Mean web object size for multiple user access in M/G/1/PS system

Yong-Jin Lee

Department of Technology Education, Korea National University of Education
250 Taesungtapyon-Ro, Heungduk-Gu, Cheongju, 28173, Korea
lyj@knue.ac.kr

Abstract - *This paper aims at estimating mean object size which affects to web service time in multiple access environments. Arrival pattern of packets arrive is described by Poisson distribution and web service time has general distribution. CPU scheduling policy in web server is based on processor (time) sharing. In order to describe the behavior of the web server system, we consider M/G/1/PS model. We derive mean web object size satisfying the constraint such that mean waiting time by round-robin scheduling in multiple access environments is equal to mean waiting delay of M/G/1/PS system. Given the system utilization and maximum segment size, we can find mean object size by varying the number of simultaneous access users. Performance evaluation shows that mean web object size increases as the system utilization increases at the given maximum segment size (MSS), but converges on the same lower bound as the number of users increase beyond any threshold. Our results can be applied to the economic web service design.*

Keywords: Mean object size, M/G/1/PS, Mean waiting delay, simultaneous user access

1 Introduction

Mean waiting delay is one of important quality of service (QoS) measures when multiple users access a web server simultaneously. This measure is affected by mean object size imbedded in the HTML file. In order to meet mean waiting delay to the end-user's defined threshold, we should first estimate mean object size. Controlling web object size leads to the least maintenance cost in satisfying the user's QoS.

End-user requests an object in the web server according to the Poisson distribution and the web service time is described by general distribution. Therefore, in the depiction of mean waiting delay or mean response time in the web system, M/G/1 model [1] is mainly used. Shi et al [2], Khayari et. al [3] and Riska et. al [4] have presented that Weibull distribution, Exponential, and Hyper-exponential distribution are suitable to describe the web service. However, more empirical researches for web services distribution related to the Internet are still needed.

Scheduling is an extremely important topic in computer and communication system. The right scheduling policy reduces mean waiting delay and mean response time remarkably without additional costs. Scheduling policies is classified into non-preemptive and preemptive. FCFS (First Come First Served), RANDOM, and LCFS (Last Come First Served) are examples of preemptive scheduling policy. PS (Processor Sharing) and PLCFS (Preemptive Last Come First Served) are examples of non preemptive scheduling policy.

Previous researches for the M/G/1 model have mainly considered the preemptive scheduling policy, especially FCFS which does not make use of object (job) size. However, since web service is affected by web object size in the multiple user environments, we should consider not only preemptive scheduling policy but also preemptive scheduling policy.

When several users simultaneously request an object in the web server, and the packed based round-robin (RR) scheduling for the web service is used, we can compute mean waiting delay by using the deterministic model composed of the number of users and object size. In the steady state, we can infer that mean waiting delay in the deterministic model is equal to mean waiting delay in the M/G/1/PS model. Therefore, we can find out mean object size to satisfying mean waiting delay which end-user can tolerate. Previous models find mean waiting delay for M/G/1/FCFS system [5-10].

In order to solve the problem, we first find mean waiting delay represented in terms of the number of users and the number of packets included in the object when multiple users access to web server simultaneously. And then we find mean waiting delay in terms of job size in M/G/1/PS system. Finally, we find mean web object size in the steady state. In this paper, the reason to find the web object size satisfying the end-user's delay requirement is why the controlling object size is the least cost method in designing web system.

The rest of this paper is organized as follows. In the next section, we first describe the deterministic model to find the mean waiting delay by round-robin scheduling in the multiple users access environment. And then we discuss mean waiting delay in M/G/1/PS system. We determine mean object size satisfying the constraint that the mean waiting delay in the deterministic model is equal to the mean waiting delay in M/G/1/PS system. In section 3, we present and analyze the performance evaluation results. Finally, in section 4, we discuss conclusions and future research.

2 Mean waiting delay for object transfer

We first describe mean waiting delay for web object transfer in the deterministic model. In the web service, m concurrent users generally require a same object such as index.html on a web server simultaneously. A web object is segmented into several packets with a maximum segment size (MSS) in a transport layer. Let denote θ the web object size. Let denote mss the MSS. Then the number of packets (n) is given by $n = \theta/mss$.

When multiple clients request a same object, each client thinks that his service time is the same as others. However, because the number of processors is less than the number of clients, each user's completion time is different according to scheduling policy. In most of operating systems, processor sharing such as round robin is mostly used as scheduling policy.

We assume the time quantum (τ) in RR scheduling policy is based on the packet service time. When a client requests an object from the server, n packets are included in the object. Job (object) size (x) represents total service time that each client expects. Since the time quantum (τ) is equal to the packet service time, thus $\tau = x/n$. Figure 1 depicts RR service in the multiple users (clients) access environment [7,10].

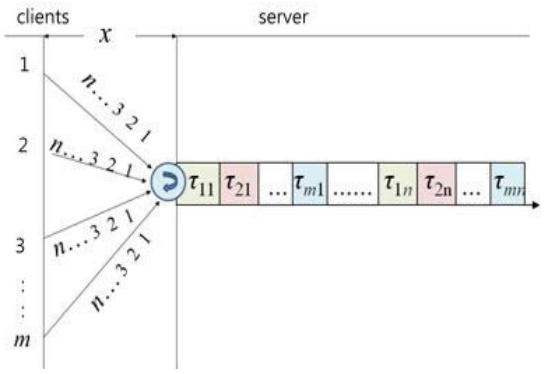

Fig. 1. Web service using RR policy based on the packet service time (τ) for multiple users (m)

In Figure 1, τ_{ij} represents j^{th} packet service time of the i^{th} user. Assuming $\tau_{ij} = \tau$ ($\forall$ i, j), mean waiting delay in the deterministic model (W_D) is given by (1).

$$\begin{aligned} W_D &= \frac{1}{m}\sum_{i=1}^{m}\left[(m-i)\tau + m(m-1)(n-1)\tau\right] \\ &= \frac{(m-1)(2n-1)\tau}{2} \\ &= \frac{(m-1)(2n-1)x}{2\theta} \times mss \end{aligned} \tag{1}$$

Next we consider mean waiting delay in M/G/1/PS system. As an example of PS(Process-Sharing) in computer system, a time-sharing CPU rotates in round-robin order between m jobs in the system, giving the first job one quantum, then the second job one quantum,…, then mth job one quantum, and then returning back to the first job to repeat. If we think of the quantum size as approaching 0, we get PS [11].

A server with service rate μ operates under Processor-Sharing (PS) service order if, whenever there are m jobs at the server, each of the job is processed at rate μ/m[11]. In Figure 1, if we regard τ as μ/m, mean waiting delay for deterministic model becomes that for M/G/1/PS system.

Mean response time (E[$T(x)$]) in M/G/1/PS system with job size (x) is given by

$$E[T(x)] = x + W_Q(x) \tag{2}$$

Here, $W_Q(x)$ is mean waiting delay in the system, and is equal to E[wasted time (x)][11].

$$\begin{aligned} W_Q(x) &= E[\text{wasted time}(x)] \\ &= E[\text{the number of times tagged job is interupted}] \\ &\quad \times E[\text{length of interupt}] \\ &= \frac{\lambda E(S)}{1-\rho} \\ &= \frac{\lambda x}{\mu(1-\rho)} \\ &= \frac{\rho x}{1-\rho} \end{aligned} \tag{3}$$

In equation (3), λ and μ are average arrival rate and average service rate, respectively. E(S) represents mean service time which means the average time required to service a job on the CPU. ρ (λ/μ) is the system utilization ($0 \le \rho < 1$).

We can infer that mean waiting delay for the deterministic model with RR policy and M/G/1/PS model in

the steady state, By letting $W_D = W_Q(x)$ in Equation (4), we can find mean object size (θ: bytes) in the steady state.

$$W_D = W_Q(x) \rightarrow \frac{(m-1)(2n-1)x}{2\theta} \times mss = \frac{\rho x}{1-\rho} \quad (4)$$
$$\rightarrow \theta = \frac{(1-\rho)(m-1) \times mss}{2[(1-\rho)(m-1)-\rho]}$$

In equation (4), since $(1 - \rho)(m - 1) - \rho$ should be positive, the number of users (m) is given by

$$m > 1 + \frac{\rho}{1-\rho} \quad (5)$$

Table 1 shows the minimum number satisfying equation (5) when we vary the system utilization (ρ).

TABLE I

MAXIMUM NUMBER OF USERS (M) SATISFYING EQUATION (5)

Utilization(ρ)	The number of users (m)
0.1 0.2 0.3 0.4	2
0.5 0.6	3
0.7	4
0.8	6
0.9	11

3 Performance evaluation

We first compute mean object size when *mss* = 1460B for various ρ. Figure 2 shows mean object size according to varying utilization (ρ). The minimum number of users (m) is used in Table 1. For example, when $\rho = 0.1$, minimum number of users (m) is two in Table 1. In equation (5), we set m (=2), *mss* (=1460), and ρ (= 0.1), we obtain mean object size, θ (=821B).

In Figure 2, as ρ increases, mean object size increases although MSS is fixed at 1460B. However, mean object size decreases at ρ=0.5 compared with at ρ=0.4. The reason is why the denominator of equation (5) fluctuates by ρ.

Now, we compute mean web object size (θ) satisfying W_D=$W_Q(x)$ for varying m when the MSS and ρ are given. Figure 3 depicts mean object size varying the number of users (m) when MSS is given by 1460B and several utilization (ρ) are given. When m is less than 100, mean object size becomes larger as ρ becomes larger. However, when m is larger than 100, mean object size converges to the *mss*/2 regardless of ρ.

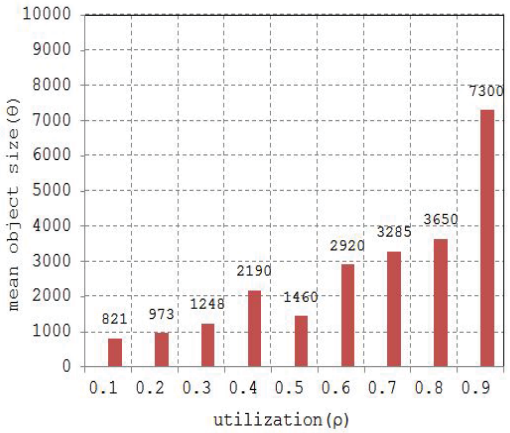

Fig. 2. Mean object size given *mss*=1460B and minimum number of users (m) varying utilization (ρ).

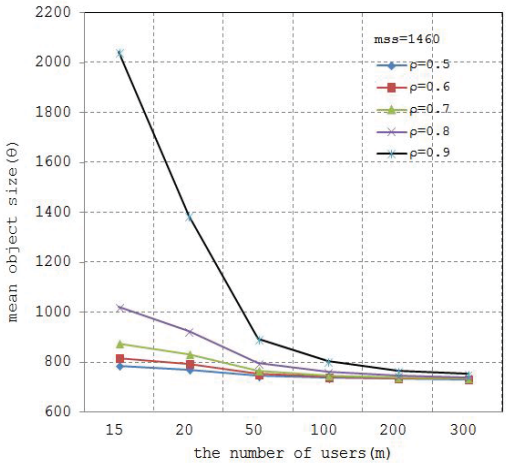

Fig. 3. Mean object size given *mss*=1460B and utilization (ρ) varying the number of users (m).

Figure 4 shows mean object size varying the number of users (m) when ρ is given by 0.5 and several MSS (*mss*) are given. Mean object size becomes larger as m and *mss* becomes larger. However, mean object size is nearly the same at the given ρ and *mss* regardless of m.

Figure 5 shows mean object size varying the system utilization (ρ) when *mss* is given by 1460B and several numbers of users (m) are given. Mean object size becomes larger as ρ becomes larger. However, mean object size is nearly the same when ρ is less than 0.5.

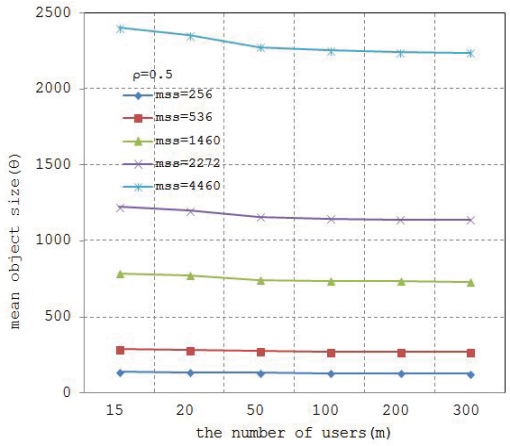

Fig. 4. Mean object size given ρ=0.5 and *mss* varying the number of users (m).

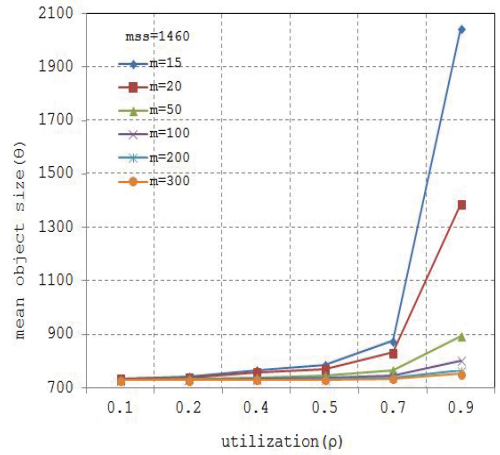

Fig. 5. Mean object size given *mss*=1460B and the number of users (m) varying utilization (ρ).

4 Conclusions

This paper presents the mean object size estimation method satisfying that mean waiting delay for the deterministic model is equal to mean waiting delay for M/G/1/PS system in the steady state. By simple inference, we can find out mean object size satisfying mean waiting delay that end-user can allow as quality of service. This mean waiting object size can be used to control the web service economically. Some computational experiences show that mean object size converges to the some constant value as the number of users become larger regardless of the system utilization factor. Future works include more exact model to describe web service behaviour and the comparison of M/G/1/PS and M/G/1/FCFS in multiple access environments.

5 References

[1] S. Ross, Introduction to probability model, Academic Press, New York, 2010.

[2] W. Shi, E. Collins, and V. Karamcheti, "Modeling Object Characteristics of Dynamic Web Content," Journal of Parallel and Distributed Computing, vol. 63, no. 10, 2003.

[3] R. Khayari, R. Sadre and B. R. Haverkort, "Fitting world-wide web request traces with the EM-algorithm, Performance Evaluation," vol. 52, no. 2, 2003.

[4] Riska, V. Diev and E. Smirni, "Efficient fitting of long-tailed data sets into hyper-exponential distributions," Proc. of IEEE Global Telecommunications Conference (GLOBECOM 2002), vol. 3, 2513-2517, 2002.

[5] Y. Lee, "Mean waiting delay for web service perceived by end-user in multiple access environment," Natural Science, Natural Science Institute of KNUE, vol. 2, 55-58, 2012.

[6] Y. Lee, "Web Object Size satisfying M/D/1 Queueing Delay in Multiple Access Web Service," Natural Science, Natural Science Institute of KNUE, vol. 3, 1-7, 2013.

[7] Y. Lee, "Web Object Size satisfying Mean Waiting Time in Multiple Access Environment," International J. of Computer Networks & Communications, vol. 6, no. 4, 2014.

[8] Y. Lee, "Maximum Web Object Size satisfying M/G/1 Queueing Delay Constraint in Multiple User Access Environment," The Journal of Korean Institute of Information Technology", vol. 12, no. 6, 2014.

[9] Y. Lee, "Novel Quality of Service Measure for Web Transaction in Multiple User Access Environments," International J. of App. Eng. Res., vol. 10, no. 16, 2015.

[10] Y. Lee, "Mean waiting time of an end-user in the multiple web access environment," Proc. of the Sixth International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ-2013), 2013.

[11] M. Harchol-Balter, Performance Modeling and Design of Computer Systems, Cambridge University Press, USA, 2013.