UNIT 6 QUEUEING THEORY

Structure

6.1 INTRODUCTION

Queueing is a common phenomenon in everyday life. We wait in queues in post offices, banks, restaurants, railways and airline reservation counters. Vehicles wait at traffic lights and aeroplanes circle around airports while waiting to land. You can think of many more examples. In all such cases, there are customers who require some sort of services after waiting in a queue for some time. The customer may be a person, machine, vehicle or anything else which requires service. In fact, waiting for service is an integral part of our daily life and that too at considerable cost most of the times.

We would like to find ways of reducing the time spent in waiting by the customer and at the same time optimising the cost to the service provider. This is where the **queueing theory**, also known as the waiting line theory helps us. It was developed in 1909 when A. K. Erlang made an effort to analyse telephone traffic congestion. The purpose of queueing analysis is to provide information to determine an acceptable level of service and service capacity since providing too much service capacity is costly (owing to idle employees or equipment). However, providing too little service capacity is also costly (owing to waiting members in the queue). For example, when a hospital consistently has a long queue in its emergency room, a large number of waiting patients may aggravate the injury or illness.

In this unit, we first discuss the basic concepts of queueing theory in Sec. 6.2, which would help you understand the techniques of queueing models. In Secs. 6.3 and 6.4, we explain the fundamental structure and operating characteristics of a queueing system. We describe the M/M/1 queueing model and its applications in Sec. 6.5.

In the next unit, we shall discuss the sequencing problems and explain the n-jobs, 2-machine and 2-jobs, m-machines, sequencing problems.

Objectives

After studying this unit, you should be able to:

- explain the Poisson process and the birth and death process in queueing theory;

- explain the fundamental structure of a queueing system;
- describe the operating characteristics of a queueing system;
- explain a single server queueing model with Poisson input and exponential service time; and
- solve problems based on the M/M/1 queueing model.

6.2 BASIC CONCEPTS OF QUEUEING THEORY

For understanding queueing systems, you have to be familiar with the probability theory. The concept of random variable and its probability distribution such as Poisson distribution, Exponential distribution and Poisson process play a significant role in describing queueing systems. You have studied probability theory, random variable, Poisson and Exponential distributions in the course MST-003 on Probability Theory. However, the Poisson process is a new concept for you and we explain it, in brief. Let us quickly recall a few basic definitions.

Consider an experiment whose outcome is not uniquely determined. In such a situation an observed outcome of the experiment is one from a set of possible outcomes. An outcome of an experiment is called a **sample point** and the set of all possible outcomes of a random experiment is called a **sample space**. Subsets of the sample space are called **events. A random variable is a function that associates a point of the sample space with a real number. A random process** or **stochastic process** is a family (or collection) of random variables. For example, if a die with six faces numbered 1, 2, …, 6 is thrown, the set of all possible outcomes is S={1, 2, 3, 4, 5, 6}. The set S is a sample space. A function X that assigns to an outcome or sample point, the number written on it, is the random variable. The event that an even number is observed corresponds to the set {2, 4, 6} which is a subset of S.

Let us consider another simple experiment such as throwing a fair die.

1. Suppose X_n is the outcome of the n^{th} throw, $n \geq 1$. Then $\{ X_n, n \geq 1\}$ is a family of random variables, such that for a distinct value of n, one gets a distinct random variable X_n. The sequence $\{ X_n, n \geq 1\}$ constitutes a random (stochastic) process known as the **Bernoulli process.**
2. Suppose X_n is the number of sixes in the first n throws. For a distinct value of n = 1, 2, …, we get a distinct **Binomial** random variable. The sequence X_n:$\{ X_n, n \geq 1\}$, which gives a family of random variables, is a random or stochastic process.
3. Suppose a telephone call is received at a switchboard. Let X_t be the random variable, which represents the number of incoming calls in an interval (0, t). Then X_t is a random variable and the family $\{ X_t, t\, \varepsilon T\}$ constitutes a stochastic process, where T is the interval $0 \leq t \leq \infty$.

In Examples 1 and 2 above we saw that the subscript n of X_n was restricted to non-negative integers n = 0, 1, 2, …. In these examples we observe the outcome of a random variable at distinct time points n = 0,1, 2, …. Here, the word ‘time’ is used in a wider sense. You can visualise an infinite family of random variables $\{ X_t, t\, \varepsilon T\}$ such that the state of the system is characterised at every instant over a finite or infinite interval. The process (or collection) is then defined for a continuous range of time and we say that we have a family of random variables. On the other hand, the family $\{ X_n, n =0, 1, 2, …\}$ is called a discrete parameter (or discrete time) stochastic process. The value of X_n for a specific realisation of the process is called the state of the process at the n^{th} step. If the random variables X_n are discrete random variables, i.e., if they take only integer values, it is called a discrete state process. Let us now explain the Poisson process used in the M/M/1 model being discussed in Sec. 6.5.

6.2.1 Poisson Process

Suppose X(t) represents the maximum temperature at a particular place. Here we deal with discrete state, continuous time stochastic processes, i.e., X(t) is a discrete random variable. The Poisson process is one of the representatives of this type of stochastic processes. It plays an important role in the study of a large number of phenomena. The Poisson process may be described as follows: Let E be a random event such as (i) incoming telephone calls at a switchboard, (ii) arrival of patients for treatment at a clinic, (iii) occurrence of accidents at a certain place.

We consider the total number N(t) of occurrences of an event E in an interval of time 't'. Suppose $P_n(t)$ is the probability that the random variable N(t) assumes the value n, i.e.,

$$P_n(t) = P[N(t) = n] \qquad \dots (1)$$

for n = 0, 1, 2, 3, …. We have

$$\sum_{n=0}^{\infty} P_n(t) = 1, \text{ for each fixed t.} \qquad \dots(2)$$

We can thus say from equation (2) that $P_n(t)$ is a probability mass function of the random variable N(t) and the family of random variables $\{N_t, t > 0\}$ is a stochastic process. From our earlier discussion, you may understand that this family is a continuous parameter (in this case, time) stochastic process with a discrete state space. This is called a Poisson process. Under certain conditions, N(t) follows a Poisson distribution with mean λt (λ being constant). This is true for most practical situations.

Assumptions in Poisson Process

i) The process has independent increments. Future changes in N(t) are independent of past changes in it, i.e., the number of customers which arrive in disjoint time intervals are statistically independent.

ii) The probability of more than one occurrence of event E between time t and $t+\Delta t$ is $o(\Delta(t))$, i.e., the probability of two or more arrivals of customers during the small interval of time Δt is negligible. Thus,

$$P_0(\Delta t) + P_1(\Delta t) + o(\Delta t) = 1 \qquad \dots (3)$$

iii) The probability that event E occurs between time t and Δt is equal to $\lambda\Delta t + o(\Delta(t))$. Thus,

$$P_1(\Delta t) = \lambda\Delta t + o(\Delta t),$$

i.e., $P_1(\Delta t)$ is approximately proportional to the length of the interval where λ is the expected average number of arrivals of customers per unit time and $\lambda > 0$. Here λ is constant and Δt is an incremental element.

Under the assumptions stated above, N(t) follows a Poisson distribution with mean λt, i.e., $P_n(t)$ is given by

$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \qquad n = 0, 1, 2, 3, \dots . \qquad \dots(4)$$

To formulate a queueing model, we have to specify the assumed form of the probability distributions of both inter-arrival times and service times. You have learnt that inter-arrival times follow the Poisson distribution. Similarly, most of the times, the service times in a queueing system follow the exponential distribution. Hence, the exponential distribution is the most important distribution in queueing theory, which we now discuss.

Suppose a random variable T represents either inter-arrival or service times. T is said to have an exponential distribution with parameter α if its probability density function is given by

$$f(t) = \begin{cases} \alpha e^{-\alpha t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \qquad \ldots(5)$$

Let N(t) denote the number of occurrences of an event E (say arrival or number of telephone calls) in duration Δt. There is an associated variable: the interval T between two successive occurrences of event E. The interval T between two successive occurrences of a Poisson process $\{N_t, t \geq 0\}$, having parameter λ, has an exponential distribution with parameter λ. If the intervals between successive occurrences of an event E are independently distributed with a common exponential distribution with parameter λ, the events follow a Poisson process with mean λt, i.e., the probability distribution of the number of times these kinds of events occur over a specified duration of time has a Piosson distribution with parameter λt.

The basic methodology discussed here was developed initially by the Russian mathematician, A. A. Markov, around the beginning of the 20th century. The Markov Process forms a sub-class of the set of all random processes. It is a sub-class with enough simplifying assumptions to make them easy to handle. Mathematically, we define the Markov Process as follows:

Let $\{ X(t), t \geq 0\}$ be a continuous time stochastic process taking on values in the set of non-negative integers. If, for all $t_n, t_{n-1}, \ldots, t_0$, satisfying $t_n > t_{n-1} > \ldots > t_0$, and non-negative integer j,

$P[X(t_n) = j_n / X(t_{n-1}) = j_{n-1}, \ldots, X(t_0) = j_0] = P[X(t_n) = j_n / X(t_{n-1}) = j_{n-1}]$,

the process has the Markov property and is called a **continuous time Markov process**. Now we confine to Markov processes in which time is measured discretely. The process is described by a sequence of random variables, $\{X_1, X_2, X_3, \ldots\}$. Here the probability distribution of each X_i must be specified. But it is not enough to specify a stochastic process, since X_is are not independent. At this point we need Markov's simplifying assumption: A **Markov chain** is a discrete time stochastic process in which each random variable, X_i depends on the previous one, X_{i-1} and affects only the subsequent one, X_{i+1}. The term **chain** suggests the linking of the random variable to neighbours in the sequence.

6.2.2 Birth and Death Process

The birth and death process is a special case of the general continuous time Markov process. A birth and death process is characterised as a Markov process in which all transitions are to be to the next state, immediately above (a "birth") or immediately below (a "death") in the natural integer ordering states, i.e., a birth and death process does not 'jump' states. The Poisson process is a special case of the birth-death process. It might be more appropriate to call it a pure birth process since the death rates (μ_i) are all zero. The birth rates, λ_i, are all equal to a constant value λ. The Poisson process is often used to model the kind of situation in which a count is made of the number of events occurring in a given time. For example, it will be used in the discussion of queueing models to represent the arrivals of customers to a service facility. The state at time t would correspond to the number of arrivals by time t. P_n (t) denotes the probability that there are n customers in the system at any time t, both waiting and being served and P_n stands for time independent probability that there are n customers in the system, both waiting and being served. Arrivals can be considered as births. If the system is in state E_n and an arrival occurs, the state is changed to E_{n+1}. Similarly, a departure can be looked upon as death. A departure occurring while the system is in state E_n changes the system to the state E_{n-1}. This type of process is generally referred to as a birth-death process.

Assumptions in the Birth-Death Process

The assumptions of the birth-death process are as follows:

i) If the system is in state E_n, the current probability distribution of the time t until the next arrival (birth) is exponential with parameter λ_n, where n = 0, 1, 2, … .

ii) If the system is in state E_n, the current probability distribution of the time t until the next departure or service completion (death) is exponential with parameter μ_n, where n = 0, 1, 2, … .

iii) Only one birth or death can occur in a small interval of time, i.e., Δt.

Having discussed the basic concepts and techniques for studying queueing systems, we now explain the fundamental structure of a queueing system.

6.3 FUNDAMENTAL STRUCTURE OF A QUEUEING SYSTEM

Queueing theory is concerned with the mathematical study of queues or waiting lines (seen in banks, post offices, hospitals, airports, etc.). The formation of waiting lines usually occurs whenever the current demand for a service exceeds the current capacity to provide that service.

In many cases, the customer's arrival and his or her service time are not known in advance or cannot be predicted accurately. Otherwise, the operation of the service facility could be scheduled in a manner that would eliminate waiting completely. Both arrival and departure phenomena are random. This necessitates mathematical modelling or queueing systems/ models to alleviate waiting. It involves reducing excessive costs that result from creating excess service capacity and at the same time ensuring that the system has enough service capacity to avoid long waiting lines. There has to be a balance between service capacity and waiting time. Therefore, an industry or an agency would like to provide such services and also maintain balance between the cost of service and the cost associated with waiting for the service.

A simple queueing system has the following fundamental structure:

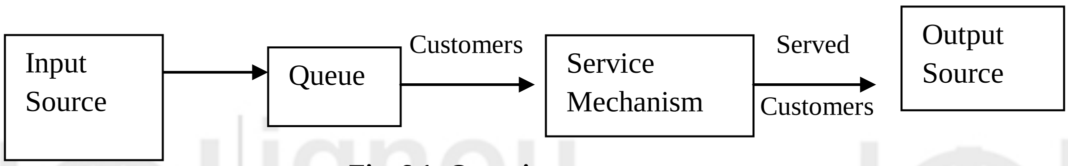

Fig. 6.1: Queueing system.

A queueing system is described by the following elements:

1. Input or arrival process of customers;
2. Service mechanism; and
3. Queue discipline.

We now give a brief description of each of these components.

1. Input or Arrival Process of Customers

This is the element concerned with the pattern in which the customers arrive and join the system. An input process is characterised by its size, the arrival time distribution, and the attitude of the customers. We describe these, in brief.

i) **Size:** It may be finite or infinite according as the arrival rate is affected or not affected by the number of customers in the service system.

ii) The **arrival time distribution:** In most cases, the arrivals occur in accordance with a Poisson distribution.

iii) **Customer behaviour:**

- The customer may stay in the system until served. Such a customer is known as **Patient Customer.**
- The customer may wait for a certain time and leave the system if service is not commenced by that time. This kind of departure from the queue without receiving the service is known as **impatient or reneging behaviour.**
- When a customer arrives at a queueing system and perceives that the number of customers already in the queue is too large, s/he does not join the queue. This behaviour of the customers is known as **balking behaviour**.
- When a customer moves from one waiting line to another because he/she thinks that his/her queue is moving slower, the behaviour of the customer is known as **queue jockeying**.

Generally, it is assumed that the customers arrive into the system one at a time. But, sometimes, customers may arrive in groups and such arrival is called **bulk arrival**.

2. Service Mechanism

Service time distributions are generally exponential distributions. A service facility can be any one of the following types:

i) Single channel facility, i.e., one queue-one service station facility;

ii) One queue-several station facilities (e.g., booking at a service station that has several service mechanisms);

iii) Several queues-one service station (e.g., railway station ticket counters);

iv) Multi-channel facility (e.g., several counters); and

v) Multistage Channel facility (e.g., various medical tests of a patient).

3. Queue Discipline

Queue discipline refers to the order in which the service station selects the next customer from the waiting line to be served. It may be any one of the following:

i) First In, First Out (FIFO) or in other words First Come First Served (FCFS);

ii) Last In, First Out (LIFO); and

iii) Service In Random Order (SIRO).

In this unit, we consider the FCFS queue discipline.

So far you have learnt the fundamental structure of a queueing system. We now describe its operating characteristics.

6.4 OPERATING CHARACTERISTICS OF A QUEUEING SYSTEM

We have set up a set of equations to study the queueing system. We have to solve these equations to determine the operating characteristics. There are two types of solutions of these equations: transient and steady state. The time dependent solutions are known as **transient solutions**. For a complete description of a queueing process, we need transient solutions. However, it is difficult to obtain such solutions. Moreover, in many practical situations, we usually need to know the system's behaviour in steady state, i.e., the behaviour of the system when it reaches its equilibrium state after being in operation for a long time. The steady state solution is independent of time and represents the probability of the system being in a particular state in the long run.

A birth and death process gradually attains steady state after a sufficiently large time has elapsed, provided that the parameters of the process permit reaching the steady state. For example, queues with arrival rate λ higher than the departure rate μ will not reach a steady state as the queue size will increase with time. The ultimate objective of analysing queueing situations is to develop measures of performance for evaluating the real system. Here we shall be analysing the system under the steady state condition since the time dependent analysis is beyond the scope of this course. Let us now list the operating characteristics of a queueing system.

6.4.1 Operating Characteristics

The operating characteristics of a queueing systems are determined by two statistical properties, namely, the probability distribution of inter-arrival times and the probability distribution of service times. In the analysis of a queueing system, we would like to compute various operating characteristics, i.e., the measures of performance. These include:

1. The average number of customers in the queueing system (L_s), i.e., those waiting to be served and those being served.
2. The average time each customer spends in the queueing system from entry into the queue to completion of the service (W_s), i.e., the time spent waiting in the queue and during the service.
3. The average number of customers in the queue waiting to get service (Lq), i.e., this excludes customers undergoing service.
4. The average time each customer spends waiting in the queue to get service (W_q), i.e., this excludes time spent during the service.
5. The relative frequency with which the service system is idle is called **service idle time**. Though the idle time is directly related to the cost, its reduction may have adverse effects on other characteristics.

These equations can be used to study the time dependent behaviour as well . We now state the equations for each of the operating characteristics listed above under steady state condition.

By definition of expectation, we have

$$L_s = \sum_{n=1}^{\infty} nP_n \qquad \dots (6)$$

$$L_q = \sum_{n=c+1}^{\infty} (n-c)P_n \qquad \dots (7)$$

Suppose that arrivals follow the Poisson distribution, and the inter-arrival times follow exponential distributions [Chapter 4 of Stochastic Process by J. Medhi]. The exponential distribution has memoryless property, which means that the future is independent of the past. It depends only on the present and hence the Markov property [Unit 15, Block 4 of MST-003: Probability Theory and Chapter 3 of Stochastic Process by J. Medhi] is satisfied by the inter-arrival times. Therefore, symbol '**M'** has been used in place of '**a**' to represent the Markov property. Also, if service times follow the exponential distribution, then the service time distribution too has the Markov property and hence symbol '**M**' can be used in place of '**b**'.

where there are c parallel servers so that c customers can be served simultaneously. With λ arrival rates for those who join the system, we have

$$L_s = \lambda W_s \quad \text{and} \quad L_q = \lambda W_q \qquad \dots (8)$$

If μ is the service rate, the expected service time is $\frac{1}{\mu}$ and we have

$$W_s = W_q + \frac{1}{\mu} \qquad \dots (9)$$

Multiplying both sides of equation (9) with λ, we get

$$L_s = L_q + \frac{\lambda}{\mu} \qquad \dots (10)$$

6.4.2 Classification of Queueing Systems

A queueing system is usually described by five symbols and denoted as **a / b / c: d / e** or separated by slashes as **a / b / c / d / e**. The first symbol '**a**' describes the arrival process. The second symbol '**b**' describes the service time distribution. The third symbol '**c**' stands for the number of servers. The symbols '**d**' and '**e**' stand for system capacity and queue discipline, respectively. If the system has FCFS queue discipline, the queueing system may be described as **a / b / c / d**, i.e., the fifth symbol may be omitted in this case. If a system has infinite capacity with FCFS queue discipline, the queueing system may be described as **a/b/c :** ∞ **/ FCFS** or simply **a/b/c.**

Note: If arrivals follow the Poisson distribution, the symbol **M** is used in place of '**a**'. If departures are exponential, the symbol **M** is used in place of '**b**'. The explanation for using the symbol M is given in the margin for the sake of interest.

So far you have learnt about the operating characteristics of a queueing system. We now discuss the M/M/1 queueing model and determine its operating characteristics.

6.5 M/M/1 QUEUEING MODEL

As the symbol indicates, the M/M/1 queueing model deals with a queueing system having a single service channel with Poisson input, exponential distribution for services and there is no limit on the system capacity while the customers are served on a "First Come First Served" basis.

For this model, arrivals follow Poisson distribution and hence inter-arrival times have an exponential distribution. Let λ denote the average arrival rate. Therefore, $1/\lambda$ is the mean arrival time. For example, if customers arrive at a rate of 15 per hour, it means that on an average, 1 customer arrives in every 4 minutes, i.e., the mean arrival time is $\frac{1}{15}$ hr or 4 minutes (the reciprocal of the arrival rate). The service times for this model follow an exponential distribution. Let μ denote the average service rate. Hence, 1/ μ is the mean service time.

Then the ratio

$$\rho = \frac{\lambda}{\mu} \qquad \ldots (11)$$

is called the **traffic intensity** or the **utilisation factor**. It is a measure of the degree to which the capacity of the service station is utilised. For example, if customers arrive at a rate of 15 per minute and the service rate is 20 per minute, the utilisation of the service facility is $\frac{15}{20} = 75\%$, i.e., the service facility is kept busy 75% of the time and remains idle 25% of the time.

6.5.1 Arrival-Departure Equations for M/M/1 Queueing Model

In this section, we obtain the arrival-departure equations for M/M/1 queueing model. But before doing so, it is necessary to state the postulates of the Poisson process, which are as follows:

i) The number of occurrences of an event in an interval of duration t are independent of the number of occurrences in an interval prior to the interval (0, t), i.e., future changes are independent of past changes.

ii) The probability of n customers in 't' depends only on the duration 't' of the interval and is independent of where this interval is situated. Let $P_n(t)$ be the probability of n customers in an interval t. Then $P_n(t)$ also gives the probability of a number of occurrences of an event in the interval (a, t + a), which is of duration t for every 'a'. For example, the interval may be (5, 5 + t) (7, 7 + t) or any other interval of duration t. For all these intervals, the probability will have homogeneity in time, i.e., it will remain $P_n(t)$ for all intervals of duration t.

iii) In an interval of infinitesimal duration Δt, i.e., for very small Δt, the probability of exactly one occurrence is $\lambda\Delta t + o(\Delta t)$ and that of more than one occurrences is $o(\Delta t)$, where $o(\Delta t)$ is a function of Δt which tends to 0 more rapidly than Δt. The function $o(\Delta t)$ contains the terms of squares and higher powers of Δt. Thus, as $\Delta t \to 0$, $\frac{o(\Delta t)}{\Delta t} \to 0$. In other words,

$$P_1(\Delta t) = \lambda\Delta t + o(\Delta t), \qquad \ldots (12)$$

and $$P_2(\Delta t) + P_3(\Delta t) + P_4(\Delta t) + P_5(\Delta t) + \ldots = o(\Delta t) \qquad \ldots (13)$$

Since the sum of probabilities for occurrence/non-occurrence of an event is 1, we have

$$P_0(\Delta t) + P_1(\Delta t) + P_2(\Delta t) + P_3(\Delta t) + P_4(\Delta t) + \ldots = 1 \qquad \ldots (14)$$

Therefore,

$$\begin{aligned} P_0(\Delta t) &= 1 - P_1(\Delta t) - \{P_2(\Delta t) + P_3(\Delta t) + P_4(\Delta t) + \ldots\} \\ &= 1 - \{\lambda\Delta t + o(\Delta t)\} - \{o(\Delta t)\} \qquad \ldots (15) \\ &= 1 - \lambda\Delta t + o(\Delta t) \end{aligned}$$

Since o(Δt) is a fuction containing squares and higher powers of (Δt), o(Δt) + o(Δt) is very small and has been written as o(Δt) in equation (15).

Now, we can write the arrival-departure equations for the M/M/1 queueing model as follows:

If n ≥ 1, the probability of n customers in the system at time t + Δt

= [{Probability of n – 1 customers in the system at time t and (1 arrival, no departure in time Δt or 2 arrivals, one departure in time Δt or 3 arrivals, two departures in time Δt or . . .)} + {Probability of n – 2 customers in the system at time t and (2 arrivals, no departure in time Δt or 3 arrivals, one departure in time Δt or 4 arrivals, two departures in time Δt or . . .)}+ {Probability of n – 3 customers in the system at time t and (3 arrivals, no departure in time Δt or 4 arrivals, one departure in time Δt or 5 arrivals, two departures in time Δt or . . .)}+ ...]

+ [{Probability of n + 1 customers in the system at time t and (1 departure, no arrival in time Δt or 2 departures, one arrival in time Δt or 3 departures, two arrivals in time Δt or . . .)} + {Probability of n + 2 customers in the system at time t and (2 departures, no arrival in time Δt or 3 departures, one arrival in time Δt or 4 departures, 2 arrivals in time Δt or . . .)} + {Probability of n + 3 customers in the system at time t and (3 departures, no arrival in time Δt or 4 departures, one arrival in time Δt or 5 departures, two arrivals in time Δt or . . .)}+ ...] + [{Probability of n customers in the system at time t and (no arrival, no departure in time Δt or 1 arrival, one departure in time Δt or 2 arrivals, 2 departures in time Δt or . . .)]

$$
\begin{aligned}
&= P_{n-1}(t).\left[\{\lambda\Delta t + o(\Delta t)\}\{1-\mu\Delta t + o(\Delta t)\} + \text{The terms equal to } o(\Delta t)\right] \\
&\quad + P_{n-2}(t).\left[\{o(\Delta t)\}\{1-\mu\Delta t + o(\Delta t)\} + \text{The terms equal to } o(\Delta t)\right] \\
&\quad + P_{n-3}(t).o(\Delta t) + \text{ The terms equal to } o(\Delta t) \\
&\quad + P_{n+1}(t).\left[\{\mu\Delta t + o(\Delta t)\}\{1-\lambda\Delta t + o(\Delta t)\} + \text{The terms equal to } o(\Delta t)\right] \\
&\quad + P_{n+2}(t).\left[\{o(\Delta t)\}\{1-\lambda\Delta t + o(\Delta t)\} + \text{The terms equal to } o(\Delta t)\right] \\
&\quad + P_{n+3}(t).o(\Delta t) + \text{ The terms equal to } o(\Delta t) \\
&\quad + P_n(t).[\{1-\mu\Delta t + o(\Delta t)\}\{1-\lambda\Delta t + o(\Delta t)\} + \{\mu\Delta t + o(\Delta t)\}\{\lambda\Delta t + o(\Delta t)\} \\
&\quad + \text{The terms which ultimately become equal to } o(\Delta t)]
\end{aligned}
$$

Notice that $\Delta t.o(\Delta t), o(\Delta t).o(\Delta t), o(\Delta t)+o(\Delta t), o(\Delta t)-o(\Delta t), \Delta t.\Delta t$ are very small and may be taken as equal to $o(\Delta t)$.

We deal with the case n = 0 separately because there cannot be any possibility of less than zero customers. However, in the above equation, the case of less than n customers is included.

The probability of 0 customers in the system at time t + Δt

= [{Probability of 1 customer in the system at time t and (1 departure, no arrival in time Δt or 2 departures, one arrival in time Δt or 3 departures, two arrivals in time Δt)} + {Probability of 2 customers in the system at time t and (2 departures, no arrival in time Δt or 3 departures, one arrival in time Δt or 4 departures, 2 arrivals in time Δt)}+ {Probability of 3 customers in the system at time t and (3 departures, no arrival in time Δt or 4 departures, one arrival in time Δt or 5 departures, two arrivals in time Δt)}+ ...] + [{Probability of 0 customers in the system at time t and (no arrival, no departure in time Δt or 1 arrival, one departure in time Δt or 2 arrivals, 2 departures in time Δt)]

$$= P_1(t).\left[\{\mu\Delta t + o(\Delta t)\}\{1-\lambda\Delta t + o(\Delta t)\}\right.$$
$$\left.+ \text{The terms which ultimately become equal to } o(\Delta t)\right]$$
$$+ P_2(t).\left[\{o(\Delta t)\}\{1-\lambda\Delta t + o(\Delta t)\}\right.$$
$$\left.+ \text{The terms which ultimately become equal to } o(\Delta t)\right]$$
$$+ P_3(t).o(\Delta t) + \text{ The terms which ultimately become equal to } o(\Delta t)$$
$$+ P_0(t).[\{\text{ Probability of no departure which is 1 as there is no customer}$$
$$\text{in system and hence no chance of departure}\}.\{1-\lambda\Delta t + o(\Delta t)\}$$
$$+\{\mu\Delta t + o(\Delta t)\}\{\lambda\Delta t + o(\Delta t)\} + \text{The terms which ultimately become equal to } o(\Delta t)]$$

Therefore, the arrival-departure equations for the M/M/1 queueing model are:

$$\begin{cases} P_n(t+\Delta t) = P_{n-1}(t).\left[\lambda\Delta t + o(\Delta t)\right] + P_{n+1}(t)\left[\mu\Delta t + o(\Delta t)\right] \\ \qquad + P_n(t)\left[\lambda\Delta t + o(\Delta t)\right]\left[\mu\Delta t + o(\Delta t)\right] \\ \qquad + P_n(t)\left[1-\lambda\Delta t\right]\left[1-\mu\Delta t\right] + o(\Delta t),\ n \ge 1 \\ P_0(t+\Delta t) = P_1(t)\left[\mu\Delta t + o(\Delta t)\right] + P_0(t)\left[\lambda\Delta t + o(\Delta t)\right] \\ \qquad + P_0(t)\left[1-\lambda\Delta t\right].1 + o(\Delta t), n = 0 \end{cases} \quad \ldots(16)$$

$$\Rightarrow \begin{cases} P_n(t+\Delta t) = P_{n-1}(t).\lambda\Delta t + P_{n+1}(t)\mu\Delta t \\ \qquad + P_n(t)\left[1-\lambda\Delta t - \mu\Delta t\right] + o(\Delta t), \quad n \ge 1 \\ P_0(t+\Delta t) = P_1(t)\mu\Delta t + P_0(t)\left[1-\lambda\Delta t\right] + o(\Delta t) \end{cases} \quad \ldots(17)$$

Dividing equation (16) by Δt and taking the limit as Δt tends to zero, we have

$$p_n'(t) = \lambda\ P_{n-1}(t) + \mu\ P_{n+1}(t) - (\lambda+\mu)P_n(t)$$
$$[\because \lim_{n\to\infty} \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = p_n'(t)\left[\text{derivative of } P_n(t)\right] \quad \ldots(18)$$

and $$p_0'(t) = -\lambda\ P_0(t) + \mu P_1(t) \quad \ldots(19)$$

When steady state, i.e., the equilibrium state, is reached, $p_n'(t)$ becomes independent of time, say p_n , and the rate of its change with respect to time becomes zero, i.e.,

$$p_n'(t) = 0$$

Therefore, the steady-state solution is given by

$$\begin{cases} \lambda P_{n-1} + \mu P_{n+1} - (\lambda+\mu)P_n = 0,\ n \ge 1 \\ \mu P_1 = \lambda P_0 \end{cases} \quad \ldots(20)$$

Now, from equation (20), we have

$$\mu P_{n+1} - \lambda P_n = \mu P_n - \lambda P_{n-1} \quad \ldots(21)$$

This implies that

$$\mu P_n - \lambda P_{n-1} = \mu P_{n-1} - \lambda P_{n-2} \text{ (changing n to } n-1)$$

$$\mu P_{n-1} - \lambda P_{n-2} = \mu P_{n-2} - \lambda P_{n-3} \text{ (again changing n to } n-1)$$

$$\mu P_{n-2} - \lambda P_{n-3} = \mu P_{n-3} - \lambda P_{n-4} \text{ (again changing n to } n-1)$$

...

...

...

$$\mu P_2 - \lambda P_1 = \mu P_1 - \lambda P_0$$

But, $\mu P_1 - \lambda P_0 = 0$ $\qquad [\because \mu P_1 = \lambda P_0]$

Therefore,
$$\begin{aligned} \mu P_{n+1} - \lambda P_n &= \mu P_n - \lambda P_{n-1} \\ &= \mu P_{n-1} - \lambda P_{n-2} \\ &= \mu P_1 - \lambda P_0 = 0 \end{aligned}$$

This implies that

$$P_n = \frac{l}{m} P_{n-1}$$

$$= \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} P_{n-2}\right)$$

$$= \frac{\lambda}{\mu}\frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} P_{n-3}\right)$$

...

$$= \frac{\lambda.\lambda \ldots \lambda (\text{n times})}{\mu.\mu \ldots \mu (\text{n times})} \; P_0$$

$$= \frac{\lambda^n}{\mu^n} \; P_0 \quad = (\lambda/\mu)^n P_0 \;\; = \rho^n P_0$$

Since $\sum_{n=0}^{\infty} P_n = 1$, it follows that $\sum_{n=0}^{\infty} P_0 \rho^n = 1$

$$\Rightarrow P_0 (1 + \rho + \rho^2 + \ldots + \rho^n) = 1$$

$$\Rightarrow P_0 \left(\frac{1}{1-\rho}\right) = 1 \Rightarrow P_0 = 1 - \rho$$

Therefore, the probability of n customers (units) in the system is given by

$$P_n = (1-\rho)\rho^n, \quad n \geq 0 \qquad \ldots (22)$$

We can now determine the operating characteristics for the M/M/1queueing model.

6.5.2 Operating Characteristics for the M/M/1 Queueing Model

1. The average number of customers in the system is given by

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

$$= \frac{\rho}{1-\rho} \text{ (on simplification)} \qquad \text{...(23)}$$

$$\text{or } L_s = \frac{\lambda}{\mu-\lambda} \qquad \text{... (24)}$$

2. The average queue length is

$$L_q = L_s - \text{Traffic intensity}$$

$$= L_s - r = \frac{r}{1-r} - r$$

$$\text{or} \qquad L_q = \frac{r^2}{1-r} = \frac{l^2}{m(m-l)} \qquad \text{... (25)}$$

3. The average time an arrival spends in the system is

$$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu-\lambda} \qquad \text{... (26)}$$

4. The average waiting time of an arrival in the queue is

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} \qquad \text{... (27)}$$

W_q can also be obtained using the following result:

$$W_q = W_s - \frac{1}{m} \qquad \text{... (28)}$$

5. The probability that the number of units waiting in the queue and the number of units being serviced is greater than k is

$$P[n > k] = \rho^{k+1} \qquad \text{... (29)}$$

6. The probability of having a queue, i.e. $[1-P_0] = \rho$ ... (30)

We now illustrate the M/M/1 queueing model with some examples.

Example 1: A TV repairman finds that the time spent on his job has an exponential distribution with mean 30 minutes. If he repairs sets in the order in which these come in, and if the arrival of sets is approximately Poisson with an average rate of 10 per 8-hour day, what is the repairman's expected idle time each day? How many jobs are ahead of the average set just brought in?

Solution: It is given that

$$\text{Arrival rate } (\lambda) = \frac{10}{8} \text{ arrivals/hr}$$

$$\text{Service rate } (\mu) = \frac{60}{30} = 2 \text{ services/hr}$$

$\therefore$ The probability of the repairman being idle is $P_0 = 1-\rho$

$$\text{or} \qquad P_0 = 1-\frac{\lambda}{\mu} = 1-\frac{5}{8} = \frac{3}{8}$$

Hence, the idle time in the 8-hour day = $\frac{3}{8} \times 8$ hours = 3 hours

Now, the average number of jobs that are ahead of the set just brought in

= Average number of units in the system

$$= \frac{\lambda}{\mu-\lambda} = \frac{\frac{10}{8}}{2-\frac{10}{8}} = 1.67$$

Example 2: Customers arrive at a window in a bank, according to a Poisson distribution with mean 10 per hour. Service time per customer is exponential with mean 5 minutes. The space in front of the window including that for the serviced customers can accommodate a maximum of three customers. Other customers can wait outside this space.

a) What is the probability that an arriving customer can go directly to the space in front of the window?

b) What is the probability that an arriving customer will have to wait outside the indicated space?

c) How long is an arriving customer expected to wait before being served?

Solution: Arrival rate λ = 10 per hour

Service rate $\mu = \frac{60}{5}$ per hour = 12 per hour

$$\rho = \frac{10}{12} = \frac{5}{6}$$

a) From equation (22), the required probability = $P_0 + P_1 + P_2$

$$= (1-r) + (1-r)\,r + (1-r)\,r^2 \approx 0.42$$

b) The required probability $= P_3 + P_4 + P_5 + \ldots\ldots\ldots$

$$= 1 - (P_0 + P_1 + P_2)$$

$$= 1 - 0.42 = 0.58$$

c) Expected waiting time before being served = $W_q = \frac{\lambda}{\mu(\mu-\lambda)}$

$$= \frac{10}{12\,(12-10)} = 0.417 \text{ hours}$$

Example 3: Arrivals of machinists at a tool crib are considered to be Poisson distributed at an average rate 6 per hour. The length of time the machinists must remain at the tool crib is exponentially distributed with average time of 0.05 hours.

a) What is the probability that a machinist arriving at the tool crib will have to wait?

b) What is the average number of machinists at the tool crib?

c) The company will install a second tool crib when convinced that a machinist would have to spend 6 minutes in waiting and being served at the tool crib. At what rate should the arrival of machinists to the tool crib increase to justify the addition of a second crib?

Solution: a) Here the arrival rate λ = 6/hr and the service rate μ = 20/hr.

Therefore, the probability of zero customer in queue is

$$P_0 = 1 - \rho \qquad \text{where } \rho = \frac{\lambda}{\mu} = \frac{6}{20}$$

or $$P_0 = 1 - \frac{6}{20} = \frac{14}{20} = \frac{7}{10} = 0.7$$

The probability that a machinist arriving at the tool crib will have to wait

= Probability that there is at least one machinist at the tool crib

= 1– Probability that there is no machinist at the tool crib

= $1 - P_0 = 1 - 0.7 = 0.3$

b) The average number of machinists at the tool crib is given by

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{6}{20-6} = \frac{6}{14} = 0.428 \approx 0.43$$

c) The company is ready to install a second tool crib when convinced that a machinist would have to spend 6 min. in waiting and being served. Let the increased arrival rate be λ'.

Waiting time in the system = 6 min. = $\frac{1}{10}$ hr. We have

$$\frac{1}{\mu - \lambda'} = \frac{1}{10}$$

or $$\frac{1}{20 - \lambda'} = \frac{1}{10} \quad \text{or} \quad 10 = 20 - \lambda'$$

$\Rightarrow$ λ' = 20 – 10= 10/hr.

The increase is, therefore, (10 – 6)/hr = 4/hr.

Example 4: A repairman is to be hired to repair machines which break down at an average rate of 3 per hour. The breakdown follows a Poisson distribution. Nonproductive time of a machine is considered to cost `10 per hour. Two repairmen have been interviewed of whom one is slow but charges less and the other is fast but more expensive. The slow repairman charges `5 per hour and services breakdown machines at the rate of 4 per hour. The fast repairman demands `7 per hour, but services breakdown machines at an average rate of 6 per hour. Which repairman should be hired?

Solution: The data given is summarised below:

Slow/Less expensive Repairman	**Fast/ More expensive Repairman**
λ = 3/hr, μ = 4/hr, Labour cost = `5/hr	λ = 3/hr, μ = 6/hr, Labour cost = `7/hr

<u>Case of Slow/Less expensive Repairman</u>

Cost of engaging slow repairman for 8 hours

= Breakdown cost + Labour cost for 8-hour working day.

= (No. of breakdowns per hr) × 8 hours × Av. time spent in the system
× (Breakdown cost/ cost for non-productive time)
+ Labour cost per hr × 8 hours

$= (\lambda) \times 8 \times W_s \times 10 + 5 \times 8$

Since the average time spent in the system is $W_s = \dfrac{1}{\mu - \lambda} = \dfrac{1}{4-3} = 1$,

the total cost for engaging the slow repairman is

$$3 \times 8 \times 1 \times 10 + 5 \times 8 = 240 + 40 = \text{`}280$$

Case of Fast/More expensive Repairman

The cost of engaging fast repairman for 8 hours

= Breakdown cost + Labour cost in 8-hour working day.

= (No. of breakdowns per hr) × 8 hours × Av. time spent in the system
× (Breakdown cost/ cost for non-productive time)
+ Labour cost per hr × 8 hours

$= (\lambda) \times 8 \times W_s \times 10 + 7 \times 8$

Since the average time spent in the system is $W_s = \dfrac{1}{\mu - \lambda} = \dfrac{1}{6-3} = \dfrac{1}{3}$,

the total cost for engaging the fast repairman is

$$3 \times 8 \times \frac{1}{3} \times 10 + 7 \times 8 = 80 + 56 = \text{`}136$$

Since the total cost of engaging the fast repairman is less, we should engage the fast repairman.

You should now apply the M/M/1 model to solve the following exercises.

E1) Customers arrive at a box office window being served by a single individual according to a Poisson input process with a mean rate of 30 per hour. The time required to serve a customer has an exponential distribution with a mean of 90 seconds. Find the average waiting time of a customer in the queue.

E2) Arrivals at a telephone booth are considered to be Poisson with an average time of 10 minutes between one arrival and the next. The length of a phone call is assumed to be distributed exponentially with mean 3 minutes.

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

b) The telephone company will install a second booth when convinced that an arrival would be expected to wait at least 3 minutes for the phone. By how much should the flow arrivals increase in order to justify the setting up of a second booth?

E3) A fertiliser company distributes its products by trucks loaded at its only loading station. Both company trucks and contractor's trucks are used for this purpose. It was noticed that on an average one truck arrived every 5 minutes and the average loading time was 3 minutes. Forty percent of the trucks belong to the contractor. Determine the expected waiting time of contractor's trucks per day.

E4) In a railway marshalling yard, goods trains arrive at a rate of 36 trains per day. Assuming that the inter-arrival time follows exponential distribution and the service time distribution is also exponential with an average of 30 minutes, calculate the following:

a) The mean line length (mean length of the system).

b) The probability that the queue size exceeds 10.

c) If the input increases to an average of 42 per day, what will the change in (a) and (b) be?

E5) A road transport company has one reservation clerk on duty at a time. She handles information of bus schedules and makes reservations. Customers arrive at a rate of 8 per hr and on an average the clerk can service 12 customers per hr.

i) What is the average waiting time of a customer in the system?

ii) The management is contemplating to install a computer system to handle the information and reservations. This is expected to reduce the service time from 5 to 3 minutes. The additional cost of having the new system works out to be `50/- per day. If the cost of the goodwill of having a customer to wait is estimated to be 12 paise per minute spent waiting before being served, should the company install the computer system? Assume an 8-hour working day.

iii) What is the expected waiting time of all the customers in an 8-hour day?

Let us summarise the concepts that we have discussed in this unit.

6.6 SUMMARY

1. The purpose of queueing theory is to provide information for determining an acceptable level of service since providing too much service capacity is costly (owing to idle employees or equipment) and providing too little service capacity is also costly (owing to waiting members of the queue). A queueing system is described by three elements: (1) Input or arrival process of customers, (2) Service mechanism, and (3) Queue discipline.

2. **Input process** is concerned with the pattern in which the customers arrive and join the system. An input process is characterised by its size, the arrival time distribution, and the attitude of the customers.

3. **Service time** distributions are generally exponential distributions. Service facility can be 'Single channel facility' or 'One queue-several station facilities' or 'Several queues-one service station', 'Multi-channel facility' or 'Multi-Stage Channel facility'.

4. **Queue discipline** refers to the order in which the service station selects the next customer from the waiting line to be served. It may be FCFS, LIFO or SIRO. In this unit, the FCFS queue discipline has been considered.

5. The time dependent solutions are known as **transient solutions**. **Steady state solution** is independent of time and represents the probability of the system being in the equilibrium state.

6. A queueing system is usually described by five symbols such as **a/b/c : d/e** or **a/b/c/d/e**. The first symbol '**a**' describes the arrival process. The second symbol '**b**' describes the service time distribution. The third symbol '**c**' stands for the number of servers. The symbols '**d**' and '**e**' stand for system

capacity and queue discipline, respectively. If arrivals are Poisson, symbol **M** is used in place of '**a**' If departures are exponential, symbol **M** is used in place of '**b**'.

7. The **M/M/1 queueing model** deals with a queueing system having single service channel, Poisson input, and exponential services. There is no limit on the system capacity while the customers are served on a "First Come First Served" basis. The probability of n customers (units) in the system for this model is

$$P_n = (1-\rho)\rho^n, \quad n \geq 0,$$

where $\rho = \frac{\lambda}{\mu}$, λ is the average arrival rate and μ, the average service rate.

Some important characteristics of this model are:

- Average number of customers in the system $(L_s) = \frac{\lambda}{\mu - \lambda}$
- Average queue length $(L_q) = L_s - \rho = \frac{\lambda^2}{\mu(\mu-\lambda)}$
- Average time an arrival spends in the system $(W_s) = \frac{L_s}{\lambda} = \frac{1}{\mu-\lambda}$
- Average waiting time of an arrival in the queue

$$(W_q) = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} \quad = W_s - \frac{1}{\mu}$$

6.7 SOLUTIONS/ANSWERS

E1) Arrival rate (λ) = 30/hr

Service rate (μ) = $\frac{1}{90}$/seconds = $\frac{60 \times 60}{90}$/hr

Therefore, the average waiting time of a customer in a queue is

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{30}{40(40-30)} = \frac{3}{40} \text{ hours}$$

$$= \frac{3}{4}{}' \ 60 \text{ minutes} = 4.5 \text{ minutes}$$

E2) We are given that

Arrival rate (λ) = $\frac{1}{10}$ per minute

Service rate (μ) = $\frac{1}{3}$ person/minute

a) The probability that a person arriving at a booth will have to wait

= Probability that there is at least one customer

= $P(n > 0) = 1 - P_0 = \lambda/\mu$

= $\frac{1}{10} \times \frac{3}{1}$ = 0.3

b) The installation of the second booth will be justified only if the arrival rate is more than the waiting time. Let λ' be the increased arrival rate.

Then the expected waiting time in queue is

$$W_q = \frac{\lambda'}{\mu(\mu-\lambda')} \quad \Rightarrow \quad 3 = \frac{\lambda'}{\frac{1}{3}(\frac{1}{3}-\lambda')}$$

Thus $\lambda' = \frac{1}{6}$ per minute

Hence, the increase in the arrival rate is

$$= \frac{1}{6} - \frac{1}{10} = \frac{1}{15} \text{ per minute}$$

E3) We are given that

Arrival rate (λ) = 1/5 per minute $= \frac{60}{5}$ = 12 per hour

Service rate $(\mu) = \frac{1}{3}$ per minute $= \frac{60}{3}$ = 20 per hour

The expected waiting time of contractor's trucks per day

= Expected waiting time for a truck × Number of contractor's trucks per day

$$= \frac{\lambda}{\mu(\mu-\lambda)} \times (\text{40 per cent of the total number of trucks per day})$$

$$= \frac{\lambda}{\mu(\mu-\lambda)} \times [\text{40 per cent of (Arrival rate per hr} \times \text{24 hours)}]$$

$$= \frac{\lambda}{\mu(\mu-\lambda)} \times \frac{40}{100} \times (12 \times 24)$$

$$= \frac{12}{20(20-12)} \times \frac{40}{100} \times (12 \times 24) = 8.64 \text{ hours}$$

E4) We are given that

Arrival rate (λ) = 36 trains/day

Service rate (μ) =1 per 30 minutes

$$= \frac{60}{30} \text{ per hour} = \frac{60}{30} \times 24 \text{ per day} = 48 \text{ trains/day}$$

Therefore, traffic intensity $\rho = \frac{\lambda}{\mu} = 0.75$

a) Mean length of system

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{36}{48-36} = \frac{36}{12} = 3 \text{ trains}$$

b) Probability that the queue size exceeds 10

= Probability that the system size exceeds 11

i.e., $P(n > 11) = \rho^{11+1} = (0.75)^{12}$

c) If the input increases to an average of 42 per day, then the arrival rate (λ')= 42/day.

Hence, the new traffic intensity $\rho' = \frac{42}{48} = \frac{7}{8}$

Therefore, in this case, $L_s = \frac{\lambda'}{\mu - \lambda'} = \frac{42}{48-42} = \frac{42}{6} = 7$ trains

and the probability that the queue size exceeds 10 is $(\rho')^{11+1} = \left(\frac{7}{8}\right)^{12}$

E5) We are given that

Arrival rate (λ) = 8/ hr,

Service rate (μ) = 12/ hr

i) Average waiting time of customers in the system is

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{12-8} = \frac{1}{4}\text{hr}$$

ii) An arrival waits for W_q hours before being served and there are λ arrivals per hour. Thus, the expected waiting time for all customers in an 8-hour day with one system

$$= 8\lambda \times W_q$$

With the existing system, the waiting time for the customers, therefore, is

$$= 8 \times 8 \times \frac{\lambda}{\mu(\mu - \lambda)} = 8 \times 8 \times \frac{8}{12(12-8)}$$

$$= \frac{64}{6} \text{ hours} = 640 \text{ min.}$$

The goodwill cost per day with the existing system

$$= 640 \times \frac{12}{100} = \text{` } 76.80$$

If a computer system is installed, the service rate is 1 per 3 minutes or 20 per hr. Hence, the expected waiting time in this case = $8\lambda \times W_q'$

$$= 8 \times 8 \times \frac{8}{20(20-8)} = 8 \times 8 \times \frac{1}{30}\text{ hr} = 128 \text{ min.}$$

The goodwill cost per day when the computer system is installed

$$= 128 \times \text{` } \frac{12}{100} = \text{` } 15.36$$

The additional cost of computer system = ` 50

Therefore, the total cost of having the computer system

= ` (50 + 15.36) = ` 65.36

Since this amount is less than the goodwill cost to be incurred with the existing system, the management should install a computer system.