SOS POLITICAL SCIENCE AND PUBLIC ADMINISTRATION

MBA FA 204

SUBJECT NAME: OPERATION RESEARCH

UNIT-IV

TOPIC NAME: ELEMENTS OF QUEUING THEORY

Elements of Queuing Systems

Figure 1 shows the elements of a single queue queuing system:

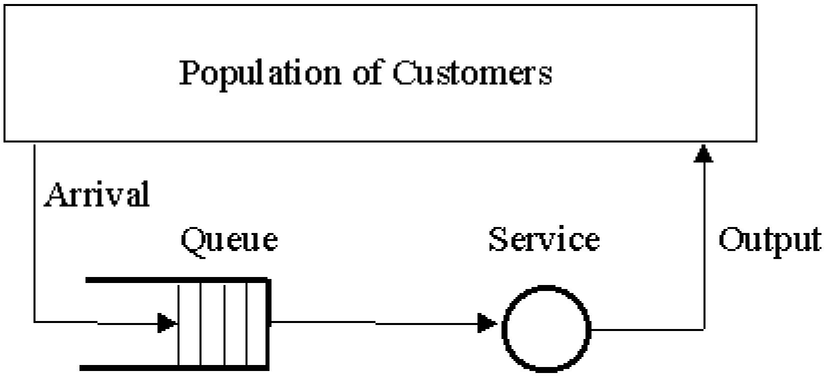

Figure 1

Population of Customers can be considered either limited (closed systems) or unlimited (open systems). Unlimited population represents a theoretical model of systems with a large number of possible customers (a bank on a busy street, a motorway petrol station). Example of a limited population may be a number of processes to be run (served) by a computer or a certain number of machines to be repaired by a service man. It is necessary to take the term "customer" very generally. Customers may be people, machines of various nature, computer processes, telephone calls, etc.

Arrival defines the way customers enter the system. Mostly the arrivals are random with random intervals between two adjacent arrivals. Typically the arrival is described by a random distribution of intervals also called *Arrival Pattern*.

Queue represents a certain number of customers waiting for service (of course the queue may be empty). Typically the customer being served is considered not to be in the queue. Sometimes the customers form a queue literally (people waiting in a line for a bank teller). Sometimes the queue is an abstraction (planes waiting for a runway to land). There are two important properties of a queue: *Maximum Size* and *Queuing Discipline*.

Maximum Queue Size (also called *System capacity*) is the maximum number of customers that may wait in the queue (plus the one(s) being served). Queue is always limited, but some theoretical models assume an unlimited queue length. If the queue length is limited, some customers are forced to renounce without being served.

Queuing Discipline represents the way the queue is organised (rules of inserting and removing customers to/from the queue). There are these ways:

1) FIFO (First In First Out) also called FCFS (First Come First Serve) - orderly queue.

2) LIFO (Last In First Out) also called LCFS (Last Come First Serve) - stack.

3) SIRO (Serve In Random Order).

4) Priority Queue, that may be viewed as a number of queues for various priorities.

5) Many other more complex queuing methods that typically change the customer’s position in the queue according to the time spent already in the queue, expected service duration, and/or priority. These methods are typical for computer multi-access systems.

Most quantitative parameters (like average queue length, average time spent in the system) do not depend on the queuing discipline. That’s why most models either do not take the queuing discipline into account at all or assume the normal FIFO queue. In fact the only parameter that depends on the queuing discipline is the variance (or standard deviation) of the waiting time. There is this important rule (that may be used for example to verify results of a simulation experiment):

The two extreme values of the waiting time variance are for the FIFO queue (minimum) and the LIFO queue (maximum).

Theoretical models (without priorities) assume only one queue. This is not considered as a limiting factor because practical systems with more queues (bank with several tellers with separate queues) may be viewed as a system with one queue, because the customers always select the shortest queue. Of course, it is assumed that the customers leave after being served. Systems with more queues (and more servers) where the customers may be served more times are called *Queuing Networks.*

Service represents some activity that takes time and that the customers are waiting for. Again take it very generally. It may be a real service carried on persons or machines, but it may be a CPU time slice, connection created for a telephone call, being shot down for an enemy plane, etc.

Typically a service takes random time. Theoretical models are based on random distribution of service duration also called *Service Pattern*. Another important parameter is the number of servers. Systems with one server only are called *Single Channel Systems*, systems with more servers are called *Multi Channel Systems*.

Output represents the way customers leave the system. Output is mostly ignored by theoretical models, but sometimes the customers leaving the server enter the queue again ("round robin" time-sharing systems).

Queuing Theory is a collection of mathematical models of various queuing systems that take as inputs parameters of the above elements and that provide quantitative parameters describing the system performance.

Because of random nature of the processes involved the queuing theory is rather demanding and all models are based on very strong assumptions (not always satisfied in practice). Many systems (especially queuing networks) are not soluble at all, so the only technique that may be applied is simulation.

Nevertheless queuing systems are practically very important because of the typical trade-off between the various costs of providing service and the costs associated with waiting for the service (or leaving the system without being served). High quality fast service is expensive, but costs caused by customers waiting in the queue are minimum. On the other hand long queues may cost a lot because customers (machines e.g.) do not work while waiting in the queue or customers leave because of long queues. So a typical problem is to find an optimum system configuration (e.g. the optimum number of servers). The solution may be found by applying queuing theory or by simulation.

Kendall Classification of Queuing Systems

The Kendall classification of queuing systems (1953) exists in several modifications. The most comprehensive classification uses 6 symbols:

$$A/B/s/q/c/p$$

where:

A is the arrival pattern (distribution of intervals between arrivals).

B is the service pattern (distribution of service duration).

s is the number of servers.

q is the queuing discipline (FIFO, LIFO, ...). Omitted for FIFO or if not specified.

c is the system capacity. Omitted for unlimited queues.

p is the population size (number of possible customers). Omitted for open systems.

These symbols are used for arrival and service patterns:

M is the Poisson (Markovian) process with exponential distribution of intervals or service duration respectively.

E_m is the Erlang distribution of intervals or service duration.

D is the symbol for deterministic (known) arrivals and constant service duration.

G is a general (any) distribution.

GI is a general (any) distribution with independent random values.

Examples:

D/M/1 = Deterministic (known) input, one exponential server, one unlimited FIFO or unspecified queue, unlimited customer population.

M/G/3/20 = Poisson input, three servers with any distribution, maximum number of customers 20, unlimited customer population.

D/M/1/LIFO/10/50 = Deterministic arrivals, one exponential server, queue is a stack of the maximum size 9, total number of customers 50.