Chapter 3

QUEUEING SYSTEM

Introduction:

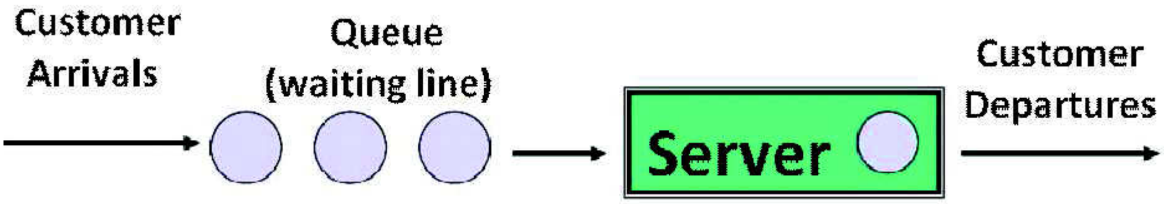
- A ***customer*** is a person or thing receiving or waiting for service.
- A ***server*** is any person or thing that gives service.
- A ***queue*** is formed when customers arrive faster than they can get served or A queue is formed when a customer waits for service.

Customers who arrive to find all servers busy generally join one or more queues (lines) in front of the servers, hence the name queuing systems.

Definition of a Queuing System:

A queuing system consists of a customer, a queue and a service facility (server) with one or more identical parallel servers that provide service of some sort to arriving customers.

In designing queuing systems we need to aim for a balance between service to customers (short queues implying many servers) and economic considerations (not too many servers).



Examples:

✓ Banks / supermarkets - waiting for service

✓ Computers - waiting for a response

✓ Failure situations - waiting for a failure to occur e.g. in a piece of machinery

✓ Service time = 10 minutes, a customer arrives every 15 minutes ---> No queue will ever be formed.

- ✓ Service time = 15 minutes, a customer arrives every 10 minutes ---> Queue will grow for ever (bad for business!)

Type of System based on Queue and Server

a. Single queue and server

   Ex: ATM for ladies.

b. Single queue and multiple servers.

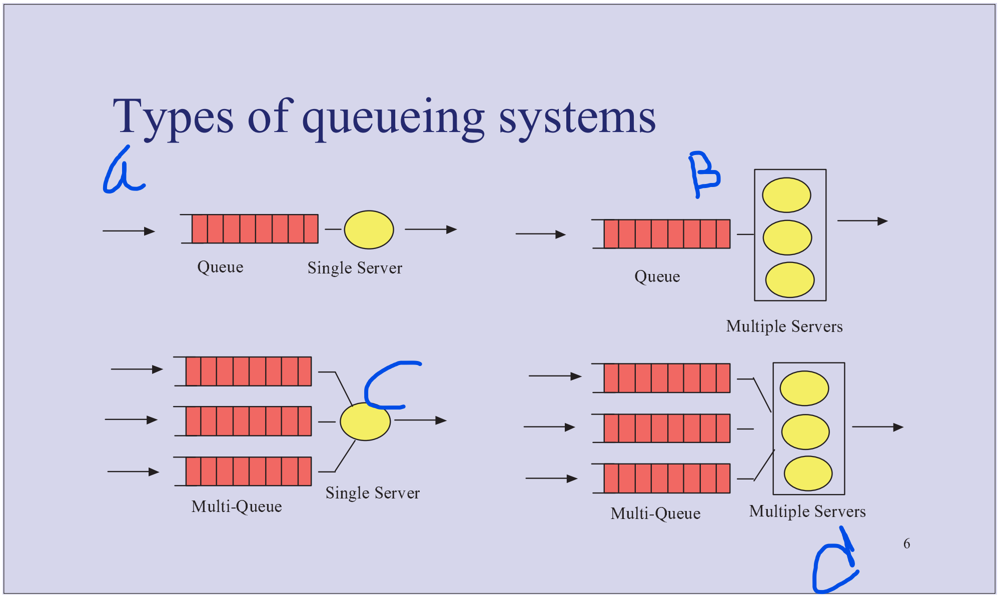
   Ex: Bank with Token Service.

c. Multiple queues and a single server.

   Ex: General ATM (separate queue for ladies and gents)

d. A network of queues and servers.

   Ex: Supermarket Billing Counters.



Types of queueing systems

Components of a Queuing System:

A queuing system is has three components:

1. **Arrival process**

2. **Service mechanism**

3. **Queue discipline.**

1. Arrival Process

Arrivals may originate from one or several sources referred to as the calling population. The calling population can be limited or 'unlimited'. An example of a limited calling population may be that of a fixed number of machines that fail randomly. The arrival process consists of describing how customers arrive to the system. If A_i is the inter-arrival time between the arrivals of the (i-1)th and ith customers, we shall denote the mean (or expected) inter-arrival time by as E(A) and Arrival rate λ.

$$\text{Arrival rate } (\lambda) = 1/\,E(A)$$

2. Service Mechanism

The service mechanism of a queuing system is specified by the number of servers (denoted by s), each server having its own queue or a common queue and the probability distribution of customer's service time. Let S_i be the service time of the ith customer, we shall denote the mean service time of a customer by E(S) and service rate μ.

$$\text{Service rate } (\boldsymbol{\mu}) = 1/\,E(S)$$

3. Queue Discipline

Discipline of a queuing system means the rule that a server uses to choose the next customer from the queue (if any).

Commonly used queue disciplines are:

FIFO - Customers are served on a first-in first-out basis.

LIFO - Customers are served in a last-in first-out manner.

Priority - Customers are served in order of their importance on the basis of their service requirements.

<u>Kendall's Notation for Queues (A/B/C/D/E).</u>

Shorthand notation where A, B, C, D, E describe the queue. It is applicable to a large number of simple queuing scenarios.

A Inter-arrival time } ⇨

M exponential
D deterministic
E_k Erlangian (order k)
G general

distribution

B Service time distribution

C Number of servers

D Maximum number of jobs that can be there in the system (waiting and in service)

E Queuing Discipline (FIFO, LIFO, Priority etc.)

Default is FIFO

M/M/1 or M/M/1/∞ Single server queue with Poisson arrivals, exponentially distributed service times and infinite number of waiting positions

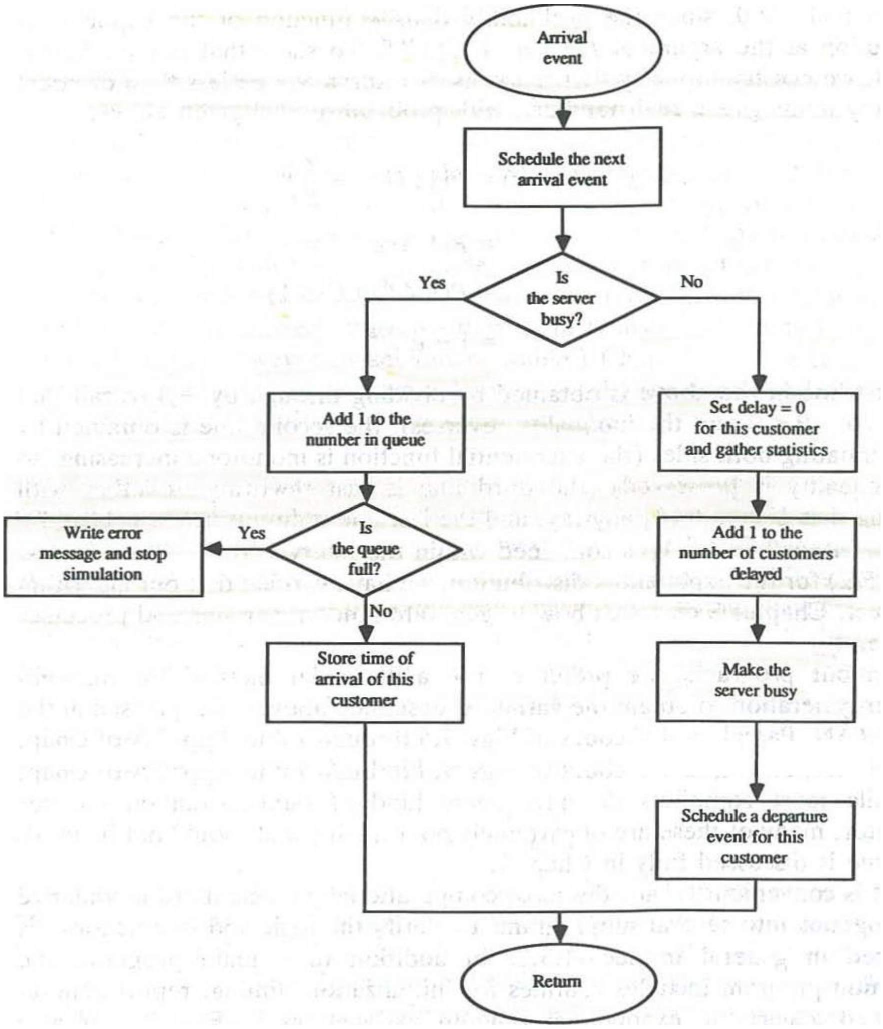
Poisson distribution:
The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).
If we let X = the number of events in a given interval, Then, if the mean number of events per interval is λ
The probability of observing x events in a given interval is given by

$$P(X = x) \quad = \quad e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, 3, 4, \ldots$$

Note: e is a mathematical constant $e \approx 2.718282$

Arrival Routine

First, the time of the next arrival in the future is generated and placed in the event list. Then a check is made to determine whether the server is busy. If so, the number of customers in the queue is incremented by one, and we ask whether the storage space allocated to hold the queue is already full. If the queue is already full, an error message is produced and the simulation is stopped; if there is still room in the queue, the arriving customer's time of arrival is put at the (new) end of the queue. On the other hand, if the arriving customer finds the server idle, then this customer has a delay of zero, which is counted as a delay, and the number of customer delays completed is incremented by one. The server must be made busy, and the time of departure from service of the arriving customer is scheduled into the event list.



Arrival event
Schedule the next arrival event
Is the server busy?
Yes
No
Add 1 to the number in queue
Set delay = 0 for this customer and gather statistics
Write error message and stop simulation
Yes
Is the queue full?
No
Add 1 to the number of customers delayed
Store time of arrival of this customer
Make the server busy
Schedule a departure event for this customer
Return

Figure: Flowchart for Arrival routine, queueing model.

Departure Routine

Recall that this routine is invoked when a service completion (and subsequent departure) occurs. If the departing customer leaves no other customers behind in queue, the server is idled and the departure event is eliminated from consideration, since the next event must be an arrival.

On the other hand, if one or more customers are left behind by the departing customer, the first customer in queue will leave the queue and enter service, so the queue length is reduced by one, and the delay in queue of this customer is computed and registered in the appropriate statistical counter. The number delayed is increased by one, and a departure event for the customer now entering service is scheduled. Finally, the rest of the queue (if any) is advanced one place.

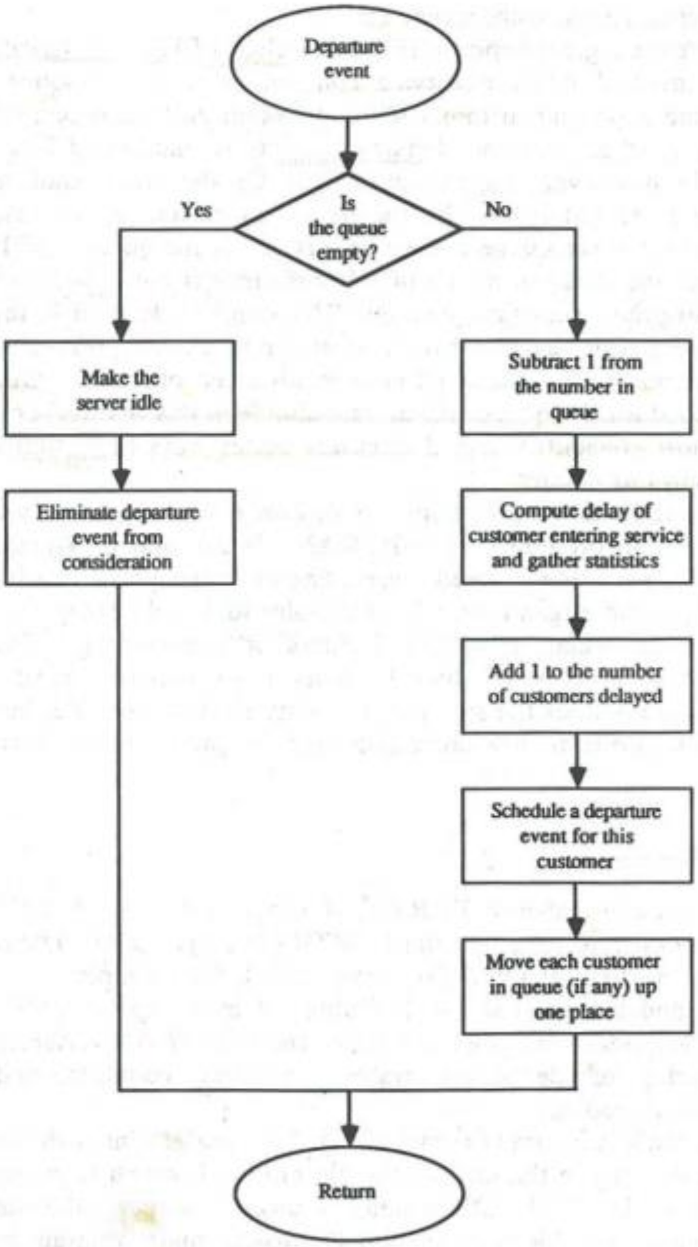
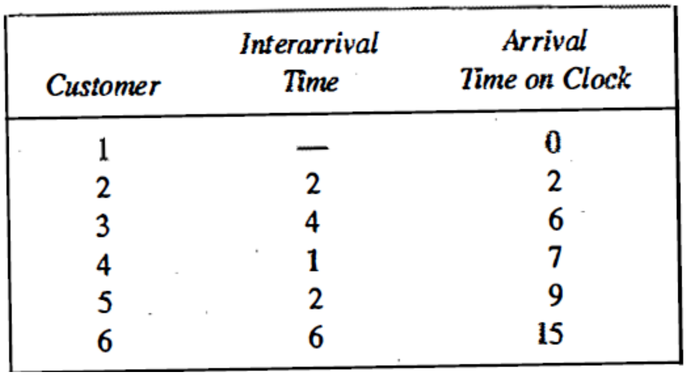

Departure event
Is the queue empty?
Yes
No
Make the server idle
Eliminate departure event from consideration
Subtract 1 from the number in queue
Compute delay of customer entering service and gather statistics
Add 1 to the number of customers delayed
Schedule a departure event for this customer
Move each customer in queue (if any) up one place
Return

Figure: Flowchart for D routine, queueing model.

Inter-arrival times and service times in Queuing System

In a single-channel queueing simulation, interarrival times and service times are generated from the distributions of these random variables. The examples that follow show how such times are generated. For simplicity, assume that the times between arrivals were generated by rolling a die five times and recording the up face. Table 2.2 contains a set of five inter-arrival times generated in this manner. These five inter-arrival times are used to compute the arrival times of six customers at the queueing system.

Table 2.2 Interarrival and Clock Times



Customer	*Interarrival Time*	*Arrival Time on Clock*
1	—	0
2	2	2
3	4	6
4	1	7
5	2	9
6	6	15

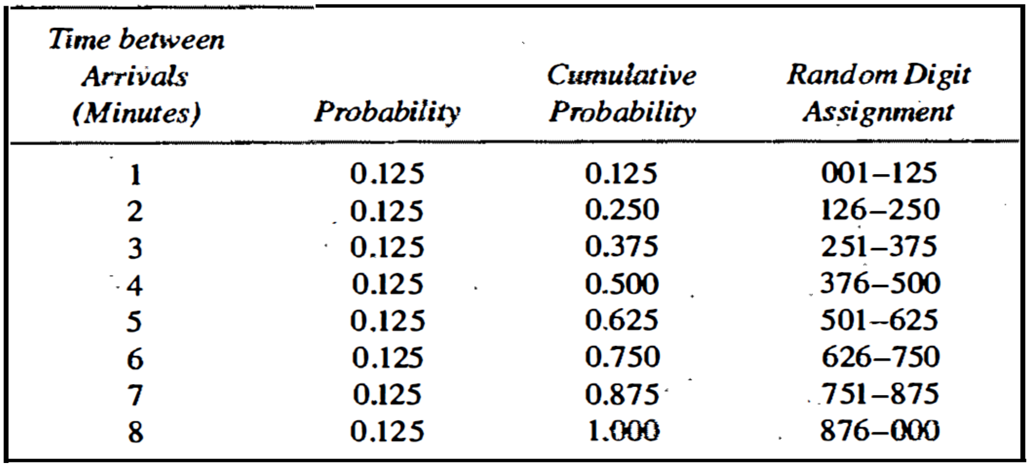
Random Number Generation

The randomness needed to imitate real life is made possible through the use of "random numbers." Random numbers are distributed uniformly and independently on the interval (0, 1). Random digits are uniformly distributed on the set {0, l, 2, ..., 9}. Random digits can be used to form random numbers by selecting the proper number of digits for each random.

Random Digit Assignment

a) **Time Distribution of Time Between Arrivals**
b) **Service Time Distribution**

Table 2.6 Distribution of Time Between Arrivals



Time between Arrivals (Minutes)	*Probability*	*Cumulative Probability*	*Random Digit Assignment*
1	0.125	0.125	001–125
2	0.125	0.250	126–250
3	0.125	0.375	251–375
4	0.125	0.500	376–500
5	0.125	0.625	501–625
6	0.125	0.750	626–750
7	0.125	0.875	751–875
8	0.125	1.000	876–000

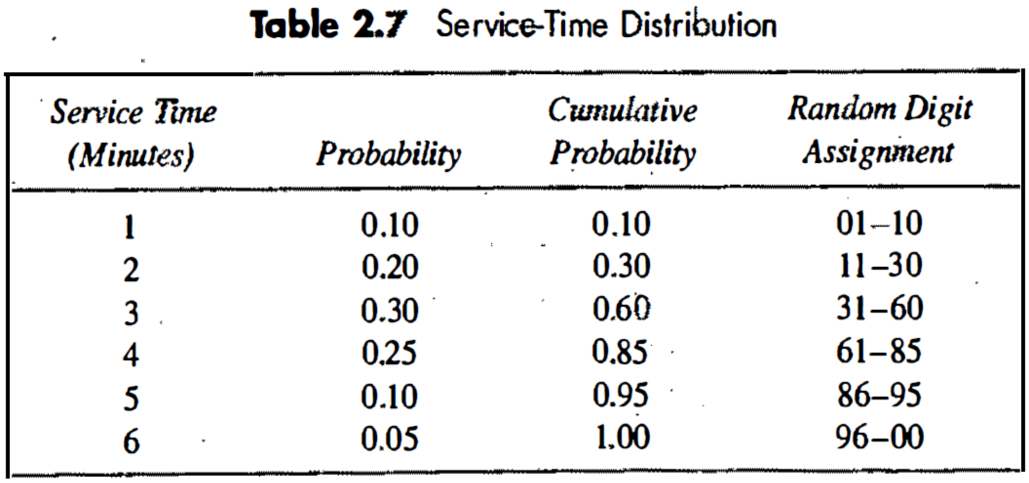


Table 2.7 Service-Time Distribution

Service Time (Minutes)	*Probability*	*Cumulative Probability*	*Random Digit Assignment*
1	0.10	0.10	01–10
2	0.20	0.30	11–30
3	0.30	0.60	31–60
4	0.25	0.85	61–85
5	0.10	0.95	86–95
6	0.05	1.00	96–00

Queuing Famous Formula

$$\begin{array}{c}\text{Average waiting time}\\ \text{(minutes)}\end{array} = \frac{\text{total timecustomers wait in queue (minutes)}}{\text{total numbersof customers}}$$

$$\text{probability(wait)} = \frac{\text{numbers of customers who wait}}{\text{total number of customers}}$$

$$\begin{array}{c}\text{probability of idle}\\ \text{server}\end{array} = \frac{\text{total idle time of server (minutes)}}{\text{total run time of simulation (minutes)}}$$

$$\begin{array}{c}\text{Average service time}\\ \text{(minutes)}\end{array} = \frac{\text{total service time (minutes)}}{\text{total number of customers}}$$

$$\text{expected service time} \qquad E(S) = \sum_{s=0}^{\infty} s p(s)$$

$$\begin{array}{c}\text{Average time between}\\ \text{arrivals (minutes)}\end{array} = \frac{\begin{array}{c}\text{sum of all times}\\ \text{between arrival (minutes)}\end{array}}{\text{number of arrivals} - 1}$$

$$\text{Average waiting time of those who wait (minutes)} = \frac{\text{total time customers wait in queue (minutes)}}{\text{total number of customers that wait}}$$

$$\text{Average time customer spends in the system (minutes)} = \frac{\text{total time customers spend in the system (minutes)}}{\text{total number of customers}}$$

Probability of zero unit in the queue (P_0)=	$1-\frac{\lambda}{\mu}$
Average queue length (L_q)=	$\frac{\lambda^2}{\mu(\mu-\lambda)}$
Average number of units in the system (L_s)=	$\frac{\lambda}{\mu-\lambda}$
Average waiting time of an arrival (W_q)=	$\frac{\lambda}{\mu(\mu-\lambda)}$
Average waiting time of an arrival in the system (W_s)	$\frac{1}{\mu-\lambda}$
Average inter arrival time =	$\frac{1}{\lambda}$
Average service time =	$\frac{1}{\mu}$

1. Complete the following table (for 10 customers) and calculate the average time a customer spends in the system, Arrival rate, service rate and percentage idle time of server.

Customer	Arrival Time	Service start time	Service finish time	Server time	Inter arrival time	Waiting time in Queue	Waiting time in System	Server Idle time
1	7:00	7:00	7:03	3-0 =3	0	0	3	0
2	7:01	7:04	7:06	6-4 = 2	1	3	5	1
3	7:04	7:06	7:09	3	3	2	5	0
4	7:07	7:09	7:12	3	3	2	5	0
5	7:09	7:13	7:16	3	2	4		1
6	7:15	7:16	7:22	6	6	1		0
7	7:17	7:22	7:25	3	2			0
8	7:24	7:25	7:27	2	7			0
9	7:28	7:28	7:33	5	4			1
10	7:30	7:33	7:35	2	2			

Solution:

Server time = Service finish time – service start time

Inter-arrival time of ith customer = Arrival time of ith customer – Arrival time of (i -1)th customer

Waiting time in Queue = Service start time – Arrival time

Waiting time in system = Service finish time – Arrival time

Server Idle time of ith customer = Service start time of ith customer – Service finish time of (i-1)th customer

Customer	Arrival Time	Service start time	Service finish time	Server time	Inter arrival time	Waiting time in Queue	Waiting time in System	Server Idle time
1	**7:00**	**7:00**	**7:03**	**3**	**0**	**0**	**3**	**0**
2	**7:01**	**7:04**	**7:06**	**2**	**1**	**3**	**5**	**1**
3	**7:04**	**7:06**	**7:09**	**3**	**3**	**2**	**5**	**0**
4	**7:07**	**7:09**	**7:12**	**3**	**3**	**2**	**5**	**0**
5	**7:09**	**7:13**	**7:16**	**3**	**2**	**4**	**7**	**1**
6	**7:15**	**7:16**	**7:22**	**6**	**6**	**1**	**7**	**0**
7	**7:17**	**7:22**	**7:25**	**3**	**2**	**5**	**8**	**0**
8	**7:24**	**7:25**	**7:27**	**2**	**7**	**1**	**3**	**0**
9	**7:28**	**7:28**	**7:33**	**5**	**4**	**0**	**5**	**1**
10	**7:30**	**7:33**	**7:35**	**2**	**2**	**3**	**5**	**0**

Average time a customer spend in the system = $\frac{\textbf{Total waiting time in system}}{\boldsymbol{Total\ number\ of\ customers}}$

= 53/10 = 5.3 minutes

Arrival rate = 10 customers per 30 minutes= 10/30 customers/ minute
Service rate = 10 customers per 32 minutes= 10/32 customers/ minute
Percentage Idle time for Server = (3/35)*100 = 8.57%

2. In a bank every 5 minutes a customer arrives. The teller takes 8 minutes to serve one customer. Complete the following table (for 5 customers) and calculate the average time a customer spends in the system, arrival rate, service rate and percentage idle time of server. (Assume the first customer arrives at 9:00 A.M and service begin immediately).

Customer	Arrival Time	Service start time	Service finish time	Server time	Inter arrival time	Waiting time in Queue	Waiting time in System	Server Idle time
1	**9:00**	**9:00**	**7:08**	**8**	**0**	**0**	**8**	**0**
2	**9:05**	**9:08**	**9:16**	**8**	**5**	**3**	**11**	**0**
3	**9:10**	**9:16**	**9:24**	**8**	**5**	**6**	**14**	**0**
4	**9:15**	**9:24**	**9:32**	**8**	**5**	**9**	**17**	**0**
5	**9:20**	**9:32**	**9:40**	**8**	**5**	**12**	**20**	**0**

Average time a customer spend in the system = $\frac{\textbf{Total waiting time in system}}{\boldsymbol{Total\ number\ of\ customers}}$

= 70/5= 14 minute

Arrival rate = 1 customer per 5 minute = 1/5 customer/minute
Service rate = 1 customer per 8 minute = 1/8 customer/minute
Percentage idle time for server = 0%

3. A self-service store employs one cashier at its counter. On an average 9 customers arrive every 5 minutes while cashier can serve 10 customers in 5 minutes. Assuming Poisson distribution for arrival rate and exponential distribution for service rate, find

i. Average number of customers in the system
ii. Average number of customers in the queue
iii. Average time a customer spends in the system
iv. Average time a customer waits before being served

Solution: **<u>If we consider the unit time as 5 minute, then</u>**

λ = Number of customers arrives in 5 minute (unit time) = 9

μ = Number of customers served in 5 minute (unit time) = 10

i) Average number of customers in the system

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{9}{10-9} = 9 \textit{ customers}$$

ii) Average number of customers in the queue

$$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{81}{10} = 8.1 \textit{ customers}$$

iii) *Average time a customer spends in a system*

$$W_s = \frac{1}{\mu-\lambda} = \frac{1}{10-9} = 1 \textit{ unit time} = 5 \textit{ mins}$$

iv) Average time a customer wait before being served

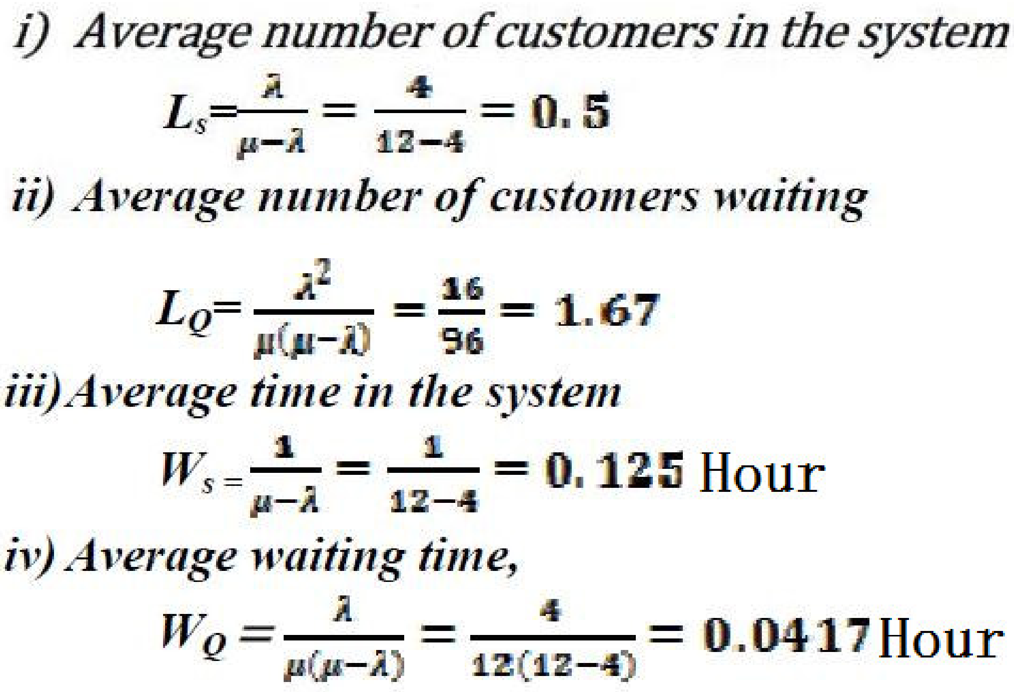
$$W_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{9}{10} = 0.9 \text{ Unit time} = 0.9 \times 5 = 4.5 \text{ minute}$$

4. The arrivals at an ATM booth are assumed to be exponentially distributed. The arrival rate is 4 per hour and the service rate is 12 per hour. Determine the following:

i) Average number of customers in the system
ii) Average number of customer waiting to be served or average queue length
iii) Average time a customer spends system
iv) Average waiting time of a customer before being served.

Solution: If we consider the unit time as 1 hour, then

λ = 4/hour and μ = 12/hour



i) Average number of customers in the system

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{4}{12-4} = 0.5$$

ii) Average number of customers waiting

$$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{16}{96} = 1.67$$

iii)Average time in the system

$$W_s = \frac{1}{\mu-\lambda} = \frac{1}{12-4} = 0.125 \text{ Hour}$$

iv) Average waiting time,

$$W_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{4}{12(12-4)} = 0.0417 \text{ Hour}$$

5. The Taj service station has a central store where service mechanics arrive to take spare parts for the job they work upon. The mechanics wait in queue if necessary and are served on FIFO basis. The store is manned by attendant who can attend 8 mechanics in an hour on an average. The arrival rate of mechanics is 6 per hour and the arrivals are in arrivals in Poisson distribution and service exponentially distributed, determine Ws, Wq, and Lq where these symbol carry their usual meaning.

Solution: If we consider the unit time as 1 hour, then

$$\mu = 8, \lambda = 6$$

i) $W_s = \frac{1}{\mu-\lambda} = \frac{1}{8-6} = 0.5 = 30\, mins$

ii) $W_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{6}{8(8-6)} = 0.375\, hr = 22.5\, mins$

iii) $L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{36}{16} = 2.25\, machines$

μ

6. A person repairing radios find that the time spent on the radio sets has an exponential distribution with mean 20 minutes. If the radios in the order in which they come in and their arrival is approximately Poisson with an average rate of 15 for 8 hour day, a what is the repairman's expected idle time each day? How many jobs are ahead of the average set just brought in (average queue length)?

λ

Solution: If we consider the unit time as 1 day(8 hours), then

λ = Number of radios arrive in a day of 8 hours = 15

μ = Number of radio repaired in a day of 8 hours = 24

Number of radios repaired in 20 minutes = 1
Number of radios repaired in 1 minute = 1/20
Number of radios repaired in 60 minutes(1 hour) = (1/20) * 60 =3
Number of radios repaired in 8 hours = 3 * 8 = 24

d Expected Idel Time

$$P_o = 1 - \frac{\lambda}{\mu} = 1 - \frac{15}{24} = 0.375$$

$$= 0.375 * 8 = 3 \text{ hours}$$

$$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{225}{216} = 1.042 \textit{ radio sets}$$

7. A typist at an office receives, on an average $\overset{\lambda}{\underline{\text{22 letters per day}}}$ of typing. The typist works 8 hours a day and it takes on an average $\underset{u}{\underline{\text{20 minutes}}}$ to type a letter.
i) What is the typist's utilization rate?
ii) What is the average numbers of letters waiting to be typed?
iii) What is the average waiting time needed for a letter before being typed?

Solution: <u>If we consider the unit time as 1 day, then</u>
λ = Number of letters coming for typing in 1 day (8 hours) (unit time) = 22

μ = Number of letters typed in 1 day = 8 hours (unit time)

Number of letters typed in 20 minutes = 1

Number of letters typed in 1 minute = 1/20

Number of letters typed in 60 minute (1 hour) = (1/20) * 60 =3

Number of letters typed in 1 day (8 hours) = 3 * 8 = 24

i) $$\text{Utilization, } \rho = \frac{\lambda}{\mu} = \frac{22}{24} = 0.917$$

ii) $$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{484}{24(24-22)} = 10.08 \textit{ letters}$$

iii) $$W_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{22}{24(24-22)} = 0.458 \textit{ days of 8hrs}$$

$$0.458X8=3.67 \textit{ hrs}$$

8. At a doctor’s clinic, on an average 6 patients arrive per hour. The doctor takes 6 minutes per patient on an average for treatment. It can be assumed that arrivals follow Poisson distribution and the doctor’s inspection time follows an exponential distribution. Determine the following:
(i) percent of times a patient can walk right inside doctor’s cabin without having to wait.
(ii) The average number of patients waiting for their turn
(iii) the average time a patient spends in the clinic.

Solution: If we consider the unit time as 1 hour, then

λ = Number of patient arrive per hour = 6

μ = Number of patient served (treatment) = 10

Number of patient served by doctor in 6 minutes = 1
Number of patient served by doctor in 1 minutes = 1/6
Number of patient served by doctor in 60 minutes (1 hour) = (1/6) * 60 =10

i) Percent of time a patient can walk, right inside doctor cabin without waiting.

Therefore, service is idle.

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{6}{10} = 0.4$$

% of time = 40%

ii) Average number of patients in doctor’s clinic

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{6}{10-6} = 1.5$$

iii) Average number of patients waiting for their turn

$$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{36}{40} = 0.9$$

iv) The average time a patient spends in clinic

$$W_s = \frac{1}{\mu-\lambda} = \frac{1}{10-6} = 0.25hr = 15mins$$