

Stochastic Modeling and the Theory of Queues
Prof. Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology - Madras

Module - 3
Lecture - 19
Introduction to Queueing (with examples)

(Refer Slide Time: 00:15)

Lec 17: Intro. to Queueing (with examples)

server

At queue is typically characterised by

- (i) Arrival process of customers
- (ii) Server characteristics
- (iii) Service discipline (FCFS, LCFS, PS, --)
- (iv) # servers
- (v) Buffer size

Buffer size

m servers

NPTEL

Now, let us give some queueing examples; let me call this Introduction to Queueing, with some examples. So, this is a course where we are going to encounter various types of queueing systems as and when appropriate. Now, I have to tell you what a queueing system is, what a queue is. So, a queue, we can think of a queue as a; so, a queue consists of a server which provides some service, and to that server, there are customer arrivals that take place, according to some counting process.

So, these arrival processes can be the arrival of customers to a ticketing counter, or arrival of phone calls to an exchange, or arrival of packets to an internet router, or any of these things. In all this, the server takes some amount of time to offer service to a customer, and when a customer is being served by the server, other customers have to wait for their turn. This is a feature that is common when you go to renew your passport or you just call customer service or your credit card or whatever; you tend to wait till your turn comes.

So, you can see that this has widespread applications and these counting discrete stochastic processes help us study various kinds of queueing systems. So, as we study Poisson process or renewal processes or Markov chains, we will constantly be applying them to learn more about queueing systems; throughout the course, we will do many examples. So, in its simplest setting, we can think of a server and some sort of a buffer which stores awaiting customers.

And arrivals occur to this queue as some counting process basically. So, this is time, and customers come in continuous time according to some counting process. And the server has some, takes some time to serve each customer. And these customers; so, there is usually, if the server may be empty, which means there is nobody being served; or there could be somebody being served, and if somebody comes during the service interval of a customer, they will just queue up and they wait for service.

Now, we have to think about what characterises a queue like this. So, a queue is typically characterised by the following: So, it matters what the arrival process of customers is. So if a lot of arrivals take place, you know that there will be a longer queue. And the statistical nature of this arrival process will obviously have some bearing on the statistical properties of how many people wait and how long you wait and so on.

So, this is clearly important, the arrival process of the customers is clearly important. Then the server characteristic; which means, how long does it take to serve a particular customer; does it take a deterministic amount of time? Do I serve everybody in 10 minutes? Well, that is one possibility, but often it takes, that is also variable, right? So, you may hold a call for longer or shorter; so, that is also often a random variable.

So, the service time; so, the amount of time that the server spends serving a particular customer is also a random variable. And that random variable also will impact the evolution of this queueing system. Of course, there could also be service discipline. So, in some situations you may serve; I mean, the logical way to serve in a queue like this is to serve the person who comes first, first come first serve.

But not all systems work like this. Sometimes you may serve the last person who comes in; sometimes you may serve all existing customers at the same rate; all these possibilities exist. That also matters. Although we will study, first come first serve; so, we will probably study first come first serve; last come first serve. There are other disciplines like processor sharing and so on.


But I think that, when nothing is specified, people usually assume that this is first come first serve. What else matters? The number of servers could also matter. So, for example, there may be multiple ticketing counters and you could just be queuing up here. So, they all serve different customers. This is like an airline check-in counter. So, the moment this guy finishes checking in, the next person in the queue will be directed to the server which just finished serving somebody.

So, there is only one queue, but the number of servers is many. So, this could be some m server system. So, you do not have to have this picture where there is only 1 server, it could have multiple servers, but just 1 queue. So, this is common in call centres; this is common in airline check-in. So, call centres all, there are multiple people who usually attend these calls, and you get routed to the first person who becomes free; you know, you must have noticed this.

You do not always get the same person; you know this, right, when you call your credit card or something. Good. Also, the number of customers who can wait, which can be called buffer size. So, buffer is a terminology that comes from the communication network's world, where in the internet routers, you store a, buffer up a certain number of packets in a router. But this could just be the size of a hall where you wait or whatever, it does not have to be a buffer in the sense of a buffer on the internet.

So, buffer size also matters. So, if the buffer gets full, you are dropped. If your booking hall is full and you cannot even enter it, you will go away. So, all these things matter in the way the system evolves. So, queueing theorists; queueing theory is a field by itself. And queueing theorists have evolved a certain pithy notation to talk about what a particular queueing system behaves like. They use a slash notation.

(Refer Slide Time: 10:00)



(i) Server characteristics (Distribution of service time)
 (ii) Service discipline (FCFS, LCFS, PS, ...)
 (iii) # Servers.
 (iv) Buffer size.


Buffer size
 ← →
 → □ □
 m servers.

Servers
 ← Buffer size

"Slash notation"
 Arrival Service Distribution

Eg (i) $G/G/m/n$
 ↑ ↑
 "General"

(ii) $G_I/G_I/1$ (Renewed arrivals, General but iid service times across customers
 Arrival process is indep of service process.)



So, the slash notation works like this. So, it goes like “. / . / .”. So, maybe here; so, this talks about the arrival process, and the second one talks about service distribution in some sense; which is item two, service time distribution, this is what I mean. And the third one is the number of servers, and this fourth is the buffer size. So, this will not make any sense to you until I actually give you an example.

By the way, this buffer size is optional, so, often you may not have the fourth dot, you may just have three; in which case, the buffer size is assumed to be infinity, there are no restrictions on how many people can wait. So, often it is just “. / . / .”. The fourth one may not be there, in which case, you assume that the buffer is infinite. So, the most general is something like $G/G/m/n$.

So, G denotes general. It could be any general arrival in any general service distribution. Now, one potential scope for confusion is, there are two G's here, right? Those two G's may not be the same. So, the arrival statistics, service statistics could be different. Just because I say $G/G/1$ queue does not mean that the arrival in inter-arrival times or whatever has the same statistics as the service time; no.

G just means general, and these two G's could be different. And of course, m servers and n is the buffer size. Usually, some authors prefer to say, when you have independent; so, when the

service times across different customers are independent across customers and the interarrival times of the various customers are also independent, like in the renewal process; some authors usually write GI/GI .

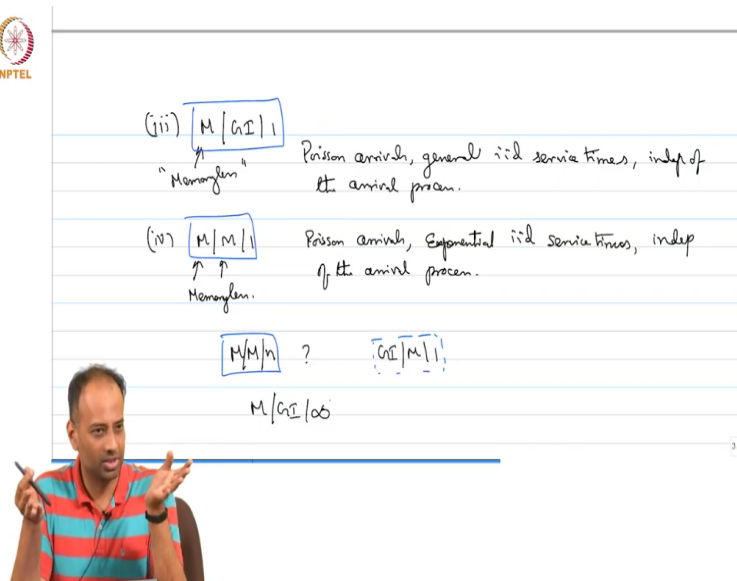
To say general, you can have any distribution, but across users. So, you could have $GI/GI/1$ queue, which means that renewal arrivals, general but IID service times, and of course one server in this case. And I have not put a fourth slash, meaning that the buffers are, I do not have any restrictions on how many people can wait. So, $GI/GI/1$ queue means renewal arrivals general, but IID service times across customers.

And also an implicit assumption in all these queueing systems is that, arrival process is independent of the service process. What does that mean? So, the amount of; so, the arrival time of a customer says nothing further about how long the customer will take at service. So, for example; so, by looking at my arrival time, you can not tell whether I am going to be a, let us say a big packet or a small packet, in a communication network.

So, it is usually in this notation $GI/GI/1$ queue, means that the arrival process and the service time random variables are independent. If they are correlated, they could; I mean, it is still a queue, it is just that this notation usually does not capture that scenario. Usually; I should mention that this service discipline is mentioned separately, it is usually not a part of this notation, slash notation.

So, you will typically say $GI/GI/1$ queue with last come first serve, or first come first serve, whatever queue. It is not included in this slash notation usually. And often, when there is no service discipline specified, you should assume first come first serve, just like when there is no buffer size specified, you assume infinity.

(Refer Slide Time: 16:25)



So, let me give you a few more examples of this notation. So, when I say $M/GI/1$ queue for example, this means memoryless; M stands for memoryless. So, in this case, then $M/G/1$ queue, we have Poisson arrivals; memoryless arrival process, which is Poisson; Poisson arrivals and general IID service times which are independent of arrival process. **"Professor - student conversation starts"** Yes.

In this notation, when I say GI/G or $M/GI/1$, the notation implies that the service times across these customers is independent and identically distributed. Sometimes, some authors do not even use I; when they say $M/G/1$ queue, they mean $M/GI/1$ queue; but potentially they could be correlated or they could even be non-identically distributed across customers. This notation does not, $M/GI/1$ means IID across customers, and independent of the arrival process.

It could be any distribution, but this is the same distribution for each customer and independent across customers and independent of the arrival process. That is what this notation means; I am not saying that this is true in all queueing systems. I am just giving some; see, if you have the; the most general queueing system where everything is correlated with everything else is impossible to study.

So, these are just examples of queueing systems we can do something mathematically about. **"Professor - student conversation ends"** The simplest queue mathematically is the so

called $M/M/1$ queue. So, this means, both are memoryless, meaning that the interarrival times are exponentially distributed as Poisson arrivals. So, this means, Poisson arrivals, exponential IID service times independent of arrival process.

$M/M/1$ queue means, arrivals are Poisson, the time it takes to serve any customer is an IID exponential random variable, which is independent of the arrival process. So, now you can make up stuff, I mean, now you can; what does $M/M/n$ mean? Poisson arrivals, exponential service times, n servers. If I had $M/M/n/m$, we mean the buffer size is m , and so on. So, these you can talk about.

What is a G; maybe I should write $GI/M/1$. What kind of a queue is that? Meaning renewal arrivals, general independent identically distributed. So, this is general renewal arrivals exponential service times and one server. So, this is how a queue is specified. I think it takes a little bit of getting used to, but once you specify these three things or four things across slashes, you know what you are talking about.

And sometimes, you may have to specify the service discipline. **"Professor - student conversation starts"** Correct. So, when I say $M/M/n$, no matter which server the job goes to, the customer goes to, he or she faces exponential service times, which are identically distributed across servers, meaning that each server is working at the same speed. Of course, it is possible that the; in a call centre, for example, if you have a situation like this;

So, her question was that; so, if you have $M/M/n$ kind of a setting like this, you could potentially have one person in the call centre who is more efficient than another person in a call centre, in which case the; if you went to the more efficient person, you will get out sooner, right? So, that is not captured in the, if I just say $M/M/n$ queue. Then it goes, I mean, it is assumed that all the servers have the same service rate, service characteristics.

See, it is not that these more complicated scenarios do not arise, in fact, they do arise. All I am saying is that this notation does not capture those things. In fact, the call centre example itself is a very important example, you want to know how many staff you put for what kind of

load. Of course, your staff are not identical, some are better than the others. So, just think about it, right?

If one staff is extremely good; so, let us say this first staff, the first person you have put in your call centre is extremely good and is clearing jobs very fast; most of the waiting jobs will go to her, which means she will be overloaded. So, you are penalising her for working very well, unfortunately. So, these kinds of problems do arise, and people have studied these kinds of issues.

So, if you want to study call centres properly, $M/M/n$ where all the servers are identical is not; it is a good first order model maybe, but you should really be considering heterogeneous servers and all that. Good. So, all I have said is that, there is some notation. So, from now on, if I tell you what a $M/G/1$ or $M/G/n$ is, you know what I am talking about, that is all.

"Professor - student conversation ends"

So, in this course, we will study, $M/M/1$, we will study a lot; $M/M/n$ also we will study; $M/G/1$ we will study; $GI/M/1$ Also, I think we will at least have homework, let us say; these queueing systems, we will cover at some point. You could also have, the number of servers could actually be infinite. So, what does that mean? There is no waiting. It is like having a call centre which is so overstaffed that nobody has to wait.

So, that kind of a queueing system is also possible. In that case, you speak of an $M/GI/\infty$.

"Professor - student conversation starts" No, it is an arrival process; now, the service process is, it could be general, IID across customers independent of service times; it is just that you do not queue up, you do not wait for service, you just spend time in service.

There is nobody that you have to wait for; as soon as you come, there are infinitely many servers, so, you go to a server straightaway; but still you can talk about how many people are being served at any given time. So, this is some sort of a limit; $M/G/\infty$ is some sort of a limit of $M/G/n$, where n becomes very large; or $M/M/\infty$ is some sort of a limit of $M/M/n$, where n becomes very large.

So, these, these kinds of systems we can study. **"Professor - student conversation ends"** Is the notation clear? Any questions? See, generally, if this notation, if there is something in the problem, which says that the servers are heterogeneous or the service discipline is not FCFS whatever, the problem or the exercise you are dealing with should say so, okay? We will say last come first serve $M/G/2$ queue in which servers have different rates, we will say so.

If you do not say anything, assume that buffer size is infinite, assume that all these servers are identical speeds and all that. And you can assume independence between arrival and service process always. And often, this GI is dropped, I is dropped. Usually people just speak of $M/G/1$ queue or $M/G/\infty$ queue, and it is assumed that the service times are independent across customers. In fact, Gallager's book just uses $M/G/1$, $M/G/\infty$.

(Refer Slide Time: 27:06)

The slide contains handwritten notes on a lined background. On the left is the NPTEL logo. The notes are as follows:

- At the top, a line with a dot above it is labeled "Homog. PP." and "Poisson arrivals, general iid service times, indep of the arrival process."
- Below that, a box contains $M/M/1$. Above it is "Homog. PP." and to the right is "Poisson arrivals, exponential iid service times, indep of the arrival process."
- Below $M/M/1$ is another box containing $M/M/n$. Below it is "Homog. PP."
- To the right of $M/M/n$ is a question mark and a box containing $GI/M/1$.
- Below $M/M/n$ is a box containing $M/G/\infty$. To its right is an arrow pointing left and the text "Can be studied using Non Homog. PP."

So, as it happens, this $M/G/\infty$ is the queue that can be; so, this can be studied using a non-homogeneous Poisson process. So, the first queue we will study is the $M/G/\infty$ queue. So, mathematically, the simplest queue, the nicest queue is $M/M/1$, which is Poisson arrivals exponential service times, but to study, it is somewhat ironic that this is something that we will study towards the very end, because we need to study continuous time Markov chains before we can study $M/M/1$; but it turns out $M/G/\infty$ can be studied just using a non-homogeneous Poisson process.

It is not the simplest queueing system from the point of view of its structure, but because it is a clever way to look at it using a non-homogeneous Poisson process, it turns out you can actually study it without doing any of these Markov chains stuff. All the other queues will involve some kind of a Markov chain, which is what we will study for the second part of this course.