Introduction to Queueing Theory

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

Lecture - 21
Queues with Bulk Service

Hi and hello, everyone. What we are seeing now are the queuing systems that can be modeled by a general continuous-time Markov chain or what we generically call General Markovian Queueing models meaning non birth-death queuing models in which the first group, we are still within the, in some sense, Poisson exponential assumptions and in which we just considered what we called it as a bulk arrival queue, and we saw how it could be analyzed. What we will see next is what we call bulk service queueing models. So, this bulk service model is similar to that in the notation wise what we are using it is $M/M^{[Y]}/1$ is what is the model that we are going to look at it we mean we are looking at single server models only. So, what do we have?

- We are considering a single-server Markovian queue with $bulk\ service$.

  So, what are the typical examples?

  ► The mass transit vehicles or mass transit carriers are very good natural examples that serve the customers, what we call these customers in batches, so this is a natural batch server.

- All the assumptions of $M/M/1$ are applicable with an additional assumption that the customers are served in groups/batches of size K.

  Now, this batch size is we are keeping it independent of any other thing that we are considering within the model. So, that is what is batch size. So, here the notation wise, we just kept it Y, which could be any random variable, but the consideration that we are making it is batch size K.

- Now, within this structure, we are considering two variations based on the system behavoiur when there are less than K in the system.

  ► **Full Batch Model**:- Here the server processes exactly K customers at a time. If less than K customers are in the system, then the server remains idle until there are K customers in the system, at which point the server processes the K customers simultaneously. The service times are $Exp(\mu)$ and is the same for all the customers in a batch.
  For example, a small ferry service crossing the river only when it is full.
  If you look at it in some places, there are small ferries that only when it is full it will cross the river or go to a

certain island and so on. It may not be the case that they are working in some time-scheduled manner; it might be if that is the case, then it is a slightly different model. But there are some private ferry services that might cross the river only if, say, if they have 10 persons, they can hold only up to 10 finish. So, I am just giving a simple example, but there could be many different situations which you can think as if it can happen in this way right in a production system or in anywhere like this can easily be taken into account because of this cost-effective manner like you will have only when there are K items are there, you start the processing of that altogether because all of them take the same time. So, that is the possible idea which we have full batch meaning, there is K, K is a fixed number here, and the server always processes K customers at once following an exponential distributed service time.

Now, as opposed to this full batch model, there could be another variation of this model, which is a partial batch model. So, here and in fact, this is what is the model that basically Bailey first introduced in 1954.

► **Partial Batch Model**:- Here the server can process partial batches up to a maximum size of K. As before, customers are served K at a time, but now if there are less than K in the system, the server begin service on these customers. Furthermore, when there are less than K in service, new arrivals immediately enter service up to the limit K and finish with the others, regardless of the entry time into service. Service times are $Exp(\mu)$.

For example, a guided tour or a puppet/movie show. Suppose, if you are looking at some show, the show will begin as long as some customers are there possibly here we can you can generalize any number of customers is there the show will start. Now, anyone who arrives late will simply join the show, and the show will finish for everyone at the same time, a movie show, a light, and laser and light show. These kinds of shows like you can think or a guided tour, for example, you are going to you some tourist place and the guide explains the significance the guide will start with some number of customers and then as people come they will join in the intermediate and then they will finish with everyone else at the same time. Of course, I mean, even in a real-life situation, you can think about many situations where this partial batch model would hold good. So, we will consider these two variations of such a single server Markovian queue with bulk service. But there are many other variants with respect to this bulk service; I mean, that is the beauty of such queueing models as you would see. When you observe, you will find that this does not really fit into any of the existing models that you have in your hand for analysis. So, what do you do? You create a new model; it is always as always you do that.

- And there are other variants; the prominent one is what is called the **general bulk service rule**; many a time, it is simply referred to as the GBS rule, which was basically initiated in 1967 by Neuts and in which the service is rendered with a minimum batch size of L and with a maximum batch size of $K(L \leq K)$.

So, you have the K like in the previous description in the previous service rules that we have followed. But if you think in a partial service partial batch model, it will start with any number of customers, even if there is 1, but it may not be economical to start with 1. So, you want it to reach at least some level, say L; that is what this would mean, but that is a difference between that; we will come to that. So, how it is being operated.

► If there are fewer than L customers in the system, then the server remains idle until there are L customers whereupon all L enters the service.

► If there are L or more, but less than K customers waiting, then all of them served together.
► If there are more than K customers waiting, then a group of K enters the service.
► The corresponding model is typically denoted as $M/M(L, K)/1$, or more common; it will be something like $M/M(a, b)/1$; this is the same as $M/M(L, K)/1$.

So, this is the same as this because I am just writing because this is the notation that you might find in many books. So, that is the model that you have, which is what a general bulk service rule a and b are the levels since we already have K here. So, we just define another one, L, to denote it in this manner; it is the same; there is nothing on that. So, this is just a notation, whichever these are the same.

► So, examples include this airport shuttle. So, when you are trying to board in the shuttle like at least some number of customers come, then it will start, it may not be full, the bus may not be full when you are getting into the flight.

► Or a lift on a building's ground floor; probably it is being operated by a person, so even just 1 person entering may not start the lift. So, he wants not to be full, but at least some number of customers come in so that he will start the lift.
Suppose, if it is a big storey some 40, 50 floors of the building and then you will not be able to operate for every 1 customer so you may wait for some time until some more customers join in.

► So, that is the lift or a traffic flow which is basically described in this newspaper itself like where the minor road merges into a major road.

So, there is a major road, and there is a minor road that comes and merges with this. So, the traffic coming from that minor road when it will be allowed inside the major road blocking the traffic of the major road. Since this is the major road, you will not block until there are at least you know L number of cars or vehicles that are coming on the minor road, at which point then the traffic is stopped on the major road. And this is minimum L has to be there, or sometimes that is what you would look at. So that is another example or many other examples also you can think.
But now, the difference between the previous model and this model is basically what we are describing here; the difference may be there may not be there depending upon how you are pitching it; what is that?

► Systems where the customers can join or access an ongoing service batch anytime before the service of the batch is completed, provided the specified maximum service batch size is not reached, are known as systems with *accessible* batch service (otherwise they are called non-accessible).

So, the previous one, the partial batch model that we have considered, is basically a system with accessible batch service. Whereas, in this particular case, for example, lift suppose, if you are thinking right of course, once it starts from the ground floor then any even with less than the full capacity then the arriving customers of the ground floor cannot access. So, the service that is being provided to the already existing customers. So, in that

case, it is basically a non-accessible, you know, batch service system. Whereas in a traffic flow, depending upon how you operate, you can think that at the point of start of the service, whatever vehicles are there only that will be allowed to go into that anything that is coming after that you have to wait for the next cycle then it is non-accessible. But during the course, suppose if some more vehicle comes and you also allow that to pass up to a maximum of K, then it becomes accessible. So, like that, you can think about the models as a variant of this; whatever it is, this is the feature. Now, the literature on both accessible batch service model and non-accessible batch service models are plenty. There are different situations that require the model to be studied under both scenarios. So, if it is accessible, then both the previous two models can be thought about as this general bulk service rule, but if it is a non-accessible batch service model, then at least the partial batch would not think about this part as a special case of this. But the fixed size, you can always think about it like this, where $L = K$ is what then will give you the fixed-size one.

- Another variant is that of the size of a batch being a random variable (depending on the unfilled capacity of the server).

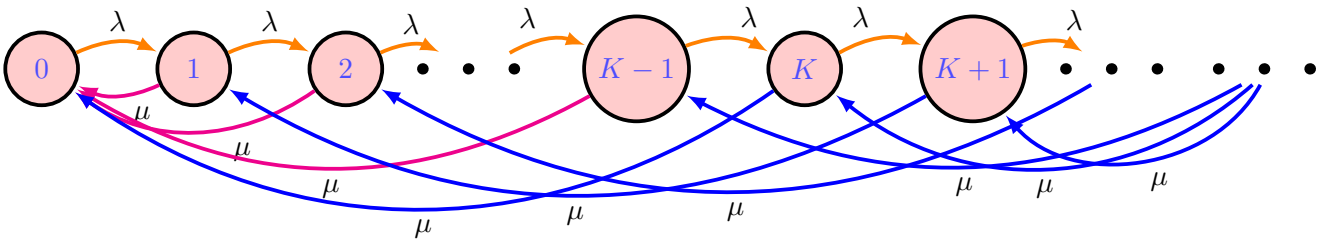
  Again you will not send all, but you may send a partial one that could be random; again, this is all studied, but we consider here only the previous two models.

  So, these are all some ideas about; what could be the general models that one can think of along those lines that we are going to discuss within the Markovian framework itself as you can think about different service policies will give rise to different models. Now, what are the kinds of policies that you can think of. Of course, it will be dependent on the system, and accordingly, you can develop.

- So, that is why to give a hint that we are mentioning this, but we will be considering only these models like $M/M^{[K]}/1$ full batch model and partial batch model. And the partial batch model with this notion of accessible batch service is basically we are looking at the partial batch model with accessible batch service, not with non-accessible batch service models.

So, these are the two models that we are going to consider next; that is our objective in this session. So, the first model is the $M/M^{[K]}/1$ partial-batch model; every assumption we have already described; it is a partial batch model.

- So, the underlying model is a CTMC model as described below.



- The stochastic balance equations are

$$\lambda p_0 = \mu p_1 + \mu p_2 + \cdots + \mu p_{K-1} + \mu p_K$$
$$(\lambda + \mu) p_n = \mu p_{n+K} + \lambda p_{n-1}, \quad n \geq 1.$$

So, this is what is the partial batch model now; we have described the partial batch model, and we have written down now the flow balance equations or the global balance equations for this particular system. Now, what we have to do is we have to solve this system in the usual as per your usual process; once we have written it down, then we have to solve this is the state balance equations. Now, we look at the general equation; we will come back to this equation in a moment, but we look at now the general equation; we can solve it by generating function method or other methods, but we will use the operator method here.

- The general equation above can be rewritten in operator notation as

$$[\mu D^{K+1} - (\lambda + \mu)D + \lambda]p_n = 0, \quad n \geq 0.$$

- If $(r_1, r_2, \ldots, r_{K+1})$ are the roots of the operator or characteristic equation, then

$$p_n = \sum_{i=1}^{K+1} C_i r_i^n, \qquad n \geq 0.$$

  Now, $\sum_{n=0}^{\infty} p_n = 1$ implies that each r_i must be less than one or $C_i = 0$ for r_i not less than one.

- From Rouche's theorem, it can be found that there is exactly one root (say, r_0) in $(0, 1)$, under the condition for the stability of the system $\rho = \frac{\lambda}{K\mu} < 1$. (Exercise! Hint: Take $g = \mu r^{K+1} + \lambda$ and $f = -(\lambda + \mu)r$.) This implies that

$$p_n = Cr_0^n, \qquad n \geq 0, \quad 0 < r_0 < 1.$$

  Now, $\sum_{n=0}^{\infty} p_n = 1 \Rightarrow C = p_0 = 1 - r_0$ and hence

$$p_n = (1 - r_0)r_0^n, \quad n \geq 0, \quad 0 < r_0 < 1.$$

- The performance metrics can be obtained in the usual manner (noting that the steady-state distribution is geometric, like in $M/M/1$).

- We get

$$L = \frac{r_0}{1 - r_0}, \quad W = \frac{L}{\lambda} = \frac{r_0}{\lambda(1 - r_0)}$$

  and, using Little's Law:

$$W_q = W - \frac{1}{\mu}, \quad L_q = L - \frac{\lambda}{\mu}.$$

  Alternatively, directly computing from the steady-state probabilities, we get

$$L_q = r_0^K L, \quad \text{which is equal to } L_q = L - \frac{\lambda}{\mu}.$$

These are all exercises in your book. But of course, you can try it is not a very complex one that you can you have to try if you want to understand the theory better.

Example. • A drive-it-through-yourself car wash facility installs a new machinery that permits the washing of two cars at once (and one if no other cars wait).

- A car that arrives while a single car is being washed joins the wash and finishes with the first car.
- There is no waiting capacity limitation.
- Arrivals are Poisson with mean 15 per hour. Time to wash a car is exponentially distributed with a mean of 6 minutes.
- The given data are: $\lambda = 15/h, \quad \mu = 10/h, \quad$ and $\quad K = 2$. The characteristic equation is

$$10r^3 - 25r + 15 = 5(2r^3 - 5r + 3) = 0.$$

Observing that one root is 1, we get

$$2r^2 + 2r - 3 = 0 \quad \implies r = (-2 \pm \sqrt{28})/4.$$

Thus, we get $r_0 = 0.8229$, choosing the root less than 1. Therefore,

$$L = \frac{0.8229}{0.1771} = 4.6458\ cars \quad L_q = 4.6458 - \frac{15}{10} = 3.1458\ cars.$$

So, we are observing that 4.6 cars will be there in the system, and 3.1458 cars will be waiting on an average at any given point of time in such a system; then, you can decide whether this is sufficient or you have to do something more to either increase the rate if this you feel this is unreasonable then you want to reduce then you can do all those things. A typical example where you can use this kind of analysis. This is the partial batch model.

Next, what we will consider $M/M^{[K]}/1$, but with now full-batch model.

- In this model, we assume that the batch size must be exactly K for the server to start the service, and if not, the server waits until such time to start.
- The stochastic balance equations are modified accordingly to yield

$$\begin{aligned} \lambda p_0 &= \mu p_K \\ \lambda p_n &= \mu p_{n+K} + \lambda p_{n-1}, \quad 1 \le n < K \\ (\lambda + \mu) p_n &= \mu p_{n+K} + \lambda p_{n-1}, \quad n \ge K \end{aligned}$$

- The third equation is identical to that of the partial-batch model and hence, proceeding on similar lines, we get

$$p_n = C r_0^n, \qquad n \ge K - 1, \quad 0 < r_0 < 1.$$

But, obtaining C (and p_0) is a bit complicated here.

- From the first equation, we have

$$p_K = \frac{\lambda}{\mu} p_0 = C r_0^K, \quad C = \frac{\lambda p_0}{\mu r_0^K} \quad \text{and therefore} \quad p_n = \frac{p_0 \lambda r_0^{n-K}}{\mu} \qquad n \ge K - 1.$$

So, I have expressed p_n in the first step in terms of p_0 that is what we have done here. So, this is one part we have done. Now, I have to look at other p_n's in this range 1 to $K - 1$; I will try to express in terms of p_0, then I can utilize the total probability equal to 1 condition to obtain this p_0 that is the idea. So, basically, I have not yet used this equation so far; the middle set of equations $\lambda p_n = \mu p_{n+K} + \lambda p_{n-1}, \quad 1 \le n < K$.

- Now, to get p_0, we use the remaining $K-1$ equations given by

$$\mu p_{n+K} = \lambda p_n - \lambda p_{n-1}, \qquad 1 \leq n < K.$$

Using the geometric form of p_n when $n \geq K-1$ and substituting for p_{n+K} in the above equation, we get

$$p_0 r_0^n = p_n - p_{n-1}, \qquad 1 \leq n < K.$$

These equations can be solved by iteration starting with $n = 1$. Alternatively, we observe that these are nonhomogeneous linear difference equations whose solutions are

$$p_n = C_1 + C_2 r_0^n.$$

Direct substitution into the above equation implies that $C_2 = -\frac{p_0 r_0}{(1-r_0)}$. The boundary condition at $n = 0$ implies that $C_1 = p_0 - C_2$. This gives

$$p_n = \begin{cases} \dfrac{p_0(1-r_0^{n+1})}{1-r_0}, & 1 \leq n < K-1 \\ \dfrac{p_0 \lambda r_0^{n-K}}{\mu}, & n \geq K-1 \end{cases}$$

Now, all the p_n's, I have expressed in terms of p_0.

- We can now determine p_0 using the usual boundary condition $\sum_{n=0}^{\infty} p_n = 1$ and we have

$$\begin{aligned} p_0 &= \left(1 + \sum_{n=1}^{K-1} \frac{1-r_0^{n+1}}{1-r_0} + \frac{\lambda}{\mu} \sum_{n=K}^{\infty} r_0^{n-K}\right)^{-1} \\ &= \left(1 + \frac{K-1}{1-r_0} - \frac{r_0^2(1-r_0^{K-1})}{(1-r_0^2)} + \frac{\lambda}{\mu(1-r_0)}\right)^{-1} \\ &= \left(\frac{\mu r_0^{K+1} - (\lambda+\mu) r_0 + \lambda + \mu K(1-r_0)}{\mu(1-r_0)^2}\right)^{-1} \end{aligned}$$

But we know that r_0 satisfy the characteristic equation and this means that

$$\mu r_0^{K+1} - (\lambda+\mu) r_0 + \lambda = 0$$

and thus we obtain finally

$$p_0 = \frac{\mu(1-r_0)^2}{\mu K(1-r_0)} = \frac{1-r_0}{K}.$$

So, I have now determined completely the steady-state distribution for this full batch bulk service model. Now, in the other performance measures of interest, whatever you want, you can obtain from this distribution, so that is always possible. Alternatively, this can also be obtained through generating function approach, which of course, we will not do, but of course, we have seen that this is the solution that you can get for this model.

So, this is the full-batch model. Now you can use these models themselves; suppose you have to use a non-accessible batch service; how you can modify it can happen only in the partial batch case. So, if it is a non-accessible case, then what one has to do is that for all the states, you need to have a parallel state where a system is not empty. But whether someone is waiting or no one is waiting means, this system is busy serving someone, or without serving someone, the server is waiting. So, that kind of expansion will happen in terms of the states of the process from here onwards, of course, things will remain the same and so ones that you have to do. So, you have to make modifications; you can see how one can do it. So, for a simple case of say two or something, and if you put, then you need to introduce only one more state, and then you can analyze that model easily. So, these kinds of variations one can do within this idea. So, this is all about the $M/M^{[K]}/1$ partial-batch model and full batch model that we have seen. So, again, as we said, there are many different service rules, and accordingly, the model can be analyzed under such scenarios as well. But things are complicated; we are not going to get into that; for us, this is sufficient that we get an idea about how one can handle the bulk-service models using these two models. So, with this, we conclude the ah discussion on these bulk queues here.

Thank you.