Operations Research: Unit 5 (Queuing Theory)
Lecture Notes

Dr K MANOJ,
Assistant Professor, Department of Statistics,
Manonmaniam Sundaranar University, Tirunelveli - 12, Tamilnadu, India

April 16, 2020

Contents

1 Introduction

Queuing theory deals with problems which involve queuing (or waiting). Before going to queuing theory, one has to understand two things in clear. They are service and customer or element. Here customer or element represents a person or machine or any other thing, which is in need of some service from servicing point. Service represents any type of attention to the customer to satisfy his need.

In essence all queuing systems can be broken down into individual sub-systems consisting of entities queuing for some activity (as shown below).

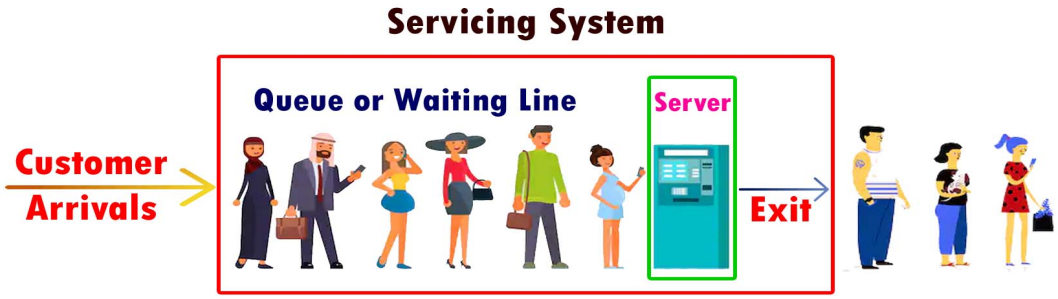


Figure 1: Queuing System Activities

For example,

1. Person going to hospital to get medical advice from the doctor is an element or a customer,
2. A person going to railway station or a bus station to purchase a ticket for the journey is a customer or an element,
3. A person at ticket counter of a cinema hall is an element or a customer,
4. A person at a grocery shop to purchase consumables is an element or a customer,
5. A bank pass book tendered to a bank clerk for withdrawal of money is an element or a customer,
6. A machine break down and waiting for the attention of a maintenance crew is an element or a customer.
7. Vehicles waiting at traffic signal are elements or customers,
8. A train waiting at outer signal for green signal is an element or a customer

2 Notations and Terminology

Basic terminology and Notations of queuing system

n = number of customers/units in the system
$p_n(t)$ = transient state probability that exactly
P_n = steady state probability of having n units in the system
λ_n = average number of customers arriving per unit of time, when there are already n units in the system
λ = average arrival rate when λ_n is constant for all n
μ_n = average number of customers being served per unit of time, when there are already n units in the system
μ = average service rate when μ_n is constant for all $n \geq 1$.
s = number of parallel service channels in the system
$1/\lambda$ = inter arrival time between two arrivals
$1/\mu$ = service time between two units or customers
ρ = traffic intensity or utilization factor for service facility, i.e., the expected fraction of time the servers are busy
N = maximum number of customers allowed in the system
L_s = average number of customers in the system
L_q = average number of customers in the queue
W_s = average waiting time in the system
W_q = average waiting time in the queue
P_w = probability of a customer having to wait for service

3 Queueing models and Classifications

Most elementary queuing models assume that the inputs / arrivals and outputs / departures follow a birth and death process. Any queuing model is characterized by situations where both arrivals and departures take place simultaneously. Depending upon the nature of inputs and service faculties, there can be a number of queuing models as shown below:

(i) Probabilistic queuing model: Both arrival and service rates are some unknown random variables.

(ii) Deterministic queuing model: Both arrival and service rates are known and fixed.

(iii) Mixed queuing model: Either of the arrival and service rates is unknown random variable and other known and fixed.

Arrival pattern / Service pattern / Number of channels / (Capacity / Order of servicing). $(A/B/S/(d/f)$.
In general M is used to denote Poisson distribution (Markovian) of arrivals and departures.
D is used to constant or Deterministic distribution.
E_k is used to represent Erlangian probability distribution.
G is used to show some general probability distribution
In general queuing models are used to explain the descriptive behavior of a queuing system. These quantify the effect of decision variables on the expected waiting times and waiting lengths as well as generate waiting cost and service cost information. The various systems can be evaluated through these aspects and the system, which offers the minimum total cost is selected.

Procedure for Solution

(a) List the alternative queuing system
(b) Evaluate the system in terms of various times, length and costs.
(c) Select the best queuing system.

4 Queuing System (or) Process

Queuing system can be completely described by:

- The input (Arrival pattern)
- The service mechanism or service pattern
- The queue discipline and
- Customer behavior.

Components of the queuing system are arrivals, the element waiting in the queue, the unit being served, the service facility and the unit leaving the queue after service. This is shown in figure 2.

Input Process

The input describes the way in which the customers arrive and join the system. In general customer arrival will be in random fashion, which cannot be predicted, because the customer is an independent individual and the service organization has no control over the customer. The characteristics of arrival are shown in figure 3.

Input to the queuing system refers to the pattern of arrival of customers at the service facility. We can see at ticket counters or near petrol bunks or any such service facility that the customer arrives randomly individually or in batches. The input process is described by the following characteristics (as shown in the figure) nature of arrivals, capacity of the system and behavior of the customers.

(a) Size of arrivals: The size of arrivals to the service system is greatly depends on the nature of size of the population, which may be infinite or finite. The arrival pattern can be more clearly described in terms of probabilities and consequently the probability distribution for inter- arrival times i.e. the time between two successive arrivals or the distribution of number of customers arriving in unit time must be defined.

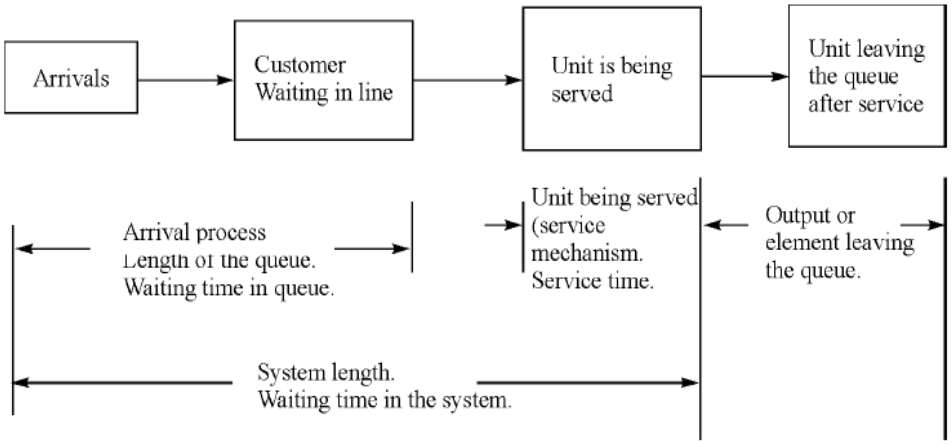


Figure 2: Components of queuing system

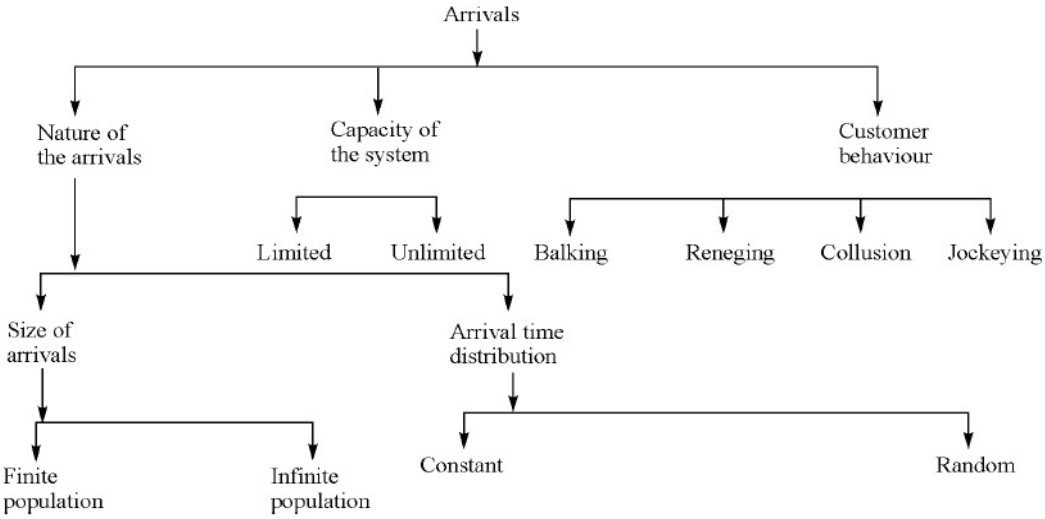


Figure 3: Characteristics of Arrivals or input

(b) Inter-arrival time: The period between the arrival of individual customers may be constant or may be scattered in some distribution fashion. Most queuing models assume that the some inter-arrival time distraction applies for all customers throughout the period of study. It is true that in most situations that service time is a random variable with the same distribution for all arrivals, but cases occur where there are clearly two or more classes of customers such as a machine waiting for repair with a different service time distribution. Service time may be constant or random variable.

(c) Capacity of the service system: In queuing context the capacity refers to the space available for the arrivals to wait before taken to service. The space available may be limited or unlimited. When the space is limited, length of waiting line crosses a certain limit; no further units or arrivals are permitted to enter the system till some waiting space becomes vacant. This type of system is known as system with finite capacity and it has its effect on the arrival pattern of the system, for example a doctor giving tokens for some customers to arrive at certain time and the present system of allowing the devotees for darshan at Tirupathi by using the token belt system.

(d) Customer behaviour: The length of the queue or the waiting time of a customer or the idle time of the service facility mostly depends on the behaviour of the customer. Here the behaviour refers to the impatience of a customer during the stay in the line. Customer behaviour can be classified as:

(i) Balking: This behaviour signifies that the customer does not like to join the queue seeing the long length of it. This behaviour may effect in loosing a customer by the organization. Always a lengthy queue indicates

insufficient service facility and customer may not turn out next time. For example, a customer who wants to go by train to his destination goes to railway station and after seeing the long queue in front of the ticket counter, may not like to join the queue and seek other type of transport to reach his destination.

(ii) Reneging: In this case the customer joins the queue and after waiting for certain time looses his patience and leaves the queue. This behaviour of the customer may also cause loss of customer to the organization.

(iii) Collusion: In this case several customers may collaborate and only one of them may stand in the queue. One customer represents a group of customer. Here the queue length may be small but service time for an individual will be more. This may break the patience of the other customers in the waiting line and situation may lead to any type of worst episode.

(iv) Jockeying: If there are number of waiting lines depending on the number of service stations, for example Petrol bunks, Cinema theaters, etc. A customer in one of the queue after seeing the other queue length, which is shorter, with a hope of getting the service, may leave the present queue and join the shorter queue. Perhaps the situation may be that other queue which is shorter may be having more number of Collaborated customers. In such case the probability of getting service to the customer who has changed the queue may be very less. Because of this character of the customer, the queue lengths may goes on changing from time to time.

Service Mechanism or Service Facility

The time required to serve the customer cannot be estimated until we know the need of the customer. Many a time it is statistical variable and cannot be determined by any means such as number of customers served in a given time or time required to serve the customer, until a customer is served completely.

Definition: Service Mechanism

Service facilities are arranged to serve the arriving customer or a customer in the waiting line is known as service mechanism.

Service facility design and service discipline and the channels of service as shown in figure 4 may generally determine the service mechanism.

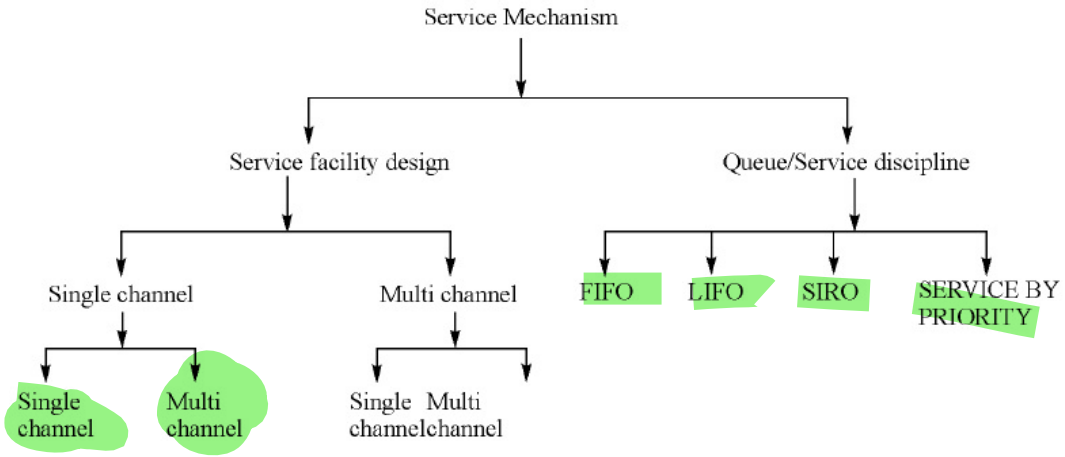


Figure 4: Service Mechanisms

(a) Service facility design: Arriving customers maybe asked to form a single line (Single queue) or multi line (multi queue) depending on the service need. When they stand in single line it is known as Single channel facility when they stand in multi lines it is known as multi channel facility.

(i) Single channel queues: If the organization has provided single facility to serve the customers, only one unit can be served at a time, hence arriving customers form a queue near the facility. The next element is drawn into service only when the service of the previous customer is over. Here also depending on the type of service the system is divided into Single phase and Multi phase service facility. In Single channel Single Phase queue,

the customer enters the service zone and the facility will provide the service needed. Once the service is over the customer leaves the system.

For example, Petrol bunks, the vehicle enters the petrol station. If there is only one petrol pump is there, it joins the queue near the pump and when the term comes, get the fuel filled and soon after leaves the queue. Or let us say there is a single ticket counter, where the arrivals will form a queue and one by one purchases the ticket and leaves the queue.

(ii) Multi Channel queues When the input rates increases, and the demand for the service increases, the management will provide additional service facilities to reduce the rush of customers or waiting time of customers. In such cases, different queues will be formed in front of different service facilities. If the service is provided to customers at one particular service center, then it is known as Multi channel Single-phase system. In case service is provided to customer in different stages or phases, which are in parallel, then it is known as multi channel multi phase queuing system.

(b) Queue discipline or Service discipline When the customers are standing in a queue, they are called to serve depending on the nature of the customer. The order in which they are called is known as Service discipline. There are various ways in which the customer called to serve. They are:

(i) First In First Out (FIFO) or First Come First Served (FCFS): We are quite aware that when we are in a queue, we wish that the element which comes should be served first, so that every element has a fair chance of getting service. Moreover it is understood that it gives a good morale and discipline in the queue. When the condition of FIFO is violated, there arises the trouble and the management is answerable for the situation.

(ii) Last in first out (LIFO) or Last Come First Served (LCFS): In this system, the element arrived last will have a chance of getting service first. In general, this does not happen in a system where human beings are involved. But this is quite common in Inventory system. Let us assume a bin containing some inventory. The present stock is being consumed and suppose the material ordered will arrive that is loaded into the bin. Now the old material is at the bottom of the stock where as fresh arrived material at the top. While consuming the top material (which is arrived late) is being consumed. This is what we call Last come first served). This can also be written as First In Last Out (FILO).

(iii) Service In Random Order (SIRO): In this case the items are called for service in a random order. The element might have come first or last does not bother; the servicing facility calls the element in random order without considering the order of arrival. This may happen in some religious organizations but generally it does not followed in an industrial / business system. In religious organizations, when devotees are waiting for the darshan of the god man / god woman, the devotees are picked up in random order for blessings. Some times we see that in government offices, the representations or applications for various favors are picked up randomly for processing. It is also seen to allocate an item whose demand is high and supply is low, also seen in the allocation of shares to the applicants to the company.

(iv) Service By Priority: Priority disciplines are those where any arrival is chosen for service ahead of some other customers already in queue. In the case of Pre-emptive priority the preference to any arriving unit is so high that the unit is already in service is removed / displaced to take it into service. A non- pre-emptive rule of priority is one where an arrival with low priority is given preference for service than a high priority item. As an example, we can quote that in a doctors shop, when the doctor is treating a patient with stomach pain, suddenly a patient with heart stroke enters the doctors shop, the doctor asks the patient with stomach pain to wait for some time and give attention to heart patient. This is the rule of priority.

5 Definition of transient and Steady-states

The distribution of customer's arrival time and service time are the two constituents, which constitutes of study of waiting line. Under a fixed condition of customer arrivals and service facility a queue length is a function of time. As such a queue system can be considered as some sort of random experiment and the various events of the experiment can be taken to be various changes occurring in the system at any time. We can identify three states of nature in case of arrivals in a queue system. They are named as steady state, transient state, and the explosive state.

Definition: Transient State

Queuing theory analysis involves the study of a system's behaviour over time. A system is said to be in '**transient state**' when its operating characteristics or behaviour are dependent on time. This happens usually at initial stages of operation of the system, where its behaviour is still dependent on the initial conditions. So when the probability distribution of arrivals, waiting time and servicing time are dependent on time the system is said to be in transient state.

Definition: Steady State

The system will settle down as steady state when the rate of arrivals of customers is less than the rate of service and both are constant. The system not only becomes steady state but also becomes independent of the initial state of the queue. Then the probability of finding a particular length of the queue at any time will be same. Though the size of the queue fluctuates in steady state the statistical behaviour of the queue remains steady. Hence we can say that a steady state condition is said to prevail when the behaviour of the system becomes independent of time.

A necessary condition for the steady state to be reached is that elapsed time since the start of the operation becomes sufficiently large i.e. $(t \to \infty)$, but this condition is not sufficient as the existence of steady state also depend upon the behaviour of the system i.e. if the rate of arrival is greater than the rate of service then a steady state cannot be reached. Hence we assume here that the system acquires a steady state as $t \to \infty$ i.e. the number of arrivals during a certain interval becomes independent of time. i.e.

$$\lim_{t \to \infty} P_n(t) \to P_n$$

Hence in the *steady state system, the probability distribution of arrivals, waiting time, and service time does not depend on time.*

6 Kendall's Notations and Classification of Queuing Models

Different models in queuing theory are classified by using special (or standard) notations described initially by D.G.Kendall in 1953 in the form (a/b/c). Later A.M.Lee in 1966 added the symbols d and c to the Kendall notation. Now in the literature of queuing theory the standard format used to describe the main characteristics of parallel queues is as follows:

$$(a/b/c) : (d/c)$$

Where

a = arrivals distribution

b = service time (or departures) distribution

c = number of service channels (servers)

d = max. number of customers allowed in the system (in queue plus in service)

e = queue (or service) discipline.

Certain descriptive notations are used for the arrival and service time distribution (i.e. to replace notation a and b) as following:

M= exponential (or markovian) inter-arrival times or service-time distribution (or equivalently poisson or markovian arrivel or departure distribution)

D= constant or deterministic inter-arrival-time or service-time.

G = service time (departures) distribution of general type, i.e. no assumption is made about the type of distribution.

GI = Inter-arrival time (arrivals) having a general probability distribution such as as normal, uniform or any empirical distribution.

E_k = Erlang-k distribution of inter-arrival or service time distribution with parameter k (i.e. if $k = 1$, Erlang is equivalent to exponential and if $k =$, Erlang is equivalent to deterministic).

For example, a queuing system in which the number of arrivals is described by a Poisson probability distribution, the service time is described by an exponential distribution, and there is a single server, would be designed by M/M/I.

The Kendall notation now will be used to define the class to which a queuing model belongs. The usefulness of a model for a particular situation is limited by its assumptions.

7 Distributions in queuing systems

The common basic waiting line models have been developed on the assumption that arrival rate follows the Poisson distribution and that service times follow the negative exponential distribution. This situation is commonly referred to as the Poisson arrival and Exponential holding time case. These assumptions are often quite valid in operating situations. Unless it is mentioned that arrival and service follow different distribution, it is understood always that arrival follows Poisson distribution and service time follows negative exponential distribution.

On queuing models have conducted careful study about various operating conditions like - arrivals of customers at grocery shops, Arrival pattern of customers at ticket windows, Arrival of breakdown machines to maintenance etc. and confirmed almost all arrival pattern follows nearly Poisson distribution. Although we cannot say with finality that distribution of arrival rates are always described adequately by the Poisson, there is much evidence to indicate that this is often the case. We can reason this by saying that always Poisson distribution corresponds to completely random arrivals and it is assumed that arrivals are completely independent of other arrivals as well as any condition of the waiting line. The commonly used symbol for average arrival rate in waiting line models is the Greek letter Lamda (λ), arrivals per time unit. It can be shown that when the arrival rates follow a Poisson processes with mean arrival rate λ, the time between arrivals follow a negative exponential distribution with mean time between arrivals of $1/\lambda$. This relationship between mean arrival rate and mean time between arrivals does not necessarily hold for other distributions. The negative exponential distribution then, is also representative of Poisson process, but describes the time between arrivals and specifies that these time intervals are completely random.

The distribution of arrivals in a queuing system can be considered as a pure birth process. The term birth refers to the arrival of new calling units in the system the objective is to study the number of customers that enter the system, i.e. only arrivals are counted and no departures takes place. Such process is known as pure birth process. An example may be taken that the service station operator waits until a minimum-desired customers arrives before he starts the service.

8 Solution of queuing models: Model-I :$(M/M/1:\infty/FCFS)$

The derivation of this model is based on certain assumptions about the queuing system:

1. Exponential distribution of inter-arrival times or poisson distribution of arrival rate.

$$\int^{\infty} \frac{\lambda}{\mu}(\mu-\lambda)\, e^{-2(\mu-\lambda)}$$

2. Single waiting line with no restriction on length of queue (i.e. infinite capacity) and no banking or reneging.
3. Queue discipline is 'first-come, first-serve
4. Single serve with exponential distribution of service time

Performance characteristics

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

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \left(1-\frac{\lambda}{\mu}\right) = \rho^n(1-\rho);\ \rho<1,\ n=0,1,2...$$

P_w = probability of server being busy (i.e. customer has to wait) $= 1-\rho_{o=\lambda\ /\ \mu}$

1. Expected (or average) number of customer in the system (customers in the line plus the customer being served)

$$L_s = \sum_{n=1}^{\infty} nP_n = \sum_{n=1}^{\infty} n(1-p)p^n, 0<p<1$$

$$= \frac{p}{1-p} = \frac{\lambda}{\mu-\lambda}; p = \frac{\lambda}{\mu}$$

2. Expected (or average) queue length or expected number of customers waiting in the queue

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=1}^{\infty} np^n - \sum_{n=1}^{\infty} p_n = L_s - (1-P_0) = L_s - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

3. Expected (or average) waiting time of a customer in the queue

$$W_q = \lambda\left(1-\frac{\lambda}{\mu}\right)\frac{1}{(\mu-\lambda)^2} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{L_q}{\lambda}$$

4. Expected (or average) waiting time of a customer in the system (waiting and service)

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

5. Expected (or average) waiting time in the queue for busy system

$$W_b = \frac{\text{Expected waiting time of acustomerinthe queue}}{\text{Prob(system being busy)}} = \frac{1}{\mu-\lambda}$$

6. Probability of k or more customers in the system

$$P(n \geq k) = \left(\frac{\lambda}{\mu}\right)^k ; P(n > k) = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

7. The variance (fluctuation) of queue length $\frac{p}{(1-p)^2} = \frac{\lambda\mu}{(\mu-\lambda)^2}$

8. Expected non-empty queue length

   $L_b = \frac{L_s}{1-P_0} = \frac{\lambda}{\mu-\lambda} \times \frac{1}{\lambda/\mu} = \frac{\mu}{\mu-\lambda}$ (System)

   $L_b = \frac{1}{\mu-\lambda} + \frac{\lambda}{\mu}$ (Queue)

9. Probability that waiting time is more than $P(x > t) = \begin{cases} (\mu-\lambda)e^{-(\mu-\lambda)t} & \text{(System)} \\ \frac{\lambda}{\mu}(\mu-\lambda)e^{-(\mu-\lambda)} & \text{(Queue)} \end{cases}$

Solved Example Problem 1

A television repairman finds that the time spent on his jobs has an exponential distribution with mean of 30 minutes. If he repairs sets in the order in which they came in, and if the arrival of sets follows a Poisson distribution approximately 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:

From the data of the problem, we have

λ = 10/8 = 5/4 sets per hour; and μ= (1/30) 60 = 2 sets per hour

(a) Expected idle time of repairman each day = Number of hours for which the repairman remains busy in an 8-hour day (traffic intensity) is given by

(8) (λ /μ) = (8) (5/8) = 5 hours

Hence, the idle time for a repairman in an 8-hour day will be: (8 – 5) = 3 hours.

(b) Expected (or average) number of TV sets in the system

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{5/4}{2-(5/4)} = \frac{5}{3} = 2(\text{approx.}) \text{ TV sets}$$

Solved Example Problem 2

On an average 96 patients per 24-hour day require the service of an emergency clinic. 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 clinic Rs 100 per patient treated to obtain an average servicing time of 10 minutes, and that each minutes of decrease in this average time would cost Rs. 10 per patient treated. How much would have to be budgeted by the clinic to decrease the average size of the queue from one and one-third patients to half patient.

Solution:

From the data of the problem, we have
$\lambda = \frac{96}{24\times 60} = \frac{1}{15}$ and $\mu = \frac{1}{10}$ patients per minute; $p = \frac{\lambda}{\mu} = \frac{2}{3}$

1. Average number of patients in the queue

$$L_q = \frac{p^2}{1-p} = \frac{(2/3)^2}{1-2/3} = \frac{4}{3}$$

2. Fraction of the time for which there no patients, $P_0 = 1-p = 1-\frac{2}{3} = \frac{1}{3}$
3. When the average queue size is decreased from $4/3$ patient, the new service rate is determined as:

   $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$ or $\frac{1}{2} = \frac{(1/15)^2}{\mu(\mu-1/15)^2}$, i.e $\mu = \frac{2}{15}$ patients per minute.

   Average rate of treatment required is: $\frac{1}{\mu} = \frac{15}{2} = 7.5$ minutes i.e. a decrease in the average rate of treatment is 2.5(= 10 – 7.5) minutes.

   Budget per patient = Rs (100 + 2.5 x 10) = Rs 125 per patient.

9 Inter-relationship between L_q, L_s, W_q and W_s

It can be proved under general conditions of arrival, departure, and service discipline that the formulae

$$L_s = \lambda W_s$$
$$\text{and} \quad L_q = \lambda W_q$$

hold. These formulae act as key points in establishing the strong relationships between W_s, W_q, L_s and L_q which can be found as follows.

By definition, we have $W_q = W_s - 1/\mu$

Then multiplying both sides by λ, we get $L_q = L_s - \lambda/\mu$.

Solved Example Problem

Telephone users arrive at a booth following a Poisson distribution with an average time of 5 minutes between one arrival and the next. The time taken for a telephone call is on an average 3 minutes and it follows an exponential distribution. What is the probability that the booth is busy? How much service rate should be increased in order to reduce the waiting time to less than or equal to half of the present waiting time?

Solution:

Solution: Given that arrival rate, $\lambda = 12$ per hour. Service rate $\mu = 20$ per hour.

Probability that the booth is busy $= 1 - P_0 = \frac{\lambda}{\mu} = \frac{12}{20} = 0.60$

Average waiting time in queue $W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{12}{20(20-12} = \frac{3}{40}$ hour.

Average waiting time in the system $W_s = \frac{1}{\mu-\lambda} = \frac{1}{20-12} = \frac{1}{8}$ hour.

In case, the waiting time is required to be reduced to half,

we have $W_s' = \frac{1}{\mu'-\lambda} \Rightarrow \frac{1}{16} = \frac{1}{\mu'-12}$ or $\mu' = 28$ per hour.

Hence the increase in service rate should be 8 users per hour.

10 Model-II - General Erlangian queuing model $(M/M/1 : FCFS/\propto/\propto)$: (Birth - Death process)

In waiting line system each arrival can be considered to be a birth i.e. if the system is in the state E_n, i.e. there are n units in the system and there is an arrival then the state of the system changes to the state E_{n+1}. Similarly when there is a departure from the system the state of the system becomes E_{n-1}. Hence whole system is thus viewed as a birth and death process. When λ is the arrival rate of the system, will never be fixed and dependent on the queue length $'n'$, then it will mean that some person interested in joining the queue may not join due to long queue. Similarly, if μ is also dependent on the queue length it may affect the service rate. Hence in this case both λ and μ cannot be taken to be fixed. Three cases may occur, which are described below.

In this model, arrival rate and service rate i.e. λ and μ do not remain constant during the queuing phenomenon and vary to $\lambda_1, \lambda_2, \ldots \lambda_n$ and $\mu_1, \mu_2, \ldots \mu_n$ respectively. Then:
$p_1 = (\lambda_0/\mu_1)\, p_0$
$p_2 = (\lambda_0/\mu_1)\,(\lambda_1/\mu_2)\, p_0$

.....
.....

$p_n = (\lambda_0/\mu_1)\,(\lambda_1/\mu_2)\ldots(\lambda_{n-2}/\mu_{n-1}) \times (\lambda_{n-1}/\mu_n)\, p_0$

But there are some special cases when:

1. $\lambda_n = \lambda$ and $\mu_n = \mu$ then,
   $p_0 = 1-(\lambda/\mu), p_n = (\lambda/\mu)^n \times [1-(\lambda/\mu)]$
2. When $\lambda_n = \lambda/(n+1)$ and $\mu_n = \mu$
   $p_0 = e^{-p}$
   $p_n = [\rho^n/(n!)] \times e^{-\rho}$, where $\rho = (\lambda/\mu)$
3. When $\lambda_n = \lambda$ and $\mu_n = n \times \mu$ then $p_0 = e^{-\rho}$ and $p_n = (\rho^n/n!)\, x e^{-\rho}$

Solved Example Problem

A transport company has a single unloading berth with vehicles arriving in a Poisson fashion at an average rate of three per day. The unloading time distribution for a vehicle with 'n' unloading workers is found to be exponentially with an average unloading time (1/2) xn days. The company has a large labour supply without regular working hours, and to avoid long waiting lines, the company has a policy of using as many unloading group of workers in a vehicle as there are vehicles waiting in line or being unloaded. Under these conditions find
(a) What will be the average number of unloading group of workers working at any time?
(b) What is the probability that more than 4 groups of workers are needed?

Solution:

Let us assume that there are 'n' vehicles waiting in line at any time. Now service rate is dependent on waiting length hence $\mu_n = 2n$ vehicles per day (when there are 'n' groups of workers in the system).

Now $\lambda = 3$ vehicles per day and $\mu = 2$ vehicles per day. (With one unloading labour group)

Hence, $p_n = (\rho^n/n!) \times e^{-\rho}$ for $n \geq 0$

Therefore, expected number of group of workers working any specified instant is

$$E(n) = \sum_{n=0}^{\infty} n \times p_n = \Sigma n \times \left(\rho^n e^{-\rho}\right)/n! = \rho \times e^{-\rho} \times \sum_{n=1}^{\alpha} \left(\rho^{n-1}\right)/(n-1) = (\lambda/\mu) = 1.5 \text{ labour group.}$$

The probability that the vehicle entering in service will require more than four groups of workers is

$$\sum_{n=5}^{\infty} p_n = 1 - \sum_{n=0}^{4} (\rho^n/n!)\, e^{-p} = 0.019$$

11 Model-III: $(M/M/1:N/FCFS)$ (Finite Queue Length Model)

This model differs from the above model in the sense that the maximum number of customers in the system is limited to N. Therefore the equations of above model is valid for this model as long as $n \leq N$ and arrivals will not exceed N under any circumstances. The various equations of the model is:

1. $p_0 = (1-\rho)/\left(1-\rho^{N+1}\right)$, where $\rho = \lambda/\mu$ and $\lambda/\mu > 1$ is allowed.
2. $p_{\mathrm{n}} = (1-\rho)\rho^n/\left(1-\rho^{N+1}\right)$ for all $n = 0,1,2,\ldots.N$
3. Average queue length $E(n) = \rho\left[1-(1+N)\rho^N + Np^{N+1}\right]/(1-\rho)\left(1-\rho^{N+1}\right)$

$$= \left[(1-\rho)/\left(1-\rho^{N+1}\right)\right] \times \sum_{n=0}^{n} n\rho^n = p_0 \times \sum_{n=0}^{N} n \times \rho^n$$

4. The average length of the waiting line $= E(L) = \left[1-N\rho^{N+1}+(N-1)\rho^N\right]\rho^2/(1-\rho)\left(1-\rho^{N+1}\right)$
5. Waiting time in the system $= E(v) = E(n)/\lambda'$ where $\lambda' = \lambda\left(1-\rho_N\right)$
6. Waiting time in the queue $= E(w) = E(L)/\lambda' = [(E(n)/\lambda'/(1/\mu)]$

Solved Example Problem

In a railway marshalling yard, good train arrives at the rate of 30 trains per day. Assume that the inter arrival time follows an exponential distribution and the service time is also to be assumed as exponential with a mean of 36 minutes. Calculate
(a) The probability that the yard is empty,
(b) The average length assuming that the line capacity of the yard is 9 trains.

Solution:

Data: $\lambda = 30/(60\times 24) = 1/48$ trains per minute. And $\mu = 1/16$ trains per minute.

Therefore $\rho = (\lambda/\mu) = 36/48 = 0.75$

(a) The probability that the queue is empty is given by $= p_0 = (1-\rho)/\left(1-\rho^{N+1}\right)$, where $N = 9$
$(1-0.75)/[1-(0.75)^{9+1}] = 0.25/0.90 = 0.28$. i.e. 28 % of the time the line is empty.

(b) Average queue length is $= \left[(1-\rho)/\left(1-\rho^{N+1}\right)\right] \times \sum_{n=0}^{N} n\rho^n$
$\left[(1-0.75)/\left(1-0.75^{10}\right)\right] \times \sum_{n=0}^{9} n(0.75)^n = 0.28 \times 9.58 = 3$ trains.

12 Model-IV: $(M/M/S):FCFS/N/N$ (Limited Popultion or Source Model)

In this model, we assume that customers are generated by limited pool of potential customers i.e. finite population. The total customer's population is M and n represents the number of customers already in the system (waiting line), any arrival must come from $M-n$ number that is not yet in the system. The formulae for this model are:

$p_0 = 1/\sum_{n=0}^{M}[M!l(M-n)!](\lambda/\mu)^n$

$p_n = [M](M-n)!] \times (\lambda/\mu)^n \times p_0 = \left\{[M|^n(M-n)!] \times (\lambda/\mu)^n\right\}/\sum_{n=0}^{M} M!/(M-n)! \times (\lambda/\mu)^n\}$

Average number of customers in the system $= E(n) = \sum_{n=0}^{M} np_n = M - (\mu/\lambda)(1-p_0)$

Average number in the queue $= E(L) = M - [(\mu+\lambda)/\lambda] \times (\mathrm{d}-p_0)$

Solved Example Problem

A mechanic repairs 4 machines. The mean time between service requirements is 5 hours for each machine and forms an exponential distribution. The mean repair time is 1 hour and also follows the same distribution pattem. Machine down time costs Rs. 25/ per hour and the mechanic costs Rs. 55/. per day. Find (a) Expected number of operating machines, (b) the expected down time cost per day, (c) Would it be economical to engage two mechanics, each repairing only two machines?

Solution

Data: Finite population, λ = Amival rate $= (1/5) = 0.2, \mu$ = Service rate $= \mu = (1/1) = 1$ Probability of the empty system $= p_0 =$

$$p_0 = 1/\sum_{n=0}^{4}[4/(4-n)](0.2/1)^n$$

$= 1/1 + (4\times 0.2) + \left(4\times 3\times 0.2^2\right) + \left(4\times 3\times 2\times 0.2^3\right) + \left(4\times 3\times 2\times 1\times 0.2^4\right) = 0.4$ ie. 40 percent of the time the system is empty and 60 percent of the time the system is busy.

(a) Expected number of breakdown machines I the system $= E(n) = M - (\mu/\lambda)(1-p_0) = 4 - (1/0.2)(1 - 0.4) = 4 - 5\times 0.6 = 4 - 3 = 1$. i.e. Expected mumber of operating machines in the system $= 4 - 1 = 3$

(b) Expected down time cost per day of 8 hours = 8 $\times$ (expected number of breakdown machines x Rs. 25 per hour) $= 8\times 1\times 25 =$ Rs. 200/ - day.

(c) When there are two mechanics each serving two machines, $M = 2, p_0 =$

$$p_0 = 1/\sum_{n=0}^{2}[21/(2-n)!](0.2/1)^n = 1/1 + (2\times 0.2) + \left(2\times 1\times 0.2^2\right) = 1/1.48 = 0.68$$ i.e. 68 percent

of the time the system is idle. It is assumed that each mechanic with his two machines constitutes a separate system with no interplay.

Expected number of machines in the system = $M - (\mu/\lambda)\times(1/p_0) = 2 - (1/0.2)\times(1 - 0.68) = 0.4$

Therefore expected down time per day $= 8\times 0.4\times$ Number of mechanics or machine in system $= 8\times 0.4\times 2 =$ 6.4 hours per day. Hence total cost involved $=$ Rs. $55\times 2 + 6.4\times$ Rs $\cdot 25/- =$ Rs $\cdot (110 + 160) =$ Rs.270 per day.

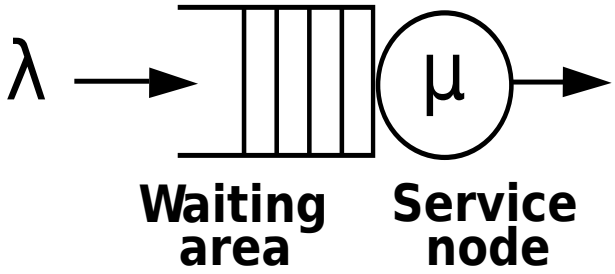
But total cost with one mechanic is Rs. (55 +200) = Rs. 255/ - per day, which is cheaper compared to the above. Hence use of two mechanics is not advisable.

13 Steady-state solutions of Markovian queuing models of M/M/1

An $M/M/1$ queue represents the queue length in a system having a single server, where arrivals are determined by a Poisson process and job service times have an exponential distribution. The model name is written in Kendall's notation. The model is the most elementary of queueing models and an attractive object of study as closed-form expressions can be obtained for many metrics of interest in this model. An extension of this model with more than one server is the $M/M/c$ queue.

An $M/M/1$ queue is a stochastic process whose state space is the set $0, 1, 2, 3, ...$ where the value corresponds to the number of customers in the system, including any currently in service.

1. Arrivals occur at rate λ according to a Poisson process and move the process from state i to $i+1$.
2. Service times have an exponential distribution with rate parameter μ in the $M/M/1$ queue, where $1/\mu$ is the mean service time.
3. A single server serves customers one at a time from the front of the queue, according to a first-come, first-served discipline. When the service is complete the customer leaves the queue and the number of customers in the system reduces by one.



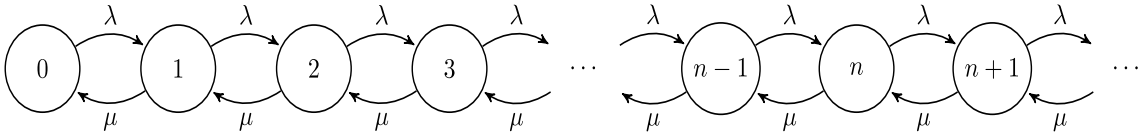
Source: By Tsaitgaist - Mm1.png, CC BY-SA 3.0
Figure 5: An M/M/1 queueing node

4. The buffer is of infinite size, so there is no limit on the number of customers it can contain.

The model can be described as a continuous time Markov chain with transition rate matrix

$$Q = \begin{pmatrix} -\lambda & \lambda & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & \\ & \mu & -(\mu+\lambda) & \lambda & & \\ & & \mu & -(\mu+\lambda) & \lambda & \\ & & & & \ddots & \end{pmatrix}$$

on the state space $0, 1, 2, 3, \ldots$. This is the same continuous time Markov chain as in a birth–death process. The state space diagram for this chain is as below.



Source: By Gareth Jones - Own work, CC BY-SA 3.0
Figure 6: State space diagram of an $M/M/1$ queue. The state space records the number of customers in the queueing system. The values lambda and mu represent the arrival and service rates of customers.

Transient solution

We can write a probability mass function dependent on t to describe the probability that the $M/M/1$ queue is in a particular state at a given time. We assume that the queue is initially in state i and write $pk(t)$ for the probability of being in state k at time t. Then

$$p_k(t) = e^{-(\lambda+\mu)t}\left[\rho^{\frac{k-i}{2}} I_{k-i}(at) + \rho^{\frac{k-i-1}{2}} I_{k+i+1}(at) + (1-\rho)\rho^k \sum_{j=k+i+2}^{\infty} \rho^{-j/2} I_j(at)\right]$$

where $\rho = \lambda/\mu$, $a = 2\sqrt{\lambda\mu}$ and I_k is the modified Bessel function of the first kind. Moments for the transient solution can be expressed as the sum of two monotone function's.

14 Multi Channel Queuing Model: M / M / c: (∝ / FCFS)

The above symbols indicate a system with Poisson input and Poisson output with number of channels $= c$, where c is ≥ 1, the capacity of line is infinite and first come first served discipline. Here the length of waiting line depends on the number of channels engaged. In case the number of customers in the system is less than the number of channels i.e. $n \leq c$, then there will be no problem of waiting and the rate of servicing will be $n\mu$ as only n channels are busy, each servicing at the rate m. In case $n = c$, all the channels will be working and when $n \geq c$, then $n - c$ elements will be in the waiting line and

the rate of service will be $c\mu$ as all the $c$channels are busy. Various formulae we have to use in this type of models are:

$$p_0 = 1/\sum_{n=0}^{c-1}[(\lambda/\mu)^n/n!] + [(\lambda/\mu)^c/c!] \times [(c\mu/c\mu-\lambda)]$$

$$OR = 1/\left[\sum_{n=0}^{c-1}(c\rho)^n/n!\right] + [(c\rho)^c/c!(1-\rho)]$$
$$p_n = \{[(\lambda/\mu)^n/n!]/n!\} \times p_0, \text{ when } 1 \le n \le c$$
$$p_n = [1/(c^{n-c} \times c!)](\lambda/\mu)^n \times p_0 \text{ when } n \ge c$$

Average number of units in waiting line of the system $= E(n) = \left[\rho p_c/(1-\rho)^2\right]$

$$= \left\{[\lambda, \mu.(\lambda/\mu)^c]/\left[(c-1)!(c\mu-\lambda)^2\right]\right\} p_0 + (\lambda/\mu)$$

Average number in the queue $= E(L) = \left[pc\rho/(1-\rho)^2\right] + c\mu = \left\{[\lambda, \mu.(\lambda/\mu)^c]/\left[(c-1)!(c\mu-\lambda)^2\right]\right\} p_0$

Average queue length = Average number of units in waiting line + number of units in service Average waiting time of an arrival

$= E(w) =$ (Average number of units in waiting line) $/\lambda = \left[(p_c\rho)/\lambda(1-\rho)^2 = \left[\rho/\lambda(1-\rho)^2 \times (1/c!) \times (\lambda/\mu)^c \times p_0\right.\right.$

$$= E(L)/\lambda = \left\{[\mu \times (\lambda/\mu)^c]/\left[(c-1)!(c\mu-\lambda)^2\right]\right\} \times p_0$$

Average time an arrival spends in system $= E(v) =$ (Average number of items in the queue) $/\lambda = [(p_c\rho)/\lambda(1-\rho)] + (C\mu/\lambda)$

$$E(n)/\lambda = \left\{[\mu \times (\lambda/\mu)^c]/\left[(c-1)!(c\mu-\lambda)^2\right]\right\} \times p_0 + (1/\mu)$$

Probability that all the channels are occupied $= p(n \ge c) = [1/(1-\rho)]p_c$

$$= [\mu \times (\lambda/\mu)^c]\, p_0/[(c-1)!(c\mu-\lambda)]$$

Probability that some units has to wait $= p(n \ge c+1) = [\rho p_c/(1-\rho)]$

$$= 1 - p(n \ge c) = 1 - [\mu \times (\lambda/\mu)^c]\, p_0/[(c-i)!(c\mu-\lambda)]$$

The average number of units which actually wait in the system $=$

$$\left[\sum_{n=c+1}^{\infty}(n-c)p_n\right] \div \sum_{n=c+1}^{\infty} p_n = 1/(1-\rho)$$

Average waiting time in the queue for all arrivals $= (1/\lambda)\sum_{n=0}^{\infty}(n-c)p_n = p_c/c\mu(1-\rho)^2$

Average waiting time in queue for those who acutely wait $= 1/(c\mu-\lambda)$

Average number of items served $= \sum_{n=0}^{c-1} np_n + \sum_{n=c}^{\infty} p_n$

Average number of idle channels $= c -$Average number of items served

Efficiency of MM/c model: = (Average number of items served) / (Total number of channels)

Utilization factor $= \rho = (\lambda/c\mu)$

Solved Example Problem

A telephone exchange has two long distance operators. The telephone company finds that during the peak load, long distance calls arrive in a Poisson fashion at an average rate of 15 per hour. The length of service on these calls is approximately exponentially distributed with mean length 5 minutes. What is the probability that a subscriber will have to wait for his long distance call during the peak hours of the day? If subscribers wait and are serviced in turn, what is the expected waiting time.

Solution: Data: $\lambda = 15$ calls per hour, $\mu = 60/5 = 12$ calls per hour.

Therefore $\rho = (15)/(2\times 12) = 5/\ 8. p_0 =$

$1/\left[\sum_{n=0}^{c-1}(c\rho)^n/n!\right] + [(c\rho)^c/c!(1-\rho)] = 1/1 + (5/4) + (1/2)\times(25/16)\times[1/(1-5/8)] = (12/52)$

(a) Probability that a subscriber has to wait $= p(n \geq 2) = 1 - p_0 - p_1 = [1-(12/52)-(15/32)] = 25/52 -$ 0.48. i.e. 48% of the time the subscriber has to wait.

Expected waiting time $= E(w) = \left[p/\lambda(1-\rho)^2\right] \times (1/cl) \times (\lambda/\mu)^c \times p_c$

$$= \left\{(5/8)/15[1-(5/8)]^2 \times (1/2!) \times (15/12)^2 \times (12/32) \text{ hours } = 3.2 \text{ minutes.}\right.$$

15 $M/G/1$ with Limited Waiting Spaces

An $M/G/1$ queue is a queue model where arrivals are Markovian (modulated by a Poisson process), service times have a General distribution and there is a single server. The model name is written in Kendall's notation, and is an extension of the $M/M/1$ queue, where service times must be exponentially distributed. The classic application of the $M/G/1$ queue is to model performance of a fixed head hard disk.

Model definition

A queue represented by a $M/G/1$ queue is a stochastic process whose state space is the set $0,1,2,3...$, where the value corresponds to the number of customers in the queue, including any being served. Transitions from state i to $i+1$ represent the arrival of a new customer: the times between such arrivals have an exponential distribution with parameter λ. Transitions from state i to $i-1$ represent a customer who has been served, finishing being served and departing: the length of time required for serving an individual customer has a general distribution function. The lengths of times between arrivals and of service periods are random variables which are assumed to be statistically independent.

Queue length

Pollaczek–Khinchine method: The probability generating function of the stationary queue length distribution is given by the Pollaczek–Khinchine transform equation.

where $g(s)$ is the Laplace transform of the service time probability density function. In the case of an $M/M/1$ queue where service times are exponentially distributed with parameter μ, $g(s) = \mu/(\mu+s)$.

$$\pi(z) = \frac{(1-z)(1-\rho)g(\lambda(1-z))}{g(\lambda(1-z))-z}$$

This can be solved for individual state probabilities either using by direct computation or using the method of supplementary variables. The Pollaczek–Khinchine formula gives the mean queue length and mean waiting time in the system.

Matrix analytic method

Consider the embedded Markov chain of the $M/G/1$ queue, where the time points selected are immediately after the moment of departure. The customer being served (if there is one) has received zero seconds of service. Between departures, there can be $0,1,2,3,...$ arrivals. So from state i the chain can move to state $i-1, i, i+1, i+2, ...$ The embedded Markov

chain has transition matrix

$$P = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where

$$a_v = \int_0^\infty e^{-\lambda u} \frac{(\lambda u)^v}{v!} \mathrm{d}F(u) \text{ for } v \geq 0$$

and $F(u)$ is the service time distribution and λ the Poisson arrival rate of jobs to the queue.

Markov chains with generator matrices or block matrices of this form are called $M/G/1$ type Markov chains, a term coined by M. F. Neuts. The stationary distribution of an $M/G/1$ type Markov model can be computed using the matrix analytic method.

Busy period

The busy period is the time spent in states $1,2,3,...$ between visits to the state0. The expected length of a busy period is $1/(\mu-\lambda)$ where $1/\mu$ is the expected length of service time and λ the rate of the Poisson process governing arrivals. The busy period probability density function $\phi(s)$ can be shown to obey the Kendall functional equation

$$\phi(s) = g[s+\lambda-\lambda\phi(s)]$$

where as above g is the Laplace–Stieltjes transform of the service time distribution function. This relationship can only be solved exactly in special cases (such as the $M/M/1$ queue), but for any s the value of $\phi(s)$ can be calculated and by iteration with upper and lower bounds the distribution function numerically computed.

Waiting/response time

Writing $W^*(s)$ for the Laplace–Stieltjes transform of the waiting time distribution, is given by the Pollaczek–Khinchine transform as

$$W^*(s) = \frac{(1-\rho)sg(s)}{s-\lambda(1-g(s))}$$

where g("s") is the Laplace–Stieltjes transform of service time probability density function.

Sample Multiple Choice Questions

1. As per queue discipline the following is not a negative behaviour of a customer:

   a) Balking b) Reneging c) Boarding d) Collusion

2. The system of loading and unloading of goods usually follows:

   a) LIFO b) FIFO c) SIRO d) SBP

3. In (M /M /1) : ($\propto$ / FCFS) model, the length of the system L_s is given by:

   a) $\rho^2/1/\rho$ b) $\rho/1-\rho$ c) $\lambda^2/(\mu-\lambda)$ d) $\lambda^2/\mu(\mu-\lambda)$

4. This department is responsible for the development of queuing theory:

   a) Railway station b) Municipal office c) Telephone department d) Health department

5. When the operating characteristics of the queue system dependent on time, then it is said to be:

   a) Steady state b) Explosive state c) Transient state d) Any one of the above

Answer the following questions

1. What are the main characteristics of Parallel queues
2. What are three axioms of Poisson distribution
3. What are the assumptions of $(M/M/1:\infty/FCFS)$ model
4. What is Expected (or average) queue length or expected number of customers waiting in the queue
5. What is the probability that waiting time is more than t.

References

[1] P. Ramamurthy, "Operations Research", New Age International; 2007

[2] System Engineering 3(3+0), http://ecoursesonline.iasri.res.in/mod/page/view.php?id=2970, 2013

[3] Single queueing nodes, https://en.wikipedia.org/wiki/M/M/1_queue, 2020

[4] Queueing models, https://en.wikipedia.org/wiki/M/G/1_queue,

[5] Bibhas C. Giri, Chapter 6:Queueing Theory, http://epgp.inflibnet.ac.in/, 2020

****