Paper I

Unit IV (Queuing Theory)

We always talk about customers and servers. However, customers can be: persons, orders, packets, ... servers can be: persons, machines, communication channels, ...

We like to answer the following questions:

• How many customers are there on average in the system?

• How long do customers on average spend in the system?

• Which part of the customers will be served and which part will be lost due to the finite capacity of the queue?

• What is the occupation rate of the server (i.e., which part of the server will be busy)?

A queueing situation is basically characterized by a flow of customers arriving at a service facility. On arrival at the facility the customer may be served immediately by a server or, if all the servers are busy, may have to wait in a queue until a server is available. The customer will then leave the system upon completion of service.

The following are some typical examples of such queueing situations:

(i) Shoppers waiting in a supermarket [Customer: shoppers; servers: cashiers].
(ii) Diners waiting for tables in a restaurant [Customers: diners; servers: tables].
(iii) Patients waiting at an outpatient clinic [Customers: patients; servers: doctors].

(iv) Broken machines waiting to be serviced by a repairman [Customers: machines; server: repairman].

(v) People waiting to take lifts. [Customers: people; servers: lifts].

(vi) Parts waiting at a machine for further processing. [Customers: parts; servers: machine].

A mathematical theory has thus evolved that provides means for analysing such situations. This is known as queueing theory (waiting line theory, congestion

theory, the theory of stochastic service system), which analyses the operating characteristics of a queueing situation with the use of probability.

A queueing system is specified by the following elements.

(i) Input Process: How do customers arrive?
Often, the input process is specified in terms of the distribution of the lengths of time between consecutive customer arrival instants (called the interarrival times).
In some models, customers arrive and are served individually (e.g. supermarkets and clinic).
In other models, customers may arrive and/or be served in groups (e.g. lifts) and is referred to as bulk queues.
• Customer arrival pattern also depends on the source from which calls for service (arrivals of customers) are generated. The calling source may be capable of generating a finite number of customers or (theoretically) infinitely many customers.

- In a machine shop with four machines (the machines are the customers and the repairman is the server), the calling source before any machine breaks down consists of four potential customers (i.e. anyone of the four machines may break down and therefore calls for the service of the repairman). Once a machine breaks down, it becomes a customer receiving the service of the repairman (until the time it is repaired), and only three other machines are capable generating new calls for service. This is a typical example of a finite source, where an arrival affects the rate of arrival of new customers.

• For shoppers in a supermarket, the arrival of a customer normally does not affect the source for generating new customer arrival and is therefore referred to as an input process with infinite source.

(ii) Service Process: The time allocated to serve a customer (service time) in a system (e.g. the time that a patient is served by a doctor in an outpatient clinic) varies and is assumed to follow some probability distribution.

(iii) Queue Discipline: The manner that a customer is chosen from the waiting line to start service is called the queue discipline.
• The most common discipline is the first-come-first-served rule (FCFS). Service in random order (SIRO), last-come-first serve (LCFS) and service with priority are also used.

$P_n(t)$

Exponential distribution and Poisson distribution in Queuing Theory

Both the Poisson and Exponential distributions play a prominent role in queuing theory. The Poisson distribution counts the number of discrete events in a fixed time period; it is closely connected to the exponential distribution, which (among other applications) measures the time between arrivals of the events. The Poisson distribution is a discrete distribution; the random variable can only take nonnegative integer values. The exponential distribution can take any (nonnegative) real value.

P (n customers during period t) = the probability that n arrivals will be observed in a time interval of length t

$$P(n, t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \quad (n = 0, 1, 2, \ldots\ldots\ldots\ldots)$$

The time between successive arrivals is called **inter-arrival time.** In the case where the number of arrivals in a given time interval has Poisson distribution, inter-arrival times can be shown to have the exponential distribution.

Arrival Rate is denoted by λ.

If the arrival rate λ = 30/hour, the average time between two successive arrivals are 1/30 hour or 2 minutes.

For example, in the following arrival situations, the average arrival rate per hour, 1 and the average inter arrival time in hour, are determined.

(i) One arrival comes every 15 minutes.

Average arrival rate, $\lambda = \frac{60}{15}$ = 4 arrivals per hour.

Average inter arrival time = 15 minutes = ¼ or 0.25 hour.

(ii) Three arrivals occur every 6 minutes.

Average arrival rate, λ = 30 arrivals per hour.

Average Inter-arrival time, $\bar{I} = \frac{6}{3} = 2$ minutes $= \frac{1}{30}$ or 0.33 hr.

(iii) Average interval between successive intervals is 0.2 hour.

Average arrival rate, $\lambda = \frac{1}{0.2}$ = 5 arrivals per hour.

Average Inter-arrival time, $\bar{I}$ = 0.2 hour.

Poisson and Exponential distribution practice problems

The practice problems of poisson and exponential distributions are given below

Example : In a factory, the machines break down and require service according to a Poisson distribution at the average of four per day. What is the probability that exactly six machines break down in two days?

Solution: Given $\lambda = 4$, n = 6, t = 2

$P(n, t) = P(6, 4)$ when $\lambda t = 4$

we know, $P(n, t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$

$$P(6,2) = \frac{(4\times 2)^6 e^{-4\times 2}}{6!}$$

$$= \frac{8^6 e^{-8}}{720}$$

$$= 0.1221$$

Example: On an average, 6 customers arrive in a coffee shop per hour. Determine the probability that exactly 3 customers will reach in a 30 minute period, assuming that the arrivals follow Poisson distribution.

Solution: Given, $\lambda = 6$ customers / hour

$t = 30$ Minutes $= 0.5$ hour

$n = 2$

we know, $P(n, t) = \dfrac{(\lambda t)^n e^{-\lambda t}}{n!}$

$$P(6,2) = \frac{(6 \times 0.5)^2 e^{-6 \times 0.5}}{2!} = 0.22404$$

Example : A manager of a fast food restaurant observes that, an average of 9 customers is served by a waiter in a one-hour time period. Assuming that the service time has an exponential distribution, what is the probability that

a. A customer shall be free within 12 minutes.
b. A customer shall be serviced in more than 25 minutes.

Solution:

(a) Given, $\lambda = 9$ customers / hour

$t = 15$ minutes $= 0.25$ hour

Therefore, p (less than 15 minutes) ~~$= 1 - e^{-\lambda t}$~~

$= 1 - e - 9 \times 0.25$

$= 0.8946$

(b) Given, $\lambda = 9$ customers / hour

$t = 25$ minutes $= 0.4166$ hour

Therefore, P (more than 25 minutes) ~~$= 1 - e^{-\lambda t}$~~

~~$= 1 - e - 9 \times 0.4166$~~

~~$= 0.0025$~~

$= \int_0^{15} \lambda e^{-\lambda t} dt$

$= \frac{\lambda}{-\lambda} \left[e^{-\lambda t} \right]_0^{15}$

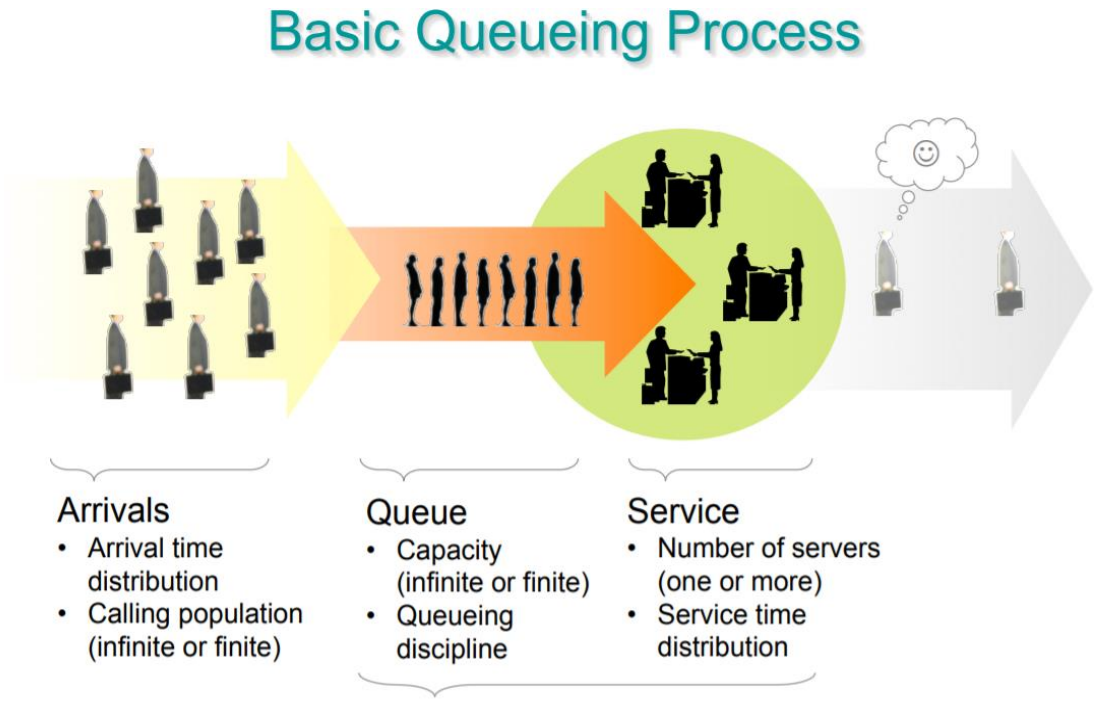
$= 1 - e^{-\lambda t}$

$= e^{-\lambda t}$

$= e^{-9 \times 0.4166}$

$=$

$\lambda_s = 11 \quad \lambda_2 = 8$



If we assume that arrivals to a queuing system follow a Poisson **process** and that service times are exponentially distributed, then the resulting queuing system is a **birth-and-death process**.

To analyse a queueing system, normally we try to estimate quantities such as the average number of customers in the system, the fluctuation of the number of customers waiting, the proportion of time that t

$P_n(t)$

(i) p_n = the probability that there are n customers in the system (waiting or in service) at an arbitrary epoch

(ii) a = offered load = $\frac{\lambda_n}{\mu_n}$

(iii) s = Number of servers

(iv) ρ = traffic intensity = utilization factor = offered load per server = a/s ($s < \infty$).

(v) Ws = mean waiting time in the system, i.e the mean length of time from the moment a customer arrives until the customer leaves the system (also called sojourn time).

(vi) Wq = mean waiting time in the queue, i.e. the mean length of time from the moment a customer arrives until the customer' service starts.

(vii) Ls = mean number of customers in the system, i.e. including all the customers waiting in the queue and all those being served.

(viii) Lq = mean number of customers waiting in the queue.

Suppose λ = 6 customers/hour and μ = 2 customers/hour •
Utilization is ρ= λ/(sμ)
If one server, s=1, ρ= λ/μ = 6/2 = 3, utilization > 1, so steady state will never be reached, queue length will increase to infinity in the long run
• If three servers, s=3, ρ= λ/(3μ) = 1 utilization = 1, queue is unstable and may never reach steady state • If four servers, s=4, ρ= λ/(4μ) = 3/4 utilization < 1, the queue will reach steady state and L is finite

Derivation of the steady-state probabilities.

In steady-state for a continuous-time Stochastic process the rate at which individuals transition into a particular state must be equal to the rate at which individuals transition out of that state.
In other words, A queueing system is in statistical equilibrium or in steady state if the probability that the system is in a given state is not time dependent:
That is, $P(X(t) = n) = p_n(t) = p_n$
If λ_n is the birth rate for state *n*, μ_n is the death rate for state *n*, and P_n is the probability of being in state *n* in steady-state, then we have the set of equations

$$\lambda_0 P_0 = \mu_1 P_1,$$
$$(\lambda_n + \mu_n) P_1 = \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}, \text{ for } n \geq 1.$$

The first equation gives us $P_1 = \frac{\lambda_0}{\mu_1} P_0$. Substituting into the second equation when $n = 1$ yields

$$P_2 = \frac{(\lambda_1 + \mu_1)P_1 - \lambda_0 P_0}{\mu_2} = \frac{(\lambda_1 + \mu_1)P_1 - \mu_1 P_1}{\mu_2} = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0.$$

This suggests that we might have $P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1}$ for $n \geq 1$. Assuming this, we have

$$P_{n+1} = \frac{(\lambda_n + \mu_n)P_n - \lambda_{n-1} P_{n-1}}{\mu_{n+1}} = \frac{(\lambda_n + \mu_n)P_n - \mu_n P_n}{\mu_{n+1}} = \frac{\lambda_n}{\mu_{n+1}} P_n.$$

Therefore, $P_n = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} P_0$.

Assuming that the birth and death rates are such that we actually have well-defined steady state probabilities, $\sum_{n=0}^{\infty} P_n = 1$, and so

$$\sum_{n=0}^{\infty} \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} P_0 = 1,$$

which implies

$$P_0 = \left(\sum_{n=0}^{\infty} \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} \right)^{-1}.$$

In general, then,

$$P_n = \left(\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) \left(\sum_{n=0}^{\infty} \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} \right)^{-1}.$$

Kendall-Lee Notation.

Since describing all of the characteristics of a queue inevitably becomes very wordy, a much simpler notation (known as Kendall-Lee notation) can be used to describe a system. Kendall-Lee notation gives us six abbreviations for characteristics listed in order separated by slashes.

A notation system for parallel server queues: A/B/c/d/N/K

- A represents the interarrival-time distribution
- B represents the service -time distribution

• c represents the number of parallel servers

- d represents queue discipline

• N represents the system capacity

• K represents the size of the calling population

Common symbols for A and B

• M Markov, exponential distribution • D Constant, deterministic

• Ek Erlang distribution of order k

• H Hyperexponential distribution

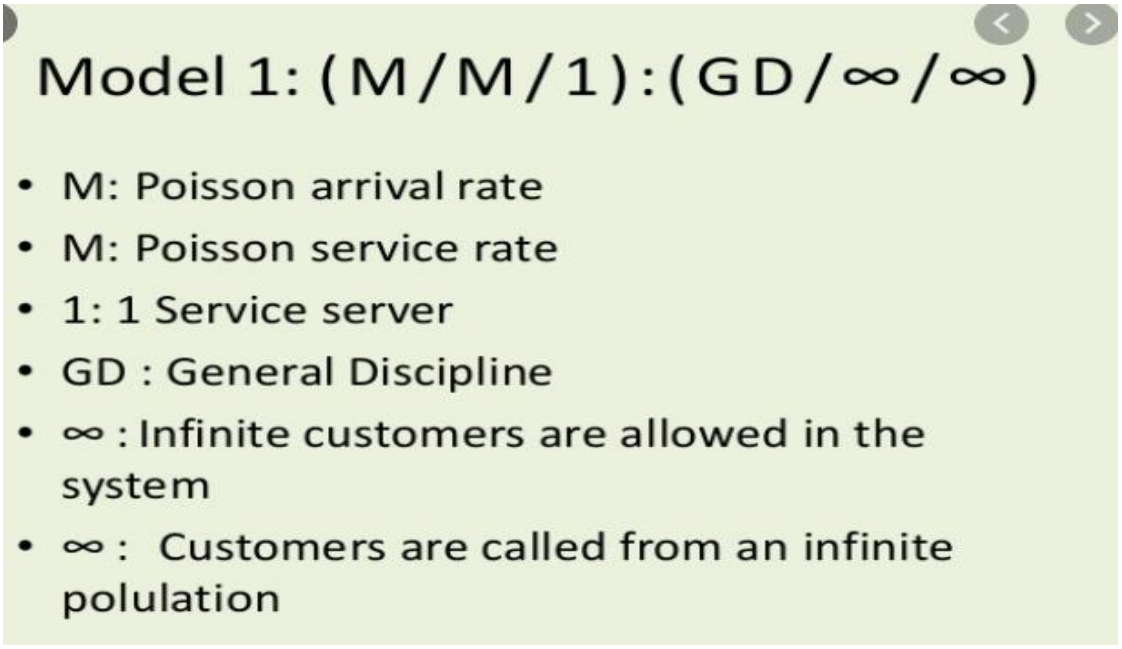
• G General, arbitrary

For example, **M/M/5/F CF S/20/** ∞ could represent a bank with 5 tellers, exponential arrival times, exponential service times, an FCFS queue discipline, a total capacity of 20 customers, and an infinite population pool to draw from.

Steady state probabilities and various average characteristics for the following models:

(i) (M/M/1) : (GD/ ∞ /∞) (ii) (M/M/1) : (GD/ N /∞) (iii) (M/M/c) : (GD/∞/∞) (iv) (M/M/c) : (GD/ N /∞) (v) (M/M/∞) : (GD/ ∞ /∞) (

(i) **(M/M/1) : (GD/ ∞ /∞)**



Application to the M/M/1 queuing system.

For an M/M/1 queuing system (i.e., a system with Markovian, or exponential, interarrival and service times, plus a single server), we have $\lambda_i = \lambda$ and $\mu_i = \mu$ for each i. In this case, the formula for P_0 simplifies nicely. We get

$$P_0 = \left(\sum_{n=0}^{\infty} \frac{\lambda^n}{\mu^n}\right)^{-1} = \left(\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n\right)^{-1} = \left(\frac{1}{1-\lambda/\mu}\right)^{-1} = 1 - \frac{\lambda}{\mu}.$$

In the second-to-last step we use the formula for a geometric series. This step requires $\lambda < \mu$; otherwise, the infinite series fails to converge. This should make sense: If the arrival rate is larger than the service rate then over time the queue will get longer and longer and so will never approach steady-state. (It is perhaps harder to see intuitively that we do not get a steady state if the arrival and service rates are equal, but that falls out of the mathematics here.)

In general, then, for the M/M/1 queuing system, the probability that there are exactly n customers in the system in steady-state is given by $P_n = \frac{\lambda^n}{\mu^n}\left(1 - \frac{\lambda}{\mu}\right)$.

Working Formulae

1. Probability of zero units in the queue $(P_0) = 1 - \frac{\lambda}{\mu}$
2. Average queue length $(L_q) = \frac{\lambda^2}{\lambda(\lambda - \mu)}$
3. Average number of units in the system $(L_S) = \frac{\mu}{\lambda - \mu}$
4. Average waiting time of an arrival $(W_q) = \frac{\lambda}{\mu(\lambda - \mu)}$
5. Average waiting time of an arrival in the system $(W_S) = \frac{1}{\lambda - \mu}$

Example 1: A Television repairman finds that the time spent on his jobs has an exponential distribution with mean 30 minuets. If he repairs the sets in the order in which they come in, and if the arrivals of sets are approximately Poisson with an average rate of 10 per 8 hours day which is the repairs man idle time each day ?Find the expected number of units in the system and in the queue?

Solution: it is a (M\M\1: (∞\FCFS) queuing system.

Where, λ= mean arrival rate =10/8 units per hour

μ=mean service rate =2mins per hour.

Therefore ρ=λ/μ=10/8.2=5/8

1. Expected number of units in the system $L_S = \frac{\rho}{1-\rho} = \frac{5}{3}$ sets

2. Expected number of units in the queue $(L_q) = \frac{\rho^2}{(1-\rho)} = (L_q) = \frac{\frac{5}{8}^2}{(1-\frac{5}{8})} =$

3. Probability of repairman being idle=probability of having no T.V sets in the system

$(p_0) = 1-\rho = 1-\frac{5}{8} = \frac{3}{8}$

4. Therefore repairman will remain idle for $\frac{3}{8} \times 8 = 3$ hours per day.

Example 2:

Customers arrive at a fast food restaurant at a rate of 100 per hour and take 30 seconds to be served. • How much time do they spend in the restaurant? How much time waiting in line?

What is the server utilization? How many customers in the restaurant?

Solution: Service rate = μ = 60/0.5=120 customers per hour

time spend in the restaurant = W_s = 1/μ−λ = 1/(120-100) = 1/20 hrs = 3 minutes

time waiting in line = W_q = W_s - 1/μ = 2.5 minutes

What is the server utilization? – ρ = λ/μ = 5/6

How many customers in the restaurant = L_s =

Example 3: 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: λ = 1/10; μ = 1/8.

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

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

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

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

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

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

6. If the arrival and departure rates in a public telephone booth with a single phone are 1/12 *and1 /14* **respectively, find the probability that the phone is busy.**

Answer:

P[Phone is busy] = 1 – P [No customer in the booth]

$$= 1 - P_0 = 1 - \left(1 - \frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu} = \frac{1}{12} / \frac{1}{4} = \frac{1}{3}$$

5. If the arrival and departure rates in a M/M/1 queue are 1/2 per minute and 2/3 per minute respectively, find the average waiting time of a customer in the queue.

(ii) (M/M/1) : (GD/ c /∞)

An M/M/1/GD/c/∞ queuing system has exponential interarrival and service times, with rates λ and μ respectively. This system is very similar to the previous system, except that whenever c customers are present in the system, all additional arrivals are excluded from entering, and are thereafter no longer considered. For example, if a customer were to walk up to a fast food restaurant and see that the lines were too long for him to want to wait there, he would go to another restaurant instead. A system like this can be modeled as a birth-death process with these parameters: $\lambda j = \lambda$ for j = 0, 1, ..., N − 1 $\lambda_N = 0$ $\mu 0 = 0$ $\mu j = \mu$ for j = 1, 2, ..., N

20.5 The *M/M/1/GD/c/∞* Queuing System

When c customers are present, all arrivals are turned away and lost to the system

$$L = \begin{cases} \dfrac{\rho\left[1-(c+1)\rho^{c}+c\rho^{c+1}\right]}{(1-\rho^{c+1})(1-\rho)} & \text{when } \lambda \neq \mu \\ \dfrac{c}{2} & \text{when } \lambda = \mu \end{cases}$$

$$L_s = 1-\pi_0$$

$$L_q = L - L_s$$

$$W = \frac{L}{\lambda(1-\pi_c)} \quad \text{and} \quad W_q = \frac{L_q}{\lambda(1-\pi_c)}$$

where

$$\pi_c = \rho^{c}\pi_0 \quad \text{and} \quad \pi_0 = \frac{1-\rho}{1-\rho^{c+1}}$$

$$\pi_c = P(c \text{ customers in the system})$$

6. A student counselling centre can accommodate at most 5 students. The students arrive according to Poisson distribution at the rate of 4 per hour. The single counsellor can counsel only 5 students per hour and counselling time is exponentially distributed. Any student overflow is directed to another centre.

   a. Determine probability distribution for number of students either waiting for or receiving treatment at given time.
   b. Determine average number of students in the counselling centre and average number of students waiting to see the counsellor.

3

20.6 The *M/M/s/GD/∞/∞* Queuing System

$$\rho = \frac{\lambda}{s\mu}, \text{ where } s\mu \text{ is the maximum service rate}$$

$$\pi_0 = \frac{1}{\sum_{i=0}^{(s-1)} \frac{(s\rho)^i}{i!} + \frac{(s\rho)^s}{s!(1-\rho)}}$$

$$\pi_j = \begin{cases} \frac{(s\rho)^j \pi_0}{j!} & (j = 1,2,\ldots s) \\ \frac{(s\rho)^j \pi_0}{s!s^{j-s}} & (j = s, s+1, s+2,\ldots) \end{cases}$$

$$L_q = \frac{A\rho}{1-\rho}$$

Let A $P(j \geq s) = \frac{(s\rho)^s \pi_0}{s!(1-\rho)}$

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

$$L_s = L_q + \rho$$

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

7. A post office has 3 windows providing the same service. It receives on an average 30 customers per hour. Arrivals are in Poisson type and service time is exponentially distributed. Each window serves on an average 12 customers per hour.

Obtain

i. The average number of customers in the system.

ii. The average total time that a customer spends in the post office.