4. Waiting Line Theory or Queueing Model

Waiting Line Theory or Queueing Model

Definitions :

$(M/N/S) : (C/D)$

M - Arrival Pattern
N - Service Pattern
S - Service Channels
C - Service Capacity
D - Service Decipline

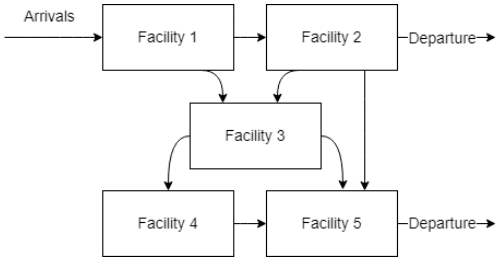
e.g., $(M/N/1) : (\infty/FIFO)$

A queue is formed when there is disbalance between number of servers and number of customers.

A flow of customers from finite or infinite population towards the service facility forms a queue on account of lack of capability to serve them all at a time.

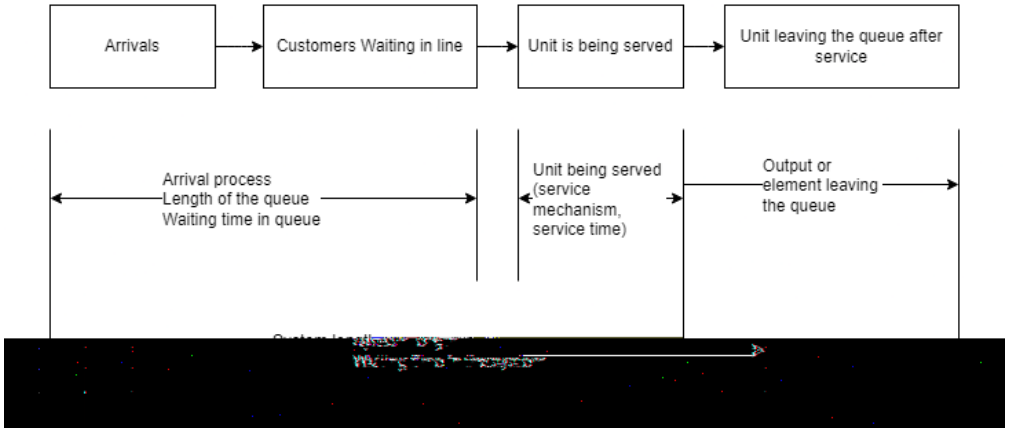
In absence of perfect balence between the service facilities and the customers, waiting time is required either for the service facilities or the customer's arrival.

Customer - An arriving unit waiting to be serviced.
Queue - No of customers waiting to be serviced.
Service Facility - The body providing the service.

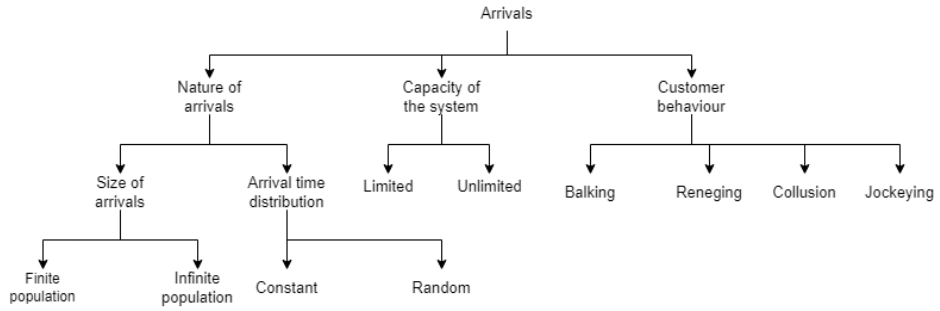


Queueing system / process

a> Input (Arrival pattern)
b> Service mechanism / service pattern
c> Queue decipline
d> Customer behaviour



Input process



characteristics of arrivals

a) Size of arrivals - Depends on the nature of size of the population (i.e. finite / infinite). More specifically described in terms of probabilities and probability distributions for inter-arrival time (time between two successive arrivals or the distribution of customers arriving in a unit time must be defined).

> **Note:**
> For simplicity only Poisson arrivals are considered at the moment.

b) Inter-arrival time - the period between two successive arrivals. Most

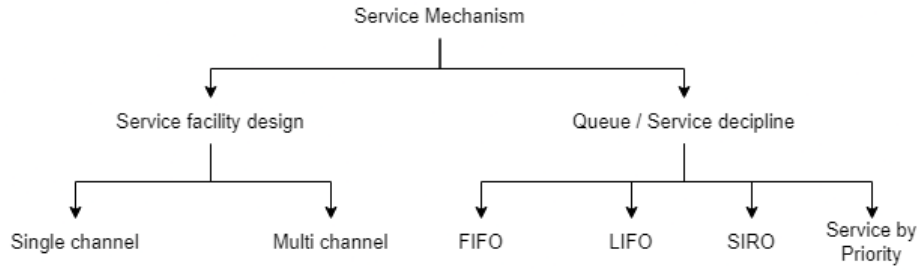
queueing models assume that some inter-arrival time distraction exists for all customers exists throught the period of study. In most situations, the service time is a random variable with the same distribution for all arrivals but there might be cases where there are clearly two or more classes of customers such as machines waiting for repair with different service time distributions. Here mostly the important service time distributions are considered (*Negative exponential* and *Erlang or Gamma Distribution*)

c) capacity of the system - Space available for arrivals to wait before being taken to service. The space available may be limited or unlimited. When the space is limited, if the length of the queue crosses a certain limit : no more arrivals are permitted to enter the system till the waiting space becomes vacant. This type of system is called *system with finite capacity* and has effects on the arrival pattern of the system.

d) customer behaviour - The length of the queue depends on the behaviour of the customer, i.e. the impatiance of the customer during their stay in the queue. Customer behaviour can be classified as

- **Balking** - When the customer does not like to join the queue seeing the length of it. This behaviour might result in losing a customer. A lengthy queue indicates insufficient service facility and the customer(s) may not turn out next time.
- **Reneging** - In this case, a customer joins the queue and after waiting for a certain time loses his patience and leaves the queue. This behaviour may also result in loss of a customer.
- **Collusion** - Several customers may collaborate and only one of them may stand in the queue. One customers represents a group of customers. In this case, the queue itself might be small but service time for an individual will be more. This may lead to other customers becoming impatient.
- **Jockeying** - If there are multiple queues depending on the number of service stations, a customer in one of the queues after seeing the other queue length which is shorter, might leave the present queue and join the other queue with the hopes of getting the service faster. Perhaps the shorter queue might have more collaborated groups which might leade to longer service time. In such case the probability of the customer who has switched queues getting the service may be very less. Because of this customer behaviour, the queue lengths may keep changing from time to time.

Service Mechanism / Service Facility



Notations

X - Inter-arrival time between two successive customers / arrivals.
Y - Service time required by a customer.
W - Waiting time for any customer before it is taken into service.
V - Time spent by the customer in the system.
n - Number of customers in the system. i.e. customers in the waiting line at any time and number of customers being served included.

$P_n(t)$ - Probability that n customers arrive in the system at time t.
$\Phi_n(t)$ - Probability that n customers are served at time t.
$U(T)$ - Probability distribution of inter-arrival time ($P[t < T]$).
$V(T)$ - Probability distribution of service time ($P[t < T]$).
$F(N)$ - Probability distribution of queue length at any time t.

λ_n - Average no. of customers arriving per unit of time when there are already n customers in the system.
λ - Average no. of customers arriving per unit of time.
μ_n - Average no. of customers being served (completion) per unit of time when there are already n customers in the system.
μ - Average no. of customers being served (completion) per unit of time.
$1/\lambda$ - average inter-arrival between two arrivals.
$1/\mu$ - average service (completion) time between two customers.
$\rho = \left(\frac{\lambda}{\mu}\right)$ - System utility or traffic intensity (the amount of time for which the system was utilized). e.g., given time is 8 hours and if $\rho = 3/8$, that implies that the system was used for 3 hours and was idle for 5 hours.

Case 1

$(M/M/1) : (\infty/FIFO)$

= (Poisson arrival / Poisson service / No. of service channels) : (Infinite capacity / FIFO model)

Given $\lambda, \mu, \rho = \left(\frac{\lambda}{\mu}\right)$,
$(\lambda < \mu)$

[the system would be idle for some time.]

i) Probability that the system is empty

$$= P_0 = (1 - \rho)$$

ii) Probability that there are n customers in the system

$$= P_n = \rho^n \cdot P_0$$

iii) Average number of customers in the system

$$= E[n] = \sum_{n=0}^{\infty} n \cdot \rho^n (1 - \rho) = (1 - \rho) \cdot \frac{\rho}{(1 - \rho)^2}$$

$$= \frac{\rho}{1 - \rho}$$

$$= \frac{\lambda/\mu}{(1 - \lambda/\mu)} = \frac{\lambda}{\mu - \lambda}$$

iv) Average number of customers in the queue

$$= \frac{\rho^2}{1 - \rho} = \frac{(\lambda/\mu)^2}{1 - \lambda/\mu} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

v) Average waiting length (mean time in the system)

$$= \frac{1}{\mu - \lambda} = \frac{1/\mu}{1 - \lambda/\mu} = \frac{1}{\mu (1 - \rho)}$$

vi) Average length of waiting line with the condition that it is always > 0

$$V(n) = \frac{\rho}{(1 - \rho)^2} = \frac{\lambda/\mu}{(1 - \lambda/\mu)^2} = \frac{\lambda/\mu}{\frac{(\mu - \lambda)^2}{\mu^2}} = \frac{\lambda \cdot \mu}{(\mu - \lambda)^2}$$