PERFORMANCE MEASURES FOR INTERNET SERVER BY USING M/M/m QUEUEING MODEL

Raghunath Y. T. N. V[1], A. S. Sravani[2]

[1]Assistant professor, Dept.of ECE, DVR & Dr. HS MIC College of Technology, Andhra Pradesh, India,
[2]P.G. Scholar, Dept. of ECE, St. Ann's College of Engg &Tech, Andhra Pradesh, India,
bobby_ytnv@yahoo.com, sravani.ayachitula@gmail.com

Abstract

This paper deals with the performance measurement of single queue multiple server model. This gives the performance measure for internet server for highly dynamic traffic conditions. Our previous work is related to performance measurement of single queue single server model. This is achieved by analyzing the performance measures and capacity planning of internet server using different queuing models by comparing the parameters like queuing length, response time, waiting time for different links.

Keywords- Internet server, waiting time, response time, queue length, queuing models

---*****---

1. INTRODUCTION

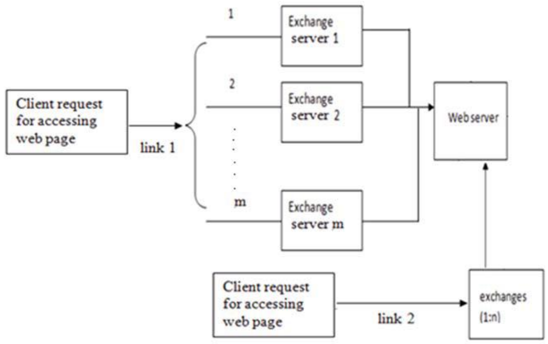

Figure 1.1: Layout of Internet Server-Data Arrival/ Data departure [1].

In Figure 1.1, the client requests for accessing the web page through different links those are link1 and link 2. These links may be either satellite link or optical link in real time. The next block is exchange servers, in the above diagram there are m exchange servers in link 1 and n exchange servers in links [1]. These exchange servers are connected to internet server. In this system, the client requests for the web page by typing the domain name of particular web page. Then the exchanges servers those are used to mapping that particular domain name with particular IP address. After that the web server will process the IP address and generates the client requested web page. Like this the exchange servers and internet servers are used process the client requests [2].

The objective of the paper is getting the performance measures and capacity estimation of the above system by using different queueing models i.e., M/M/1, M/G/1, M/D/1 and M/Ek/1.The above system is modeled in queueing network by using the above queuing models.

Analyzing a Queue System:

A full queuing situation involves arrivals and service[3], so we need some more Greek letters:
λ: "lambda" is the average customer arrival rate per unit time
μ : "mu" is the average customer service rate (when customers are waiting) =1/(average service time)
ρ :"rho" is the server utilization factor

It uses the Poisson and exponential distributions to model both arrival times and service times. As mentioned, the Poisson and exponential distributions are mathematically related. If the number of service completions per unit of time, when there is a backlog of customers waiting, has the Poisson distribution, then service time has the exponential distribution. It's conventional to think of service in terms of the length of service time.

The standard simple queuing model assumes that [4]:
1. Arrivals have the Poisson distribution
2. Service times have the exponential distribution
3. Arrivals and service times are all independent.
(Independence means, for example, that: arrivals don't come in groups, and the server does not work faster when the line is longer.)

The performance measures those can be done by using queuing models are utilization, queue length, waiting time and response time [8].

Utilization factor :- The utilization gives the fraction of time that the server is busy. It is defined as ratio of arrival rate to service rate.

Queue length :- It defines the maximum capacity of the queue or number of customers in the queue.

Waiting time :- It is defined as the ratio of average queue length to arrival rate or it defines the amount of time the customer has to be waited in the queue.

Response time :- It defines the sum of waiting time and service time for a particular customer [8].

2. LITERATURE REVIEW

In [1], this paper deals with an improved scheme for autonomous performance of gateway servers under highly dynamic traffic loads. The most wide spread contemplation is performance, because gateway servers offer cost-effective and high performance modeling and predictions. This paper describes possible queuing models that can be applied in capacity planning analysis. This is achieved by utilizing the internal queue length measurements. Extensive simulation study shows that the new scheme can provide smooth performance control and better tracking ability in web server systems.

In [2], this paper presents a workload characterization study for Internet Web Servers. Six different data sets are used in the study: three from academic environments, two from scientific research organizations, and one from a commercial internet provider. These data sets represent three different orders of magnitude in server activity, and two different order of magnitude in time duration, ranging from one week of activity to one year of activity.

In [3], this paper provides information about High performance web site design techniques. Performance and high availability are critical at web sites that receive large numbers of requests. This paper presents several techniques including redundant hardware, load balancing. Web server acceleration, and efficient management of dynamic data that can be used at popular sites to improve performance and availability. It describes how several techniques are deployed at the official web site.

3. SINGLE QUEUE SINGLE SERVER MODELS

3.1 Queuing network model for Internet Server using/M/1 model:

The M/M/1-Queue has interarrival times [6], which are exponentially distributed with parameter and also service times with exponential distribution with parameter . The system has only a single server and uses the FIFO service discipline. The waiting line is of infinite size.

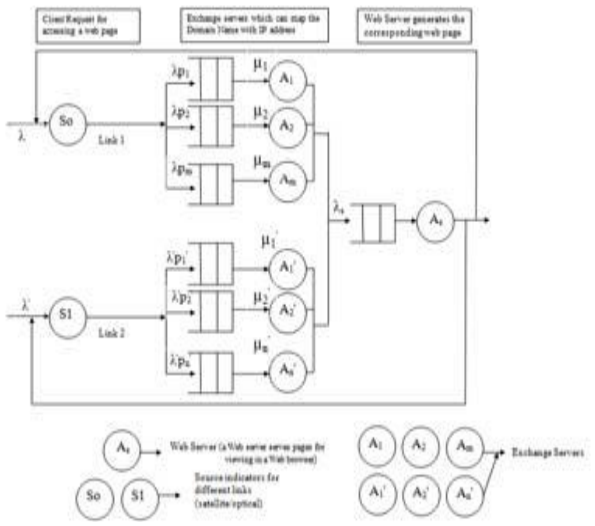
Client Request for accessing a web page
Exchange server which can map the Domain Name with IP address
Web Server generates the corresponding web page
λ
So
Link 1
λp1
λp2
λpm
μ1
μ2
μm
A1
A2
Am
λs
As
λ'
S1
Link 2
λp1'
λp2'
λpn'
μ1'
μ2'
μn'
A1'
A2'
An'
As
Web Server (a Web server serves pages for viewing in a Web browser)
So
S1
Source indicators for different links (satellite/optical)
A1
A2
Am
A1'
A2'
An'
Exchange Servers

Fig 3.1 Block diagram for queuing network model for Internet server using M/M/1.

Different Queueing models [6]:
1. M/M/1
2. M/D/1
3. M/G/1
4. M/Ek/1

Previous work is related to the comparison of server performance with different queueing models. This paper deals with the performance of single queue multiple server model.

4. SINGLE QUEUE MULTIPLE SERVER MODEL

4.1 M/M/m queueing model:

In this M/M/m queueing model, the 1st M indicates the interarrival time distribution arrivals follows Poisson distribution with parameter λ, and 2nd M indicates service time distribution [15]. Here, it is exponential distribution with parameter μ and 3rd m indicates number of servers available those are in parallel. The system has multiple server and uses the FIFO service discipline. The waiting line is of infinite size.

In the case M/M/1 queueing model, there is only one single server. It means that the system can process a single request at a time. But in this M/M/m queueing model, there are m number of servers that are connected in parallel [15]. That means, this model can process m requests at a time. So compared with the M/M/1 queueing model, this model gives better performance measures.

In figure 3.1, the M/M/m queueing model comprises m internet servers. In this system, the client requests for accessing the web page is by using domain name of that particular web page. then the request arrives at one of the m exchange servers. These exchange servers are used to map the client request domain name with the corresponding IP address and that result is arrived at any one of the internet servers, because this system uses multiple servers that are connected in parallel. Based on that arrived IP address, these internet servers that generate the web page which is displayed in the client's monitor.

The performance measures those can be done by using this queueing model are ulilization, queue length, waiting time and response time [15].

Utilization factor :- The utilization gives the fraction of time that the server is busy. It is defined as ratio of arrival rate to service rate.

Queue length :- It defines the maximum capacity of the queue or number of customers in the queue.

Waiting time :- It is defined as the ratio of average queue length to arrival rate or it defines the amount of time the customer has to be waited in the queue.

Response time :- It defines the sum of waiting time and service time for a particular customer.

4.2. Queueing network model for Internet Server using M/M/m model:

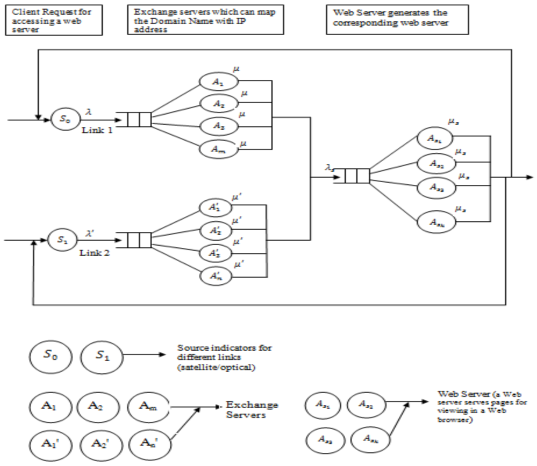
Client Request for accessing a web server
Exchange servers which can map the Domain Name with IP address
Web Server generates the corresponding web server
S_0
λ
Link 1
A_1
A_2
A_3
A_m
μ
S_1
λ'
Link 2
A_1'
A_2'
A_3'
A_n'
μ'
λ_s
A_{s_1}
A_{s_2}
A_{s_3}
A_{s_k}
μ_s
Source indicators for different links (satellite/optical)
Exchange Servers
Web Server (a Web server serves pages for viewing in a Web browser)

Fig 4.1. Block diagram for queueing network model for Internet server using M/M/m.

The analysis of queueing network for internet server is described as follows:

In figure 4.1, there are two links which are connected to the internet server to provides services to the users. In figure 4.1, S_o, S_1 are two source indicators that are used to request the web pages. Let $A_1, A_2, \ldots, A_m$ be the m parallel exchange servers in link 1 and let $A_1', A_2', \ldots, A_n'$ be the n parallel exchange servers and $A_{s_1}, A_{s_2}, \ldots, A_{s_k}$ be the k parallel internet servers. Let λ be the arrival at link 1 and λ' be the arrival rate at link 2. After the request arrives from the source, it arrives into the queue which is connected to the m exchange servers in link1 and n exchange servers in link 2. These exchanges are used for mapping the domain name with the corresponding IP address. Later, it arrives at any one of the k parallel internet servers that are used to process the IP address and generate the web page.

4.3. Performance measures for M/M/m model:

The performance measure for single queue multiple server system can be obtained by finding the utilization factor, queue length, waiting time, response time.

Utilization factor: - The utilization gives the fraction of time that the server is busy. It is defined as ratio of arrival rate to service rate.

According to figure 4.1, the utilization factor can be expressed as follows:

$$\rho = \frac{\lambda}{m\mu} + \frac{\lambda_s}{\kappa\mu_s} \text{ , (for link 1)} \qquad (4.1)$$

$$\rho' = \frac{\lambda'}{n\mu'} + \frac{\lambda_s}{\kappa\mu_s} \text{ . (for link 2)} \qquad (4.2)$$

Queue length: - It defines the maximum capacity of the queue or number of customers in the queue.

According to figure 4.1, the total queue length can be expressed as follows:

$$E\,[N_{tot}] = \mathbf{E\,[N]} + E\,[N_s], \quad \text{(for link 1)} \qquad (4.3)$$

$$E\,[N_{tot}'] = \mathbf{E}\,[N'] + \mathbf{E}\,[N_s]. \quad \text{(for link 2)} \qquad (4.4)$$

Where,

$$E\,[N] = m\,(\frac{\lambda}{m\mu}) + (\frac{\lambda}{m\mu})\frac{(\frac{\lambda}{\mu})^m}{m!}\frac{1}{\left[\sum_{x=0}^{m-1}\frac{\left(\frac{\lambda}{\mu}\right)^x}{x!}+\frac{\left(\frac{\lambda}{\mu}\right)^m}{m!}\frac{1}{\left(1-\frac{\lambda}{m\mu}\right)}\right](1-\frac{\lambda}{m\mu})^2},$$

(for link 1) (4.5)

$$E\,[N'] = n\,(\frac{\lambda'}{n\mu'}) + (\frac{\lambda'}{n\mu'})\frac{(\frac{\lambda'}{\mu'})^n}{n!}\frac{1}{\left[\sum_{x=0}^{n-1}\frac{\left(\frac{\lambda'}{\mu'}\right)^x}{x!}+\frac{\left(\frac{\lambda'}{\mu'}\right)^n}{n!}\frac{1}{\left(1-\frac{\lambda'}{n\mu'}\right)}\right](1-\frac{\lambda'}{n\mu'})^2},$$

(for link 2) ……………………………….(4.6)

$$E\,[N_s] = k\,(\frac{\lambda_s}{k\mu_s}) + (\frac{\lambda_s}{k\mu_s})\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{1}{\left[\sum_{x=0}^{k-1}\frac{\left(\frac{\lambda_s}{\mu_s}\right)^x}{x!}+\frac{\left(\frac{\lambda_s}{\mu_s}\right)^k}{k!}\frac{1}{\left(1-\frac{\lambda_s}{k\mu_s}\right)}\right](1-\frac{\lambda_s}{k\mu_s})^2}.$$

(for internet server) (4.7)

Waiting time: - It is defined as the ratio of average queue length to arrival rate or it defines the amount of time the customer has to be waited in the queue.

According to figure 4.1, the total waiting time can be expressed as follows:

E [W_{tot}] = **E [W]** + E [W_s] (for link 1)
E [$W_{tot}^{'}$] = **E [$W^{'}$] + E [W_s]** (for link 2)

Where,

$$E\,[W] = m\,(\frac{1}{m\mu}) + (\frac{1}{m\mu})\frac{(\frac{\lambda}{\mu})^m}{m!}\,\frac{1}{\left[\sum_{x=0}^{m-1}\frac{\left(\frac{\lambda}{\mu}\right)^x}{x!}+\frac{\left(\frac{\lambda}{\mu}\right)^m}{m!}\,\frac{1}{\left(1-\frac{\lambda}{m\mu}\right)}\right](1-\frac{\lambda}{m\mu})^2}\ ,$$

(for link 1) (4.8)

$$E\,[W^{'}] = n\,(\frac{1}{n\mu'}) + (\frac{1}{n\mu'})\frac{(\frac{\lambda^{'}}{\mu'})^n}{n!}\,\frac{1}{\left[\sum_{x=0}^{n-1}\frac{\left(\frac{\lambda^{'}}{\mu'}\right)^x}{x!}+\frac{\left(\frac{\lambda^{'}}{\mu'}\right)^n}{n!}\,\frac{1}{\left(1-\frac{\lambda^{'}}{n\mu'}\right)}\right](1-\frac{\lambda^{'}}{n\mu'})^2}\ ,$$

(for link 2) (4.9)

$$E\,[W_s] = k\,(\frac{1}{k\mu_s}) + (\frac{1}{k\mu_s})\frac{(\frac{\lambda}{\mu})^k}{k!}\,\frac{1}{\left[\sum_{x=0}^{k-1}\frac{\left(\frac{\lambda_s}{\mu_s}\right)^x}{x!}+\frac{\left(\frac{\lambda_s}{\mu_s}\right)^k}{k!}\,\frac{1}{\left(1-\frac{\lambda_s}{k\mu_s}\right)}\right](1-\frac{\lambda_s}{k\mu_s})^2}$$

(for internet server) (4.10)

Response time: - It defines the sum of waiting time and service time for a particular customer.

According to figure 4.1, the total response time can be expressed as follows:

E [R_{tot}] = **E [R]** + E [R_s] (for link 1)
E [$R_{tot}^{'}$] = **E [$R^{'}$] + E [R_s]** (for link 2)

Where,

$$E\,[R] = m\,(\frac{1}{m\mu}) + (\frac{1}{m\mu})\frac{(\frac{\lambda}{\mu})^m}{m!}\,\frac{1}{\left[\sum_{x=0}^{m-1}\frac{\left(\frac{\lambda}{\mu}\right)^x}{x!}+\frac{\left(\frac{\lambda}{\mu}\right)^m}{m!}\,\frac{1}{\left(1-\frac{\lambda}{m\mu}\right)}\right](1-\frac{\lambda}{m\mu})^2} + \frac{1}{\mu}$$

, (for link 1) (4.11)

$$E\,[R^{'}] = n\,(\frac{1}{n\mu'}) + (\frac{1}{n\mu'})\frac{(\frac{\lambda^{'}}{\mu'})^n}{n!}\,\frac{1}{\left[\sum_{x=0}^{n-1}\frac{\left(\frac{\lambda^{'}}{\mu'}\right)^x}{x!}+\frac{\left(\frac{\lambda^{'}}{\mu'}\right)^n}{n!}\,\frac{1}{\left(1-\frac{\lambda^{'}}{n\mu'}\right)}\right](1-\frac{\lambda^{'}}{n\mu'})^2} + \frac{1}{\mu'}$$

, (for link 2) (4.12)

$$E\,[R_s] = k\,(\frac{1}{k\mu_s}) + (\frac{1}{k\mu_s})\frac{(\frac{\lambda}{\mu})^k}{k!}\,\frac{1}{\left[\sum_{x=0}^{k-1}\frac{\left(\frac{\lambda_s}{\mu_s}\right)^x}{x!}+\frac{\left(\frac{\lambda_s}{\mu_s}\right)^k}{k!}\,\frac{1}{\left(1-\frac{\lambda_s}{k\mu_s}\right)}\right](1-\frac{\lambda_s}{k\mu_s})^2} + \frac{1}{\mu_s}\,.$$ (for internet server)

(4.13)

5. Numerical Results

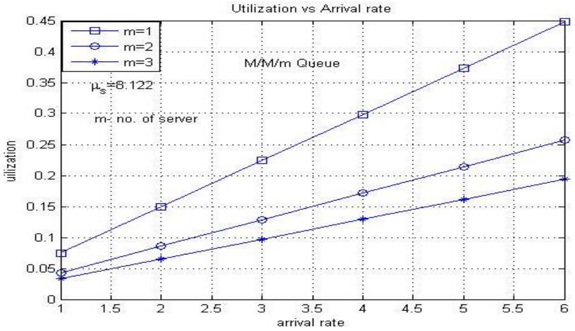

Fig 4.2: Plot for finding the Utilization factor for internet server using M/M/m queuing model.

Figure 4.2 shows the curves for utilization versus arrival rate with different m values. Here, m represents the number of servers. In the above figure, there are three different curves for three different m values. Higher the value m, lower is the utilization factor because, if the number of servers increases, the incoming arrivals will be shared for service with different servers. Then, the utilization factor for each internet server decreases.

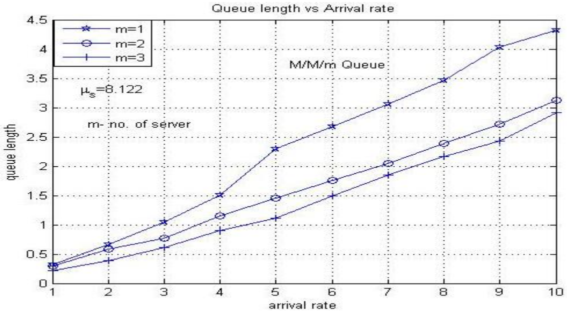

Fig 4.3: Plot for finding the queue length for internet server using M/M/m queuing model.

Figure 4.3 shows the curves for queue length versus arrival rate with different m values. Here, m represents the number of servers. In the above figure, there are three different curves for three different m values. Higher the value m, lower is the queue length, if the number of servers increases, the incoming arrivals will be waited in the queue, they will be shared for service with different servers. Then, the queue length decreases.

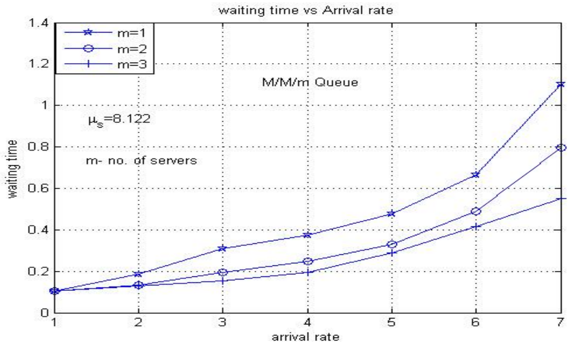

Fig 4.4: Plot for finding the waiting time for internet server using M/M/m queuing model.

Figure 4.4 shows the curves for waiting time versus arrival rate with different m values. Here, m represents the number of servers. In the above figure, there are three different curves for three different m values. Higher the value m, lower is the waiting time, if the number of servers increases, the incoming arrivals will be waited in the queue, they will be shared for service with different servers. Then, the waiting time decreases.

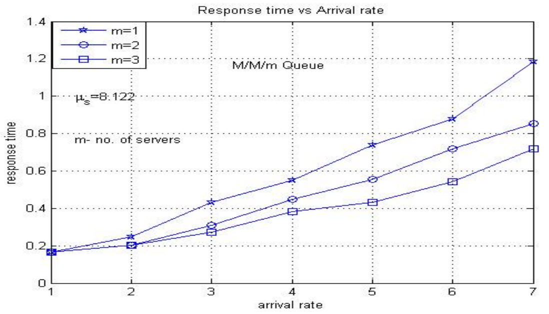

Fig 4.5: Plot for finding the response time for internet server using M/M/m queuing model.

Figure 4.5 shows the curves for response time versus arrival rate with different m values. Here, m represents the number of servers. In the above figure, there are three different curves for three different m values. Higher the value m, lower is the response time, if the number of servers increases, the incoming arrivals will be waited in the queue, they will be shared for service with different servers. Then, the response time decreases.

CONCLUSIONS

In this paper, performance measures for Internet server such as average queue length, average response times and average waiting times are derived and plotted by using M/M/m queuing model.

REFERENCES

[1] Dr. L.K. Singh, Riktesh Srivastava, “Estimation of Buffer Size of Internet Gateway Server via G/M/1 Queuing Model”, International Journal of Applied Science, Engineering and Technology, Volume 4, No.1, pp. 474-482, January 2007.

[2] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization And Performance Implications", IEEE/ACMTransactions on Networking, Vol. 5, No. 5, pp. 631-645, Oct. 1997.

[3] A. Iyengar, et al., "High-Performance Web Site Design Techniques", IEEE Internet Computing, Vol. 4, No. 2, pp. 17-26, 2000.

[4]. Kishor S.Trivedi, *Probability & Statistics with Reliability, Queuing, andComputer Science Applications*. Prentice-Hall of India Private Limited. New Delhi-110 001 2004.

[5] J.P.Buzzen, A Queueing network model of MVS, *ACM Computing surveys*, Sept 1978, Vol. 10, No 3, pp-319-331.

[6] Anderson, Darrell et. al.(1999), “A Case for Buffer Servers”, pp. 82-88, IEEE Seventh Workshop on Hot Topics in Operating Systems.

[7] Dimitri Bertsekas, Robert Gallager, *Data Networks*, Second Edition, Prentice Hall of India Private Limited, 1997.

[8] A. Iyengar, et al., "High-Performance Web Site Design Techniques", IEEE Internet Computing, Vol. 4, No. 2, pp. 17-26, 2000.

[9] D. Dias, et al., "A Scalable And Highly Available Web Server", in COMPCON '96. Technologies for the Information Superhighway Digest of Papers., San Jose, CA., pp. 85-92, 1996.

BIOGRAPHIES:

Raghunath Y. T. N. V, Assistant Professor, ECE dept. in DVR & Dr. HS MIC College of Technology, Vijayawada, Andhra Pradesh, India.

A.S.Sravani, P.G Scholor, ECE Dept. in St. Anns College of Engineering and Technology, Chirala, Andhra Pradesh, India