International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 2 Issue 2 ǁ February. 2013 ǁ PP.01-11

QUEUEING THEORY APPR OACH WITH QUEUEING MODEL: A STUDY

Ajay Kumar Sharma[1], Dr. Rajiv Kumar[2] ,Dr. Girish Kumar Sharma[3]

[1] Research Scholar[1], Deptt.of Mathematics, DCRUST,Murthal ,Sonepat, India
[2]Professor Department of Mathematics, DCRUST,Murthal,Sonepat,India
[3] Associate Professor, Department of Computer Application, ,BPIBS,Delhi,India

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: *Queueing theory, Queueing system, Queueing network, Queueing model,.*

I. 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.ls and. 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).

II. BACKGROUND OF LITERATURE REVIEW

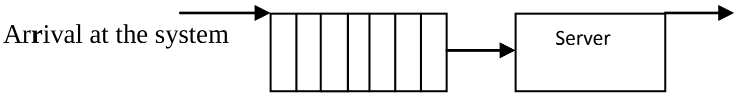
Queues (waiting lines) are a part of everyday life. Every human being wait in queues to buy a ticket of railway ticket, make a bank deposit in the bank counter, start a ride in an paly ground park, etc. We have become familiar to huge amounts of waiting, but still get upset by unusually long waits. The amount of time that a nation's of common people wastes by waiting in queues is a major factor in both the value of life there and the competence of the nation's economy. For example, India has a large number of before people in the country. So most of the people face the problem of any kind of the services in their daily life.. Even in the United States today, it has been estimated that Americans spend Approximately 37,000,000,000 hours per year waiting in queues. Even this staggering figure does not tell the whole story of the impact of causing excessive waiting. Great inefficiencies also occur because of other kinds of waiting than people standing in line. For example, making machines wait to be repaired may result in lost production. Vehicles (ships and trucks Buses, Cars) that need to wait to be unloaded may delay subsequent shipments. Airplanes waiting to take off or land may disrupt later travel schedules. Delays in telecommunication transmissions due to soaked lines may cause data glitches. Causing manufacturing jobs to wait to be performed may disrupt subsequent production. Delaying service jobs beyond their due dates may result in lost future business. Queueing theory is the study of waiting in all these various fields. Queueing models to provide the various types of queueing systems (systems that involve queues of some kind) that arise in daily practice. Formulas for each model indicate how the corresponding queueing system should perform the service, including the average amount of waiting that will occur, under a diversity of circumstances. Therefore, these queueing models are very helpful for determining how to operate a queueing system in the most effective way. Providing too much service capacity to operate the system involves excessive costs. But not providing enough service capacity results in excessive waiting and all its unfortunate consequences. The models also provide to decision an appropriate balance between the cost of service and the amount of waiting.

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.

Basic System of Queueing Process:



Departure from the system
Figure 1 Basic Queue Process

A Basic queuing 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

III. QUEUEING NETWORKS

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 queuing systems in the network may be other queuing systems in the same network (Figure 2).

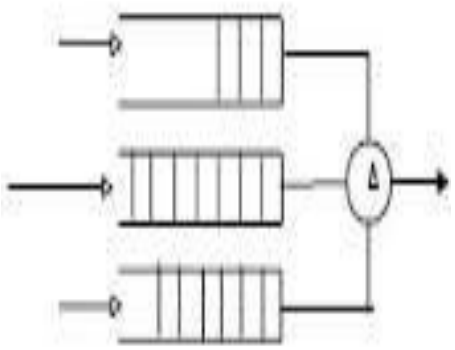


Figure 2 – Queuing network

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

Queue model:

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

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

Arrival Process

- Jobs/customer arrival pattern
- τ form a sequence of Independent and Identically Distributed(IID) random variables
  - Arrival times : t1, t2, …, tj
  - Interarrival times : T_j=tj-tj-1
- Arrival models
  - Exponential + IID (Poisson)
  - Erlang
  - Hyper-exponential
  - General : results valid for all distributions
  -

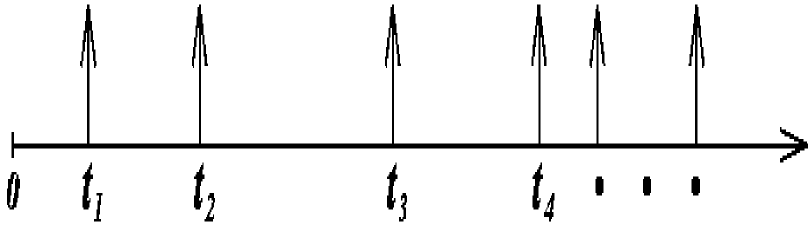


Figure 3

Service Time Distribution

- Time each user spends at the terminal
- IID
- Distribution model
  - Exponential
  - Erlang
  - Hyper-exponential
  - General
- cf.
  - Jobs = customers
  - Device = service center = queue
  - Buffer = waiting position

Number of Servers

Single Server Queue

Multiple Server Queue

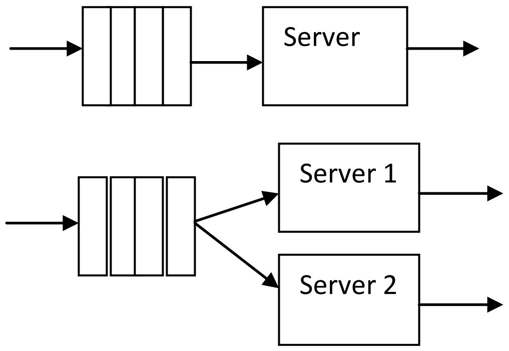


Figure 4

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

Common Distributions

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

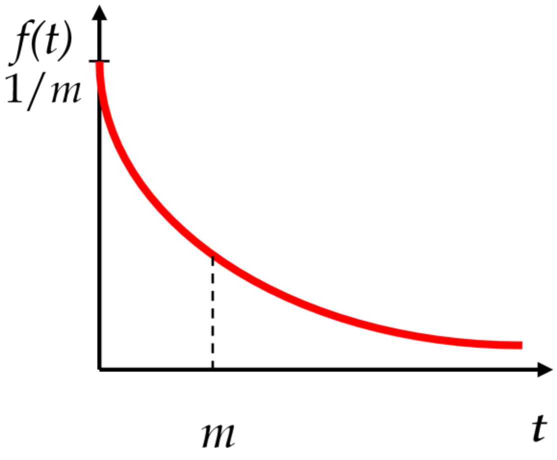


Fig .5(a) Probability density function for the exponential distribution

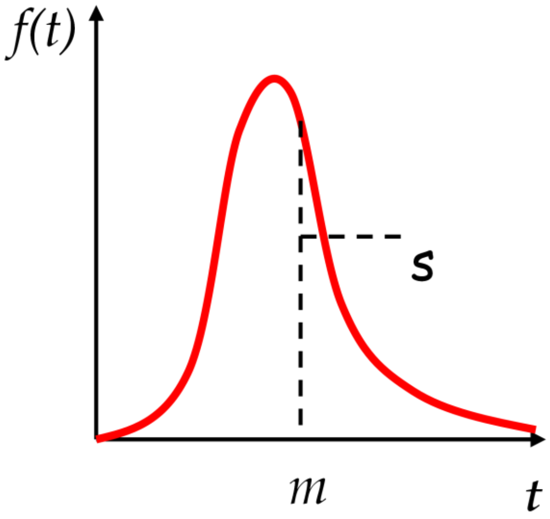


Figure 5(b) Probability density function for the exponential distribution General

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
- – G/G/1 ⇔ G/G/1/ $\infty/\infty/\mathrm{FCFS}$

Key Variables used in following process

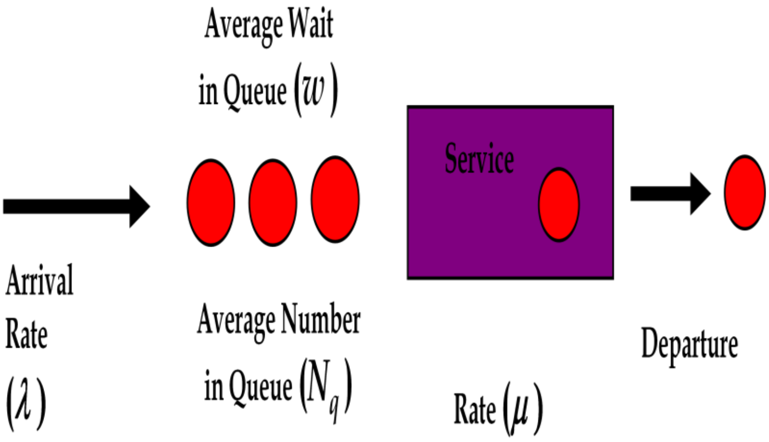


Figure 6

- 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
-

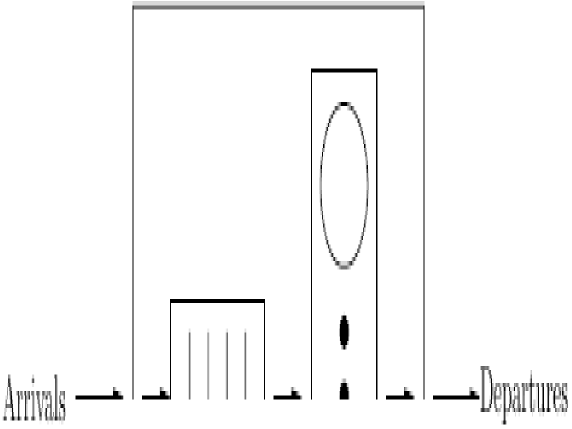


Figure 7

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

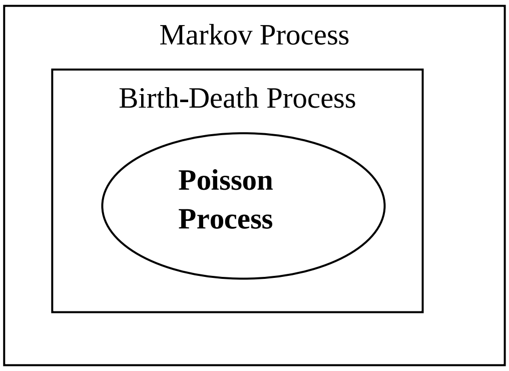
• A queuing system is characterized by three elements:

– 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,…

• 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

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

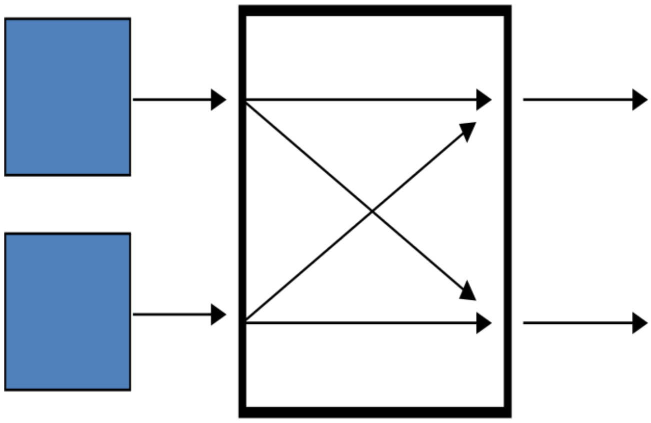


Figure 9

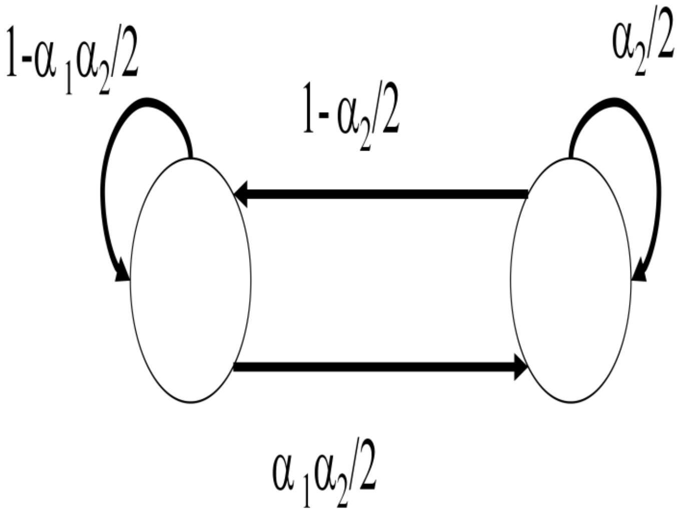


Figure 10

Birth-Death Processes

❖ The foundation of many of the most commonly used queuing models

✓ Birth – equivalent to the arrival of a customer or job

✓ 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)

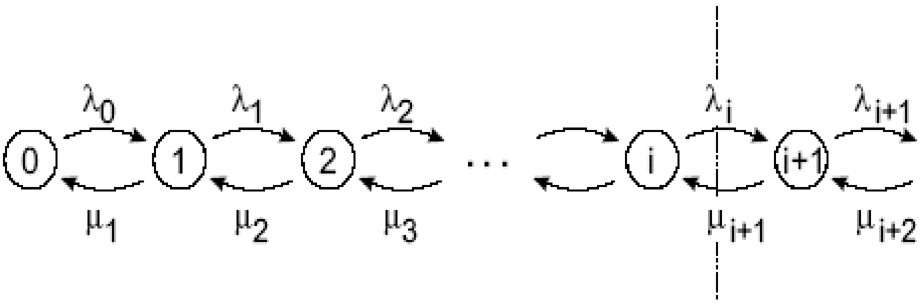


Figure 11

Poisson Processes

- Interarrival time s = Independent and Identically Distributed(IID) + exponential
- Birth death process that λk = λ, μk = 0 for all k

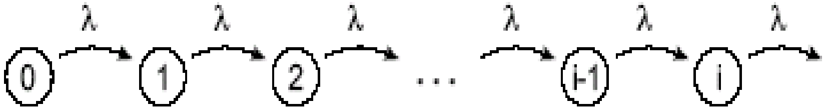


Figure 12

- 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 interarrival 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 λ → exponentially distributed

– Service rate of μ → 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

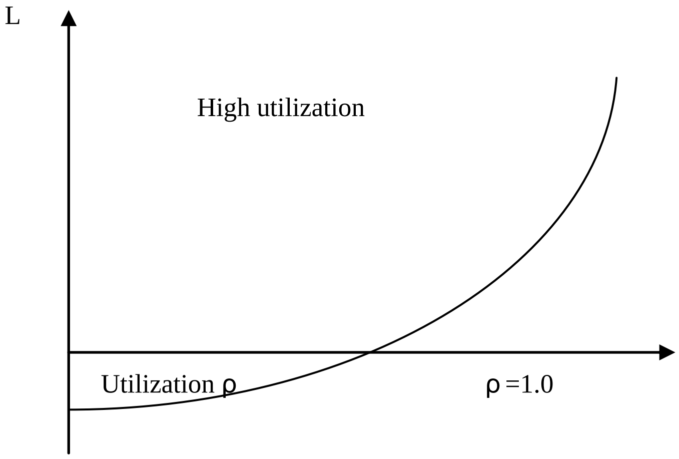
- Utilization(fraction of time server is busy)
  - $\rho = \lambda/\mu$
- Average waiting times
  - $W = 1/(\mu - \lambda)$
  - $Wq = \rho/(\mu - \lambda) = \rho W$
- Average number waiting
  - $L = \lambda /(\mu - \lambda)$
  - $Lq = \rho \lambda /(\mu - \lambda) = \rho L$

Flexibility/Utilization Trade-off

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

Lq
W
W
Low ops costs
Low flexibility
Poor service
Low utilization
High ops costs
High flexibility
Good service

ρ =0.0
Figure 13



Cost Trade-offs

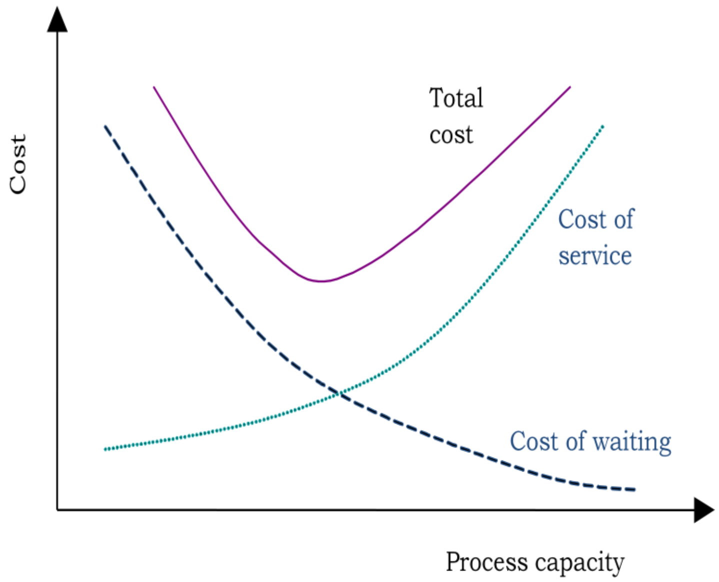


Figure 14

M/M/1 Queue Example

On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per seconds (pps) and the gateway takes about two milliseconds to forward them. Using an M/M/1 Queue model, analyze the gateway. What is the probability of buffer overflow if the gateway had only 13 buffers? How many buffers do we need to keep packet loss below one packet per million.

- Arrival rate λ = 125pps
- Service rate μ = 1/.002 = 500 pps
- Gateway utilization $\rho = \lambda/ \mu = 0.25$
- Probability of n packets in the gateway

– (1- ρ) ρ n = 0.75(0.25)n
- Mean number of packets in the gateway

– ρ/(1- ρ) = 0.25/0.75 = 0.33
- Mean time spent in the gateway

– (1/ μ)/(1- ρ) = (1/500)/(1-0.25) = 2.66 milliseconds
- Probability of buffer overflow
- P(more than 13 packets in gateway) = ρ13 = 0.2313 =1.49 X 10^{-8} ≈ 15 packets per billion packets
- To limit the probability of loss to less than 10^{-6}

– ρ n < 10^{-6}

– n > log(10^{-6})/log(0.25) = 9.96

– Need about 10 buffers

IV. CONCLUSIONS

Queueing systems are useful throughout society. The capability of these systems can have an important result on the quality of human life and productivity of the process. Queueing theory studies queueing systems by formulating mathematical models of their operation and then using these models to derive measures of performance. This analysis provides fundamental information for successfully designing queueing systems that achieve an appropriate balance between the cost of providing a service and the cost associated with waiting for that service. Queueing Models are useful to evaluate the performance of Networking in Parallel & Distributed System Models.

REFERENCES

[1]. Queueing Theory with Application to Packet Telecommunication: Solution Manual Prof. of Electrical Engineering ,The
[2]. University of Mississippi,University, MS 38677
[3]. Cooper, R. B.: Introduction to Queueing Theory, 2d ed., Elsevier North-Holland, New York, 1981.
[4]. Queue Network,- Customers Signals and Product form Solution, by X. Chao, M. Miyazawa, and
[5]. M. Pincdo Journal of Applied Mathematics and Stochastic Analysis, 14:4 (2001), 421-426
[1]. Basic Queueing Theory: Debrecen, 2011 by :Dr. János Sztrik University of Debrecen, Faculty of Informatics
[2]. Karlin, S., and Taylor, H. Stochastic process (in Hungarian). Gondolat Kiadó, Budapest, 1985.
[6]. Karlin, S., and Taylor, H. An introduction to stochastic modeling. Harcourt, New York, 1998.
a. Khintchine, A. Mathematical methods in the theory of queueing. Hafner, New
[1]. Kleinrock, L. Queueing systems. Vol. I. Theory. John Wiley & Sons, New York, 1975
[7]. Walrand, J.: An Introduction to Queueing Networks, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[8]. Wolff, R. W.: Stochastic Modeling and the Theory of Queues, Prentice-Hall, Englewood Cliffs,NJ, 1989.
[9]. I.G. Boldur-LăŃescu, I. Suciu, E. łigănescu, Operational Research with Applications in Economy, A.S.E., 1990
a. The art of computer systems performance analysis. Raj Jain Gh. Dodescu, Operating Systems, ASE 1997
[1]. Gh. Dodescu, B. Oancea, M. Raceanu, Parallel Processing, Ed. Economica, Bucharest, 2002
b. D. Gross, C. M. Harris, Fundamentals of Queuing Theory, Wiley, New York, 2003 J. Joseph, C. Fellenstein, Grid Computing, Prentice Hall, 2003