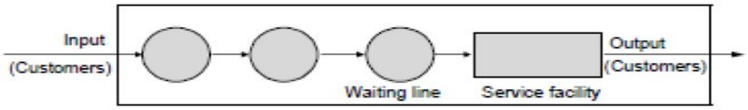
Chapter 2
Queuing system and Markov Chains

1. Queuing system: Introduction

Most systems of interest in a simulation study contain a process in which there is a demand for services. The system can service entities at a rate which is greater than the rate at which entities arrives. The entities are then said to join waiting line. The line where the entities or customers wait is generally known as queue.

The combination of all entities in system being served and being waiting for services will be called a **queuing system.**

The general diagram of queuing system can be shown as



A queuing system involves customers arriving at a constant or variable time rate for service at a service station. Customers can be students waiting for registration in college, aero -plane queuing for landing at airfield, or jobs waiting in machines shop.

If the customer after arriving can enter the service center, it is good, otherwise they have to wait for the service and form a queue i.e. **waiting line**. They remain in queue till they are provided the service. Sometimes queue being too long, they will leave the queue and go, it results a loss of customer. Customers are to be serviced at a constant or variable rate before they leave the service station.

2. Characteristics or elements of queuing system

In order to model queuing systems, we first need to be a bit more precise about what constitutes a queuing system. The three basic elements common to all queuing systems are:

1. Arrival Process or patterns

2. Service process or patterns

3. Queuing discipline

a) Arrival Process or patterns

Any queuing system must work on something – customers, parts, patients, orders, etc. We generally called them as **entities or customers**. Before entities can be processed or subjected to waiting, they must first enter the system. Depending on the environment, entities can arrive smoothly or in an unpredictable fashion. They can arrive one at a time or in clumps (e.g., bus loads or batches). They can arrive independently or according to some kind of correlation.

A special arrival process, which is highly useful for modeling purposes, is the **Markov** arrival process. Both of these names refer to the situation where entities arrive one at a time and the times between arrivals are **exponential** random variables. This type of arrival process is *memoryless*, which means that the likelihood of an arrival within the next *t* minutes is the same no matter how long it has been since the last arrival.

Examples where this occurs are phone calls arriving at an exchange, customers arriving at a fast food restaurant, hits on a web site, and many others.

b) Service Process

Once entities have entered the system they must be served. The physical meaning of "service" depends on the system. Customers may go through the checkout process. Parts may go through machining. Patients may go through medical treatment. Orders may be filled. And so on. From a modeling standpoint, the operational characteristics of service matter more than the physical characteristics. Specifically, we care about whether service times are long or short, and whether they are regular or highly variable. We care about whether entities are processed in first-come-first-serve (FCFS) order or according to some kind of priority rule. We care about whether entities are serviced by a single server or by multiple servers working in parallel etc.

Markov Service Process

A special service process is the **Markov** service process, in which entities are processed one at a time in FCFS order and service times are independent and **exponential**. As with the case of Markov arrivals, a Markov service process is memoryless, which means that the expected time until an entity is finished remains constant regardless of how long it has been in service.

For example, in the Marcrohard example, a Markov service process would imply that the additional time required to resolve a caller's problem is 15 minutes, no matter how long the technician has already spent talking to the customer. While this may seem unlikely, it does occur when the distribution of service times looks like the case shown in Figure 1. This depicts a case where the average service time is 15 minutes, but many customers require calls much shorter than 15 minutes (e.g., to be reminded of a password or basic procedures) while a few customers require significantly more than 15 minutes (e.g., to perform complex diagnostics or problem resolution). Simply knowing how long a customer has been in service doesn't tell us enough about what kind of problem the customer has to predict how much more time will be required.

c) Queuing Discipline:

The third required component of a queuing system is a queue, in which entities wait for service. The number of customer can wait in a line is called **system capacity**.

The simplest case is an unlimited queue which can accommodate any number of customers. It is called system with unlimited capacity.
But many systems (e.g., phone exchanges, web servers, call centers), have limits on the number of entities that can be in queue at any given time.

Arrivals that come when the queue is full are rejected (e.g., customers get a busy signal when trying to dial into a call center). Even if the system doesn't have a strict limit on the queue size,

The logical ordering of customer in a waiting line is called queuing discipline and it determines which customer will be chosen for service. We may say that queuing discipline is a rule to chose the customer for service from the waiting line.

The queuing discipline includes:

a) **FIFO (First in First out)** : According to this rule, Service is offered on the basis of arrival time of customer. The customer who comes first will get the service first. So in other word the customer who get the service next will be determine on the basis of longest waiting time.

b) **Last in First Out(LIFO):** It is usually abbreviated as LIFO, occurs when service is next offered to the customer that arrived recently or which have waiting time least. In the crowded train the passenger getting in or out from the train is an example of LIFO.

c) **Service in Random order (SIRO):** it means that a random choice is made between all waiting customers at the time service is offered. I.e a customer is picked up randomly form the waiting queue for the service.

d) **Shortest processing time First(SPT)**: it means that the customer with shortest service time will be chosen first for the service. i.e. the shortest service time customer will get the priority in the selection process.

e) **Priority**: a special number is assigned to each customer in the waiting line and it is called priority. Then according to this number, the customer is chosen for service.

Queuing Behavior
Customers may balk(refuse) at joining the queue when it is too long (e.g., cars pass up a drive through restaurant if there are too many cars already waiting). It is called balking.

Customer may also exit the system due to impatience (e.g., customers kept waiting too long at a bank decide to leave without service) or perishability (e.g., samples waiting for testing at a lab spoil after some time period). It is called *reneging.*

When there is more than one line forming for the same service or server, the action of moving customer from one line to another line because they think that they have chosen slow line. It is called Jockeying.

3) Queuing Notations (or KENDALL'S NOTATION)

We will be frequently using notation for queuing system, called Kendall's notation,
i.e A/B/c/N/K,

where, A, B, c, N, K respectively indicate arrival pattern, service pattern, number of servers, system capacity, and Calling population.

The symbols used for the probability distribution for inter arrival time, and service time are, D for deterministic, M for exponential (or Markov).

If the capacity Y is not specified, it is taken as infinity, and if calling population is not specified, it is assumed unlimited or infinite

For example
a) M/D/2/5/∞ stands for a queuing system having exponential arrival times, deterministic service time, 2 servers, capacity of 5 customers, and infinite population.
b) If notation is given as M/D/2 means exponential arrival time, deterministic service time, 2 servers, infinite service capacity, and infinite population.

4) Single server queuing system

For the case of simplicity, we will assume for the time being, that there is single queue and only one server serving the customers. We make the following assumptions.

- First-in, First-out (FIFO): Service is provided on the first come, first served basis.

- Random: Arrivals of customers is completely random but at a certain arrival rate.

- Steady state: The queuing system is at a steady state condition. (*Steady state is a condition of a system or a process where the variables that define its behavior are constant in time.*)

The above conditions are very ideal conditions for any queuing system and assumptions are made to model the situation mathematically.

First condition only means irrespective of customer, one who comes first is attended first and no priority is given to anyone.

Poison arrival Patterns

Second condition says that **arrival of a customer** is completely random. This means that an arrival can occur at any time and the time of next arrival is independent of the previous arrival. With this assumption it is possible to show that the distribution of the inter-arrival time is exponential. This is equivalent to saying that the number of arrivals per unit time is a random variable with a Poisson's distribution. This distribution is used when chances of occurrence of an event out of a large sample is small.

i.e. if X = number of arrivals per unit time, then, probability distribution function of arrival is given as

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad \begin{cases} x = 0, 1, 2, \ldots \\ \lambda > 0 \end{cases}$$

$$E(X) = \lambda$$

where λ. is the average number of arrivals per unit time (1/ τ), and x is the number of customers per unit time. This pattern of arrival is called **Poisson's arrival pattern**.

Illustrative example

In a single pump service station, vehicles arrive for fueling with an average of 5 minutes between arrivals. If an hour is taken as unit of time, cars arrive according to Poison's process with an average of λ = 12 cars/hr.

The distribution of the number of arrivals per hour is,

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-12}12^{x}}{x!}, \quad \begin{cases} x = 0, 1, 2, .. \\ \lambda > 0 \end{cases}$$

$$E(X) = 12 \text{ cars/hr}$$

5) Measure of Queues

We have already defined the mean inter arrival time Ta and the mean service time Ts and the corresponding rates;

Arrival rate λ=1/Ta

Service rate μ=1/Ts

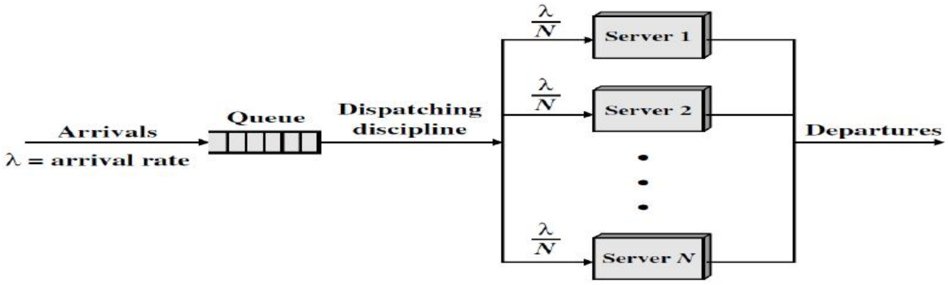
The following measures are used in the analysis of queue system

Traffic intensity: the ratio of the mean service time to the mean inter arrival time is called traffic intensity. i.e. $u= \lambda$‟Ts or $u=Ts/Ta$
If there is any balking or reneging, not all arriving entities get served. It is necessary therefore to distinguish between actual arrival rate and the arrival rate of entities that get served.
Here λ‟ denoted the all arrivals including balking or reneging.
Server utilization: It consists of only the arrival that gets served. It is denoted by and defined as
$= \lambda Ts= \lambda/ \mu$ (server utilization for single server).
This is also the average number of customers in the service facility.
Thus probability of finding service counter free is $(1 – \rho)$. That is there is zero customers in the service facility.

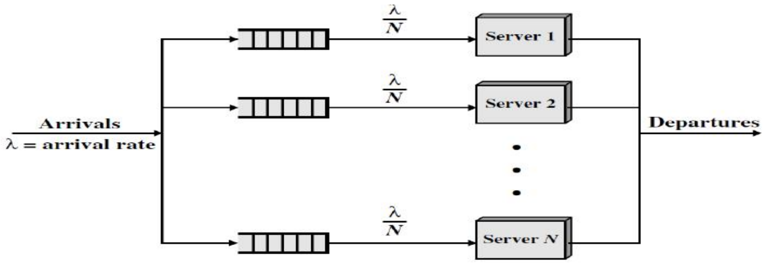
6) Concept of Multi-server Queue

In the following figure shows a generalization of the simple model we have been discussing for multiple servers, all sharing a common queue. If an item arrives and at least one server is available, then the item is immediately dispatched to that server. It is assumed that all servers are identical; thus, if more than one server is available, it makes no difference which server is chosen for the item. If all servers are busy, a queue begins to form. As soon as one server becomes free, an item is dispatched from the queue using the dispatching discipline in force.

The key characteristics typically chosen for the multi-server queue correspond to those for the single-server queue. That is, we assume an infinite population and an infinite queue size, with a single infinite queue shared among all servers. Unless otherwise stated, the dispatching discipline is FIFO. For the multi-server case, if all servers are assumed identical, the selection of a particular server for a waiting item has no effect on service time.



The total server utilization in case of Multi-server queue for N server system is
$\rho=\lambda/c\mu$ where μ is the service rate and lemda is the arrival rate.
There is another concept which is called multiple single server queue system as shown below



7) Some notation or Formula used to Measure the different parameter of queue

Two principal measures of queing system are;

a) The mean number of customers waiting and

b) The mean time they spend waiting

Bothe these quantities may refer to the total number of entities in the system, those waiting and those being served or they may refer only to customer in the waiting line.

Average number of customers in the queue L_Q is same as expected number in the system – the expected number in the service facility:

$$\overline{L}_Q = \overline{L}_S - \rho = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{(1-\rho)}$$

Average time a customer spends in the system is denoted by WS , and is equal to expected number of customers in the system at time t, divided by number of customers arrived in unit time i.e.,

$$\overline{W}_S = \frac{\lambda}{\mu-\lambda} \cdot \frac{1}{\lambda} = \frac{1}{(\mu-\lambda)}$$

Average time a customer spends in the queue (WQ) is same as average time a customer spends in the system – average time a customer spends in the server i.e.,

$$\overline{W}_Q = \overline{W}_S - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Example

At the ticket counter of football stadium, people come in queue and purchase tickets. Arrival rate of customers is 1/min. It takes at the average 20 seconds to purchase the ticket.

(a) If a sport fan arrives 2 minutes before the game starts and if he takes exactly 1.5 minutes to reach the correct seat after he purchases a ticket, can the sport fan expects to be seated for the tip-off ?

Solution:

(a) A minute is used as unit of time. Since ticket is disbursed in 20 seconds, this means, three customers enter the stadium per minute, that is service rate is 3 per minute.

Therefore,

λ = 1 arrival/min
μ = 3 arrivals/min
WS = waiting time in the system=1/(μ- λ)=0.5
The average time to get the ticket and the time to reach the correct seat is 2 minutes exactly, so the sports fan can expect to be seated for the tip-off.

Example2

Customers arrive in a bank according to a Poisson"s process with mean inter arrival time of 10 minutes. Customers spend an average of 5 minutes on the single available counter, and leave. Discuss

(a) What is the probability that a customer will not have to wait at the counter?
(b)What is the expected number of customers in the bank?
(c) How much time can a customer expect to spend in the bank?

Solution: We will take an hour as the unit of time. Thus

λ = 6 customers/hour,
μ = 12 customers/hour.
The customer will not have to wait if there are no customers in the bank. Thus
P0 = 1 – λ/μ= 1− 6/12 = 0.5
Expected numbers of customers in the bank are given by
LS = λ /(μ - λ)=6/6=1
Expected time to be spent in the bank is given by
WS=1/(μ – λ)= 1/(12-6) = 1/6 hour = 10 minutes.

8) Markov Chains and its applications

a) Markov chains and Markov Process

Important classes of stochastic processes are Markov chains and Markov processes. A Markov chain is a discrete-time process for which the future behavior, given the past and the present, only depends on the present and not on the past. A Markov process is the continuous-time version of a Markov chain. Many queuing models are in fact Markov processes. This chapter gives a short introduction to Markov chains and Markov processes focusing on those characteristics that are needed for the modeling and analysis of queuing problems.

A Markov chain

A Markov chain, named after Andrey Markov, is a mathematical system that undergoes transitions from one state to another, between a finite or countable number of possible states. It is a random process characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. This specific kind of "memorylessness" is called the Markov property. Markov chains have many applications as statistical models of real-world processes.

Formally

A Markov chain is a sequence of random variables X_1, X_2, X_3, ... with the Markov property, namely that, given the present state, the future and past states are independent. i.e

$$\Pr(X_{n+1}=x|X_1=x_1,X_2=x_2,\ldots,X_n=x_n)=\Pr(X_{n+1}=x|X_n=x_n).$$

Example; A simple whether model

The probabilities of weather conditions (modeled as either rainy or sunny), given the weather on the preceding day, can be represented by a transition matrix:

$$P=\begin{bmatrix}0.9 & 0.1\\ 0.5 & 0.5\end{bmatrix}$$

The matrix P represents the weather model in which a sunny day is 90% likely to be followed by another sunny day, and a rainy day is 50% likely to be followed by another rainy day. The columns can be labelled "sunny" and "rainy" respectively, and the rows can be labeled in the same order.

(P)i j is the probability that, if a given day is of type i, it will be followed by a day of type j.

Notice that the rows of P sum to 1: This is because P is a stochastic matrix.

The weather on day 0 is known to be sunny. This is represented by a vector in which the "sunny" entry is 100%, and the "rainy" entry is 0%:

$$\mathbf{x}^{(0)}=\begin{bmatrix}1 & 0\end{bmatrix}$$

The weather on day 1 can be predicted by:

$$\mathbf{x}^{(1)}=\mathbf{x}^{(0)}P=\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}0.9 & 0.1\\ 0.5 & 0.5\end{bmatrix}=\begin{bmatrix}0.9 & 0.1\end{bmatrix}$$

The weather on day 2 can be predicted in the same way:

$$\mathbf{x}^{(2)}=\mathbf{x}^{(1)}P=\begin{bmatrix}0.9 & 0.1\end{bmatrix}\begin{bmatrix}0.9 & 0.1\\ 0.5 & 0.5\end{bmatrix}=\begin{bmatrix}0.86 & 0.14\end{bmatrix}$$

Or

$$\mathbf{x}^{(2)}=\mathbf{x}^{(1)}P=\mathbf{x}^{(0)}P^2=\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}0.9 & 0.1\\ 0.5 & 0.5\end{bmatrix}^2=\begin{bmatrix}0.86 & 0.14\end{bmatrix}$$

In general

$$\mathbf{x}^{(2)}=\mathbf{x}^{(1)}P=\begin{bmatrix}0.9 & 0.1\end{bmatrix}\begin{bmatrix}0.9 & 0.1\\ 0.5 & 0.5\end{bmatrix}=\begin{bmatrix}0.86 & 0.14\end{bmatrix}$$

General rules for day *n* are:

$$\mathbf{x}^{(n)} = \mathbf{x}^{(n-1)}P$$
$$\mathbf{x}^{(n)} = \mathbf{x}^{(0)}P^n$$

b) Markov chain or process Applications

Physics

Marko chain systems appear extensively in thermodynamics and statistical mechanics, whenever probabilities are used to represent unknown or unmodelled details of the system, if it can be assumed that the dynamics are time-invariant, and that no relevant history need be considered which is not already included in the state description.

Queueing theory

Markov chains are the basis for the analytical treatment of queues (queuing theory). Agner Krarup Erlang initiated the subject in 1917. This makes them critical for optimizing the performance of telecommunications networks, where messages must often compete for limited resources (such as bandwidth).

Internet applications

The Page Rank of a webpage as used by Google is defined by a Markov chain. It is the probability to be at page i in the stationary distribution on the following Markov chain on all (known) web pages

Statistics

Markov chain methods have also become very important for generating sequences of random numbers to accurately reflect very complicated desired probability distributions, via a process called Markov chain Monte Carlo (MCMC) And many more.

9) Differential and partial deferential equations

Continuous model

When continuous system is modeled mathematically, the variables of model representing the attribute of system are controlled by continuous functions. The distributed lag model is an example of a continuous model. Since in continuous system, the relationship between variables describe the rate at which the value of variable change, these system consist of differential equations.

Continuous system simulation uses the notation of differential equation to represent the change on the basic parameter of the system with respect to time. Hence the Mathematical model for continuous system simulation is usually represented by differential and partial differential equations.

Differential Equations

An example of a linear differential equation with constant coefficients to describe the wheel suspension system of an automobile can be given as

Here the dependent variable x appears together with first and second derivatives single dot and double dot respectively.

The simple differential equation can model the simplest continuous system and they an have one or more linear differential equation with constant coefficients. It is then often possible to solve the model without using simulation technique i.e. we can solve such equations using analytical methods as (we have done in Numerical methods)

However when non linearity involves into the model, it may be impossible or at least very difficult to solve such model without simulation.

Partial Differential Equations

When more than one independent variable occurs in a differential equation the equation is sait to be partial differential equations. It can involve the derivatives of the same dependent variable with respect to each of the independent variable.

Differential equations both linear and nonlinear occur frequently in scientific and engineering studies. The reason for this is that most physical and chemical process involves rates of change, which require differential equation to represent their mathematical descriptions.

Since partial deferential equation can also represent a growth rate, continuous model can also be applied to the problems of a social or economic nature.

Analog Computer

Before the invention of digital computer, there existed devices whose behavior is equivalent of mathematical operation such as addition or subtraction ot integration. Putting together these device in a manner specified by a mathematical model or equation of a system , allowed us to simulate the system.

Some devices have been created for simulation continuous system and called analog computer or differential analyzer.

Digital analog simulators

To avoid the disadvantages of analog computers, many digital computer programming language have been written to produce digital-analog simulators. They allow or facilitate a continuous model to be programmed on a digital computer in essentially the same way as it is solved on analog computer. The language contains micro instructions that carry the action of addition, integration and sign changer. A program is written to link together these micro instructions in the same way as operational amplifiers are connected in analog computer.

Since more powerful digital computer and programming language have been developed for this purpose of simulating continuous system on digital computer, the digital-analog simulators are now in extensive use.

Question: Why is the server utilization should not be close to 1 or 100 percent.

Solution: The server utilization is measure of server business when the server is always busy, the server utilization is 100 % or 1. initially at the opening of the service center, the server is idle and the customer who arrive first get the service immediately but when the server become busy serving the continuous flow of the customer, the queue is formed. When the server engages continuously in service, it could not get free time on idle time. Therefore a long queue may be formed and the customer service time also increase. similarly more server utilization decreases the efficiency of the server and the waiting time of the customer also increases. the server which is always busy or overloaded server means an inefficiency system. therefore the server utilization should always be kept below 1. the server utilization more than 1 indicates a long queue and overloading of the server and it should always be avoided.

The efficiency of a queuing system can be measured using traffic intensity.

Let Ta be the mid-internal arrival time, mean arrival rate λ, Ts mean service time and μ be the mean service rate. The traffic intensity can be defined as the ratio of mean service time to the mean interval time.

i.e. T. I(μ) = Ts/Ta

It is measured in Eriang.

Due to bulking an reneging, the actual arrival rate and the arrival rate of the entities that get served is different. In that case the traffic intensity μ = λ'Ts ; Where λ' is all arrivals including loses. The actual server utilization can be calculated by removing the primes which includes the arrival rate and get served and represented as ρ.

Therefore, $\mu = \lambda' T_s$ ....(i)

$\rho = \lambda T_s$ ....(ii)

The equation (ii) gives the server utilization single server. Both μ ρ can be greater than 1. therefore a single server system is not enough to provide service in a queuing system. If n be number of server then the server utilization

$\rho = \lambda / n.\ \mu$ ....(iii)

The server utilization in equation (iii) should always be less than 1 in order to keep up with traffic flow. If it reaches 1 or near to 1 then it includes a long queue and it should be avoided.