CS-417
COMPUTER SYSTEMS MODELING

Spring Semester 2020

Batch: 2016-17

(LECTURE # 18)

FAKHRA AFTAB
LECTURER
DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING
NED UNIVERSITY OF ENGINEERING & TECHNOLOGY

Recap of Lecture # 17

Periodicity Properties of Markov Chain

Example Problems

Counting Processes (Poisson Process as a Counting Process)

Merging and Splitting of Poisson Processes

Chapter # 6

FUNDAMENTALS OF QUEUING MODELS

Prepared by: Ms. Fakhra Aftab (Lecturer, CISD, NEDUET)

What is Queuing Theory?

- *Queuing theory* is the study of waiting in various guises.

- It deals with quantifying the phenomenon of waiting in lines using representative measures of performance, such as
  - *average queue length*,
  - *average waiting time* in queue, and
  - *average service time*.

- It uses ***queuing models*** to represent the various types of ***queuing systems*** that arise in practice.

What is Queuing Theory?

- **Advantages of queuing models**: very helpful for determining how to operate a queuing 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 operating characteristics of queuing systems are determined largely by two statistical properties, namely,
  - the probability distribution of *inter-arrival times* and
  - the probability distribution of *service times*.

The Basic Queuing Process

- *Customers* requiring service are generated over time by an ***input source***.
- These customers enter the ***queuing system*** and join a ***queue***.
- At certain times, a member of the queue is selected for service by some rule known as the ***queue discipline***.
- The required service is then performed for the customer by the ***service mechanism*** after which the customer leaves the queuing system.

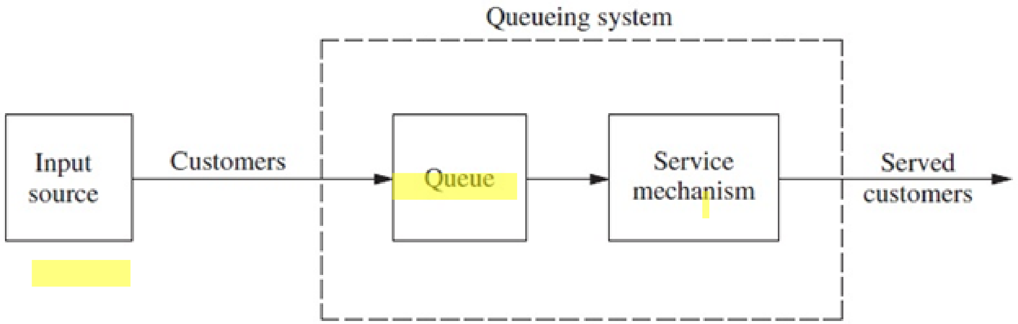


Fig 1: The basic queuing process

Input Source (Calling Population)

- **Calling population**: population from which arrivals come.

- One characteristic of the input source is its size.
  - total number of distinct potential customers that might require service from time to time.

- The size may be assumed to be either ***infinite*** or ***finite***.

- An infinite source is forever abundant (e.g., calls arriving at a telephone exchange).

- Because the calculations are ***far easier*** for the infinite case, this assumption often is made even when the actual size is some relatively large finite number.

Input Source (Calling Population)

- The finite case is ***more difficult analytically.***

- The statistical pattern by which customers are generated over time must also be specified.

- **The common assumption**: ***Poisson process***; i.e.,
  - the number of customers generated until any specific time has a Poisson distribution.
  - arrivals to the queuing 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*.

- **An equivalent assumption**: the probability distribution of the time between consecutive arrivals is an *exponential distribution*.

Queuing Behavior

- The queuing behavior of customers plays a role in waiting-line analysis.
- Human customers may ***jockey*** from one queue to another in the hope of reducing waiting time.
- They may also ***balk*** from joining a queue altogether because of anticipated long delay, or
- They may ***renege*** from a queue because they have been waiting too long.

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 queuing models.
- However, for queuing systems where this upper bound is small enough that it actually would be reached with some frequency, it becomes necessary to assume a *finite queue.*

Queue Discipline

The order in which members are selected from a queue

- An important factor in the analysis of queuing models.
- First-come-first-served (FCFS) usually is assumed by queuing models, unless stated otherwise.
- Other disciplines include Last Come, First Served (LCFS) and Service In Random Order (SIRO).
- Customers may also be selected from the queue based on some order of priority.
- For example, rush jobs at a shop are processed ahead of regular jobs.

Service Mechanism

- The service mechanism consists of
  - one or more ***service facilities***,
  - each of which contains one or more ***parallel service channels***, called **servers.**
- If there is more than one service facility, the customer may receive service from a sequence of these (*service channels in series*).
- At a given facility, the customer enters one of the parallel service channels and is completely serviced by that server.
- A queuing model must specify the arrangement of the facilities and the number of servers (parallel channels) at each one.
- Most elementary models assume one service facility with either one server or a finite number of servers.

Service Mechanism

- **Service time (or *holding time*)**: The time elapsed from the commencement of service to its completion for a customer at a service facility.

- The service-time distribution that is most frequently assumed in practice
  - (largely because it is far more tractable than any other)
  - is the *exponential* distribution, and most of our models will be of this type.

- Other important service-time distributions are
  - the *degenerate* distribution (constant service time) and
  - the *Erlang (gamma)* distribution.

Cost-based Queuing Decision Model

- **Cost optimization model**: we seek the minimization of the sum of the two costs:-
  - the cost of offering the service and the
  - cost of waiting.
- Fig 2 depicts a typical cost model (in Rs per unit time).

- **The main obstacle**: difficulty of obtaining reliable estimates of the cost of waiting, particularly when human behavior is involved.

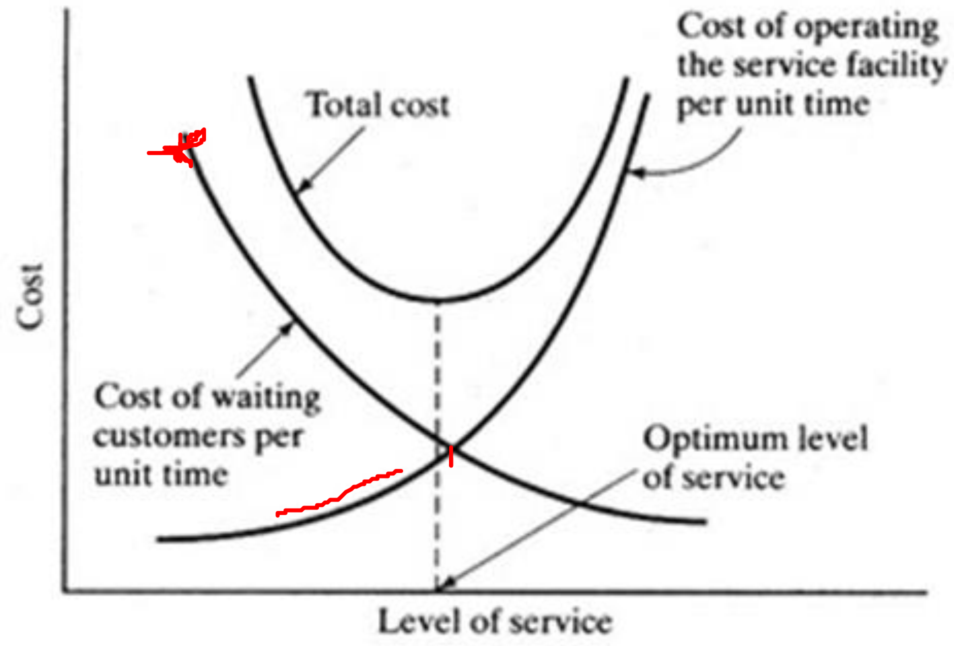


Fig 2

An Elementary Queuing Process

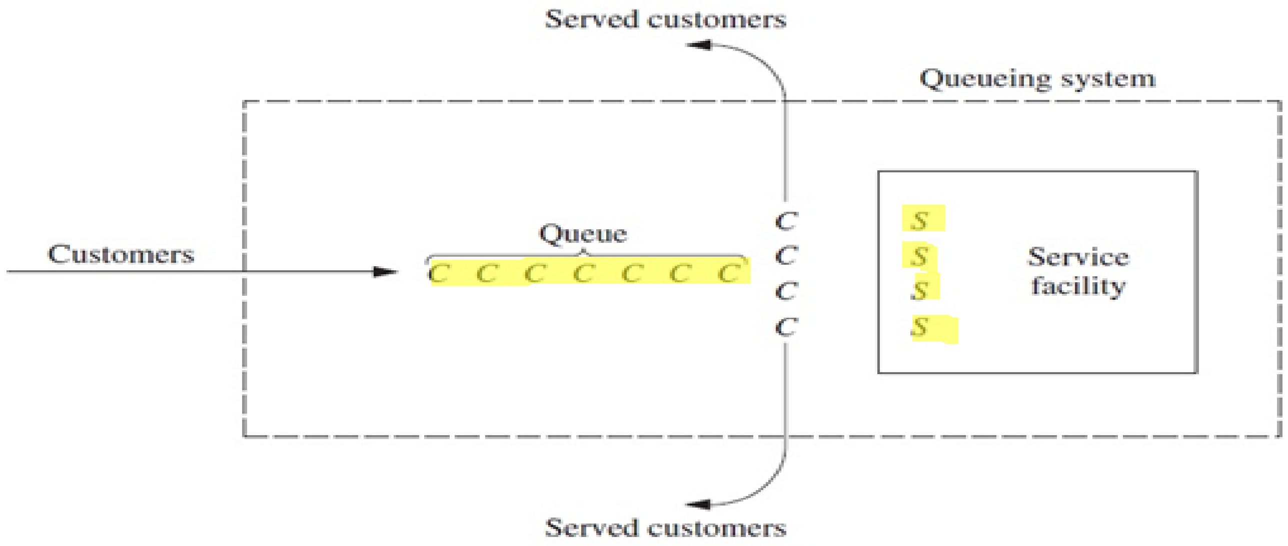
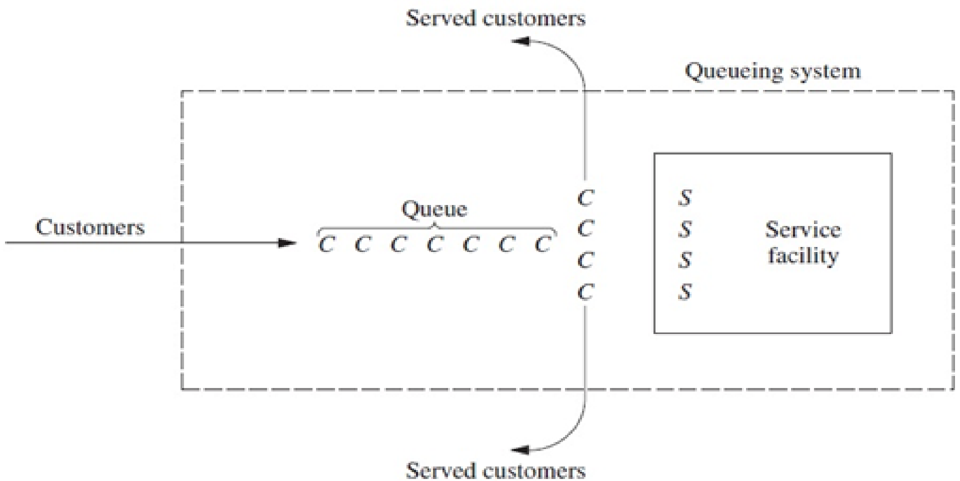


Fig 3: An elementary Queuing System

- Queuing theory has been applied to many different types of waiting-line situations.
- The most prevalent type of situation is the following:
  - A single waiting line (which may be empty at times) forms in the front of a single service facility, within which are stationed one or more servers.
  - Each customer generated by an input source is serviced by one of the servers, perhaps after some waiting in the queue (waiting line).

An Elementary Queuing Process

- Furthermore, servers need not even be people.
- In many cases, a server can instead be a machine, a vehicle, an electronic device, etc.
- By the same token, the customers in the waiting line need not be people.
- For example, they may be items waiting for a certain operation by a given type of machine, or they may be cars waiting in front of a tollbooth.



Examples

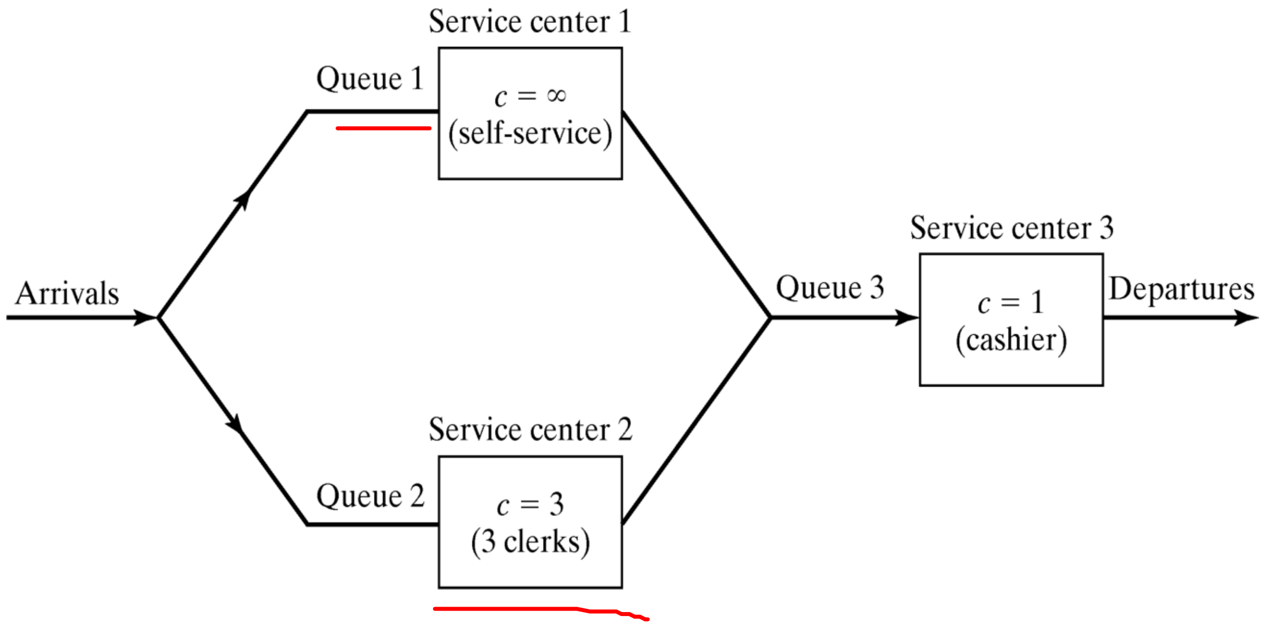
System	Customers	Server
Reception desk	People	Receptionist
Hospital	Patients	Nurses
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Computer	Jobs	CPU, disk, CD
Network	Packets	Router

Example Problem

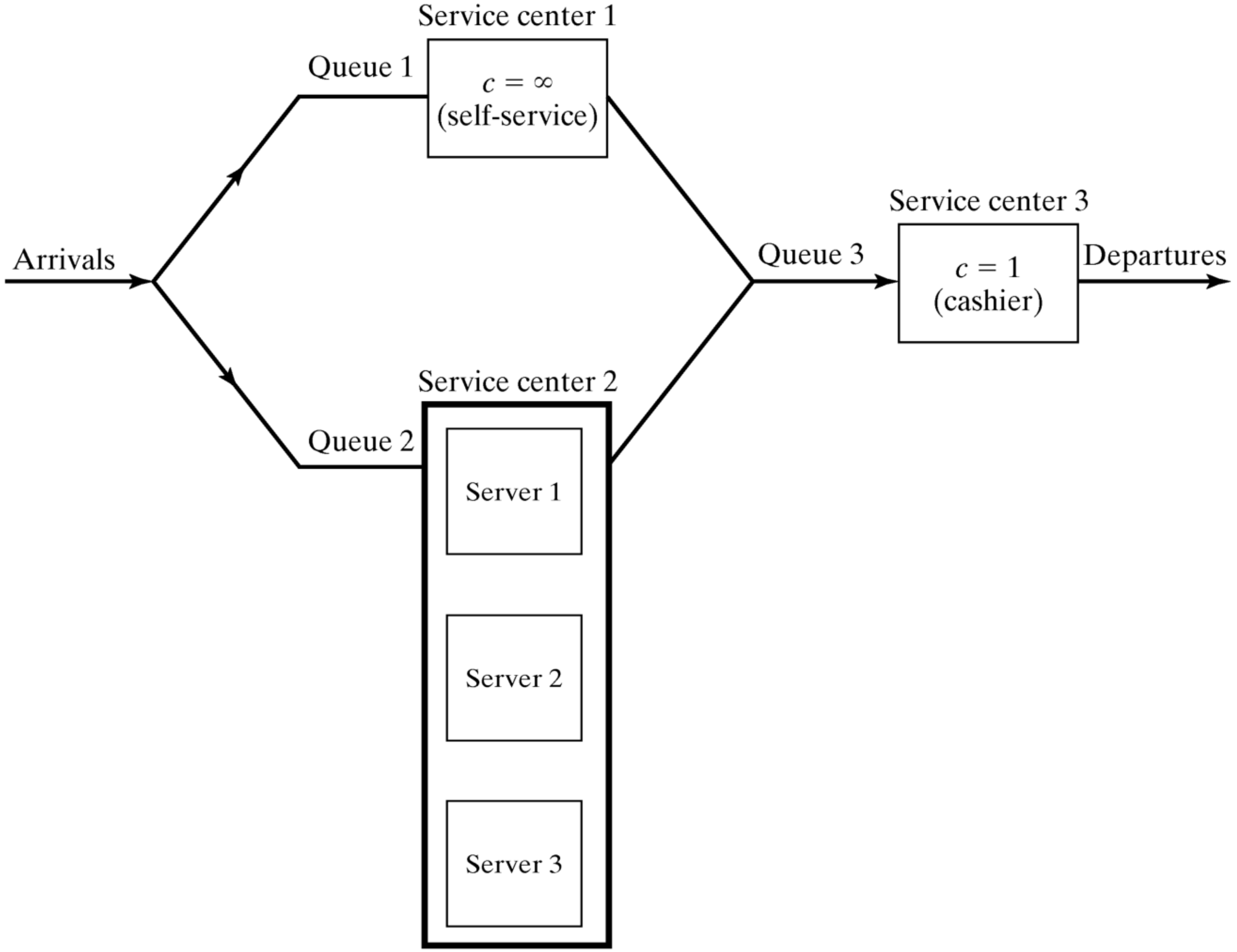
Consider a discount warehouse where customers may:

- serve themselves before paying at the cashier (service center 1) or
- served by a clerk (service center 2)

Solution:



Solution:



Service center 1
Queue 1
$c = \infty$
(self-service)
Arrivals
Service center 3
Queue 3
$c = 1$
(cashier)
Departures
Service center 2
Queue 2
Server 1
Server 2
Server 3