Queuing Theory

Introduction

Waiting lines are the most frequently encountered problems in everyday life. For example, queue at a cafeteria, library, bank, etc. Common to all of these cases are the arrivals of objects requiring service and the attendant delays when the service mechanism is busy. Waiting lines cannot be eliminated completely, but suitable techniques can be used to reduce the waiting time of an object in the system. A long waiting line may result in loss of customers to an organization. Waiting time can be reduced by providing additional service facilities, but it may result in an increase in the idle time of the service mechanism.

Definition

Queuing theory is based on mathematical theories and deals with the problems arising due to flow of customers towards the service facility.

The waiting line models help the management in balancing between the cost associated with waiting and the cost of providing service. Thus, queuing or waiting line models can be applied in such situations where decisions have to be taken to minimize the waiting time with minimum investment cost.

Basic Terminology

The present section focuses on the standard vocabulary of Waiting Line Models.

Queuing Model

It is a suitable model used to represent a service oriented problem, where customers arrive randomly to receive some service, the service time being also a random variable.

Arrival

The statistical pattern of the arrival can be indicated through the probability distribution of the number of the arrivals in an interval.

Service Time

The time taken by a server to complete service is known as service time.

Server

It is a mechanism through which service is offered.

Queue Discipline

It is the order in which the members of the queue are offered service.

Poisson Process

It is a probabilistic phenomenon where the number of arrivals in an interval of length t follows a Poisson distribution with parameter λt, where λ is the rate of arrival.

Queue

A group of items waiting to receive service, including those receiving the service, is known as queue.

Waiting time in queue

Time spent by a customer in the queue before being served.

Waiting time in the system

It is the total time spent by a customer in the system. It can be calculated as follows:

Waiting time in the system = Waiting time in queue + Service time

Queue length

Number of persons in the system at any time.

Average length of line

The number of customers in the queue per unit of time.

Average idle time

The average time for which the system remains idle.

FIFO

It is the first in first out queue discipline.

Bulk Arrivals

If more than one customer enter the system at an arrival event, it is known as bulk arrivals.
Please note that bulk arrivals are not embodied in the models of the subsequent sections.

Components of Queuing System

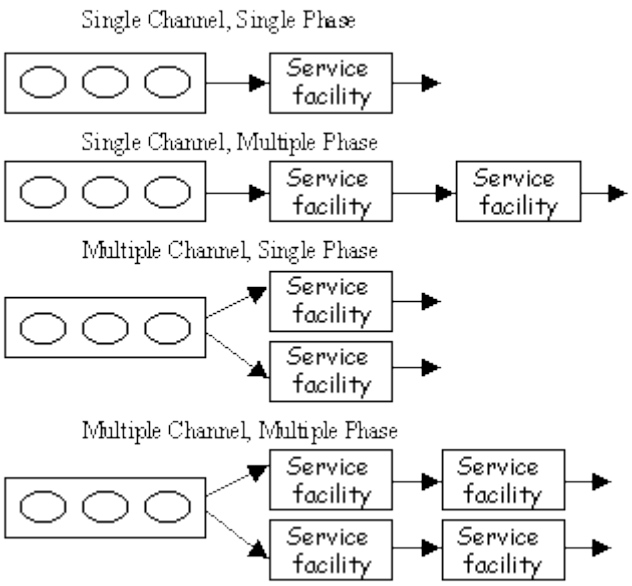
1. **Input Source:** The input source generates customers for the service mechanism. The most important characteristic of the input source is its size. It may be either finite or infinite. Please note that the calculations are far easier for the infinite case, therefore, this assumption is often made even when the actual size is relatively large. If the population size is finite, then the analysis of queuing model becomes more involved.
   The statistical pattern by which calling units are generated over time must also be specified. It may be Poisson or Exponential probability distribution.

   Usually the source population is considered as unlimited.

2. **Queue:** It is characterized by the maximum permissible number of units that it can contain. Queues may be infinite or finite.
3. **Service Discipline:** It refers to the order in which members of the queue are selected for service. Frequently, the discipline is first come, first served.
   Following are some other disciplines:
   - LIFO (Last In First Out)
   - SIRO (Service In Random Order)
   - Priority System

4. **Service Mechanism:** A specification of the service mechanism includes a description of time to complete a service and the number of customers who are satisfied at each service event. The service mechanism also prescribes the number and configuration of servers. If there is more than one service facility, the calling unit may receive service from a sequence of these. At a given facility, the unit enters one of the parallel service channels and is completely serviced by that server. Most elementary models assume one service facility with either one or a finite number of servers.The following figure shows the physical layout of service facilities.



Unusual Customer/Server Behaviour

Customer's Behaviour

- **Balking.** A customer may not like to join the queue due to long waiting line.
- **Reneging.** A customer may leave the queue after waiting for sometime due to **impatience**.


"My patience is now at an end." - *Hitler*

- **Collusion.** Several customers may cooperate and only one of them may stand in the queue.
- **Jockeying.** When there are a number of queues, a customer may move from one queue to another in hope of receiving the service quickly.

Server's Behaviour

- **Failure.** The service may be interrupted due to failure of a server (machinery).
- **Changing service rates.** A server may speed up or slow down, depending on the number of customers in the queue. For example, when the queue is long, a server may speed up in response to the pressure. On the contrary, it may slow down if the queue is very small.
- **Batch processing.** A server may service several customers simultaneously, a phenomenon known as batch processing.

- The source population has infinite size.

- The inter-arrival time has an exponential probability distribution with a mean arrival rate of l customer arrivals per unit time.
- There is no unusual customer behaviour.
- The service discipline is FIFO.
- The service time has an exponential probability distribution with a mean service rate of m service completions per unit time.
- The mean arrival rate is less than the mean service rate, i.e., l < m.
- There is no unusual server behaviour.

In this section and the subsequent sections of this chapter, we explain several models. In presenting the models below, we start slowly and provide several examples, so that you can acquire a better feeling for waiting line models. Be patient and give yourself plenty of time to study these sections; otherwise, you may easily get confused.

The M/M/1 (∝/FIFO) system

It is a queuing model where the arrivals follow a Poisson process, service times are exponentially distributed and there is only one server. In other words, it is a system with Poisson input, exponential waiting time and Poisson output with single channel.

Queue capacity of the system is infinite with first in first out mode. The first M in the notation stands for Poisson input, second M for Poisson output, 1 for the number of servers and ∝ for infinite capacity of the system.

Formulas

Probability of zero unit in the queue (P_o) = $1 - \frac{\lambda}{\mu}$

Average queue length (L_q) = $\frac{\lambda^2}{\mu(\mu - \lambda)}$

Average number of units in the system (L_s) = $\frac{\lambda}{\mu - \lambda}$

Average waiting time of an arrival (W_q) = $\frac{\lambda}{\ldots}$

$$\mu(\mu - \lambda)$$

Average waiting time of an arrival in the system (W_s) =

$$\frac{1}{\mu - \lambda}$$

Example 1

Students arrive at the head office of www.universalteacher.com according to a Poisson input process with a mean rate of 40 per hour. The time required to serve a student has an exponential distribution with a mean of 50 per hour. Assume that the students are served by a single individual, find the average waiting time of a student.

Solution.

Given
$\lambda = 40$/hour, $\mu = 50$/hour

Average waiting time of a student before receiving service (W_q) =

$$\frac{40}{50(50 - 40)} = 4.8 \text{ minutes}$$

Example 2

New Delhi Railway Station has a single ticket counter. During the rush hours, customers arrive at the rate of 10 per hour. The average number of customers that can be served is 12 per hour. Find out the following:

- Probability that the ticket counter is free.
- Average number of customers in the queue.

Solution.

Given
$\lambda = 10$/hour, $\mu = 12$/hour

Probability that the counter is free = $1 - \frac{10}{12} = 1/6$

Average number of customers in the queue (L_q) = $\frac{(10)^2}{12\,(12 - 10)} = 25/6$

Example 3

At Bharat petrol pump, customers arrive according to a Poisson process with an average time of 5 minutes between arrivals. The service time is exponentially distributed with mean time = 2 minutes. On the basis of this information, find out

1. What would be the average queue length?
2. What would be the average number of customers in the queuing system?
3. What is the average time spent by a car in the petrol pump?
4. What is the average waiting time of a car before receiving petrol?

Solution.

Average inter arrival time = $\frac{1}{\lambda} = 5\text{minutes} = \frac{1}{12}$ hour

$\lambda = 12/\text{hour}$

Average service time = $\frac{1}{\mu} = 2\text{ minutes} = \frac{1}{30}$ hour

$\mu = 30/\text{hour}$

Average queue length, L_q = $\frac{(12)^2}{30(30 - 12)} = \frac{4}{15}$

Average number of customers, L_s = $\frac{12}{30 - 12} = \frac{2}{}$

3

Average time spent at the petrol pump = $\frac{1}{30 - 12}$ = 3.33 minutes

Average waiting time of a car before receiving petrol = $\frac{12}{30(30 - 12)}$ = 1.33 minutes

Example 4

Universal Bank is considering opening a drive in window for customer service. Management estimates that customers will arrive at the rate of 15 per hour. The teller whom it is considering to staff the window can service customers at the rate of one every three minutes.

Assuming Poisson arrivals and exponential service find

1. Average number in the waiting line.
2. Average number in the system.
3. Average waiting time in line.
4. Average waiting time in the system.

Solution.

Given
$\lambda = 15$/hour,
$\mu = 3/60$ hour
or 20/hour

Average number in the waiting line = $\frac{(15)^2}{20(20 - 15)}$ = 2.25 customers

Average number in the system = $\frac{15}{20 - 15}$ = 3 customers

$$\text{Average waiting time in line} = \frac{15}{20(20 - 15)} = 0.15 \text{ hours}$$

$$\text{Average waiting time in the system} = \frac{1}{20 - 15} = 0.20 \text{ hours}$$

Example 5

Chhabra Saree Emporium has a single cashier. During the rush hours, customers arrive at the rate of 10 per hour. The average number of customers that can be processed by the cashier is 12 per hour. On the basis of this information, find the following:

- Probability that the cashier is idle
- Average number of customers in the queuing system
- Average time a customer spends in the system
- Average number of customers in the queue
- Average time a customer spends in the queue

Solution.

Given
$\lambda = 10$/hour, $\mu = 12$/hour

$$P_o = 1 - \frac{10}{12} = 1/6$$

$$L_s = \frac{10}{12 - 10} = 5 \text{ customers}$$

$$W_s = \frac{1}{12 - 10} = 30 \text{ minutes}$$

$$L_q = \frac{(10)^2}{12(12 - 10)} = 25/6 \text{ customers}$$

$$W_q = \frac{10}{12(12 - 10)} = 25 \text{ minutes}$$

The M/M/1 (N/FIFO) system

It is a queuing model where the arrivals follow a Poisson process, service times are exponentially distributed and there is only one server. Capacity of the system is limited to N with first in first out mode.

The first M in the notation stands for Poisson input, second M for Poisson output, 1 for the number of servers and N for capacity of the system.

$\rho = \lambda/\mu$

$$P_o = \frac{1-\rho}{1-\rho^{N+1}}$$

$$L_s = \frac{\rho}{1 - \rho} - \frac{(N + 1)\rho^{N+1}}{1-\rho^{N+1}}$$

$$L_q = L_s - \lambda/\mu$$

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

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

Example

Students arrive at the head office of www.universalteacher.com according to a Poisson input process with a mean rate of 30 per day. The time required to serve a student has an exponential distribution with a mean of 36 minutes. Assume that the students are served by a single individual, and queue capacity is 9. On the basis of this information, find the following:

- The probability of zero unit in the queue.
- The average line length.

Solution.

$$\lambda = \frac{30}{60 \text{ X } 24}$$

= 1/48 students per minute

μ = 1/36 students per minute

ρ = 36/48 = 0.75

N = 9

$$P_o = \frac{1- 0.75}{1- (0.75)^{9+1}}$$

= 0.26

$$L_s = \frac{0.75}{1 - 0.75} - \frac{(9 + 1)(0.75)^{9+1}}{\text{----------}}$$

$$1 - (0.75)^{9+1}$$

= 2.40 or 2 students.

The M/M/C (∝/FIFO) system

It is a queuing model where the arrivals follow a Poisson process, service times are exponentially distributed and there are C servers.

$$\frac{1}{P_0} = \sum_{n=0}^{C-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!} \times \frac{1}{1-\rho}$$

$$\text{Where } \rho = \frac{\lambda}{c\mu}$$

$$L_q = P_0 \times \frac{(\lambda/\mu)^c}{c!} \times \frac{\rho}{(1-\rho)^2}$$

$$W_q = \frac{1}{\lambda} \times L_q$$

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

$$L_s = L_q + \frac{\lambda}{\mu}$$

Example 1

The Silver Spoon Restaurant has only two waiters. Customers arrive according to a Poisson process with a mean rate of 10 per hour. The service for each customer is exponential with mean of 4 minutes. On the basis of this information, find the following:

- The probability of having to wait for service.
- The expected percentage of idle time for each waiter.

Solution.

This is an example of M/M/C, where c = 2

λ = 10 per hour or 1/6 per minute.
μ = 1/4 per minute
ρ = 1/3

$$\frac{1}{P_0} = \sum_{n=0}^{C-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!} \times \frac{1}{1-\rho}$$

$$\frac{1}{P_0} = \sum_{n=0}^{1} \frac{(2/3)^n}{n!} + \frac{(2/3)^2}{2!} \times \frac{1}{1-1/3}$$

$$\frac{1}{P_0} = 1 + \frac{2}{3} + \frac{1}{3}$$

$$P_0 = \frac{1}{2}$$

The expected percentage of idle time for each waiter.

$1 - \rho =$ $1 - 1/3 = 2/3 = 67\%$

Example 2

Universal Bank has two tellers working on savings accounts. The first teller handles withdrawals only. The second teller handles deposits only. It has been found that the service times distributions for both deposits and withdrawals are exponential with mean service time 2 minutes per customer. Deposits & withdrawals are found to arrive in a Poisson fashion with mean arrival rate 20 per hour. What would be the effect on the average waiting time for depositors and withdrawers, if each teller could handle both withdrawers & depositors?

Solution.

Given
$\lambda = 20$ per hour or $1/3$ per minute, $\mu = 1/2$ per minute, $c = 2$

Case I - Treating depositors and withdrawers as unit of M/M/1 system.

Average waiting time of an arrival (W_q) =

$$\frac{\lambda}{\mu(\mu - \lambda)}$$

$$W_q = \frac{1/3}{1/2\,(1/2 - 1/3)} = 4 \text{ minutes}$$

Case II - If each teller handles both depositors and withdrawers.

$P_0 = 1/2$

$L_q = 1/12$

$$W_q = \frac{1}{\lambda} \times L_q$$

$W_q =$ 1/4 minutes

Hence, when both tellers handle both withdrawals & deposits, then expected waiting time is reduced.

The $M/E_k/1$ (∞/FIFO) system

It is a queuing model where the arrivals follow a Poisson process, service time follows an Erlang (k) probability distribution and the number of server is one.

Queue capacity of the system is infinite with first in first out mode. The first M in the notation stands for Poisson input, k for number of phases, 1 for the number of servers and ∞ for infinite capacity of the system.

An Erlang Family (E_k)

E_k of a probability distribution is the probability distribution of a random variable, which can be expressed as the sum 'k' independently, identically distributed exponential variables.

The expected numbers of customers in the queue, L_q =

$$\frac{(1+k)}{2k} \times \frac{\lambda^2}{\mu(\mu-\lambda)}$$

The expected waiting time before being served, W_q =

$$\frac{(1+k)}{2k} \times \frac{\lambda}{\mu(\mu-\lambda)}$$

The expected time spent in the system, W_s =

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

The expected numbers of customers in the system, L_s =

$$\lambda W_s$$

Example 1

The registration of a student at www.universalteacher.com requires three steps to be completed sequentially. The time taken to perform each step follows an exponential distribution with mean 30/3 minutes and is independent of each other. Students arrive at the head office according to a Poisson input process with a mean rate of 25 per hour. Assuming that there is only one person for registration. On the basis of this information, find the following:

a. expected waiting time
b. expected numbers of students in the queue.

Solution.

This is an $M/E_k/1$ system.

Here $k = 3$, $\lambda = 25$ per hour.

$$\text{Service time per phase} = \frac{1}{3\mu} = \frac{30}{3}$$

Therefore, $\mu = 30$ per hour.

$$\text{The expected numbers of students in the queue, } L_q = \frac{1+3}{2 \times 3} \times \frac{(25)^2}{30(30-25)} = 2.78 \text{ students or 3 students}$$

$$\text{The expected waiting time before being served, } W_q = \frac{1+3}{2 \times 3} \times \frac{25}{30(30-25)} = 1/9 \text{ hour or 6.67 minutes}$$

Example 2

Repair of a certain type of machine requires three steps to be completed sequentially. The time taken to perform each step follows an exponential distribution with mean 20/3 minutes and is

independent of each other. The machine breakdown follows a Poisson process with rate of 1per 2 hours. Assuming that there is only one repairman, find out

a. The expected idle time of a machine.
b. The average waiting time of a broken down machine in a queue.
c. The expected number of broken down machines in the queue.
d. The average number of machines which are not in operation

Solution.

This is an $M/E_k/1$ system.

Here $k = 3$, $\lambda = 1/2$ per hour.

Service time per phase $= \frac{1}{3\mu} = \frac{20}{3}$

Therefore, $\mu = 3$ per hour.

The expected numbers of customers in the queue, $L_q = \frac{1+3}{2 \times 3} \times \frac{(1/2)^2}{3(3 - 1/2)} = 1.33$ minutes

The expected waiting time before being served, $W_q = \frac{1+3}{2 \times 3} \times \frac{1/2}{3(3 - 1/2)} =$ 2 minutes 40 seconds

The expected time spent in the system, $W_s = \frac{2}{45} + \frac{1}{3} =$ 22 minutes 40 seconds

The expected numbers of customers in the system, $L_s = \frac{1}{2} \times \frac{17}{45} =$ 11.33 minutes

As k approaches to infinity, the expressions of L_q, W_q, W_s and L_s are given by

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

$$W_q = \frac{\lambda}{2\mu(\mu - \lambda)}$$

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

$$L_s = L_q + \frac{\lambda}{\mu}$$

Example 3

At Indira Gandhi airport, it takes exactly 6 minutes to land an aeroplane, once it is given the signal to land. Although incoming planes have scheduled arrival times, the wide variability in arrival times produces an effect which makes the incoming planes appear to arrive in a Poisson fashion at an average rate of 6 per hour. This produces occasional stack-ups at the airport, which can be dangerous and costly. Under these circumstances, how much time will a pilot expect to spend circling the field waiting to land?

Solution.

Here, service time is fixed being equal to 6 minutes.
The service distribution is last member of Erlang family, i.e., for $k = \infty$
Mean arrival rate of aeroplanes, $\lambda = 6$ per hour
Mean landing rate of planes, $\mu = (1/6) \times 60 = 10$ per hour

$$W_q = \frac{6}{2 \times 10 \times (10 - 6)} = 3/40 \text{ hours} = 4.5 \text{ minutes}$$