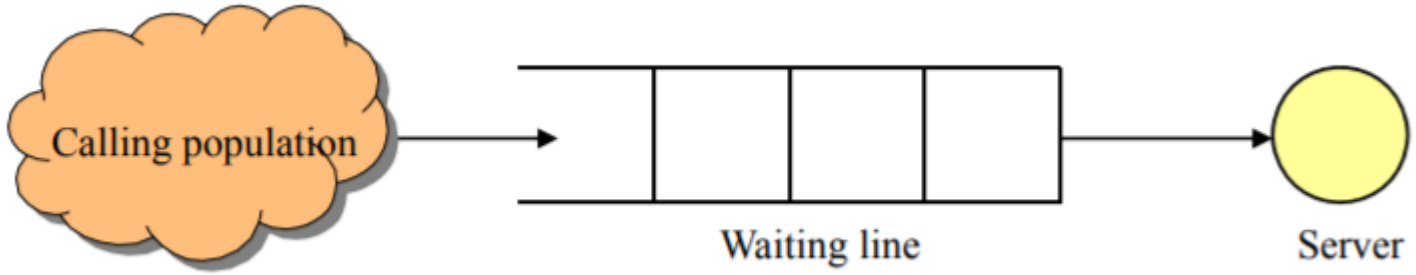
Queuing Theory



Introduction

Queuing Theory is a branch of applied probability which tries to analyse this phenomenon and find methods to minimise the inconvenience.

There are many situations in daily life when a queue is formed. For example, people waiting to buy tickets, patients waiting in Doctor's room, cars waiting for a traffic signal, machines waiting to be repaired, passengers waiting at bus stop etc.

A queue is formed when a customer does not get the service required immediately that is when the current demand for service is more than capacity to provide the service.

Queues may be decreased in size by providing additional service facilities which results in drop of the profit but excessively long queues may result in lost sales and lost customers. Thus the problem is to find the optimum size of queue such that the profit earned in giving the service is maximum and cost involved is minimum.

Examples

System	Customers	Server
Reception desk	People	Receptionist
Hospital	Patients	Nurses
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Computer	Jobs	CPU, disk, CD
Network	Packets	Router

Queueing System

- A Queueing System may be defined as any facility at which a customer arrives, stays there for a certain period of time and departs after being served.
- Structure of Queueing System may be given as follows:

(i) Calling Population/ input source: customers arriving per unit time

(ii) Queuing Process: The manner the customers arrive at the service facility

(iii) Queue Discipline: The method of admitting the customer for service

(iv) Service Process: Service facility given to the customer

(i)Calling Population/ input source

- Size-

  finite (queue for doctor in government hospital)

  Infinite (queue for specialist doctor where patients come through appointment)

- Behaviour-

  Patient (customer arriving at service waits in the queue until served eg. Machines for maintenance)

  Impatient (customer waits in the queue for a certain time and leaves the service system without getting service eg. Customer at crowded grocery shop)

- Pattern of arrival in the system-

  Customers may arrive in batches (eg. Family at restaurant) or individually (eg. Train at a platform) These customers may arrive at a service facility either on scheduled time(by prior information) or on unscheduled time.

(ii) Queuing Process

- It refers to number of queues-single, multiple or priority ques
- In certain cases a service system is unable to accommodate more than the required number of customers at a time.
- No further customers are allowed to enter until more space is made available to accommodate new customers

(iii) Queue Discipline

- It is the order in which customers from a queue are selected for service. For this there are a number of ways. Some of them are:
- Static Queue Discipline-

    FCFS (first come first serve) eg prepaid taxi at airport

    LCFS (last come first serve) eg cargo handling – last item loaded is removed first

- Dynamic Queue Discipline- service at random

    Priority service-payment by cheque or cash

    Emergency service-hospitals

    VIP priority

(iv) Service Process

- Service channels may be in series, in parallel or in mixed manner.

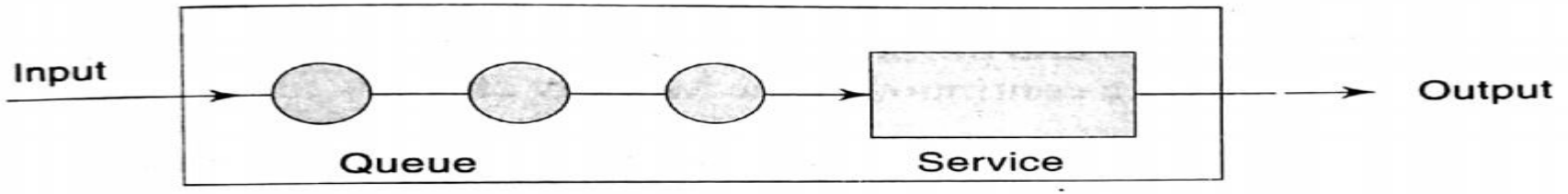


Fig. 9.1 Single server queueing system

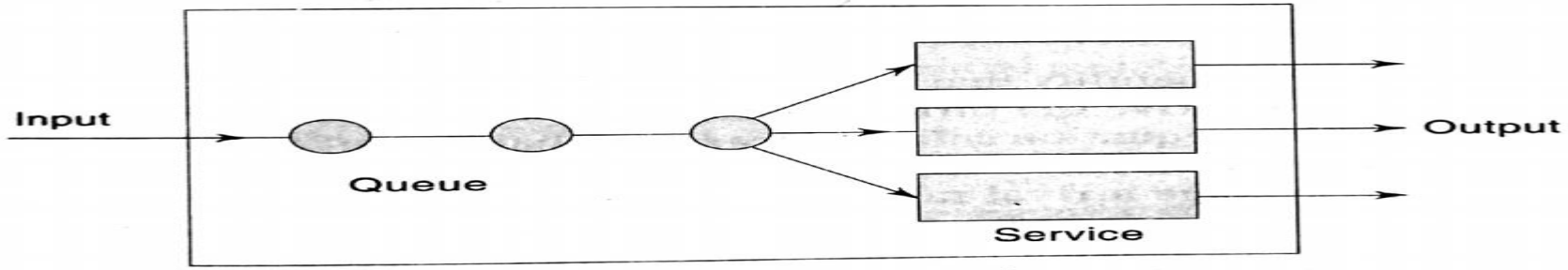


Fig. 9.2 Multiple servers (in parallel) queueing system

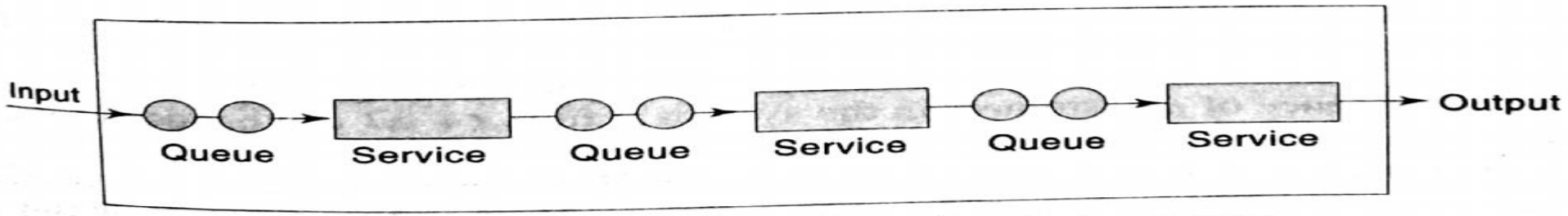


Fig. 9.3 Multiple servers (in series) queueing system

Symbolic Representation for Queuing Model -

- (a / b/ c):(d/e)

Where

a – type of distribution of the number of arrivals per unit time

b – type of distribution of the service time

c – number of services

d – capacity of system that is maximum queue size

e – queue discipline

For example (M/M/1):(∞,FCFS), (M/M/1):(N,FCFS), (M/M/s):(∞,FCFS), (M/M/s):(N,FCFS)

M stands for Markov indicating that the number of arrivals in time t and number of completed services in time t follow Poisson distribution which is continuous time Markov chain.

We will deal with only those Queueing Systems in which

(i) The number of customers arriving per unit time has a Poisson distribution with mean (λ). This means the interval of time between two consecutive customers has an exponential distribution with mean $\frac{1}{\lambda}$.

(ii) The number of customers serviced per unit time has a Poisson distribution with mean (μ). This means the time required to give the full service to the customer is an exponential distribution with mean$(1/\mu)$.

(iii) Service discipline- This means the manner the customers form the queue and they are selected for service. The most common discipline are FCFS or FIFO.

Performance measures of Queuing System

W_q - average time an arriving customer has to wait in a queue before being served

W_s - average time an arriving customer spends in the system including waiting in queue and being served

L_q - average number of customers has to wait in a queue before being served

L_s - average number of customers in the system including waiting in queue and being served

Other Notations

n-number of customers in the system including waiting in queue and being served

P_n- probability of n customers in the system

P_0- probability of no customers in the system (idle time)

$1 - P_0$ - probability that an customer has to wait in the system (system is busy)

λ- average number of arrivals per unit time in the system

μ - average number of customers served per unit time in the system

$$\rho = \frac{average\ service\ completion\ time(1/\mu)}{average\ interarrival\ time(1/\lambda)} = \frac{\lambda}{\mu}$$

$=$ traffic intensity or service utilization factor

s – number of service channels (servers)

N – maximum number of customers allowed in the system

Model-1
Infinite Queuing Model
(M/M/s):(∞,FCFS)

This model is based on certain assumptions about the queuing system

- No limit on queue length (infinite capacity)
- First come first serve discipline
- Average service rate is more than the average arrival rate

  $(\mu > \lambda \Longrightarrow \rho < 1)$

Formulae for (M/M/s):(∞,FCFS) model

- $\rho = \frac{\lambda}{\mu} =$ traffic intensity or service utilization factor
- $P_0 = 1 - \rho$ =probability of no customers in the system (idle time)
- $P_n = \rho^n P_0 = \rho^n (1 - \rho)$
- $P(n \geq k) = \rho^k$ and $P(n > k) = \rho^{k+1}$
- $L_s = \frac{\rho}{1-\rho}$
- $L_q = L_s - \rho = \frac{\rho^2}{1-\rho}$
- $W_s = \frac{L_s}{\lambda}$
- $W_q = \frac{L_q}{\lambda}$
- $P(W_s > t) = e^{-\mu(1-\rho)t}$
- $P(W_q > t) = \rho\, e^{-\mu(1-\rho)t}$
- We can derive that $W_s = W_q + \frac{1}{\mu}$ and $L_s = L_q + \frac{\lambda}{\mu}$

Ex 1 Find the average number of customers in the system and in the queue if the system is (M/M/1/∞) and $\mu = 15, \lambda = 10$

$\rho = \frac{\lambda}{\mu} = \frac{10}{15} = \frac{2}{3} =$ service utilization factor

L_q- average number of customers has to wait in a queue before being served $= \frac{\rho^2}{1-\rho}$

$= \frac{4/9}{1-2/3} = 1.33 \approx 1$

L_s- average number of customers in the system including waiting in queue and being served $= \frac{\rho}{1-\rho}$

$= \frac{2/3}{1-2/3} = 2$

Ex 2 Find service utilization factor, the average waiting time per customer in the queue and in the system for (M/M/1/∞) model if $\mu = 15, \lambda = 9$ per hour. Also find the probability that (i)a customer has to wait in the system (ii)there are more than 8 customers in the system.

$\rho = \frac{\lambda}{\mu} = \frac{9}{15} = \frac{3}{5} =$ service utilization factor

W_q= average time an arriving customer has to wait in a queue before being served

$= \frac{L_q}{\lambda} = \frac{1}{\lambda}\frac{\rho^2}{(1-\rho)} = \frac{9/25}{9(1-3/5)} = 0.1\ hrs.$

W_s- average time an arriving customer spends in the system including waiting in queue and being served

$= \frac{L_s}{\lambda} = \frac{1}{\lambda}\frac{\rho}{(1-\rho)} = \frac{3/5}{9(1-3/5)} = \frac{1}{6}\ hrs.$

$(i) P_0$- probability of no customers in the system (idle time) $= 1 - \rho$

probability that a customer has to wait in the system $= 1 - P_0 = \rho = 3/5$

(ii) $P(n > k) = \rho^{k+1}$

$P(n > 8) = \rho^9 = (\frac{3}{5})^9 = .01008$

Ex 3 Find the traffic intensity of the system (M/M/1/∞) model if $\mu = 11\ per\ hour, \lambda = 8$ per hour. Also find the probability that a customer has to wait for more than 20 minutes to be out of the service station.

$\rho =$ traffic intensity or service utilization factor $= \frac{\lambda}{\mu} = 8/11$

$t = 20\ min = \frac{20}{60}\ hrs$

a customer has to wait for more than 20 minutes to be out of the service station

$= P(W_s > t) = e^{-\mu(1-\rho)t}$

$= P(W_s > 1/3) = e^{-11\left(1-\frac{8}{11}\right)\frac{1}{3}} = 0.3679$

Ex 4 A customer arrives at a clinic according to a poisson process with a mean interval of 25 minutes. The doctor needs on an average 20 minutes for a patient to examine. Find

(i) the expected number of patients in the clinic and in the queue

(ii) percentage of patients who are not required to wait

(iii) on an average how much time is spent by a patient in the clinic

(iv) the doctor will appoint another doctor if the patient's time in the clinic exceeds 2 hours. How much must the rate of arrivals increase so that another doctor is appointed?

(v) average time a patient has to be in queue before the doctor examines him.

(vi) probability that the total waiting time of patient in the system is greater than 1 hour.

(vii) percentage of patients who have to wait before they are called by the doctor for examination

(viii) probability that there are more than 4 patients in the queue

(ix) it is desired that fewer than 5 patients are in the queue for 99% of the time. How fast the service rate should be?

Given

λ- average number of arrivals per unit time in the system=1/25 patients per minute

μ - average number of customers served per unit time in the system=1/20 patients per minute

$\rho =$ traffic intensity or service utilization factor $= \frac{\lambda}{\mu}$=4/5

(i) the expected number of patients in the clinic and in the queue

$L_s = \frac{\rho}{1-\rho} = 4$ $\quad L_q = \frac{\rho^2}{1-\rho} == 3.2 \approx 3$

(ii) percentage of patients who are not required to wait

prob(no patient in the system) $= P_0 = 1 - \rho = \frac{1}{5} = 0.2$

$\Longrightarrow$ percentage of patients who are not required to wait= 0.2*100= 20%

(iii) on an average how much time is spent by a patient in the clinic

$W_s = \frac{1}{\lambda}\frac{\rho}{(1-\rho)}$ =100 min

(iv) the doctor will appoint another doctor if the patient's time in the clinic exceeds 2 hours. How much must the rate of arrivals increase so that another doctor is appointed?

New doctor is appointed if W_s>2 hrs=120 min $\Longrightarrow \frac{1}{\lambda}\frac{\rho}{(1-\rho)}$>120 $\Longrightarrow \frac{1}{\mu-\lambda}$>120

$\Longrightarrow \frac{1}{1/20-\lambda}$>120 $\Longrightarrow \lambda$>1/24 $\Longrightarrow$increase in arrival rate=1/24 – 1/25 =1/600 per min

λ- average number of arrivals per unit time in the system=1/25 patients per minute
μ - average number of customers served per unit time in the system=1/20 patients per minute

$\rho =$ traffic intensity or service utilization factor $= \frac{\lambda}{\mu}$=4/5

(v) average time a patient has to be in queue before the doctor examines him. $W_q = \frac{1}{\lambda}\frac{\rho^2}{(1-\rho)}$ =80 min

(vi) probability that the total waiting time of patient in the system is greater than 1 hour $= P(W_s > t) = e^{-\mu(1-\rho)t}$, t=1 hr = 60 min
$P(W_s > 60) = e^{-\mu(1-\rho)t}$= 0.5488

(vii) percentage of patients who have to wait before they are called by the doctor for examination

=prob(system is busy) $= 1 - P_0 = \rho = 0.8 = 80\%$

(viii) probability that there are more than 4 patients in the queue
$P(n > k) = \rho^{k+1} \Longrightarrow P(n > 4)$ =0.3277

λ- average number of arrivals per unit time in the system=1/25 patients per minute
μ - average number of customers served per unit time in the system=1/20 patients per minute
$\rho =$ traffic intensity or service utilization factor $= \frac{\lambda}{\mu}$=4/5

(ix) it is desired that fewer than 5 patients are in the queue for 99% of the time. How fast the service rate should be?

$$P(n < 5) \geq 99\% \Longrightarrow P(n \leq 4) \geq 99\%$$

$$P(n > k) = \rho^{k+1} \Longrightarrow P(n \leq k) = 1 - \rho^{k+1} \Longrightarrow 1 - \rho^{5} \geq 99\%$$

$$\Longrightarrow \left(\frac{\lambda}{\mu}\right)^{5} \geq 0.01 \Longrightarrow \left(\frac{1}{25\mu}\right)^{5} \geq \frac{1}{100} \Longrightarrow \mu \geq 0.1105 \text{ patients per min}$$

Ex 5-Trucks arrival at a factory for collecting finished goods that are supposed to be transported to distant markets. As and when they come they are required to join a waiting line and are served on first come, first served basis Trucks arrive at the rate of 10 per hour whereas the loading rate is 15 per hour It is also given that arrivals are Poisson and loading is exponentially distributed. Transporters have complained that their trucks have to wait for nearly 12 mins at the plant. Examine whether the complaint is justified Also determine the probability that the loaders are idle in the above problem.

Ex 6-At what average rate must a clerk at a super market work in order to ensure a probability of 0.90 so that the customer will not have to spend more than 12 min.? It is assumed that there is only one counter at which customers arrive in a Poisson fashion at an average rate of 15 per hour The length of service by the clerk has an exponential distribution

Model II
finite Queuing Model
(M/M/1):(N,FCFS)

This model is also based on the same assumptions of model 1except a limit on the capacity of the system to accommodate only N customers.

In this model, P_0- probability of no customers in the system is given by

$$P_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}} & if\ \rho \neq 1 \\ \dfrac{1}{N+1} & if\ \rho = 1 \end{cases}$$

Formulae for (M/M/s):(N,FCFS) model

- $\rho = \frac{\lambda}{\mu} =$ traffic intensity or service utilization factor
- $P_0 = \begin{cases} \frac{1-\rho}{1-\rho^{N+1}} & if \rho \neq 1 \\ \frac{1}{N+1} & if \rho = 1 \end{cases}$ =probability of no customers in the system (idle time)
- $P_n = \rho^n P_0$
- $L_s = \begin{cases} \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}} & if \rho \neq 1 \\ \frac{N}{2} & if \rho = 1 \end{cases}$
- $L_q = L_s - \rho$
- $W_s = \frac{L_s}{\lambda(1-P_N)}$
- $W_q = \frac{L_q}{\lambda(1-P_N)}$
- λ_{eff}=effective arrival rate= $\lambda/(1 - P_N)$
- $\rho_{eff} = \frac{\lambda_{eff}}{\mu}$

Ex 1- Consider a single server queuing system with Poisson input and exponential service times. Suppose the mean arrival rate is 3 calling units per hour, the exponential service time is 0.25 hrs and a maximum permissible calling units I the system is 2.Derive a steady state probability distribution of the number of calling units in the system and then calculate the expected number of calling units in the system

- Transient state and steady state

At the beginning of service operations a queuing system is influenced by initial conditions such as no of customers waiting for service and the time when the servers are busy etc. This initial state is termed as Transient state. However after a certain period of time the system becomes independent of initial conditions and enters into steady state.

Given

λ- average number of arrivals per unit time in the system= 3 units per hour
μ - average number of customers served per unit time in the system=1/0.25=4 units per hour
$\rho =$ traffic intensity or service utilization factor $= \frac{\lambda}{\mu} = \frac{3}{4} \neq 1,$ N=2.

steady state probability distribution of the number of calling units in the system

$= P_n = \rho^n P_0 = \frac{\rho^n(1-\rho)}{1-\rho^{N+1}} = \frac{(0.75)^n(1-0.75)}{1-(0.75)^3}$, n = 0, 1, 2

$P_0 = 0.431, P_1 = 0.3243, P_2 = 0.2432$

expected number of calling units in the system $= L_s = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}} = 0.81 \approx 1$

Ex 2 The local one person barber shop can accommodate maximum of 5 people at a time (4 waiting and 1 getting haircut). Customers arrive according to a PD with mean 5 per hour. The barber cuts hair according to a ED at an average rate of 4 per hour.
(i) What percentage of time is the barber idle?
(ii) What fraction of potential of customers are turned away?
(iii) What is the expected number of customers waiting for a haircut?
(iv) How much time can a customer expect to spend in the barber shop?

SOLUTION

$\lambda = 5$ customers per hour, μ = 4 customers per hour, $\rho = \frac{\lambda}{\mu} = \frac{5}{4} \neq 1$, N=5

(i) Idle time of barber = $P_0 = \frac{1-\rho}{1-\rho^{N+1}} = 0.088 \approx 8.8\%$

(ii) P(customers are turned away)= potential customer loss= $P_n = \rho^n P_0$ for n=N =0.2711

(iii) expected number of customers waiting for a haircut $= L_q = L_s - \rho = \left\{\frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}\right\} - \rho = 1.88$

(iv) time a customer expect to spend in the barber shop $= W_s = \frac{L_s}{\lambda(1-P_N)}$=0.9466

Ex 3- At a railway station, only one train is handled at a time Railway yard is sufficient only for 2 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 railway station can handle them on an average of 6 per hour. Assuming Poisson input and exponential service distribution, find the probabilities for the number of trains in the system. Also find the average waiting time of the new train coming to the yard. If the handling rate is doubled, how will the above results get modified?

(i) $\lambda = 6$ trains per hour, μ = 6 trains per hour, $\rho = \frac{\lambda}{\mu} = \frac{6}{6} = 1$, N=2+1=3

- $P_0 = \frac{1}{N+1} = \frac{1}{4}$, $P_n = \rho^n P_0, \text{n} \leq N = \frac{1}{4} \; \forall \; n$
- $W_s = \frac{L_s}{\lambda(1-P_N)} = \frac{N/2}{\lambda(1-P_N)} = 0.33hrs$

(ii) If the handling rate is doubled

- $\mu = 6 \times 2$ = 12 trains per hour $\therefore \; \rho = \frac{\lambda}{\mu} = \frac{6}{12} = \frac{1}{2} \neq 1$
- $P_0 = \frac{1-\rho}{1-\rho^{N+1}}$=8/15=0.53, $P_n = \rho^n P_0, \text{n} \leq N \therefore P_1 = 0.27, P_2 = 0.13, P_3 = 0.07$
- $W_s = \frac{L_s}{\lambda(1-P_N)} = \frac{\frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}}{\lambda(1-P_N)}$=0.13 hrs or 7.9 min

Ex 4 Patients arrive at a clinic according to a PD at a rate of 30 patients per hour. The waiting room does not accommodate mora than 14 patients. Examination time per patient is exponential with mean rate of 20 per hour. Find (i) effective arrival rate at a clinic and effective traffic intensity (ii) probability that an arriving patient will not wait (iii) expected waiting time until a patient is discharged from the clinic.

Ex 5 If in a period of 2 hours, in a day (08:00 to 10:00 am), trains arrive at the yard every 20 min. but the service time continues to remain 36 min. then calculate, for this period

(a) The probability that the yard is empty, and

(b) The average number of trains in the system, on the assumption that the line capacity of the yard is only limited to 4 trains.