*International Journal of Internet, Broadcasting and Communication Vol.14 No.3 1-7 (2022)*
*http://dx.doi.org/10.7236/IJIBC.2022.14.3.1*

**IJIBC 22-3-1**

# On the comparison of mean object size in M/G/1/PS model and M/BP/1 model for web service

Yongjin Lee

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

***Abstract***

*This paper aims to compare the mean object size of M/G/1/PS model with that of M/BP/1 model used in the web service. The mean object size is one of important measure to control and manage web service economically. M/G/1/PS model utilizes the processor sharing in which CPU rotates in round-robin order giving time quantum to multiple tasks. M/BP/1 model uses the Bounded Pareto distribution to describe the web service according to file size. We may infer that the mean waiting latencies of M/G/1/PS and M/BP/1 model are equal to the mean waiting latency of the deterministic model using the round robin scheduling with the time quantum. Based on the inference, we can find the mean object size of M/G/1/PS model and M/BP/1 model, respectively. Numerical experiments show that when the system load is smaller than the medium, the mean object sizes of the M/G/1/PS model and the M/BP/1 model become the same. In particular, when the shaping parameter is 1.5 and the lower and upper bound of the file size is small in the M/BP/1 model, the mean object sizes of M/G/1/PS model and M/BP/1 model are the same. These results confirm that it is beneficial to use a small file size in a web service.*

***Keywords:*** *mean object size, M/G/1/PS, M/BP/1, deterministic model, web service, multiple access users*

## 1. Introduction

The mean object size is one of the most important service quality measures on the Internet when multiple users want to transfer objects from a server at the same time. This measurement affects the mean waiting latency perceived by the end user. The mean object size must be estimated to meet the mean waiting latency desired by the end user.

The mean waiting latency of a web service is important performance measure in managing a web server. This time increases fast as the number of concurrent users increases. Typically, end users request objects from the web server based on the Poisson distribution, and the service time has the general distribution. Thus, the mean waiting latency in the communication network is formulated by the M/G/1 model [1]. As a general distribution of web services on the Internet, Weibull distribution, exponential distribution, and hyper exponential distribution have been proposed, respectively [2,3,4].

Manuscript Received: June. 30, 2022 / Revised: August. 8, 2022 / Accepted: August. 9, 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

On the other hand, the correct scheduling policy significantly reduces mean waiting latency at no additional cost. Processor Sharing (PS) is the most common used service policy on web server. Thus, we consider M/G/1/PS model [5,6,7].

The file size distribution with the lower bound and the upper bound is known as the Bounded Pareto distribution. Therefore, we also consider the M/BP/1 model [7, 8, 9] to describe the web service. If multiple users request objects from a web server at the same time and round-robin scheduling policy is used, the deterministic model can find the mean waiting latency. It can be inferred that in a steady state, the mean waiting latency of the deterministic model is equal to the mean waiting latency of the M/G/1/PS model and the M/BP/1 model. Thus, we can find mean object size satisfying that mean waiting delay of deterministic model is equal to that of M/G/1/PS model and M/BP/1 model, respectively.

This paper is based on the previous papers [5, 8] and compares the mean web object size between M/G/1/PS model and M/BP/1 model. When system load is less than 0.6, mean object size of M/G/1/PS model and M/BP/1 model is same regardless of the number of simultaneous access users. Maximum object size difference ratio between M/G/1/PS and M/BP/1 model is less than 1%. However, system load is greater than 0.7, difference ratio increases up to about 2 % ~ 4%. However, this difference may decrease by adjusting the shape parameter of M/BP/1 model.

This paper is organized as follows. Section 2 first describes the mean waiting latency in the deterministic model. It then explains the mean object size for M/G/1/PS model and M/BP/1 model when the mean waiting latency of the deterministic model is equal to the mean waiting latency of M/G/1/PS model and M/BP/1 model, respectively. Section 3 presents and analyzes the comparison of mean object size between M/G/1/PS model and M/BP/1 model. Finally, section 4 discusses the conclusions and future research.

## 2. Mean web object size in M/G/1/PS and M/BP/1 model

This section describes the results for the mean object size derived from the deterministic model, M/G/1/PS model, and M/BP/1 model used as web services in the previous papers [5, 8].

### 2.1. Mean object size in deterministic model

The deterministic model outlines the mean waiting latency for web object transmission [10]. In most object delivery service, $\alpha$ concurrent users access the same objects such as home page of web server simultaneously. The transport layer divides the object into several packets having the maximum segment size (MSS). If $\theta$ and $S$ represent the object size and the maximum segment size respectively, the number of packets ($\beta$) is equal to $\theta/S$.

When $\alpha$ concurrent users request the same object, each user thinks the service time is the same as the other users. However, service completion time depends on the scheduling policy. Most operating systems use a round-robin scheduling policy.

We assume that the time quantum of the round-robin scheduling policy is equal to a packet service time. When a user requests an object from a server, the object contains $\beta$ packets. The task size ($x$) is the total service time each user expects. The time quantum is equal to the packet service time ($\delta$), so $\delta = x/\beta$. Figure 1 depicts the relationship between service time and task size in a multiple user access environment [5].

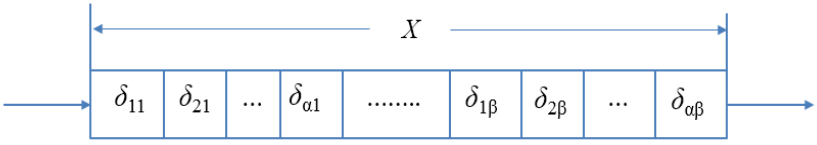


**Figure 1. packet service time ($\delta$) and Job size($x$) for multiple users ($\alpha$)**

In Figure 1, $\delta_{ij}$ is service time of the $j^{th}$ packet for the $i^{th}$ user. If we letting $\delta_{ij} = \delta$ $(\forall i, j)$, mean waiting latency in the deterministic model (E($W_D$)) is given as follows

$$E(W_D) = \frac{1}{\alpha}\sum_{i=1}^{\alpha}[(\alpha - i)\delta + \alpha(\alpha - 1)(\beta - 1)\delta] = \frac{(\alpha-1)(2\beta-1)E(X)\times S}{2\theta} \quad (1)$$

**2.2. Mean object size in M/G/1/PS system**

In this section, we consider the mean waiting latency of M/G/1/PS model. In process-sharing (PS) systems, the time-division CPU rotates in round-robin order between $\alpha$ tasks in the system that provide time quantum. When time quantum approaches to zero, PS [7] is obtained.

In the processor sharing service, when the service rate is $\mu$ and there are $\alpha$ jobs on the server, each job is processed at a rate of $\mu/\alpha$. In Figure 1, because $\delta$ is equal to $\mu/\alpha$, the mean waiting latency of the deterministic model can be regard as the mean latency of the M/G/1/PS model.

$W_Q(x)$ is mean waiting latency and is equal to E[wasted time ($x$)] [7].

$$\begin{aligned} W_Q(x) &= \mathrm{E}[\text{wasted time(x)}] \\ &= \mathrm{E}[\text{the number of times tagged job is interrupted}] \times \mathrm{E}[\text{length of interrupt}] \\ &= \frac{\lambda E(Y)}{1-\rho} = \frac{\lambda x}{\mu(1-\rho)} = \frac{\rho x}{1-\rho} \end{aligned} \quad (2)$$

In Eq. (2), $\lambda$ and $\mu$ are mean arrival rate and mean service rate, respectively. $\rho$ ($\lambda/\mu$) is the system load ($0 \leq \rho < 1$). E($Y$) is mean service time of a job on the CPU.

We can infer that the mean waiting latency of the deterministic model using round-robin policy and that of M/G/1/PS model would be the same in the steady state. By letting E($W_D$) = E($W_Q(x)$) in Eq. (3), we can find the mean object size ($\theta$: bytes) in the steady state [5].

$$\begin{aligned} E(W_D) = \mathrm{E}(W_Q(\mathrm{x})) \rightarrow \frac{(\alpha-1)(2\beta-1)E(x)\times S}{2\theta} = \frac{\rho E(x)}{1-\rho} \\ \rightarrow\ \theta_{M/G/1/PS} = \frac{(\alpha-1)(1-\rho)\times S}{2[(\alpha-1)(1-\rho)-\rho)]} \\ \text{where} \quad \alpha > 1 + \frac{\rho}{(1-\rho)} \end{aligned} \quad (3)$$

Because $(1 - \rho)(\alpha - 1) - \rho$ should be positive in Eq. (3), the number of users ($\alpha$) is given by

$$\alpha > 1 + \frac{\rho}{1-\rho} \tag{4}$$

### 2.3. Mean object size in M/BP/1 model

In this section, we consider the mean waiting latency of M/BP/1 model [7, 8]. Probability density function (pdf) of file size in Bounded Pareto distribution is given by

$$f_x(x) = \frac{kL^k x^{-k-1}}{1-\left(\frac{L}{U}\right)^k}, \quad L \le x \le U \tag{5}$$

Here, $k$ is the shape parameter, $L$ and $U$ are the lower bound and the upper bound of file size, respectively. Mean and second moment of the Bounded Pareto distribution are given by

$$\begin{cases} E_x(x) = \frac{L^k}{\left(1-\left(\frac{L}{U}\right)^k\right)}\left(\frac{k}{k-1}\right)\left(\frac{1}{L^{k-1}} - \frac{1}{U^{k-1}}\right) \\ E_x(x^2) = \frac{L^k}{\left(1-\left(\frac{L}{U}\right)^k\right)}\left(\frac{k}{k-2}\right)\left(\frac{1}{L^{k-2}} - \frac{1}{U^{k-2}}\right) \end{cases} \tag{6}$$

When $\lambda$ and $\rho$ are the arrival rate and the system load, respectively and $X$ is a random variable of the service time in an M/G/1 model, the mean waiting latency is given by

$$E(W) = \frac{\lambda E(X^2)}{2(1-\rho)} \tag{7}$$

We let E($X$) and E($X^2$) to be the mean and the second moment of the service time distribution respectively, $\rho = \lambda$E($X$). When the link capacity is $C$, by using the file size distribution in Eq. (5), we obtain following E($X$) and E($X^2$).

$$\begin{cases} E(X) = \frac{E_x(X)}{C} \\ E(X^2) = \frac{E_x(X^2)}{C^2} \end{cases} \tag{8}$$

Now, we may infer that in the steady state, the mean waiting latency (E($W_D$)) of the deterministic model becomes the mean waiting latency (E($W$)) of the M/BP/1 model.

By using $\beta = \theta/S$, we can find the mean object size, $\theta_{M/BP/1}$ when E($W_D$) = E($W$). Since Eq. (1) and Eq. (7) are the same, we obtain $\theta_{M/BP/1}$ as the follows [8].

$$\begin{aligned} \mathrm{E}(W_D) \;=\; \mathrm{E}(W) &\rightarrow \frac{(\alpha-1)(2\beta-1)E(X)\times S}{2\theta} = \frac{\lambda E(X^2)}{2(1-\rho)} \\ \theta_{M/BP/1} &= \frac{(\alpha-1)E(X)(1-\rho)\times S}{2(\alpha-1)E(X)(1-\rho) - \lambda E(X^2)} \\ &\text{where} \quad \alpha > 1 + \frac{\lambda E(X^2)}{2(1-\rho)E(X)} \end{aligned} \tag{9}$$

In Eq. (9), $\alpha$ is the number of multiple access users that satisfies the positive denominators of $\theta_{M/BP/1}$.

## 3. Mean object size comparison between M/G/1/PS model and M/BP/1 model

We present and analyze the comparison of mean object size between M/G/1/PS model and M/BP/1 model in this section. We first investigate the mean object size of M/G/1/PS model and M/BP/1 model when lower bound ($L$) is 50KB and upper bound ($U$) is 1MB for varying system load ($\rho$). The number of concurrent users ($\alpha$) are 15 and 35. Link capacity ($C$) is 10 Mbps. Shaping parameter of Bounded Pareto distribution ($k$) is 1.1 for each load. Maximum segment size($S$) is 1460 B.

Table 1 compares the mean object size of M/G/1/PS ($\theta_{M/G/1/PS}$) with the mean object size of M/BP/1 ($\theta_{M/BP/1}$). For a more quantitative comparison, we introduce the difference ratio as follows.

$$diff\_ratio = \frac{\theta_{M/BP/1} - \theta_{M/G/1/PS}}{\theta_{M/G/1/PS}} \times 100\ (\%) \tag{10}$$

When $\rho$ is less than 0.4, the diff_ratio is zero. That is, $\theta_{M/G/1/PS}$ is equal to $\theta_{M/BP/1}$. As $\rho$ increases, the diff_ratio also increases, but the maximum diff_raio is 4%. Average difference ratio (Avg. diff_ratio) is 1.2 % for $\alpha$=15 and 0.39% for $\alpha$=35, respectively. That is, as the number of users ($\alpha$) increases, $\theta_{M/G/1/PS}$ becomes equal to $\theta_{M/BP/1}$.

**Table 1. Mean object size comparison for $\alpha$=15 and $\alpha$=35 varying $\rho$**

| | $\alpha$=15, $k$=1.1, $C$=10Mbps, $L$=50KB, $U$=1MB | | $\alpha$=35, $k$=1.1, $C$=10Mps, $L$=50KB, $U$=1MB | |
|---|---|---|---|---|
| $\rho$ | $\theta_{M/BP/1}$ | $\theta_{M/G/1/PS}$ | $\theta_{M/BP/1}$ | $\theta_{M/G/1/PS}$ |
| 0.1 | 736 | 736 | 732 | 732 |
| 0.2 | 744 | 743 | 735 | 735 |
| 0.3 | 755 | 753 | 740 | 739 |
| 0.4 | 770 | 767 | 746 | 745 |
| 0.5 | 792 | 786 | 754 | 752 |
| 0.6 | 827 | 818 | 767 | 764 |
| 0.7 | 894 | 876 | 789 | 784 |
| 0.8 | 1065 | 1022 | 838 | 827 |
| mean | 823 | 813 | 763 | 760 |
| *Avg. diff_ratio* | 1.2 % | - | 0.39 % | - |

Table 2 shows numerical comparison results of the mean object size of M/G/1/PS and M/BP/1 model when lower bound ($L$) is 50KB and upper bound ($U$) is 1MB for varying shaping parameter ($k$) of Bounded Pareto distribution. The number of concurrent users ($\alpha$) and link capacity ($C$) are 20 and 10 Mbps, respectively. Maximum segment size($S$) is 1460 B.

**Table 2. Mean object size comparison for $\alpha$=20 and varying $\rho$ and $k$**

| | $\alpha$=20, $C$=10Mbps, $L$=50KB, $U$=1MB | | | |
|---|---|---|---|---|
| $\rho$ | $\theta_{M/BP/1}$ | | | $\theta_{M/G/1/PS}$ |
| | $k$ = 0.7 | $k$ = 1.1 | $k$ = 1.5 | |
| 0.1 | 734 | 734 | 734 | 734 |
| 0.2 | 741 | 740 | 740 | 740 |
| 0.3 | 749 | 748 | 747 | 747 |
| 0.4 | 761 | 759 | 757 | 757 |
| 0.5 | 777 | 774 | 772 | 771 |
| 0.6 | 803 | 799 | 795 | 793 |
| 0.7 | 851 | 844 | 836 | 832 |
| 0.8 | 966 | 950 | 933 | 925 |
| mean | 798 | 794 | 789 | 787 |
| *Avg. diff_ratio* | 1.39 % | 0.88 % | 0.25 % | - |

The difference ratio is smallest when $k$ = 1.5, followed by when $k$ = 1.1. When $k$ = 0.7, the difference ratio is the largest. When $k$=1.5 and $\rho$ is less than 0.7, $\theta_{M/G/1/PS}$ is equal to $\theta_{M/BP/1}$.

Average difference ratio is 1.39 % for $k$ = 0.7 and 0.88 % for $k$ = 1.1, and 0.25 % for $k$ = 1.5, respectively. These small difference ratio means that $\theta_{M/G/1/PS}$ becomes equal to $\theta_{M/BP/1.}$

## 4. Conclusions

In this paper, we compare the mean object size of each model when using the M/G/1/PS model and the M/BP/1 model on a web server. We infer that the mean waiting latency of M/G/1/PS and M/BP/1 models would be equal to the mean waiting latency of the deterministic model.

Numerical experiments show that the mean object size of the M/G/1/PS model and the M/BP/1 model is the same when the system load is below medium even though the number of concurrent users increase. In this case, the maximum object size difference between M/G/1/PS model and M/BP/1 model is less than 1%.

When the system load is greater than medium, the object size difference between M/G/1/PS model and M/BP/1 model increases slightly. But this difference can be reduced by modifying the shape parameter of the bounded pareto distribution in the M/BP/1 model.

Especially, when the shaping parameter is 1.5 and the lower and upper bounds of the file size are small in the bounded pareto distribution, the mean object size of the M/G/1/PS model and the M/BP/1 model is the same.

According to the comparison presented in this paper, a small file size is advantageous for efficient web services. In the future, it is necessary to compare the mean object sizes of this paper with the mean object sizes of other service models.

## References


[1] S. Ross, *Introduction to Probability Model*, 12th Ed., Academic press, New York, 2019, USA.

[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.
DOI: https://doi.org/10.1016/j.jpdc.2003.05.001

[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.
DOI: https://doi.org/10.1016/S0166-5316(02)00179-7

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

[5] 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

[6] S. Aalto, U. Ayesta, E. Nyberg-Oksanen, "M/G/1/MLPS compared to M/G/1/PS", *Operations Research Letters*, vol. 33, no. 5, pp. 519-524, 2005.
DOI: https://doi.org/10.1016/j.orl.2004.09.009

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

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

[9] F. Grassi and A. Coluccia, "On the sum of random samples with bounded Pareto distribution", *Signal Processing*, vol. 192, 2022.
DOI: https://doi.org/10.1016/j.sigpro.2021.108389

[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