Introduction to Queueing Theory
Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 03
Little's Law, General Relationships

Hello, everyone. Now that we have learned how to denote a particular queueing system in terms of Kendall's notation, we will see some "General Relationship" that holds across various queueing systems in general. So, that means what? For example, if we say $G/G/1$, we mean that generally distributed inter-arrival times and generally distributed service times with a single server or with multiple server queues irrespective of the distribution, there will be certain results that hold in general. Of course, as you narrow down to smaller groups of systems, again, you may have some more relationships that might hold, but what we are concerned about now is the main ones that hold in such generality, either single server scenario or multiple server scenario in general. Before we get into that, let us give the notations that we might use throughout the course, and the notations that we are introducing will be used throughout the course. So, and hence you pay attention to this because later on, you will not say what these quantities are because this is where you get started.

The first one is $N(t)$. So, this is essentially the number of customers in the queueing system at time t is what we denoted by $N(t)$, the number in the system at time t. So, henceforth we may not say the number of customers in the system; when we say that number, it always means it is the customers. And, when we say N, it is a number in the system; as I already pointed out, system means queue plus service, and $N(t)$in some limiting sense is what we call it as N without the dependency on this t is what this $N(t)$ would be. Sometimes instead of this $N(t)$, like, you might also write N_t just for notational convenience because you do not need to write one more bracket to write within this. So, this would also mean exactly the same thing that you might use here. And, these random variables in some limiting sense, so, here we have written down in this form; so, that means, one needs to make sense of what is this limiting means. So, basically, we are looking at the distribution of $N(t)$ in the limiting sense is what would be the distribution of N, so that is the N that we will be interested in this which means that when the system is in operation for a long time, then you are looking at the number of the system at that point of time. So, that is what is this N in some sense. And, as opposed to the number in the system, if I am looking at the number in the queue, then the notation would be $N_q(t)$ which means the number of customers in the queue means excluding the ones that are getting served currently at time t is what this $N_q(t)$ and its limiting this quantity is what simply will be N_q. And, T will be the time a typical customer spends in the system; T means the time a typical customer spends in the system. T_q is again a time a typical customer spends waiting in the queue.

So, this is only waiting in the queue; this is the system. So, this means this is basically the sojourn time that we talked about or system time that we talked about, and T_q is queueing time or waiting time in the queue; T is waiting time in the system, T_q is waiting time in the queue and its average quantities. The N number in the system this average is represented by L, the number N_q in the queue average is L_q, the sojourn time T the average is W and the queueing time T_q the average is W_q is what would be the case. Then $p_n(t)$ would be the $P\{N(t) = n\}$ is what we denoted by $p_n(t)$ and this as t tends to ∞ what you would get what you call p_n and that would correspond to the $P\{N = n\}$. So, this is what you would say in the limiting sense or when the system is in some steady-state or equilibrium after a long time of things; that is what you have here. Now, there will be other notations that we will pick up as we go along, but the main goal is to specify and analyze $\{N(t), t \geq 0\}$, the number in the system process, and its related processes because that is how one starts the analysis. So, whenever you are able to describe it for $N(t)$, which is a stochastic process, so, whenever you are able to describe it for $N(t)$, then you have a complete description, but many a time, that is difficult. So, what one looks at is the distribution of N, its limiting behavior, which means rather than working at $p_n(t)$, you try to get p_n. If you are able to get this $p_n(t)$ and that is what is called the transient solution or time-dependent solution because there is a dependency on this time t, transient solution, and if you are getting this, then that is what is referred to as a steady-state solution or the equilibrium solution, and that is what would be called in that scenario.

So, how the system behaves, whether you want to look at the transient behavior of the system or you look at the steady-state behavior of the system, for most of our course, we will be concerned only with the steady-state behavior of the system. Briefly, for the simplest of the model, we will highlight how the transient behavior or how the transient solution can be obtained, and from there, you know you will know how complex it is whereas, you know it is easy to much easier to work with $Steady - state\ behaviour$. And, many a time, as you know, it would be sufficient that if you know the steady-state behavior then what you wanted to know when the system is in operation for a long time things would have settled down to some level, and steady-state would exist, and in such scenario, you would be interested mainly in $Steady - state\ behaviour$. So, our main interest would be in steady-state behavior, pointing out at one point in time for the simplest model what these kinds of things would actually mean how you know one would do that thing. But, again, this is only like is much easier as compared to this ok, but many a time things may not be sufficient that if you know only the steady-state behavior alone that you do not have the complete picture of what is going on in the system. Because there are many systems that never reach a steady-state, and there are or even if it does not reach, then how it is getting exploded in some sense can be studied only if you know the transient solution. So, there are reasons why you know one needs to know the transient analysis or transient solution, but you know we will not go into that aspects, but our concern would be on the steady-state behaviors that we might see. Now, a very fundamental relationship that exists in queueing theory and which is being extensively used throughout is what is known as "Little's law," or "Little's formula" or "Little's theorem," whatever you want to call it. So, that "Little's law" what does that mean is what is given by this very simple equation. We all remember we have already defined what this L is. L is the mean number of customers in the system, W is the mean sojourn time in the system, and we have not said what this λ is. λ is the average number of customers in the, sorry, average rate of arrivals of customers to the system. So, the average rate of arrivals per unit, how many arrivals have come this what on an average is what is we denoted by λ. So, this Little's law relates these three quantities using this simple relationship. So, as you know, this L and λ and W what that means is what is given here. And this is a very general result and can be applied to many systems, many stochastic systems beyond queues too. And it is basically the conservation of any conservative system with inputs and outputs; you can use this result. A Conservative system means one that does not create or destroy users in between. So, the generic setting for the applicability of the Little's law, of course, holds under certain assumptions, but there is a lot of work on with different sets of assumptions and the relaxation of those basic sets of assumptions under which this might hold; anyway for our systems, this should not pose a problem. So, we will take it as given.

Now, you have a system which is given by this blue dashed line; there is some black box, and there is arrival happening with rate λ, and then something happens to the entities which arrive to get service within the system, and then they depart. So, under the very generic assumption that the system has been in operation for a long time, it is in equilibrium, and no customer is getting destroyed inside or created inside this system, and they leave the way order they come and so on. So, you will get, I mean, depending upon the system, some generic assumptions when it holds, so the generic setting under such a setting this Little's law holds. So this was developed in 1958 by Moore, but it was his student who gave the first proof of this law by the name Little, and hence this is known as Little's law, a Little's rule, a Little's formula, and so. Of course, there have been afterward that happened around 1961, and since then, there has been a lot of work extending these and generalizing these or relaxing the assumptions and so on. Now, going further like to take this up further, let us denote the certain notations furthermore we will introduce.

So, we denote by $A^{(k)}$ the time of arrival of customer k in the system where obviously, the $A^{(k+1)}$ customer would arrive after k^{th} customer. So, that is what the relationship that it will hold. So, this is basically the k^{th} customer time point. So, imagine a timeline here and then some 0 and ∞. So, you are looking at the time. Now, $A(t)$ is the cumulative number of arrivals to the system by time t. So, suppose if you are starting assuming that you are starting at time 0. So, in an interval of length $[0, t]$, how many arrivals have happened to the system is what is $A(t)$. And, $D(t)$ is the cumulative number of departures from the system by time t. So, in that same interval $[0, t]$, what is the number of departures that have happened from the system up to time t. And, W^k is the time that the k^{th} customer spends in the system; obviously, it needs to be non-negative. And, $N(t)$, which we already defined is the number of customers in the system at time t, and what would be this? This would be the number of indices k number of such $k's$ such that for that particular k, he has arrived before time t, but he has not yet left by time t. So, he will be inside the system. So, that is what these $A^{(k)} \leq t$ and $A^{(k)} + W^{(k)} \geq t$ represent. So, this number of such $k's$, if you count that, will give you the number of customers in the system at time t. So, you take the first customer whether this is satisfied; if it is satisfied, he is there in the queue, the second customer, the third customer, and so on. So, the number of such customers you count and then you will get this. $\lambda(t)$ is the average arrival rate during $[0, t]$, and $W(t)$ is the mean waiting time in the system of the customers $[0, t]$; is a mean waiting time $W(t)$.

$L(t)$ is the mean number of customers in the system $[0, t]$. This is what is the notation that we are introducing further. From this above, you can see that

$$\lambda(t) = \frac{A(t)}{t}$$

because $A(t)$ will give you the cumulative number of arrivals to the system by time t divided by time length t. So, this will be the average arrival rate, and

$$N(t) = A(t) - D(t) \quad (\text{with } N(0) = 0)$$

. Obviously, we are assuming that the system starts with 0 customers. So, $A(t) - D(t)$ is what would be this $N(t)$. Now, we will also define the following limits whenever they exist.

So,

$$\lambda = \lim_{t\to\infty} \lambda(t), \quad L = \lim_{t\to\infty} \frac{1}{t}\int_0^t N(s)ds, \quad W = \lim_{t\to\infty} \frac{1}{A(t)}\int_0^t N(s)ds$$

we will understand how exactly it comes to this. So, we are defining these three limits to exist.

Now, when this is the case, we can now precisely state the theorem, which is basically the "Little's rule" or "Little's formula," or "Little's law."

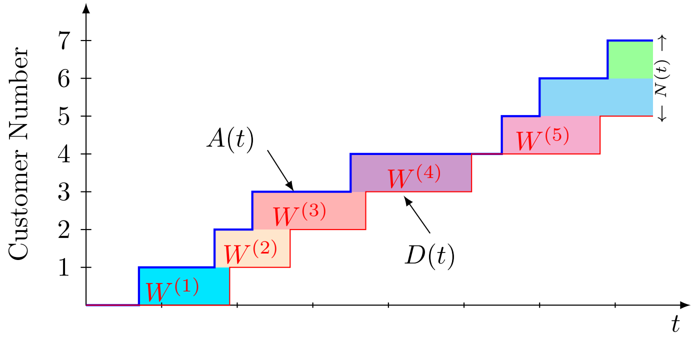
Theorem. *(Little's Law)*
If the limits λ and W exist and finite, then the limit L exists and $L = \lambda W$.

So, we remember L, λ, and W is what then we have defined whenever the limits λ and W exist, that means

$$\lambda = \lim_{t\to\infty} \lambda(t) \quad \text{and} \quad W = \lim_{t\to\infty} \frac{1}{A(t)}\int_0^t N(s)ds$$

then $L = \lim_{t\to\infty} \frac{1}{t}\int_0^t N(s)ds$ also exists, and $L = \lambda W$; this is what this result says.

So, of course, for the detailed, precise proof, we require certain assumptions and so on. So, we will not get into that. So, we will just see like a geometric illustration in some sense you can think about as a geometric proof also of this Little's law in this manner by looking at this figure.



Now, consider the figure where on the $x - axis$, you have the time, and on the $y - axis$, you have the customer number, which is the first customer or second customer, and so on. So, at time 0, you start with 0 customers, and you do not have any customer level until this point in time. At this point, a customer arrives, so this one becomes 1 here where the upper the blue bar here. So, I have to point out here that the blue bar here this part denotes the cumulative number of arrivals. So, you can see this here. And this red bar below denotes the cumulative number of departures that we had defined to be $D(t)$. So, this part is $D(t)$; the upper one is $A(t)$, which is what we are denoting by these two. Now, at this point, the first arrival happens. So, this cumulative number of arrivals is 1, and then it remains 1, until this point of time and then at this point of time the second customer comes here so, it moves to 2 and so on this. So, the first customer arrives here, the second customer arrives here, the third customer arrives here, the fourth customer arrives here, the fifth customer arrives at this point, and so on. So, this figure you will get it. Now, the first customer was in service until he arrived at this point, and since there is no one in the system, he immediately gets into the service, and then his service time, the time he spent in the system totally, is essentially up to this. And, at this point, he departed from the system. So, he departed from the system at this point of time. So, the number of departures became 1 here, and until this point of time, from this point to this point, this customer was waiting in the queue. So, now his service gets started, and then he spent the remaining possibly in the service.

So, at this point, he departs. So, the number of departures becomes 2, and the third customer departs at this point, so it goes here, and the fourth customer departs here, and at this point of time, there is nobody in the system, you remember? This means there is nobody in the system here, then arrival comes, and then things will start all over again. So, this rectangle is basically the area if you think is what represents the $W^{(1)}$ means the waiting time because it is the unit length height. So, this is what is the length of time that he spends in the system. So, we call this $W^{(1)}$, and this is $W^{(2)}$, this is $W^{(3)}$, and this is $W^{(4)}$, and this is $W^{(5)}$ and $W^{(6)}$, $W^{(7)}$, and so on it goes. Now, the difference between these two is what would be the number in the system at any point of time. So, at this point, if you see, there is only one in the system. At this point, if you look at it like, the difference between these two is 2. So, at this point, the number at the system is 2, and at this point again, it is 1, then it is 2, then like it is 1 and 2, and so on it goes. So, that is what you are seeing as the difference in the number in the system.

Now, what you observe here is that observe that the total time all customers have spent in the system $[0, t]$, that means if there were $A(t)$ customers, the total time this $A(t)$ customers spend in the system is $\sum_{i=1}^{A(t)} W^{(i)}$ which will be equal to $\int_0^t N(s)ds.$ because you see that the difference I can represent it as $N(t)$. So, if I integrate this gap so, if I want this particular thing, I can either take $W^{(1)}$, which is this, or I can integrate $N(t)$ from this point to this point up to this level with 1, so, then I will get the same quantity as $W^{(1)}$. So, I can integrate this with respect to this. So, this equality I have written, but there will be exact equality if the system becomes empty by time t again. Suppose, here, it is not the case, so, just to give the description. So, I have kept this $N(t)$ here; otherwise, suppose I am looking at such instances where the system has started with an empty system; it has ended with an empty system; this will be exact equality. If it is not, it will not be exact equality, but since we are going to look at the limiting cases as time t tends to infinity, the difference would be minimal, and the proof will go through actually.

So, do not worry about that part, but if you want to understand this, that this is how this is exact, you look at instances where the system becomes empty again, and you would assume this t to be one such instance. So, this will be exact equality; otherwise, it will not be exact equality, but things will work out even then because we are looking at the limits. Now, then $W(t)$, which is basically the average waiting time of any typical customer in the system during $[0, t]$, is basically $W(t) = \dfrac{1}{A(t)} \displaystyle\sum_{i=1}^{A(t)} W^{(i)}$. Now, $\sum_{i=1}^{A(t)} W^{(i)}$ again, we have shown that $\sum_{i=1}^{A(t)} W^{(i)} = \frac{1}{A(t)} \int_0^t N(s)ds.$. So, this is what we meant in the previous one when we said that you look at this thing. Then the average number of customers in the system during $[0, t]$ is basically this

$$L(t) = \frac{1}{t} \int_0^t N(s)ds = \frac{A(t)}{t} \; \frac{1}{A(t)} \int_0^t N(s)ds = \lambda(t)W(t).$$

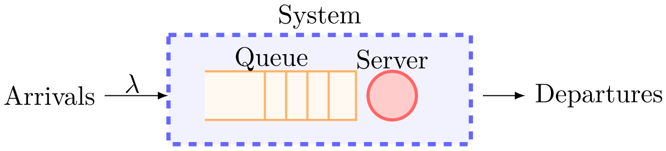
If the limits λ and W exists, then the stated relation $L = \lambda W$ follows for t tending to infinity.

So, this is just an illustration of this Little's law that you might use; you might see it here. So, the proof can be found in different ways in different places; of course, it is more than 50 years. So, you can see many places it is being mentioned here, and some variations also people have worked on, and that has been available into this. Now, let us make some general remarks about this particular Little law. First of all, you see that this is a statement about long-run averages. So, the quantities are defined as infinite limits since most of the result results that we are going to see in the course are about infinite time limits. So, this is also applicable in such scenarios, and it requires that the limits for λ and W exist. So, this precludes the scenario in which the system is growing without bound. So, the system is in equilibrium, meaning that the system has not grown without bound; it is in equilibrium which means it is a stable system in some sense. So, that only then and this would exist. So, the requirement that these imply that the condition of stability for the system is in place; that is what it would mean. The third is basically a bit more general that this requires only a system to which entities arrive and from which they depart. Now, it does not worry about what is happening to the system or the customer in the system; the so-called "system ."

In terms of how you are defining the system, a different relationship can be derived from Little's law that we will come in a moment. However, before that, we will just highlight a couple of things that this formula is, as we said, valid in great generality. If you know any two of the quantities, then the third quantity can be estimated, and you will see like it how it can be applied in a more general setup.

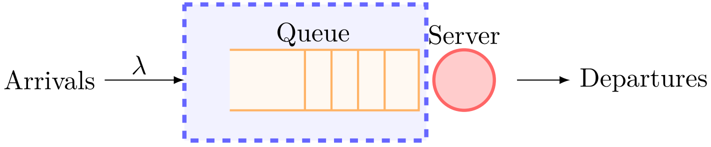
So, this has also been a special case of a more general relationship; it has been proved and relates to the most conservative system. So, for example, there is a relationship called $h = \lambda g$, which is called rate conservation law, and this is a special case of it. So, that is a generalization, and similarly, here, these moments are all first-order moments that are being related between the waiting time and the number in the system; the first-order moments are related. Now, what if the second-order moments or higher-order moments can be related, what if you can relate the distributions between these two? These are all the generalizations, but we may not require that, and as and when we require it, I will just point out what is the relationship that is needed, and it will not require much in our course at least. Of course, if you want to know more about that, of course, we can look into that.

Now, come back to what this generalization means: depending upon how the system is defined, different relationships can be derived from Little's law. That is what we said. So, the most common one, which is what we have written and we have understood so far, is that system includes both queue and the server.



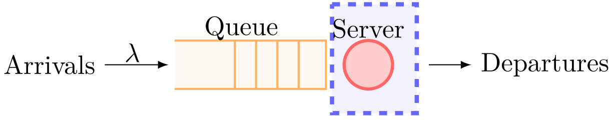
So, you have an arrival a system that has a queue and a server and to which it departs. So, $L = \lambda W$, where L here is the average number of customers in the system, meaning this whole thing, and W is the average time spent in the system. So, this is one way of looking at the most general setup here.

Now, look at it this way.

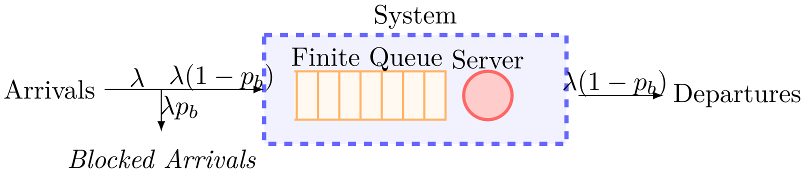


Suppose, if I look at the system as only including the queue, then $L = \lambda W$ relationship can be brought down to ${}_qL = \lambda W_q$ relationship where you are looking at scenarios where customers arriving in the queue, they are waiting in the queue, and they are leaving for service. So, the departure is actually is getting into the service, leaving for the service. So, if you define the system to be this, then the same relationship holds within this structure as well-meaning for the queue. So, Little's law implies that in this particular case, the average number in the queue is related to the average waiting time in the queue by a customer through $L_q = \lambda W_q$. So, this is only a queue when you describe the system here.

In the third case, suppose you define the system to include only the server.



So, which means that there is arrival happens, there is queueing happens you are not interested, and you are looking at the rate of arriving arrival to the server, and its departure. Suppose, if I look at this particular part as my system, whatever, we are looking at the dotted lines like this. Now, Little's law would hold true here for this particular system as well, and here L would represent the average number of customers in service, which equals $1 - p_0$. p_0 here; remember, $p_n = P\{N = n\}$. So, $n = 0$ means there is nobody in the system. So, p_0 is the fraction of time the system is empty is basically there is no customer. So, in the long-run average, it is what is the fraction of time the system is empty. So, its average number of customers in service will be equal to $1-p_0$ there is a single server here. So, it is simply $1-p_0$ that would be the average number of customers in service. Now, W would represent the average time a customer spends in service. Now, what is that? This is essentially $E(S)$, where S is the service time, the random service time that you have. So, in this particular case, $L=\lambda W$. So, this is my L, this is the W. So, $L=\lambda W$ so; that means, this is what is the case $1-p_0$ would be equal to $\lambda E(S)$. This is a very generic relationship. So, you use Little's Law to arrive at $1-p_0=\lambda E(S)$; we will use some other way also to arrive at the same result a little later. So, this is what will happen.



Now, you can introduce more complexity, say, for example, in this particular case, if there is blocking, meaning this queueing system has a finite capacity waiting systems waiting space this particular queueing system. So, it is what the finite one. So, we are just keeping it a closed thing here; otherwise, you know, we could leave it as open in this particular strategy. So, now, there is a finite queue, and there is a server. So, if this is what your system, now, out of all those arrivals that happen at the rate λ, there will be some proportion of those arrivals will not be able to enter into the system, and that is what we call blocked arrivals, and $p_b$is the fraction of arrivals that are blocked and hence does not enter the system.

So, λp_b is the proportion out of this λ. So, the remaining proportion $\lambda(1-p_b)$ of this λ would enter into the system. So, they will be waiting into the system and then go into service and departure; departure would also be this is the rate that is is what we call it simply as a throughput to the system. Now, this W and L here remember like depending upon the scenario; you have to interpret the meaning of W and L here; it is not always a number in the system number. So, L and W, you have to interpret here. So, here L and W would correspond to only those customers who actually enter the system. See out of λ only $\lambda(1-p_b)$ enter into the system now this L and W corresponds to it corresponds to only $\lambda(1-p_b)$. Suppose, if I have to take care of this for *Blocked Arrivals*, they do not enter, and you know if I have to compute waiting time or anything number in the system also they do not count and so on. So, you need to be careful when you are trying to interpret what is this W and L here.

Now, let us look at a couple of very simple examples. So, let us take a case of a maternity ward in a hospital, and what is your objective? The objective is to determine the size of the ward, and you want to look at the staffing issues, staffing scheduling of staff or whatever the staff, how many staffs are needed, and so on. Now, you have past history data available and which shows that on an average of 5 births per day, this is what the record shows. It also shows in the normal cases when there are not many complications, the 90 percent of the mothers stay for 2 days and leave this 2 could also be taken as an average, but you know we are taking it as 2, everyone is 2 and 10 percent of the cases which have little complications involved. They stay much longer, and the average length of these 10 percent of these mothers is 7 days. So, overall if I look at it, if you put together all, then my W will be 2.5 days which is what is the time the mother an expecting mother spends in the maternity ward of a particular hospital, and lambda is 5 mothers arrive per day right 5 births per day that is what you may assume that 5 per day. So, now what would be the average number of mothers who are there in the hospital at any point of time in equilibrium when the system has been in operation for a long time is 12.5 mothers.

So, this may not give you know immediately like; I will have 13 beds in my maternity ward, things will be ok. No, it is not the case. Again, you have to remember that this is an average case scenario that we are looking at. But, this could be a very well starting point for analysis. So, it gives you that you do not need to count at each time on each day how many mothers were there in the ward and keep that data with you, and then you know you run it for a long time and take its average. Of course, you will get that, but you know you do not need it because from $\lambda = 5$ *per day* and $W = 2.5$ *days* you can obtain $L = 12.5$ *mothers* very nicely. Many a time, that is how this relationship will be very useful. So, this is one particular case of an operations management problem where you are looking for this. So, one needs to keep in mind the peak arrival rates and the variability in the arrival rates before you conclude anything regarding how many beds you need and how many staff you need. But you also do not want to have like 25 beds available at all points of time. So, you may or may not need it, but it is very highly unlikely that you will need double the average number; you have to find out the optimal number. Again trade-off between utilization of the beds and whenever the patient comes, you want to make a bed available to the patient. So, it is a basic trade-off between these two quantities that you want to have. So, this is one simple example that you can use. Of course, there could be plenty of examples.

Another simple one is let us look at a semiconductor factory where semiconductors are being manufactured in extremely capital intensive fabrication facilities, so high costs business. This manufacturing of semiconductor devices starts with a silicon wafer and then building the electronic circuitry on that silicon wafers through many process steps; it is not a few it is many process steps so that to make it finally, the electronic device that you want to achieve. So, the objective here could be to determine the flow time; for example, that is, the time between a job starts and finishes. So, that is what flow time is. So, if you look at in our queueing context, this would be called a sojourn time or waiting time in the system; that is what it is. So, the past record shows that the factory on each day started on an average with a thousand wafers per day, and it does not get over within a day that the whole process is completed. So, this keeps in waiting, but every day thousand around thousands on an average is what we are taking it up. So, the work in process inventory varies between 40000 to 50000 so, with an average of 45000. So, λ is 1000 per day the arrival, L which is a number of wafers that are being manufactured semiconductors that are being manufactured is basically is 45000 is the L.

Then W is 45 days which means you start with one wafer and produce one of these semiconductors; it takes about 45 days. So, knowing of this W is critical in the planning and scheduling of these machines and whenever there is a demand and to make delivery commitments. So, you need to know this flow time or lead time. It may be like 45 days here, and there are certain processes for which the flow time could be even years. For example, if you look at the steel manufacturing process or anything of that sort, it is a very long period starting from the basic ingredient of making any product. So, it is a long time. So, it may have to be sourced from elsewhere, and everything to be counted and so on. So, in all such scenarios, getting to know what you want, getting your hands on some information about the system so that you can play with it, and then do to your advantage to make certain progress in the analysis and make a better system efficient one. So, for all those things. So, this is a very simple relationship, as we said. But, it is very useful, and you will see that throughout when you will use this kind of thing in some such systems like handling L or handling W may be very complex, but the average quantity can be obtained through the Little's law that is the beauty of this quantity. So, now, let us look at some more general results that might be of interest; in particular, this is for both $G/G/1$ and $G/G/c$ systems.

So, again you know, I could see that the lambda is what we already defined it, but you know, we could take it to be the average arrival rate in equilibrium. S, we already said that it is a random service time and $E(S)$. So, $1/E(S)$ will be the average service rate. So, this λ and μ will be throughout used straight throughout our course we will use it. So, it basically means that it is the average arrival rate, and the average service rate is what λ's and μ's. c would be the number of servers, and $r = \lambda/\mu$, is called offered load to the system. So, this is basically the amount of work arriving in the system per unit of time because λ and μ are per unit of time. So, the λ/μ is basically the offered load, basically the amount of work arriving in the system per unit of time. And, ρ, which is $\lambda/c\mu$ when there are c servers here $\lambda/c\mu$ is known as the traffic intensity or utilization of the server. Sometimes this is also called offered load to the server, offered load to the system is, or offered load to the server is ρ, but we will call traffic intensity or utilization following that. Now, with this notation, this Little's law can be used to establish a relationship among the performance measure. These are the four major performance measures for a typical simple system. L, Lq, W, Wq; if that is the case, you can see that you can relate these four quantities in this manner. Suppose L and W, you know, in the normal sense written by $L = \lambda W$. Now, if I look at W, which is basically the $E(T)$, the system time, the mean sojourn time consisting of $W_q + 1/\mu$ because my T is basically the time I spent in the queue which is T_q and the time I spent in service getting service S. So, this $W_q + 1/\mu$ is what is my T. So, $W = W_q + 1/\mu$. So, that is what you get this. So, from W, I can move to W_q using $W = W_q + 1/\mu$. W to W_q or whichever way it is, you can use this relationship. Now, once I am here, I can view only the queue alone and apply Little's law; then, I will get $L_q = \lambda W_q$ as we have already seen.

Now, once I have this L_q and this L_q again, I use the same idea here that $L = L_q + r$ because $L = \lambda W$, and $W = W_q + 1/\mu$.

So, that means, $L = \lambda(W_q + 1/\mu)$, $\lambda W_q = L_q$ and $\lambda/\mu = r$. So, $L = L_q + r$. So, this is what is the relationship between L and L_q.

Now, if I know the parameters for the model, and if I get, say, L or L_q or W or W_q one of those, then I can get the remaining three using

$$\begin{tikzcd}
L \arrow[r, leftrightarrow, "L = \lambda W"] \arrow[d, leftrightarrow, "\begin{array}{c} L = L_q + r \\ (\text{from } L = \lambda(W_q + 1/\mu)) \end{array}"'] & W \arrow[d, leftrightarrow, "\begin{array}{c} W = W_q + 1/\mu \\ (\text{from } T = T_q + S) \end{array}"] \\
L_q \arrow[r, leftrightarrow, "L_q = \lambda W_q"'] & W_q
\end{tikzcd}$$

So, this relationship is very important in that sense very useful generic one, and now whenever this $c = 1$, there is a single server system, then my r from $L = L_q + r$ I can write as $L - L_q$, which will be equal to simply $1 - p_0$ since $r = \rho$ in this particular case because my $r = \lambda/\mu$ and $\rho = \lambda/\mu$ when $c = 1$ this is one and the same.

So, my $p_0 = 1 - \rho$ which is earlier also we have obtained if you recall. So, that is the relationship that we have obtained. Now, this interpretation the ratio so, this again, and I will say that all these results are applicable for a $G/G/c$ system whereas, the last line For $c = 1$, we have $r = L - L_q = 1 - p_0$. Since $r = \rho$ here, $p_0 = 1 - \rho$. is for a single server system.

Now, this $r = \lambda/\mu$, can be interpreted as the expected number of customers in service or the average number of busy servers. It is obvious; you can see that you could give an interpretation. So, this, in many situations like this particular offered load, basically represents the minimum number of servers needed to meet particular traffic demand. So, this is what is offered load, which means this is the offered load. So, the number of customers number of servers needed to serve this particular stream of arrivals must be at least the offered load because otherwise, you cannot handle but due to variability in the offered load. So, you need to buffer up a little bit more, but how much is that little bit more is again, in a study of a different regime and different scenarios, you will arrive at different results, and one can show that that is a different matter anyway. But at least this represents the minimum number of servers needed to meet a particular traffic demand so that stability of the queueing system can be ensured. So, similarly, this could be represented; ρ, which is $\lambda/c\mu$, can be represented as the fraction of time each server is busy. Suppose, if this $\rho = 0.5$, which means that only 50 percent of the time, all the servers, whether a single server or multi-server each server, is how much is the proportion of time that this particular server is free or busy it turns out. So, it turns out that for the steady-state to exist, $\rho < 1$ or $\lambda < c\mu$, that is what is offered load should be a minimum number of servers. So, $c \geq \lambda/\mu$ which is essentially what you need to achieve stability. Now, on the other hand, if $rho = 1$ unless things are deterministic, if $rho = 1$ and things are deterministic, of course, you are meeting exactly like server is never free, and arrivals are always happening exactly when it is needed, just like just in the concept that you have. So, that is $rho = 1$ case.

But, if there is randomness, then $rho = 1$ is not going to work because due to randomness, the queue will start building up, and then things will be not in a stable situation, and obviously, $rho > 1$ means the number of arrivals if you look at here the arrivals and this is c server together like what is the service rate. So, this is the number of arrivals; when it exceeds; the average rate of arrivals exceeds the maximum rate of service of the system, then obviously, things are not going to be stable, and things will explode. So, that is what. So, for stability, or the steady-state as we call it, to exist for the system, we must have $\rho < 1$. So, that is the thing that obviously, in reality, think queue will not grow forever because queues have a certain inner mechanism by which it will turn out, even if it is you are designing a telephone network or a communication network where you are not given enough bandwidth, but you just go. What will happen when the things are full remaining calls are blocked. Customers will be complaining that my call is not getting connected, but for the system is in operation. So, what do you have a scenario? You have a blocking scenario here. So, there is a natural phenomenon that prevents the system from exploding. The system is not going to explode in any situation, even if it is traffic, even if 10-kilometer lane, cars are waiting after that there is no more waiting because people will go away; this is impatience, balking things. So, there will be other features that will happen naturally, which will make things not to explode. But what will happen if things can explode and how quickly and all those stuff one would need to look at here. So, you will naturally incorporate those features, and then you will make it the feature so that things will become stable, and you will analyze a stable situation which is a stable queue where a steady-state will exist, and then you will analyze that kind of scenarios. So, that is what we will do when we take up the queueing models one by one. These are some generic results that we have; of course, there could be some more, but it does not matter; we will come back when things would be done. So, there could be some special classes you will have some special generic results; this is a much more generic result because it is applicable for the $G/G/c$ queue. That is, as and when we encounter that, we will look into that. So, this is what we have today.
And thank you. Bye.