Queuing Theory:

This is used in situations where a queue is formed (for example, customers waiting for service, aircraft waiting for landing, jobs waiting for processing in the computer system, etc.). The objective here is to minimize the cost of waiting without increasing the cost of servicing.

Simulation: Simulation is a procedure that studies a problem by creating a model of the process involved in the problem and then attempting to determine the best solution through a series of organized trials and error solutions. Sometimes, this is a difficult/time-consuming procedure. Simulation is used when actual experimentation is not feasible or the solution of the model is not possible.

Queueing Theory Introduction:

A flow of customers from finite or infinite population towards the service facility forms a **queue (waiting line)** an account of lack of capability to serve them all at a time. In the absence of a perfect balance between the service facilities and the customers, **waiting time** is required either for the service facilities or for the customers arrival. In general, the **queueing system** consists of one or more queues and one or more servers and operates under a set of procedures. Depending upon the server status, the incoming customer either waits at the queue or gets the turn to be served. If the server is free at the time of arrival of a customer, the customer can directly enter into the counter for getting service and then leave the system. In this process, over a period of time, the system may experience " Customer waiting" and /or "Server idle time"

5.1.2 Queueing System:

A queueing system can be completely described by

(1) the input (arrival pattern)
(2) the service mechanism (service pattern)
(3) The queue discipline and
(4) Customer's behaviour

5.1.3. The input (arrival pattern)

The input described the way in which the customers arrive and join the system. Generally, customers arrive in a more or less random manner which is not possible for prediction. Thus the arrival pattern can be described in terms of probabilities and consequently the probability distribution for **inter-arrival** times (the time between two successive arrivals) must be defined. We deal with those Queueing system in which the customers arrive in poisson process. The mean arrival rate is denoted by

The Service Mechanism:

This means the arrangement of service facility to serve customers. If there is infinite number of servers, then all the customers are served instantaneously or arrival and there will be no queue. If the number of servers is finite then the customers are served according to a specific order with service time a constant or a random variable. Distribution of service time follows '**Exponential distribution'** defined by

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0$$

The mean Service rate is $E(t) = 1/\lambda$

5.1.5 Queueing Discipline:-

It is a rule according to which the customers are selected for service when a queue has been formed. The most common disciplines are

1. First come first served – (FCFS)
2. First in first out – (FIFO)
3. Last in first out – (LIFO)
4. Selection for service in random order (SIRO)

5.1.6 Customer's behaviour

1. Generally, it is assumed that the customers arrive into the system one by one. But in some cases, customers may arrive in groups. Such arrival is called **Bulk arrival.**

2. If there is more than one queue, the customers from one queue may be tempted to join another queue because of its smaller size. This behaviour of customers is known as **jockeying.**

3. If the queue length appears very large to a customer, he/she may not join the queue. This property is known as **Balking** of customers.

4. Sometimes, a customer who is already in a queue will leave the queue in anticipation of longer waiting line. This kind of departare is known as **reneging.**

5.1.7 List of Variables

The list of variable used in queueing models is give below:

n - No of customers in the system
C - No of servers in the system
P_n (t) – Probability of having n customers in the system at time t.
P_n - Steady state probability of having customers in the system
P_0 - Probability of having zero customer in the system
L_q - Average number of customers waiting in the queue.
L_s - Average number of customers waiting in the system (in the queue and in the service counters)
W_q - Average waiting time of customers in the queue.
W_s - Average waiting time of customers in the system (in the queue and in the service counters)
δ - Arrival rate of customers
μ - Service rate of server
ϕ - Utilization factor of the server
δ eff - Effective rate of arrival of customers
M - Poisson distribution
N - Maximum numbers of customers permitted in the system. Also, it denotes the size of the calling source of the customers.
GD - General discipline for service. This may be first in first – serve (FIFS), last-in-first serve (LIFS) random order (Ro) etc.

5.1.8 Traffic intensity (or utilization factor)

An important measure of a simple queue is its traffic intensity given by

$$\text{Traffic intensity } \phi = \text{Mean } \frac{\text{arrival time}}{\text{Mean service time}} = \frac{\delta}{\mu} \quad (< 1)$$

and the unit of traffic intensity is Erlang

5.1.9 Classification of Queueing models

Generally, queueing models can be classified into six categories using Kendall's notation with six parameters to define a model. The parameters of this notation are

P- Arrival rate distribution ie probability law for the arrival /inter – arrival time.

Q - Service rate distribution, ie probability law according to which the customers are being served.

R - Number of Servers (ie number of service stations)
X - Service discipline
Y - Maximum number of customers permitted in the system.
Z - Size of the calling source of the customers.

A queuing model with the above parameters is written as
(P/Q/R : X/Y/Z)

Single-server queueing problems:

Example 1: Consider the following single-server queue: the inter-arrival time is exponentially distributed with a mean of 10 minutes and the service time is also exponentially distributed with a mean of 8 minutes, find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle.

Solution: We have an M/M/1 system. We also have: $\lambda = 1/10$; $\mu = 1/8$. Hence, $\rho = 8/10$. Then:

$$\text{Number in the Queue } = L_q = \frac{\rho^2}{1-\rho} = \frac{0.8^2}{1-0.8} = 3.2.$$

$$\text{Wait in the Queue } = W_q = L_q/\lambda = 32 \; mins.$$

$$\text{Wait in the System } = W = W_q + 1/\mu = 40 \; mins.$$

$$\text{Number in the System } = L = \lambda W = 4.$$

$$\text{Proportion of time the server is idle } = 1 - \rho = 0.2.$$

Example 2: Consider the following single-server queue: the inter-arrival time is exponentially distributed with a mean of 10 minutes and the service time has the uniform distribution with a maximum of 9 minutes and a minimum of 7 minutes, find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle.

Solution: We have an M/G/1 system. We also have: $\lambda = 1/10$; the mean service time will be $(7+9)/2 = 8$, i.e., $\mu = 1/8$. The variance of the service time, σ_s^2 will equal $(9-7)^2/12 = 1/3$. Also, $\rho = 8/10$. Then:

$$\text{Number in the queue } = L_q = \frac{\lambda^2\sigma_s^2 + \rho^2}{2(1-\rho)} = 1.608.$$

$$\text{Wait in the queue } = W_q = L_q/\lambda = 16.08 \; mins.$$

$$\text{Wait in the system } = W = W_q + 1/\mu = 24.08 \; mins.$$

$$\text{Number in the system } = L = \lambda W = 2.408.$$

$$\text{Proportion of time the server is idle } = 1 - \rho = 0.2.$$

Example 3: Consider the following single-server queue: the inter-arrival time has a gamma distribution with a mean of 10 minutes and a variance of 20 min^2. The service time has the normal distribution with a mean of 8 minutes and a variance of 25 min^2 , find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle. Simulation results indicate W_q to be about 8.1 minutes.

We have a G/G/1 system. We also have: $\lambda = 1/10$; the variance of the inter-arrival time is 20. The mean service time will be 8, i.e., $\mu = 1/8$. The variance of the service time, σ_s^2 is

25. Also, $\rho = 8/10$. Then,

$$C_a^2 = \frac{\sigma_a^2}{(1/\lambda)^2} = 0.2; C_s^2 = \frac{\sigma_s^2}{(1/\mu)^2} = 0.3906.$$

Now using Marchal's approximation:

$$\text{Number in the Queue via Equation (2)} = L_q = \frac{\rho^2(1+C_s^2)(C_a^2+\rho^2 C_s^2)}{2(1-\rho)(1+\rho^2 C_s^2)} = 0.8010.$$

Wait in the queue $= W_q = L_q/\lambda = 8.01\ mins \approx 8.1\ mins$, which is the simulation estimate.

$$\text{Wait in the system} = W = W_q + 1/\mu = 16.01\ mins.$$

$$\text{Number in the system} = L = \lambda W = 1.601.$$

$$\text{Proportion of time the server is idle} = 1 - \rho = 0.2.$$

Multi-server queueing problems:

We will only consider the identical (homogenous) server case in which there are c identical servers in parallel and there is just one waiting line (i.e., the queue is a single-channel queue). Let c denote the number of identical servers. Here

$$\rho = \frac{\lambda}{c\mu}$$

For the M/M/c queue (Ross, 2014),

$$L_q = \frac{P_0\left(\frac{\lambda}{\mu}\right)^c \rho}{c!(1-\rho)^2} \tag{6}$$

where

$$P_0 = 1/\left[\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]. \tag{7}$$

Note that P_0 denotes the probability that there are 0 customers in the system.

Hence, W_q can be obtained as follows:

$$W_q = L_q/\lambda.$$

Then, for the $G/G/c$ queue, we have the following *approximation* (Whitt, 1976; Medhi, 2003):

$$W_q^{G/G/c} \approx W_q^{M/M/c} \frac{C_a^2 + C_s^2}{2}, \tag{8}$$

where $W_q^{A/B/c}$ denotes the waiting time in the queue for the A/B/c queue. The above works well for M/G/c queues, but does not always work well when the inter-arrival time is not exponentially distributed. For multi-server queues, it has been shown that data on two moments is usually not sufficient to generate good approximations for the mean waiting time or queue length (Gupta et al , 2010). When the distributions are known, it is often possible to deduce expressions for these metrics, but they often involve calculus and computational methods (see Kahraman and Gosavi (2011) for one such situation in bulk queues and Medhi (2003) for general discussions, including the Lindley equation).

Example 4: Consider the following scenario: the inter-arrival time has an exponential distribution with a mean of 10 minutes. There are two servers, and the service time of each server has the uniform distribution with a maximum of 20 minutes and a minimum of 10 minutes, find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle. Results from discrete-event simulation, which are known to be very accurate, show that the mean waiting time in the queue is 9.5693 minutes. Compute the error in the G/G/c approximation.

Solution: This is an $M/G/2$ system. We have $\lambda = 1/10$; the $C_a^2 = 1$ as a result. The mean service time will be $(10 + 20)/2 = 15$, i.e., $\mu = 1/15$. The variance of the service time, σ_s^2 will equal $(20 - 10)^2/12 = 8.33$. Also, $\rho = 15/(2 \times 10) = 0.75$. Then:

$$C_s^2 = \frac{\sigma_s^2}{(1/\mu)^2} = 8.33/(15)^2 = 0.03.$$

Using the $G/G/c$ approximation, we first assume the queue to be an $M/M/c$ queue and compute its L_q: Now using the formula above in Eqn. (7): $P_0 = 0.1453$. Then, using Eqn. (6), we have that $L_q = \frac{0.1453(\frac{1/10}{1/15})^2 0.75}{2!(1-0.75)^2} = 1.929$. Then, $W_q = L_q/\lambda = 1.929 \times 10 = 19.29$.

Now, we need to transform this to an $G/G/2$ queue using the approximation in Eqn. (8):

$$W_q^{G/G/c} \approx W_q^{M/M/c} \frac{C_a^2 + C_s^2}{2} = (19.29)(1 + 0.03)/2 = 9.93.$$

Then, $L_q^{G/G/c} = W_q^{G/G/c} \times \lambda = 9.93 \times 1/10 = 0.993$. The error in the approximation is:

$$\frac{|9.9376 - 9.5693|}{9.9376} \times 100\% = 3.07\%.$$

$$\text{Wait in the System} = W = W_q + 1/\mu = 9.93 + 15 = 24.93 \; mins.$$

$$\text{Number in the System} = L = \lambda W = 2.493.$$