Broadband Networks

Prof. Karandikar

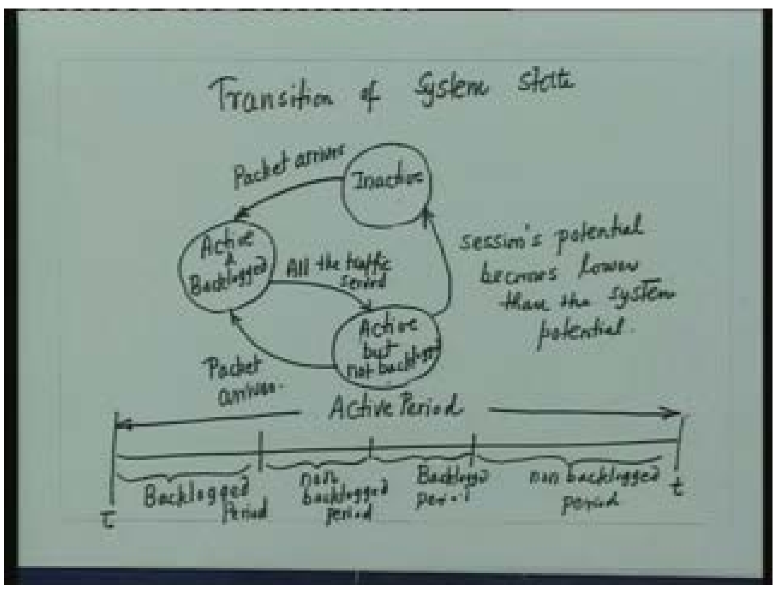
Electrical Engineering Department

Indian Institute of Technology, Bombay

Lecture - 14

So, we were discussing about rate proportional servers and we saw that in the rate proportional servers, session can be in any one of these 4 states. That is it could be either in the ideal states or it could be in the active state or it could be in the backlogged or non backlogged states. So, we were discussing in the previous lecture about the state transitions of the session's various states. So, let me just revise that.

(Refer Slide Time: 1:53)



So, what we were saying is that session could be in the inactive state that is the ideal state and when a packet arrives, the session becomes active and backlogged. So, we say that a session is in active state if the session's potential is greater than or equal to the system potentials. So, when the session is inactive and when a packet arrives, the session's potential is updated and may be kept equal to the system potential. So, in that case it is active and also the backlogged.

Now, as the session starts receiving service, whenever a session starts receiving service; then the session's potential increases and when all the traffic has been served, the session remains active because its potential will be more than the system potentials. But since, all the traffic has been served; it will be now not backlogged.

Now, if the session continuous to be in this state that is the active but not backlogged state; then what will happen is that the session's potential, the session's potentials will remain constant now because the session is not receiving any service. But on other hand, the system potential will keep on increasing. So, as a result, a time may come when the session's potential may become lower than the system potentials. Now, when that happens, as we had seen that the state moves from the active non backlog state to in active state; the state moves here.

But on other hand, if the session potential was greater than the system potential and the session was not receiving any service that is it was not having any traffic also that means it was not in backlog state and then suddenly if some traffic arrives and the queue becomes non empty; then the session can even move to the active and the backlog state.

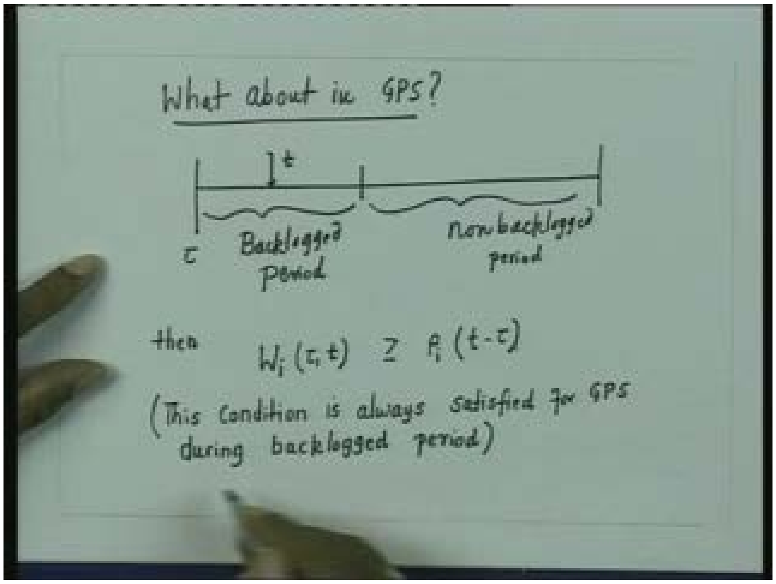
So, you can see in the various cycles that the active period starts here at tau for the first time here, the state was inactive here; so, when you become from inactive to active and the active period starts here at tau, your backlogged period also starts because the traffic has arrived and at this point all the traffic which was there here or which arrived during this period has been served and the session therefore enters a non backlog period.

Now, during this period, the session's potential is still higher than the system potential and at this point, some traffic might arrive and therefore the session moves into the backlogged period. Again, session's potential keeps on increasing. At this point, all the traffic that arrived during this period has been served and therefore the session moves into a non backlogged period.

Now, in this non backlogged period, the session's potential is not increasing. But a system potential keeps on increasing so that at point at time t, the session potential will become lower than the system potentials and the session then moves into here and inactive states. Now, this is what we have seen - the state transitions in the rate proportional servers.

Now, we had asked this question that what about in the fluid flow where queuing algorithm or in the GTS server?

(Refer Slide Time: 5:35)



So, we see that in the GPS server in the GPS server the session will either be in the backlogged period or it will be in the non backlogged period. So, when the session is in the backlogged period at time tau, the session becomes backlogged and the session starts receiving service and during this period, the session's potential is also equal to the system potentials because the system potential in the case of a fluid flow server is always equal to the session's potentials in the fluid flow GPS servers.

At this point, all the traffic that arrived during this period has been served and therefore the session enters the non backlogged period. Now, during this point, the system potential will keep on increasing and it will be again equal to the potentials of those sessions which are backlogged during this interval. But since, this session is ideal its potential is not increasing.

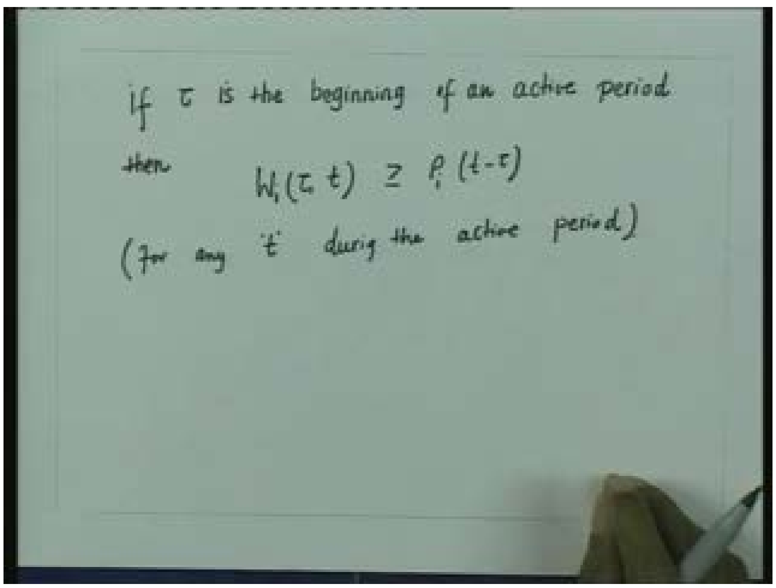
So, now the question is that let us say that t is in any time period during this backlog period in the GPS server, then it is clear that the service received by the session i during the period tau to t is always greater than or equal to rho i, where rho i is the rate allocated to the session i into t minus tau. Now, this is always satisfied, this condition is always satisfied during the backlogged period. So, I should say that this condition is always satisfied during for GPS server during backlog period, for GPS during backlog period.

That is what we are saying? That the service received by any session during sub interval, and this sub during a sub interval of the backlog period is always greater than or equal to rho i into t minus tau. Whether it this is whether this is true for rate proportional servers? Whether this is true for rate proportional servers? No, this will not be true always for rate proportional servers. That is for rate proportional servers; if we say that tau to t is any sub interval of a backlog period, then for rate proportional servers the amount of traffic that is W_i (tau, t) will not be greater than or equal to rho i into t minus tau.

So, why it is? because, it may be just possible that a session can be backlogged, the sessions can be backlogged in a rate proportional server. But, still it may not be receiving still it may not be receiving any service in the rate proportional servers. So, that is possible because it is possible that service of a particular session may get temporarily suspended in the rate proportional server and this will happen if the total service if the total service received by this particular session has become greater than the minimum service and its potential is not minimum so that the other sessions who have minimum potentials, they will be served.

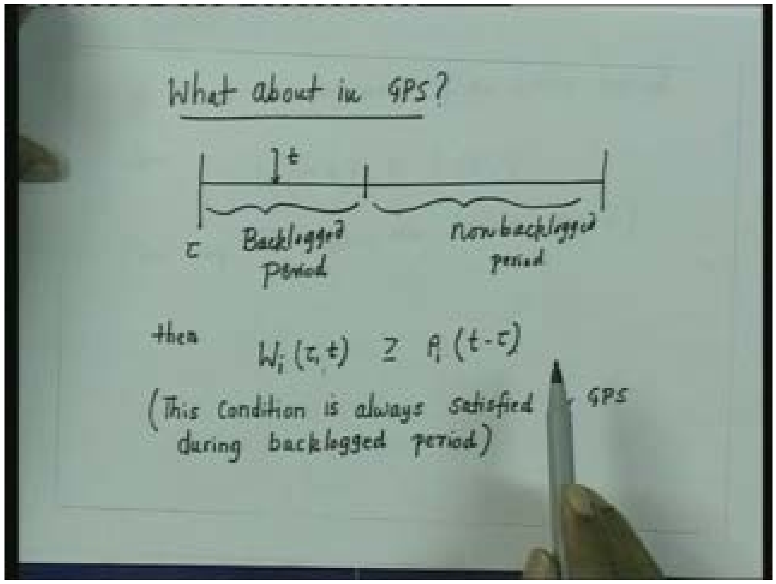
So, it might be quiet possible that a particular session, even though it is backlogged, its service may be temporarily suspended in the rate proportional servers. So, as a result, these particular relationships will not hold good. Then what kind of relationships holds good in the RPS? So, what happens in the RPS? So, let us see. If this relationship does not hold, then what happens in the RPS?

(Refer Slide Time: 9:48)



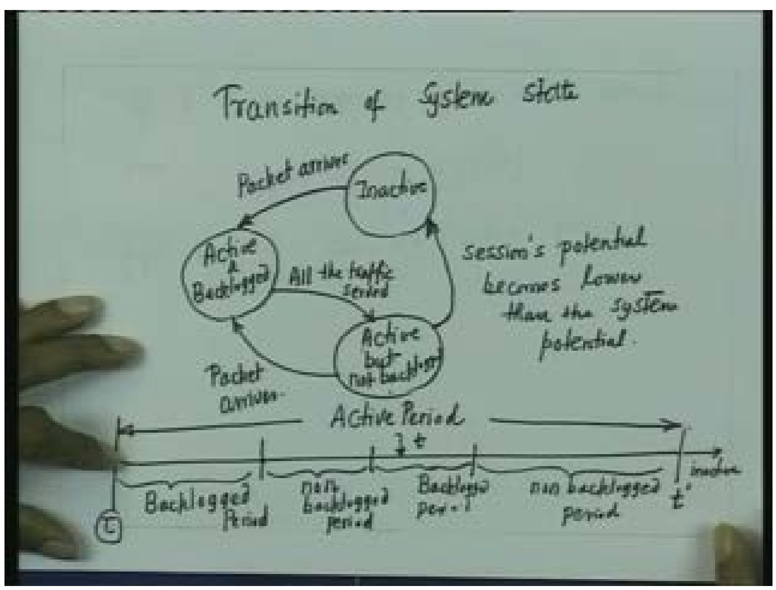
So, in RPS if we see that if tau is the beginning of an active period, if tau is the beginning of an active period, then this result will be guaranteed. So, if tau is the beginning of an active period and then for, this is for any t during this active period; so, this result needs to improved.

(Refer Slide Time: 10:51)



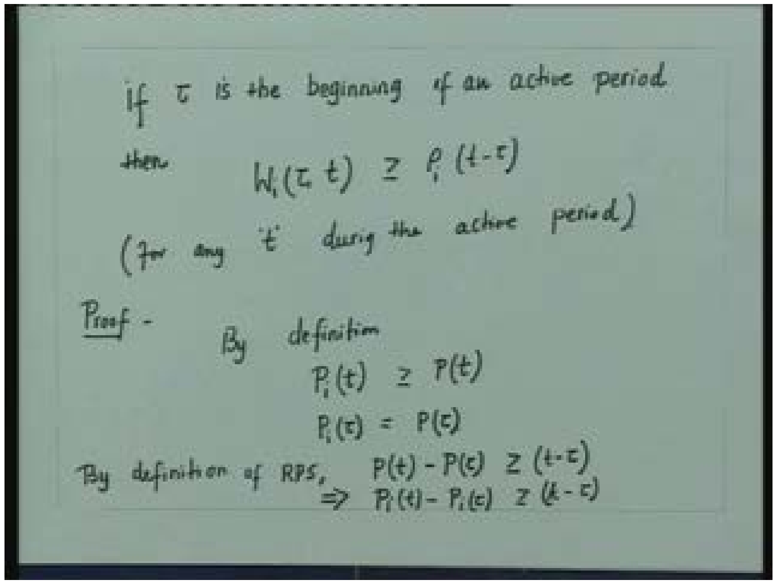
So, on other hand as we had seen that in the case of a GPS server, this result will hold true for any tau to t, any interval tau to t which is sub interval of the backlog period. But this result does not hold good for the rate proportional servers. However, for the rate proportional servers if tau is the beginning of an active period, then this result will hold good.

(Refer Slide Time: 11:12)



So, if you can see here that this is the rate proportional server state transition diagram. So, if tau happens to be the beginning of an active period and if t is any time here and let us say the time at which this end is tau prime; if t is any time here, then the amount of service received by a particular session i during this period of tau to t will be greater than or equal to rho i into t minus tau.

(Refer Slide Time: 11:36)

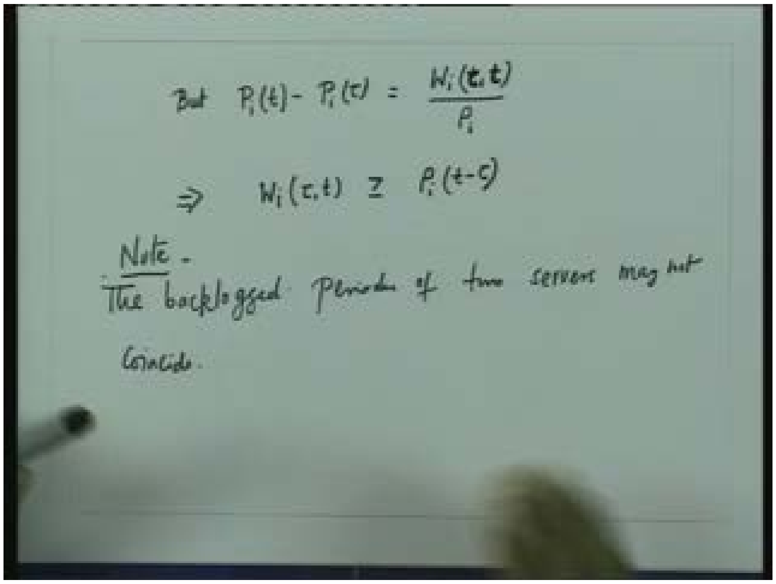


Let us prove this result. So, I will try to give a sketch of this proof. So, we know by definition the potential, the session's potential, the potential of the session i will be always greater than or equal to the system potential at time t where t is any time during the active period. So, during the active period by definition, the session's potential is always greater than or equal to the system potential - that is by the definition of the active period.

Now, tau is the beginning of the active period. That is when the system, the session has just changed its state from an inactive period to the active period. So, tau is the beginning of that active period. So therefore, at that point - the tau, the session's potential will be updated and it will be made equal to the system potential. So therefore, P $_i$ (tau) is actually equal to P (tau).

Now, you know by definition of rate proportional server, by definition of rate proportional server; the difference in the system potential that is P (t) minus P (tau) will be greater than or equal to t minus tau. But P (t) is actually P $_i$ (t) is actually greater than or equal to P (t) and P $_i$ (tau) is actually equal to P (tau). So therefore, this implies that P $_i$ (t) minus P $_i$ (tau), they are greater than or equal to t minus tau. But during the active period during the active period; the session potential, the increase in the session's potential that will be equal to the normalized service the session has received.

(Refer Slide Time: 14:08)



So, P_i (tau) minus P_i (tau); that is the increase in the session's potential is always equal to W_i (t, tau) upon rho i. Now, this is the normalized service received by the session i during this active period. So therefore, W_i sorry this should be W_i (tau, t) will be then greater than or equal to rho i into t minus tau and hence we have proved a result.

So, what does it say is that service received by a particular session i during a sub interval, during a sub interval of the active period, starting from the active period and during any interval of the active period; the service received by a particular session i will be greater than or equal to its allocated rate into the durations. So, this is the result we have proved.

Note that this result holds true for any subinterval of the backlog period in case of a GPS server - generalized processor sharing or fluid flow fair queuing server, that we have been discussing. However, this result does not hold true in the case of a rate proportional servers that the service received by a particular session i will be greater than or equal to its allocated rate times the time durations because in the rate proportional server, a particular session even though it is backlogged may not be receiving any service. Its service might have been temporarily suspended because its system potential may not be the minimum and other backlogged sessions which are having minimum system potentials, they are being served.

So therefore, the service of a particular session may be temporarily suspended also and therefore it is possible that a particular session is not receiving service and that will happen only if a particular session has received more service than its minimum guaranteed service rate. Now, other thing is we also see that the backlog periods of 2 servers need not coincide. This also shows that the backlogged periods, the backlogged periods of 2 servers may not coincide.

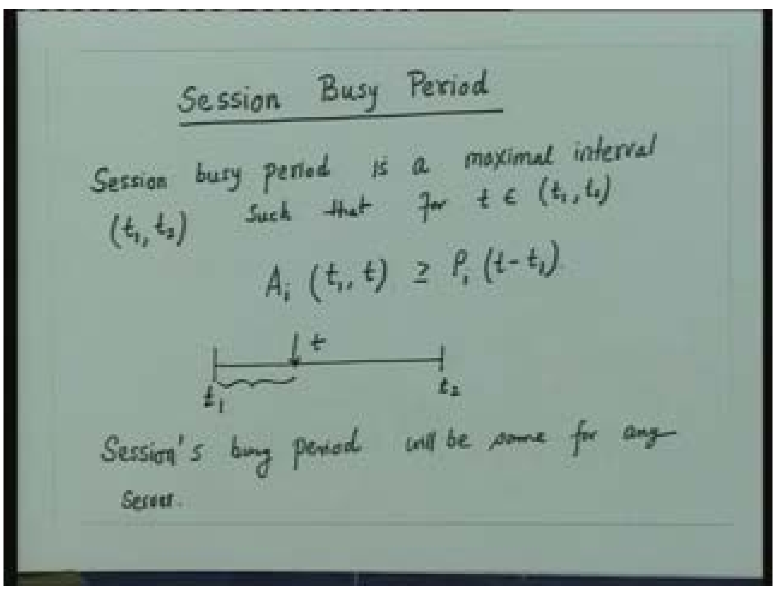
Why? because in the GPS server that is in the fluid flow fair queuing server, if a session is backlogged; then it will always receive its service and the system potential or as it is called system virtual time in the GPS servers, that will increase in proportion to the normalized service received by backlogged session. And, the normalized service received by 2 sessions that are backlogged in the GPS is always equal during any sub interval of the backlogged period.

So, the service received by the backlogged sessions is always equal during the backlogged period for the sessions in the GPS servers. But this may not be true in the case of a rate proportional server and therefore as we have seen that a backlogged period of the rate proportional server may not be the same as the backlogged period in the case of a GPS server.

So, now if this is so, then now we suppose we want to compare the 2 scheduling algorithms in terms of fairness or the delay properties; then these questions obviously arises that is the backlogged periods of 2 servers are not coinciding, then how do we compare the 2 servers.

Now, towards this purposes, to compare the properties of 2 scheduling algorithms; we define some notion of a session busy period. We have already seen the notion of a system busy period. What is the notion of a system busy period? The system busy period is any interval of time during which the server is always busy, the new server is always busy. It is servicing some packets or some traffic. This is what is called as a system busy period. We now introduce the notion of a session busy period. So, let us see what is a notion of a session busy period.

(Refer Slide Time: 18:52)



Session busy period: so, the session busy period is a maximal interval is a maximal interval (t_1,t_2) such that for any t during this interval such that for any t during this interval of (t_1, t_2), we have the arrivals that is A_i (t_1 into t) is greater than or equal to row $_i$ (t minus t_1). What about v? By these relationships it means that the arrivals which have occurred at during the interval t_1 to t_2 that is the number of arrivals.

So, session busy period actually starts here at t_1 and it ends we are saying that it ends at t_2. Now, this t is a time interval here. What is it see, it says the number of arrivals that will occur during this intervals that is t_1 to t_2 is always greater than or equal to rho into t minus t_1. It means that if the session was served with exactly the rate as rho $_i$; not more, not less, if the session was served with exactly the row rate row $_i$, then this session will remain continuously backlogged during the interval t_1 to t.

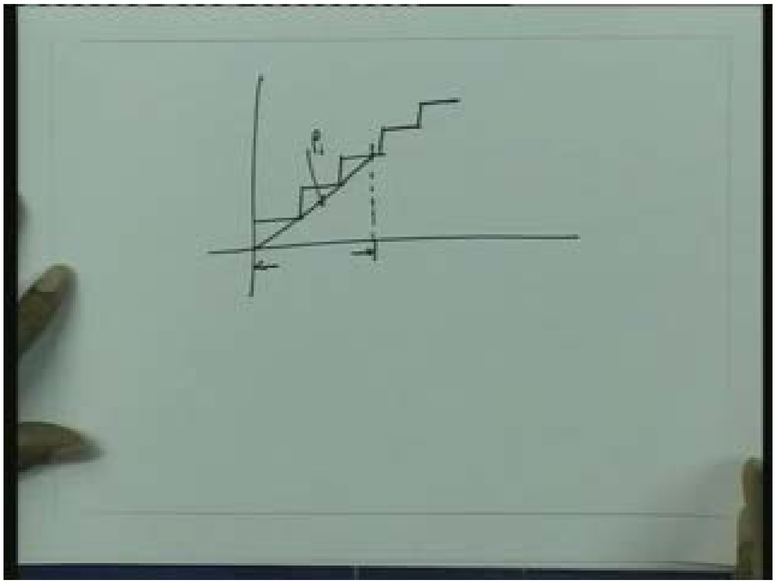
Now, during this interval, it shows that if this relationship is true, then this session is backlogged and obviously if this relationship is true for over this interval t_1 to t_2, then we say that the session is in busy period. This is the called a session's busy period.

Now, note that in the actual server, the session may actually be receiving more service than its minimum guaranteed or allocated rate of rho i. It may be quite possible because if other sessions are empty or non backlogged, if they do not have any packets; then this i'th particular session may be receiving more service than its guaranteed allocated minimum rate also. That would be possible in a particular scheduling algorithm but we are now defining a hypothetical server.

We are saying that there is an hypothetical server and in this hypothetical server, suppose if you serve the sessions with the rate rho $_i$, then if we can say that the number of arrivals during any sub interval of a particular period is always greater than or equal to row $_i$ times that interval, that is the service received by this particular sessions. If the number of arrivals is greater than or equal to the service received or the number of traffic or the traffic, the amount of traffic that has been served; then this particular session will remain backlogged.
So, this is session busy period is therefore defined with respect to a hypothetical server and therefore the busy periods of all the sessions will be equal irrespective of the fact that we have used different scheduling algorithms in different cases. So, the session's busy period session's busy period will be same for all scheduling algorithms.

Now, this is an important this is an important notion and therefore what we would like to see is that we would like to understand the behavior of a particular scheduling algorithms during a session's busy period. That is what we would like to do.

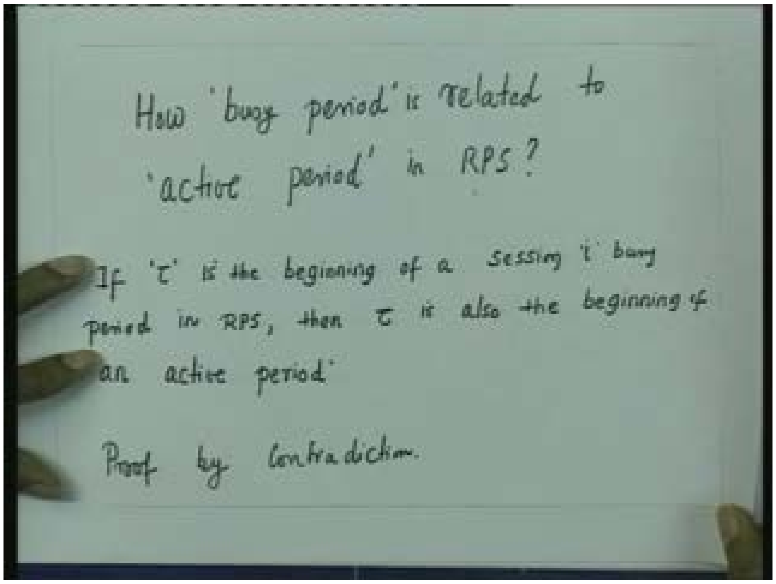
(Refer Slide Time: 23:30)



Now, session's busy period, we just draw it graphically. It looks something like this that here are the arrivals, arrivals are the stair case because the arrival increases whenever the last bit of a particular packet has arrived. So, here is a session's busy period. This curve has a slope rho $_i$. So, you can see that during this interval, the arrival is always greater than or equal to rho $_i$ into t. After this the arrivals are going to become greater, sorry the service is going to become greater and therefore this will not be the session's busy period. So, this is the session's busy period.

Now, we would like to study, we would like to ask this question that what how the busy period is related?

(Refer Slide Time: 24:45)



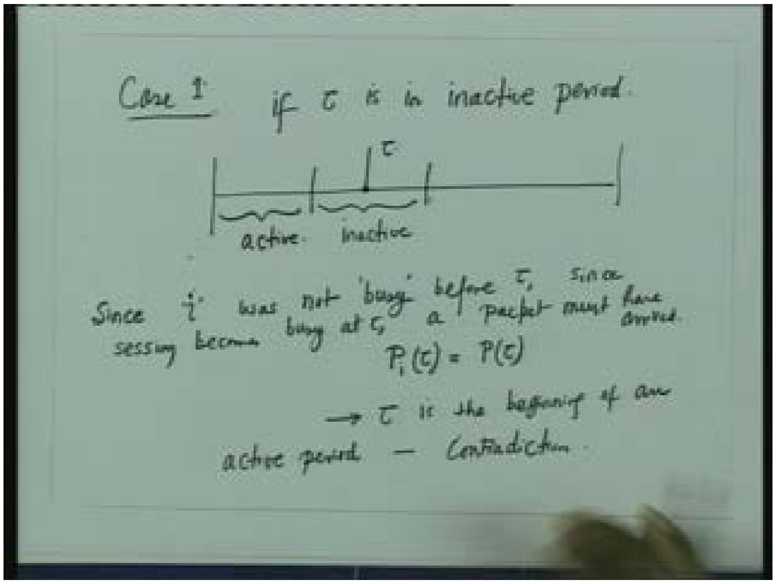
So, how busy period is related to active period? How busy period is related to active period in rate proportional servers, this is the question that we would like to ask. And, answer of this question is… basically, we would like to ask that in what way we can relate the session's busy period to an active period in the rate proportional servers?

Now again, I want to emphasis here that the session's busy period is defined with respect to a hypothetical server in which the server is serving with a rate rho $_i$ and what we are saying is that if the server serves with a rate rho $_i$; then the session's busy period is that maximal interval of time during which the number of traffic arrivals exceeds the amount of service received by the session i. So, that is how we define the session's busy period.

The system's busy period is any interval of time during which the system will remain busy. So, note that system's busy period, the system's busy period for all scheduling algorithms will coincide. We have seen that a backlog period for different scheduling algorithms may not coincide. However, session's busy period will also coincide. The question now that we are asking is that in what manner the session's busy period is related to the session's active period in a rate proportional server.

So, the answer is that if tau is the beginning of a session i busy period tau is the beginning of a session i busy period in rate proportional servers, then tau is also the beginning of an active period. So, we say that if tau is the beginning of a session i busy period in the rate proportional servers, then tau is also the beginning of an active period. So, we will try to prove this by contradictions. So, we have a proof by contradictions.

(Refer Slide Time: 27:56)



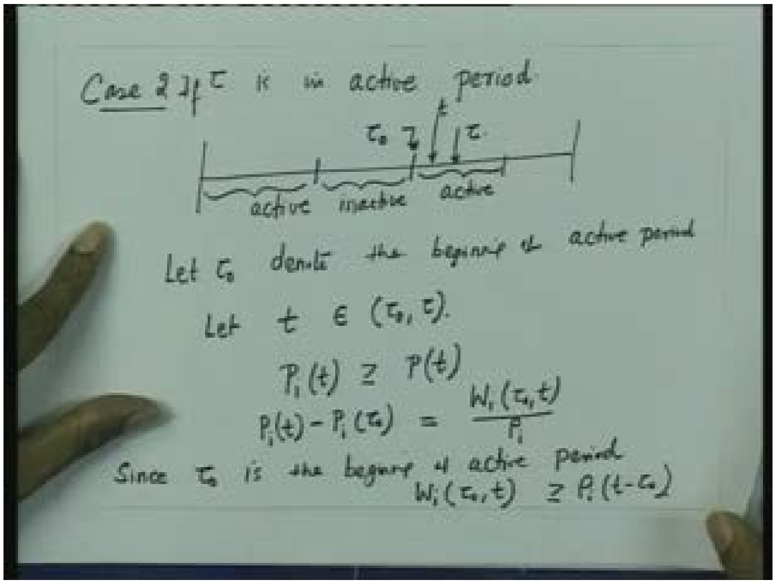
So, we take a case. So, case one is if tau is in inactive period, so the tau what we are saying is the tau is the beginning of a busy period. But tau lies in the inactive period, assume. So, something like this - if tau is the beginning of a session i busy period in rate proportional servers, so by contradiction, so what we are saying is the tau lies in an inactive period. So, this period is active and this period is inactive. Let us say that tau is here.

Now, since session i was not busy before tau and the session becomes busy at t and the session becomes busy at tau, then a packet must have arrived. So therefore, P_i (tau) should be equal to P (tau) and that means that tau is the beginning of an active period also. So, which is a contradiction and therefore this is not possible, the tau cannot be in the active period. So, this is the first case. So therefore, tau becomes the beginning of an active period if we try to prove that tau is in the active period.

Since, what we are saying is that since session i was not busy before tau, it became busy at t only and here it was in inactive period; then here the packet must have arrived. So, that means tau must be here, there inactive period must end here actually. So, since i was not busy before tau and the session became busy at tau since session becomes tau busy at tau; so, a packet must have arrived and therefore p_i tau should be equal to p tau. So, this is first case where tau is in the inactive period.

So, we will now try to prove the second case; when tau case 2 - when the tau is in, somewhere in the active period, something like this.

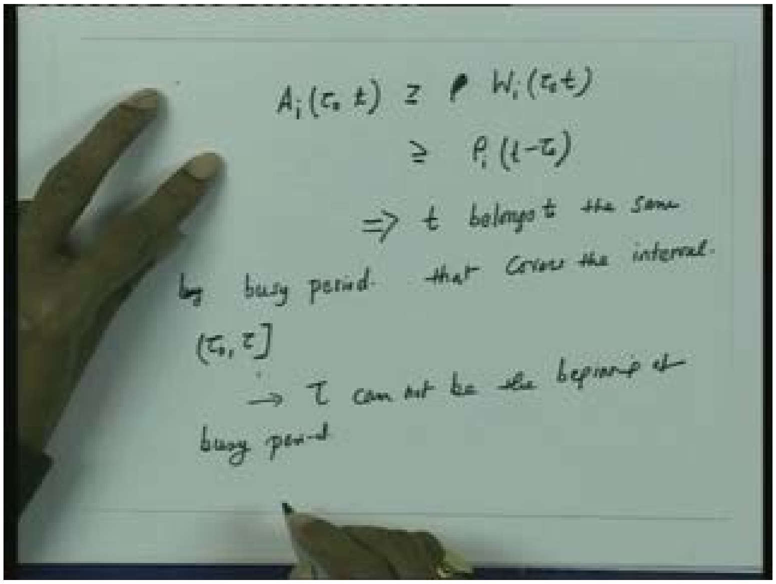
(Refer Slide Time: 31:15)



So, case 2 is when tau is in active period. So, this is something like something like this that here is an active period, again inactive, active, inactive so on. So, let us say that this is the tau and let us say and let us say that tau $_0$, let tau $_0$ denote the beginning of active period. Now, let us say any t belonging to tau $_0$ to t. Now, since t is in active period, P i (t) that is the sessions i potential at time t should be greater than or equal to the system potentials. Now, since tau $_0$ is the beginning of the active period, we should also have that P $_i$ (t) minus P $_i$ (tau $_0$) during this interval, it should always be equal to W $_i$ (tau$_0$, t) upon rho $_i$. This is by definition of the active period.

Now, since tau $_0$ is the beginning of an active period so since tau $_0$ is the beginning of active period, it also means that w $_i$ (tau $_0$, t) should be greater than or equal to rho $_i$ into t minus tau $_0$. Now, since before tau $_0$, the system was not backlogged; before tau $_0$ the system was active, therefore we also have this relationship before tau $_0$ the system was not backlogged.

(Refer Slide Time: 34:07)



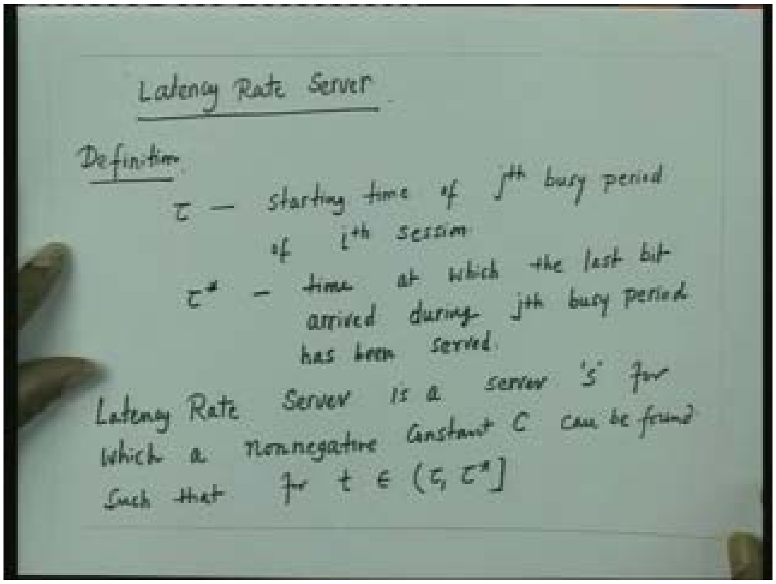
A $_i$ (tau $_0$, t) should be greater than or equal to W $_i$ (tau $_0$, t). Now, which also means that this is sorry this should be greater than or equal to rho $_i$ t minus tau which also means that t belongs to the same busy period that covers the interval tau $_0$ to tau and therefore which means tau cannot be the beginning of the busy period. It cannot be or the tau $_0$ cannot be the beginning of the active period. Then the tau $_0$ has to be equal to tau.

So therefore, sessions what we are really try to prove here is that if tau is the beginning of a busy period, then it is also the beginning of an active period in a rate proportional server. So, in a rate proportional servers, what we have tried to prove is that if tau is the beginning of a sessions i busy period, then it is also the beginning of an active period in a rate proportional servers.

So, we have proved this result by contradictions. So, two cases we had considered: whether the tau is in the inactive period or tau is in the active period. So, by considering these two cases we have tried to prove this result that if tau is the beginning of session i busy period in rate proportional servers, then it is also the beginning of an active period.

So, we now introduce the notion of latency rates servers. So, I will just define the notion of latency rate servers.

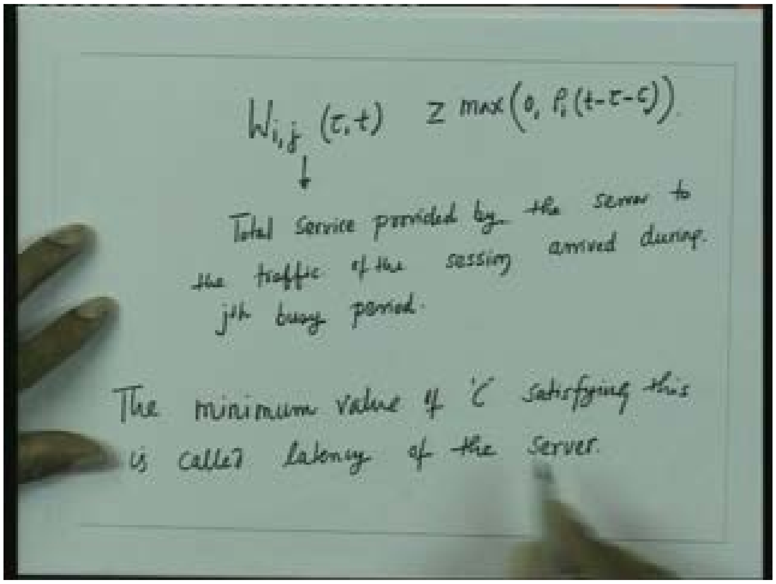
(Refer Slide Time: 37:03)



Now, the definition of the latency rate servers is that so let tau be the starting time or the beginning of j'th busy period of i'th session, j'th busy period of i'th session and then tau star be the time at which the last bit the last bit arrived during this j'th busy period has been served during j'th busy period has been served.

Then, we define a latency rate server, is the definition by definition we define a latency rate server is the server for which a non negative constant can be found, non negative constant Ci C can be found such that for any t which lies between tau to tau star, we have this relationships; for any t between these two things.

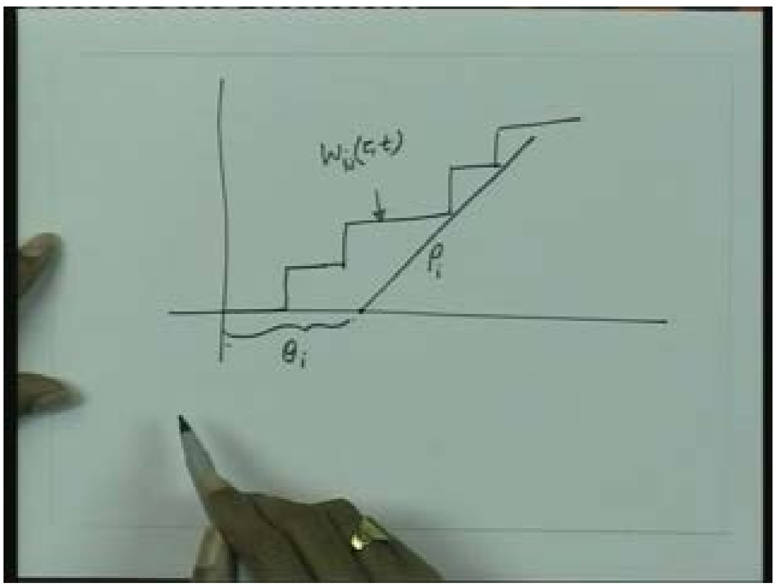
(Refer Slide Time: 39:15)



W $_{ij}$ tau t that is the amount of service received by i'th session in the j'th busy period during the interval of tau to t, this interval of tau to t of the j'th busy period that is greater than or equal to max of 0 or rho $_i$ times t minus tau minus the constant C.

Now, so this includes, this gives you total service provided by the server to the traffic of the session arrived during j'th busy period. Now, the minimum value of C that satisfies this is called a latency of the server. So, the minimum value of C satisfying this is called latency of the server which we call it to be latency theta $_i$. So, just to show you how the latency concept looks like, it will be something like this that this is let us say the service W $_{i\,j}$ tau.

(Refer Slide Time: 41:07)

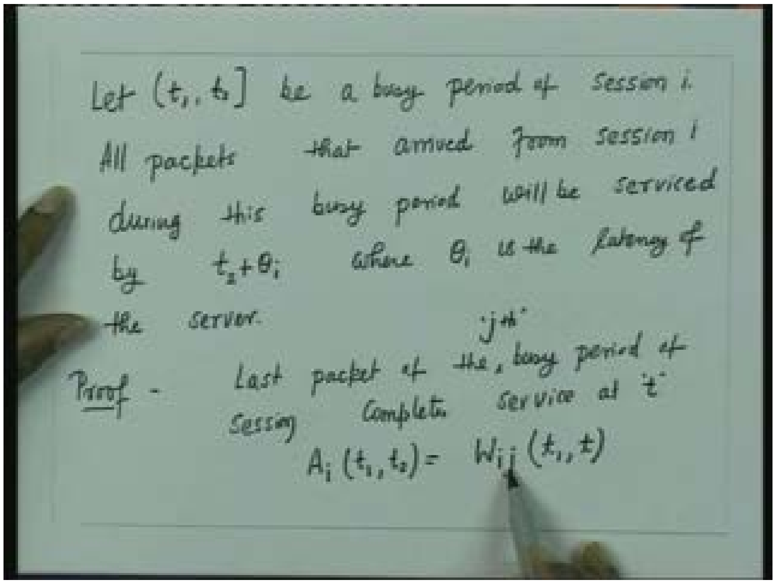


So, here is what we are saying is the minimum, the duration minimum theta $_i$ and this has a rate of rho $_i$ and this denotes W $_i$ (tau, t). So, this is what is called as the latency of the server. Now, this latency also has the significance that the minimum time that a particular session has to wait before it actually starts getting its service. So, the importance on the latency is that the minimum amount of time a particular traffic session may have to wait before it starts getting service by the server.

Now, we will actually prove that a RPS, a fluid RPS server, rate proportional server is a 0 latency server. We will also prove that a fluid GPS server, a generalized processor sharing server is also a 0 latency server. We will then prove the latencies of, then we will try to determine the latencies of various servers and from the latencies of the various servers, we will try to prove the delay bounds packet scheduling algorithms.

So, now let us prove first an important result with respect to the latency rate servers.

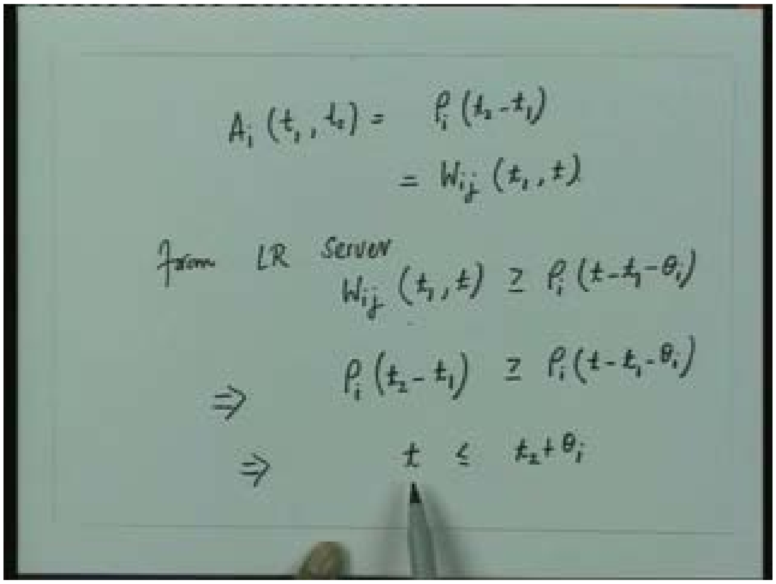
(Refer Slide Time: 43:05)



Now, this result states that let t_1 to t_2 be a busy period of session i. Then all packets that arrived all packets that arrived from session i during this period, during this busy period will be serviced by the time t_2 plus theta $_i$, where theta $_i$ is the latency of the server. So, t_1 to t_2 is the busy period of the session i and all packets which have arrived from session i during this busy period, they will be serviced by t_2 plus theta $_i$ at most, where theta i is the latency of the servers.

So, the proof is let us say that the last packet let us say that the last packet of the busy period of this session, it completes service at time t. Then, it is clear that the amount of traffic which came during t_1 to t_2 that is equal to W $_{i j}$ and let us say that this is the j'th busy period – W $_{i j}$ t_1 to t because by definition W $_{i j}$ denotes the amount of service received by session i during the j'th j'th busy period for all that traffic which arrived during this j'th busy period.

(Refer Slide Time: 45:49)



So, from the definition of the busy period, now you know that A $_i$ (t_1 to t_2) is actually equal to rho $_i$ t_2 to t_1 which therefore is equal to W $_{i\ j}$ t_1 to t_2. But from the definition of the latency rate server from the definition of the latency rate server W $_{ij}$ t_1 to t is greater than or equal to rho $_i$ (t minus t_1 minus theta) - this is by the definition which if you substitute here which also means that rho $_i$ t_2 minus t_1 rho $_i$ into (t_2 minus t_1) is greater than or equal to rho i into (t minus t_1 minus theta i) which also implies that t is less than or equal to t_2 plus theta i.

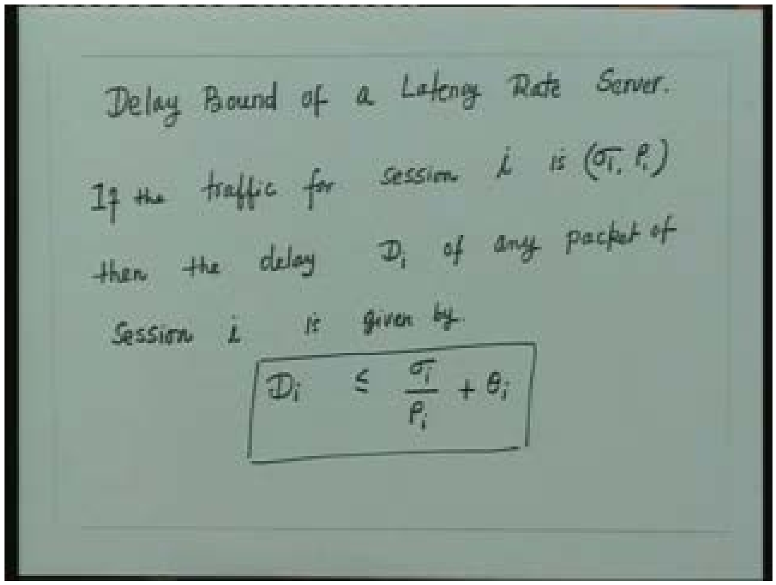
What is that t? t is the time at which the last packet completes its service. So, what we are saying is that the time at which the last packet completes its service is less than or equal to t_2 plus theta $_i$, t_2 is the time at which this particular busy period ends and theta $_i$ is the latency of the server. So, this is an important concept from which we would try to derive the delay bounds of a particular packet scheduling algorithms.

Now, what we are trying to, what we are trying to derive now is that if the input traffic to these rate proportional server which are also latency rate servers by the way; the rate proportional servers are also latency rate servers and we will shortly prove that the latency of a fluid rate proportional server is 0. So, these rate proportional servers which are latency rate servers; if the input traffic to these latency rate servers is rho sigma regulated where the allocated rate to the server rho i is now equal to or proportional to the average rate rho of that rho sigma traffic, if the two are equal; then we will try to derive the delay bounds for the particular packet scheduling algorithms.

Now, we proved a very important result for the delay bounds of a packet scheduling algorithms which are latency rate servers. Now, since most packet scheduling algorithms that we have studied like WFQ - waited fair queuing algorithms or generalized processor sharing algorithms or rate proportional servers that we have studied; they are actually latency rate servers, all these scheduling algorithms are latency rate servers. So therefore, this delay bound that we are deriving will be general enough and its result will be applicable to all the packet scheduling algorithms.

So, we will now state the result for the delay bound of a packet scheduling algorithm. So, that result we now state and that result is now, if as we say latency rate servers…

(Refer Slide Time: 49:33)



So, this is a delay bound of a latency rate server. So, we will try to prove that if the traffic in a latency rate server for session i is rho sigma regulated is rho i sorry sigma $_i$ rho $_i$, then the maximum delay D $_i$ of any packet of session i is given by D $_i$ is less than or equal to sigma $_i$ upon rho $_i$ plus the latency.

Now, this is an important result which is trying to say that if we know the latency of a particular packet scheduling algorithms, then we can determine what is the maximum delay that the packet of a particular session will suffer if being served by the particular packet scheduling algorithms. All we need to know is the latency of that packet scheduling algorithms. So, what we will do is that we will first try to prove this result and then we will try to derive the latencies of the standard scheduling algorithms like rate proportional servers that we have derived and the waited fair queuing scheduling algorithms. So, we will try to do this in our next lecture.

(Refer Slide Time: 51: 38)