Queuing theory

Lecture 7

Introduction

- Queues (waiting lines) are a part of everyday life. We all wait in queues to buy a movie ticket, make a bank deposit, pay for groceries, mail a package, obtain food in a cafeteria, start a ride in an amusement park, etc.
- However, having to wait is not just a petty personal annoyance. The amount of time that a nation's populace wastes by waiting in queues is a major factor in both the quality of life there and the efficiency of the nation's economy.

Introduction

- Great inefficiencies also occur because of other kinds of waiting than people standing in line. For example, making machines wait to be repaired may result in lost production. Vehicles (including ships and trucks) that need to wait to be unloaded may delay subsequent shipments.
- Queueing theory is the study of waiting in all these various guises. It uses queueing models to represent the various types of queueing systems

Introduction

- Therefore, these queueing models are very helpful for determining how to operate a queueing system in the most effective way. Providing too much service capacity to operate the system involves excessive costs.
- But not providing enough service capacity results in excessive waiting and all its unfortunate consequences.
- The models enable finding an appropriate balance between the cost of service and the amount of waiting.

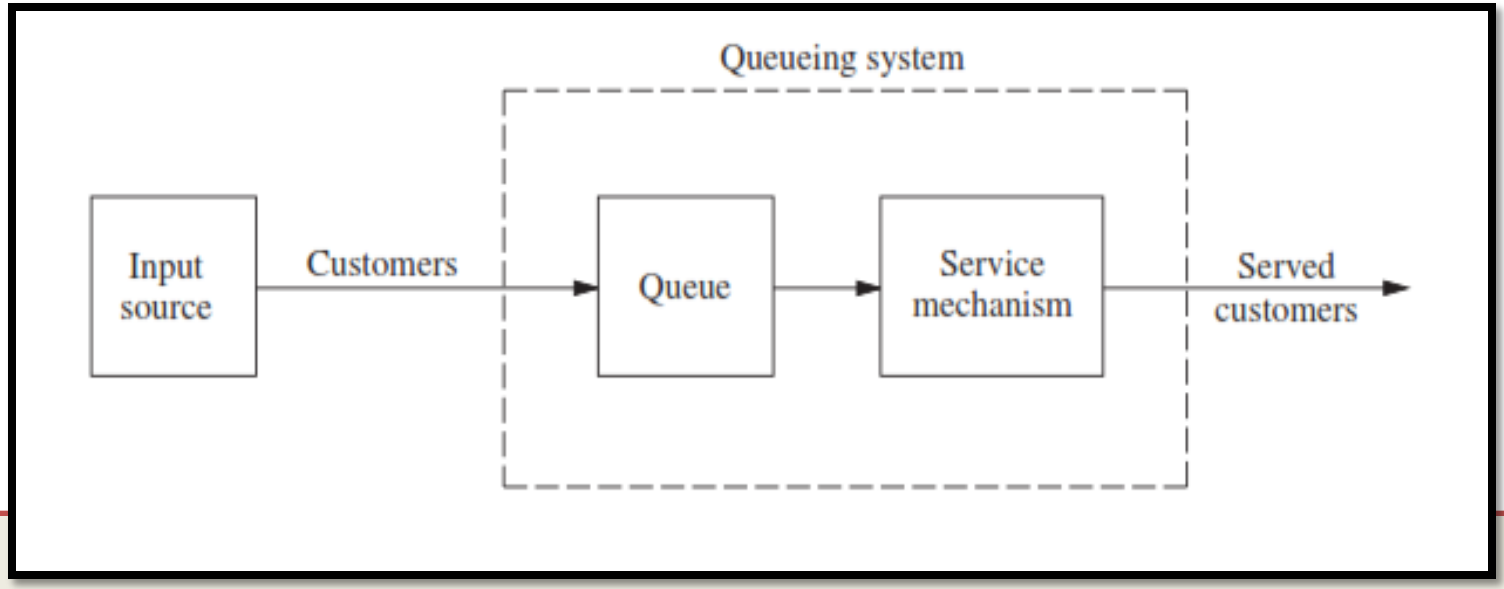
Basic structure of queueing models

- The basic process assumed by most queueing models is the following. Customers requiring service are generated over time by an input source.

- These customers enter the queueing system and join a queue if service is not immediately available. At certain times, a member of the queue is selected for service by some rule known as the queue discipline.

Basic structure of queueing models

- The required service is then performed for the customer by the service mechanism, after which the customer leaves the queueing system.



Input Source (Calling Population)

- One characteristic of the input source is its **size**. The size is the total number of customers that might require service from time to time.
- This population from which arrivals come is referred to as the **calling population**.
- The size may be assumed to be either infinite or finite (so that the input source also is said to be either unlimited or limited).

Basic structure of queueing models

- The statistical pattern by which customers are generated over time must also be specified.
- The common assumption is that they are generated according to a Poisson process;
- This case is the one where arrivals to the queueing system occur randomly but at a certain fixed mean rate, regardless of how many customers already are there (so the size of the input source is infinite).

Queue

- The **queue** is where customers wait before being served. A queue is characterized by the maximum permissible number of customers that it can contain. Queues are called infinite or finite, according to whether this number is infinite or finite.
- The assumption of an infinite queue is the standard one for most queueing models, even for situations where there actually is a (relatively large) finite upper bound on the permissible number of customers.

Queue discipline

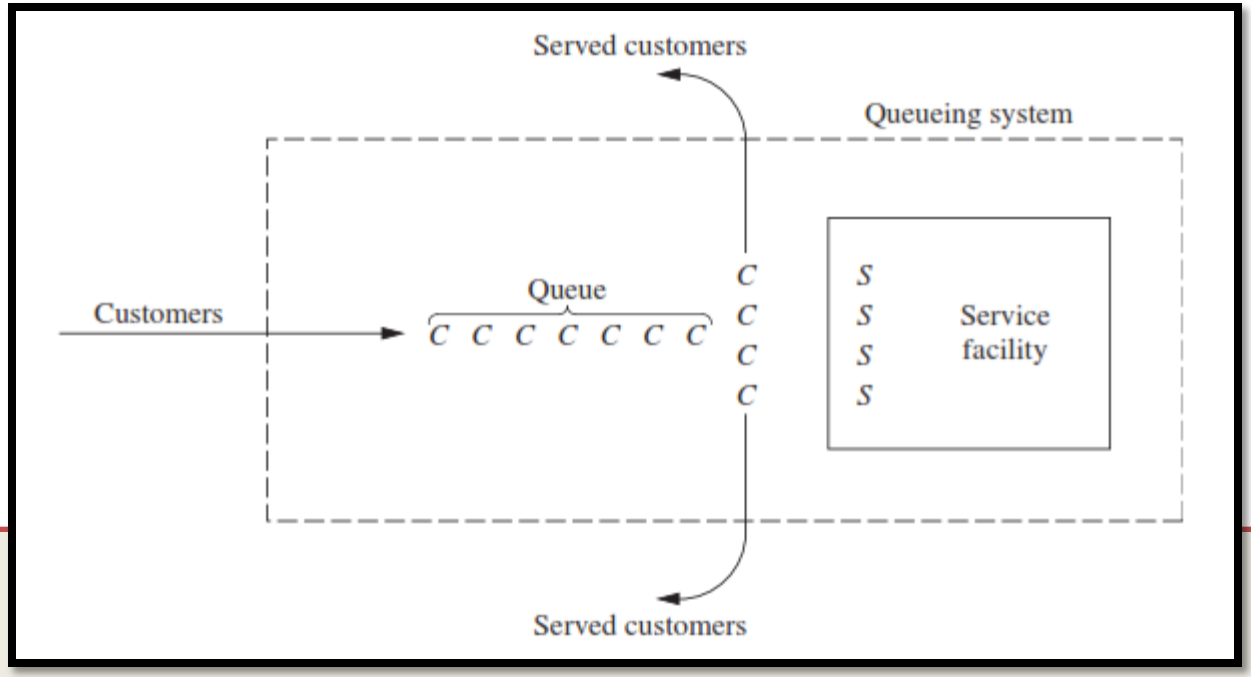
- The **queue discipline** refers to the order in which members of the queue are selected for service.
- For example, it may be first-come-first-served, random, according to some priority procedure, or some other order. First-come-first-served usually is assumed by queueing models, unless it is stated otherwise.

Service Mechanism

- The service mechanism consists of one or more service facilities, each of which contains one or more parallel service channels, called **servers**.
- The time elapsed from the commencement of service to its completion for a customer at a service facility is referred to as the **service time** (or holding time).

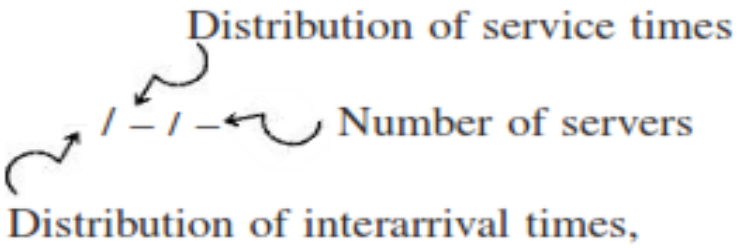
An Elementary Queueing Process

- A model of a particular queueing system must specify the probability distribution of service times for each server, although it is common to assume the same distribution for all servers



Queuing process

- Such models conventionally are labeled as follows:



Where

- M exponential distribution (Markovian),
- D degenerate distribution (constant times)
- E Erlang distribution (shape parameter k)
- G general distribution (any arbitrary distribution allowed)

Queuing process

- For example, the **M/M/s** model assumes that both interarrival times and service times have an exponential distribution and that the number of servers is s (any positive integer).
- The **M/G/1** model assumes that interarrival times have an exponential distribution, but it places no restriction on what the distribution of service times must be, whereas the number of servers is restricted to be exactly 1.

Notations

State of system = number of customers in queueing system.

Queue length = number of customers waiting for service to begin.
= state of system *minus* number of customers being served.

$N(t)$ = number of customers in queueing system at time t ($t \geq 0$).

$P_n(t)$ = probability of exactly n customers in queueing system at time t, given number at time 0.

s = number of servers (parallel service channels) in queueing system.

λ_n = mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in system.

μ_n = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in system. *Note:* μ_n represents *combined* rate at which all *busy* servers (those serving customers) achieve service completions.

λ, μ, ρ = see following paragraph.

Notations

L = expected number of customers in queueing system $= \sum_{n=0}^{\infty} nP_n$.

L_q = expected queue length (excludes customers being served) $= \sum_{n=s}^{\infty} (n-s)P_n$.

$\mathcal{W}$ = waiting time in system (includes service time) for each individual customer.

$W = E(\mathcal{W})$.

$\mathcal{W}_q$ = waiting time in queue (excludes service time) for each individual customer.

$W_q = E(\mathcal{W}_q)$.

THE BIRTH-AND-DEATH PROCESS

- Most elementary queueing models assume that the inputs (arriving customers) and outputs (leaving customers) of the queueing system occur according to the birth-and-death process.
- This important process in probability theory has applications in various areas. However, in the context of queueing theory, the term birth refers to the arrival of a new customer into the queueing system, and death refers to the departure of a served customer.

The *M/M/s Model*

- The model assumes that both interarrival times and service times have an exponential distribution and that the number of servers is s (any positive integer). For example (M/M/1)

$$L=\frac{\lambda}{\mu-\lambda} \qquad L_q=\frac{\lambda^2}{\mu(\mu-\lambda)} \qquad W=\frac{1}{\mu-A} \qquad W_q=\frac{\lambda}{\mu(\mu-\lambda)}$$

The M/G/1 model

- The model assumes that interarrival times have an exponential distribution, but it places no restriction on what the distribution of service times must be, whereas the number of servers is restricted to be exactly 1.

$$
\begin{aligned}
P_0 &= 1 - \rho, \\
L_q &= \frac{\lambda^2 \sigma^2 + \rho^2}{2(1-\rho)}, \\
L &= \rho + L_q, \\
W_q &= \frac{L_q}{\lambda}, \\
W &= W_q + \frac{1}{\mu}.
\end{aligned}
$$