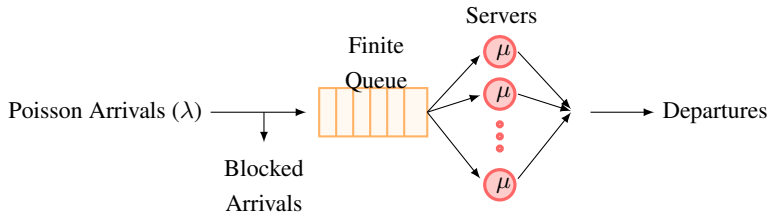
Introduction to Queueing Theory

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

Lecture - 16
M/M/c/K Queues

Hi and hello, everyone. What we will see next is a queue with the feature that it can accommodate only a finite number of customers in the system at any given point of time. So, it is referred to simply as queues with truncation or finite capacity queues or, in short, $M/M/c$, which is what we have seen in the previous model that we have seen $M/M/c$. But now we are putting a restriction on the capacity, which is K. So, this is $M/M/c/K$ queues. So, meaning here, K is for the total system is, not just for the queue, so; that means if there are c servers when we are saying K is the total number of customers that can be there in the system at any given point of time; that means, there is a waiting space for $K - c$ customers that is what it would mean. So, that is the situation.

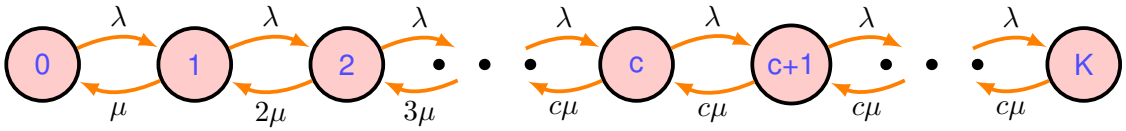
- Similar to $M/M/c$ with the condition that a limit K is placed on the number allowed in the system at any time.



So, this is then the pictorial representation of such a queueing system that we have here, except that for the finite one, we will just make it the boxes; as this is a closed box at this end as well; otherwise, you would have seen the previous figure that these things were open meaning that you know this is infinite capacity queues are there.

♦ Blocked customers is the new feature.

- Modelled by a BDP as in an $M/M/c$ except that $\lambda_n = 0$ for $n \geq K$.



- Steady state system size probabilities are given by:

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0, & 0 \le n < c \\ \dfrac{\lambda^n}{c^{n-c} c!\mu^n} p_0, & c \le n \le K \end{cases}$$

- Since $\sum_{n=0}^{K} p_n = 1$, we get $p_0 = \left(\sum_{n=0}^{c-1} \dfrac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{K} \dfrac{\lambda^n}{c^{n-c} c!\mu^n} \right)^{-1}$.

$$\Rightarrow \quad p_0 = \begin{cases} \left[\dfrac{r^c}{c!} \left(\dfrac{1-\rho^{K-c+1}}{1-\rho} \right) + \sum_{n=0}^{c-1} \dfrac{r^n}{n!} \right]^{-1}, & \rho \neq 1 \quad \{\text{with } r = \frac{\lambda}{\mu}, \rho = \frac{r}{c}\} \\ \left[\dfrac{r^c}{c!} (K-c+1) + \sum_{n=0}^{c-1} \dfrac{r^n}{n!} \right]^{-1}, & \rho = 1 \end{cases}$$

Note: Both the series in the computation of p_0 are finite and thus there is no requirement on ρ. So, I have the complete solution of steady-state probabilities.

- One can show that taking the limit as $K \to \infty$ with the restriction that $\rho = \frac{\lambda}{c\mu} < 1$ yields the $M/M/c$ results. You can try a few of these results whenever we fully develop. Of course, currently, you have this steady-state distribution themselves. You can see that what happens to that when you have with restriction $\rho < 1$ what happens to $\left[\dfrac{r^c}{c!} \left(\dfrac{1-\rho^{K-c+1}}{1-\rho} \right) \right]$ whether it leads to the same expression as you had it for an $M/M/c$. Similarly for other expressions also like you can try the same thing by writing this. That is one exercise that you can carry out and see.

- If we take $c = 1$, we get results for an $M/M/1/K$ model. Many results gets simplified in this case, but this is left as an exercise, for all the results that we develop for $M/M/c/K$.

- Expected queue length (for $\rho \neq 1$):

$$\begin{aligned} L_q &= \sum_{n=c+1}^{K} (n-c) p_n = \frac{p_0 r^c}{c!} \sum_{n=c+1}^{K} (n-c) \frac{r^{n-c}}{c^{n-c}}, \quad \{\text{by putting the value of } p_n\} \\ &= \frac{p_0 r^c \rho}{c!} \sum_{i=1}^{K-c} i\rho^{i-1} = \frac{p_0 r^c \rho}{c!} \frac{d}{d\rho} \left(\sum_{i=0}^{K-c} \rho^i \right) = \frac{p_0 r^c \rho}{c!} \frac{d}{d\rho} \left(\frac{1-\rho^{K-c+1}}{1-\rho} \right) \\ &= \frac{p_0 r^c \rho}{c!(1-\rho)^2} \left[1 - \rho^{K-c+1} - (1-\rho)(K-c+1)\rho^{K-c} \right] \end{aligned}$$

  For $\rho = 1$, L_q can be obtained similarly.

- For this finite capacity waiting space model, $L = L_q + r$ is not true, since a fraction p_K of arrivals do not join the system (when there is no waiting space).

- Effective arrival rate seen by the servers is $\lambda_{eff} = \lambda(1 - p_K)$ (by PASTA property).

So, this is the word that will be used; from now onwards when anything when you are looking at the actual input which goes into service. In the earlier model, that was not a problem because every arrival was going to the service after waiting in the queue, but here it is not the case. So, that means this quantity which we call an effective arrival rate, is actually the one that determines the arrivals that go into the service and gets serviced by the server. So we will be using $\lambda_{eff} = \lambda(1 - p_K)$, because p_K is the proportion of arrivals that are turned away, and the $\lambda(1 - p_K)$ the rate of arrivals. So, this is out of those lambdas; then only this much proportion actually gets into the service.

So, this is out of those lambdas; then only $\lambda(1 - p_K)$ actually gets into the service. So, this will be again; this is the effective rate seen by the servers, and because of the pasta property here, because the arrival see time averages, and then the time average is p_K is what is the proportion that will not be entering into the system and $1 - p_K$ is. So, that will also be the arrival rate seen by the servers. And this would actually be an effective arrival rate. So, in this r, whatever we had earlier as a λ, now we have to modify that λ with this λ_{eff}, and we have to derive the results; that is what we have to do.

- The relation between L and L_q for this model is modified as

$$L = L_q + \frac{\lambda_{eff}}{\mu} = L_q + \frac{\lambda(1 - p_K)}{\mu} = L_q + r(1 - p_K)$$

- By Little's law

$$W = \frac{L}{\lambda_{eff}} = \frac{L}{\lambda(1 - p_K)}$$
$$W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda_{eff}}$$

- In such finite capacity queues, an important performance measure is p_K, the probability of blocking (or the proportion of arrivals turned away as they could not enter the system).

For example, when you are trying to model the mobile networks, where if you have to describe a model like you have base stations and you have that will handle a particular call, and then the customer who may earn a call is moving from one place to the other. So, the call gets transferred from one base station to the other base station. Then this is what is handoff calls, and new calls are originating from that center. So, all these quality of parameters when you are establishing a network want that there is a certain up value is given, and the probability of blocking, the blocking probabilities for handoff calls for the new call should be below this. So, that is the standard that you want to maintain. Then only when you actually make a call; you will be able to connect it; rather, you will not get a busy tone or lines are not available; this kind of message that you will be able to avoid. So, that kind of parameter specifications, basically, what is given is given in terms of blocking probabilities. So, this blocking probability is effectively in this particular case is essentially p_K, but here we are assuming that there is a queueing component also, not just that you are getting into that. So, you may not be able to connect, but you can be queued. But really, this will make sense when we see the next model, but nevertheless, this is also one of the quantities that we always look at when we analyze the model, which is called the blocking probabilities, which gives us an idea about the proportion of arrivals turned away as they could not enter the system, so, you can increase the capacity if you want to reduce that, you have to meet certain requirements. So, to study that, you will need this p K because this is a new feature that we have introduced. So, this is a new quantity that will come. So, as we introduce more and more features, there may be certain quantities associated with that feature that will come as performance measures, and one can analyze that in terms of, say, model parameters.

So, of course, if you want to analyze this p_K alone, you go back to your this one $p_K = \frac{\lambda^K}{c^{K-c}c!\mu^K}p_0$,, now this p_K alone you can analyze it as a function of λ, μ and how much if you have to increase you say you know all these things server or K or whatever is the case. So, these are all parameters that one can analyze. So, that one can do for every performance measure that we develop, but we will not do, but of course, that can also be done, and that is what performance analysis is all about. Now, as we said that when there is $c = 1$ with a single server, so, the results are for the special case of $M/M/1/K$.

- For $M/M/1/K$, the expressions are simpler as given below.

$$p_0 = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{1}{K+1}, & \rho = 1 \end{cases}$$

$$p_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{1}{K+1}, & \rho = 1 \end{cases}, n \geq 0$$

$$L_q = \begin{cases} \frac{\rho}{1-\rho} - \frac{\rho(K\rho^K+1)}{1-\rho^{K+1}}, & (\rho \neq 1) \\ \frac{K(K-1)}{2(K+1)}, & (\rho = 1) \end{cases}$$

$$L = L_q + (1 - p_0)$$

▶ Note: Here, $L = L_q + (1 - p_0) \implies \mu(1 - p_0) = \lambda(1 - p_K)$
(that is, the system's effective output rate equals its effective input rate)

We said that $\lambda(1 - p_K)$ is the effective arrival rate; $\mu(1 - p_0)$ is the effective departure rate you can call because when the departure happens, it will be of these things. So, that is what you would see here in this particular situation. Now what we will do next is that we will try to do the waiting time distribution, which is a bit complicated here because of the finiteness, but for which now what we have to observe one important thing is that this part that we have we have to look at the arrival point probabilities or the arrival time probabilities which are what earlier we have considered as a_n. In earlier cases, a_n's were equal to p_n's. So, we were happily using it now because of this finiteness because of this blocking that is happening.

- It is necessary to derive the arrival-point probabilities $\{a_n\}$, since the input is no longer Poisson because of size truncation at K, and $a_n \neq p_n$.

$$\begin{aligned} a_n &= P\left\{N = n \text{ in the system} \mid \text{arrival about to occur}\right\} \\ &= \frac{P\left\{\text{arrival about to occur} \mid N = n\right\} p_n}{\sum_{n=0}^{K} P\left\{\text{arrival about to occur} \mid N = n \text{ in system}\right\} p_n} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{\left[\lambda \Delta t + o(\Delta t)\right] p_n}{\sum_{n=0}^{K-1} \left[\lambda \Delta t + o(\Delta t)\right] p_n} \right\} = \frac{\lambda p_n}{\lambda \sum_{n=0}^{K-1} p_n} \\ \Rightarrow\ a_n &= \frac{p_n}{1 - p_K}, \quad n \leq K - 1 \end{aligned}$$

So, now, you see here a situation where $a_n \neq p_n$ not because the arrival process is not Poisson, but because of the truncation that happens at K for this process that this is actually the arrival point probabilities and which is again you

can express in terms of p_n, but with normalized by this $1 - p_K$ is what then will give you the arrival point probabilities. So, this is the difference that you need to keep in mind because we are using this, as you have seen in the earlier model that the arrival point probabilities we are using it in the analysis of waiting time distribution so, that we have to keep that in mind here. Another thing which is a fact which can be used in different places, later on also we will be using it. So, mainly what we want is that $\int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x} dx$ is equal to $\sum_{i=0}^m \frac{(\lambda t)^i e^{-\lambda t}}{i!}$ is what we need it. But what is this integral quantity we can give an interpretation.

- For a Poisson arrival process with rate λ, we have that

$$\begin{aligned} P\{N(t) \leq m\} &= P\{\text{sum of } m+1 \text{ interarrival times } > t\} \\ &= \int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x} dx \\ &= \sum_{i=0}^m \frac{(\lambda t)^i e^{-\lambda t}}{i!}. \end{aligned}$$

  The working:

$$\begin{aligned} \int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x} dx &= \int_0^\infty \frac{\lambda^{m+1}(u+t)^m}{m!} e^{-\lambda t} e^{-\lambda u} du = \int_0^\infty \frac{\lambda^{m+1} e^{-\lambda t} e^{-\lambda u}}{m!} \sum_{i=0}^m \binom{m}{i} u^{m-i} t^i du \\ &= \sum_{i=0}^m \frac{\lambda^{m+1} e^{-\lambda t} t^i}{i!(m-i)!} \int_0^\infty e^{-\lambda u} u^{m-i} du = \sum_{i=0}^m \frac{\lambda^{m+1} e^{-\lambda t} t^i}{i!(m-i)!} \frac{(m-i)!}{\lambda^{m-i+1}} \\ &= \sum_{i=0}^m \frac{(\lambda t)^i e^{-\lambda t}}{i!}. \end{aligned}$$

So, this is simple; the interpretation of $\int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x} dx$ is exactly $\sum_{i=0}^m \frac{(\lambda t)^i e^{-\lambda t}}{i!}$, which is exactly $P\{N(t) \leq m\}$ as here, but what we want is $\int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x} dx = \sum_{i=0}^m \frac{(\lambda t)^i e^{-\lambda t}}{i!}$ is what we will need it. So, that is what we are going to use in our analysis of waiting time distribution.

- The CDF $F_{T_q}(t)$ for the line delays can be obtained as earlier:

$$\begin{aligned} F_{T_q}(t) &= P\left\{T_q \leq t\right\} \\ &= F_{T_q}(0) + \sum_{n=c}^{K-1} P\left\{n - c + 1 \text{ service completions in} \leq t \mid \text{arrival finds } n \text{ in system}\right\} \cdot a_n \\ &= F_{T_q}(0) + \sum_{n=c}^{K-1} a_n \int_0^t \frac{c\mu(c\mu x)^{(n-c)}}{(n-c)!} e^{-c\mu x} dx \\ &= F_{T_q}(0) + \sum_{n=c}^{K-1} a_n \left\{1 - \int_t^\infty \frac{c\mu(c\mu x)^{(n-c)}}{(n-c)!} e^{-c\mu x} dx\right\} \end{aligned}$$

  Using the fact we have seen just earlier (with $\lambda = c\mu$ and $m = n - c$), we get

$$\begin{aligned} F_{T_q}(t) &= F_{T_q}(0) + \sum_{n=c}^{K-1} a_n - \sum_{n=c}^{K-1} a_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!} \\ &= 1 - \sum_{n=c}^{K-1} a_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!} \end{aligned}$$

Example.

Again if you want to see what is the probability of delay, what is the probability of no delay, or any other quantities which we have done for $M/M/c$. Here also you can talk about it, there is no harm, or there is no problem with that, but anyway see this will be a repetition you will not do that actually; it will be similar the case. Now, let us look at a simple example to see how this certain results can be used to analyze this. Exactly similar to what we have seen earlier, but just that we want to have some feel of how these examples work in our case.

A vehicle pollution testing center has three inspection stalls (each with room for only one car).

The station can accommodate at most four cars (seven in the system).

The arrival pattern is Poisson with a mean of one car every minute during the peak periods.

The service time is exponential with mean 6 minutes.

From the data, we have $\lambda = 1$ and $\mu = 1/6$ (with time in minutes).

And, the system is $M/M/3/7$ with $r = 6$ and $\rho = 2$.

We find that $p_0 = \frac{1}{1141} = 0.00088$.

$L_q = \frac{3528}{1141} = 3.09$ *cars and* $L = L_q + r(1 - p_K) = \frac{9606}{1141} = 6.06$ *cars.*

And, $W = \frac{L}{\lambda_{eff}} = 12.3$ minutes.

The expected number of cars per hour that cannot enter the station is $60\lambda p_K = 60p_7 = 30.4$ cars per hour.

Some serious thinking is required! (50% cars not able to enter the station.)

So, this gives us an idea that ok there is there need to be some serious thinking on the way this testing center has been functioning because, in an hour on an average, 60 car comes, one per minute, but out of which you see 30 cars are being turned away. So, even with 4 waiting spaces, now what do you do? You want to increase the inspection stalls, or you want to increase the waiting space, suppose if we increase the waiting space, we can think about it. Suppose if we increase the waiting space from 4 cars to say 6 cars or 7 cars, then what will be the proportion of cars that cannot enter into the system one you can think. And in that case, what will be the effect of this L_q, L, and everything? Now suppose if you feel that this itself is not acceptable that someone comes and he waits for the total time that is required is this W is 12 minutes, which means 6 minutes on an average is what is the service time, but more than 6 minutes he waits in whether that is acceptable, you can think.

If that is not acceptable, you can think of putting up one more inspection stall like that. So, which direction do you want to go? So, you have to analyze. So, once you get hold of these measures, then increasing this is what theoretically you can do. This is you do not need actually to create one stall and to see how much the performance is. So, with this understanding, you can get an idea of; suppose if I increase to 4 inspection stall what would be the effect of in all this L_q, p_0, W and this quantity how much are being turned away the block blocked cars what the proportion of the blocked cars is.

Or if I increase the waiting space from 7 to say 10, then what is the effect of all these things you can analyze and decide accordingly what kind of decision that you want to make. Again we keep saying this is the this is how this had to be utilized. Our objective is to obtain this, which is not very complex so far, whatever we have encountered. But how it is to be used is in this manner. So, most of the things like practically, if you want to utilize it, you have to look at the little closely the examples and see how this information can be utilized to improve the system further.

So, this is a simple example of an $M/M/c/K$ model. So, this is another model which is again based on the birth-death process. So, it is a birth-death queueing system., but that feature that we have added here is the finite capacity queues, and then we have seen how one can do the analysis. We have not done the complete analysis as you see as I said, but you know what are the important ones that we have considered with respect to this, but other aspects like we have done for $M/M/c$ or $M/M/1$ model. Similar analysis can also be done for these kinds of models, but we are not going to do it. We will just what is the model and what is the performance measure that we can derive, and we will try to see how it can be utilized in this; it will give you an idea on that.
So, we will restrict ourselves to this kind of behavior only. So, with that, we will end our discussion on this finite capacity queueing model. We will come back with more of BDP in the following lectures.
Thank you, bye.