Introduction to Queueing Theory

Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 25
Nonpreemptive and Preemptive Priority Queues

Hi and hello, everyone recall that what we have been seeing in the previous lecture was this priority queuing system. Basically, we have considered the Nonpreemptive Priority systems with two classes, much like $M/M/1$ now with that there are two arrival rates; one for higher priority, one for lower priority or priority-1 and priority-2. And we assume that the first model that we considered was the equal service rate case which assumed that for both the priority classes, the service rate was mu. We call this a two priorities single-rate model, and we wrote the balance equation; and very difficult to get a complete solution for the steady-state system size probabilities. So, what we have defined is we defined this partial generating functions, and then we constructed the full generating function or the joint generating function, and by differentiating that, we obtained the mean performance measures, which are number in the system, waiting time in the system, waiting time in the queue number in the queue for priority class $1, 2$. And these are the quantities that one can obtain. So we just finally said that this boxed thing is what we are keeping it boxed because we want to make some comparison with using these things. So, that is why we kept it in this form. That is the performance measures that we could achieve in that, and then we made some observations with regard to the behaviour of these performance measures. Of course, much more we can have if you analyze more deeply, but these are some specific observations specific to the priority system. Now, let us consider model B, in which we now relax the single rate for this service of both the priorities to two rates; essentially, everything remains the same. So, the service rates of the two classes are not necessarily equal in general, and that is where this is going to be useful.

- Assume that the service rates of the two classes are not necessarily equal and μ_1 is the service rate for priority-1 customer $\&$ μ_2 for priority-2 customers. Define

$$\rho_1 = \frac{\lambda_1}{\mu_1}, \quad \rho_2 = \frac{\lambda_2}{\mu_2} \text{ and } \rho = \rho_1 + \rho_2.$$

- A similar analysis one be done for this model too.
  Now, but rates have to be appropriated whether it is μ_1 or μ_2 depending upon whether the service completion happening is essentially what we call a priority-1 customer or priority-2 customers accordingly that only the change would be mu would be replaced by either μ_1 or μ_2 depending upon the scenario. Then you have all similar analyses one can do; you can define the partial generating function, you can analyze everything,

  One gets finally

$$\boxed{L_q^{(1)} = \frac{\lambda_1(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{(1-\rho_1)}, \quad L_q^{(2)} = \frac{\lambda_2(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{(1-\rho_1)(1-\rho)}, \quad L_q = L_q^{(1)} + L_q^{(2)}.}$$

- Extra (again for interested, refer Miller (1981)): The probabilities for priority-1 customers are

$$p_{n_1} = (1-\rho)\left(\frac{\lambda_1}{\mu_1}\right)^{n_1} + \frac{\lambda_2}{\lambda_1+\mu_2-\mu_1}\left[\left(\frac{\lambda_1}{\mu_1}\right)^{n_1} - \frac{\mu_1\lambda_1^{n_1}}{(\lambda_1+\mu_2)^{(n_1+1)}}\right] \qquad (n_1 \geq 0).$$

Now, let us consider another model, model C.

- There are two customer classes with respective arrival rates λ_1 and λ_2 and with respective service rates μ_1 and μ_2.
- Service times are exponential and customers are served on an FCFS basis. There are no priorities.

So, it is merely that there are two classes of customers who are being merged into a single queue, but when they are taken up for service, they are served at different rates, μ_1 and μ_2. But the picking of customers for service is on the basis of FCFS, so there is no priority. So, this is the simple two-class model. Remember that whenever you say multiple classes, it does not always mean priority customers; multiple classes could still follow an FCFS business as given here.

- This two-class FCFS model can be viewed as single-class $M/H_2/1$ queue, where customers are grouped into a single arrival stream and the service distribution is a mixture of two exponential distribution.

But each customer is being picked one of these distributions depending upon the class, but here once you group them, in this, as such, you do not have the distinction between the arrival units, but in this, there is. So, there is some difference, but this can be viewed in this manner as a mixture of two exponential distributions is what the service time and a single class arrival.

For this model, we are now using this because $M/H_2/1$ we have not dealt with in any way, but this could be dealt with after having seen the $M/G/1$ model, which we will do later. If we have the knowledge of $M/G/1$, then this is very easy to see how we are obtaining this performance measures this average performance measures. But right now, you take this as if it is something for granted, but you can obtain even otherwise with the current material, but that is a little tedious process, but if you have knowledge about the analysis of the $M/G/1$ model, then this is very easy to obtain these qualities very very easily one can obtain. So, you can see that later, but if you want to analyze again, you can go back, and then you can start from the basics, and you can get the analysis, but that is a very tedious job anyway. Now, what we are getting here is this as the number of priority-1 customers, the mean number of priority-2 customers, both of them in the queue, and this is the total number.

$$\boxed{L_q^{(1)} = \frac{\lambda_1(\frac{\rho_1}{\mu_1}+\frac{\rho_2}{\mu_2})}{1-\rho}, \quad L_q^{(2)} = \frac{\lambda_2(\frac{\rho_1}{\mu_1}+\frac{\rho_2}{\mu_2})}{1-\rho}, \quad L_q = \frac{\lambda(\frac{\rho_1}{\mu_1}+\frac{\rho_2}{\mu_2})}{1-\rho}.}$$

Note: The L_q above is always greater than that of the standard $M/M/1$ model with mean service time equal to the weighted average of the respective means, namely, $\frac{1}{\mu} = \frac{(\frac{\lambda_1}{\lambda})}{\mu_1} + \frac{(\frac{\lambda_2}{\lambda})}{\mu_2}$ (due to the higher variability in the service times).

So, this is the two-class model. So, we have considered now A, B, C three models one rate two priorities one rate which is model A, two priorities two rates which is model B; now this is two class no priority two-class model this is two class FCFS. Now, let us make some kind of comparisons and some analysis we can do with respect to these parts. First, let us consider the last two models, (B), and (C).

- (B) Vs. (C) : Priority queues (unequal service rates) with the nonpriority queue.
  ► Imposition of priorities decreases the mean number of priority-1 customers $(L_q^{(1)})$ and increases the mean number of priority-2 customers $(L_q^{(2)})$. This result is quite intuitive.

  Well, this is nothing unusual this is also quite what you would expect because in the one case, you are treating all of them are equal; in the other case, you are giving some priority though it is a non-preemptive model, at least if there is anything any lower priority customers are there he jumps the queue. Some customers, at least he will jump. So, he gets service ahead. So, the number of priority-1 customers should obviously, decrease, and the mean number of priority customers will also increase for the lower priority customers; this is obvious. But then why the quantification? The quantification is always essential for proving that, now you have the expression, you can show it with expression this is true. The additional benefit is that if you want to measure how much it is increasing or how much it is decreasing, then you would need the quantification just that intuitive idea is not sufficient; this happens with any mathematical model, for that matter. So, it also happens here also. So, you could think about it in that way. So, this is about priority-1 and priority-2.

  Now, how does this overall system performance compares? Between these two. For which then we can pick say L q again you can do analysis with W_q, L or anything because they are all equivalent if you know one, you can get the remaining three as well. So, we can do with one.

- Comparison of average overall number in the queue L_q between the two models $(B \ \& \ C)$.
  ► They (and W_q's) differ by a factor $\dfrac{\lambda - \lambda_1\rho}{\lambda - \lambda\rho_1}$ as evident from the below mentioned equations:

$$L_q \ of \ B = \left(\frac{\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2}}{1-\rho} \right) \frac{\lambda - \lambda_1\rho}{1-\rho_1}$$

$$L_q \ of \ C = \left(\frac{\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2}}{1-\rho} \right) \lambda.$$

  ► Thus there are fewer customers waiting in the priority queue of B when

$$\frac{\lambda - \lambda_1\rho}{\lambda - \lambda\rho_1} < 1 \iff \lambda_1\rho > \lambda\rho_1 \iff \lambda_1 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \right) > (\lambda_1 + \lambda_2)\frac{\lambda_1}{\mu_1} \iff \mu_2 < \mu_1.$$

So, when does this happen? $\mu_2 < \mu_1$; that is the idea finding that you are making in this particular case. So, the fewer there will be fewer customers in the whole system when there is priority provided this is satisfied, which means that the priority customers have a higher service rate that is what $\mu_2 < \mu_1$ means.

- This gives rise to an optimal design rule called "the shortest processing time (SPT) rule".

So, what we are seeing here from this comparison, just this comparison of L_q between these two models that $L_q\ of\ B < L_q\ of\ C$ if $\mu_2 < \mu_1$; that is what we are saying.

► The priority queue results in less overall waiting (compared with corresponding FCFS) when the first-priority customers have a faster service rate (or shorter service times).
► Conversely, the priority queue results in more overall waiting when the priority customers have longer service rates.

So, if you want, what is your objective then? When you are designing a queuing system, you will have a certain objective in mind. If your objective is the overall reduction of the total number of customers who are waiting, or equivalently the overall mean delay, whatever equivalent words you can use, then what should you do?

► You should give priority to the group of class of customers that has a faster service rate.

So, how will you achieve it? So, that is what is this rule shortest processing time rule. So, the shortest processing time means a faster service rate. So, the basis for this is basically you can establish by this analysis that you have done so far.

Some things may be intuitive, some things may not be intuitive that much, but still, you can do it by analysis that is typical and here also that holds. Yeah, things may be intuitive, or it does not matter, but you can show it in this way. You see precisely when that can happen in this case. So, this is what you can do if you compare the two-rate priority with the two-rate no priority model. Now, let us compare the priority-two rate with the priority-one rate model.

- (B) Vs. (A) : Priority $two - rate$ model of B with the priority $one - rate$ model of A.
  ► We must make some choice for the value of the single rate μ. For example, choose μ so that

$$\frac{1}{\mu} = \left(\frac{\lambda_1}{\lambda}\right)\frac{1}{\mu_1} + \left(\frac{\lambda_2}{\lambda}\right)\frac{1}{\mu_2}.$$

  or one can choose μ to lie somewhere between μ_1 & μ_2.

  You may not make much justification if you go beyond this rate, but at least it will be this must be between these two that is clear when you make this one rate comparison. Now, one can show, so it can be shown, but of course, these are all some of the exercise problems in your book, but otherwise also one can show.

  ► If $\mu = max\{\mu_1, \mu_2\}$ then $L_q^{(1)}, L_q^{(2)}, and\ L_q$ of A are less than that of B.
  ► If $\mu = min\{\mu_1, \mu_2\}$ then the reverse happens.
  ► If μ is strictly between μ_1 & μ_2 then the comparison would depend on the parameter values.

Of course, on the one extreme, say if it is maximum, then the one-rate model is better; if it is minimum, then the two-rate model is better, but what if in between? Obviously, somewhere it has to make a transition from one model to

the other model to show that this is better, but that depends really on the parameter values. So, this you can do some such things. And this is the summary of these kinds of comparisons.

- $(A^{'})$: $M/M/1$ with $\lambda = \lambda_1 + \lambda_2$, $\frac{1}{\mu} = \frac{\lambda_1}{\lambda}\frac{1}{\mu_1} + \frac{\lambda_2}{\lambda}\frac{1}{\mu_2}$
- (A): Two priorities, one service rate with $\mu \in (\mu_1, \mu_2)$
- (B): Two priorities, two service rates
- (C): No priority, two service rates

Versus	$(A^{'})$	(A)	(B)
(A)	$L_q : (A) = (A^{'})$		
(B)	$L_q : (B) < (A^{'})$ iff $\mu_1 > \mu_2$	depends on parameters	
(C)	$L_q : (C) \geq (A^{'})$	N/A	$L_q^{(1)} : (B) < (C)$ $L_q^{(2)} : (C) < (B)$ $L_q : (B) < (C)$ iff $\mu_1 > \mu_2$

So, this is the summary of the kind of comparisons that one can make with respect to these models; these are the priority models that we have. So, priority models priority systems, and analysis wise is difficult, but one can obtain certain mean performance measures, and from that mean performance measures, you can for managerial aspects for managerial purposes you can draw certain conclusions, and that is otherwise what the use of studying such model is. So, you can draw certain conclusions and the by making a comparison of these measures under different scenarios and so on. To draw some insights, see this performance analysis or analysis; it is basically it is up to, in some sense, the analyst how much he puts his brain behind and then how much he goes deeper into understanding the system and how much he can unearth the insights from the analysis it is up to him. So, the deeper and the closer you look at it, you will get more insights typical of any model, any mathematical model per se. But depending upon what your objective is, you will target such kind of analysis. So, in this case, suppose if you have it in mind, something like you have two classes there, but you do not know you are not assigned. Suppose if I assign priority, then what will happen? Whether I am going to reduce my overall number or not? Now, you can see, for example, if you go to a supermarket, customers with less than five items are allowed in one particular queue. And customers with their bags full of cart, full of items like they will be they are in the other ones here probably they are not allowed because its specifically for these kinds of people. The reason behind this is they are expected to take less service time, shorter service time. So, you see, the designing in the design rule that we talked about in that "SPT" rule is essentially applied there. Because, but justification now like today you ask like what is justification for that, you have to do this analysis and show that yes this is the justification.

So, that is what is pretty much happening there. So, to do that, to justify that, you need to have this in your hand to say, and that is precisely what you have in your hand. Fine, one can go on. Now, there could be some extensions that one can think of with respect to non-preemptive, of course; there are many directions. So, we have dealt with only two classes, but more classes; obviously, one can handle theoretically in principle; principle remains the same, but we have seen even in the case of two priority obtaining the steady-state probabilities is very, very difficult.

- Now, for more than two priorities, it is almost impossible in a way; it is near impossible because of their multidimensional nature, and you should define the multidimensional generating function and what not like things become complex here.
- If you are interested only in the average performance measures ($L_q \; and \; W_q$), you have a direct approach expected value procedure that can be used to obtain the mean value measures.

  But this is not in our scheme, so we are not going into that, but a similar thing we are going to do it for an $M/G/1$ model later. So, at that time same things could be applied similar to these kinds of models as well. The principle will one has to think, and then we have to adapt to the situation.
- But of course, the textbook has this thing also, so one can have an idea and further extension could be like, for example, continuous priority classes. So, this is essentially in computer and communication system this is important, these are based on the actual service times.

  Here we know exactly how much is the service requirement. Then that is what we say assumed to be known; this is if you know the actual service time now; this leads to what is called the "Shortest Job First" SJF rule. Now, one can analyze using a similar thing and then the multi-server model also, one can think with respect to that. Again direct expected value procedure is what one might be interested in doing in this particular case. Some non-preemptive systems and whatever we did over that that gives the flavour of how one can analyze it on a priority system which is non-preemptive in nature, and one can think about extensions on that.

Now, the other system, we said there are two kinds of priorities; one is non-preemptive and preemptive. So, non-preemptive we have seen. Now, let us brief about what could be done for a system with preemptive priorities. Now, in this case, what do we have? We take the same model that we have considered; it is a two-priority model, higher priority and lower priority. There are two rates, μ_1 and μ_2 same Markovian assumptions; everything is there, meaning that the arrival process is the Poisson process, service times are exponential distribution, and so on. The difference only now is that there is preemption now.

The no priority non-preemptive only we have seen earlier, but now with preemption is what now we are looking at here. So, what happens now? Suppose that a higher priority customer is undergoing a service, a higher priority comes customer comes he queues, a lower priority customer comes he also gets into the queue. Suppose a lower priority customer is being served currently in the service, then a lower priority customer comes he gets into the queue, but a higher priority customer comes what he will do is that he will interrupt the ongoing service of the lower priority customer and he will get into the service, he will get into. The fact that earlier lower priority customer is in service means that there are no higher priority customers waiting in the queue; that is clear. So, whenever an arrival happens, if it is a higher priority, then he will interrupt the service of the lower priority, and he will get into the service, and his service will start. So, lower priority units that are ejected from the service cannot reenter the service until the system is free of all higher priority units.

Now, what will happen to the ejected units? They must start over losing all the partial work already completed; this is preemptive nonresume, or the ejected units resume service from the point of interruption; this is a preemptive resume regime. So, you have a preemptive resume and preempt non-resume; there are two cases here, but here because the

service times are exponential, there is no distinction between these things; whenever the service time is exponential, you know that if you want to compute what is the remaining service time of this customer, it is still exponential with the same parameter.

So, whenever it is interrupted from that point, if you look at it, what is going to be the remaining service time that needs to be completed, it is still exponential. So, whether you start all over or you start from there in the exponential case is one and the same. So, there is no difference, but this difference will come into play in the case of service time distributions which are not exponential that much we have to remember.

- The state space for this preemptive priority two-class system is $S = \{(m, n) : m, n \geq 0\}$ with their steady state system size probability given by

$$p_{mn} = P\{m \text{ units of priority-1 } \& \ n \text{ units of priority-2 in the system in steady-state}\}$$

  ♦ (λ_1, μ_1) and (λ_2, μ_2) are the corresponding arrival and service rates (of the two classes).
  ♦ $\lambda = \lambda_1 + \lambda_2, \qquad \rho = \dfrac{\lambda_1}{\mu_1} + \dfrac{\lambda_2}{\mu_2} < 1 \qquad$ (assume)

- As earlier, one can proceed to write the balance equations (there will be $2^2 = 4$ sets of equations) as

$$\begin{aligned}
\lambda p_{00} &= \mu_1 p_{10} + \mu_2 p_{01} \\
(\lambda + \mu_1) p_{m0} &= \lambda_1 p_{m-1,0} + \mu_1 p_{m+1,0}, \quad m \geq 1 \\
(\lambda + \mu_2) p_{0n} &= \mu_1 p_{1,n} + \lambda_2 p_{0,n-1} + \mu_2 p_{0,n+1}, \quad n \geq 1 \\
(\lambda + \mu_1) p_{mn} &= \lambda_1 p_{m-1,n} + \lambda_2 p_{m,n-1} + \mu_1 p_{m+1,n}, \quad m, n \geq 1
\end{aligned}$$

- After deriving various partial generating functions, one can obtain the moments of the number of units in the system. This gives us

$$\begin{aligned}
L^{(1)} &= \frac{\rho_1}{1 - \rho_1} \\
L^{(2)} &= \frac{\rho_2 - \rho_1\rho_2 + \rho_1\rho_2(\frac{\mu_2}{\mu_1})}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}
\end{aligned}$$

  look at the expressions for $L^{(1)}$. Isn't it look like what you would see in a typical $M/M/1$ queue? And if you look at with preemption, that priority-1 customer will behave as if it is a $M/M/1$ queue for him. And whenever the system is free, he is free to serve the lower priority customer; thus and whenever the system is busy, he has to serve the priority-1 customer. Here busy means that in the $M/M/1$ context, which we are looking at it.

  Here, $L^{(i)}$ is the average number of class-i customers in the system in steady state.

- Class-1 customers are not affected by the presence of the class-2 customers, this is true with the preemptive case.

  But whereas, in a non-preemptive case, the class-1 customer would also be affected by the presence of a class-2 customer, whereas, in a preemptive priority queue, the class-1 customer or the higher priority customer is not affected by the presence of lower priority customers.

- Thus the class-1 customers are effectively operating as if they were in an $M/M/1$.

So, you have a simple $M/M/1$ queue; now you are putting one more class of customers whenever this server is free, like you decide to serve for someone else stream that could be of lower priority. Now, whenever your customer comes, you will stop him in between, and then you start doing this; that is the model that you can think of in this particular situation. So, this is preemptive. Again you see that more than two again, things become complex, but again one can get in some sense the expected value measures in a reasonable analysis level. So, these are all about priority queuing systems; of course, this can be studied in a general framework, but we have studied it in the Markovian framework. So, you know more about how a general Markovian system can be handled in that sense. So, this is what we have done. So, we have mentioned, of course, we have not derived all results; of course, you have to do some kind of derivations by yourself, and you can see that this is what one encounters when you have general Markovian queues. For example, in a priority, things become complex; that is what you have to understand, that is it. So, these are all about these priority queuing systems that we are going to see in this course as of now.
Thank you bye.