APPLICATION OF QUEUING THEORY FOR INTERNET SERVER

Win Lei Lei Aung
Faculty of Computing, University of Information Technology, Hlaing, Yangon
wllaung@gmail.com

ABSTRACT

This paper deals with the performance measurements of the single queue multiple server models by the method of queuing theory. This gives the performance measure for highly dynamic traffic conditions. This paper analyzes the highly dynamic conditions by the methods of Queuing theory, set up a corresponding mathematical model and compares the results with different servers. This is achieved by analyzing the performance measures in capacity planning of internet server using different queuing models by comparing the parameters like queue length, response time and waiting time for different links.

KEYWORDS

Queuing Model, Queue Length, Response Time, Utilization Factor & Waiting Time

1. INTRODUCTION

With the development of industry, Queuing Theory is one of popular method to solve various problems. In the early 20th century, queuing theory originated from the Danish engineer Erlang's study of telephone exchange efficiency of communication system. After the Second World War, especially with the rapid development of computer and communication technology, queuing theory gets attention and developed fast. It also became an important branch of operation research and its corresponding disciplines theory and reliability theory were developed.

In the mid-1930-s, queuing theory was recognized one important subject when W.Feller recommended birth and death process. In the early 1950s, D.G.Kendall researched queuing theory systematically by the methods of Markov chain and made it develop further. In the 1960, the projects studied complicatedly in queuing theory, it is so difficult to get the exact solution that people began to study the approximate method. At present, queuing system models have been widely used in all kind of management system. Such as production management, transportation, banking, medical services, computer design and performance evaluation and so on.

In [1], this paper deals with an improved scheme for autonomous performance of gateway servers under highly dynamic traffic loads. The most widely 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 orders 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 can be used at popular sites to improve performance and availability.

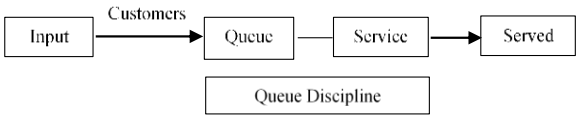

Figure 1. Basic Structure of Queuing Model

In Figure 1, the client requests for accessing the web page through different links those are link 1 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 2. 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].

2. BASIC KNOWLEDGE

The Queuing theory is the mathematical theory and method of queuing system (stochastic system). In daily life, people will encounter all sorts of queuing problems, such as, standing at bus stops, going to hospital, and going to the ticket office to buy the tickets and so on. In these problems, bus and passengers, doctor and patients, conductor and the buyers forms a queuing system or service system respectively; the former can be regarded as service agencies and the latter can be regarded as customers. The queue can be tangible queue, may also be intangible queue. For example, several passengers make telephone call to order train tickets at the same time, if a passenger is on the phone, can only wait for the other passengers, this form of queue is invisible. The people or some objects can be the queue, such as semi-finished products for processing in the production line, machine waiting for maintenance, and the information waiting for computing center to process, etc. Queuing theory consists of three parts: input process, queuing rules and service agencies. The schematic diagram as follows:

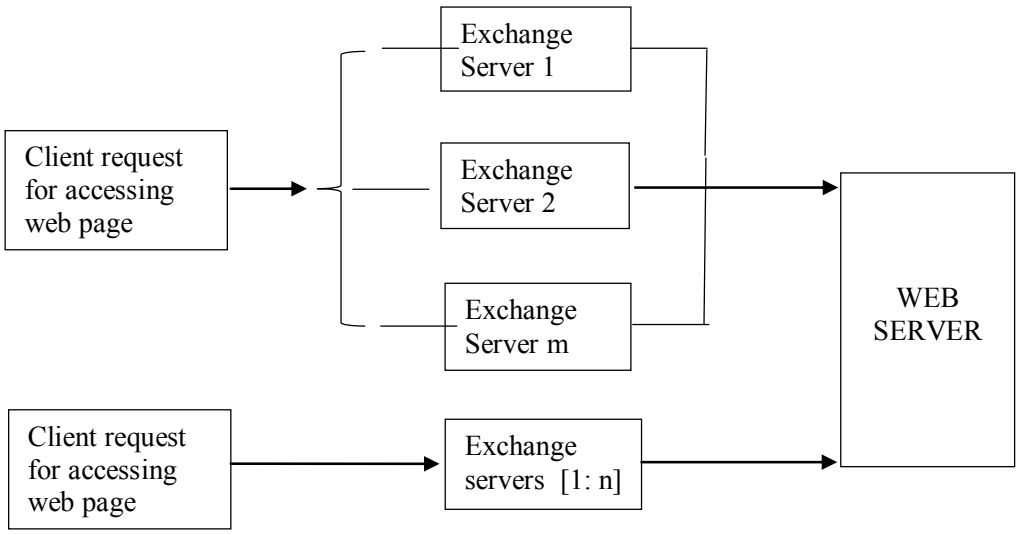

Figure 2. Layout of Internet Server

The objective of the paper is getting the performance measures and capacity estimation of the above system by using different queuing models i.e., M/M/1, M/G/1, M/D/1 and M/E_k/1. The above system is modeled in queuing network by using the queuing models. A Queue System analyzes a full queuing situation involving 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), where 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
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 gives the fraction of time that the server is busy. It is defined as ratio of arrival rate to service rate. Queue length defines the maximum capacity of the queue or number of customers in the queue. Waiting time 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 defines the sum of waiting time and service time for a particular customer.

3. MODELING

Queuing theory can be divided into single channel queuing system and multi-channel queuing system. This paper mainly researches the performance index under the steady state.

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

The M/M/1-Queue has inter-arrival 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.

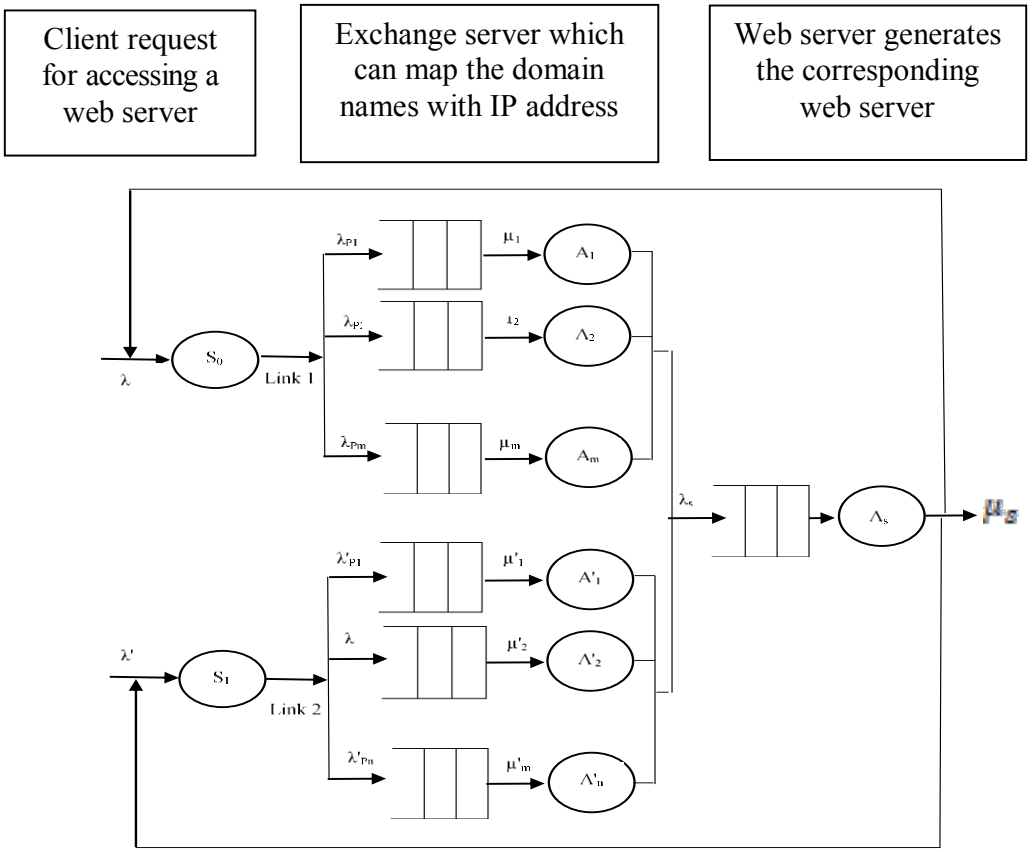

Figure 3. Block diagram for queuing server using M/M/1 system

In Figure.3, S_0 and S_1 are source indicators for different links (satellite/optical), A_1 to A_m and A'_1 to A'_n are exchange servers. A_s is also web server (a web server serves pages for viewing in a web browser) in Figure.3.

3.2. Single queue multiple server model

3.2.1. M/M/m queuing model

In this M/M/m queuing model, the 1st M indicates the inter-arrival time distribution arrivals follows Poisson distribution with parameter λ, and 2nd M indicates service time distribution [5]. Here, it is exponential distribution with parameter μ and 3rd m indicates number of servers available those are in parallel. The system has multiple servers and uses the FIFO service discipline. The waiting line is of infinite size.

In the case M/M/1 queuing 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 queuing model, there are m numbers of servers that are connected in parallel [5]. That means, this model can process m requests at a time. So compared with the M/M/1 queuing model, this model gives better performance measures.

In Figure 3, the M/M/m queuing model comprises m internet servers. In this system, the client requests for accessing the web page are 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 generate the web page which is displayed in the client's monitor. The performance measures those can be done by using this queuing model are utilization, queue length, waiting time and response time [5].

3.2.2. Queuing network model for Internet Server using M/M/m model

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

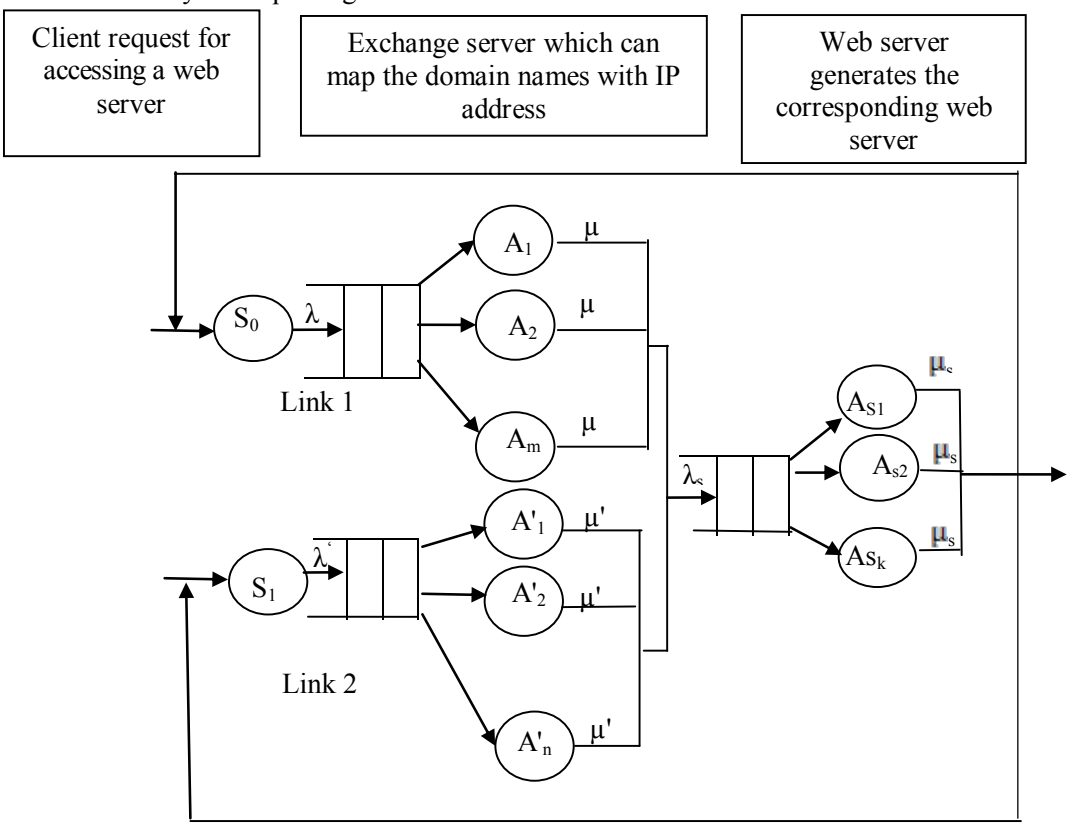

Figure 4.Block diagram for queuing network model for internet server using M/M/m model

In Figure 4, there are two links which are connected to the internet server to provide 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…, A_m be the m parallel exchange servers in link 1 and let A_1', A'_2 …, A'_3 be the n parallel exchange servers and A_{s1}, A_{s2} …, A_{sk} be the k parallel internet servers. Let λ be the arrival at link 1 and λ' is 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 link 1 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.

3.2.3. 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, the utilization factor can be expressed as follows:

For link 1,

$$\rho = \frac{\lambda}{m\mu} + \frac{\lambda}{k\mu_s} \quad , \qquad (1)$$

For link 2,

$$\rho = \frac{\lambda}{n\mu} + \frac{\lambda}{k\mu} \quad , \qquad (2)$$

For internet web server,

$$\rho = \frac{\lambda}{k\mu} \qquad (3)$$

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

According to Figure 4, the total queue length can be expressed as follows:

For link 1,

$$\text{The total queue length} = L'_q + L_q \qquad (4)$$

For link 2,

$$\text{The total queue length} = L''_q + L_q \qquad (5)$$

Where for link 1,

$$L'_q = \frac{\lambda}{m\mu}\left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}\left[\sum_{n=0}^{m-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^m}{m!} * \left(1-\frac{\lambda}{m\mu}\right)^{-1}\right]^{-1}\left(1-\frac{\lambda}{m\mu}\right)^{-2} \qquad (6)$$

For link 2,

$$L''_q = \frac{\lambda'}{m\mu}\left(\frac{\lambda'}{\mu}\right)^m \frac{1}{m!}\left[\sum_{n=0}^{m-1}\frac{\left(\frac{\lambda'}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda'}{\mu}\right)^m}{m!} * \left(1-\frac{\lambda'}{m\mu}\right)^{-1}\right]^{-1}\left(1-\frac{\lambda'}{m\mu}\right)^{-2} \qquad (7)$$

For internet web server,

$$L_q = \frac{\lambda_s}{m\mu}\left(\frac{\lambda}{\mu}\right)^m \frac{\lambda_s}{m!}\left[\sum_{n=0}^{m-1}\frac{\left(\frac{\lambda_s}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda_s}{\mu}\right)^m}{m!} * \left(1-\frac{\lambda_s}{m\mu}\right)^{-1}\right]^{-1}\left(1-\frac{\lambda_s}{m\mu}\right)^{-2} \qquad (8)$$

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, the total waiting time can be expressed as follows:
For link 1,
The total waiting time = $W_q' + W_q$ (9)

For link 2,
The total waiting time= $W_q'' + W_q$ (10)
Where, for link 1,

$$W_q' = \frac{L_q'}{\lambda} \quad (11)$$

For link 2, $W_q'' = \frac{L_q''}{\lambda}$ (12)

For internet web server,

$$W_q = \frac{L_q}{\lambda_s} \quad (13)$$

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

According to Figure 4, the total response time can be expressed as follows:

For link 1, the total response time = $W' + W$ (14)

For link 2, the total response time = $W'' + W$ (15)

Where, for link 1,

$$W'' = W_q'' + \frac{1}{\mu'} \quad (16)$$

For link 2,

$$W'' = W_q'' + \frac{1}{\mu'} \quad (17)$$

For internet web server,

$$W = W_q + \frac{1}{\mu_s} \quad (18)$$

4. NUMERICAL RESULTS

Assume the service rate, $\mu = \frac{40}{3}$. Table I to Table IV show the numerical results for queuing model.

Table I: The Utilization factor for internet server using M/M/m queueing model

Λ	Utilization Factor					
	1	2	3	4	5	6
M/M/1	0.7500	0.1500	0.2250	0.3000	0.3750	0.4500
M/M/2	0.0375	0.0750	0.1125	0.1500	0.1875	0.2250
M/M/3	0.0250	0.0500	0.0750	0.1000	0.1250	0.1500

Table II: The queue length for internet server using M/M/m queuing model

λ	The total queue length					
	1	2	3	4	5	6
M/M/1	0.0811	0.1765	0.2903	0.4286	0.6000	0.8182
M/M/2	0.0011	0.0008	0.0029	0.0069	0.0137	0.0240
M/M/3	0.0000	0.0000	0.0001	0.0004	0.0009	0.0020

Table III: The waiting time for internet server using M/M/m queuing model

λ	The total waiting time					
	1	2	3	4	5	6
M/M/1	0.0811	0.0883	0.0968	0.1072	0.1200	0.1364
M/M/2	0.0011	0.0004	0.0010	0.0017	0.0027	0.0048
M/M/3	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003

Table IV: The response time for internet web server using M/M/m queuing model

λ	The response time					
	1	2	3	4	5	6
M/M/1	0.1561	0.1633	0.1718	0.1822	0.1950	0.2114
M/M/2	0.0751	0.0754	0.0760	0.0767	0.0777	0.0788
M/M/3	0.0750	0.0750	0.0750	0.0751	0.0752	0.0753

Table I shows for utilization versus arrival rate with different m values. Table II shows for queue length versus arrival rate with different m values. Table III shows for waiting time versus arrival rate with different m values. Table IV shows for response time versus arrival rate with different m values. Here, m represents the number of internet web servers. In the above Tables, there are three different values for three different m values.

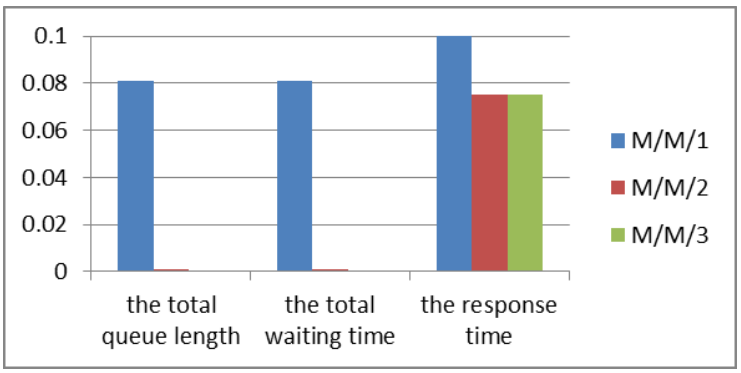

Figure 5.the total queue length, the total waiting time and the response time using M/M/m queuing model

Higher the value m, lower is the utilization factor because of 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. In Figure 5, assume the arrival rate is one unit per time, we can see that the changes in the total queue length, the total waiting time and the response time with the increasing value m. Higher the value m, lower is the queue length as shown in Figure 5, 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. 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. 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.

5. CONCLUSION

In this paper, performance measures for Internet server such as average queue length, average response times and average waiting times are derived and data in Tables by using M/M/m queuing model. If the number of server increases, the utilization factor for each internet server and the total queue length decrease. Higher the value m, lower are the response time and the waiting time for web server. The service indicators of system are in decline with the increase of the internet web server. If we design the internet web server, we consider the measurements of M/M/m queuing model from the above tables. Three server queuing model is better than two sever. In this service rate, three server queuing models are the best because of the least of performance measures. The increase of the web server has a positive impact on internet server system. From this paper, the theoretical data is consistent with the reality.

ACKNOWLEDGEMENT

Firstly, I would like to thank Dr. Pyke Tin who shares ideas and helpful suggestion. I am also grateful to Dr. Swe Swe Kyaw who motivates me to do this. Finally, I appreciate to my mother for her patient, understanding and encouragement during my work that has to successful finish.

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/ACM Transactions 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 and Computer Science Applications. Prentice Hall of India, Private Limited, New Delhi-110 001 2004.

[5] Raj Jian, “The Art Of Computer Systems Performance Analysis”, John Wily & Sons Inc, pp 592-604, 1992 computer press award winner.

[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 Paper, San Jose, CA. , pp. 85-92.

BIOGRAPHY

I am Dr. Win Lei Lei Aung from University of Information Technology. I got BSc (Maths.Hons :) from Yangon University in 2002 and had Master of Science in 2004 from Yangon Technological University. After that I have Ph.D. degree with Applied Mathematics in 2007. I am a professionally qualified teacher with seventeen year experiences. I am Assistant Professor at Faculty of Computing.