Queuing Theory: An Introduction

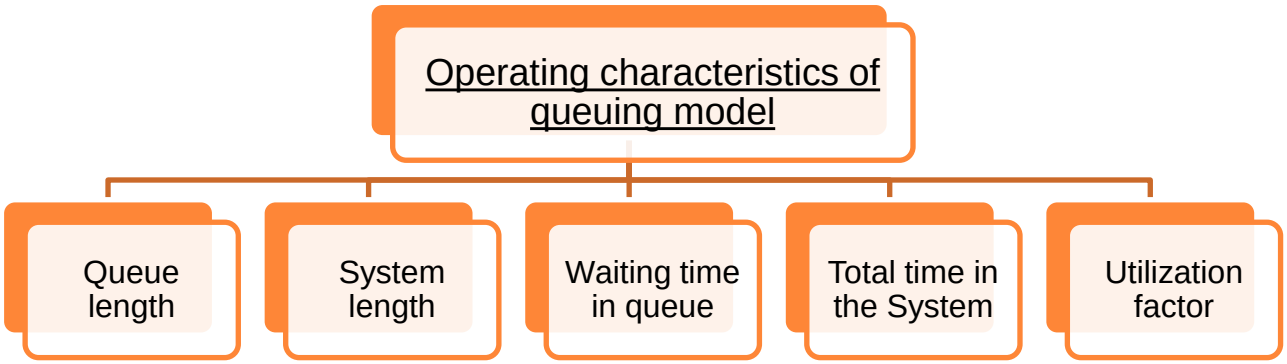
"Delay is the enemy of efficiency" and "Waiting is the enemy of utilization"

Overview

- Transient & Steady State of the system
- Waiting and Idle time costs
- The Queuing Cost Trade-off
- Kendall Notations
- A Commonly Seen Queuing Model
- Queuing Model with Key Variables
- Steady State Performance Measures

Transient & Steady State of the system

- Queuing analysis involves the system's behavior over time. If the operating characteristics vary with time then it is said to be transient state of the system.
- If the behavior becomes independent of its initial conditions (no. of customers in the system) and of the elapsed time is called Steady State condition of the system.



Waiting and Idle time costs

Cost of waiting customers

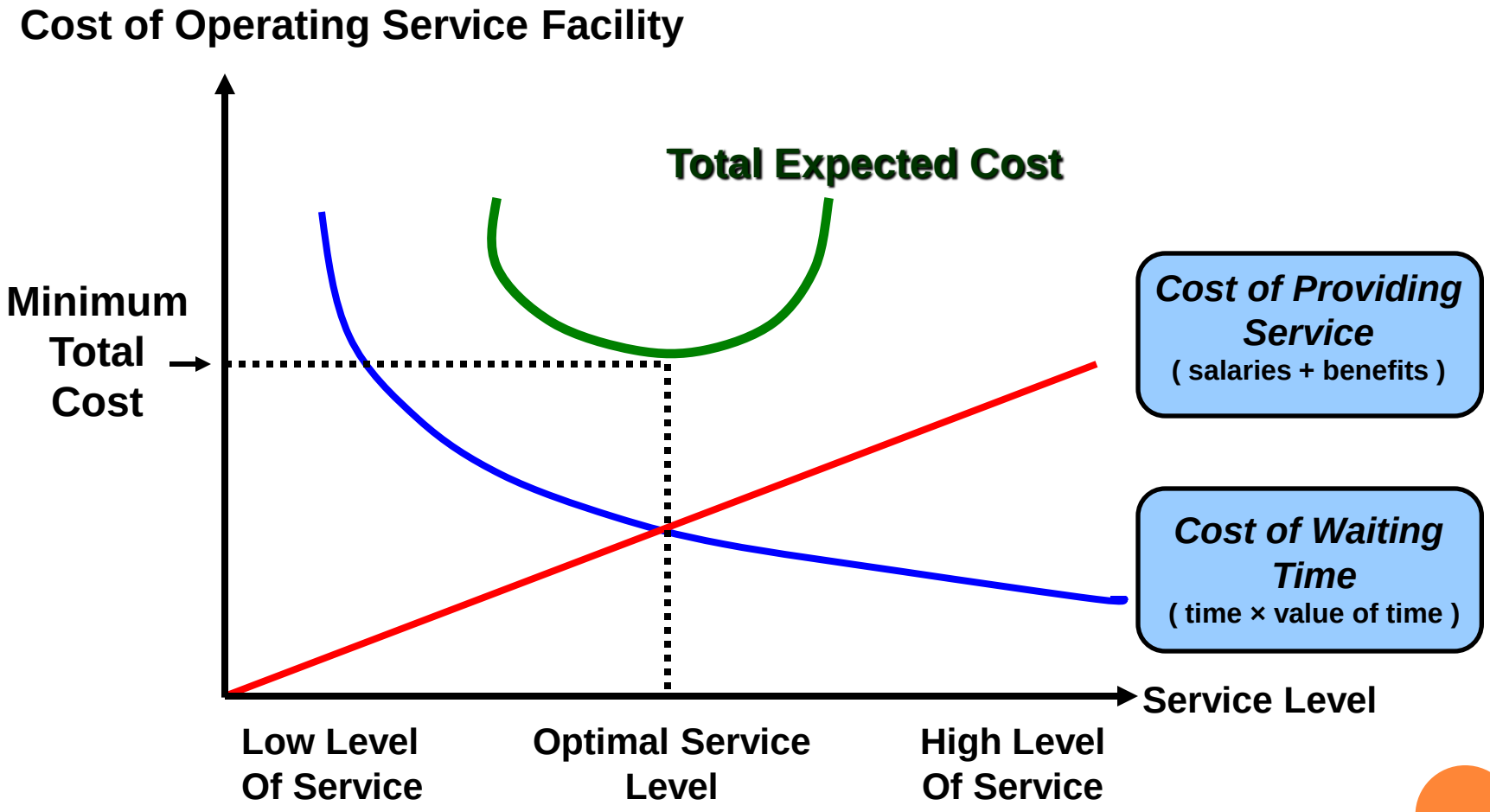
- Indirect cost of business loss
- Direct cost of idle equipment or person.

Cost of idle service facility

- Payment to be made to the servers(engaged at the facilities),for the period for which they remain idle.

The optimum balance of costs can be made by **scheduling the flow of units** or **providing proper number of service facilities** .

The Queuing Cost Trade-off



Total expected cost = cost of waiting time + cost of providing service

KENDALL NOTATIONS

- The Kendall classification of queuing systems (1953) exists in several modifications.
- The most comprehensive classification uses 6 symbols:

A/B/c/K/m/Z

where:

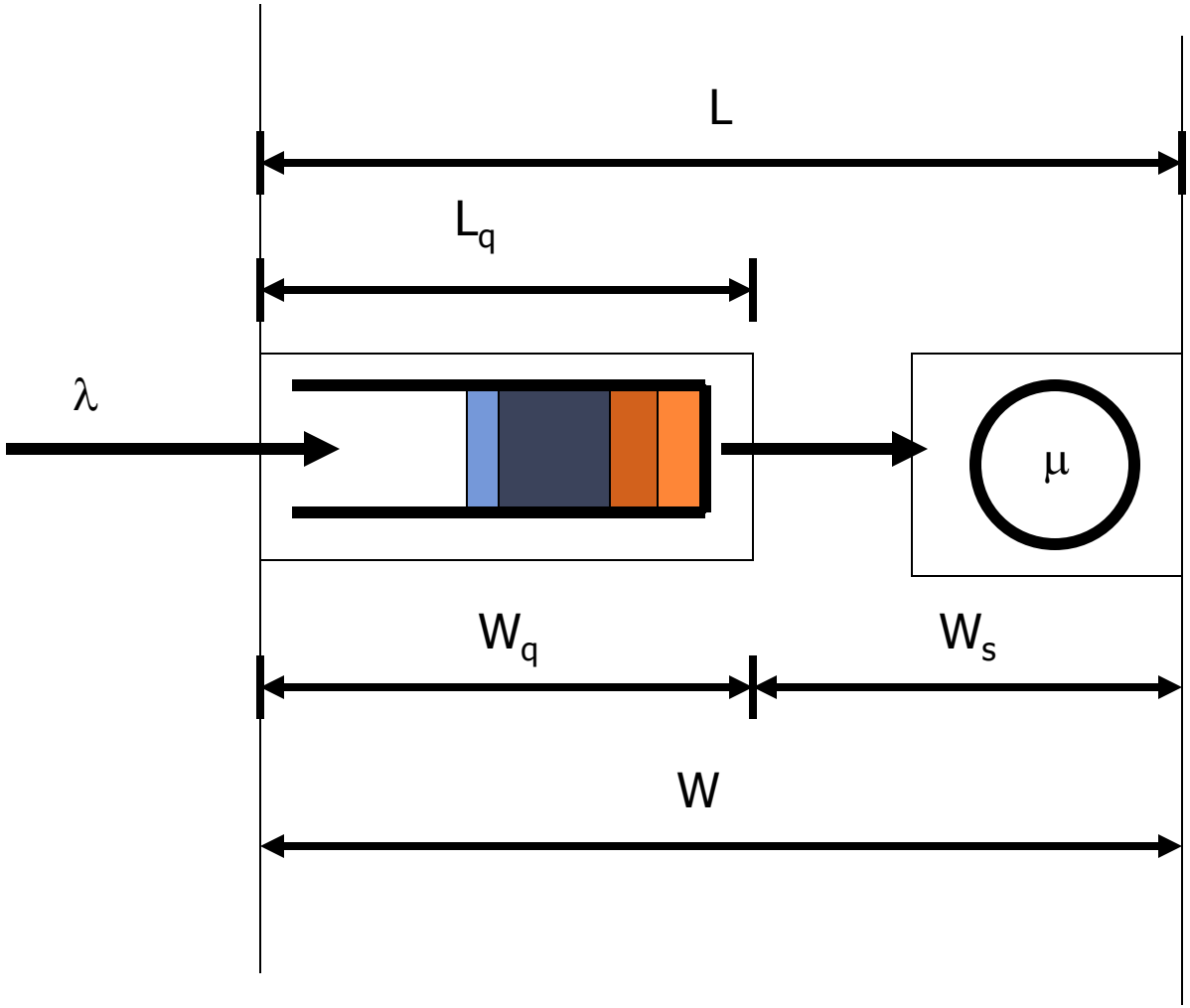
- ***A*** is the arrival pattern (distribution of intervals between arrivals).
- ***B*** is the service pattern (distribution of service duration).
- ***c*** is the number of servers (e.g., 1, 2, 3, ……..).
- ***K*** is the system capacity (e.g., 1, 2, 3, ……. ∞). Omitted for unlimited queues.
- ***m*** is the population size (number of possible customers) (e.g., 1, 2, 3, ……. ∞). Omitted for open systems.
- ***Z*** is the queuing discipline (FIFO, LIFO, ...). Omitted for FIFO or if not specified.

Shorter Notation: A/B/c

KENDALL NOTATIONS

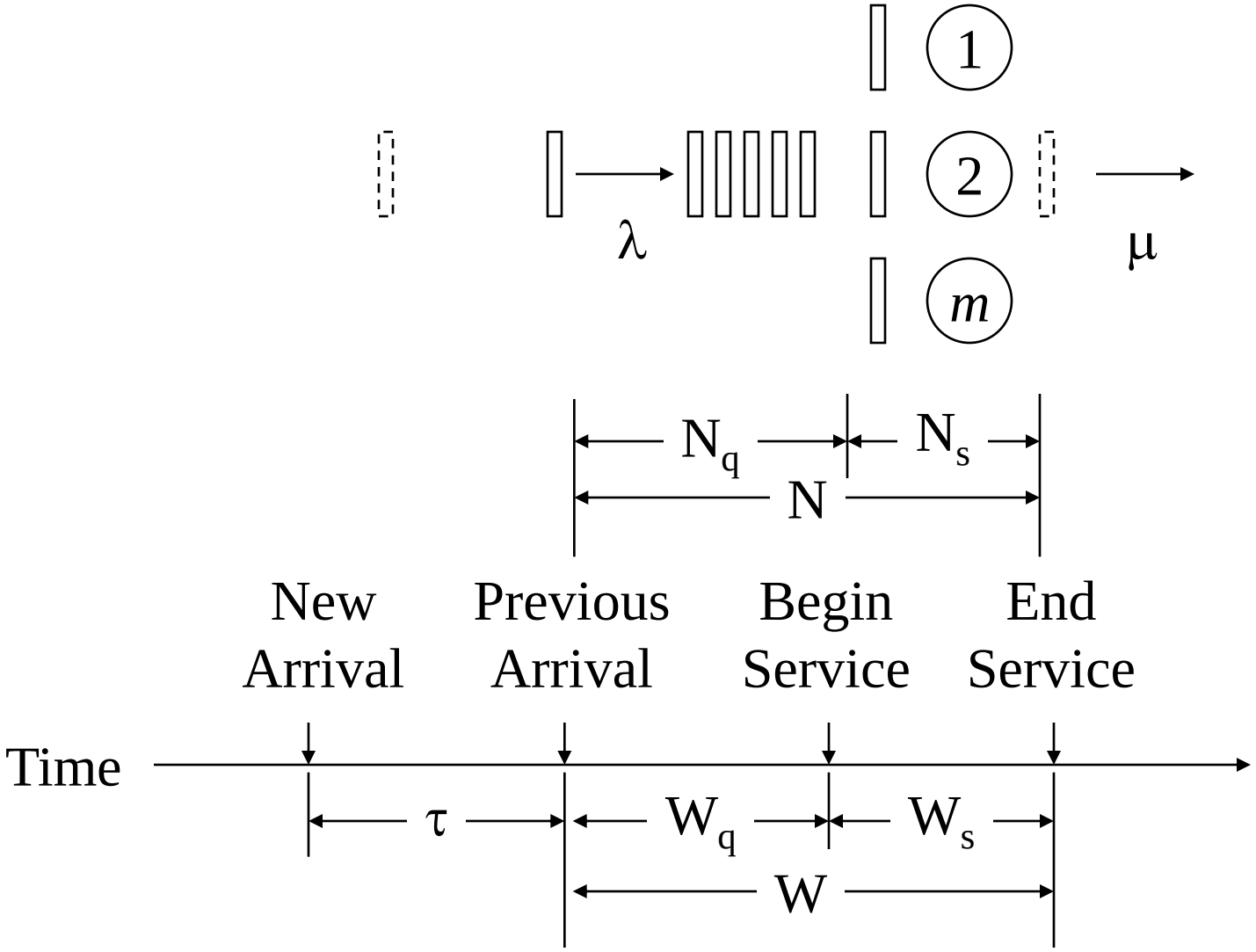
- **These symbols are used for arrival and service patterns:**
  - ***M*** is the Poisson (Markovian) process with exponential distribution of intervals or service duration respectively.
  - E_m is the Erlang distribution of intervals or service duration.
  - ***D*** is the symbol for deterministic (known) arrivals and constant service duration.
  - ***G*** is a general (any) distribution.
  - ***GI*** is a general (any) distribution with independent random values.
- **Examples:**
  - **D/M/1** = Deterministic (known) input, one exponential server, one unlimited FIFO or unspecified queue, unlimited customer population.
  - **M/G/3/20** = Poisson input, three servers with any distribution, system capacity 20 customers (3 in service and 17 in the queue), unlimited customer population.
  - **D/M/1/10/50/LIFO** = Deterministic arrivals, one exponential server, queue is a stack of the maximum size 9, total number of customers 50.

A Commonly Seen Queuing Model



The Queuing System

Queuing Model with Key Variables



Queuing Model with Key Variables

- τ = Inter-arrival time = time between two successive arrivals
- λ = Mean arrival rate = $1/E[\tau]$
  May be a function of the state of the system,
  e.g., number of jobs already in the system
- W_s = Service time per job
- μ = Mean service rate per server = $1/E[W_s]$
- Total service rate for m servers is $m\mu$
- N = Number of customers/jobs in the system at time t
- N_q = Number of jobs waiting for service
- N_s = Number of jobs receiving service
- W = Response time or the time in the system
  = time waiting + time receiving service
- W_q = Waiting time = Time between arrival and beginning of service

Queuing Model with Key Variables

- The state of the system = the number of customers in the system
- Queue length = (The state of the system) – (number of customers being served)
- $\mathbf{P_n(t)}$ =The probability that at time t, there are n customers/jobs in the system
- ρ =The utilization factor for the service facility

  (The expected fraction of the time that the service facility is being used)

Steady State Performance Measures

P_0 = Probability that there is no customer in the system.

P_n = Probability that there are "n" customers in the system (in steady state, i.e., when $t \to \infty$).

L = Average number of customers in the system.

L_q = Average number of customers in the queue.

W = Average time a customer spends in the system.

W_q = Average time a customer spends in the queue.

P_w = Probability that an arriving customer must wait for service.

ρ = Utilization rate for each server (the percentage of time that each server is busy).


Thank You!