# QUEUING THEORY: AN INTRODUCTION

*"Delay is the enemy of efficiency"  and "Waiting is the enemy of utilization"*
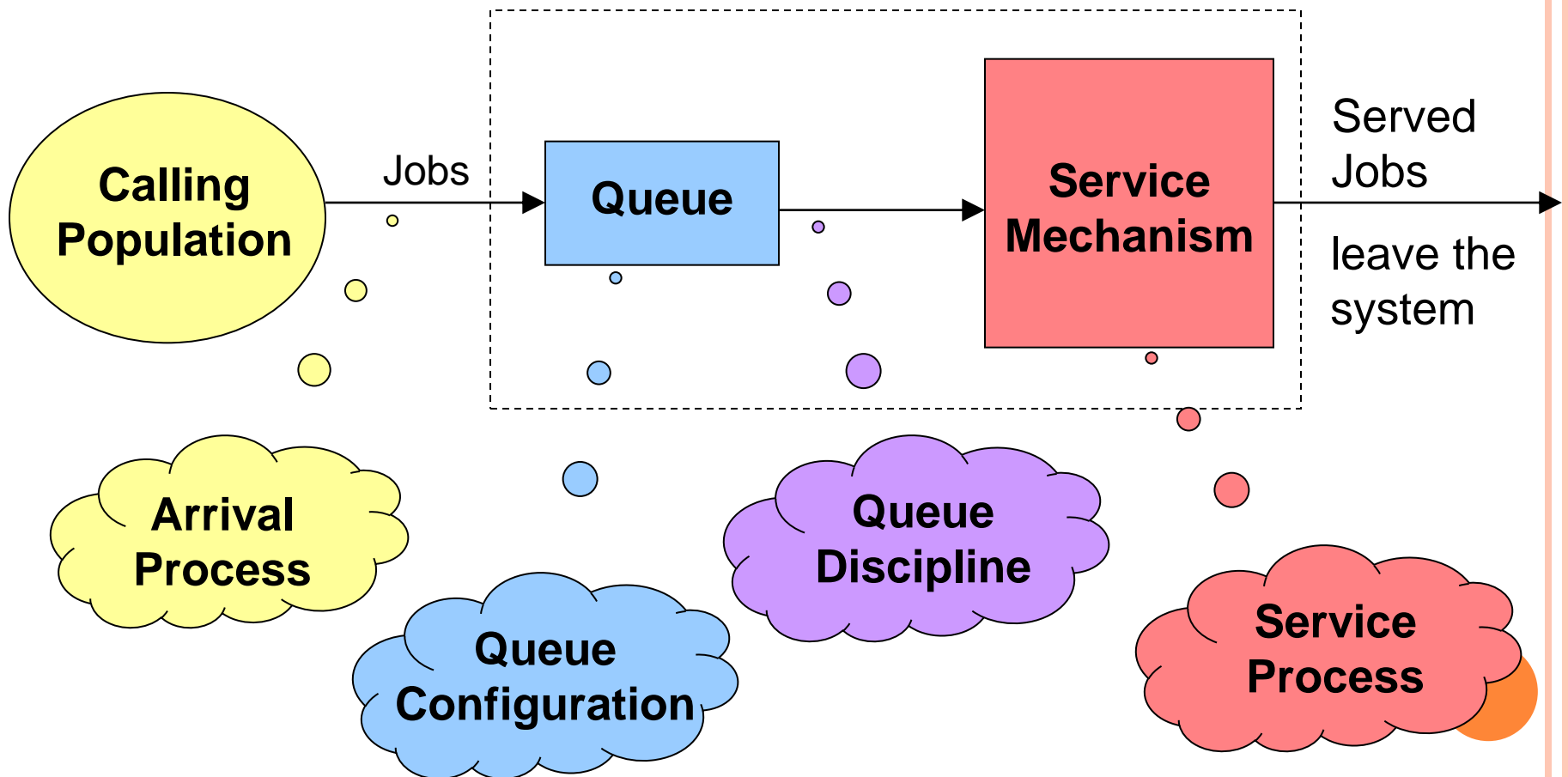
# OVERVIEW

- **Components of a Basic Queuing Process**

- **The Calling Population**

- **Arrival Pattern**

- **Service Pattern**

- **Number of Servers**

- **Service Discipline**

- **System Capacity**

- **Customer's Behavior**

- **Service Facility Behavior**

# COMPONENTS OF A BASIC QUEUING PROCESS

**Input Source**                    **The Queuing System**

**Calling Population** → Jobs → **Queue** → **Service Mechanism** → Served Jobs leave the system

**Arrival Process**

**Queue Configuration**

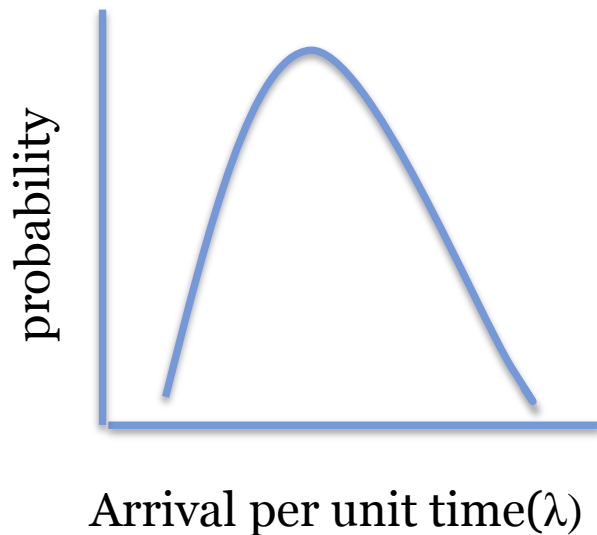**Queue Discipline**

**Service Process**

# THE CALLING POPULATION

- Population of customers or jobs

- The size can be finite or infinite
  - The latter is most common

- Can be homogeneous
  - Only one type of customers/ jobs

- Or heterogeneous
  - Several different kinds of customers/jobs

# ARRIVAL PATTERN

Poisson Distribution



probability

Arrival per unit time($\lambda$)

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

- In what pattern do jobs / customers arrive to the queuing system?
  - ➤ Scheduled, or
  - ➤ Random fashion

- The time duration between each customer's arrival is known as **interarrival time**.

- We assume it to follow **Poisson Distribution.**

- Other distribution: Constant, Erlang-k**,** Hyper-exponential, etc.

# SERVICE PATTERN

Exponential Distribution



probability

Customer served per unit time (μ)
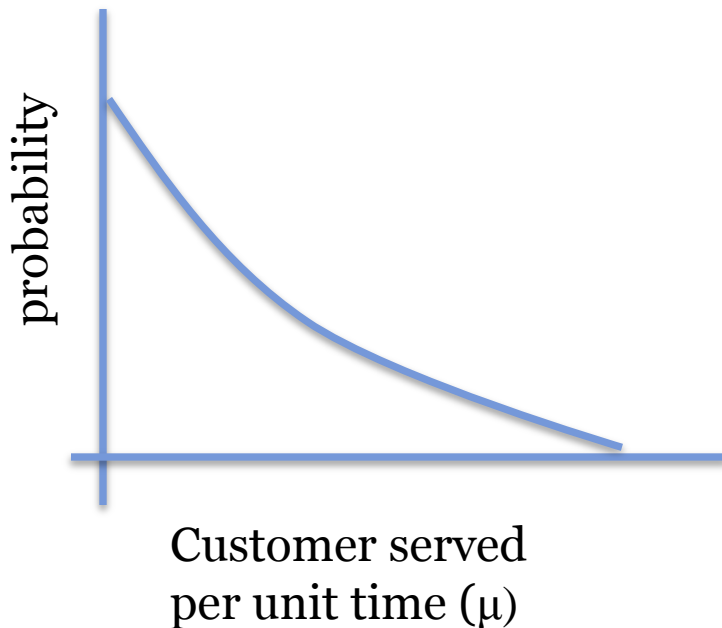
$$P(X \leq t) = 1 - e^{-\mu t}$$

Assumption: $\lambda < \mu$

- **How long does it take to service a job or customer?**
- Number of servers and speed of service to be considered.
- The time taken by a server to service a customer is known as **Service Time**.
- It is represented by **Exponential Distribution.**
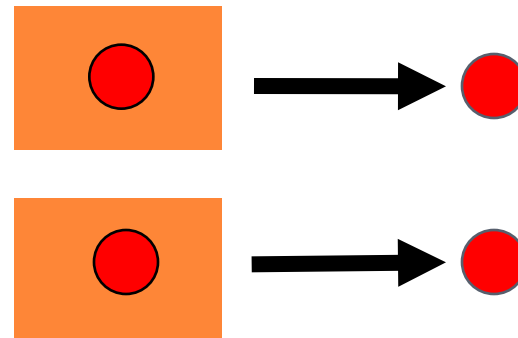- Other distribution: Constant, Erlang-k, Hyper-exponential, etc.

# NUMBER OF SERVERS

- **How many servers are available?**

**Single Server Queue**

**Multiple Server Queue**

**Infinite Server Queue**

# SERVICE/QUEUE DISCIPLINE

- How are jobs / customers selected from the queue for service?
  - FCFS (First-Come-First-Served)/FIFO
  - LCFS (Last-Come-First-Served)/LIFO
  - SIRO (Service in random order)/RSS
  - PRI (Priority service)



National Library of Australia · nla.pic-an12942889-30-v

# System Capacity

- Maximum number of customers that can be accommodated in the queue.
- Assumed to be of infinite capacity.
- Other queuing systems, called "loss systems," have zero queue capacity, e.g. dial-up telephone systems.
- K: to represent maximum no. of customers.

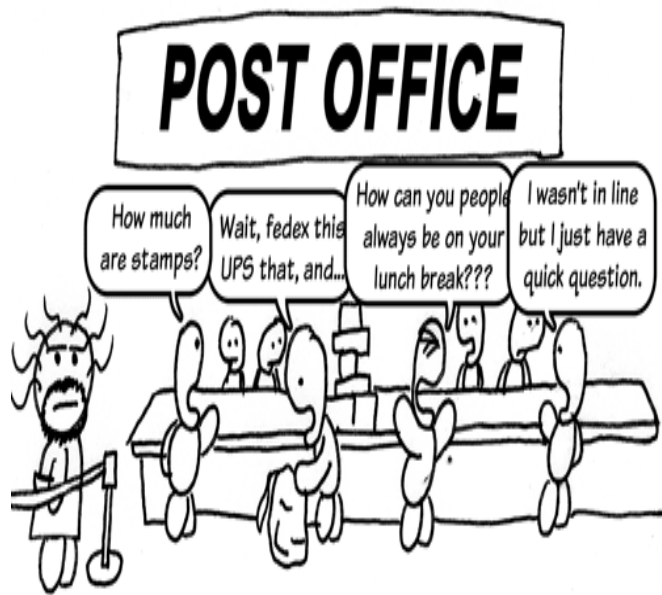Number of customers

Finite     Infinite
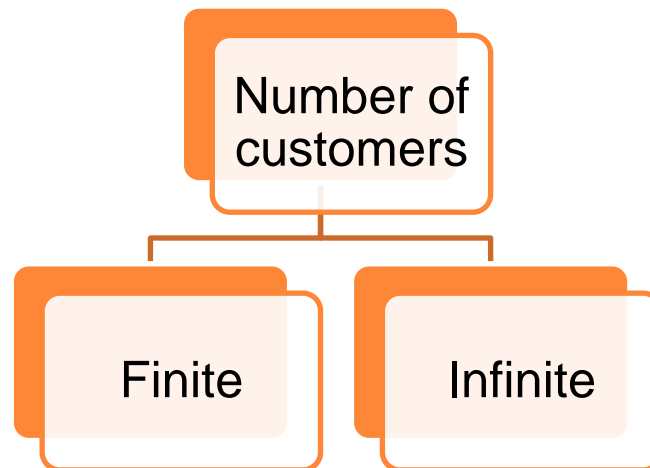
# Customer's Behavior

- **Balking-** When a customer leaves the queue because it is too long, has no time to wait, no space to stand, etc.

- **Reneging-** When a customer leaves the queue because of his impatience.

- **Jockeying-** When a customer shifts from one queue to another.

# SERVICE FACILITY BEHAVIOR



POST OFFICE

How much are stamps?

Wait, fedex this UPS that, and...

How can you people always be on your lunch break???

I wasn't in line but I just have a quick question.

2/9/05

www.kickinthehead.org    © 2003-05 Derek Lieu

- **Failure**: A server may fail while serving a customer, thereby interrupting service until a repair can be made.
- **Changing service rate**: A server may speed up or slow down, depending on the number of customers in the queue.
- **Batch processing:** A server may service several customers simultaneously.