Chapter 7

Queuing Theory

Babita Goyal

Key words: queue, waiting time, customer, server, service time, queue discipline, expected waiting time, expected service time, arrival process, service time distribution, departure process, steady-state system.

Suggested readings:

1. Swarup, K., Gupta, P.K. and Man Mohan (2001), Operations research, Sultan Chand and Sons.
2. Medhi, J. (1996), Stochastic Processes, New Age International (P) Ltd.
3. Feller, W.(1968), An introduction to Probability Theory and its Applications, Vol. I ,Wiley, New York, 3rd edition.
4. Karlin, S. and Taylor, H.M.(1975), A first course in Stochastic Processes, Academic Press.
5. Parzen, E.(1962), Introduction to Stochastic Processes, Universal Book Stall.
6. Ross, S.M.(1983), Stochastic Processes, John Wiley.

7.1 Introduction

As discussed earlier, a flow of customers from an infinite/finite source (population) towards a service facility forms a queue (waiting line) on account of lack of capability to serve them all at a time. Thus a queuing system may be described as the flow of units towards a system in the hope of getting service, forming or joining a queue, if service is not immediately available and leaving the system after being served.

Waiting phenomenon is the direct consequence of randomness in the phenomenon of interaction between the system and the customers and in the operation of service facilities. In general, the customers' arrival and their service times are not known in advance, for otherwise the operation of the facility could be scheduled in such a manner that would eliminate waiting completely.

The objective of the queuing theory is the study of the operations of a service facility under random conditions in order to secure some characteristics that measure the performance of the system under study. For example, a logical measure of performance is how long a customer is expected to wait before being served. Another measure is the percentage of time the service facilities are not in use. The first measure looks at the system from the customers' point of view whereas the second measure evaluates the degree of utilization of the service facility, i.e., from the management's point of view. Obviously, the larger the customers' waiting time, the smaller is the percentage of time the facility will remain idle, and vice-versa. A longer waiting time may have a negative impact on the new arrivals as the customers may be reluctant to very long waiting periods. However, if the service facility is not utilized to its optimum capacity, it may increase the cost of service operations. Then the objective of the queuing theory is to strike a balance between the two.

7.2 Queuing systems: Some general concepts

Basic elements of a queuing model: A queuing model may be described as follows:

A customer arrives at a service facility and joins a queue. The server chooses a customer from the queue to begin service. Upon the completion of a service, the process of choosing a new customer from the queue is repeated. It is assumed that no time is lost between the release of a serviced customer and admission of a new one from the waiting line.

Two principal components of a queuing model, viz., the customer and the server interact only for the period of time the customer needs to complete his service. The basic elements of a queuing model depend on the following factors:

1. **Input processes:**
    1. Arrival distribution (single or bulk arrivals).
    2. Queue size (finite or infinite).
    3. Calling source (finite or infinite).
    4. Human behavior (jockeying, balking or reneging).

2. **Service processes:**
    1. Service-time distribution (single or bulk service).
    2. Design of service facility (series, parallel or network stations).
    3. Service discipline (FCFS, LCFS, SIRO) and service priority.

1. **Input processes:** Various components of input processes of a queuing model can be described as follows:

    1. **Arrival distribution:** In queuing models, the customer arrivals are summarized in terms of a probability distribution, which is known as the arrival distribution. The customers may arrive at the service facility in batches of fixed sizes or of variable sizes or one-by-one. If two or more customers arrive the system simultaneously, the arrival is said to be in bulk.

    2. **Queue size:** The queue size may be finite or infinite depending upon the capacity of the system. If the queue size is finite and the queue fills to the capacity, new arrivals are denied service. These arrivals may be lost forever.

3. **Calling source:** This is the source from which calls for service (arrival of customers) are generated. The calling source may be of finite or infinite capacity. In case of a finite source, an arrival affects the rate of arrival of new customers.

4. **Human behavior:** Queuing models representing the situations in which human beings take the role of customers and/or servers must be designed to account for the effect of human behavior. A "human" server may speed up the rate of service when the queue builds up in size. A "human" customer may jockey from one queue to another in the hope of reducing waiting time. Some "human" customers may also balk from joining a queue because they anticipate a long waiting or they may renege after being in queue for a while because their wait has been too long. However, a long wait for one person may not be as long as for another. All customers in the queue are expected to behave equally while in the facility.

2. **Service processes:** These are the processes related to the service providing mechanism of the system and can be described as follows:

1. **Service-time distribution:** Service-time distribution represents the service offered in terms of service time. The service may be offered individually (as in banks) or in groups (as in restaurants). The queues in the later case are called the bulk queues.

2. **Design of the service facility:** The service facility may include one or more servers. If there are infinite number of servers then all the arriving customers are serviced instantaneously on arrival and there will be no queue. In case all the servers offer the same service, the system is said to have parallel channels. On the other hand, the service may comprise of a number of series stations through which the customer has to pass before the service is completed (e.g., processing of a product on a sequence of machines). In this case, waiting lines may or may not be allowed between the stations. Such queues are called the queues in series or the queues in tandem. The most general design of a service facility includes both series and parallel processing stations. Such systems are called network queues.

3. **Service discipline:** This is the rule according to which customers are selected for the service when a queue has been formed. The most common disciplines are the FCFS (first come, first served) and SIRO (service in random order). Another service discipline is LCFS (last come, first served). The customers arriving at the facility may be put in priority queues such that those with a higher priority will receive preference to start service with (pre-emptive service) or a customer of low priority is serviced before a customer of higher priority (non pre-emptive service). Another service discipline is processor sharing, when m customers receive service simultaneously at the rate $\frac{1}{m}$ each.

In case of parallel channels, FSR (fastest service rule) is adopted, i.e., on customer's arrival if only one server is free, then the incoming customer is assigned to that server and if more than one server are free, then the customer is assigned to a server having the largest service rate among the free ones.

7.3 Stochastic nature of queuing models

As we discussed earlier also, queuing models are stochastic models and they need probability distributions to be explained and analyzed.

(i) Probability distributions associated with the queuing models:

We have seen that if a sequence of events (arrivals or departures in case of queuing models) occur according to a Poisson process, then the inter-event time is distributed exponentially with the same parameter as that of the occurrence process. A queuing system may be identified as a birth (-death) process.

Arrival process: In this model, events are pure arrivals and no departures are there, i.e., once a unit joins the system as a customer, it never departs. This is a pure birth process. Examples are registrations of birth, deaths or other vital events. If the rate of arrival is λ, then

$$p_n(t) = P(n \text{ arrivals in an interval of length } t) = e^{-\lambda t}\frac{(\lambda t)^n}{\underline{\lfloor n}} \quad , n = 0, 1, 2...$$

Departure process: This is a pure death process, which assumes that the system starts with a given number of customers who leave the facility at a rate μ after being serviced while no fresh arrivals are allowed to join the system. As an example, consider an inventory stock where N items have been made available at the beginning of the stocking period. No fresh stock are added to the inventory till the next period starts and existing stock is used sequentially at a rate of μ units per unit of time, i.e., if

$$q_n(t) \;=\; P(n \text{ departures in an interval of length } t),$$

then

$$\begin{aligned}
q_0(h) &= e^{-\mu h} \approx 1-\mu h, \text{ for very small } h. \\
q_1(h) &= 1-q_0(h) \approx \mu h \\
q_n(t+h) &= q_n(t)(1-\mu h)+q_{n-1}(t)\mu h;\;\; 1 \le n \le N \\
&= q_N(t)+q_{N-1}(t)\mu h;\;\; n = N \\
\Rightarrow\; q_n(t) &= (\mu t)^n \frac{e^{-\mu t}}{\underline{|n}};\;\; n = 0,1,2,\ldots N-1 \\
&= 1-\sum_{n=0}^{N-1} q_n(t);\;\; n = N
\end{aligned}$$

and

$$\begin{aligned}
p_n(t) &= P(n \text{ customers after time } t) \\
&= q_{N-n}(t) \;=\; (\mu t)^{N-n} \frac{e^{-\mu t}}{\underline{|N-n}};\;\; n = 1,2,\ldots,N \\
&= 1-\sum_{n=1}^{N} p_n(t);\;\; n = 0
\end{aligned}$$

Arrival-departure process: This is the general birth-death process where arrivals and departure occur simultaneously. Customers arrive (births occur) at a rate λ and after getting service, they depart (deaths occur) at rate μ per unit of time.

(ii) Operating characteristics of a queuing model:

These characteristics provide a measure for the evaluation of efficiency of a queuing model and are needed to design a new system. A queuing model can be a transient system or a steady-state system.

Transient system: If the operating characteristics of the system are functions of time, then the system is called a transient system

Steady-state system: If the operating characteristics of the system are independent of time, then the system is called a steady-state system. Most of the queuing systems are designed to be steady-state systems, i.e., the systems will attain stability after a sufficiently long period of time.

Following characteristics are used to study the performance of the system under steady-state conditions:

1. **Expected number of customers in the system:** Denoted by $E(n)$ or L_s, expected number of customers in the system is the average number of customers, both in the queue and the service, where n is a realization of the random variable N, the number of customers in the system.

2. **Expected number of customers in the queue:** Denoted by L_q, expected number of customers in the queue is the average number of customers in the waiting or in the queue.

3. **Expected waiting time in the system:** Denoted by W_s, expected waiting time in the system is the average total time a customer spends in the system. It includes his waiting as well as service time.

4. **Expected waiting time in the queue:** Denoted by W_q, expected waiting time in the queue is the average total time a customer spends in the waiting on joining the system before being served.

5. Server utilization factor (busy period or traffic intensity): Denoted by ρ, the server utilization factor is the proportion of time that a server is actually busy in providing service to the customers.

If p_n is the steady-state probability of n customers in the system, then we have the following relations:

(i) $$L_s = \sum_{n=0}^{\infty} np_n$$

$$L_q = \sum_{n=c}^{\infty} (n-c)\, p_n \; ; \; c \text{ is the number of servers in the system.}$$

(ii) For arrival rate λ,

$$L_s = \lambda W_s$$
$$L_q = \lambda W_q$$

(iii) If the system capacity is finite, then the effective arrival rate for those who join the system is not same as the arrival rate as the pressure of new arrivals diminishes once the system tends to approach saturation point and if the effective arrival rate is denoted by λ_{eff}, then in general $\lambda_{eff} = \beta\lambda \ (0 < \beta < 1)$. Then

$$L_s = \lambda_{eff} W_s$$
$$L_q = \lambda_{eff} W_q$$

(iv) If μ is the service rate, then the expected waiting time in the service is $\frac{1}{\mu}$, so

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

$$\text{or,} \quad \lambda W_s = \lambda W_q + \frac{\lambda}{\mu}$$

$$\Rightarrow \quad L_s = L_q + \frac{\lambda}{\mu} \quad \text{or,} \quad \lambda = \mu(L_s - L_q)$$

and as the relation holds for λ_{eff} in case of finite system capacity also so, ,

$$\lambda_{eff} = \mu(L_s - L_q)$$

(v) All these basic measures can be combined as

$$p_n \rightarrow L_s = \sum_{n=0}^{\infty} np_n \rightarrow W_s = \frac{L_s}{\lambda} \rightarrow W_q = W_s - \frac{1}{\mu} \rightarrow L_q = \lambda W_q$$

7.4 Representation of a queuing system: Some standard notations

As we have seen earlier, a queuing system can be identified with its major characteristics. Kendall developed a standard set of notations to represent a queuing system, which was further modified by Lee. These notations are now-a-days known as Kendall-Lee notations. The major characteristics of a queuing system are represented as

$$(a/b/c);(d/e/f)$$

where,

$a \equiv$ arrival distribution

$b \equiv$ service time (departure) distribution

$c \equiv$ number of parallel servers $(c = 1, 2, ...\infty)$

$d \equiv$ service discipline (FCFS, LCFS, SIRO etc.)

$e \equiv$ system capacity

$f \equiv$ size of the calling source.

The arrival and departure distributions can be according to the following rules:

1. If arrivals (departures) are according to a Poisson (Markovian) distribution i.e., exponential inter-arrival (service) time, then the distribution is denoted by the letter *M*.
2. If inter-arrival (service) time is deterministic, then the distribution is denoted by the letter *D*.
3. If inter-arrival (service) time is an Erlang (gamma with parameter *k*) distribution, then the distribution is denoted by the letter E_k.
4. In case, the inter-arrival times are according to some general distribution but are independent, then the letter *GI* denotes the distribution.
5. In case, the inter-arrival times are according to some general distribution, then the letter *G* denotes the distribution.

For example, the notation $(M / D / 10);(GD / N / \infty)$ can be interpreted as follows:

The queuing system follows Poisson arrivals, service-time is deterministic, 10 parallel servers are there in the system, general service discipline, system capacity is finite and the system can accommodate at the most N customers including those in the service and arrivals are being generated from a source of infinite capacity.

7.5 Some queuing systems

Now, we study some queuing systems, which are Poisson processes. In these systems, the service discipline is general except for the cases where waiting time distribution have been obtained in which case, the service discipline is FCFS.

7.5.1 $(M/M/1);(GD/\infty/\infty)$

This is a single server model with infinite system capacity, where both arrival and departure process are Poisson processes. For this system, we want to obtain steady-state probability and study the operational characteristics of the system.

Let $p_n(t) = P(n$ customers in the system in an interval of length $t)$

and $p_n = \lim_{t\to\infty} p_n(t)$ is the steady-state probability of the system.

This is a birth-death process with birth rate, say, λ and death rate, say, μ. Then the differential-difference equations for this process are

$$p_n'(t) = -(\lambda+\mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t); \quad n > 0$$
$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t)$$

Under steady-state conditions, as $n \to \infty$

$$p_n'(t) \to 0, \text{ and } p_n(t) \to p_n$$

and the differential-difference equations tend to

$$\lambda p_{n-1} + \mu p_{n+1} - (\lambda+\mu)p_n = 0; \quad n > 0$$
$$\lambda p_0 - \mu p_1 = 0$$

or,

$$(\rho+1)p_n = \rho p_{n-1} + p_{n+1}; \quad n > 0 \tag{7.1}$$

$$(\rho+1)p_0 = p_0 + p_1 \tag{7.2}$$

where $\rho = \dfrac{\lambda}{\mu}$ is the traffic intensity.

Multiplying (7.1) by s^n, summing over all possible values of n and adding the resultant sum to (7.2), we have

$$\Rightarrow (\rho+1)\sum_{n=1}^{\infty} p_n s^n + (\rho+1)p_0 = \rho\sum_{n=1}^{\infty} p_{n-1}s^n + \sum_{n=1}^{\infty} p_{n+1}s^n + p_0 + p_1$$

$$\Rightarrow (\rho+1)P(s) = \rho s P(s) + \frac{1}{s}P(s) + \frac{s-1}{s}p_0$$

or, $$\left(\rho+1-\rho s-\frac{1}{s}\right)P(s) = \frac{s-1}{s}p_0$$

or, $$P(s) = \frac{s-1}{(s-1)-\rho s(s-1)}p_0 = \frac{1}{1-\rho s}p_0 = p_0(1-\rho s)^{-1}$$

$$\Rightarrow p_n = p_0\rho^n,\ n = 0,1,2...$$

and as $$\sum_{n=0}^{\infty} p_n = 1 = p_0\sum_{n=0}^{\infty}\rho^n$$

$$\Rightarrow \frac{p_0}{1-\rho} = 1 \text{ provided } \rho<1$$

$$\Rightarrow p_0 = 1-\rho = 1-\frac{\lambda}{\mu}$$

and, $$p_n = p_0\rho^n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n$$

Performance of the model:

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

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

$$= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} = E(n)$$

(ii) $$L_q = L_s-\frac{\lambda}{\mu} = \frac{\rho^2}{(1-\rho)} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

(iii) $$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}$$

(iv) $$W_q = \frac{L_q}{\lambda} = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Some interesting properties of the model:

(ii) The probability that the queue size will be greater than or equal to n, the number of customers in the system is given by

$$P(\text{queue length } \geq n) = \sum_{k=n}^{\infty} p_k = \sum_{k=n}^{\infty} (1-\rho)\rho^k$$

$$= \frac{(1-\rho)\rho^n}{1-\rho} = \rho^n$$

(iii) Average length of non-empty queue is

$$E(n-1 \mid n-1>0) = \frac{E(n-1)}{P(n>1)} = \frac{\lambda^2}{\mu(\mu-\lambda)} \frac{1}{\left(\frac{\lambda}{\mu}\right)^2} = \frac{\mu}{\mu-\lambda}$$

$$\because \; P(n>1) = \sum_{n=0}^{\infty} p_n - p_0 - p_1 = \left(\frac{\lambda}{\mu}\right)^2$$

(iv) The fluctuation (variance) of the number of customers in the system is given by

$$\text{Var}(n) = \sum_{n=0}^{\infty} n^2 p_n - E^2(n) = \sum_{n=0}^{\infty} n^2 \rho^n (1-\rho) - \frac{\rho^2}{(1-\rho)^2}$$

$$= \frac{\rho}{(1-\rho)^2} = \frac{\lambda\mu}{(\mu-\lambda)^2}$$

(v) We now find the waiting time of the customer, based upon FCFS discipline, who finds n customers ahead of him when he reaches the system.

Let τ be the amount of time a person just arriving must wait in the system (until his service is completed). Then

$$\tau = t_1' + t_2 + ... + t_{n+1}$$

where t_1' is the time needed for the customer in service to complete the service and $t_2, \ldots, t_n$ are the service times of n-1 customers ahead in the queue. t_{n+1} is the service time of the customer who has just joined the queue.

Let $\varpi(\tau \mid n+1)$ be the conditional p.d.f. of τ given n customers ahead in the system upon the new arrival. Then $t_i \sim \exp(-\mu)\ \forall i$ and due to lack of memory property $t_1' \sim \exp(-\mu) \Rightarrow \tau \sim Gamma(\mu, n+1)$

$$\Rightarrow \varpi(\tau) = \sum_{n=0}^{\infty} \varpi(\tau \mid n+1) p_n = \sum_{n=0}^{\infty} \mu(\mu\tau)^n \frac{e^{-\mu\tau}}{\lfloor n}(1-\rho)\rho^n$$

$$= (1-\rho)\mu e^{-\mu\tau} \sum_{n=0}^{\infty} \frac{(\mu\rho\tau)^n}{\lfloor n}$$

$$= (1-\rho)\mu e^{-\mu(1-\rho)\tau}; \quad \tau > 0$$

which is an exponential distribution with mean

$$E(\tau) = \frac{1}{\mu(1-\rho)} \quad \left(= \frac{L_s}{\lambda}\right)$$

An interesting feature of the waiting time distribution is that the probability p_n for an $(M/M/1);(GD/\infty/\infty)$ system is completely independent of the service discipline and expected waiting times in the queue as well as system are completely independent of the service discipline, still the probability distribution of the waiting time depends on the type of service discipline used. (See Feller). However, whatever is the service discipline, the expected waiting time must remain the same.

(vi) $P($ a customer has to wait mote than $W_s) = P(\tau > W_S) = 1 - \int_0^{W_S} \varpi(\tau) d\tau$

$$= e^{-\mu(1-\rho)W_S} = e^{-1} \approx 0.368$$

i.e., under FCFS scheme, 36.8% customers will have to wait more than the average waiting time.

(vii) However, the waiting time of a customer before he gets the service is given by the probability distribution of W_q, which on using a similar argument, is given by

$$\psi(\tau') = \sum_{n=0}^{\infty} \psi(\tau' \mid n) p_n = (1-\rho)\lambda e^{-(\mu-\lambda)\tau}; \quad \tau > 0$$

Examples

(1) A worker in an assembling unit of a factory finds that a typical assembling job, on average requires 30 minutes whereas the jobs are independently distributed exponentially. The jobs are to be performed on a FCFS basis. If jobs arrive on an average rate of 10 per 8-hour shift according to a Poisson rule, find (i) his expected busy time per day (ii) the number of jobs ahead of the job just arrived.

Sol: The arrival rate λ = 10 jobs per day and the service rate μ = 16 jobs per day.

$$\therefore\ P(\text{ the worker remains busy}) = \rho = \frac{\lambda}{\mu} = \frac{10}{16} = 0.625$$

(i) Expected busy time per day = duration of the day $\times\ \rho$ = 8 $\times$ 0.625 = 5 hrs/day.

(ii) The number of jobs ahead of the job just arrived is the expected number of jobs in the system

$$\therefore\ E(n) = \frac{\rho}{1-\rho} = \frac{0.625}{0.375} \approx 2 \text{ (Jobs cannot be in fractions)}$$

(2) The rate of arrival of customers at a public telephone booth follows Poisson distribution with an average inter-arrival time of 10 minutes. The duration of a phone call is assumed to be an exponential variable with mean time of 3 minutes.

(a) What is the probability that a person arriving at the booth will have to wait?

(b) What is the average length of the non-empty queue that forms from time to time?

(c) The owner company of the booth is planning to install a second booth when convinced that the expected waiting time for the customers is at least 3 minutes. What should be the pattern of flow of customers in order to have a justification for the second booth?

(d) Estimate the fraction of the day the phone will be in use.

(e) What is the probability that a new arrival will have to spend 10 minutes in all in the system?

Sol: In this case,

$$\lambda = \frac{1}{10} \times 60 = 6 \text{ per hour and } \mu = \frac{1}{3} \times 60 = 20 \text{ per hour}$$

$$\text{and } \rho = \frac{\lambda}{\mu} = \frac{6}{20} = 0.3$$

(a) $P(\text{a new arrival will have to wait }) = P(\tau' > 0) = 1 - p_0$

$$= 1 - \left(1 - \frac{\lambda}{\mu}\right) = 0.3$$

(b) Average length of non-empty queue $E(m \mid m > 0) = \dfrac{\mu}{\mu - \lambda} = 1.43$

(c) The second booth will be installed if the expected waiting time is at least 3 minutes, i.e., the arrival rate has to be more than the current arrival rate. Let λ' be the new arrival rate. Then

$$E(\tau') = \frac{\lambda'}{\mu(\mu - \lambda')} \Rightarrow \frac{3}{60} = \frac{\lambda'}{20(20 - \lambda')}$$

$$\Rightarrow \lambda' = 10 \text{ per hour.}$$

(d) The fraction of time the phone is busy is the traffic intensity $\rho = 0.3$.

(e) $P(\tau \geq 0) = \int_{10}^{\infty} \lambda \left(1 - \frac{\lambda}{\mu}\right) e^{-(\mu - \lambda)t} dt = 0.03$

i.e., on average 3% of the arrivals will have to spend 10 or more minutes before leaving the system after making a call.

(3) In the production shop of a company, the break down of machines is found to be Poisson with an average rate of 3 machines per hour. Break down time at one-machine costs *Rs*. 40 per hour to the company. There are two choices before the company for hiring the repairman. One of the repairman is slow but cheap, the other is fast but expensive. The slow-cheap repairman demands *Rs*.20 per hour and will repair the broken down machines exponentially at a rate of 4 per hour. The fast-expensive repairman demands *Rs*. 30 per hour and will repair the machines exponentially at a rate of 6 machines per hour. Which repairman the company should hire?

Sol: **Case I:** Slow-cheap repairman

$\lambda = 3$ machines / hour ; $\mu = 4$ machines / hour $\Rightarrow \rho = \frac{3}{4}$

and average down-time of a machine $= L_q = \frac{1}{\mu(1-\rho)} = \frac{1}{4\left(\frac{1}{4}\right)} = 1hr.$

So average down-time of 3 machines which arrive in 1 $hr.$ $= 3\times 1 = 3hrs.$

$\Rightarrow$ down-time cost $= Rs.40\times 3 = Rs.120$

charges paid to the repairman $= Rs.20\times 3 = Rs.60$

$\Rightarrow$ Total cost $= Rs.120 + Rs.60 = Rs.180$

Case II: Fast-expensive repairman

$\lambda = 3$ machines / hour ; $\mu = 6$ machines / hour $\Rightarrow \rho = \frac{1}{2}$

and average down-time of a machine $= L_q = \frac{1}{\mu(1-\rho)} = \frac{1}{6\left(\frac{1}{2}\right)} = \frac{1}{3}hrs.$

So average down-time of 3 machines which arrive in 1 $hr.$ $= 3\times\frac{1}{3} = 1hr.$

$\Rightarrow$ down-time cost $= Rs.40\times 1 = Rs.40$

charges paid to the repairman $= Rs.30\times 1 = Rs.30$

$\Rightarrow$ Total cost $= Rs.40 + Rs.30 = Rs.70$

So the company should hire the second repairman.

7.5.2 $(M/M/1);(GD/N/\infty)$

This is a single server model with finite system capacity N, i.e., maximum permissible queue length is N-1 and all the new arrivals either balk or are not allowed to join the system. As a result, the effective arrival rate λ_{eff} becomes a function of number of customers present in the system and is less than the rate λ at which the customers are generated. Arrival and departure, both processes are Poisson processes. For this system, we want to obtain steady-state probability and study the operational characteristics of the system.

The differential-difference equations for this process are

$$
\begin{aligned}
p_0'(t) &= -\lambda p_0(t) + \mu p_1(t);\;\; n=0 \\
p_n'(t) &= -(\lambda+\mu) p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t);\;\; 0<n\le N-1 \\
p_N'(t) &= -\mu p_N(t) + \lambda p_{N-1}(t);\;\; n=N \\
&\qquad (\text{ as } p_N(t+h) = p_N(t)(1-\mu h) + p_{N-1}(t)(\lambda h)(1-\mu h))
\end{aligned}
$$

Under steady-state conditions, as $n \to \infty$, the differential-difference equations tend to

$$
\begin{aligned}
-\rho p_0 + p_1 &= 0 \\
-(1+\rho) p_n + p_{n+1} + \rho p_{n-1} &= 0;\;\; 0<n\le N-1 \\
-p_N + \rho p_{N-1} &= 0;\;\; n=N
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow\;\; (\rho+1)\sum_{n=1}^{N} p_n s^n &= \sum_{n=1}^{N-1} p_{n+1} s^n + \rho \sum_{n=1}^{N} p_{n-1} s^n + p_0 + \rho p_N s^N \\
\Rightarrow\;\; (\rho+1)P(s) &= \frac{1}{s}\Big(P(s)-p_0\Big) + \rho s\Big(P(s)-p_N s^N\Big) + p_0 + \rho p_N s^N \\
&= \left(\frac{1}{s}+\rho s\right)P(s) + \left(1-\frac{1}{s}\right)p_0 + (s-1)\rho p_N s^N
\end{aligned}
$$

or,

$$
\begin{aligned}
P(s) &= p_0(1-\rho s)^{-1} - \rho p_N s^{N+1}(1-\rho s)^{-1} \\
\Rightarrow\;\; p_n &= \text{coefficient of } s^n \text{ in } P(s) \\
&= \begin{cases} p_0 \rho^n, & n = 0,1,2\ldots < N \\ p_0\rho^n - p_N \rho^{n-N}, & n \ge N+1 \end{cases}
\end{aligned}
$$

and as $$p_0\rho^n - p_N\rho^{n-N} = p_0\rho^n - p_0\rho^N\rho^{n-N} = 0$$

$$\Rightarrow p_n = p_0\rho^n;\ n = 0,1,2,\ldots,N$$

and as $p_0 \sum_{n=0}^{N}\rho^n = 1 \Rightarrow p_0\left(\frac{1-\rho^{N+1}}{1-\rho}\right) = 1$

$$\Rightarrow p_0 = \frac{1-\rho}{1-\rho^{N+1}};\ \text{and}\ p_n = \left(\frac{1-\rho}{1-\rho^{N+1}}\right)\rho^n;\ n = 0,1,2,\ldots,N$$

$$\Rightarrow p_n = \begin{cases}\left(\frac{1-\rho}{1-\rho^{N+1}}\right)\rho^n; & \rho\neq 1,\\ & \qquad n = 0,1,2,\ldots,N\\ \frac{1}{N+1}; & \rho = 1\end{cases}$$

In this case, $\rho = \frac{\lambda}{\mu}$ need not be less than 1 as ρ is controlled the number of customers in the system and not by the relative sizes of λ and μ.

Performance of the model:

(i) $$L_s = \sum_{n=0}^{\infty} np_n = \frac{1-\rho}{1-\rho^{N+1}}\sum_{n=0}^{N} n\rho^n = \frac{1-\rho}{1-\rho^{N+1}}\rho\frac{d}{d\rho}\left(\frac{1-\rho^{N+1}}{1-\rho}\right)$$

$$= \begin{cases}\frac{\rho\left(1-(N+1)\rho^N + N\rho^{N+1}\right)}{(1-\rho)(1-\rho^{N+1})}; & \rho\neq 1\\ \frac{N}{2}; & \rho = 1\end{cases}$$

Now, $P($ a customer does not join the system $) = p_N$

$$\Rightarrow \lambda_{eff} = \lambda(1-p_N)$$

(ii) $$L_q = L_s - \frac{\lambda_{eff}}{\mu} = L_s - \frac{\lambda(1-p_N)}{\mu}$$

(iii) $$W_q = \frac{L_q}{\lambda_{eff}} = \frac{L_q}{\lambda(1-p_N)}$$

(iv) $$W_s = W_q + \frac{1}{\mu} = \frac{L_s}{\lambda(1-p_N)}$$

Example:

(4) At a railway station only one train is handled at one time. The railway yard is sufficient only for two trains to wait while the other is given signal to leave the station. Trains arrive at the station at an average rate of 6 per hour and the railway station can handle them at a rate of 12 per hour. Assuming Poisson arrivals and exponential service time distributions, find the steady-state probabilities for various numbers of trains in the system and the average waiting time of a new train coming into the yard.

Sol: In this case,

$$\lambda = 6, \text{ and } \mu = 12 \Rightarrow \rho = 0.5$$
$$N = 2+1 = 3$$

$$p_0 = P(\text{ no train in the system}) = \frac{1-\rho}{1-\rho^{N+1}} = \frac{0.5}{1-(0.5)^4} = 0.53$$

$$\text{and } p_N = p_0 \rho^n$$

$$\Rightarrow L_s = \sum_{n=0}^{3} np_n = p_0 \rho(1+2\rho+3\rho^2) = 0.74$$

$$\text{and } W_s = \frac{0.74}{6(1-p_3)} = 3.8 \text{ minutes.}$$

7.5.3 $(M/M/c);(GD/\infty/\infty)$

This is a multi-server model with infinite system capacity with arrival rate λ. Arrival and departure, both processes are Poisson processes. The mean service time for a customer is $\frac{1}{\mu}$. The c servers simultaneously serve c customers, provided c or more customers are there in the system. In that case, the service rate will be $c\mu$. If the number of customers n in the system is less than c, then n servers will be busy and remaining c-n servers will be idle. In that case, the service rate will be $n\mu$. For this system, we want to obtain steady-state probabilities and study the operational characteristics of the system.

Under steady-state conditions, the differential-difference equations for this process are

$$-(\lambda_n + \mu_n) p_n + \mu_{n+1} p_{n+1} + \lambda_{n-1} p_{n-1} = 0; \quad n > 0$$

$$-\lambda_0 p_0 + \mu_1 p_1 = 0; \quad n = 0$$

$$\Rightarrow \quad p_1 = \frac{\lambda_0}{\mu_1} p_0$$

and, $p_{n+1} = \dfrac{\lambda_n + \mu_n}{\mu_{n+1}} p_n - \dfrac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1}; \quad n > 0$

Now, $$p_2 = \frac{\lambda_1 + \mu_1}{\mu_2} p_1 - \frac{\lambda_0}{\mu_2} p_0 = \frac{\lambda_1 + \mu_1}{\mu_2} \frac{\lambda_0}{\mu_1} p_0 - \frac{\lambda_0}{\mu_2} p_0$$

$$= \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0$$

Let $$p_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0$$

Then, $$p_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}} \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} - \frac{\lambda_{n-1}}{\mu_{n+1}} \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-2}}{\mu_1 \mu_2 \ldots \mu_{n-1}} p_0$$

$$= \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1} \lambda_n}{\mu_1 \mu_2 \ldots \mu_n \mu_{n+1}} p_0$$

Thus, for $n \geq 1$ $$p_n = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} p_0$$

And $$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 \left(1 + \sum_{n=0}^{\infty} \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} \right) = 1$$

$$\Rightarrow \quad p_0 = \frac{1}{\left(1 + \sum_{n=0}^{\infty} \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} \right)}$$

Now, for this model

$$\lambda_n = \lambda \ \ \forall n$$

$$\mu_n = \begin{cases} n\mu, \ n < c \\ c\mu, \ n > c \end{cases}$$

$$\therefore \ p_n = \begin{cases} \dfrac{\lambda^n}{\mu(2\mu)\ldots(n\mu)} p_0, \ \ n \le c \\ \dfrac{\lambda^n}{\mu(2\mu)\ldots((c-1)\mu)\underbrace{((c\mu)(c\mu)\ldots(c\mu))}_{(n-c)\text{times}}} \ p_0, n > c \end{cases}$$

$$= \begin{cases} \dfrac{\lambda^n}{\mu^n \lfloor n} p_0, \ \ n \le c \\ \dfrac{\lambda^n}{\mu^n c^{n-c} \lfloor c} p_0, \ \ n > c \end{cases}$$

$$= \begin{cases} \dfrac{\rho^n}{\lfloor n} p_0, \ \ n \le c \\ \dfrac{\rho^n}{c^{n-c} \lfloor c} p_0, \ \ n > c \end{cases}$$

and $$p_0 = \left(\sum_{n=0}^{c-1} \frac{\rho^n}{\lfloor n} + \frac{\rho^c}{\lfloor c \left(1 - \dfrac{\rho}{c}\right)} \right)^{-1}$$ where $\dfrac{\rho}{c} < 1 \Rightarrow \dfrac{\lambda}{\mu c} < 1 \Rightarrow \lambda < \mu c$

Performance of the model:

(i) $$L_q = \frac{\rho^{c+1}}{\lfloor (c-1)(c-\rho)^2} p_0 = \frac{c\rho}{(c-\rho)^2} p_c$$

(ii) $$L_s = L_q + \rho$$

(iii) $$W_q = \frac{L_q}{\lambda}$$

(iv) $$W_s = W_q + \frac{1}{\mu}$$

For ρ much smaller than 1, $$p_0 \approx 1-\rho; \ \ L_q = \frac{\rho^{c+1}}{c^2}$$

And for $\dfrac{\rho}{c}$ close to 1, $$p_0 \approx \frac{(c-\rho)\lfloor (c-1)}{c^c}; \ \ L_q \approx \frac{\rho}{c-\rho}$$

Example

(5) A small town is being serviced by two cab companies. Both the companies own two cabs and are known to share the market almost equally. The calls arrive at each company's office at a rate of 10 per hour. The average time per ride is 11.5 minutes. Arrivals of calls follow a Poisson distribution, whereas ride times are exponential. In an effort to consolidate two companies, it was observed that utilization factor for each company (ratio of hourly arriving calls to rides) is

$$100 \times \frac{\lambda}{c\mu} = 100 \times \frac{10}{2 \times \frac{60}{11.5}} = 95.8\%$$

Is consolidation worth?

Sol: Before consolidation:

We have two independent $(M/M/2);(GD/\infty/\infty)$ with λ = 10 calls/hr. and μ = 5.217 rides/hr. The utilization factor for each of the queue is given by

$$100 \times \frac{\lambda}{c\mu} = 100 \times \frac{10}{2 \times \frac{60}{11.5}} = 95.8\% \text{ i.e., it is the same for both the queues.}$$

The expected waiting time in the queue W_q for any of the queue is obtained as follows:

$$\rho = \frac{\lambda}{\mu} = \frac{10}{5.217} \approx 1.917$$

$$p_0 = \left[\left(\frac{1.917}{\lfloor\underline{0}}\right)^0 + \left(\frac{1.917}{\lfloor\underline{1}}\right)^1 + \frac{(1.917)^2}{\lfloor\underline{2}\left(1 - \frac{1.917}{2}\right)}\right]^{-1} \approx 0.0212$$

$$\Rightarrow W_q = \frac{1}{10}\left(\frac{(1.917)^3 \times 0.0212}{\lfloor\underline{1}(2 - 1.917)^2}\right) \approx 2.16 \text{ hrs.}$$

After consolidation:

We have one $(M/M/4);(GD/\infty/\infty)$ with $\lambda = 2 \times 10 = 20$ calls/hr. and μ = 5.217 rides/hr. The utilization factor for the queue is

$$100 \times \frac{\lambda}{c\mu} = 100 \times \frac{20}{4 \times \frac{60}{11.5}} = 95.8\%$$

which is same as the utilization factor for the earlier system. The expected waiting time in the queue W_q for the queue is

$$
\begin{aligned}
\rho &= \frac{\lambda}{\mu} = \frac{20}{5.217} \approx 3.83 \\
p_0 &= \left[\left(\frac{3.83}{\lfloor 0}\right)^0 + \left(\frac{3.83}{\lfloor 1}\right)^1 + \left(\frac{3.83}{\lfloor 2}\right)^2 + \left(\frac{3.83}{\lfloor 3}\right)^3 + \frac{(3.83)^4}{\lfloor 4\left(1-\frac{3.83}{4}\right)}\right]^{-1} \approx 0.0042 \\
\Rightarrow W_q &= \frac{1}{20}\left(\frac{(3.83)^5 \times 0.0042}{\lfloor 3(4-3.83)^2}\right) \approx 1.05\,\text{hrs.}
\end{aligned}
$$

Thus consolidation will reduce the expected waiting time by almost 50% whereas the utilization factor remains unchanged. Hence it is worth to consolidate.

7.5.4 $(M/M/c);(GD/N/\infty)$

This is a multi-server model with finite system capacity N. The maximum queue size is N- c. Arrival and departure, both processes are Poisson processes. For the system

$$
\lambda_n = \begin{cases} \lambda, & 0 \le n \le N \\ 0, & n \ge N \end{cases} \quad \text{and} \quad \mu_n = \begin{cases} n\mu, & 0 \le n \le c \\ c\mu, & c \le n \le N \end{cases}
$$

$$
\Rightarrow p_n = \begin{cases} \dfrac{\rho^n}{\lfloor n} p_0\,, & 0 \le n \le c \\[2ex] \dfrac{\rho^n}{c^{n-c}\lfloor c} p_0\,, & c \le n \le N \end{cases}
$$

For this system, we want to obtain steady-state probabilities and study the operational characteristics of the system.

$$
p_0 = \begin{cases} \left(\displaystyle\sum_{n=0}^{c-1} \frac{\rho^n}{\lfloor n} + \frac{\rho^c\left(1-\left(\frac{\rho}{c}\right)^{N-c+1}\right)}{\lfloor c\left(1-\frac{\rho}{c}\right)}\right)^{-1}; & \dfrac{\rho}{c} \ne 1 \\[4ex] \left(\displaystyle\sum_{n=0}^{c-1} \frac{\rho^n}{\lfloor n} + \frac{\rho^c (N-c+1)}{\lfloor c}\right)^{-1}; & \dfrac{\rho}{c} = 1 \end{cases}
$$

Performance of the model:

(i) $$L_q = \sum_{n=c+1}^{N} (n-c)p_n = \sum_{j=1}^{N-c} jp_{j+c}$$

$$= p_0 \frac{\rho^c}{\underline{|c}} \frac{\rho}{c} \sum_{j=1}^{N-c} j\left(\frac{\rho}{c}\right)^{j-1}$$

$$= p_0 \frac{\rho^{c+1}}{\underline{|(c-1)}(c-\rho)^2}\left[1-\left(\frac{\rho}{c}\right)^{N-c} - (N-c)\left(\frac{\rho}{c}\right)^{N-c}\left(1-\frac{\rho}{c}\right)\right]$$

(ii) Again, $$L_q = \sum_{n=0}^{N} np_n - \sum_{n=0}^{c} np_n - c\left(1-\sum_{n=0}^{c} p_n\right)$$

$$\Rightarrow L_s = L_q + (c-\bar{c}) \quad \text{where, } \bar{c} = \sum_{n=0}^{c} (c-n)\, p_n$$

$$\Rightarrow \lambda_{eff} = L_s - L_q = c - \bar{c} = c - \sum_{n=0}^{c} (c-n)\, p_n = \lambda(1-p_N)$$

(iii) $$W_q = \frac{L_q}{\lambda_{eff}}$$

(iv) $$W_s = W_q + \frac{1}{\mu}$$

Example

(6) If in the earlier example, the owner refuses prospective customers once the waiting list reaches 16 customers, by what extent the expected waiting time be affected?

Sol: In this case,

$$p_0 = \left[\left(\frac{3.83}{\underline{|0}}\right)^0 + \left(\frac{3.83}{\underline{|1}}\right)^1 + \left(\frac{3.83}{\underline{|2}}\right)^2 + \left(\frac{3.83}{\underline{|3}}\right)^3 + \ldots + \frac{(3.83)^4\left(1-\frac{3.83}{4}\right)^{17}}{\underline{|4}\left(1-\frac{3.83}{4}\right)}\right]^{-1} \approx 0.00753$$

$$L_q = \frac{(3.83)^5 \times 0.00753}{\underline{|3}(4-3.83)^2}\left[1-\left(\frac{3.83}{4}\right)^{16} - 16\left(\frac{3.83}{4}\right)^{16}\left(1-\frac{3.83}{4}\right)\right]^{-1} \approx 5.85$$

$$p_{20} = \frac{(3.83)^{20} \times 0.00753}{\underline{|4}(4)^{16}} \approx 0.03433$$

$$\lambda_{eff} = \lambda(1 - p_{20}) = 20(1 - 0.03433) \approx 19.31$$

$$\Rightarrow W_q = \frac{L_q}{\lambda_{eff}} = \frac{5.85}{19.31} \approx 0.303 \text{ hrs.}$$

Thus by loosing 3.4% of potential customers (p_{20}), the waiting line can be cut to one-third.

7.5.5 $(M/M/\infty);(GD/\infty/\infty)$

This is a self-service model with infinite system capacity where each arriving customer serves himself. For the system

$$\begin{aligned}
\lambda_n &= \lambda, \quad \forall\ n \geq 0 \\
\text{and} \quad \mu_n &= n\mu \quad \forall\ n \geq 0 \\
\Rightarrow p_n &= \frac{\lambda^n}{\mu^n \underline{|n}} p_0 = \frac{\rho^n}{\underline{|n}} p_0 \\
\text{where,} \quad p_0 &= \frac{1}{1+\rho+\frac{\rho^2}{\underline{|2}}+\ldots} = e^{-\rho} \\
\Rightarrow p_n &= \frac{\rho^n}{\underline{|n}} e^{-\rho}
\end{aligned}$$

Performance of the model:

(i) $L_q = W_q = 0$

(ii) $L_s = \rho$

(iii) $W_s = \dfrac{1}{\mu}$

Example

(7) Problems arrive at a computing center according to a Poisson distribution at an average rate of five per day. Any person waiting to get his problem solved aids the man whose problem is being solved. If the

time to solve problem with one person has an exponential distribution with mean time of $\frac{1}{3}$ day and if the average solving time is inversely proportional to the number of people working on the problem, approximate the expected time in the center for a person entering the line.

Sol: In this case

$$\lambda = 5 \text{ problems / day}$$
$$\mu = 3 \text{ problems / day}$$
$$\mu_n = n\mu \quad \forall \; n \geq 0$$
$$\Rightarrow p_n = \frac{\rho^n}{\lfloor n} e^{-\rho}$$
$$\Rightarrow L_s = \sum_{n=1}^{\infty} np_n = \sum_{n=1}^{\infty} n\frac{\rho^n}{\lfloor n} e^{-\rho}$$
$$= \rho e^{-\rho} e^{\rho} = \rho = \frac{5}{3} \text{persons}$$

$$\Rightarrow \text{average solving time of a problem} = \frac{1}{3} \times \frac{3}{5} = \frac{1}{5} \text{day/problem.}$$

(Average solving time is inversely proportional to n).

$$\Rightarrow W_s = \frac{1}{5} L_s = \frac{1}{3} \text{day} = 8 \text{hrs.}$$

Problems

1. Find the steady-state probability p_n for the model $(M / M / c); (GD / c / \infty)$. Also find the probability that all the servers are busy.

2. For the system $(M / M / 1); (GD / N / \infty)$, show that

$$\lambda_{eff} = \lambda(1 - p_N) = \mu(L_s - L_q)$$

3. A small barbershop, operated by a single barber has room for at most two customers. Potential customers arrive at a Poisson rate of three per hour. The successive service times are independent exponential variables with mean rate of $\frac{1}{4}$ hour.

   (a) What is the average number of customers in the shop?

   (b) What is the proportion of potential customers that enter the shop?

(c) If the barber could work twice as fast, how much more business would he do?

4. For $(M/M/1);(GD/\infty/\infty)$ system, obtain Var(n) and Var(m), where m is the random variable denoting the queue length. Show that

$$\text{Var}(m) + \text{Var}(s) = \text{Var}(n)$$

where s is the random variable denoting the service time.

5. Consider that two identical $(M/M/1);(GD/\infty/\infty)$ with the same parameters λ and μ are operating. Show that the steady-state distribution for both the systems jointly is

$$p_n = P(N = n) = (n+1)(1-\rho)^2 \rho^n; \quad n \geq 0$$

Now, suppose that it is decided to from a single queue instead of two queues but the two servers will be retained. Show that the steady-state distribution for the combined system is

$$\bar{p}_n = \begin{cases} \dfrac{2(1-\rho)\rho^n}{1+\rho}; & n \geq 1 \\ \dfrac{1-\rho}{1+\rho}; & n = 0 \end{cases}$$

Compare the performance of the two models.

6. On an average 96 patients per 24-hour require the emergency service of a hospital. Also, on an average, a patient requires 10 minutes of active attention. Assume that the facility can handle only one emergency at a time. Suppose that it costs the hospital Rs. 100 per patient treated to obtain an average servicing time of 10 minutes, and if the time is to be decreased, each minute of decrease in this average time would cost Rs. 10 per patient treated. How much would have to be budgeted by the hospital to decrease the average size of the queue from $1\frac{1}{3}$ patients to $\frac{1}{2}$ patients?

7. A maintenance service has Poisson arrival rates, negative exponential service times, and operates on a first come, first served queue discipline. Breakdowns occur at an average rate of three per day with a range of

zero to eight. The maintenance crew can service at an average rate of six machines per day with a range from zero to seven. Find the

(a) utilization factor of the service facility.

(b) mean time in the system.

(c) mean number in the system in breakdown or repair.

(d) mean waiting time in the queue.

(e) probability of finding two machines in the system, and

(f) expected queue size.

8. Assume that the goods train are coming to a yard at the rate of 30 trains per day and suppose that inter-arrival time follows a negative exponential distribution. The service time for each train is exponential with an average of 36 minutes. If the service yard can admit 9 trains at a time, calculate the probability that the yard is empty. Find also the average queue length.

9. Patients arrive at a clinic according to a Poisson distribution at a rate of 30 patients per hour. The waiting room does not accommodate more than 14 patients. Examination time per patients is exponential with mean rate 20 per hour.

(a) Find the effective arrival rate at the clinic.

(b) What is the probability that an arriving person will not wait?

(c) What is the expected waiting time until a patient is discharged from the clinic?

10. A shipping company has a single unloading berth with ships arriving in a Poisson fashion at an average rate of three per day. The unloading time distribution for a ship with n unloading crew is found to be exponential with average unloading time $\frac{n}{2}$ days. The company has a large labour supply without regular working hours, and to avoid long waiting line, the company has a policy of as many unloading crews on a ship as there are ships waiting in line or being unloaded.

(a) Find the average number of unloading crews working at any time?

(b) What is the probability that more than four crews will be needed?

11. A bank has two tellers working on savings accounts. The first teller handles withdrawals only and the second teller handles the deposits only. It has been found that the service time distribution for both deposits and withdrawals is exponential with mean three minutes per customers. Depositors are found to arrive in a Poisson manner through out the day at an average rate of 16 customers per hour whereas depositors are found to arrive in a Poisson manner at an average rate of 14 customers per hour. What would be the effect on the average waiting time for depositors and withdrawers if each teller could handle both deposits and withdrawals? What could be the effect if this could be accomplished by increasing the mean service time to 3.5 minutes?

12. A car servicing station has two bays where service can be offered simultaneously. Due to space limitation, only four cars can be accepted for servicing. The arrival pattern is Poisson with 12 cars per day. The service time in both the bays is exponential with average rate of 8 cars per day per bay. Find the average number of cars in the service station, the average number of cars waiting to be serviced and the average time a car spends in the system.

13. A company has five machines, which suffer from breakdown at an average rate of 2 per hour. There are two servicemen. However, only one can work on a machine at a time. If n machines are out of order where $n > 2$, then $n - 2$ of them will wait until at least one of the servicemen is free. Once a serviceman starts working on a machine the time to complete the repair has an exponential distribution with mean 5 per minutes. Find the distribution of the number of machines out of action at a given time. Find also the time an out of action machine has to spend waiting for the repairs to start.