Introduction to Queueing Theory

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

Lecture - 24
Nonpreemptive Priority Queues

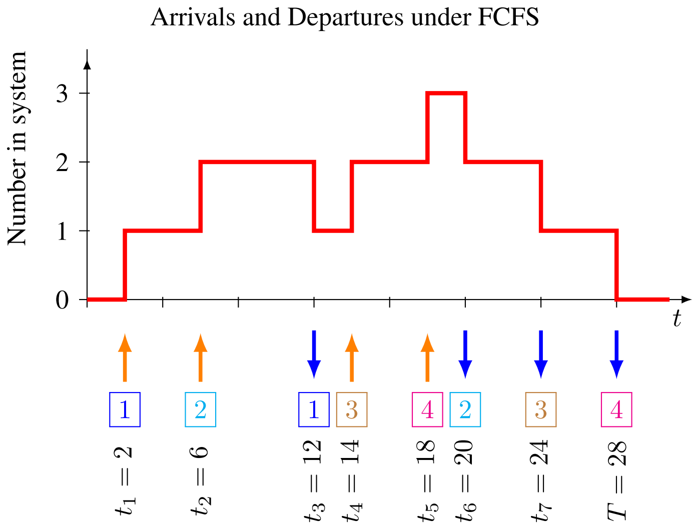
Hi and hello, everyone. We will see next what are called Priority Queues. So, these kinds of systems occur very frequently or very often you would see in real-life systems. So, these are also models which have been studied long back. What we are going to see is that we are going to see this queue in the Markovian framework, though it might be, sometimes, easier to get the required performance measures using a more general approach. But, we will be viewing it from the Markovian perspective. Recall that what we have considered so far are all birth-death queueing systems and Markovian queueing systems. But they all have queue discipline as first come first serve or first in first out models. So, the customers were selected for service on a first come first serve basis or first in first out basis. But, there do exist alternative queueing disciplines which include last come first serve like in an inventory system or selection in random order or priority, etcetera. But you may not see all, but what we are going to look at is this particular queue discipline is what we are going to look at it. So, in priority schemes, when in priority queueing discipline, when priority is being employed, the higher priorities are selected for service ahead of those with lower priorities. This is done in two ways: one is with preemption and without preemption. And what is that? So, in a preemptive case, the customer with a higher priority either gets into service immediately on arrival if no customers are there or if a lower priority customer is getting service. If this customer who is currently getting service is of the same priority or higher priority, obviously, he will get into the queue. He will interrupt any service that is ongoing for any lower priority customer to start his or her service is what is the preemptive case. Because he preempts or interrupts the service of any lower priority customer that might be ongoing to start his service; otherwise, he will get into the queue. Now, depending upon what happens to the customer whose service has been interrupted in the middle by a higher priority customer, this could further be divided into two cases: one is called preemptive resume; in this particular case, the service can be resumed from the point of interruption. So, that is one case, or it can be started afresh. So, you could have things like situations in both may be possible here because it is some sort of paperwork that is being done as a service; then wherever you stopped probably, you will start from there for the lower priority customer that is one way. If it is something that requires some startup things, and then say, for example, if you are looking at some kind of heating process is involved.

Then you have to start afresh because the already done one is not of no use kind of thing. So, that could be a preemptive-non-resume. So, in general, this could be, of course, you could think of more general situations as well; this is a preemptive case. A preemptive case means he preempts any ongoing service of a lower priority customer; that is the key. In a non-preemptive case, the higher priority customer goes into the queue. But either if you have a system where separate queues are there, of course, he will join the queue of the appropriate one. But for the system as a whole, we might consider this is a single queue. But then where will he move? He will move ahead of any lower priority customer, but he will be behind any higher priority customer or of the same level priority; the ones who arrived before him will be ahead of him. So, that is how the thing; within that priority, of course, they may follow an FCFS basis in that case. So, even if there is a lower priority customer who is getting service, he will not preempt the ongoing service. So, he is not going to disrupt any ongoing service that is being given to a lower priority customer; he will wait and then, but he will move ahead of any of those lower priority customers in the queue when selection for service is coming. But, must wait for any customer whoever is there, whether higher or lower currently ongoing service he is not going to interrupt, that is what we said irrespective of priority any customer service to complete the service he will wait. So, that is the non-preemptive case, and its convention to use the lower numbers to denote the higher priorities. So, we may use 1 for the highest priority and 2 second highest, 3 the third highest, and so on; in the case of two priority like, we might use simply 1 and 2, where this 1 could also be referred to as higher priorities, and 2 would be referred to as lower priorities; higher and lower. Since there are only two so, you could also call this for convenience that you have seen. Now, these priority queues, because of this priority discipline that you have here, are, in general, difficult to analyze compared to the non-priority case. But, if you recall that for the $M/M/1$ queue, for example, and some other related models or anywhere, the derivation of the steady-state system size probabilities did not depend on the queue discipline. We pointed out this while discussing $M/M/1$ also and at which point the queue discipline comes into the picture also; we have discussed at that point of time. So, that is what you may recall. And, it can be shown that indeed like, it can be shown that as long as the selection of customers for service is independent of the relative size of the service time, $\{p_n\}$'s would be independent of the queue discipline.

What does that mean? That would mean that, for example, if you are looking at a situation where the shorter service time requirements are taken up first for service. So, those kinds of things are not happening; that is the case where the selection would depend on the relative size of the service time, shortest service time first. This is the one queuing discipline that exists, and we may not see that, but of course, that is one such discipline. So, in such cases, this is not the case as long as such situations are not there, the selection of customers for service is independent of the relative size of the service time. Then $\{p_n\}'s$ are independent of the queue discipline, when as long as this is true, the Little's formula remains unchanged. And hence the average number in the system and hence the average waiting time. But we know there will be changes in the waiting time distribution; again, you may recall our $M/M/1$ analysis where we said that the first come first serve is what we are assuming while deriving the waiting time distributions. If not, what would happen you can imagine in that situation.

So, that is what we can see, and we can now make a claim which says that waiting times are stochastically smallest under FCFS with all other things being equal; an introduction of any scheme of priorities which does not depend on the service time makes the higher-order moments worse than what was it under the FCFS discipline, that is what mean. So, higher-order moments under this is stochastically smaller. And, it will be better under the FCFS scheme; that is what we are meaning, which means that you will have a scenario where the FCFS scheme produces a lower variance for the waiting times under the FCFS scheme than any other scheme. And, what would that lower variance mean? As for any data, mean and variance are important; mean gives what the center point of the data is, and the spread or the dispersion is given by the variance. Now, variance smaller means the points are more closely clustered around the mean, that is, the meaning of lower variance. So, in this case, there is greater equality among the waiting times of various customers; that is what it would mean. Because any customer that you pick it up, you know it will be they will have closer waiting time durations, there is no undue wait. So, in that case, this could also be lower variance can also be associated with, in some sense, the fairness of the system because it is fair to everyone. So that, almost everyone gets almost equal waiting time rather than very someone getting 5-time units, someone getting 20-time units, and if you someone getting

10 units, someone getting 13 units. So, this is what is more reasonable than your thinking. So, this is what you can think of these, or you can view it as if it is a more fair scheme, the FCFS scheme, than the other scheme. So, that is the implication when you look at variance as a measure of fairness in a way. So, a lower variance would give you fairer waiting times for all the customers. Now, this claim is very difficult to prove at this point with all the materials that we have at hand or even we are going to see later or also including that. So, you will need it because you need to know what you mean by a random variable being stochastically smaller and so on and so many other results connected with that. But, we can get an idea about how why this is true or how this is true; let us just take an illustration. So, let us look at this diagram that we have put here, which depicts the arrivals and departures under the FCFS scheme.



- Under FCFS, the waiting times are $10, 14, 10$ and 10, with mean as 11 and unbiased variance as 4.
- Under LCFS, the waiting times are $10, 14, 6$ and 14, with mean as 11 and unbiased variance as $44/3 = 14.67$. [customer-4 departs at 24 and customer-3 at 28]
- Variance higher under LCFS.

And, this is what the claim typically is what what we made it is basically the claim is in terms of the distributions you are talking about. But, that would imply that it will have higher higher-order moments worse, and in this any LCFS or priority schemes than it was under FCFS, that is what it would mean, and; that means that it becomes some sort of unfair system. But, there is always a question about fairness in the system and the queue discipline; this is always conflicting in nature. So, if you want to introduce this, then basically, you are compromising on the fairness of all customers. But, if it has to be for a certain reason that if you have to create a system where this particular group has to be given priority, you have to live with that unfairness in that situation; that is what it will turn out to be.

- Furthermore, the remaining total service or work required for a single server at any point during an arbitrary busy period is independent of the order of service as long as the system is conservative (i.e., no service needs are created or destroyed with the system).

Whatever it comes, for example, there is no reneging happens in the middle of the service. So that means the system need is getting destroyed for the remaining period, and the one customer also is lost. Now how do you count whether this is served or not served is also another question, but his needs are destroyed. So, this destroyed, or you might look at the case where again when in priority when the preemption happens, what happens to the preempted service which starts all over; then again some portion which server has already put in that is getting destroyed, that service needs have been destroyed. And, when you start afresh, then you are creating additional needs as well. And also, like when there are customers, you are not forcing the server to be idle; it is not the situation. So, if you do not have any such behavior, then that system is what is generally called a conservative system. So, in such conservative systems, the workload for the server is independent of any order of service. So, that is also what one can see. Now, what we do next is we will consider; starting with the priority queueing system, but of course, as I said that this is the Markovian system that we are thinking about, and in that framework, we are going to look at that.

- Customers arrive as a Poisson process to a single exponential channel.

  And this is the non-preemptive priority system with two classes that we are going to consider. So, you can think of it as if it is a typical $M/M/1$ system only, but with certain changes.

- A customer, upon arrival, is assigned to one of the two priority classes.

  You can think that you know it is independent of everything else that any particular customer can be assigned with certain probability class 1 and certain probability class 2, you think it that way that would be equivalent to.

- First or higher priority customers arrive as PP(λ_1) and the second or lower priority customers arrive as PP(λ_2).
  ▶ The total arrival rate is $\lambda = \lambda_1 + \lambda_2$.

- Assume that there is no preemption.

  So, this is a non-preemptive priority system. So, there are two priority classes, the higher priority one arrives at the rate λ_1, and the lower priority arrives at the rate λ_2. Now, such a system can be modelled with the assumption of exponential service and the arrival process being Poisson. The arrival could be with a certain probability at any given point of time, the arrival could be a higher priority, and with the remaining probability, the arrival could be of the lower priority. And, every other assumption connected with such a Markovian system is in place.

- The system can be modelled by a CTMC with state space $S = \{0\} \cup \{(m, n, r) : m, n = 0, 1, 2, \dots, \max\{m, n\} > 0, r = 1, 2\}$ and the corresponding steady state probabilities (where m & n not both 0) are denoted by

$$\begin{aligned} p_{mnr} &= P\{m \text{ priority-1 customers in the system,} \\ &\qquad n \text{ priority-2 customers in the system, and} \\ &\qquad \text{the customer in service is of priority } r = 1 \text{ or } 2\} \\ p_0 &= P\{\text{the system is empty}\} \end{aligned}$$

♦ Let $L^{(i)}, L_q^{(i)}, W_q^{(i)}, W^{(i)}$ denote the measures of effectiveness for class-i customers.

Now we will consider three models, and we will compare them to make certain inferences. So, **model *A*** would be equal service rates meaning that both the class demands an equal service needs from the server; that is, the service rate, service distribution is exponential we already mentioned. But, the service rate or the mean service time that is required to serve either priority-1 or priority-2 is equal to the rate μ.

- Assume that the service rates of both the classes are equal to μ. Define

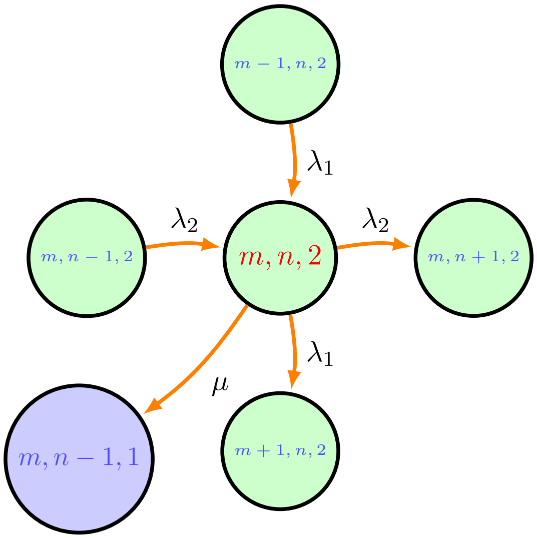
$$\rho_1 = \frac{\lambda_1}{\mu}, \quad \rho_2 = \frac{\lambda_2}{\mu}, \quad \rho = \rho_1 + \rho_2 = \frac{\lambda}{\mu}$$

We assume that $\rho < 1$ for the stability of the system for the underlying Markov chain to be positive recurrent, and so on. So, all these λ, $\lambda_1, \lambda_2, \mu$, of course, we assumed to be strictly greater than 0. So, $\rho < 1$ is what is the stability condition that is required for this steady-state to exist. Now, once we have $S = \{0\} \cup \{(m,n,r) : m,n = 0,1,2,\ldots, \max\{m,n\} > 0, r = 1,2\}$ as the state space, now we can write the balance equations corresponding to the different states in the state space.

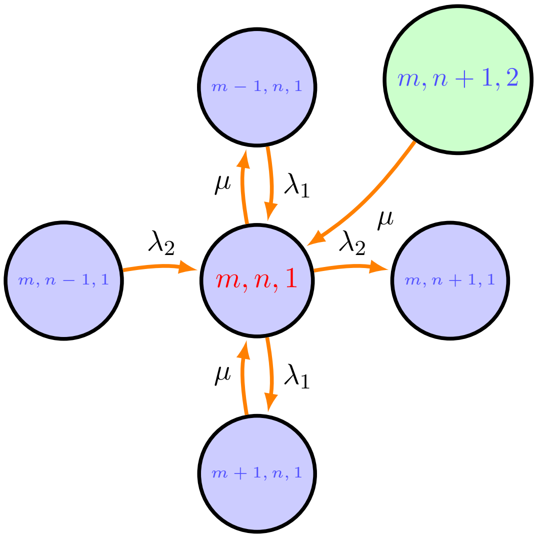
- The balance equations are:

$$
\begin{aligned}
(\lambda+\mu)p_{mn2} &= \lambda_1 p_{m-1,n,2} + \lambda_2 p_{m,n-1,2} && (m \geq 1, \quad n \geq 2) && \text{(Eq.A)}\\
(\lambda+\mu)p_{mn1} &= \lambda_1 p_{m-1,n,1} + \lambda_2 p_{m,n-1,1} + \mu(p_{m+1,n,1} + p_{m,n+1,2}) && (m \geq 2, \quad n \geq 1) && \text{(Eq.B)}\\
(\lambda+\mu)p_{m12} &= \lambda_1 p_{m-1,1,2} && (m \geq 1) &&\\
(\lambda+\mu)p_{1n1} &= \lambda_2 p_{1,n-1,1} + \mu(p_{2n1} + p_{1,n+1,2}) && (n \geq 1) &&\\
(\lambda+\mu)p_{0n2} &= \lambda_2 p_{0,n-1,2} + \mu(p_{1n1} + p_{0,n+1,2}) && (n \geq 2) &&\\
(\lambda+\mu)p_{m01} &= \lambda_1 p_{m-1,0,1} + \mu(p_{m+1,0,1} + p_{m12}) && (m \geq 2) &&\\
(\lambda+\mu)p_{012} &= \lambda_2 p_0 + \mu(p_{111} + p_{022}) &&&&\\
(\lambda+\mu)p_{101} &= \lambda_1 p_0 + \mu(p_{201} + p_{112}) &&&&\\
\lambda p_0 &= \mu(p_{101} + p_{012}) &&&&
\end{aligned}
$$

State Transitions for $(m,n,2)$ (Eq.A above)



State Transitions for $(m,n,1)$ (Eq.B above)



These are the balance equations, and how we have obtained them is what we have depicted and explained through these two diagrams.

- By Little's law (to the server), ρ is the fraction of time the server is busy, or equivalently $p_0 = 1 - \rho$.

- Similarly, the fraction of time the server is busy with a priority-r customer is ρ_r. Thus,

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} p_{mn1} = \rho_1 \quad \text{and} \quad \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} p_{mn2} = \rho_2.$$

- And, since the service times are $Exp(\mu)$ for both the priority classes, the system size steady state distribution for the number of customers is

$$p_n = \sum_{m=0}^{n-1} (p_{n-m,m,1} + p_{m,n-m,2}) = (1-\rho)\rho^n, \quad n > 0.$$

- Obtaining a complete analytical solution is very difficult here. We will try to obtain the expected value measures.

So, what we will try to do is we will try to obtain certain expected value measures using generating functions which is what is typically done in such complex models; at least the expected value measures you can obtain explicitly so that you can do the performance analysis of that. For which what do we do? We define because of this three-dimensional nature of the process, because basically, the CTMC is of three-dimension and because of this three-dimensional nature like we have to have more complex generating functions, but because the third dimension is either 1 or 2. So, you can slightly make it in two dimensions in the following way.

- Define two-dimensional generating functions as:

$$P_{m1}(z) = \sum_{n=0}^{\infty} p_{mn1} \qquad (m \geq 1), \qquad\qquad P_{m2}(z) = \sum_{n=1}^{\infty} z^n p_{mn2} \qquad (m \geq 0),$$

$$H_1(y,z) = \sum_{m=1}^{\infty} y^m P_{m1}(z) \quad (\text{with } H_1(1,1) = \rho_1),$$

$$H_2(y,z) = \sum_{m=0}^{\infty} y^m P_{m2}(z) \quad (\text{with } H_2(1,1) = \rho_2),$$

$$\begin{aligned} \text{and} \quad H(y,z) &= H_1(y,z) + H_2(y,z) + p_0 \\ &= \sum_{m=1}^{\infty}\sum_{n=0}^{\infty} y^m z^n p_{mn1} + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} y^m z^n p_{mn2} + p_0 \\ &= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} y^m z^n (p_{mn1} + p_{mn2}) + \sum_{m=1}^{\infty} y^m p_{m01} + \sum_{n=1}^{\infty} z^n p_{0n2} + p_0, \end{aligned}$$

where $H(z,y)$ is the joint generating function for the two classes, regardless of which type is in service.

- Note that $H(y,y) = \dfrac{p_0}{(1-\rho y)}$, with $H(1,1) = 1$, since $H(y,z)$ collapses to the generating function of an $M/M/1$ queue when $z = y$ and thus no priority distinction is made. Also,

$$\left.\frac{\partial H(y,z)}{\partial y}\right|_{y=z=1} = L^{(1)} = L_q^{(1)} + \rho_1 = \lambda_1 W^{(1)},$$

$$\left.\frac{\partial H(y,z)}{\partial z}\right|_{y=z=1} = L^{(2)} = L_q^{(2)} + \rho_2 = \lambda_2 W^{(2)}$$

- Multiplying the balance equations by the appropriate powers of y and z and summing, we have

$$(1 + \rho - \rho_1 y - \rho_2 z - \frac{1}{y})H_1(y,z) = \frac{H_2(y,z)}{z} + \rho_1 y p_0 - P_{11}(z) - \frac{P_{02}(z)}{z},$$

$$(1 + \rho - \rho_1 y - \rho_2 z)H_2(y,z) = P_{11}(z) + \frac{P_{02}(z)}{z} - (\rho - \rho_2 z)p_0$$

- In order to determine H_1 and H_2 completely, we need the values of $P_{11}(z), P_{02}(z)$and p_0. An equation giving the relationship between these quantities can be found by summing z^n $(n = 2, 3, \dots)$ times the balance equations involving p_{0n2}, and then using the final three balance equations. This gives

$$P_{11}(z) = \left(1 + \rho - \rho_2 z - \frac{1}{z}\right) P_{02}(z) + (\rho - \rho_2 z)p_0.$$

- Substituting the above equation for $P_{11}(z)$ into the previous equations gives H_1 and H_2 as functions of p_0 and $P_{02}(z)$.

- We can thus obtain $H(y, z)$ also in terms of p_0 and $P_{02}(z)$ as

$$\begin{aligned} H(y,z) &= H_1(y,z) + H_2(y,z) + p_0. \\ &= \frac{(1-y)p_0}{1 - y - \rho y + \rho_1 y^2 + \rho_2 yz} + \frac{(1 + \rho - \rho z + \rho_1 z)(z - y)P_{02}(z)}{z(1 + \rho - \rho_1 y - \rho_2 z)(1 - y - \rho y + \rho_1 y^2 + \rho_2 yz)}. \end{aligned}$$

- From $H(1, 1) = 1$, we get $P_{02}(1) = \dfrac{\rho_2}{1 + \rho_1}$. We can then determine $L^{(1)}$ from $L^{(1)} = \left.\dfrac{\partial H(y, z)}{\partial y}\right|_{y=z=1}$ (the partial derivative cannot be evaluated directly at $(1, 1)$ so a limit must be taken). Fortunately, in these steps, only $P_{02}(1)$ is required, and not the function $P_{02}(z)$.

- Since the total number of customers in the system is the same as that of the $M/M/1$ system, we have that $L^{(1)} + L^{(2)} = \dfrac{\rho}{1-\rho}$ and hence $L^{(2)} = \dfrac{\rho}{1-\rho} - L^{(1)}$.

- The other measures can be obtained from

$$L_q^{(i)} = L^{(i)} - \rho_i, \quad L_q^{(i)} = \lambda_i W_q^{(i)}, \quad L^{(i)} = \lambda_i W^{(i)}$$

- The final results for $L_q^{(i)}$ are (recall that $\rho = \lambda_1/\mu + \lambda_2/\mu$)

$$\boxed{L_q^{(1)} = \frac{\lambda_1 \rho}{\mu - \lambda_1}, \qquad L_q^{(2)} = \frac{\lambda_2 \rho}{(\mu - \lambda_1)(1 - \rho)}, \qquad L_q = \frac{\rho^2}{1 - \rho}}$$

- Extra (for interested, refer Miller (1981)): The actual probabilities for priority-1 customers can be shown to be given by

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

Now, with these mean value results, we have obtained what? Not steady-state probabilities, but certain mean value results; with these mean value results, we can make some observations here, which are of importance; though we do not have the complete distributions.

1. The first part is that lower priority customers always wait in queue longer (on average) than the higher priority customers. This can be seen as follows:

$$W_q^{(2)} = \frac{\rho}{(\mu - \lambda_1)(1 - \rho)} = \frac{\left(\frac{\rho}{\mu - \lambda_1}\right)}{1 - \rho} = \frac{W_q^{(1)}}{1 - \rho} > W_q^{(1)} \quad (\text{when } \rho < 1).$$

However, it is not always the case that $L_q^{(2)} > L_q^{(1)}$.

$W_q^{(2)}$ is stochastically greater than $W_q^{(1)}$, but $L_q^{(2)}$ is not strictly greater than $L_q^{(1)}$, meaning the average waiting time for the lower priority customers is longer, but the number in the system need not be so.

The number in the system for the lower priority customers may not be longer than this particular case. You can show with certain parameter values by picking λ_1, λ_2, and μ; you can show that $L_q^{(2)} > L_q^{(1)}$ and $L_q^{(2)} < L_q^{(1)}$; both are possible. So, that is an exercise problem in the text, but of course, that is one can make an observation here.

2. As $\rho \to 1, L_q^{(2)} \to \infty$ (and so do $W_q^{(2)}, W^{(2)},$ and $L^{(2)}$). However, $L_q^{(1)}$ approaches a finite limit, if $\frac{\lambda_1}{\mu} < 1$ is held constant. The first-priority means go to ∞ only when $\frac{\lambda_1}{\mu} \to 1$.
   ► Possible that higher priority customers do not accumulate, even when an overall steady state does not exist.

   This means overall system size is exploding; even then, it is possible that you may have a system where the higher priority customers do not accumulate; they exhibit in some sense a steady-state behavior or a stable behavior.

3. The presence of class-2 customers still creates delays for class-1 customers (because of nonpreemptiveness). In particular,

$$\{L_q^{(1)} \text{ when } \lambda_2 = 0\} < \{L_q^{(1)} \text{ when} \lambda_2 > 0\}.$$

   However, if the class-1 customers have the power of preemption, then the class-2 customers do not effect the class-1 customers.

4. The average number in queue is the same as an $M/M/1$ queue. Similarly, the unconditional average wait, $W_q = \left(\frac{\lambda_1}{\lambda}\right) W_q^{(1)} + \left(\frac{\lambda_2}{\lambda}\right) W_q^{(2)}$ is the same as an $M/M/1$ queue.

These are some observations that you can make with respect to this non-preemptive priority queueing system with two classes of customers, with equal rates, which is what is model A that we have considered. We have written down the balance equations; we cannot solve them completely; it is very difficult. So, what we did is expected value measures that we have obtained which we have given in the boxed one

$$\boxed{L_q^{(1)} = \frac{\lambda_1 \rho}{\mu - \lambda_1}, \qquad L_q^{(2)} = \frac{\lambda_2 \rho}{(\mu - \lambda_1)(1-\rho)}, \qquad L_q = \frac{\rho^2}{1-\rho}}$$

which we are going to refer back to in the later lecture because of some comparison that we want to make. So, of course, we will end here, and we will continue the same priority queues with further discussion on this model in the next lecture.
Thank you bye.