Introduction to Queueing Theory

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

Lecture - 26
M/M/1 Retrial Queues

Hi and hello, everyone. What we have been seeing is the General Markovian queues with different features when you add how the system gets complex. So, we are within the Markovian framework we are looking at it. We have seen priority queues earlier. What we will see next is a feature that is also quite important with respect to many practical applications. Especially in computer and communication systems or in a network scenario or in a mobile network situation, or even otherwise even in other places, this is a feature which has lots of importance and needs to be tackled or need to be incorporated into our modeling framework. And this is called the retrial phenomena. So, what is this feature? This feature captures the phenomena of customers who make repeated attempts to access the service facility when they are not getting into the service in the first place and their immediate ah arrival time. A simple example could be a customer calling a local service centre. Say you have something to be done and you want to get connected to that; for example, you have your car which needs to be serviced. So, you are calling the car service centre there. So, what you do is you make a call, and in the first place, you see that it is a busy tone. So, which means that you are not able to get connected to that service facility. So, what do you do? You hang up. You try again after some time. You may try immediately; that is also a possibility, but when an ongoing call is there so you might think that at least 1 or 2 minutes gap, I will give, and then I will make a call again. So, you make a call again after 2 minutes or 3 minutes, or 5 minutes. Again you might get a busy signal, or you are able to get a free signal; the other side is they are picking up the phone, and then you can get the service. The second time when you get a busy signal, what you will do, fine, then I will try after sometime; like again you, disconnect and then may call after sometime. This you may do indefinitely, or after some time, you will lose your patience, and you said like I will try afterward at a later point of time. So, everything is possible. So, this is the repeated attempts to gain service are what we call as retrial, and the queues which have this feature is what a retrial queues.

So, the basic idea we will depict here now. So, what do you have? You have a service facility. So, you have a service facility here to which customers come. So, which we call it as primary customers come and if they find the service facility free or there is at least one server available to get their customers leave. It is a normal phenomena that you might have. Then, when this particular case, when if the service facility is busy. In that case, what might happen? That you would have what we call this is, orbit. So the customers in orbit, customers in orbit, they get into that. So, it could be the scenario that when the service facility is busy, they can get into orbit in this way, or they can leave, leave; this is what we say as impatience also. So, this is a general depiction of the scenario that might happen. In the scenario that you have a service facility to which the primary customers come, or customers arrive; and if the service facility is free or if there is at least one server available, then they get their service they leave it is a normal phenomenon. Now, this new phenomenon comes in only when they are not able to access this service facility immediately; what they do. They can do differently, like either they can leave; without joining, or they can get into what we call an orbit. This

means that they are inside this orbit, and after some time, after some time, then they will make a call again so, which we denote as in this way. So, from here, then they will rearrive. So, customers rearrive from orbit; that is what this phenomenon can happen. So, they can rearrive here. So, these are primary customers, and there are customers in orbit who try to access the service facility. So, as such, you see here because, whenever the service is occupied, a server is occupied. Since this happened, there has been no queueing happens here. So, that is what you have to understand here; there is no queue in front of the service facility; that is what you need to look at. So, this is the general depiction. So, the main characteristic is what we are writing it here, what we just explained.

- Main characteristics are:
  ► An arriving customer enters the service immediately if a server is available.
  ► If all servers are busy, then the customer may either
  (a) leave the system completely (impatience), or
  (b) temporarily leaves the service facility and return later to the service facility. While away, the customer is said to be in orbit.

  So, this is the word that will be used in the retrial queue to mean the customers who did not get access to the service yet but are waiting somewhere it is not in front of the service facility, unlike the other queueing system. They are waiting somewhere, and they will make an attempt to get into the service facility at a later point of time. So, those customers are what we generally termed as that they will be in orbit.

  ► Now, customers in orbit cannot see the status of the service facility that you have to remember because they will not know while in orbit whether the server is free in the middle in between before they make the retrial again.

  Say, for example, when you are making the call to your car service centre you decide to try after 5 minutes, maybe the customer I mean the service the other side, the service agent got free in 30 seconds from the time you disconnected the call. So, you do not know that that is what is the case.
  So, they will check only by rearriving. Such an event is called a retrial.

  ► Customers go back and forth from the orbit to the service facility until either service is received or they abandon the system.

- The orbit is like a queue (customers wait for the service) but the customer cannot see the status of servers.
  ♦ Servers may be idle while there is a customer in orbit.
  ♦ No concept of queueing order (service is in random order), not an FCFS discipline.

  Because a customer may come, he may be waiting in orbit, but in between, another customer might come, and at that point of time, the server may be idle, so he might get into the service immediately. So, there is no FCFS order also. And there may be 5 customers who will be waiting in orbit and depending upon who makes the next move of retrial. So, anyone might get access to the service. So, it is not that ok. Again within the orbit, there is no FCFS or any order; it is random in this case.

- One limiting case is when the time spent in orbit for each customer is instantaneous (goes to the orbit and return instantly back to the service facility).

  ■ In this case, the orbit behaves like a queue within the random service discipline.

  It is not again a first come first serve again because any one of them can gain access to the server. Though they keep, each one keeps trying on an instantaneous level.

- Now, as expected, you have some ideas with different features when you are trying to add a system that becomes more complex. If you go too much generality again, the system becomes very complex, but except for a few simple models, the other queues are generally difficult.

  I mean difficult, we say, but the difficulty level varies, and it is not that you cannot do anything with that system; no, it is not like that, but it becomes difficult and for our course when we are looking at some the how to incorporate such features into the queue that is what our main aim is. So, it is still quite a bit easy in some sense for a few simple models, but it may not be so for more general systems that you might encounter in reality and that you might want to consider it. So, you have to simplify it a bit.

So, what we will do is we will take it up one particular model, which we call as $M/M/1$ retrial queue.

- Customers arrive at a queueing system according to a Poisson process with rate λ.
- There is a single server and the service times are exponentially distributed with rate μ.

  This is what is the $M/M/1$ part, of course, with the assumption of independence between all those service times, arrival times, and so on.

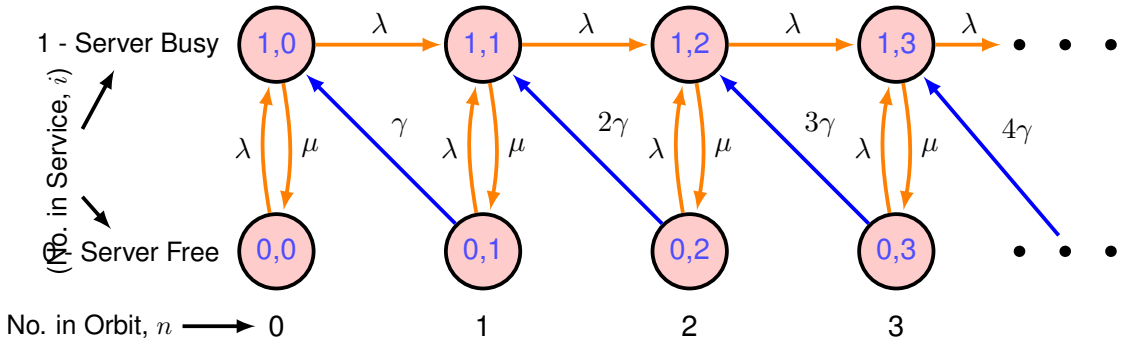
- Any arriving customer, upon finding the server busy, enters the orbit and spend an $Exp(\gamma)$ distributed time in orbit before a retrial attempt.
- Now, we assume that the customers will retry until they are served, which means there is no impatience that you are incorporating.

  If you incorporate it, then this model becomes a bit more complex and then that you can analyze, but for our simpler model, we say that every customer who has come to the system will leave the system only after the getting the service and till that time he will be in orbit, and he will keep trying to get the service.

- All interarrival times (primary arrivals), service times and orbit times are all independent.
- Let $N_s(t)$ be the number of customers in service at t. And,$N_s(t) \in \{0, 1\}$, as there is only a single server.
- Let $N_o(t)$ be the number of customers in orbit at time t. And, $N_o(t) \in \{0, 1, 2, \dots\}$.
- Then $(N_s(t), N_o(t))$ is a two-dimensional CTMC with state space $S = \{(i, n) : i = 0, 1; n = 0, 1, 2, \dots\}$.

- The total number of customers in the system at time t is $N(t) = N_s(t) + N_o(t)$.

So, $(N_s(t), N_o(t))$ is the two-dimensional CTMC that can describe the system state, but of course, we are looking at it not in a time-dependent fashion but in equilibrium. So, in equilibrium, you will have an equilibrium version of this two-dimensional with this.



- Let $p_{i,n}$ be the steady state probability of the system being in state (i, n). Then, they satisfy

$$(\lambda + n\gamma)p_{0,n} = \mu p_{1,n}, \qquad n \geq 0 \tag{1}$$

$$(\lambda + \mu)p_{1,n} = \lambda p_{0,n} + (n+1)\gamma p_{0,n+1} + \lambda p_{1,n-1}, \qquad n \geq 1 \tag{2}$$

$$(\lambda + \mu)p_{1,0} = \lambda p_{0,0} + \gamma p_{0,1}. \tag{3}$$

Now, once we write down this flow balance equation, it is up to us what we want to get out of this. Whether you want to get a complete solution of $p_{0,n}, p_{1,n}$ for all these states, or you want to get certain expected measures that depend on the complexity. In this case, it is not that difficult to obtain $p_{0,n}$ and $p_{1,n}$ quantities. So, let us see how we can do this.

- We will use generating functions to obtain the solution. Define

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{0,n} \quad \& \quad P_1(z) = \sum_{n=0}^{\infty} z^n p_{1,n}.$$

- Multiplying eqn. (1) by z^n and summing over $n \geq 0$, we get

$$\lambda \sum_{n=0}^{\infty} z^n p_{0,n} + \gamma \sum_{n=0}^{\infty} n z^n p_{0,n} = \mu \sum_{n=0}^{\infty} z^n p_{1,n}. \tag{4}$$

$$\implies \lambda P_0(z) + z\gamma P_0'(z) = \mu P_1(z). \tag{5}$$

Similarly, multiplying eqn. (2) by z^n, summing over $n \geq 1$, and adding eqn. (3), we obtain

$$(\lambda + \mu)P_1(z) = \lambda P_0(z) + \gamma P_0'(z) + \lambda z P_1(z). \tag{6}$$

- Using eqn. (5) in eqn. (6), we get,

$$P_0'(z) = \frac{\lambda\rho}{\gamma(1-\rho z)} P_0(z), \qquad \rho = \frac{\lambda}{\mu}$$

$$\implies \frac{P_0'(z)}{P_0(z)} = \frac{\lambda\rho}{\gamma(1-\rho z)}, \quad \implies \ln P_0(z) = -\frac{\lambda}{\gamma}\ln(1-\rho z) + C_1$$

$$\implies P_0(z) = C(1-\rho z)^{-\frac{\lambda}{\gamma}} \qquad \text{where} \qquad C = e^{C_1}, \text{ a constant} \tag{7}$$

- By plugging $P_0(z)$ into eqn. (5), we have

$$P_1(z) = \rho P_0(z) + \frac{\gamma}{\mu} z P_0'(z) = C\rho(1-\rho z)^{-\frac{\lambda}{\gamma}-1} \tag{8}$$

- The constant C can be found from $P_0(1) + P_1(1) = 1 \quad \Longrightarrow C = (1-\rho)^{\frac{\lambda}{\mu}+1}$.

- We finally get the partial generating functions as

$$\boxed{\begin{aligned} P_0(z) &= (1-\rho z)\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\gamma}+1} \\ P_1(z) &= \rho\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\gamma}+1} \end{aligned}}$$

- Expand $P_0(z)$ and $P_1(z)$ in a power series using the binomial formula

$$(1+z)^m = \sum_{n=0}^{\infty} \binom{m}{n} z^n = \sum_{n=0}^{\infty} \frac{z^n}{n!} \prod_{i=0}^{n-1} (m-i).$$

[The product is assumed to be 1 when $n = 0$.] Expanding $P_0(z)$, from eqn. (7), gives

$$P_0(z) = C(1-\rho z)^{-\frac{\lambda}{\gamma}} = C\sum_{n=0}^{\infty} \frac{(-\rho z)^n}{n!} \prod_{i=0}^{n-1} (-\frac{\lambda}{\gamma} - i) = \sum_{n=0}^{\infty} \left[C\frac{\rho^n}{n!\gamma^n} \prod_{i=0}^{n-1} (\lambda + i\gamma) \right] z^n,$$

The coefficient of z^n is $p_{0,n}$.

- In a similar manner, one can get $p_{1,n}$ from $P_1(z)$.

- Finally, we obtain the equilibrium probabilities as

$$p_{0,n} = (1-\rho)^{\frac{\lambda}{\gamma}+1} . \frac{\rho^n}{n!\gamma^n} \prod_{i=0}^{n-1} (\lambda + i\gamma), \quad n \geq 0$$

$$p_{1,n} = (1-\rho)^{\frac{\lambda}{\gamma}+1} . \frac{\rho^{n+1}}{n!\gamma^n} \prod_{i=1}^{n} (\lambda + i\gamma), \quad n \geq 0.$$

In this particular case, it is not so difficult to obtain this expression, but as you would see, suppose if you introduce impatience here, then you will see that things become quite complex, even this one, but it is still doable. So, this is what it is in this simple $M/M/1$ retrial queue, the steady-state system size probabilities. Now, once we have this in your hand, now, you can answer questions related to the steady-state performance measure.

- For example, the fraction of time the server is busy is $\sum_{n=0}^{\infty} p_{1,n} = P_1(1) = \rho$.

■ One can obtain this result by applying Little's law to the server as well and that will also be equal to the average number of customers in the service.
So, this ρ is actually also it will be equal to the average number of customers in service because $P_1(1)$ is what then you will get from there.

- The PGF for the number of customers in orbit is

$$P(z) = \sum_{n=0}^{\infty} z^n (p_{0,n} + p_{1,n}) = P_0(z) + P_1(z). \tag{9}$$

If L_o denotes the mean number of customers in orbit, then

$$L_o = P'(1) = \frac{\rho^2}{1-\rho} \frac{\mu+\gamma}{\gamma}. \tag{10}$$

This is a product of two terms: the average number in queue for an $M/M/1$ and a term that depends on the retrial rate γ.

► If γ is large, customer spends little time in orbit before making a retrial attempt. As $\gamma \to \infty$, they spend no time in orbit and hence are continuously able to monitor the status of the server.

So, in such a situation, as you can see, as $\gamma \to \infty$, $\frac{\mu+\gamma}{\gamma} \to 1$, so you will get only $\frac{\rho^2}{1-\rho}$ term which is basically what you would have had in the case of an $M/M/1$ queue.

This L_o is the mean number of customers in orbit. Now that we have steady-state probabilities or generating functions, we can obtain all these quantities very nicely; there is no problem.

- The mean time spent in orbit W_o (i.e., the mean time in orbit until finding the server idle and beginning service) can be obtained as

$$W_o = \frac{L_o}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} \frac{\mu+\gamma}{\gamma} = \frac{\rho}{\mu-\lambda} \frac{\mu+\gamma}{\gamma}.$$

- The average time in system W and the average number of customers in the system L can be determined similarly.

$$W = W_o + \frac{1}{\mu} = \frac{\rho\mu(\mu+\gamma) + \gamma(\mu-\lambda)}{\mu\gamma(\mu-\lambda)} = \frac{\lambda\mu + \gamma\mu}{\mu\gamma(\mu-\lambda)} = \frac{1}{\mu-\lambda} \frac{\lambda+\gamma}{\gamma},$$
$$L = \lambda W = \frac{\rho}{1-\rho} \frac{\lambda+\gamma}{\gamma} = L_o + \rho.$$

- In all cases, the service measures are the product of the analogous measure for the $M/M/1$ queue and a term that goes to 1 as $\gamma \to \infty$.

- Conversely, as $\gamma \to 0$, the expected service measures go to ∞ because blocked customers spend an extremely long period of time in orbit before trying.

Now, to see how exactly this will look like, we can just depict one; of course, we are not trying to draw in an exact manner, but just to get the idea. Suppose I look at this particular graph where γ is on the x-axis and say L_o, which is easy to depict here. So, what the figure will look like is essentially it will be something like this it will come, and this would be the $M/M/1$ case, and this is the $M/M/1$ retrial case. So, you can see here how it will behave; one can depict if you just plot it whatever you have it L here. So, the corresponding $M/M/1$ case is essentially this. Again it is intuitive that it should be like this, but again how exactly whether it is coming like this or it is coming like this. So, you have to understand. So, for which you need to look at exactly the rate. So, for which you need the expression, that is

what our expressions would help us to understand. So, this is the lower value of γ; for example, γ here results in a large number of customers in orbit because each customer now waits a long period before making a retrial attempt. Whereas $\gamma \to \infty$, it is something like they are making at every instant. In that case, it will be closer to your $M/M/1$ system; that is what you will and also, as you would see how this approaches and so on. This is, you know simple one, but then if you have more parameters, then, of course, you can have much more interesting phenomena which can be brought out from this.

So, this is all about retrial; of course, you have the steady-state distribution. So, you can talk about some more measures if you are interested; I mean, if there is anything ah what is it this many numbers beyond in orbit and so on suppose if it is the case, but at least the basic performance measure we can obtain in this manner. And this performance can be studied with respect to the retrial rate, not just the retrial rate, even with respect to other parameters you can do, but since there is a new feature that we have introduced. So, how exactly this retrial rate plays out vis-a-vis without any such retrial phenomena. So, this is another feature that can be added.

Of course, here we have studied with respect to the $M/M/1$ system, but this can be extended to, for example, in the same particular case, so you can say that you can add the impatient feature. In that case, what will happen, or you could have multi-server then, how how much is the complexity you can think or any other generalization like in terms of the interarrival time, distribution or service time distribution or any other feature. Like this one can study, and this is how the simple queues with $M/M/1$, for example, the moment you introduce retrial, you are no longer in some sense of a BDP model itself; it has become a more general Markovian model.

In the same way, in a BDP model, when you introduce more features, say, for example, you could think of server failure, the server being interrupted, the scheduling there could be, or the server may not start work until a certain number of customers in the system. Like this, there is n number of features that you can observe in reality, and in each of these cases, the moment you want to incorporate into the feature, the simple Markovian model might become a more general, a bit more complex Markovian model, and any such model like what we have exhibited so far in this general Markovian system type that you can, in principle analyze.

But, how much how far you can go and how much you will be able to handle depends on the scenario that you have seen. For example, in the retrial case, getting the exact distributions, even in a very simple situation, is not that easy. In the retrial, you could still do that, even if with a simple model retrial imposed on it. Like this, these things would vary. So, accordingly like, you have to base on that. So, this is what we you know see in the Markovian general setup. This is how one can incorporate the different features into the elementary models and see how the system performs and by incorporation of that how much the system is being affected in which way everything can be studied in that feature. So, this is; basically, we are still within the Markovian framework incorporation of features. Of course, we are we have not incorporated all the features so far. There are plenty of other features which we will not be able to cover, but of course, if you are interested, you can look deeper into this case. So, we end our discussion of this retrial here we will see in the next lecture.
Thank you. Bye.