Introduction to Queueing Theory

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

Lecture - 42
M/G/c, M/G/∞ and M/G/c/c Queues

Hi and hello, everyone. What we have seen so far in the Semi-Markovian Queueing System is basically the $M/G/1$ model and its finite version that, is $M/G/1/K$ model. What we have obtained was this equilibrium system size probabilities and waiting time distributions, which, if you put the service time distribution as exponential, would reduce to the corresponding $M/M/1$ model as expected. Of course, there are other variations that you can bring in, even with respect to that single server model itself, meaning that suppose if you want to add retrial or anything of that sort, impatience. So, those kinds of things can also be brought in, but we are not looking at that. What we will do next is that it is not a very detailed study, but it is an overview study of the same general service, but now we have the setup of multi-server queues. If we have multi-server queues and general service, we are still retaining the Poisson process arrivals as far as the arrival process is concerned. And then we just see how far or whether we can extend the ideas that we have gained with respect to the single server model in any way. But, if you look at little closely, what you would see is that we are basically starting with a disadvantage in this multi-server M/G model.

The reason is that these models, whether it is $M/G/c$ or $M/G/c/K$ models, they do not possess the embedded Markov chains that we have used in the case of among $M/G/1$ model. They and here, this multi-server version of them, do not possess such an embedded Markov chain in the useful sense. That is so because, as you can imagine, what can happen at departure points at departure epochs, what we had earlier? We had departure epochs; you can look at what was the system size at the previous departure epoch, and in between, how many arrivals have come during this one service time was sufficient enough to determine what is the departure system size departure point, system size probabilities at this departure epoch. So, in the multi-server case, this is not just sufficient because the number of arrivals during any interdeparture period is dependent on more than just the system size at the immediate departure predecessor; that is because of these multiple server cases. So, it poses, I mean, you have to expand much more in order to capture, bring it to your Markov chain basis. So, because of this reason, as you can imagine what might happen during this departure epoch and if you are looking for the next departure epoch, I mean there are multiple servers, how many servers are busy, which server completes the service, the next service is completed by which server. And, during that duration, like how many arrivals have come, you have to count, but this is not so easy as we are seeing it here.

Because here, the service time distribution is no longer exponential; if it is exponential, of course, you can be memoryless, you can use, and then you can get something, but here it is a general distribution. So, you need to keep track of all the c servers, how much time the service takes, the remaining service time, and which one will end next. And that number or that duration of the next departure where it is going to happen, now during that period how many has come? It is to say that it is easy to visualize, but the thing is to do it, it is not that easy. So, that is why we may not

be able to get some neat and clean results about the queue length distribution of such systems, but certain mean value results are easy to obtain. But, if you put certain restrictions and so on, there are multi-server systems that do possess some structure to get some kind of results that could be even complete results also it is possible to get. But, anyway, we will look at a few of them, and whatever is we can obtain, and we can easily give at this level is what then we will be concentrating on it. So, this is what we will do with respect to the multi-server M/G type queues. Let us take first this $M/G/c$ queues. As we just said, it is not that easy to get the departure point system size distributions, but what one can easily obtain is that the mean value formulas or the moments of these two quantities could be related in some way.

- Recall that, for $M/G/1$, we had

$$L_{(k)} = \lambda^k W_k.$$

that is what we said that the factorial moment of the system size distribution and the ordinary moment of the waiting time distribution was related by this relationship which we call as a generalization of Little's law.

Now, what we can do for an $M/G/c$ model, we can obtain a line version of this relationship. Remember, $L_{(k)} = \lambda^k W_k$ is for the system since there was a single-server; it was easier to obtain it for even the system. Of course, you can have a line version as well, but in the case of multi-server, it is very difficult to get the system version, but it is easier to get the line version. Because this you can count like when the next departure or how many it leaves behind and so on. So, that is what we are going to use. So, when we obtain now a queue version or the line version of $L_{(k)} = \lambda^k W_k$, this is for the number in the system; now, our number in the queue is what then we look at it. Because, multi-server again, you have to see which server, how many servers are occupied, which server will service complete the service that earlier also there. But, in a queue, if you consider as yow that it is first come first serve queue; so, the departure from the queue you are looking at it which means he is getting into the service. So, the next person will get. So, at least capturing that is easier. We know that

$$\pi_n = \text{P}\left\{n \text{ in the system just after the departure}\right\} = \frac{1}{n!}\int_0^\infty (\lambda t)^n e^{-\lambda t} dF_T(t).$$

This is valid in modified form for $M/G/c$ if we consider the quantities in terms of the queue and not the system. Then

$$\pi_n^q = \text{P}\left\{n \text{ in queue just after a departure }\right\} = \frac{1}{n!}\int_0^\infty (\lambda t)^n e^{-\lambda t} dF_{T_q}(t)$$

The mean queue length at departure points $L_q^{(D)}$ is given by

$$L_q^{(D)} = \sum_{n=1}^\infty n\pi_n^q = \int_0^\infty \lambda t dF_{T_q}(t) = \lambda W_q.$$

- Denote the k^{th} factorial moment of the departure-point queue size by $L_{q(k)}^{(D)}$. Then

$$\begin{aligned} L_{q(k)}^{(D)} &= \sum_{n=1}^\infty n(n-1)\ldots(n-k+1)\pi_n^q \\ &= \int_0^\infty dF_{T_q}(t) \sum_{n=1}^\infty \frac{n(n-1)\ldots(n-k+1)(\lambda t)^n e^{-\lambda t}}{n!}. \end{aligned}$$

The summand is the kth factorial moment of the Poisson and equal to $(\lambda t)^k$. Therefore

$$L_{q(k)}^{(D)} = \lambda^k W_{q,k}$$

where $W_{q,k}$ is the ordinary k^{th} moment of the line waiting time.

So, $L_q^{(D)} = \sum_{n=1}^{\infty} n\pi_n^q = \int_0^{\infty} \lambda t dF_{T_q}(t) = \lambda W_q$. is a special case, this is for $k = 1$ case, and here $L_{q(k)}^{(D)} = \lambda^k W_{q,k}$ is for any general k. So, what does it relate to? It is basically the relationship that exists between the kth factorial moment of the queue size and the kth ordinary moment of line delay or the queue delay, which is what it is. So, this is what you get the result for this $M/G/c$.

So, you see here $M/G/c$; you can still obtain a Little's law type of relationship with respect to queue length; that is what you can easily obtain in this particular case of the $M/G/c$ model. So, this is the first thing that we are looking at it. Again, we said that it is very difficult to get the system says probabilities, but the mean results in this particular case one can give. Of course, is it the only result that is available for the $M/G/c$ model? No, there are other results available, but those are all more complex, and they are beyond the level of what we have been doing. So, we will just highlight some of the results for some of these multi-server model cases; in the case of an M/G type queue, that is all we are trying to do.

Next, what we will consider is this $M/G/\infty$ model because we said something when we talked about the $M/M/\infty$ model. So, we will now try to relate that with respect to this $M/G/\infty$ model.

- So we will see two results for $M/G/\infty$ model
  ▶ The transient distribution for the number of customers in the system at time t
  ▶ The transient distribution for the number of customers who have completed service by time t, i.e., departure counting process.

- Let $N(t)$ be the overall system-size process, $Y(t)$ be the departure process, and $X(t) = Y(t) + N(t)$ be the input process or $N(t) = X(t) - Y(t)$ where, $X(t)$ is the input process, that is the relationship that you have between these three random variables. Now, you are looking at the overall system size probabilities for this $M/G/\infty$ model. So, we have Poisson arrival, general service time distribution, and an infinite number of servers is what we have; in this, we are looking at the number in the system. Then

$$P\{N(t) = n\} = \sum_{i=n}^{\infty} P\{N(t) = n | X(t) = i\} \frac{e^{-\lambda t}(\lambda t)^i}{i!}$$

- The probability that a customer who arrives at time x will still be present at time t is $1 - B(t - x)$, with $B(u)$ being the service-time CDF.
  Then, the probability that an arbitrary one of these customers is still in service is given by

$$q_t = \int_0^t P\{\text{service time} > t - x \mid \text{arrival at time } x\}\ P\{\text{arrival at } x\} dx$$

  Now what is this $P\{\text{arrival at } x\}$? So, given that there is arrival by time t, there is one arrival by time t, what is the probability that it would have arrived at time x? As you know, by the conditional arrival time property of this Poisson process that we have seen earlier, this is this arrival could have happened anywhere between $[0, t]$, according to a uniform random variable that is what that property is true. So that means $P\{\text{arrival at } x\}$ is simply $1/t$, and $P\{\text{service time} > \ t - x \mid \text{arrival at timex}\}$ is basically $[1 - B(t - x)]$

Using the conditional arrival time property of the Poisson process, we have

$$q_t = \frac{1}{t}\int_0^t [1 - B(t-x)]dx = \frac{1}{t}\int_0^t [1 - B(x)]dx$$

which is independent of any other arrival. Therefore, by the binomial law,

$$P\{N(t) = n | X(t) = i\} = \binom{i}{n} q_t^n (1-q_t)^{i-n}, \quad n \geq 0,$$

and the transient distribution is

$$\begin{aligned} P\{N(t) = n\} &= \sum_{i=n}^{\infty} \binom{i}{n} \frac{q_t^n (1-q_t)^{i-n} e^{-\lambda t}(\lambda t)^i}{i!} \\ &= \frac{(\lambda q_t t)^n e^{-\lambda t}}{n!} \sum_{i=n}^{\infty} \frac{[\lambda t(1-q_t)]^{i-n}}{(i-n)!} \\ &= \frac{(\lambda q_t t)^n e^{-\lambda t} e^{\lambda t - \lambda q_t t}}{n!} = \frac{(\lambda q_t t)^n e^{-\lambda q_t t}}{n!}, \end{aligned}$$

a nonhomogeneous Poisson with mean $\lambda q_t t$.

So, this is the number in the system that you have obtained. So, now you have seen here that the number in the system is given by this $\frac{(\lambda q_t t)^n e^{-\lambda q_t t}}{n!}$. Now, you can think, suppose if $B(x)$ were to be an exponential distribution, what would have been $\frac{1}{t}\int_0^t [1 - B(x)]dx$ and that is what would come here, and that is what we have derived it as the transient solution of $M/M/\infty$ model. So, now, for in general q_t also, this is what is true.

- The equilibrium solution is obtained by taking $t \to \infty$ to yield

  $$\lim_{t\to\infty} (\lambda q_t t) = \lambda \int_0^{\infty} [1 - B(x)]dx = \frac{\lambda}{\mu},$$

  and hence the equilibrium solution is Poisson with mean $\lambda E[S] = \frac{\lambda}{\mu}$.

  This is what we said it as the insensitivity to B, but now you are actually seeing it here; by this argument, you are shown that now this is actually the property you have proved it, now that is what you have seen. Now, this is about the number in the system; that is what one of the first two results we said we would look at for $M/G/\infty$. The first one is this; the second one is the distribution of the departure counting process, the number of service completion that have happened by time t.

- The distribution of the departure-counting process $Y(t)$ can be found using the same argument and by using $1 - q_t = \int_0^t B(x)dx/t$ instead of q_t. The result is

  $$P\{Y(t) = n\} = \frac{[\lambda(1-q_t)t]^n e^{-\lambda(1-q_t)t}}{n!}$$

- As $t \to \infty$, we see that $q_t \to 0$, and thus the interdeparture process is Poisson in the steady state, which is precisely the same as the arrival process.

- $M/M/c$ is the only $M/G/c$ with Poisson output with c must be finite.

So, this is basically what we have Burke's theorem; if you recall, it is basically giving that kind of result only. So, basically, $M/M/c$ is Poisson for any value of c; that is what we said there. So, then $M/M/c$ is the only $M/G/c$ with Poisson output with c must be finite is the case where the output process remains as the input process is what you are looking at here. So, this derivation is much like the previous derivation that we have obtained here; that is what you can see.

Now, that is about the $M/G/\infty$ model. Next, what we will see is what we call the loss system, but now the service duration for the last system is a generic distribution rather than an exponential distribution. Recall what we set for the $M/M/c/c$ model, which is also called Erlang's loss system. We obtained the steady-state distribution, which was a truncated Poisson, and we obtained what is called Erlang's loss formula or Erlang's B-formula. And, we said that this formula is insensitive to the service type distribution. And that is what; we might see now how exactly that is happening, but again this is a sketch; this is not a complete one. But, one has to look at the proper research papers for a complete one. So,

$$p_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!}, \quad 0 \leq n \leq c,$$

is what we obtained as the equilibrium steady-state size distribution for an $M/M/c/c$ model, and we said that it is valid for $M/G/c/c$ independent of the form of the service time distribution G.

- The specific value from this for p_c is called Erlang's loss or B-formula , and the result extends to $M/G/\infty$ (as seen above), where $p_n = e^{-\lambda/\mu}(\lambda/\mu)^n/n!$ for any form of G.
- Now let us sketch the proof of the general assertion for $M/G/c/c$.

It is not complete proof again, but anyway, we just want to see how exactly this is happening. So, this was the argument that we have given; of course, Erlang himself noticed that this is the case, $M/G/c/c$ also this is what is going to be. But, there was no proof until, say, 1956, when the complete proof was given.

- For $c = 1$ case is simple, $p_0 = 1 - \rho_{\text{eff}} = 1 - p_1$ for any $M/G/1/1$ queue.
  Since $\rho_{\text{eff}} = \dfrac{\lambda(1-p_1)}{\mu}$, the required result follows:

  $$p_1 = \frac{\lambda/\mu}{1+\lambda/\mu}, \quad p_0 = \frac{1}{1+\lambda/\mu}$$

  So, this is so trivial for the single-server case because this is anyway there are two states; it is very simple to find out here, and because of this result $p_0 = 1 - \rho_e f f$. So, it is very easy to see that this is true for the single-server case.

  Now, for $c > 1$ case, this formula, whatever which derived as the Erlang's loss formula, is basically one can derive by a combination of certain observations connecting the Markov process, reversibility, and the product from solutions; that is the idea. Now, we know that this system size here when you look at an $M/G/c/c$ model;

they are not Markovian by definition; that is what we know in this particular case of any G based model, M/G or G/M or whatever G/G or anything else that is not the system sizes are not Markovian by definition.

But then we need to define a Markov process. So, what one does is that one uses a certain number of here, basically c number of supplementary variables to denote or to define a Markov chain corresponding to this $M/G/c/c$ model. So, it is basically the supplementary variable technique that is being used here to extract a Markov process out of this non-Markov process. So, your original system state was simply the number in the system n, but since that is not a Markov process by definition.

► So expand the model state from n to the multidimensional vector $(n, u_1, u_2, \ldots, u_c)$, where $0 \leq u_1 \leq u_2 \leq \cdots \leq u_c$ are the c ordered service ages (i.e., completed service times so far, ranked smallest to largest, recognizing that $u_1, \ldots, u_{c-n}$ will be zero when the system is in n).

Now, this one what it tells you is that with respect to each server, if there is a customer who is getting currently being serviced by the corresponding server, what is the age of service is what is given here. Then, once you give what the completed service time is, then one can determine the remaining service time because the total service time is according to the distribution B. This is given for each one of these servers. Remember, in $M/G/1$ model; the supplementary variable was the technique to get the Markov process out of this semi Markov process; basically, you add the completed service time variable along with the number in the system; this pair was a Markov chain. So, here you need a $c+1$ dimensional Markov chain to describe the system, which is a Markovian.

► Here future state is clearly a function of only its current position, with change over infinitesimal intervals depending on the fact that the instantaneous probability of a service completion when the service age is u depends solely on u. So, hence the state vector together in $c+1$ dimension is a Markovian process; if you take n alone, it is not a Markov process; together, only we are talking about this being a Markov process.

So, this is what is the Markov process that you define for this $M/G/c/c$ model.

- Then, it turns out that the augmented process is reversible.
  ► A hint of this was evident earlier in $M/G/\infty$ where the output process is Poisson.

- The reversibility property implies that the limiting joint distribution of $(n, u_1, u_2, \ldots, u_c)$ has the proportional product form

$$p_n(u_1, u_2, \ldots, u_c) = C a_n \bar{B}(u_1)\bar{B}(u_2)\ldots\bar{B}(u_c)/n!,$$

where

► $u_1, \ldots, u_{c-n}$ are zero and C is proportional to the zero state probability,
► a_n is proportional to the probability that n servers are busy.
► $B(u_i)$ is the service-time distribution function [thus $\bar{B}(u)$ is the probability that a service time is at least equal to u].
► $n!$ accounts for all of the rearrangements possible for the n order statistics within the $\{u_1, u_2, \ldots, u_c\}$

corresponding to the n busy servers. So, because again, you are using a property, but anyway, we will not go into the details here.

So, this is what is the proportional form or product form solution is what we are much like your Jackson network is that you are handling it here is also, exactly those ideas are only used here also.

- Using the boundary conditions of the problem and the Poisson input assumption, we see that

$$\begin{aligned} C = p_n(0,0,\ldots,0) &= \lambda p_{n-1}(0,0,\ldots,0) = \lambda[\lambda p_{n-2}(0,0,\ldots,0)] \\ &= \cdots = \lambda^n p_0(0,0,\ldots,0) = \lambda^n p_0. \end{aligned}$$

- It follows that the product form result given earlier is a completely separable product form, since we know that $\mu\bar{B}(u)$ is a legitimate PDF (of the residual service time) and each term will independently integrate out to 1.

- Thus, the marginal system-size probability function is

$$p_n = p_0 \frac{(\lambda/\mu)^n}{n!}$$

It then follows that $p_0 = \left(\sum_{n=0}^{c} \frac{(\lambda/\mu)^n}{n!} \right)^{-1}$ and

$$p_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!}, \quad 0 \le n \le c.$$

Again, this is what we say; we are saying that as the equilibrium system size probabilities for the case of the $M/G/c/c$ model is what is given here $p_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!}, \quad 0 \le n \le c.$. Again, this is not a complete proof; this is a sketch of the proof; like how this is true, I mean whether what is the equations and how this satisfies this because one has to write the Chapman Kolmogorov equation of this Markov process to see all these things and then like how this is true and so on is a bit you have to look into, but that is why we said that this is a sketch. So, the idea is that you define a Markov process and see to it that it is reversible because you are observing that it is a reversible property that gives rise to your product from solution, where the product form is basically a product form of distributions themselves. And, hence you will obtain $p_n = p_0 \frac{(\lambda/\mu)^n}{n!}$, the number in the system, very easily like this. And that proves the equilibrium system size probabilities for the $M/G/c/c$ model are also the same as what one would obtain in the case of an $M/M/c/c$ model. And, hence the Erlang loss formula is also insensitive to the service time distribution is what you are seeing here. This is for the $M/G/c/c$ model; this is all multi-server models. Now, this is from $M/G/1$; we moved to $M/M/c$ models basically, but that c is infinity or c is finite like we are looking at the multi-server model.

- The steady-state probabilities for the $M/G/c/c$ are insensitive to the choice of G means that these probabilities will always satisfy the $M/M/c/c$ birth-death recursion

$$\lambda p_n = (n+1)\mu p_{n+1}, \quad n = 0, \ldots, c-1.$$

- It also turns out that the insensitivity to G can be retained when the arrival process is generalized to a state-dependent birth process (with rate λ_n).

Table 1: **Insensitivity results for $M/G/c/K$ models**

$M/G/c/c$ vs. $M/M/c/c$	Steady state probabilities and output process independent of form of G
$M/G/\infty$ vs. $M/M/\infty$	Steady state probabilities and output process independent of form of G
$M/G/c$ vs. $M/M/c$	Output process equal if and only if $G = M$.

Now, from $M/G/1$, what about the other direction that we have taken which is basically the bulk and other stuff? One can do, but again we are not going into the detail except that we will just highlight this particular queue which is $M^{[X]}/G/1$ queues, this bulk arrival queues.

- Apart from other assumptions, assume that the $\{c_j\}$ is the batch size distribution with PGF $C(z)$ and mean $E(X)$.
  ▶ The total arrivals constitute a compound Poisson process with PGF $e^{-\lambda[1-C(z)]}$.

  Then, with generally distributed service time, if you follow the similar line along what we have done for an M G 1 model, one can follow in a similar way.

- The PGF of the total number of arrivals during a service time of a customer is $K(z) = B^*(\lambda - \lambda C(z))$.

- Assume that the traffic intensity $\rho = \lambda E(X)/\mu < 1$.

- Then the PGF of the number in the system at departure epochs in steady state is then given by

$$\Pi(z) = \frac{(1-\rho)(1-z)B^*(\lambda - \lambda C(z))}{B^*(\lambda - \lambda C(z)) - z}$$

  which is the PK transform formula for $M^{[X]}/G/1$.

- There are results one can obtain in a relatively easier manner for the bulk-service queues too, much like the bulk-arrival queue above.

So, that is all we have for these semi-Markovian queues of the type M/G, meaning Poisson arrivals, but generally distributed service time. This is what we have seen so far; we will end this discussion on this M/G because one can go on with different variations being brought in, like you can introduce impatience and so on, one can go on, but we are not going to do that. So, we will stop here with respect to the discussion of these M/G-type models. Now, what we will take up next is basically when the service remains as Poisson, but the arrival process is general, arrival process then how one can analyze, that is a very briefly we will see that in the subsequent lectures.
Thank you, bye.