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

Lecture - 27
Discrete-Time Queues: *Geo/Geo/1* (EAS), *Geo/Geo/1* (LAS)

Hi and hello, everyone; what we have studied so far in the queueing models that we have considered up to now is either whether it is a BDP-based model or a General Markovian model. We have considered only continuous-time models because we have considered the time to be only continuous-time. In which what that the t the time that we are talking about was always a nonnegative real number and the interarrival times and service times were exponential or something related to exponential so far or whatever it is they are all nonnegative random variables. But they are continuous random variables; that is how we have considered, and this is how the initial development of queueing theory was that every model was considered only in continuous time and the continuous-time Markov chain theory to start with as a first level like was applied to study them and so on. Most queueing systems until the early 90s, in a way, were developed only in continuous time, there have been some works, but they were far less as compared to the continuous-time version. Because the researchers did not see compelling major reasons to study the model in discrete time, but if you look at it in one way, the continuous-time model is only an approximation in some sense because things happen in discrete time in most situations in real life. So, at a micro-level, if you go and look at it, it is always a discrete quantity. So, no matter whatever, even if you pick a rupee or time, it goes to second, or if you go to millisecond or anything like it is again, you can think of it as a discrete. But since it is so small, the difference between one unit and the next unit that you assume to be a continuous-time model for the analysis becomes easier. But what happened was that there were many situations like, for example, a telecommunication system. There all analyzed in discrete-time these days; in most of the cases, because it is based on mainly the discrete technology, the technology itself has become from analog to digital it has become. So, it becomes like a discrete one; the quantities have become discrete points. So, for practical measurement purposes, it is considered discrete, especially in modern communication systems, which are more digital and where we work in the time slot. So, the time is basically slotted; suppose you will look at an ATM network, and you have the packets or cells with a certain number of bytes that are created and sent, so the duration is taken to be one slot. So, to transmit to one packet like how much it takes, it is basically slotted in that way. So, naturally, the time is slotted in such situations or even otherwise; if you look at it, you may not observe any queueing system in a continuous manner; you will observe it every minute, every hour, every day. So, things become discrete, so the discrete-time models are appropriate in many situations when you want to see them. The reason why initially like much attention was not paid was that one could analogously analyze a discrete-time model; for every continuous-time model, there is a discrete-time version of it one can create, and the analogous analysis can be done, and if in some sense the limiting quantity if you take it then it will become the continuous-time model. So, that is how you know it was viewed, but there are reasons as to why this may not be really the case; we will come up later. So, the reason in the last, say 30 years or so, because of the nature of the problem that you encounter required that you look at the system as it is. That means, in discrete-time rather than as an approximating continuous-time model. So, and hence more emphasis more analysis was done in the last say 30 years majority of the discrete-time model works were in that and still continuing because of the demand in that. So, to analyze that so we have to get into the framework of discrete-time queueing models, that is what we call it again; only in this lecture we will concentrate on the discrete-time queues; we are not going to consider anything further for any of the other models we will come back to the continuous-time at a later stage after this lecture. But this gives you an idea about how the analysis could be done in discrete time. Whether it is analogous or not analogous to the continuous-time and just to give a flavor of what is a discrete-time queue and how one looks at it in that scenario. So, that is what is the objective of this lecture. So, to start, we have to take this simple model again in discrete-time, which is a discrete-time birth-death process.

- In the discrete-time version of a BDP, we assume the following:
  ► the birth-process is a state dependent Bernoulli process, and
  ► the death-process follows a state dependent geometric distribution.

  Of course, it is not really a process; basically, the duration follows a geometric distribution.

  [Note: A simple process $\{S_n, n = 0, 1, 2, \dots\}$ is called a **Bernoulli process**, if $S_0 = 0$ and $S_n = \sum_{i=1}^{n} X_i$, where the increments X_i's are IID Bernoulli random variables.]

- The parameters are :
  ▶ $\lambda_i =$ Probability that a birth occurs when there are $i \geq 0$ customers in the system.
  So, the time to birth, then naturally, will become a geometric distribution because, in each time unit, it has a probability λ_i of there is a birth; if you look for the occurrence of birth as a success, so, how many failures preceding that. So, you will generally have a geometric distribution comes into play.

  ▶ $1 - \lambda_i = \bar{\lambda}_i =$ Probability that no birth occurs.

  So, in this lecture, wherever the bar is there, it means $1 - \lambda_i$ which is the probability that no birth occurs in that slot.

  ▶ $\mu_i =$ Probability that a death occurs when there are $i \geq 1$ customers in the system.
  ▶ $1 - \mu_i = \bar{\mu}_i =$ Probability of no death occurs.

- Assumptions for the model:
  - The arrivals are governed by the Bernoulli process with parameter λ_i, i.e., the interarrival times are IID geometric random variables with parameter λ_i.
  - The service-times, independent of arrivals, are IID geometric random variables with parameter μ_i.

- Let the state space $S = \{i : i \geq 0\}$ denote the number of customers in the system.

- Then the transition probability matrix P is:

$$P = \begin{bmatrix} \bar{\lambda}_0 & \lambda_0 & & & \\ \bar{\lambda}_1\mu_1 & \bar{\lambda}_1\bar{\mu}_1 + \lambda_1\mu_1 & \lambda_1\bar{\mu}_1 & & \\ & \bar{\lambda}_2\mu_2 & \bar{\lambda}_2\bar{\mu}_2 + \lambda_2\mu_2 & \lambda_2\bar{\mu}_2 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

- Define $\pi_i(t)$ (So, many a time, it is preferred to use n rather than t, but for this lecture, it will be assumed that t to be the same just like λ and μ; we

have taken the same notation) as the probability that there are i customers in the system at time t (note that t is discrete now, though we use the same notation).
And, $\pi(t) = [\pi_0(t), \pi_1(t), \pi_2(t), \dots]$.

- If P is irreducible, aperiodic and positive recurrent then you will have this $\pi = \lim_{t\to\infty} \pi(t)$ and will be satisfying the stationary equations given by$\pi P = \pi, \pi\mathbf{e} = 1$, which are all the normal theory that you have here.

- Then, $\pi = [\pi_0, \pi_1, \pi_2, \dots]$ satisfies

$$\pi_0 = \pi_0\bar{\lambda}_0 + \pi_1\bar{\lambda}_1\mu_1$$
$$\pi_1 = \pi_0\lambda_0 + \pi_1(\bar{\lambda}_1\bar{\mu}_1 + \lambda_1\mu_1) + \pi_2\bar{\lambda}_2\mu_2$$
$$\pi_i = \pi_{i-1}\lambda_{i-1}\bar{\mu}_{i-1} + \pi_i(\bar{\lambda}_i\bar{\mu}_i + \lambda_i\mu_i) + \pi_{i+1}\bar{\lambda}_{i+1}\mu_{i+1}, \quad i \geq 2$$

We get

$$\pi_1 = \frac{\lambda_0}{\bar{\lambda}_1\mu_1}\pi_0, \quad \pi_i = \prod_{j=0}^{i-1} \frac{\lambda_j\bar{\mu}_j}{\bar{\lambda}_{j+1}\mu_{j+1}}\pi_0, \quad i \geq 1,\ \mu_0 = 0$$

From the normalizing condition $\pi\mathbf{e} = 1$, we obtain

$$\pi_0 = \left[1 + \sum_{i=1}^{\infty}\prod_{j=0}^{i-1} \frac{\lambda_j\bar{\mu}_j}{\bar{\lambda}_{j+1}\mu_{j+1}}\right]^{-1}$$

♦ Note that for stability, we require that $1 + \sum_{i=1}^{\infty}\prod_{j=0}^{i-1}(\lambda_j\bar{\mu}_j)\left[\bar{\lambda}_{j+1}\mu_{j+1}\right]^{-1} < \infty$.

So, the discrete-time BDP theory is almost similar to what you would have had for the continuous-time model. Now here, as opposed to continuous time in discrete time, you have to worry about at what point of time you are looking at the system; accordingly, the analysis would vary.

- There are two types of systems, namely "**Early Arrival**" system and "**Late Arrival**" system based on different assumptions made on arrival times and departure times in relation to time slot boundaries.

So, what is the Early Arrival system and Late Arrival system this is depending upon different assumptions made on arrival times and service times in relation to the slot boundaries like how things are slotted and where arrival happens, where departure happens, and at what point of time you are looking at the system that is what is important. In continuous-time, this does not matter because, in an infinitesimal interval, there is only one thing that is possible, and it does not matter if you are looking at it continuously. So, things will happen and no issue, but here you are observing the system only at discrete time points. Now, suppose if the arrivals have happened or departure happens in between your observation point of time, how you are going to count it at what place you are going to add and what place you are going to count accordingly that this description is what we give here.

- To make our life easier, we first divide the time axis into a sequence of contiguous slots $(0, 1], (1, 2], \ldots, (n-1, n], \ldots$.

Now let us consider what an Early Arrival system is.

- In "**Early Arrival**" system, arrivals join the system immediately at the beginning of a slot and departures are recorded immediately before the end of a slot.

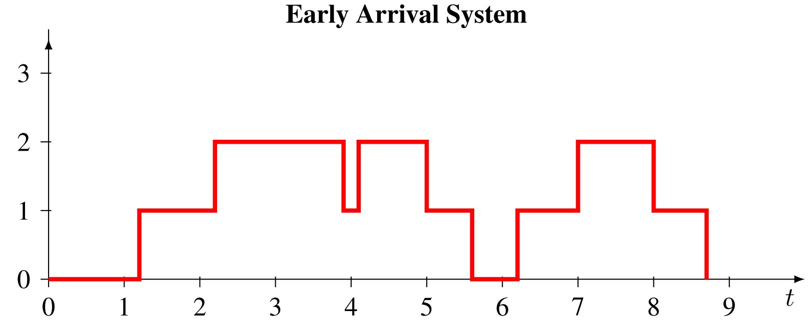
So, what happens is that suppose if this is my say n here, and this is n minus and n plus if I look at as two time adjacent to this slot. So, somewhere here is n plus 1, and this is basically this difference; this between n to n plus 1 is what one particular slot. So, now I am looking at it here. So, in the Early Arrival system, what happens is that arrivals join the system immediately at the beginning of a slot, and departures are recorded immediately before the end of this slot. So, where is the potential arrival point, the arrival point is basically here. So, these are the potential arrival points at the beginning of the interval. Now, these are, so these are basically potential departure points. So, that is what we are assuming here. So, basically, what Early Arrival means, so this is what it is basically. So, for the nth one, when you are looking at it, the arrivals up happened just after this one, and the departures happen just prior to this is what is the EAS system or Early Arrival system.

► A customer completing service at the end of slot j leaves behind those arrived in the jth slot and those waiting at the beginning of the slot. A customer starts being served always at the beginning of a slot.

- It is important to know exactly when the system is noticed.
  Let $X_n = X(n), Y_n = X(n-)$ and $Z_n = X(n+)$.

Consider a sample path example with arrival instants $1, 2, 4, 6, 7$ and service times $3, 1, 1, 2, 1$.

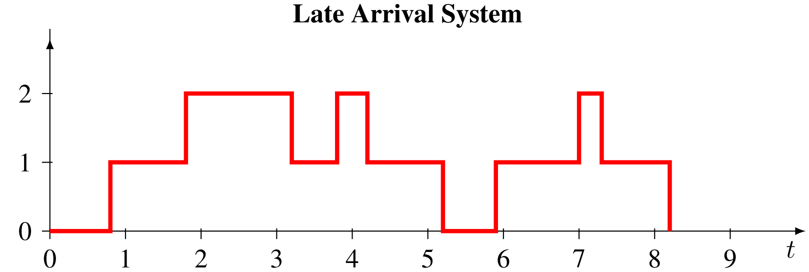
Early Arrival System



X_n :	0	0	1	2	1	1	0	1	1	0
Y_n :	0	0	1	2	2	2	1	1	2	0
Z_n :	0	1	2	2	2	1	1	2	1	0

- In the "**Late Arrival**" system, the arrivals appear just prior to the end of a slot and departures at the beginning of a slot.

Late Arrival System



X_n :	0	1	2	2	2	1	1	2	1	0
Y_n :	0	0	1	2	1	1	0	2	1	0
Z_n :	0	1	2	1	1	0	1	1	0	0

- Observe that $Z_n = Y_{n+1}$.

So this is called the LAS system or EAS system. So, it depends on when you are looking at the system and how you are sequencing the events that can happen in a particular slot, that is what these two systems can do. Now, with this idea of the basic thing, we can look at the first or the simplest discrete-time queueing model, which is called as $Geo/Geo/1$ or $Geometric/Geometric/1$, which is typically called a $Geo/Geo/1$ under EAS. Of course, this may not be written there when queueing is mentioned because in many books, say, for example, in alphas book, which we have referenced in the beginning, it is always assumed that everything, the whole treatment, is based upon one particular type which is LAS system. But when you want to distinguish, you can make this kind of distinction.

- The simplest discrete-time queueing model is the $Geo/Geo/1$ system which is analogous to the $M/M/1$ model in continuous time (and hence similar assumptions holds here too).

  It has infinite capacity FIFO discipline one server is there. The only change is that the times are now slotted in discrete time units and the interarrival time and service times are all exponential distributions.

- This system can be modelled as a special case of discrete BDP with $\lambda_i = \lambda$ and $\mu_i = \mu$.

- Arrivals and departures are according to the early arrival system.

- Let queue-size at (discrete) time t be $X(t)$, for $t = 0, 1, 2, \ldots$.
  ▶ Then $X(t)$ is a discrete-time Markov Chain with state space $\{0, 1, 2, \ldots\}$.

- Consider the case of viewing the system exactly at the point of division in the EAS system. Observe that

  $$X(t+1) = X(t) + N(1),$$

  where $N(1)$ represents the net addition to the system during the interval $(t, t+1]$.

- $N(1)$ is independent of t and has the following distribution:

  ▶ For $X(t) \geq 1$, $N(1) = \begin{cases} 1 \text{ with probability } \lambda\bar{\mu}, \\ 0 \text{ with probability } \bar{\mu}\bar{\lambda} + \lambda\mu, \\ -1 \text{ with probability } \bar{\lambda}\mu. \end{cases}$

  ▶ For $X(t) = 0$, $N(1) = \begin{cases} 1 \text{ with probability } \lambda\bar{\mu}, \\ 0 \text{ with probability } 1 - \lambda\bar{\mu}. \end{cases}$

If you change the LAS system, then the change will be only on this $N(1)$; one would see that.

- Remembering that $X_n = X(n)$, we find that $\{X_n\}$ is a homogeneous MC with transition probability matrix P given by

$$P = \begin{bmatrix} 1-\lambda\bar{\mu} & \lambda\bar{\mu} & & & \\ \bar{\lambda}\mu & \bar{\lambda}\bar{\mu}+\lambda\mu & \lambda\bar{\mu} & & \\ & \bar{\lambda}\mu & \bar{\lambda}\bar{\mu}+\lambda\mu & \lambda\bar{\mu} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

- It can be shown that the steady state solution for this MC exists when $\lambda < \mu$. The stationary equations are

$$\pi_0 = (1-\lambda\bar{\mu})\pi_0 + \bar{\lambda}\mu\pi_1$$
$$\pi_i = \pi_{i-1}\lambda\bar{\mu} + \pi_i(\bar{\lambda}\bar{\mu}+\lambda\mu) + \pi_{i+1}\bar{\lambda}\mu, \quad i \geq 1$$

Starting from the first equation, we can obtain recursively

$$\mu\bar{\lambda}\pi_{i+1} = \lambda\bar{\mu}\pi_i, \quad i \geq 0$$

- This leads to

$$\pi_i = \alpha^i \pi_0, \quad i \geq 1,$$

where $\alpha = \dfrac{\lambda\bar{\mu}}{\mu\bar{\lambda}}$.

- Using the normalizing condition $\sum_{i=0}^{\infty} \pi_i = 1$, we get $\pi_0 = 1 - \alpha$, provided $\alpha < 1$ which is equivalent to $\lambda < \mu$. Therefore,

$$\pi_i = (1-\alpha)\alpha^i, \quad i = 0, 1, 2, \ldots,$$

► The stationary distribution is geometric as in the continuous-time case (but with parameter α rather than $\rho = \lambda/\mu$).

- Performance metrics and their analysis can now be carried out in the usual manner.

We will not go into that, but we will highlight some other points here. Now, some remarks are in order with respect to this model.

So, here we viewed it at the point of, say, for example, what did we start to view the system exactly at the point of division in the EAS system, which means we are looking at these points, not at this point not at this point and so on. Like we are looking at this point, and that is typically the case in all queueing models. Once you sequence it out, then you will sequence it out in such a way that something happens before something happens after or either way, and you will always look at this, n that is a normal thing to do.

- Now, instead of viewing the system at the points of division, consider viewing the system just prior to it, i.e., we are studying $\{Y(t)\}$. In this case, $N(1)$ (for $Y(t) = 0$) changes to $N(1) = \begin{cases} 1 \text{ with probability } \lambda, \\ 0 \text{ with probability } 1-\lambda. \end{cases}$

  Denoting the steady state probability distribution by $\{\tilde{\pi}_i\}$, writing the balance equations and proceeding in an analogous manner, it can be shown that, provided $\rho < 1$,
$$\tilde{\pi}_i = \begin{cases} \rho(1-\alpha)\alpha^{i-1}, & i \geq 1 \\ 1-\rho, & i = 0 \end{cases}$$

  ► Here, the distribution is not geometric, but conditionally so given that the system is not empty.

- The case of viewing the system just after the points of division (i.e., the process $\{Z(t)\}$) does not need any new treatment and the stationary distribution is the same as the one given above since $Z_n = Y_{n+1}$.

- The stationary distribution for the continuous-time is derivable (with conditions on the parameters as $t \to \infty$) and, in all the above three cases, we obtain $(1-\rho)\rho^i$ of the $M/M/1$ model.

So, whichever way you look at it in discrete-time, it matters, but when you take it to the continuous-time limit, all of them will lead to the same limit, which is what one would expect to happen because it cannot go differently. So, that is what we have. So, normally this is only for our illustration purpose that we have given here, but normally once you define whether EAS, LAS, or any other type of system that you have. Suppose you have vacation or other retrials, then you have to sequencing it out the events that can happen in a particular time slot, and then your analysis will proceed accordingly. So, this is only for the illustration purpose of what would happen if you look at the system at another point in time and how this will vary.

- We now consider the $Geo/Geo/1$ under LAS scheme. Let the queue size at time t be $X(t)$.

- Consider the case of viewing the system exactly at the point of division. As before,

$$X(t+1) = X(t) + N(1)$$

where $N(1)$ represents the net addition to the system during the interval $(t, t+1]$. $N(1)$ is independent of t and has the following distribution:

► For $X(t) \geq 1$, $N(1) = \begin{cases} 1 \text{ with probability } \lambda\bar{\mu}, \\ 0 \text{ with probability } \bar{\mu}\bar{\lambda} + \lambda\mu, \\ -1 \text{ with probability } \bar{\lambda}\mu. \end{cases}$

► For $X(t) = 0$, $N(1) = \begin{cases} 1 \text{ with probability } \lambda, \\ 0 \text{ with probability } 1 - \lambda. \end{cases}$

- The TPM of the MC is

$$P = \begin{bmatrix} \bar{\lambda} & \lambda & & & \\ \bar{\lambda}\mu & \bar{\lambda}\bar{\mu} + \lambda\mu & \lambda\bar{\mu} & & \\ & \bar{\lambda}\mu & \bar{\lambda}\bar{\mu} + \lambda\mu & \lambda\bar{\mu} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

with the balance equations

$$\begin{aligned} \pi_0 &= \pi_0\bar{\lambda} + \pi_1\bar{\lambda}\mu \\ \pi_1 &= \pi_0\lambda + \pi_1(\bar{\lambda}\bar{\mu} + \lambda\mu) + \pi_2\bar{\lambda}\mu \\ \pi_i &= \pi_{i-1}\lambda\bar{\mu} + \pi_i(\bar{\lambda}\bar{\mu} + \lambda\mu) + \pi_{i+1}\bar{\lambda}\mu, \quad i \geq 2 \end{aligned}$$

Solving the above, with $\alpha = \dfrac{\lambda\bar{\mu}}{\bar{\lambda}\mu}$, we have

$$\pi_i = \left(\frac{\alpha^i}{\bar{\mu}}\right)\pi_0, \quad i \geq 1$$

- Using the normalizing condition, we get

$$\pi_0 = \frac{\mu - \lambda}{\mu} = 1 - \rho, \qquad \rho = \lambda/\mu.$$

and therefore, we get

$$\pi_i = \left(\frac{\mu - \lambda}{\mu\bar{\mu}}\right)\alpha^i, \quad i \geq 1, \quad \lambda < \mu$$

So, you see here in the EAS system, if you are looking at the prior to time point, the similar solution is what you are getting in a LAS system kind of thing. So, you define what system you are following, EAS or LAS, and look at the point of division to get the solution that is normally done here.

Now, at least for these LAS, one can obtain the performance measures that are the major four:

- **Mean number in the system:**

$$E[X] = \sum_{i=1}^{\infty} i\pi_i = \frac{\mu - \lambda}{\mu\bar{\mu}}\left[\alpha + 2\alpha^2 + 3\alpha^3 + \dots\right] = \frac{\lambda\bar{\lambda}}{\mu - \lambda}.$$

- **Mean number in the queue:** Let the number in the queue be Y, then the mean queue length is obtained as

$$E[Y] = \sum_{i=2}^{\infty}(i-1)\pi_i = E[X]\,\frac{\lambda}{\mu}.$$

Now, mean waiting time in the queue; of course, once you have generating functions defined and more general, one can easily get from the generating functions. But here we are trying to obtain it directly.

- **Mean waiting time in the queue:** We assume that queue discipline is FCFS. Let T_q be the waiting time in the queue for a customer and let $w_i^q = P\{T_q = i\}$, then

$$w_0^q = \pi_0 = \frac{\mu - \lambda}{\mu}$$

$$w_i^q = \sum_{j=1}^{i} \pi_j \binom{i-1}{j-1} \mu^j (1-\mu)^{i-j}, \quad i \geq 1$$

■ Reasoning: You are viewing the system in FCFS. So, you look at a customer who is arriving here; if we have j customers in the system, as arriving customer will wait i units of time if in the first $i-1$ time units exactly $j-1$ services are completed and the service completion of the jth item occurs at time i then his waiting time would be equal to i. This is then summed over j from 1 to i.

And that is precisely what you can look into this argument; this is your usual probabilistic argument that you adopt here, or one can obtain this using generating functions and things like that. But we will not go into that.

The mean time in the queue $E(T_q)$ is obtained as $E(T_q) = \dfrac{\lambda\bar{\lambda}}{\mu(\mu - \lambda)}$.

- **Mean waiting time in system:** Let T be the waiting time in the system and T equals the sum of T_q and the service time. Therefore, the mean waiting time in the system is $E\left[T\right] = E\left[T_q\right] + \dfrac{1}{\mu}$.

So, this is for the LAS system; similar things for the EAS system also you can obtain. Now, some remarks are in order. Now, we have seen here we have considered the $Geo/Geo/1$ model. We saw that the discrete-time BDP was applied to that, and the analysis was something similar to what we had done for the $M/M/1$ model in continuous time you are simply applying. You are; basically, it looked like as if you are simply replicating or doing a similar and analogous analysis in discrete time only. That is true, but that is not true for any general cases in general, but in this particular case for $Geo/Geo/1$ model, that might be true.

- The case of $Geo/Geo/1/K$ can also be handled in a similar manner as a special case of discrete BDP.

  But what is important is that the next point that we are highlighting.

- Unlike the continuous time case, multi-server discrete time queues such as $Geo/Geo/c$ model is not a discrete BDP model.

  But in continuous time, when $M/M/c$ was a multiserver model or even infinite server model, you could model with BDP. But in discrete-time, it can no longer be modelled with the discrete version of the BDP, the discrete-time BDP. So, $Geo/Geo/c$ model, the reason is that now, when more than one customer is there, and each of these servers is independent, then it is possible that there could be more than one service completion happen during a time slot when you have multiple servers in the place.

  So, it will no longer be the BDP as long as a single server is there, of course; you can look at it as a BDP special case. But again, if you want to complicate except the Geo/Geo, also if you remove it, then, of course, you are not going to get the BDP. But maybe a discrete-time Markov chain that you will have.

But here, if you have multiple; directly, you will reach, and you are no longer into the BDP framework. So, it is true that $M/M/1$ is continuous-time BDP, $Geo/Geo/1$ is discrete-time BDP. So, you can just mirror the analysis you can think of, but not every model will fit into that framework. Because we just said that $M/M/c$ is in continuous time, it can be modelled through a continuous-time BDP, but here in the discrete-time $Geo/Geo/c$ model, that c server discrete-time model cannot be done with the BDP models.

► This key difference makes the discrete time multiserver queues more involved in terms of analysis.

There are other reasons why you look at, of course, from the application point of view, this is more appropriate. So, and hence the discrete-time analysis is also relevant in most cases.

So one does, but it is not that every model also you have; what you have to keep is that you keep in mind is that every model is not just the discrete-time version of the continuous-time model and vice versa because the model could itself change in a way so, that you have to keep that in mind. So, with this, then we will end this discussion on discrete-time queues. Of course, we are not going to come back to discrete-time queues anymore just to give an overview of how a discrete-time queue and, again, the Markov chain within that you are having. So, this is we are putting it under the Markovian queueing system analysis. But as we said, more can be heard from there are many books exclusively for discrete-time queues and their application. So, one can look into that for further analysis.
Thank you bye.