QUEUING THEORY

A queue, in general, is formed at any place when a customer (human beings or physical entities) that requires services is made to wait due to the fact that the number of customers exceeds the number of service facilities or when service facilities do not work efficiently and take more time than prescribed to serve a customer

INTRODUCTION:

A common situation that occurs in everyday life is that waiting in a line either at bus stop, petrol pump, restaurants, ticket booths, doctor's clinics, bank counters' traffic lights and so on. Queues (waiting line are also found in workshops where the machines wait to be repaired; at a tool crib where the machine wait to receive tools; in a warehouse where items wait to be used; incoming calls wait to mature in telephone exchange, trucks wait to be unloaded, airplanes wait to take off or land and so on

The study of **queuing theory** helps to determine the balance between

(a) cost of offering the service, and

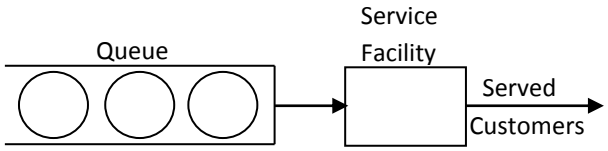
(b) cost incurred due to delay in offering service

The first cost is associated with the service facilities and their operation, and the second represent the cost of customers waiting for services

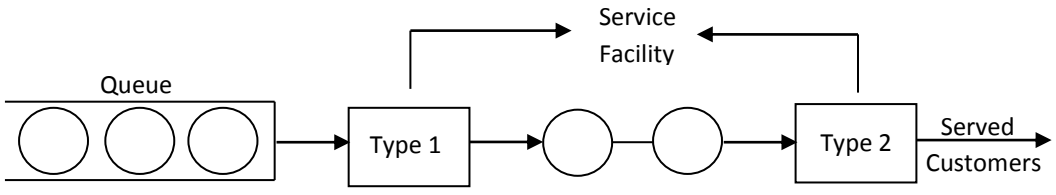
Many real life situations in which study of queuing theory can provide solution to waiting line problems in Table

SITUATION	CUSTOMERS	SERVICE FACILITIES
Petrol pump (stations)	Automobiles	Pumps/ Passionel
Hospital	Patients	Doctors/ Nurses/ Rooms
Airport	Aircraft	Runways
Post office	letters	Sorting system
Job interviews	Applicants	Interviewers
Cargo	Trucks	loader/unloader
Workshop	Machines/Cars	Machines/ Floor space
Factory	Employees	Cafeteria/ Punching Machine

ARRANGEMENT OF SERVICE FACILITIES:

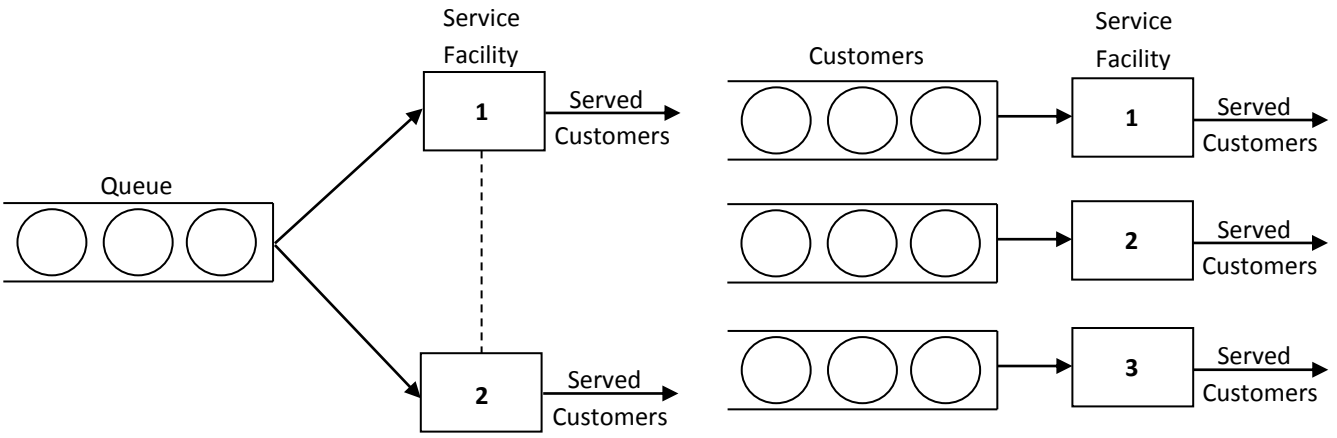


(a) Single Queue, Single Service Facility



(b) Single Queue, Multiple Service Facility

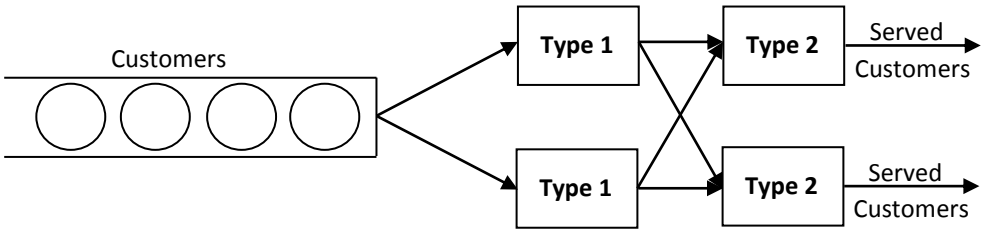
Arrangements of service facilities in series



(a) Single Queue, Multiple Service Facility in Parallel

(b) Multiple Queue, Multiple Service Facility in Parallel

Arrangements of service facilities in parallel



Single queue, Multiple Service Facilities in Parallel and in series

PERFORMANCE MEASURE OF A QUEUING SYSTEM

The performance measure (operating characteristics) for the evaluation of the performance of an existing queuing system, and for designing a new system in terms of the level of service a customer receives as well as the proper utilization of the service facilities are listed as follows:

1. **Average (or expected) time spent by a customer in the queue and system**

   W_q: Average time an arriving customer has to wait in queue before being served,

   W_s: Average time an arriving customer spends in the system, including waiting and service

2. **Average (expected) number of customers in the queue and system**

   L_q: Average number of customers waiting for service in the queue (queue length)

   L_s: Average number of customers in the system (either waiting for services in the queue or being served)

3. **Value of time both for customers and servers**

P_w: Probability that an arriving customer has to wait before being served (also called locking probability)

$\rho = \frac{\lambda}{\mu}$: Percentage of time a server is busy serving customers, i.e., the system utilization

P_n: Probability of n customers waiting for service in the queuing system

P_d: Probability that an arriving customer is not allowed to enter in the queuing i.e., system capacity is full

4. **Average cost required to operate the queuing system**

(i) Average cost required to operate the system per unit of time?

(ii) Number of servers (service centres) required to achieve cost effectiveness?

Service time is the elapsed time form the beginning to the end of a customer's service

Steady state condition is the normal condition that a queuing system is in after operating for some time with a fixed utilization factor less than one

NOTATIONS:

The notation used for analyzing of a queuing system are as follows:

$n =$ number of customers in the system

$P_n =$ probability of n customers in the system

$\lambda =$ average customer arrival rate or average number of arrivals per unit of time in the queuing system

$\mu =$ average service rate or average number of customers served per unit time at the place of service

$\frac{\lambda}{\mu} = \rho = \frac{\text{Average service completion time } (1/\mu)}{\text{Average interarrival time } (1/\lambda)}$

$=$ traffic intensity or server utilization factor

$P_0 =$ probability of no customer in the system

$s =$ number of service channels (service facilities or servers)

$N =$ maximum number of customers allowed in the system

$L_s =$ average number of customers in the system (waiting and in service)

$L_q =$ average number of customers in the queue (queue length)

$W_s =$ average waiting time in the system (waiting and in service)

$W_q =$ average waiting time in the queue

$P_w =$ probability that an arriving customer has to wait (system being busy), $1 - P_0 = (\lambda/\mu)$

For achieving a steady state condition, it is necessary that, $\lambda/\mu < 1$ (i.e., the arrival rate must be less than the service rate) .

utilization factor is the average fraction of time that the serves are being utilized while serving customers

RELATION AMONG PERFORMANCE MEASURES:

The following basic relationship holds for all infinite source queuing models

$L_s = \sum_{n=0}^{\infty} nP_n$ and $L_q = \sum_{n=s}^{\infty}(n-s)P_n$

The general relationship among various performance measure is as follows:

(i) Average number of customers in the system is equal to the average number of customers in queue (line) plus average number of customers being served per unit of time (system utilization)

$$L_s = L_q + \text{Customer being served}$$
$$= L_q + \frac{\lambda}{\mu}$$

The value of $\rho = \frac{\lambda}{\mu}$ is true for a single server finite source queuing model

(ii) Average waiting time for a customer in the queue (line)

$$W_q = \frac{L_q}{\lambda}$$

(iii) Average waiting time for a customer in the system including average service time

$$W_s = W_q + \frac{1}{\mu}$$

(iv) Probability of being in the system (waiting and being served) longer than time t is given by:

$P(W_s > t) = e^{-(\mu-\lambda)t}$ and $P(W_s \leq t) = 1 - P(W_s > t)$

where $W_s =$ time spent in the system

$t =$ specified time period

(v) Probability of only waiting for service longer than time t is given by:

$$P(W_q > t) = \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}$$

(vi) Probability of exactly n customers in the system is given by

$$P_n = P_0 \left(\frac{\lambda}{\mu}\right)^n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$$

(vii) Probability that the number of customers in the system, n exceeds a given number, r is given by

$$P(n > r) = \left(\frac{\lambda}{\mu}\right)^{r+1}$$

The general relationship among various performance measures are:

(a) $L_s = \lambda W_s$

(b) $W_s = W_q + \frac{1}{\mu} = \frac{1}{\lambda} L_s$

(c) $L_q = L_s - \frac{\lambda}{\mu} = \lambda W_q$

(d) $W_q = W_q - \frac{1}{\mu} = \frac{1}{\lambda} L_q$

CLASSIFICATION OF QUEUING MODELS:

Queuing theory models are classified by using special (or standard) notations described initially by D.G. Kendall in the form $(a/b/c)$ Later A.M. Lee added the symbol d and c to the Kendall's notation In the literature of queuing theory, the standard format used to describe queuing models is as follows:

$$\{(a/b/c):(d/c)\}$$

where $a =$ arrivals distribution

$b =$ service time distribution

$c =$ number of servers (service channels)

$d =$ capacity of the system (queue plus service)

$e =$ queue (or service) discipline

In place of notation a and b, the following descriptive notations are used for the arrival and service time distribution:

$M =$ Markovian (or Exponential) interarrival time or service time distribution

$D =$ Deterministic (or constant) interarrival time or service time

$G =$ general distribution of service time, i.e., no assumption is made about the type of distribution with mean and variance

$GI =$ General probability distribution - normal or uniform for inter arrival time

$E_k =$ Erlang - k distribution for interarrival or service time with parameter k (i.e., if $k = 1$, Erlang is equivalent to exponential and if $k = \infty$, Erlang is equivalent to deterministic)

For example, a queuing system in which the number of arrivals is described by a Poisson probability distribution, the service time is described by an exponential distribution, and there is a single server, would be represented by $M/M/1$

SINGLE SERVER QUEUING THEORY:

Model 1: $\{(\boldsymbol{M/M/1}): (\infty/\boldsymbol{FCFS})\}$ **Exponential Service - Unlimited Queue**

This model is based on certain assumption about the queuing system:

(i) Arrivals are described by Poisson probability distribution and come from an infinite calling population

(ii) Single waiting line and each arrival waits to be served regardless of the length of the queue (i.e., no limit on queue length - infinite capacity) and that there is no balking or reneging

(iii) Queue discipline is 'first come, first served'

(iv) Single server or channel and service times follows exponential distribution

(v) Customer arrival is independent but arrival rate (average number of arrivals) does not change over time

(vi) The average service rate is more than the average arrival rate

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

$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) = \rho^n(1-\rho);\ \rho < 1, n = 0, 1, 2, \ldots \ldots$

2. **(a)** Expected number of customers in the system (customer in the line plus the customer being served)

$L_s = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda};\ \rho = \frac{\lambda}{\mu}$

(b) Expected number of customers waiting in the queue (i.e., queue length)

$$L_q = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)};$$

3. **(a)** Expected waiting time for a customer in the queue

$$W_q = \lambda\left(1-\frac{\lambda}{\mu}\right)\frac{1}{(\mu-\lambda)^2} = \frac{\lambda}{\mu(\mu-\lambda)} \text{ or } \frac{L_q}{\lambda}$$

(b) Expected waiting time for a customer in the system (waiting and service):

$$W_s = W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} = \frac{1}{\mu-\lambda} \text{ or } \frac{L_s}{\lambda}$$

4. The variance (fluctuation) of queue length

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

5. Probability that the queue is non empty

$$P(n>1) = 1 - P_0 - P_1$$
$$= 1 - \left(1-\frac{\lambda}{\mu}\right) - \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right) = \left(\frac{\lambda}{\mu}\right)^2$$

6. Probability that the number of customers, n is the system exceeds a given number k

$$P(n \geq k) = \left(\frac{\lambda}{\mu}\right)^k \text{ and } P(n>k) = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

Model 2: $\{(\boldsymbol{M/M/1}):(\boldsymbol{N/FCFS})\}$ **Exponential Service - Finite (or Limited Queue)**

This model is based on all assumption of Model 1, expect a limit on the capacity of the system to accommodate only N customers This implies that once the line reaches its maximum length of N customers, no additional customer will be allowed to enter into the system

A finite queue may arise due to physical constraint such as emergency room in hospital; one man barber shop with certain number of chairs for waiting customers, etc.

PERFORMANCE MEASURES FOR MODEL 2:

$$\boldsymbol{P_0 = \frac{1-\rho}{1-\rho^{N+1}}} \quad \boldsymbol{,\rho \neq 1,} \quad \boldsymbol{\rho = \frac{\lambda}{\mu} < 1}$$

$$\boldsymbol{P_n} = \begin{cases} \left(\dfrac{\boldsymbol{1-\rho}}{\boldsymbol{1-\rho^{N+1}}}\right)\boldsymbol{\rho^n} \; ; \; \boldsymbol{n \leq N}; \; \dfrac{\boldsymbol{\lambda}}{\boldsymbol{\mu}} \; \boldsymbol{\neq 1} \\ \dfrac{\boldsymbol{1}}{\boldsymbol{N+1}} \qquad\qquad ; \dfrac{\boldsymbol{\lambda}}{\boldsymbol{\mu}} \boldsymbol{= 1} \end{cases}$$

1. **Expected number of customers in the system:**

$$L_S = \begin{cases} \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}; & \rho \neq 1(\lambda \neq \mu) \\ \frac{N}{2}; & \rho = 1(\lambda = \mu) \end{cases}$$

2. **Expected number of customers waiting in the queue:**

$$L_q = L_S - (1 - P_0)$$

3. **Expected waiting time of customer in the system (waiting + service):**

$W_s = \frac{L_s+1}{\mu}$

4. **Expected waiting time of a customer in the queue:**

$W_q = W_s - \frac{1}{\mu} = \frac{L_s}{\mu}$

5. **Potential customer lost** $(= \textbf{time for which sustem is busy})$

$P_N = P_0\rho^N$

Effective arrival rate, $\lambda_{\text{eff}} = \lambda(1 - P_N)$

Effective traffic intensity, $\rho_{\text{eff}} = \lambda_e/\mu$