# Introduction to Queueing Theory

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

## Lecture - 41
## M/G/1/K Queues, Additional Insights on M/G/1 Queues

Hi and hello, everyone. What we have seen so far is the base classical $M/G/1$ model where we assumed that the Poisson process arrivals and generally, but IID distributed random service times single server. The capacity of the system is infinite, and the service discipline or the queueing discipline that is being adopted is this FCFS which is basically the analysis. We are trying to make an analogous analysis corresponding to, as opposed to the case of $M/M/1$. We want to look at the number in the system; we want to look at the waiting time; we want to look at the busy period. Now we have given certain results; whether it is PK mean value formulas, or for the waiting time distributions we have given in terms of transform for this busy period also, we are given in terms of transforms. So, that is the core of the analysis that one does for any queuing system. So, the popularity or the relevance of that model is that it is applicable in a much, much wider class of situations as compared to a purely Markovian based model, but we see here that the analysis can still be done using the Markov process analysis or the Markov chain analysis that we have acquainted with ourselves by considering only the Markovian process models. So, that is the advantage. Once you do the complete Markov process analysis-based models, you increase your familiarity or comfortness with the Markov process-based analysis. When you come to $M/G/1$ also especially, I mean, what are you trying to do? Basically, whether it is supplement variable technique or embedded Markov chain technique. You are somehow bringing back to the Markov and then applying the Markovian theory and then trying to connect the dots. Say, for example, in this case, you consider the departure epochs, and then you connect the dots means that you ensured that that is the same as at an arbitrary time point. So, that is all one would be interested in doing here.

So, let us look at the $M/G/1$ framework itself; of course, one can have variations just like we had variations in $M/M/1$, but not all of them we are going to consider. We are going to highlight certain portions and certain portions we are leaving open. So that you can explore further when you actually want to look at how exactly that is being done. But, one can do, and we will just highlight what are the things that are doable and at what level what would be which is easier which is a little messier, or complex analysis one has to do is what then we will try to highlight in the remaining portions of this $M/G$ type models. So, in that process, first, one that we will take it up is the capacity limitation. Suppose instead of the infinite capacity model and if you find that the capacity is limited to, like in our case ordinary case that we limit our capacity to $K$. So, basically then what we have we have an $M/G/1/K$ model.

- An $M/G/1/K$ model, the analysis of such a model is very similar to that of an $M/G/1$. So, we will just look at the main ideas or main results or what we need to worry about in that case.
- Now, one thing is clear PK mean value formulas that we or that we directly derived based upon this are no longer applicable here, since the expected number of (joined) arrivals during a service period must be conditioned on

the system size.

Because any number of arrivals that comes during a service period of a particular customer may not be able to join. If there is a space in the system, then only they will be able to join; if not, they will not be able to join. So, it is in some way related to the system size. The current system size also you need to worry if you have to capture that portion, but we will not worry about this mean formula obtaining it directly in this by looking at the system size and so on; that is an idea, but we will not look at that aspect, in this particular case since there is only a finite number of states, because $M/G/1/K$, so there is an only finite number of customers can be there in the system. So, it has only a finite number of states if you think the number of customers has the states. Then one can find the steady-state probabilities directly, I mean, at first basically in a sense, and then obtain the mean value results. That will be both ways you have obtained, like the distribution and the mean value result. So, in this case, now what happens that the basic step that is different is that we have a similar Markov chain, but now like what will leave behind everything will come, but now the transition probability matrix of the embedded Markov chain. Now, this must be truncated at $K-1$. Remember here that we are not truncating at $K$ as what we are observing? We are observing the system just after a departure. So, maximum just prior to departure maximum number that could have been in the system was or is $K$. Since one customer is leaving, the number of customers who are left behind the departing customer is at most $K-1$. So that is what then you will have in this Markov chain. So, this is truncated at $K-1$; if you truncate it, then this is what, and we also assume $K>1$ just to avoid that triviality because later on, when we consider $M/G/c/c$ model, it will become a special case of that just to avoid that we will also assume that $K>1$, $K=1$ also will equally hold good. So, sometimes for simpler problems that when you want to do, you can always assume there is no harm, but for this analysis, we will assume that $K>1$, it does not make a difference, so do not worry about that. So, basically, this was the one-step transition probability matrix; everything is there.

$$P=((p_{ij}))=\begin{bmatrix} k_0 & k_1 & k_2 & \dots & 1-\sum_{n=0}^{K-2} k_n \\ k_0 & k_1 & k_2 & \dots & 1-\sum_{n=0}^{K-2} k_n \\ 0 & k_0 & k_1 & \dots & 1-\sum_{n=0}^{K-3} k_n \\ 0 & 0 & k_0 & \dots & 1-\sum_{n=0}^{K-4} k_n \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1-k_0 \end{bmatrix}$$

If I want to obtain the steady-state solution or the equilibrium solution for this Markov chain, we will use the same stationary equations.

- The stationary equations now become

$$\pi_i = \begin{cases} \pi_0 k_i + \sum_{j=1}^{i+1} \pi_j k_{i-j+1}, & i = 0, 1, 2, \ldots, K-2 \\ 1 - \sum_{j=0}^{K-2} \pi_j, & i = K-1 \end{cases}$$

- These $K$ (consistent) equations in $K$ unknowns can then be solved for all the probabilities

  Because we have removed from $\pi P = \pi$ one equation, and we have added $1 - \sum_{j=0}^{K-2} \pi_j$. So, then this is a linearly independent system of equations that you can solve for these quantities.

- The average system size at points of departure is thus given by $L = \sum_{i=0}^{k-1} i\pi_i$.

  So, once you have the distribution, you can obtain the means corresponding to that. Now, that is what you are seeing it here.

- The first portion of the stationary equation is identical to that of the unlimited $M/G/1$.
  ► The respective stationary probabilities $\{\pi_i\}$ for $M/G/1/K$ and $\{\pi_i^*\}$ for $M/G/1/\infty$ must be at worst proportional for $i \leq K-1$; that is, $\pi_i = C\pi_i^*$, $i = 0, 1, \ldots, K-1$.

  ►The usual condition that the probabilities sum to one implies that $C = \frac{1}{\sum_{i=0}^{K-1}}\pi_i^*$.

So, that is a way you can easily obtain the connection between or establish the relationship between the stationary distribution or the steady-state distribution between $M/G/K$ and $M/G/1$. Because you see this equation is the same; that means that it must be that $\pi_i = C\pi_i^*$ must be the case where $pi_i^*$ is the corresponding quantity in the $M/G/1$ model and this constant $C$ then I can obtain it as the one which is normalizing $C = \frac{1}{\sum_{i=0}^{K-1}}\pi_i^*$ to 1. So, that is a way we can do this.

- Also, the probability distribution for the system size encountered by an arrival will be different from $\{\pi_i\}$, since now the state space must be enlarged to include $K$.

  Because now the state space must be enlarged to include $K$ because an arriving customer can also see $K$.

- Let $a_n'$ denote the probability that an arriving customer finds a system with n customers.
  ► Here, we are talking about the distribution of arriving customers whether or not they join the queue, as opposed to only those arrivals who join, denoted by $a_n$.
  ► The distribution $\{a_n'\}$ also has its own significance. Because you just want to look at the system on to those time points.

Now, to do this to obtain this a n dash in this particular case.

- Recall that in the proof that $\pi_n = p_n$, the equality holds as long as arrivals occur singly and service is not in bulk then the limiting proportions are all equal is what then we have seen.
- Similar is the case with $a_n\prime$ except that the state space are different. This difference is taken care of by first noting that

$$\begin{aligned}\pi_n &= P\left\{\text{arrivals finds } n \mid \text{customer does in fact join}\right\}\\ &= a_n = \frac{a_n'}{1-a_K'}, \quad 0 \le n \le K-1\\ \text{therefore } a_n' &= (1-a_K')\pi_n, \quad 0 \le n \le K-1\end{aligned}$$

- To get $a_K'$, we use an approach similar to Markovian model where we equate the effective arrival rate with the effective departure rate, i.e.,

$$\lambda\left(1-a_K'\right) = \mu\left(1-p_0\right)$$

Therefore

$$\begin{aligned}a_n' &= \frac{\left(1-p_0\right)\pi_n}{\rho}, \quad 0 \le n \le K-1\\ a_K' &= \frac{\rho-1+p_0}{\rho}\end{aligned}$$

- But, since the original arrival process is Poisson, $a_n' = p_n$ for all $n$. Thus,

$$a_0' = p_0 = \frac{\left(1-p_0\right)\pi_0}{\rho} \Rightarrow p_0 = \frac{\pi_0}{\pi_0+\rho}$$

Finally,

$$a_n' = \frac{\pi_n}{\pi_0+\rho}$$

Now, you can think about what would happen if there was no limit or anything; what would have been these expressions you can think about it a little bit. So, this is what you could consider. So, $M/G/1/K$ model, what one can do is that the analysis is almost similar to the $M/G/1$ model; there is not much difference. The only thing is now you have a finite state space which gives rise to some complexities that you need to take care of when you are working out. Now, in this case, directly, PK mean value formulas do not hold. So, you try to obtain the system size distribution first, and from there, you obtain the mean value results, and that is what one can do. Of course, other aspects one can look at it, but anyway, it is almost similar to what one would do for an $M/M/1/K$ model in the particular scenarios.

So, this is on the finite queuing systems. Now, there are some more additional things that we will highlight or say where we will not even go to even this many details in that situation. We will study some additional results connecting with impatience, output, transience, finite source matching, and so on. We just highlight that we are not going to go into any detail in any of those.

- Now let us take the case of impatience; we said that the impatience could be of different forms, but one can easily introduce balking into the $M/G/1$ queue by prescribing a probability $b$ that an arrival decides to actually join the system.

  And this is not just in this particular case; in any model, you have considered so far, you can always easily introduce this balking. So, here it is also very easy. So, then the true input process would be then the filtered process with the rate now then it will be $\lambda b$. This b is the probability of joining each arrival with probability b only it joins; otherwise, it will balk. Then $b\lambda$ would be that rate of arrival for such arrivals who actually come into the system. This arrival process is a Poisson process with $b\lambda t$ mean or with the rate $b\lambda$ and hence these $k_n$'s which is basically when you want to consider the number of arrivals during a service time.

$$k_n = \int_0^\infty \frac{e^{-b\lambda t}(b\lambda t)^n}{n!} dB(t)$$

- The rest of the analysis goes through parallel to that for the regular $M/G/1$ queue, with the probability of idleness, $p_0$, now equal to $1 - b\lambda/\mu$.

  So, this form is not just in M G 1; this kind of change can be effected in any model that you can think of.

Because you are filtering away a portion in the beginning itself without coming into the system, and because this is randomly, you are picking the customer. Every customer has a probability of $b$ of joining the queue or leaving or balking one with a probability $1 - b$. So, this is the random split of the Poisson process, which is again a Poisson process. So, which is what so one need not consider that, but if it is then one can think there are situations where you these kinds of models are also required to be handled, but the other kind of thing which is basically the reneging kind of impatience if you want to then it requires some amount of work. Now, that distribution also matters there actually, what is the distribution of it whether that is general or that is still you are keeping it as exponential. In $M/G/1$, you are still retaining one exponential and one thing you are generalizing. When you bring in the reneging, how long after joining the queue is that customer thinking about reneging. If you think so, you need to bring in that distribution of that, whether that is also $G$ or an $M$, which is exponential. If it is exponential, then you know that things will be slightly easier, but because then you do, then you do not have only one sort of non-exponential distribution to handle; we can handle it. So, the complexity level varies depending upon what you are assuming, as you say, but here in the balking case, you do not bring in any other distribution, but in the reneging case, you will bring in another kind of distribution. Now, what kind of distribution would also give rise to the complexity. So, that is about impatience.

Now, as far as the output process is concerned, we have already seen in the case of $M/M/1$ that particular system, and in fact, $M/M/c$ itself also has Poisson output, but if you ask the question, is there any other type of queues that has this same property? The answer is no. So, again $M/G$, in particular, of course, $M/G/1$ queues do not possess this output process because Burke's theorem; when we also proved, we said that this is the only system that has this property. So, because such processes are not reversible processes. For the output to be the same as the input, what should be intuitively it is clear that whether I am looking at the process starting from this side if I look at this and starting and pick a time point at the end and then starting from this if I look at the process. So, this is a probabilistic replica of the forward process, the forward moment. It is not that it has to match exactly. So, it is a that whatever we say, it is a probabilistic replica meaning that there is a path from this side to this, and this path, when you look back it, is also one of the paths which you would have otherwise obtained in as one of the paths in the forward process. So, that is what you can do. So, if that is the case, then only like you will get this property to hold, but that is what made this $M/M/1$ process $M/M/c$ queue process as having that kind of reversibility property leading to the output process being Poisson, but $M/G/1$ does not have that, then what can we say about the distribution of the inter departure time in an $M/G/1$ queue in steady-state. We might still like to look at that because this is one of the important models. So, one wants to look at what is that ok, because the reason also will clearly be known to you because why in the queueing network context like we assumed the Poisson arrivals Poisson service. Poisson service really we need to keep in because output we want to make it Poisson so, that you get the network which gives rise to your product form network for which analysis can be done easily. If you bring in $G$ now, we just said that this does not have the Poisson output; then what kind of output is this whether we can characterize. If you can characterize it to some extent, at least that one can use as an input to the next one, but the next one will become a really $G/G$ model kind of thing. So, that is what then one has to handle. However, even in any case, sometimes it is necessary because when the network is small, like 2 nodes, 3 nodes, then one can easily characterize the distribution. One can still analyze to some extent. So, if that is the case, if you are interested in the CDF of the interdeparture time, let us call this $C(t)$ with $B(t)$, as usual, the service time CDF.

$$\begin{aligned} C(t) &= P\left\{\text{interdeparture time} \leq t\right\} \\ &= P\left\{\text{system experienced no idleness during interdeparture period}\right\} \\ &\times P\left\{\text{interdeparture time} \leq t | \text{ no idleness}\right\} \\ &+ P\left\{\text{system experienced some idleness during interdeparture period}\right\} \\ &\times P\left\{\text{interdeparture time} \leq \ t | \text{ some idleness}\right\} \\ &= \rho B(t) + (1-\rho)\int_0^t B(t-u)\lambda e^{-\lambda u} du \end{aligned}$$

since the length of an interdeparture period with idleness is the sum of the idle time and service time.

**Exercise.** Exponentiality of $C(t)$ implies exponentiality of $B(t)$

So, in our context, $C(t) = \rho B(t) + (1-\rho)\int_0^t B(t-u)\lambda e^{-\lambda u} du$ is an important result in the sense that it characterizes that gives us the interdeparture time distribution in terms of the service time distribution when arrivals are Poisson; that is what you have here. So, this is a quite an important result in that context, but then this fact that $M/M/1$ is the only $M/G/1$ with exponential output has serious repercussions for the solution of, say, for example, series a tandem network models because the output of the first stage will be exponential which would like to be only if it is $M/M/1$ otherwise it is not going to be exponential.

But still, some small $M/G/1$ tandem network kinds of problems can be handled numerically with the help of $C(t)$, where $C(t)$, that is where $C(t) = \rho B(t) + (1-\rho)\int_0^t B(t-u)\lambda e^{-\lambda u} du$ is being used. When I said that, for a small network, two or three networks, which is basically, say, for example, in many of the applications related to supply chain management or anything of inventory management and so on. You would always in production systems or manufacturing systems you will have two, three systems or two, three stages only like you will have in such situations which can still be handled within this framework using this particular results, one can at least handle numerically. So, that is what is the advantage that one can where $C(t) = \rho B(t) + (1-\rho)\int_0^t B(t-u)\lambda e^{-\lambda u} du$ will be of very high utility in that scenario.

Now, by putting a capacity restriction on $M/G/1$ at $K = 1$ with $M/G/1/1$, it can be seen that such queues also have IID departure times, but because now this is because the successive departure epochs are identical to busy cycle which is found as the sum of ideal time paired with an adjacent service time is what then the busy cycle. So, that is what it will come out to be in the case of $M/G/1/1$ system. So, this is about the output process. Now, with respect to the transient results, though we said that we are not going to talk about. So, we will just highlight the point that is all. In this transient result of $M/G/1$ queue, we will again take the embedded Markov chain and appeal to Markov chain theory and Chapman Kolmogorov equations to obtain the transient state distributions.

$$p_j^{(m)} = \sum_k p_k^{(0)} p_{kj}^{(m)}$$

(where $p_j^{(m)}$ is then the probability that the system state is in state $j$ just after the $m$th customer has departed)

But then here, this $k$ is basically the state space is infinite, but computation wise, you need to get, you cannot handle it in that sense, but then you have to truncate at some level. So, careful, and there have been studies about how one can truncate and so, on not just now, even 30, 40 years before itself, think that those things have started. Because in an infinite-dimensional matrix, you cannot really code it. So, it needs to have some finite dimension. So, then what would be that how will you determine what would be the impact of that and the further analysis all those things have to be kept in mind. So, what are the points to note, and how can one truncate carefully so that you are not losing anything in your analysis? You are not leading to wrong results in that. So, there are papers which are written on that. So, a careful truncation is what is needed in such situations when actually, not just in this case, any numerical procedure that you want to apply when you have an infinite state-space system you will have to truncate at some point then the where you will truncate is the question always that you will. So, using that idea, then transit analysis can also be done.

The finite source $M/G/1$ is essentially the machine repairmen problem with arbitrarily distributed repair time; the repair time is $G$ and has been solved in the literature again using an embedded Markov chain approach. So, it is not very difficult to do that. Now with above, with respect to bulk queues, the bulk-input $M/G/1$ and which is denoted by $M^{[X]}/G/1$, and the bulk-service $M/G/1$, which is denoted by $M/G^{[Y]}/1$, can also be solved with the use of Markov chains.

While the bulk-input model is relatively easy, bulk-input because one can do a similar analysis, but with certain differences are there, and it is available in the text itself the, which is easier to handle. But the bulk-service problem is a bit more complex, but it is still doable, no problem. Now with respect to priorities, again, the mean value measures or expected value measures can be obtained in an easier manner in certain models, but other than that, even you have seen even with Markovian structure itself how complex that has become. So, here also there are not done but is also there. So, anyone interested can always look at that further into this case. So, these are certain points that we wanted to see with different features how one gets added to this and whether those kinds of models are still doable or it has been done, and in what kind of complexity levels that you want to you just want to have a certain idea that is what we have given here fine. So, we will stop here, and we will continue in the next lecture.
Thank you bye.