SIMULATION OF CONTINUOUS SYSTEMS

UNIT 3

Queuing system:

- **Queuing system are the waiting lines in which the system attribute are waiting for a service.**
- The queue may be of the customer waiting for the **server** or **server waiting** for customer.
- The waiting line situation arises either there is too much demand on the service facility so that customer have to wait for getting service or there is too less demand in which service facility have to wait for the customer.

Queuing system:

- The line where the entities or customers wait is generally known as queue.
- The combination of all entities in system being served and being waiting for services will be called a queuing system.
- The general diagram of queuing system can be shown as a queuing system involves customers arriving at a constant or variable time rate for service at a service station.
- Customers can be students waiting for registration in college, airplane queuing for landing at airfield, or jobs waiting in machines shop.
- ***They remain in queue till they are provided the service. Sometimes queue being too long, they will leave the queue and go, it results a loss of customer.***

Queuing system:

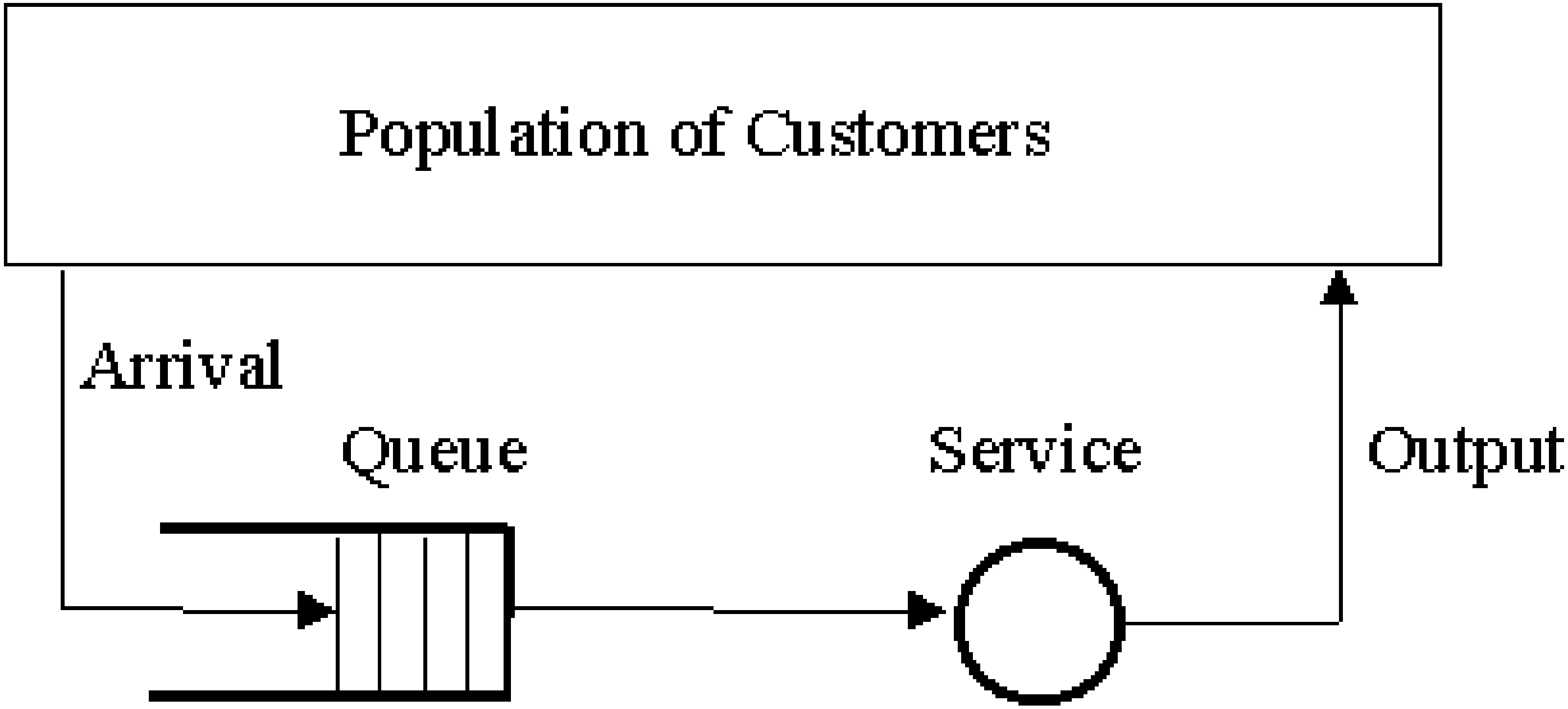


Figure 1

Queuing system:

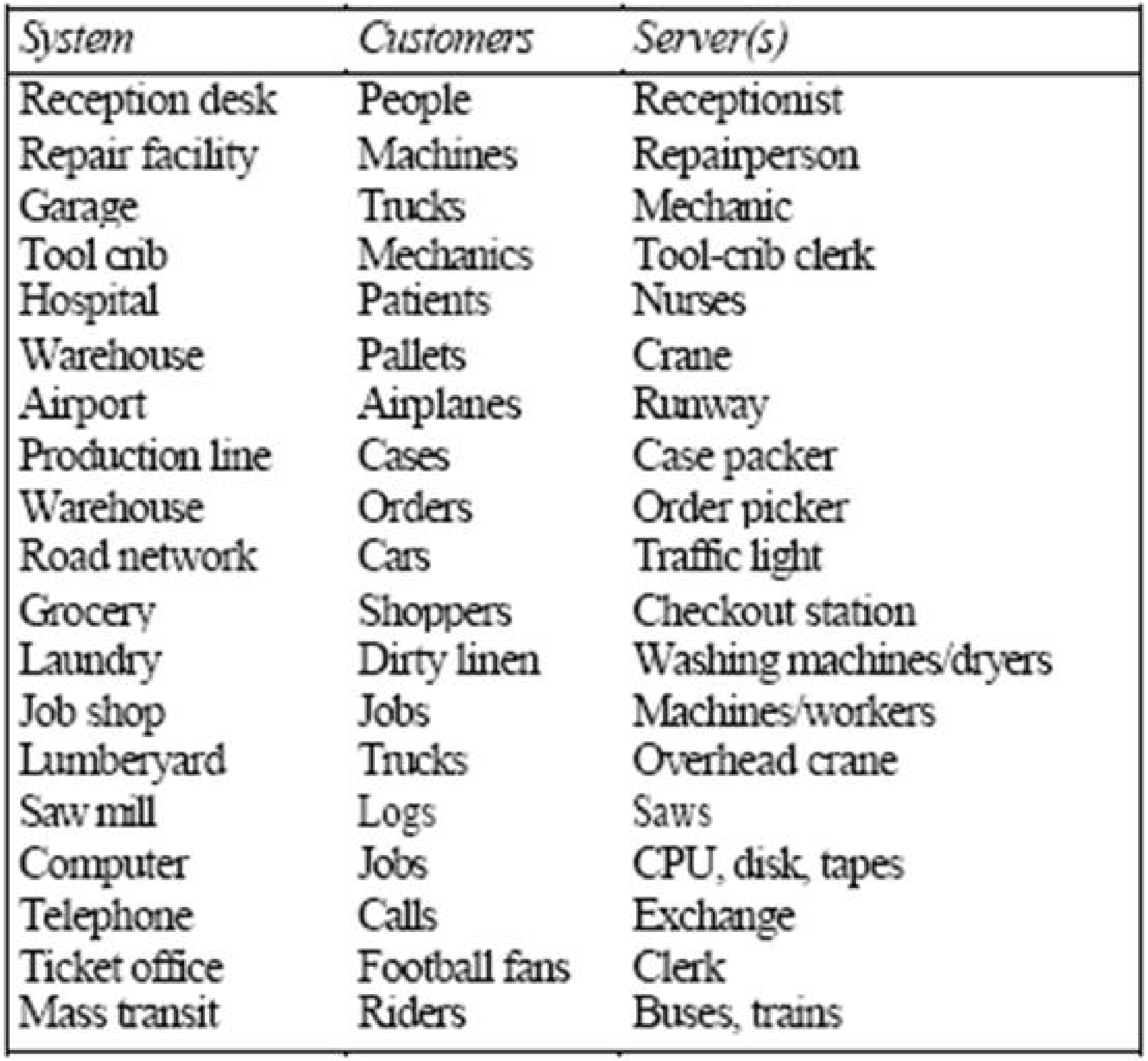
Queuing system:

- The basic concept of queuing theory is the optimization of **wait time**, **queue length**, and the **service available** to those standing in a queue.
- Cost is one of the important factors in the queuing problem.
- Waiting in queues incur cost, whether human are waiting for services or machines waiting in a machine shop. On the other hand if service counter is waiting for customers that also involves cost.
- **In order to reduce queue length, extra service centers are to be provided but for extra service centers, cost of service becomes higher.**

Characteristics of Queuing Systems

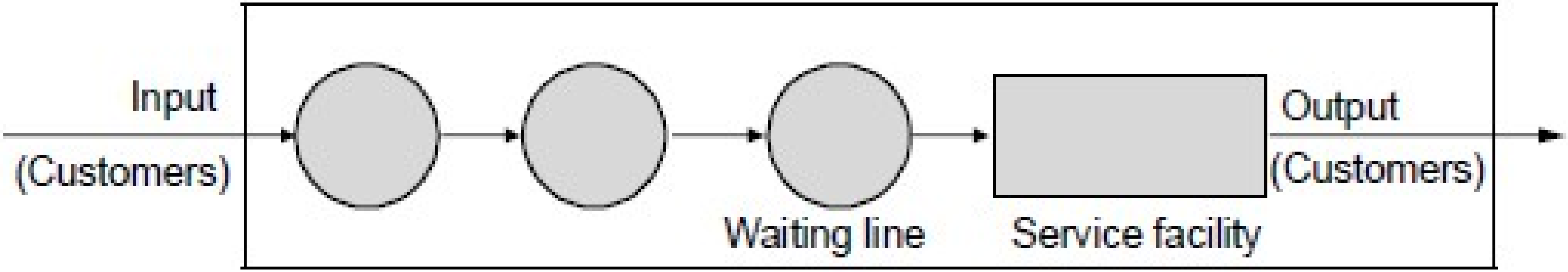
- The key elements, of a queuing system are the ***customers*** and ***servers***. The term "***customer***" can refer to people, machines, trucks, mechanics, patients— anything that arrives at a facility and requires service.
- The term "***server***" might refer to receptionists, repairpersons, CPUs in a computer, or washing machines….any resource (person, machine, etc. which provides the requested service.
- Table 1 lists a number of different queuing systems.

Table 1: Examples of Queuing Systems



System	*Customers*	*Server(s)*
Reception desk	People	Receptionist
Repair facility	Machines	Repairperson
Garage	Trucks	Mechanic
Tool crib	Mechanics	Tool-crib clerk
Hospital	Patients	Nurses
Warehouse	Pallets	Crane
Airport	Airplanes	Runway
Production line	Cases	Case packer
Warehouse	Orders	Order picker
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Laundry	Dirty linen	Washing machines/dryers
Job shop	Jobs	Machines/workers
Lumberyard	Trucks	Overhead crane
Saw mill	Logs	Saws
Computer	Jobs	CPU, disk, tapes
Telephone	Calls	Exchange
Ticket office	Football fans	Clerk
Mass transit	Riders	Buses, trains

Elements of queuing system:



Elements of queuing system:

1 . Population of customer:

- **Customer are the entities who wants service from the server. It can be considered either limited (closed system) or unlimited (open system).**
- In systems with a large population of potential customers, the calling population is usually assumed to be finite or infinite. Examples of infinite populations include the potential customers of a restaurant, bank, etc.
- **The main difference between finite and infinite population models is how the arrival rate is defined.**
- In an infinite-population model, the arrival rate is not affected by the number of customers who have left the calling population and joined the queuing system.
- On the other hand, for finite calling population models, the arrival rate to the queuing system does depend on the number of customers being served and waiting.

Elements of queuing system:

2 . Arrival:

- It defines the way that customer enter the system.
- **Mostly arrivals are random with random intervals between two adjacent parameters.**
- Typically the arrival is described by random distribution of intervals also called arrival pattern.
- Arrival process for infinite-population models is usually characterized in terms of inter arrival times of successive customers. Arrivals may occur at scheduled times or at random times. When at random times, the inter arrival times are usually characterized by a probability distribution. The most important model for random arrivals is the Poisson arrival process.

Elements of queuing system:

3 . Queue or waiting line:

- It especially represents a certain number of customers waiting for service. Two important properties of queue are:
  - Maximum size
  - Queuing discipline
  - Maximum size is the maximum number of customers that may be waiting in a queue.
  - Common queue disciplines include first-in, first-out (FIFO); last-in first out (LIFO); service in random order (SIRO); shortest processing time first |(SPT) and service according to priority (PR).

Elements of queuing system:

4 . The service times and service mechanism

- **It represents some activity that takes time and that the customers are waiting for.**
- It may be not only be real service carried on person or machines but it may also be CPU time slice, connection created for telephone calls.

Elements of queuing system:

4 . The service times and service mechanism….

Theoretical models are based on random distribution of service duration also called service patterns.

- System with one server is single channel system and with more servers in multichannel servers.
- The service times of successive arrivals are denoted by S1, S2, S3, .. . They may be constant or of random duration.
- Sometimes services may be identically distributed for all customers of a given type or class or priority, while customers of different types may have completely different service-time distributions. In addition, in some systems, service times depend upon the time of day or the length of the waiting line. For example, servers may work faster than usual when the waiting line is long, thus effectively reducing the service times.
- Each service center consists of some number of servers, c, and working in parallel; that is, upon getting to the head of the line, a customer takes the first available server. Parallel service mechanisms are either single **server (c = 1), multiple server (l < c < ∞**), or **unlimited servers (c = ∞**). A self-service facility is usually characterized as having an unlimited number of servers.

Elements of queuing system:

5 . 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.

Application of queuing system:

- Telecommunication
- Traffic control
- Computer process
- Manufacturing process

Queuing discipline:

- It explains how the customer is solved by the server or the way in which queue is organized. It is the rule by which customer enters and exits the queue. Some queuing discipline are
  - FIFO – First In First Out
  - LIFO – Last In First Out
  - SIRO – Serial In Random Out
  - SPTF – Shortest Processing Time First
  - PR – Service According to Priority

Queuing Notation

- Recognizing the diversity of queuing systems, Kendall [1953] proposed a notational system for parallel server systems which has been widely adopted.
- Kendall classify a queuing notation system as

$$A\ /\ B\ /\ s\ /\ q\ /\ c\ /\ P$$

Where,

A is the Arrival pattern

B is the Service pattern

s is No. of server

q is Queuing discipline

c is System capacity

P is Population size

Queuing Notation..

Arrival and service pattern uses the following notations.

D − *Deterministic or constant*

G − *General Distribution*

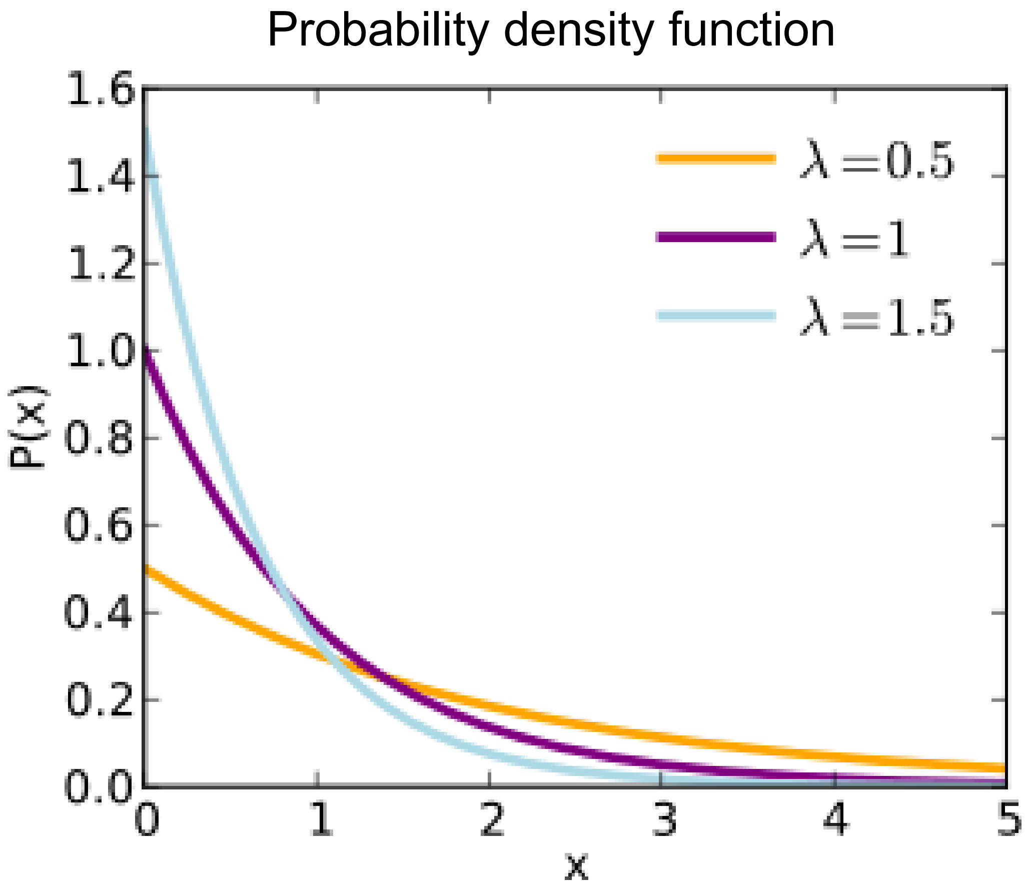
GI − *General Distribution with Independent Random Values*

M − *Poisson (Markovian) process or Exponential Distribution*

Em − *Erlang Distribution*

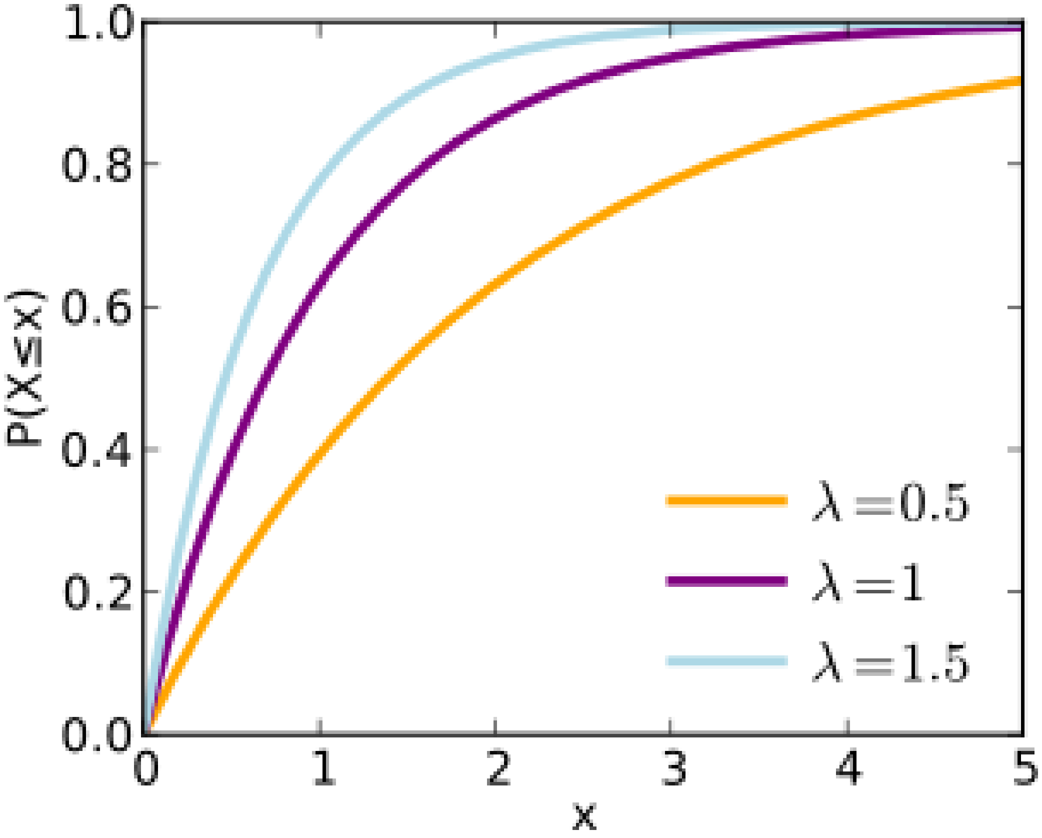
H − *Hyper exponential distribution*

ExponentiaDistribution

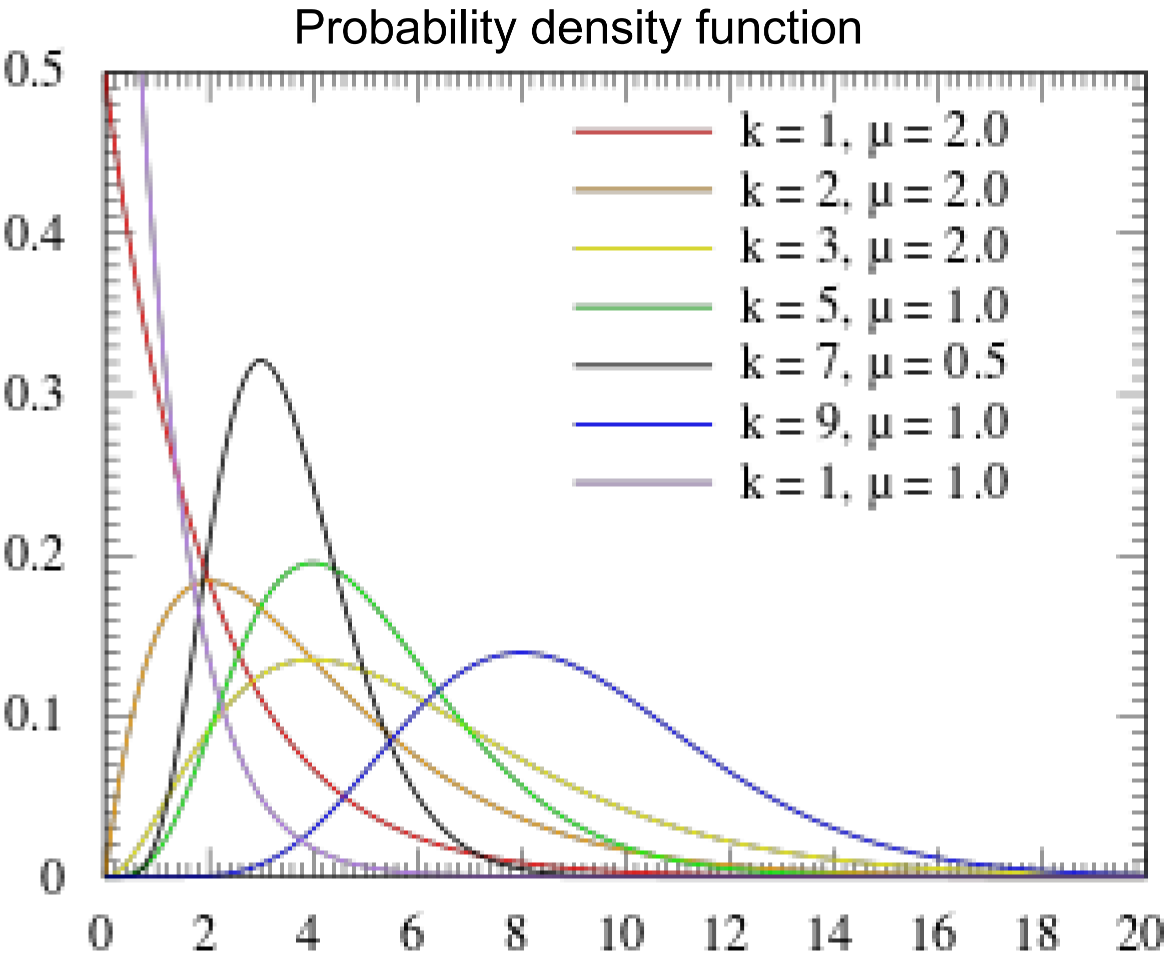


ExponentiaDistribution

Cumulative distribution function

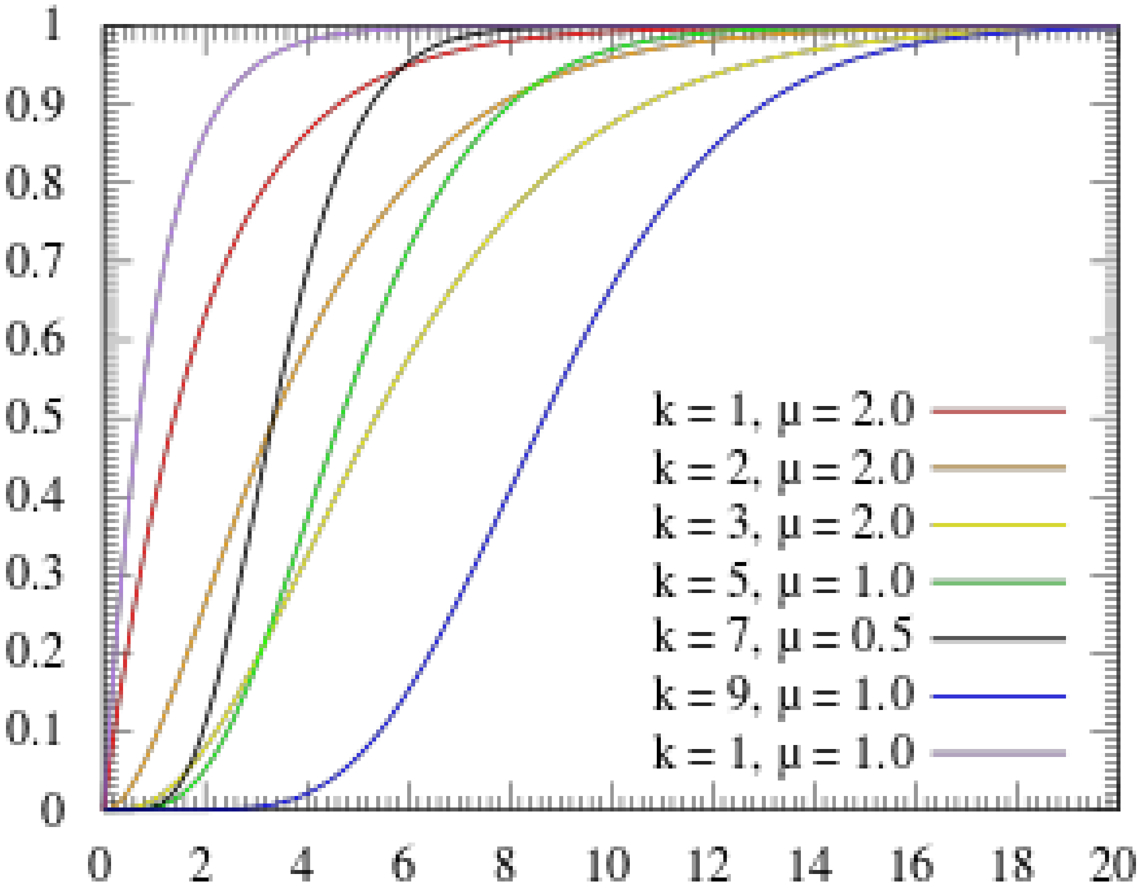


ErlangDistribution



ErlangDistribution

Cumulative distribution function



Queuing Notation..

Example:

1. D / M / 1

$Arrival\ pattern - Deterministic$

$Service\ pattern - Exponential\ Distribution$

$No.\ of\ Server - 1$

$Queuing\ discipline - FIFO$

$System\ capacity - Infinite$

$Population\ size - Infinite$

Note : systems will be assumed to have a FIFO queue discipline.

Queuing Notation..

Example:

2. M **/** D **/ 2 /** $LIFO$

Arrival pattern − Exponential Distribution

Service pattern − Deterministic

No. of Server − 2

Queuing discipline − LIFO

System capacity − Infinite

Population size − Infinite

Queuing Notation..

Example:

3. G / E_m / 1 / 20

$Arrival\ pattern - General\ Distribution$

$Service\ pattern - Erlang\ Distribution$

$No.\ of\ Server - 1$

$Queuing\ discipline - FIFO$

$System\ capacity - 20$

$Population\ size - Infinite$

Queuing Notation..

Example:

4. M / M / 1 / œ /œ

indicates a single-server system that has unlimited queue capacity and an infinite population of potential arrivals. The inter arrival times and service times are exponentially distributed. When C and P are infinite, they may be dropped from the notation. For example, M / M / 1 / œ / œ is often shortened to M/M/l.

Queuing Notation..

Example

a) **M/D/2/5/∞** stands for a queuing system having exponential arrival times, deterministic service time, 2 servers, capacity of 5 customers, and infinite population.

b) If notation is given as **M/D/2** means exponential arrival time, deterministic service time, 2 servers, infinite service capacity, and infinite population.

Queuing Notation..

Examples:

i. **D/M/1** =

Deterministic (known) input, one exponential server, one unlimited FIFO or unspecified queue, unlimited customer population.

i. **M/G/3/20** =

Poisson input, three servers with any distribution, maximum number of customers 20, 32 unlimited customer population.

ii. **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.

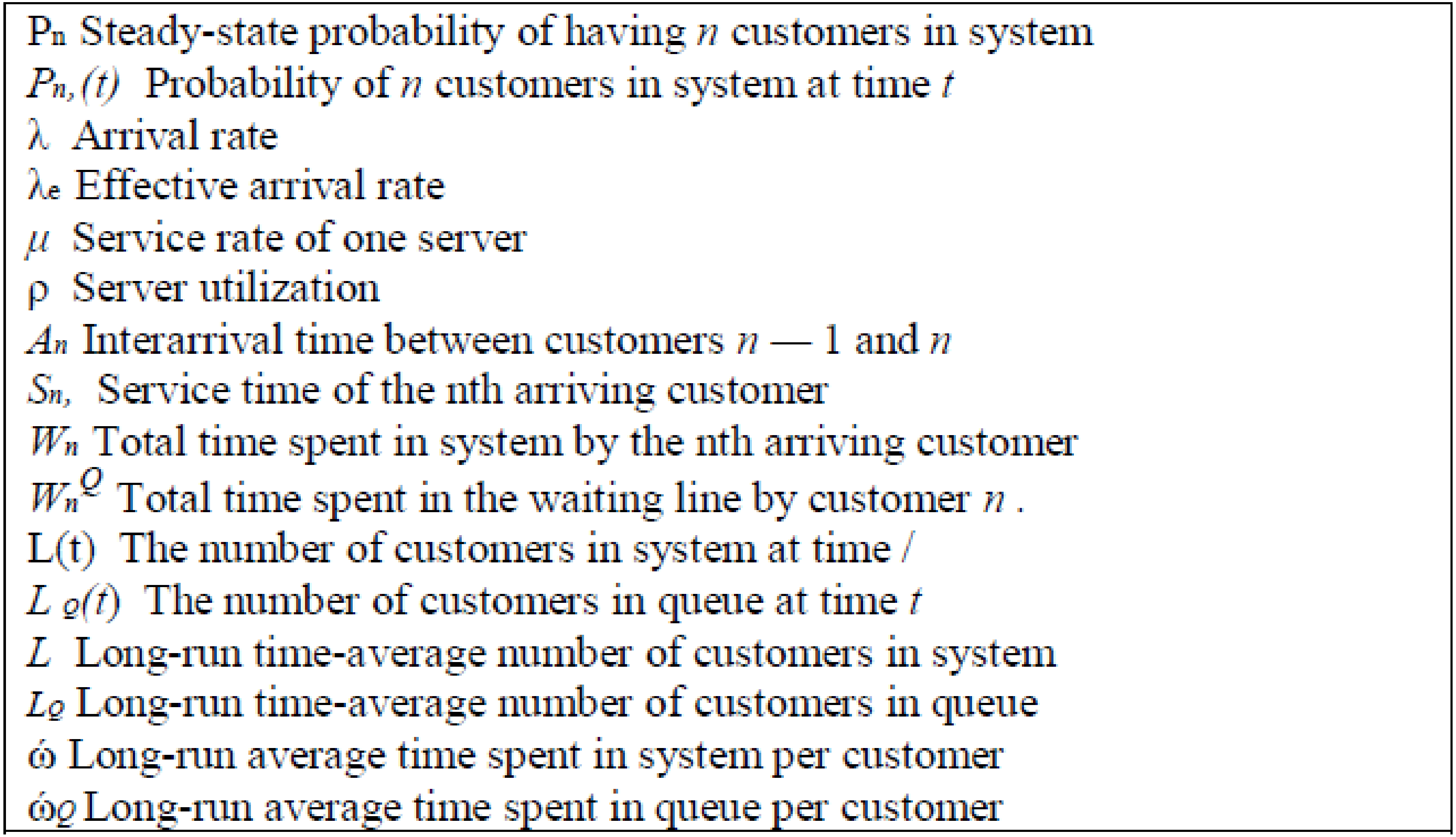
Queuing Notation..

Examples:

1. D / D / 2 / LIFO
2. D / M / 1 / 2
3. Gi / H / 2 / SIRO / ∞ / 20
4. D / G / 3 / 20
5. H / Em / 2 / FIFO / 15 / 20
6. Gi / G / 4
7. D / M / 1 / 2 / 30
8. Gi / H / 2 / LIFO / 20

Queuing Notation..

Table: Queuing Notation for Parallel Server Systems:

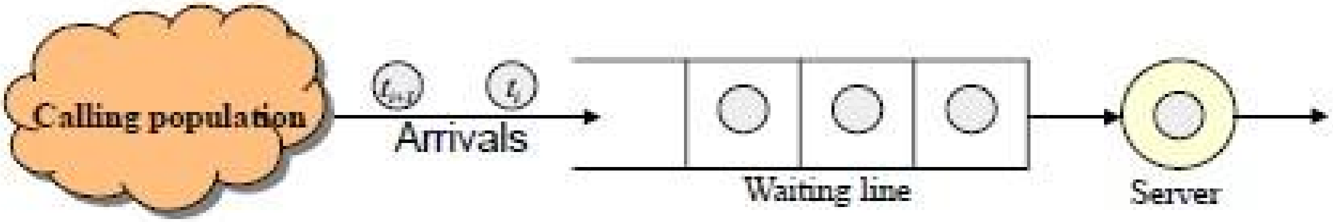


P_n	Steady-state probability of having n customers in system
$P_n,(t)$	Probability of n customers in system at time t
λ	Arrival rate
λ_e	Effective arrival rate
μ	Service rate of one server
ρ	Server utilization
A_n	Interarrival time between customers n — 1 and n
S_n,	Service time of the nth arriving customer
W_n	Total time spent in system by the nth arriving customer
W_n^Q	Total time spent in the waiting line by customer n .
L(t)	The number of customers in system at time /
$L_Q(t)$	The number of customers in queue at time t
L	Long-run time-average number of customers in system
L_Q	Long-run time-average number of customers in queue
ώ	Long-run average time spent in system per customer
$ώ_Q$	Long-run average time spent in queue per customer

Simulation of queuing system

Queuing system state:

- System
  - Server
  - Units (in queue or being served)
  - Clock
- State of the system
  - Number of units in the system
  - Status of server (idle, busy)
- Events
  - Arrival of a unit
  - Departure of a unit



Queuing Model

- **Single Server**
- **Multiple Server**

Within these single server and multiple server there are two types:

- **Finite queue length** : restriction in queue length
- **Infinite queue length** : no restriction in queue length

Queuing Model

Balking:

Balking is a queue behavior wherein people leave as soon as they realize that they will to wait.

So **if an arrival doesn't join the system and leave is said to be Balking.**

Balking can also be two types

- Forced balking
- Unforced balking

Reneging:

Reneging refer to a queue behavior wherein **people leave a queue after they are tired of waiting .**

Retrial or Jockeying Queue:

As the name suggests, this particular queue behavior refer to **customers' response to rejoin a queue that they had left earlier due to balking or reneging.**

Queuing Model

Polling:

When there are more than one queue forming (establishing) for same service, the action of sharing service between the queues is called polling.

A bus picking up passengers from different stoppage along its route is an example of polling service.

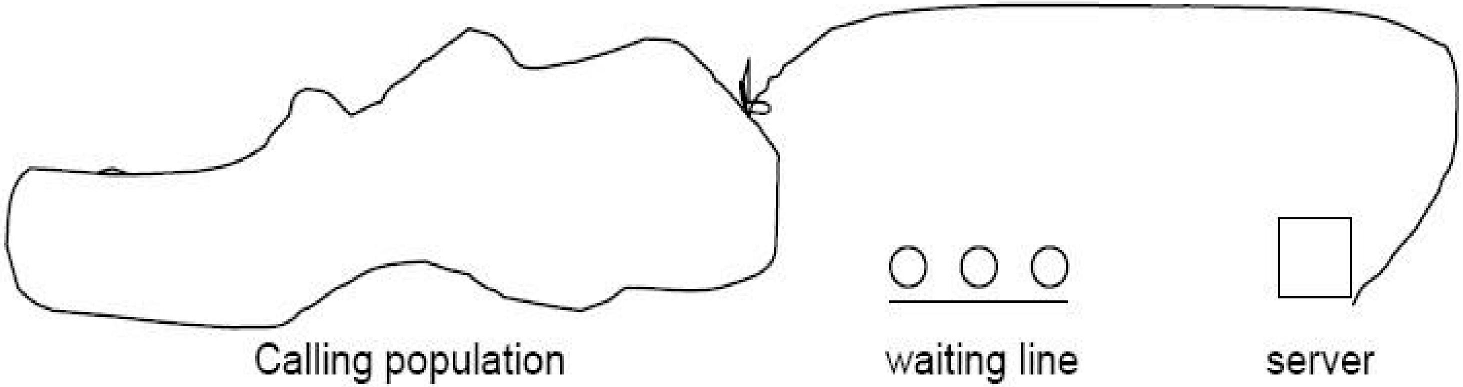
Separate queue for ladies and gents at ticket window, is another example of polling service.

Queuing Model

1) Single server queue:

A queuing system is described by its calling population, the nature of arrivals, the service mechanism, system capacity and the queuing discipline.

A single channel queuing system is portrayed in fig below.



Queuing Model

So in a **single server queue**,

- **Calling population is infinite**
  - Arrival rate does not change
- **Units are served according to FIFO**
- Arrivals are defined by the distribution of time between arrivals
  - Inter-arrival time
- Service time are according to distribution
- **Arrival rate must be less than service rate**
  - Stable system
- Otherwise waiting line will grow unbounded
  - Unstable system

Queuing Model

Arrival event:

- If server idle unit gets service, otherwise unit enters queue.

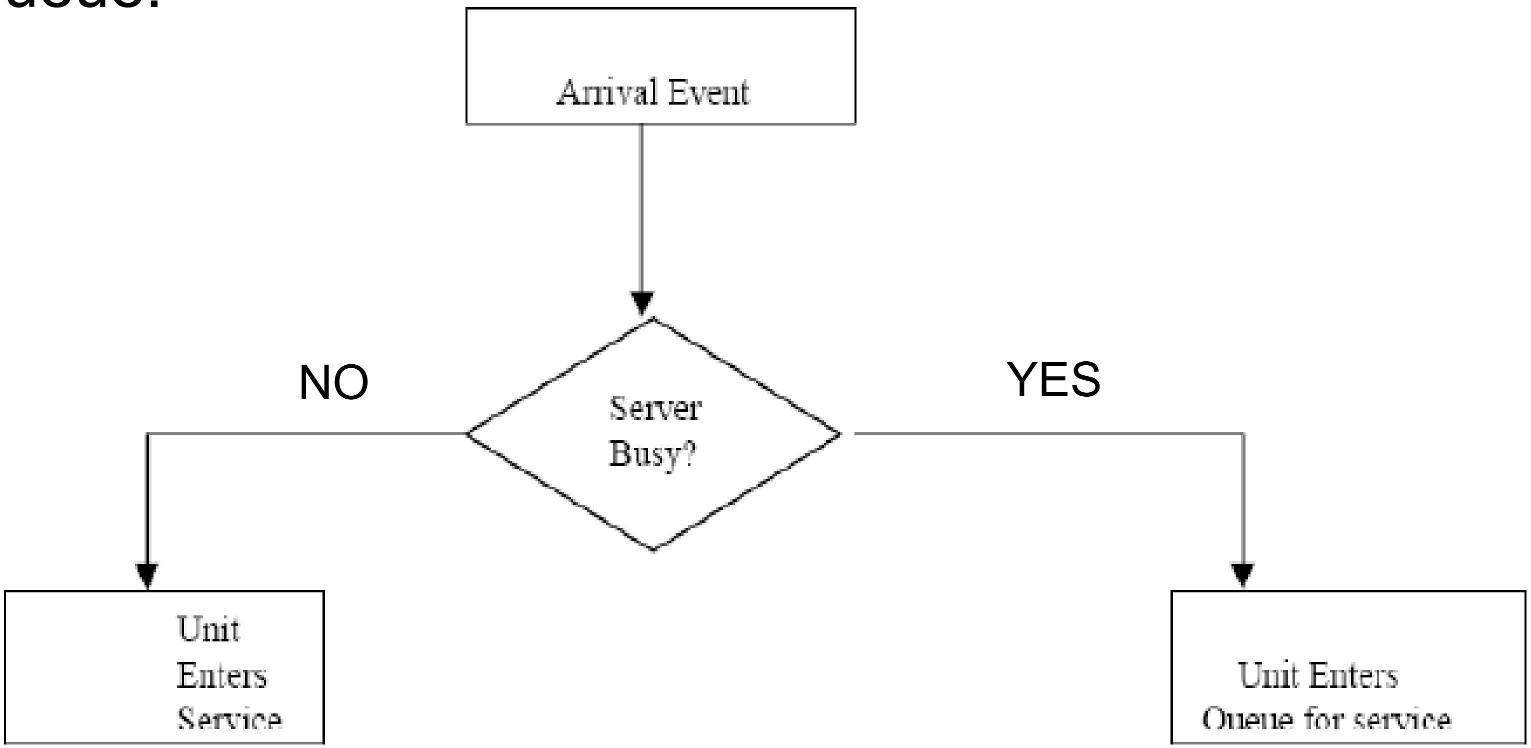


Fig: Unit entering system-flow diagram

Queuing Model

Departure event:

- If queue is not empty begin servicing next unit, otherwise service will be idle.

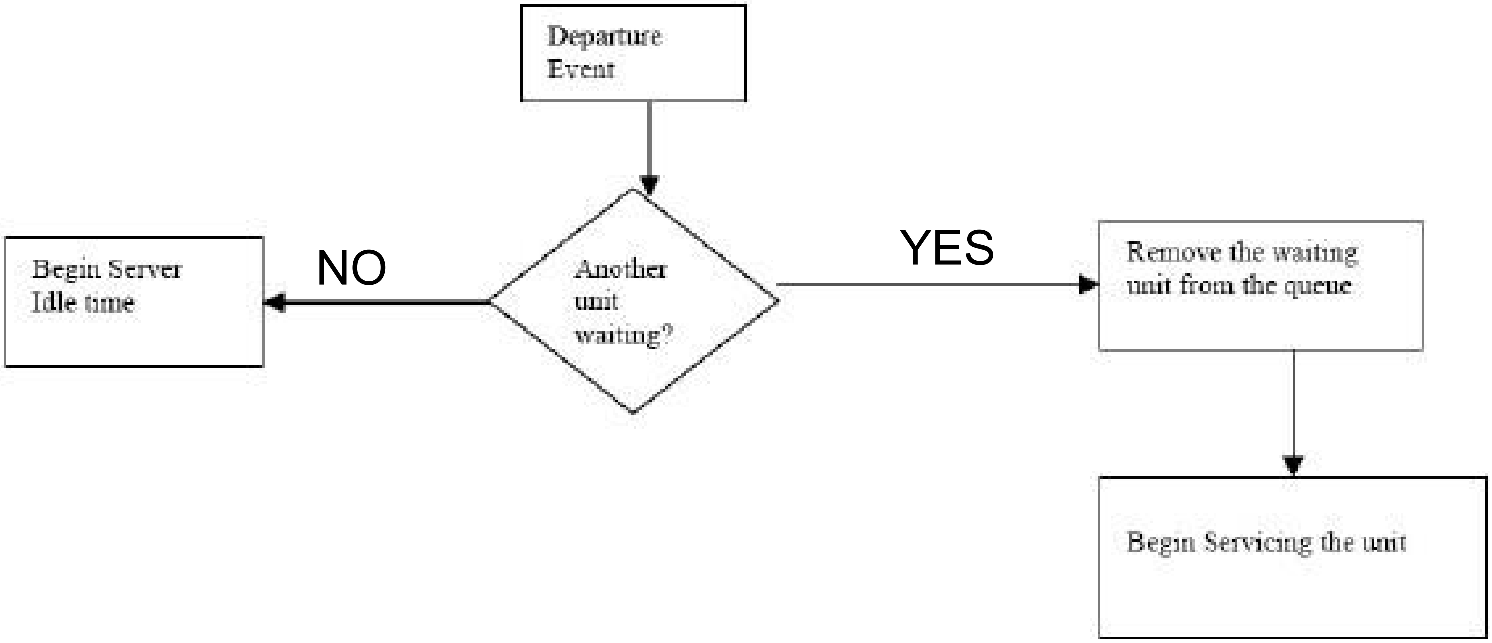
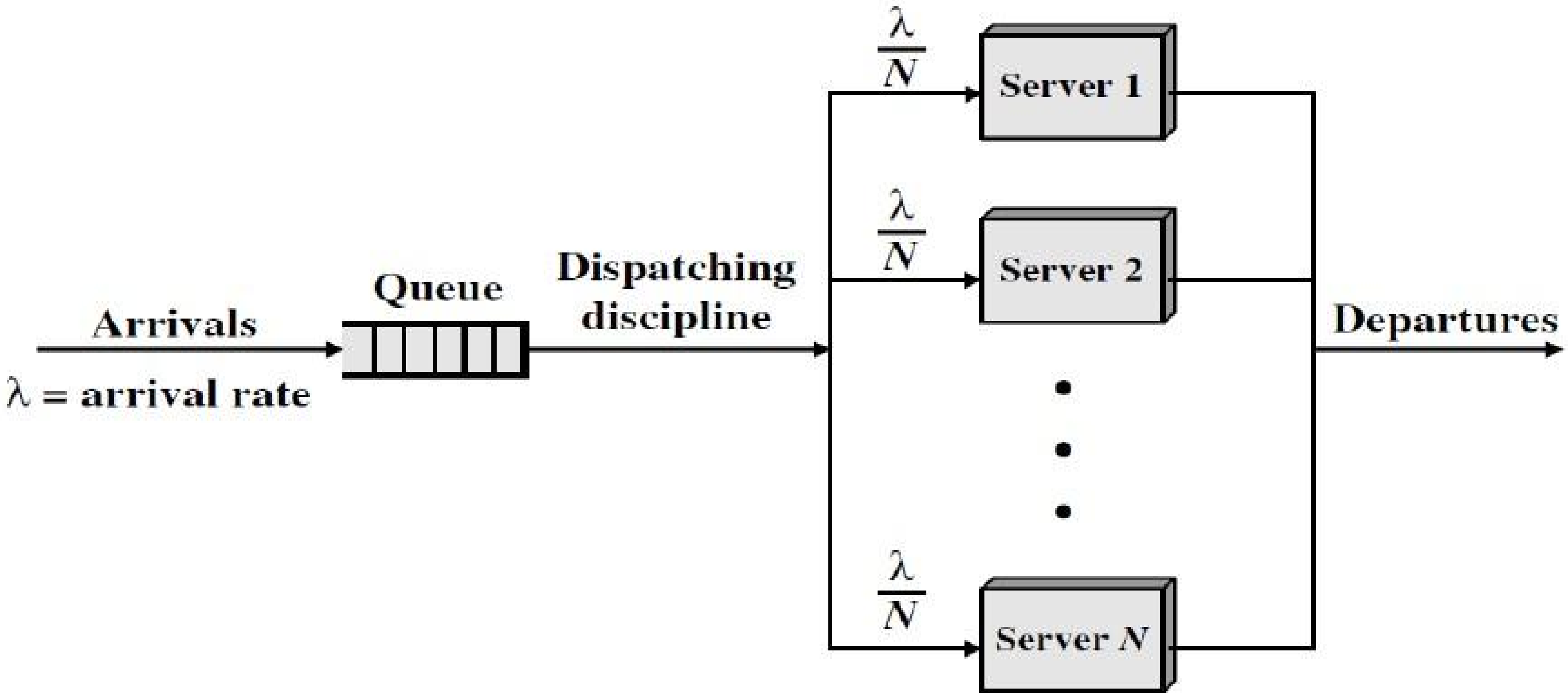


Fig: Service just-completed flow diagram

Queuing Model

2) Multi-server Queue



Queuing Model

Multi-server Queue

Figure shows a generalization of the simple model we have been discussing for multiple servers, all sharing a common queue.

If an item arrives and at least one server is available, then the item is immediately dispatched to that server.

It is assumed that all servers are identical; thus, if more than one server is available, it makes no difference which server is chosen for the item.

If all servers are busy, a queue begins to form. As soon as one server becomes free, an item is dispatched from the queue using the dispatching discipline in force.

Multi-server Queue…

The key characteristics typically chosen for the multi-server queue correspond to those for the single-server queue.

That is, we assume an infinite population and an infinite queue size, with a single infinite queue shared among all servers.

Unless otherwise stated, the dispatching discipline is FIFO.

For the multi-server case, if all servers are assumed identical, the selection of a particular server for a waiting item has no effect on service time.

Queuing Model

Multi-server Queue

The total server utilization in case of Multi-server queue for N server system is

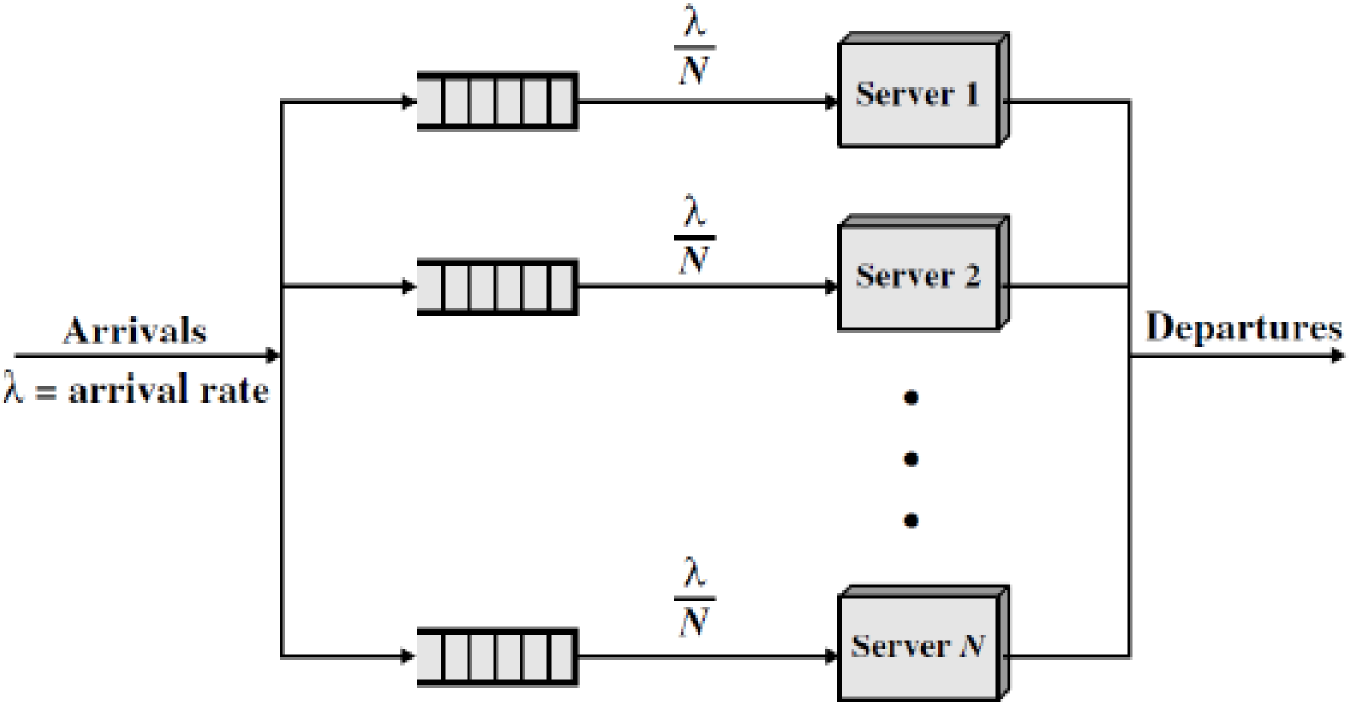
(server utilization) $\rho = \lambda / \mu$

Where μ is the service rate and λ is the arrival rate.

Queuing Model

Multi-server Queue

There is another concept which is called multiple single server queue system as shown below



Queuing Model

Some notation or Formula used to Measure the different parameter of queue

Two principal measures of queuing system are;

1. The mean number of customers waiting and
2. The mean time the customer spend waiting

Both these quantities may refer to the total number of entities in the system, those waiting and those being served or they may refer only to customer in the waiting line.

Queuing Model

Average number of customers in the System $\overline{L_S} = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\lambda}{\mu-\lambda}$

Average number of customers in the Queue $\overline{L_Q}$

= Average number of customers in the System − Server Utilization

$$= \overline{L_S} - \frac{\lambda}{\mu} = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Queuing Model

$$\text{Average waiting time in the System } \overline{W_S} = \frac{\textit{Average number of customer in the system}}{\textit{Mean arrival rate}}$$

$$= \frac{\overline{L_S}}{\lambda} = \frac{\frac{\lambda}{\mu-\lambda}}{\lambda} = \frac{1}{\mu-\lambda}$$

$$\text{Average waiting time in the Queue } \overline{W_Q} = \frac{\textit{Average number of customer in the Queue}}{\textit{Mean arrival rate}}$$

$$= \frac{\overline{L_Q}}{\lambda} = \frac{\frac{\lambda^2}{\mu(\mu-\lambda)}}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Queuing Model

Example 1

Q.N > At the ticket counter of football stadium, people come in queue and purchase tickets. Arrival rate of customers is 1/min. It takes at the average 20 seconds to purchase the ticket.

(a) If a sport fan arrives 2 minutes before the game starts and if he takes exactly 1.5 minutes to reach the correct seat after he purchases a ticket, can the sport fan expects to be seated for the kick-off?

Solution:

(a) A minute is used as unit of time. Since ticket is disbursed in 20 seconds, this means, three customers enter the stadium per minute, that is service rate is 3 per minute.

Therefore,

λ = 1 arrival/min

μ = 3 arrivals/min

$\boldsymbol{WS}$ = waiting time in the system = 1/(μ- λ)=0.5 minutes

The average time to get the ticket plus the time to reach the correct seat is 2 minutes exactly, so the sports fan can expect to be seated for the kick-off.

Queuing Model

Example 2

Q.N > At the Bank counter, people come in queue for service. Arrival rate of customers is 2/min. It takes at the average 15 seconds to take service.

(a) If bank will close 2 minutes and if he takes exactly 1 minutes to reach the door, can the customer leave bank in time?

Solution:

(a) A minute is used as unit of time. Therefore,

λ = 2 arrival/min

μ = 4 arrivals/min

$\boldsymbol{WS}$ = waiting time in the system=1/(μ- λ)= 1/(4-2) = 0.5 minutes

- Now total time will be 1+0.5 = 1.5 minutes.

Queuing Model

Example 3

Q.N > At the orchid college, student come in queue for service. Arrival rate of customers is 1/min. It takes at the average 22 seconds to take service.

(a) If college will close 3 minutes and if he takes exactly 1.5 minutes to reach the door, can the student leave college in time?

Solution:

(a) A minute is used as unit of time. Therefore,

λ = 1 arrival/min

μ = 2 arrivals/min

$\boldsymbol{WS}$ = waiting time in the system=1/(μ- λ)= 1/(2-1) = 1 minutes

- Now total time will be 1+1.5 = 2.5 minutes.

Queuing Model

Example 4

Q.N > Customers arrive in a bank according to a Poisson's process with mean inter arrival time of 10 minutes. Customers spend an average of 5 minutes on the single available counter, and leave.

(a) What is the probability that a customer will not have to wait at the counter?

(b) What is the expected number of customers in the bank?

(c) How much time can a customer expect to spend in the bank?

Solution:

We will take an hour as the unit of time. Thus, λ = 6 customers/hour,

μ = 12 customers/hour.

The customer will not have to wait if there are no customers in the bank.

Thus, $P0 = 1 – \lambda/\mu = 1 - 6/12 = 0.5$

Expected numbers of customers in the bank are given by

$\boldsymbol{LS} = \lambda /(\mu - \lambda)=6/6=1$

Expected time to be spent in the bank is given by

$\boldsymbol{WS} =1/(\mu – \lambda)= 1/(12-6) = 1/6$ hour = 10 minutes.

Queuing Model

Example 5

Q.N. > At the Banking, people come in queue . Arrival rate of customers is 1/min. It takes at the average 30 seconds to take the token. If a customer arrives 5 minutes before the bank closed and if he takes exactly 4.5 minutes to reach the correct counter after he take a token, can the customer expects to take banking service?

Solution :

A minute is used as unit of time. Since token takes 30 seconds, this means, two customers enter the bank per minute, that is service rate is 2 per minute.

Therefore,

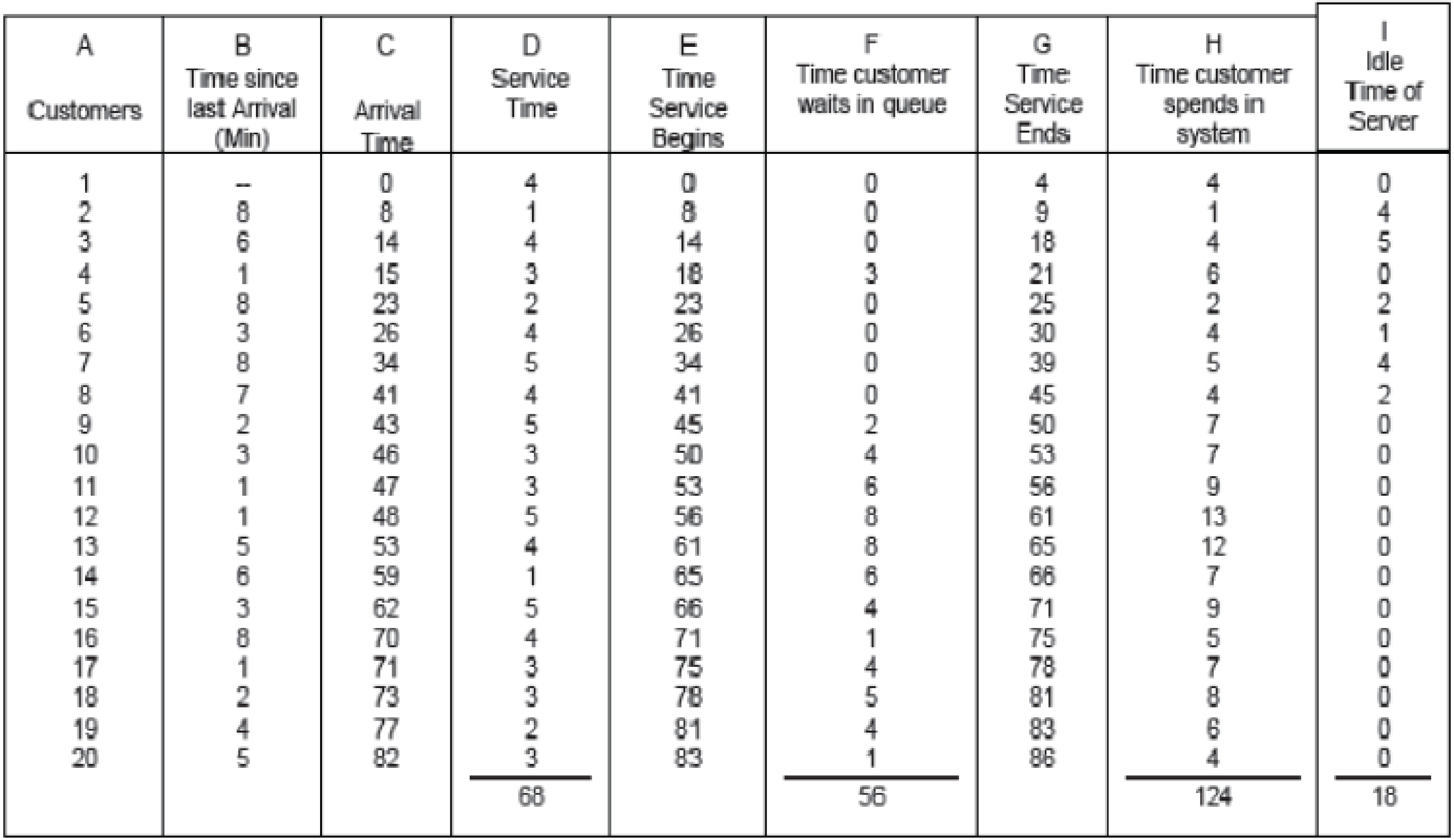
λ = 1 arrival/min

μ = 2 arrivals/min

$\boldsymbol{WS}$ = waiting time in the system=1/(μ- λ)=1 minutes

and if he takes exactly 4.5 minutes to reach the correct counter after he take a token, so bank is closed because 4.5+1=5.5

Queuing Model



A Customers	B Time since last Arrival (Min)	C Arrival Time	D Service Time	E Time Service Begins	F Time customer waits in queue	G Time Service Ends	H Time customer spends in system	I Idle Time of Server
1	--	0	4	0	0	4	4	0
2	8	8	1	8	0	9	1	4
3	6	14	4	14	0	18	4	5
4	1	15	3	18	3	21	6	0
5	8	23	2	23	0	25	2	2
6	3	26	4	26	0	30	4	1
7	8	34	5	34	0	39	5	4
8	7	41	4	41	0	45	4	2
9	2	43	5	45	2	50	7	0
10	3	46	3	50	4	53	7	0
11	1	47	3	53	6	56	9	0
12	1	48	5	56	8	61	13	0
13	5	53	4	61	8	65	12	0
14	6	59	1	65	6	66	7	0
15	3	62	5	66	4	71	9	0
16	8	70	4	71	1	75	5	0
17	1	71	3	75	4	78	7	0
18	2	73	3	78	5	81	8	0
19	4	77	2	81	4	83	6	0
20	5	82	3	83	1	86	4	0
			68		56		124	18

Queuing Model

1. $$\text{Average waiting time} = \frac{\text{total time customers wait in queue}}{\text{total number of customers}} = \frac{56}{20} = 2.8 \text{ min}$$

2. $$\text{Probability of wait} = \frac{\text{Number of customers who wait}}{\text{total number of customers}} = \frac{13}{20} = 0.65$$

3. $$\text{Probability of idle server} = \frac{\text{Total idle time of server(minutes)}}{\text{Total run time of simulation(minutes)}} = \frac{18}{86} = 0.21$$

4. The average service time is 3.4 minutes, determined as follows:

$$\text{Average service time (minutes)} = \frac{\text{Total Service time(minutes)}}{\text{Total run timTotal Number of Customers)}} = \frac{68}{20} = 3.4 \text{ minutes}$$

Measures of System Performance

The performance of a queuing system can be evaluated in terms of a number of response parameters, however the following four are generally employed.

- ✓ ***Average number of customer in the queue or in the system***
- ✓ ***Average waiting time of the customer in the queue or in the system***
- ✓ ***System utilization (Server utilization)***
- ✓ ***The cost of waiting time and idle time***

Measures of System Performance

- Each of these measures has its own importance.
- 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 good quality to attract customers.**

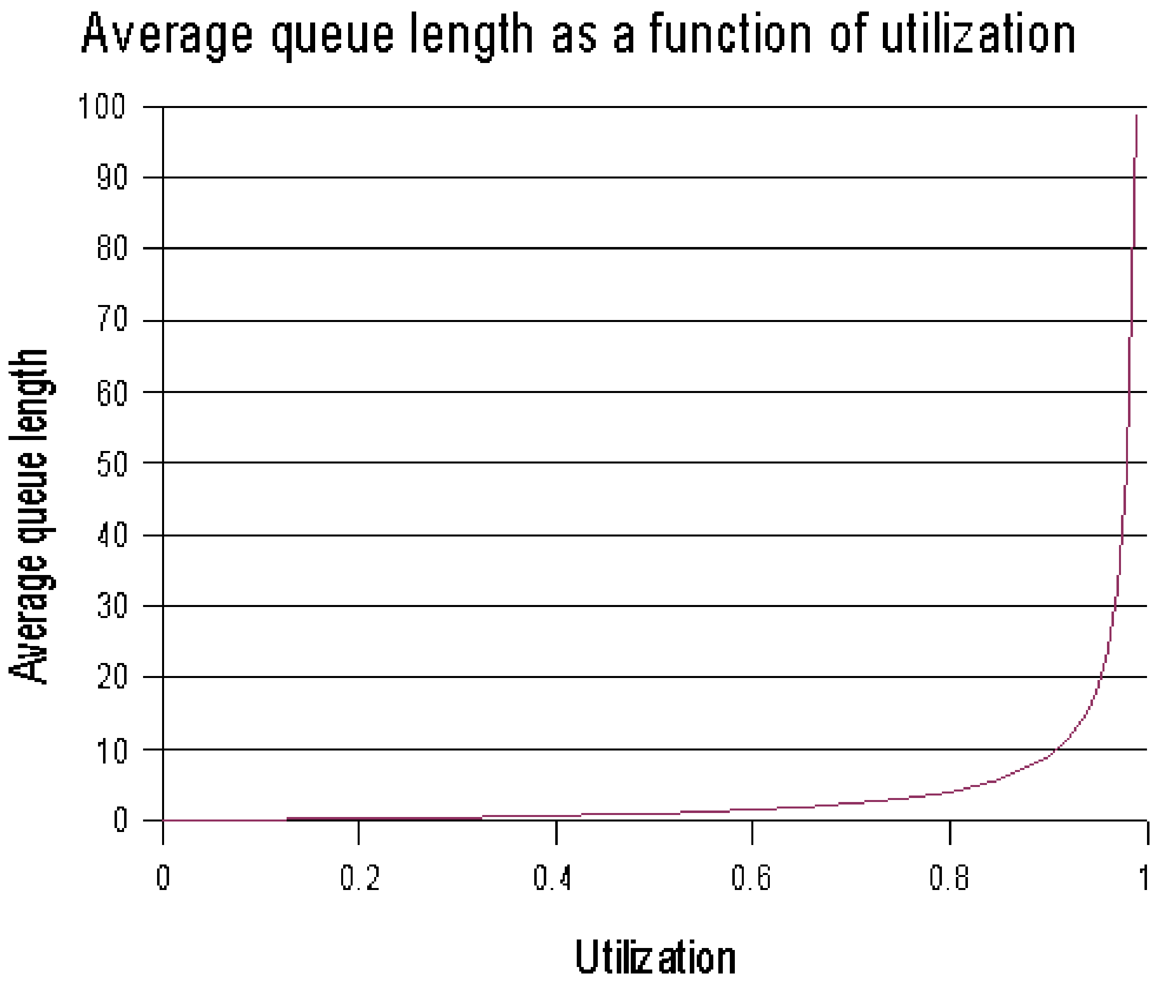
Measures of System Performance

- Let Ta be mean arrival time, Ts be service time, λ be arrival rate and μ be the service rate then, the ratio of mean service rate and mean inter-arrival rate is called the **traffic intensity** ($\boldsymbol{u}$).

$$u = \frac{T_s}{T_a}$$

- The probability that an entity have to wait more than a given time is known as **delay distribution**.
- The knowledge of average waiting time in the queue is necessary for determining the cost of waiting in the queue.

Measures of System Performance



Measures of System Performance

System utilization, that is, the percentage capacity utilized reflects the extent to which the facility is busy rather than idle. System utilization factor (S) is the ratio of average arrival rate (λ) to the average service rate (μ).

$S = \lambda / \mu$, in case of single server model

$S = \lambda / n\mu$, in case of ‘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 is above figure. Under normal circumstances 100% system utilization is not a realistic goal.

Measures of System Performance

Conservation Law

An important law in queuing theory states

$$L = \lambda w$$

Where, L is the long-run in the system, λ is the arrival rate and w is the long-run time in the system.

Often called as **Little's equation**.

CHAPTER 3

Finished !!!