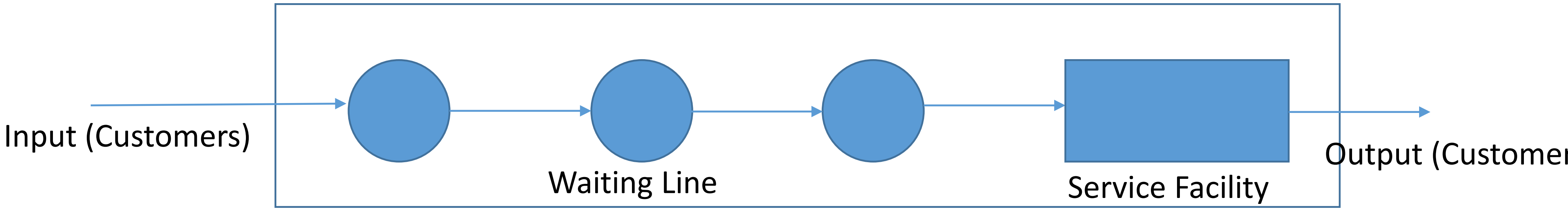
Queuing system

Queuing System

- Most systems in a simulation study contain a process in which there is a demand for services.
- The system can serve entities at a rate which is greater than the rate at which entities arrive. The entities are then said to join waiting line. The line where entities or customers wait is generally known as queue.
- The combination of all entities in system being served and waiting for services is called a queuing system



Characteristics of Queuing System

- 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, repair persons, CPU in computer or washing machines – any resource which provides the requested services.
- In order to model queuing systems, we first need to be a bit more precise about what constitutes a queuing system. The three basic elements to all queuing systems which (also called *congestion*) are:
- *Arrival process or patterns*, which describes the statistical properties of the arrivals.
- *Service process or patterns*, which describes how the entities are served.
- *Queuing discipline and behavior* , which describes how the next entity to be served is selected.

1. Arrival process or patterns

- Any queuing system must work on something – customers, parts, patients, orders, etc.
- We generally call them as entities or customers.
- Before entities can be processed or subjected to waiting, they must first enter the system.
- Depending on the environment, entities can arrive smoothly or in an unpredictable fashion.
- They can arrive one at a time or in clumps.
- They can arrive independently or according to some correlation.

Cont..

- A special arrival process, which is highly useful for modeling purposes, is the Markov arrival process.
- It refers to the situation where entities arrive one at a time and the times between arrivals i.e. inter-arrival time are exponential random variables.
- There are theoretical results showing that if a large population of customers makes independent decisions of when to seek service, the resulting arrival process will be Markov.
- Examples where this occurs are phone calls arriving at an exchange, customers arriving at a fast food restaurant, hits on a web site, and many others.

2. Service process or patterns

- Once entities have entered the system they must be served. The physical meaning of “service” depends on the system.
- Customers may go through the checkout process.
- Parts may go through machining. Patients may go through medical treatment. Orders may be filled. And so on.
- From a modeling standpoint, the operational characteristics of service matter more than the physical characteristics.
- Specifically, we care about whether service times are long or short, and whether they are regular or highly variable.

Cont..

- We care about whether entities are processed in first-come-first-serve (FCFS) order or according to some kind of priority rule. We care about whether entities are serviced by a single server or by multiple servers working in parallel.

- A special service process is the Markov service process, in which entities are processed one at a time in FCFS order and service times are independent and exponential.

3.a Queuing discipline

- The third required component of a queuing system is a queue, in which entities wait for service.
- The number of customers that can wait in a line is called *system capacity*.
- The simplest case is an unlimited queue which can accommodate any number of customers.
- It is called *system with unlimited capacity*. But many systems (e.g., phone exchanges, web servers, call centers), have limits on the number of entities that can be in queue at any given time.
- Arrivals that come when the queue is full are rejected (e.g., customers get a busy signal when trying to dial into a call center).

Cont..

- The logical ordering of customer in a waiting line is called queuing discipline and it determines which customer will be chosen for service.
- We may say that queuing discipline is a rule to choose the customer for service from the waiting line.

Queuing discipline

- **FIFO (First In First Out):** According to this rule, service is offered on the basis of arrival time of customer. The customer who comes first will get the service first.
- **LIFO (Last In First Out):** It occurs when service is next offered to the customer that arrived recently or which have least waiting time. In the crowded train, the passengers getting in or out from the train is an example of LIFO.
- **SIRO (Service in Random Order):** It means that a random choice is made between all waiting customers at the time service is offered i.e. a customer is picked up randomly from the waiting queue for the service.
- **SPTF (Shortest Processing Time First):** It means that a customer with the shortest service time will be chosen first for the service i.e. the shortest service time customer will get the priority in the selection process.
- **Priority:** A special number is assigned to each customer in the waiting line and it is called priority. Next, according to this number, the customer is chosen for service.

Queuing behavior

- Even if the system doesn’t have a strict limit on the queue size, customers may shy away at joining the queue when it is too long (e.g., cars pass up a drive through restaurant if there are too many cars already waiting). It is called *balking.*
- Customers may also exit the system due to impatience (e.g., customers kept waiting too long at a bank decide to leave without service) or perishability (e.g., samples waiting for testing at a lab spoil after some time period). It is called *reneging*.

Cont..

- When there is more than one line forming for the same service or server, the action of moving of customer from one line to another line because they think that they have chosen slow line is called *jockeying*

Example: A candy manufacturer has a production line which consists of three machines separated by inventory-in-process buffers. The first machine makes and wraps the individual pieces of candy, the second packs 50 pieces in a box, and the third seals and wraps the box. The two inventory buffers have capacities of 1000 boxes each. As illustrated by

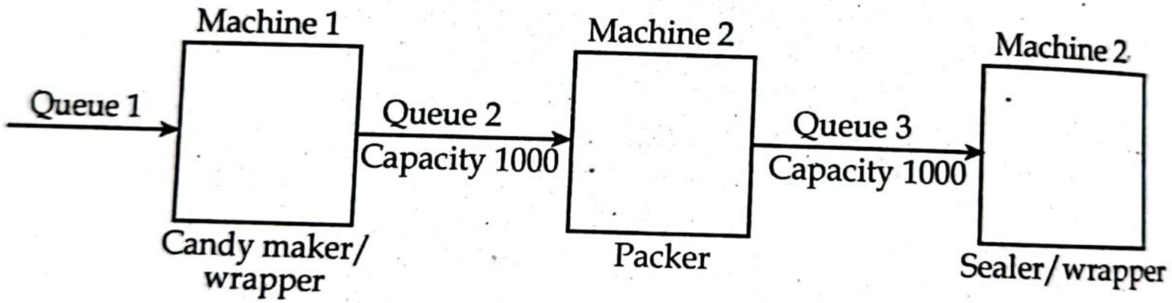


Figure 3.5: Candy-production line

Figure 3.5, the system is modeled as having three service centers, each center having c = 1 server (a machine), with queue-capacity constraints between machines. It is assumed that a sufficient supply of raw material is always available at the first queue. Because of the queue-capacity constraints, machine 1 shuts down whenever the inventory buffer fills to capacity, while machine 2 shuts down whenever the buffer empties. In brief, the system consists of three single-server queues in series with queue-capacity constraints and a continuous arrival stream at the first queue.

QUEUEING N

Queuing notation (Kendall’s notation)

- Different notations are frequently used in queuing system and are called Kendall’s Notation.
- Kendall’s notation is the standard system used to describe and classify a queuing node.
- It can be represented in the form, *A/B/c/D/N/K*, where, *A*, *B*, *c*, *N, K* respectively indicate arrival pattern, service pattern, queuing discipline, number of servers, system capacity and calling population.
- The symbols used for the probability distribution for inter arrival time, and service time are, *D* for deterministic, *M* for exponential and E_k for Erlang distribution.
- Similarly, FIFO, LIFO, etc is used for queue discipline.
- If the N and *K* are not specified, it is taken as infinity, and if queuing discipline D is not specified, it is FIFO.

Example

- $M/D/2/5$ stands for a queuing system having exponential arrival times, deterministic service time, 2 servers, capacity of 5 customers, unlimited population and first in first out discipline.
- If notation is given as $M/D/2$ means exponential arrival time, deterministic service time, 2 servers, infinite service capacity, infinite population and FIFO queue discipline.

Question:

- What do you mean by M/M/1/N/∞? Suppose an office working 8 hrs per day for 5 days a week gets about 800 telephone calls a week. Find out the number of calls per minute? What is the mean inter-arrival time of calls?
- The notation M/M/1/N/∞ is called Kendall's notation for queuing system, which respectively indicate arrival pattern, service pattern, number of servers, system capacity and calling population. Here the queuing system has random arrival and service times, one server, capacity of handling N customers, an unlimited population and FIFO queuing discipline.

solution

- Total working hours in a week = 40 hours.
- Total working minutes in a week = 40*60 minutes.
- Total calls = 800.
- Thus, mean number of calls per minute = 800 / (40*60) = 0.33 calls per minute.
- That is mean arrival rate of calls, denoted by λ, is 0.33 calls per minute and inter-arrival time of calls, denoted by T_a, is = (40*60) / 800 = 3 minutes per call.

Single server queuing system (M/M/1)

- It is a queuing system with only one server for any number of customers.
- It is a FIFO queuing system with Kendall notation, M/M/1 with poisson input, exponential service time and unlimited waiting positions.
- The model is based on following assumptions.
    - The arrival follow Poisson distribution with a mean arrival rate λ (lambda)
    - The service time has exponential distribution, average service rate μ
    - Arrivals are infinite population
    - Customers are served on FIFO basis
    - There is only a single server
- In a single server queuing system, there is an infinite number of waiting positions in the queue. Hence there can be any number of customers in the queue, so it becomes a challenge to maintain the service rate in such a way as to match up with the continuous arrival of customers.

Poison arrival Patterns

- It says that arrival of a customer is completely random.
- This means that an arrival can occur at any time and the time of next arrival is independent of the previous arrival.
- With this assumption it is possible to show that the distribution of the inter-arrival time is exponential.
- This is equivalent to saying that the number of arrivals per unit time is a random variable with a Poisson's distribution.
- This distribution is used when chances of occurrence of an event out of a large sample is small.

Cont…

That is if X = number of arrivals per unit time, then, probability distribution function of arrival is given as

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \begin{cases} x = 0, 1, 2, \ldots \\ \lambda > 0 \end{cases}$$

$$E(X) = \lambda$$

Where λ. is the average number of arrivals per unit time (1/ τ), E(X) is the expected number, and x is the number of customers per unit time. This pattern of arrival is called Poisson’s arrival pattern. τ is inter arrival time.

Illustrative example

- In a single pump service station, vehicles arrive for fueling with an average of 5 minutes between arrivals. If an hour is taken as unit of time, cars arrive according to Poison’s process with an average of λ= 12 cars/hr. The distribution of the number of arrivals per hour is,

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-12}12^{x}}{x!}, \begin{cases} x = 0, 1, 2, .. \\ \lambda > 0 \end{cases}$$

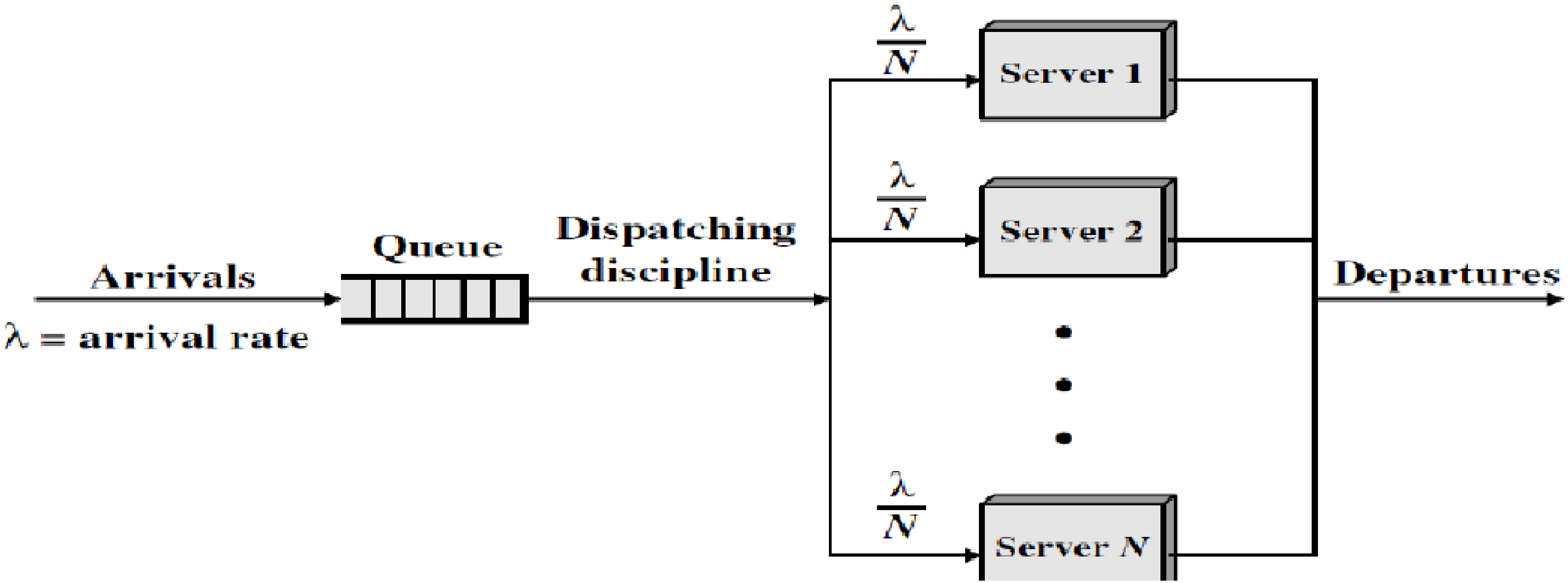
Measure of Queues

- The ratio of the mean service time (Ts) to the mean inter arrival time (Ta) is called traffic intensity.
- I.e. u= λ"Ts or u=Ts/Ta If there is any balking or reneging, not all arriving entities get served. It is necessary therefore to distinguish between actual arrival rate and the arrival rate of entities that get served.
- Here λ" denote all arrivals including balking or reneging

Server utilization

- It consists of only the arrival that gets served. It is denoted by and defined as ρ = λTs= λ/ μ (server utilization for single server).
- This is also the average number of customers in the service facility.
- Thus probability of finding service counter free is (1 – ρ) That is there are zero customers in the service facility.

Concept of Multi-server Queue



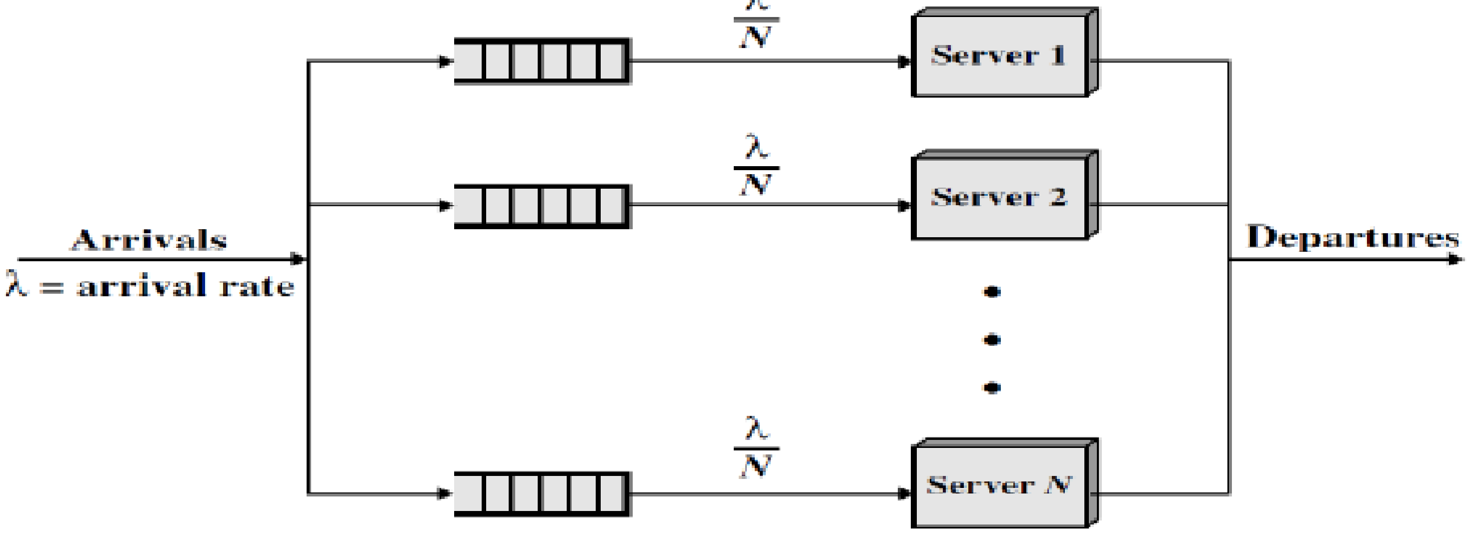
Cont..

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

Cont..

$$\rho = \lambda / c\mu$$

Where μ is the service rate and λ is the arrival rate. There is another concept which is called multiple single server queue system as shown below



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

- Two principal measures of queuing system are;

a) The mean number of customers waiting and

b) 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.

Cont..

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

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

= Average number of customers in the System − Server Utilization

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

Cont..

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

$$= \frac{\overline{L}_S}{\lambda} = \frac{\frac{\lambda}{\mu - \lambda}}{\lambda} = \frac{1}{\mu - \lambda}$$

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

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

Example

- 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
  $\boldsymbol{W_s}$ = 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.

Example2

- 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, $P_0 = 1 – \lambda/\mu = 1 - 6/12 = 0.5$

Expected numbers of customers in the bank are given by
$\boldsymbol{L_S} = \lambda /(\mu - \lambda)=6/6=1$

Expected time to be spent in the bank is given by
$\boldsymbol{W_S} =1/(\mu – \lambda)= 1/(12-6) = 1/6$ hour = 10 minutes.

Example

- 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
  $\boldsymbol{W_S}$ = 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.

Example2

- 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 – λ/μ= 1− 6/12 = 0.5
- Expected numbers of customers in the bank are given by $\boldsymbol{L_S}$ = λ /(μ - λ)=6/6=1
- Expected time to be spent in the bank is given by $\boldsymbol{W_S}$ =1/(μ – λ)= 1/(12-6) = 1/6 hour = 10 minutes.

Network of Queues

- Many systems are naturally modeled as network of single queues in which customers departing from one queue may be routed to another.
- The following results assume a stable system with infinite calling population and no limit on system capacity.

i. Provided that no customers are created or destroyed in the queue, then the departure rate out of the queue is the same as the arrival rate into the queue over the long run.

ii. If customers arrive to queue I at rate λ_i and a fraction $0 <= P_{ij} <= 1$ of them are routed to queue j upon departure, then the arrival rate from queue I to queue j is $\lambda_i p_{ij}$ over the long run.

iii. The overall arrival rate in to queue j , λ_j is the sum of the arrival rate from all sources. If customers arrive from outside the network at rate a_j, then $\lambda_j = a_j + \sum_{for\ all\ i} \lambda_i p_{ij}$

iv. If queue j has $C_j < \infty$ parallel servers, each working at rate μ_j , then the long run utilization of each server is,

$\rho_j = \lambda_j / (C_j\ \mu_j)$

and $\rho_j < 1$ is required for the queue to be stable.

Applications of Queuing System

Queuing systems are used in our daily life in every aspect. Some of the common applications are

1. Commercial queuing systems: commercial organizations serving external customers. Eg. Dental , bank, ATM, Gas station, Plumber
2. Transportation service system: Vehicles are customers or servers. Eg. Vehicles waiting at traffic lights, buses, taxi, cabs, etc
3. Business internal service systems: customers receiving service are internal to the organization providing the service. Eg: inspection stations, conveyor belts, customer support
4. Social service systems: Eg: judicial process, hospital, waiting list for organ transplants etc.