

A Study on Queuing Model

Dr. S. Chandrasekaran¹, P. Velusamy²

¹KhadirMohideen College, Adirampattinam

²Research Scholar, Department of Mathematics, KhadirMohideen College, Adirampattinam

Abstract: *Queuing theory is the mathematical study of waiting lines and it is very useful to define Modern information technologies require innovations that are based on modelling, analyzing, designing to deals as well as the procedure of traffic control of daily life of human like telecommunications, reservation counter, super market, big bazaar, Picture Cinema hall ticket window and also to determining the sequence of computer operations, computer performance, health services, airport traffic, airline ticket sales. In the field of computer Parallel System and Distributed system are also have the base of Queue models. In this paper we are discussing the approach of Queueing theory and queueing model.*

Keywords: Mathematics Study, Queueing System, traffic control, airline ticket

1. Introduction to Queueing Theory

Queueing theory introduces by A.K. Erlang a Danish mathematician who studied telephone traffic congestion problems in the first decade of the 20th century. Queueing theory very useful in many practical applications in areas such as, e.g., telephone exchange, traffic control, manufacture systems, inventory systems and communication systems, telephone exchange, supermarket, at a petrol station, at computer systems, etc. Queueing theory is a set of mathematical tools for the analysis of probabilistic systems of customers and servers. Queueing theory, also known as the theory of overcrowding, is the branch of operational research that explores the relationship between demand on a service system and the delays suffered by the users of that system.

Queueing theory is generally considered a branch of operations research because the results are often used when making business decisions about the resources needed to provide service. There are many valuable applications of the theory, most of which have been well documented in the literature of probability, operations research, management

science, and industrial engineering. Some examples are traffic flow (vehicles, aircraft, people, communications), scheduling (patients in hospitals, jobs on machines, programs on a computer), and facility design (banks, post offices, amusement parks, fast-food restaurants).

1.1 Basic elements of Queue

The analysis of queue is based on building a mathematical model representing the process of arrival of Item who joins the queue, the rules by which they are allowed into service, and the time it takes to service. Queueing theory embodies the full scope of such models cover all perceivable systems which incorporate characteristics of a queue. We identify the unit demanding service, whether it is human or otherwise.

Queue: queue is a file or line of persons. „Queue“ means to form a line while waiting for something or a waiting line, involves arriving items that wait to be served at the facility that provides the service they seek.

1.2 Basic System of Queueing Process:

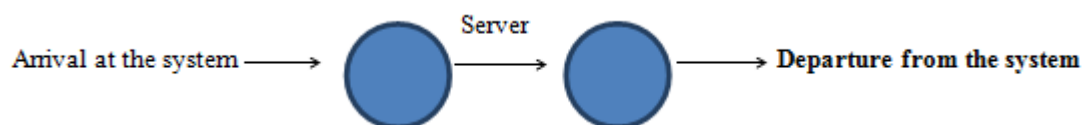


Figure 1: Basic Queue Process

Basic Queue Process A Basic queueing system is formed from three general elements (Figure 1)

- 1) The arrival process of users in the system;
- 2) The order in which users obtain access to the service facility, once they join the queue.
- 3) The service process and departure from the system.
 - Arrival refers to the average number of customers who require service within a specific period of time.
 - Customers can be people, work-in-process inventory, raw materials, incoming digital messages, or any other entities that can be modeled who are to wait for some process to take place it may be infinite or finite also is said size of queue..

- A Server can be a human worker, a machine, or any other entity that can be Processor as executing some process for waiting customers.

Queue Discipline refers to the priority system by which the next customer to receive service is selected from a set of waiting customers. One common queue discipline is first-in-first-out, or FIFO.

Service Rate (or Service Capacity) refers to the overall average number of customers a system can handle in a given time period.

Stochastic Processes are systems of events in which the times between events are random variables. In queueing models, the patterns of customer arrivals and service are modeled as stochastic processes based on probability distributions.

Utilization refers to the proportion of time that a server (or system of servers) is busy handling customers.

Scheduling adopt by server: refer scheduling how to work server. Server follows the following discipline. The server in accepting customers for service. In this context, the rules such as “first-come, first-served” (FCFS), “last-come, first-served” (LCFS), and random selection for service” (RS) are self-explanatory. Others such as “round robin” and “shortest processing time” may need some elaboration. In many situations customers in some classes get priority in service over others. There are many other queue disciplines which have been introduced for the efficient operation of computers and communication systems. Also, there are other factors of customer behavior such as balking, reneging, and jockeying that require consideration as well.

The Notation use for Scheduling discipline is

FCFS: first-come, first-served

LCFS: last-come, first-served

RS: Random selection for Service

RRS: Round Robin Faison

SPT: shortest processing time

2. Queueing Systems

A queueing network is Networks of queues are systems which contain an arbitrary, but finite, number of queues. Customers, sometimes of different queue travel through the network and are served at the node. The user sources for some of the queueing systems in the network may be other queueing systems in the same network (Figure 2).

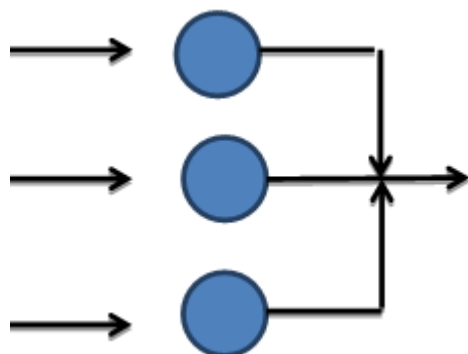


Figure 2: Queueing Systems

Queueing networks, networks of service facilities where customers must receive service at some of or all these facilities. It is therefore necessary to study the entire network to obtain such information as the expected total waiting time, expected number of customers in the entire system, and so forth. Because of the importance of queueing networks, this is wide are of network of parallel and distributed computing, research into this area has been very active.

To describe a queueing network, further information must be provided on how the queueing systems are interconnected, how they interact and how users are assigned to the queueing systems.

Busy period service delays must occur in case of the services that respond to unpredictable demands whose time and location of occurrence are governed by probabilistic laws. The cost of providing sufficient capacity to avoid all delays under all circumstances would be impossible. The role of the analysis is to design service systems that achieve an acceptable balance between system operating costs and the delays suffered by the users of that system.

2.1 Queue Model

- 1) Queueing Model are used to estimate desired performance measures of the system
- 2) Provide rough estimate of a performance measure
- 3) Typical measures
 - Server utilization
 - Length of waiting lines
 - Delays of customers
- 4) Applications
 - Determine the minimum number of servers needed at a service center
 - Detection of performance bottleneck or congestion
 - Evaluate alternative system designs

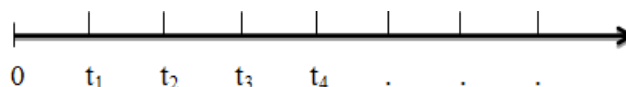
2.2 Kendall Notation

A/S/m/B/K/SD

- A: arrival process
- S: service time distribution
- m: number of servers
- B: number of buffers(system capacity)
- K: population size
- SD: service discipline

2.3 Arrival Process

- 1) Jobs/customer arrival pattern
- 2) τ form a sequence of Independent and Identically Distributed(IID) random variables
 - Arrival times : t_1, t_2, \dots, t_j
 - Interarrivaltimes : $T_j = t_j - t_{j-1}$
- 3) Arrival models
 - Exponential + IID (Poisson)
 - Erlang
 - Hyper-exponential
 - General : results valid for all distributions

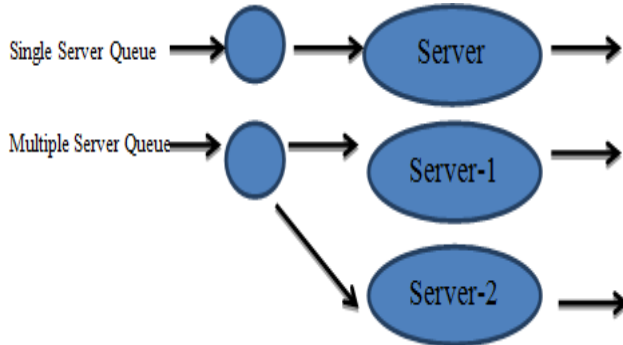


2.4 Service Time Distribution

- 1) Time each user spends at the terminal
- 2) IID
 - Exponential
- 3) Distribution model

- Erlang
 - Hyper-exponential
 - General
- 4) cf.
- Jobs = customers
 - Device = service center = queue
 - Buffer = waiting position

2.5 Number of Servers



2.6 Service Disciplines

FCFS: first-come, first-served
LCFS: last-come, first-served
RS: Random selection for Service
RRS: Round Robin Faison
SPT: shortest processing time

2.7 Common Distributions

- M : Exponential
- Ek :Erlang with parameter k
- Hk :Hyperexponential with parameter k(mixture of k exponentials)
- D : Deterministic(constant)
- G : General(all)

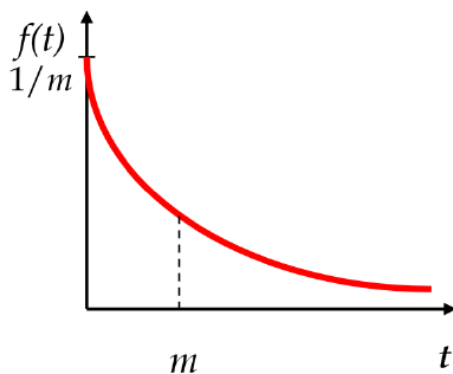


Figure 5 (a): Probability density function for the exponential distribution

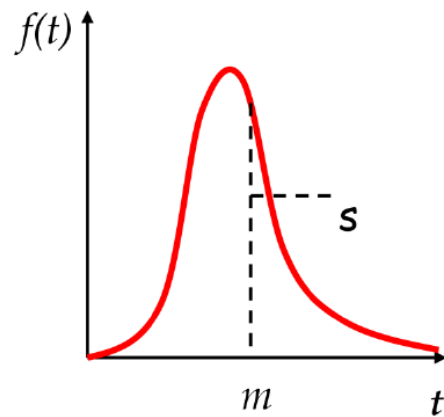


Figure 5(b): Probability density function for the exponential distribution General

Example

- M/M/2/10/1000/FCFS
- Time between successive arrivals is exponentially distributed
- Service times are exponentially distributed
- Three servers
- 10 buffers = 2service + 10 waiting
- After 10, all arriving jobs are lost
- Total of 1000 jobs that can be serviced
- Service discipline is first-come-first-served

Default Input Source means Calling Process Size of buffer

- Infinite buffer capacity
- Infinite population size
- FCFS service discipline
- Example

Key Variables used in following process

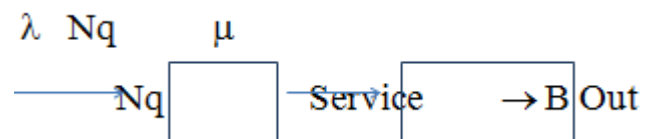
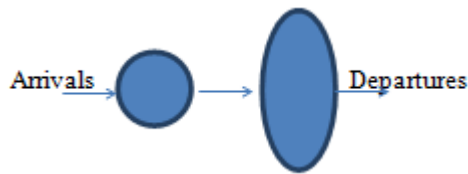


Figure 6

- T : interarrival time
- λ : mean arrival rate = $1/E[T]$
- s : service time per job
- μ : mean service rate per server = $1/E[s]$
- n : number of jobs in the system(queue length) = $nq + ns$
- nq : number of jobs waiting
- ns : number of jobs receiving service
- r : response time – time waiting + time receiving service
- w : waiting time – Time between arrival and beginning of service

Little's Law

- Waiting facility of a service center
- Mean number in the queue = arrival rate X mean waiting time
- Mean number in service = arrival rate X mean service time



Example

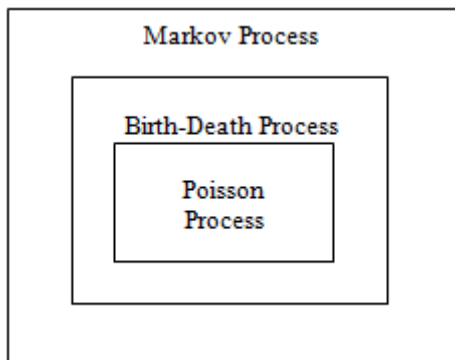
A monitor on a disk server showed that the average time to satisfy an I/O request was 100msecs. The I/O rate was about 100 requests per second. What was the mean number of request at the disk server?

– Mean number in the disk server
 = arrival rate X response time
 = (100 request/sec) X (0.1 seconds)
 = 10 requests

– Stochastic Processes: systems of events in which the times between events are random variables. In queueing models, the patterns of customer arrivals and service are modeled as stochastic processes based on probability distributions i.e process with random events that can be described by a probability distribution function

- Process : function of time
- A queuing system is characterized by three elements:
 - A stochastic input process
 - A stochastic service mechanism or process
 - A queuing discipline

Types of Stochastic Process



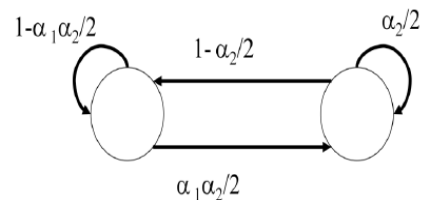
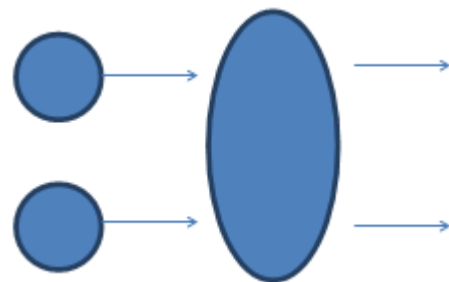
Discrete/Continuous State Processes

- Discrete = finite or countable
- Discrete state process
 - Number of jobs in a system $n(t) = 0, 1, 2, \dots$
- Continuous state process
 - Waiting time $w(t)$
- Stochastic chain : discrete state stochastic process

Markov Processes

- Future states are independent of the past
- Markov chain : discrete state Markov process
- Not necessary to know how long the process has been in the current state – State time : memory less(exponential) distribution
- M/M/m queues can be modeled using Markov processes
- The time spent by a job in such a queue is a Markov process and the number of jobs in the queue is a Markov chain
- The transition probability matrix

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \dots \\ P_{10} & P_{11} & P_{12} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$



Birth-Death Processes

The foundation of many of the most commonly used queueing models

- 1) Birth – equivalent to the arrival of a customer or job
- 2) Death – equivalent to the departure of a served customer or job

- The discrete space Markov processes in which the transitions are restricted to neighboring states
- Process in state n can change only to state $n+1$ or $n-1$

Example

The number of jobs in a queue with a single server and individual arrivals(not bulk arrivals)

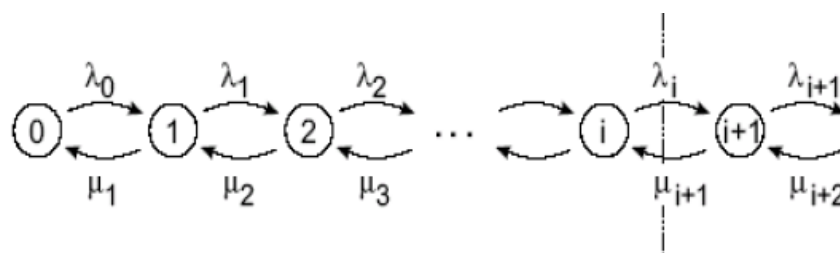


Figure 11

Poisson Processes

Interarrival time s = Independent and Identically Distributed (IID) + exponential

- Birth death process that $\lambda_k = \lambda$, $\mu_k = 0$ for all k

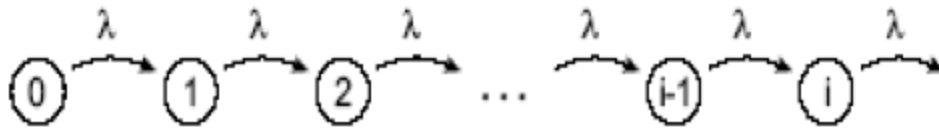


Figure 12

- Probability of seeing n arrivals in a period from 0 to t

$$\frac{(\lambda t)^n}{n!} \exp(-\lambda t)$$

t : interval 0 to t

n : total number of arrivals in the interval 0 to t

λ : total average arrival rate in arrivals/sec

- Pdf of inter arrival time

$$p(\tau_n) = \lambda e^{-\lambda \tau_n}$$

M/M/1 Queue (Single Queue Model)

- The most commonly used type of queue.
- Used to model single processor systems or individual devices in a computer system.
- Assumption
 - Interarrival rate of $\lambda \rightarrow$ exponentially distributed
 - Service rate of $\mu \rightarrow$ exponentially distributed
 - Single server
 - FCFS
 - Unlimited queue lengths allowed

- Infinite number of customers

- Need to know only the mean arrival rate(λ) and the mean service rate μ

- State = number of jobs in the system

M/M/1 Operating Characteristics

- Utilization(fraction of time server is busy)

$$\rho = \lambda / \mu$$

- Average waiting times

$$W = 1 / (\mu - \lambda)$$

$$W_q = \rho / (\mu - \lambda) = \rho W$$

- Average number waiting

$$L = \lambda / (\mu - \lambda)$$

$$L_q = \rho \lambda / (\mu - \lambda) = \rho L$$

Flexibility/Utilization Trade-off

Must trade off benefits of high utilization levels with benefits of flexibility and service

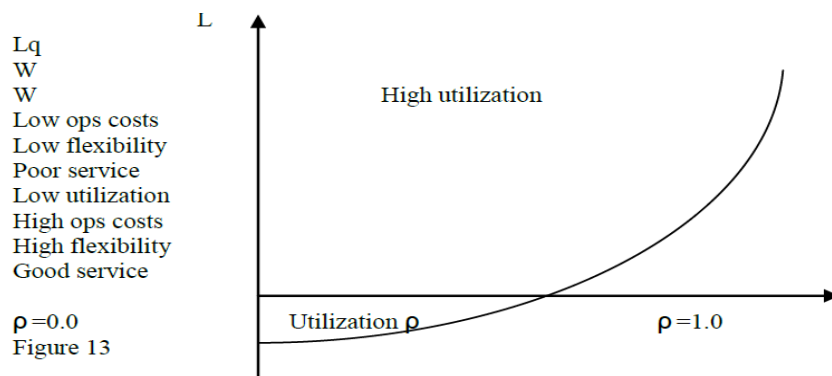


Figure 13

Cost Trade-offs

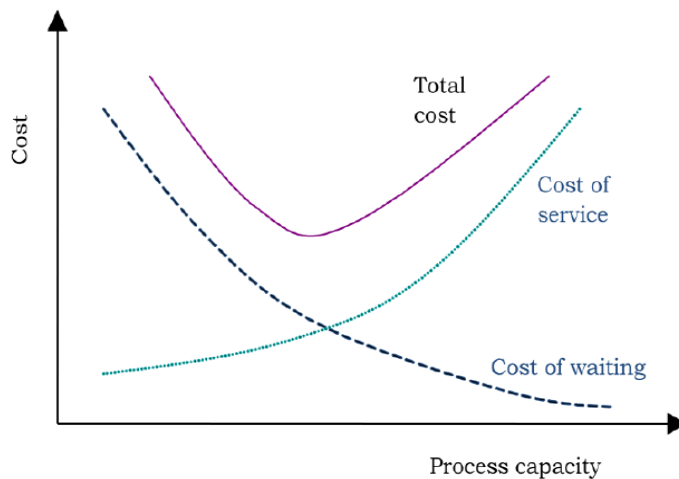


Figure 14

