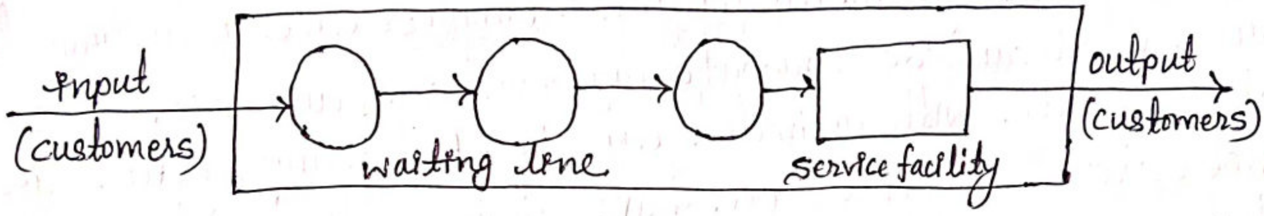
UNIT-3

Queuing System

[Imp]

Most systems 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 entities or customers wait is generally known as queue. The combination of all entities in system being served and being waiting for services is called a queuing system.



⊛. Characteristics or elements of queuing system: [Imp]

Following are the ~~three~~ basic elements common to all queuing systems:

1) **Arrival Process or patterns:** Any queuing system must work on something - customers, parts, patients, orders etc. We generally call them as entities or customers. Depending on the environment, entities can arrive smoothly or in a unpredictable fachion. They can arrive one at a time or in groups. A special arrival process, which is highly useful for modeling purposes, is the Markov arrival process.

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.

2) **Service Process:** Once the entities have entered the system they must be served. The physical meaning of "service" depends on the system. Customers may go through checkout process. Patients may go through medical treatment and so on. From a modeling point of view, we care about whether service times are long or short, and 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 etc.

escape this only if characteristics for 5 marks asked

Markov Service Process: It is a special service process in which entities are processed one at a time in FCFS order and service time are independent and exponential. It is a memoryless service process which means that expected time until an entity is finished remains constant regardless of how long it hass been in service.

3> Queuing Discipline: 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 like web servers, call centers etc. have limits on the number of entities that can be in queue at any given time. Arrivals that come when queue is full are rejected.

The logical ordering of customer in a waiting line is called queing discipline and it determines which customer will be choosen for service.

* Queuing Disciplines:

if characteristics asked for 5 marks escape this in exam. Write only points 1 to 3.

i) First in First out (FIFO): 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.

ii) Last in First out (LIFO): It occurs when service is next offered to the customer that arrived recently or which have least waiting time. In crowded train passanger getting in or out from train is an example of LIFO.

iii) Service in Random order (SIRO): It means that a random choice is made between all waiting customers at the time service is offered.

iv) Shortest processing time first (SPT): It means that the customer with shortest service time will be choosen first for the service.

v) 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 choosen for service.

⊛. Kendall's notation for queuing system: [Imp]

Different notations are frequently used in queuing system and are called Kendall Notation. Kendal Notation is the standard system used to describe and classify a queuing node. The Kendall Notation can be represented in the form A/B/c/N/K.

where, A indicates arrival pattern, B indicates service pattern, c indicates number of server, D indicates queuing discipline, N indicates system capacity and K indicates calling population.

The symbols used for the probability distribution for inter arrival time, and service time are, D for deterministic, M for exponential and E_k for Erlang distribution.

If parameters D, N & K are not specified then, N = ∞ (infinity), K = ∞ (infinity), D = FIFO.

Example: M/D/2/FIFO/5/∞, queuing system having exponential arrival pattern, deterministic service time, 2 servers, FIFO queuing discipline, capacity of 5 customers and infinite population.

⊛. Single server queuing system:

It is a queuing system with only one server for any number of clients. It is a FIFO queuing system with Kendall Notation, M/M/1 with poisson input, exponential service time and unlimited waiting positions. The model is based on following assumptions:

i) The arrival follow poisson distribution with a mean arrival rate λ.

ii) The service time has exponential distribution, average service rate μ.

iii) Arrivals are infinite population.

iv) Customers are served on First-in First-out (FIFO) basis.

v) There is only a single server.

In a single server queuing system, there is an infinite number of waiting positions in the queue. Hence there can be any number of customers in the queue. It becomes a challenge to maintain the service rate in such a way as to match up with the continuous arrival of customers.

The model can be described as a continuous time Markov chain with transition matrix:

$$Q = \begin{pmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \mu & -(\mu+\lambda) & \lambda & \\ & & \mu & -(\mu+\lambda) & \lambda \\ & & & & \vdots \end{pmatrix}$$

All formulas below like utilization, no. of customers in system, others all are performance measures of (M/M/1) queuing system. [Imp]

The model is considered stable only if $\lambda < \mu$. We write $\rho = \lambda/\mu$ for the utilization of buffer and require $\rho < 1$ for the queue to be stable. It represents the average portion of time which the server occupied.

Number of customers in the system:

The probability that the stationary process is in state i is $\pi_i = (1-\rho)\rho^i$. The number of customers in the system is geometrically distributed with parameter $1-\rho$. Thus the average number of customers in the system is $\rho(1-\rho)^2$.

Response Time:

Scheduling discipline can be computed as $1/(\mu-\lambda)$.
The average time spent waiting is $\frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\rho}{(\mu-\lambda)}$.

Expected number of customers in system $L_s = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$

" " in queue, $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{(1-\rho)}$

Average waiting time in system, $W_s = \frac{1}{\mu-\lambda}$

" " in queue, $W_q = \frac{\lambda}{\mu(\mu-\lambda)}$

Average waiting time for customer $= \frac{1}{\mu-\lambda}$.

Probability that there are n customers in system

$$P_n = \left[\frac{\lambda}{\mu}\right]^n \qquad P_0 = \left[\frac{\lambda}{\mu}\right]^n \left[1 - \frac{\lambda}{\mu}\right].$$

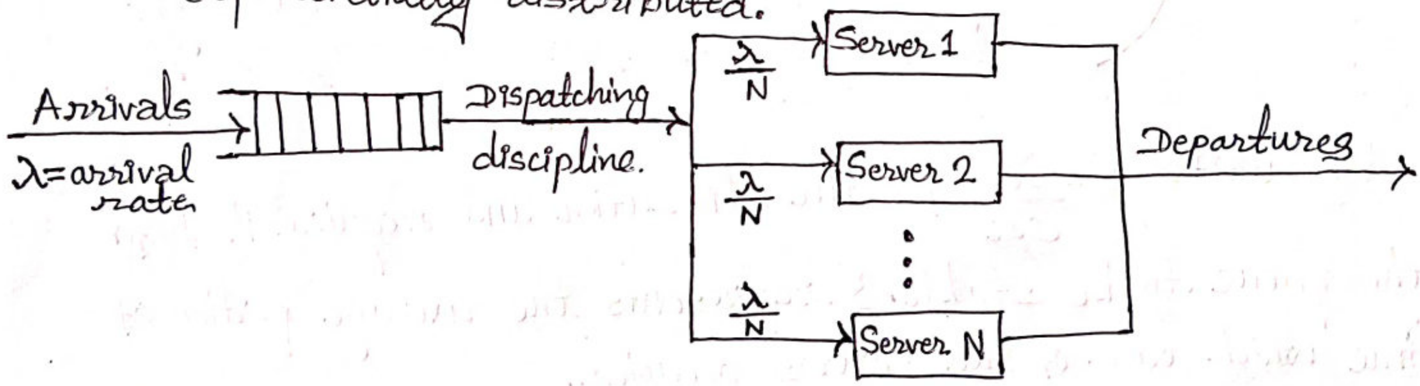
Probability that there is nobody in system.

$$P_0 = \frac{1-\lambda}{\mu}.$$

⊛. Multi-Server queuing system:

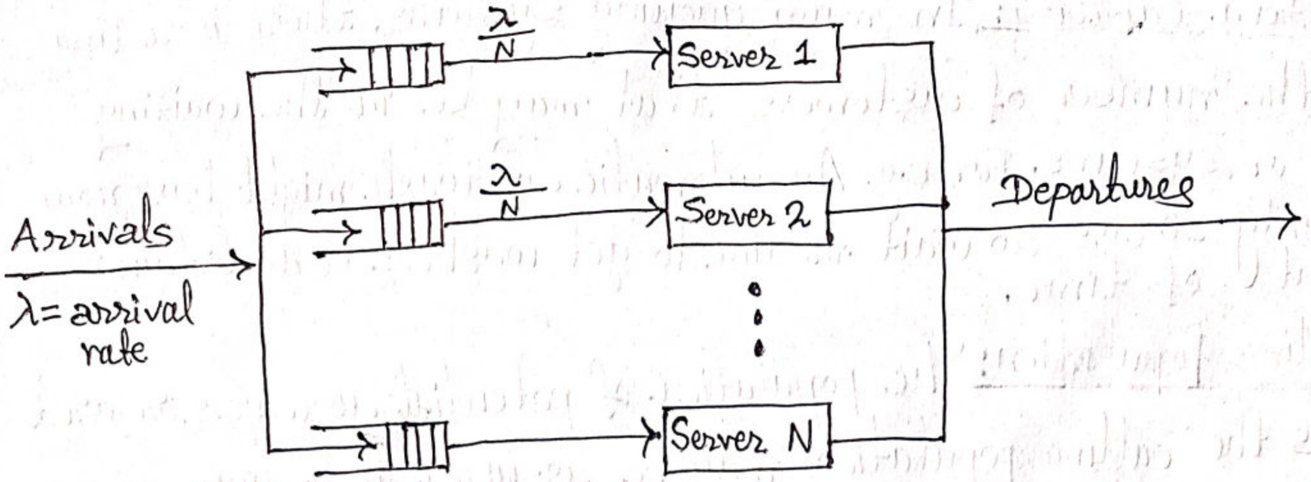
It is queuing system with more than one server. In this system all share a common queue. If an item arrives and at least one server is available, then the item is immediately dispatched to the server. It is assumed that all servers are identical, it makes no difference which server is chosen for the item. If all servers are busy, a queue begins to form.

Multi-server queuing system is represented by M/M/c where arrivals form single queue and are governed by a poisson process, there are c-servers and job service times are exponentially distributed.



The total server utilization in case of Multi-server queue for N server queue 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:



Numerical Examples:

Example1: 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. 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 minute is used as unit of time. Since ticket is purchased 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

Waiting time in the system $(W_s) = \frac{1}{(\mu-\lambda)} = \frac{1}{(3-1)} = 0.5$

The average time to get the ticket and the time to reach the 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:

$\lambda = 6$ customers/hour

$\mu = 12$ customers/hour.

(a) The customer will not have to wait at the counter. Thus,

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{6}{12} = 0.5$$

(b) Expected numbers of customers in the bank are given by,

$$L_S = \frac{\lambda}{(\mu - \lambda)} = \frac{6}{(12-6)} = 1$$

(c) Expected time to be spent in the bank is given by

$$W_S = \frac{1}{\mu - \lambda} = \frac{1}{12-6} = \frac{1}{6} \text{ hour} = 10 \text{ minutes.}$$

⊛. Network of Queues:

Many systems are naturally modeled as networks of single queues in which customers departing from one queue may be routed to another. The following results assume a stable system with infinite calling population and no limit on system capacity:

i) Provided that no customers are created or destroyed in the queue, then the departure rate out of the queue is the same as the arrival rate into the queue over the long run.

ii) If customers arrive to queue i at rate λ_i and a fraction $0 \le P_{ij} \le 1$ of them are routed to queue j upon departure, then the arrival rate from queue i to queue j is $\lambda_i P_{ij}$ over the long run.

iii) The overall arrival rate into queue j, λ_j is the sum of the arrival rate from all sources. If customers arrive from outside the network at rate a_j, then $\lambda_j = a_j + \sum_{\text{all } i} \lambda_i P_{ij}$.

iv) If queue j has $c_j < \infty$ parallel servers, each working at rate μ_j, then the long-run utilization of each server is

$$\rho_j = \frac{\lambda_j}{c_j \mu_j}$$

and $\rho_j < 1$ is required for the queue to be stable.

Applications of queuing system:

Queuing systems are used in our daily life in every aspect. Some of the common applications are:

1) Commercial Queuing Systems:
- → Commercial organizations serving external customers.
- → Eg: Dentist, Bank, ATM, Gas Station, Plumber.

2) Transportation Service System:
- → Vechiles are customers or servers
- → Eg: Vechiles waiting at traffic lights, buses, taxi cabs etc.

3) Business internal service systems:
- → Customers receiving service are internal to the organization providing the service.
- → Eg: Inspection stations, conveyor belts, customer support etc.

4) Social service systems:

Eg: Judicial process, hospital, waiting list for organ transplants or students dorm rooms etc.