The resampling M/G/1 non-preemptive LIFO queue and its application to systems with uncertain service time*

Illés Horváth

MTA-BME Information Systems Research Group, Budapest, Hungary

Rostislav Razumchik

Peoples' Friendship University of Russia (RUDN), Moscow, Russsia

Miklós Telek

Dept. of Networked Systems and Services, Budapest University of Technology and Economics, Budapest, Hungary

Abstract

We introduce and analyze the $M/G/1$ resampling queue with non-preemptive LIFO policy, then we use it to provide bounds on the performance characteristics of an $M/G/1$ processor sharing queue with inaccurate service time information.

Keywords: inaccurate service time information, resampling queue, sojourn time distribution

1. Introduction

In this paper we present the detailed analysis of the main stationary characteristics of an $M/G/1$ queue with a new scheduling policy, referred to as non-preemptive LIFO with resampling, and show how these results may be useful in the performance evaluation of queueing systems, which are fed by customers with inaccurate job size information. Besides the analytical description of the considered queueing system (in Section 2) the only ingredient here, which needs clarification, is the notion of inaccurate job size information and the motivation behind it.

*The authors thank the insightful comments of the reviewers which helped to improve the paper. I. Horváth and M. Telek are supported by the OTKA 123914 project and R. Razumchik by the RFBR 18-07-00692 and 19-07-00739 projects.

Email addresses: horvath.illes.antal@gmail.com (Illés Horváth), rrazumchik@gmail.com (Rostislav Razumchik), telek@hit.bme.hu (Miklós Telek)

It is well-known that size-based scheduling policies allow one to greatly improve a system's performance compared to size-oblivious policies (see, for example, [1, Chapter 31]) and this gain comes almost for free in the sense that it is simple and inexpensive to alter the scheduling policy according to the job size information. Yet, those good size-based policies require the *exact* or *true* knowledge of the customers[1] service time which is not always possible in practice. In contrast to systems, with fixed job size (and thus the exact job execution time for a given job execution rate, is known), there are systems in which the relation between the job size (or any information about the job) and its execution time is not so straightforward. For example, in web servers the time required to service the requests can be only approximated by the file sizes, which are known to the server (see [2]). Similarly, [3] reports that for server scheduling in peer-to-peer networks estimates of request service times must be used. In systems where the file sizes do not serve as any indication of the request service times (as in MapReduce-like systems [4]), estimating procedures become more involved and lead only to approximate values. Whenever the *approximate* values of the service times are used for scheduling instead of the exact ones, we say that the job size information is inaccurate. If a size-based policy is fed by the inaccurate job size information, the system's performance may decrease drastically compared to the system's performance under policies which are not size-based (like Processor Sharing, PS). The numerical evidences for this intuitively expected result are given in [5], where it is shown that size-based policies (specifically SRPT and Fair Sojourn Protocol [6]) perform poorer and poorer (compared to PS, which is the commonly assumed size-oblivious policy to compare with size-aware policies) as the uncertainty about the job size distribution increases.

During the last two decades there appeared a number of research papers that studied the behaviour of scheduling policies in systems with inaccurate job size information. To our best knowledge in [7, 8, 4, 9, 10, 11] one can find the most recent results on the topic, including some reviews. The question, which is being usually addressed, is how to get rid of the inaccuracy and how to schedule the jobs properly.

Another question, which, to our best knowledge has not been addressed in the previous studies, but may provide a useful insight into the *true* system's behaviour, is to provide some meaningful upper and lower bounds for the true performance characteristics of the system based only on knowledge about the inaccurate service time distribution such that the bounds are better than simply treating the assumed service time distribution as accurate. We claim that there are conditions when such improved approximation is possible. In what follows we give these conditions and present the methodology which leads to the analytical expressions for the bounds. To our knowledge this is the first study in this direction and the conditions that we state are somewhat restrictive (we give bounds only for the true mean sojourn time and true mean service time under Poisson arrivals and single server queues). But we believe that the presented

[1] Throughout the paper we will use the terms customer, job, request interchangeably.

methodology may be a fruitful instrument in the study of more complex systems with inaccurate job size information.

The next section is devoted to the analysis of the introduced $M/G/1$ resampling queue, while Section 3 discusses how it can help in the approximation of the $M/G/1-PS$ queue with inaccurate service time information.

2. $M/G/1$ resampling queue with non-preemptive LIFO service

We consider non-preemptive LIFO with resampling in a usual M/G/1 environment: there is a single server and an infinite buffer, jobs arrive according to a Poisson arrival process with rate λ and the service time distribution is $B(x) = Pr(S < x)$. Resampling means that arriving customers resample the service time of the customer under service (if any) from distribution $B(x)$. This resampling policy is referred to as Preemptive-repeat-different in [12]. That is, if the remaining service time of the customer under service is θ at the arrival of a new customer then after the arrival the remaining service time is going to be S with distribution $B(x)$ (independent of any other elements of the model, including θ). The probability density function (PDF) of the service time is $b(x) = \frac{d}{dx}B(x)$ and their Laplace transforms are $B^*(s) = \int_0^\infty e^{-sx}B(x)dx$ and $b^*(s) = \int_0^\infty e^{-sx}b(x)dx$. Due to the fact that the service time is positive, we have $b^*(s) = sB^*(s)$. For later use we introduce the probability that an inter-arrival period is larger than the service time,

$$Pr(S < A) = \int_0^\infty Pr(S < A|S = x)b(x)dx = \int_0^\infty e^{-\lambda x}b(x)dx = b^*(\lambda),$$

which turns out to be an essential quantity in the analysis of our resampling queue.

2.1. Condition of stability

Due to the fact that an arrival renews the system we compute the mean number of customers served during an inter-arrival period, $E(W)$, assuming that a service starts right at the arrival and the server is always busy during the inter-arrival period. Let $W(\tau)$ be the number of served customers during an inter-arrival of length τ. For $E(W(\tau))$ we have

$$\begin{aligned} E(W(\tau)) &= \int_{x=0}^\infty E(W(\tau)|S = x)b(x)dx = \int_{x=0}^\tau (1 + E(W(\tau - x)))b(x)dx \\ &= B(\tau) + \int_{x=0}^\tau E(W(\tau - x))b(x)dx \end{aligned}$$

and for $W^*(s) = \int_{\tau=0}^\infty e^{-s\tau}E(W(\tau))d\tau$ we have

$$W^*(s) = B^*(s) + W^*(s)b^*(s) = B^*(s) + W^*(s)sB^*(s) = \frac{B^*(s)}{1 - sB^*(s)},$$

since $b^*(s) = sB^*(s)$. The unconditional measure is

$$E(W) = \int_{\tau=0}^{\infty} \lambda e^{-\lambda\tau} E(W(\tau))d\tau = \lambda W^*(s)|_{s=\lambda} = \left.\frac{\lambda B^*(s)}{1 - sB^*(s)}\right|_{s=\lambda}$$
$$= \frac{\lambda B^*(\lambda)}{1 - \lambda B^*(\lambda)} = \frac{b^*(\lambda)}{1 - b^*(\lambda)}. \quad (1)$$

The necessary and sufficient condition of stability is $E(W) > 1$, which gives $b^*(\lambda) > 1/2$. Since $b^*(\lambda) = E(e^{-\lambda S}) < 1$, the valid range of $b^*(\lambda)$ is $b^*(\lambda) \in (1/2; 1)$. For a given service time distribution $B(x) = Pr(S < x)$, $b^*(\lambda) = E(e^{-\lambda S})$ is a monotone decreasing function of λ, for $\lambda \geq 0$, where $b^*(0) = 1$. Let λ^* be such that $b^*(\lambda^*) = 1/2$. The valid range of the load for a stable resampling queue is $\lambda \in (0, \lambda^*)$, which is equivalent to $b^*(\lambda) \in (1/2; 1)$.

The results for $E(W(\tau))$ and $W^*(s)$ are also known from renewal theory, since $E(W(\tau))$ is the renewal function of the ordinary renewal process where the distribution of the inter-event time is the service time distribution. The same stability condition was obtained in [13] based on a branching process based approach.

In the special case when the service time is exponential with parameters μ and $E(e^{-sS}) = \frac{\mu}{s+\mu}$ we have $b^*(\lambda) = E(e^{-\lambda S}) = \frac{\mu}{\lambda+\mu}$ and $E(W) = \frac{\mu}{\lambda}$.

***Remark* 1.** *The stability region of non-resampling work conserving queues with the same arrival process and service time distribution is $\lambda \in (0, 1/E(S))$, which is different from $\lambda \in (0, \lambda^*)$, in general. Intuitively, resampling increases the overall service time if the distribution is ageing (the hazard rate function is monotone increasing) and decreases it if the distribution is de-ageing (the hazard rate function is monotone decreasing). In the former case $\lambda^* \leq 1/E(S)$, while in the latter case $\lambda^* \geq 1/E(S)$. When the hazard rate function of the service time is not monotone a detailed analysis of λ^* is required based on the particular service time distribution.*

2.2. Joint stationary distribution of the number of customers and the remaining service time

Let $\nu(t)$ be the number of customers in the system at time t and $\xi(t)$ be the remaining service time of the customer under service, which is resampled upon each customer arrival. For $n \geq 1$, we define

$$P_n(t, x) = Pr(\nu(t) = n, \xi(t) < x),$$

$p_n(t, x) = \partial P_n(t, x)/\partial x$ and

$$p_n(t) = Pr(\nu(t) = n) = \int_{x=0}^{\infty} p_n(t, x)dx.$$

For $n = 0$, there is no customer in the system and we define $p_0(t) = Pr(\nu(t) = 0)$.

Theorem 1. *Assuming the stationary measures* $\lim_{t\to\infty} p_n(t) = p_n$ *(for* $n \geq 0$*) and* $\lim_{t\to\infty} p_n(t,x) = p_n(x)$ *(for* $n \geq 1$*) exist, then*

$$p_0 = \frac{2b^*(\lambda) - 1}{b^*(\lambda)}, \tag{2}$$

and for $n \geq 1$*,*

$$p_n(x) \quad = \quad \frac{p_0}{b^*(\lambda)} \left(\frac{1 - b^*(\lambda)}{b^*(\lambda)} \right)^{n-1} \int_x^\infty \lambda e^{-\lambda(u-x)} b(u)\, du. \tag{3}$$

Proof. For $n \geq 1$, $p_n(t,x)$ satisfies the forward argument

$$p_n(t+\Delta, x) = p_n(t, x+\Delta)(1 - \lambda\Delta) + \lambda\Delta p_{n-1}(t)b(x) + p_{n+1}(t,0)\Delta b(x) + o(\Delta),$$

where $o(\Delta)$ is such that $\lim_{\Delta\to 0} \frac{o(\Delta)}{\Delta} = 0$. Subtracting $p_n(t, x+\Delta)$, dividing by Δ and making the $\Delta \to 0$ limit gives

$$\frac{\partial}{\partial t} p_n(t,x) - \frac{\partial}{\partial x} p_n(t,x) = -\lambda p_n(t,x) + \lambda p_{n-1}(t)b(x) + p_{n+1}(t,0)b(x). \tag{4}$$

As t tends to infinity and the process converges to the stationary distribution (4) becomes

$$-\frac{d}{dx} p_n(x) = -\lambda p_n(x) + \lambda p_{n-1} b(x) + p_{n+1}(0)b(x). \tag{5}$$

In stationary regime, the stationary transition rates from n to $n+1$ and from $n+1$ to n are equal, that is

$$\lambda p_n = p_{n+1}(0), \tag{6}$$

from which

$$-\frac{d}{dx} p_n(x) = -\lambda p_n(x) + \lambda p_{n-1} b(x) + \lambda p_n b(x). \tag{7}$$

We look for the solution in the form $p_n(x) = g_n(x)e^{\lambda x}$. Substituting it into (7) we get

$$g_n'(x) = -\lambda(p_{n-1} + p_n)e^{-\lambda x} b(x), \tag{8}$$

whose solution is

$$g_n(x) = \lambda(p_{n-1} + p_n) \int_x^\infty e^{-\lambda u} b(u) du. \tag{9}$$

Using $g_n(0) = p_n(0) = \lambda p_{n-1}$, for $g_n(0)$ we have

$$\lambda p_{n-1} = \lambda(p_{n-1} + p_n) \int_0^\infty e^{-\lambda u} b(u) du = \lambda(p_{n-1} + p_n) b^*(\lambda), \tag{10}$$

from which $p_n = p_{n-1}\frac{1-b^*(\lambda)}{b^*(\lambda)}$ and

$$p_n = p_0 \left(\frac{1-b^*(\lambda)}{b^*(\lambda)}\right)^n. \tag{11}$$

$\sum_{n=0}^{\infty} p_n = 1$ yields (2) and (9) with $p_n(x) = g_n(x)e^{\lambda x}$ yields (3). □

***Remark* 2.** *Some of the above proved results can be obtained from a simple balance argument. In the Markov chain embedded into the resampling non-preemptive $M/G/1/LIFO$ queue at population change points there might be two kinds of transitions: new arrival before service completion with probability $1 - b^*(\lambda)$ and service completion before a new arrival, with probability $b^*(\lambda)$. For the stationary distribution of this embedded Markov chain, $\tilde{p}_n$, we have $\tilde{p}_{n-1}(1-b^*(\lambda)) = \tilde{p}_n b^*(\lambda)$. The mean time while the resampling non-preemptive $M/G/1/LIFO$ queue has n ($n > 0$) customers is independent of n, which yields $p_{n-1}(1 - b^*(\lambda)) = p_n b^*(\lambda)$ and (11).*

Similarly, using the memoryless property and the probability that a new arrival occurs before service completion is $1-b^*(\lambda)$ we have $E(W) = (1-b^*(\lambda))0+ b^*(\lambda)(1 + E(W))$ which yields (1).

2.3. Stationary sojourn time

In the resampling non-preemptive $M/G/1/LIFO$ queue we denote the customer sojourn time by T, the time a customer spends in the server by V and the length of a busy period by U. Their Laplace transforms (LTs) are denoted by $\chi(s) = E(e^{-sT})$, $\psi(s) = E(e^{-sV})$ and $u(s) = E(e^{-sU})$, respectively.

Theorem 2. *$\chi(s)$ satisfies*

$$\chi(s) = \psi(s)p_0 + u(s)\psi(s)(1 - p_0), \tag{12}$$

where p_0 is given in (2),

$$\psi(s) = \frac{b^*(\lambda + s)(\lambda + s)}{s + \lambda b^*(\lambda + s)}, \tag{13}$$

and

$$u(s) = \frac{\lambda + s - \sqrt{[\lambda + s]^2 - 4\lambda[1 - b^*(s + \lambda)]b^*(s + \lambda)[\lambda + s]}}{2\lambda[1 - b^*(s + \lambda)]}. \tag{14}$$

Proof. Since the LIFO service is non-preemptive, if a customer arrives to an idle system, whose probability is p_0, its system time is identical with the time a customer spends in the server, V. If a tagged customer arrives to a busy system, whose probability is $1 - p_0$, its system time is identical with the time to reduce the number of customers in the system by one, U_b, plus the time the tagged customer spends in the server, V. (12) reflects this relation with the following consideration.

Since the system renews at each arrival the time to reduce the number of customers in the system by one after an arrival does not depend on the number of customers in the system and consequently has the same distribution as the busy period, which is the period when after an arrival there is one customer in the system until the number of customers reduces to zero, i.e., $U_b \equiv U$.

Let $\psi(s|S=\tau)$ be the LT of the time a customer spends in the server whose initial work requirement at arrival is τ. This work requirement is resampled upon each new arrival. Depending on the time of the first arrival A, for $\psi(s|S=\tau)$ we can write

$$\begin{aligned}\psi(s|S=\tau) &= e^{-s\tau}\underbrace{e^{-\lambda\tau}}_{A>\tau} + \int_{x=0}^{\tau} e^{-sx}\psi(s)\underbrace{\lambda e^{-\lambda x}}_{A=x<\tau}dx \\ &= e^{-(s+\lambda)\tau} + \lambda\psi(s)\frac{1-e^{-(s+\lambda)\tau}}{s+\lambda}\ ,\end{aligned}$$

where the first term represents the case when the arrival happens after completing the service of length τ and the second term represents the case that an arrival at time x resamples the service requirement of the customer in the server. Unconditioning according to the service time distribution provides

$$\begin{aligned}\psi(s) &= \int_{\tau=0}^{\infty}\psi(s|S=\tau)b(\tau)d\tau \\ &= \int_{\tau=0}^{\infty} e^{-(s+\lambda)\tau}b(\tau)d\tau + \lambda\psi(s)\int_{\tau=0}^{\infty}\int_{x=0}^{\tau} e^{-(s+\lambda)x}dx b(\tau)d\tau \\ &= b^*(s+\lambda) + \lambda\psi(s)\frac{1-b^*(s+\lambda)}{s+\lambda},\end{aligned}$$

which is equivalent with (13).

Let $u(s|S=\tau)$ be the LT of the busy period when the work requirement of the customer is τ. Depending on the time of the first arrival, denoted by A, for $u(s|S=\tau)$ we can write

$$u(s|S=\tau) = e^{-s\tau}\underbrace{e^{-\lambda\tau}}_{A>\tau} + \int_{x=0}^{\tau} e^{-sx}u^2(s)\underbrace{\lambda e^{-\lambda x}}_{A=x<\tau}dx$$

from which

$$\begin{aligned}
u(s) &= \int_{\tau=0}^{\infty} u(s|S=\tau)b(\tau)d\tau \\
&= \int_{\tau=0}^{\infty} e^{-(s+\lambda)\tau}b(\tau)d\tau + \lambda u^2(s)\int_{\tau=0}^{\infty}\int_{x=0}^{\tau} e^{-(s+\lambda)x}dx b(\tau)d\tau \\
&= b^*(s+\lambda) + \lambda u^2(s)\int_{x=0}^{\infty} e^{-(s+\lambda)x}\underbrace{\int_{\tau=x}^{\infty} b(\tau)d\tau}_{1-B(x)}dx \\
&= b^*(s+\lambda) + \lambda u^2(s)\left(\frac{1}{s+\lambda} - B^*(s+\lambda)\right) \\
&= b^*(s+\lambda) + \lambda u^2(s)\frac{1-b^*(s+\lambda)}{s+\lambda}.
\end{aligned}$$

This equation has two solutions

$$u_{1,2}(s) = \frac{\lambda + s \mp \sqrt{[\lambda+s]^2 - 4\lambda[1-b^*(s+\lambda)]b^*(s+\lambda)[\lambda+s]}}{2\lambda[1-b^*(s+\lambda)]}.$$

The valid solution should satisfy $u(0) = 1$ and $u(s) = E(e^{-sU}) \leq 1$ for real and positive s, from which $u(s) = u_1(s)$ is the valid solution when $b^*(\lambda) \in (1/2; 1)$. □

Corollary 1.

$$E(V) = -\psi'(0) = \frac{1-b^*(\lambda)}{\lambda b^*(\lambda)}. \tag{15}$$

Proof. The derivative of the LT at $s = 0$ provides the expression after some algebra. □

Corollary 2.

$$E(T) = E(U) = \frac{1-b^*(\lambda)}{\lambda(2b^*(\lambda)-1)}. \tag{16}$$

Proof. (16) can also be obtained from the corresponding LTs, but we present another proof that provides a better insight to the relation to $E(V)$ using a coupling and an insensitivity argument.

Consider an M/G/1 preemptive LIFO queue with resume policy with service time distribution according to V. Denote the response time and length of the busy period for this queue by T' and U', respectively. We argue that

$$E(T) = E(T') \qquad \text{and} \qquad E(U) = E(U'), \tag{17}$$

where $E(T)$ and $E(U)$ are the mean response and mean length of the busy period for the M/G/1 non-preemptive resampling LIFO queue as in (16).

To prove (17), for each realization of the busy period in the M/G/1 non-preemptive resampling LIFO queue, we couple a realization of the busy period of the M/G/1 preemptive LIFO queue with resume policy with service time distribution according to V. The coupling has the following properties:

- arrival times are identical;
- service completion times are also identical, but possibly belong to different jobs;
- the next job in the server (after a service completion or an arrival) depends on the service policy of either queue.

An example is presented in Figure 1.

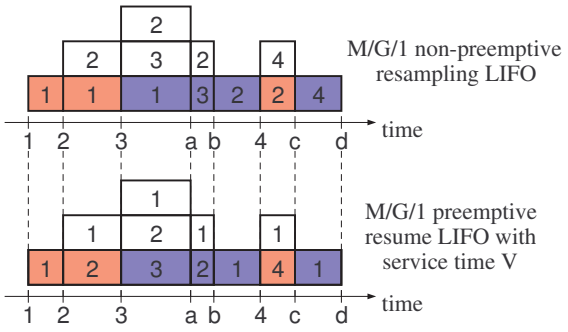

Figure 1: Coupling of M/G/1 non-preemptive resampling LIFO queue and M/G/1 preemptive LIFO queue without resampling with service time V

1, 2, 3 and 4 in Figure 1 mark the times of arrival of jobs 1, 2, 3 and 4, respectively, while a, b, c and d mark service completions. For each queue, jobs in the server are in the bottom row and jobs waiting in queue are in rows above.

All endpoints of the intervals are renewal points; each interval will either end by an arrival (red intervals in Figure 1) or a service completion (blue intervals). The lengths of red intervals have the same distribution, and the lengths of blue intervals also have the same distribution (but different from red). Red intervals are only coupled to red intervals and blue intervals are only coupled to blue intervals.

The total server time of a job in either queue has the same distribution: it is equal to the total length of a geometric number of intervals (the last of which is blue and the rest are red). The parameter of the geometric distribution is the probability that an arrival occurs before service is finished. In the top queue, the distribution of the total server time of a job is according to V, so in the bottom queue, the service time distribution is also according to V.

In Figure 1, the total width corresponds to the length of the busy period. This is equal for both queues per the coupling, hence U and U' are identically distributed and $E(U) = E(U')$.

The total area corresponds to the total sojourn time of all jobs in the busy period. This is also equal for the coupled queues, along with the total number of jobs, hence the average sojourn times (within the busy period) are also equal, so $E(T) = E(T')$. Note that since the coupling does not match the sojourn times of individual jobs, the sojourn time distribution can be different.

With (17) proven, we compute $E(T')$ and $E(U')$ next. The M/G/1 queue with preemptive LIFO policy is insensitive [14], and the classic M/M/1 queue with service rate $\mu = \frac{1}{E(V)}$ is a special case for which $E(T') = E(U') = \frac{1}{\mu - \lambda}$, so for the general M/G/1 queue with preemptive LIFO policy, we have

$$E(T') = E(U') = \frac{1}{\frac{1}{E(V)} - \lambda} \tag{18}$$

accordingly. Putting (15) into (18) gives (16). □

***Remark* 3.** *Based on* (11) *the mean number of customers in the system is equal to* $\sum_{n=1}^{\infty} np_n = \frac{1-b^*(\lambda)}{2b^*(\lambda)-1}$*, from which we see that Little's law holds in this system. The other way to make sure that this conservation law applies here is to use the general sample-path results stated, for example, in [15, Chapter 6]. Since* λ*,* $\pi'(1)$ *and* $E(T)$ *are finite and the system becomes empty infinitely often when* $b^*(\lambda) \in (1/2, 1)$*, then according to [15, Theorem 6.1]* $\pi'(1) = \lambda E(T)$*.*

***Remark* 4.** *For the comparison of LIFO policy with resampling and FIFO policy with resampling, we note that*

- *the stationary distribution for the number of customers in the system,* p_n*, remains the same as in* (11)*;*
- *the server time (and its LT* $\psi(s)$*) remains the same as in* (13)*;*
- *the customer sojourn time is different; for FIFO policy, instead of* (12) *we have*

$$\chi^{FIFO}(s) = \sum_{n=0}^{\infty} p_n \psi^{n+1}(s) = \frac{(2b^*(\lambda) - 1)\psi(s)}{b^*(\lambda) - (1 - b^*(\lambda))\psi(s)}. \tag{19}$$

  (19) *represents the fact that if a customer finds* n *other customers in the system upon its arrival (whose probability is* p_n *according to the PASTA property) its sojourn time is composed by the service of the previous* n *customers and its own service. Where we also used that upon a customer arrival the remaining time the customer in service spends in the server is distributed according to* $\psi(s)$ *independent of the past.*

- *the mean customer sojourn time is the same as for the LIFO policy*

$$-\chi^{FIFO'}(0) = \frac{1 - b^*(\lambda)}{\lambda(2b^*(\lambda) - 1)}. \tag{20}$$

We define the *slowdown* as the ratio of the mean sojourn time of a customer whose work requirement upon arrival is x and x, that is $\frac{E(T|S=x)}{x}$.

Corollary 3. *The slowdown can be computed from*

$$\chi(s|S=x) = \psi(s|S=x)p_0 + u(s)\psi(s|S=x)(1-p_0),$$

where

$$\psi(s|S=x) = E(e^{-sT}|S=x) = e^{-(\lambda+s)x} + \frac{\lambda}{\lambda+s}\psi(s)[1-e^{-(\lambda+s)x}] \ ,$$

and

$$\begin{aligned} E(T|S=x) &= -[\psi(s|S=x)p_0 + u(s)\psi(s|S=x)(1-p_0)]'_{s=0} \\ &= \frac{b^*(\lambda)}{\lambda(2b^*(\lambda)-1)} - e^{-\lambda x}\frac{1}{\lambda b^*(\lambda)}. \end{aligned} \tag{21}$$

Proof. The LTs are obtained already in the proof of theorem 2. For $E(T|S=x) = -\chi'(0|S=x)$, the derivatives of the LT at $s=0$ give the mean values with straightforward algebra. □

***Remark* 5.** *The slowdown of resampling queues essentially differs from the slowdown of non-resampling queues. According to* (21)*,*

$$\lim_{x\to 0}\frac{E(T|S=x)}{x} = \infty \ \textit{and} \ \lim_{x\to\infty}\frac{E(T|S=x)}{x} = 0,$$

while for PS queues the slowdown is independent of x and it is

$$\frac{E(T|S=x)}{x} = \frac{1}{1-\lambda E(S)} > 1.$$

3. Approximating $M/G/1$ PS queue with inaccurate service time information

We intend to investigate the sojourn time in the $M/G/1/PS$ queue with inaccurate knowledge on the service time distribution. The $M/G/1/PS$ queue is one of the basic models used to study the impact of inaccurate service time in applications. Additionally, we are going to relate this sojourn time of interest with the sojourn time spent in the $M/G/1$ resampling queue with non-preemptive LIFO service, which was studied in the previous section. Following [4] we apply the next assumptions.

- the theoretical service time is S with CDF, PDF and LT, $B(x)$, $b(x)$ and $b^*(s) = E(e^{-sS})$, respectively.

- the inaccurate knowledge on the service time is $\hat{S} = SX$, where S and X are independent and X has a *log-symmetric* distribution with PDF $\varepsilon(x)$. Log-symmetric is defined as $X = e^Y$, where Y has a symmetric distribution with PDF $g(y) = g(-y)$. We note that any log-symmetric distribution has $E(X) = E(1/X) \geq 1$.

  Our assumption about the multiplicative model is based on the recently reported results that the job sizes (in MapReduce-like systems) and errors are related in a multiplicative way [4, Section 5.3] and the error distribution is log-normal [16][2]. For the theoretical bounds presented in Theorems 3, 4 and 5, we allow for any log-symmetric error distribution (not just log-normal).

  We denote the CDF, PDF and LT of $\hat{S}$ by $\hat{B}(x)$, $\hat{b}(x)$, and $\hat{b}^*(s)$, respectively.

In the rest of this section we assume that S is exponentially distributed, but due to the multiplicative error model $\hat{S}$ is not exponential.

3.1. Bounding the mean sojourn time

Let T^{PS} denote the stationary sojourn time in the M/G/1 system with arrival rate λ service time S and processor sharing discipline, and similarly, let $\hat{T}^{Re}$ ($\hat{T}^{PS}$) denote the stationary sojourn time in the M/G/1 system with arrival rate λ, service time $\hat{S}$ and non-preemptive LIFO resampling (processor sharing) discipline.

Theorem 3. *Independent of the mean service time ($E(S)$), the mean sojourn times obey the following relation*

$$E(T^{PS}) \leq E(\hat{T}^{Re}) \leq E(\hat{T}^{PS}). \tag{22}$$

Proof. The mean sojourn time in the $M/G/1/PS$ queue with the original and the observed service time are $E(T^{PS}) = \frac{E(S)}{1-\lambda E(S)}$ and $E(\hat{T}^{PS}) = \frac{E(SX)}{1-\lambda E(SX)}$. From (16), we have $E(\hat{T}^{Re}) = \dfrac{1-\hat{b}^*(\lambda)}{\lambda(2\hat{b}^*(\lambda)-1)}$. That is, we need to show that

$$\frac{E(S)}{1-\lambda E(S)} \leq \frac{1-\hat{b}^*(\lambda)}{\lambda(2\hat{b}^*(\lambda)-1)} \leq \frac{E(SX)}{1-\lambda E(SX)}, \tag{23}$$

which is equivalent to

$$\frac{1}{1+\lambda E(S)E(X)} \leq \hat{b}^*(\lambda) \leq \frac{1}{1+\lambda E(S)}. \tag{24}$$

[2]Consequently, the distribution of the product of the job size and the error is long-tailed.

To show that $\hat{b}^*(\lambda) \geq \frac{1}{1+\lambda E(S)E(X)}$, using Jensen's inequality we write

$$\begin{aligned}\hat{b}^*(\lambda) &= E(e^{-\lambda SX}) = \int_0^\infty E(e^{-\lambda SX}|S=y)b(y)dy = \int_0^\infty E(e^{-\lambda yX})b(y)dy \\ &\geq \int_0^\infty e^{-\lambda yE(X)}b(y)dy = \int_0^\infty e^{-\lambda yE(X)} \underbrace{\frac{1}{E(S)}e^{-\frac{1}{E(S)}y}}_{S:Exp(1/E(S))} dy \\ &= \frac{1}{\lambda E(S)E(X)+1}.\end{aligned}$$

To prove $\hat{b}^*(\lambda) \leq \frac{1}{1+\lambda E(S)}$, we write

$$\begin{aligned}\hat{b}^*(\lambda) &= E(e^{-\lambda SX}) = \int_0^\infty E(e^{-\lambda XS}|X=x)\varepsilon(x)dx = \int_0^\infty E(e^{-\lambda xS})\varepsilon(x)dx \\ &= \int_0^\infty \underbrace{\frac{1/E(S)}{1/E(S)+\lambda x}}_{S:Exp(1/E(S))}\varepsilon(x)dx.\end{aligned}$$

Now we change variables as $\ln(x) = y$, by which $x = e^y$ and $dx = e^y dy$, and get

$$\hat{b}^*(\lambda) = \int_{-\infty}^\infty g(y)\frac{1}{1+\lambda E(S)e^y}dy. \tag{25}$$

Since $g(y)$ is an even function of y we can reduce the integration to $(0,\infty)$ as

$$\begin{aligned}\hat{b}^*(\lambda) &= \int_0^\infty g(y)\left(\frac{1}{1+\lambda E(S)e^{-y}} + \frac{1}{1+\lambda E(S)e^y}\right)dy \\ &= \int_0^\infty g(y)\left(\frac{2+\lambda E(S)[e^y+e^{-y}]}{1+(\lambda E(S))^2+\lambda E(S)[e^{-y}+e^y]}\right)dy \\ &= \int_0^\infty g(y)\left(\frac{2+2\lambda E(S)\cosh(y)}{1+(\lambda E(S))^2+2\lambda E(S)\cosh(y)}\right)dy,\end{aligned}$$

where $\cosh(y)$ denotes the hyperbolic cosine, $\cosh(y) = \frac{1}{2}(e^{-y}+e^y)$.

It is more convenient to rewrite the last expression in the following form

$$\begin{aligned}\hat{b}^*(\lambda) &= \frac{1}{1+\lambda E(S)}\int_0^\infty 2g(y)\left(\frac{[1+\lambda E(S)\cosh(y)][1+\lambda E(S)]}{1+(\lambda E(S))^2+2\lambda E(S)\cosh(y)}\right)dy \\ &= \frac{1}{1+\lambda E(S)}\int_0^\infty 2g(y)\underbrace{\left(\frac{1+\lambda E(S)+\cosh(y)[\lambda E(S)+(\lambda E(S))^2]}{1+(\lambda E(S))^2+2\lambda E(S)\cosh(y)}\right)}_{h(y)}dy \\ &= \frac{1}{1+\lambda E(S)}\int_0^\infty 2g(y)h(y)dy.\end{aligned}$$

Now we show that $0 \leq h(y) \leq 1$ for $y \geq 0$. The non-negativity of $h(y)$ is given from the fact that both the numerator and the denominator are non-negative.

From $1 \leq \cosh(y) = \frac{1}{2}(e^{-y} + e^{y})$, $\lambda E(S) < 1$ and $\lambda E(S) - (\lambda E(S))^2 > 0$ we have

$$\lambda E(S) - (\lambda E(S))^2 < \cosh(y)[\lambda E(S) - (\lambda E(S))^2].$$

Adding $1 + \lambda E(S)\cosh(y)$ to both sides we get

$$1 + \lambda E(S) + \cosh(y)[\lambda E(S) + (\lambda E(S))^2] < 1 + (\lambda E(S))^2 + 2\lambda E(S)\cosh(y),$$

which shows that the denominator of $h(y)$ is greater than its numerator, that is, $0 \leq h(y) \leq 1$ for $y \geq 0$.

Thus, since $\int_0^\infty 2g(y)dy = 1$, we have

$$\begin{aligned} \hat{b}^*(\lambda) &= \frac{1}{1+\lambda E(S)} \int_0^\infty 2g(y)h(y)dy \\ &\leq \frac{1}{1+\lambda E(S)} \int_0^\infty 2g(y)dy = \frac{1}{1+\lambda E(S)}. \end{aligned} \tag{26}$$

□

***Remark* 6.** *The importance of Theorem 3 comes from the fact that when the theoretical service time is exponentially distributed then the mean sojourn time in the $M/G/1/PS$ queue, $E(T^{PS})$, is better approximated with the mean sojourn time in the $M/G/1$ non-preemptive LIFO resampling queue with the observed sojourn time, $E(\hat{T}^{Re})$, than with the mean sojourn time in the $M/G/1/PS$ queue with the observed sojourn time distribution, $E(T^{PS})$.*

Since Little's law holds for the PS system, and according to (16), also for the resampling queue, we have a double inequality, similar to (22), for the mean number of customers in the system

$$E(N^{PS}) \leq E(\hat{N}^{Re}) \leq E(\hat{N}^{PS}). \tag{27}$$

Theorem 4. *The PDF of $\hat{S} = SX$ at zero satisfies*

$$\frac{1}{\hat{b}(0)} \leq E(S).$$

Proof. The CDF and the PDF of $\hat{S} = SX$ are

$$\begin{aligned} \hat{B}(x) = Pr(SX < x) &= \int_0^\infty \varepsilon(z) Pr(SX < x | X = z) dz = \int_0^\infty \varepsilon(z) Pr(S < x/z) dz \\ &= \int_0^\infty \varepsilon(z) \left[1 - e^{-\frac{x/z}{E(S)}}\right] dz = 1 - \int_0^\infty \varepsilon(z) e^{-\frac{x}{E(S)z}} dz. \end{aligned}$$

$$\hat{b}(x) = \hat{B}'(x) = \frac{1}{E(S)} \int_0^\infty \frac{\varepsilon(z)}{z} e^{-\frac{x}{yE(S)}} dy,$$

and, for $\hat{b}(0)$, we have

$$\hat{b}(0) = \frac{1}{E(S)} \int_0^\infty \frac{1}{z}\, \varepsilon(z) dz = \frac{E(1/X)}{E(S)}.$$

Since X has log-symmetric distribution, $1/X$ has the same distribution and $E(1/X) = E(X) \geq 1$. As a result,

$$\hat{b}(0)E(S) = E(1/X) \geq 1. \tag{28}$$

□

Corollary 4. *The mean sojourn time, $E(T^{PS})$, in $M/G/1/PS$ queue with inaccurate service time distribution, $\hat{B}(x)$, is bounded by*

$$\frac{1}{\hat{b}(0) - \lambda} < E(T^{PS}) = \frac{E(S)}{1 - \lambda E(S)} < \frac{1 - \hat{b}^*(\lambda)}{\lambda(2\hat{b}^*(\lambda) - 1)}. \tag{29}$$

Proof. The right inequality in (29) virtually repeats the statement of Theorem 3, more precisely the left inequality of (23). The left inequality is a straightforward consequence of Theorem 4. □

3.2. Bounds on further measures

Denote by $F^{PS}(x)$, $\hat{F}^{PS}(x)$ the CDF of the stationary number of customers in the $M/G/1/PS$ queue with service time S and $\hat{S} = SX$, respectively, and by $\hat{F}^{Re}(x)$ the CDF of the total number of customers in the $M/G/1$ non-preemptive LIFO resampling queue with service time $\hat{S}$.

Corollary 5. *The following stochastic order applies*

$$F^{PS}(x) \succ \hat{F}^{Re}(x) \succ \hat{F}^{PS}(x). \tag{30}$$

Proof. The $M/G/1/PS$ queue is known to be insensitive [17] and the stationary distribution of the number of customers is geometric [18, Section 5]. From (11) we also have that the stationary distribution of the number of customers in the $M/G/1$ non-preemptive LIFO resampling queue is geometric. Their stochastic ordering is determined by their mean in (27). □

An other consequence of Corollary 5 is

$$E\left(N^{PS\,k}\right) \leq E\left(\hat{N}^{Re\,k}\right) \leq E\left(\hat{N}^{PS\,k}\right),\ k \geq 1.$$

Inspired by Theorem 3, a similar result is proved for the variance in the next theorem.

Theorem 5. *Independent of the mean service time ($E(S)$), the mean sojourn times satisfy*

$$\mathrm{Var}(T^{PS}) \leq \mathrm{Var}(\hat{T}^{Re}). \tag{31}$$

Proof. The variance of the sojourn time in an $M/M/1/PS$ queue is provided in [19]

$$\mathrm{Var}(T^{PS}) = \frac{(ES)^2}{(1-\lambda ES)^2}\frac{2+\lambda ES}{2-\lambda ES}, \tag{32}$$

while $\mathrm{Var}(\hat{T}^{Re})$ is obtained from $\mathrm{Var}(\hat{T}^{Re}) = \hat{\chi}''(0) - \hat{\chi}'(0)^2$, with $\hat{\chi}(s)$ coming from (12), such that the service time, with Laplace transform $b^*(s)$, corresponds to the inaccurate service time Laplace transform $\hat{b}^*(s)$. Expanding $\hat{\chi}$, we get

$$E\left(\hat{T}^{Re^2}\right) = \hat{\chi}''(0) = \tag{33}$$
$$= \frac{2\left(1-3\hat{b}^*(\lambda)+3\hat{b}^*(\lambda)^2\right)\left(\hat{b}^*(\lambda)^2(1-\hat{b}^*(\lambda))+\lambda(2\hat{b}^*(\lambda)-1)\hat{b}^{*'}(\lambda)\right)}{\lambda^2\hat{b}^*(\lambda)^2(2\hat{b}^*(\lambda)-1)^3}$$

We remark that $\hat{b}^{*''}(\lambda)$ is missing from (33); this is in accordance with $(b^*)'(\lambda)$ missing from (16).

To estimate (33), we need bounds on $\hat{b}^*(\lambda)$ and $\hat{b}^{*'}(\lambda)$. From (26) we already have an upper bound on $\hat{b}^*(\lambda)$, whose obvious lower bound is zero. The next lemma provides a similar bound for $\hat{b}^{*'}(\lambda)$.

***Lemma* 1.**

$$-\frac{1}{4\lambda} \le \hat{b}^{*'}(\lambda) \le 0. \tag{34}$$

Proof. We give an estimate for $\hat{b}^{*'}(\lambda)$ using a technique similar to the proof of Theorem 3. From (25) we get that

$$\begin{aligned}\hat{b}^{*'}(\lambda) &= \frac{d}{d\lambda}\int_{-\infty}^{\infty} g(y)\frac{1}{1+\lambda E(S)e^y}dy = \\ &= \int_{-\infty}^{\infty} g(y)\frac{-E(S)e^y}{(1+\lambda E(S)e^y)^2}dy = \\ &= -\int_0^{\infty} g(y)\left(\frac{E(S)e^y}{(1+\lambda E(S)e^y)^2}+\frac{E(S)e^{-y}}{(1+\lambda E(S)e^{-y})^2}\right)dy. \end{aligned}\tag{35}$$

Next we prove that

$$0 \le \frac{E(S)e^y}{(1+\lambda E(S)e^y)^2}+\frac{E(S)e^{-y}}{(1+\lambda E(S)e^{-y})^2} \le \frac{1}{2\lambda}. \tag{36}$$

The first inequality is trivial; for the second inequality, we use the straightforward inequality

$$\frac{x}{(1+x)^2} \le \frac{1}{4}$$

and set $x = \lambda E(S)e^y$ and $x = \lambda E(S)e^{-y}$ respectively to obtain both terms in (36).

Putting (36) into (35) we obtain (34). □

Going on with the proof of Theorem 5, we examine the sign of the coefficient of $\hat{b}^{*'}(\lambda)$ in (33).

- the denominator $\lambda^2\hat{b}^*(\lambda)^2(2\hat{b}^*(\lambda)-1)^3$ is positive;
- $1-3\hat{b}^*(\lambda)+3\hat{b}^*(\lambda)^2$ is also positive since $1-3x+3x^2>0$ on the entire real line, and
- $2\hat{b}^*(\lambda)-1$ is also positive since $\hat{b}^*(\lambda)>1/2$.

Altogether, the coefficient of $\hat{b}^{*'}(\lambda)$ in (33) is positive, so by replacing $\hat{b}^{*'}(\lambda)$ by $-\frac{1}{4\lambda}$ according to Lemma 1, we obtain a lower bound on $\mathrm{Var}(\hat{T}^{Re})$:

$$\begin{aligned}\mathrm{Var}(\hat{T}^{Re}) &= E\left(\hat{T}^{Re^2}\right) - E\left(\hat{T}^{Re}\right)^2 \\ &\geq \frac{1}{\lambda^2\hat{b}^*(\lambda)^2(2\hat{b}^*(\lambda)-1)^3}\Bigg[(2\hat{b}^*(\lambda)-1)(1-\hat{b}^*(\lambda))^2\hat{b}^*(\lambda)^2 \\ &\qquad + 2\left(1-3\hat{b}^*(\lambda)+3\hat{b}^*(\lambda)^2\right)\left(\frac{1}{4}-\frac{\hat{b}^*(\lambda)}{2}-\hat{b}^*(\lambda)^2+\hat{b}^*(\lambda)^3\right)\Bigg] \\ &= \frac{1}{\lambda^2}\cdot\underbrace{\frac{1-5x+3x^2+18x^3-34x^4+16x^5}{2x^2(2x-1)^3}}_{f_1(x)}\Bigg|_{x=\hat{b}^*(\lambda)}. \end{aligned} \tag{37}$$

We transform (32) in a similar manner:

$$\mathrm{Var}(T^{PS}) = \frac{(ES)^2}{(1-\lambda ES)^2}\frac{2+\lambda ES}{2-\lambda ES} = \frac{1}{\lambda^2}\cdot\underbrace{\frac{(1-y)^2(1+y)}{(2y-1)^2(3y-1)}}_{f_2(y)}\Bigg|_{y=\frac{1}{1+\lambda E(S)}}. \tag{38}$$

All that remains to prove Theorem 5 is to check that

$$f_2(y)\leq f_1(x) \quad \text{holds for any} \quad \frac{1}{2}<x\leq y<1. \tag{39}$$

This can be done by elementary calculations:

$$f_2'(y) = \frac{(1-y)(3-8y+y^2)}{(1-3y)^2(2y-1)^3} < 0 \quad \text{for } \frac{1}{2}<y<1$$

implies that $f_2(y)$ is decreasing, and

$$f_1(x)-f_2(x) = \frac{1-8x+20x^2+3x^3-86x^4+124x^5-52x^6}{2x^2(2x-1)^3(3x-1)} > 0;$$

implies that $f_1(x)>f_2(x)$ for $\frac{1}{2}<x<1$ holds (since the numerator does not have any roots in $\left(\frac{1}{2},1\right)$), from which (39) and Theorem 5 follows. $\square$

***Remark* 7.** *In [19] Theorem 2.13 states that the processor sharing discipline is monotone with respect to the variance of the sojourn time. Thus* $\mathrm{Var}(S) \leq \mathrm{Var}(\hat{S})$ *directly implies*

$$\mathrm{Var}(T^{PS}) \leq \mathrm{Var}(\hat{T}^{PS}).$$

As a result, Theorem 5 can be interpreted as

$$\mathrm{Var}(T^{PS}) \leq \min(\mathrm{Var}(\hat{T}^{Re}), \mathrm{Var}(\hat{T}^{PS})),$$

but the computation of $\mathrm{Var}(\hat{T}^{PS})$*, which is provided as a solution of an integral equation in [19], is not that simple.*

The slowdown for the $M/G/1/PS$ and the $M/G/1/Re$ queues are defined as $E(T^{PS}/S)$, $E(\hat{T}^{PS}/\hat{S})$, $E(\hat{T}^{Re}/\hat{S})$. For $E(T^{PS}/S)$ in $M/G/1/PS$ queue we have [20]

$$E(T^{PS}/S) = \int_x E(T^{PS}|S=x)\frac{1}{x}b(x)dx = \int_x \frac{x}{1-\lambda E(S)}\frac{1}{x}b(x)dx = \frac{1}{1-\lambda E(S)}$$

and, similarly, $E(\hat{T}^{PS}/\hat{S}) = \frac{1}{1-\lambda E(S)E(X)}$, from which

$$E(T^{PS}/S) \leq E(\hat{T}^{PS}/\hat{S}). \tag{40}$$

For the $M/G/1$ resampling queue

$$E(\hat{T}^{Re}/\hat{S}) = \int_x E(\hat{T}^{Re}|\hat{S}=x)\frac{1}{x}\hat{b}(x)dx = \infty,$$

because $\lim_{x\to 0} E(\hat{T}^{Re}|S=x)$ and $\lim_{x\to 0}\hat{b}(x)$ are positive constants according to (21) and (28).

4. Conclusion

In this work we studied a single server processor sharing queue with inaccurate service time information and we obtained a very special property, namely that the M/G/1 non-preemptive LIFO queue with inaccurate service time information approximates better the processor sharing queue with accurate service time, than the associated processor sharing queue with inaccurate service time, when the accurate service time distribution is exponential.

An intuitive interpretation of this property comes from the nature of resampling. When the service time whose distribution has a decreasing (increasing) hazard rate is resampled then the resampled service time is smaller (larger), in stochastic ordering sense, than the remaining service time at resampling.

As a consequence, on one hand, for a service time distribution with strictly monotone increasing hazard rate, the resampling queue overestimates the mean response time of the PS queue with accurate service time. On the other hand, above a certain load, the resampling queue underestimates the mean response

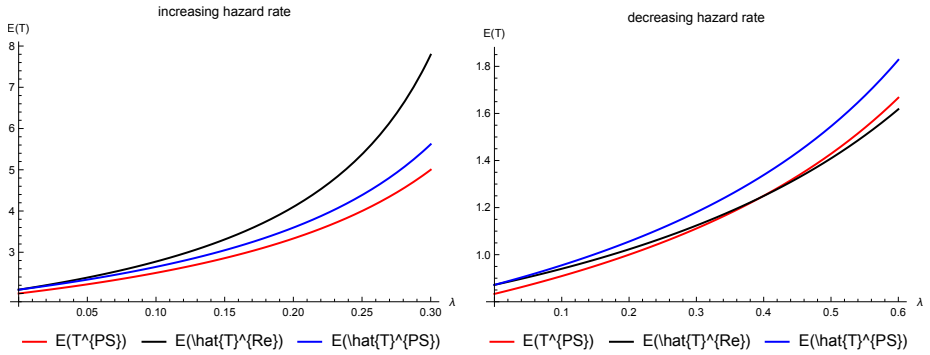

Figure 2: Mean system time of the three queueing systems as a function of the load when the service time distribution has strictly increasing and deceasing hazard rates (with PDF xe^{-x} and $2(e^{-x}+e^{-2x})/3$, respectively) and log-normal multiplicative error

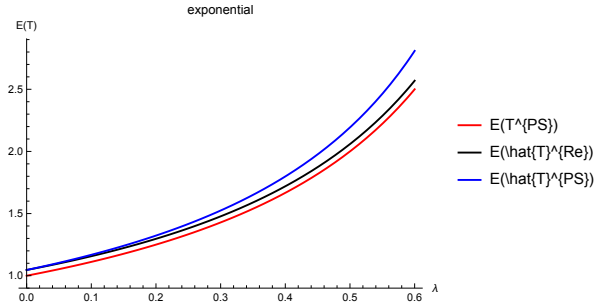

Figure 3: Mean system time of the three queueing systems as a function of the load when the service time is exponentially distributed with mean 1

time of the accurate PS queue when the hazard rate of the service time distribution is strictly monotone decreasing. In the second case the effect of the increased mean service time due to $E(\hat{S}) = E(SX) > E(S)$ dominates over the effect of resampling at low utilization levels, as it is depicted in Figure 2.

In case the service time $\hat{S} = SX$ has a decreasing hazard rate, this ensures that $E(\hat{T}^{(Re)} < E(\hat{T}^{(PS)})$, but the effect of resampling is so moderate that it never dominates over the effect of the increased mean service time (c.f. Figure 3).

References

[1] M. Harchol-Balter, Performance Modeling and Design of Computer Systems: Queueing Theory in Action, Cambridge University Press, 2013.

[2] M. Harchol-Balter, N. Bansal, B. Schroeder, M. Agrawal, SRPT scheduling for web servers, in: Job Scheduling Strategies for Parallel Processing: 7th International Workshop, Springer LNCS, 2001, pp. 11–20.

[3] Y. Qiao, F. Bustamante, P. Dinda, S. Birrer, D. Lu, Improving peer-to-peer performance through server-side scheduling, ACM Trans. Comput. Syst. 26 (4) (2008) 10:1–10:30.

[4] M. Dell'Amico, D. Carra, P. Michiardi, PSBS: Practical size-based scheduling, IEEE Transactions on Computers 65 (7) (2016) 2199–2212.

[5] L. Dong, S. Huanyuan, P. Dinda, Size-based scheduling policies with inaccurate scheduling information, in: Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS), 2004, pp. 31–38.

[6] E. Friedman, S. Henderson, Fairness and efficiency in web server protocols, SIGMETRICS Perform. Eval. Rev. 31 (1) (2003) 229–237.

[7] H. Chang, M. Kodialam, R. R. Kompella, T. V. Lakshman, M. Lee, S. Mukherjee, Scheduling in MapReduce-like systems for fast completion time, in: 2011 Proceedings IEEE INFOCOM, 2011, pp. 3074–3082.

[8] Y. Chen, J. Hasenbein, Staffing large-scale service systems with distributional uncertainty, Queueing Systems 87 (1) (2017) 55–79.

[9] M. Harchol-Balter, B. Schroeder, N. Bansal, M. Agrawal, Size-based scheduling to improve web performance, ACM Trans. Comput. Syst. 21 (2) (2003) 207–233.

[10] A. Wierman, M. Nuyens, Scheduling despite inexact job-size information, SIGMETRICS Perform. Eval. Rev. 36 (1) (2008) 25–36.

[11] Y. Xu, A. Scheller-Wolf, K. Sycara, The benefit of introducing variability in single-server queues with application to quality-based service domains, Operations Research 63 (1) (2015) 233–246.

[12] D. P. Gaver, A waiting line with interrupted service, including priorities, Journal of the Royal Statistical Society. Series B (Methodological) 24 (1) (1962) 73–90.

[13] S. Asmussen, P. W. Glynn, On preemptive-repeat lifo queues, Queueing Systems 87 (2017) 1–22. doi:10.1007/s11134-017-9532-3.

[14] D. Y. Burman, Insensitivity in queueing systems, Advances in Applied Probability 13 (4) (1981) 846–859.
URL http://www.jstor.org/stable/1426976

[15] M. El-Taha, S. Stidham, Sample-path analysis of queueing systems, Springer US, 1999.

[16] M. Pastorelli, D. Carra, M. Dell'Amico, P. Michiardi, HFSP: Bringing size-based scheduling to Hadoop, IEEE Transactions on Cloud Computing 5 (1) (2017) 43–56.

[17] T. Bonald, A. Proutiere, Insensitivity in processor-sharing networks, Performance Evaluation 49 (1) (2002) 193 – 209.

[18] S. Yashkov, Processor-sharing queues: Some progress in analysis, Queueing Systems 2 (1) (1987) 1–17.

[19] S. Yashkov, Mathematical problems in the theory of shared-processor systems, Journal of Soviet Mathematics 58 (2) (1992) 101–147.

[20] M. Harchol-Balter, A. Downey, Exploiting process lifetime distributions for dynamic load balancing, ACM Trans. Comput. Syst. 15 (3) (1997) 253–285.