Performance evaluation of computer systems

Prof. Krishna Moorthy Sivalingam

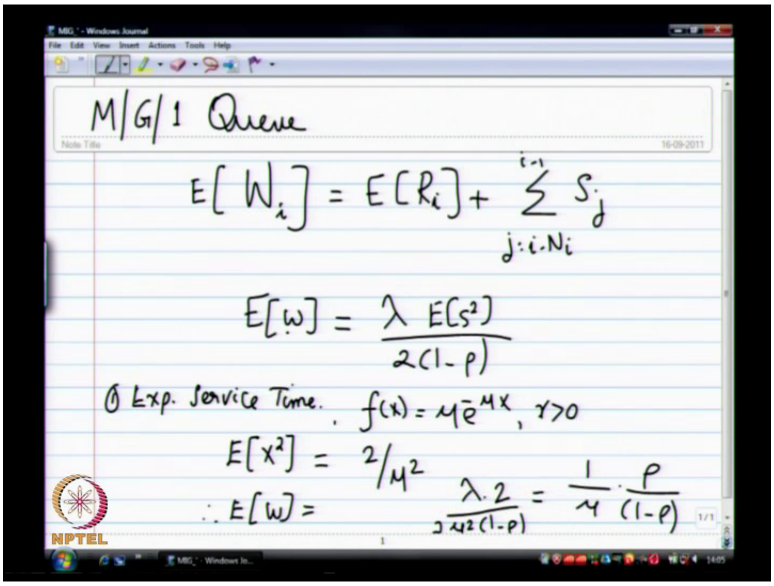
Department of Computer Science and Engineering

Indian Institute of Technology, Madras

Lecture No. # 17

Queuing theory-VI

(Refer Slide Time: 00:10)



Should we drive this waiting time right, wait for a particular process for a particular customer, wait the expected time is equal to (( )) standard of last time. Then true that any question about that we are finally right, so one mistake (( )) sometime we try to over view (( )) property. I will (( )) d of r i right, is also one over mu that technically not correct E of r i is the remaining service - remaining timing service for a given customer right. And that at any point in time expected time for a particular customer is always right, the remain the expected time is one over mu esperential, but this is this is not one over mu. We have to followed same derivation whether it is experential service time non esperential service time, E of r i simply E of r i. What we derive electron is going to apply this is cannot, this can be approximate one over mu, otherwise (( )) in corrected

(( )). So finally, we derived our E of r, sorry at least E of w to what is that?

lamda E…

Lamda E of…

s square…

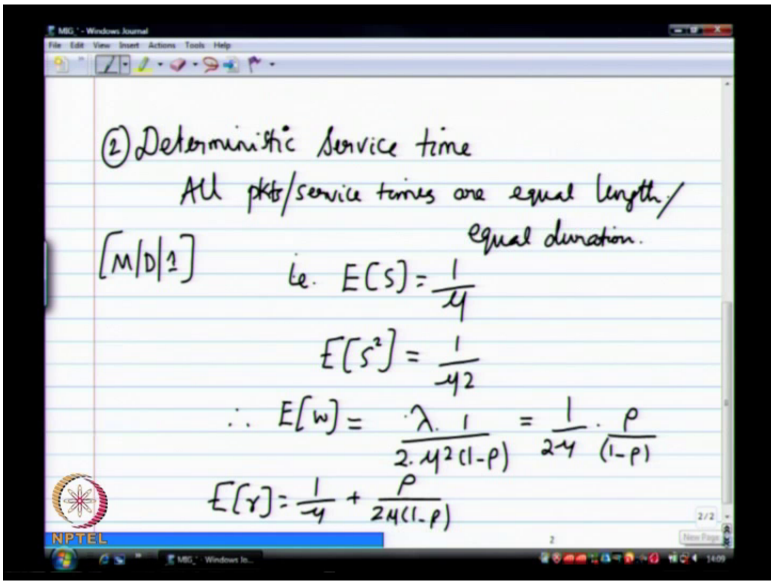
s square divided by…

(( )

This is what we had. And E of t etcetera versus E of r simply this (( )). So, this is the generic expresion for any kind of distribution, whether it is Poisson, (( )) exponential service time and so on. So, let us look at for exponential service time, what this reduces to the… Same the cases exponential variable will use again our parameter is mu, right this is following f of x equals mu E power minus (( )). So, in this case what is the f, what is the second moment 2 by mu square, to derive this.

Therefore, a plugged in then what you get right, so the E of x square E of x square is the same. So therefore E of w lambda lambda into 2 mu square, so there cut becomes one over mu into o by one minus l, so this plus one my one more I will give you f or that more

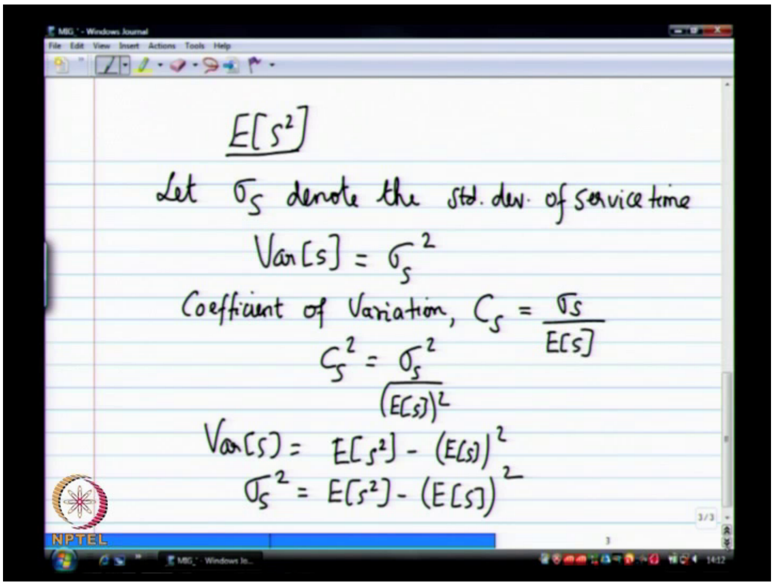
(Refer Slide Time: 04:06)



Now we come to our special case right, so the look at deterministic service time right, so how would you find the deterministic service time, lets the service time is fixed as one over mu, it is only one value for service time all pockets of explain right; all pockets or all customer service time right. So, therefore that is it, so the E of s is simply one over mu, the constant there is no variations, and what is E of s square now, E of s square is simply square right it all possible values, if there is only simply one over mu square is only one possible value is propability one. E of s square every value f of x equal x into probability of x right in this case f is simply x square right x square probability, x is the one over mu; therefore, one over mu square into only probability is one right. This is from basic derive principles, but there is nothing; therefore, a E of w is now lambda one by 2 mu square one minus four. So that now becomes one over 2 mu into so this is the basically the formula for this is m d one q right , this is so m d one acquire when ever we have fixling pockets of atm switch in all pockets of explain.

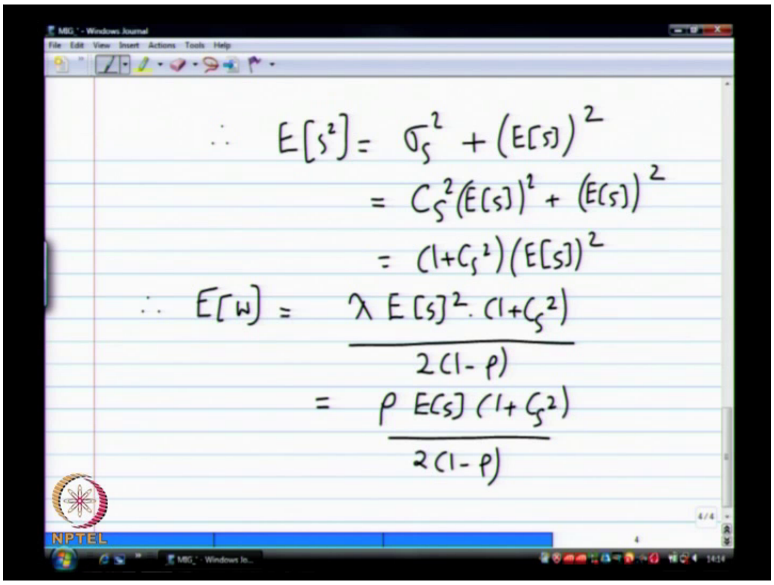
So, then what is the delay an atm switch E of w is this E of r is simply one over mu plus, so the special thing about m d one is, that is the one the distribution, this given that poison available of there only for m d one will you have the least waiting time. The waiting time is least in the case of m d one q. How do we prove that from this expression previously is the offered (( )) row by mu into one minus row in this case row by 2 b into one minus row, why cannot I go belive that are at least mathamatically right. Let us the instutional first mathamatically what you, so the we can not do the better than that, so the other wait do is that look at the co efficient of variation right. So, we can derive we can define the same thing in terms of c o v and then we can show that c s equals c o v equal to 0 only for m d one for another distribution is c o v going to be equal to 0 is, lets look at that.

(Refer Slide Time: 07:58)



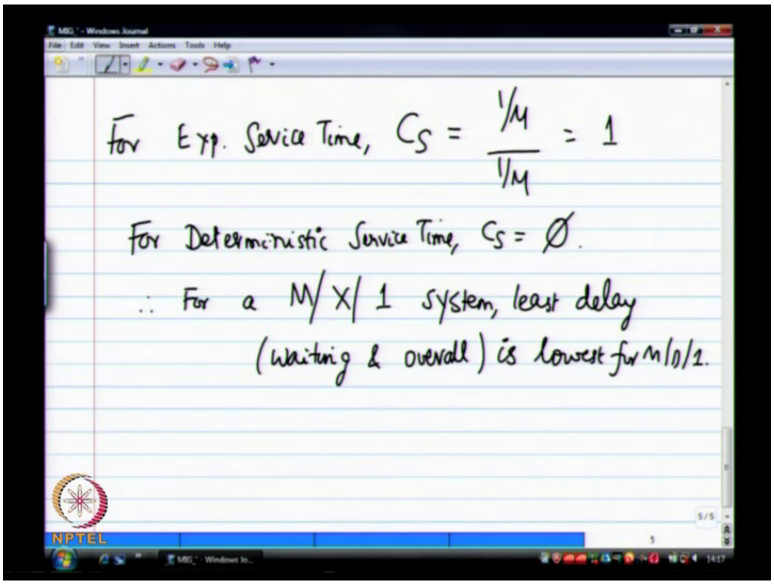
So, we look at e of s square let us to define some first let sigma s right, denote the standard deviation of service time right. Therefore, the variance s is simply sigma s square, then we can defined then this co efficient of variation, we defined by this, c s is the ratio of right mean E of s right. E of s is the mean and sigma is the standard deviation, so the ratio of the standard deviation is to mean is the co efficient of variation; therefore, c s square equal to now, we any connect our varience with is the c s square or with is the E of s square right. So, varience of s is also equal to in other way of defining the E of s square minus E of s hole square you would take this one step also this varience of s is also sigma s square right.

(Refer Slide Time: 10:32)



And sigma s square is equal to that some more right c s square definations basically c of s square into E of s square right, c of s square into E of s square all on all we get in this one plus c s square E of s square. Therefore, E of w the generic expression for derive is now simply of for plugging in are E of s square for this is time that solve, which I cannot lambda E of s what lambda in the derived service time simply the utilisation right row lambda is this row. So, this is row E of s one plus c of s square divided by 2 into one minus thus the formula in law of james book, which the expreses that in terms of the co efficient of variation questions And this of an… Now, we can see that, so questionery variation is going to be 0 right; only if the variables constant only for m d one for the deterministic service time alone c is going to be 0 right.

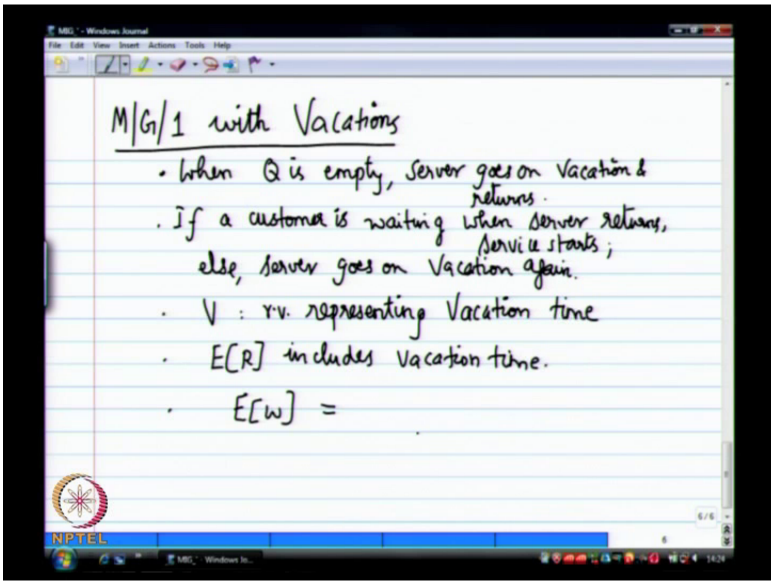
(Refer Slide Time: 13:12)



And therefore, lets going to be then least any thing is then behave greater than one greater than 0 right. So, for exponential service time - exponential service time distribution what is c of s c of s is the ratio of the variance to the mean sorry, the standard deviation it is the main right. In the case of, so in the case of poison the variance equals mean in the case of exponential, the standard deviation equals mean right; E of s square is 2 by mu square minus one by mu square. Therefore, the variation equals to one by mu square right, for exponential, variance is one by mu square again go back out of something go back in figure it out. And therefore, standard deviation this is one by mu only case of c s equals one and then of course, one plus one by 2 that formula will applied right for deterministic service time.

What is standard deviation variables have been single value 0; therefore, the waiting time is the least and the service time is the independent right, and the service time independent or waiting time that is always the pronoun the service capacity. So, therefore, the least delay over all delay not only waiting delay the least over all delay will also be seen for a given m slash x slash one distribution, the least total delay will be for the m d one system every thing is will be more than that right. Therefore and convenient thing is if the serve waiting time, in the case of exponential is twice set of the deterministic that all some terms see, if you suddenly see a two mark question; it is tells you m d one system this is a delay waiting time is two milli seconds. Tell me the delay for the other one no that simply multiply. So, that is the m g one deviation questions. So, now we look at the one variation right, one can go endlessly on the some q system m g 1 m g m 1 and so on.

(Refer Slide Time: 16:28)



But we will probability stop this m g 1 for the single q system, because I want to getting to q's of net works a net works of q's. Now, we look at the system m g 1 with vacations, and there is the application for this; its not made up changes to satisfied, so the defination of vacation. In this particular derivation it, we going to look at this when ever the system is idle right when there is no customer to be save in the q the serve goes on vacation that solve when the serve comes, there from vacation is this q is the idle it again goes on vacation.

So that is that vacation time it self is a random variable with some distribution right, so that could be deterministic exponential, what ever this right some distribution for the vacation period, but this is definition of vacation vacation is what when will sometimes goes to the bank right. There is already three people in the cube still, and guy and go up for the break right; that is not the vacation here, the idea is the cube is should be idle for serve to take a vacation. So in terms of there should be applicable is there, if you are thinking about machine right; on a machine on a machine which are on a lath what ever it is which are using it forward some jobs

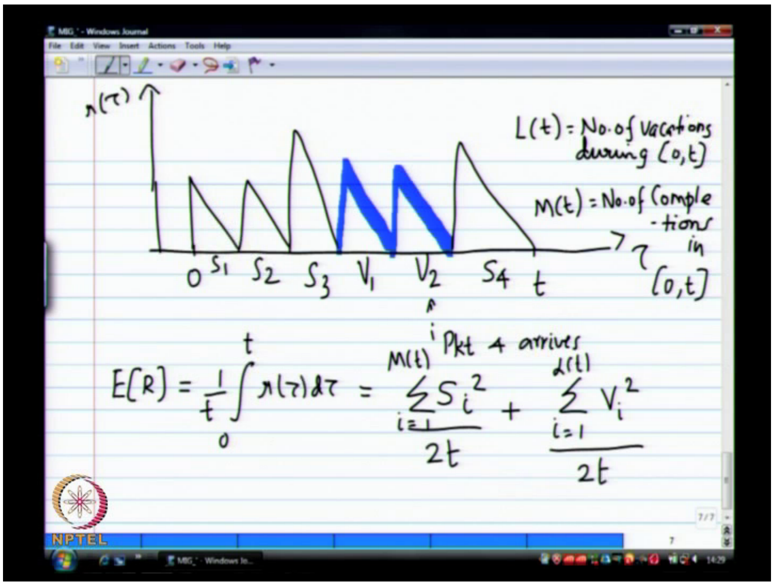
Then when ever the no jobs for done, you simply take down the machine for maintenance it is the vacation is simply means maintenance period, where will do some maintenance for some one day activities and an so on or of it is. In the case of net works, you will selled on the do some control pockets are some maintenance is as to send you want to wait untiled then q m t f, all data pockets before is send some of this control pockets right, specificly for diagnostic purposes something like that. So, that is for the vacations will be applicable, which is empty an the server is an vacation, and the packet has to wait untiled the server packet on the vacation there is the customer waiting, and end of vacation you would serve the customer. And that customer there comes to server on vacation as to wait untiled server french pack right, you do not own that server does not maintenance routined than the server why cannot be stopped.

So, there is some waiting time as per as the customer is concern even it is that concern on m d q, therefore difference between this 1 on the previous mark E of the q right, m m 1 q this is the m g 1, but it can server it can go vacation. In incase of m g of in any if it is m g man in more complicate it, because the server can go but, there m minus 1 serve is in the system right. And of course, that case if i look at shade q, then only the q is m t right; there is nobody in the no packet an all service can go on vacation single q single server than the q is idle. Server takes the vacation that solve thats the difference this will be added this will be, so residual waiting time depend upon not only the service time in pockets in front of you, but also the the the vacation time of the server of its that is the same thing.

Same derivation is applying now, so if I come to an m d q on the server is on vocation the residual waiting time residual vacation time is what will be that. So from user perspective committed idle q as good as as there is one customer in front of you, whose being serviced whether it is right theoritically there is no service, but is also waiting for the server to come back. After finishing some other, some other job or in this case of maintenance of in this case, lets right, this down when the q is empty when the q is empty the server goes on vacations and returns, if the customer is waiting when the server returns then thus the customer is serviced right. Then service starts else server takes a vacation again, and then returns of some time let s v be the random variable representing the vacation time.

So that r now also includes vacation time, so that original derivation for w that is in change right E of r includes the vacation time has from the user perspective, he does not matter whether it is server is busy with in other customer simply not available, because it is an vacation right. Pockets is simply q, it is an infinite q, so all determines in pockets will be q; if there is going to be no service stop for this particular, you just waiter and server look at starting service, and pockets right. That is a is as good as that a only thing a vacational distribution difference from the service time distribution right, that can be a different that that could be longer. For example, in look at be say on average takes ten minutes service system to maintaining the system right; that is different for this service time distribution, so this equations is till holds E of w is till equal to E of r by 1 minus row. Only thing it also compute my E of r differently, so if you look at the picture E s it is similar to what a same that, it is no different from yet another customer.

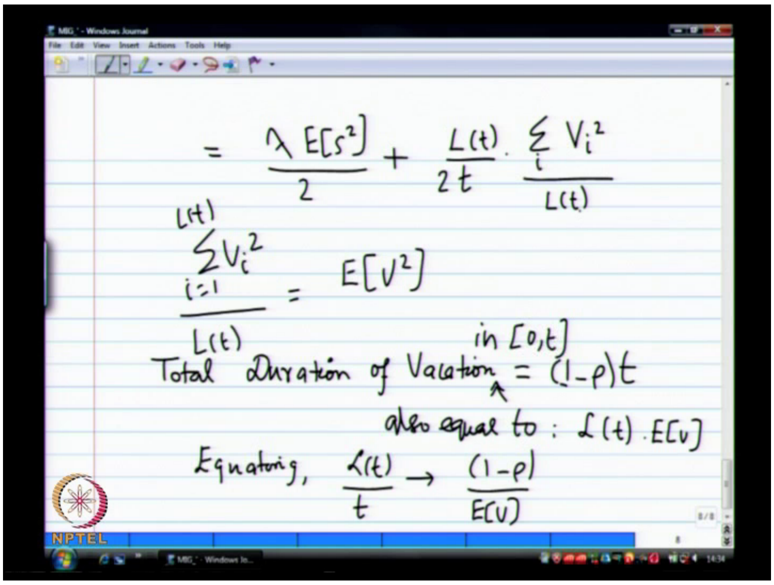
(Refer Slide Time: 23:24)



So will you the same rotation, and this is the set of r of tau this is tau and this is time 0, this is time in t, so this is the service one right, service 2; then service three. And now look at straight different colour; that is vacation, we will choose something; that is 1 v he own will draw one more v next the high light treser thing, I am not changing the then colour mearly like that. You have a vacation service right, and this is s four this is v 1 and this is v 2; it just show that in case of server comes back in this system is still empty. Then will till continue taking in vacation right, and this packet four could how arrived any were not only in between the during the second vacation it would arrived right some like hear we are packet four arrived. Now, again my expected value residual time, mean residual time is given by 1 over t 0 at t r tau d tau should we least the time m of t to denote the number of completions. Like last time will you is the term l of t is the number of vacations during this period is the number of completions.

So, what do you we have? Now, what is going to be simply last time is 1 over sigma x square right or sigma s square, in this case sigma s square plus sigma v square again this right is v 1 that with is also v 1. Therefore, this is equal to by 2 t is there I goes from 1 to, so that first of in that we alright derived last time, so that is simply lambda x i square x are lambda E of s square by 2, what did you say s sorry, there should be s. So, E of by 2 is m t by t m t by t was lambda right that solve, this is the last time we added multiply divided m of t simply say that this is equal to lambda. Then let as again shall we do that l of t by t then this should be sigma v i square all is divided by l of t is again 2, so multiply divided by l of t like with in before this was the term right v i square 2 t was there simply add in this term.

(Refer Slide Time: 27:42)



Simply say that this is equal to lambda, then let us again do that l of t by t, then this should be sigma v i square or all these divided by L of t is again 2, so multiply divided by L of t like we did before, this was the term right v i square 2 t was there, we simply add in this term. So what is v i square by L of t? E of v square right, second moment of vacation time. So, this is so sigma v i square from 1 to L of t divided by L of t is nothing but the second moment of the vacation time.

Now what is L of t by t? L of t by, L of t is the number of vacations taking in time, this is the average number of vacations with some of we do not have, in that case we got it is lambda; so that is number of service completions was equal to the arrival therefore, that was m of t by t is little bit easier; l of t by t one extra work is little bit on work is for there. So let as look at this way; what is the total time spent by the server vacational in some interval 0, what is the total time spent? We you can look at in two different way.

(( ))

What is the probability of queue being empty?

In the time span of…

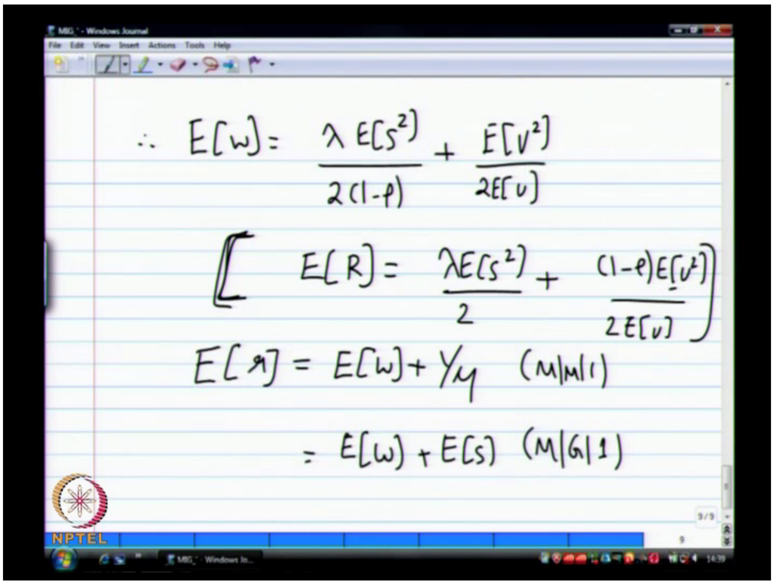
In the time span.

Therefore, that is simply; so what is the duration, total duration of vacation right it is total duration. This is given by excense, what is the server utilisation row, right?

And when rho is less than 1 that is also fraction that this server is busy, and 1 minus rho remember is p naught, fraction of the server is idle right; so p naught into t that is the total time that will expect you would take the vacation on average right; this again average so 1 minus rho into t is one way they can derive the total server was on vacation. Agreed? In other way, this also equals to it also equal to something else. The number of vacations taken into average time for a vacation right. They can look at it both ways, it is simply then given interval took for vacation each vacation for the ten minutes average therefore, the total average time taken right.

So this is respect to the interval 0 t right therefore, we get the two follows right equating so L t by t in the long run approaches; 1 minus rho by E of v right L t into E of v equal to be 1 minus rho L t by t approaches 1 minus rho divided by E of v questions on this pa when the customer comes when he is on vacation he is on the server that is true so if a when somebody, when the server is on vacation somebody arrives that is, that is approximation that we are trying to take that from. So in that case, queue is technically not empty, but at instant that the server went on vacation in the queue was empty right.

(Refer Slide Time: 32:41)



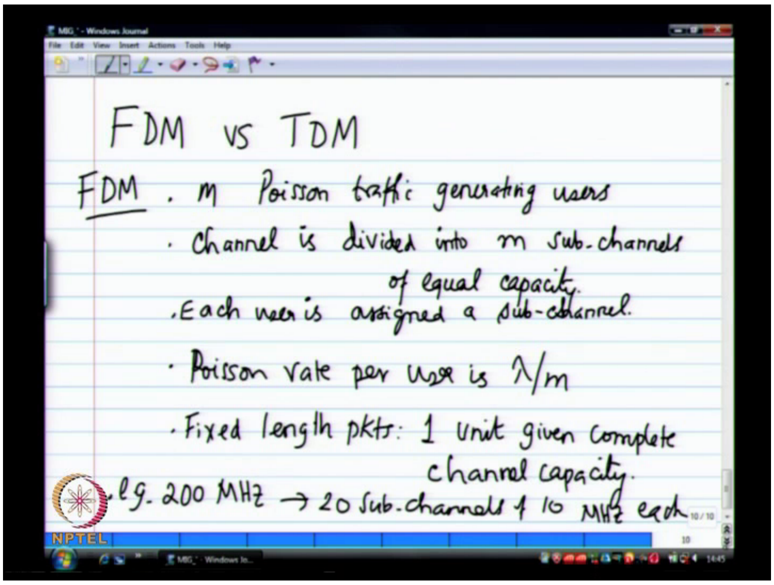
So that is the that is enough for the approximation therefore, our E of w now translates to lambda, it is an 1 minus rho on top right remember we have E of w equals I have said to gone be a E of w E of r has E of r divided 1 minus rho is over E of w right, so skip one step in the derivation if it complete our E of r E of r is the lambda E s square divided by 2 plus 1 minus rho by 2 E of v into E of v square. So 1 minus rho factors vanishes, when i should divided that when w equations, so this is our derived expression right this what the E of r translate to divide this by 1 minus rho.

Now I have m g 1 with vacations. So if I want to plug whatever I want say, if an m m 1 verifications no big deal only need to take my m m 1 delay plus the service vacations. So as long is no that first and second moments of server's vacation time and that I get the mean right not other parameters what the mean of the service time vacant at with vacation questions. So the best application are at least one easy application that we can relate to this TTM system. Now I look that I want model time deviation multiplexing; so time deviation multiplexing is defined as you have set of n users, you want share a medium single medium that is mean share TTM idea is that for every users there is a lot and therefore, what is the average delay of TTM that user will see every user as a q right and then the q when ever the user gets is of our lot send the pocket once he send the pocket when also server issue again after m minus 1 time units right so technically m m minus technically you are seeing the server vacation time is m that so will try to model the TTM system is m g 1 with vacations and see if again and get then that correctly. E of v square 1 v 0 only is a E of v square is simply, simply the vacation square that solve m square m square by 2 m; mean the second moment does not go to 0 only the standard deviation is to be 0 for deterministic.

No, no after the server comes to you is when you are right so this is E of r includes the time which is spend waiting for the server also to come; waiting for the server is also good as the server serving somebody else. So you are E of r so once my waiting time is over, all I do wait for my service that is still right in the case of m m 1 or it is E of w plus E of s M G 1. So that is why we included in service time in residual, the residual service time it includes service time vacations. So, here is one standard multiplexing problem; we see this in wireless networks, wireless networks you have a large band with right say 20 mega hertz or 200 mega hertz for available that is what you paid for in auxtion award our vacations right and you want to service a customers.

So then let us you know that number of customers fixed right only 20 customers of there you have 200 mega hertz of bandwidth, then the two choices are, two three chargers are there right how do you effectively share this; one is to break down 200 mega hertz in to 10 mega hertz chunks and give each user one 20 eth of the total bandle right. So what is that system called FDM right Frequency Division Multiplexing, if where is simply share bandwidth equally, so every body has their you break it into sub channels, m sub channels, every user has their own sub channels only thing is the capacity of the sub channels is share of the total sub channels right.

(Refer Slide Time: 38:20)



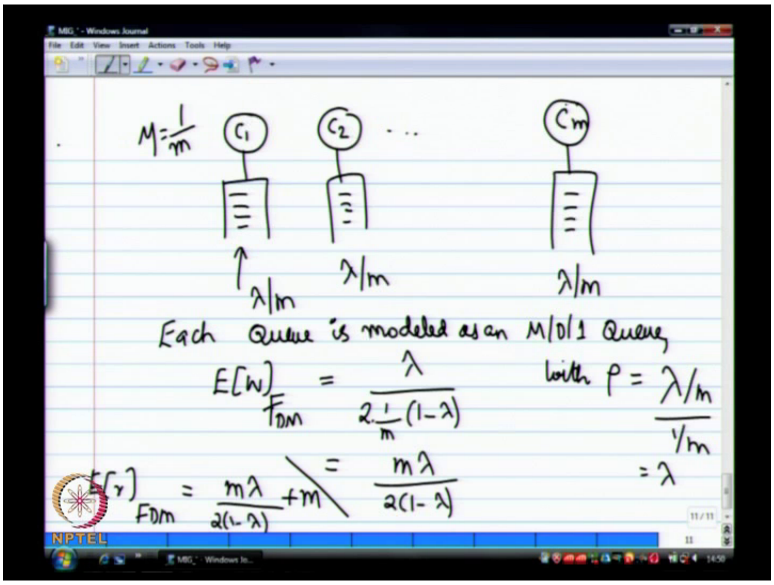
So let us tried to that comparison FDM vises TDM; so you move the channel itself as the server right you we look at the single user, each user has a queue and then they going have a corresponding server right. So if I were to so here let us define the system first. So there are m right poisson traffic generating users; and the channel band with channel is divided into when is a channel that is the entire capacity that you wait for that you want 200 mega watts you can use however you want right is divided into m sub channels of equal capacity. So that what is the disadvantage disadvantage of FDM is that if I send the pocket is go to take m times longer 200 mega hertz right only 10 mega hertz channels, my bitrate is prepotioned by band with therefore, the larger band with higher the bit rate higher the bit rate less the time going to take right. So the service time will be longer in the case of FDM, they only have 1 m of the total bandwidth right, so will take m times the time to send with the entire, the entire capacity agreed agreed.

They have 200 mega hertz this I do 1 bit per heads again 200 mega hertz per second right. I have 1 kilo bit pocket taking 1 kilo bit bite 200 mega bites is the time taken, but break it down to 20 users each gets only 10 mega hertz, then I get 10 mbps. But 10 mbps will take longer compared to 200 mbps channels thatsolve right therefore, the service time only thing is I get the channel entire my sub right my sub channel is entirely right each user is given a, so this FDM is the TDM divide goes is be along for long time. Even today our silver systems right some of the FDM, what is GSM have who is derived for WCM what is GSM do? FDM TDM or something else; so actually hybrid both FDM and TDM. FDM, TDM right you have FDM, and then you have slots with in each frequencies. But you have this notion of frequencies, and then that is time share that is the next step right initial are simply FDM or TDM. And of course, CDM I came along (( )) right that is the different thing is the hole spectrum, but you are using codes to let them act some are simultaneously, but it is differentiate but there that is your complete access to the spectrum right.

So that is the your FDM right; this is are FDM definition and then let us assume that each users the poison right per user is just make to like this simple its lambda by m packets or lambda by m units per second. Assume that all are fixed length packets right, server and server is the fixed lengh packets; and this is fixed length packet takes one unit given the complete channel capacity that is my definition of unit, fully 200 mega hertz right. So right this for this example 200 mega hertz which is broken in over to if it twenty channels of ten mega hertz each.

So this is lambda by n right packets per unit time that is over. So now, how do we model this system, what kind of queuing model will be applicable? How many servers are there, how many queues are there? There are young servers right, three sub channels is the server and there are m queue, m seperate queue and each queue on the server ofcapacity that is right. So what is the time, what is the capacity of the server it self will be 1 by m capacity of server is 1 by m.

(Refer Slide Time: 43:48)



So I have c 1, 2, c m there is the queue for a every reserve and here packets right. So the capacity here so mu minus 1 by m book is correct let us for the book it is now this steps m seperate for actually, since that fixed length packets each of this m d 1 queue, independent queue, so right.

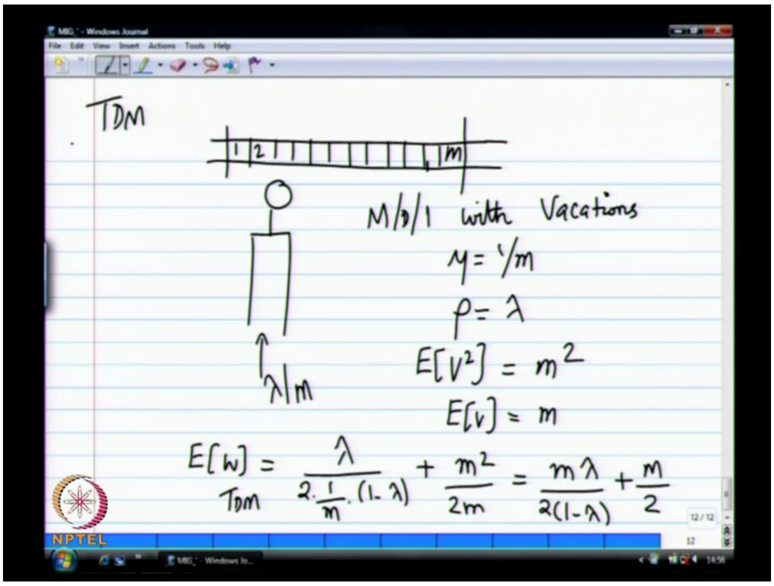
(( ))

Let us because the just you assume will total is convenience subcription.

The total rate is lambda and each user gendrading for rate lambda by m I could lambda per user and set m lambda that is also; we just assuming that all at them and generating at the same rate and the total rate is the some of all there that solve. The lambda m just for the sake of convenience that solve so the total offered rate to the system is what we some times you know, what is total capacity that the system is expected to serve; you use that lambda and that saying equalated each of the guys generatingly equal share of the total vacations.

Now this is rate each queue is model as an m d 1 queue. Therefore, the let us defind rho so rho is rho equals lambda right, lambda by m by 1 by 1 m that equals lambda each queue is lambda by m arrivals 1 by m service rate. As it takes m units, service packets therefore, per pocket each unit, each time unit serving one in packet this say… So what is that rho divided by 2 b into 1 minus rho right that is the waiting time therefore, rho is this divided by 2 and mu is 1 by m right therefore, we should be 1 by m 1 minus rho so given lambda, now that is why do this lambda right; you know lambda you know m I can simply say that the this is the expected waiting time, and what is the expected service time and total response time?... Plus service time; service time is m units. So that is how (( )).

(Refer Slide Time: 48:44)



So that was simpler, independent queues therefore, the delays are independent; there is nothing (( )). Questions? Now now our TDM systems. So the TDM system is modeled as follows. So there is m users 1 2 m; it too many length as many length. So difference so 1 to m, m is (( )). Now is it seperate queue or combined queue TDM? Each users again generating traffic at rate lambda by m right therefore, is again seperate queues on q right. So this is user i, but the channel like models likely differently, again it is single server system, again m d 1, so there is one server service time is fixed, because slots are fixed length right and it is m d 1.

But the only problem is if a packet comes to this queue right this queue is empty currently, queue is empty right, and just after 1 plus first slot is starting at time right, 1 plus epsilon the packet comes to the queue so technically queue's on the vacation right. How long it goes for vacation and goes for m time variation, because of next slot only one unit time. So that is 1. The service time, what is the service time the service time for each packet is also m time unit rate, because the guy in front of you because the one pocket front of you serviced here, when will you that the service server next of the m time units right. You get the server every m time unit one definition therefore, the service time once again m time units right therefore, your mu is again 1 by m.

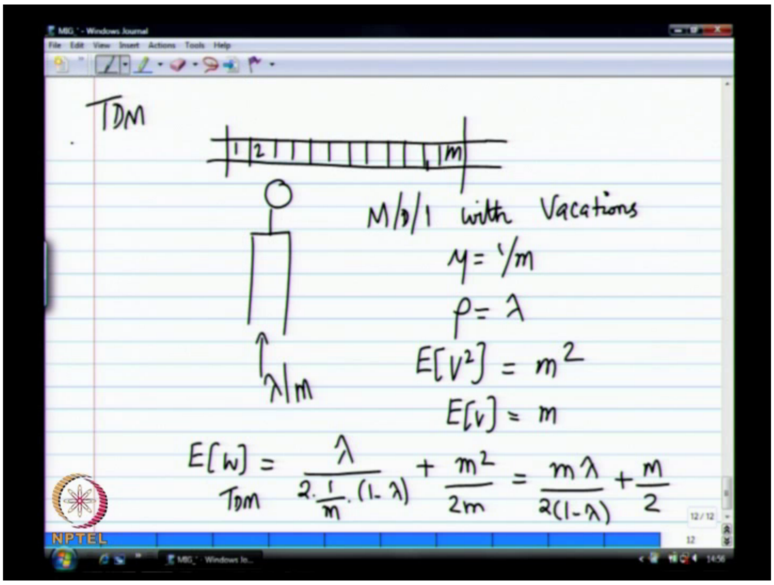
But there is the unfortunate for that even if you come to empty queue right , you want to service rate away; if you come to empty queue its server is gone vacation basically gone to the other m minus one users right. So this is an m d 1 with vacations

(( ))

That will come next, that will we do that service time the actual packet transmission time is only 1 but the time between two consigative pockets is actually m. So from this from users perspective I am getting the server once every m slots right, but when will do E of r it will that plus 1 they i do not we plus m that b plus 1 that will be a difference there E of w not see that; only in terms of waiting each pocket a head of you takes m time units it that solve of often it units.

So rho is again lambda; now the vacation rate, so E of v square vacation is again fixed time rate, it always m units return miss the server, it goes of on service system units so E of v square is m square E of v is m rate fixed. So, therefore E of w is basically rho by 2 mu into 1 minus rho, so rho is lambda, mu is again 1 by m plus so the waiting time here is more than the waiting time of they are mean waiting time higher in the case of TDM, because weather in the case of FDM you come to empty queue weather the channel is always there your server is always there serve you right that is what the difference what with you see here so this is m lambda 2 into. So when lambda is small, value of lambda rate average waiting time always going to be higher in the case of FTDM system

(Refer Slide Time: 53:52)



This is the TDM and the response time is the waiting time plus 1,1 you do not read m; in this case is only m in this case only one and one times are divide them. So that all the queue have now I can actually simplify this (( ))

1 by m to denote that to transpicable slot now, only after m slot again traspered in the thing and an also modeling vacation.

You look at the one packets perspective rate and one packet in front of me right and that pocket just now pocket of started in service right, when will you see the server next after m time units. So the time between two consecutive packets getting service at the given queue is m that is why your service rate is 1 by m. The actual transmission time is only one only 1, one slot you will done right, where is in the case of FDM you needed m slots actually send the packet, because band with is divided rate hear having 200 mega hertz channel will divided in time there is divided in frequencies; that is well. If you want we can simplify this for the 2 m by 2 into 1 minus lambda, and now which is better TDM which will always lower total response time.

So, if you simply compare one had E of w, this 1 i E of w plus m by 2; that had E of w plus m right. This had the E of w of TDM plus m by 2, but this is one; therefore, the difference is simply m by 2 minus 1. If you want the right that E of r TDM is E of r f d m minus, so therefore will always in this case of TDM …

(( ))

Even is two arrivals goes to each other unit, if you than the if you miss your slot the server goes on vacation that is what the verification right. If you arrived there is the packet in the queue and that beginning of the slot that is the alloted to you, then the server cannot go on vacation the server service the packet.

So, even if two packets come just after this slot is started the server comes and checks the queue m t is goes of, because it cannot slot transmission after this slot after that period, (( )) come before this slot, then that is fine. One of them we will will get serviced right, when the server arrives.

After that…

After that server only come back after to l time units, server will not server two packet at the time; only one packet at the time; that is the difference system only m g 1 (( )) like that. Here is only one packet service at the time, the two-reservation system also we can do analysis right, where you reserve multiple packets, where more than one packet can be transmited; that is with reservation you will try to do right. But in this case that is the…

(( )

There is the plus m by there is the see what you have, here if you have plus m (( )) m by 2, and then have this is m by 2 plus 1 is here that have plus m right. So, it is m minus do this is…

(( ))

Plus this is m by 2 plus 1 and this side, and here you have plus m right, rest of the term is common other than only these two are different right, other two terms of this time. So, it will be there is m minus 2, and then wait a minute m that that term is going to be same.

Be simplify this m by 2 into 1 minus lambda, that two terms there is m by 2 here right, this m lambda by 2 m 1 minus lambda is same for a growth right, there is m by 2 plus 1 here, and there is plus m there right. Therefore, this is equal to yeah, I want to bring this below this side. So, this shows that in the long run right as are in general even though the waiting time is slightly higher service time will make a different. So, if you m is very large then you are improve mu better, m is very small not much it; m equals 2 there is what is the difference 2 by 2 minus 1, so 0 right. Is only two channels two users it is in make it difference when you go to more than two users (( )).