Introduction to Queueing Theory

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

Lecture - 38
M/G/1 Queues, The Pollaczek-Khinchin Mean Formula

Hi and hello, everyone; what we have studied so far are all Markovian queueing systems, as we have said. And then, what we will consider in next are what we called as semi Markovian queueing systems, for which a little bit of the background material required in terms of the underlying stochastic processes is what we have seen in the context of renewal processes and Semi Markov processes. Now with those ideas in mind, let us start our discussion of this semi Markovian queueing system, as we said. So far, whatever we have considered, they are all assumed in some way or other with these distributions, whether it is inter-arrival times or service time distributions, or any other type of distributions like retrial distributions and so on. The retrial duration distributions or anything of that was all either exponential or something that can be expressed in terms of exponential. And hence finally, it led to a model that is basically a continuous-time Markov chain. So, the Markovian analysis or Markov process analysis, like writing down the Chapman Kolmogorov equations, forward Kolmogorov equation, solution of that would give you the system size probabilities and so on like was employed starting from the balance equation identifying the Markov process. The balance equations can be written down immediately. And then, the solution will give you the solution to the queueing system from which the performance measures were obtained; this is what we have been doing here. But now, what we will do is relax this exponential assumption or anything exponential; we will not assume that. We can assume a general distribution that may be expressed as exponential; then it will become a Markov chain; if it cannot be expressed in terms of exponential distributions like an Erlang case or phase-type case, and so on, if it is not of that type then still it can be handled under this framework. And for many, the queueing theory actually the real power or the real utility or the real use comes from the analysis of models of semi Markovian type because this is what really, you are going beyond a simple Markov process analysis. So, the model becomes complex. So, you will not be able to explicitly obtain the distributions of the system sizes in many cases. And then the associated problems, whatever the complexities that are there with the analysis of systems, would also play a role. So, when we relax, we involve a general distribution, so the queues are no longer continuous-time Markov chain models, and the methodology or approach that we have used to analyze such systems is no longer applicable here. That is what will happen in such a scenario, but what we will do is that, for now, at least for the next few lectures. What we will do is we will relax the exponential assumption in a phase-by-phase manner, meaning that we will relax; for example, we are typically assuming that it is a simple type of queue where no other features are there, say $M/M/1$ type kind of. Then what we will do; we will relax the assumption of exponential only either in the arrival process or with respect to the service process. Meaning that either the inter-arrival time or service time distribution we will assume to be a general distribution that is not exponential or anything related to exponential; it could be general, but not both. So, that is the first step; as always, you go step by step.

So, what we will do first is relax this assumption only in one of those, either in the arrival process or in the service process. So, the resulting queueing models would be called semi Markovian queueing models because you still retain some amount of Markov nature within the structure. You still have on one side, either during the service time or during the inter-arrival time, you still have an exponential distribution; whether that is of some help, we will figure it out. So, because of that nature because we are relaxing in only one of them, so we have a semi Markov models. And within this context of this semi Markovian queues, we will come across what we call embedded Markov chains. We have already exhibited this embedded Markov chain even within the Markovian process in the beginning itself. What would be the because we defined CTMC in terms of these embedded Markov chains. And this embedded Markov chain is embedded within a continuous-time, but now a non-Markov process we have already seen in the semi Markov process how the Markov chain we are getting it as an embedded in a semi Markov process. And we can then employ some of the theories of Markov chains to analyze the queues; that is what we are going to do. And we are going to use this embedded Markov chain technique for our analysis, but there are other equally efficient methodologies available for the analysis of such a system. One such quantity is what is called supplementary variables techniques. This is also quite popular; this is a way of extracting a Markov process from a non-Markov process. You look at the non-Markov process, define what the additional random variables are, and make the dimension of the process a bit more, like from one dimension to two dimensions or three-dimension and so on. And then, in two dimensions, that will be a Markov process, so and then you analyze that. So, possibly complexity would increase, but the analysis can be done using Markov. So, basically, that is what supplementary variables you add; you supplement additional random variables or random processes to define. And then, together, they will form a Markov, and then you can analyze the Markov, so that is the basic idea of the supplement variable technique. And also, there are other methods like common material approaches, even the integral equation approach, which is applicable for a general; general model in both sides general. That kind of thing can also be, used for such analysis of such semi Markovian queueing systems. So, we will be considering this embedded Markov chain technique only. And note here is sometimes, some authors would call this an embedded Markov chain, but it is one and the same. So, we are using this spelling anyway; this is always a choice; even with queueing the spelling, there is a choice in the same way. So that is the background semi Markovian queueing systems. So, where the natural processes that arise are semi Markov processors, and hence we are calling this as semi Markovian queueing systems. And within that semi Markov process, we know that there is an embedded Markov chain. And using that, we will try to analyze and give an analysis of the entire queueing system using such kind of embedded Markov chain technique; that is what we are going to do.

In that analysis of the semi Markovian queues, the first and the most important queueing system in the whole of queueing theory, in terms of its simplicity, maybe $M/M/1$. But in terms of its importance, the most applicable or most useful queueing model is these $M/G/1$ queues. So, that is the model that we consider. So, what are the assumptions; it has all the assumptions of the $M/M/1$ model except that the service time distribution is now a general nonnegative distribution. I mean a general distribution, but which is with support in the nonnegative real line.

Since we are in the continuous-time model, that is what it says rather than the exponential, which is also a nonnegative random variable case. So, now we are picking any general distribution you can think about it; you need not even name it. Because in statistics and probability and statistics like there are plenty of distributions anything that comes up, you can take it as it is and try to analyze that is what is the advantage when you consider or give the theory for with keeping in mind G. Again, you have the arrival process is Poisson process which means the inter-arrival times are IID exponential. The service time distributions are IID, but they now follow a general distribution, but still, they are

IID. Now, if you count the number of service completions in $[0, t]$ in such a situation, then what you are going to get is a renewal process. So, that is why renewal service time, renewal service process is what renewal process is what is coming out as service process. Because we are just relaxing the exponential assumption, we are keeping all other things within that service process. So, that is what you would see here. And these two things, the arrival process, and service process, are independent, and you have an infinite capacity for queueing, and you are assuming a first come first serve queueing discipline. All these assumptions are there, like in an $M/M/1$ system in this $M/G/1$ system as well.

- Let λ be the arrival rate (of the Poisson process).
- Let S denote a (nonnegative) random service time with a general distribution and $\mu = \dfrac{1}{E[S]}$ be the service rate.

  So, $1/\mu$ is its mean like with the similarity with $M/M/1$ or in exponential case we will call $1/\mu$ as its mean not μ as the mean, so you remember that.
- The assumption $\rho = \dfrac{\lambda}{\mu} < 1$ is essential for our equilibrium analysis of the $M/G/1$ queue.

  So, the assumption that we need to make for the equilibrium analysis; we said that we are not going to look at the transient distribution, but we are looking at the equilibrium distributions or the steady-state analysis of the system, for which you need the stability condition which is basically $\rho = \frac{\lambda}{\mu} < 1$ which it can be proved later. We may not even do that, but it is quite easy. And this is easy to understand that you need such a condition even from the very simple intuitive idea that the arrival rate and service rate, with respect to that itself, you can see that this condition is necessary. So, this condition will be assumed to hold for the equilibrium analysis to hold.

  Now, this analysis of $M/G/1$ is a bit substantially in a way is difficult as compared to the simple $M/M/1$ model. The reason is the state description in the $M/M/1$ system; when you want to analyze an $M/M/1$ system, you just need to define one quantity, and that is the number of customers in the system. That at an arrival point is what one has to look at, but since there, any other point is like any other point in some sense in the Markovian system because of this Markovian system. And that is because of the memoryless property of the exponential distribution like when the customer arrives, what is the remaining duration of the customer who suppose some customer is undergoing service and if you are looking at what is the remaining duration, that will also be exponential, like how long he has been in service, that will also be an exponential. So, because of such things as you just need, you do not need to keep looking at the service; how much service has been done with the server which is currently in service. But because of the exponential, you could treat it as if the service had started just then. So, so you just need the number in the system alone, which was much easier.

- For an $M/G/1$ queue, its general state description would require specification of both the number in the system and the amount of service already provided to a customer being served when an arrival takes place.

Because now, you no longer have the memoryless property because that exponential distribution is the only distribution with this memoryless property. So, now you need to keep, so typically even in a very simple sense like it will become a two dimensional one, one is the number in the system the other is the amount of service already provided to the customer being served when an arrival takes place. So, that is what would happen here. So, because of I mean, the

main part of why this analysis becomes difficult is because you need to keep additional information in mind if you want to describe the state. So, that is the reason.

Now what we will do first is that we will not directly look at the system state distributions immediately; what we will do is we will try to derive the expected value measures or the mean value measures of effectiveness directly or mean performance measures or mean performance metrics which in this case, basically what we have in L, L_q, W, W_q like what we had in the case of $M/M/1$ queue this is what we are interested.

So, we will directly derive these results for these four quantities and these formulas, which is a collection in a way because it can be expressed in different forms. These formulas are known as Pollaczek-Khinchin formula or PK formula. And since there are two such formulas that are typically there, we will call this as Pollaczek-Khinchin (PK) mean formula because this connects the mean values of the system. There is another one which we will later we will simply refer to that also as PK formula, but that will be the PK formula, but that will also be called as PK transform formula because it relates to the transforms of the distributions. So, here it is only mean. So, we can, to be specific, we may call this PK mean formula Pollaczek Khinchin mean formula is what we have.

What we will do; we will obtain one of them; you can obtain, for example, any one of them, and the others can be obtained using the usual relationship that exists between these four quantities. We will derive the first derivation by considering the system when the customers arrive at the system. But there could also be a second derivation where you look at the system at the point of departure of customers, and the same thing can also be derived. We will first derive this part which is basically at the arrival instance. So, what we will do now is we will derive this PK mean formula using arrival times are looking at the system at the arrival instance.
Now consider a customer arriving to this queueing system; we will call the customer she. So, her delay is determined by the customers already in the system when she arrives. So, look at an arriving customer. So, how much is going to be her delay in the system will be determined by the customers who are already in the system when she arrives. In particular, there may be customers in the queue, and there may be a customer already in service; there are two possibilities. Now let us consider the customers in the queue at the time of her arrival. So, what is the delay that each of these customers contributes to her delay in the system. So, each customer in the queue ahead of her contributes to an average $E[S]$, the mean service time to her delay. So, how many are there ahead of her, there are, on average, L_q customers in the queue when she arrives.

Now, remember this is L_q is the number of customers at an arbitrary point of time. Now, because the arrivals are Poisson in this particular system, and Poisson arrivals see time averages, that means the PASTA property. So, the number of customers that an arriving customer sees in front of him is also the same as the number of customers at an arbitrary time. And hence the mean number of customers that an arriving customer sees in front of her in the system is also L_q. So, this property holds only under the Poisson arrivals or wherever this PASTA property holds. So, arrival also sees L_q much like an arbitrary customer, arbitrary time point view of the system. So, since each of these customers contributes $E[S]$ and there is L_q number of customers in the system when she arrives. So, the total average delay due to these customers waiting in front of her when she arrives since we are following FCFS discipline will be $L_q E[X]$; this is one component of the waiting time. Now the customer who is in service (if there is anyone in the service) when she arrives contributes a different amount to her delay. So, this customer has already completed part of his service. So, his contribution to her delay is his remaining service time, not his total service time. This would have been the case

total service time in the case of exponential, but now since that is not the case. So, whatever the remaining service time, that is the contribution of that customer to her delay in the system.

- Combining these two, the average queue wait for the arriving customer is

$$W_q = L_q E[S] + P\{\text{server busy}\} \cdot E[\text{residual service time} \mid \text{server busy}].$$

If the server is free, obviously, you do not have anything here because that is 0. Because they do not contribute anything, and that will happen only if the queue is also empty. But if the server is busy, then only like this residual service time comes, so this one server is busy.

Using $L_q = \lambda W_q$ to eliminate L_q and then rearranging terms gives

$$W_q = \frac{P\{\text{server busy}\} \cdot E\left[\text{residual service time} \mid \text{server busy}\right]}{1-\rho}.$$

So, we know ρ is λ/μ or $\lambda E[S]$ is what we already know. So, that is what is this quantity is because you just have to this is what will give you this one actually there is nothing great about it. So, this is what we have here.

- In the above, $P\{\text{server busy}\}$ is the probability that the arriving customer finds the server busy. By the PASTA property, this is the same as the fraction of time the server is busy, so $P\{\text{server busy}\} = \rho$.

So, this is again; the PASTA property is used here to interpret this quantity as equal to rho because of the PASTA property. So, $P\{\text{server busy}\}$ we have now computed to be ρ. Now what is remaining here is $E[\text{residual time}|\text{server is busy}]$.

- We need to find the expected residual service time, conditional on the arrival finding the server busy. It can be shown that

$$E\left[\text{residual service time} \mid \text{server busy}\right] = \frac{E[S^2]}{2E[S]}$$

You recall we did this when we studied the renewal process; we considered the excess lifetime or the residual lifetime or remaining lifetime, and we showed that in the limiting case. The mean of either residual lifetime or one of them, the residual lifetime or the age in for both of them were equal, and it is given by $\frac{E[S^2]}{2E[S]}$. So, that is a standard result in renewal theory; that is what we said, and we have given that even we have given the distribution as well, and from there, one can easily obtain this. But one can also obtain an alternative way does not matter, but we have already seen this result. So, $\frac{E[S^2]}{2E[S]}$ is what is obtained to be $E\left[\text{residual service time} \mid \text{server busy}\right]$. Now, you can write it by multiplying and dividing by $E[S]$; you can easily show that

$$E\left[\text{residual service time} \mid \text{server busy}\right] = \frac{E[S^2]}{2E[S]} = \frac{1+C_B^2}{2}E[S],$$

where C_B^2 is the squared coefficient of variation (SCV) of the service distribution, namely $\frac{Var(S)}{E^2(S)}$.
♦ Recall that the above result is the average excess or average residual time of a renewal process.

Now you see here $\frac{E[S^2]}{2E[S]} = \frac{1+C_B^2}{2}E[S]$ leads to some counter-intuitive idea. What you expect in an intuitive way is the residual service time when the server is busy. You expect that to be $\frac{E[S]}{2}$, this quantity alone or the first term $E[S]/2$. So, in general, $\frac{E[S^2]}{2E[S]} = \frac{1+C_B^2}{2}E[S] > E[S]/2$, that is,

- Unless $C_B^2 = 0$ (in the case of a deterministic situation where the variance would be equal to 0 or a degenerate distribution case), the expected remaining service time as seen by a customer arriving to a busy server is more than half of the expected service time (this is an example of the inspection paradox or paradox of residual life).

  ▶ This counter-intuitive result is because customers are more likely to arrive during long service intervals compared with shorter ones, bringing the average above $E[S]/2$.

So, this is a bit counter-intuitive, of course; there can be another I mean, many reasoning and why this happening can be given here. But this is what you could observe; you could observe that this one expected residual service time when server is busy of the customer who is undergoing service, as seen by an arriving customer, is basically more than the average or the half of the average waiting time. So, that is what paradox is; of course, probability theory is full of paradoxes, and this is also possibly one. But this is a bit counter-intuitive, but in many such results, it holds true in queueing context or in any other case as well. So, that is what you are observing.

- Finally, combining the preceding results gives us

$$W_q = \frac{1 + C_B^2}{2} \cdot \frac{\rho}{1 - \rho} \cdot E(S)$$

- This formula has three terms: a variability term (because this coefficient of variation is really a measure of variation), a utilization term (connected with the server utilization), and a time scale term (which is E[S]).

- *The first term* $(1 + C_B^2)/2$ involves the squared coefficients of variation of the service distribution C_B^2. For exponential service, $C_B^2 = 1$ so $(1 + C_B^2)/2 = 1$ and, in this case, formula for W_q reduces to the analogous formula for the $M/M/1$ queue.

  So, this is reduced to the analogous formula for an $M/M/1$ queue. So, if $\frac{\rho}{1 - \rho} \cdot E[S]$ is the analogous formula for $M/M/1$ queue, then you can also think about think of this W_q as if it composing of two terms, one is $\frac{1 + C_B^2}{2}$ which is connected with the variability the other one $\frac{\rho}{1 - \rho} \cdot E[S]$ is the corresponding formula of an $M/M/1$ queue. So, you can also think about this as if the product of these two terms.

- *The second term* $\rho/1 - \rho$ involves the queue utilization and increases to infinity as $\rho \to 1$.

- *The last term* $E(S)$ has units of time and can be thought of as a time-scale factor.

- Thus, W_q is the product of two time quantities that are independent of the time scale chosen and the time-dependent quantity $E(S)$.

- The formula for W_q is a powerful result. Only three parameters are needed to compute W_q:
  – The arrival rate λ.
  – The mean $E(S) = 1/\mu$ of the service distribution.
  – The squared coefficient of variation (SCV) C_B^2.

- Other measure of effectiveness can easily be obtained from W_q.
  Using Little's law and/or $W = W_q + E(S)$, we obtain

$$L_q = \lambda W_q, \quad W = W_q + 1/\mu, \quad L = \lambda W = L_q + \rho.$$

Now, once we do this, then we can write it out or put this in this table form which we will refer to as the table of PK mean formulas or Pollaczek Khinchin mean formula for the $M/G/1$ queue.

The following table shows several different ways to express the results. The first column gives the performance measures and the second column express the measure using the SCV of the service distribution C_B^2, the third column uses the second moment of the service distribution $E(S^2)$, and fourth column uses the variance of the service distribution σ_B^2.

L_q	$\frac{1+C_B^2}{2} \cdot \frac{\rho^2}{1-\rho}$	$\frac{\lambda^2 E\left[S^2\right]}{2(1-\rho)}$	$\frac{\rho^2 + \lambda^2 \sigma_B^2}{2(1-\rho)}$
W_q	$\frac{1+C_B^2}{2} \cdot \frac{\rho}{\mu-\lambda}$	$\frac{\lambda E\left[S^2\right]}{2(1-\rho)}$	$\frac{\rho^2/\lambda + \lambda \sigma_B^2}{2(1-\rho)}$
W	$\frac{1+C_B^2}{2} \cdot \frac{\rho}{\mu-\lambda} + \frac{1}{\mu}$	$\frac{\lambda E\left[S^2\right]}{2(1-\rho)} + \frac{1}{\mu}$	$\frac{\rho^2/\lambda + \lambda \sigma_B^2}{2(1-\rho)} + \frac{1}{\mu}$
L	$\frac{1+C_B^2}{2} \cdot \frac{\rho^2}{1-\rho} + \rho$	$\frac{\lambda^2 E\left[S^2\right]}{2(1-\rho)} + \rho$	$\frac{\rho^2 + \lambda^2 \sigma_B^2}{2(1-\rho)} + \rho$

You do not need to remember this; you just need to understand the one component, which is basically $W_q = \frac{1+C_B^2}{2} \frac{\rho}{1-\rho} E[S]$ and how this is obtained here. And this approach is known as the residual lifetime approach, and this is applicable only to apply to get the mean results. If you are interested in distributions of the number in the system or waiting time distributions, then you cannot employ this approach. So, this is a way of obtaining, and this is you see that from the renewal theory something we are using here to arrive at the mean value. So, in queueing theory, many a time, I mean employ the mean value approaches meaning to directly obtain the mean values as you have done for a queueing network case. The mean value analysis algorithm such kind of thing. So, here also we obtained directly, without going into the distributions we directly obtain the mean value measures. So, this is a quite powerful tool or the idea that you can employ very effectively if you are interested only in the mean value measures. So, that is what we did here.

Let us look at quickly like a special case; it is not really an example; it is a special case.

Example.
Consider an $M/E_k/1$ queue. The SCV of an E_k distribution equals $1/k$. Therefore,

$$W_q = \frac{1+1/k}{2} \frac{\rho}{1-\rho} E(S)$$

which coincides with earlier results.

Similarly, we can obtain the results of $M/D/1$ either from this by letting $k \to \infty$ or by taking $C_B^2 = 0$ in the PK formula in the table.

So, accordingly, you will get the measures from here which for the $M/D/1$ queue or any other distribution now, what you need is mean and variance or mean and squared coefficient of variation or mean and the second moment, the first moment, and second moment of the service time distribution. So, these are the two things you need to get the mean value measures; that is the bottom line when you employ these ideas that we have here.

So, this is what is called the residual lifetime approach to obtain the mean value measures in the case of an $M/G/1$ queue. This is where we will stop, and then we will then take it up later the case of obtaining the distribution of in the case of an $M/G/1$ type. This is only the mean value measure, which one can directly apply to get the things. That is what our intention is to show at this point. So, we will see later.
Thank you, bye.