

Operations Research

Unit 4: Queuing Theory

(6 Hours)

1. Basis of Queuing Theory

Queuing Theory (also called Waiting Line Theory) is the mathematical study of waiting lines, or queues, that form whenever the demand for a service temporarily exceeds the capacity available to provide it. It was founded by the Danish engineer **A. K. Erlang** in 1909, who studied the fluctuating demand for telephone facilities at the Copenhagen Telephone Company and developed formulas to predict waiting times and call congestion.

The central problem addressed by queuing theory is a trade-off: providing too little service capacity causes long queues, customer dissatisfaction, and lost business, while providing too much service capacity is wasteful and increases idle-time cost. Queuing theory provides the mathematical tools to find the balance that minimises total cost — the sum of the cost of waiting and the cost of service — or to meet a desired service-level target at least cost.

Example 4.1 — Why Queuing Theory Matters

A supermarket has 4 checkout counters. During peak hours customers arrive at an average rate of 60 per hour, and each counter can serve 20 customers per hour. Without queuing analysis, the manager may open too few or too many counters, causing either long queues or idle cashiers. Queuing theory helps determine the minimum number of counters needed to keep the average waiting time below an acceptable limit, such as 2 minutes.

2. Elements of a Queuing System

Every queuing system, regardless of context, can be described using the same basic set of structural elements:

Element	Description
Input / Arrival Process	Describes how customers arrive at the system: arrival pattern (random or scheduled), arrival rate (λ), and source population (finite or infinite).
Queue / Waiting Line	The line of customers who have arrived but are not yet being served. Characterised by its capacity — finite or infinite.
Queue Discipline	The rule used to select the next customer for service, e.g. First-Come-First-Served (FCFS), Last-Come-First-Served (LCFS), Priority, or Random Selection for Service (RSS).
Service Mechanism	Describes how service is provided: number of servers, arrangement of servers (single, parallel, series), and the service-time distribution (service rate μ).
Output / Departure	Customers leave the system after service is completed, either returning to the source population or exiting altogether.

Together these elements determine the “shape” of a queuing system and are the building blocks used to classify and solve it mathematically.

3. Kendall's Notation

Proposed by **D. G. Kendall** in 1953, Kendall's Notation is a compact shorthand used to classify and describe a queuing model. The general form is:

$$(A / B / C) : (D / E)$$

Symbol	Meaning
A	Probability distribution of the arrival (inter-arrival time) process.
B	Probability distribution of the service time process.
C	Number of parallel servers (channels) in the system.
D	Maximum capacity of the system (queue + service), finite (N) or infinite (∞).
E	Queue discipline, e.g. FCFS, LCFS, Priority.

Common symbols used for A and B include: **M** (Markovian / exponential, memoryless), **D** (Deterministic, constant time), **E_k** (Erlang-k distribution), and **G** (General, any

arbitrary distribution).

Example 4.2 — Reading Kendall's Notation

The notation $(M / M / 1) : (\infty / \text{FCFS})$ describes a single-server queue with Poisson (Markovian) arrivals, exponential (Markovian) service times, infinite queue capacity, and First-Come-First-Served discipline — the simplest and most widely used queuing model, often called the **M/M/1 model**.

4. Operating Characteristics of a Queuing System

Once a queuing model is specified, its performance is measured using a standard set of operating characteristics (also called performance measures):

Symbol	Operating Characteristic
λ	Mean customer arrival rate (customers per unit time).
μ	Mean service rate per server (customers per unit time).
$\rho = \lambda/\mu$	Utilisation factor / traffic intensity — the fraction of time the server is busy.
L_s	Expected (average) number of customers in the system (waiting + being served).
L_q	Expected (average) number of customers in the queue (waiting only).
W_s	Expected (average) time a customer spends in the system.
W_q	Expected (average) time a customer spends waiting in the queue.
P_0	Probability that the system is empty (no customers present).
P_n	Probability that exactly n customers are present in the system.

For the basic $(M/M/1):(\infty/\text{FCFS})$ model, these characteristics reduce to simple closed-form formulas:

$$\rho = \lambda/\mu \quad L_s = \lambda/(\mu - \lambda) \quad W_s = 1/(\mu - \lambda) \quad L_q = \rho \times L_s$$

Example 4.3 — Computing Operating Characteristics

A bank counter (single server) has customers arriving at $\lambda = 8$ per hour (Poisson) and is served at $\mu = 10$ per hour (exponential). Then $\rho = 8/10 = 0.8$ (the teller is busy 80% of the time), the average number of customers in the system $L_s = 8/(10-8) = 4$ customers, and the average waiting time in the system $W_s = 1/(10-8) = 0.5$ hour, i.e. 30 minutes per customer.

5. Classification of Queuing Models

Queuing models are classified mainly on the basis of the number of servers and the capacity of the system. The commonly studied models are:

Model	Description
(M/M/1) : (∞ /FCFS)	Single server, infinite queue capacity, Poisson arrivals, exponential service — the classical and simplest queuing model.
(M/M/1) : (N/FCFS)	Single server with finite system capacity N; customers arriving when the system is full are turned away (balking).
(M/M/c) : (∞ /FCFS)	Multiple (c) parallel servers with a common queue and infinite capacity — e.g. multi-counter bank or call centre.
(M/M/c) : (N/FCFS)	Multiple servers with finite system capacity N.
(M/D/1)	Single server with Poisson arrivals but constant (deterministic) service time, e.g. an automated machine with fixed cycle time.
(M/E _k /1)	Single server with service time following an Erlang-k distribution, used when service has k sequential sub-stages.
(M/G/1)	Single server with a general (arbitrary) service-time distribution; solved using the Pollaczek–Khinchine formula.
Finite-Source Models	Used when the population generating arrivals is itself limited, e.g. the machine-repair (machine-interference) problem with a fixed number of machines and repairmen.

6. Applications of Queuing Theory

Sector	Typical Application
Banking	Number of teller counters / ATMs required to keep waiting time within limits.
Telecommunications	Call-centre staffing, trunk-line sizing, network packet buffering.
Healthcare	Outpatient appointment scheduling, emergency-room triage and bed allocation.
Transportation	Toll-booth and runway capacity planning, traffic-signal design.
Manufacturing	Machine-repair problems, buffer sizing between production stages.
Retail & E-commerce	Checkout-counter planning, website/server load and request-queue management.
Computer Systems	CPU scheduling, network router buffer design, cloud-server load balancing.

7. Limitations of Queuing Theory

- **Distributional Assumptions:** Most closed-form models assume Poisson arrivals and exponential service times, which may not hold in every real situation.
- **Steady-State Focus:** Many formulas describe long-run, steady-state behaviour and may not capture short-term or transient congestion accurately.
- **Behavioural Factors Ignored:** Customer behaviour such as balking (refusing to join), reneging (leaving before being served), and jockeying (switching queues) complicates basic models.
- **Complexity for Realistic Systems:** Multi-stage, multi-server, finite-population, or priority-based systems often have no simple closed-form solution and require simulation.
- **Data Requirements:** Accurate estimates of arrival and service rates are needed; poor estimates lead to misleading conclusions.

Unit 4 — Key Takeaways

- Queuing theory studies waiting lines and was founded by A. K. Erlang in 1909 for telephone traffic problems.

- A queuing system has five key elements: arrival process, waiting line, queue discipline, service mechanism, and departure.
- Kendall's Notation, (A/B/C):(D/E), gives a compact way to classify any queuing model.
- Key operating characteristics are λ , μ , ρ , L_s , L_q , W_s , and W_q .
- Queuing models are classified by number of servers and system capacity — e.g. M/M/1, M/M/c, M/D/1, M/G/1, and finite-source models.
- Applications span banking, telecommunications, healthcare, transportation, manufacturing, and computer systems.

— End of Unit 4: Queuing Theory —