End-to-End Delay caused by M/M/1 Queuing Model

Kiran Gowda.C[1], K.C.Gouda[2], K.Raghuveer[3], Kumuda.M.N[4]

[1]*Dept. of PGS-CEA, The National Institute of Engineering, Mysore, Karnataka, India*
abgowda@gmail.com

[2]*CSIR-Centre for Mathematical Modelling and Computer Simulation,*
Wind tunnel Road, Bangalore-37 Karnataka, India
kcgouda@cmmacs.ernet.in

[3]*Dept of ISE, The National Institute of Engineering, Mysore, Karnataka, India*
raghunie@gmail.com

[4] *Dept. of ISE, Don Bosco Institute of Technology, Bangalore, Karnataka, India*
kumudapapu@gmail.com

Abstract **One of the most important process is the prediction of congestion in a system, as measured by delays caused by waiting in line for a service in a network. Even if a service system has the capacity to provide service at a faster rate than the rate at which customers arrive, waiting lines can still form if the arrival and service processes are random. Some queuing models assume that customer arrivals and service completions follow the Poisson process. Queuing theory provides a mathematical basis for understanding and predicting the behaviour of communication networks. End-to-End delay refers to the time taken for a packet to be transmitted across a network from source to destination. This delay can be minimized using some algorithms. In this paper a system is proposed which shows, how a system is utilized to overcome end-to-end delay using queuing model, this is done depending on arrival rate and service rate of the customer. Finally the results are presented in terms of the probability of the system utilization.**

Keywords— **Delay, Queuing theory, arrival and service.**

Introduction

I. INTRODUCTION

Queuing theory is the mathematical study of waiting lines, or queues [1]. In queuing theory a model is constructed so that queue lengths and waiting times can be predicted. This theory deals with the analysis of queues (or waiting lines) where customers wait to receive a service. Network delay is an important design and performance characteristic of a computer network or telecommunications network. The delay of a network specifies how long it takes for a bit of data to travel across the network from one node (or endpoint) to another. It is typically measured in multiples or fractions of seconds. Delay may differ slightly, depending on the location of the specific pair of communicating nodes. Although users only care about the total delay of a network, engineers need to perform precise measurements. Thus, we report both the maximum and average delay, and they divide the delay into several parts [3]:

1. Processing delay - time routers take to process the packet header
2. Queuing delay - time the packet spends in routing queues
3. Transmission delay - time it takes to push the packet's bits onto the link
4. Propagation delay - time for a signal to reach its destination

Delay happens due to all of the above four characteristics.

However, in a non-trivial network, a typical packet will be forwarded over many links via many gateways, each of which will not begin to forward the packet until it has been completely received. In such a network, the minimal latency is the sum of the minimum latency of each link, plus the transmission delay of each link except the final one, plus the forwarding latency of each gateway. In practice, this minimal latency is further augmented by queuing and processing delays. Queuing delay occurs when a gateway receives multiple packets from different sources heading towards the same destination. Since typically only one packet can be transmitted at a time, some of the packets must queue for transmission, incurring additional delay. Processing delays are incurred while a gateway determines what to do with a newly received packet. A new and emergent behaviour called Buffer bloat can also cause increased latency that is an order of magnitude or more. The combination of propagation, serialization, queuing, and processing delays often produces a complex and variable network latency profile.

Wait time is affected by the design of the waiting line system. A waiting line system (or queuing system) is defined by two elements: the population source of its customers and

the process or service system itself. In this supplement we examine the elements of waiting line systems and appropriate performance measures. Performance characteristics are calculated for different waiting line systems. We conclude with descriptions of managerial decisions related to waiting line system design and Performance.

II. ARRIVAL AND SERVICE PATTERNS

Waiting line models require an arrival rate and a service rate. The arrival rate specifies the average number of customers per time period. For example, a system may have ten customers arrive on average each hour. The service rate specifies the average number of customers that can be serviced during a time period. The service rate is the capacity of the service system. If the number of customers you can serve per time period is less than the average number of customers arriving, the waiting line grows infinitely. It is the variability in arrival and service patterns that causes waiting lines. Lines form when several customers request service at approximately the same time. This surge of customers temporarily overloads the service system and a line develops. Waiting line models that assess the performance of service systems usually assume that customers arrive according to a Poisson probability distribution, and service times are described by an exponential distribution. The Poisson distribution specifies the probability that a certain number of customers will arrive in a given time period (such as per hour).The exponential distribution describes the service times as the probability that a particular service time will be less than or equal to a given amount of time.

Waiting Line Priority Rules: A waiting line priority rule determines which customer is served next. A frequently used priority rule is first-come, first-served [4].

III. MODEL DESCRIPTION

The simplest view of a queuing model is shown in fig1. Here we represent a node that is characterized by what is known as a single-server queue. This node has a buffer associated with it in which arriving jobs queue up and wait for service from a single processing element (a single server).

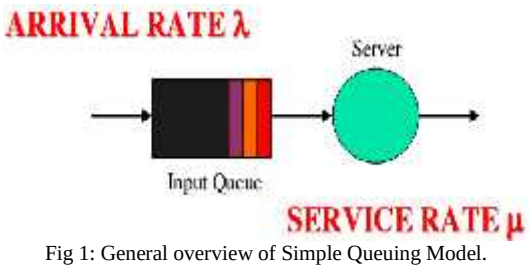

Fig 1: General overview of Simple Queuing Model.

The customer who arrives at the buffer can be coming either from a group of customer sources that are directly connected to the node or they can arrive on an external line that is connected to another node. The source of customer can be finite or infinite. A finite-source system cannot have an arbitrarily long service queue, but the number of customer sources in the system affects the arrival rate; that is, the larger the number of sources, the higher the customer arrival rate. For an infinite-source system the length of the service queue is unlimited, and the arrival rate is not affected by the number of customers in the system. If the source Population is finite but large, an infinite-source system is usually assumed to simplify the mathematics.

IV. EXPERIMENTAL RESULTS

The customer arrives at the buffer at a rate of λ customer per second and they have a length of X data units. Functions at a node are characterized by a service rate and a queuing discipline. The service rate is expressed as the number of jobs leaving the node per unit servicing time. This service rate may be load-dependent, meaning that the rate at which the mode processes jobs may depend on the length of the associated queue. The queuing discipline is the rule used to determine the order in which the queued-up jobs receive service. For example, queues in supermarkets, banks, or airports process jobs in the order of arrival. This is known as a first-come-first-served (FCFS) discipline. In other situations, such as in hospital emergency rooms, for example, there are certain customers (jobs), which obviously have priority over others.

V. END-TO-END PATH DELAY CALCULATION AND DISCUSSION

Each queue delay is formed by queuing delay (wait delay) and transmission delay in the queuing network for WSNs. Transmission delay is linked with the arrival rate of packets and the service rate of nodes.[2]

The utilization of node *i* is given by

$\rho_i = \frac{a_i}{\mu_i}$, Let λ be the average arrival rate and μ be the average service rate[5]. Then:

1. Processor utilization: $\rho = \lambda / \mu$
2. Average queue length: $L = \rho / 1 - \rho$
3. Average number waiting: $L_q = \rho^2 / 1 - \rho$
4. Average waiting time: $W_q = \rho / \mu (1-\rho)$
5. Average service time: $W_s = 1 / \mu$
6. Average response time:
$W = \rho / \lambda (1-\rho) = 1 / \mu (1-\rho) = W / 1 - \rho$

End-to-end delay caused by M/M/1 queuing models of the N-level nodes are presented by $E(T) = \sum_{i=1}^{N} \frac{1}{\mu_i - a_i} = \sum_{i=1}^{N} \frac{1/\mu_i}{1/\rho_i}$

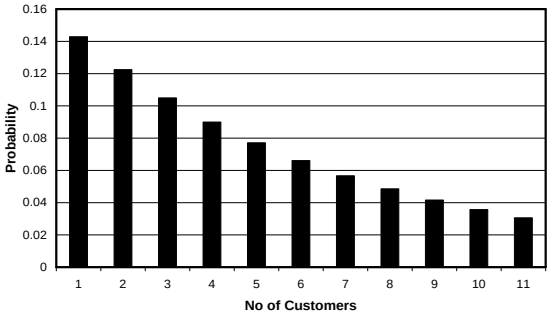

Fig: 2 Variation of probability with the number of customers.

Figure 2 describes the probability trend and it is obvious that the probability decreases with increase in the customer number. In the present work the probability analysis carried out with two approach, in the first (second) case λ(μ) is changed with fixed μ(λ). Figure 3 presents the probability of first scenario for fixed μ (=35) and its found that both the probability (pro) and utilization (row) increases with increase in λ. Similarly figure 4 presents the pro and row variation for fixed λ (=30) and varying μ and its observed that for μ value 60 both the pro and row are same (0.5). Figure 5 presents results with λ (=90) and higher μ (i.e 95 to 130) and in all cases pro and row shows reverse relation with each other.

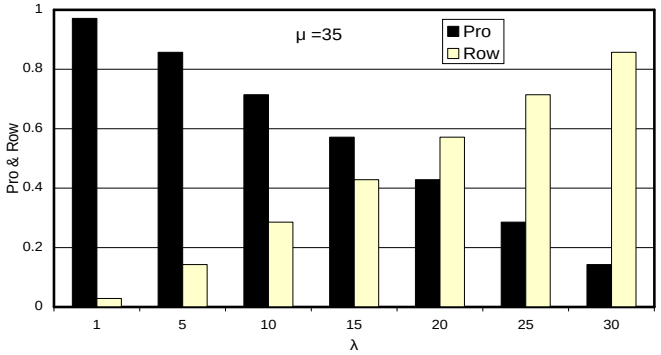

Fig : 3 Probaility and utilisation analysis with changing λ and fixed μ=35.

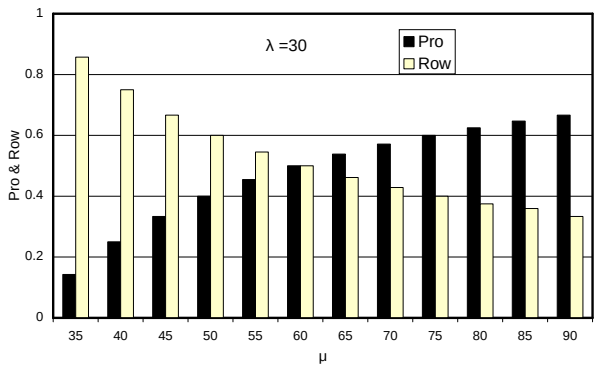

Fig : 4 Probaility and utilisation analysis with changing μ for fixed λ (30).

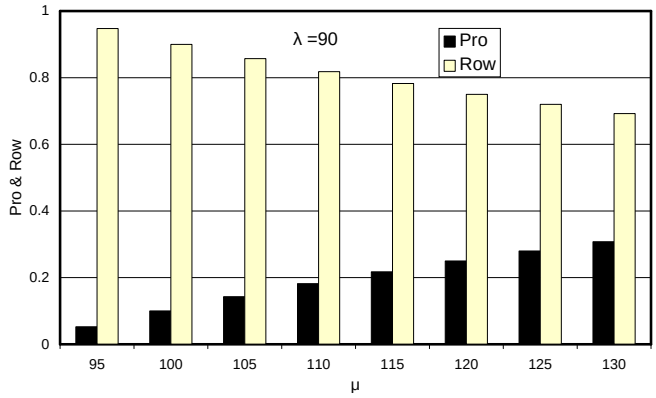

Fig : 5 Probaility and utilisation analysis with changing μ. for constant λ (90).

To know the sensitivity of λ and μ in the simulation, we carried out some more experiments. Figure 6 represents the reverse relationship of node utilisation and probability which are computed from several experiments with fixed μ(=45). Similarly figure 7 presents the average values from 50 simulation carried out with a smaller value of λ (=5). There is a clear indication of the sensitivity of results with the value of λ and μ.

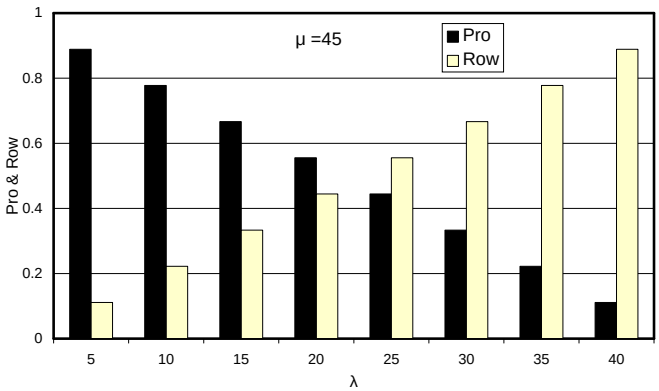

Fig : 6 Average Probaility and utilisation analysis from 50 simulations as a function of λ and fixed μ=45

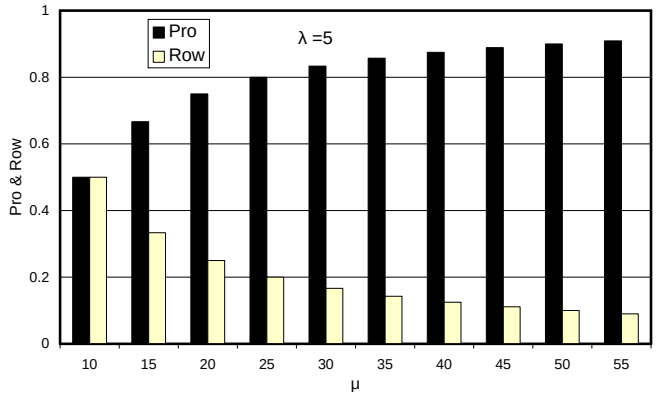

Fig : 7 Average Probaility and utilisation analysis from 50 simulations as a function of μ and fixed λ=5.

V. CONCLUSIONS

This paper has presented a method of modeling and delay analysis for queuing model. Based on the principles of data transmisssion queuing network model is established. Using that model an approximate method is designed to calculate arrival and servide rate. If the network is very large the calculation is very difficult, so the queuing network of a single server M/M/1 type system is discussed in this paper. Also the average delay calculation for entire queuing network is carried out and the optimization analysis of queuing network for delay analysis needs to be done in the future.

REFERENCES

[1]. Introduction to Waiting Line Models_ ©David W. Ashley, 2000.
[2]. Tie QIU, Feng XIA, Lin FENG, Guowei WU, Bo JIN, Queueing theory-based path delay analysis of wireless sensor networks. Advances in Electrical and Computer Engineering Volume 11, Number 2, 2011.
[3]. Muriel Medard, EECS, LIDS, MIT, Delay Models and Queueing.
[4]. csus.edu/indiv/b/blakeh/mgmt/OPM101 SupplC.pdf Waiting Line Models Waiting Line Models .
[5]. Thin-Yin Leong, Simpler Spreadsheet Simulation of Multi-Server Queues. Vol. 7, No. 2, January 2007, pp. 172–177 ISSN1532-0545 - 07 – 0702- 0172.

Kiran Gowda C presently pursuing his M.Tech (Computer Networking) at the National Institute of Engineering, Mysore, Affiliated to Visvesvaraya Technological University, Karnataka, India. He joined as an Asst. Prof. in Rajarajeswari College of Engineering, Bangalore, India since 2011. His Research interests includes Wireless adhoc networks, Image Processing, Big data analysis, Simulation and Computer Networks. He has about 6 year of teaching experience. He has presented several papers in the International and national conferences.

K C Gouda is currently working as a Scientist at CSIR Centre for Mathematical Modeling and Computer Simulation (CSIR C-MMACS). His research and professional career spans about twelve years of research and capacity building in modeling and computer simulation, satellite data processing, numerical modeling, Data mining, Data assimilation, cloud computing knowledge engineering and related subjects. His expertise is primarily in the domains of Software development for modeling and simulation. He is presently involved in several international and national projects related to HPC enabled modeling for weather and climate forecasting and analysis. He has published about 75 peer-reviewed papers as journal articles, book chapters, Technical reports and contributions to conference proceedings. He is a member of IMS, IEEE, AGU, AOGS and EGU. He is also a member in the board of studies of Department of Computer Science in the Jain University and Dayananda Sagar Institutes, Bangalore. He obtained his M.Sc., M.Phil, from Berhampur University, MCA from IGNOU, New Delhi and completed PhD from Mangalore University. He has Guided 15 M.Tech., 55 Masters and 30 B.E students for their academic project.

K Raghuveer has been at the Department of Computer Science & Engineering , National Institute of Engineering Mysore since 1984. He has completed M.E from Devi Ahilya VishwaVidyalaya , Indore (1991-93) and Ph.D from VTU (2003-2007). Currently he is working as Professor and Head, Department of Information Science & Engineering. He has worked as Chairman , Board of Studies in Computer and Information Science. He has participated in EDUSAT, a distance based satellite education and delivered number of courses. He was a Board Member for Examination, VTU, Board of Studies Member for many committees. He is a executive member for Center for Sericulture Board, a Government of India. He has presented number of papers in national and international journals/conferences. He was a Visiting Professor for Hunghuai University, China. He has guided 40 M.Tech students. Currently he is guiding two students for Ph.D . He is a Life member for Indian Society for Technical Education and Fellow of Institute of Engineers.

Kumuda M N obtained her M.Tech. Degree from SJCE, Mysore, Affiliated to Visvesvaraya Technological University, Karnataka, India. Presently working as an Asst. Prof. in the Dept of Information Science and Engineering, Don Bosco Institute of Technology, Bangalore, India. Her Research interests are in the field of Wireless adhoc networks, Image Processing and Computer Networks. He has about 6 year of teaching experience. She has presented several papers in the International and national conferences.