LECTURE - 3

M/M/1 Queuing Model

Learning objective

- **To explain the state of the system**
- **To analyze M/M/1 queues**

9.6 M|M|1 queuing model

Features of the M|M|1 queuing system are presented in Table 9.3

Calling Population	An infinite population with independent arrivals and not influenced by the queuing system
Arrival Process	Poisson distribution of arrival rate
Queuing configuration	Single waiting line with unlimited space
Queue discipline	First come, First serve
Service Process	Exponential service time distribution

Table 9.3: Features of M|M|1 queuing system

M|M|1 Queuing model following Poisson distribution of arrival (λ) and service rate (μ) with single server is presented in Figure 9.4. There is no balking and reneging in this model. The numbers of customers in a queue as well as in the system are represented by Lq and Ls respectively.

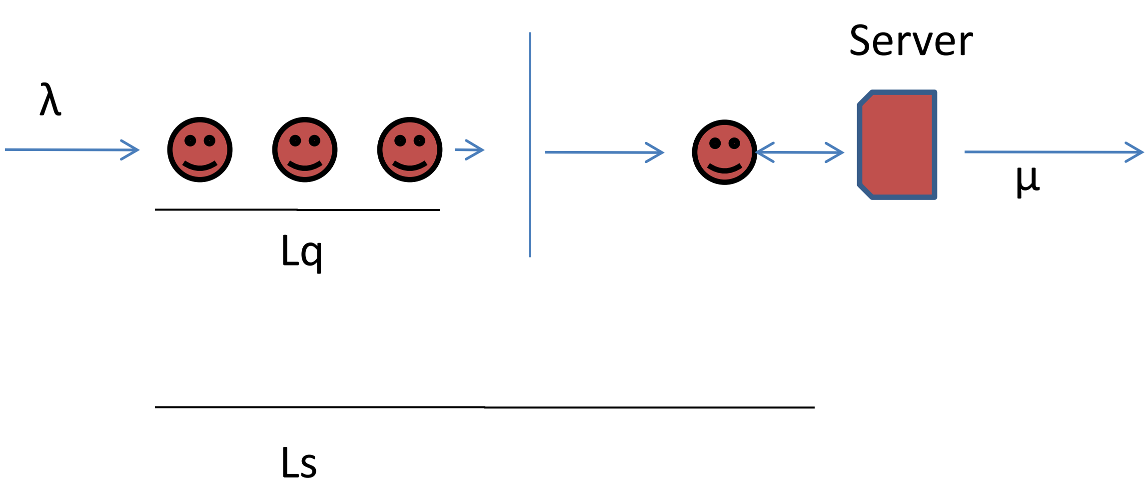

Figure 9.4: M|M|1 queuing system

To analyze M|M|1 Queuing model we require information regarding the number of customers currently present in the system, which is represented by the state of the system.

9.6.1 State of the system

- The state of queuing system is represented by a single number *n*, the number of customers currently in the system.
- It utilizes memory-less property of exponential distribution. As per this property the time since the last arrival and the time the current customer has been in the service process are irrelevant to the future behavior of the system.
- Consider the system to be in steady state, which means that the system has been running for so long that the current state doesn't depend on the starting condition.
- By computing the long run probabilities of being in each state, we will determine the performance measures of queuing models as long term steady state performance measures
- Hence, the customers arrive only one customer at a time. The system state can change only by one unit at a time.

9.6.2 Transition from one state to another state in a queuing system

- If, currently there are *n* customers in the system, then the following changes can happen in the system
  - The state of the system increases from *n* to *n +1*, if arrival occurs in to the service system. The rate of increase is represented by λ, the arrival rate.
  - The state of the system decreases from *n* to *n-1*, if departure to the system occurs. The rate of decrease is represented by μ, the service rate.

The transition from one state to other state can take place as shown in Figure 9.5.

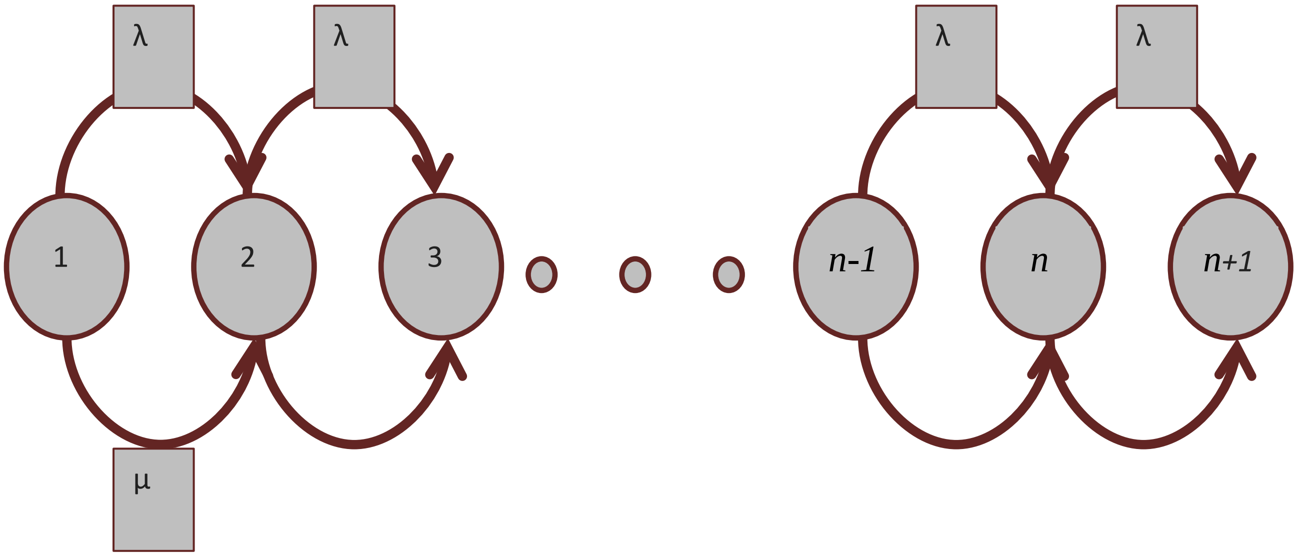

Figure 9.5: Transition of state in a queuing system

- At any given point,
  - If the system is in the state *n*,
    1. The state of the system moves from *n-1* to *n* at the rate of $(P_{n-1})(\lambda)$
    2. The state of the system moves from *n* to *n-1* at the rate of $(P_n)(\mu)$

Where, $(P_n - 1)$ and (P_n) are the probabilities of being in *n-1* and *n* state respectively as shown in Figure 9.6.

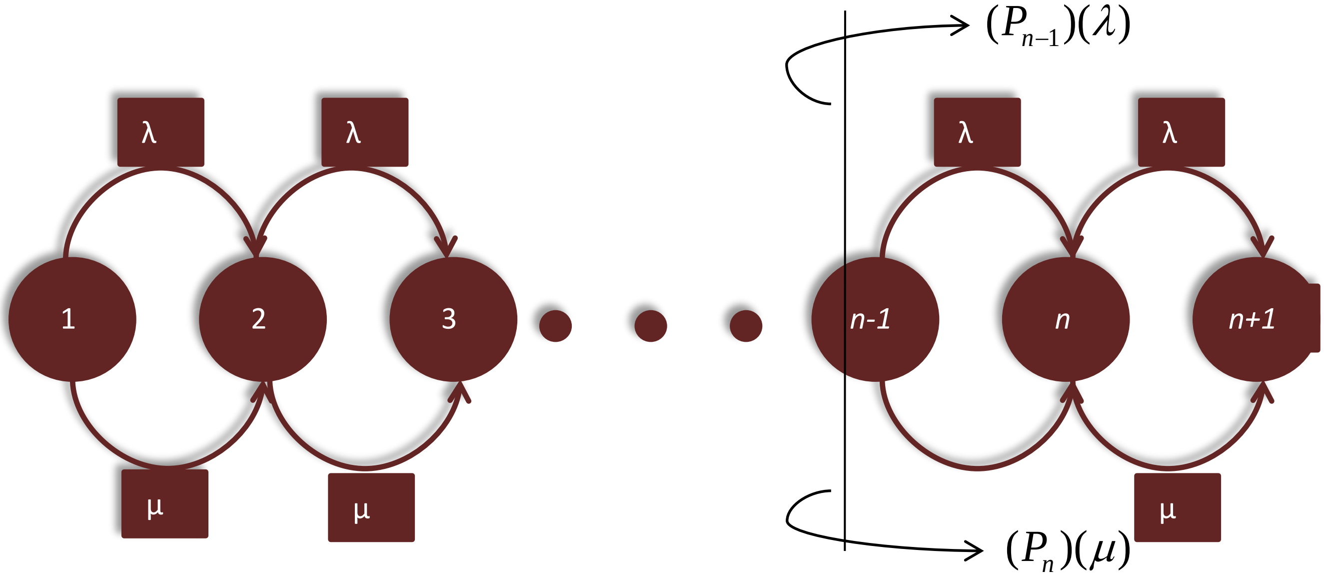

Figure 9.6: Probabilities associated with state of the queuing system

- As per the stability condition

$$(P_{n-1})(\lambda) = (P_n)(\mu)$$

or

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

or

$$(P_n) = (\rho)(P_{n-1})$$

9.6.3 Performance measures of M|M|1 queues

The memory less property is utilized to define the state of the queuing system. To determine the performance measures, first we will find the probability of having *n* number of customers in the queuing system.

Probability of having 1 customer (i.e. n=1) in the service system is:

$$P_1 = \rho P_0$$

Similarly,

$$P_2 = \rho P_1$$

$$P_2 = \rho^2 P_0$$

.

.

$$P_n = \rho^n P_0$$

- We also know that the sum of the probabilities is 1 i.e.,

$$P_0 + P_1 + P_2 + ... = 1$$

or

$$P_0 + \rho P_0 + \rho^2 P_0 + ... = 1$$

or

$$(1 + \rho + \rho^2 + ...)P_0 = 1$$

$$\textit{where } (1 + \rho + \rho^2 + ...) \textit{ is an infinite series,}$$

$$\textit{Sum of infinite series can be written as,} \left(\frac{1}{1 - \rho} \right)$$

$$\textit{Hence, } P_0 = 1 - \rho$$

i.e. Probability of no customer in the system

To determine performance measures L_s, L_q, W_s, W_q in the queuing system we need to determine the average number of customers in the system.

- The average number of customers, L_s in the system can be written as,

$$L_s = \sum_{n=0}^{\infty} n \times P_n$$

$$Where, P_n = \rho^n P_0$$

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

$$L_s = (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$

$$= \rho(1-\rho)\sum_{n=0}^{\infty} n\rho^{n-1}$$

$$We\ know\ that\ \frac{\partial(\rho^n)}{\partial\rho} = n\rho^{n-1},$$

$$L_s = \rho(1-\rho)\frac{\partial}{\partial\rho}\left(\sum_{n=0}^{\infty}\rho^n\right)$$

$$= \rho(1-\rho)\frac{\partial}{\partial\rho}\left(\frac{1}{1-\rho}\right)$$

$$or$$

$$L_s = \frac{\lambda}{\mu - \lambda}$$

- Using Little's law, according to which the average number of customers in the service system is the product of arrival rate and average time a customer spends in the system.
  - Average time a customer spends in the system, W_S can be written using Little's law as given below

$$W_s = \frac{L_s}{\lambda}$$

- We have determined L_s and we know λ, hence

$$W_s = \frac{L_s}{\lambda}$$

$$W_s = \left(\frac{\lambda}{\mu - \lambda}\right) x \left(\frac{1}{\lambda}\right) = \left(\frac{1}{\mu - \lambda}\right)$$

- Average time a customer spends in the queue, W_q, can be determined by subtracting expected service time or average service time from average time a customer spends in the system, W_s

$$W_q = \mathrm{W_s} - \text{Expected service time}$$

$$= \mathrm{W_s} - \mu$$

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

$$= \frac{\mu - (\mu + \lambda)}{\mu \times (\mu - \lambda)}$$

$$= \frac{\lambda}{\mu \times (\mu - \lambda)}$$

- Using Little's law to determine the average number of customers in the queue,

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

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

The Performance measures of a M|M|1 Queue can be written as below.

- Probability of n customers in the service system $(1 - \rho) \times \rho^n$
- Average number of customers in the system $L_s = \dfrac{\lambda}{\mu - \lambda}$
- Average number of customers in the queue $= \dfrac{\lambda^2}{\mu \times (\mu - \lambda)}$

- Average time a customer spends in the system = $\dfrac{1}{\mu-\lambda}$

- Average time a customer spends in the queue = $\dfrac{\lambda}{\mu\times(\mu-\lambda)}$

Example 1

- The mean time between arrivals of customers in a bank is 3 minutes. Write the expression for the exponential distribution for average time between arrivals for any time t (t>=0).

- If a customer has already arrived in the bank, what is the probability that the next customer will come after 10 minutes?

- What is the probability that 5 customers will arrive in the one hour interval?

Solution:

- The mean time between arrivals = 3 mins

- Average arrival rate λ = (1/3)

  = 0.333 arrivals/min

  = 20 arrivals/hour

- The exponential distribution for average time between arrivals for any time t, f(t) is

$$f(t) = 0.333 \times e^{-0.333t}, \ t \geq 0$$

- The cumulative distribution F(t) is

$$F(t) = 1 - e^{-0.333t}, \ t \geq 0$$

Solution:

- The probability that the next customer will come after 10 minutes is

$$F(10) = 1 - e^{-0.333(10)}$$
$$= 0.9643$$

- There is 96.4 percent chance that another customer will arrive in the next 10 minute interval.
- Probability that 5 customers will arrive in one-hour interval will follow Poisson distribution with $\lambda = 20$

S0, $f(5) = \dfrac{(20)^5 \times e^{-20}}{5!}$

$$= 5.49 \times 10^{-5}$$

Example 2:

- In a health clinic, the average rate of arrival of patients is 12 patients per hour. On an average, a doctor can serve patients at the rate of one patient every four minutes.
- Assume, the arrival of patients follows a Poisson distribution and service to patients follows an exponential distribution.
- Find the average number of patients in the waiting line and in the clinic.
- Find the average waiting time in the waiting line or in the queue and also the average waiting time in the clinic.

Solution:

- Poisson arrivals, Exponential service and a single doctor on service, follows a M|M|1 Queuing model.
- The average rate of patient arrival, $\lambda = 12$ patients per hour
- The average rate of serving a patient , $\mu = 1$ in 4 minutes = 15 patients per hour

- The utilization factor ρ = (12/15) = 0.8
- Average number of patients in the clinic

$$L_s = \frac{\lambda}{\mu - \lambda} = 4 \text{ patients}$$

- Average number of patients in the waiting line

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = 3.2 \text{ patients}$$

- Average waiting time in the clinic,

$$W_s = \frac{1}{\mu - \lambda} = 0.267 \text{ hours}$$

- Average waiting time in the queue,

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = 0.333 \text{ hours}$$