Performance Measures of Queueing Systems

Lecture 4

Communication Theory III

Eng. (Mrs.) PN Karunanayake

Introduction

- To characterize a queueing system we have to identify the probabilistic properties of the incoming **flow of requests**, **service times** and **service disciplines.**
- The arrival process can be characterized by the distribution of the **interarrival times of the customers**, denoted by A(t),

$$A(t) = P(Interarrivaltime < t)$$

- The interarrival times are usually assumed to be **independent** and **identically distributed** random variables.
- The distribution function of the random variable **service time**, also known as **service request**, is denoted by B(x),

$$B(x) = P(servicetime < x)$$

- The **structure of service and service discipline** tell us the **number of servers**, the **capacity of the syste.**
- The **service discipline** determines the rule according to the next customer is selected. The most commonly used laws are.

1. FIFO - First In First Out: who comes earlier leaves earlier, FCFS - First Come First Served

2. LIFO - Last Come First Out: who comes later leaves earlier, LCFS - Last Come First Served

3. RS - Random Service: the customer is selected randomly, SIRO - Service In Random Order

4. Priority without Preemption or Head of Line (HOL), Priority with Preemption / Resume or Repeat

5. PS - Processor Sharing

Summary

The **aim of all investigations in queueing theory** is to get the **main performance measures of the system** which are the probabilistic properties such as **distribution function**, **density function**, **mean**, **variance** of the following random variables:

- Number of customers in the system
- Number of waiting customers
- Utilization of the server/s
- Response time of a customer
- Waiting time of a customer
- Idle time of the server
- Busy time of a server

- Concerning the distribution of **interarrival times**, **service times**, **number of servers**, **capacity** and **service discipline**.

Kendall's Notation

- To describe a queueing systems, ***Kendall*** has introduced a notation.

A / B / m / K / n / D

where

A : distribution function of the interarrival times

B : distribution function of the service times

m : number of servers,

K : capacity of the system, the maximum number of customers in the system including the one being serviced

n : population size, number of sources of customers

D : service discipline.

Kendall's Notation

- Exponentially distributed random variables are denotated by M , meaning Markovian or memoryless.
- If the **population size** and the **capacity** is **infinite**, the **service discipline is FIFO**, then they are omitted.
- **M/M/1** denotes a system with **Poisson arrivals**, **exponentially distributed service times** and a **single server**.
- **M/G/m** denotes an **m - server system** with **Poisson arrivals** and **generally distributed** service times.

Kendall's Notation

- **M/M/r/K/n** stands for a system where the customers **arrive from a finite-source** with **n elements** where

  - They stay for an **exponentially distributed time**

  - The service times are exponentially distributed

  - The **service is carried out according to the request's arrival by r severs**

  - The **system capacity is K.**

Example

Consider as an example the case of travelers who arrive at a train station and purchase tickets at **one of six ticket distribution machines**. It is reasonable to assume that the **time taken to purchase a ticket is constant** whereas it may be observed that **arrivals follow a Poisson process.**

Number of Customers

Let *N* be the random variable that describes the number of customers in the system at steady state. The probability that at steady state the number of customers present in the system is *n* is denoted by p_n

$$p_n = \text{Prob}\{N = n\}$$

The average number in the system at steady state is

$$L = E[N] = \sum_{i=0}^{\infty} np_n.$$

Number of Customers

Let *Nq* be the random variable that describes the number of customers waiting in the queue and we shall denote its mean by *Lq*

$$L_q = E[N_q].$$

System Time and Queueing Time

The **time that a customer spends in the system**, from the instant of its arrival to the queue to the instant of its departure from the server, is called the ***response time*** or *sojourn time.*

The response time E[R] = The waiting time in the queue (Wq)

+

the service time

System Utilization

In a queueing system with a single server ($c = 1$), **the utilization *U* is defined as the fraction of time that the server is busy.**

If the rate at which customers arrive at the queueing facility is λ and if μ is the rate at which these customers are served,

Utilization $\rho = \lambda / \mu$

For system to be stable (the queue does not grow without bound),

$$\lambda/\mu < 1$$

any time interval, the average number of customers that arrive must be strictly less than the average number of customers that the server can handle.

System Utilization

For a queueing systems with multiple servers ($c > 1$), the utilization is defined as the average fraction of servers that are active.

If the rate at which customers arrive at the queueing facility is λ and if μ is the rate at which these customers are served,

Utilization $\rho = \lambda / C\mu$

System Throughput

The average number of customers that are processed per unit time (denoted by X).

Traffic Intensity

The rate at which work enters the system.

In a queueing system in which all customers that arrive are eventually served and leave the system, the throughput is equal to the arrival rate, λ.

- This is not the case in queueing systems with finite buffer, since arrivals may be lost before receiving service.

$$\lambda \bar{x} = \lambda / \mu$$

Traffic Intensity

$$\lambda \bar{x} = \lambda / \mu$$

- In single-server systems the traffic intensity is equal to the utilization.
- For multiple servers, the traffic intensity is equal to *cU*

Little‘s Law

The number of customers in a system is equal to the product of the effective arrival rate and the time spend in the system.

$$L = \lambda W$$

L – Mean number of customers in the system

λ - the arrival rate

W – Response time

Little‘s Law

Can be applied to the different parts of the system.

$$L_q = \lambda W_q$$

L_q – Mean number of customers in the queue

λ - the arrival rate

W – average time spent in the queue

$$L_s = \lambda \bar{x}$$

L_s – Mean number of customers receiving the service

λ - the arrival rate

W – average time spent receiving the service

$$L = L_q + L_s = \lambda W_q + \lambda \bar{x} = \lambda (W_q + \bar{x}) = \lambda W$$

Little‘s Law

Is independent of following:

- Specific assumptions regarding the arrival distribution A(t).
- Specific assumptions regarding the service time distribution B(t).
- The number of servers.
- The particular queueing discipline.

Example

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

1. Prob{Time between arrivals ≤ τ hours}

- 2. Prob{k arrivals in τ hours}

suppose 45 minutes have passed without an arrival