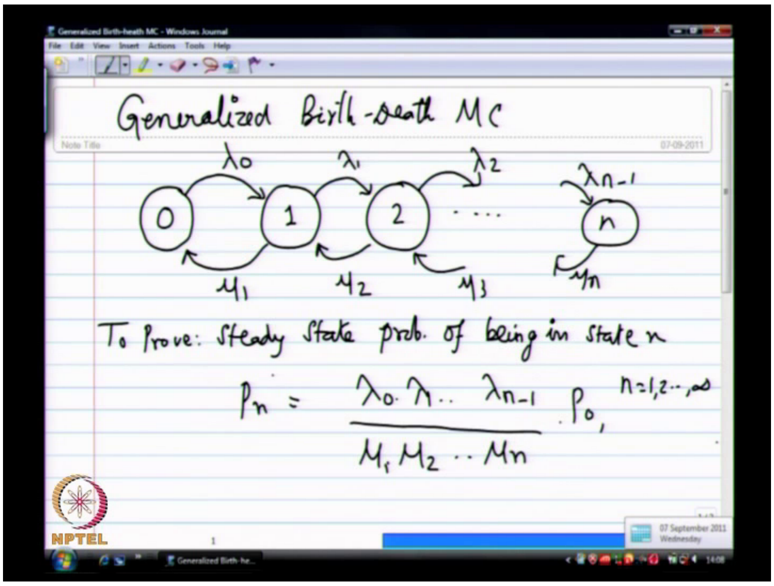
Performance Evaluation of Computer Systems
Prof. Krishna Moorthy Sivalingam
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture No. # 13
Queuing theory – II

So, last class we had finished off with of some of the properties of mm 1 queue, the performance analysis based on deriving the steady state probabilities. So, let us look at some other variance of this mm 1 would be, before that we look at a generalized version of a birth death process.

(Refer Slide Time: 00:31)



So, this is generalized. So, the characterization is that we have again a set of finite or countable set of states, and the transition rate from state 0 is given by the lambda 0, this is mu 1 and so on. So, the arrival and departure rates are state dependent.

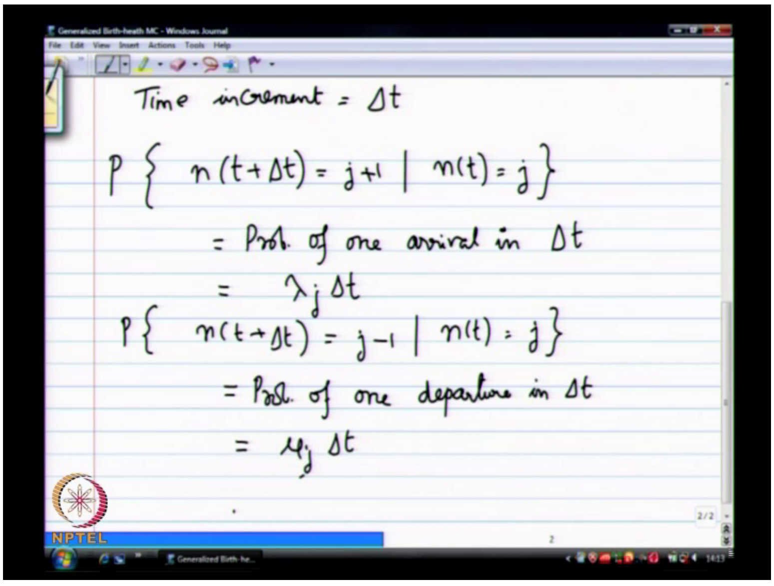
In the previous example, we saw lambda was equal to mu, all the lambda was the same and sorry not lambda equal to all the lambdas where the same, 1 equal to lambda and all the departure rate for the same n equal to, but because this particular formulation appears often that is try to derive 1 common expression from that we can always derive the subsequent for the all the different system that we look at. So, we will derive this, but the basic thing that we want to prove is, the steady state probability of being in state n which is given by P n. So, if we look at some state n so, what comes n is at the rate lambda n minus 1 and what you leave it with the rate.

So, this is simply the product of all the arrival rate from lambda naught to lambda 1, we move here the probably cannot see for n trial. So, just all the lambda 0 into lambda 1 up to lambda n minus 1, where it by the product of all the rates the departure rates mu 1 or mu 2, that is about it times P naught, where n goes from 1, 2 to infinity. So, P naught is the probability of being in state 0. So, everything is expressed as a probability of as a multiple of P naught and because all the P i some all the P n some 2, 1.

We can always solve for equation that will want so, this is in easier way of remembering what we last time. So, at simply product of all the gets birth rates divided by the product of all the death rates up to the state n. So, that is what we would write to prove and we will try to prove it slightly differently this time. So, last time we want to see how we want to see this treat, it as a discrete processing as the discrete time process so, will try to look at this is as a discrete time process.

So, suppose that at so, n of t is the variable that I am trying to, I am interested in n of t the number of customers in the system at time t and. So, what they we need to look at this. So, assume that we are looking at the increments of delta so, looking at the system in terms of t, t plus delta t, t plus 2 del t and so on, 0 delta 2 delta t and so on.

(Refer Slide Time: 04:26)



So, the basic time increment is delta t discrete yeah the process it will continuous but, what if we look modal tried as the look at the probabilities in terms of a discrete time intervals. So, this is what we would try to find out so, probability. So, given that at time t now, was in state j what is the probability that a time t plus delta t a will be in state j plus 1 rumor.

I can only go from j to j plus 1 or to j minus 1 and delta t is so small that only one of those 2 as possible or if you get of course, state in state j itself for just looking at what is the probability of that going to state j plus 1. So, what is that probability going to be? So, this the probability of exactly one arrival until, thirty and what is that derived to be will go back to our lambda j to delta t, you are assuming that there is at most one arrival or one departure delta t so, small.

So, that is dependent on the birth rate at state j likewise the probability that I go to j minus 1. So, probability of amusing the term arrival and birth interchangeably so, this is again mu j into delta t that.

delta t into probability of 1 being no departure in the other

No only 1 of those 2 is assume is going to take place you are

delta t

That there is no departure,

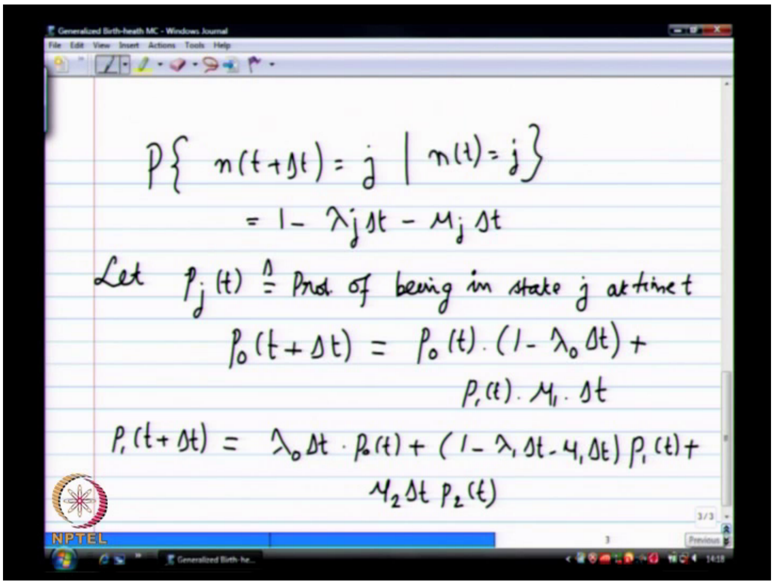
m m

only 1 can take place.

m m let see how to get the intuition so in delta t has the 2 different processor.

And what you see those, the I will come back on there just why we do not need to quantify that. So, this time the fact that there is no departure from the system.

(Refer Slide Time: 08:54)



And the probability of n of t plus delta t equals j given that. So, that is simply the probability of no departure, no arrival in the small time t. So, the one explanation is that delta t so, small that you cannot have even two arrivals or even two.

Evens cannot place only one of the two can take place so, either an arrival on departure.

I am sorry.

Probability, such that

That the which will go to 0 so this is the probability of.

Staying in the same state delta t is very small this will be is probably hard fairly large closer to one. If depending on the product of these two now, let t naught of t here, this is the probability of being in state j at time t. Now, we can define the probability of say P naught at time t plus delta t. So, what is the steady state or at least an probability of being at some that says just look as t n at time t plus delta t. I will be in state p not when two cases happen this where as you are in a question will come in. So, either I was in state t or state 0, at time t into no departure from this state it is 1 minus lambda j as there is no mu there and then P 1 of t, what is the probability the state that is P 1 but, there is a departure from that state.

So, P naught no departure or no births or P 1 and there being exactly one departure and one death. So, this is the probability that I will stay in state t. So, I was in state 0 and therefore, a state 1 state 0 and state 1, I move to state 1 in this probability, that is what you derived in the previous and likewise I will write the same thing for P 1 also, P 1 on words it is slightly different. So P 1, t plus delta t is so, first from P naught I will move in P 1 with lambda Del.

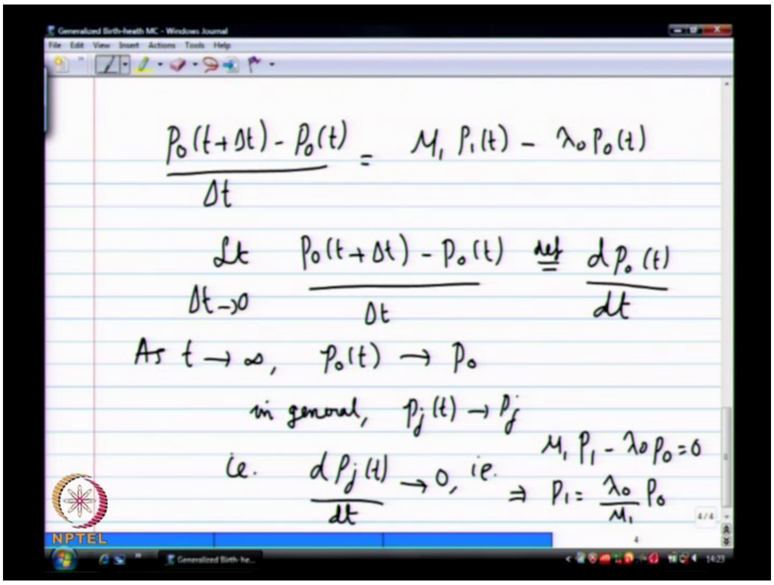
Lambda 0 delta t into P naught of t plus, I would stay in and then I would have been in state 2 therefore, the probability of coming from state 2, is this questions of have been there are routinely deriving.

The probability of being steady state will be

When t was to infinity that will be this steady state is that is what you to derive that this is.

Instantaneously, probability of being in state j at or time t plus delta t.

(Refer Slide Time: 14:02)



Now let us derive P naught. So, what is add equal to so, I expect this is tell me a you have the expressions in front of you sorry.

mu 1 into

Mew 1 into p 1 of t

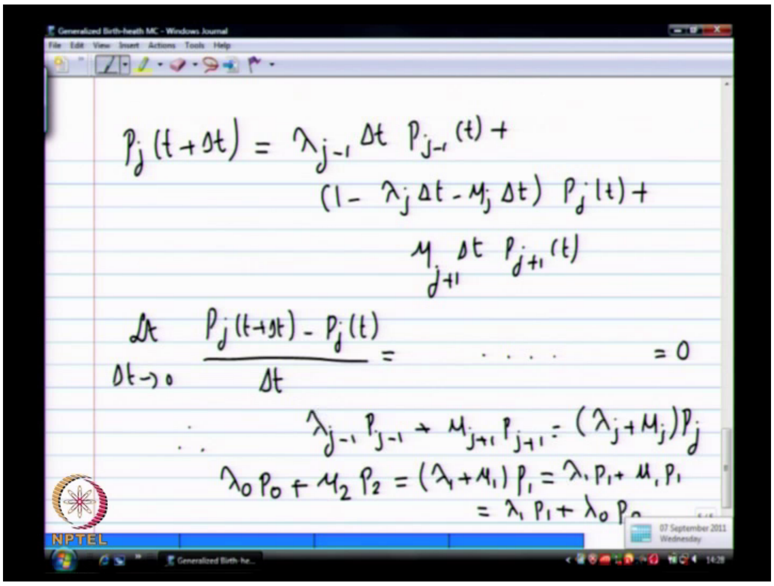
Then.

Minus p not of t lambda

Into lambda so, this is basically so, I will let you work it out. So, when this limit goes to infinity, this goes to infinity what are what you get 7 (no audio from 15:37 to 16:15) sorry.

Jumping there has the definition as t was to infinity. So, this 1 so, this is have definition of derivative. So, then so, assume that it is has t goes to infinity approaches, the steady state probability P naught and general. So, as steady state approaches your P naught of t approaches the steady state, probability P naught. So, what happens as t goes to infinity the differential of all the P j is equal to 0 therefore, so, all you are and so plugging in here will come back to our famous. So, the last time, we derived this, we looked at that queue matrix and set we can came arrive at the same this time there going back to basic definitions of change with respect to time.

(Refer Slide Time: 18:28)

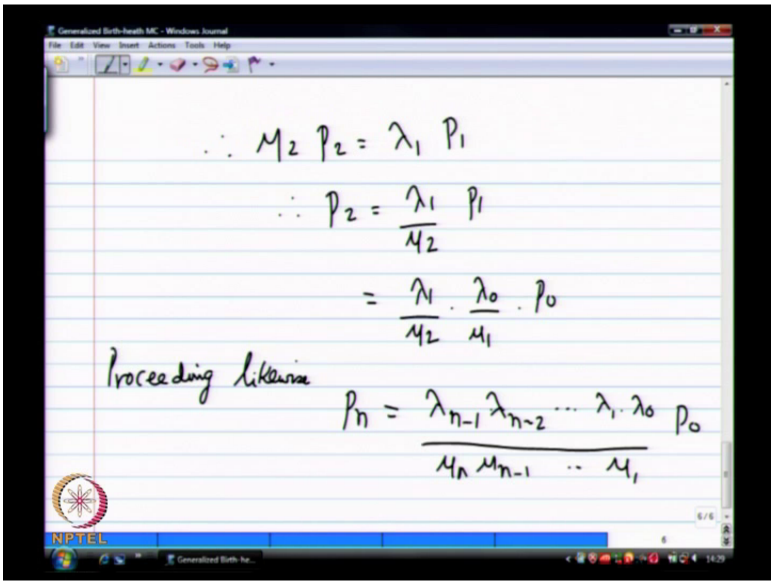


So, likewise you can derive the same thing for P j of. So, that was a special case for P naught and p 1 but, when you go to other P j is then the expression is slightly different. So, this is lambda j minus 1 delta t, P j minus 1 of t, plus 1 minus lambda j minus 1. So, this is a fairly easy to remember so, I can now again apply the same thing. So, I can derive P j of t can apply said that is 0's what will that we. So, finally, what is the balance equation we will get I am not doing that I will ask you do that. So, do the same thing that it before. So, limit equals something that equal to 0 so, dot, dot, dot, dot.

Therefore, what you we get lambda j minus 1, v j minus 1, minus plus mu j plus 1 equals again this remember, this balance equation you are coming in from the previous with rate lambda, coming from the next state with rate mu that equals to probability of staying in the same state, did you get that correct yeah. So, since we did the relationship between P naught and P 1 now, this is able to get this. So, substitute j equals 1, then what you get this we have P naught, P 1 relation, we need P 1, P 2 we did thus a last time it will go to the same thing. So it is going to be lambda not P naught.

It will be same thing 1 is lambda mu. So, once we get you that the same thing but, just Want to show another way of getting to the same thing. So, this is which is basically lambda 1, P 1 plus mu 1, P 1 but, we also know that lambda 1 or mu 1, P 1 is the same as. So, what you we get to at the end mu 2. P 2 equals lambda 1, P 1 agree.

(Refer Slide Time: 22:36)



So, this two cancel out so, proceeding likewise, we come back to our P n which is P 1 that solve. So, that is what we want to discussed. We only had it as lambda 1, lambda 0 and lambda 1 but, we consist with the equation on this page and this derivative. So, that is our basic derivation for any lambda mu for a birth death process it give any set of rates should be able to just apply this and derive this get the answer, you turn and go back to basics an there derive each time questions.

Steady state, they will go down more edition on this equation. steady state of probability of being in a particular state yes.

lambda mu also

You teel me increase an.

lambda 3 over mew to into we focus the 3 4th again more deliverable

yeah that is yeah that is

that is the that is the formula for for P n in general depends upon the birth and death rates of all the previous states that yeah so, state to state the.

So, you are thing, your saying that steady state you will just reach 1 state and stay there but, whats happening it, this is the process that is going back and forth, that is a purpose.

Of trying to do work out your that a programming assignment one, which a presume that all of you have seen, where your are trying to do essentially a monticorlesimlution of the state such a moving. So, whats happening thus the probability of going from one state to another state assume that, is like a random one where you starting one point when you go back and forth, you go back and you forth with the different states but, what you want is a long term what is the probability of being in one state your not going to be stuck in a state, that means you are that means in a obserbing state you never have an observing state this case there is in no there is no obserbing state, then there is a law the probability of being is the differentite what you try to compute probability of say 0.1 will be state 0.2 will be in state one and so on.

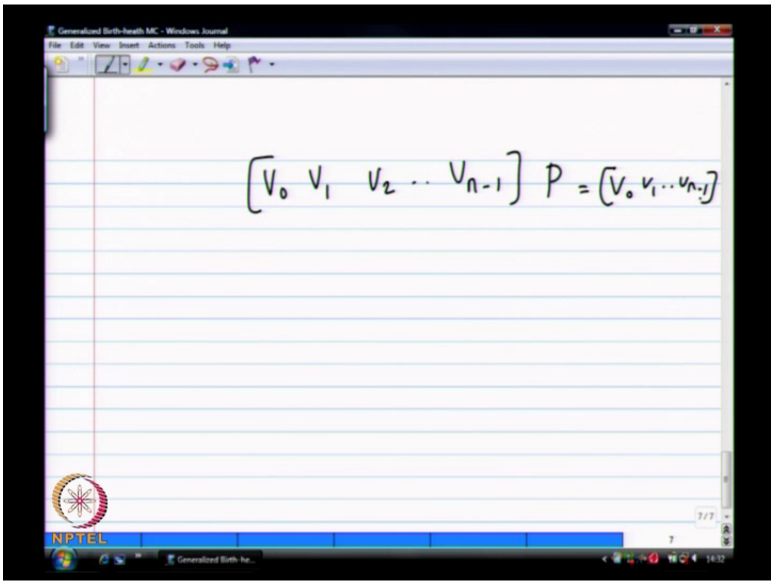
That is not mean your system your system is dynamic system in every time instant you are changing states but, if you keep running this for a million time units the two million time units then you will then you look at the actual distribution what is. So, if you run it for a million time units and that is what you will count and then trying to measure what fraction of time was I in state i, there is your steady state probability. So, you run it for a million times and then see how long was I instance how many time instance was I there an state i, that is your steady state probability not that you will say get stuck in state. So, your ultimately you will to converge to that if that limit exist for this particular system that which you looking at if it.

So that is also…

Yes same thing yes.

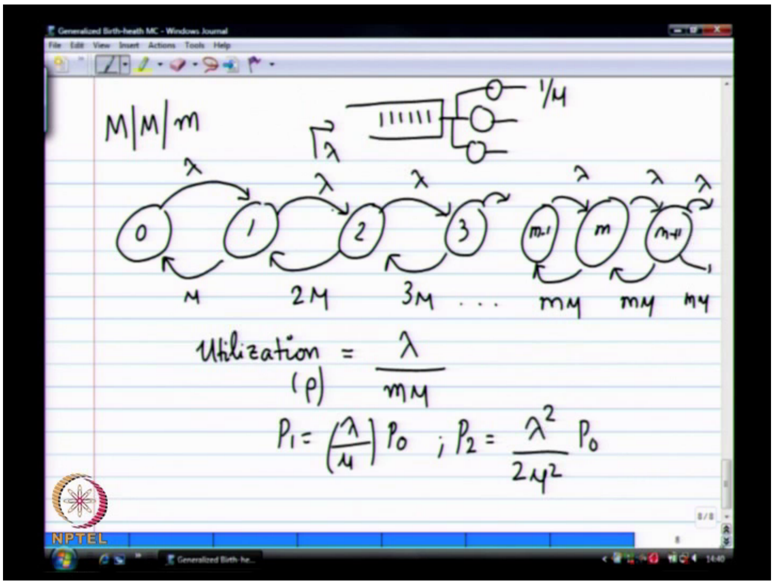
V p equals v means that i started with sub state v if a some probability distribution.

(Refer Slide Time: 26:58)



So, if you are saying that this is the current distribution of my different states I have a single step I know the transition this is your transition probability matrix. So, whatever, be this I will end up in being the steady state probability is being is the different states does not change after just one more transition, that is are that is what it say you will be different states are different points in time but, the long term distribution of the probabilities is what your trying to calculate which one should be do first mm m or mm b 1, mm m both coming up.

(Refer Slide Time: 27:38)



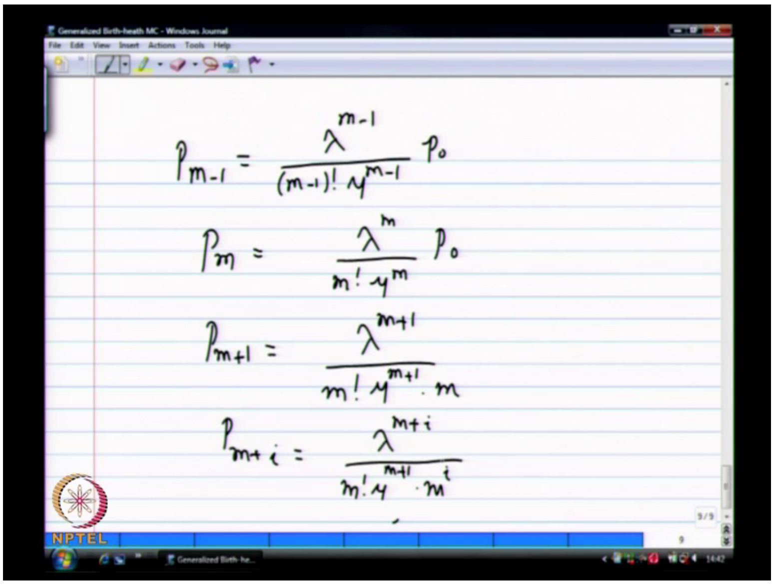
So one way of want to see both just a matter of when M M m; so, this multi server system; so, same Poisson arrival same exponential departure service time, but there are m servers in the system.

So, this what we started off last time, then you trying see what is the intuition by saying this is the these are departure this service rate rated, which I leave a state is proportional to the number of active servers. So, this are single common queue so, that the representation for this is there is 1 queue to which all the packets arrive and then there are a set of parallel servers and each server is operating at the rate 1 over mu or with service time 1 over mu and rate mu and the arrival it is lambda. So, only one common away so, compare to before we do the analysis compare this to a single mm 1, single server mm 1 queue to multi server mm 1 queue, both of them as a same capacity mu packets per second or customers per second of the e of t or the e of r response time number of customers and so on. Which will be less than the system, response time will be less in the system how about waiting time, waiting time will be less.

Less, because your multiple servers. So, when will compare this to a system which a where I have capacity of m queue, m mu rather. So, I have m servers in parallel with total capacity m over mu but, we look at a system which as single system single server with capacity m mu and then we will try to see where which one is better that will be the next analysis. So, for now let us just derive this routinely then will start applying.

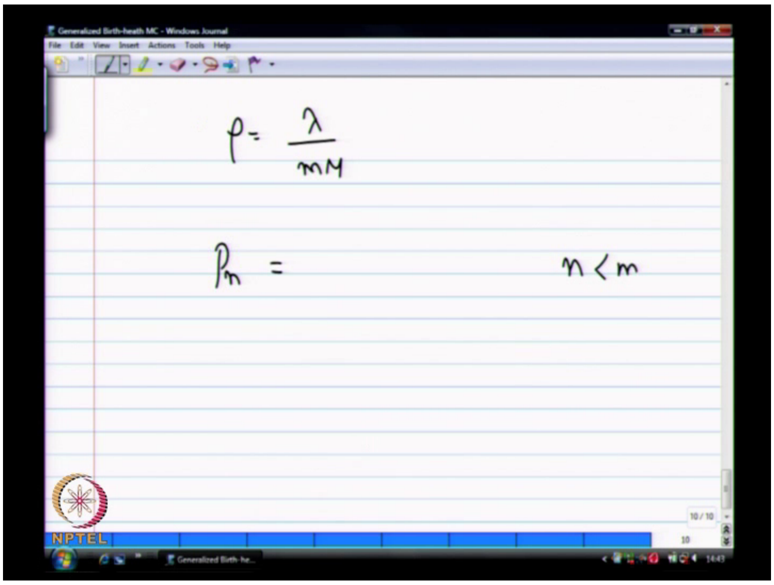
Some numbers to see the difference (no audio from 30:11 to 30:48) and so on. So, you return with the rate n mu for all states less than m, m greater than or equal to m, it is m mu. So, will define this utilization factor so, utilization is the ratio of the total arrivals to the total capacity of the system. So, what is the total arrivals that set lambda packets per second, total system capacity is m into mu so, when use this the rho is same way as before, where m equals 1, then the definition does not change. So, will rho equals lambda by m mu. So, when use this rho as we go along now that we derive that is a generic birth death process I simply out to blindly apply that here. So, there will be two cases for all cases of for P n, I am to derive P n the steady state probability less than or equal to m, allow one expression greater than m, I will have different explanation. So, if we want to again do this set of so, 1 step at a time so, P 1 is lambda by rho into P naught.

(Refer Slide Time: 32:46)



So, what is a P m minus 1 lambda by? So, there are m minus 1 terms here and then you have m minus 1, factorial mu m minus 1, p not. So, it is simply mu 1 to all the way up to m minus 1 mu, then P m is again just and just waiting that 1 by 1 so, we can come up with the generalized expression that make sense. So, of how to we generalize that m factorial mu plus 1, 1 extra m is there. So, we can generally make this m plus i, if you want is lambda m plus i, m factorial mu m plus i, P naught is always them. So, you can simply this two something more remember able something is yet remember. So, try to use lambda rho or lambda equals rho equals.

(Refer Slide Time: 34:59)



So, if we use this definition than what is P naught, P n become for n less than m (no audio from 35:05 to 35:41) so, lambda by um is m rho yeah.

in the denominator m plus 1

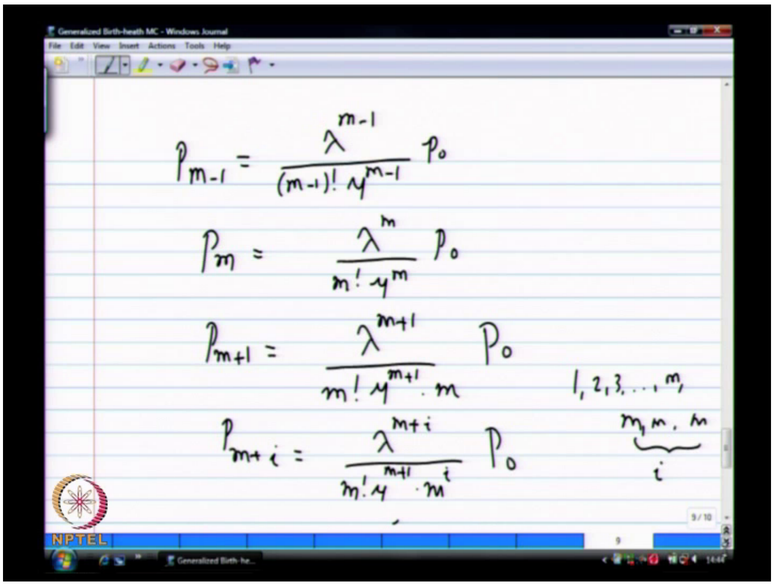
Yeah this 1 that is because is you have m [in-to] m mew into m mew into m.

Minus 1 mew and so on so it is 1 all they are put to m than becomes m factorial.

m 1 m 2 n plus 1 log

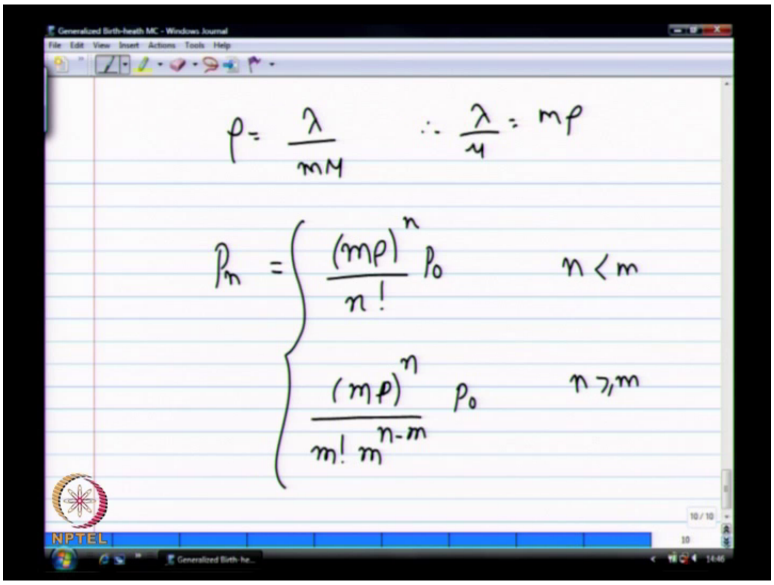
This 1 p m plus i so, this is state i so, up to m you have 1, 2, 3 up to then you have how to get simply m for the last i sates.

(Refer Slide Time: 36:22)



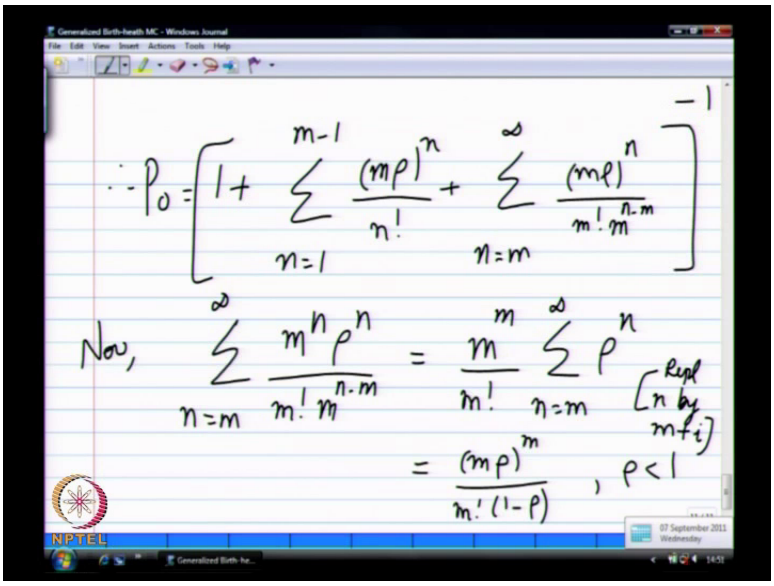
Your multiplicities terms are 1, 2, 3 all the way up to m and then its m repeating I terms.

(Refer Slide Time: 36:36)



So, how do you simply simplify that therefore, we can say lambda by mu is equal to m rho. So, will talk of m rho, everything in a m rho, and then for n greater than or equal to m. So, what is the corresponding expression for n greater than or equal to m, M rho will use m rho P naught divided by there is n factorial and then there is m i, n i is basically since, we used simply n minus 1.

(Refer Slide Time: 38:13)



Therefore, P naught is so, basically we can derive the against steady state probability thus before so, this now simply which use an mm so, this.

Corresponds to P naught this corresponds to all the m 1 to m minus 1 states and this is m rho to the power n by n factorial then of course, only the towards, can we simplify this any further.

second term can be.

Second term can be or cannot be

if rho is less than

if rho is less than 1

second term.

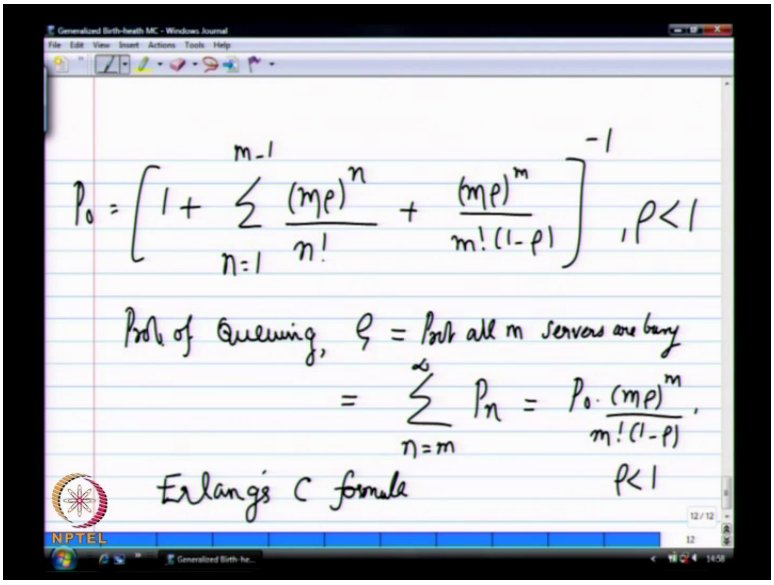
Third term

Third term, yeah mm go it and simplify by the third which one got the answers. So, we want to simplify this fellow will separate this out m m divided by m factorial m to the n minus 1, n going from. So, what are the things that are unrelated to oh no I got this all wrong than to be careful, in that n there yeah there we. So, what comes are the summation I need you to work it out and m factorial comes out, than what else what is this what is you got it so, what is this simplified to rho for n by 1 minus rho.

1 by 1 minus

1 by 1 minus will have to extract the, we can have to replace n by m plus i and then i goes to 0 infinity. So, any degree of get back to 0 to infinity so, this is basically m rho to the m divided by. So, that get that summation this is 1 over 1 minus rho we need the rho less than 1, only at that point we can simplify make sense 0 in get you have to go back in complete the derivation, if you are in a vague on that if you want I will simply say. So, replace n by m plus i, than rho m comes out than you only have the y of the and the indexes go to i equal to 0 to infinity so, I want so, that is 1 simplification.

(Refer Slide Time: 42:36)



Therefore, P naught is that will be very careful then we put n equals 0 in the previous sum no not and if rho is greater than or equal to 1, this also become unstable, because that summation in the large its keeps going to infinity and then P it is a P 0 goes to 0, than everything was to 0 therefore, this condition for the q to be stable is again that the rho should be less than 1. Now, we have this now we need to find out the expected number of customers e of n, e of n, q e of t all that is term. So, no different queuing systems, it is easier to find different metrics easily. So, in the case of mm 1, the e of r was easy to remember just 1 over mu.

Minus lambda but, in this case we will find that it is easy to remember, what is e of n, q in the case of mm 1 e of n, q was rho square by 1 minus rho, that rho square by 1 minus rho is what we will be actually. You try to get something equivalent here only think that rho square will be split into two parts, that rho has two meanging. One is utilization and it also means the probability of an arriving packet having to wait In the quueue. So, rho is probability of queuing Pq thats its called which means the server is busy. In that case in the case of mm 1 so, like wise will derive a parameter call probability of queuing here or metric called probability of queuing that will use the same way. So, instead of rho square by 1 minus rho, it will be rho into probability of queuing divided by 1 minus rho that is what e of n, q will be for an mm 1 system. So, to do derive that we need to derive the probability of queuing. So, what is the probaility in arriving packet will be queued and the book has this annoyingly looking variable at are preferred P q but, this is what it uses and yeah I am not going to spell i simply call it is P q we.

Yeah i know that but, the that time of this book was the side was the was the side refernce now it is a.

Main book so, will stick this book this is P q. So, let us derive this now so, what is the probability of queueing in an mm 1 system, when will a packet be queued.

all the servers are

If all the servers

In all the server are busy so, that means that the how do you describe that so, this is a

Probability that all m servers are busy alright, if we go to a bank the servers pretting to be busy. So, there always we looking down try to do something on this system, because do not to handle your. So, what is that in probabilistic terms in terms of our state problem steady state probabilities, the probability system being a state greater than the greater than equal to m so, P n is a probability of being in a state. So, whenever, the number of customer number of customers in the system is more than equal to m, that means all m servers of busy and therefore, you will end up being queue there are m customers. In the system already you are going to queued there are m plus 5 customers you want get queued so, that is your probability of queuing is simplify.

You get a probability of can a packet arrives

Yeah yeah arrive

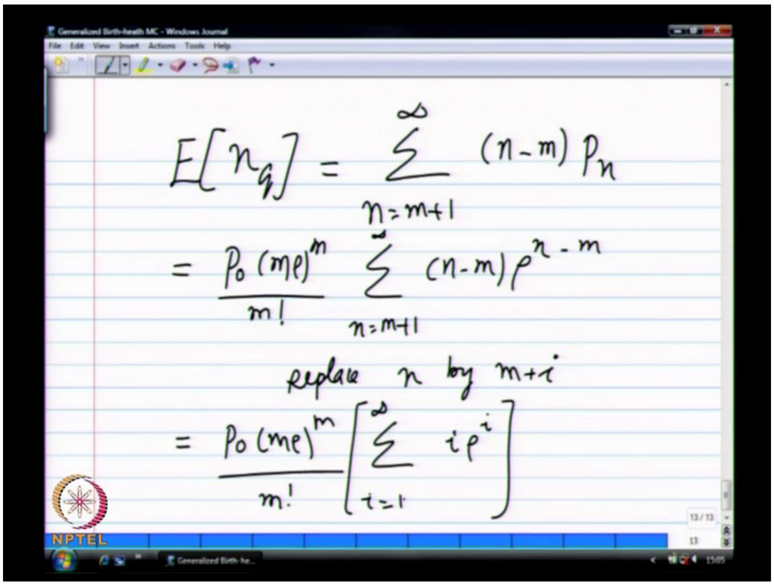
yeah can arrive at a whatever be the current state of the system and arriving packet will get queue and this whether we just now derive so, this is simply m plus 1.

Even if also, if the busy have to wait m server is the one server m server is I am so, you are coming to the system, the system state is m. So, if you come to the system at that time you are going to be you going to be queued. If your coming to this is m minus 1 you are the mth customer therefore, the customer is not there is 1 server to survive the m plus 1th customer is arriving to system, because you are starting of state 0 the state 0 you go to state 1. So, when there are already m customers in the system you are this plus 1 is always going to be bothers what for now just take that and you can question always and you can verify always, when you do have next step what are the code which will be the simulation. In verify this in discrete even simulation that this is actually for mm 1 it is fairly, it is close to what we see in theory only we going to other distribution that is have to verify, it may not be a close match.

So, this is my probability of queuing. So, this is a 1 of Erlang's C.formula called an Erlang C form. Let as be y but, Erlang was some guy was in the early 19th century or 20 th century was working on designing, this telephone networks they were just for 30 years old and we want to make sure that the customers did not face lot of block calls. In the olden days, you simply called up, the exchange inside give me a connection two and that put the phone on than call you back and there is no there is even there is direct dial is can of it is not there even 20 years ago, direct dial was not all make at book at trunk call is only used to do, because you only wait for a.

So, you want to do reduce the waiting time, you also you want reduce thus time that the customers would get lock. So, in this case you are getting queued, we look at the generalized mm m system then the customers get dropped and that is your next Erlang formula, what is the probability that in for examples in a cell phone system, this a cellular network you try to make a call all the channels are busy then you end up getting drop calls.

(Refer Slide Time: 49:45)



Now let us try to derive as a set E of n q is easy to derive in this case and then we can derive everything else from that. So, what is the number of customer queued, expected number of customers queued. In the system again, how do you post that using the steady state probabilities. Now, you look at each state look at every state where is queuing taking place from state m plus 1, on 1 the state m there is no queue there is 0 plus 1 but, we can still. So, m plus 1, m plus 2 and so on, the number of customers queued is 1, 2, 3 and so on. So, will have queuing only when you go from m plus 1 to infinity, the number of customers queued is n minus n times to n that is all.

So, will again use the definition for P n; so, what is P n, because it is greater than m plus 1, we have to use this formula rho m by m factorial and state n m plus 1, there are m customers being saved and therefore, one guy is waiting. So, the number of waiting customer is want. The n q is a number of waiting customers so, when you have state m plus i there are i customers waiting n minus, m is a number of yes n minus m is a number of customers waiting. So, if we to derive to e of n we did the P n, n into p, n but, now it is slightly different if this where mm 1 queue does not state 2 to infinity, it is always n minus 1 whatever, this one this is exception yes.

Probability the number of customers into…

Probability of mm.

Probability of number of customers there…

Probability of queuing into e of n.

Total number of customers into…

n minus n.

Number of customers.

Yeah this is looking at every state.

Expected.

This is this is expected. So, it is every state you are looking at the number of customers at all waiting for service and the other way that you are saying is I am not tried that you can tried that in the tutorial, but. So, you are looking at the delta that the whatever, that P q that we derived. So, P q is the probability that the customer will arrive to a queue that into the.

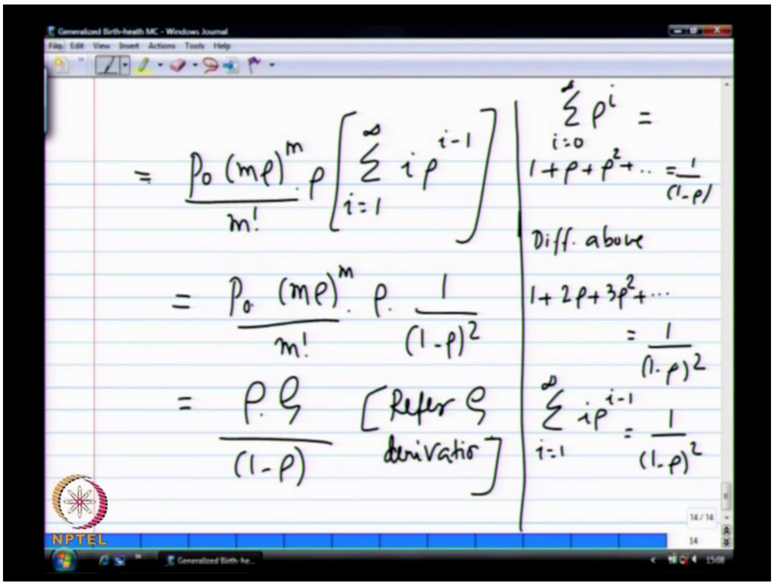
Total number of customers…

Total no no no

Number of waiting customer that is what I am trying to derive, you derive the that is it is n minus 1, it is you are doing that so you are saying expect the expected value of arriving to queue, that is already full is the that the P q value that we whether it is m plus time. So, that into that is simply the probability, but you need the number of customers actually, in the queue at given point in time at the expected number over average over all instance of time, that alternative we definitely look at, but for now if we can proceed with this will have one result. In I have must have the definition to that. We get this rho in factor out we said m to the m and then rho to the n it was. So, I took out the rho m outside if you go back to the definition. So, replace n by m plus i, and then how do your detection? So, you should try to do this way.

So replace n by m plus i than change all your variables. So, what you we get so, when n n minus 1 will become i. So, i into rho to the i and this index will now go from one when n is m plus 1 than i equals 1 and this is something that we have seen before you know and here this again should be familiar as you get. We see many of the summation as we know that there is some fact on that we see here thus you know the geometric is i rho to the i minus 1. We have done those kind of when we did there summations before so, if this is better this if this as i minus 1.

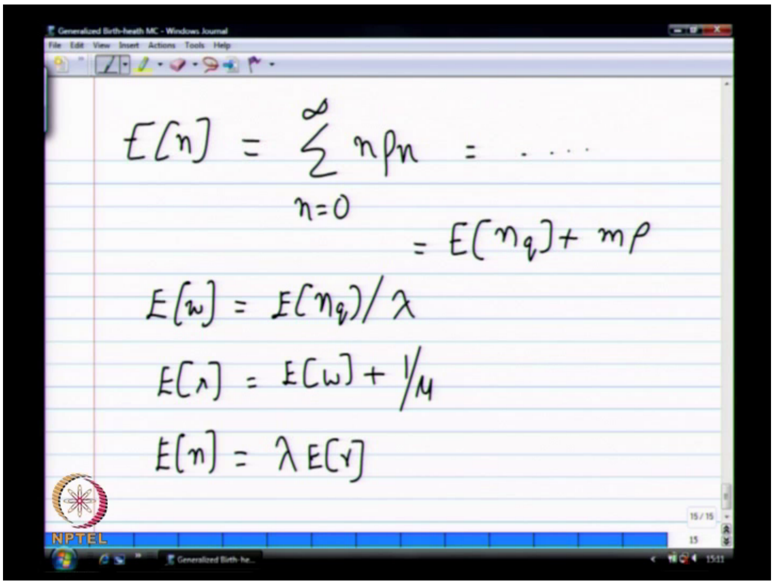
(Refer Slide Time: 55:37)



So, I am go pull out the 1 rho here. So, then I have one nice expression. So, we know that, we can derive or differentiate all of you know what I am talking about getting the derivation done this is the differentiation done was not so, this is now basically P naught. So, what is this equal to 1 over 1 minus rho square no funny business there, because sometimes this index is going from 1 and 0 meet some difference (no audio from 56:40 to 57:23) this we know this is your sigma rho I this is our there is put in notes. So, what is this now simplified to so, this is m rho, m rho all the business should look can a familiar. So, m rho m by m factorial by P naught 1 minus rho all that we saw in our P q so, this is an extra factor rho so, and rho is taken out.

My probability of queuing comes here, they divided by other one is rho this one by the rho square 1, 1 minus rho goes up to that P q other one stays here. So, that is why rho square by 1 minus rho. So, you looks familiar in the case of mm 1 that this probability of queuing is simply rho utilization itself, but here it is slightly different not for utilization little bit more complex. Now, I get e of n queue than everything else is easily derivable I know what my e of n queue is I can find e of w, then I can find e of r then I can find e of n everything. Once I get this expression, I can simply derive I can also go back in derive e of n if I want to like crazy.

(Refer Slide Time: 59:18)



So, I can also derive them n into p, n this is also one way to derive that from scratch. So, we can continue that, but that will be E of n, q plus m rho that m rho can derive. So, one thing is enough, but we know that E of w is E of n q by lambda. So, what is E of r waiting time plus service time; what is service time mm m, what is the average service time? Mm m average service time it depends up on the service capacity, which is mu say average service time is still the same one by mu by that, it is single server or multiple servers. Once you get to the server you going to be service, based on the servers capacity and then my E of n was simply, then if you go the table on 3, 1, 2. You get more detailed expressions for the variance of number of jobs in the system and variance of number of jobs in their queue and their CDF of the response time CDF of the waiting time all those things are available.

So, there are one thing that we do not derive, but we just notice that in case of a single mm 1 system the CDF of the rest, the CDF of the response time is also the response is basically, exponential; whereas, in the case of mm m the queuing time is exponential not the response time. Response time is different yeah those things are there, things for those really interested should can go through those ran as a pass time to derive those our look at some old queuing theory books that it is go for this at all moldy for now, we are looking at the first order parameters only not much of second order.