MEAN OBJECT SIZE CONSIDERING AVERAGE WAITING LATENCY IN M/BP/1 SYSTEM

Y. –J. Lee

Department of Technology Education, Korea National University of Education, Cheongju, South Korea

ABSTRACT

This paper deals with the web object size which affects to the service time in multiple access environments. The M/BP/1 model can be considered because packets arrival and web service are Poission and Bound Pareto (BP) distribution respectively. We find mean object size which satisfies that the average waiting latency by deterministic model equals the mean queueing delay of the M/BP/1 model. Performance evaluation shows that the mean web object size is affected by file size bounds and shape parameter of BP distribution, however, the impact of link capacity is not significant. When the system load is low, web object size converges on half the maximum segment size (MSS). Our results can be applied to find mean web object size in the economic web service design.

KEYWORDS

M/BP/1system, average waiting latency, multiple web access, mean queueing delay

1. INTRODUCTION

Average waiting latency for web services is one of the most important control factors when managing a web server. It increases rapidly as the number of concurrent user's increases. To solve the problem, we must first accurately calculate the average waiting latency of the end user.

Generally, users' requests for web servers per unit time follow a Poisson distribution and Web service time follows a general distribution instead of an exponential distribution. The M/G/1 model is known to be suitable for describing web services affected by the web object size [1, 2]. Khayari et al. and Riska et al. presented the fitting algorithm of empirical and hyper-exponential distribution [3, 4]. Si et al. suggests that Weibull and exponential distributions are suitable for statistical distributions describing web services [5]. Meanwhile, the number of concurrent users meeting average latency has been found when web services are provided by a hyper-exponential distribution in a steady state [6]. Y. Lee estimated the average web object size in multi-user services for M/D/1 and M/H2/1 systems [7].

The file size distribution of Internet traffic with many small files and a small number of large files using the TCP protocol is known as the Pareto distribution [8]. Therefore we use the M/BP (bounded pareto)/1 model to describe the web service in this paper.

When multiple users simultaneously request web objects from a web server and round-robin scheduling is used for web services, we can determine the average waiting latency in the deterministic model. In a steady state, we can deduce that the average waiting latency of deterministic model is almost equal to the mean queueing delay of M/BP/1 model.

This study aims to find mean web object size satisfying that the average waiting latency for deterministic model is equal to the mean queueing delay for M/BP/1 model. We also find the number of concurrent users satisfying this assumption. The reason for obtaining web object size that satisfies end-user delay constraints is why its control is most economical in the design of web services.

The remainder of this paper is composed of followings. Next section first discusses the deterministic model to find the average waiting latency in the deterministic model. It then explains the M/BP/1 model, and estimate mean web object size if average waiting latency is the same as the mean queueing delay for M/BP/1. Section 3 presents and analyzes the computational results. Finally, Section 4 presents conclusions and future research.

2. AVERAGE WAITING LATENCY FOR DETERMINISTIC MODEL AND MEAN QUEUEING DELAY FOR M/BP/1 MODEL

2.1. Average Waiting Latency for Deterministic Model

The deterministic model describes the average latency for web object transmission [9]. In most object transfer services, m concurrent users typically require the same object at the same time, such as index.html on a web server. The object is split into multiple packets with the maximum segment size (MSS) in the transport layer. θ represents the object size and mss denotes maximum segment size. The number of packets (n) is then given $n = \theta/mss$.

When multiple clients request the same object, each client thinks its service time is the same as the other. However, because the number of clients is larger than the number of processors, service completion time varies depending on scheduling policy. In most operating systems, processor is shared by round-robin (RR) scheduling policy.

We assume that the time quantum of the RR scheduling policy is equal to the packet service time. When a client requests an object from the server, the object contains n packets. The job size(x) represents the total service time each client expects. Because the time quantum equals the packet service time, $\tau = x/n$. Figure 1 shows the relationship between service time and job size in a multi-user access environment [9].

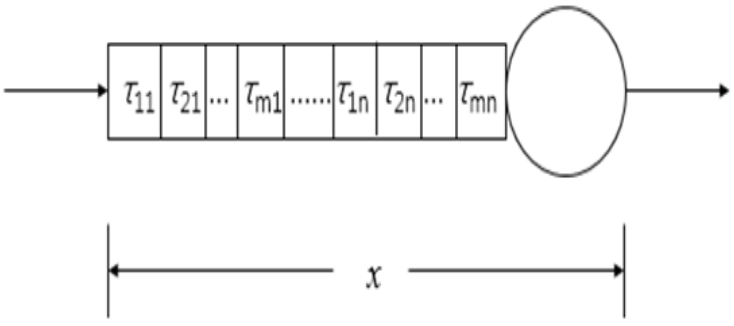

Figure 1. Job size(x) and packet service time (τ) for multiple users (m)

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

$$E(W_D) = \frac{1}{m}\sum_{i=1}^{m}\left[(m-i)\tau + m(m-1)(n-1)\tau\right] \\ = \frac{(m-1)(2n-1)E(X)\times mss}{2\theta} \tag{1}$$

2.2. Mean Queueing Delay for M/BP/1 Model

In this section, we describe the mean queueing delay for M/BP/1 model based on [10, 11, 12]. Probability density function (pdf) of file size for Bounded Pareto distribution is given by

$$f_x(x) = \frac{kL^k x^{-k-1}}{1-\left(\frac{L}{U}\right)^k}, \qquad L \le x \le U \tag{2}$$

Here, k is the shape parameter, L is the minimum file size, and U is the maximum file size. Mean of the Bounded Pareto distribution is given by

$$E_x(x) = \int_{-\infty}^{\infty} x \cdot f_x(x)dx = \frac{kL^k}{1-\left(\frac{L}{U}\right)^k}\frac{1}{k-1}\left(L^{-k+1} - U^{-k+1}\right) \\ = \frac{L^k}{\left(1-\left(\frac{L}{U}\right)^k\right)}\left(\frac{k}{k-1}\right)\left(\frac{1}{L^{k-1}} - \frac{1}{U^{k-1}}\right) \tag{3}$$

Second moment of the Bounded Pareto distribution is given by

$$E_x(x^2) = \int_{-\infty}^{\infty} x^2 \cdot f_x(x)dx = \frac{kL^k}{1-\left(\frac{L}{U}\right)^k}\frac{1}{k-2}\left(L^{-k+2} - U^{-k+2}\right) \\ = \frac{L^k}{\left(1-\left(\frac{L}{U}\right)^k\right)}\left(\frac{k}{k-2}\right)\left(\frac{1}{L^{k-2}} - \frac{1}{U^{k-2}}\right) \tag{4}$$

Generally, j^{th} moment of the Bounded Pareto distribution is given by

$$E_x(x^j) = \int_{-\infty}^{\infty} x^j \cdot f_x(x)dx = \begin{cases} \dfrac{kL^k}{1-\left(\frac{L}{U}\right)^k}\dfrac{\left(L^{j-k} - U^{j-k}\right)}{k-j} & if\ \ k \ne j \\ \dfrac{L}{1-\left(\frac{L}{U}\right)}\left(\ln U - \ln L\right) & if\ k = j = 1 \end{cases} \tag{5}$$

Variance of the file size distribution is given by

$$\sigma_x{}^2 = \int_{-\infty}^{\infty} (x - E_x(x)^2 \cdot f_x(x) dx = E_x(x^2) - \left(E_x(x)\right)^2 \tag{6}$$

If λ is the arrival rate and X is an arbitrary variable representing service time of the M/G/1 model, the mean queue delay of the system is as follows:

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

Here, system load is ρ (=λE(X)) and second moment is E(X^2). By using the file size distribution in Eq. (2), and the link capacity (C), E(X) and E(X^2) are given by

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

The mean queueing delay through system (E(T)) is given by

$$\mathrm{E}(T) = \mathrm{E}(W) + \mathrm{E}(X) \tag{9}$$

By the Little's formula, mean number of calls (E(N)) in the system is given by

$$\mathrm{E}(N) = \lambda \mathrm{E}(T) \tag{10}$$

2.3. Mean Web Object Size

Now, we can infer that the average waiting latency(W_D) in the deterministic model will be equal to mean queueing delay in the system(E(W)) or mean queueing delay through the system (E(T)) in the M/BP/1 model in the steady state.

By using $n = \theta/mss$, we can obtain mean object size(θ_w) when E(W_D) = E(W) and the mean object size(θ_T) when E(W_D) = E(T).

By letting Eq. (1) and Eq. (7) to be equal

$$\frac{(m-1)(2m-1)E(X) \times mss}{2\theta} = \frac{\lambda E(X^2)}{2(1-\rho)} \tag{11}$$

Solving the above equation, θ_w is given by

$$\theta_w = \frac{(m-1)E(X)(1-\rho)\times mss}{2(m-1)E(X)(1-\rho)-\lambda E(X^2)}$$

$$where \quad m_w > 1+\frac{\lambda E(X^2)}{2(1-\rho)E(X)} \tag{12}$$

In the same way, by letting Eq. (1) and Eq. (9) to be equal

$$\frac{(m-1)(2m-1)E(X)\times mss}{2\theta} = \frac{\lambda E(X^2)}{2(1-\rho)} + E(X) \tag{13}$$

Solving the above equation, θ_T is given by

$$\theta_T = \frac{(m-1)E(X)(1-\rho)\times mss}{2(m-1)E(X)(1-\rho)-[\lambda E(X^2)+(1-\rho)E(X)]}$$

$$where \quad m_T > 1+\frac{\lambda E(X^2)+(1-\rho)E(X)}{2(1-\rho)E(X)} \tag{14}$$

In Eq. (12) and Eq. (14), m_w and m_T represent minimum number of concurrent users satisfying the denominators of θ_w and θ_T are positive respectively. Thus, m is set the greater value than maximum (m_w, m_T).

Upper bound of object size is given by

$$\theta_w^{UB} = \lim_{m\to\infty} \theta_w = \lim_{m\to\infty} \theta_T = \theta_T^{UB} = \frac{mss}{2} \tag{15}$$

3. PERFORMANCE EVALUATION

In this section, we look at the changes in the mean object size as the shape parameter(k), the minimum file size(L), and the maximum file size(U) vary.

We first compute the mean object size when lower bound(L) = 50KB, upper bound(U) = 1MB for various load(ρ). The number of concurrent users (m) = 15 and link capacity(C) is 10Mbps. Shaping parameter(k) is 1.1 for each load and mss is 1460B. Table 1 shows computational result. For all ρ, θ_w is less than θ_T. When $\rho \leq 0.7$, θ_w and θ_T are nearly same as shown in Table 1.

Although we do not show the mean object size for varying link capacity(C) in Table 1, θ_w is equal to θ_T for all loads when C = 10Mbps, 100Mbps, 1Gbps, and 10Gbps. The reason is that because E(X) is very small for above link capacities, E(W) and E(T) have no differences for all link capacities.

Table 1. Mean object size (θw, θ_T) for k =1.1, m=15, and C=10Mbps

ρ	E(X)	$E(X^2)$	E(W)	E(T)	E(N)	m_w	θ_w	m_T	θ_T
0.0	0.108	0.026	0.000	0.108	0.000	2	730	2	757
0.1	0.108	0.026	0.013	0.121	0.112	2	736	2	763
0.2	0.108	0.026	0.030	0.137	0.255	2	744	2	772
0.3	0.108	0.026	0.051	0.159	0.442	2	755	2	784
0.4	0.108	0.026	0.079	0.187	0.694	3	770	3	800
0.5	0.108	0.026	0.119	0.226	1.051	3	792	3	824
0.6	0.108	0.026	0.178	0.286	1.592	4	827	4	862
0.7	0.108	0.026	0.277	0.385	2.500	5	894	5	935
0.8	0.108	0.026	0.475	0.582	4.327	6	1065	6	1124
0.9	0.108	0.026	1.068	1.176	9.828	11	2505	12	2854
mean	0.108	0.026	0.229	0.337	2.080	4	982	4	1048

Now, we compute mean object sizes(θ_w and θ_T) for varying lower bound(L) and upper bound(U) when shaping parameter(k) is equal to 1.1 and the number of concurrent users (m) is set to 35. When C=10Mbps, Table 2 shows mean object sizes (θ_w and θ_T).

In Table 2, for all ρ, θ_w is less than θ_T. When $\rho \leq 0.8$, θ_w and θ_T are nearly same except when ρ = 0.9, L=500KB, and U=100MB. That is, upper bound is very large and system load approaches to 1, then mean object size increases sharply.

Table 2. Mean object size (θw, θ_T) for varying L and U when k =1.1, m=35, and C=10Mbps

ρ	L=50KB U=1MB		L=500KB U=100MB		L=5MB U=10MB	
	θ_w	θ_T	θ_w	θ_T	θ_w	θ_T
0.0	730	740	730	740	730	740
0.1	732	743	738	749	735	746
0.2	735	747	749	760	741	752
0.3	740	751	763	775	749	761
0.4	746	757	783	796	761	773
0.5	754	766	813	827	777	790
0.6	767	779	863	878	804	817
0.7	789	802	961	980	852	867
0.8	838	853	1242	1274	968	987
0.9	1030	1052	10096	12674	1634	1690
mean	786	799	1774	2045	875	892

Table 3 represents numerical computation results of mean object sizes(θ_w and θ_T) for varying shaping parameter(k) when lower bound(L) = 50KB, upper bound(U) = 1MB, the number of concurrent users (m) = 20, and C=10Mbps.

For all ρ, θ_w is less than θ_T. When shaping parameter (k) increases, both of θ_w and θ_T decreases. When system load (ρ) = 0.9, difference of θ_w and θ_T is largest. When $\rho \leq 0.5$, mean web object sizes of θ_w and θ_T converge to the half to maximum segment size.

Table 3. Mean object size (θw, θ_T) for varying k when m=20, L=50KB, U=1MB and C=10Mbps

ρ	k =0.3		k = 0.7		k =1.1		k = 1.5	
	θ_w	θ_T	θ_w	θ_T	θ_w	θ_T	θ_w	θ_T
0.0	730	749	730	749	730	749	730	749
0.1	736	756	734	755	734	754	734	754
0.2	745	765	741	761	740	761	740	760
0.3	756	777	749	770	748	769	747	768
0.4	771	794	761	782	759	780	757	778
0.5	794	818	777	800	774	797	772	794
0.6	831	857	803	827	799	823	795	818
0.7	901	931	851	878	844	870	836	862
0.8	1082	1126	966	1001	950	984	933	966
0.9	2736	3035	1625	1726	1527	1616	1434	1512
mean	1008	1061	874	905	861	890	848	876

From Table 1 ~ Table 3, we find that mean object size is affected by shaping parameter and lower and upper bound of BP distribution. However, the impact of link capacity is not significant.

4. CONCLUSIONS

We present an analytical model for finding mean web object size that meet constraints so that the average waiting latency in the deterministic model is equal to the mean queueing delay in the M/BP/1 model. We derive mean web object size and also find out feasible number of users satisfying the constraint. Numerical computation results show that bounds of file size and shape parameter of BP distribution mainly affect to web object size. It is also found that mean web object size converges on half the maximum segment size of TCP when system load is low. Our result can be applied to control web service of end-users. Future works include more exact model to describe web service pattern more exactly.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

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

[2] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, Cambridge University Press, New York, 2013, pp. 354-358, pp. 395-404, USA.

[3] R. Khayari, R. Sadre and B. R. Haverkort, "Fitting world-wide web request traces with the EM-algorithm, *Performance Evaluation*," Vol. 52, pp. 175-191, 2003.

[4] A. 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.

[5] W. Shi, E. Collins, and V. Karamcheti, "Modeling Object Characteristics of Dynamic Web Content," *Journal of Parallel and Distributed Computing*, Elsevier Science, pp. 963-980, 1998.

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

[7] 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.

[8] W. J. Reeds and M. Jorgensen, "The Double Pareto-Lognormal Distribution – A New Parametric Model for Size Distributions," Communications in Statistics – Theory and Methods. Vol. 33, No. 8. pp. 1733–53, 2004.
[9] Y. Lee, "Mean Object Size Comparison of M/G/1/PS and TDM System," ICIC Express Letters, Vol. 12, No. 5, pp. 417-423, 2018.
[10] http://www.ece.virginia.edu/~mv/edu/715/matlab-files/MG1-bounded-Pareto/m_bp_1.htm.
[11] Y.M. Tripathi, C. Petropoulos, and M. Jha, "Estimation of the shape parameter of a Pareto distribution," Communications in Statistics- Theory and Methods, Vol. 47, NO. 18, pp. 4459-4468, 2018.
[12] H. Chen, W. Cheng, J, Zhao, and X. Zhao, "Parameter estimation for generalized Pareto distribution by generalized probability weighted moment-equations," Communications in Statistics- Simulation and Computation, Vol. 46, No. 10, pp. 7761-7776, 2017.