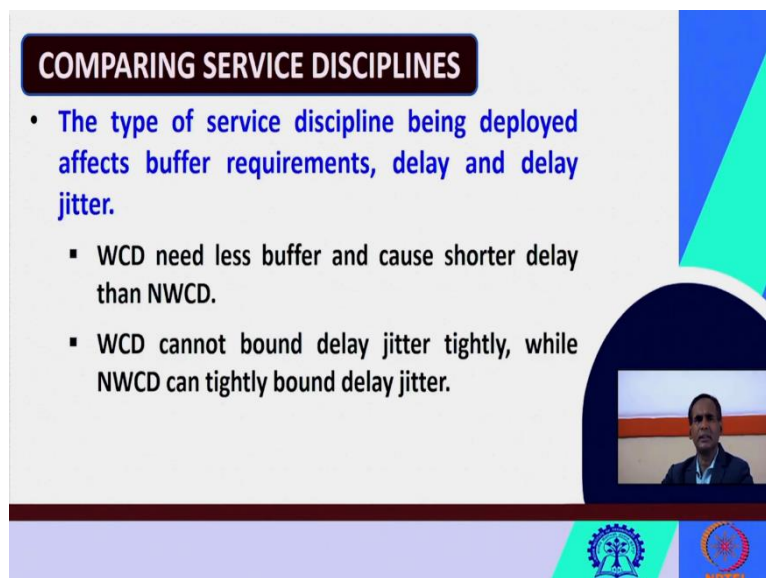
If the outgoing link is busy, then what it will do? Simply then the packet is queued, the packet has to wait. Thus in work conserving discipline, no traffic shaping is achieved. Examples of work conserving disciplines are WFQ, multi-level FCFS queue, delay-EDD, etc. So, we will take up some of these examples in detail just after few minutes. So let us see, how does it differ in non-work conserving discipline?

So here, each packet is assigned an eligibility time. So every packet will be assigned an eligibility time, after which a packet is eligible to be serviced. Before that, that packet cannot be serviced. So in non-work conserving discipline, every packet is assigned an eligibility time, after which a packet is eligible to be serviced or after which the packet can be serviced. Before that the packet cannot be serviced. Therefore, an outgoing link can be idle, if there is no eligible package.

So if at a point of time, there is no eligible packet, then what will happen? The outgoing link will remain idle. This is the most important difference between work conserving discipline and non-work conserving discipline. I have already told you, in work conserving discipline, a server is never kept idle, if there is a packet that has to be sent, but in case of non-work conserving discipline, it is possible that an outgoing link may be idle, if there is no eligible packet. So a server is never idle as long as there are eligible packets to be sent, but if there are the eligible packets to be sent, the server is never kept idle.

But, if there are no eligible packets, the outgoing link can be idle, which is the not here, here in work conserving discipline, a server is never kept idle, if there is a packet to be sent. But here, in non-work conserving discipline, it is possible that on outgoing link may be idle, if there is no eligible packet. The examples of non-work conserving disciplines are Jitter-EDD, Stop-and-Go, HRR, RCSP, etc. So those things will see later on.

(Refer Slide Time: 19:51)



COMPARING SERVICE DISCIPLINES

- The type of service discipline being deployed affects buffer requirements, delay and delay jitter.
 - WCD need less buffer and cause shorter delay than NWCD.
 - WCD cannot bound delay jitter tightly, while NWCD can tightly bound delay jitter.

The slide features a dark blue header with the title in white. The main content is on a light grey background. A small video inset of a man is in the bottom right. Logos for IIT Bombay and NPTEL are at the bottom.

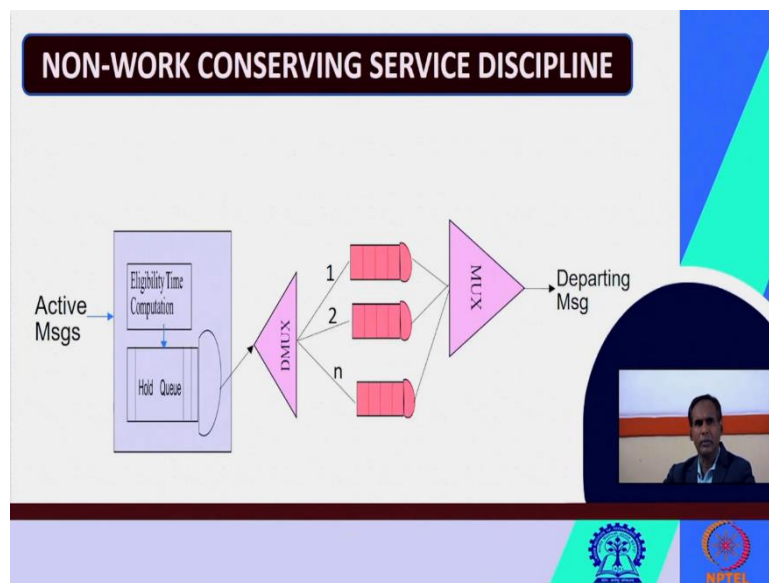
So let us first see, the comparison between these two service disciplines, this work conserving discipline and non-work conserving discipline. The type of service disciplines being deployed,

affects the buffer requirements, delay and delay jitter. So, what type of service discipline you are deploying, whether it is work conserving discipline or non-work controlling discipline, this affects the buffer requirements, this affects the delay and this affects the delay jitter.

So let us see a small comparison. If you are using work conserving discipline, so work conserving discipline need less amount of buffer and the cause shorter delay than NWCD, non-work conserving discipline. Please see, work conserving discipline, it requires less amount of buffer and causes shorter delay than the non-work conserving discipline.

On the other hand, work conserving discipline, it cannot bound delay jitter tightly, whereas non-work conserving discipline, it can tightly bound the delay jitter. So WCD cannot bound the delay jitter tightly, whereas, NWCD can tightly bound to the delay jitter. These are some of the differences between WCD and NWCD.

(Refer Slide Time: 21:13)



So this is, how the non-work conserving service discipline, it works. As I have already told you, the active messages first, they are entering into the system, then there is a mechanism to compute the eligibility time, and every packet will be assigned with the eligibility time, and then, if a packet is eligible, then it will be hold in a queue, then it will be put in a queue, then there is a DMUX and demultiplexer then again, there will be mods, there will be multiplexer and finally the message will be departing. This is the working of non-work conserving service discipline, can be represented in a figure.

(Refer Slide Time: 21:51)

IMPORTANT WORK CONSERVING SERVICE DISCIPLINES

- MULTI-LEVEL FCFS QUEUE
- WEIGHTED FAIR QUEUING (WFQ)

The slide features a dark blue header with white text. The main content area is white with blue text. A small video inset of a man is in the bottom right. The footer contains logos for a university and NPTEL.

Now today we will discuss only some of the important work conserving service disciplines and non-work conserved service disciplines, we will take up in the next class. Today we will take up two important work con work conserving service disciplines, one is a multi-level FCFS queue another is a WFQ, which stands for Weighted Fair Queuing. So I hope FCFS, that you have already known in normal operating system paper, so will now see multi-level FCFS Queue. And then we will see the weighted fair queuing or WFQ.

(Refer Slide Time: 22:25)

MULTI-LEVEL FCFS QUEUE

- It is used to implement static priority schedulers.
 - Each level of the queue corresponds to a different priority level.
- Each connection is assigned a priority
 - Packets from that connection are inserted into the corresponding FCFS queue.
 - Multiple connections can be assigned the same priority.

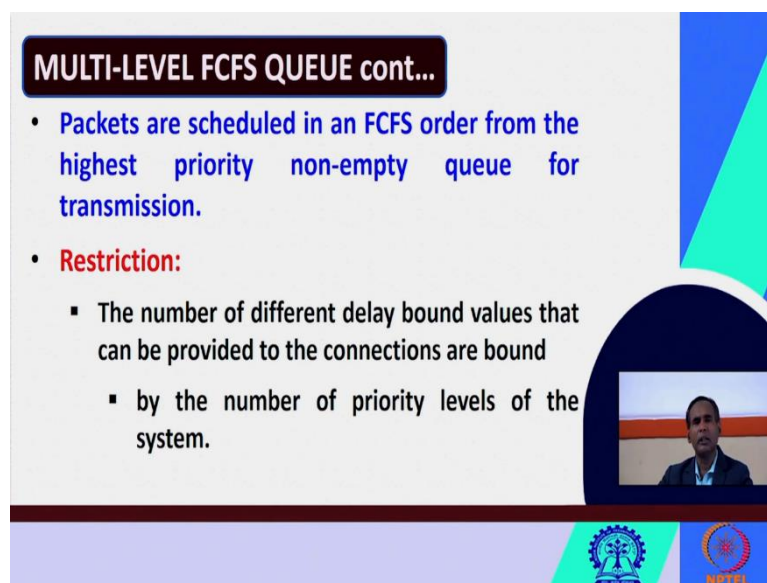
The slide features a dark blue header with white text. The main content area is white with blue text. A small video inset of a man is in the bottom right. The footer contains logos for a university and NPTEL.

Let us first see little bit about the multi-level FCFS queue. Why this is used? Multi-level FCFS queue is used to implement static priority schedulers. You have already known, what is a static

priority scheduler and what is a dynamic priority scheduler. So multi-level FCFS queue is used to implement the static priority schedulers. Here each level of the queue corresponds to a different priority level. So here, in this multi-level FCFS queue, each level of the queue, it corresponds to a different priority level.

Now each connection is assigned a priority. So in this multi-level FCFS queue, each connection is assigned a priority, the packets from that connection they are inserted into the corresponding FCFS queue. In this approach, each connection is assigned a specific priority then the packets from that connection, they are inserted into the corresponding FCFS queue. So it is also possible that multiple connections can be assigned at the same priority. So if there are more number of connections, multiple connections can be assigned to the same priority also. So this is a little bit about the multi-level FCFS queue.

(Refer Slide Time: 23:36)



MULTI-LEVEL FCFS QUEUE cont...

- Packets are scheduled in an FCFS order from the highest priority non-empty queue for transmission.
- **Restriction:**
 - The number of different delay bound values that can be provided to the connections are bound
 - by the number of priority levels of the system.

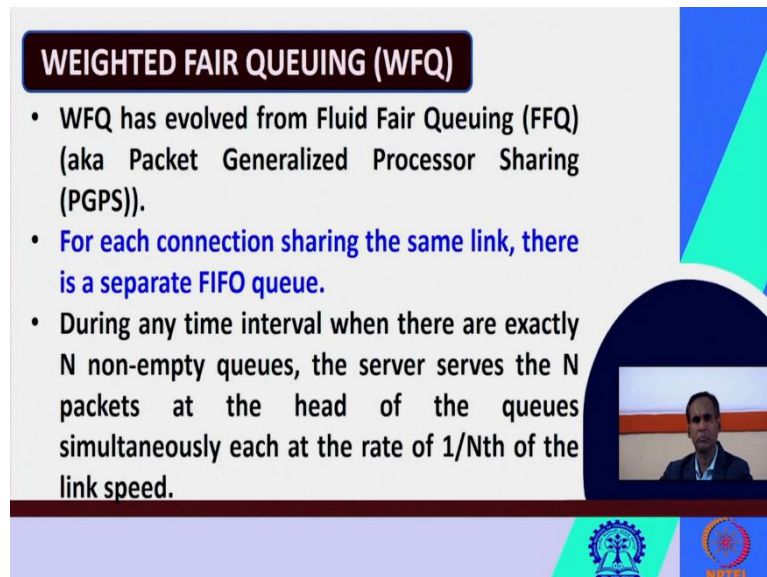
The slide features a video inset of a speaker in the bottom right corner and logos for IIT Bombay and NPTEL at the bottom.

Then let us see, how the packets are scheduled. The packets are scheduled in an FCFS order. So first come first service order. So in multi-level FCFS queue, the packets are scheduled in an FCFS order, from where, starting from the highest priority non-empty queue for the transmission. So in this approach, the packets are scheduled in an FCFS order, from the highest priority non-empty queue for transmission.

But there is a restriction here. That restriction says that the number of different delay bound values which can be provided to the connections are bound. Are bound by which value? They are bound by the number of priority levels of the system. So this is the restriction. This is the constraint you should remember. In multi-level FCFS queue the number of different delay

bound values which can be provided to the connections they are bound. They are bound by what value? They are bound by the number of priority levels of the system. So this restriction is imposed on this multi-level FCFS queue.

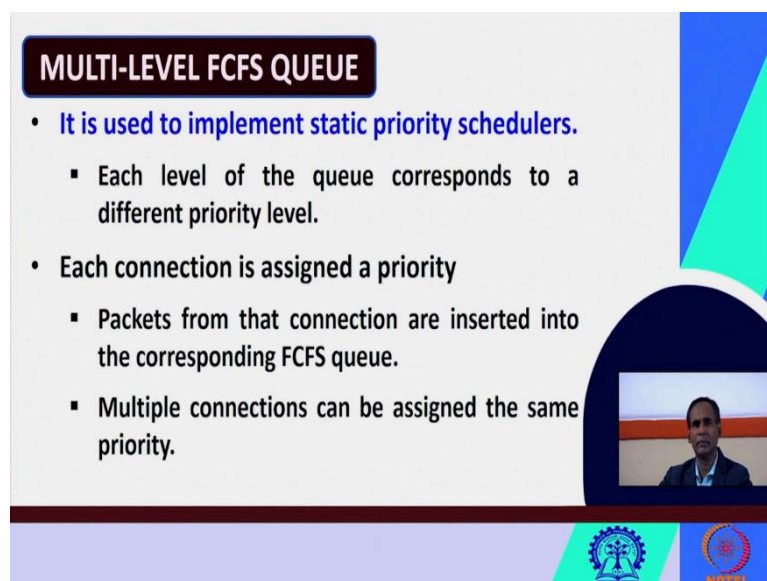
(Refer Slide Time: 24:42)



WEIGHTED FAIR QUEUING (WFQ)

- WFQ has evolved from Fluid Fair Queuing (FFQ) (aka Packet Generalized Processor Sharing (PGPS)).
- For each connection sharing the same link, there is a separate FIFO queue.
- During any time interval when there are exactly N non-empty queues, the server serves the N packets at the head of the queues simultaneously each at the rate of $1/N$ th of the link speed.

The slide features a video inset of a speaker in the bottom right corner and logos for IIT Bombay and NPTEL at the bottom.



MULTI-LEVEL FCFS QUEUE

- It is used to implement static priority schedulers.
 - Each level of the queue corresponds to a different priority level.
- Each connection is assigned a priority
 - Packets from that connection are inserted into the corresponding FCFS queue.
 - Multiple connections can be assigned the same priority.

The slide features a video inset of a speaker in the bottom right corner and logos for IIT Bombay and NPTEL at the bottom.

We will go to the next approach that is weighted fair queuing or WFQ. WFQ has evolved from this Fluid Fair Queueing, which is known as FFQ or this is also known as Packet Generalized Processor Sharing or PGPS. So in a WFQ scheme, for each connection sharing the same link, there is a separate FIFO queue. Please mark the difference between this FCFS queue, where each level of the queue corresponds to a different priority level. In case of WFQ, for each connection sharing the same link, there is a separate FIFO queue.

Now during any time interval, say t , when there are exactly N number of non-empty queues, the server, it can serve or the server serves the N number of packets at the head of the queues simultaneously, each at the rate of one by N th of the link speed. So, at which speed? It will serve the different packets.

It will serve the different packets at the rate of 1 by N th of the link speed, where N number of, N is the number of packets are in the queue. So during any time interval, when there are exactly N number of non-empty queues, then the server serves the N number of packets, where, at the head of the queues simultaneously. Each at the rate, which rate, each at the rate of 1 by N th of the link speed.

(Refer Slide Time: 26:15)

WEIGHTED FAIR QUEUING cont...

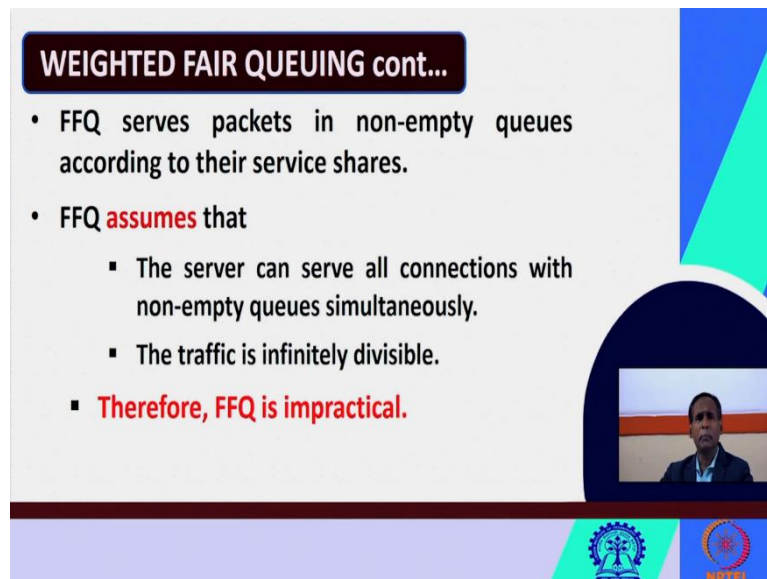
- FFQ allows different connections to have different services shares.
- An FFQ is characterized by N positive real numbers $\phi_1, \phi_2, \dots, \phi_N$ for each queue.
- Let $B(\pi)$ be the set of non-empty queues and C be the link speed. Then, the service rate for a non-empty queue i is: $(\phi_i / \sum_{j \in B(\pi)} \phi_j) * C$

FFQ allows different connections to have different service shares, it is possible. So FFQ, it allows different connections to have different service shares. An FFQ is characterized by N number of positive real numbers. An FFQ is characterized by N positive real numbers such as $\phi_1, \phi_2, \dots, \phi_N$, for each queue. Let $B(\pi)$ be the set of non-empty queues. So let the B of π , it be the set of non-empty queues and see the C be the link speed.

So $B(\pi)$ represents the set of non-empty queues and C represents the link speed. Then the service rate for the non-empty queue, i , is defined as ϕ_i divided by summation ϕ_j , where j belongs to $B(\pi)$. And $B(\pi)$ what? $B(\pi)$ is the set of non-empty queues, and into C . What is C ? C the link speed. Very simple. So suppose, the $B(\pi)$ be the set of non-empty queues. C the link speed.

Then the service rate for a non-empty queue i is given by the equation as follows. The service rate is equal to ϕ_i divided by summation ϕ_j , where j belongs to B_i , where B_i is the set of empty queues, multiplied by C where C is the link speed. So in this way you can determine the service rate for a non-empty queue i .

(Refer Slide Time: 27:57)



WEIGHTED FAIR QUEUING cont...

- FFQ serves packets in non-empty queues according to their service shares.
- FFQ **assumes** that
 - The server can serve all connections with non-empty queues simultaneously.
 - The traffic is infinitely divisible.
 - **Therefore, FFQ is impractical.**

The slide features a video inset of a man speaking, a blue and green geometric design on the right, and logos for IIT Bombay and NPTEL at the bottom.

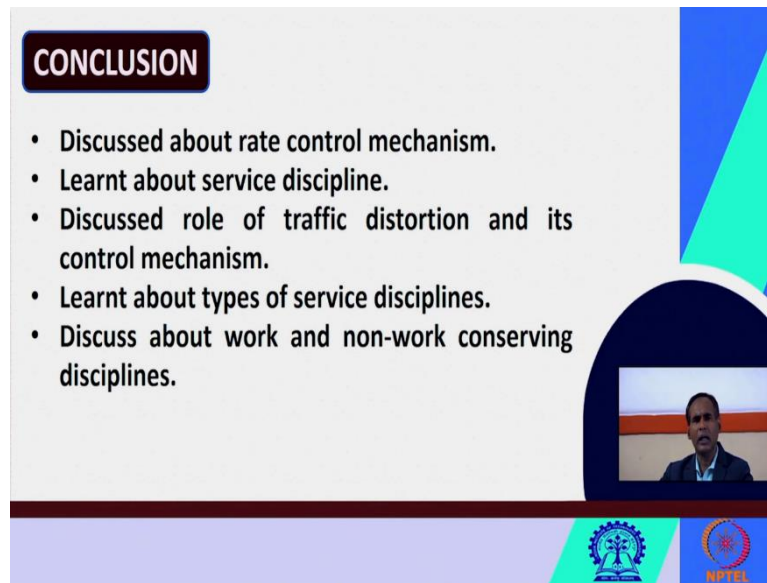
FIFO Queue, it serves packets in non-empty queues according to their service shares. So this is another important issue that you should remember. FFQ, it serves the packets in non-empty queues. According to what? According to their service shares. So according to service shares, they will be served. There is an important assumption in FFQ. FFQ assumes that the server can serve all the connections with non-empty queues simultaneously.

So in an FFQ, it is assumed that the server, it can serve all the connections with non-empty queues simultaneously. The traffic is infinitely divisible. It, FFQ, assumes that the traffic is infinitely divisible. So FFQ makes these two assumptions, which are very very difficult to achieve. Therefore, FFQ is impractical. These assumptions are very much difficult to get or to realize, therefore FFQ is impractical.

(Refer Slide Time: 29:05)

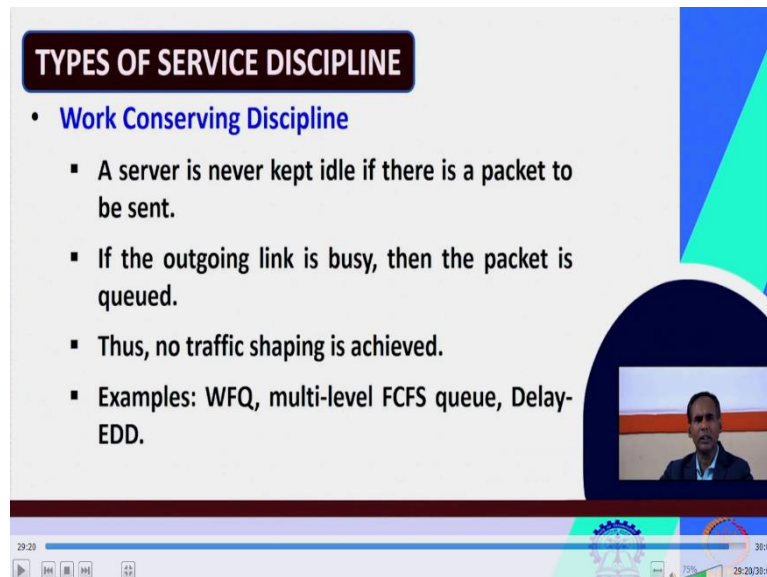
CONCLUSION

- Discussed about rate control mechanism.
- Learnt about service discipline.
- Discussed role of traffic distortion and its control mechanism.
- Learnt about types of service disciplines.
- Discuss about work and non-work conserving disciplines.



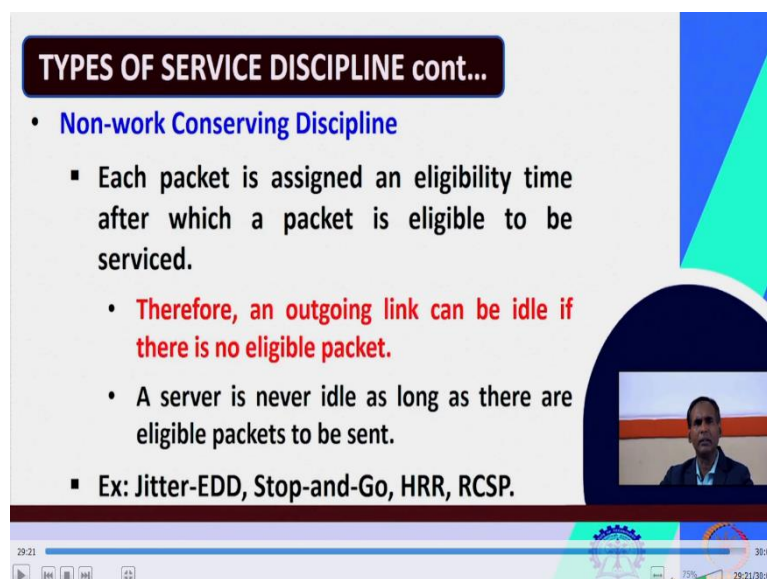
TYPES OF SERVICE DISCIPLINE

- **Work Conserving Discipline**
 - A server is never kept idle if there is a packet to be sent.
 - If the outgoing link is busy, then the packet is queued.
 - Thus, no traffic shaping is achieved.
 - Examples: WFQ, multi-level FCFS queue, Delay-EDD.



TYPES OF SERVICE DISCIPLINE cont...

- **Non-work Conserving Discipline**
 - Each packet is assigned an eligibility time after which a packet is eligible to be serviced.
 - **Therefore, an outgoing link can be idle if there is no eligible packet.**
 - A server is never idle as long as there are eligible packets to be sent.
 - Ex: Jitter-EDD, Stop-and-Go, HRR, RCSP.



Today I have discussed the fundamentals of rate control mechanism. We have learnt about service discipline. We have seen two important service disciplines, one is, we have seen, one is, work conserving discipline, another is non-work conserving discipline. We have seen examples from each category.

We have discussed the role of traffic distortion and its control mechanism. We have also learnt the different types of service disciplines that I have already told you. We have seen these two important types of service disciplines, work conserving discipline and non-work conserving discipline.

We have discussed two important types of works or work conserving service disciplines. So next class, we will discuss about the examples of the non-work conserving service disciplines.

(Refer Slide Time: 30:01)



REFERENCES

1. Rajib Mall, Real-Time Systems: Theory and Practice, 1st Edition, 2007, Pearson Education
2. C. M. Krishna & K. G. Shin, Real-Time Systems, 2017, Tata McGraw Hill Education

The slide features a dark blue header with the word 'REFERENCES' in white. Below the header, two references are listed in black text. To the right of the text is a video inset showing a man in a dark suit and white shirt speaking. The slide is decorated with a blue and green geometric pattern on the right side and a dark blue footer containing the logos of IIT Bombay and NPTEL.

We have taken from these books. Thank you very much for your presenceful hearing.