ICIC Express Letters
Volume **12**, Number **5**, May **2018**

pp. 417–423

MEAN OBJECT SIZE COMPARISON OF M/G/1/PS AND TDM 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

Received October 2017; accepted January 2018

Abstract. *This paper aims at comparing mean object size in M/G/1/PS system with M/G/1/FCFS with TDM system in multiple access environments. Arrival pattern of packets is described by Poisson distribution and service time has general distribution. CPU scheduling policy in the server is considered processor sharing (PS) and time division multiplexing (TDM). We thus consider M/G/1/PS model and M/G/1/FCFS with TDM model. We derived mean object size satisfying the constraint such that mean waiting delay by round-robin scheduling in the deterministic model is equal to mean waiting delay of M/G/1/PS system and that of M/G/1/FCFS with TDM, respectively. 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 object size increases as the system utilization increases at the given maximum segment size, and lower bound of mean object size in M/G/1/PS system is less than that of M/G/1/FCFS with TDM system. Our results can be applied to the economic service design in the communication network.*

Keywords: Mean object size, M/G/1/PS, M/G/1/FCFS with TDM, Mean waiting delay, Simultaneous user access

1. **Introduction.** Mean object size is one of important service quality measures in the communication network including the Internet when multiple users want to transfer object in a server simultaneously [1,2]. This measure affects the mean waiting delay which end-user perceives. In order to satisfy the mean waiting delay that the end-user wants, mean object size should be first estimated. Controlling object size is to minimize the maintenance cost in the communication network.

Generally, end-user requests an object in the server according to the Poisson distribution and the service time is described by general distribution. Therefore, mean waiting delay in the communication network is formulated by M/G/1 model [3]. For the general distribution of web service in the Internet, Shi et al. [4], Khayari et al. [5] and Riska et al. [6] have proposed Weibull distribution, Exponential, and Hyper-exponential distribution, respectively. However, more exact service distribution is still expected.

Scheduling is an extremely important topic in computer and communication system. The right scheduling policy reduces mean waiting delay remarkably without additional costs. Scheduling policies are classified into non-preemptive and preemptive ones. FCFS (first come first served), RANDOM, and LCFS (last come first served) are examples of non-preemptive scheduling policy. PS (processor sharing) and PLCFS (preemptive last come first served) are examples of preemptive scheduling policy.

Previous researches for the M/G/1 model [7,8] have mainly considered the non-preemptive scheduling policy. Especially, FCFS policy does not make use of object size. However, since the service is affected by object size in the multiple user environments, we should consider object size in the not only non-preemptive scheduling policy but also the preemptive scheduling policy.

Time division multiplexing (TDM) is very similar to PS system excluding that the service time quantum is constant. To apply the TDM scheme to the queueing system can be described as an M/D/1 with vacations model in which the service distribution (G) is given by the constant (D). Thus, statistical model for TDM can be described by M/D/1/FCFS with TDM.

When several users simultaneously request an object in the server, and the packed based round-robin (RR) scheduling for the service is used, we can find the 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 or M/D/1/FCFS with TDM. Therefore, we can find out mean object size satisfying mean waiting delay which end-user wants in M/G/1/PS and M/D/1/FCFS with TDM system, respectively. Mean object size for M/G/1/PS [9] and M/D/1/FCFS with TDM [10] is derived in the earlier versions of this paper.

The main contribution of this paper is to find out the least cost object size satisfying the end-user's delay requirement in the server system by comparing mean object size between M/G/1/PS and M/G/1/FCFS with TDM. The controlling object size is the simple and easy method in server management. Our performance evaluation is based on the analytical model; therefore, the extended evaluation by real measurements is necessary in the further study.

The rest of this paper is organized as follows. In Sections 2 and 3, 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 and M/G/1/FCFS with TDM system, respectively. In Section 4, we present and analyze the performance evaluation results. Finally, in Section 5, we discuss conclusions and future research.

2. Mean Object Size in M/G/1/PS System.

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

When multiple clients request a same object, each client thinks that his response (service) time is the same as others. However, because the number of processors is less than the number of clients, each user's service completion time is different according to the used 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 equal to the packet service time. When a client requests an object from the server, n packets are included in the object. Job 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 access environment.

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

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

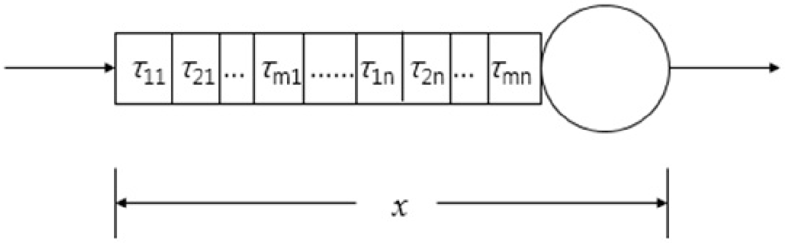

Figure 1. RR policy based on the packet service time (τ) for multiple users (m)

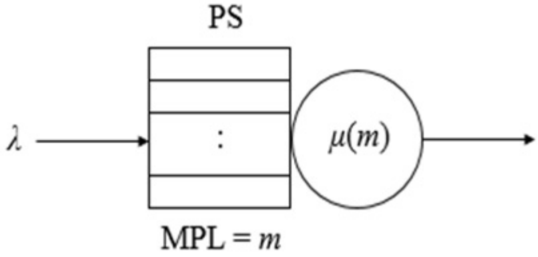

Figure 2. Processor sharing with multiprogramming level, MPL = m

Next we consider mean waiting delay in M/G/1/PS system. As an example of PS 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 service order if, whenever there are m jobs at the server, each of the job is processed at rate μ/m. In Figure 1, if we regard τ as μ/m, mean waiting delay for deterministic model becomes that for M/G/1/PS system presented in Figure 2. MPL means multiprogramming level.

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

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

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

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

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 serve a job on the CPU. ρ (λ/μ) is the system utilization ($0 \leq \rho < 1$).

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

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

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

$$m > 1 + \frac{\rho}{1-\rho} \tag{5}$$

TABLE 1. Maximum number of users (m) satisfying Equation (5)

Utilization (ρ)	The number of users (m)
$0 \leq \rho < 0.4$	2
$0.5 \leq \rho \leq 0.6$	3
0.7	4
0.8	6
0.9	11

Table 1 shows the maximum number of users (m) satisfying Equation (5).

The lower bound for object size in M/G/1/PS system is given by

$$\lim_{m\to\infty} \theta_{PS} = \theta_{PS}^{LB} = \frac{(2-\rho)\times mss}{2(1-\rho)} \tag{6}$$

3. **Mean Object Size in M/D/1/FCFS with TDM System.** In the TDM system depicted in Figure 3, m fixed-length packets with each λ/m arrival rate are multiplexed and arrive into the system according to the Poisson distribution. Total traffic is λ, the service rate is $1/m$, and the load on the system is $\rho = \lambda$.

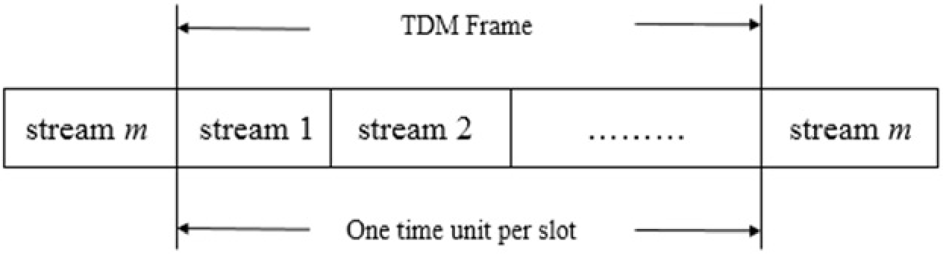

FIGURE 3. TDM system

In the M/D/1/FCFS system, the service times are identical for all requests. The expected mean queueing delay is given by

$$W_{M/D/1/FCFS} = \frac{\rho}{2\mu(1-\rho)} \tag{7}$$

By letting $\mu = 1/m$ and $\rho = \lambda$, we can obtain the mean queueing delay per packet for frequency division multiplexing (FDM).

$$W_{FDM} = \frac{\rho m}{2(1-\rho)} \tag{8}$$

In the TDM, m traffic streams are time division multiplexed in a scheme, whereby the time axis is divided in m slot frames with one slot dedicated to each traffic stream in Figure 3. Thus the mean queueing delay in TDM is given by [10,12]

$$W_{TDM} = \frac{m}{2(1-\rho)} \tag{9}$$

We assume that the mean waiting delay for the deterministic model with RR policy is equal to that for M/D/1/FCFS with TDM model. By letting $W_D = W_{TDM}$ in Equation (4), we can find mean object size (θ: bytes) for M/D/1/FCFS with TDM.

$$\begin{aligned} W_D = W_{TDM} &\to \frac{(m-1)(2n-1)x}{2\theta} \times mss = \frac{m}{2(1-\rho)} \\ &\to \theta_{TDM} = \frac{[(1-\rho)(m-1)+m]\times mss}{2(1-\rho)(m-1)} \end{aligned} \tag{10}$$

The lower bound for object size in M/G/1/FCFS with TDM system is given by

$$\lim_{m\to\infty}\theta_{TDM} = \theta_{TDM}^{LB} = \frac{(2-\rho)\times mss}{2(1-\rho)} \tag{11}$$

4. **Performance Evaluation.** We first compute mean object size when $mss = 1460$B for various ρ. Table 2 shows mean object size according to varying utilization (ρ). In Table 2, as ρ increases, both mean object sizes of M/G/1/PS (θ_{PS}) and M/D/1/FCFS with TDM (θ_{TDM}) increase although MSS is fixed at 1460B. When m is larger than 100, mean object size converges to the $mss/2$ for M/G/1/PS system and $2(1-\rho) * mss/[2(1-\rho)]$ for M/D/1/FCFS with TDM system regardless of ρ.

Table 2. Mean object size for M/G/1/PS and M/D/1/FCFS with TDM with varying utilization given $mss = 1460$B

The number of users (m)	$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.4$	
	θ_{PS}	θ_{TDM}	θ_{PS}	θ_{TDM}	θ_{PS}	θ_{TDM}
15	736	1599	743	1708	767	2034
20	734	1584	740	1691	757	2011
50	732	1558	734	1661	740	1971
100	731	1549	732	1652	735	1959
200	730	1545	731	1647	732	1953
300	730	1544	731	1646	732	1951
The number of users (m)	$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
	θ_{PS}	θ_{TDM}	θ_{PS}	θ_{TDM}	θ_{PS}	θ_{TDM}
15	786	2294	876	3337	2044	8551
20	771	2267	832	3291	1387	8414
50	745	2220	767	3213	894	8179
100	737	2205	748	3188	803	8104
200	734	2197	739	3176	765	8067
300	732	2195	736	3171	753	8054

Table 3 shows mean object size varying the number of users (m) when ρ is given by 0.5 and several MTUs (maximum transfer unit) are given. Mean object size becomes larger as m and MTU become larger. Mean object size of M/G/1/PS system is less than that of M/D/1/FCFS with TDM system. The ratio of θ_{TDM}/θ_{PS} increases as MTU and m increase.

Table 3. Mean object size for M/G/1/PS and M/D/1/FCFS with TDM with varying MTU given utilization = 0.5

The number of users (m)	MTU = 536		MTU = 2312		MTU = 4500	
	θ_{PS}	θ_{TDM}	θ_{PS}	θ_{TDM}	θ_{PS}	θ_{TDM}
15	289	842	1245	3633	2423	7071
20	283	832	1220	3590	2375	6987
50	274	815	1180	3515	2297	6842
100	271	809	1168	3491	2273	6795
200	269	807	1162	3480	2261	6773
300	269	806	1160	3476	2258	6765

Figure 4 depicts mean object size ratio of M/G/1/PS and M/D/1/FCFS with TDM varying the system utilization (ρ) when mss is given by 1460B and several numbers of users (m) are given. Mean object size ratio is the nearly same when the utilization is less than 0.5; however it becomes larger as m becomes larger and ρ approaches to 0.9.

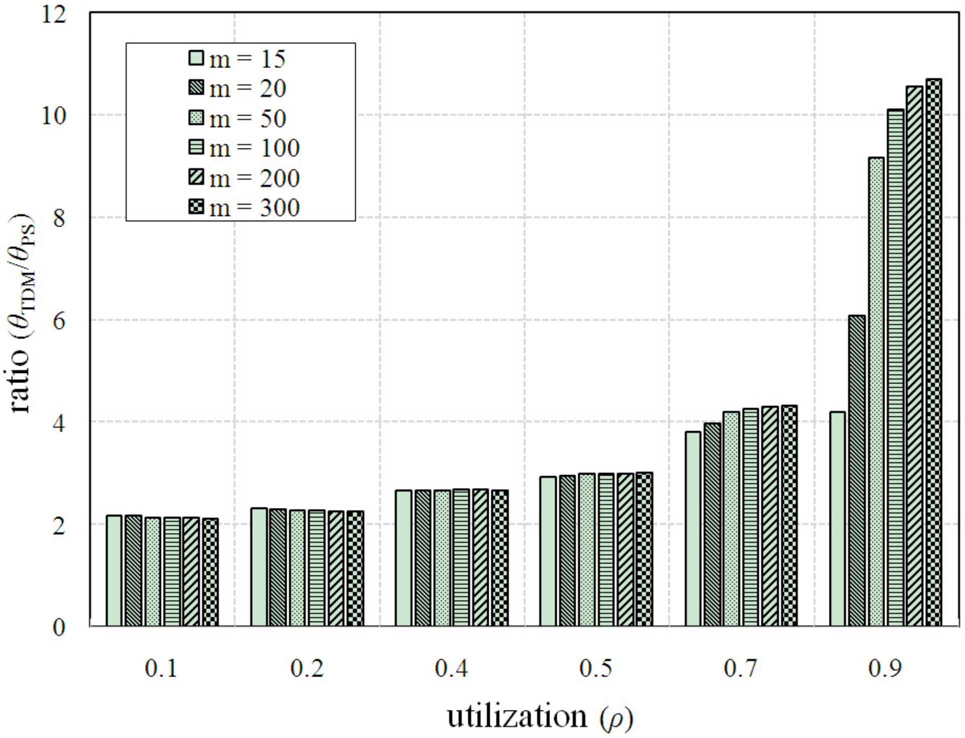

Figure 4. Mean object size ratio of M/G/1/PS and M/D/1/FCFS with TDM

As shown in above results, mean object size in M/G/1/PS system is less than that of M/D/1/FCFS with TDM system for varying utilization, MSS, and MTU. This means that small object size in time sharing server is desirable for reducing the waiting delay of end-user.

5. **Conclusions.** This paper presents the mean object size comparison between M/G/1/PS and M/G/1/FCFS with TDM system. We inferred that mean waiting delay for the deterministic model is equal to mean waiting delay for M/G/1/PS and M/G/1/FCFS with TDM system in the steady state. We can find out mean object size satisfying mean waiting delay that end-user can allow as quality of service. This means waiting object size can be used to control the object transfer service economically. Some computational experiences show that mean object size of M/G/1/PS system is less than that of M/G/1/FCFS with TDM system when the simultaneous access number of users becomes very larger regardless of the system utilization factor. Future works include more exact model to describe service behaviour and the distribution in multiple access environments.

REFERENCES

[1] Y. Lee, Lower bound for mean object transfer latency in the narrowband IoT environment, *International Journal of Applied Engineering Research*, vol.12, no.2, pp.3365-3369, 2017.

[2] Y. Lee, Novel quality of service measure for web transaction in multiple user access environments, *International Journal of Applied Engineering Research*, vol.10, no.16, 2015.

[3] S. Ross, *Introduction to Probability Model*, Academic Press, New York, 2010.

[4] 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.

[5] 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.

[6] A. Riska, V. Diev and E. Smirni, Efficient fitting of long-tailed data sets into hyper-exponential distributions, *Proc. of IEEE Global Telecommunications Conference*, vol.3, pp.2513-2517, 2002.

[7] G. A. Vidura, Performance analysis of EDF scheduling in a multi-priority preemptive M/G/1 queue, *IEEE Trans. Parallel and Distributed Systems*, vol.25, no.8, pp.2149-2158, 2014.

[8] S. Gao, An M/G/1 queue with single working vacation interruption under Bernouii schedule, *Applied Mathematics Modelling*, vol.37, no.3, pp.1564-1579, 2013.

[9] Y. Lee, Mean web object size for multiple user access in M/G/1/PS system, *Proc. of the 2017 World Congress on Computer Science, Computer Engineering and Applied Computing*, Lasvegas, USA, 2017.

[10] Y. Lee, Web object size satisfying mean waiting time in multiple access environment, *International Journal of Computer Networks & Communications*, vol.6, no.4, pp.1-9, 2014.

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

[12] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, New Jersey, 2007.