

Introduction to Queueing Theory

1 What is Queueing Theory?

Queueing theory is the mathematical study of waiting lines (queues). It analyzes systems where customers or requests arrive, wait if necessary, receive service, and then depart.

This theory is widely used in various sectors including:

- **Informatics:** Servers, networks, and data centers.
- **Industry:** Production lines.
- **Transportation:** Traffic and logistics.
- **Telecommunications.**

1.1 Applications in Computer Science

In the field of Computer Science, queueing theory is essential for:

- Designing servers and data centers.
- Analyzing computer networks (packet flow in routers).
- Operating Systems (process scheduling).
- Cloud computing and distributed systems.

2 Basic Elements of a Queueing System

A typical queueing system consists of the following components:

1. **Source:** The population source that generates requests.
2. **Customers/Jobs:** The entities arriving for service.
3. **Queue (Buffer):** The area where customers wait before being served.
4. **Servers:** The resources that perform the service.
5. **Queue Discipline:** The rule determining the order in which customers are served. Common disciplines include:
 - **FIFO:** First In - First Out.
 - **LIFO:** Last In - First Out.
 - **Priority:** Customers with higher priority are served first.

3 Performance Metrics

To evaluate the efficiency of a system, we use specific metrics:

- λ (lambda): Arrival rate (customers per unit of time).
- μ (mu): Service rate (customers per unit of time).
- L : Average number of customers in the system (queue + service).
- L_q : Average number of customers in the queue.
- W : Average time a customer spends in the system.
- W_q : Average time a customer waits in the queue.
- ρ (rho): Utilization factor (percentage of time the server is busy).

4 Kendall's Notation

Kendall's notation is a standard system used to describe and classify queueing nodes. It is written as **A/B/s/K/N/D**.

A - Arrival Process: Describes how customers arrive.

- **M (Markovian):** Exponential distribution (Poisson arrival process).
- **D (Deterministic):** Fixed time intervals between arrivals.
- **G (General):** Any general distribution.

B - Service Process: Describes the time required for service.

- **M:** Exponential service time distribution.
- **D:** Fixed service time.
- **G:** Any distribution.

s - Servers: The number of servers operating in parallel (e.g., 1 or 2).

K - System Capacity: The maximum number of customers allowed in the system (queue + service). If infinite, it is denoted as ∞ .

N - Population Size: The total number of potential customers. Usually considered infinite (∞).

D - Queue Discipline: The service order (e.g., FIFO, LIFO).

Note: The most common model is **M/M/1**, which implies Poisson arrivals, exponential service times, and a single server.

5 Little's Law

Little's Law is a fundamental theorem valid for all queueing systems. It states that the long-term average number of customers in a stable system is equal to the long-term average effective arrival rate multiplied by the average time a customer spends in the system.

$$L = \lambda \cdot W \tag{1}$$

Similarly, for the queue: $L_q = \lambda \cdot W_q$.

Example: Consider a bank teller:

- Customers arrive at a rate of $\lambda = 30$ per hour.
- Each customer waits 10 minutes and is served for 10 minutes. Total time $W = 20$ minutes (1/3 hour).
- Using Little's Law: $L = 30 \times (1/3) = 10$.
- Interpretation: On average, there are 10 customers in the bank (in queue + being served).

6 The Classical M/M/1 Model

The M/M/1 model is the simplest and most widely used queueing model.

6.1 Assumptions

- Arrivals follow a Poisson process (rate λ).
- Service times follow an Exponential distribution (rate μ).
- Single server ($s = 1$).
- Infinite queue capacity.
- FIFO discipline.

Note on Distributions: The Poisson distribution counts "how many events" occur in a time period, while the Exponential distribution measures "how much time" until the next event.

6.2 Key Formulas

For a stable system, we require $\lambda < \mu$ (or $\rho < 1$). The steady-state results are:

$$\text{Utilization Factor: } \rho = \frac{\lambda}{\mu} \tag{2}$$

$$\text{Avg. Customers in System: } L = \frac{\rho}{1 - \rho} \tag{3}$$

$$\text{Avg. Customers in Queue: } L_q = \frac{\rho^2}{1 - \rho} \tag{4}$$

$$\text{Avg. Time in System: } W = \frac{1}{\mu - \lambda} \tag{5}$$

$$\text{Avg. Time in Queue: } W_q = \frac{\rho}{\mu(1 - \rho)} \tag{6}$$

7 Practical Exercise

Scenario: A helpdesk operates under the following conditions:

- Customer arrivals follow a Poisson distribution with $\lambda = 4$ customers/hour.
- Service time is exponential with $\mu = 6$ customers/hour.
- There is 1 server (M/M/1).

Calculations:

1. **Utilization (ρ):**

$$\rho = \frac{4}{6} \approx 0.667$$

2. **Average Customers in System (L):**

$$L = \frac{0.667}{1 - 0.667} = 2 \text{ customers}$$

3. **Average Customers in Queue (L_q):**

$$L_q = \frac{0.667^2}{1 - 0.667} \approx 1.33 \text{ customers}$$

4. **Average Time in Queue (W_q):**

$$W_q = \frac{1.33}{4} \approx 0.333 \text{ hours (20 minutes)}$$

5. **Average Time in System (W):**

$$W = \frac{1}{6 - 4} = 0.5 \text{ hours (30 minutes)}$$