Introduction

- Waiting line queues are one of the most important areas, where the technique of simulation has been extensively employed.
- The waiting lines or queues are a common site in real life.
- People at railway ticket window, vehicles at a petrol pump or at a traffic signal, workers at a tool crib, products at a machining center, television sets at a repair shop are a few examples of waiting lines.
- The waiting line situations arise, either because,
    - There is too much demand on the service facility so that the customers or entities have no wait forgetting service, or
    - There is too less demand, in which case the service facility have to wait for the entities
- The objective in the analysis of queuing situations is to balance the waiting time and idle time, so as to keep the total cost at minimum.
- The queuing theory its development to an engineer A.K.Earlang, who in 1920, studied waiting line queues of telephone calls in Copenhagen, Denmark.
- The problem was that during the busy period, telephone operators were unable to handle the calls, there was too much waiting time, which resulted in customer dissatisfaction.

State Variables

server

State:

- **InTheAir:** number of aircraft either landing or waiting to land
- **OnTheGround:** number of landed aircraft
- **RunwayFree:** Boolean, true if runway available

Discrete Event Simulation Computation

schedules

Events that have been scheduled, but have not been simulated (processed) yet are stored in a pending

event list

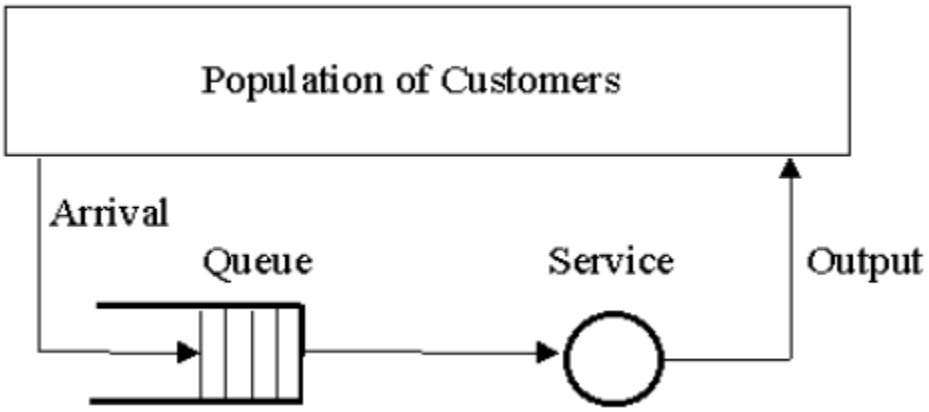
example: air traffic at an airport

events: aircraft arrival, landing, departure,

Model of the physical system

Independent of the simulation application

Elements of Queuing Systems



- Population of Customers or calling source 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

- 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 or waiting line

- Queue or waiting line 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

Applications of Queuing Theory

- Telecommunications
- Traffic control
- Determining the sequence of computer operations
- Predicting computer performance
- Health services (eg. control of hospital bed assignments)
- Airport traffic, airline ticket sales
- Layout of manufacturing systems.

Characteristics of queuing systems

a. Arrival Process

The distribution that determines how the tasks arrives in the system.

b. Service Process

The distribution that determines the task processing time

c. Number of Servers

Total number of servers available to process the tasks

Queuing Discipline

It represents the way the queue is organized (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 importantrule (that may be used for example to verify resultsof 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

- 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

- 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

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).

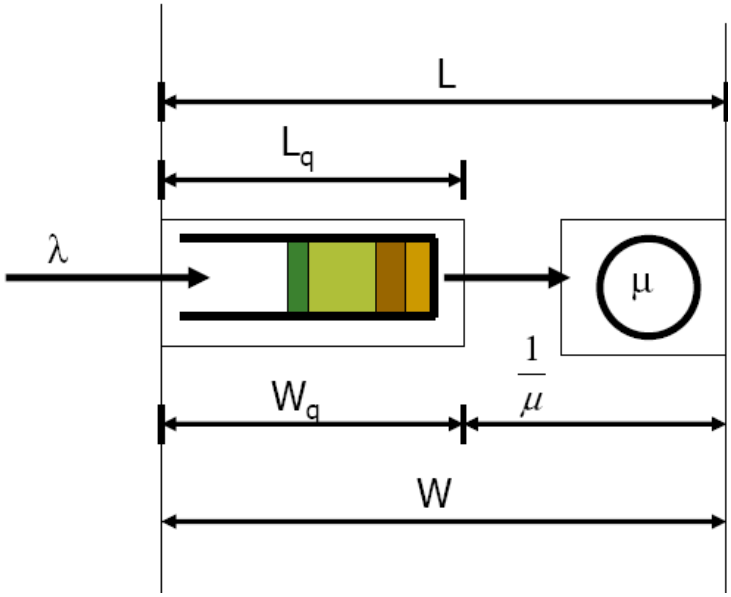
Many systems (especially queuing networks) are not soluble at all, so the only technique that may be applied is simulation.

- 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 .

Analysis of M/M/1 queue

Given:

λ : Arrival rate of jobs (packets on input link)
μ : Service rate of the server (output link)



Solve:

L : average number in queuing system
Lq : average number in the queue
W : average waiting time in whole system

Wq : average waiting time in the queue

Six parameters in shorthand

First three typically used, unless specified

1. **Arrival Distribution**
2. **Service Distribution**
3. **Number of servers**
4. **Total Capacity (infinite if not specified)**
5. **Population Size (infinite)**
6. **Service Discipline (FCFS/FIFO)**

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.

Em 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.

Simulation of Queuing Systems

- The knowledge of average number of customers in the queue or in the system helps to determine the space requirements of the waiting entities. Also too long a waiting line may discourage the prospectus customers, while no queue may suggest that service offered is not of good quality to attract customers.
- The knowledge of average waiting time in the queue is necessary for determining the cost of waiting in the queue.
- System Utilization that is the percentage capacity utilized reflects that extent to which the facility is busy rather than idle.
- System utilization factor (s) is the ratio of average arrival rate (ë) to the average service rate (ì).
- S= λ/μ in the case of a single server model
- S= λ/μn in the case of a “n” server model
- The system utilization can be increased by increasing the arrival rate which amounts to increasing the average queue length as well as the average waiting time, as shown in fig. Under the normal circumstances 100% system utilization is not a realistic goal.

Time Oriented Simulation

A factory has large number of semi automatic machines. On 50% of the working days none of the machines fail. On 30%of the days one machines fails and on 20%of the days two machines fail. The maintenance staff on the average puts 65% of the machines in order in one day, 30% in two days and remaining 5% in three days.

Simulate the system for 30 days duration and estimate the average length of queue, average waiting time and server loading that is the fraction of time for which server is busy.

Solution:

The given system is a single server queuing model. The failure of the machines in the factory generates arrivals, while the maintenance staff is the service facility. There is no limit on the capacity of the system in other words on the length of waiting line. The population of machines is very large and can be taken as infinite.

Arrival pattern:

On 50%of the days arrival=0

On 30%of the days arrival=1

On 20%of the days arrival=2

Expected arrival rate=0*.5+1*.3+2*.2=0.7 per day.

Service pattern:

65% machines in 1 day

30% machines in 2 days

5% machines in 3 days

Average service time: 1*.65+2*.3+3*.05=1.4 days

Expected service rate=1/1.4=0.714 machines per day

The expected arrival rate is slightly less than the expected service rate and hence the system can reach a steady state. For the purpose of generating the arrivals per day and the services completed per day the given discrete distributions will be used.

Random numbers between 0 and 1 will be used to generate the arrivals as under.

0.0<r<=0.5 Arrivals=0

0.5<r<=0.8 Arrivals=1

0.8<r<=1.0Arrivals=2

Similarly, random numbers between 0 and 1 will be used for generating the service times (ST)

0.0<r<=0.65ST=1day

0.65<r<=0.95ST=2days

0.95<r<=1.0 ST=3 days

- In the time-oriented simulation, the timer is advanced in fixed steps of time and at each step the system is scanned and updated.
- The time is kept very small, so that not many events occur during this time.
- All the events occurring during this small time interval are assumed to occur at the end of the interval.
- At start of the simulation, the system that is the maintenance facility can assumed to be empty, with no machine waiting for repair.
- On day 1, there is no machine in the repair facility.
- On day 2 there are 2 arrivals, the queue is made 2.
- Since service facility is idle, one arrival is put on service and queue becomes 1.
- Server idle time becomes 1 day and the waiting time of customers is also 1 day. Timer is advanced by one day.
- The service time, ST is decreased by one and when ST becomes zero facility becomes idle.
- Arrivals are generated which come out to be 1, it is added to the queue.
- Facility is checked, which is idle at this time.
- One customer is drawn from the queue, its service time is generated.
- Idle time and waiting time are updated.
- The process is continued till the end of simulation.

The following statistics can be determined.

Machine failures(arrivals) during 30 days=21

Arrivals per day=21/30=0.7

Waiting time of customer=40 days

Waiting time per customer=40/21=1.9 days

Average length of the queue=1.9

Server idle time=4 days=4/3* 100=13.33 %

Server loading=(30-4)/30=0.87