

Queueing Syst (2007) 56: 169–181
DOI 10.1007/s11134-007-9026-9

Sojourn time asymptotics in Processor Sharing queues with varying service rate

Regina Egorova · Michel Mandjes · Bert Zwart

Received: 20 September 2006 / Revised: 1 February 2007 / Published online: 22 May 2007
© Springer Science+Business Media, LLC 2007

Abstract This paper addresses the sojourn time asymptotics for a GI/GI/· queue operating under the Processor Sharing (PS) discipline with stochastically varying service rate. Our focus is on the logarithmic estimates of the tail of sojourn-time distribution, under the assumption that the job-size distribution has a light tail. Whereas upper bounds on the decay rate can be derived under fairly general conditions, the establishment of the corresponding lower bounds requires that the service process satisfies a sample-path large-deviation principle. We show that the class of allowed service processes includes the case where the service rate is modulated by a Markov process. Finally, we extend our results to a similar system operation under the Discriminatory Processor Sharing (DPS) discipline. Our analysis relies predominantly on large-deviations techniques.

Keywords Processor Sharing · Sojourn-time asymptotics · Large deviations · Change of measure · Varying capacity

Mathematics Subject Classification (2000) Primary 60K25 · Secondary 68M20 · 90B22

R. Egorova (✉) · M. Mandjes
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
e-mail: r.egorova@cwi.nl

M. Mandjes
Korteweg-de Vries Institute for Mathematics, University of
Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam,
The Netherlands
e-mail: mmandjes@science.uva.nl

B. Zwart
Stewart School of Industrial and Systems Engineering, Georgia
University of Technology, 765 Ferst Drive, Atlanta, GA 30332,
USA
e-mail: bertzward@gatech.edu

1 Introduction

Based on the traffic characteristics and Quality-of-Service requirements, traffic flows in communication networks can be roughly divided into two categories: *streaming* flows (voice, video, etc.) and *elastic* flows (data files, Web pages, etc.), see e.g. [20]. Streaming flows require strict delay guarantees for the duration of its connection time, whereas elastic traffic is less demanding. One way of handling both types of traffic is to meet these Quality-of-Service requirements by prioritizing streaming traffic. The bandwidth remaining from the transmission of streaming traffic is made available to elastic traffic. It is widely agreed upon that the protocols for handling elastic traffic are such that each elastic flow obtains roughly an equal share.

Motivated by the above application, one could consider the following model. Let elastic flows (we use the word *jobs* throughout) arrive at a queueing resource, according to a renewal process, and let these jobs be independent samples from some common distribution. The jobs are served in a Processor Sharing (PS) manner, but the capacity available (to be interpreted as the service rate left over by the streaming flows) fluctuates in time. The streaming flows ‘do not see’ the elastic flows, so their performance can be evaluated by using traditional models. The performance experienced by the elastic flows, however, can be regarded as a GI/GI/· queue with a service rate that varies in time (according to some stochastic process), operating under PS, and is more involved. In this paper we study the asymptotic properties of the sojourn-time distribution of the elastic flows.

It is worth noting that in the case of processor-sharing queues with *constant* service capacity, the sojourn time has been studied in many different settings, and this has already proven to be a rather challenging task. The (conditional) sojourn time distribution in the M/G/1-PS queue

was analyzed in terms of Laplace–Stieltjes transforms (LST) by e.g. Yashkov [23], Schassberger [21], Ott [18], Núñez-Queija [16], and Zwart and Boxma [24]. Unfortunately, analytic inversion of these LSTs has appeared to be hard, and only partial results are available.

Another classical subject of research is the derivation of the *asymptotic* behavior of sojourn times in PS-queues. Notably, one of the major insights is that there is a fundamental difference between sojourn-time asymptotics under heavy-tailed and light-tailed jobs. A so-called reduced-load approximation for queues with heavy-tailed distributed job sizes was proven in different settings by, e.g., Zwart and Boxma [24], Núñez-Queija [16] and Jelenković and Momčilović [13]; importantly, long sojourn times are essentially due to the tagged job itself being large. For PS queues with a light-tailed job-size distribution long sojourn times are predominantly caused by the jobs that arrive during the sojourn time of the tagged job. For the light-tailed case, exact asymptotics are known in a few special cases, see [4, 10, 11]; for a survey, see [6].

Recent work by Mandjes and Zwart [15] addressed the logarithmic asymptotics of the sojourn time in the GI/GI/1-PS queue, under technical assumptions which guarantee that the tail distribution of the service time is not too light and not too heavy. More precisely, they proved under specific conditions that the sojourn time V obeys

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{\theta \geq 0} (\alpha(\theta) - \theta), \quad (1.1)$$

where $\alpha(s)$ is the so-called (asymptotic) cumulant function of total amount of work fed to the queue, i.e.,

$$\alpha(\theta) = \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{E}[e^{\theta A(0,x)}],$$

with $A(0, x)$ the amount of traffic offered to the system in $(0, x]$. The goal of the present paper is to generalize the result (1.1) of [15] to a setting in which the available service capacity varies according to some stochastic process. Again the job sizes should be from a light-tailed distribution (but not too light). We extend (1.1) by constructing asymptotic lower and upper bounds, which coincide as x becomes large. The upper bounds can be established under rather general conditions, whereas the lower bound requires that the service process obeys a sample-path large-deviation principle. More specifically, the main result of our work is that we can express the exponential decay rate of $\mathbf{P}(V > x)$ through

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{\theta \geq 0} (\alpha(\theta) + c(-\theta)), \quad (1.2)$$

where $\alpha(\cdot)$ and $c(\cdot)$ are the cumulant functions of the arrival and service processes, respectively. The exact statement of the result is given in Theorem 3.4.

As a special case, we study service processes that have a so-called Markov-fluid structure. Under the extra assumption of the interarrival times of the jobs being exponential (rather than renewal), we derive for these service processes an explicit upper bound on the tail probability, rather than just an upper bound on the exponential decay rate.

Our proofs predominantly rely on large-deviation tools, such as the classical Chernoff bound, as well as the application of sample-path large deviations principles. An important role, however, is also played by the insight that, for PS systems with load larger than 1, the queue length increases roughly at a linear rate. As a by-product, the proofs also show that the sojourn-time asymptotics resemble busy-period asymptotics (in the sense that their exponential decay rates coincide). Although our results are an extension of the results in [15], we have succeeded to simplify the proofs; in particular, we have eliminated the need to use detailed fluid-limit results for overloaded PS queues, as used in [15].

Finally, our methods allow us to obtain an extension of the result to the system operating under the Discriminatory Processor Sharing (DPS) discipline. As for the single-class case, we allow the service process to be random, but note that this result is also new for the standard DPS queue with a fixed service rate. More specifically, we show that the decay rate of the sojourn-time is weight-independent (and hence the same for jobs of any class).

The literature on the analysis of the sojourn time asymptotics in PS queues with time-varying service is quite sparse. Assuming the job-size distribution being heavy-tailed, different extensions of the reduced-load approximations (as derived for the situation with constant service rate) were found. Núñez-Queija [16] studied the M/G/1-PS system in which the service rate is an On-Off process with exponential On-periods. Later, the sojourn time asymptotics for a general DPS queue with time-varying service rate were analyzed in [7]; the authors considered the service rate process which does not fluctuate too wildly compared to the size of a job. Other versions of the reduced-load approximation for queues with varying service rate are given in e.g. [1, 3, 5]. Somewhat related results for workloads in a GPS queue can be found in [25].

The organization of this paper is as follows: The model is described in Sect. 2. In Sect. 3 we present our main results on the logarithmic asymptotics for the queue with general service rate. In addition, we consider the special case in which the service rate varies according to a Markov-fluid process. The proofs can be found in Sect. 4. In Sect. 5 we generalize the result to the DPS queue. Concluding remarks can be found in Sect. 6.

2 Model description and preliminaries

We now specify the GI/GI/· queue operating under the PS discipline, with varying service rate. Jobs arrive according to a renewal process, and the job sizes constitute an i.i.d. sequence; our focus is on the sojourn time, say V , of a ‘tagged’ job (with job size B_0), that we assume to arrive at time 0. The service is according to the PS discipline; all jobs present in the system are served simultaneously and receive an equal share of the available capacity. A complication, however, is that we allow the available capacity to fluctuate in time. The main goal of the paper is to describe the asymptotic behavior of $\mathbf{P}(V > x)$ as $x \rightarrow \infty$.

In this paper we use the following notation. Let $A_n, n \in \mathbb{N}$, be the time between the $(n - 1)$ -st and n -th arrival after time zero. To emphasize that an arrival occurred in the past, we also use the notation $A_{-n}, n \in \mathbb{N}$, for the time between the $(n - 1)$ -st and n -th arrival before time zero. Furthermore, let $B_n, n \in \mathbb{Z}$, be the request size of the n -th job; recall that B_0 corresponds to the tagged job. We assume that $(A_n)_n$ and $(B_n)_n$ are mutually independent sequences, each consisting of i.i.d. random variables. We introduce the random walks $S_n^A = A_1 + \dots + A_n$ and $S_n^B = B_1 + \dots + B_n$, and similarly, with respect to events in the past, $S_{-n}^A = A_{-n} + \dots + A_{-1}$, $S_{-n}^B = B_{-n} + \dots + B_{-1}$. We denote the random variable corresponding to a generic interarrival time (service time) by A (B , respectively).

We set

$$N(t) := \max\{n \in \mathbb{N} : S_n^A \leq t\}$$

representing the number of arrivals in the time interval $(0, t]$. Denote by $A(0, t), t > 0$, the total amount of work fed into the queue in the time interval $(0, t]$, i.e.,

$$A(0, t) = \sum_{i=1}^{N(t)} B_i.$$

Analogously, $C(t_1, t_2)$ is defined as the total service available during the time interval $(t_1, t_2]$ with $t_2 > t_1$,

$$C(t_1, t_2) = \int_{t_1}^{t_2} R(u)du,$$

where $R(u)$ denotes the (random, non-negative) service rate available at time u . Later we also consider the system in the past, i.e., before time zero; then we use the notation $A(-t, 0)$ for the total amount of work fed into the system on $[-t, 0)$. Note that we do not include the tagged arrival which occurred at time 0 into either $A(0, t)$ or $A(-t, 0)$. The cumulative arrival and service processes are assumed to be independent of each other.

Throughout the paper we assume the cumulative service process to satisfy the following conditions:

1. the cumulative service process has *stationary increments*, i.e., the distribution of $C(t_1 + \delta, t_2 + \delta)$ does not depend on δ ;
2. the *service rate $R(\cdot)$ is bounded from above*, i.e. there exists r_{\max} such that $R(u) \leq r_{\max}$ for all u ;
3. the asymptotic cumulant function of $C(0, x)$ exists:

$$c(s) := \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{E}[e^{sC(0,x)}].$$

Furthermore, the system is assumed to be stable, i.e., the long-run average work offered to the system, say α , is smaller than the average offered service, say c , where

$$\alpha := \lim_{t \rightarrow \infty} \frac{\mathbf{E}A(0, t)}{t}, \quad c := \lim_{t \rightarrow \infty} \frac{\mathbf{E}C(0, t)}{t}.$$

Define the moment generating functions (mgfs) $\Phi_B(s) := \mathbf{E}[e^{sB}]$ and $\Phi_A(s) := \mathbf{E}[e^{sA}]$. Since both $\Phi_A(\cdot)$ and $\Phi_B(\cdot)$ are strictly increasing and strictly convex functions, the inverse functions $\Phi_A^{\leftarrow}(\cdot)$ and $\Phi_B^{\leftarrow}(\cdot)$ are well defined. We assume that either A or B does not have a deterministic distribution. A useful result is that the cumulant function of the amount of work fed to the system can be expressed explicitly in terms of the moment generating functions of A and B .

Lemma 2.1 *For $s \geq 0$, the asymptotic cumulant function $\alpha(s)$ of $A(0, x), x > 0$, is given by*

$$\alpha(s) := \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{E}[e^{sA(0,x)}] = -\Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(s)}\right). \tag{2.1}$$

If either A or B is non-deterministic, then $\alpha(\cdot)$ is strictly convex.

The result of Lemma 2.1, as stated by Whitt [22], was proved in [15]. In light of the above result, we require throughout the paper that $\Phi_B(s) < \infty$ for some positive $s > 0$. This way we exclude jobs with a heavy-tailed distribution. Importantly, there are no restrictions imposed on the interarrival times; these can be heavy-tailed or light-tailed: for $s \geq 0, 1/\Phi_B(s) \in (0, 1]$, and hence $\Phi_A^{\leftarrow}(1/\Phi_B(s))$ is well-defined, irrespective of the shape of the distribution of the interarrival times.

In the sequel, we separately consider the special case in which the service process is given by a Markov-fluid process. Such a process can be described as follows. Consider a continuous-time Markov chain on a finite state space $\{1, 2, \dots, d\}$. The transition rate matrix is denoted

by $Q := (q_{ij})_{i,j=1,2,\dots,d}$, where $q_{ij} \geq 0$ ($i \neq j$) and $q_{ii} = -\sum_{j \neq i} q_{ij}$. We assume that the Markov chain is irreducible, and π denotes its unique steady-state distribution. When the Markov chain is in state i , the server provides service at constant rate $r_i \geq 0$. Let R be the diagonal matrix with coefficients r_i on the diagonal. Denote the mean rate by $c := \sum_{i=1}^d r_i \pi_i$. We denote this class of processes by $\text{Mf}(Q, R)$; if the service process is of this type, we write $C(\cdot, \cdot) \in \text{Mf}(Q, R)$. Results from Kesidis et al. [14] yield the following standard properties.

Property 2.1 *Let $C(\cdot, \cdot) \in \text{Mf}(Q, R)$. Then the following statements hold:*

1. *The mgf of the service provided to the jobs in an interval of length x is given by*

$$\mathbf{E}e^{sC(0,x)} = \pi e^{(Q+sR)x} \mathbf{1},$$

where $\mathbf{1}$ the all-one vector of dimension d .

Denote by $c_1(s), \dots, c_d(s)$ the eigenvalues of matrix $Q + sR$. Hence, the mgf can be represented, for appropriate numbers m_1, \dots, m_d , as

$$\mathbf{E}e^{sC(0,x)} = \sum_{i=1}^d m_i e^{c_i(s)x}.$$

2. *For all real s there exists a limiting mgf:*

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{E}e^{sC(0,x)} = c(s).$$

Moreover, $c(s) = \max(c_1(s), \dots, c_d(s))$, i.e., $c(s)$ is the largest real eigenvalue of $Q + sR$; the corresponding eigenvector is componentwise positive.

3. *There exists a finite K such that*

$$\mathbf{E}e^{sC(0,x)} \leq K e^{c(s)x}.$$

For instance, $K = \sum_{i=1}^d m_i$.

3 Main results

In this section we present the main results of the paper. These characterize the logarithmic asymptotic behavior of the tail probability $\mathbf{P}(V > x)$ as $x \rightarrow \infty$, under the assumption that the job size has a light-tailed distribution.

To put things in perspective, we first recall the result for asymptotic behavior of the sojourn time in a PS queue with constant (rather than fluctuating) service capacity. Mandjes and Zwart [15] performed a large-deviation analysis of the steady-state sojourn time distribution in the GI/GI/1-PS queue, and derived the following logarithmic estimates under the assumption that the job-size distribution has a light tail.

Theorem 3.1 (See [15]) *Consider the GI/GI/1-PS queue (with constant service rate of, say, 1). If there exists a solution $v^* > 0$ to $\alpha'(s) = 1$, and for each constant $c > 0$*

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(B > c \log x) = 0,$$

then

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{s \geq 0} (\alpha(s) - s) = \alpha(v^*) - v^*. \quad (3.1)$$

The main goal of the present paper is to derive a generalization of the above result for a queue with varying service rate. We will show that under similar assumptions on the arrival and job-size processes, and in addition certain assumptions on the service process, we can prove the following extension of (3.1):

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{s \geq 0} (\alpha(s) + c(-s)). \quad (3.2)$$

Despite the simple form, the proof of the above result is quite technical. The proof consists of two parts, derivation of the upper bound (i.e., (3.2) with “=” replaced by “ \leq ”) and derivation of the lower bound (i.e., (3.2) with “=” replaced by “ \geq ”) which coincide.

The proof of the upper bound is rather elementary, and is essentially based on classical Chernoff-bound arguments, and applies without imposing additional conditions on the service process.

The proof of the lower bound, however, is substantially harder. There we first truncate the job-size distribution (and then let the truncation threshold increase to ∞), so that we enforce linearly bounded queue length growth. Thus, the problem is reduced to finding the corresponding busy-period asymptotics, which effectively can be decomposed into two independent terms representing a large deviation of the arrival process and a large deviation of service process. This argument, in particular, intuitively explains the occurrence of long busy periods (and hence also of long sojourn times) as the result of a ‘conspiracy’: the arrival rate is higher than usual, the size of the jobs are larger than usual, and in addition less service is available than usual. We remark that analysis of the large deviation of service process requires an additional assumption: the service process should obey a so-called sample-path large-deviation principle (more precisely: only the large-deviations lower bound is required here).

In the following subsections we will present results for the system with general service process, but also (more explicit) results for the case the service process is Markov fluid. The proofs are deferred to Sect. 4.

3.1 Upper bound

We first present the upper bound for a GI/GI/· system with a generally distributed service process. We need to make the following assumption.

Assumption 3.1 There exists $\nu > 0$ such that $\alpha(\nu) + c(-\nu) < 0$.

This assumption ensures that service times are light-tailed and that the system is stable. To be more precise, what the assumption states is that in some neighborhood to the right of the origin the sum stays finite. This implies (due to Lemma 2.1) that there exists a neighborhood of the origin in which the mgf $\Phi_B(\cdot)$ is well defined (as an aside, note that this implies that B is light tailed). Since the function $g(s) = \alpha(s) + c(-s)$ is strictly convex and equals 0 at $s = 0$, the assumption implies that $g(\cdot)$ has a negative derivative at $s = 0$, $\alpha - c < 0$, and hence, the system is stable.

Due to strict convexity of the cumulant function, we can define $\omega^* > 0$ such that

$$\omega^* = \arg \inf_{s \geq 0} (\alpha(s) + c(-s)).$$

Since $\alpha(s) + c(-s)$ equals zero at $s = 0$ and has a strictly negative derivative at $s = 0$, we also have $\inf_{s \geq 0} (\alpha(s) + c(-s)) = \alpha(\omega^*) + c(-\omega^*) < 0$.

The next theorem gives the logarithmic upper bound for $\mathbf{P}(V > x)$ in terms of the cumulant functions.

Theorem 3.2 *If Assumption 3.1 is valid, then*

$$\limsup_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) \leq \inf_{s \geq 0} (\alpha(s) + c(-s)). \tag{3.3}$$

Besides the general upper bound on the exponential decay rate, as presented in Theorem 3.2, we have a tighter result (namely bounds on the probability $\mathbf{P}(V > x)$ itself, uniformly in x) for an important special case. This result requires an additional assumption; it implies Assumption 3.1 and existence of ω^* .

Assumption 3.2 There exists a solution $\nu^* > 0$ to $\alpha(\nu^*) + c(-\nu^*) = 0$.

In the special case we consider, we specialize to Poisson interarrival times (rather than renewal arrivals; the arrival process is thus a compound Poisson process) and Markov-fluid service. We remark that the constant K , as used in Theorem 3.3, will be explicitly given in the proof of the result.

Theorem 3.3 *Suppose the arrival process is given by a compound Poisson process (with arrival rate λ) and the service*

process is in $\mathbb{Mf}(Q, R)$. Then, under Assumption 3.2, uniformly in $x \geq 0$,

$$\mathbf{P}(V > x) \leq K e^{(\alpha(\omega^*) + c(-\omega^*))x}, \tag{3.4}$$

and $\alpha(\omega^) = \lambda(\Phi_B(\omega^*) - 1)$.*

3.2 Lower bound

Let us now turn to stating the results for the lower bound on $\mathbf{P}(V > x)$. Here we need the following assumption.

Assumption 3.3 For each constant $c > 0$, we have

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(B > c \log x) = 0.$$

It is readily checked that this assumption is satisfied by most distributions of interest, such as phase-type, Gamma, Weibull distributions, etc. However, it is noted that it is violated by distributions with *extremely light* tails. For instance, the assumption does not hold for service times for which $\mathbf{P}(B > x)$ is of the form $\exp(-e^x)$, and by job size distributions with bounded support (for instance the deterministic distribution). The merit of this assumption has been illustrated in [15], where it was shown that service times with bounded support need a completely different treatment, see also [10] for a detailed treatment of the M/D/1-PS queue.

The derivation of the lower bound is considerably more involved than the corresponding upper bound. Importantly, it requires extra structure of the process $C(\cdot, \cdot)$, namely that the process $C(\cdot, \cdot)$ must satisfy the lower bound of a *sample-path large-deviation principle*.

Definition 3.1 Denote by \mathcal{AC} the space of all absolutely continuous functions (see e.g. [9], p. 176). Consider the space $\Omega := \{f : [0, 1] \rightarrow \mathbb{R}, f \in \mathcal{AC}, f(0) = 0\}$ equipped with supremum norm topology $\|f\| = \sup_{u \in [0, 1]} f(u)$.

Let the process $Z_x(\cdot)$ be given through

$$Z_x(u) := \frac{1}{x} \int_0^{ux} R(s) ds = \frac{1}{x} C(0, ux), \quad u \in [0, 1].$$

The process $Z_x(\cdot)$ obeys a *sample-path large-deviation principle* (sp-LDP) if

$$\limsup_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(Z_x(\cdot) \in S) \leq - \inf_{f \in S} \int_0^1 \Lambda(f'(t)) dt, \tag{3.5}$$

for all closed $S \subset \Omega$,

$$\liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(Z_x(\cdot) \in T) \geq - \inf_{f \in T} \int_0^1 \Lambda(f'(t)) dt, \tag{3.6}$$

for all open $T \subset \Omega$,

where $\Lambda(t) := \sup_{s \in \mathbb{R}}(st - c(s))$. We say that (3.5) is the upper bound of the sp-LDP, and (3.6) is the lower bound of the sp-LDP.

Assumption 3.4 The process $Z_x(\cdot)$, defined through $Z_x(u) := C(0, ux)/x$, $u \in [0, 1]$, satisfies the lower bound of the sp-LDP (3.6).

The next theorem presents the main result of the paper; its upper bound was already stated in Theorem 3.2.

Theorem 3.4 *If Assumptions 3.1, 3.3 and 3.4 are valid, then*

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{s \geq 0} (\alpha(s) + c(-s)).$$

Although, to our best knowledge, no sp-LDP was established for Markov fluid, we were still able to prove the corresponding logarithmic lower bound.

Theorem 3.5 *If $C(\cdot, \cdot) \in \mathbb{Mf}(Q, R)$, and Assumptions 3.1 and 3.3 are valid, then*

$$\liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) \geq \inf_{s \geq 0} (\alpha(s) + c(-s)). \tag{3.7}$$

Thus, combining the results in Theorems 3.3 and 3.5, we conclude that if the service process is of Markov fluid type, the logarithmic asymptote (3.2) holds under Assumptions 3.1 and 3.3.

Remark 3.1 Our results allow us to compare the performance of the systems with varying service rate and with constant rate (where the mean service rate is the same in both systems). It is a quite general phenomenon that performance nearly always improves if a random process is replaced by a deterministic process with the same mean.

Therefore, we now consider the GI/GI/1-PS system with fixed service rate c (recall that this is the mean service rate of the system considered in this paper). Applying Jensen’s inequality we obtain that

$$\begin{aligned} c(s) &= \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{E}[e^{sC(0,x)}] \geq \lim_{x \rightarrow \infty} \frac{1}{x} \log e^{\mathbf{E}[sC(0,x)]} \\ &= \lim_{x \rightarrow \infty} \frac{1}{x} \mathbf{E}[sC(0, x)] = sc. \end{aligned}$$

Hence,

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) &= \inf_{s \geq 0} (\alpha(s) + c(-s)) \\ &\geq \inf_{s \geq 0} (\alpha(s) - sc), \end{aligned}$$

where the latter is the exponential decay rate in the system with constant service rate c . If the function $c(-s)$ is strictly

convex, it can be shown that the above inequality is strict. Thus, we conclude that, informally speaking, the random service rate increases the probability of a long sojourn time.

4 Proofs

We now provide the proofs of the results presented in the previous section.

4.1 Upper bound

We start by proving the upper bound.

Proof of Theorem 3.2 The event $\{V > x\}$ implies that the queue does not empty before time x . Evidently, as we assume the system to be in steady state (with respect to job arrivals), the workload present at time 0, say W , can be identified with the FIFO waiting time. In other words, W has the representation $W = \sup_{t \geq 0} (A(-t, 0) - C(-t, 0))$. Hence, we can write

$$\begin{aligned} \mathbf{P}(V > x) &\leq \mathbf{P}(W + B_0 + A(0, x) - C(0, x) > 0) \\ &= \mathbf{P}\left(\sup_{t > 0} (A(-t, 0) - C(-t, 0)) + B_0 \right. \\ &\quad \left. + A(0, x) - C(0, x) > 0\right). \tag{4.1} \end{aligned}$$

Now note that the process $A(0, x)$ jumps at the arrival epochs and is constant in between, whereas we assumed the process $C(0, x)$ to be non-decreasing. Hence, the difference $A(0, x) - C(0, x)$ increases with positive jumps at arrival epochs and is non-increasing in between. Therefore, the supremum can only be attained at arrival epochs. This yields that (4.1) is equivalent to

$$\begin{aligned} \mathbf{P}\left(\sup_{n \in \mathbb{N}} (A(-S_{-n}^A, 0) - C(-S_{-n}^A, 0)) \right. \\ \left. + B_0 + A(0, x) - C(0, x) > 0\right). \end{aligned}$$

Now applying the standard union bound, this expression is further bounded by

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbf{P}(A(-S_{-n}^A, 0) - C(-S_{-n}^A, 0) + B_0 + A(0, x) \\ - C(0, x) > 0) \\ = \sum_{n=1}^{\infty} \mathbf{P}(A(-S_{-n}^A, 0) + B_0 + A(0, x) \\ - C(-S_{-n}^A, x) > 0), \end{aligned}$$

where we recall that $-S_{-n}^A$ denotes the time of the n -th arrival in the past. Now we can apply the Chernoff bound to

(each term in) the last expression, so that (recall independence of $A(-S_{-n}^A, 0)$ and $A(0, x)$) we arrive at

$$\begin{aligned} \mathbf{P}(V > x) &\leq \sum_{n=1}^{\infty} \mathbf{E}[e^{\omega^*(A(-S_{-n}^A, 0) + B_0 + A(0, x) - C(-S_{-n}^A, x))}] \\ &= \sum_{n=1}^{\infty} \int_0^{\infty} \mathbf{E}[e^{\omega^*(A(-S_{-n}^A, 0) + B_0 + A(0, x) - C(S_{-n}^A, x))} | \\ &\quad S_{-n}^A = y] d\mathbf{P}(S_{-n}^A \leq y) \\ &= \sum_{n=1}^{\infty} \int_0^{\infty} (\mathbf{E}[e^{\omega^* B}])^{n+1} \mathbf{E}[e^{\omega^* A(0, x)}] \\ &\quad \times \mathbf{E}[e^{-\omega^* C(-y, x)}] d\mathbf{P}(S_{-n}^A \leq y), \end{aligned}$$

where in the last equality $A(-S_{-n}^A, 0)$ is interpreted as the sum of n job sizes. Now applying the definition of the cumulant function $c(\cdot)$, we obtain that for any $\varepsilon > 0$ for x large enough the expression in the previous display is bounded from above by

$$\begin{aligned} &\sum_{n=1}^{\infty} \int_0^{\infty} (\mathbf{E}[e^{\omega^* B}])^{n+1} e^{(\alpha(\omega^*) + \varepsilon)x} \\ &\quad \times e^{(c(-\omega^*) + \varepsilon)(x+y)} d\mathbf{P}(S_{-n}^A \leq y). \end{aligned}$$

Evaluating the integral and using the definition of S_{-n}^A , we see that the last expression equals

$$\begin{aligned} &\sum_{n=1}^{\infty} (\mathbf{E}[e^{(c(-\omega^*) + \varepsilon)A}])^n e^{(\alpha(\omega^*) + c(-\omega^*) + 2\varepsilon)x} (\mathbf{E}[e^{\omega^* B}])^{n+1} \\ &= \mathbf{E}[e^{\omega^* B}] e^{(\alpha(\omega^*) + c(-\omega^*) + 2\varepsilon)x} \\ &\quad \times \sum_{n=1}^{\infty} (\Phi_B(\omega^*) \Phi_A(c(-\omega^* + \varepsilon)))^n. \end{aligned}$$

Now observe that the summation over n does not depend on x ; we therefore now verify whether this sum is finite. Note that (apply Lemma 2.1)

$$\alpha(\omega^*) + c(-\omega^*) = -\Phi_A^{\leftarrow} \left(\frac{1}{\Phi_B(\omega^*)} \right) + c(-\omega^*) < 0.$$

Hence, due to continuity of the mgfs, we see that for ε small enough the product under the sum is less than one, and hence the geometric series is converging. Furthermore, $\mathbf{E}[e^{\omega^* B}] < \infty$. Thus, we conclude that $\mathbf{P}(V > x)$ can be bounded from above by

$$\mathbf{P}(V > x) \leq M e^{(\alpha(\omega^*) + c(-\omega^*) + 2\varepsilon)x},$$

where $M < \infty$ is some positive constant. Taking logarithms, dividing by x , letting $x \rightarrow \infty$ and $\varepsilon \downarrow 0$, we obtain

$$\limsup_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V > x) \leq \alpha(\omega^*) + c(-\omega^*).$$

This proves the upper bound. \square

We now turn to the proof of Theorem 3.3. Let us first state the basic result for the workload distribution which is useful for our proof. Denote by $X(t)$ the state of the underlying Markov chain at time t ; $X(t) \in \{1, 2, \dots, d\}$.

Proposition 4.1 *If $C(\cdot, \cdot) \in \mathbb{Mf}(Q, R)$ and Assumption 3.2 is valid, then there exists a constant $K > 0$ such that for any initial state of service process $X(0) = i$, $i \in \{1, 2, \dots, d\}$, uniformly in x*

$$\begin{aligned} \mathbf{P} \left(\sup_{t \geq 0} A(-t, 0) - C(-t, 0) > x | X(0) = i \right) \\ \leq K e^{-v^* x}. \end{aligned} \tag{4.2}$$

Proof of Proposition 4.1 We present a proof that is based on a change-of-measure argument; there are several alternative approaches possible. This change of measure is such that the event $\{W > x\}$ becomes more likely than under the old measure. We introduce a process

$$T(x) := \inf\{t : A(-t, 0) - C(-t, 0) > x\}.$$

Then we can write

$$\mathbf{P}(W > x) = \mathbf{P}(T(x) < \infty).$$

Let us first twist the interarrival-time and job-size distributions. Define a new probability measure \mathbf{P}_ω for $\omega > 0$ such that

$$\mathbf{P}_\omega(A \in dx) = \mathbf{P}(A \in dx) e^{-\alpha(\omega)x} / \Phi_A(-\alpha(\omega)),$$

$$\mathbf{P}_\omega(B \in dx) = \mathbf{P}(B \in dx) e^{\omega x} / \Phi_B(\omega).$$

In order to construct the change of measure for the service process let us first define the largest real eigenvalue of the matrix $Q + \omega R$, which coincides with $c(\omega)$, where the corresponding right eigenvector $(v_1, \dots, v_d)^T$ is componentwise positive, see Property 2.1(2). Note that the eigenvector also depends on ω , but for compactness we suppress this. With the new probability measure we associate the modified Markov chain with transition matrix Q^* defined as (for $i \neq j$)

$$q_{ij}^* = q_{ij} v_j / v_i,$$

$$q_{ii}^* = q_{ii} + r_i \omega + c(-\omega).$$

It is not hard to verify that these rates indeed constitute a generator matrix (use that $c(\omega)$ is eigenvalue of $Q + \omega R$).

We have the following fundamental identity

$$\mathbf{P}(W > x) = \mathbf{E}_\omega[LI\{T(x) < \infty\}], \tag{4.3}$$

see e.g. Theorem XIII.3.2 in [2]; here \mathbf{E}_ω denotes expectation under the new measure \mathbf{P}_ω , and L is the likelihood ratio process stopped at $T(x)$, which we specify below.

Instead of giving a detailed description of this likelihood ratio process, we prefer to construct L directly, exploiting the fact that the number of events form a simple discrete set. We take the parameter ω (the ‘exponential twist’) to be equal to ν^* . Suppose that in $[-T(x), 0)$ there were n arrivals; denote by $a_i, b_i, i = 1, \dots, n$, the interarrival times and corresponding job sizes. Also suppose that there were m transitions of the Markov chain governing the service process; let, in time interval $[-T(x), 0)$, the Markov chain $X(\cdot)$ visit states i_0, i_1, \dots, i_m . Define by $t_{ij}, j = 1, \dots, m$, the time service process spends in state i_j .

Denote by L_j the likelihood after the j -th event. Let $L_0 \leftarrow 1$. The value of L_j is updated as follows:

$$L_j \leftarrow L_{j-1} e^{-c(-\nu^*)a_k} e^{-\nu^*b_k} \Phi_A(-\alpha(\nu^*)) \Phi_B(\nu^*),$$

if the j -th event is the k -th arrival ($k = 1, \dots, n$), and

$$L_j \leftarrow L_{j-1} \frac{v_{i_{k-1}}}{v_{i_k}} e^{\nu^*r_{i_k}t_{i_k} + c(-\nu^*)t_{i_k}},$$

if the j -th event is the k -th transition of the Markov chain ($k = 1, \dots, m$). Observe that L_j is the product of (unit-mean) random variables U_1, \dots, U_j , that form a martingale with respect to the measure \mathbf{P}_ω .

Now consider this likelihood ratio stopped at the stopping time $T(x)$, which we simply denote by L . It is readily verified that we can write

$$\begin{aligned} L &= \left(\frac{v_{i_0}}{v_{i_m}} \right) \times \left(e^{\nu^* \sum_{j=1}^m r_{i_j} t_{i_j} + c(-\nu^*) \sum_{j=1}^m t_{i_j}} \right) \\ &\times \left(e^{\alpha(\nu^*) \sum_{i=1}^n a_i} \right) \times \left(e^{-\nu^* \sum_{i=1}^n b_i} \right) \\ &\times \left(\Phi_A(-\alpha(\nu^*)) \Phi_B(\nu^*) \right)^n. \end{aligned}$$

As $-T(x)$ corresponds to an arrival epoch, we have that $\sum_{i=1}^n a_i = T(x)$, $\sum_{i=1}^n b_i = A(0, T(x))$. Also, recall from Lemma 2.1 that $\Phi_A(-\alpha(\nu^*)) \Phi_B(\nu^*) = 1$. Recall the new measure was chosen so that the event $\{T(x) < \infty\}$ occurs with probability 1. We thus find

$$\begin{aligned} L &\leq \left(\frac{v_{i_0}}{v_{i_m}} \right) \times \left(e^{-\nu^*(A(0, T(x)) - C(0, T(x)))} \right) \\ &\times \left(e^{\alpha(\nu^*)T(x) + c(-\nu^*) \sum_{j=1}^m t_{i_j}} \right). \end{aligned}$$

Taking into account that $\{I\{T(x) < \infty\} = 1\}$ implies $A(-T(x), 0) - C(-T(x), 0) > x$, in conjunction with $\alpha(\nu^*) = -c(-\nu^*)$, we have identified a $K > 0$ such that

$$LI\{T(x) < \infty\} \leq Ke^{-\nu^*x}.$$

We conclude that the identity (4.3) implies that indeed $\mathbf{P}(W > x) \leq Ke^{-\nu^*x}$, irrespective of the value of $X(0) = i$. \square

Proof of Theorem 3.3 Since the event $\{V > x\}$ implies that the queue does not empty before time x , we obtain by using the Chernoff bound

$$\begin{aligned} \mathbf{P}(V > x) &\leq \mathbf{P}(W + B_0 + A(0, x) - C(0, x) > 0) \\ &\leq \mathbf{E}[e^{\omega^*(W + B_0 + A(0, x) - C(0, x))}] \\ &= \mathbf{E}[\mathbf{E}[e^{\omega^*(W + B_0 + A(0, x) - C(0, x))} | X(0)]]]. \end{aligned}$$

Conditioning on the state of the Markov chain at time 0 provides the independence between the workload process and the arrival and service process after time 0. Therefore, the last expression in the previous display is equal to

$$\begin{aligned} &\mathbf{E}[e^{\omega^*B_0}] \mathbf{E}[\mathbf{E}[e^{\omega^*W} | X(0)]] \mathbf{E}[e^{\omega^*(A(0, x) - C(0, x))} | X(0)]] \\ &= \mathbf{E}[e^{\omega^*B_0}] \sum_{i=1}^d \mathbf{E}[e^{\omega^*W} | X(0) = i] \\ &\quad \times \mathbf{E}[e^{\omega^*(A(0, x) - C(0, x))} | X(0) = i] \pi_i, \end{aligned}$$

where we recall that π is the equilibrium distribution of $X(\cdot)$. Since $\alpha(s) + c(-s)$ equals zero at $s = 0$, and has a strictly negative derivative at $s = 0$, it follows that $\omega^* < \nu^*$. Then, Proposition 4.1 implies that there is a K_1 such that

$$\begin{aligned} \mathbf{E}[e^{\omega^*W} | X(0)] &= \int_0^\infty \mathbf{P}(e^{\omega^*W} > x | X(0)) dx \\ &= \int_0^\infty \mathbf{P}(W > \log x / \omega^* | X(0)) dx \\ &\leq 1 + \int_1^\infty \mathbf{P}(W > \log x / \omega^* | X(0)) dx \\ &\leq 1 + \int_1^\infty K_1 e^{-\nu^*/\omega^* \log x} dx \\ &< 1 + K_1 \int_1^\infty x^{-\nu^*/\omega^*} dx =: K_2 < \infty. \end{aligned}$$

Consequently,

$$\mathbf{P}(V > x) \leq K_2 \cdot \mathbf{E}[e^{\omega^*B_0}] \mathbf{E}[e^{\omega^*A(0, x)}] \mathbf{E}[e^{\omega^* - C(0, x)}]. \tag{4.4}$$

Note that due to Assumption 3.2, $\mathbf{E}[e^{\omega^*B}] < \infty$. Since the process $A(0, x)$ is a compound Poisson process we have

$$\mathbf{E}[e^{\omega^*A(0, x)}] = e^{\alpha(\omega^*)x} = e^{\lambda x (\Phi_B(\omega^*) - 1)}.$$

Due to Property 2.1(3), there exists $K_3 < \infty$ such that

$$\mathbf{E}[e^{\omega^*C(0, x)}] \leq K_3 e^{c(\omega^*)x}.$$

Combining this with (4.4), we have identified a $K > 0$ such that, uniformly in $x \geq 0$, $\mathbf{P}(V > x) \leq Ke^{(\alpha(\omega^*) + c(-\omega^*))x}$, where $\alpha(\omega^*) = \lambda(\Phi_B(\omega^*) - 1)$, as desired. \square

4.2 Lower bound

We now proceed with proving the lower bound results.

Proof of Theorem 3.4

Our proof consists of five steps: (i) we truncate the job-size distribution to find a lower bound on $\mathbf{P}(V > x)$ that, by virtue of Assumption 3.3, reduces the problem to finding a lower bound on a related busy period problem for the system with truncated jobs; (ii) next, we show that long busy periods are due to a large deviation of both the arrival process and the service process; (iii) after that, we analyze the large deviations of the arrival process, and pay special attention to the technicality of dealing with the truncated job sizes; (iv) we then invoke the sp-LPD lower bound (3.6) to analyze the large deviations of the service process; (v) finally, we combine all the above to establish the result.

Step (i). We truncate the job-size distribution, by introducing a new stochastic process $A_k(0, x)$, $k > 0$, as follows:

$$A_k(0, x) := \sum_{i=1}^{N(x)} B_i I\{B_i < k\}.$$

By definition of the PS queue with varying service capacity,

$$\mathbf{P}(V > x) = \mathbf{P}\left(B_0 > \int_0^x \frac{1}{1 + Q(u)} dC(0, u)\right),$$

where $Q(u)$ is the number of jobs in the system at time u excluding the tagged job.

If we have $A_k(0, u) - C(0, u) > \varepsilon u$, then also $A_k(0, u) > \varepsilon u$, and as all jobs are at most of size k , we find a linear lower bound on the number of jobs present at time u : $Q(u) \geq \varepsilon u/k$. We thus obtain

$$\begin{aligned} \mathbf{P}(V > x) &\geq \mathbf{P}\left(B_0 > \int_0^x \frac{1}{1 + Q(u)} dC(0, u), \right. \\ &\quad \left. A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)\right) \\ &\geq \mathbf{P}\left(B_0 > \int_0^x \frac{1}{1 + \varepsilon u/k} dC(0, u) \mid \right. \\ &\quad \left. A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)\right) \\ &\quad \times \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)). \end{aligned}$$

By applying integration by parts and standard calculus,

$$\begin{aligned} &\int_0^x \frac{1}{1 + \varepsilon u/k} dC(0, u) \\ &= \frac{C(0, x)}{1 + \varepsilon x/k} + \frac{\varepsilon}{k} \int_0^x C(0, u) \frac{1}{(1 + \varepsilon u/k)^2} du \\ &\leq \frac{C(0, x)}{1 + \varepsilon x/k} + \frac{\varepsilon}{k} r_{\max} \int_0^x \frac{u}{(1 + \varepsilon u/k)^2} du \end{aligned}$$

$$\begin{aligned} &\leq \frac{r_{\max} x}{1 + \varepsilon x/k} + \frac{r_{\max} k}{\varepsilon} \left(\frac{1}{1 + \varepsilon x/k} - 1 + \log\left(1 + \frac{\varepsilon}{k} x\right) \right) \\ &= \frac{r_{\max} k}{\varepsilon} \log\left(1 + \frac{\varepsilon}{k} x\right). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbf{P}(V > x) &\geq \mathbf{P}\left(B_0 > \frac{r_{\max} k}{\varepsilon} \log\left(1 + \frac{\varepsilon}{k} x\right) \mid \right. \\ &\quad \left. A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)\right) \\ &\quad \times \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)). \end{aligned}$$

Now observe that in the first probability in the right-hand side of the previous display, the value of B_0 does not depend on the condition, so that we finally arrive at the lower bound

$$\begin{aligned} &\mathbf{P}\left(B_0 > \frac{kr_{\max}}{\varepsilon} \log\left(1 + \frac{\varepsilon}{k} x\right)\right) \\ &\quad \times \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)). \end{aligned} \tag{4.5}$$

Due to Assumption 3.3 we conclude that the first probability in (4.5) asymptotically behaves as $e^{\rho(x)}$. Therefore, we are left with analyzing the second probability, which could be interpreted as the probability of a busy period exceeding x in the system with truncated jobs and a service rate perturbed by ε .

Step (ii). We bound the second factor in (4.5) as follows:

$$\mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in [0, x]) \geq \mathbf{P}_1(x) \cdot \mathbf{P}_2(x);$$

here $\mathbf{P}_1(x) := \mathbf{P}(A_k(0, u) - bu > 0, u \in (0, x))$, $\mathbf{P}_2(x) := \mathbf{P}(C(0, u) < (b - \varepsilon)u, u \in (0, x))$, and $b < c$ is any fixed number. We have thus decomposed the probability of a long busy period into a large deviation of the arrival process and a large deviation of the service process; the intuitive explanation is that the occurrence of a long busy period is the result of both the arrival process generating traffic at a *higher* rate than usual and the service process offering service at a *lower* rate than usual. We emphasize that the value of b is free now, but in Step (v) we choose an appropriate value. We now deal with each of the probabilities separately; in Step (iii) we analyze $\mathbf{P}_1(x)$, and in Step (iv) $\mathbf{P}_2(x)$.

Step (iii). Consider $\mathbf{P}_1(x)$. Denote by P_k the busy period in the system with truncated job size (at threshold k) and constant service rate b . In [19] the asymptotics for large busy periods in this system were derived; it is readily checked that the corresponding conditions apply for truncated job sizes. We thus find

$$\begin{aligned} &\liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(A_k(0, u) - bu > 0, u \in (0, x)) \\ &= \liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(P_k > x) = \inf_{s \geq 0} (\alpha_k(s) - bs) = \gamma_b^k < 0, \end{aligned}$$

where

$$\alpha_k(s) := \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{E}[e^{sA_k(0,x)}].$$

We now show that $\gamma_b^k \rightarrow \gamma_b := \inf_{s \geq 0} (\alpha(s) - bs)$ as $k \rightarrow \infty$. To this end, define $f_k(s) := \alpha_k(s) - bs$. Clearly, $f_k(s) \rightarrow f(s) = \alpha(s) - bs$ pointwise as $k \rightarrow \infty$ and $f_k(s)$ is increasing in k . Consequently, we have that the limit of γ_b^k for $k \rightarrow \infty$ exists and that

$$\gamma_b^* := \lim_{k \rightarrow \infty} \gamma_b^k = \lim_{k \rightarrow \infty} \inf_{s \geq 0} f_k(s) \leq \inf_{s \geq 0} f(s) = \gamma_b.$$

It remains to be shown that the reverse inequality holds. For this we follow an argument similar to the proof of Proposition 2.2 in Nuyens and Zwart [17].

Note that the function $f_k(\cdot)$ is continuous in s . Moreover, it is non-decreasing in k , and thus so is $\gamma_b^k \equiv \inf_{s \geq 0} f_k(s)$. Clearly, $\inf_{s \geq 0} f_k(s) \leq f_k(0) \leq f(0) = 0$, and hence $\gamma_b^* \equiv \lim_{k \rightarrow \infty} \inf_{s \geq 0} f_k(s) \leq 0$.

Now denote by B^k the job size truncated at k . Take k_0 such that $\mathbf{P}(B^k > bA) > 0$ for $k > k_0$. Then there exist $\delta, \eta > 0$ such that $\mathbf{P}(B^k - bA \geq \delta) \geq \eta > 0$ for $k > k_0$. Hence, for $k \geq k_0$,

$$\begin{aligned} \Phi_{B^k}(s)\Phi_A(-bs) &= \mathbf{E}[e^{sB^k}]\mathbf{E}[e^{-sbA}] \\ &= \mathbf{E}[e^{s(B^k - bA)}] \geq \eta e^{s\delta}, \end{aligned}$$

and consequently, for s large enough,

$$\Phi_A(-bs) \geq \frac{1}{\Phi_{B^k}(s)}.$$

Since $\Phi_A^{-1}(-s)$ is increasing in s , we find that for s and k large enough,

$$\begin{aligned} \alpha_k(s) - bs &= -\Phi_A^{-1}\left(\frac{1}{\Phi_{B^k}(s)}\right) - bs \\ &\geq -\Phi_A^{-1}(\Phi_A(-bs)) - bs = 0, \end{aligned}$$

and $\gamma_b^* > -\infty$. Therefore, the level sets $L_k = \{s \geq 0 : f_k(s) \leq \gamma_b^*\}$ are non-empty, compact sets that are nested with respect to k , which implies that there exists at least one point, say s_0 , in their intersection. By definition of s_0 , we have $f_k(s_0) \leq \gamma_b^*$ for every k . Since f_k converges pointwise, we find

$$\gamma_b = \inf_{s \geq 0} f(s) \leq f(s_0) = \lim_{k \rightarrow \infty} f_k(s_0) \leq \gamma_b^*.$$

Thus, we conclude that $\gamma_b^k \rightarrow \gamma_b$ as $k \rightarrow \infty$, and

$$\liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}_1(x) = \inf_{s \geq 0} (\alpha(s) - bs).$$

Step (iv). We now analyze the asymptotic behavior of $\mathbf{P}_2(x)$. First observe that we can rewrite $\mathbf{P}_2(x)$ as follows:

$$\begin{aligned} \mathbf{P}_2(x) &= \mathbf{P}(C(0, u) < (b - \varepsilon)u, u \in (0, x)) \\ &= \mathbf{P}(C(0, ux) < (b - \varepsilon)ux, u \in (0, 1)) \\ &= \mathbf{P}\left(\frac{1}{x}C(0, ux) < (b - \varepsilon)u, u \in (0, 1)\right) \\ &= \mathbf{P}\left(\frac{1}{x}C(0, \cdot x) \in S\right), \end{aligned}$$

where $S := \{f \in \Omega : f(u) < (b - \varepsilon)u, u \in (0, 1)\}$. As we assumed that $C(0, \cdot x)/x$ obeys the lower bound of the sp-LDP (Assumption 3.4), we have

$$\begin{aligned} \liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}\left(\frac{1}{x}C(0, ux) \in S\right) &\geq -\inf_{f \in S^o} I(f) =: -I^*, \\ I(f) &:= \int_0^1 \Lambda(f'(t))dt, \end{aligned}$$

where we recall that $\Lambda(t) = \sup_{s \in \mathbb{R}} (st - c(s))$. Since the infimum of $I(f)$ over all $f \in S^o$ is not larger than $I(f^*)$ for any particular $f^* \in S^o$, taking $f^*(u) := (b - \bar{\varepsilon})u$ with $\bar{\varepsilon} := \varepsilon(1 + \delta)$ for some small $\delta > 0$, we obtain the lower bound

$$-I^* \geq -\sup_{s \in \mathbb{R}} ((b - \bar{\varepsilon})s - c(s)).$$

Observe that since the constant b is chosen such that $b < c$, the supremum is attained for $s \leq 0$. Hence, we may write

$$\begin{aligned} \liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}_2(x) &\geq -I^* \geq -\sup_{s \leq 0} ((b - \bar{\varepsilon})s - c(s)) \\ &= -\sup_{s \geq 0} (-(b - \bar{\varepsilon})s - c(-s)) \\ &= \inf_{s \geq 0} ((b - \bar{\varepsilon})s + c(-s)). \end{aligned}$$

Step (v). By combining the results for $\mathbf{P}_1(x)$ and $\mathbf{P}_2(x)$ we find that, for any $b < c$,

$$\begin{aligned} \liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)) \\ \geq \inf_{s \geq 0} (\alpha(s) - bs) + \inf_{s \geq 0} ((b - \bar{\varepsilon})s + c(-s)). \end{aligned} \tag{4.6}$$

Take $\varepsilon > 0$ sufficiently small and note that $\log \Phi_B(\cdot)$ is convex, $(\log \Phi_B(\cdot))' = \Phi'_B(\cdot)/\Phi_B(\cdot)$ is increasing and, due to Assumption 3.2, is finite and continuous in a neighborhood of ω^* . Similar arguments yield that $\Phi'_A(\cdot)/\Phi_A(\cdot)$ is an increasing, finite and continuous function as well, and $\alpha(\cdot)$ is continuous and increasing. Thus, there exists an $\varepsilon > 0$ for which there is $\omega = \omega_\varepsilon$ such that $\Phi_B(\omega_\varepsilon) < \infty$, $\Phi'_B(\omega_\varepsilon) < \infty$, and $\alpha'(\omega_\varepsilon) - c'(-\omega_\varepsilon) = \bar{\varepsilon}$. Since $\alpha(\cdot) + c(\cdot)$ is a strictly convex function (this follows from the fact that

$\alpha(\cdot)$ is strictly convex and $c(\cdot)$ is convex), $\alpha'(\cdot) - c'(\cdot)$ is increasing and hence, ω_ε is the unique solution. The continuity properties imply that $\lim_{\varepsilon \rightarrow 0} \omega_\varepsilon = \omega^*$.

Let us now take $b := \alpha'(\omega_\varepsilon)$ in (4.6). Note that this choice satisfies requirement $b < c$: since the cumulant function $c(\cdot)$ is a convex function, its derivative is increasing, and consequently, for $\bar{\varepsilon}$ small, $b = c'(-\omega_\varepsilon) + \bar{\varepsilon} < c'(0) = c$.

Now consider the first optimization in (4.6): $\inf_{s \geq 0} (\alpha(s) - \alpha'(\omega_\varepsilon)s)$. It is readily checked that its first order condition is $\alpha'(s) = \alpha'(\omega_\varepsilon)$, which is obviously met for $s = \omega_\varepsilon$ (and there is just at most one solution, so ω_ε is the unique minimizer). The first order condition for the second optimization in (4.6) is then $\alpha'(\omega_\varepsilon) - c'(-s) = \bar{\varepsilon}$, which is by definition solved for $s = \omega_\varepsilon$. We conclude that

$$\begin{aligned} & \inf_{s \geq 0} (\alpha(s) - bs) + \inf_{s \geq 0} ((b - \bar{\varepsilon})s + c(-s)) \\ &= \inf_{s \geq 0} (\alpha(s) + c(-s) - \bar{\varepsilon}s). \end{aligned}$$

Now let $\varepsilon \rightarrow 0, \delta \rightarrow 0$ (and hence also $\bar{\varepsilon} \rightarrow 0$). Due to continuity we have that $\omega_\varepsilon \rightarrow \omega^*$, and consequently,

$$\begin{aligned} & \liminf_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)) \\ & \geq \inf_{s \geq 0} (\alpha(s) + c(-s)). \end{aligned}$$

This completes the proof. \square

Proof of Theorem 3.5 The proof strongly resembles that of Theorem 3.4. We leave it to the reader to check that only the argumentation in Step (iv) needs to be modified. This step relies on the validity of the lower bound of the sp-LDP, and to our best knowledge, an sp-LDP for the processes in $\mathbb{Mf}(Q, R)$ is not available from the literature. Therefore we need a different approach to analyze the large deviation $\mathbf{P}_2(x)$ of the service process $C(\cdot, \cdot)$. The main idea of this modification is to apply results of Chang [8] for Markov-type processes in discrete time. For that we need to cast our model into Chang’s framework. This is done as follows.

Consider, as before, $\mathbf{P}_2(x) = \mathbf{P}(C(0, u) < (b - \varepsilon)u, u \in (0, x))$. For any fixed $M < x$ and $C_M < (b - \varepsilon)M$,

$$\begin{aligned} \mathbf{P}_2(x) & \geq \mathbf{P}(C(0, u) < (b - \varepsilon)u, u \in (0, x), \\ & C(0, M) < C_M, X(M) = j), \end{aligned}$$

as the event in the right hand side is fully contained in that of the left hand side. Now consider separately the intervals $(0, M]$ and (M, x) . By using the conditional independence and a straightforward time-shift, we have that the previous probability is not smaller than

$$\begin{aligned} & \mathbf{P}(C(0, u) < (b - \varepsilon)u, u \in (0, M), C(0, M) < C_M, \\ & X(M) = j) \times \bar{\mathbf{P}}_2(x), \quad \text{where} \end{aligned}$$

$$\begin{aligned} \bar{\mathbf{P}}_2(x) & := \mathbf{P}(C(0, u) < (b - \varepsilon)u + (b - \varepsilon)M - C_M, \\ & u \in (0, x - M) \mid X(0) = j). \end{aligned}$$

Observe that the former probability is constant in x ; therefore we need to concentrate just on $\bar{\mathbf{P}}_2(x)$. Now the fact that the service rate is bounded by r_{\max} entails

$$C(0, u) \leq C\left(0, \left\lfloor \frac{u}{\delta} \right\rfloor \delta\right) + r_{\max} \delta,$$

for any δ . As a consequence, $\bar{\mathbf{P}}_2(x)$ majorizes

$$\begin{aligned} & \mathbf{P}\left(C(0, i\delta) + r_{\max} \delta < (b - \varepsilon)i\delta + (b - \varepsilon)M - C_M, \right. \\ & \left. i = 0, \dots, \left\lfloor \frac{x - M}{\delta} \right\rfloor \mid X(0) = j\right). \end{aligned}$$

Let us take $\delta < ((b - \varepsilon)M - C_M)/r_{\max}$. Then the probability in the previous display is not smaller than

$$\begin{aligned} & \mathbf{P}\left(C(0, i\delta) \leq (b - \varepsilon)i\delta, \right. \\ & \left. i = 0, \dots, \left\lfloor \frac{x - M}{\delta} \right\rfloor \mid X(0) = j\right). \end{aligned}$$

The process $C_i := C(0, i\delta) = \sum_{u=0}^{i\delta} R(u) = \sum_{u=0}^{i\delta} r_{X(u)}(u)$ is now a (discrete-time) process that fits in the framework considered in Example 2.5 in Chang [8] and proven to satisfy sp-LDP with rate function as given in Definition 3.1. Applying the sp-LDP lower bound on the last probability, it is straightforward that the decay rate (in x) of the latter probability is indeed

$$-\sup_{s \geq 0} ((b - \varepsilon)s - c(s)),$$

as desired. Proceeding with Step (v) as before completes the proof. \square

Remark 4.1 An interesting implication of our results is the following. Denoting by P_r the residual busy period in the corresponding GI/GI/· queue (with the service rate varying as specified before), we have

$$\begin{aligned} \mathbf{P}(V > x) & \leq \mathbf{P}(P_r > x) \\ & \leq \mathbf{P}(W + B_0 + A(0, x) - C(0, x) > 0). \end{aligned}$$

Thus, from the proof of the lower bound in the next section, it follows that the decay rates of P_r and V coincide under Assumptions 3.1, 3.3 and 3.4. Similar methods as in the present paper can be applied to show that the decay rate of the busy period P equals the decay rate of the residual busy period P_r . Consequently, under Assumptions 3.1, 3.3 and 3.4,

$$\log \mathbf{P}(P > x) \sim (\alpha(\omega^*) + c(-\omega^*))x.$$

The most likely way a large busy period occurs is by changing the drift in such a way that the system becomes critical. This means that the average arrival rate is increased, and the service rate decreased in such a way that they become equal. In our setting, the arrival rate is increased from $\alpha'(0)$ to $\alpha'(\omega^*)$, and the service rate is decreased from $c'(0)$ to $c'(-\omega^*)$; the definition of ω^* ensures that both values are equal.

5 Extension to Discriminatory Processor Sharing

We now consider the extension of our analysis to the GI/GI/queue operating under Discriminatory Processor Sharing (DPS) with varying service rate. The proof indicates that essentially the same argumentation can be used as in the case of PS with varying service rate (as dealt with in the previous sections).

Suppose that there are M job classes sharing the available capacity. The aggregate job arrival process is assumed to be a renewal process as considered in Sect. 2. An arriving job is of type k with probability p_k , $k = 1, \dots, M$. The service times B_n in Sect. 2 denote the unconditional service times of jobs (for our purposes, we do not need to specify the conditional service time distributions). Thus, the asymptotic cumulant generating function of the aggregate arrival process is still given by $\alpha(s)$.

All jobs present in the system are served simultaneously with rates controlled by a vector of weights $\{g_k > 0; k = 1, \dots, M\}$. If there are Q_j jobs of class j present in the system, $j = 1, \dots, M$, each class- k job is served at rate $g_k / \sum_{j=1}^M g_j Q_j$. When all the weights are equal, this is equivalent to the standard (i.e., egalitarian) Processor Sharing (PS) system. By changing the DPS weights, one can effectively control the instantaneous service rates of different job classes.

The proofs of the previous section show that the logarithmic sojourn-time asymptotics coincide with the logarithmic busy-period asymptotics. The following theorem states that the same result holds in the DPS queue, regardless of the specific values of the weight factors.

Suppose the tagged job belongs to class 1. Denote by V_1 its sojourn time, and by B_0^1 its job size.

Theorem 5.1 *If Assumptions 3.1, 3.3 and 3.4 are valid, then*

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbf{P}(V_1 > x) = \inf_{s > 0} (\alpha(s) + c(-s)).$$

Thus, the large deviations estimate does not change when one assigns different weights to the various customer classes. This may not be surprising since we already obtained the insight that on a large deviations scale, a large sojourn-time resembles a large busy period. The decay rate

of the latter is obviously weight-independent (as the length of a busy period is the same for all work conserving service disciplines, such as DPS).

On the one hand, this asymptotic insensitivity might be considered as a negative fact. It says that independent of the particular weights assignment, the DPS discipline does not reduce the likelihood of extremely long sojourns. Long sojourn times are inevitable, since they are typically caused by the large amount of work brought by customers during the service of the tagged job. On the other hand, the insensitivity property may be regarded as a positive result, because it implies that preferential treatment of classes with large weights does not carry the penalty of increasing the occurrence of long sojourn times for classes with smaller weights.

Proof of Theorem 5.1

The proof of the upper bound uses the same arguments as for the single class PS queue, which we will not repeat here. The proof of the lower bound is similar to that of Theorem 3.4. We truncate the work process by accepting into the system only jobs of size smaller than k and proceed in the similar fashion as before. The only extra step involves the maximal weight $g_{\max} = \max_{j=1, \dots, M} g_j$,

$$\begin{aligned} \mathbf{P}(V_1 > x) &\geq \mathbf{P}\left(B_0^1 > \int_0^x \frac{g_1 dC(0, u)}{1 + \sum_{j=1}^M g_j Q_j(u)} \mid \right. \\ &\quad \left. A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)\right) \\ &\quad \times \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)) \\ &\geq \mathbf{P}\left(B_0^1 > \int_0^x \frac{g_1 dC(0, u)}{1 + g_{\max} \sum_{j=1}^M Q_j(u)} \mid \right. \\ &\quad \left. A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)\right) \\ &\quad \times \mathbf{P}(A_k(0, u) - C(0, u) > \varepsilon u, u \in (0, x)). \end{aligned}$$

Since the jobs are not larger than k , under present condition, the total number of jobs

$$\sum_{j=1}^M Q_j(u) \geq \frac{\varepsilon u}{k}.$$

It is now straightforward to verify that the first probability behaves as $e^{\rho(x)}$ when $x \rightarrow \infty$. The second probability gives the desired decay rate. For details see Theorem 3.4. \square

6 Concluding remarks

We have considered a PS queue with fluctuating service rate and light-tailed service times and established logarithmic

tail asymptotics for the sojourn time in a Processor Sharing queue. Our results indicate that on a large-deviations scale, the event of a large sojourn time is equivalent to the event of a large busy period. In particular, the system is critically loaded during the sojourn time of a customer. Our proof method turned out to be powerful enough to extend to Discriminatory Processor Sharing. As shown in Sect. 4, all that is required is to be able to compute the decay rate of the busy period distribution. It is rather surprising that this decay rate has not been found yet under assumptions of the type considered in [12]; we were not able to deal with this level of generality, and consider this to be an interesting open problem. Another open problem is to show that the sample-path LDP holds for the Markov fluid service process. As mentioned before, we could not find any direct reference in the applied probability literature. Finally, we expect that the results in this paper are useful to develop importance sampling techniques.

References

1. Antunes, N., Fricker, C., Guillemin, F., Robert, P.: Perturbation analysis of a variable M/M/1 queue: a probabilistic approach. *Adv. Appl. Probab.* **38**, 263–283 (2006)
2. Asmussen, S.: *Applied Probability and Queues*. Springer, New York (2003)
3. Bekker, R., Borst, S.C., Núñez-Queija, R.: Performance of TCP-friendly streaming sessions in the presence of heavy-tailed elastic flows. In: *Performance Evaluation Review. Proceedings of MAMA 2004 Workshop* (2004)
4. Borst, S.C., Boxma, O.J., Morrison, J.A., Núñez-Queija, R.: The equivalence between processor sharing and service in random order. *Oper. Res. Lett.* **31**, 254–262 (2003)
5. Borst, S.C., Núñez-Queija, R., van Uiter, M.J.G.: User-level performance of elastic traffic in a differentiated-services environment. *Perform. Eval.* **49**, 507–519 (2002)
6. Borst, S.C., Núñez-Queija, R., Zwart, A.P.: Sojourn time asymptotics in processor-sharing queues. *Queueing Syst.* **53**, 31–51 (2006)
7. Borst, S.C., van Ooteghem, D., Zwart, A.P.: Tail asymptotics for discriminatory processor-sharing queues with heavy-tailed service requirements. *Perform. Eval.* **61**, 281–298 (2005)
8. Chang, C.-S.: Sample path large deviations andintree networks. *Queueing Syst.* **20**, 7–36 (1995)
9. Dembo, A., Zeitouni, O.: *Large Deviation Techniques and Applications*. Springer, New York (1998)
10. Egorova, R., Zwart, B., Boxma, O.: Sojourn time tails in the M/D/1 processor sharing queue. *Probab. Eng. Inf. Sci.* **20**, 429–446 (2006)
11. Flatto, L.: The waiting time distribution for the random order service M/M/1 queue. *Ann. Appl. Probab.* **7**, 382–409 (1997)
12. Glynn, P.W., Whitt, W.: Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab. A* **31**, 131–156 (1994)
13. Jelenković, P., Momčilović, P.: Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Math. Oper. Res.* **28**, 587–608 (2004)
14. Kesidis, G., Walrand, J., Chang, C.-S.: Effective bandwidths for multi-class Markov fluids and other ATM sources. *IEEE/ACM Trans. Netw.* **1**, 424–428 (1993)
15. Mandjes, M., Zwart, A.P.: Large deviations for sojourn times in processor sharing queues. *Queueing Syst.* **52**, 237–250 (2006)
16. Núñez-Queija, R.: *Processor-sharing models for integrated-services networks*. PhD thesis, Eindhoven University of Technology (2000)
17. Nuyens, M., Zwart, A.P.: A large deviations analysis of the GI/GI/1 SRPT queue. *Queueing Syst.* **54**, 85–97 (2006)
18. Ott, T.J.: The sojourn-time distribution in the M/G/1 queue with processor sharing. *J. Appl. Probab.* **21**, 360–378 (1984)
19. Palmowski, Z., Rolski, T.: On the exact asymptotics of the busy period in GI/GI/1 queues. *Adv. Appl. Probab.* **38**, 792–803 (2006)
20. Roberts, J.: Traffic theory and the internet. *IEEE Commun. Mag.* **39**, 94–99 (2001)
21. Schassberger, R.: A new approach to the M/G/1 processor sharing queue. *Adv. Appl. Probab.* **16**, 802–813 (1984)
22. Whitt, W.: Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommun. Syst.* **2**, 71–107 (1993)
23. Yashkov, S.F.: A derivation of response time distribution for an M/G/1 processor-sharing queue. *Probl. Control Inf. Theory* **12**, 133–148 (1983)
24. Zwart, A.P., Boxma, O.J.: Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Syst.* **35**, 141–166 (2000)
25. Zhang, Z.-L.: Large deviations and the generalized processor sharing scheduling for a two-queue system. *Queueing Syst.* **26**, 229–254 (1997)