QUEUING SYSTEM

Yetunde Folajimi, PhD

Introduction

- What is a queue?
  - A line of people or things waiting to be handled, usually in sequential order starting at the beginning or top of the line or sequence.
- Queue in computer technology:
  - Sequence of work objects that are waiting to be processed.

Queuing Theory

- The possible factors, arrangements, and processes related to queues.
- can be studied in terms of:
  - The source of each queued item,
  - how frequently items arrive on the queue,
  - how long they can or should wait,
  - whether some items should jump ahead in the queue,
  - how multiple queues might be formed and managed, the rules by which items are enqueued and dequeued

Queuing Theory in Computer Science

- the study of queues as a technique for managing processes and objects in a computer, for example in operating system design.
- The queues that a computer manages are sometimes viewed as being in stacks.
  - In programming, a queue is a data structure in which elements are removed in the same order they were entered. This is often referred to as FIFO (first in, first out). In contrast, a stack is a data structure in which elements are removed in the reverse order from which they were entered. This is referred to as LIFO (last in, first out).

Basic Terminology of Queueing Theory

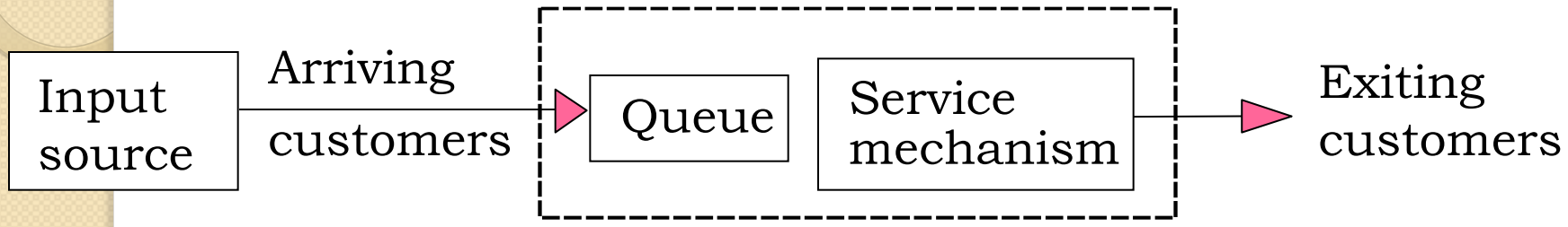
- The three main concepts in queueing theory are customers, queues, and servers (service mechanisms).
- Input Source

- The input source:
  - A population of individuals, (calling population).
  - calling population size is the number of potential customers to the system(finite or infinite).
  - Most queueing models assume that the population is infinite.

Basic Terminology (contd.)

- **QUEUE**
  - Queues can be either infinite or finite.
  - if the maximum queue size is significantly larger than the likely number of customers at any one time, then to all intents and purposes it is infinite in size.
  - The amount of time which is a customer waits in the queue for is called the queueing time.
  - The number of customers who arrive from the calling population and join the queue in a given period of time is modelled by a statistical distribution.
- **QUEUE DISCIPLINE**
  - the method by which customers are selected from the queue for processing by the service mechanisms (also called servers).
  - Most queueing models assume FCFS as the queue discipline,
- **SERVICE MECHANISM**
  - The way that customers receive service once they are selected from the front of a queue.
  - More commonly called a **SERVER**
  - **SERVICE TIME** is the amount of time which a customer takes to be serviced by the server
  - A statistical distribution is used to model the service time of a server
  - most queueing models assume that the system has either a single server or allow the number of servers to become a variable.

Structure of Single Queuing Systems



Note

1. Customers need not be people; other possibilities include parts, vehicles, machines, jobs.
2. Queue might not be a physical line; other possibilities include customers on hold, jobs waiting to be printed, planes circling airport.

Applications Of Queueing Theory

- **Traffic Flow**
  - concerned with the flow of objects around a network, avoiding congestion and trying to maintain a steady flow, in all directions.
- **Queueing on roads**
  - Queues at a motorway junction,
  - queueing in the rush hour
- **Scheduling**
  - Computer scheduling
- **Facility Design and Employee Management**
  - Queues in a bank
  - A Mail Sorting Office
- **Other Examples**
  - Design of a garage forecourt
  - Airports - runway layout, luggage collection, shops, passport control etc.
  - Hair dressers
  - Supermarkets
  - Restaurants
  - Manufacturing processes
  - Bus scheduling
  - Hospital appointment bookings
  - Printer queues
  - Minimising page faults in computing

Examples of Queuing Applications

System	Arrival Process	Service Process
Bank	Customers Arrive	Tellers serve customers
Pizza parlor	Orders are phoned in	Orders are driven to customers
Blood bank	Pints of blood arrive via donation	Patients use up pints of blood
Shipyard	Damaged ships sent to shipyard for repair	Ships are repaired & return to sea
Printers	Jobs arrive from computers	Documents are printed

Typical Performance Questions

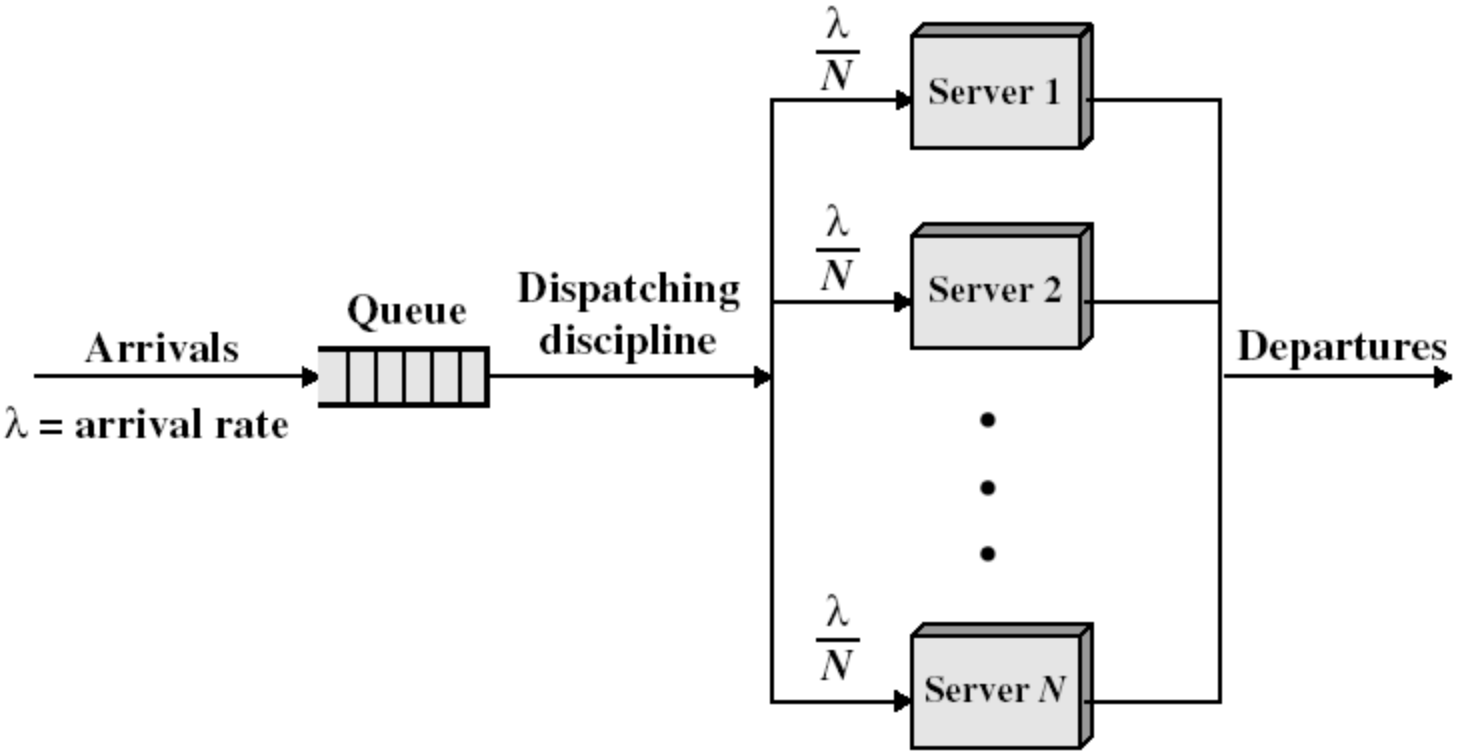
What is the …

1. average number of customers in the system?
2. average time a customer spends in the system?
3. probability a customer is rejected?
4. fraction of time a server is idle?

These questions are aimed at
characterizing complex systems.
Analyses used to support decision-making.

In queuing (and most analyses of complex stochastic systems), design takes the form of asking “what if ” questions rather than trying to optimize the design.

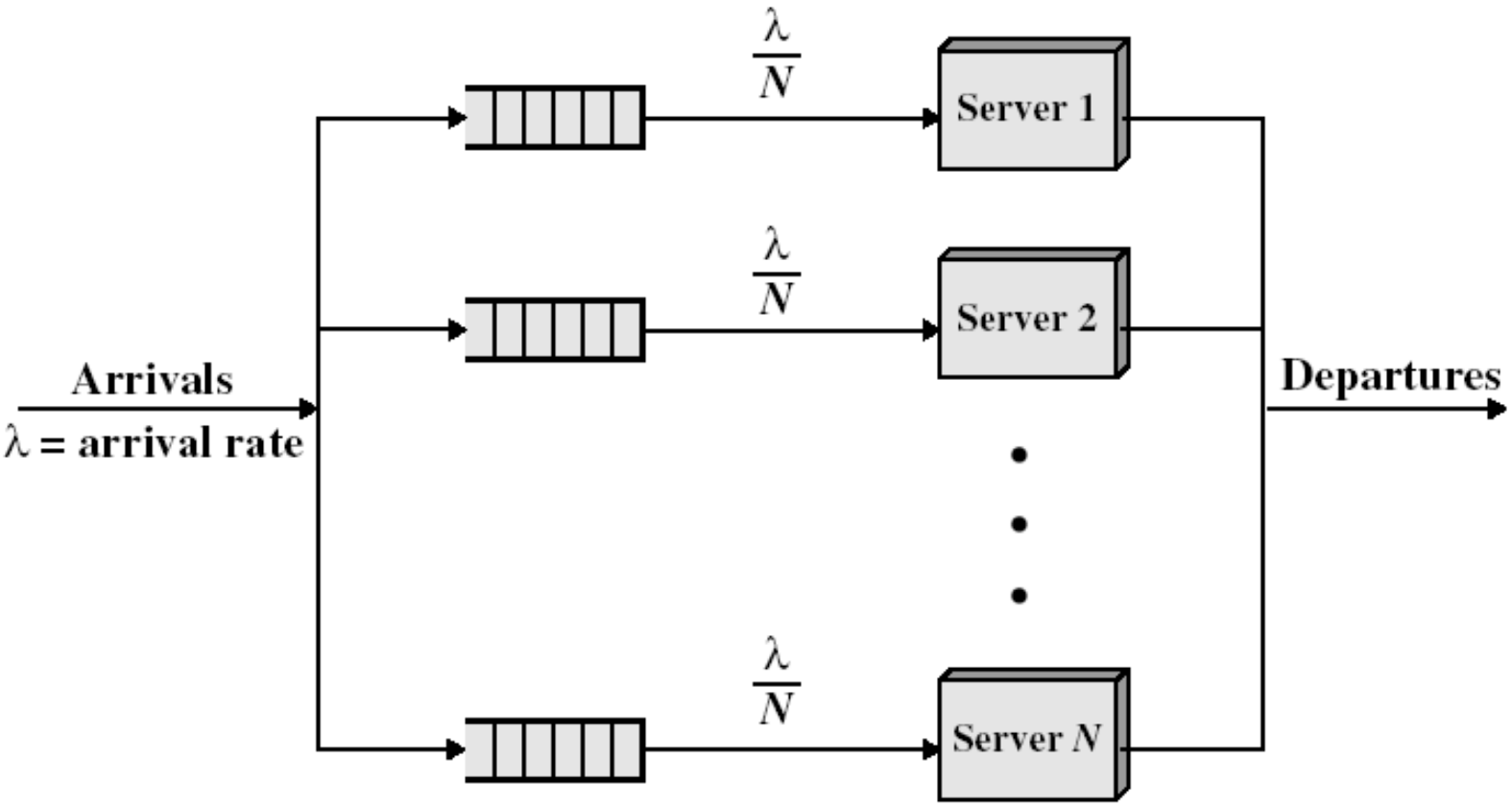
Multiple Servers, Single Queue



What is average wait in the queue?

What is average time in the system?

Multiple Servers, Multiple Queues



What is average wait in the queue?

What is average time in the system?

Elements of Queuing Systems

- **Population of Customers**
- Customers may be people, machines of various nature, computer processes, telephone calls, etc.
  - limited population (closed systems)
    - a number of processes to be run (served) by a computer
    - certain number of machines to be repaired by a service man.
  - Unlimited Population (open systems): theoretical model of systems with a large number of possible customers
    - a bank on a busy street,
    - a motorway petrol station.

Elements of Queuing Systems

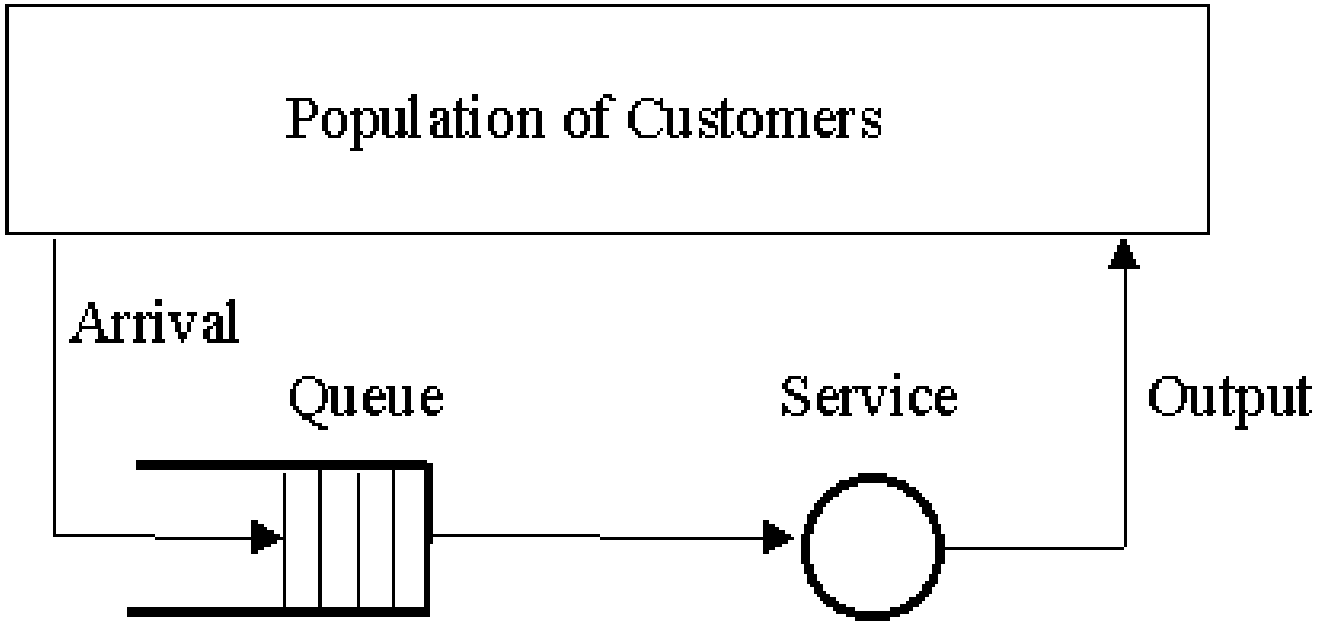


Figure 1

Elements of Queuing Systems (contd.)

- **Arrival**
  - defines the way customers enter the system.
  - Mostly the arrivals are random with random intervals between two adjacent arrivals.
  - Typically the arrival is described by a random distribution of intervals also called *Arrival Pattern*.
- **Queue**
  - represents a certain number of customers waiting for service (of course the queue may be empty).
  - Typically the customer being served is considered not to be in the queue.
  - Sometimes the queue is an abstraction (e.g planes waiting for a runway to land).
  - There are two important properties of a queue: *Maximum Size* and *Queuing Discipline*.

Elements of Queuing Systems (contd.)

- ***Maximum Queue Size*** (also called *System capacity*) is the maximum number of customers that may wait in the queue (plus the one(s) being served).
  - If the queue length is limited, some customers are forced to renounce without being served.
- ***Queuing Discipline*** represents the way the queue is organised (rules of inserting and removing customers to/from the queue).:
  - 1) FIFO (First In First Out) also called FCFS (First Come First Serve) - orderly queue.
  - 2) LIFO (Last In First Out) also called LCFS (Last Come First Serve) - stack.
  - 3) SIRO (Serve In Random Order).
  - 4) Priority Queue, that may be viewed as a number of queues for various priorities.
  - 5) Many other more complex queuing methods that typically change the customer's position in the queue according to the time spent already in the queue, expected service duration, and/or priority. These methods are typical for computer multi-access systems.

Elements of Queuing Systems (contd.)

- Queuing Discipline (Contd.)
  - Last-Come-First-Served with Preempt and Resume (LCFS-PR)
  - Round-Robin (RR) with a fixed quantum.
  - Small Quantum ⇒ Processor Sharing (PS)
  - Infinite Server: (IS) = fixed delay
  - Shortest Processing Time first (SPT)
  - Shortest Remaining Processing Time first (SRPT)
  - Shortest Expected Processing Time first (SEPT)
  - Shortest Expected Remaining Processing Time first (SERPT).
  - Biggest-In-First-Served (BIFS)
  - Loudest-Voice-First-Served (LVFS)

Elements of Queuing Systems (contd.)

- Most quantitative parameters (like **average queue length**, **average time spent in the system**) do not depend on the queuing discipline.
  - Most models either do not take the queuing discipline into account at all or assume the normal FIFO queue.
  - The two extreme values of the waiting time variance are for the FIFO queue (minimum) and the LIFO queue (maximum).
- Theoretical models (without priorities) assume only one queue.
- bank with several tellers with separate queues may be viewed as a system with one queue, because the customers always select the shortest queue.
- Systems with more queues (and more servers) where the customers may be served more times are called *Queuing Networks*.

Elements of Queuing Systems (contd.)

- **Service**
  - some activity that takes time and that the customers are waiting for.
  - Theoretical models are based on random distribution of service duration also called ***Service Pattern***.
  - Systems with one server only are called ***Single Channel Systems*, systems** with more servers are called *Multi* ***Channel Systems*.**
- **Output**
- The way customers leave the system.
- mostly ignored by theoretical models,
- sometimes the customers leaving the server enter the queue again ("round robin" time-sharing systems).

balking, reneging and jockeying

- **Balking**
  - leaving the system without joining the queue.
    - **Unforced balking:** Not joining the queue as a matter of self-will.
    - **Unforced balking:** Not joining the queue because the system doesn't permit
- **Reneging**
  - Quitting the queue after joining
- **Jockeying**
  - Shifting from one line to another