UDC 519.87

Estimating mean sojourn time in the processor sharing M/G/1 queue with inaccurate job size information

I. Horváth*, **R. Razumchik**†, **M. Telek**‡

* *MTA-BME Information Systems Research Group, Magyar Tudosok krt 2, Budapest, 1117, Hungary, pollux@math.bme.hu*

† *Institute of Informatics Problems FRC CSC RAS, 44k2 Vavilova street, Moscow, 119333, Russian Federation, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation, rrazumchik@ipiran.ru, razumchik_rv@rudn.university*

‡ *Department of Networked Systems and Services, Budapest University of Technology and Economics, Magyar Tudosok krt 2, Budapest, 1117, Hungary, telek@hit.bme.hu*

Abstract.

We seek to approximate the mean sojourn time in the processor sharing M/G/1 queue with inaccurate job size information. Suppose we are given the arrival rate λ and random service time $\hat{S} = SX$ where $X \sim LN(0, \sigma)$ represents the inaccuracy. Denote the mean sojourn time in an M/G/1 queue with processor sharing with service time $\hat{S}$ by $E(\hat{T}^{PS})$ and with service time S by $E(T^{PS})$. Finally, $E(\hat{T}^{Re})$ denotes the mean sojourn time of an M/G/1 queue with resampling service policy and service time distribution according to S. It can be shown that for exponential service time S, $E(T^{PS}) < E(\hat{T}^{Re}) < E(\hat{T}^{PS})$ holds for any $\sigma > 0$.

Keywords: inaccurate job size, long tails, service policy, size-based scheduling.

1. Introduction

Inaccurate job size information, when encountered in practice (in systems such as, for example, mapreduce), can lead to performance degradation of the size-based scheduling policies (see [1–3]). We understand the inaccurate job size information as described in [3]. Suppose a job arrives at the system. Upon arrival its (future) service time, say $\hat{S}$, is sampled from the known distribution $\hat{B}(x)$ and the value of $\hat{S}$ is used for scheduling the job. After the job has received service, it turned out that its service time was $S \neq \hat{S}$. The same thing happens with the next job etc. It means that the (true) service time is sampled from another unknown distribution with cumulative distribution function (cdf) $B(x)$. And the scheduler has to use the values of S instead of $\hat{S}$ when scheduling the jobs. The use of such inaccurate information by the scheduler may impact the system considerably. For example, the mean sojourn time in the M/G/1/ processor sharing (PS) queue with the cdf $B(x) = 1 - e^{-x}$ is given in Fig. 1 by the lowest solid line. The other two solid lines show the values of the mean sojourn in the M/G/1 PS queue, when $\hat{B}(x) \neq B(x)$ and the mean of $\hat{B}(x)$ is greater than $B(x)$.

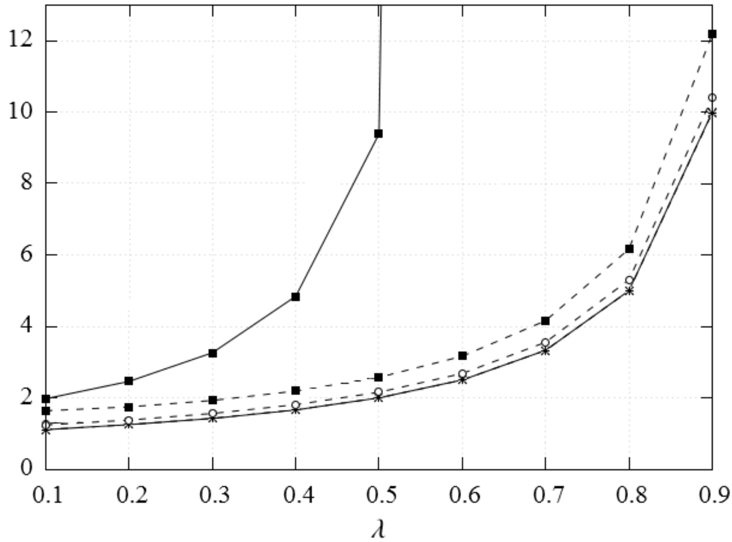

Figure 1. Mean sojourn times for different values of σ. Upper solid line represents the value of $E(\hat{T}^{PS})$, the lowest solid line — the value of $E(T^{PS})$. Dotted lines are the values of $E(\hat{T}^{Re})$.

Even without this example it is clear that the performance characteristics calculated from the mathematical models of the systems with inaccurate job size information will be biased. But is it possible to tune the mathematical model so as to reduce the bias? We claim that there are conditions when such tuning is possible.

2. Main results

Let us put everything into the mathematical perspective. Consider the following three systems running in parallel:

— classical M/G/1 PS queue with Poisson arrivals of rate λ and the service time S distributed as $B(x)$;

— classical M/G/1 PS queue with Poisson arrivals of rate λ and the service time $\hat{S}$ distributed as $\hat{B}(x)$;
— M/G/1 non-preemptive LIFO queue with resampling. Arrivals are Poisson of rate λ, the service time distribution is $S(x)$. Resampling means that upon arrival, the job currently in service resamples its service time according to $\hat{B}(x)$.

We assume that the service time S is exponential with mean μ and the service time $\hat{S}$ is $\hat{S} = SX$, where S and X are independent, X has a log-normal distribution with with parameters 0 and σ. If M/G/1 PS queue is the model of the real-life system, then the service times from the cdf $\hat{B}(x)$ we will call inaccurate. On average they are greater than those from the cdf $B(x)$.

Denote by $E(T^{PS})$, $E(\hat{T}^{PS})$ and $E(\hat{T}^{Re})$ denote correspondingly the job's mean sojourn times in each of the systems. It is known that in the classical M/G/1 PS is insensitive to the distribution of the service time, the total number of jobs depends only on the system load and, by Little's law, the mean sojourn time depends on the arrival rate and the system load:

$$E(T^{PS}) = \frac{\lambda E(S)}{1 - \lambda E(S)}, \quad E(\hat{T}^{PS}) = \frac{E(SX)}{1 - \lambda E(SX)}.$$

The expression for the $E(\hat{T}^{Re})$, which can also be derived from Little's law, depends on the whole distribution and follows from the paper [4]:

$$E(\hat{T}^{Re}) = \frac{1 - \hat{\beta}(\lambda)}{\lambda(2\hat{\beta}(\lambda) - 1)}.$$

Here $\hat{\beta}(\lambda)$ is the LST $\hat{\beta}(s)$ of $\hat{B}(x)$ at point $s = \lambda$.

It can be shown that for any $\sigma > 0$, the following inequalities always hold:

$$E(T^{PS}) < E(\hat{T}^{Re}) < E(\hat{T}^{PS}).$$

Moreover $E(\hat{T}^{Re}) \to E(T^{PS})$ in the limit as $\sigma \to 0$. The dotted lines in Fig. 1 show the behaviour of $E(\hat{T}^{Re})$ for two different values of σ.

Acknowledgments

This work is partially supported by the Russian Foundation for Basic Research (grants 15-07-03007 and 15-07-03406) and by grant K123914 of Hungarian Scientific Research Fund (OTKA).

References

1. Lu, D., Sheng, H., Dinda, P. 2004. Size-based scheduling policies with inaccurate scheduling information. Modeling, Analysis, and Simulation of Computer and Telecommunications Systems. (MASCOTS 2004). Proceedings. The IEEE Computer Society's 12th Annual International Symposium on. Pp.31–38.
2. Wierman, A., Nuyens, M. 2008. Scheduling despite inexact job-size information. In ACM SIGMETRICS Performance Evaluation Review. Vol. 36. Issue 1. Pp.25–36.
3. DellAmico, M., Carra, D., Michiardi, P. 2015. PSBS: Practical Size-Based Scheduling. IEEE Transactions on Computers. Issue. 99. Pp. 1–15.
4. Meykhanadzhyan, L., Razumchik, R. 2016. New Scheduling Policy For Estimation Of Stationary Performance Characteristics In Single Server Queues With Inaccurate Job Size Information. ECMS 2016: 710-716.