International Journal of Advanced Smart Convergence Vol.11 No.3 1-6 (2022)
http://dx.doi.org/10.7236/IJASC.2022.11.3.1

IJASC 22-3-1

Estimating the mean number of objects in $M/H_2/1$ model for web service

Yongjin Lee

Professor, Dept. of Technology Education, Korea National University of Education, Korea
lyj@knue.ac.kr

Abstract

In this paper, we estimate the mean number of objects in the $M/H_2/1$ model for web service when the mean object size in the $M/H_2/1$ model is equal to that of the M/G/1/PS and M/BP/1 models. To this end, we use the mean object size obtained by assuming that the mean latency of deterministic model is equal to that of $M/H_2/1$, M/G/1/PS, and M/BP/1 models, respectively. Computational experiments show that if the shape parameter of the M/BP/1 model is 1.1 and the system load is greater than 0.35, the mean number of objects in the $M/H_2/1$ model when mean object size of $M/H_2/1$ model is the same as that of M/G/1/PS model is almost equal to the mean number of objects in the $M/H_2/1$ model when the mean object size of $M/H_2/1$ model is the same as that of M/BP/1 model. In addition, as the upper limit of the M/BP/1 model increases, the number of objects in the $M/H_2/1$ model converges to one, which increases latency. These results mean that it is efficient to use small-sized objects in the web service environment.

Keywords: *mean number of objects, deterministic model, $M/H_2/1$ model, M/G/1/PS model, M/BP/1 model, web service*

1. Introduction

The object size and the number of objects are important parameters in the design of web service. These parameters depend on the mean waiting latency in the used web service model. They are also affected by the number of concurrent users accessing the web server. The M/G/1 model [1] is used as the web service model and the probability distribution of the service includes exponential, hyper-exponential, and Weibull [2,3]. FIFO is mostly used as a service policy, but process sharing (PS) is also commonly used web service policy. This case is called an M/G/1/PS model [4,5,6].

The object size affects the mean waiting latency in the web service model and for this case, the M/BP/1 model [6,7,8] is considered. Here, BP means the Bound Pareto distribution, which has the upper bound and lower bound parameters of the file size and shaping parameter.

The number of objects should be also found in the web design. In general, web objects are classified into a static object and several dynamic objects. $M/H_2/1$ model describes this feature [9,10].

When multiple users access objects on a web server simultaneously by round-robin, we can find the mean waiting latency by the deterministic model [10]. By inferring that at steady state, mean waiting latency in the deterministic model is equal to that of $M/H_2/1$, M/G/1/PS, and M/BP/1 models respectively, mean object size with the same mean waiting latency can be found. By inference from the same logic that mean object size for $M/H_2/1$ model is equal to that for M/G/1/PS and M/BP/1 models, we find the mean number of objects.

This paper is based on the related research [4,6,8,10] and finds the mean number of objects for $M/H_2/1$ model in web service environment. Numerical experiments show that mean number of objects when the object

Manuscript Received: September. 4, 2022 / Revised: September. 18, 2022 / Accepted: September. 19, 2022
Corresponding Author: lyj@knue.ac.kr
Tel: +82-43-230-3774, Fax: +82-43-230-3610
Professor, Dept. of Technology Education, Korea National University of Education, Korea

size of M/H_2/1 model is equal to that of M/G/1/PS model is about the same as mean number of objects when the object size of M/H_2/1 model is equal to that of M/BP/1 model.

The rest of this paper is organized as follows. Section 2 describes mean waiting latency and mean object size in the related web service models. Section 3 derives the mean number of objects for M/H_2/1 model. Finally, section 4 presents the conclusion.

2. Related works

In this section, we describe mean waiting latency and mean object size in the deterministic model, M/H_2/1 model, M/BP/1 model, and M/G/1/PS model used in the related papers [4, 6, 8, and 10].

The deterministic model describes mean waiting latency in the web object transfer [10]. In most object transfer service, m concurrent users access the same object on the web server simultaneously. In the transport layer, an object is divided into several packets with the maximum segment size (MSS). If we set S and θ to the maximum segment size and the object size respectively, the number of packets (n) is equal to θ/S.

In the web service environment, service completion time of each user is affected by the scheduling policy. Most operating systems uses a round-robin scheduling policy.

We can assume that a packet service time is equal to the time quantum in the round-robin scheduling policy. When a user access an object on the server, the object contains n packets. The task size (x) is equal to the total service time of each user. Since the time quantum is the packet service time (γ), thus $\gamma = x/n$. If γ_{ij} is the j^{th} packet service time of the i^{th} user and $\gamma_{ij} = \gamma$ for all i, j, we determine mean waiting latency in the deterministic model (E(W_D)) [4].

$$E(W_D) = \frac{(m-1)(2n-1)E(X)S}{2\theta} \tag{1}$$

Now, we consider M/H_2/1 model for web service [10]. Web objects typically consist of two types. A static object is the first requested home page. N dynamic objects are embedded in the static object and requested after the home page is parsed. This situation can be represented by the Hyper-exponential distribution [10], which chooses the i^{th} negative exponential distribution. The density function is given by

$$f(x) = \sum_{i=1}^{2} p_i \lambda_i \, e^{-\lambda_i x} \qquad x \geq 0 \tag{2}$$

For a static object, access probability (p_1) is 1/(N+1) and arrival rate (λ_1) is λ. For N dynamic objects, access probability (p_2) is N/(N+1) and arrival rate (λ_2) is $N\lambda$.

E(X) and E(X^2) represents the first and the second moment for the object service time, respectively. They are obtained by

$$E(X) = \frac{2}{(N+1)\lambda} \qquad E(X^2) = \frac{2}{N\lambda^2} \tag{3}$$

Using M/G/1 queueing theory [1, 10], mean waiting latency in M/H_2/1 model is given by

$$W_H = \frac{N+1}{\lambda(N-1)N} \tag{4}$$

By letting E(W_D) = E(W_H), we find mean object size of M/H_2/1 model ($\theta_{M/H2/1}$) [10].

$$\theta_{M/H_2/1} = \frac{(N-1)N(m-1)S}{2(N-1)N(m-1)-(N+1)^2} \quad (5)$$

$$\text{where} \quad m > 1+\frac{(N+1)^2}{2N(N-1)}$$

M/G/1/PS model utilizes processor sharing service. If we set the service rate to μ and there are m jobs in the server, each job can be managed at a rate of μ/m. Because γ is equal to μ/m, mean waiting latency in the deterministic model can be considered as mean waiting latency in M/G/1/PS model ($W_Q(x)$). $W_Q(x)$ is given by [4,6].

$$W_Q(\mathrm{x}) = \frac{\lambda E(Y)}{1-\rho} = \frac{\lambda x}{\mu(1-\rho)} = \frac{\rho x}{1-\rho} \quad (6)$$

In Eq. (6), λ is mean arrival rate and ρ ($0 \le \rho < 1$) is the system load and equal to λ/μ. E(Y) is mean service time of job. We can assume that mean waiting latency in the deterministic model using round-robin policy and that in M/G/1/PS model become the same at the steady state. By setting E(W_D) = E($W_Q(x)$), we find mean object size of M/G/1/PS model ($\theta_{\mathrm{M/G/1/PS}}$) [4,6].

$$\theta_{M/G/1/PS} = \frac{(m-1)(1-\rho)S}{2[(m-1)(1-\rho)-\rho]} \quad (7)$$

$$\text{where} \quad m > 1+\frac{\rho}{1-\rho}$$

Now, we describe mean waiting latency in M/BP/1 model [6, 8]. We set E(X) and E(X^2) to first and second moment for service time distribution, respectively. If the link capacity is C, we obtain E(X) and E(X^2) by Eq. (8). Here, α is the shape parameter. L and U are file size's lower bound and upper bound, respectively. $E_x(x)$ and $E_x(x^2)$ represent mean and second moment of Bounded Pareto distribution.

$$\begin{cases} E(X) = \frac{E_x(X)}{C} = \frac{L^\alpha}{C\left(1-\left(\frac{L}{U}\right)^\alpha\right)}\left(\frac{\alpha}{\alpha-1}\right)\left(\frac{1}{L^{\alpha-1}}-\frac{1}{U^{\alpha-1}}\right) \\ E(X^2) = \frac{E_x(X^2)}{C^2} = \frac{L^\alpha}{C^2\left(1-\left(\frac{L}{U}\right)^\alpha\right)}\left(\frac{\alpha}{\alpha-2}\right)\left(\frac{1}{L^{\alpha-2}}-\frac{1}{U^{\alpha-2}}\right) \end{cases} \quad (8)$$

In an M/G/1 model, if λ, ρ and X are the arrival rate, the system load, and the service time respectively, mean waiting latency in M/BP/1 model is given by

$$E(W_{BP}) = \frac{\lambda E(X^2)}{2(1-\rho)} \quad (9)$$

Now, we may infer that at the steady state, mean waiting latency in the deterministic model (E(W_D)) becomes mean waiting latency in the M/BP/1 model (E(W_{BP})).

We find mean object size of M/BP/1 model ($\theta_{\mathrm{M/BP/1}}$) by using $n = \theta/S$ when E(W_{D}) = E(W_{BP}). Since Eq. (1) and Eq. (9) are the same, we find $\theta_{\mathrm{M/BP/1}}$ as the following [6,8].

$$\theta_{M/BP/1} = \frac{(m-1)E(X)(1-\rho)S}{2(m-1)E(X)(1-\rho)-\lambda E(X^2)} \quad (10)$$

$$\text{where} \quad m > 1+\frac{\lambda E(X^2)}{2(1-\rho)E(X)}$$

3. Mean number of objects in M/H$_2$/1 model

This section finds mean number of objects in M/H$_2$/1 model by using the mean object sizes of M/H$_2$/1 model, M/G/1/PS model and M/BP/1 model.

3.1. Mean number of objects when $\theta_{M/H2/1} = \theta_{M/G/1/PS}$

Assuming that $\theta_{M/H2/1}$ in Eq. (5) is equal to $\theta_{M/G/1/PS}$ in Eq. (7),

$$\theta_{M/H_2/1} = \theta_{M/G/1/PS} \rightarrow \frac{(N-1)N(m-1)S}{2(N-1)N(m-1)-(N+1)^2} = \frac{(m-1)(1-\rho)S}{2[(m-1)(1-\rho)-\rho]} \tag{11}$$

Rewriting Eq. (11) for N, we obtain the following equation.

$$(1-3\rho)N^2 + 2N + 1 - \rho = 0 \tag{12}$$

By the quadratic formula, N is given by

$$N = \frac{-1 \pm \sqrt{1-(1-3\rho)(1-\rho)}}{1-3\rho}, \quad N > 0 \tag{13}$$

3.2. Mean number of objects when $\theta_{M/H2/1} = \theta_{M/BP/1}$

Assuming that $\theta_{M/H2/1}$ in Eq. (5) is equal to $\theta_{M/BP/1}$ in Eq. (10),

$$\theta_{M/H_2/1} = \theta_{M/BP/1} \rightarrow \frac{(N-1)N(m-1)S}{2(N-1)N(m-1)-(N+1)^2} = \frac{(m-1)E(X)(1-\rho)S}{2(m-1)E(X)(1-\rho)-\lambda E(X^2)} \tag{14}$$

To summarize Eq. (14) for N, we have a following quadratic equation.

$$aN^2 + bN + \mathrm{c} = 0 \tag{15}$$

$$\begin{aligned} where\ a &= (1-\rho)E(x) - \lambda E(X^2) \\ b &= 2(1-\rho)E(x) + \lambda E(X^2) \\ c &= (1-\rho)E(x) \end{aligned}$$

Using the quadratic formula, we find N. Since N is the mean number of objects, it must be positive.

$$N = \frac{-b \pm \sqrt{b^2-4ac}}{2a}, \quad N > 0 \tag{16}$$

3.3. Mean number of objects analysis for M/H$_2$/1 model

We investigate the mean number of objects when the object size of M/H$_2$/1 model is the same as that of M/G/1/PS model. In addition, when the object size of M/H$_2$/1 model and that of M/BP/1 model are the same, we compare the mean number of objects in M/H$_2$/1.

Table 1 shows the comparison results when the shaping parameter (α) are 0.3, 0.7, 1.1, and 1.5 respectively. Lower bound (L) and upper bound (U) are 50KB and 1MB respectively for varying system load (ρ) from 0.35 to 0.9.

The mean number of objects is the largest when α is 1.5, and the smallest when α is 1.1. Especially, when α is equal to 1.1 and system load (ρ) is greater than 0.35, the number of objects when $\theta_{M/H2/1} = \theta_{M/BP/1}$ is almost equal to the number of objects when $\theta_{M/H2/1} = \theta_{M/G/1/PS}$.

Table 1. The number of objects (*N*) comparison varying *α*

when $\theta_{M/H2/1} = \theta_{M/G/1/PS}$ and $\theta_{M/H2/1} = \theta_{M/BP/1}$

ρ	*N when* $\theta_{M/H2/1} = \theta_{M/G/1/PS}$	*N when* $\theta_{M/H2/1} = \theta_{M/BP/1}$			
		α = 0.3	*α* = 0.7	*α* = 1.1	*α* = 1.5
0.35	40	-	51	39	136
0.37	18	50	21	18	27
0.40	10	16	11	10	13
0.45	6	8	6	6	7
0.50	4	5	4	4	5
0.60	3	3	3	3	3
0.70	2	2	2	2	2
0.80	2	2	2	2	2
0.90	1	1	1	1	1
mean	9.6	10.9	11.2	9.4	21.8

Table 2 shows the comparison results when the shaping parameter (α) is 1.1 and lower bound (L) is 50KB for varying system load (ρ) from 0.35 to 0.9. The upper bound (U) varies from 1MB to 100MB.

Table 2. The number of objects comparison varying *ρ* and *U*

when $\theta_{M/H2/1} = \theta_{M/G/1/PS}$ and $\theta_{M/H2/1} = \theta_{M/BP/1}$

ρ	*N when* $\theta_{M/H2/1} = \theta_{M/G/1/PS}$	*N when* $\theta_{M/H2/1} = \theta_{M/BP/1}$		
		L=50KB, *U*=1MB	*L*=50 KB, *U*=10MB	*L*=50KB, *U*=100MB
0.35	40	39	2	1
0.37	18	18	2	1
0.40	10	10	2	1
0.45	6	6	2	1
0.50	4	4	2	1
0.60	3	3	1	1
0.70	2	2	1	1
0.80	2	2	1	1
0.90	1	1	1	1
mean	9.6	9.4	1.5	1.0

When the system load is less than 0.35 and U is 10MB or more, the mean number of objects is greater than zero unlike when U=1MB. The mean number of objects for U=100MB is less than that for U=10MB. In particular, the number of objects for U=100MB becomes one, thus a large file is required.

The mean number of objects is the largest when L is 50KB and U=1MB and the smallest when L is 50KB and U=100MB.

4. Conclusions

In this paper, we estimate the mean number of objects for M/H₂/1 model by using mean object sizes of

$M/H_2/1$ model, M/G/1/PS model, and M/BP/1 model. In the previous research, mean object sizes of service models are found assuming that mean waiting latency in the deterministic model is equal to that in $M/H_2/1$, M/G/1/PS, and M/BP/1 models, respectively. Numerical experiments show that when the shaping parameter of M/BP/1 model is 1.1 and the system load is greater than 0.35, the mean number of objects for the $M/H_2/1$ model when mean object size of $M/H_2/1$ is the same as that of M/G/1/PS model is almost equal to the mean number of objects for the $M/H_2/1$ model when the mean object size of $M/H_2/1$ model is the same as that of M/BP/1 model. As the upper bound of Bounded Pareto distribution increases, the number of objects for $M/H_2/1$ model becomes one. Therefore, it is proposed to use multiple web objects of small size in order to reduce the upper bound in the M/BP/1 model. Future work includes the generalization of the mean number of objects model.

References

[1] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems, Cambridge University Press, USA, 2013.* pp. 353-358.

[2] 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, pp. 2513-2517, 2002.
DOI: https://doi.org/10.1109/GLOBECOM.2002.1189083

[3] Shi, E. Collins, and V. Karamcheti, "Modeling object characteristics of dynamic web content," *Journal of Parallel and Distributed Computing*, vol. 63, no. 10, pp. 963-980, 2003.
DOI: https://doi.org/10.1016/j.jpdc.2003.05.001

[4] Y. Lee, "Mean object comparison of M/G/1/PS and TDM system," *ICIC Express Letters*, vol. 12, no. 5, pp. 417-423, 2018.
DOI: https://doi.org/10.24507/icicel.12.05.417

[5] J. Cao; M. Andersson; C. Nyberg; M. Kihl, "Web server performance modeling using an M/G/1/K*PS queue", *Proceedings of 10th International Conference on Telecommunications* (*ICT* 2003), pp. 1501-1506, 2003.
DOI: https://doi.org/10.1109/ICTEL.2003.1191656

[6] Y. Lee, "On the comparison of mean object size in M/G/1/PS model and M/BP/1 model for web service," International Journal of Internet, Broadcasting and Communication, vol. 14, no. 3, pp. 1-7, 2022.
DOI: http://dx.doi.org/10.7236/IJIBC.2022.14.3.1

[7] Y. M. Tripathi, C. Petropoulos, and M. Jha, "Estimation of the shape parameter of a Pareto distribution," Communications in Statistics- Theory and Methods, vo. 47, no. 18, pp. 4459-4468, 2018.
DOI: https://doi.org/10.1080/03610926.2017.1376088

[8] Y. –J. Lee, "Mean object size considering average waiting delay in M/BP/1 system," *International Journal of Computer Networks and Communications*, vol. 12, no. 5, pp. 73-80, 2020.
DOI: https://dx.doi.org/10.2139/ssrn.3724774

[9] V.N. Tarasov, "Analysis of queues with hyperexponential arrival distribution," *Problems of Information Transmission* volume, vol. 52, pp. 14–23, 2016.
DOI: https://doi.org/10.1134/S0032946016010038

[10] Y. –J. Lee, "Web object size satisfying mean waiting time in multiple access environment," *International Journal of Computer Networks and Communications*, vol. 6, no. 4, pp.1-9, 2014.
DOI: https://doi.org/10.5121/ijcnc.2014.6401