Chapter Two

Introduction

A common phenomenon occurring in everyday life is that of queuing or the waiting line. A queue is formed when either units requiring services (customers) wait for service or the service facilities stand idle and wait for customers. Some customers wait when the total number of customers requiring service exceeds the number of service facilities. Some service facilities stand idle when the number of service facilities exceeds the number of customers. If waiting lines cannot be economically completely eliminated, they can at least be reduced by optimizing the number of service stations or by adjusting service times in one or more stations.

Queuing theory is a valuable tool in the solution of queuing problems. For better understanding of queues, this theory attempts to formulate, interpret and predict in order to introduce remedies such as adequate service with tolerable waiting.

Queuing theory provides models that are capable of influencing arrival patterns of customers or determining the most appropriate amount of service or number of service stations. Queuing models essentially relate to study of the behavior of waiting lines and attempt to strike a balance between server utilization and customer waiting time.

Queuing theory originated from efforts made by A. K. Erlang in 1909, for analyzing telephone traffic congestion, with the objective of meeting uncertain demands for services within the Copenhagen Telephone System.

Today queuing theory can be applied to a wide variety of operational situations, for example:-

Scheduling of aircraft during landing and takeoff at busy airports, shipping, transportation, inventory control, work scheduling, industrial production and process control, communication systems, and in public services like hospitals and banks.

Queuing Theory

Consider a queuing system that contain processes wherein there is a demand for service. The system can service activities at a rate which is, in general, greater than the rate at which entities arrive. There are often random fluctuations either in the rate of arrival, rate of service or both. As a result, there are certain times when more entities arrive than can be served, and some entities must therefore wait for service. These waiting entities are then said to join a waiting line. The combination of all entities in the system - those being served, and those waiting for service is termed a "queue".

Queuing System Model:

The queuing system model can be described in terms of three main characteristics. These are:-

(a) Arrival Pattern - which describes the statistical properties of the arrivals.

(b) Service Pattern - which describes how entities are being served.

(c) Queuing Discipline - which describes how entities are selected for service.

The service process, in turn, is described by two main factors:-

i. Service Time

ii. Service Capacity

Service time is the time required to serve an individual entity.

Service capacity is the number of entities that can be served.

To model a system, probability functions that describe arrival patterns and service times must be determined. Measurement of queues is among the most important output of a simulation.

Arrival Pattern:

The usual way of describing an arrival pattern is in terms of inter arrival time (the interval between successive arrivals.) For an arrival pattern that has no variability, inter arrival time is a constant. For arrival patterns that vary stochastically (randomly), it is necessary to define the probability function of the inter arrival time. Two or more arrivals may be simultaneous. If n arrivals are simultaneous, (n-1) of them have zero inter arrival time.

If T_a = mean inter arrival time

and β = mean arrival rate

Then $\beta = 1/T_a$

When describing arrival patterns, it is a common practice to express the distribution in terms of the probability that an arrival time is greater than a given time.

Let $A_o(t)$ be the probability of an inter arrival time being greater than t. Let F(t) be the probability of an inter arrival time less than t. Then we can say:

$$A_o(t) = 1 - F(t)$$

$A_o(t)$ will take a maximum value of 1, when time = 0, and it cannot increase as t increases.

Poisson Arrival Patterns:

This is a situation where the arrivals are said to be completely random. This means that an inter arrival can occur at

any time, subject only to the restriction that the mean arrival rate be some given value.

More formally, it is assumed that the time of the next arrival is independent of the previous arrival and that the probability of an arrival in an interval Δt, is proportional to Δt. If β is the mean number of arrivals/unit time, then the probability of an arrival in Δt is $\beta\Delta t$.

With these assumptions, it is possible to show that the distribution of the inter arrival time is exponential. The probability function of the inter arrival time is given by:

$f(t) = \beta e^{-\beta t}$ $(t >= 0)$

It follows the arrival distribution of

$A_o(t) = e^{-\beta t}$

The probability of n arrivals occurring in a period of length t is given by:-

$$P(n) = \frac{(\beta t)^n e^{-\beta t}}{n!} \qquad (n=0, 1, 2, 3, \ldots)$$

This distribution is called the Poisson Distribution.

Service times:

Frequently the service times of a process is constant; but when it varies stochastically, it must be described by a probability function.

If the service time is considered to be completely random, it may be represented by an exponential distribution.

Queuing discipline:

The third factor for describing congestion is the queuing discipline that determines how the next entity is selected from a waiting line. The most common queuing disciplines are:-

1. First In First Out (FIFO): service is offered to the entity that has waited longest.
2. Last In First Out (LIFO): service is offered to entities that have arrived most recently.
3. Random discipline: a random choice is made between all waiting entities at the time that service is offered.
4. Reneging: entities leave the system before their time of service comes up. This depends on queue length.
5. Polling: when there is more than one line forming for the same service.
6. Service on priority basis: entities are serviced based on priorities.

Measures of queue

Measures of queue are:

1) Traffic intensity: The ratio of mean service time to mean inter arrival time, denoted by "U".

2) Service Utilization: Denoted by $\Gamma = \beta T_s = \beta/\mu$. It is the ratio of mean arrival rate to mean service rate.

Both traffic intensity and server utilization can be greater than 1. If so, a single server cannot keep up with the flow of traffic. In fact, one way of describing the figure for traffic intensity is to say the minimum number of servers needed to handle the traffic without making customers wait or turning any customers away. If more than one server is used, say n servers, the server utilization is redefined to reflect the load on each individual server as follows:-

If n = number of servers, then

$\Gamma = \beta/n\mu$ server utilization

A server utilization of 1 or less implies that the system can keep up with the traffic flow. However, values of server utilization approaching 1 are undesirable as it indicates long waiting queues. Another method of calculating server utilization is by dividing the total time the server is used by the total time the server operates.

$$U = \frac{1}{T}\sum_{x=1}^{N} (t_f - t_b)_x$$

t_f = Service finish time

t_b = Service begin time

Two principal measures of queuing systems are the mean number of entities waiting and the mean time they spend in waiting. Both these quantities may refer to the total number of entities in the system, those waiting and those being served.

Queuing Statistics

In many stochastic simulations, the most important information obtained is the statistics on the behavior of queues. The type of statistics commonly used to describe queue behavior include average and maximum queue length and the average waiting time for an entity in the queue.

In order to compute mean queue length, the time and magnitude of change in queue length are noted either at each increment of simulated time or whenever a queue arrival or departure takes place.

If t_k is the time the queue had a length of K and T be the total simulation time, then:

$$T-\sum_{k=0}^{\infty} t_k$$

The mean queue length " Q " is the waited average of the monitored queue lengths with respect to time, i.e.

$$\bar{Q}-\frac{1}{T}\sum_{k=0}^{\infty} t_k$$

where

Average waiting time of a customer in the system:-

$W_s = 1/ (\mu - \beta)$

where W_s = average waiting time of the customer in the system

μ = average rate of service

β = average rate of arrival

$$W_s = \frac{\text{Total waiting time of all customers}}{\text{Total number of customers}}$$

Average service time (S_s)

$$S_s = \frac{\text{Total service time of all customers}}{\text{Number of customers served}}$$

$$\text{Average counter idle time/day} = \frac{\text{Total idle time of counters for all days}}{\text{Number of days}}$$

Queuing theory provides techniques for determining measures of effectiveness, such as queue length, average waiting time, etc. when the distribution of inter arrival and service times are known. If costs are assigned to waiting time of customers and idle time of the service facility, the problem of establishing a proper balance between the two can be determined.

Many queuing problems cannot be solved explicitly by analytical methods. In such cases, the only method of solution is to simulate the situation.

In case of queuing systems, simulation is one of the most appropriate techniques. It makes the study of the system under observation much simpler and reduces the time factor. Computerized simulation allows the researcher to study the system for any period of time. These observations may take a few hours as opposed to many weeks or months of real time. Further, computer simulation is an ideal technique since the probability of customer arrival is random in nature.

System Simulation

Simulation may be defined as a quantitative technique that uses a computer model in order to represent actual decision making in uncertain conditions. This is used to determine alternative courses of action based upon fact and assumption.

Simulation is a method of solving problems of wide variety. To "simulate" is to copy the behavior of a system or phenomenon under study. In order to simulate a system, relevant information about it must first be gathered. This collection of data is termed the "system model".

Simulation is basically an experimental technique. It is a fast and relatively inexpensive method of carrying out an experiment using the computer.

Simulation is a method of solving problems by designing, constructing and manipulating a model of the real system. It is defined as the act of performing experiments on a model of a given system. It duplicates the essence of a system (or activity) without actually entering reality. A system is defined as a collection of entities or components which act and interact in unison towards achievement of some goal.

Simulation involves construction of a symbolic model that describes the system. This description is in terms of :-

1. individual events and components.
2. dividing the system into smaller components and combining them in their natural and logical order.
3. analyzing effects of their interaction with one another.
4. studying various specific alternatives with reference to the performance of the model and choosing the best one.

The construction of an appropriate model of a system is a delicate and important affair in simulation.

In theory any system that can be simulated on a digital computer can be simulated manually. In practice, however, simulation as an analytic tool is useful only when done on a computer. This is because practical problems that require simulation are complex and need a very large number of simple repetitive calculations which are time consuming.

Experimental nature of simulation:

Simulation technique makes no specific attempt to isolate relationships between any particular variables of the model changing with time. Relationships between these variables must be derived from these observations. Simulation is a very general method of studying problems.

System Modeling:

The critical step in a simulation study is the development of the system model. A system is composed of objects called entities that have properties or attributes. One of the key operators in the system modeling is abstraction. Abstraction entails eliminating all but the significant attributes from the entities.

Simulation models describe the changes in system. Processes that cause system changes are called activities. The state of a system is a description of all entities, attributes and activities at any given time. It is the purpose of simulation models to describe changes in the state of the system. System modeling can be considered as a two-step process:-

1) Structural modeling:
    a) define boundaries between the system and environment.
    b) identify entities.
    c) abstract critical attributes of entities.
    d) define activities.
2) Data modeling:
    a) describe relationship of activities and attributes.

b) specify means of obtaining values for attributes.

Simulation steps are as follows:-

1) select measure of effectiveness
2) describe variables which influence the measure of effectiveness significantly.
3) determining the cumulative probability distribution for each variable.
4) generate a set of random numbers.
5) consider each random number as a decimal value of the cumulative probability distribution.
6) insert the values so generated and find corresponding values of variables.
7) repeat (5) and (6) until sample is large enough for satisfaction of decision makers.

Application of simulation technique:

Simulation technique is applied to a variety of problems in various areas like management, inventory, queuing, physics, chemistry, etc.

Types of queuing models

Modeling of queues is an essential part of nearly every simulation study. An understanding of the mechanism of queue management aids the modeler in interpretation of the results obtained. In general, the task of analyzing simulation statistics can be performed effectively with the knowledge of the procedures used to compute statistics.

There are two basic operations of queues in simulation. The arrival and departure of entities. The exact form of a queuing model depends upon the number of servers and the number of queues holding entities waiting for service.

1) The most basic model is the Single Server Single Queue situation. In this context, entities requiring service form only one queue. There is only one server that provides service to all entities in the queue based on a queuing discipline (FIFO, LIFO, priority, etc.)
2) Single Queue Multi Server: in this situation, a single queue is formed and multiple servers are available to provide service to entities of this queue based on queue discipline.
3) Multi Queue Multi Server: there are queues formed for a number of servers. Entities arriving select servers on their own choices and await turn in order to receive service.
4) Multi Queue Single Server: all entities form different queues (depending probably on type of service) for various services from a single server. Service is provided based on some queue discipline.

The above mentioned situations are only with respect to a single stage of service. However, queues of the above situation do exist for multi stage services. In these cases, the output of one server becomes the input to the next stage.

Monte Carlo Simulation Technique

A simulation model need not be a deterministic one and may include some elements of uncertainty. For example, in the waiting line model, inter arrivals and service times are usually probabalistic rather than deterministic.

The problem in all types of simulations is of generating a sequence of numbers that can be considered as being typical observations from a given probability distribution. Further more, the generated numbers must be randomly drawn from the uniform distribution so that if the sequence is tested statistically, it will be found to follow the uniform distribution.

The method takes its name from the Monte Carlo gambling establishment because the samples are selected in a purely random sequence.

This technique has become so much a part of simulation models that the terms are often considered to be synonymous. The procedure of Monte Carlo involves the selection of random observations within the simulation model and comprises the following two steps:-

1) Generation of random observations from a uniform distribution.
2) Generation of random observation from any derived probability distribution.

The Monte Carlo technique was adopted by the researcher for solution of this queuing problem.

Generation of Random Numbers from a Uniform Distribution

There are three methods of generation of uniformly distributed random numbers as follows:

1) Use of random number tables.
2) Use of physical procedures such as spinner, bowl of chips,etc.
3) Computer generated random numbers.

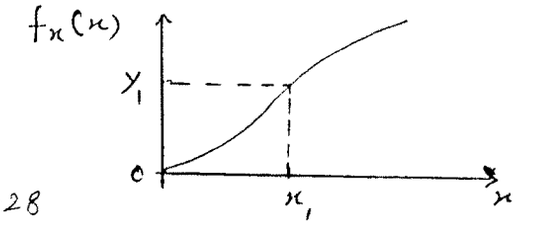
The first two methods are very slow.

Generation of random observations from a probability distribution:

To generate a sequence of random observations from any specified probability distribution when a sequence of random numbers is known. This technique is called Inverse or Probability Transformation Method. There are three steps involved which are as follows:-

1) Construction of cumulative distribution function of the variable x.
2) Generate normalized (0,1) random numbers y_1, y_2, ...
3) Set $F_x(x)$ equal to the random decimal number and solve for x. The value of x thus obtained will be the desired random observation from the probability distribution.

To obtain x_1, the first random observation corresponding to $f_x(x)$, we simply enter the ordinate with Y_1, project over and down as shown in the figure. Then the resulting value of x_1 can be obtained.



Simulation techniques

The two basic techniques of simulation are :-

1) Fixed Time Step

2) Next Event

Fixed Time Step Simulation:

Here the simulation clock is advanced in fixed steps and at each step, the control routine determines the activities to be started or terminated. The routine also updates parameters when necessary.

This approach is simple but has the disadvantage of lacking chronological order of completed activities. The value of time step is also very critical for too small a value will result in wasteful scans and too large a value could miss the occurrence of some important events.

Next Event Simulation:

In this simulation, events are listed in chronological order and are executed accordingly.

A combination of these two techniques is also in vogue. Here the control routine advances to the next unconditional event and executes all the events scheduled at that time. Next it scans the list and executes all events whose conditions are satisfied. The entire procedure is then repeated. The next event simulation method has been used in this thesis.

Advantages & Disadvantages of Simulation

Advantages:

1) It is useful in solving problems where all values of variables are not known (or only partly known) in advance and there is an easy way to find these values.
2) The model of the system, once constructed may be employed as often as desired to analyze different situations.
3) Simulation methods are handy for analyzing proposed systems in which information is sketchy at best.
4) Using a model for observation without requiring a real life situation.
5) Data for further analysis can be easily generated from a simulation model.
6) Simulation methods are easier to apply than pure analytical methods.

Disadvantages:

1) Adequate knowledge of the parts of a system in no sense guarantees adequate knowledge of system behavior.
2) Simulation model is run rather than solved.
3) Simulation does not produce optimal results. It only provides a satisfactory approach.
4) Each simulation run is like a single experiment conducted under a given set of conditions. A number of simulation runs are necessary and thus time consuming.
5) People develop the tendency of using it even when analytical techniques are better suited in a situation.

It should be noted that simulation is indeed a versatile tool. It provides only statistical estimates rather than exact results and it only compares the alternatives rather than generating an optimal one. It is a slow and costly way to study a problem. Despite limitations, it is an invaluable tool in Operation Research.

Computer Simulation Languages

A number of computer languages have been produced to simplify the task of writing system simulation programs. Each language is based upon a set of concepts used to describe the system. Some of the dedicated languages available are SIMSCRIPT, DYNAMO, GPSS (General Purpose System Simulation) developed by IBM Corporation, SIMPAC and CLS. Several high level languages like ALGOL, FORTRAN, COBOL, BASIC and PASCAL have also been used.

GPSS is written specifically for users with little or no programming experience, while SIMSCRIPT requires some programming skill (to the level where a user is conversant with FORTRAN or ALGOL). GPSS is less flexible and can be used in simple cases whereas SIMSCRIPT is more versatile and is able to handle complex model simulations.

DYNAMO is similar to SIMSCRIPT except that it is used to simulate continuous models rather than discrete (models involving individual events).

The objective of all these languages is to speed up conversion of a simulation model into a computer program. The reason for several simulation languages is that each can be applied to specific types of problems.

High level languages are also used for developing programs to simulate systems. These high level languages can be used to develop a program for any type of simulation, either discrete or continuous. However, the level of programming skill required is very high.

The language selected for simulation program development is Microsoft Quick BASIC Version 4.0. This high level language is a highly upgraded version of BASIC. It includes a large number of

structured functions which ease the task of complex programming. The program is compiled into executable code in order to further speed up the program operation.

The program developed in this thesis provides analysis of queuing problems of the following types:

a) Single Server Single Queue

b) Multi Server Multi Queue

c) Single Queue Multi Server

d) Queues with token system.