ISSN No. 0976-5697

Volume 3, No. 5, Sept-Oct 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Multiple Server Performance Modeling using M/M/n Queue

M.Murali*
Assistant Professor, School of Computing
SRM University Chennai 603 203, India.
mmurali.srm@gmail.com

Dr. R. Srinivasan
Professor Emeritus, School of Computing
SRM University Chennai 603 203, India
rsv38@yahoo.co.in

Abstract: We develop a framework for optimizing the queuing system in a multiple server environment. The motivation is the servicing problem of large no. of servers, which we have modeled as M/M/n queues with N, the number of servers in this work. The subject of the present work is the M/M/n queue. This queue is characterized by Poisson arrivals at rate λ, exponential service time at rate μ, n servers and unlimited waiting positions of packets in the queue.. The model is applied in the multiple server environment, and it finds the service quality of multiple server

Keywords: MMN queue, server occupancy, waiting time

I. INTRODUCTION

It is a challenge in designing a queue and managing a service operation in general, to achieve a desired service quality in multiple server environment. In this work, we consider the aspect of service quality having the number of servers to process the packets with minimum waiting time. Thus it can save the operational costs and avoiding packets to wait in the queue and dropping. In a multiple server environment, servers can process the packets per unit of time could exceed the maximum level. To reach the level of good performance, would require stochastic queuing models. In a large systems, all are identical servers. The servers can handle the packets and the speed in which they can process the packets may also same.

II. RELATED WORK

Gnedenko and Kovalenko[1] analyzed the M/M/n +D queueing system. Jurkevic[2] applied their methods to the general M/M/n +G system. The M/M/*n* +G queue was analyzed by Baccelli and Hebuterne[3] and Haugen and Skogan[4]. Boxma and de Waal[5] developed several approximations for the probability to abandon in the M/M/n +G queue and tested them via simulation. Brandt and Brandt[6,7] considered the more general M(*k*)/M(*k*)/*n* +G system where arrival and service rates are allowed to depend on the number *k* of calls in the system.

III. M/M/N QUEUEING MODEL

The classical M/M/n queueing model[8], also called Erlang-C, is the model used in multiple server environment. Erlang-C assumes Poisson arrivals at constant rate λ, exponentially distributed service times with a rate μ, and n independent identical servers. The Fig. 1 is Showing the M/M/n queue, has the following features: An arriving packet encounters an offered waiting time, defined as the time this packet would have to wait, given that this waiting time is infinite. If the offered waiting time exceeds the process's waiting time, the packet is then dropped otherwise the packet can wait for service.

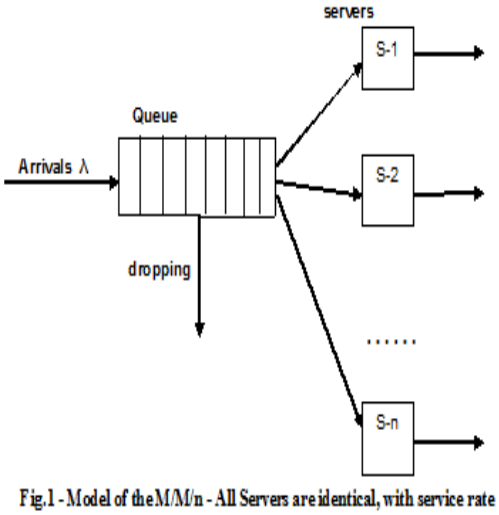

Fig.1 - Model of the M/M/n - All Servers are identical, with service rate μ

Figure. 1 Model of the M/M/n – servers are identical , service rate μ

The measures of performance are, average waiting time for a packet, traffic intensity, server occupancy, Probability of a packet has to wait in a queue, required no. of servers. Consider 10 computer users, which produces, average 300 packets per minute. The inter arrival times of packets are exponentially distributed. The lengths of packets are also exponentially distributed, with mean 128k size.

Erlang-C Calculations : 300 packets per minute, with an average packet size 128k, and 7 servers.

λ average arrival rate of packets =

300 packets / per minute (1)

= 5 packets / second

T_s = packet service duration = 0.02/ minute (2)

= 0.02/ minute

= 1.2 seconds

m = number of servers =7

Traffic intensity = u = λ . T_s

u = (5 packets / seconds) . (1.2 seconds / packet) (3)

= 6

$$\text{Server Occupancy} = \rho = \frac{\mu}{m} = \frac{6}{7} = 0.85 \quad (4)$$

a) Erlang-C formula :

$$Ec(m,u) = \frac{\frac{u^m}{m!}}{\frac{u^m}{m!} + (1-\rho)\sum_{k=0}^{m-1}\frac{uk}{k!}} = 0.6138 \quad (5)$$

E_C (m,u) = Probability, that a packet has to wait in the queue

= 0.6138

$$Tw = \text{Average Waiting Time} = \frac{Ec(m,u).Ts}{m.(1-\rho)} \quad (6)$$

$$= \frac{0.6138 * 1.2}{7 * 0.15} = \frac{0.736}{1.00} = 0.73 \text{ seconds}$$

Arrival rate of packets = 300 packets per minute
(i.e., 5 packets /second)

Packet service duration = 1.2 seconds

Traffic intensity = 6.0 packets

IV. RESULT AND DISCUSSION

Table 1 Packet Distribution Report

No. of Servers	Parameters		
	Server Occupancy	Probability a packet to wait in the Queue	Average Waiting Time (m.secs)
7	**0.85**	**0.61**	**0.7300**
8	0.75	0.35	0.2100
9	0.66	0.19	0.0780
10	0.60	0.10	0.0300
11	0.54	0.04	0.0118

The Table I is showing the packet distribution report. In this table arrival rate of packets is 300 / per minute. The server occupancy value should be between 0 and 1. If the server occupancy value is greater than 1, the servers are overloaded. This is the parameter to suggest the optimum no. of servers required. And the average waiting time of a packet also shown in the Table I. In addition to that, the probability a packet has to wait in the queue is also shown in the Table I. In Fig. 2, the graph is showing the average waiting time of a packet to get the service. The Fig. 3 is showing the probability, a packet has to wait in the queue. When all servers are busy, the offered waiting time is high in the system.

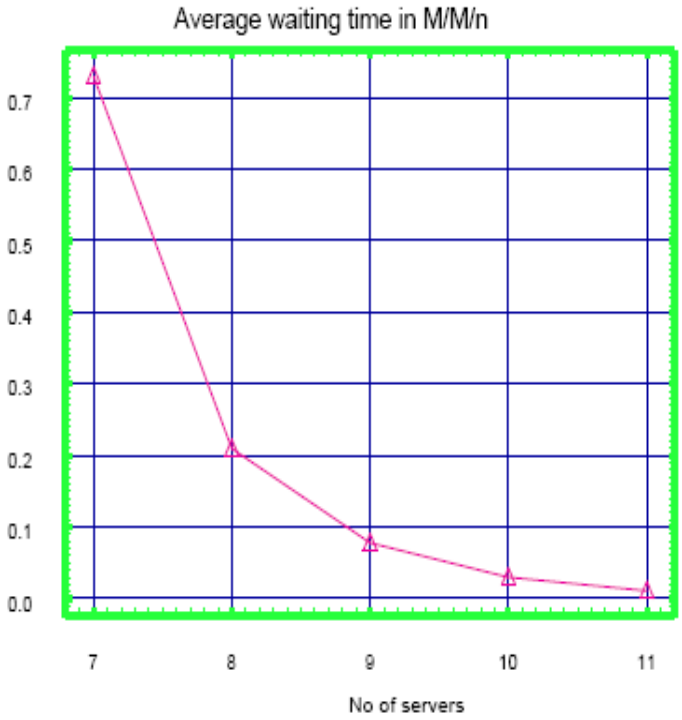

Figure **2.** Average waiting time of a packet in M/M/n model

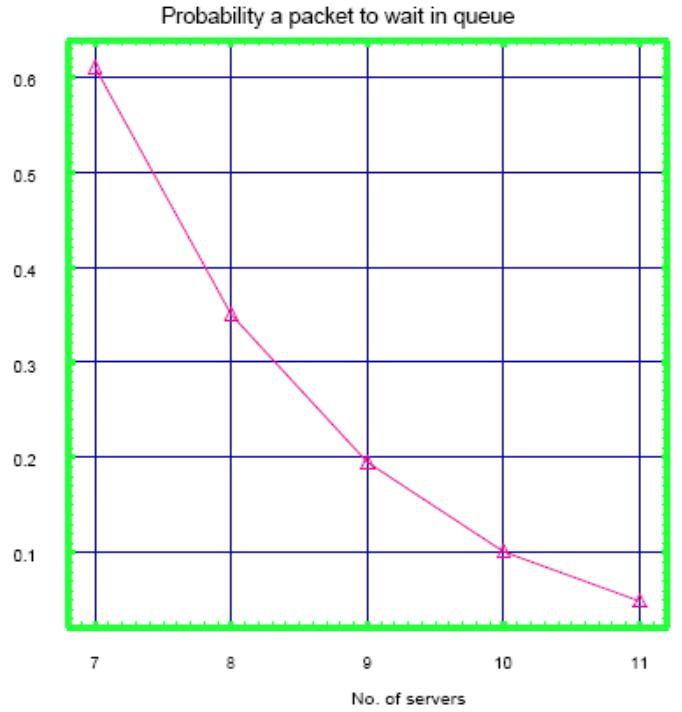

Figure 3. Probability, a packet has to wait in the Queue

V. CONCLUSION

From the above, the conclusion is that the server occupancy , probability of a packet has to wait in the queue should be between 0 and 1. When the arrival rate of packets is increased, the server occupancy exceeds 1. This value indicates that the servers have been overloaded. At this point, we can suggest to increase the no. of servers. So that the packet dropping could be avoided. The optimum no. of arrival rate of packets is 5/second, server occupancy is 0.85 and the probability of a packet has to wait in the queue is 0.6138 is suggested. Since the servers are identical, there is no difference in the service time of a packet. And the effect of the waiting time depends on the load that the system is working under. It is very important, that the offered load (server occupancy) is significantly below 1 or around 1. We hope that the given analysis is sufficient to understand the performance of multiple servers..

VI. REFERENCES

[1] B.W. Gnedenko and I.N. Kovalenko, Introduction to Queueing Theory (Jerusalem, Israel) (1968) Program for Scientific Translations.

[2] O.M. Jurkevic, On many-server systems with stochastic bounds for the waiting time (in Russian), Izv. Akad. Nauk SSSR Techniceskaja kibernetika 4 (1971) 39–46.

[3] F. Baccelli and G. Hebuterne, On queues with impatient customers, in: Kylstra F.J. (ed.), Performance ’81, North-Holland Publishing Company, 1981) pp. 159–179.

[4] R.B. Haugen and E. Skogan, Queueing systems with stochastic time out. IEEE Trans. Commun.COM-28, (1980) 1984–1989.

[5] O.J. Boxma and P.R. deWaal, Multiserver queues with impatient customers, ITC 14 (1994) 743–756.

[6] Brandt.A and Brandt.M, On the M(n)/M(n)/s queue with impatient calls, Performance Evaluation 35 (1999) 1–18.

[7] Brandt.A and Brandt.M, Asymptotic results and a Markovian approximation for the M(n)/M(n)/s+ GI system, Queueing Systems, Theory and Applications (QUESTA) 41 (2002) 73–94.

[8] Leon Garcia.A and I. Widjaja, “Communication Networks,” McGraw Hill, 2000, First Edition.