Computer Modeling and Simulation

Lectures 5&6

Queuing System

- Any system where the customer requests a service for a finite - capacity resource may be considered to be a queuing system.
- Grocery stores, theme parks, banks and fast - food restaurants are well - known examples of queuing systems.
- Even a door or a toilet can be an example of a self - service queuing system.
    - For example, McNickle used a queuing model to estimate the required number of toilets in New Zealand buildings based on the estimated number of people entering the buildings
- In computers, the number of processes need to be run at a specific time can be greater than the number of processing cores available, so some of the processes may need to wait in a queue.
- In cloud computing, you have to wait for your request of a service to be fulfilled by a server somewhere else.

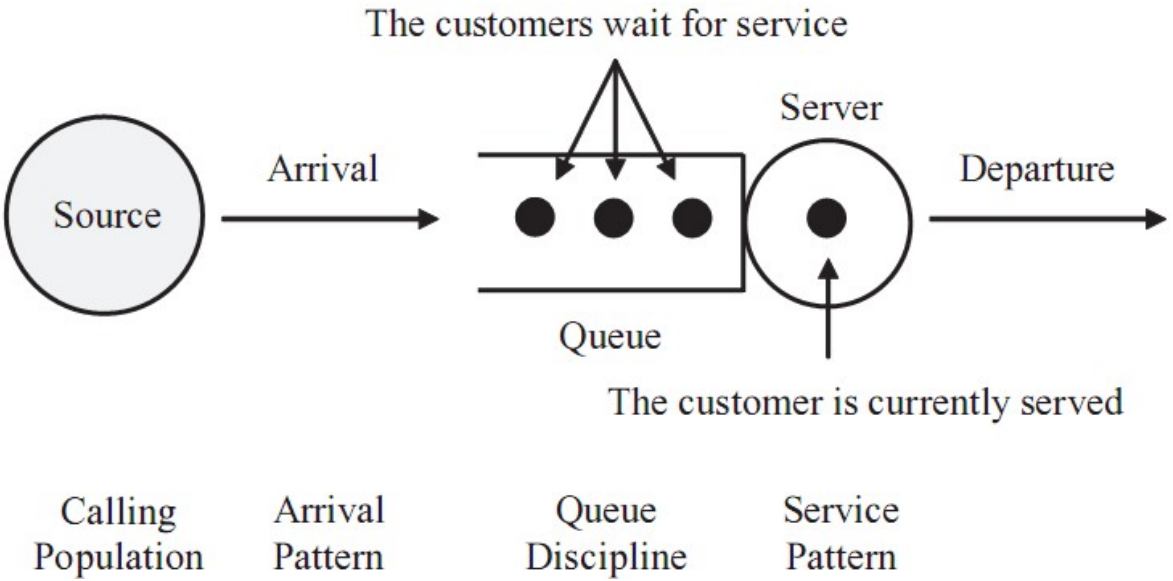
Queuing Model

- Queuing models are constructed to analyze the performance of a dynamic system where waiting can occur i.e. a Queuing System.
- The goals of a queuing model are to minimize the average number of waiting customers in a queue and to predict the estimated number of facilities in a queuing system.
- The performance results of queuing model simulation are produced at the end of a simulation in the form of aggregate statistics.

Queuing Model

- Three basic elements within a queuing model are
  - Entities
  - Servers
  - Queues.
- Entities can represent either customers or objects, servers can represent persons or production stations that treat or interact with the entity, and queues are the holding or waiting position of entities.

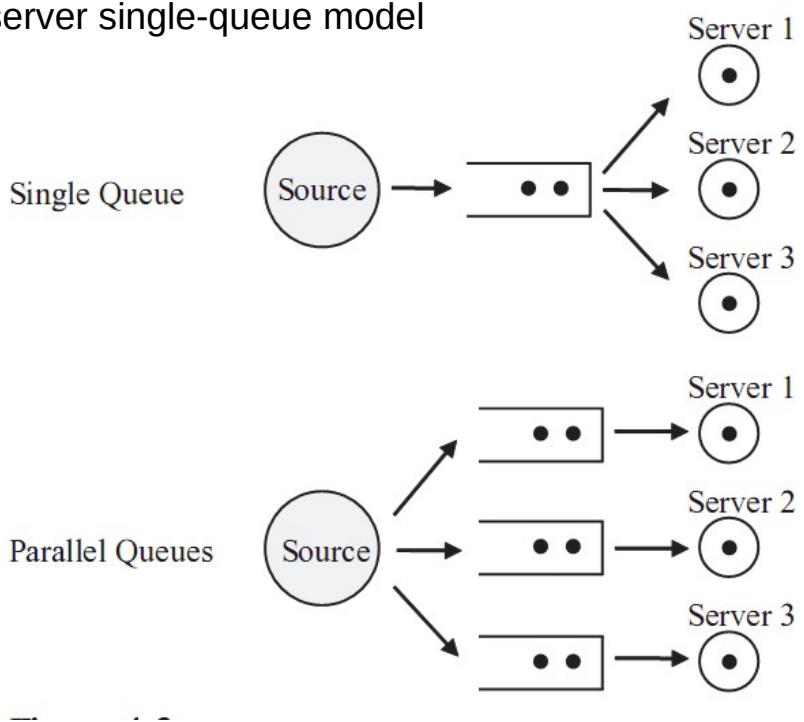
Different configurations of a Queuing Model



A single-server single-queue model

Different configurations of a Queuing Model

A multi-server single-queue model



A multi-server multi-queue model

Attributes of a Queuing Model

- A queuing model is described by its attributes:
  - customer population,
  - arrival and service pattern,
  - queue discipline,
  - queue capacity,
  - the number of servers.

Attributes of Queuing Model
1. Calling Population

- The calling population , which can be either finite or infinite, is defined as “the pool of customers who possibly can request the service in the near future”.
- If the size of the calling population is infinite, the arrival rate is not affected by others.

Attributes of Queuing Model
2. Arrival and Service Pattern

- Arrival and service patterns are the two most important factors determining behaviors of queuing models.
- A queuing model may be deterministic or stochastic .
- For the stochastic case, new arrivals occur in a random pattern and their service time is obtained by probability distribution.
- The arrival and service rates, based on observation, are provided as the values of parameters for stochastic queuing models.

Attributes of Queuing Model
2. Arrival and Service Pattern

- The arrival rate is defined as the mean number of customers per unit time, and the service rate is defined by the capacity of the server in the queuing model.
- If the service rate is less than the arrival rate, the size of the queue will grow infinitely. T
- The arrival rate must be less than the service rate in order to maintain a stable queuing system.
- The randomness of arrival and service patterns cause the length of waiting lines in the queue to vary.

Attributes of Queuing Model
2. Arrival and Service Pattern

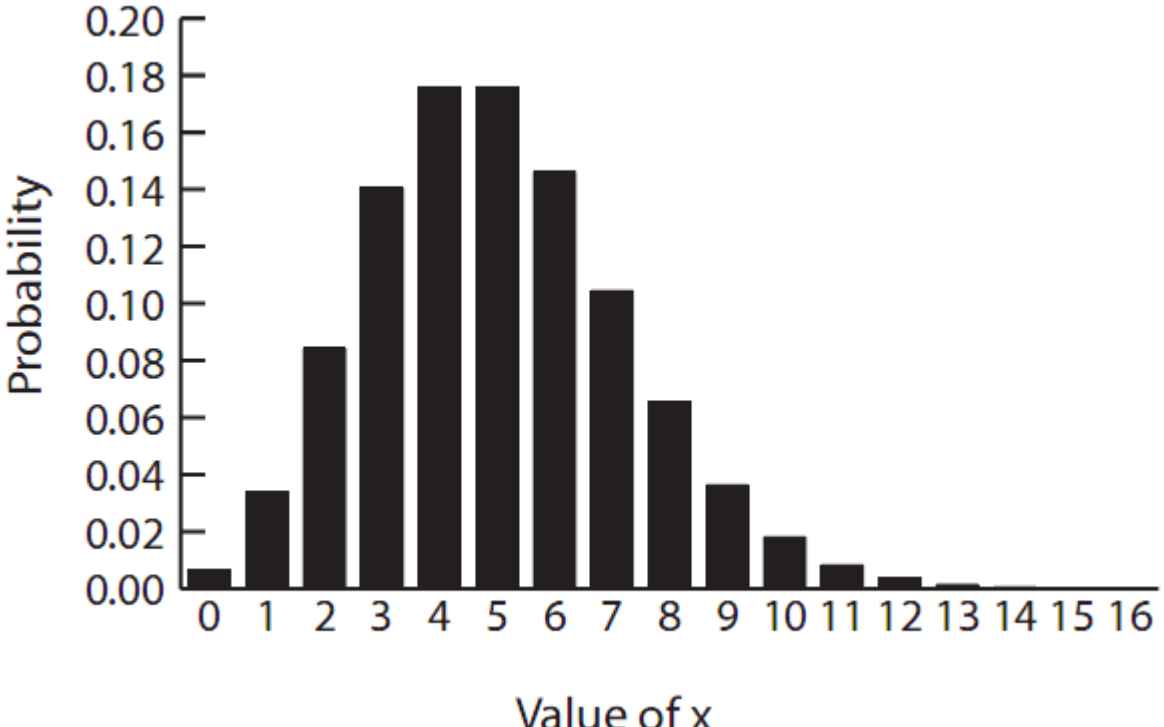
- Some customers in the real world may not stay in the queue upon arrival and leave the system.
- If the length of waiting line is too long or they find a shorter line, they may leave the queue.
- Impatient customer behaviour:
  - Balking,
  - Reneging,
  - Jockeying
- Balking occurs when an arriving customer does not enter the queue due to the limited queue capacity.
- Reneging occurs when a customer leaves the queue after waiting in a queue upon arrival.
- Jockeying occurs when a customer decides to switch the queue for earlier service.
- The decision of balking is deterministic, whereas those of reneging and jockeying are considered as probabilistic.

Attributes of Queuing Model
2. Arrival and Service Pattern

- **Arrival Process:**
- Suppose λ is the average arrival rate (e.g., customers arrive at a rate of λ per hour) and x is the number of customers.
- The probability that x customers arrive in one hour is given by a Poisson statistical distribution.

$$P(x) = (\lambda^{x} e^{-\lambda}) / x!$$, e= Euler's constant and x! = factorial of x

Arrival Process Probability Distribution (Poisson Distribution)



Inter-arrival Time Probability Distribution

- **Inter-arrival Time Probability Distribution:**
- Interarrival Time is the time between customer arrivals.
- Interarrival Time indicate the random time at which the next customer will arrive at the queue.
- This number will need to be computed for the simulation to account for the arrival of each customer.
- The Interarrival Time is exponentially distributed and follows the probability function of equation

$$P = e^{-\lambda t}$$

Service Time Probability Distribution

- The service process is described by a different probability distribution which gives the service time interval prob ıwill be between t1 and t2 .

$$P(t_1 \leq T \leq t_2) = e^{-\mu t_1} - e^{\mu t_2} \quad \text{for } t_1 < t_2$$

- The probability of a specific service time is computed in a similar manner to the inter-arrival time and follows the same function as that for inter-arrival time.

$$P(t) = e^{-\mu t}$$

- Where μ is the average service time.

Attributes of Queuing Model
3. Queue Pattern

- When a server becomes idle, the next customer is selected among candidates from the queue. The selection of strategy from the queue is called queue discipline.
- The common algorithms of queue discipline are
  - first - in first - out (FIFO),
  - last - in first - out (LIFO),
  - service in random order (SIRO),
  - priority queue.
- The most common queue discipline is FIFO.

Attributes of Queuing Model
3. Queue Pattern

- In a priority queue discipline, each arrival has its priority.
  - The priority scheme may be either preemptive or non-preemptive .
  - In a preemptive scheme, the customer currently being served is placed back at the front of the queue if the incoming customer has the higher priority than the customer has who is currently being served.
  - The displaced customer ’ s service may be either restarted or resumed.
  - In a non-preemptive scheme, the continuous service is provided for the customer currently being served until it ends.

Attributes of Queuing Model
4. Queue Capacity

- The queue size may be assumed to be either finite or infinite.
- For example, the physical confines of a buffer area (i.e., queue) between two
- workstations along a production line make this queue finite.
- In contrast, a call center may have the capacity to queue, or place on hold, up to 1000 incoming calls unable to be serviced by the three service representatives; since there is no history of the call center approaching its holding queue capacity, this queue can be assumed to be infinite.

Kendall’s Notation for a Queuing Model

- Kendall ’ s notation , A/B/c/N/K, is used to concisely define a queue and its parameters.
- “ A ” and “ B ” represent the inter - arrival and service distribution, respectively; “ D ” (deterministic), “ M ” (Poisson), “ G ” (general), and
- “ Ek “(Erlang) are used to represent “ A ” and “ B ” .
- “ c ” represents the number of servers.
- “ N ” represents the queue capacity;
- “ K ” represents the size of the calling population.
- Usually the A/B/c notation is used when “ N ” and “ K ” are infinite.
- For example, M/M/1 represents a single server queuing model, and the inter - arrival and service time are exponentially distributed.
- The queue discipline is often added to describe the system.
- Here we will address only the M/M/1 type of queue.

References

- Chapter 4 of the book “Principles of Modelling and Simulation: A Multi-disciplinary Approach by John A. Sokolowski and Catherine M. Banks.”, (already uploaded on SLATE).