Chapter 3
Queuing system

3.1. 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 less 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, airplane 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.

3.2. 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

3.2.1 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 memory-less, 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.

3.2.2. 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.

3.2.2.1. 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 memory-less, 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 a Markov service process would imply that the additional time required resolving a caller's problem is 15 minutes, no matter how long the technician has already spent talking to the customer. It is explained in 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.

3.2.3. 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 choose 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 forms 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.

3.2.3. Queuing Behavior

Customers may balk 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 perishable (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.

3.3. 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 potential customers to a system is known as calling pouplation:Finite and infinite)
The symbols used for the probability distribution for inter arrival time, and service time are, D for deterministic, M for exponential (or Markov) and Ek for Erlang.

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

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.

3.4. 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.

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.

3.5. 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 (T$_a$ id denoted by tou (τ))

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= λ"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

= λTs= λ/ μ (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 – ρ)

That is there are zero customers in the service facility.

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

Two principal measures of queuing system are;

a) The mean number of customers waiting and

b) The mean time the customer 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 System $\overline{\boldsymbol{L}}_{\boldsymbol{S}} = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\lambda}{\mu-\lambda}$

Average number of customers in the Queue $\overline{\boldsymbol{L}}_{\boldsymbol{Q}}$

= Average number of customers in the System − Server Utilization

$$= \overline{\boldsymbol{L}}_{\boldsymbol{S}} - \frac{\lambda}{\mu} = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Average waiting time in the System $\overline{\boldsymbol{W}}_{\boldsymbol{S}} = \frac{Average\ number\ of\ customer\ in\ the\ system}{Mean\ arrival\ rate}$

$$= \frac{\overline{L}_S}{\lambda} = \frac{\frac{\lambda}{\mu-\lambda}}{\lambda} = \frac{1}{\mu-\lambda}$$

Average waiting time in the Queue $\overline{\boldsymbol{W}}_{\boldsymbol{Q}} = \frac{Average\ number\ of\ customer\ in\ the\ Queue}{Mean\ arrival\ rate}$

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

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 kick-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,

$\lambda = 1$ arrival/min

$\mu = 3$ arrivals/min

$\overline{\boldsymbol{W}}_{\boldsymbol{S}}$ = waiting time in the system=1/(μ- λ)=0.5 minutes

The average time to get the ticket plus the time to reach the correct seat is 2 minutes exactly, so the sports fan can expect to be seated for the kick-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.

(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,

$P_0 = 1 - \lambda/\mu = 1 - 6/12 = 0.5$

Expected numbers of customers in the bank are given by

$\overline{L_S} = \lambda /(\mu - \lambda) = 6/6 = 1$

Expected time to be spent in the bank is given by

$\overline{W_S} = 1/(\mu - \lambda) = 1/(12-6) = 1/6$ hour = 10 minutes.

3.6. Concept of Multi-server Queue

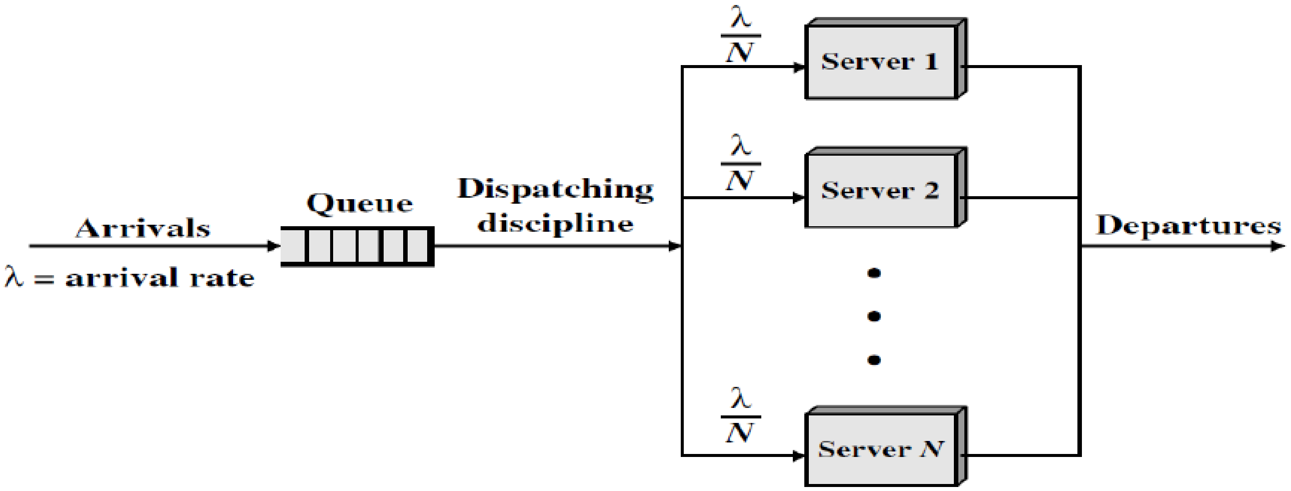


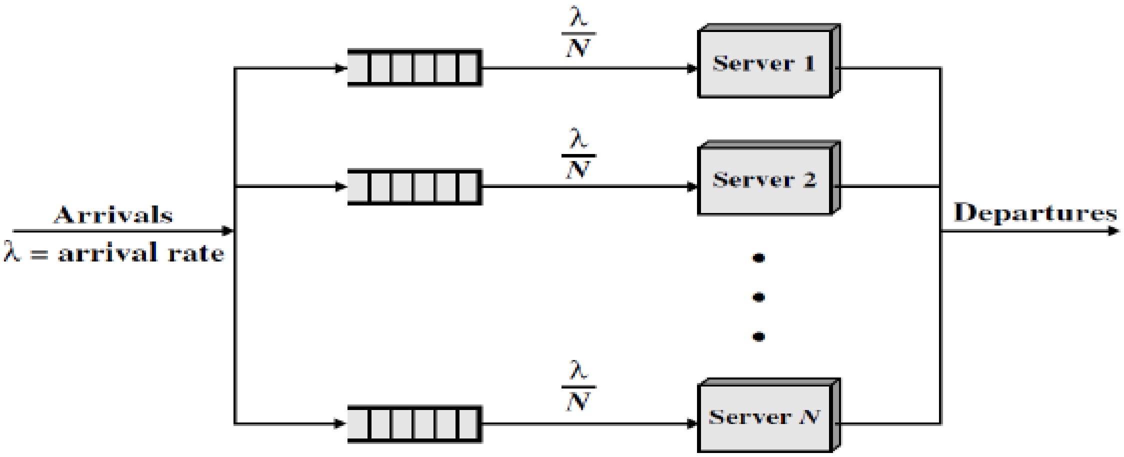
Figure shows a generalization of the simple model we have been discussing for multiple servers, all sharing a common queue. If an item arrives and at least one server is available, then the item is immediately dispatched to that server. It is assumed that all servers are identical; thus, if more than 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

$$\rho = \lambda / c\mu$$

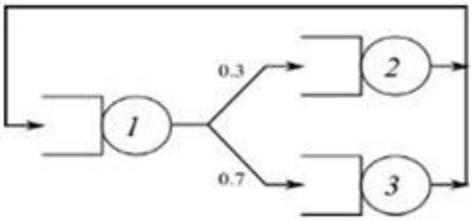
Where μ is the service rate and λ is the arrival rate.

There is another concept which is called multiple single server queue system as shown below



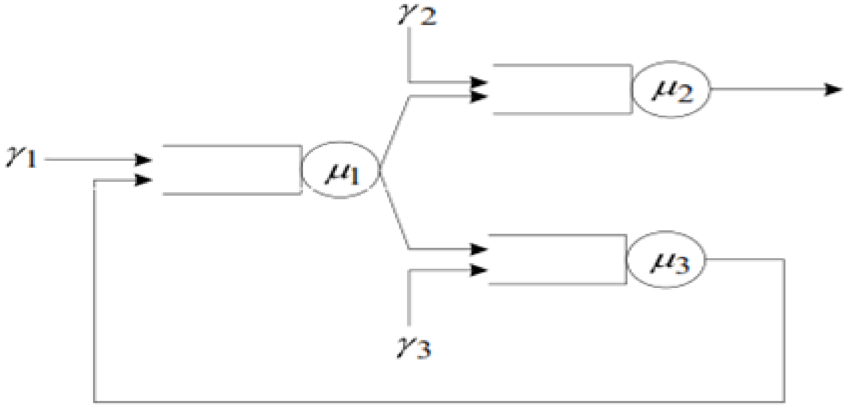
3.7 Network of Queues

- Many communication systems must be modeled as a set of interconnected queues – which is termed a queueing network.
- Systems modeled by queueing networks can roughly be grouped into four categories
  - Open networks
  - Closed networks
  - Networks with population constraints (Loss Networks)
  - Mixed network



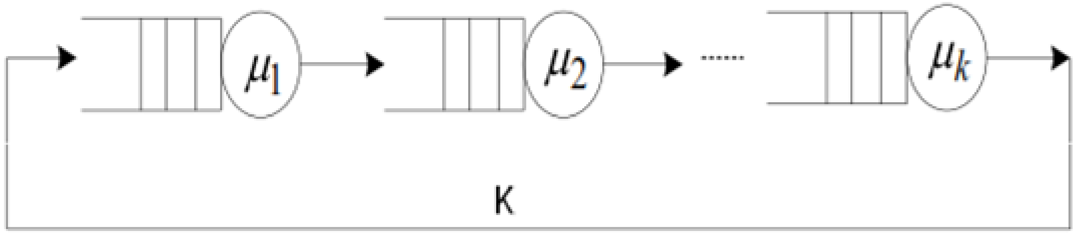
3.7.1. Open Networks

- Customers arrive from outside the system are served and then depart.
- Example: Packet switched data network.



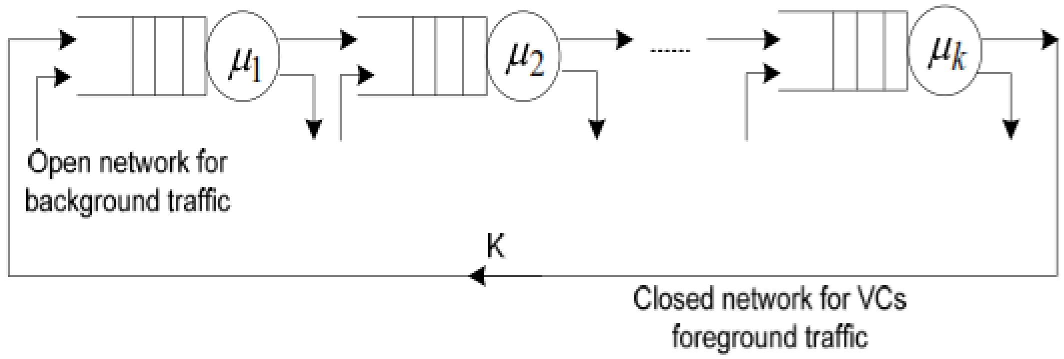
3.7.2. Closed Networks

- Fixed number of customers (*K*) are trapped in the system and circulate among the queues.
- Example: CPU job scheduling problem



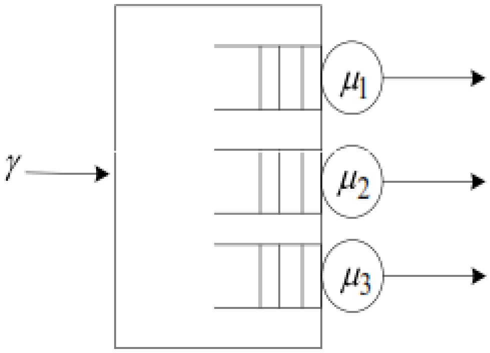
3.7.3. Mixed Networks

- Any combination of previous types.
- Example: simple model of virtual circuit that is window flow controlled.


μ_1
μ_2
......
μ_k
Open network for background traffic
K
Closed network for VCs foreground traffic

3.7.4. Loss Networks with Population Constraints

- Customers arrive from outside the system if there is room in the system. They enter, served and then depart.
- Example: queues sharing a common buffer pool – customers are lost when arriving to full system



3.8. Network Queueing

Network queueing is a very important application of queueing theory. The term 'network of queues' describes a situation where the input from one queue is the output from one or more others. This is true in many situations from telecommunications to a PC. Below is a description of some of the broad applications of network queueing:

3.8.1. Computer Networks

A simple example of network queueing is the central server network. This consists of a CPU (Central Processing Unit), storage units it can access and input devices to access it. The tasks the CPU performs are placed on queues on different criteria. Also, the storage units could have their own individual queues.

3.8.2. Network communication

There are several broad methods connected to network communication:

- **Circuit Switching**
  When a call is made from a source to a destination it must traverse several nodes along the way. Which nodes it traverses is determined by the availability of free channels along the way. Each node has a queue for calls requesting a channel. Once a channel has been opened the call can progress to the next node and wait for a channel there. The channel remains open until the source or destination (once reached) closes the call.

- **Packet Switching**
  Messages are transmitted through intermediate stages and the route a message takes depends entirely upon the current load on the system. The route allocation

is dynamic. Each stage requires a random amount of time reflecting the length of the queue at that stage.

3.8.3 Broadcasting

- **Radio Communication**
  Considering the nodes as transmitters/receivers you can treat each as having a queue for their channels. Without going into great detail of the various systems used: it is always necessary to consider the fact that to open a channel you must check to see if the two adjacent channels are also free as interference blocks transmissions. When the channels are not free it may be necessary to re-allocate communications that already have channels to make room.

- **Digital Communication**
  This is done on the basis of time slots. For a given communication link it could have several or all slots filled and no interference would take place making allocation far simpler. The aspect of nodes with queues still applies however.