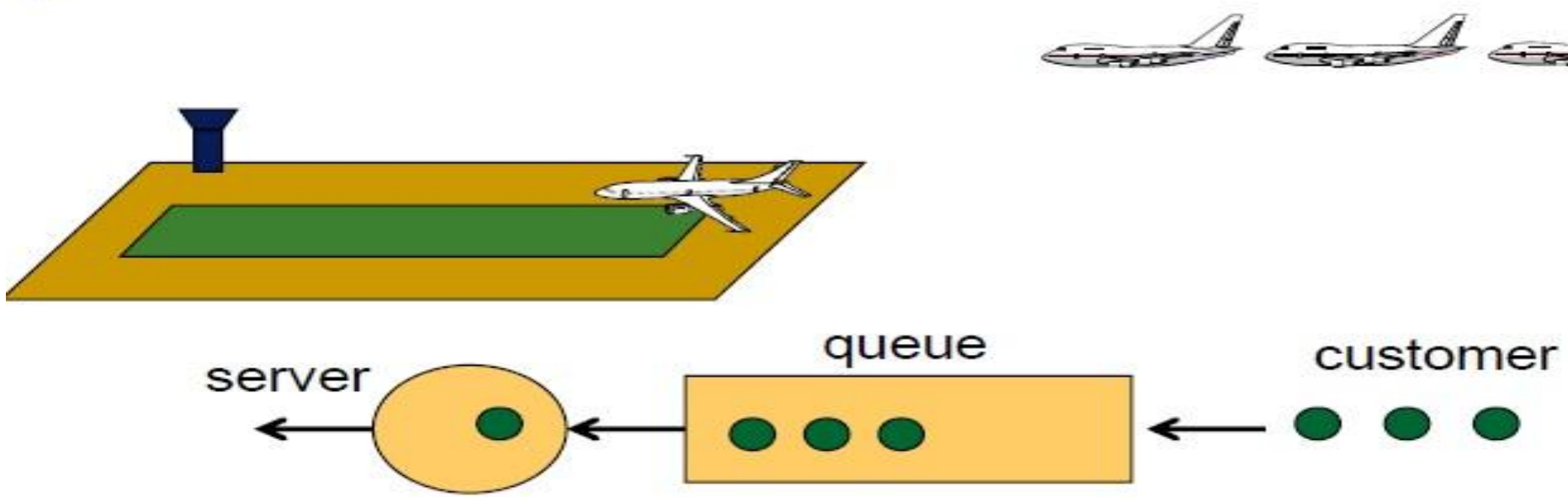
Chapter 4
Queuing System

State Variables

- A state variable is one of the set of variables that are used to describe the state of a dynamical system.



- The state variables for above system can be:

1. InTheAir: Number of aircrafts either landing or waiting to land.
2. OnTheGround: Number of landed aircraft.
3. RunWayFree: Can be a Boolean value which is true if runway available.

Queuing System

- Waiting line queues are one of the most important areas, where the technique of simulation has been extensively employed.
- People at bank for service, 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 problem with the queue is that if it is not managed properly, the customers should wait a long time.
- To prevent customer from being unsatisfied with the provided service, queuing system is managed. For eg: During cash withdraw in bank, you have to stay in queue and if it is not managed properly then you will surely be disappointed from the bank service even if that bank is one of the finest one in the city.

 The waiting situation arise because of any of the following reasons:

1. There is too much demand on the service facility so that the customers or entities have to wait for getting service.

2. There is too less demand, in which case the service facility have to wait for the entities.

 The main objective for the analysis of queuing situations is to balance the waiting time and idle time so as to balance the **waiting time** and **idle time**, so that the total cost will be minimized.

- Major elements in a Queuing System are:

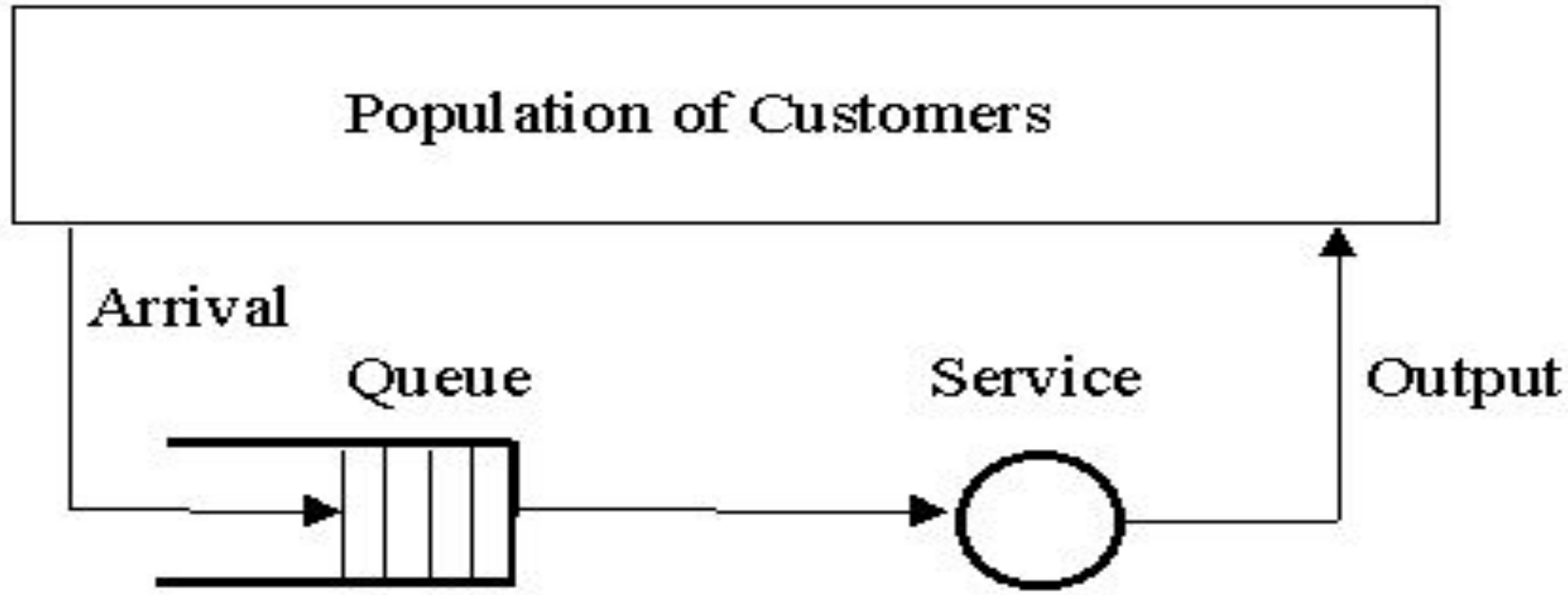


Figure 1

Elements of a Queuing System

1. Population of Customers or Calling Population

- The population of potential customers of the service is called calling population.
- 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. The system of the restaurant or bank, a motorway petrol station and so on are considered to be open system i.e. with infinite calling population.
- Example of a limited population may be a number of processes to be run (served) by a computer or a certain processes to be run (served) by a computer, system of repairing certain number of machines by a service man, etc.
- The term customer must be taken in general which may be people, machines, computer processes, telephone calls,etc.

2. Arrival

- Arrival is defined as the way in which the customers arrives into the system.
- In most of the cases, the arrival of the customer is random.
- So the inter-arrival between two customers is described by a random distribution of interval known as arrival pattern.

3. Queue

- Queue represents the number of customers that have entered into the system and are waiting for the service.
- Maximum Queue Size (also called System capacity) is the maximum number of customers that may wait in the queue.
- The two main properties of queue are as follows:
  a) Maximum Size:
  - Queue, in practice, is always limited.
  - Maximum size represents the maximum number of customers that can accommodate in the queue.

b) Queue Discipline:

- Queue discipline represents the rules in which the customers are inserted or removed to or from the queue.
- It can be organized in various ways like FIFO, LIFO, Serve In Random Order(SIRO), Priority Queue, etc.

4. Service Time

- Service time represents the time needed to provide service to a customer by a server.
- Service time may be of constant duration or of random duration.

5. Number of servers

- Servers represent the entity that provides service to the customer.
- A system may consist of single server or multiple servers.
- A system with multiple servers is able to provide parallel services to the customers.

6. 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.

Applications of Queuing Theory

1. Telecommunication
2. Traffic control
3. Determining the sequence of computer operations
4. Predicting computer performance
5. Health services (e.g. control of hospital bed)
6. Airport traffic
7. Airline ticket sales
8. Layout of manufacturing systems.

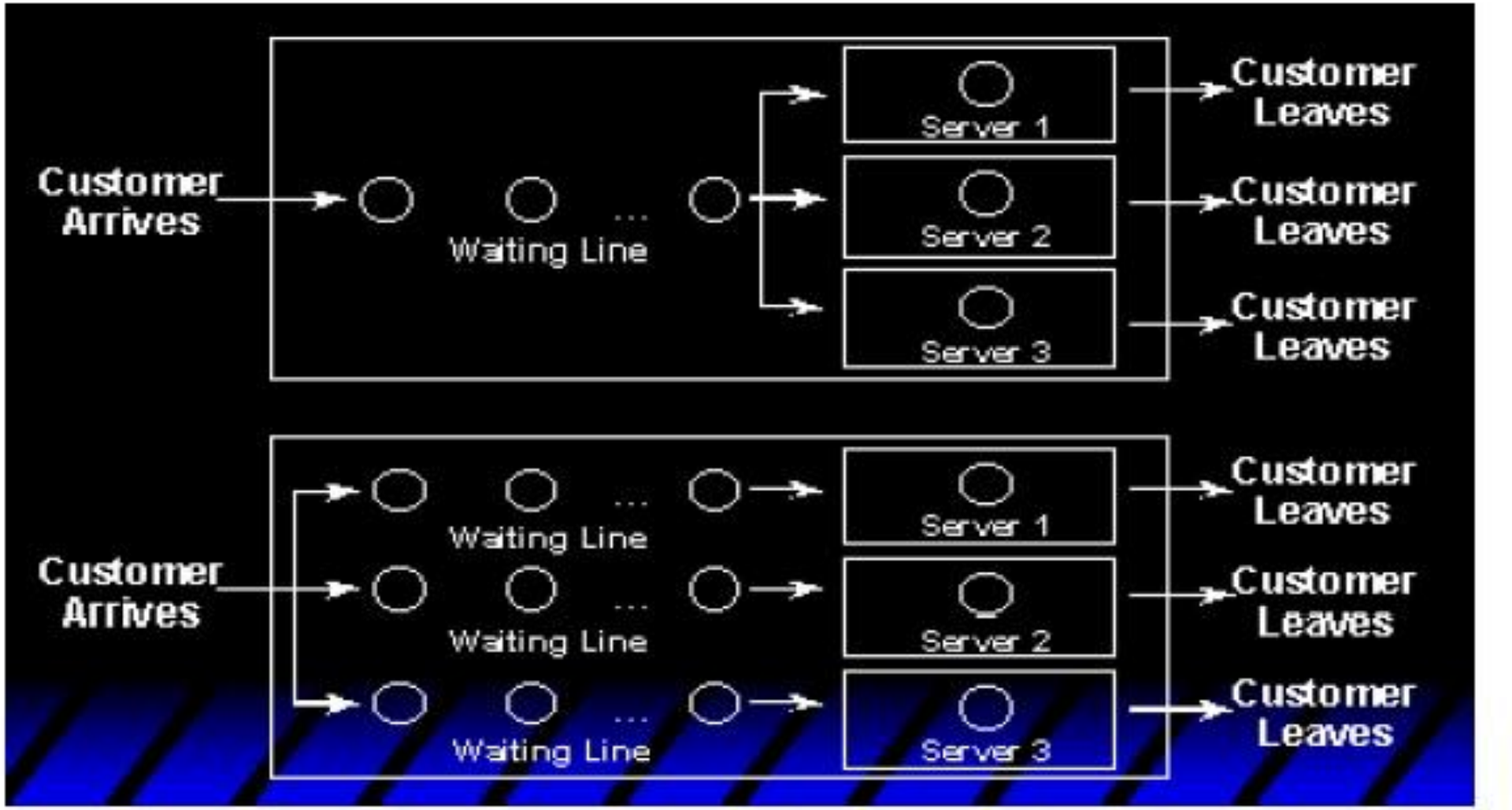


Figure: Example of Application of Queuing Theory

Examples of Some Real World Queuing System

1. **Commercial Queuing Systems:** Commercial organizations serving external customers. For example Queuing Systems in Dental Service, Bank, Garage, Gas stations, etc.
2. **Transportation Queuing Systems:** Queuing System for vehicles waiting at toll stations and traffic lights, trucks or ships waiting to be loaded, taxi cabs, etc.
3. **Business-Internal Service Systems:** Queuing systems at Inspection Stations, conveyor belts, computer support, etc.
4. **Social Service Systems:** Queuing systems at Judicial process, hospitals, waiting list for organ transplant, etc.

Characteristics of Queuing System

1. **Arrival Process:** It is a distribution that determines how the tasks arrive in a system.
2. **Service Process:** It is a distribution that determines the task processing time.
3. **Number of servers:** It is the total number of servers available to process the tasks.
4. **Queuing Discipline:**

- It is the discipline that represents the way the queue is organized.
- Queuing Discipline is the rule for inserting or removing customers to or from the queue.
- There are various discipline for inserting and removing customers to and from queue.

a. FIFO(First In First Out)

- Also called as First Come First Serve(FCFS).
- The customer that enters the queue first will be served first.

b. LIFO(Last In First Out)

- Also called as Last Come First Serve(LCFS).
- The customer that enters the queue last will be served first.

c. SIRO(Serve In Random Order)

- The customer are served in random fashion.

d. Priority Queue

- It can be viewed as a queue with various priority.
- The customer are served as per the priorities.

e. 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.

5. Number of customers: It is the number of customers waiting to be served.

Continue of Queuing System

- 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 since it is used 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:** Service represents some activity that takes time and that the customers are waiting for. 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.
- **Service Pattern:** 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** whereas systems with more servers are called **Multi Channel Systems.**

Queuing Theory

- Queuing Theory is a collection of mathematical models of various queuing systems that take inputs parameters and provide quantitative parameters describing the system performance.
- It is the mathematical study of waiting lines or queues.
- Queuing Theory refers to the mathematical models used to simulate these queues.

 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 .

Types of Queuing System

1. Single Line with Single Server Queuing System

- There is a single line of customers to be served which is served by a single server.

2. Single Line with Multiple Server Queuing System

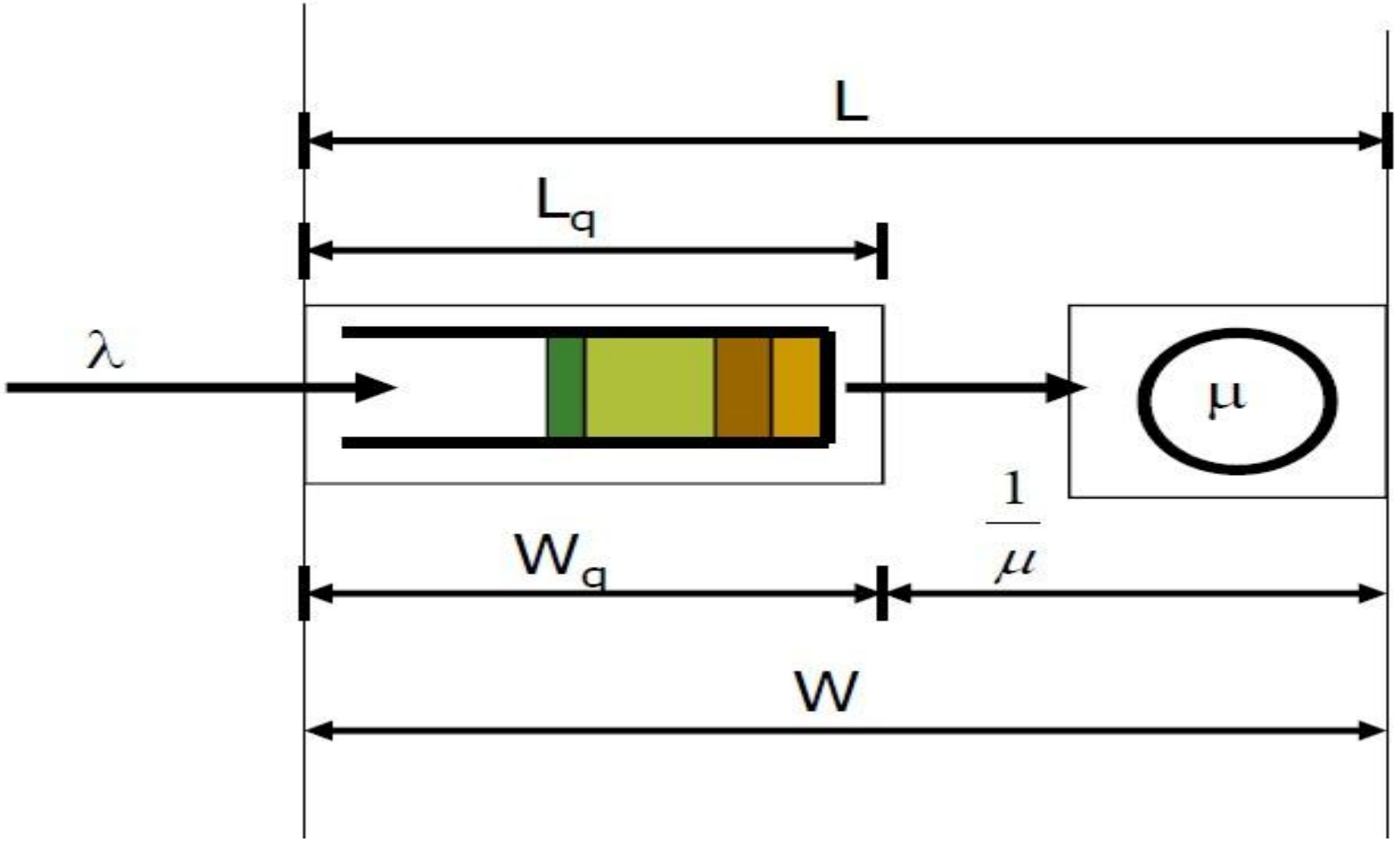
- There is a single line of customers to be served which is served by multiple or more than one server.

3. Multiple Line with Multiple Server Queuing System

- There are multiple or more than one line of customers to be served which is served by There is a single line of customers to be served which is served by a single server.

Queuing Model

M/M/1 Queuing Model



 In the above figure:

1. λ represents arrival rate of jobs
2. μ represents service rate of server
3. L represents length of the queuing system
4. Lq represents length of queue
5. W represents average waiting time in the whole system
6. Wq represents average waiting time in the queue

- M/M/1 Queue is the most widely used queue.
- This queue is used to model single processor systems or individual systems in a computer system.
- In this queue model following assumptions are made:

  -Inter-arrival rate is exponentially distributed

  - Service rate of server is exponentially distributed

  - Contains a single server

  -Follows FCFS Discipline

  - Unlimited queue length is allowed

  - Infinite number of customers

Kendall Notation

- Also called as Queuing Notation
- Six parameters are used.
- The basic format of this notation is of form: A / B / c / D / N / K

  - A represents the inter-arrival time distribution.
  - B represents the service time distribution.
  - c represents the number of parallel servers.
  - D represents the queue or service discipline.
  - N represents the maximum size of queue.
  - K represents the size of the calling population or the population size .

 The symbols used for A and B are :

a. M is the Poisson (Markovian) process

- If arrival time is Poisson Distribution, then the inter-arrival time is exponential distribution

Poisson Distribution	Exponential Distribution
Number of events in a time interval	Time between two events.
Discrete	Continuous on an interval

- So M here denotes Exponential Inter-Arrival Time Distribution

b. D is the symbol for deterministic (known) arrivals and constant service duration/ Deterministic Service Time Distribution

c. E_k represents Erlang distribution of intervals or service duration

d. G is Arbitrary or general distribution

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

f. PH (Phase type)

 If arrival time is Poisson Distribution, then the inter-arrival time is exponential distribution.

- The Kendall classification of queuing systems exists in several modifications.
- Another form is 1/2/3(/4/5/6) where:
  - 1 indicates the inter-arrival time distribution
  - 2 indicates the service time distribution
  - 3 indicates the number of servers
  - 4 indicates the maximum size of queue.
  - 5 indicates the size of the calling population or the population size
  - 6 indicates the queue or service discipline

Kendall Classification of Queuing System Examples:

1. D/M/1

- The provided notation indicates that the system has a single server with Deterministic inter-arrival time distribution and Exponential service time distribution.
- The system has unlimited population and unspecified queuing discipline.

2. M/G/3/20

- The notation indicates that the system has 3 servers with Exponential inter-arrival time distribution and has General service distribution.
- It has a maximum queue size of 20 customers and unlimited customer population can be served.

3. D/M/1/LIFO/20/510

- The notation indicates that the system has a single server with Deterministic inter-arrival time distribution and Exponential service time distribution.
- It has a maximum queue size of 20 customers and the queue follows LIFO discipline. Total of 510 customers can be served.

4. M/M/3/20/1500/FCFS

- The notation indicates that the system has a 3 servers with Exponential inter-arrival time distribution and Exponential service time distribution.
- It has a maximum queue size of 20 customers and the queue follows FCFS(FIFO) discipline. Total of 1500 customers are served.

Network of Queues

- Queuing Network are the systems in which single queues are connected by routing network.
- Network of queues are used to model queuing when a set of resources is shared.
- Such a network can be modeled by a set of service centers where each service center may contain one or more servers.
- In the study of queue networks, one queue typically tries to obtain the distribution of the network.
- In network of queues, when a customer is connected to one node it can join another node or queue for service or can leave the network.
- For a network of m-nodes, the states of the system can be described by m-dimensional vector.
- Network of queues is widely used in the field of telecommunication, computer network, etc.

Measurement of System Performance

- It is the analysis and measurement of how well the queuing system performs.
- The various parameters used for measuring the system performance are:

1. Average number of customers in the system or in the queue

- 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.

2. Number of servers
3. Average waiting time of the customers in the queue or in the system.

- The knowledge of average waiting time in the queue is necessary for determining the cost of waiting in the queue.

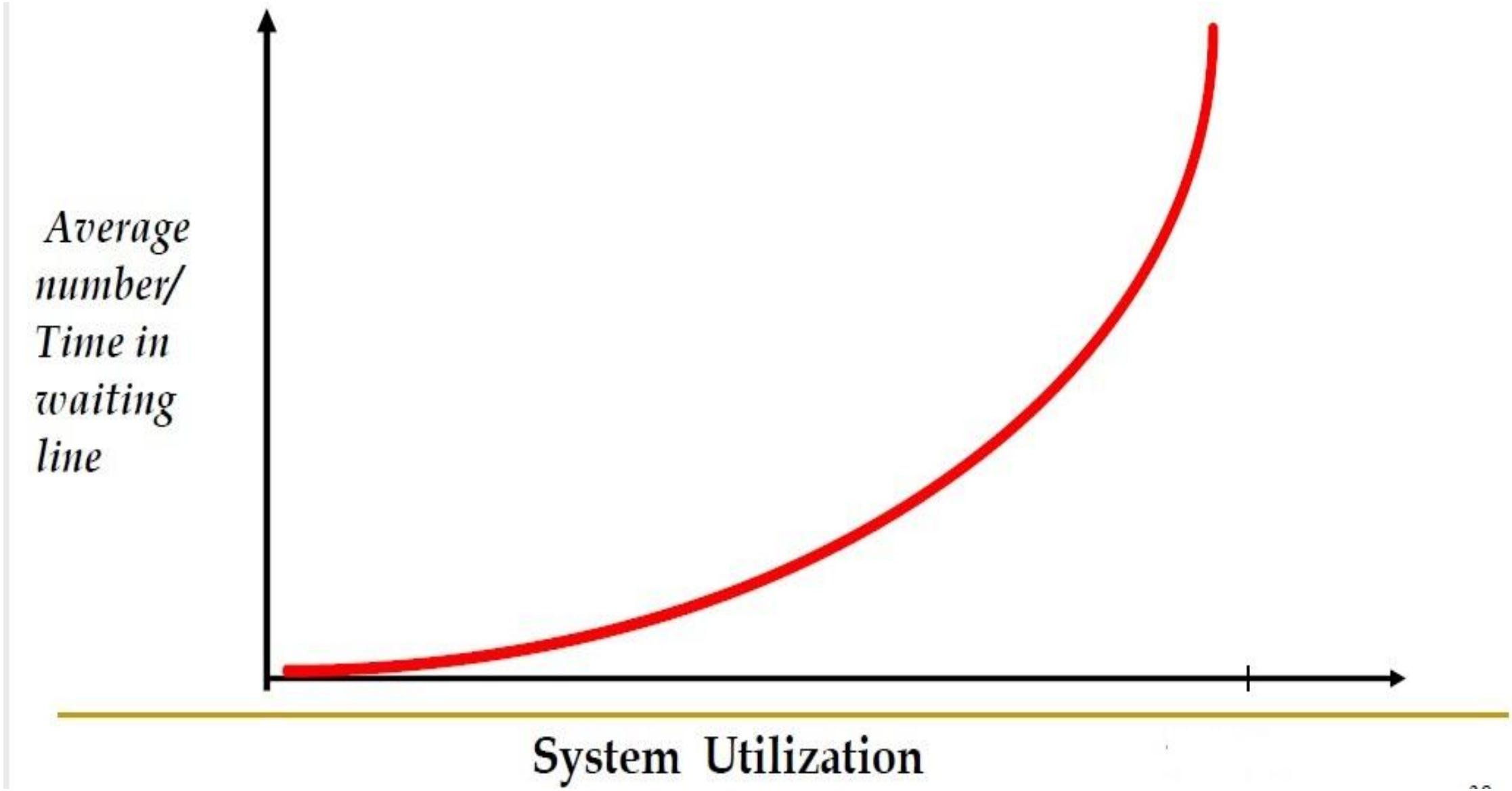
4. Length/Size of queue
5. The cost of waiting/idle time
6. System/Server Utilization

System/Server Utilization of Queuing System

- System/Server Utilization is defined as the extent to which a system/server is busy rather than idle.
- It is defined as the percentage of time during which system/server is busy processing jobs during system.
- System utilization factor denoted by S is the ratio of average arrival rate to the average service rate (μ).

$$S = \frac{\lambda}{\mu}$$

- In the case of n-server model, $S = \frac{\lambda}{n\mu}$
- Under the normal situations, 100% system utilization is not a realistic goal.
- 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 figure below.


Average number/ Time in waiting line
System Utilization

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.

- 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.
- According to the scenario given, arrival pattern of machine is:

Arrival pattern:

On 50%of the days arrival=0

On 30%of the days arrival=1

On 20%of the days arrival=2

- So the expected arrival rate can be calculated as:

Expected Arrival Rate = 0*0.5+0.3*1+0.2*2

= 0.7 per day

- Again according question, Service pattern is:

65% machines in 1 day

30% machines in 2 days

5% machines in 3 days

- So the expected service time = 0.65*1+0.3*2+0.05*3

= 1.4 days

- Hence the expected service rate is = $\frac{1}{1.4} = 0.714$ 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<=1.0 Arrivals=2

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

0.0<r<=0.65 ST=1day

0.65<r<=0.95 ST=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=4/30* 100=13.33 %

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

Timer	Random Number	Arrivals	Queue	Random Number	Service Time	Idle Time	Waiting Time
0		0	0		0	0	0
1	.273	0	0		0	0	0
2	.962	2	1	.437	1	1	1
3	.570	1	1	.718	2	1	2
4	.435	0	1		1	1	3
5	.397	0	0	.315	1	1	3
6	.166	0	0		0	2	3
7	.534	1	0	.964	3	2	3
8	.901	2	2		2	2	5
9	.727	1	3		1	2	8
10	.158	0	2	.327	1	2	10
11	.720	1	2	.776	2	2	12
12	.569	1	3		1	2	15
13	.308	0	2	.110	1	2	70
14	.871	2	3	.469	1	2	20
15	.678	1	3	.462	1	3	23
16	.470	0	2	.631	1	2	25
17	.794	1	2	.146	1	2	27
18	.263	0	1	.801	2	2	28
19	.065	0	1		1	2	29
20	.027	0	0	.86	1	2	29
21	.441	0	0		0	3	29
22	.152	0	0		0	4	29
23	.998	2	1	.160	1	4	30
24	.508	1	1	.889	2	4	31
25	.771	1	2		1	4	33
26	.115	0	1	.538	1	4	3 4
27	.484	0	0	.989	3	4	34
28	.700	1	1		2	4	35
29	.544	1	2		1	4	37
30	.903	2	3	.813	2	4	40
		21	40				

Important Formula for Numerical

1. System/Server Utilization(S) $= \frac{average\ arrival\ rate\ (\lambda)}{average\ service\ rate\ (\mu)}$
2. Fraction time Busy = S
3. Fraction Time idle = 1-S
4. Average Waiting Time $= \frac{S}{\mu - \lambda}$
5. Average Number of Customers in System(N) $= \frac{Fraction\ Time\ Busy}{Fraction\ Time\ Idle} = \frac{S}{1-S}$
6. Average Time Customer Spends in time(T) $= \frac{N}{\lambda}$
7. Probability of Zero Customer(P_0) = 1 - S
8. Probability of n Customer(P_n) = S^nP_0 for n>0

Numerical 1

Consider a database system with an average service time of 450 msec. As database requests are initiated by large number of clients, a random arrival pattern may be assumed. Thus the arrival process is assumed to be Poisson. On the average, a new database query arrives every 500 msec. Service time are assumed to be exponentially distributed, the queuing discipline is assumed to follow FCFS pattern. Calculate:

1. System Utilization
2. Fraction time busy
3. Fraction time idle
4. Average Waiting time
5. Average Number of customers in system
6. Average time customers spends in the system

Numerical 2

In a petrol pump, Customer arrival time is given by Poisson Distribution with an arrival rate of 2 customer/hour and they get exponentially served at the rate of 3 customer/hour. Find:

1. System Utilization
2. Probability of Zero Customer
3. Probability of 1 Customer
4. Probability of 4 or more Customers
5. Average Waiting time
6. Average Number of customers in system
7. Average time customers spends in the system

Hint for No. 4

Probability of 4 or more customers = 1 – probability of zero customers – probability of one customer – probability of two customers – probability of three customers

i.e. $P_{cust>=4} = 1 - P_{cust=0} - P_{cust=1} - P_{cust=2} - P_{cust=3}$