Queuing Theory

Introduction

Queuing Theory is the mathematical study of waiting lines or queues. It analyzes the behavior of queues formed by entities (like customers, jobs, or data packets) waiting for service from one or more servers. The goal is to understand and optimize system performance, such as minimizing wait times and maximizing service efficiency.

It is widely used in operations research, telecommunications, manufacturing, health services, and computing.

Definition of Terms in Queuing Models

- **Arrival Rate (λ):** The average number of arrivals per time unit.
- **Service Rate (μ):** The average number of services completed per time unit.
- **Queue Discipline:** The rule by which waiting customers are served (e.g., First In First Out - FIFO).
- **System Capacity:** The maximum number of customers allowed in the system (including those waiting and being served).
- **Population Source:**
    - Finite: A limited number of potential arrivals.
    - Infinite: An unlimited or very large number of potential arrivals.
- **Utilization Factor (ρ):** Given by $\rho = \lambda / \mu$, it indicates how busy the server is.
- **Queue Length:** The average number of customers waiting in line.
- **System Length:** The total number of customers in the queue and in service.
- **Waiting Time:** The time a customer spends in the queue or in the system.

Single-Channel Infinite Population Model (M/M/1)

Assumptions:

- One server (single channel).
- Infinite population of customers.
- Poisson arrivals and exponential service times.
- Unlimited queue capacity.
- First-Come-First-Served service discipline.

Performance Metrics:

- Average number in the system (L):
  $L = \lambda / (\mu - \lambda)$
- Average number in the queue (Lq):
  $Lq = (\lambda^2) / [\mu(\mu - \lambda)]$
- Average time in the system (W):
  $W = 1 / (\mu - \lambda)$
- Average waiting time in queue (Wq):
  $Wq = \lambda / [\mu(\mu - \lambda)]$

Multi-Channel Service Infinite Queue Model (M/M/c)

Assumptions:

- Multiple servers (c channels).
- Infinite queue capacity and population.
- Poisson arrivals and exponential service times.

- FIFO service discipline.

This model applies to environments like banks, help desks, and hospitals with multiple service counters.

Note: Calculations require Erlang-C formulas, which consider the number of servers, arrival rate, and service rate.

Finite Population Model

Features:

- Only a limited number of customers (N) in the population.
- The arrival rate depends on the number of customers not currently in the system.
- Suitable for repair stations, machine shops, or clinics with a small number of users.

Implications:

- Arrival rate is not constant—it decreases as more customers are in the system.
- Special models and formulas (like birth-death processes) are used for analysis.

Applications of Queuing Theory

- Service centers (e.g., banks, customer support)
- Assembly lines in manufacturing
- Hospital emergency rooms
- Cloud computing and data networks
- Traffic systems and toll booths
- Call centers and help desks