Introduction to Queueing Theory

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

Lecture - 15
M/M/c Queues, Erlang Delay Formula

Hi and hello, everyone; after having seen the analysis of what one does with an $M/M/1$ Queueing System, we will now move to our next queueing model that we consider, which is the $M/M/c$ Queues.

- Similar to $M/M/1$ queue except that there are now c servers with each server having an IID $Exp(\mu)$ distributed service times.

  Now, since this is also like exactly similar like $M/M/1$ queue, $M/M/c$ queue can also be modelled by a birth-death process because at any point of time, the state space is going to change only by 1, either by an arrival or by a departure. Since the time is continuous simultaneous departure, the probability is 0; that is how you know things are the time is spanned out in a way.

- So, this queue can be modelled as a BDP with rates, since arrivals happen one at a time; according to a Poisson process, there is no dependency; it does not depend on the number of customers in the state. So, the arrival rate would remain

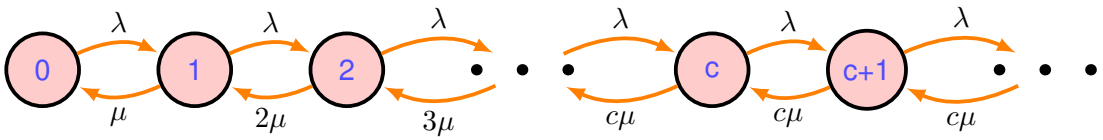
$$\lambda_n = \lambda, \quad \forall n \geq 0 \text{ like in } M/M/1 \text{ queue}$$

  But now, since there are c servers here; so, the service rate or the death rates we have to determine. Now, if you look at when there are c servers, suppose if there are n customers where $n \leq c$, then what happens? Suppose assume that $n \leq c$; then there are n customers, so; that means, what? Suppose if there are say n is 10 and c is 20. So, then there are only 10 servers that are giving service to the customers; the remaining 10 servers are idle. So, now what is the rate of departure now? Because 10 servers are working independently of each other, now what is that that you are looking at? A departure will happen according to what? If you look at little closely like you will see that, out of these 10 servers who are giving service currently, any one of them can finish the service. So, when any one of them finishes the service, then there is a departure happens. So, what is the rate of departure? Then the time that you have for these 10 servers, the minimum of these 10 servers' service times, service time distributions. Each one of them is exponential. So, the minimum is also exponential. So, the minimum time that will be required for a departure to happen would be 10 times mu.

$$\mu_n = \begin{cases} n\mu, & 1 \leq n < c \\ c\mu, & n \geq c \end{cases}$$

So, that is what will be the death rate for this BDP, which is what we are trying to model using this BDP for this $M/M/c$ queue.

- The transition rate diagram for $M/M/c$ queues



Now, once we have these birth and death rates, our objective is to obtain the steady-state distribution for the number of customers in the system at equilibrium.

- Using the prior theory developed earlier for BDPs (the iterative method), we insert the values for λ_n and μ_n to obtain the steady-state probabilities p_n:

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

Now, to find p naught; how do we find p naught?

- To find p_0, we use the condition $\sum_{n=0}^{\infty} p_n = 1$, which gives

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{\infty} \frac{\lambda^n}{c^{n-c} c!\mu^n} \right)^{-1} = \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c} c!} \right)^{-1} \quad \left\{ \text{here } r = \frac{\lambda}{\mu}, \rho = \frac{r}{c} = \frac{\lambda}{c\mu} \right\}$$

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!} \frac{1}{1-\rho} \right)^{-1}, \qquad \frac{r}{c} = \rho < 1.$$

$$\left[\text{using } \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c} c!} = \frac{r^c}{c!} \sum_{n=c}^{\infty} \left(\frac{r}{c} \right)^{n-c} = \frac{r^c}{c!} \sum_{m=0}^{\infty} \left(\frac{r}{c} \right)^m = \frac{r^c}{c!} \frac{1}{1-\rho}, \quad \rho < 1 \right]$$

Now, what can we do? We can try to obtain the performance measures. So, the first one that we will consider is this quantity of L_q, which is here; it is easier rather than obtaining L first. L_q is easier because then you will need to deal with only one summation. So, in any model this judgment, you have to make like which quantity you will tackle out of these four primary quantities that you have L, Lq, W, W_q. So, accordingly, you handle that and then get one, and then you can get the remaining three.

- Expected queue size L_q

$$L_q = \sum_{n=c+1}^{\infty} (n-c)p_n = \sum_{n=c+1}^{\infty} (n-c)\frac{r^n}{c^{n-1}c!}p_0$$
$$= \frac{r^c \rho p_0}{c!} \sum_{n=c+1}^{\infty} (n-c)\,\rho^{n-c-1} = \frac{r^c \rho p_0}{c!} \sum_{m=1}^{\infty} m\rho^{m-1} = \frac{r^c \rho p_0}{c!}\frac{1}{(1-\rho)^2}$$
$$i.e.,\ L_q = \left(\frac{r^c \rho}{c!(1-\rho)^2}\right) p_0$$

- Employ Little's formula to get W_q, then use W_q to find $W = W_q + 1/\mu$, and finally employ Little's formula again to calculate $L = \lambda W$.

$$W_q = \frac{L_q}{\lambda} = \left(\frac{r^c}{c!(c\mu)(1-\rho)^2}\right) p_0$$
$$W = \frac{1}{\mu} + W_q = \frac{1}{\mu} + \left(\frac{r^c}{c!(c\mu)(1-\rho)^2}\right) p_0$$
$$\text{and } L = \lambda W = r + \left(\frac{r^c \rho}{c!(1-\rho)^2}\right) p_0 \qquad (\text{Note: } L = L_q + r \text{ for any } G/G/c)$$

Now, again if you want to see as a parameter of ρ how this behaves, as a parameter of c how it behaves, you can analyze the quantities of interests depending upon your requirement. So, those things you can do.

- Consider $F_{T_q}(0) = P\{\text{a customer has zero delay in queue before receiving service}\}$ and $1 - F_{T_q}(0) = P\{\text{a customer has a}$

- As in the case of $M/M/1$ queue,

$$F_{T_q}(0) = P\left\{T_q = 0\right\} = P\left\{N \le c-1 \text{ in system}\right\} = \sum_{n=0}^{c-1} p_n = p_0 \sum_{n=0}^{c-1} \frac{r^n}{n!} = 1 - \frac{r^c p_0}{c!(1-\rho)}$$
$$\left[\text{Using } \sum_{n=0}^{c-1} \frac{r^n}{n!} = \frac{1}{p_0} - \frac{r^c}{c!(1-\rho)}, \text{from the expression for } p_0\right]$$

- The probability that a customer has a nonzero delay in queue

$$C(c,r) = 1 - F_{T_q}(0) = \frac{r^c}{c!(1-\rho)} \Bigg/ \left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right) = \frac{p_c}{1-\rho}$$

This is called the Erlang-C formula (or the Erlang second formula or the Erlang delay formula).

All three are one and the same as far as we are concerned; we can use any of these three interchangeably "Erlang-C formula or Erlang second formula, or Erlang delay formula." Erlang delay formula makes much more sense directly when you use this because what is the meaning? That what is the probability of delay or Erlang delay probability sometimes, that is also it is being referred to this. So, why is this? Because this is one which has been found by Erlang himself and used till today. Because in a multiserver system, if you want to compute what is the probability of

delay, what is the probability that arriving customer has to wait, then what you imply is the Erlang-C formula or the Erlang delay formula. So, this is what is Erlang delay formula; is very, very important; as we said, when Erlang did the analysis, he found that the $M/M/c$ model fits well into the situation of multiserver Markovian arrival or Poisson process arrival and exponential service time model fits well in most situations, and that is true till today. In most of the communication systems, when you want to do the analysis, what you will use is this Erlang delay formula to get that in the multiserver situation. So, that is what is relevant even today. And $C(c, r)$ is what is being used, and that is why it is what we call an important one, called the Erlang delay formula in this case, which is basically the probability of delay in a multiserver system that is what it is. So, this is an important quantity, as we said. So, that is what $C(c, r)$ is. Now, let us look at a little deeper into this Erlang-C; there is much more you can do, but we will just hint a bit on that.

- The Erlang-C formula is an important measure of congestion used in many situations (e.g. call centers).

  So, many times, the quality of service has to be specified in terms of certain parameters or certain requirements that you pose. So, the requirement in a particular situation, say, for example, in a call center. Imagine a call center situation where there are multiple agents who are working on it, and the calls arrive according to the Poisson process; each agent will take an exponential amount of time to serve one particular customer. And there are multiple agents. So, now, for this, if you want to design a system, if you want to design a call center, if your requirement is that if you want to bound the probability of delay for an incoming call, it means the incoming call has to wait. So, this basically like depends on the congestion that you experience on that call center network at that point of time. So, you want to do that analysis. So, this Erlang-C formula is what is a measure of congestion used in such a situation.

  ♦ This gives the probability that a customer is not able to immediately access a server.

  So, in a call center situation, you are not able to talk to the agent immediately. You are put on the wait, and some music is played, or some announcement is made; these are other things that is going on, but you are waiting essentially.

Example. *[Call Center - Erlang-C Formula]*
Assume that for a call center, an $M/M/c$ model is applicable with $\lambda = 30, \mu = 12, c = 3$.
We then have $r = 2.5$ and $\rho = 5/6$. From the Erlang-C formula, we get $C(c, r) = 0.702$.
$\Longrightarrow$ *The probability of no delay is* 0.298.

Suppose you wish that not more than 10% of the callers experience positive delay. This means that you want $C(c, r) \leq 0.10$ and to achieve this, you would need c to be at least 6.

So, this c will be 6; you need 6 servers or 6 agents you need in order that this is true. So, this is what we call the quality of service parameter. So, then you can say that the quality of service parameter here; is a measure of congestion that not more than 10% of the callers will experience delays when connecting to the agents. So, now, how many c you need is this is what it is we can see. So this helps in the analysis of deciding the number of servers.

- The above example is a typical example of the problem of determining of an appropriate number of servers according to a criteria.

So, this is again a large class of analysis that one can do with respect to determining the appropriate, in what sense? So, you have to determine under what criteria you are looking at it.

- The approach of choosing c according to a measure of congestion (Erlang-C formula is one) is often referred to as 'quality and efficiency domain' (QED).
  Under the QED regime, an approximate formula is $c \approx r + \beta\sqrt{r}$, where r is the offered load and β is a constant related to $C(c, r)$.

  Suppose, if this 0.10 if I call it α, there is a formula that connects α and β. So, this is a constant which is connected to that r. So, what this tells roughly is that, forget about this β, you see if r is your offered load. So, here r is say, 2.5 say for example. Now, you add an additional $\sqrt{r}$ server along with multiplication with some β to be under QED domain.

  So, this is what is often called the square root rule or in the call center thing that how many staffs you need is essentially square root staffing rule.

So, determining the number of servers is important in many contexts. So, even in this case, when there are call taxis, or there like these kinds of situations where how many drivers you need or how many cars you need depending upon what is your server in this particular case. So, all these things would require these kinds of ideas, of course; there are much more sophisticated rules are there, but this is one very simple rule which is an approximation that works quite good in a different situation; that is what you would see. So, this kind of thing, the Erlang-C formula, helps us to get immediately certain managerial insights as to what to do to improve the system.

Now, having done that, $F_{T_q}(0)$ or $1 - F_{T_q}(0)$; now, let us look at the delay time distribution.

- Similar to the case of $M/M/1$, we determine the probability distribution of the waiting times T_q assuming FCFS and using PASTA property. For $T_q > 0$, the CDF is

$$F_{T_q}(t) = F_{T_q}(0) + \sum_{n=c}^{\infty} P\left\{n - c + 1 \text{ completions in } \leq t \mid \text{arrival found } n \text{ in system}\right\} \cdot p_n$$

  When $n \geq c$, the system output is Poisson with rate $c\mu$ and the distribution of the time for the $n - c + 1$ completions is Gamma.

$$\begin{aligned}
F_{T_q}(t) &= F_{T_q}(0) + p_0 \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c}c!} \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x} dx \\
&= F_{T_q}(0) + \frac{r^c p_0}{(c-1)!} \int_0^t \mu e^{-c\mu x} \sum_{n=c}^{\infty} \frac{(\mu r x)^{n-c}}{(n-c)!} dx \\
&= F_{T_q}(0) + \frac{r^c p_0}{(c-1)!(c-r)} \int_0^t \mu(c-r) e^{-\mu(c-r)x} dx \\
&= F_{T_q}(0) + \frac{r^c p_0}{c!(1-\rho)} \left(1 - e^{-(c\mu-\lambda)t}\right) \\
&= 1 - \frac{r^c p_0}{c!(1-\rho)} e^{-(c\mu-\lambda)t}, \quad \text{by putting the value of } F_{T_q}(0)
\end{aligned}$$

- From the CDF of T_q,

$$P\left\{T_q > t\right\} = 1 - F_{T_q}(t) = \frac{r^c p_0}{c!(1-\rho)} e^{-(c\mu-\lambda)t}$$

$$\text{and therefore } \; P\left\{T_q > t | T_q > 0\right\} = e^{-(c\mu-\lambda)t}$$

- Verify the formula for the mean line delay W_q:

$$W_q = E(T_q) = \int_0^\infty \left(1 - F_{T_q}(t)\right) dt = \left(\frac{r^c}{c!(c\mu)(1-\rho)^2}\right) p_0$$

Now, the other one is the "sojourn time distribution" that we want.

- To find the formula for CDF of waiting time (T) in the $M/M/c$ system, we first split the situation into two separate possibilities,

  ▶ Customers having no wait in queue (occurring with probability $F_{T_q}(0)$) .

  $\implies$ System time = Time in service = $Exp(\mu)$.

  ▶ In the other group of customers, they have a positive wait in the queue (occurring with probability $1 - F_{T_q}(0)$).

  $\implies$ Time in system = Wait in queue + Time in service = $Exp(c\mu - \lambda) + Exp(\mu)$

  Result.

  If X_1, X_2 are independent random variables and $X_1 \sim Exp(\lambda_1), X_2 \sim Exp(\lambda_2)$ then

$$f_{X_1+X_2}(x) = \lambda_1\lambda_2 \left[\frac{e^{-\lambda_2 x} - e^{-\lambda_1 x}}{\lambda_1 - \lambda_2}\right]$$

We have given it in that particular form as hypo-exponential like you can go and check that is what is the case here, or if you want to write in terms of transforms for such quantities, you can simply take the transform of λ_1, which is say Laplace transform. So, $\frac{\lambda_1}{s+\lambda_1}$ and this is $\frac{\lambda_2}{s+\lambda_2}$. The Laplace transform of $X_1 + X_2$ is simply the product of these two Laplace transforms. So, that is very easier to understand and see. So, this sum nevertheless is given by this expression here which is what is the density. So, how can one handle you can recall which, we need to obtain $Exp(c\mu - \lambda) + Exp(\mu)$. So, this is one exponential with one parameter; this is another exponential with a different parameter. So, to obtain this, this is in the second class now.

- The required CDF for the second class of customers:

$$P\left\{T \le t | T_q > 0\right\} = \frac{c(1-\rho)}{c(1-\rho)-1}\left(1 - e^{-\mu t}\right) - \frac{1}{c(1-\rho)-1}(1 - e^{-(c\mu-\lambda)t})$$

- Putting together, we get the overall CDF of the $M/M/c$ system waiting time T as:

$$\begin{aligned} F_T(t) &= F_{T_q}(0)[1 - e^{\mu t}] + [1 - F_{T_q}(0)] \times \left[\frac{c(1-\rho)}{c(1-\rho)-1}\left(1 - e^{-\mu t}\right) - \frac{1}{c(1-\rho)-1}(1 - e^{-(c\mu-\lambda)t})\right] \\ &= \frac{c(1-\rho) - F_{T_q}(0)}{c(1-\rho)-1}\left(1 - e^{-\mu t}\right) - \frac{1 - F_{T_q}(0)}{c(1-\rho)-1}\left(1 - e^{-(c\mu-\lambda)t}\right) \end{aligned}$$

So, there could be some other quantities that are also of interest with $M/M/c$; it does not matter, but we might require; these are certain fundamental performance measures that you obtain for the $M/M/c$ model here. Now, let us take an example and see what kind of questions can arise in the case of an $M/M/c$ model.

Example.

A hospital system is modelled through an $M/M/c$ system with $\lambda = 6, \mu = 3, c = 3$.

The hospital planners are interested in knowing

- the average number of patients waiting (L_q)
- the average amount of time a patient spends in the hospital (W)
- the average percentage of idle time of each of the doctors ($1 - \rho$)

From the data, we have $r = 2, \rho = 2/3$ and $p_0 = 1/9$.

From the formulae, we compute $L_q = 8/9$ and $W = \frac{1}{\mu} + \frac{L_q}{\lambda} = \frac{13}{27}$ hours (28.9 min.)

The long-term average fraction of idle time for any server equals $1 - \rho = 1/3$. Given that there are three doctors on duty, two of them will by busy at any time (on average), since $r = 2$.

Also, the fraction of time that there is at least one idle doctor can be computed here as $p_0 + p_1 + p_2 = P\{T_q = 0\} = 5/9$.

Suppose if you want at least two doctors to be free, at least one doctor free; then all these things you can compute from this distribution from the stationary system size distributions that you can do. So, again you see that there may be some questions, there may be some more questions with respect to this, but if that can be answered through whatever we have obtained, we can do. But if there is some other thing that is you want, suppose if you want to know what is the duration of the time the servers are busy; all servers are busy, but at least one servers are busy, then you need some more analysis some more measure that you obtained which is again it is not very simple one they are all little complex ones that you obtain. So, these are the very basic ones that we are dealing with, and this is the analysis of the $M/M/c$ model; as you see, it mirrors most of the time $M/M/1$ models with additional features like c server. So, the question is the number of c servers we can determine and certain performance measures that you try to get with respect to the additional features inside these models. So, we will continue our discussion on this BDP different kinds of BDP models as we go further, at least some of some more models also we will deal with it. So, here we end the discussion on $M/M/c$. We will see another model in the next lecture.

Thank you bye.