Queueing Theory

Lecture 3

Communication Theory III

Eng. (Mrs.) PN Karunanayake

Introduction

- Queueing theory sets out to take account of the phenomena of **waiting** and **congestion** that are linked to the random and unforeseeable nature of the events encountered.
- The importance of this theory will easily be understood from the viewpoint of the essential aim: **to assess the performances of systems and networks.**
- **Limited resources** and **unpredictable demands** imply a conflict for the use of the resource and hence **queues of waiting customers**.

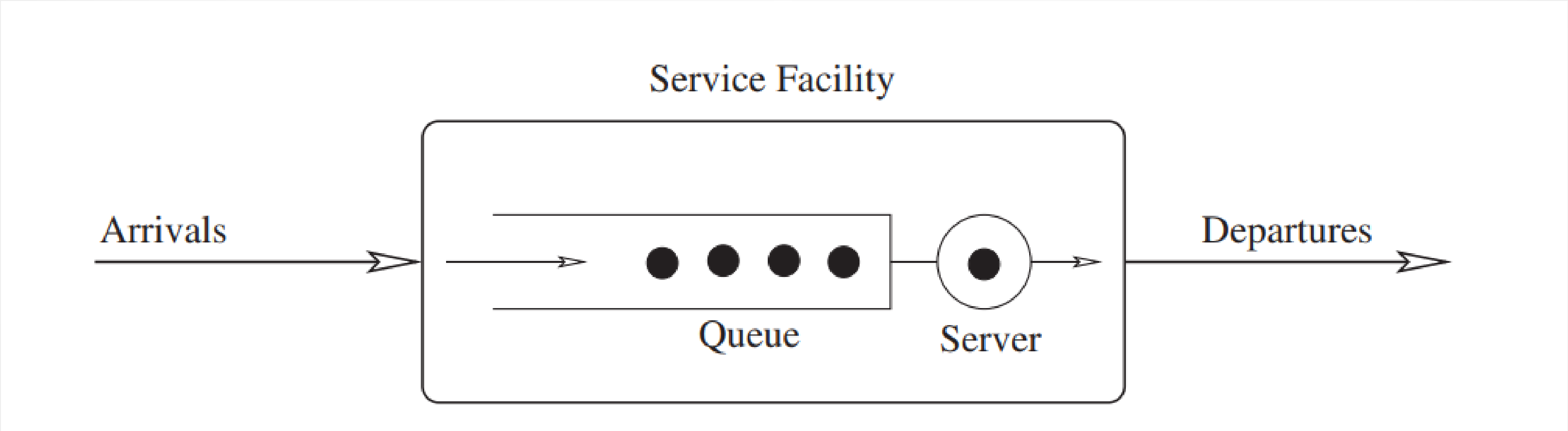
Analyzing Queues! Why?

- Modeling and analyzing systems of queues helps to minimize their inconveniences and maximize the use of the limited resources.

- An analysis will provide information on **expected time that a resource will be in use**, or the **expected time that a customer must wait**.

- This information may then be used to make decisions as to when and how to upgrade the system.

Example



Assumptions of Queuing Model

- If the server is free, (not already serving a customer) an **arriving customer goes immediately into service.** No time at all is spent in the queue.

- If the server is not free (busy), then the customer joins a queue of waiting customers and stays in the queue until entering service.

- When the server becomes free, a customer is chosen from the queue **according to a scheduling policy** and immediately enters into service.

Assumptions

- **The time between the departure of one customer and the start of service of the next customer is zero.**

- **Customers remain in the facility until their service is completed and then they depart.**

- **Customers do not become impatient and leave before they receive service.**

The basic service station: clients and servers

- **The clients** are entities that move in a network of servers, where they receive processing.
- When several clients try to obtain a service simultaneously, some of them must **wait in queues, or buffers.**

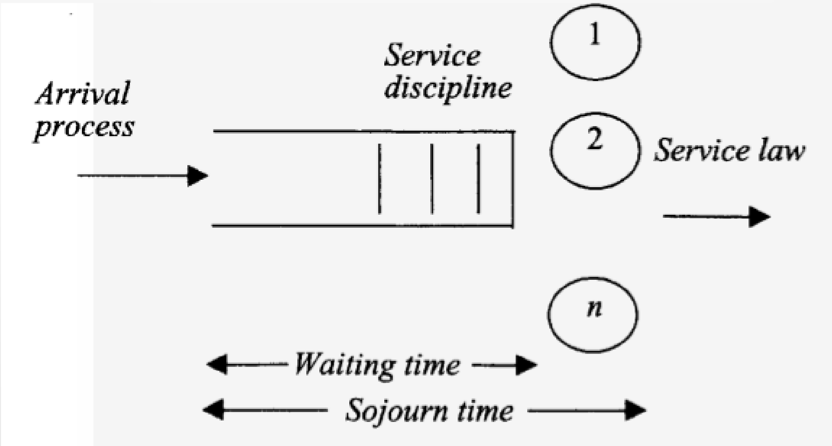
  Example:

  1. Clients may be people queueing up at a ticket office to obtain a ticket.

  2. May represent packets that are forwarded to a transmission line and awaiting the availability of the line.

The basic service station: clients and servers

- **The server** is the transmission link.
- **Service time** is the transmission time.
- **The client** can consists of requests for circuits, or bandwidth, in a telephone system.
- The most general service station consists of **a queue,** of finite or non-finite capacity, emptied by one or more servers, and supplied by a flow of clients.



The basic service station

To describe the system , it must be possible to specify:

- The mechanism according to which **clients arrive** in the buffer (probability law of the arrival process).
- **Service time** (the probability distribution of its duration)
- **Service discipline** (when a server becomes idle, which client does it choose?).

Arrival process

To describe the phenomenon of the arrival law, following can be used.

1. The time interval between successive arrivals (the interarrival time 1/ λ)

2. The number of arrivals in a given time interval (the arrival rate λ)

During the time T, n(T) arrivals occur. The flow intensity is then expressed as a number, the arrival rate, whose intuitive definition is

$$\lambda = \lim_{T \to \infty} \frac{n(T)}{T}$$

Arrival process

$$A(t) = \text{Prob}\{\text{time between arrivals } \le t\}$$

$$\frac{1}{\lambda} = \int_0^\infty t \, dA(t),$$

- Where $dA(t)$ is the probability that the interarrival time is between t and $t + dt$.

- Assume that these interarrival times are independent and identically distributed, which means that only $A(t)$ is of significance.

- An arrival pattern that does not change with time is said to be a ***homogeneous* arrival process**. If it is invariant to shifts in the time origin, it is said to be a ***stationary* arrival process.**

Arrival process

- The length of a queue depends on the complete probabilistic description of the arrival and service processes.

What will happen if the **average service demands** of arriving customers are **greater than** the **system service capacity** ?

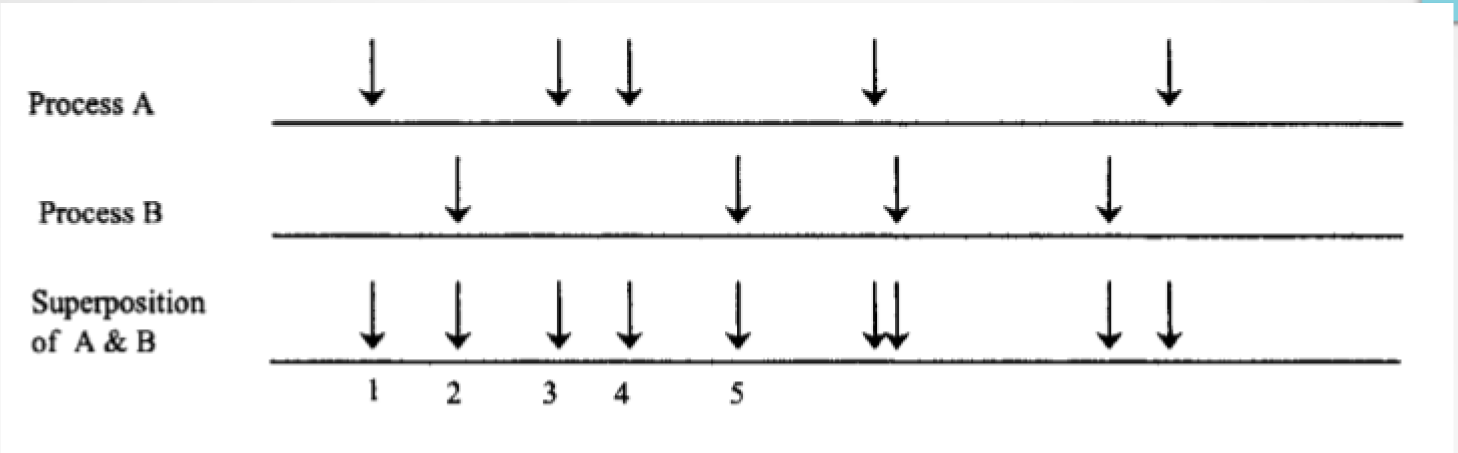
What will happen if the **average arrival rate is less** than the **system service capacity** ?

What will happen if the **space available** to hold customers waiting for service is **limited**?

Renewal process

- A renewal process is an idealized stochastic model for events that occur randomly in time.
- Assume that the situation is described as follows:

Each time an arrival takes place, the interval until the next arrival is drawn according to a given law, in such a way that **successive intervals are independent of each other**.



Superimposing processes A and B does not result in a renewal process:

- The arrival of "4 " is related to "3 ", but "5 " is related to "2 ".
- To predict the arrival of the 4th client, reference must be made to the 3rd, by using the inter-arrival time.
- But the arrival of the 5th client can only be predicted by reference to the 2nd client, and not to the 4t

Poisson process - Conditions

1. $N(0) = 0$

2. Events that occur in nonoverlapping time intervals are mutually independent

3. The number of events that occur in any interval depends only on the length of the interval and not on the past history of the system.

Poisson process

Let us assume that the arrival process complies with the following rules:

1. The probability of an arrival in an interval [t , t + Δt] does not depend on what happened before the instant t. This is the so-called **memoryless property**.

2. The probability of the arrival of a client is proportional to Δt, and the **probability of more than one event is "negligible"** (he proportionality factor is rated λ (process intensity).

Poisson process

$P_k(t)$ is the probability of k arrivals in the interval [0, t].

Description:

- k clients have been observed in [0,t], and no arrivals have been observed in $[t, t + \Delta t]$.
- (k - 1) clients have been observed in [0,t] and one arrival occurred in $[t, t + \Delta t]$.
- (k - n), n > 1 arrivals have been observed in [0, t] and n arrivals in $[t, t + \Delta t]$.

Poisson process

The probability of observing k arrivals in an interval of length t:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

$$P(X = k) = \frac{(A)^k}{k!} e^{-A}$$

A is then the mean traffic offered during the period considered.

Poisson process

- The distribution function of the probability law for the interval for two successive arrivals is derived from the distribution:

The probability of an interval between arrivals greater than t is the probability that no arrival occurs between 0 and t.

$$A(t) = 1 - e^{-\lambda t}$$

The average number of arrivals observed in any interval of length t is:

$$m = \lambda t$$

Variance:

$$\sigma^2 = \lambda t$$

Service process

Assume that each service time is independent of the others, and that all of them comply with the same distribution function.

$$B(x) = P\{\text{Service time} \leq x\}$$

- Need to deal with systems handling clients whose service laws obey different distributions.
- This is the case for example with multi-service systems, in which each type of service has its own characteristic.

Service process

The cumulative distribution function for an exponential random variable, X, with parameter $\lambda > 0$, is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

Corresponding probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Service process

Mean $$E[X] = \frac{1}{\lambda}$$

Second moment $$E[X^2] = \frac{2}{\lambda^2}$$

Variance $$\text{Var}\,[X] = \frac{1}{\lambda^2}$$

Example

1. The arrival of jobs to a supercomputing center follows a Poisson distribution with a mean interarrival time of 15 minutes.

(a) Calculate the rate of arrivals.
(b) Probability of time between arrivals $\leq \tau$ hours
(c) Probability of k arrivals in τ hours

Exponential Distribution Function

In the previous example, suppose 45 minutes have passed without an arrival. Then the expected time until the next arrival is still just 15 minutes.

Since the arrival process is Poisson, we have,

$$p_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \ldots,$$

Exponential Distribution Function

Let *X* be the random variable that denotes the time between successive events (arrivals). Its probability distribution function is given by,

$$A(t) = \text{Prob}\{X \le t\} = 1 - \text{Prob}\{X > t\}.$$

But $\text{Prob}\{X > t\} = \text{Prob}\{0 \text{ arrivals in } (0, t]\} = p_0(t)$. Thus

$$A(t) = 1 - p_0(t) = 1 - e^{-\lambda t}, \quad t \ge 0,$$

by differentiation, corresponding density function

$$a(x) = \lambda e^{-\lambda t}, \quad t \ge 0.$$

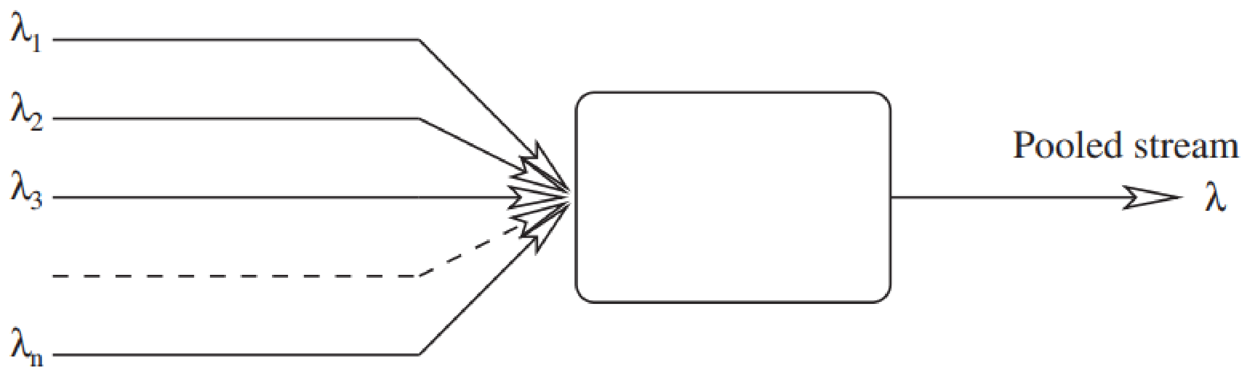
Superposition/Decomposition of Poisson Streams

- When two or more *independent* Poisson streams merge, the resulting stream is also a Poisson stream.

- Multiple arrival processes to a single service center can be merged to constitute a single arrival process whose interarrival times are exponentially distributed, so long as the interarrivals times of the individual arrival processes are exponentially distributed.

Superposition/Decomposition of Poisson Streams

$$\lambda = \sum_{i=1}^{n} \lambda_i .$$



Superposition/Decomposition of Poisson Streams