

Electrical Engineering Department

Broadband Networks

Prof. Karandikar

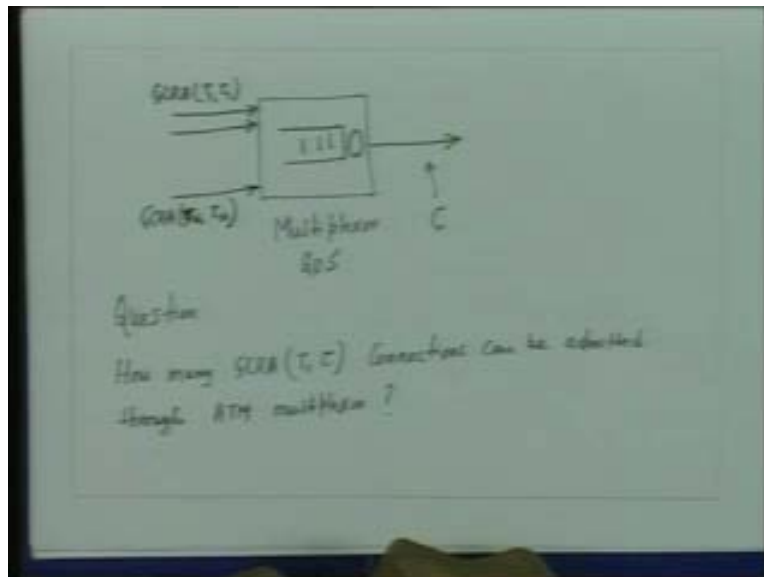
Indian Institute of Technology, Bombay

Lecture - 4

Effective Bandwidth - 1

So, in the previous lecture we have seen how the ATM traffic can describe itself through the deterministic traffic descriptor which is GCRA (T, τ). And, today we will see how a network can have a simple admission control policy to determine whether these traffic sources with GCRA (T, τ) parameters can be admitted or not.

(Refer Slide Time: 1:15)



So essentially, we would be asking this question that here is the stat mux, as we were saying. Here is say multiplexer which wants to give the quality of service guarantees and we would be asking this question that how many GCRA (T, τ) traffic, let us say that this is like $T_1 \tau_1$ and GCRA (T_2, τ_2) so on till say GCRA $n_1 n$ sorry (T_n, τ_n).

How many such sources can be admitted on to this output link and let us assume that this output link has a capacity of C . And, let us assume that this multiplexer has a buffer and it may have a scheduler which for the simplicity may be a C for scheduler.

So, the question that we are asking is how many GCRA (T, τ) connections can go through these transmitters, can be admitted through this ATM multiplexers?

Now, before we can answer this question; we of course need to know that what are the quality of service attributes and so let us assume for the time being to give you an example that the quality of service attributes are the delay d .

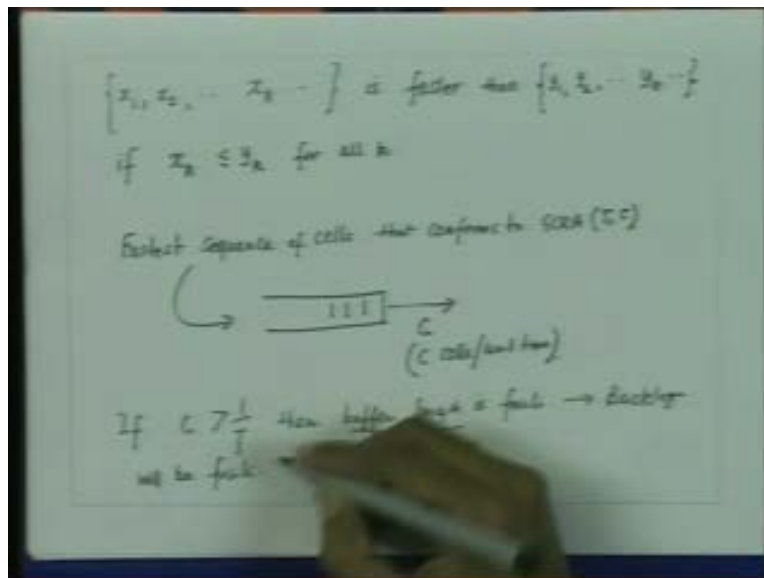
So in other words, we are asking the question that how many such GCRA (T, τ) traffic can be admitted through a multiplexer which is essentially a buffer with first in first out kind of a scheduling discipline. How many of such sources I can admit such that **the packet delay** the packet delay is less than certain quantity, let us say capital D ?

Right now, we are not assuming any other quality of service attributes like packet loss or so. Moreover, we are also assuming that these sources are homogeneous in nature. By homogeneous in nature, I mean that all the sources are described by the same GCRA T, τ parameters and at the same time all the sources have uniform quality of service requirements that is the capital D .

Now, before we answer this question, we have to actually determine what is the fastest GCRA, what is the fastest sequence of cell that confirms to a certain GCRA (T, τ) parameters. Now, why fastest? Because, if I transmit the fastest sequence of cells through an ATM multiplexer which is basically a buffer with a transmitter which is transmitting at the rate of C cells per unit of time; then if I transmit the fastest sequence of cells, I will have the maximum backlog. So, I will have the maximum backlog and whenever I have the maximum backlogs, I will suffer the maximum delay.

So essentially, I would try to find out what is the fastest sequence of cells that **you know** I can generate which will confirm to this GCRA (T, τ) parameters.

(Refer Slide Time: 5:18)



So, all the results which I will be deriving will be for this fastest sequence of cells. So, now what is the fastest sequence of cells **that a cell sequence of cell which is cell let us** which is denoted by

x_1, x_2 so on x_k , where x_1, x_2, x_k these denote the arrival times. Now, we say that this is faster than another sequence y which is having the arrival times as y_1, y_2 so on y_k .

If x_k is less than or equal to y_k for all k ; then, I say that this sequence is fastest. Now, how can you construct the fastest sequence which confirms to GCRA (T, tau) parameters? So, what you do? We start a time 0 with our bucket full of T plus tau units of fluids and then I generate a cell. So, my capital T units will be decremented. Now, the fluid accumulates in my bucket at a unit rate. Now, as soon as a capital T unit of fluid gets accumulated, I generate another sequence and so on. So, if I keep doing this, then I will generate **you know** the fastest sequence of cell which will confirm to the GCRA (T, tau) parameters.

Now, **I am assuming so** I am assuming that I have a fastest sequence of cells that confirms **that confirms to that confirms** to GCRA (T, tau) parameters. Now, this fastest sequence of cells, let us say if I transmit through a buffer which is equipped with a transmitter which can transmit with a capacity of c . So, we are saying c cells per unit of time, it transmits.

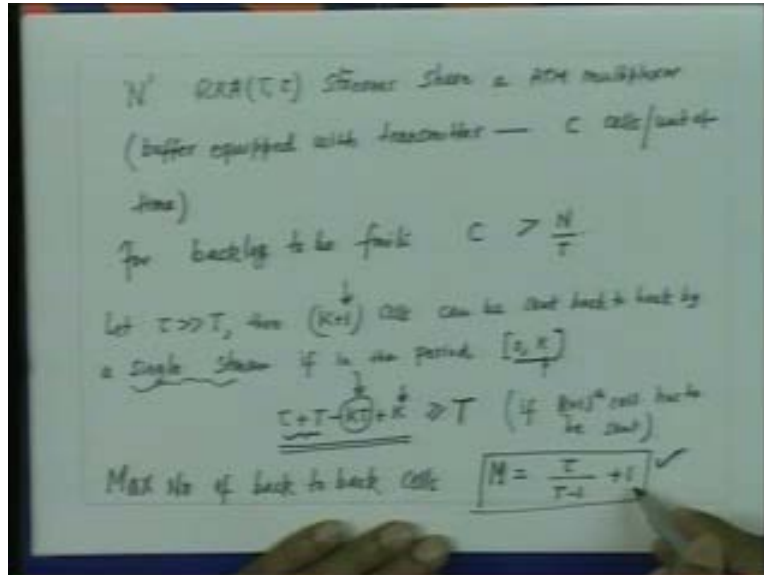
Now obviously, this fastest sequence of cell will generate **you know** the maximum backlogs. Now, we assume that if c is greater than $1/T$, then the q length or the buffer length will always be finite. That means the backlog will always be finite. What I am trying to say is that the backlog will be finite because the transmitter is transmitting at a rate faster than the nominal rate of generation of the cells. Now, we assume to determine **the to determine you know** a simple admission control policy.

Now, we assume that there are n streams, n such GCRA (T, tau) streams and assume all these T, tau streams are the fastest streams. So, they are being generated **you know** as per my explanation just I have given you that the bucket is first filled at T plus tau units. Then you generate a cell, you remove capital T units of fluids and of course, the fluid is continuously accumulating in the bucket at a unit rate. As soon as a capital T unit of fluids gets accumulated, you generate another cell and so on. So, this is the fastest sequence of cells.

Now, assume there are n such streams which are generating such faster sequence of cells and now they are transmitting through this ATM multiplexer which is basically a buffer equipped with a transmitter which is capable of transmitting at C cells per unit of time. Now, just now we saw that if we are transmitting a single stream, then in order for the backlog to be finite; our requirement was that C that is the capacity of the buffer or the transmitter should be greater than $1/T$. C should be greater than $1/T$.

Now, n streams are transmitting the same buffer. So, naturally in order that the backlog is finite, our requirement should be c should be greater than n/T .

(Refer Slide time: 9:58)



So now, we would like to determine what is **you know** an admission control policy? So, we are saying that N such GCRA (T, τ) streams share an ATM multiplexer or which is basically a buffer equipped with transmitter which transmits at C cells per unit of time. Remember, our unit of time here is the one cell transmission time. So, that we have been always using it. So, for backlog to be finite, **for backlog to be finite**, we know that our requirement is that C should be greater than N by T .

So now, we would like to determine that **you know** what or how the network can admit such N GCRA (T, τ) streams such that the **you know** a sudden delay guarantees capital D can be given. So, in other words our objective is to determine what is the maximum value of N ? That is how many numbers of such connections can be admitted such that the delay is less than some quantity, let us say capital D ?

So now, let us assume to prove this or to determine this N ; let us assume that τ is very greater than capital T . Then k plus 1 cell can be sent back to back by a single stream if in the period 0 to k , we have this quantity T plus τ minus KT plus K should be greater than or equal to T . Now, why this? Remember, T plus τ - this was the amount of fluid which was there in the bucket at the beginning. So, T plus τ was the way. Now, k amount of fluid entered into the bucket **into the bucket** during the time 0 to k .

Now, since **you know** K cells have been transmitted; KT amount of fluid would have been **you know** decremented. So, that is why we have the minus. Now, this is the amount of fluid which will be there in the bucket at the K th instance. This should be greater than are equal to T if you want to send k plus 1th cell. That means **you know** this should be greater than or equal to T if k plus 1th cell has to be sent.

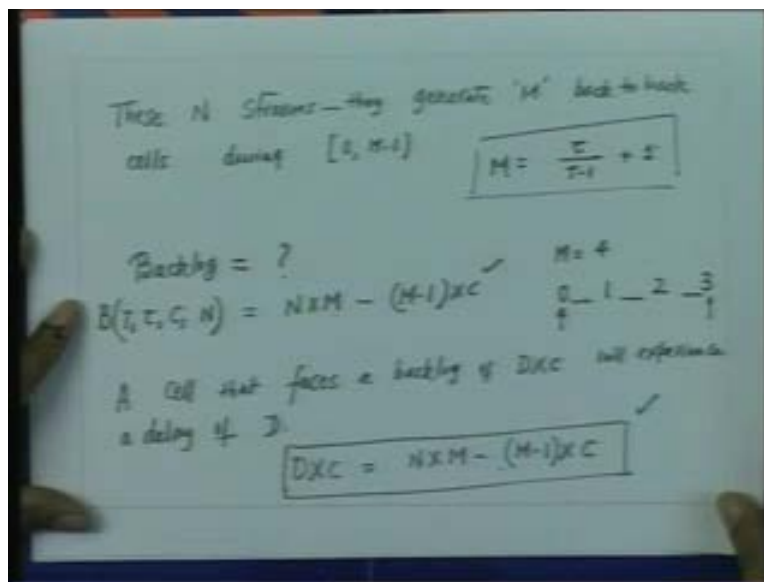
So this explains that k plus 1 cell can be sent back to back by single stream if during this period 0 to k , if it satisfies this quantity. In other words, we can say that by rearranging this equation **you**

know we can say that maximum number of back to back cells which can be maximum number of back to back cells which can be sent by a single stream M will be tau upon T minus 1 plus 1. Now, this comes after rearranging the equation. Remember, after this single stream has sent these K plus 1 cell back to back or in other words n cells back to back, after that you know the cells will come with a periodicity of capital T because you know the fluid will get accumulated at a unit rate into the bucket.

So, the maximum number of back to back cells which can be sent by a single stream is this. Now, let us say, this is by a single stream and seems that there are n streams are there. Now, each of these streams, now maximum number of back to back cells is M which I have just shown that it is by given tau upon T minus 1 plus 1.

Now, there are n streams remember. Now, these n streams, each of them can send these back to back cells during the interval 0 to M minus 1. So, these back to back cells are generated during the interval 0 to M minus 1.

(Refer Slide Time: 15:18)



So, what we were saying that each of these streams, these N streams these N streams, they generate they generate these M back to back cells during this interval of 0 to M minus 1, where M is of course given by tau upon T minus 1 plus 1. Now, how much will be then backlog in the buffer? How much will be the backlog?

Now, the backlog will be how many packets have been generated. How many packet have been generated or how many cells have been generated minus you know how many cells have been transmitted by the buffer. How many cells have been generated? Each of the streams has generated M cells. There are N streams, so how many cells have been generated? N into M. How many cells have been transmitted during the interval 0 to M minus 1 by the transmitter? The transmitter is transmitting C cells per unit of time. So therefore, the transmitter has transmitted M minus 1 into C.

So, let us see how much is the backlog? Therefore, the backlog is N cross M minus $(M$ minus 1 into $C)$. Let us **you know**, let us see an example where M could be let us say, 4 ; so 0 1 2 3 . So, the back to back cells have been generated are 0 1 2 3 and they have been generated **at this you know** during the interval of 0 to 3 . On the other hand, during the interval of 0 to 3 ; **you know** 1 to 3 that is M minus 1 into C cells **you know** have been transmitted where M is given by τ upon t minus 1 plus 1 .

Now, suppose we want to say that we want to maintain a delay of capital D ; now a cell that **will face a backlog of** that faces backlog of D cross C will experience a delay of **will experience a delay of** D . Now, suppose the backlog is D cross C ; **so a cell which** see, there is the backlog of D cross C , now the transmitter is transmitting that C cells per unit of time; so obviously that particular cell will have to wait for the time of capital D .

Now, if the maximum backlog is D cross C , then the cell will face a maximum delay of capital D . So, now the maximum backlog, we have just determined that the maximum backlog is N cross M minus C into M minus 1 . Why that is a maximum backlog? The maximum backlog will be generated when these N streams are the fastest and they are also transmitting maximum number of back to back cells.

So, when the streams are the fastest and they are transmitting maximum number of back to back cells, then you will generate the maximum backlog and if this maximum backlog is equal to D cross C , then a particular cell will experience a delay which will be a maximum delay of capital D . So, let us now equate that. So, what we are saying is that D cross C should be equal to N cross M minus $(M$ minus 1 cross $C)$.

So, this is our governing equation because remember that this is what is called as the maximum backlog. I can write this maximum backlog to be a function of T , τ , C and N . So, this is the maximum backlog and this maximum backlog is equal to this.

(Refer Slide Time: 20:07)

Handwritten notes on a whiteboard:

$$N = c \cdot \frac{D(T-1)+E}{T+\tau-1} \quad \left[C > \frac{N}{T} \right]$$

In other words: (1)

$$N \times \alpha(T, \tau, D) \leq C$$

where

$$\alpha(T, \tau, D) = \max \left\{ \frac{T-1}{D(T-1)+E}, \frac{1}{T} \right\}$$

$\alpha(T, \tau, D)$ - effective bandwidth of traffic SCRA(T, tau)

So, if I just substitute the value of M from this equation into this equation and rearrange these equations; then I will get this, N is given by I get this equation that N is approximately equal to C into D T minus 1 plus tau upon T plus tau minus 1. So, this is just by or rearranging this equation, I am putting M or the value of M here and then rearranging; taking this C here and then by rearranging, you will get these equations.

Now remember, we had also said that these N streams, you know these n streams are confirming or these n streams are transmitting to the buffer and in order that this maximum backlog is finite, we also have the condition that C is greater than N by T. So now, let us look at these conditions what we were saying. So, N is given by this C into D. At the same time, we have that C is greater than N by T. So, these are the, these are the two conditions.

So in other words, we can write these 2 conditions as or in other words we can say that we need to satisfy the conditions; we can admit you know N number of users for satisfying a delay of T. N into alpha of T tau and D should be less than or equal to C, where alpha is equal to alpha (T, tau, D) is equal to maximum of D into T minus 1 plus sorry this should be maximum of T plus tau minus 1 upon D into T minus 1 plus tau into 1 by T.

Now, how does it come? See remember, this is the quantity. This quantity is inverse of inverse of this. So, if you see here, what we are saying that if you have to satisfy both these constrains, in order to satisfy a delay maximum delay of D with N streams sharing the GCRA (T, tau) parameters; then the maximum number of N that can be there is this. At the same time, C has to be greater than N by T.

In other words, N should be such that N multiplied by this quantity alpha which should be less than or equal to C. Now, what is the interpretation of this? The interpretation of this is what we trying to say is N multiplied by alpha should be less than or equal to C. Now, let us say that these n streams where constant bit rate streams. They are all transmitting at the constant rate of r cells

per unit of time all. They are transmitting at r cells per unit of time and there are N such streams. So, in order that **you know** the maximum backlog is finite; obviously, **you know** we have to satisfy that N into r **you know** has to be less than or equal to C , if they were the constant bit rate traffic.

Now here, on the other hand, what we are determining? We are saying that N into α should be less than or equal to C . So, what is the significance of α ? Significance of α is, α is some kind of resources or the bandwidth, we can say; some kind of an abstract bandwidth which needs to be reserved **which needs to be reserved** such that when these GCRA (T, τ) n streams share this buffer which is equipped with a transmitter transmitting at C cells per unit of time, the maximum backlog remains finite and that maximum backlog is such that the maximum delay is bounded by capital D .

So therefore, α **you know** interprets the notion of some kind of bandwidth which needs to be reserved per traffic. α is a kind of bandwidth which needs to be reserved per traffic flow **per traffic flow** because there are n traffic flows, so α is a bandwidth which is to be reserved for each traffic flow, α multiplied by N should be less than or equal to c because c is the output link rate or the output capacity.

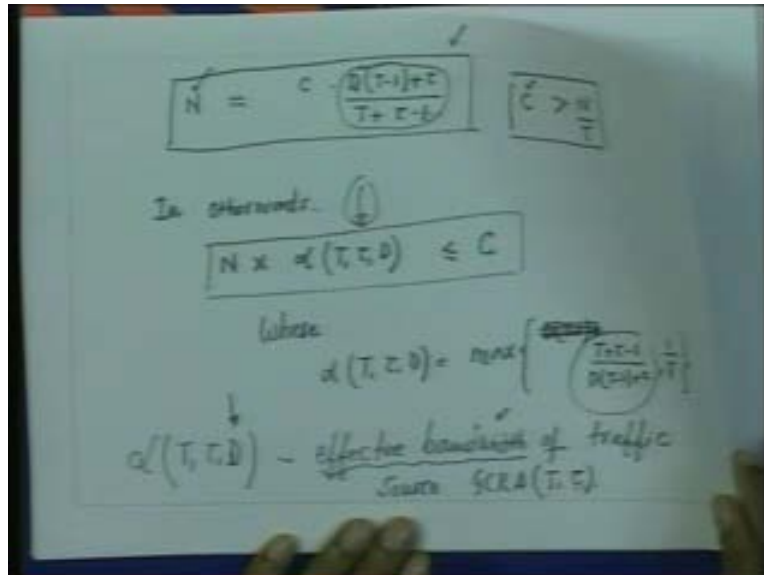
So, it means that α is some kind of a reserved bandwidth which is to be there for these variable bit rate bursty GCRA (T, τ) traffic. So, we say that α is nothing but an effective bandwidth. So, we call $\alpha(T, \tau, D)$ is some kind of an effective bandwidth of the source, effective bandwidth of the traffic source which is having the parameters of GCRA (T, τ).

Now, this notion of effective bandwidth is very important in the communications networks. What we are saying that this effective bandwidth of course is a function of the traffic descriptor which is GCRA (T, τ). But it is also a function of the quality of service attributes which is delay which is quite natural because what essentially we are asking is that how much bandwidth we must allocate out of our C - total bandwidth which is available, how much bandwidth we must allocate per traffic source which is having certain characteristics, in this case GCRA (T, τ) characteristic and at the same time, it has certain quality of service requirements which is **you know** capital D ?

So therefore, **you know** it admits the notion of some kind of reservations per traffic source and hence the name which is called as the effective bandwidth. So, what we are essentially trying to say is that a network can have a very simple admission control policy if the network can estimate or if the network can determine what is the effective bandwidth for the various traffic sources which are there. But remember that this is effective bandwidth is a function of the traffic characteristics and also is the function of the quality of service attributes that each flow desires to have.

Now, if given both these things; if the effective bandwidth of a traffic source can be determined and let us say the effective bandwidth of each source is represented or i th sources is represented by α_i ; then the network admits a very simple admission control policy and that admission control policy is that some of the effective bandwidths **some of the effective bandwidths** of all the sources must be less than or equal to the total bandwidth which is available with the system.

(Refer Slide Time: 27:27)


$$N = C \cdot \frac{D(T-1)+\tau}{T+\tau-1} \quad \checkmark$$
$$C > \frac{N}{T}$$

In other words (1)

$$N \times \alpha(T, \tau, D) \leq C$$

where

$$\alpha(T, \tau, D) = \max \left\{ \frac{T+\tau-1}{D(T-1)+\tau} \right\}$$

$\alpha(T, \tau, D)$ - effective bandwidth of traffic source GCRA(T, τ)

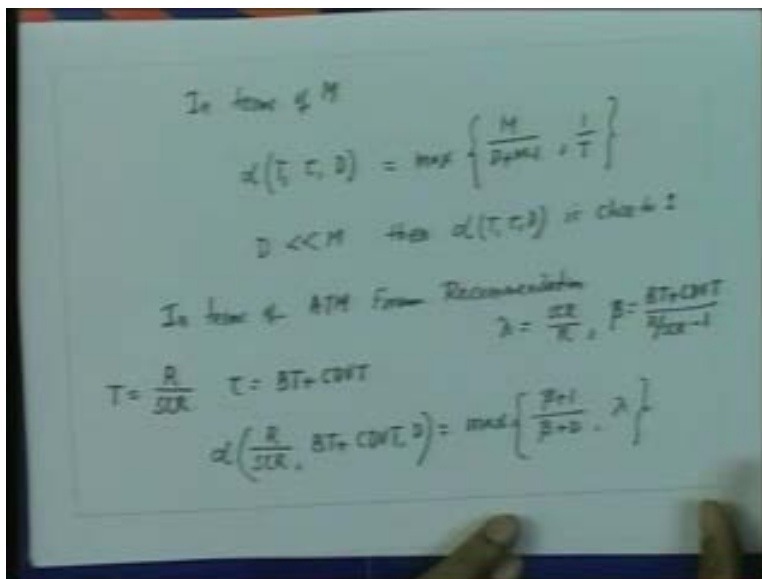
So, we have a very simple admission control policy. So, if we say that alpha i is the effective bandwidth of the traffic flow, effective bandwidth of flow i; then we have a very simple admission control policy which says that some of alpha i should be less than or equal to **you know** C, where alpha i where the effective bandwidth alpha i is definitely a function of traffic characteristics **and traffic characteristics and** as well as the quality of service quality of service attribute.

Now, remember that this is very similar, this admission control policy is very similar to the admission control policy which we follow in a circuit switch networks. If these traffic sources were to share a circuit switch networks, then obviously we need to guaranty that some of their peak rates is less than or equal to C. So, which is P i is actually the peak rate, so in circuit switch networks the policy will look something like this.

Now obviously, you can see that the number of traffic flows which can be admitted in a statistical multiplexer will be much more than in the circuit switch networks because we are admitting the sources based on their peak rates. So, naturally because the effective bandwidth is little less than the peak rates and slightly more than the average rate; **you know** typically the effective bandwidth will lie between the peak rates and the average rates.

And, the challenge really in the networks admission control policy is to determine the effective bandwidth of each traffic sources. So, now let us come back again to the effective band width for of a GCRA (T, tau) parameters that we had just determined which is given by T plus tau minus 1 D into T minus 1 plus tau and 1 by T. Max of this we need to take and this is the effective bandwidth of the GCRA (T, tau) parameter.

(Refer slide Time: 30:01)



In terms of **in terms of** M , we can rewrite this equation as **in terms of** M I can rewrite this as alpha (T, τ, D) is given by max of max of M upon D plus M minus 1 and 1 by T . So, if you say that if D is very less then M that means our **delay requirements is relaxed, then sorry** if the delay requirements is stringent because D is much less; then you know alpha (T, τ, D) is close to close to close to 1.

So, you can see actually that the effective bandwidth of a traffic source **the effective bandwidth of a traffic source** increases if the delay requirement decreases. So, as the delay requirement goes down, the effective bandwidth increases. On other hand, if the delay requirement increases; then the effective bandwidth decreases.

Now, in terms of **ATM Forum Recommendations**, just to say that in terms of **ATM Forum Recommendations**, remember that in an ATM Forum, VBR source is represented by... So, T has an interpretation of R by - this is line rate divided by the sustained cell rate, τ has an interpretation of burst tolerance plus cell delay variation tolerance and of course we have the D .

So, we can write the effective bandwidth in terms of these parameters: R by SCR , burst tolerance plus cell delay variation tolerance into D is given by max of beta plus 1 upon beta plus D is lambda, **where beta** where lambda is really nothing but SCR by R and the beta is nothing but burst tolerance plus cell delay variation tolerance divided by R upon SCR minus 1. So, here you can see that as the delay is very small if the delay is very small, then **you know** we have the requirement which will be close to lambda. That is **you know** SCR by R .

Now, we have seen essentially that how we have determined the effective bandwidth of a GCRA (T, τ) traffic source. Now, the problem however is that most of these traffic sources are statistical in nature. Now, what we saying basically is that a source first determines what are the

parameters of T and τ which suits them. So, this is the responsibility of the traffic source characterization. Once he determines that; then **he asks** he supplies these traffic descriptors to the networks **admission control** admission controller and also the quality of service attributes.

Now, from these the network determines the whether the call can be accepted or not by determining the effective bandwidths. Now, the source the traffic source actually puts this GCRA (T , τ) a traffic shaper or so in front of its traffic source at a source and at the same time it will put it at the network to monitor the traffic whether the sources confirming to these advertised traffic descriptors or not.

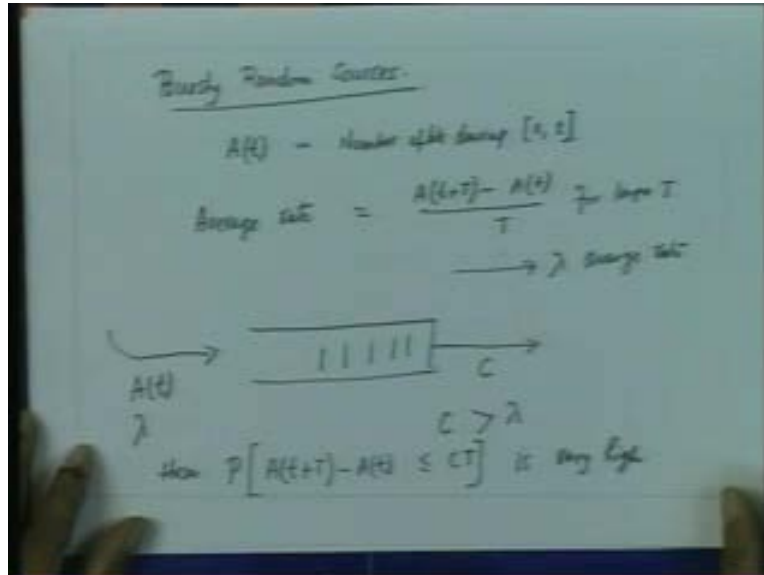
But the problem really is that how does the source choose these (T , τ) parameters because that is clearly **you know** an unanswered questions by now. We have not yet answered how does the source will choose these parameters T and τ such that it accurately represents its traffic characteristic. The difficulty is that you can always choose the parameters of T and τ such that **you know** the network will be over proficient because of that. Because you know, you can always choose those values closer to the peak rate representation of the traffic source, but we do not want to really do that.

But on the other hand, **we can** if you choose such a value such values of T and τ that the traffic characteristics itself gets distorted substantially because the traffic has to pass through this GCRA (T , τ) traffic shapers; then obviously, there will be a distortion that will get introduced the traffic source itself. We will answer these questions later in the context when we discuss the issue of quality of service guarantees in the internet.

I will take up these problems later, but for the time being let us ask this question that if the source is statistical in nature itself; then how do we determine the effective bandwidth of such a source? Because we just now saw that the network will admit a very simple admission control policy if we can determine the effective bandwidth of the traffic source, because if we determine the effective bandwidth of the traffic source; then in order to determine how many number of users can be admitted, the policy is very simple you just determine that the sum of the effective bandwidth of the traffic sources is less than the general capacity.

So, now assume that the source is actually statistical or bursty in nature. So, we will see that how do we handle now the bursty random sources.

(Refer Slide Time: 36:39)

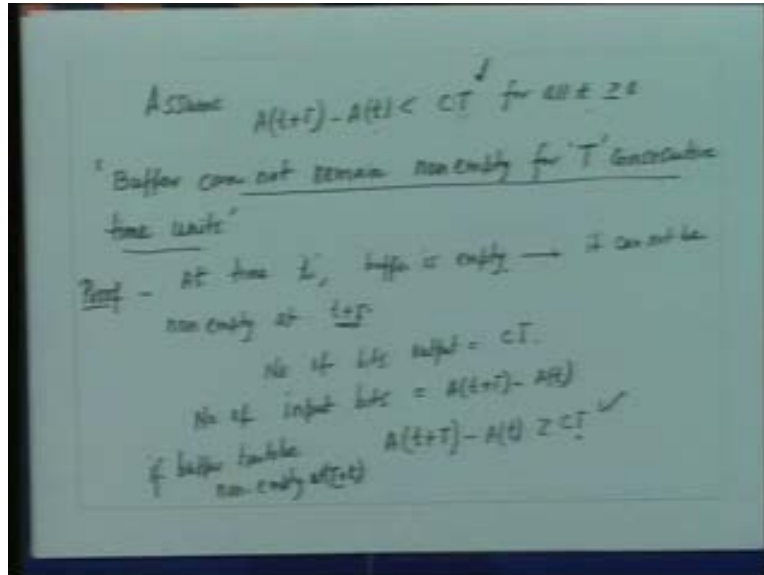


Now, our Bursty Random Source is a source which is represented by $A(t)$, where $A(t)$ denotes the number of bits or the number of cells generated during 0 to t interval. This could be **you know** number of cells also if we discuss this in the context of the ATM. So, what will be the average rate of a source? The average rate of a source will be approximately equal to $A(t+T)$ minus $A(t)$ upon T for very large T .

So, if T is very large, then the average rate will be closer to the **will be closer to** $A(t+T)$ minus $A(t)$ upon T and let us say that λ is this average rate. So, now what we are saying is that here is this fluctuating traffic source which is coming, **you know** essentially $A(t)$ whose average rate is λ and it is sending traffic here and we have C bits per unit of time or C cells per unit of time, whatever it may be and let us say that C , the transmission rate is greater than λ .

So, the transmission rate is greater than λ ; then the probability that $A(t+\tau)$ minus $A(t)$ is less than or equal to $C\tau$ for some large values of t , where t is very large, will be very high. That is the number of bits that have been generated during this interval t is less than or equal to $C\tau$ will be extremely large.

(Refer Slide Time: 38:44)



Now, assume in our case, assume that CT for all T greater than 0 ; then we would try to prove that buffer **you know** which is this buffer, cannot remain non empty, cannot remain non empty for buffer - we would try to prove this proposition that buffer cannot remain non empty for T consecutive time units.

Now, how do you prove this? At time t let us say, buffer is empty. Let us say that at time t , this buffer **you know** is empty. Then we would like to prove that it cannot be non empty at t plus T , where t is such that $A(t + T) - A(t)$ is less than CT for all t . So, we are already admitting substantially a large value of T . Now, during this interval, how many number of bits that have been transmitted? Number of bits which have been outputted is C into T . Now, since the buffer is non empty at T plus t , assume that the buffer is non empty at C plus t ; then the number of input bits which is equal to $A(t + T) - A(t)$.

Now, if the buffer is non empty at t plus t ; then obviously, if buffer has to be non empty, then $A(t + T) - A(t)$ has to be less than or equal to CT . But this is a contradiction because we have assumed $A(t + T) - A(t)$ is less than C .

So obviously, the buffer cannot remain non empty for capital T consecutive time units. That means what is the implication of this? This scheduler would delay. So, what we trying to say is that at time t , the buffer is empty. So, it cannot be non empty at time t plus capital T .

Now, the number of bits which have been outputted during this interval is C into T and the number of bits which have been inputted are $A(t + T) - A(t)$. Now, if buffer has to remain non empty during this interval, if buffer has to be non empty at capital T plus t ; then obviously, the total number of input dates which is $A(t + T) - A(t)$ has to be greater than or equal to CT . **But you know which this is less than** but, given is $A(t + T) - A(t)$ is less than CT . Now, this is therefore a contradiction and hence we can say that buffer cannot remain non empty for capital T , consecutive time units.

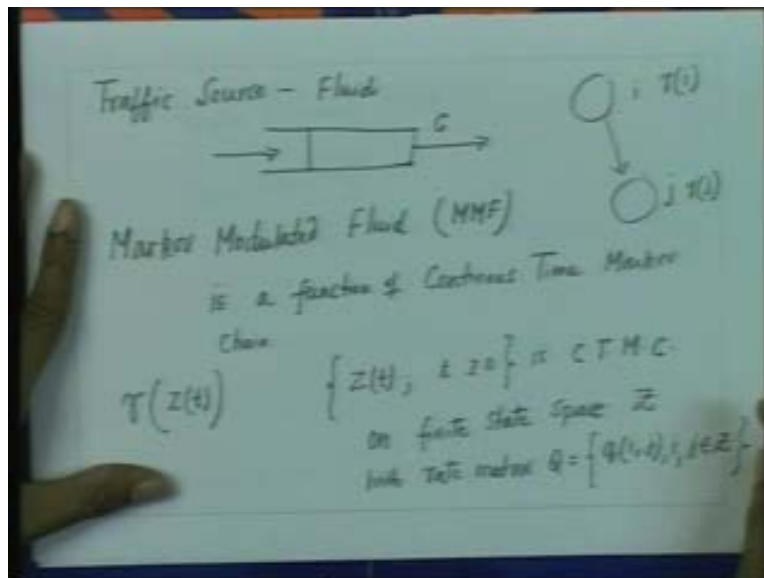
So, what are the implications of this? What we are trying to say is that our bursty random source which was transmitting which was getting transmitted through a buffer which is equipped with a transmitter which is transmitting at the rate of C ; then the scheduler or this transmitter with delay the packets by at most capital T . It will delay the packet by at most capital T , where we have carefully chosen the transmitter rate C in such a manner that it is greater than the average rate.

And, what is the definition of the average rate? The definition of the average rate is that average rate is approximately equal to $A(t + \text{capital } T) - A(t)$ upon capital T , so for a large value of capital T .

Now, what we were trying to say therefore is that it may be possible to reserve a certain portion of the bandwidth, a little larger than the average rate such that we can guarantee certain delay bounds on the traffic which is coming into this buffer which is transmitting at a rate slightly larger than the average rate. Now, slightly larger than the average rate, how much in the average rate is actually as I have already told you that we need to determine basically the effective bandwidth of the traffic source.

Now, the effective bandwidth of the traffic source definitely depends upon the traffic characteristics and so on. So now, I can just give you some example to understand actually how these effective bandwidths are derived. I can just give you one small example and try to explain how we can determine although we will not go into the detailed proofs of the derivations of these effective bandwidths and then we will come back again to the discussion on how to select appropriate parameters of the GCRA (T, τ).

(Refer Slide Time: 44:56)



So, let us consider a model where we are the considering the traffic source to be a fluid traffic source. Traffic source is a fluid traffic source. Now, this is clearly an approximation for the packetized traffic source and **this is so** this fluid is actually transmitting into this buffer which is

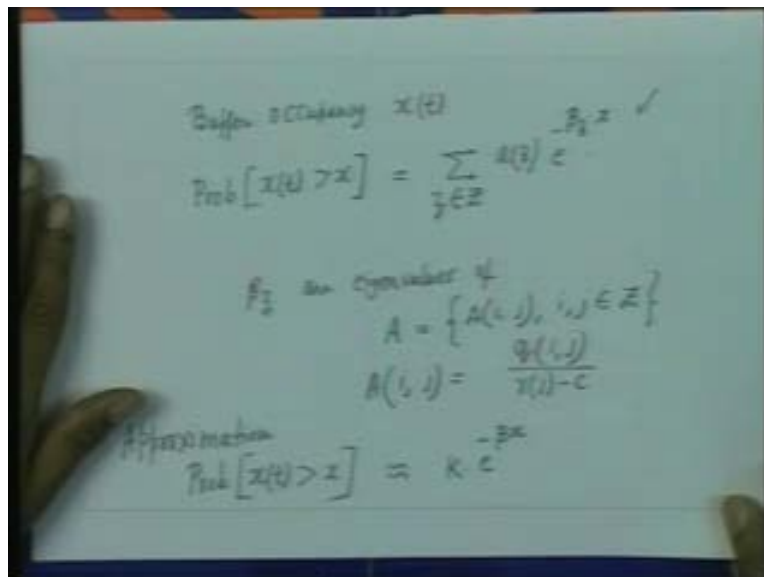
equipped again with a transmitter which is transmitting at may be at C cells per unit of time and so on.

Now, we assume this fluid to be Markov modulated fluid. So, this fluid is Markov modulated fluid which is called as MMF. So, a Markov modulated fluid is actually where the rate is actually a function of continuous time Markov chain. It is a function of continuous time **function of continuous time continuous time** Markov chain. So, we have this rate r which is a function of Z(t) and this Z (t) is a continuous time Markov chain on finite state space Z with certain rate matrix Q - q (i, j) where i, j belong to this Z.

So, what we were saying is that this Z (t) is actually a continuous time Markov chain. It may go transition from state i to state j with rate matrix which is given by q (i, j) and when it is in the state i, the fluid may be transmitting at a rate which is a function of that state of i. That is with this we are saying that the fluid or the rate which is transmitting at the state is actually as a function of that state and the state itself is evolving as a continuous time Markov chain and that is why we call it to be a Markov modulated fluid.

Now, this Markov modulated fluid is transmitting through this buffer.

(Refer Slide Time: 47:54)



So, naturally our question really is that if this buffer occupancy is denoted by x (t); so our problem really is we would like to determine that what is the buffer occupancy and let us say that this buffer occupancy is denoted by x (t) then the question that we were asking is that what is the probability that x (t) this buffer occupancy is greater than **is greater than** (x). So, this distribution we would like to know.

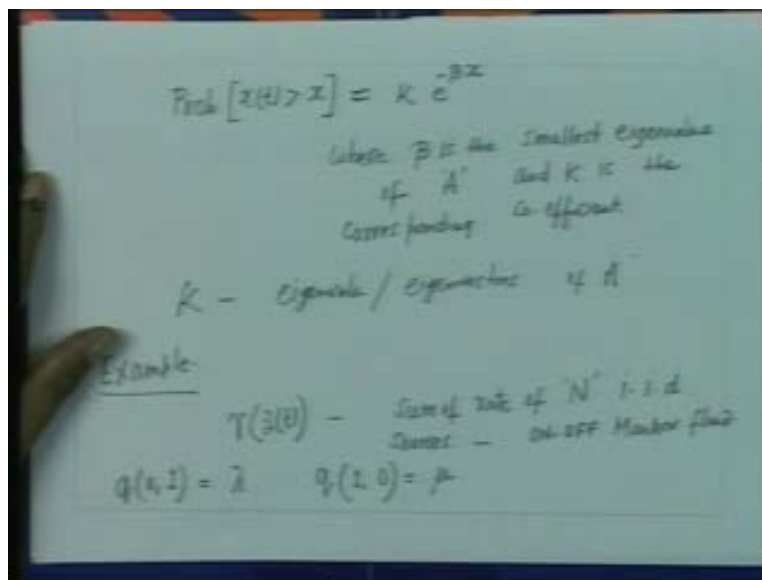
Now again, I would like to show you, this is the Markov modulated fluid. Now, this fluid is like this that the state changes from i to j. This is the state r and this is the state j; when in the state i the source may be generating the fluid at a rate of r(i), when in the state j the source may

generating fluids at a rate of $r(j)$. This is the transmitter which is transmitting at rate of c and the question that we are asking is that **what is the** how the buffer occupancy evolves.

Now, I am not going to give you the proof of this. But it can be shown that this distribution, the occupancy distribution of the buffer length q length is given by summation of z , a z into e raised to power minus beta z into x where a z of course the coefficients which are functions of the state z and beta z happens to be the Eigen values of this matrix A and this matrix A happens to be having the components as $a(i, j)$ for i and j belonging to this state space z and $a(i, j)$ itself is given by $q(i, j)$ where $q(i, j)$ is the transition rate $r(j)$ minus C .

So now, this gives the complete distributions of the buffer occupancy. Obviously, this is a little **this is a little** complicated expression. So, we can we can look for certain approximations. So, we say that let us approximate this buffer occupancy distribution by **we say that** probability that $x(t)$ is greater than x that is approximated by k into e raised to power minus beta x , where beta is the smallest of the Eigen values of this and k is the corresponding coefficients.

(Refer Slide Time: 50:50)



So, what is beta, beta here? So we are saying, in the approximation probability that $x(t)$ is greater than x is given by k into e raised to for minus beta x , where beta is the **smallest of the** smallest Eigen value, smallest Eigen value of this matrix A and K is the corresponding coefficient.

Now, how do determine K ? The K itself **you know** will depend upon all the Eigen values of **all the Eigen values of** the A . So, K itself **you know**, the coefficient K itself depends on all Eigen values and Eigen vectors. So, to determine K itself we will require all the Eigen values on Eigen vectors of A , of this matrix A .

Now, just take an example, just let us take a small example to give you a specific idea. Now, let us say that this $r(Z(t))$ **you know** which is the sum of the rates of **sum of the rates of** N

independent and identically distributed sources and i. i. d sources, independent and identically distributed. Each source is an on off **on off** Markov fluid. So, each of these sources are independent. By independent I mean, each of these sources are independent, there is no dependency in between them and each of these sources are identically distributed which is an on off Markov source. So, we say that $q(0, 1)$; so there are two states 0 and 1. The transition rate from 0 to 1 is given by λ and the transition rate from 1 to 0 is given by μ .

So now, our $q(i, j)$ matrix **you know** that we had considered, our $q(i, j)$ matrix is given by this $q(0, 1)$. So, $q(0, 1)$ is λ and $q(1, 0)$ is μ . So here, z was considered to be the finite state space **the z was considered to be finite state space**. Here, it has only 2 states 1 and 0. Now, it can be shown that and I am not again going into the detailed derivation of this sources.

(Refer Slide Time: 53:54)

$$\beta = \frac{1 + \lambda - N\lambda/C}{1 - C/N}$$

$$K = \left(\frac{N\lambda}{C(1+\lambda)} \right)^N \prod_{i=1}^{N-1} \frac{\beta_i}{\beta_i - \beta}$$

It can be shown that for this when we want to multiplex the N independent and identically distributed sources which is again as a set on off Markov fluid where the transition rate from 0 to 1 is given by λ and the transition rate from 1 to 0 is given by a μ ; we can determine the value of K and β explicitly and the value of the K and β s are given by this $1 + \lambda - N\lambda$ upon C upon $1 - C$ by N and the value of k is given by this.

Now, remember that to compute the value of K , we require all the Eigen values of the matrix A that we had just seen. But if you compute the value of K and β , then we know that what is the probability distribution of the buffer occupancy? Now, note that the probability distribution of the buffer occupancy that this probability that q length is greater than x will decrease exponentially as x increases.

So, if you can fix the value of x , if you know that what is the maximum buffer length that we want; then from that we can determine the value of β and if you know the value of β that we should keep, then from that we can determine that what is the value of C that we should have.

In other words, we can actually determine what is the effective bandwidth of a Markov modulated fluid source are which in this particular example, we have said on off sources.

So, **you know** what we are trying to say is that there is a Markov modulated fluid, MMF source which is transmitting its fluid into a buffer which is equipped with a transmitter and the transmitter is transmitting at the rate of C , then we would like to know that what should be the value C and what is this C is actually the transmitter rate that we should keep such that the buffer occupancy remains less than or **is less than or** equal to certain quantity or the probability that a buffer occupancy increases above certain limit is extremely small, it is within the tolerable limits.

And **you know**, that way we can determine the effective bandwidth. But we have just seen, to determine this we were relying upon the traffic characteristics of the source. We knew that a source is an on off Markov source which is mostly an approximate description of the VBR traffic particularly the video sources have been modeled using on off sources. We will go into the discussion of what are the traffic modelings for various sources later. But I thought I would just give you an example of how we can determine the effective bandwidth of a statistical source.

We had seen how we can determine the effective bandwidth of a deterministic GCRA (T, τ) source very clearly and we can also see how we can determine the effective bandwidth of a statistical source.

(Refer Slide Time: 57:21)

