Queuing Theory: An Introduction

"Delay is the enemy of efficiency" and "Waiting is the enemy of utilization"

OVERVIEW

- Service Utilization Factor & Traffic Intensity
- Little's Formulas
- Classification of Queuing Systems
- Types of Queues of Interest
- Assumption in Queuing Model
- Limitations of Queuing Model

Service Utilization Factor

- Consider an M/M/1 queue with arrival rate = λ and service intensity = μ
- λ = Expected capacity demand per time unit
- μ = Expected capacity per time unit

$\Rightarrow$

$$\rho = \frac{\text{Capacity Demand}}{\text{Available Capacity}} = \frac{\lambda}{\mu}$$

- Similarly, if there are c servers in parallel, i.e., an M/M/c system but the expected capacity per time unit is then c*μ

$\Rightarrow$

$$\rho = \frac{\text{Capacity Demand}}{\text{Available Capacity}} = \frac{\lambda}{c * \mu}$$

Traffic Intensity

- The ratio λ/μ is called the traffic intensity or the utilization factor and it determines the degree to which the capacity of service station is utilized.

$$\rho = \frac{Mean\ Rate\ of\ Arrival\ in\ the\ Queue(\lambda)}{Mean\ Service\ Rate(\mu)}$$

Little's Formulas

- Little's Formulas represent important relationships between L, L_q, W, and W_q.
- These formulas apply to systems that meet the following conditions:
  - Single queue systems,
  - Customers arrive at a finite arrival rate λ, and
  - The system operates under a steady state condition.

$$L = \lambda W$$

$$L_q = \lambda W_q$$

$$L = L_q + \lambda/\mu$$

For the case of an infinite population

Little's Formulas

EXAMPLE:

- A monitor on a disk server showed that the average time to satisfy an I/O request was 100 milliseconds. The I/O rate was about 100 requests per second. What was the mean number of requests at the disk server?

- **Using Little's law:**

  Mean number in the disk server

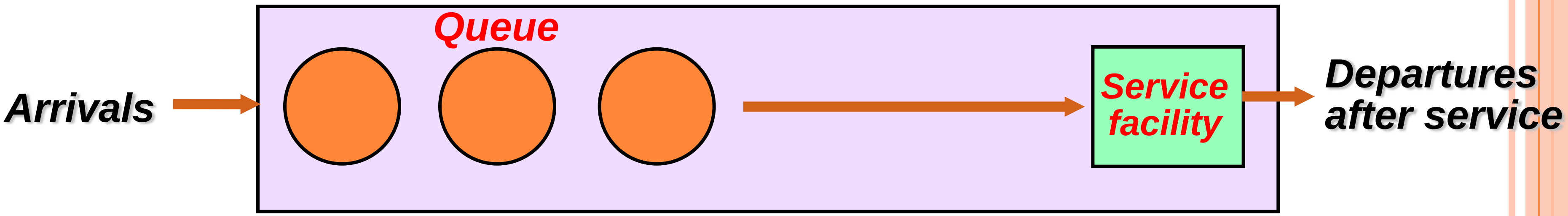
  = Arrival rate × Response time

  = 100 (requests/second) ×(0.1 seconds)

  = 10 requests

Classification of Queuing Systems

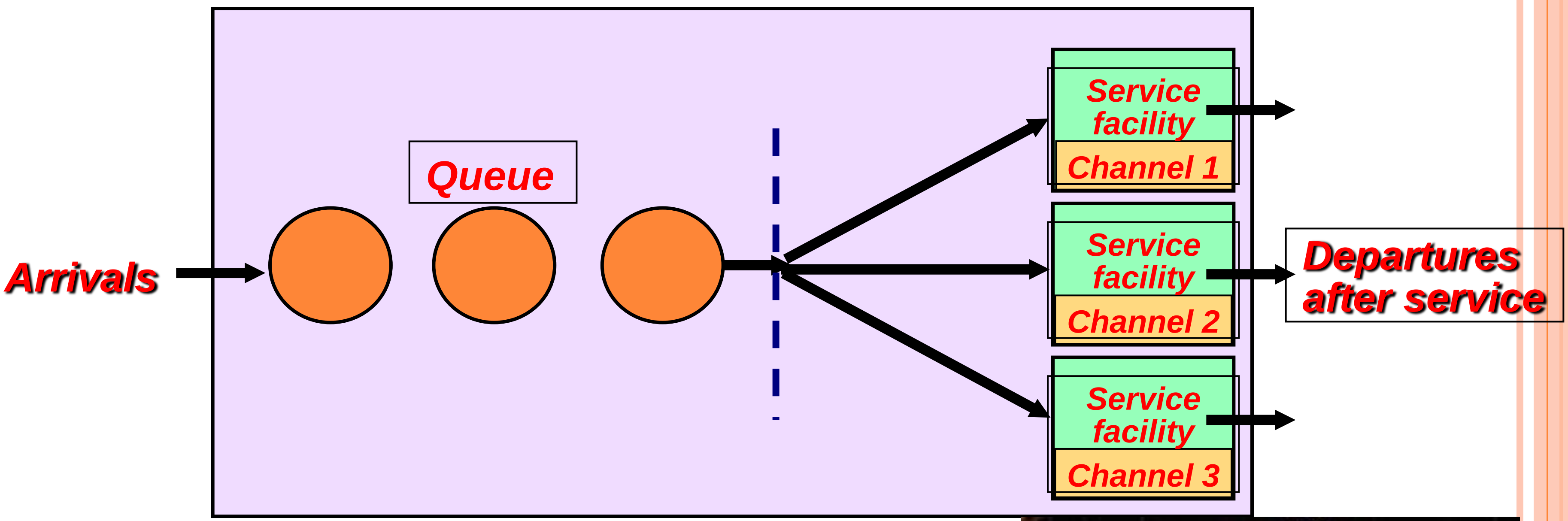
1. SINGLE-SERVER SINGLE-STAGE QUEUE



e.g., Your family dentist's office, Library counter, etc.

Classification of Queuing Systems

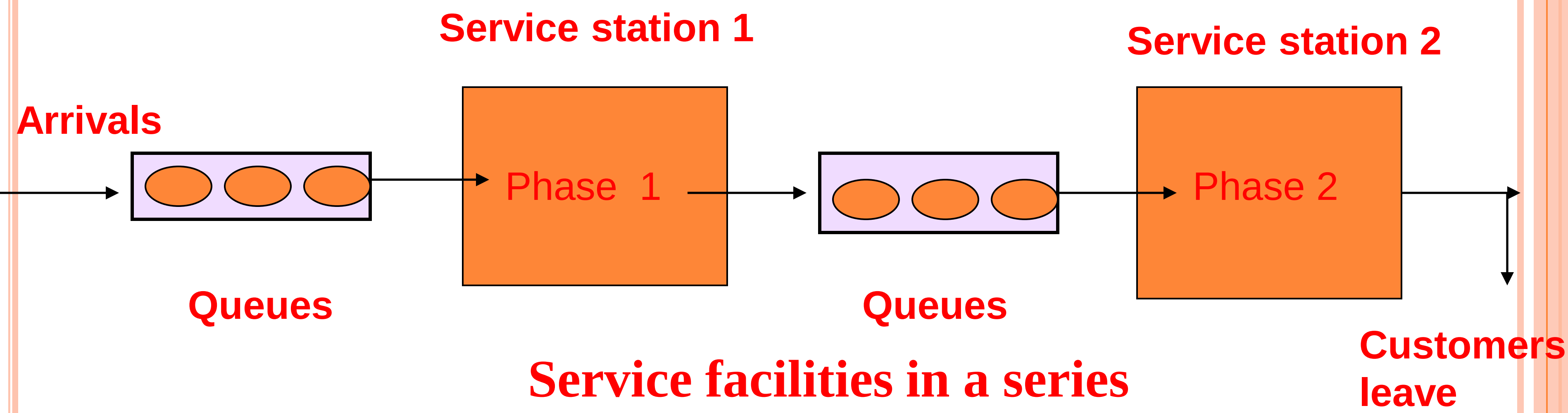
2. MULTIPLE-SERVER SINGLE-STAGE QUEUE



e.g., Booking at a service station

Classification of Queuing Systems

3. SINGLE-SERVER MULTIPLE-STAGE QUEUE



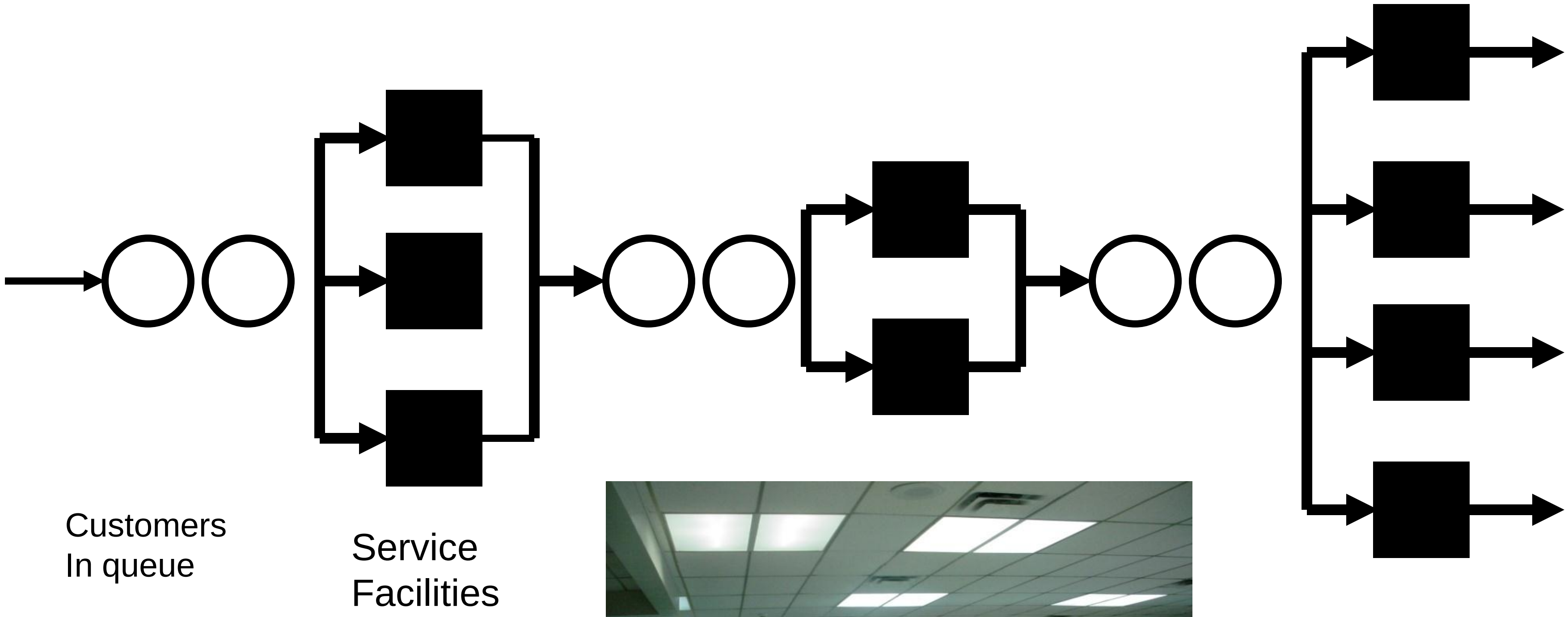
Service facilities in a series

***e.g.**, Cutting, turning, knurling, drilling, grinding, packaging operation of steel*

Pharmacy Conveyor System >>>>>

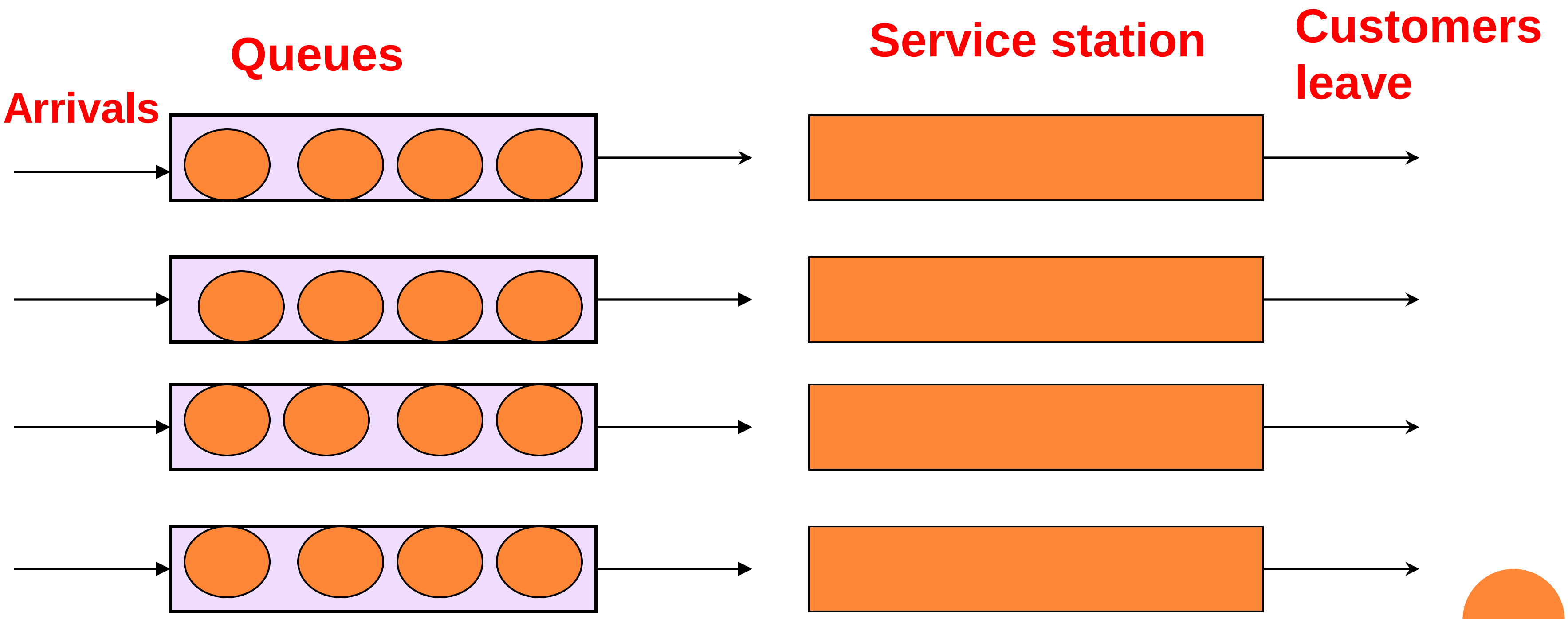
Classification of Queuing Systems

4. MULTIPLE-SERVER MULTIPLE-STAGE QUEUE



Classification of Queuing Systems

5. MULTIPLE, PARALLEL FACILITIES WITH MULTIPLE QUEUES MODEL



e.g., Different cash counters in electricity office

Types of Queues of Interest

- Analytical models for estimating capacity and related metrics
  - Single Server
    - M/M/1, M/G/1, M/D/1, G/G/1, etc.
  - Multiple Server
    - M/M/s, M/G/∞, etc.
  - Multiple Stage
    - Markov Chain models

Assumption in Queuing Model

- The customer arrive for service at a single service facility at random according to Poisson distribution with mean arrival rate λ.
- The service time has exponential distribution with mean service rate μ.
- The service discipline followed is First Come First Served.
- Customer Behavior is Normal
- Service facility behavior is Normal
- The calling source has infinite size
- The mean arrival rate is less than the mean service rate
- The waiting space available for customer in the queue is infinite

Limitations of Queuing Model

- The waiting space for the customer is usually limited
- The arrival rate may be state dependent
- The arrival process may not be stationary
- The population of customers may not be infinite and the queuing discipline may not be First Come First Serve
- Services may not be rendered continuously
- The Queuing system may not have reached the steady state. It may be, instead, in transient state


Thank You!