Introduction to Queueing Theory

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

Lecture - 14
M/M/1 Queues: Performance Measures, PASTA property, Waiting Time Distributions

Hi and hello, everyone; what we will do now is let us continue our discussion of M M 1 Queues that we have seen. So, we have obtained the equilibrium system size distribution in the previous lecture by three different methods; those three different methods you remember because that is what is applicable for different queueing models, depending upon the situation depending upon the convenience that you will apply any of them to get the solution. Now, just recall:

- For $M/M/1$ queue, the steady state system size probabilities are given by a geometric distribution.
- p_n depends on λ and μ only through $\rho = \lambda/\mu$.
- The steady state solution for the $M/M/1$ system exists under the condition that $\rho < 1$, or equivalently $\lambda < \mu$, or equivalently $\frac{1}{\mu} < \frac{1}{\lambda}$ (intuitive!).
  Q: Why no equilibrium solution when $\lambda = \mu$?
- $\rho < 1$ is the condition for ergodicity of the process and the stability condition for the $M/M/1$ system.
- The steady state probability distribution for the system size $p_n = P\{N = n\} = (1-\rho)\rho^n, n = 0, 1, 2, \ldots, \rho = \frac{\lambda}{\mu} < 1$ allows us to calculate various equilibrium performance measures of the system.

  So, many a time, it will be simply referred to as the equilibrium solution for the queue; I mean, for the queue, we simply mean it is always that it is the system size distribution that is of interest to us.

Now let us look at certain performance measures so that we can tell more about the system on hand and depending upon the quantities of interest. So, that is what one calls performance measures like you have the distribution and from which we try to obtain these quantities. First, what we will pick it up is this random variable N, which represents the number of customers in the system in equilibrium and its distribution only we have obtained.

- Recall that N represents the random variable "number of customers in the system in steady-state" and L is its expected value.

$$L = E(N) = \sum_{n=0}^{\infty} n p_n = (1-\rho)\rho \sum_{n=0}^{\infty} n\rho^{n-1} = (1-\rho)\rho \frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}.$$

$$\left[\text{using } \sum_{n=0}^{\infty} n\rho^{n-1} = 1 + 2\rho + 3\rho^2 + \cdots = \frac{d}{d\rho}\left(1 + \rho + \rho^2 + \cdots\right) = \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{1}{(1-\rho)^2}\right]$$

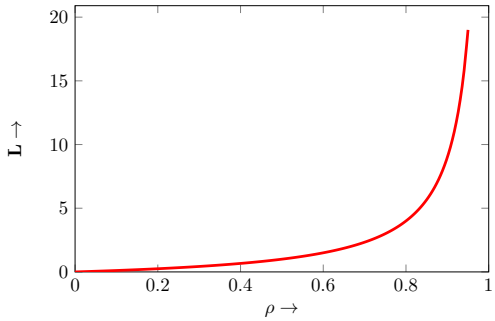
So, this is what is your expected system size as the average system size that one gets to observe.

- Similarly, we can obtain $E(N^2) = \dfrac{\rho + \rho^2}{(1-\rho)^2}$ and hence $Var(N) = \dfrac{\rho}{(1-\rho)^2}$.
- Expected steady state system waiting time W can be obtained from Little's Law as

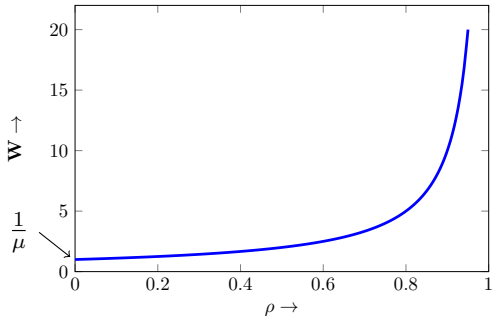
$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda}.$$

Now the purpose of getting these performance measures is typically to analyze the sensitivity of these performance measures with respect to the parameters of the model. That is what we will exhibit now. But, we will not do it for every other parameter, but that is what is typically done when you want to analyze the performance measures. In performance modeling of the systems, this is what one does to try to analyze the sensitivity. Either analytically, if possible, say, for example, I can look at now how L behaves as a function of ρ or as a function of λ alone if I have $\frac{\lambda}{\mu-\lambda}$ expression. So, I can try to get analytically, like, what is the change in L if I change λ? What is the change in the variance if I change my ρ? What is the change in the expected system time if I change the λ or μ, and so on. So, this is what is what-if analysis that one needs to do, and for that purpose, only you try to obtain these kinds of simplified measures rather than dealing with the distributions themselves.

Mean Number in the System



Mean Waiting Time in the System



- Observe that as $\rho \to 1$, both L and W grow in an unbounded fashion.
  - ♦ The behaviour is characteristic of almost every queueing system!

So, you could see for both probably up to 0.8 or something like that; you are in a good position. Things are not exploding in seen any sense. Whereas, beyond a certain point and very close to 1 for $\rho = 1$, what you are observing is that both L and W grow in an unbounded fashion, an exponential increase. And this behaviour is characteristic of almost every queueing system; it goes without saying you will not analyze it as has it already for every system. But this is what one does with respect to a system if you want to analyze the system better. Now, if you know exactly because if you know exactly ρ, you will know exactly what L is; that is what you will obtain here by using this formula simply. But you know you want to analyze; basically, that is what you know you want to do a performance analysis of the system means; that in terms of model parameters, how these performance measures behave is what is the objective here.

- Also, as $\rho \to 1$, we also see that both $p_0 \to 0$ and $Var(N) \to \infty$.

  ♦ Large variance implies that a randomly observed system size is likely to be very different from the expected system size (and hence the system predictability degrades).

  If the variance is small, I mean, that is why this variance quantity plays a role in any statistical study. Because the variance is small, you have data clustered around the mean; then, the mean will be a good measure to see. But if the variability is large, then the data will have values that could be very much different from the mean, that is what this implication and that is precisely like here also you would find.
  If the variance is large, then the system predictability degrades, which means then there is a large variation in the number of customers in the system that you get to observe at different points of time. And hence you need to account for it if you are designing the system or if you want to do an analysis or if you want to make any improvement in the system. So, that is all you have to keep in mind. So, you need because of this, this grows exponentially; it is like you for values beyond 0.98 or 0.99. Of course, you would like to use that because then you will see that you are trying to use the server to the maximum; that is what you are trying to do.

  But that comes at the cost where this is degrading:- the system predictability degrades, this mean number in the system explodes and hence the mean waiting time in the system; whether that is desirable, you have to see. You have to find an appropriate balance for this ρ in that particular case.

- Need for a sophisticated theory for ρ near 1 (heavy-traffic theory)!

  So, this kind of behaviour is typical of almost every queueing system that you would see, but we will not analyze it in such detail. But since this is the first model, what kind of analysis one can do with respect to performance measures is what we are trying to show with this kind of behaviour here.

Of course, this whole thing can be shown analytically as well, but it is much more appealing if you can exhibit it through such analysis such you know graph, it is very appealing to do. So, the similar kind of thing you can do with every performance measure that how it is sensitive with respect to parameters, but we are not going to take that, but that is free will. We will try to obtain the performance measures. One other performance measure is basically this L q which is nothing but the expected value of N q; what is that?

- The distribution of N_q (the number in the queue in steady state) can be obtained from that of N and therefore

$$L_q = E(N_q) = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n$$
$$= L - (1 - p_0) = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

  Note: $L_q = L - (1 - p_0)$ holds for all single channel one-at-a-time service queues.

- The expected queue size of nonempty queues, which we denote by $L_q^{'}$.

$$L_q^{'} = E[N_q | N_q \neq 0] = \sum_{n=1}^{\infty}(n-1)p_n' = \sum_{n=2}^{\infty}(n-1)p_n'$$
$$= \sum_{n=2}^{\infty}(n-1)\frac{p_n}{\rho^2} = \frac{1}{1-\rho} = \frac{\mu}{\mu-\lambda},$$

where p'_n is the conditional probability distribution of n in the system given that the queue is not empty and is obtained as

$$p'_n = P\left\{N = n | N \geq 2\right\} = \frac{P\left\{N = n, N \geq 2\right\}}{P\left\{N \geq 2\right\}} = \frac{p_n}{\sum_{n=2}^{\infty} p_n} = \frac{p_n}{\rho^2}, \quad n \geq 2$$

- Another performance measure of interest is finding the probability that the system is loaded with at least a particular number of customers. That is,

$$P\left\{N \geq n\right\} = \rho^n, \quad \forall n.$$

So, if there is if you are looking at what is the probability that at any given point of time, in equilibrium, what is the probability the system will have at least 10 customers, at least 20 customers. So, it is at least n customer, which means this probability is equal to rho n, which again you can sum up the tail of this geometric distribution, and from the geometric sum, you can obtain this quantity. So, this is not a very difficult one. So, you can do that.

- Now from Little's law, the steady state expected line delay W_q (expected waiting time in the queue) is

$$W_q = \frac{L_q}{\lambda} = \frac{1}{\mu} \frac{\rho}{1-\rho} = \frac{\rho}{\mu - \lambda}.$$

Again you can analyze what will happen to this $P\{N \geq n\} = \rho^n$ as $\rho \to 1$, and what will happen to, for example, $\frac{\mu}{\mu - \lambda}$ as $\rho \to 1$, and so on like you know many things you can analyze in a similar way.

- So, all the above formulas are very simple and useful. But they were developed keeping in mind that they were developed under specific assumptions of Poisson arrivals, exponential service time, and whatever assumption that we have made with respect to $M/M/1$ and also the steady-state condition and queue stability to exist that is what we are assuming.
  So, the transient effects or temporary effects are not captured by them; what we mean is that, suppose if you are looking at an airport and there is some kind of weather not good for flights to land at that particular point of time. So, temporarily, what you will see is that this $\rho < 1$ may not hold true. So, there will be some congestion; that is what will happen in the queue in the system.

So, those things and all it's not captured by the equilibrium analysis, but then the arrival rate, service rate things will get change; that is only temporary though, but still, one cannot handle that. So, these kinds of analyses are very smooth; at the first level, things are in equilibrium, and there is no abnormal behavior that happens. So, transient effects, which means when you start the system, there may be some hiccups to get the system moving and then comes to your steady state. So, basically, what you have is that the steady-state condition is established after that, then all these results will hold true; that is what you have to keep them. These are all the constraints; one needs to keep in mind that any mathematical model, any analysis that you do with respect to a particular model, is only an abstract of essential features of this real system. And the underlying assumptions have to be validated or kept in mind while trying to apply the results. If that is not done, you will not be successful in applying the mathematical model to any real system. So, that is

the basic principle that holds here for all things. So, you always have to keep in mind the underlying assumptions that we have with respect to this. We will now see some more distribution of waiting time distribution in equilibrium, but before we do like this kind of queueing system with Poisson arrivals have an interesting property which you will use quite often in our all our analysis; that is what is called as the Poisson arrivals see time averages, Poisson arrivals see time averages, or in short "PASTA" property.

- PASTA is an interesting property of Poisson arrivals. This means that an observer from a Poisson arrival stream sees the same system size distribution as a random observer from outside.

  So, at a random point of time, you look at the system, at the arrival instance where the arrivals happen according to a Poisson process, at the arrival instances, you look at the system, and looking at the system at a random point across the timeline, this behavior of the system will exactly be the same, and that property is what this PASTA.

- For Poisson arrival streams, we have $a_n = p_n$ for $n \geq 0$, provided the state of the system changes by at most one (Result holds in transient state as well).

  p_n we already know, is the probability of observing the system at a random point of time. It does not specify at what time you look at this, and then you will get to observe n in the system. The system is in equilibrium, and at a random observer at a random point of time, you are looking at the system is what will give you the probabilities p_n, a_n is basically the arrival instant observations of the system, and the corresponding probability is what this a_n.

  Here, a_n denotes the probability (in steady state) that an arriving customer finds n in the system (just prior to arrival). We determine a_n using Bayes' theorem.

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

  So, this is what you see here, $a_n = p_n$. So, it is basically the independent increment property of the Poisson process that makes this whole true. That is what we have.

By PASTA, we mean: In a queue with Poisson arrivals, the limiting proportion of arrivals that find the system in some state n is equal to the limiting proportion of time the system spends in that state n.

So, the long run proportion of the time system spends in n is what is p_n. So, the equality of these two is what we mean when we say a PASTA property. So, PASTA property means the equality of a_n's and p_n's; that is what you have. So, this is the PASTA property, which we will use now. Where will we use it? In this computation. We want to know the waiting time distribution.

- Recall that the random variable T_q denotes the time spent waiting in the queue in steady state and let $F_{T_q}(t) = P\left\{T_q \leq t\right\}$ represent its cumulative probability distribution.
- We are assuming that the queue discipline is FCFS (first come first served).

  Till now, we did not worry about that, but for waiting time distribution, this is required; what is the queue discipline? That plays a critical role, as you will see.
- Let a_n denote the probability (in steady state) that an arriving customer finds n in the system (just prior to arrival).

$$F_{T_q}(0) = P\left\{T_q \leq 0\right\} = P\left\{T_q = 0\right\} = P\left\{\text{system empty at an arrival}\right\} = a_0$$

- The probability p_n and a_n are not always the same, but for Poisson input, $p_n = a_n$ by "***PASTA***" property. Thus $F_{T_q}(0) = p_0 = 1 - \rho$.
- We need to determine $F_{T_q}(t)$ for $t > 0$.
  Observe that if there are n units in the system upon arrival of a customer, then in order for the customer to go into service at a time between 0 and t, all n units must have been served by time t, that is why this discipline is important. How the service takes place from the customers waiting in the queue?

So, if he has to go to service at a time between 0 and t, only when if he goes to service between 0 and t, so, we are computing $P\{T_q \leq t\}$ then all n units who are ahead of him or customers who arrived the ahead of him must have been served by time t.

- Since the service distribution is memoryless, the distribution of the time required for n completions is independent of the time of the current arrival and is the convolution of n exponential random variables and has a Gamma distribution. Also, note that $a_n = p_n$.

$$\begin{aligned}
F_{T_q}(t) &= P\left\{T_q \leq t\right\} \\
&= F_q(0) + \sum_{n=1}^{\infty} P\left\{n \text{ service completions in } \leq t | \text{arrival found } n \text{ in system}\right\} \cdot p_n \\
&= 1 - \rho + (1-\rho)\sum_{n=1}^{\infty} \rho^n \int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} dx \\
&= 1 - \rho + \rho \int_0^t \mu(1-\rho) e^{-\mu x} \sum_{n=1}^{\infty} \frac{(\rho\mu x)^{n-1}}{(n-1)!} dx \\
&= 1 - \rho + \rho \int_0^t \mu(1-\rho) e^{-\mu(1-\rho)x} dx
\end{aligned}$$

- T_q is a random variable that equals 0 with probability $1 - \rho$ and follows an $Exp(\mu(1-\rho))$ distribution with probability ρ. Thus,

$$F_{T_q}(t) = 1 - \rho e^{-(\mu-\lambda)t}, \quad t \geq 0.$$

So, T_q thus you see is a mixed random variable, so there is a mass at 0, and there is a continuous distribution beyond that. So, T_q is a random variable that equals 0 with probability $1-\rho$ and follows an exponential distribution with probability ρ. So, it is a mixture of discrete and continuous; discrete could be the 1-point degenerate random variable at point 0. And that it takes that random variable with $1-\rho$, and it takes this $Exp(\mu(1-\rho))$ with ρ, so that is what pretty much you have here. Now we obtained earlier this W q which is the waiting time in the queue and which you can now obtain from $F_{T_q}(t) = 1 - \rho e^{-(\mu-\lambda)t}, \quad t \geq 0.$ as well.

- Verification of the result $W_q = \rho/(\mu-\lambda)$ using the waiting time distribution:

$$W_q = \int_0^\infty \left[1 - F_{T_q}(t)\right] dt = \int_0^\infty \rho e^{-\mu(1-\rho)t} dt = \frac{\rho}{\mu-\lambda}.$$

  So, you can verify that result which is the same as the one we earlier obtained through Little's formula. But then the advantage of using Little Little's formula is that if you are not interested in the distribution of the waiting time and if you are interested only in the expected waiting time, then you need not do this analysis to obtain $F_{T_q}(t)$. But because in very complex situations, as you know, this will be very difficult to obtain, waiting time distribution may be very difficult to obtain. So, in that case, like if you are happy if you are interested only in the average value, then Little's law can help you to get the average waiting time. And that is a great help in that sense because when you are not able to do this, how will you compute the expected value of this distribution? But then Little's law helps you to do that. So, that is where the power or the utility of Little's law comes into the picture. This quantity you can obtain directly without obtaining the distribution that is what you know you would see.

- Similarly, by the exact same process, we can obtain the CDF of T, which is the waiting time in the system, or sojourn time or system time.

  T_q is called as waiting time in queue or delay time or queue time whatever is what T_q. For T, it is waiting time in the system or system time or sojourn time. Now, this is an Exercise!, only thing you have to note here is that if this whole system time has to be less than or equal to t. If he finds on his arrival n in front of him then including himself the service should be over by time t. So, you need $n+1$ service completion by time t. That is the only change that you have to make; once you make a similar argument, then you will obtain

$$F_T(t) = P\{T \leq t\} = 1 - e^{-(\mu-\lambda)t}, \quad t \geq 0$$

  giving us $T \sim Exp(\mu-\lambda)$, which will now be purely exponential; there is no mixed distribution here. So, it is a pure exponential distribution with the parameter $\mu-\lambda$ or $\mu(1-\rho)$; if you want to call it in that way, $\mu(1-\rho)$ or $\mu-\lambda$ is what the exponential distribution is.

  So, the system time or sojourn time in an $M/M/1$ queue is exponential, whereas the delay time in an $M/M/1$ queue is a mixture of exponential along with a degenerate random variable. So, remember, because of this argument that how many service completion happened before his service completion, we are looking at it. I will be looking at a typical customer who arrives at some point of time and when his service gets done.

- Note that $F_{T_q}(t)$ and $F_T(t)$ depend on queue discipline (as opposed to the expected measures W_q and W which are valid for a general discipline).

So, then queue discipline plays an important role, whereas the expected measures like this W_q do not need queue discipline; you can obtain W also from here by using exactly $F_T(t) = P\{T \leq t\} = 1 - e^{-(\mu-\lambda)t}, \quad t \geq 0$ here. So, for those measures, you do not need the queue discipline, whereas, for this particular analysis, waiting time distribution, you need the queue discipline. Remember, if suppose service in random models then what happens is that, anytime suppose if 10 customers are waiting on his arrival, you could imagine that any time, whose could be his service or his or her service can start. That is the situation that you haven't had. So, then very difficult to obtain this. So, for that queue discipline, it is not going to be that whatever you obtain $1 - e^{-(\mu-\lambda)t}$ is true, or last come first served, whatever is the case. It is, again, easier, but it is different from this; that is what you need to see.

- The results are applicable for a system in steady state. They are not applicable to the initial few customers who arrive soon after the start.

. So, the system in equilibrium is what we are assuming, and in that situation, this is what we get. Because only when the system is in equilibrium, the probability of observing the n customer in the system is p n, or on arrival, you will find n in the system is a_n. If the system is not in equilibrium, then that is not true, and hence the whole results will change. So, this is what; you have to keep it in mind. Now let us look at a simple example and see like what how these results can help us to analyze the thing.

Example. *(Beauty Parlour)*

- Ms. Deepa runs a one-person beauty parlour and she follows a FCFS pattern. She wants to analyze the performance of her parlour to serve her customers better. She does some book-keeping to collect the data.
- Customers arrive according to a Poisson Process with a mean arrival rate of 5/h.
  The service times are exponentially distributed with an average of 10 min.
- From the data, $\lambda = 5, \mu = 6, \rho = 5/6$. Hence, $L = 5, L_q = 4\frac{1}{6}$ *and* $L_q^{'} = 6$.
- The percentage of time an arrival can walk right in for service without having to wait at all, which happens when no one is in the shop is given by $p_0 = 1 - \rho = \frac{1}{6}$.
  That means that 16.7% of the time Deepa is idle and a customer can get into the chair without waiting.
  In other words 83.3% of the customers must wait prior to getting into the chair.
- Deepa's waiting room has only four seats at present and customers have to stand when there is no seat.
  $P\{$*an arriving customer finds no seat*$\} = P\{N \geq 5\} = \rho^5 = 0.402$.
  This says that a little over 40% of the time a customer cannot find a seat and also that 40% of the customers will have to stand upon arrival.
- The average system waiting time and line delay are given as $W = \frac{1}{(\mu-\lambda)} = 1h$ and $W_q = \frac{\rho}{(\mu-\lambda)} = \frac{5}{6}h$ (Not good!).
- The probability that the line delay is more than 45 min is determined from $F_{T_q}(t)$ as $\frac{5}{6}e^{-3/4} = 0.3936$.
- We can seek answers to many other questions from our results (with constraints, of course).

- If we assume that the data above represent the whole of the week, then Deepa needs to do some serious thinking of her service. She needs to move into a larger space for people to get a seat because you see 40 percent of the people will stand not even able to sit in there and also need to add a server (!?).

- If the data is a consolidation of different days' behaviour, she may look at the data of particular days (say, Saturdays and Sundays). Then, she may need to hire a part-time assistant only on those days (!?).
- We can only quantify the performance of the system, but it is ultimately up to the manager/designer of the system to make a decision (like with all business decisions).

So, it is all what-if analysis that generally you know you refer to that is what you know we do, but other than that, it is up to the manager though whoever is the authority to study or to analyze to improve the system. Like, how he will use it, so these all is a tool that is help us to improve the system. This is all we have for the $M/M/1$ system. We have given a broad overview that will be applicable across the different queueing models, but we will not analyze those queueing models in such detail. The main points alone we will highlight, and we will try to obtain the quantitative measures, is what would be our objective in the following the queueing models that we consider later on. So, understand that the $M/M/1$ model is so simple here, but it gives you so many insights into the system that one can analyze and discuss it. That is the part that you need to go through carefully and understand the analysis of an $M/M/1$ system so that it helps us to understand and analyze systems that are much more complex. A little complex or much more complex the similar things also what we are going to do.
Thank you bye.