UNIT 2
QUEUING THEORY

LESSON 21

Learning Objective:

- **Examine situation in which queuing problems are generated.**
- **Introduce the various objectives that may be set for the operation of a waiting line.**
- **Explain standard queuing language.**

Hello Students,

You all know what is a queue? So here we are going to study How things work in a queue?

What is queuing theory?

Queuing Theory is a collection of mathematical models of various queuing systems.
It is used extensively to analyze production and service processes exhibiting random variability in market demand (arrival times) and service times.

Can you tell why queues form?

Queues or waiting lines arise when the demand for a service facility exceeds the capacity of that facility, that is, the customers do not get service immediately upon request but must wait, 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 total number of service facilities exceeds the number of customers requiring service.

Waiting lines, or queues are a common occurrence both in everyday life and in variety of business and industrial situations. Most waiting line problems are centered about the question of finding the ideal level of services that a firm should provide.

For example

- Supermarkets must decide how many cash register check out positions should be opened.
- Gasoline stations must decide how many pumps should be opened and how many attendants should be on duty.
- Manufacturing plants must determine the optimal number of mechanics to have on duty in each shift to repair machines that break down.
- Banks must decide how many teller windows to keep open to serve customers during various hours of the day.

Evolution of queuing theory

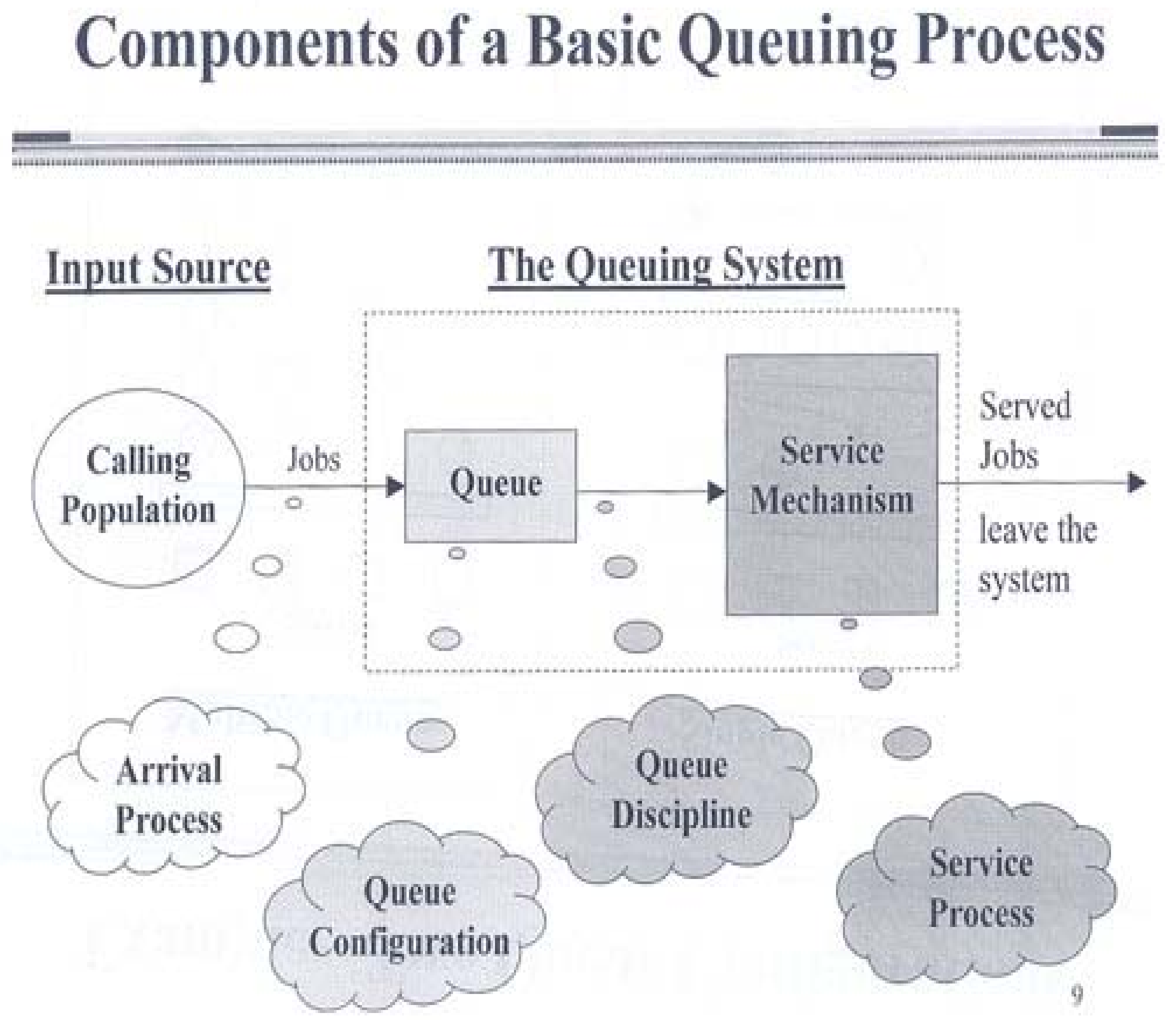
Queuing Theory had its beginning in the research work of a Danish engineer named A. K. Erlang. In 1909 Erlang experimented with fluctuating demand in telephone traffic. Eight years later he published a report addressing the delays in automatic dialing equipment. At the end of World War II, Erlang's early work was extended to more general problems and to business applications of waiting lines.

Some more examples of waiting lines are given in the following table :-

Queuing Examples

Situation	Arriving Customers	Service Facility
Passage of customers through a supermarket checkout	Shoppers	Checkout counters
Flow of automobile traffic through a road network	Automobiles	Road network

Manually placed assembly line	Parts to be assembled	Assembly Line
Inventory of items in a warehouse	Order for withdrawal	Warehouse
Banking Transactions	Bank Patrons	Bank tellers
Ships entering a port	Ships	Docks
Maintenance and repair of machines	Machine break-down	Repair crew
Scheduling of patients in a clinic	Patients	Medical Care
Number of runways at an airport	Airplanes	Runways
Parking lot	Automobiles	Parking spaces
Capacity of a motel	Motorists	Lodging facilities
Arrival of automobiles at a Garage	Automobiles	Repair of automobiles
Transfer of electronic messages	Electronic messages	Transmission lines
Flow of computer programmes through a computer system	Computer programmes	Central Processing unit
Sale of theatre tickets	Theatre-goers	Ticket windows
Arrival of trucks at central market	Trucks	Loading crews
Registration of unemployed at an employment exchange	Unemployed personnel	Registration assistants
Calls at police control room	Service calls	Policemen



Components of a Basic Queuing Process

Input Source

The Queuing System

Calling Population

Jobs

Queue

Service Mechanism

Served Jobs

leave the system

Arrival Process

Queue Configuration

Queue Discipline

Service Process

9

Firstly there are some basic components in every queuing system

BASIC COMPONENTS OF A QUEUING SYSTEM

INPUT SOURCE OF QUEUE

An input source is characterized by

- Size of the calling population
- Pattern of arrivals at the system
- Behaviour of the arrivals

Customers requiring service are generated at different times by an input source, commonly known as population. The rate at which customers arrive at the service facility is determined by the arrival process.

Size of the calling population—
The size represents the total number of potential customers who will require service.

According to source
The source of customers can be finite or infinite. For example, all people of a city or state (and others) could be the potential customers at a supermarket. The number of people being very large, it can be taken to be infinite. Whereas there are many situations in business and industrial conditions where we cannot consider the population to be infinite—it is finite.

According to numbers
The customers may arrive for service individually or in groups. Single arrivals are illustrated by patients visiting a doctor, students reaching at a library counter etc. On the other hand, families visiting restaurants, ships discharging cargo at a dock are examples of bulk, or batch arrivals.

According to time
Customers arrive in the system at a service facility according to some known schedule (for example one patient every 15 minutes or a candidate for interview every half hour) or else they arrive randomly. Arrivals are considered random when they are independent of one another and their occurrence cannot be predicted exactly. The queuing models wherein customers' arrival times are known with certainity are categorized as deterministic models. (insofar as this characteristic is concerned) and are easier to handle. On the other hand, a substantial majority of the queuing models are based on the premise that the customers enter the system stochastically, at random points in time.

Pattern of arrivals at the system—

The arrival process (or pattern) of customers to the service system is classified into two categories: ***static*** *and* ***dynamic***. These two are further classified based on the nature of arrival rate and the control that can be exercised on the arrival process.

In **static arrival process**, the control depends on the nature of arrival rate (random or constant). Random arrivals are either at a constant rate or varying with time. Thus to analyze the queuing system, it is necessary to attempt to describe the probability distribution of arrivals. From such distributions we obtain average time between successive arrivals, also called inter-arrival time (time between two consecutive arrivals), and the average arrival rate (i.e. number of customers arriving per unit of time at the service system).

The **dynamic arrival process** is controlled by both service facility and customers. The service facility adjusts its capacity to match changes in the demand intensity, by either varying the staffing levels at different timings of service, varying service charges (such as telephone call charges at different hours of the day or week) at different timings, or allowing entry with appointments.

Frequently in queuing problems, the number of arrivals per unit of time can be estimated by a probability distribution known as the Poisson distribution, as it adequately supports many real world situations.

Behavior of arrivals—

Another thing to consider in the queuing structure is the behavior or attitude of the customers entering the queuing system.

On this basis, the customers may be classified as being

(a) patient, or

(b) impatient.

If a customer, on arriving at the service system stays in the system until served, no matter how much he has to wait for service is called a patient customer.

Machines arrived at the maintenance shop in a plant are examples of patient customers.

Whereas the customer, who waits for a certain time in the queue and leaves the service system without getting service due to certain reasons such as a long queue in front of him is called an impatient customer.

Now, Let us see some interesting observations of human behavior in queues :

- Balking – Some customers even before joining the queue get discouraged by seeing the number of customers already in service system or estimating the excessive waiting time for desired service, decide to return for service at a later time. In queuing theory this is known as balking.

- Reneging - customers after joining the queue, wait for sometime and leave the service system due to intolerable delay, so they renege.

For example, a customer who has just arrived at a grocery store and finds that the salesmen are busy in serving the customers already in the system, will either wait for service till his patience is exhausted or estimates that his waiting time may be excessive and so leaves immediately to seek service elsewhere.

- Jockeying - Customers who switch from one queue to another hoping to receive service more quickly are said to be jockeying.

SERVICE SYSTEM

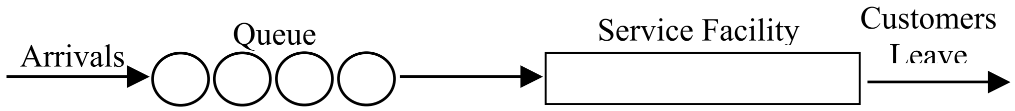
The service is provided by a service facility (or facilities). This may be a person (a bank teller, a barber, a machine (elevator, gasoline pump), or a space (airport runway, parking lot, hospital bed), to mention just a few. A service facility may include one person or several people operating as a team.

There are two aspects of a service system—(a) the configuration of the service system and (b) the speed of the service.

a) Configuration of the service system

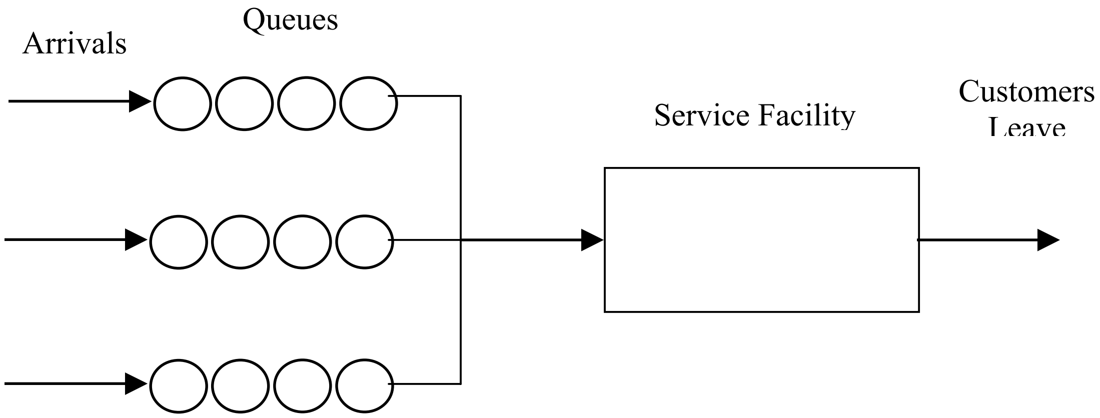
The customers' entry into the service system depends upon the queue conditions. If at the time of customers' arrival, the server is idle, then the customer is served immediately. Otherwise the customer is asked to join the queue, which can have several configurations. By configuration of the service system we mean how the service facilities exist. Service systems are usually classified in terms of their number of channels, or numbers of servers.

i **Single Server – Single Queue** -- The models that involve one queue – one service station facility are called single server models where customer waits till the service point is ready to take him for servicing. Students arriving at a library counter is an example of a single server facility.



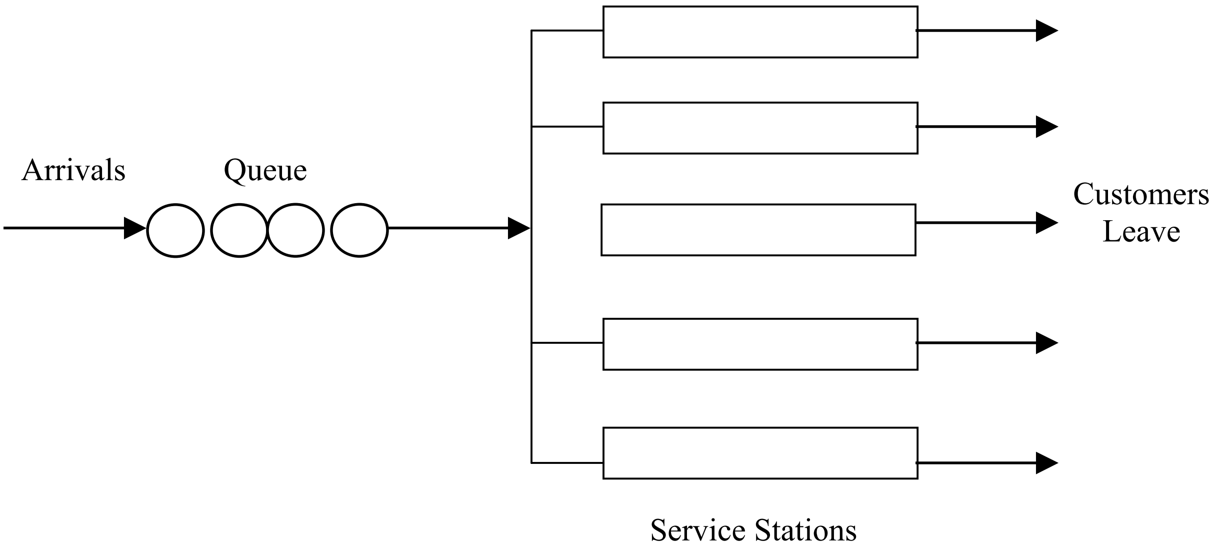
Single Server – Single Queue Model

ii **Single Server – Several Queues** – In this type of facility there are several queues and the customer may join any one of these but there is only one service channel.



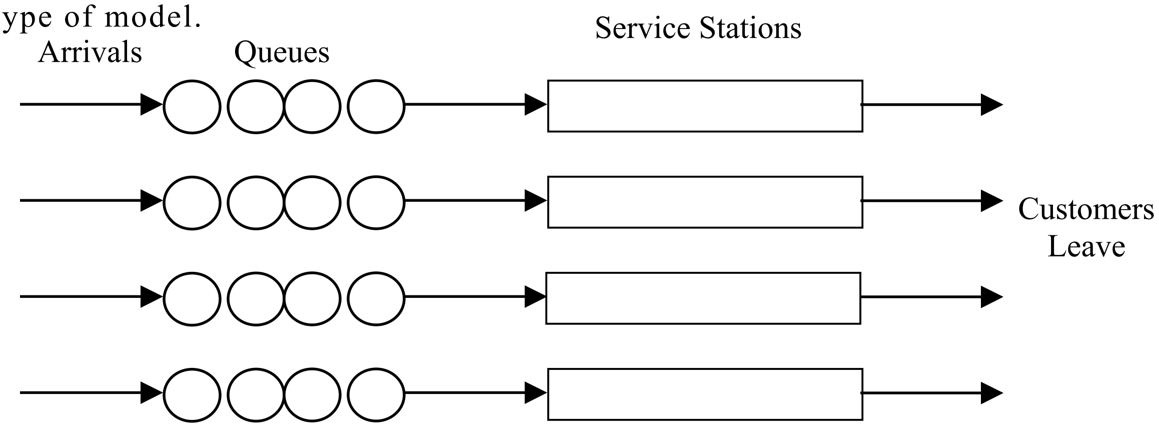
Single Server – Single Queue Model

iii **Several (Parallel) Servers – Single Queue** – In this type of model there is more than one server and each server provides the same type of facility. The customers wait in a single queue until one of the service channels is ready to take them in for servicing.



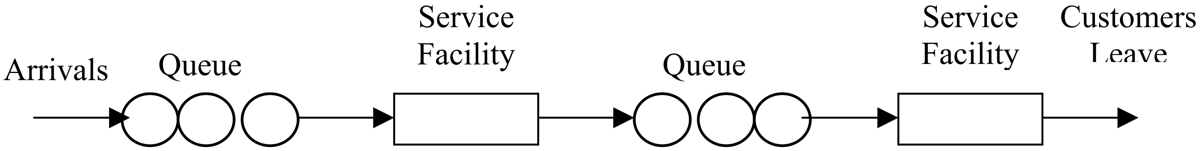
Several, Parallel Servers – Single Queue Model

iv **Several Servers – Several Queues** – This type of model consists of several servers where each of the servers has a different queue. Different cash counters in an electricity office where the customers can make payment in respect of their electricity bills provide an example of this type of model.



Several, Parallel Servers – Several Queues Model

v **Service facilities in a series** – In this, a customer enters the first station and gets a portion of service and then moves on to the next station, gets some service and then again moves on to the next station. …. and so on, and finally leaves the system, having received the complete service. For example, machining of a certain steel item may consist of cutting, turning, knurling, drilling, grinding, and packaging operations, each of which is performed by a single server in a series.



Multiple Servers in a Series

b) Speed of Service

In a queuing system, the speed with which service is provided can be expressed in either of two ways—as service rate and as service time.

- The service rate describes the number of customers serviced during a particular time period.
- The service time indicates the amount of time needed to service a customer.
- Service rates and times are reciprocal of each other and either of them is sufficient to indicate the capacity of the facility.

Thus if a cashier can attend, on an average 5 customers in an hour, the service rate would be expressed as 5 customers/hour and service time would be equal to 12 minutes/customer.

Generally, we consider the service time only.

If these service times are known exactly, the problem can be handled easily. But, as generally happens. if these are different and not known with certainty, we have to consider the distribution of the service times in order to analyze the queuing system. Generally, the queuing models are based on the assumption that service times are exponentially distributed about some average service time.

QUEUE CONFIGURATION

The queuing process refers to the number of queues, and their respective lengths. The number of queues depend upon the layout of a service system. Thus there may be a single queue or multiple queues.

Length (or size) of the queue depends upon the operational situation such as

- physical space,
- legal restrictions, and
- attitude of the customers.

In certain cases, a service system is unable to accommodate more than the required number of customers at a time. No further customers are allowed to enter until space becomes available to accommodate new customers. Such type of situations are referred to as finite (or limited) source queue.

Examples of finite source queues are cinema halls, restaurants, etc.

On the other hand, if a service system is able to accommodate any number of customers at a time, then it is referred to as infinite (or unlimited) source. queue.

For example, in a sales department, here the customer orders are received, there is no restriction on the number of orders that can come in, so that a queue of any size can form.

In many other situations, when arriving customers experience long queue(s) in front of a service facility, they often do not enter the service system even though additional waiting space is available. The queue length in such cases depends upon the attitude of the customers.

For example, when a motorist finds that there are many vehicles waiting at the petrol station, in most of the cases he does not stop at this station and seeks service elsewhere.

QUEUE DISCIPLINE

In the queue structure, the important thing to know is the queue discipline. The queue discipline is the order or manner in which customers from the queue are selected for service.

There are a number of ways in which customers in the queue are served. Some of these are:

(a) **Static queue disciplines** are based on the individual customer's status in the queue. Few of such disciplines are:

i If the customers are served in the order of their arrival, then this is known as the **first-come, first-served (FCFS)** service discipline. Prepaid taxi queue at airports where a taxi is engaged on a first-come, first-served basis is an example of this discipline.

ii **Last-come-first-served (LCFS)**-- Sometimes, the customers are serviced in the reverse order of their entry so that the ones who join the last are served first. For example, assume that letters to be typed, or order forms to be processed accumulate in a pile, each new addition being put on the top of them. The typist or the clerk might process these letters or orders by taking each new task from the top of the pile. Thus, a just arriving task would be the next to be serviced provided that no fresh task arrives before it is picked up. Similarly, the people who join an elevator last are the first ones to leave it.

(b) **Dynamic queue disciplines** are based on the individual customer attributes in the queue. Few of such disciplines are:

i **Service in Random Order (SIRO)**-- Under this rule customers are selected for service at random, irrespective of their arrivals in the service system. In this every customer in the queue is equally likely to be selected. The time of arrival of the customers is, therefore, of no relevance in such a case.

ii **Priority Service**-- Under this rule customers are grouped in priority classes on the basis of some attributes such as service time or urgency or according to some identifiable characteristic, and FCFS rule is used within each class to provide service. Treatment of VIPs in preference to other patients in a hospital is an example of priority service.

For the queuing models that we shall consider, the assumption would be that the **customers are serviced on the first-come-first-served basis**.

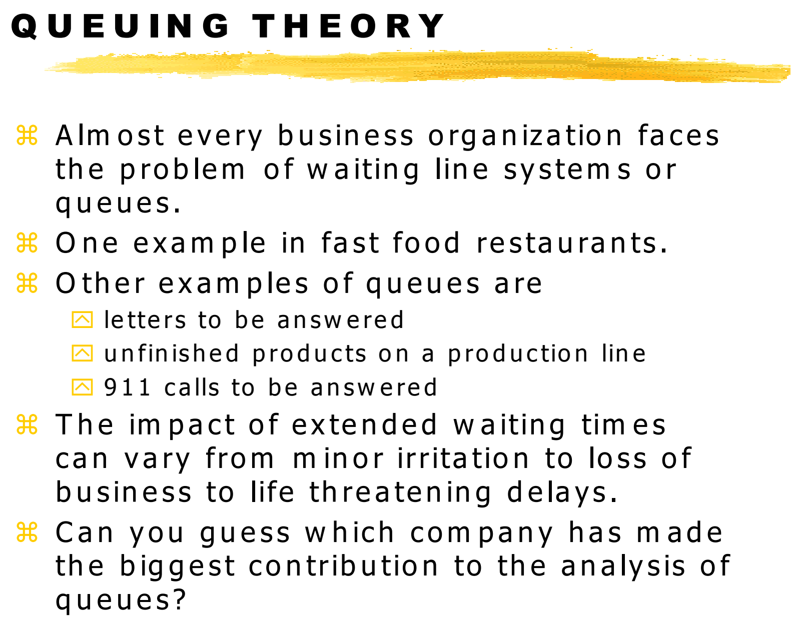
So, now let us summarize today's discussion:

Summary

We have discussed in details about

- Formation of queues.
- Basic components of queues

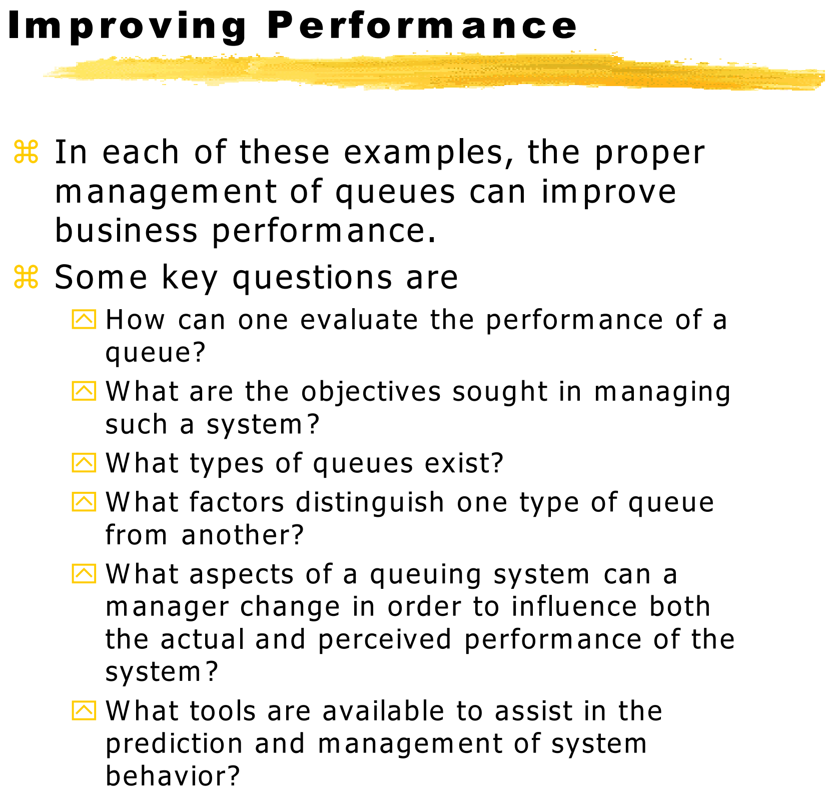
Slide 1



QUEUING THEORY

- Almost every business organization faces the problem of waiting line systems or queues.
- One example in fast food restaurants.
- Other examples of queues are
  - letters to be answered
  - unfinished products on a production line
  - 911 calls to be answered
- The impact of extended waiting times can vary from minor irritation to loss of business to life threatening delays.
- Can you guess which company has made the biggest contribution to the analysis of queues?

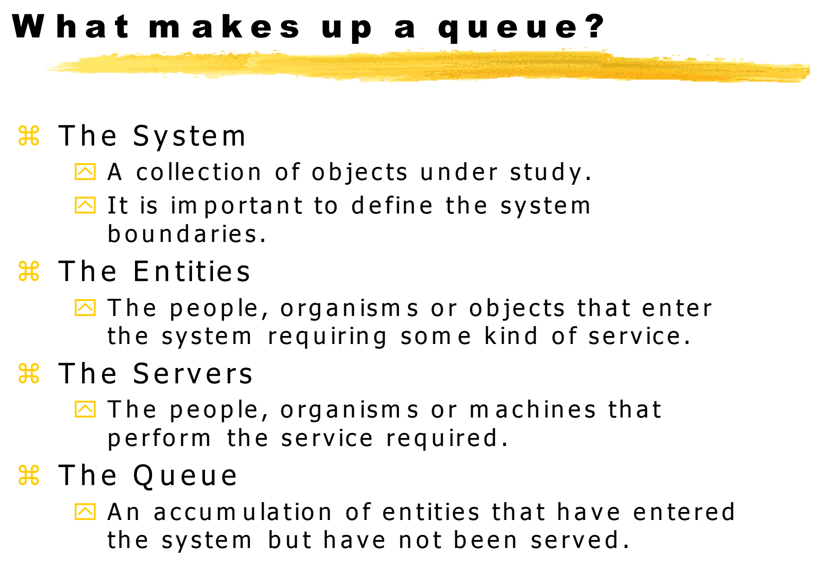
Slide 2



Improving Performance

- In each of these examples, the proper management of queues can improve business performance.
- Some key questions are
  - How can one evaluate the performance of a queue?
  - What are the objectives sought in managing such a system?
  - What types of queues exist?
  - What factors distinguish one type of queue from another?
  - What aspects of a queuing system can a manager change in order to influence both the actual and perceived performance of the system?
  - What tools are available to assist in the prediction and management of system behavior?

Slide 3

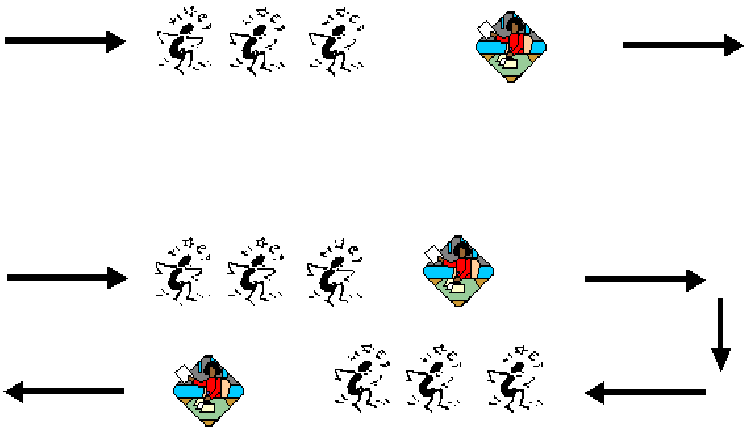


What makes up a queue?

- The System
  - A collection of objects under study.
  - It is important to define the system boundaries.
- The Entities
  - The people, organisms or objects that enter the system requiring some kind of service.
- The Servers
  - The people, organisms or machines that perform the service required.
- The Queue
  - An accumulation of entities that have entered the system but have not been served.

Slide 4

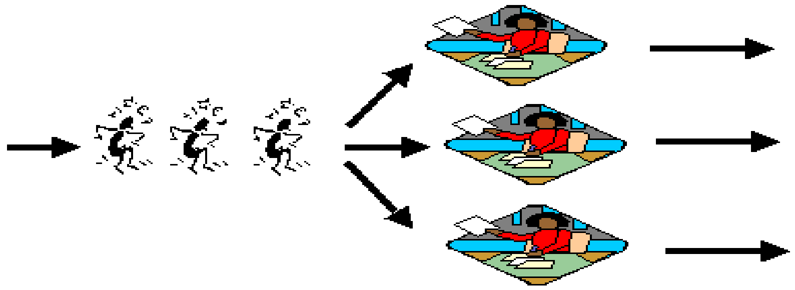
Types of Queues



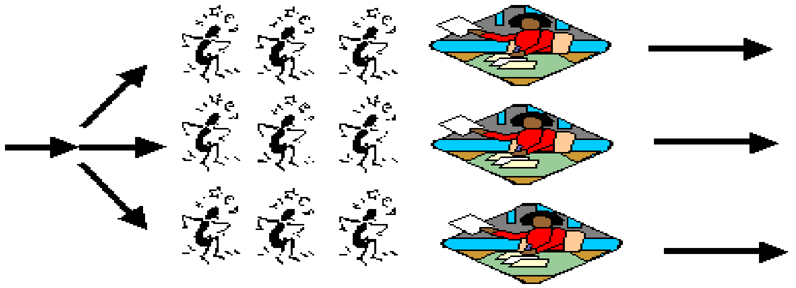
Multiple Stage - Manufacturing Plant

Slide 5

Types of Queues



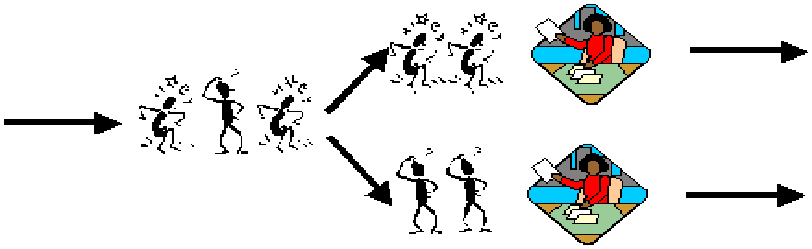
Multi-channel Single Stage - Bank



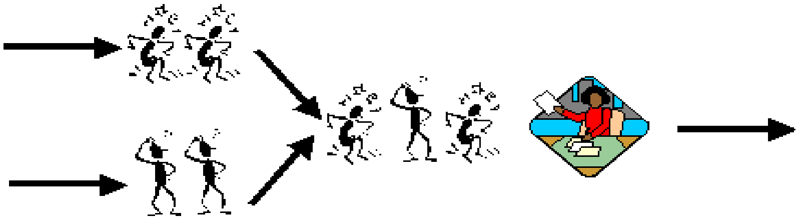
Parallel Single Stage - Supermarket

Slide 6

Types of Queues

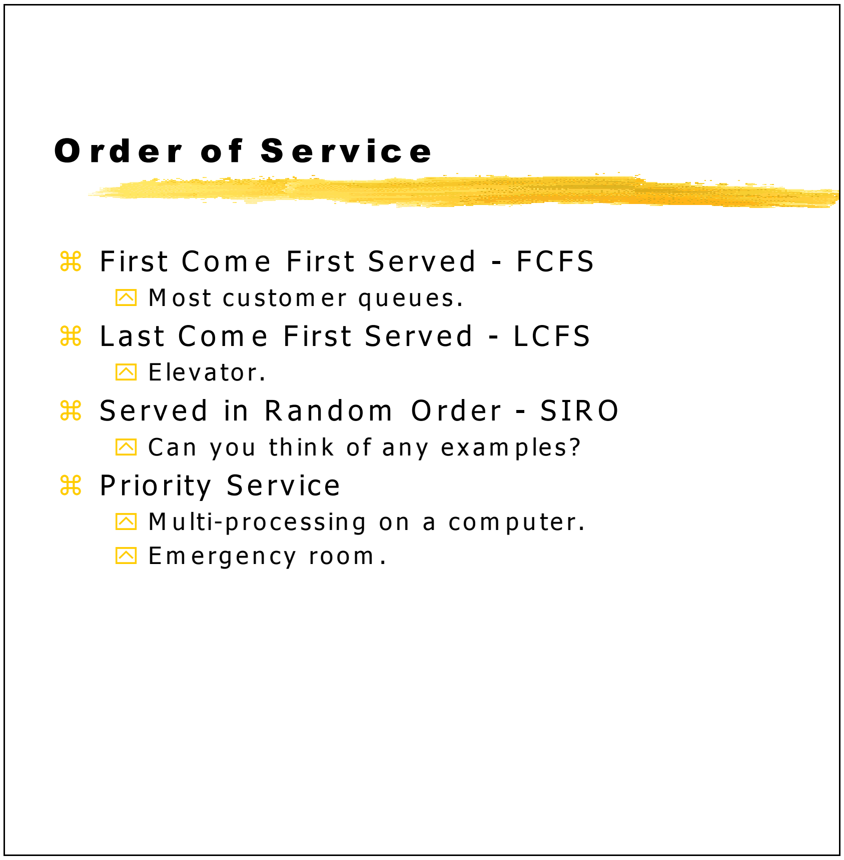


Customer Discrimination - INS



Converging Arrivals - Walk-in and Drive-thru

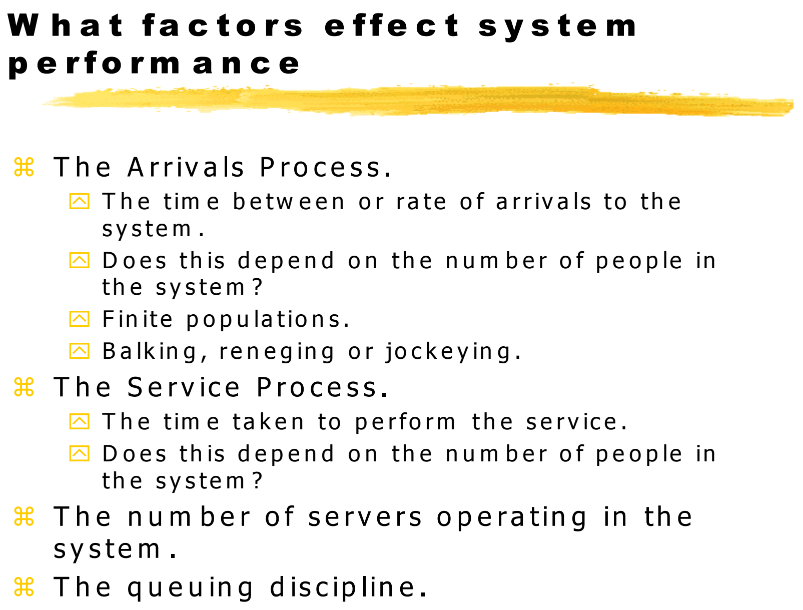
Slide 7



Order of Service

- First Come First Served - FCFS
  - Most customer queues.
- Last Come First Served - LCFS
  - Elevator.
- Served in Random Order - SIRO
  - Can you think of any examples?
- Priority Service
  - Multi-processing on a computer.
  - Emergency room.

Slide 8



What factors effect system performance

- The Arrivals Process.
  - The time between or rate of arrivals to the system.
  - Does this depend on the number of people in the system?
  - Finite populations.
  - Balking, reneging or jockeying.
- The Service Process.
  - The time taken to perform the service.
  - Does this depend on the number of people in the system?
- The number of servers operating in the system.
- The queuing discipline.