Broadband Networks
Electrical Engineering Department
Prof. Karandikar
Indian Institute of Technology, Bombay
Lecture - 5
Effective Bandwidth- II

In the previous lectures we were trying to understand how we can provide quality of service guarantees to various traffic sources. Now, first we saw that the traffic sources are essentially statistical in natures. Now, if we have to provide quality of service guarantees to the traffic sources which are essentially statistical in nature; then in order to get a full statistical multiplexing gain, we had earlier suggested a hypothetical admission control where each source specifies its probability distribution function. That is the probability that its bit rate varies and if he knows that these probability distributions of all the traffic sources; then we can determine the probability of distribution of the multiplexed source and from that we can determine what is the probability that the multiplexed bit rate will exceed the output link bit rate and from that we can determine what is the probability of the packet loss.

Then, we saw that since it is not feasible for each traffic source to characterize its probability distribution function; we said let us consider some kind of a deterministic traffic descriptors and the ATM forum has specified these deterministic traffic descriptors in terms of the peak cell rate or the sustained cell rate, the burst tolerate and the cell delay variation tolerance. And, we saw that these traffic descriptors can be ensured by some kind of a leaky bucket or a generalized cell rate algorithm GCRA (t, tau).

So, what essentially we are trying to do is that a source which is basically statistical in nature; we are trying to constrain the source by putting some kind of a traffic shaper or a leaky bucket shaper in front of the traffic source such that the output conforms to a certain peak cell rate, certain average cell rate or sustained cell rate, certain burst tolerance and cell delay variation tolerance.

Now, it is a difference matter that how a source chooses the values of these peak cell rates, average cell rates, burst tolerance and cell delay variations that accurately captures the statistical behavior of the traffic source; that is a different matter. But suppose, we put these traffic shapers in front of a traffic source which is statistical in nature, then by looking at these parameters of the peak cell rate, the average cell rate, the burst tolerance and the side delay variations tolerance; the network can determine whether the call can be accepted or not.

And, we formulated this problem in the following way that suppose that there are N traffic sources. Each of these traffic sources can be characterized by the GCRA (T, tau) algorithm, by GCRA (T, tau) parameters. Let us say that there are N sources. Each of these has parameters as T_1 tau_1, T_2 tau_2, T_n tau_n; then we asked this question or we posed this problem that how many such number of traffic sources can be admitted? That is what is the value of N such that the delay when they are served by a FIFO scheduler with certain transmitter rate which is transmitting, let us say, C ATM packets per unit of time; then what is maximum number of sources that can be multiplexed such that the maximum delay is bounded by certain tolerable value.

And, we found out that this problem can be posed in terms of ensuring that we can admit as many number of sources such that the sum of the effective bandwidths of each of these GCRA

(T, tau) parameters is less than are equal to the output link rate or the transmitters link rate which is we have assumed it to be in our formulations to be C ATM packets or ATM cells per unit of time.

Now, then we again came back to our original thing that essentially a traffic source is statistical in nature. So, can we define something like the effective bandwidth of a source? Now, we ask this question that let us say that there is 1 FIFO scheduler - first in first out kind of a scheduler, there is a buffer and a traffic source which is statistical in nature is transmitting into this buffer and this buffer is having a FIFO scheduler which is scheduling at a certain rate of say C ATM cells per unit of time or C bits per unit of time, whatever maybe the formulations equations; then we ask this question that can we define the buffer occupancy distribution of this.

We solved this problem for a very simple traffic source characterization which is by assuming that the traffic source is a Markov modulated fluid. Markov modulated fluid means that the rate at which the traffic source is transmitting depends upon the state and this state is essentially evolving as some kind of a continues time Markov chain.

So, the rate at which the source will be pumping bits into this or will be pumping fluid into this buffer depends upon the state of the underlying Markov chain. So, that is why we are saying that this underlying Markov chain is essentially modulating. It is a Markov modulated fluid. So, this underlying Markov chain is essentially modulating the rate at which the traffic source is transmitting. And then, we saw that we can actually characterize the buffer occupancy distribution and we can ask this question that what is the probability that a buffer occupancy exceeds certain value, say x.

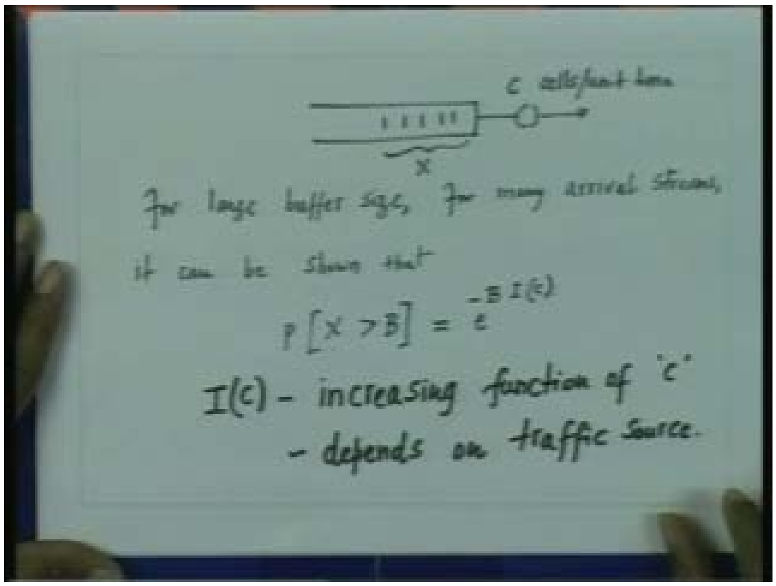
Now essentially, we were interested in finding out what is the probability that the fluid will be lost or the packets or bits will be dropped? Now, to find out this we actually need to determine that what is the probability that our finite buffer actually becomes full. For most of the problems or for most of the traffic sources, it can be shown that if we consider another problem where we assume that the buffer is basically infinite in nature and we ask this problem that what is the probability that the buffer occupancy distribution exceeds a certain value is equivalent to saying that if we have a finite buffer with that value, then what is the probability that this buffer will overflow.

So, even though in practice we are interested in finding out what is the probability of a finite buffer getting overflowed; we will solve or we will address this problem for the case when the buffer is infinite. But we are asking this question that what is the probability that a buffer occupancy distribution will exceed a certain value. Then, we post this problem that if this is indeed true that if this is indeed is true if formulated that what is probability that the buffer occupancy distribution exceeds a certain value; then, how this buffer occupancy distribution evolves and we found out that for Markov modulated fluid, this buffer occupancy distribution decays exponentially.

So from this, now we will try to find out what is then the effective bandwidth of a source? Now, for if you need to really find out the effective bandwidth of a source or rather if you need to find out the buffer occupancy distribution; then it naturally depends upon the statistical characterization of the traffic sources. And, for most of the statistical characterization of the traffic sources except for some simple models like Markov modulated fluids or on off Markov traffics; it is very difficult to determine the buffer occupancy distribution for an underlying traffic source characterizations.

However, it has been found that if you assume that the buffer if the buffer length is very large; then for most of the traffic sources, it can be shown that…

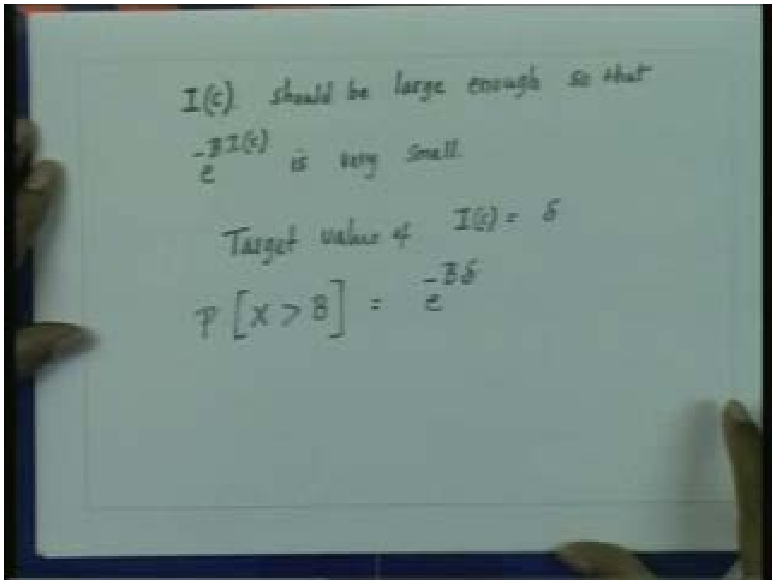
(Refer Slide Time: 10:01)



So, if you assume that here is our buffer; now this buffer is having a scheduler which is transmitting let us say at c cells per unit of time. Now, when the buffer size is when the buffer size is when the buffer size is large; then for many arrival streams - so for large buffer size for large buffer size for many arrival streams - it can be shown that the probability that X is greater than certain B. What is X? X is the queue length of the buffer length. This is X, this is evolving. This probability that the buffer length is greater than B can be approximated as e raised to power minus B into some I(c), where I is an increasing function of the transmitter rate.

So, I is an increasing function of the transmitter rate, I(c) is an increasing function of C and obviously it depends upon the traffic source. Now so, this of course, this form the exact form depends really upon the traffic source and we had actually seen it for the Markov modulated fluid. But if you assume that B is very large and for many arrival streams for most of the traffics characteristics, it can be shown that this buffer occupancy distribution will behave like this.

(Refer Slide Time: 12:52)



Now, what really we want is that I(c) should be large enough. So, what really we need is that this I(c) should be large enough so that this quantity e raised to power minus B into I(c) is very small. Now, you can see here, if we want that this quantity should be very small so that the probability that X exceeds B is a sort of very small because, this actually will determine the packet loss rate.

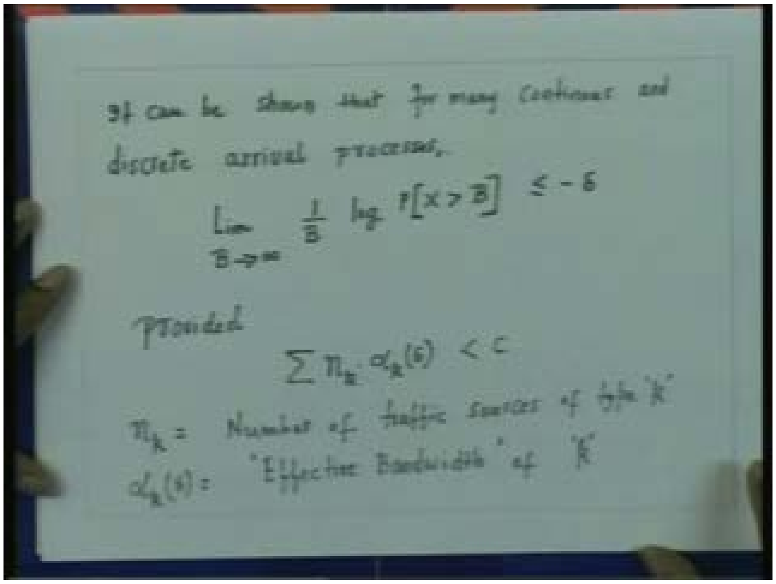
Now, let us say that this target value of this function I(c) is some delta; then we can write that the probability that X is greater than B is actually equal to e raised to power minus beta delta. Now, it can be shown that for many continuous and discreet arrival processes and through a very complex analysis; the following result actually can be proved and let me just tell you that what we can actually prove.

Although I am not giving you the detailed proof of this, I am just trying to tell you the sketch of the proof to drive home this point that for a situation where there is a buffer and this buffer is being served by a transmitter which is transmitting at a certain fixed rate and the statistical nature of the traffic source means that the buffer occupancy evolves statistically; we are interested in finding out the what is the probability that a buffer overflows.

So basically, that is our simple problem. Now, we are saying that, this is equivalent to saying that if this buffer was infinite buffer; then we would like to ask these questions that what is the probability that buffer occupancy will exceed a certain value, let us say B. Both these problems can be equivalent for most of the traffic arrival processes and for most of the problem formulations, even though the two problems are theoretically different.

So, we are now trying to ask these questions that how this buffer occupancy distribution looks like and what I have just said is that for most of the traffic arrival processes, this buffer occupancy distribution actually decays exponentially.

(Refer Slide Time: 15:57)

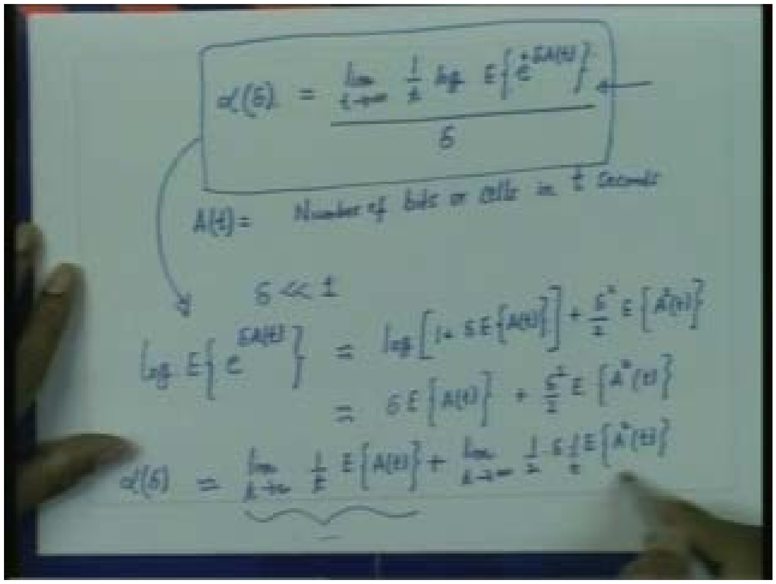


So, it can be shown that for many continuous and discrete arrival processes, it can be shown that the limit - B tends to infinity 1 upon B log of p, X is greater than B is less than or equal to minus delta. It is nearly we are trying say that if B is extremely large, then this result will hold true. So, this is what this result is trying to say that if B is very large; then this result will hold true provided - if this buffer is shared by some j traffic sources, then provided we have n_k into alpha k delta is less than C, where n_k is the number of traffic sources of type k and alpha k delta is the effective bandwidth of source k.

So, what we are trying to say is that if the sum of the effective bandwidths you know if the sum of effective bandwidths is less than or equal to C is less than C, then the loss rates, the packet loss rate decays exponentially; that is what we are trying to say. We can also define the effective bandwidth in the other words. We say that effective bandwidth of a source is that rate with which if the transmitter serves then the packet loss rate will decay exponentially. So, that 2 definitions are equivalent.

Now, we will try to see what exactly is the interpretation of the effective bandwidth? Now, this alpha k delta is actually given by… so we are saying that this is the alpha delta which is the effective bandwidth. This is equal to certain quantity which we defined it to be limit t tends to infinity 1 by t log of expected value of e raised to power delta A (t) this divide by delta.

(Refer Slide Time: 19:28)



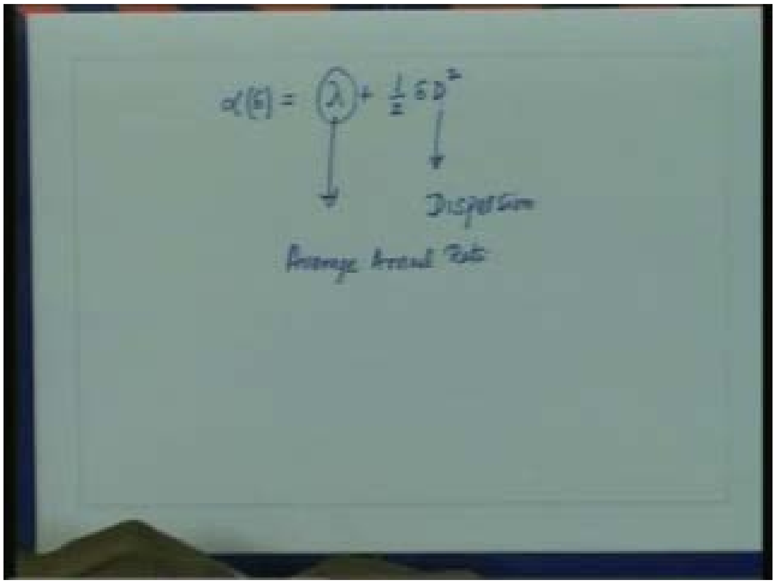
Now, what is A(t) here? A (t) denotes the number of bits or cells, whatever maybe the unit in t seconds. So, what we are trying to say is that in this result we can prove that for many continuous and discrete arrival processes, this result will hold good provided, this condition is satisfied, where this alpha k delta for each individual traffic source is given by this quantity.

Now, to further get an insight into this alpha delta, we assume let us say that this delta is very small. Let us say that delta is very less than 1. Then this quantity which is log of sorry this is not e raised to power minus delta log of expected value of e raised to power delta A(t) can be approximated by log of 1 plus delta expected value of A (t) plus delta square by 2 expected value of A square t. We can approximate this quantity by this which can be further approximated as delta expected value of A(t) plus delta square by 2 expected value of A square t.

Now, if you take limit t tends to infinity 1 by t; then what you get? You get alpha delta. That means alpha delta, if you substitute this into here; then alpha delta is approximately equal to limit t tends to infinity 1 by t expected value of A (t) plus limit t tends to infinity 1 by 2 into delta into e raised to power A squared.

Now, if you look at this, what is this? Expected value of A(t) is the average number of bits in the interval t. Divide by this t and then if you take the limit t tends to infinity, this gives you the notion of the average arrival rate and this gives you the notion of limit t tends to infinity. This should be 1 by t also here, this gives you the notion of some kind of a dispersion.

(Refer Slide Time: 23:35)



So therefore, we write this alpha delta into the form as alpha delta is given by lambda plus 1 by 2 delta D square, where we D we call this D square to be the dispersion - second order notion and this has the interpretation of the average arrival rate.

So, what we were saying is that if our delta is very less than 1, then alpha our delta can be approximated by lambda plus 1 by 2 delta D square, where lambda is the average arrival rate. So, the effective bandwidth is essentially determined by the dispersion for small values of for small values of delta.

So, what we are trying to say is that if we… Now, those small values of delta means what? The small values of delta, if you look at this; if this is this is delta - limit B tends to infinity 1 by B log of probability x greater than B. Now, this delta is very small; then what it means is that we are willing to lose quite a few cells, quite a few numbers of bits we are willing to lose. If this is the case, then what we are trying to prove is that the effective bandwidth will be largely determined 1 by the average arrival rate and around the average arrival rate, the dispersion that is a second order moment.

However, if you make the delta to be large, then we cannot have this approximation and in that case, the other higher order moments will also be called into question. So, what essentially means is that this dispersion for the small values of delta, these values of dispersion where we have said lambda plus 1 by 2 delta D square, these small values for small values of delta, this is actually capturing the traffic streams. So, this is basically capturing the burstiness of the traffic stream.

In other words, what we have actually shown is that… So, what is our conclusion? Our conclusion is that the effective bandwidth of a source is basically that rate with which the transmitter should serve such that the loss rate or the buffer overflow rate decays exponentially and if we have certain target value of the packet loss rate; then by determining the effective bandwidth we can actually determine how much ah bandwidth we should reserve at a particular node.

So therefore, for this random bursty sources effective bandwidth have the same notion which for a constant bit rate source, its bit rate, its constant bit rate or its peak rate will have that notion. So, the 2 notions are actually similar. So, effective bandwidth play the same role which a constant bit rate will play the role in the case of a constant bit rate source. So, what does it mean?

It means that if we have to provide a quality of service guarantees to the large number of the bursty traffic sources which are basically statistical in nature and all of these traffic sources are sharing a buffer and this buffer is being served by a transmitter with a constant transmission rate; then in order to determine how many such numbers of sources can be comfortably admitted such that the buffer over flow rate, the packet loss rate does not exceed a certain tolerable value, then all the admission control has to determine that the sum of the effective bandwidth of these traffic sources should be less than the transmitters capacity or the output link rate or the output bit rate of this transmitter link.

So, that is all we need to determine and therefore the admission control problem simply reduces to determining the effective bandwidth of a traffic source. Now, we can also post the other problem that we had earlier posted; of course, in this problem also. So now, in this problem what we are assuming is that we are assuming that there is a buffer. Whenever the combined output rate exceeds the link capacity, then the packets will be buffered. All we are assuming that either the packet loss rate does not exceed a certain tolerable value or the delay does not exceed a certain allowable value. So, that is what we can ensure by giving quality of service guarantees when have a buffer.

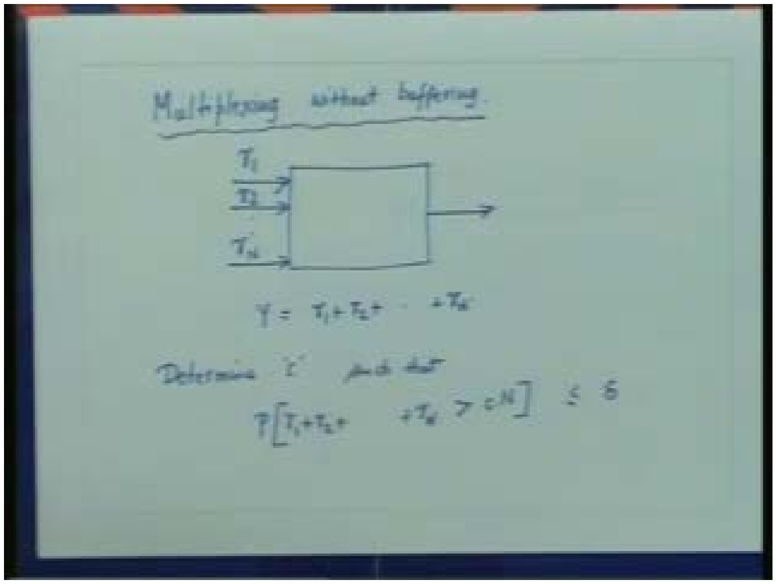
Now, we earlier also considered that let us assume that there is no buffer. That means these sources really want no delays but may be they can tolerate certain packet loss. So basically, there is no buffer but only multiplexing is there. So, the other case is at the other extreme, there is no buffer but there is a multiplexing.

Now, in that case, we need to determine how many such sources we can multiplex such that the packet loss rate remains below a certain value. Now, to determine that I had earlier said that we can determine provided, we know the traffic characteristics of these sources that is their complete distribution functions.

However in practice, again it is not possible to determine. So, one way that has been suggested in the literature is to apply some kind of a bound on these multiplexed output rates. So, let me just tell you one such example of that bound and then we will conclude this discussion by throwing some light on how these things can be actually done in practice because it turns out that if you have to have a real statistical multiplexing gain, then the traffic characterization of a source in some form or the other is required in order to actually capture the full statistical multiplexing gain and unfortunately to have the complete traffic characterization in whatever forms in whatever forms is difficult in practice. So, I will just throw some light on how the things can be done in practice for admitting these numbers of sources.

So, let me just now illustrate you the case of multiplexing, we had considered a case of buffering where we tried to determine the effective bandwidth or tried to explain the concept of effective bandwidth. Now, let me just explain you the case of multiplexing, multiplexing without buffering. So, let me just call it… Even in the case of buffering, there is a multiplexing. But we are now taking of multiplexing without buffering.

(Refer Slide Time: 31:40)



So, what I am saying is multiplexing without buffering. Now, the objective of this multiplexing… What is the objective of this? Now remember, this different traffic sources are fluctuating randomly and you would like to reserve certain bandwidth for these traffic sources such that your packet loss rate is small.

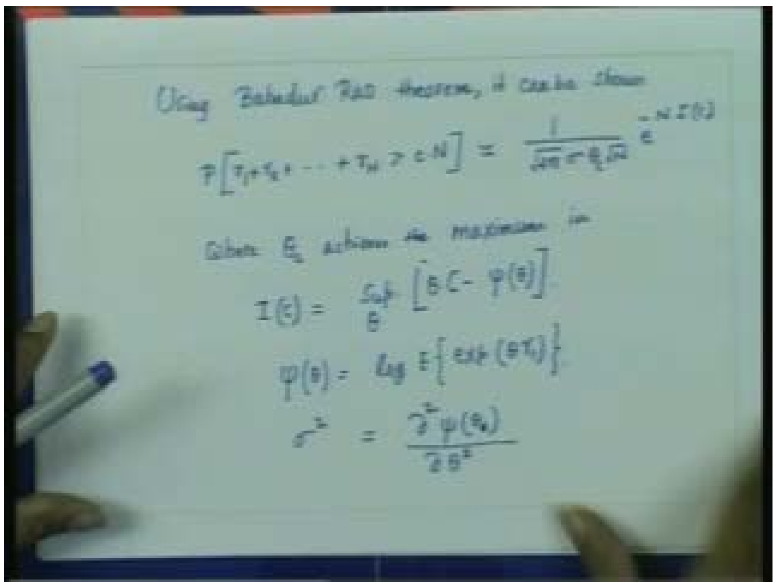
Now of course, if you know the complete traffic characterization of the source in terms of its distribution function as we have seen earlier also the problem is very simple. But otherwise, since we do not know the complete traffic characterizations; we have to reserve how much bandwidth? Typically, we have to reserve the bandwidth very close to the average rate, slightly maybe more than slightly more than the average rate such that our packet loss rate is below a target limit.

So now, let us say that we want to multiplex these N sources. So, our problem essentially is that we want to multiplex these N sources whose bit rates are r_1 r_2 so on upto r_n. Now, these bit rates are fluctuating randomly. However, we are assuming that each of these sources is stationary and therefore these bit rates are can be considered as random variables and we are also assuming that these different traffic sources are heavy, are independent and therefore we are considering the case of an independent and identically distributed random variable.

Now, when we multiplex; we get Y which is basically equal to r_1 plus r_2 plus r_n. Now, we want to determine this problem, so we want to determine c such that the probability that this r_1 plus r_2 plus r_n that is our Y is greater than c times n is less than or equal to certain target value. Let us call it delta. So, that is what we want to determine.

Now typically, the value of this c will be slightly larger than the average rate of each traffic sources, closer to the average rate. So, there are N sources. So, we ask this question that what is the probability that if you had reserved this c or each of these traffic sources, then we have reserved total c into N for all N traffic sources and we want to ask this question that what is the probability that some of these rates is greater than that c times N is less than or equal to certain target value which we have put which may be 10 raised to the power 9, 10 raised to the power ten raised to the power minus 9, 10 raised to the power minus 8, 10 raised to the power minus 10, whatever is the quality of service attributes.

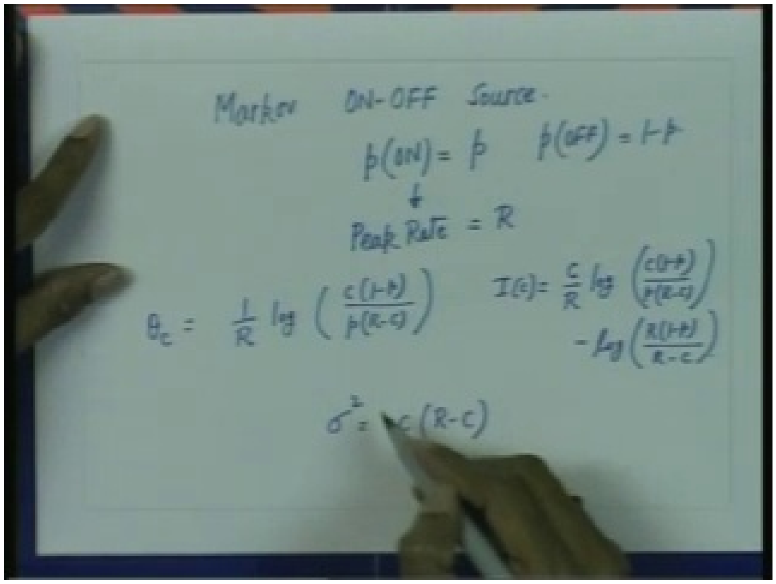
(Refer Slide Time: 35:18)



So now, we can apply the Bahadur Rao theorem for this. So now, using Bahadur Rao theorem, it can be shown that the probability that r_1 plus r_2 plus r_n is greater than c times N is approximately equal to 1 upon root 2 by sigma theta c root N e raised to power minus N into I(c), where theta c is achieving the maximum in this functions, supremum of theta, theta into C minus psi theta, where what is psi theta? Psi theta is nothing but a logarithmic moment generating function expected value of theta r_1 and sigma square is… So where, theta c achieves the maximum or the supremum in this I(c). Psi theta is actually a logarithmic moment generating functions and sigma square is given by this.

So, what does it mean? It means that if we know the logarithmic moment generating function, if you know this quantity, so if you know the traffic characterization in such a manner, if you know this quantity; then we can compute this and can compute this. And, if we can compute this, we can put into this and from that we can determine what is the value of c which satisfies a certain target limit which is given which is a delta.

(Refer Slide Time: 38:00)



So, let us take a simple example and a simple example is if you assume that we have a Markov on-off source, we assume that we have a Markov on-off source with the probability that source is ON is some p and the probability that source is OFF is of 1 minus p. In the p ON, the peak rate of transmission is some quantity let us say R and when it is OFF, it does not transmit anything. Then it can be shown that theta c is 1 by R log of c into 1 minus p upon p into R minus c. I(c) is given by the c by R log of c into 1 minus p, p into R minus c minus log of R into 1 minus p upon R minus c and sigma square is given by c into R minus c.

So, if you put all these quantities into here, you can actually compute this probability distributions, the bound on that and if you know this, then we can actually determine what is the value of c that I have to choose such that if I have to multiplex N such sources, the packet loss rate is you know less than a certain quantity. Just to give you a… So basically, we can determine all these quantities if you have a Markov on off sources.

So, what does it mean that even in this case, we at least need to have a certain characterization of the traffic source. If we have this characterization of a traffic source, we can apply the Bahadur Rao theorem and then we can determine that how much bandwidth must be reserved per source such that the packet loss rate is bounded by certain or a target value.

So in practice however, the question is really that how do you apply these techniques and in practice, one can also do some kind of an online estimations or online measurements. What when can do as I said is that one can start by reserving the bandwidth based on the peak rates initially and when the traffic source is actually start transmitting the data, one can make online measurements on the traffic sources and then apply the Bahadur Rao theorem to actually determine what is the value of c which is actually a closer to the average rate of these various fluctuating traffic sources.

So, in practice really what will be really done is that we may not be knowing the apriori traffic characterization of the source, one can actually be more conservative by admitting the users based on the peak rates and then one can do an online measurements and then apply either the Bahadur Rao theorem or even the Gaussian approximations that we had considered in the previous lectures to the combined output source.

Now, we have considered the two scenarios. Let me just briefly summarize that. The two scenarios are; one is the case of a scenario where we are multiplexing these various sources which has no buffer and therefore they cannot tolerate any delay, there will be zero delays but we want to ask this question that how many such sources can be multiplexed such that the packet loss rate is certain tolerable limit – so, that was one scenario.

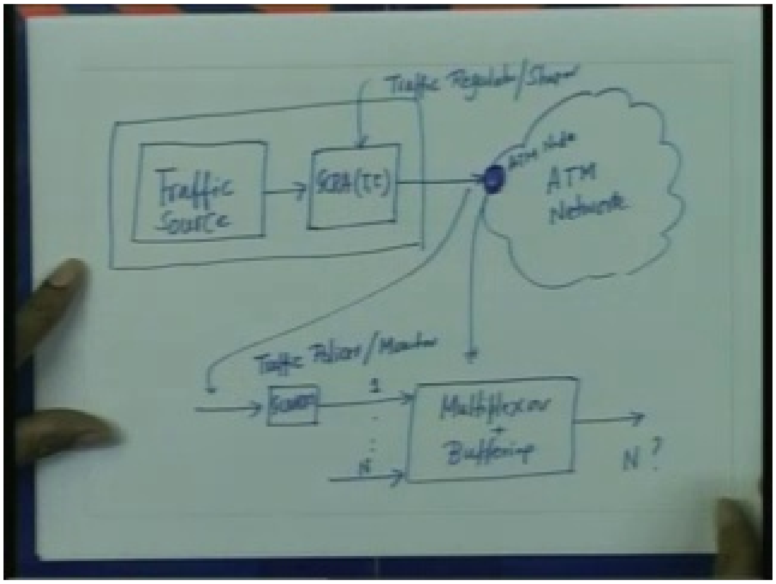
The other scenario was that these sources can tolerate certain delays and however and therefore they can be buffered. But since these buffers are also finite, when this buffer overflows or the buffer exceeds a certain limit; there will be a packet loss and we want to limit that. We want to limit the maximum delay or the q length, average q length and we also want to limit the probability of these buffer overflows, then we ask this question that how many these sources can be multiplexed.

Of course, one can consider a combination of these multiplexing with buffering and so on and can determine how many such number of sources can be multiplexed. But as we have seen that in both these cases, either we apply the centre limit theorem that combined traffic characterization looks like a Gaussian or a normal probability distribution function or we apply the Bahadur Rao theorem on the one hand and on the other hand, in the case of buffering, we try to determine the effective bandwidth of a traffic source. In both these cases we need to know the traffic characterization of a source which in general is difficult to achieve in practice.

So, what is done in practice? The best thing that can be done in practice is to follow a certain kind of deterministic approach and the ATM forum has suggested that that deterministic approach can be done by bounding, you know; by bounding the peak rate, the average rate, the burst tolerance and the cell delay variation tolerance of these sources.

So, we can bound that and so therefore for that we need to put some kind of a shaper in front of a traffic source and the shaper that has been suggested by the ATM forum standards is a leaky bucket shaper. So, what you do is that a traffic source which is essentially statistical in nature, you put a leaky buckets shaper in front of the traffic source and make sure that the output of these leaky bucket traffic shaper confirms to the peak cell rate, the average cell rate, the burst tolerance and the cell delay variation tolerance. And then, the network determines whether by admitting such deterministically bounded traffic, deterministically bounded traffic whether the sources can be multiplexed or not.

(Refer Slide Time: 44:55)



So in practice, basically, what we are doing is we are putting. So, here is a traffic source here is a traffic source. This traffic source passes through a GCRA (T, tau) and here is ATM traffic. So, this now of course to some kind of an ATM network.

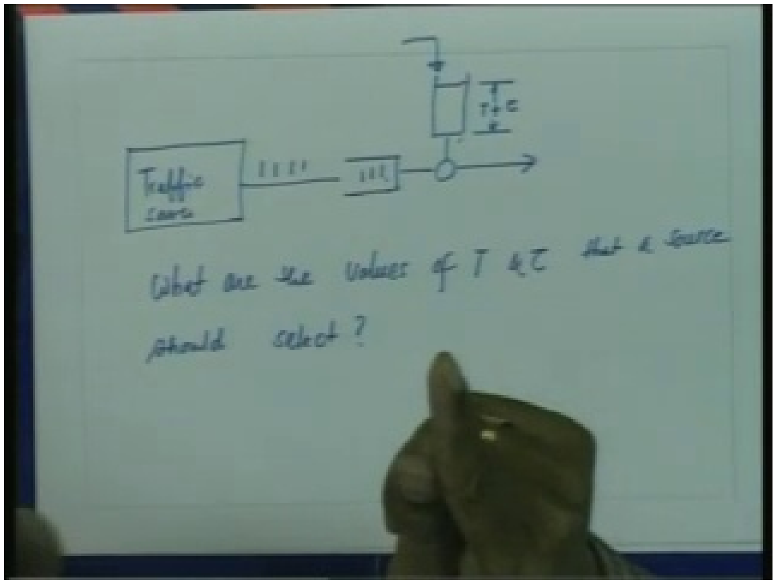
Now, this make sure this makes sure that this traffic source is now confirming to the GCRA (T, tau) parameters. Now inside if you look at, if this is an ATM node if this is an ATM node; then this ATM node if you expand this will look something like this that each of this ATM node will put a GCRA (T, tau) monitor and this will be a multiplexer buffer.

So, this is while this GCRA (T, tau) which has been put in front of the traffic source, we can call it to be a traffic shaper or a regulator, traffic regulator or a shaper and this GCRA (T, tau) parameter which we have put at a network node; so this is the node, ATM node, so this is my ATM node where I have put this GCRA (T, tau) is actually a traffic pleaser or a monitor.

Now, this is making sure that this traffic source is confirming to these GCRA (t, tau) parameters. This is necessary because otherwise, the traffic source might just remove these GCRA (T, tau) parameters and then start sending data which is violating the advertised traffic contract. So, this is how we put and then the node determines how many of these traffic sources like from 1 to N can be admitted, so basically the NO determines what is a value of N.

Now, the question really is that, we just spend some time on this traffic source. Now, if you look at this, now this traffic source essentially is statistical in nature and this traffic regulator is making sure that the output of this traffic regulator is confirming to this advertised traffic contracts T and tau.

(Refer Slide Time: 47:33)



So now, what is the situation here? The situation here is that here is a traffic source which is transmitting these cells and they make pass, this traffic GCRA (T, tau) is essentially is a leaky bucket and as we had considered this leaky bucket has a depth of T plus tau; the fluid is accumulating at a unit rate and then the sources is transmitting and it has a bucket depth of T plus tau.

Now, the question really is that whenever an ATM cell comes and if it finds that there are not enough fluid in the bucket, the cell will have to be buffered here. So, the way in which it will work is that whenever a cell comes, it takes away capital T units of fluids and if a cell comes and if it finds that the bucket does not have enough fluid that is capital T units of fluid, it will be buffered.

So, note that this regulator or shaper is basically changing the traffic characteristics also to suit the output of T and tau parameter. So, the question really is that what are the values of T and tau that a source should select? Here ultimately, what is happening is that there is a delay at the source side itself. So, what is happening really is that a traffic source is statistical in nature, you are putting some kind of traffic regulator or shaper such that the output of a traffic regulator confirms to these parameters.

But what are those parameters? How does the source determine that these parameters are such that a delay which will happen in the buffer right at the leaky bucket is not so high? It should not distort the traffic characteristics completely. So, what are those values of the T and tau parameters?

Now typically, it is very difficult to determine the values of T and tau parameters. The network what will do is that will offer to a source of certain range of parameters of T and tau and from which the traffic source will be expected to choose from these range of parameters and then the network or the source can determine that what parameters suits it in terms of both the pricing and the distortion that a traffic source will suffer because each of these parameters may have a certain cost associated with it that calls, maybe priced based on the traffic descriptors and also the quality of service attributes.

So, the pricing of a call in a quality of service environment will be a function of both the traffic descriptors as well as the quality of service attributes. So, by a combination of the distortion that a traffic will undergo if you put a traffic regulator or shaper and the pricing of the call which a network is offering, a source will be able to determine or we will be able to be judge what parameters it should choose - the T and tau so that you can also achieve a statistical multiplexing gain from the networks point of view, at the same time you can achieve quality of service guarantees and also at same the time, the traffic distortion which is occurring by putting this regulator is minimum.