Learning objective

- To discuss single server finite queues (M/M/1/N queues)
9.8 Queuing model with a finite queue

- In real cases, queues never become infinite, but are limited due to space, time or service operating policy. Such queuing model falls under the category of finite queues.

- Examples:
  - Parking of vehicles in a supermarket is restricted to the space of the parking area
  - Limited seating arrangement in a restaurant

- Finite queue models restrict the number of customers allowed in service system. That means if \( N \) represents the maximum number of customers allowed in the service system, then the \((N+1)\)th arrival will depart without being part of the service system or seeking service.

9.8.1 Probability of having no customers in M|M|1|N Queuing Model

- The service system can accommodate \( N \) customers only. The \((N+1)\)th customer will not join the queue.

We know that, probability of having one customer in the queuing system is as mentioned below.

\[
P_1 = \left( \frac{\lambda}{\mu} \right) \times P_0
\]

Similarly, probability of having \( N \) customers in the service system will be

\[
P_N = \left( \frac{\lambda}{\mu} \right)^N \times P_0
\]

Since it is finite queuing model, the sum of probabilities till \( N \)th customer will be 1 as written below.
\[ \sum_{n=0}^{N} P_n = 1 \]
\[ \sum_{n=0}^{N} \left( \frac{\lambda}{\mu} \right)^n P_0 = 1 \]
\[ P_0 = \frac{1}{\sum_{n=0}^{N} \left( \frac{\lambda}{\mu} \right)^n} \]
where \( \sum_{n=0}^{N} \left( \frac{\lambda}{\mu} \right)^n \) represents finite Geometric Progression series

Hence, \( P_0 = \frac{1}{1 - \rho^{N+1}} \)

\[ P_0 = \frac{1 - \rho}{1 - \rho^{N+1}} \]

**9.8.2 Probability of having N customers in finite queuing model**

- Due to restriction in the allowable number of customers in the system, the service system will be interested to know the expected number of customers who are lost. This can be found out by determining the probability of customers not joining the system.

- In \( P_N \) represents the probability of not joining the system, then the arrival rate (\( \lambda \)) times \( P_N \) will represent the expected number of customers who are lost after the service system becomes full.

- Finite Queues can be M|M|1 type or M|M|c type.

- In M|M|1|N type queuing model, the utilization factor \( \rho \), can exceed unity. Also, in a single server case, N-1 represents the maximum number of customers in the queue.

- In M|M|c|N type queuing model, N must be equal to or greater than the total number of servers c.
An arriving customer will not be entertained or allowed in the system if the number of customers in the system equals $N$ or the length of the queue is $N-c$.

\[ P_N = \frac{1 - \rho}{1 - \rho^{N+1}} \times \rho^N \]

9.8.3 Performance measures of $M|M|1|N$ queuing system

Average number of customers in the system, $L_s$ can be determined using probability of having finite, $N$, customers in the service system.

\[
L_s = \sum_{n=0}^{N} n \times P_N
= P_0 \times \sum_{n=0}^{N} n \rho^n
= \left( \frac{1 - \rho}{1 - \rho^{N+1}} \right) \times (\rho) \times \left( \sum_{n=0}^{N} \frac{\partial}{\partial \rho} (\rho^n) \right)
= \left( \frac{1 - \rho}{1 - \rho^{N+1}} \right) \times (\rho) \times \left( \frac{\partial}{\partial \rho} \left( \frac{1 - \rho^{N+1}}{1 - \rho} \right) \right)
= (\rho) \times \left( \frac{1 - (N+1) \rho^N + N \rho^{N+1}}{(1 - \rho)(1 - \rho^{N+1})} \right) \quad \rho \neq 1
\]

We know that for a finite queue system, probability of having $n$ customers with $(n>0)$ can be written as

\[ P(n>0) = 1 - P_0 \]

Hence, the number of customers in queue can determined as given below.

\[ L_q = L_s - (1 - P_0) \]
As mentioned in section 9.8.2, $P_N$ represents the probability for the customers of not joining the system. Hence, the $\lambda P_N$ represents the average customers lost due to finite queue. So, the effective arrival rate, $\lambda_e$, in such queues is represented as given below:

$$\lambda_e = \lambda - \lambda P_N$$

or $\lambda_e = \lambda(1-P_N)$

Using effective arrival rate and Little’s law we can determine other performance measures of M|M|1|N queuing system, which are given below.

Average waiting time in the queuing system, $W_s$, comes out to be as mentioned below.

$$W_s = \frac{L_q}{\lambda(1-P_N)} + \frac{1}{\mu}$$

Average waiting time in the queue, $W_q$, can be written as given below.

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

**Example**

The arrival rate of cars at GRAND jewelers is 20 cars per hour. The service rate at jeweler shop is 8 cars per hour. The only parking area at GRAND jewelers is restricted to 6 cars only. The arrival rate and service rate follows Poisson distribution. Identify the type of queue and then determine the performance measures of this queue. Analyze the result if service rate goes to 20 cars per hour.

**Solution**

Since, arrival and service rate follows Poisson distribution and there is one parking area with restricted capacity of cars, this is
**M|M|1|N Queue**

*Arrival rate, \( \lambda = 20 \text{ cars per hour} \)*

*Service rate, \( \mu = 18 \text{ cars per hour} \)*

\[ N = 6 \text{ cars} \]

Utilization ratio \[ \frac{\lambda}{\mu} = \frac{20}{18} = 1.11 \]

\[
P_{N=6} = \left[ \frac{1 - \rho}{1 - \rho^{N+1}} \right] \times \rho^N
\]

\[
= \left[ \frac{1 - 1.11}{1 - (1.11)^7} \right] \times (1.11)^6
\]

\[
= (0.1019) \times 1.88
\]

\[
= 0.1917
\]

\[
L_s = \rho \times \left[ \frac{1 - (N + 1) \rho^N + N \rho^{N+1}}{(1 - \rho)(1 - \rho^{N+1})} \right]
\]

*Given, \( \rho = 1.11 \)*

\[
= 1.11 \times \left[ \frac{1 - (7)(1.11)^6 + 6(1.11)^7}{1 - (1.11)[1 - (1.11)^7]} \right]
\]

\[
= 1.11 \times \frac{0.3728}{0.1212}
\]

\[
= 1.11 \times 3.075
\]

\[
= 3.41 \text{ cars}
\]

\[
\lambda_e = \lambda (1 - P_N)
\]

\[
= 20 (1 - 0.1917)
\]

\[
= 16.166
\]
\[ L_q = L_s - (1 - P_0) \]
\[ P_0 = \left( \frac{1 - \rho}{1 - \rho^{N+1}} \right) \]
\[ P_0 = \left( \frac{1 - 1.11}{1 - (1.11)^7} \right) \]
\[ = 0.1019 \]

\[ L_q = 3.41 - (1 - 0.1019) \]
\[ = 2.512 \text{ cars} \]

\[ W_s = \frac{L_q}{\lambda_c} + \frac{1}{\mu} \]
\[ = \frac{2.512}{16.166} + \frac{1}{18} \]
\[ = 0.2109 \text{ hour or 12.66 minutes} \]

\[ W_q = W_s - \frac{1}{\mu} \]
\[ = 0.2109 - \frac{1}{18} \]
\[ = 0.1553 \text{ hour or 9.32 minutes} \]

*If the service rate becomes 20 cars per hour then;*
\[ \mu = \lambda \]
\[ \text{hence, } \rho = 1 \]

\[ L_s \text{ can be written as given below.} \]
\[ L_s = \frac{N}{2} = \frac{6}{2} \]
\[ = 3 \text{ cars} \]
Hint

We know for finite queue,

$$\sum_{n=0}^{N} P_N = 1$$

which can be written as

$$P_0 + \rho P_0 + \rho^2 P_0 + \ldots + \rho^N P_0 = 1$$

If $\rho = 1$

then $P_0 = \frac{1}{N+1}$

$$L_s = \sum_{n=0}^{N} n \times P_N = P_0 \sum_{n=0}^{N} n \rho^n$$

$\rho = 1$, substitute $P_0 = \frac{1}{N+1}$

$$L_s = \frac{1}{N+1} \left[ 1 + 2 + 3 + \ldots + N \right]$$

$$= \frac{1}{N+1} \frac{N(N+1)}{2}$$

$$L_s = \frac{N}{2}$$

9.9 Utility of Queuing models in service capacity planning

Service organization faces a common trade-off between cost of providing service and cost of customer waiting. The queuing models can help in determining the desired service level which can manage these two types of costs.

9.9.1(a) Determine the service capacity for desired expected customer waiting time.

The service organization can utilize $M|\infty|c$ model; that is multiple servers queuing model to determine the service capacity. If we set a specific time of waiting i.e. customers should wait no more than the specific time. ($W_q$ known), we can determine other performance measures including optimal number of servers.

9.9.2 (b) Determine the number of servers based on the probability that X or more percent of all customers should experience a delay of less that D time units.
We will again take help of $M|M|c$ model. In $M|M|c$ queue model, we know that probability of having $n$ customers is as given below

$$P_n = \begin{cases} \frac{1}{n!} \left[ \frac{\lambda}{\mu} \right]^n P_0 & \text{if } n \leq c \\ \frac{1}{c!} \left[ \frac{\lambda}{\mu} \right]^c \left[ \frac{\lambda}{c\mu} \right]^{n-c} P_0 & \text{if } n > c \end{cases}$$

We can utilize the above relation to determine $P(n \geq c)$

After writing the expression for $P(n \geq c)$, we can keep varying the value of $c$ till we achieve the probability of fewer than $X$ percent of arriving customers have to wait.