Queueing Syst (2008) 60: 227–246
DOI 10.1007/s11134-008-9095-4

Steady state approximations of limited processor sharing queues in heavy traffic

Jiheng Zhang · Bert Zwart

Received: 28 March 2008 / Revised: 9 September 2008 / Published online: 12 November 2008

Abstract We investigate steady state properties of limited processor sharing queues in heavy traffic. Our analysis builds on previously obtained process limit theorems, and requires the interchange of steady state and heavy traffic limits, which are established by a coupling argument. The limit theorems yield explicit approximations of the steady state queue length and response time distribution in heavy traffic, of which the quality is supported by simulation experiments.

Keywords Limited processor sharing · Measure-valued process · Steady state · Heavy traffic · Queue size · Delay probability · Response time

Mathematics Subject Classification (2000) Primary 60K25 · Secondary 68M20 · 90B22 · 68M07

1 Introduction

We consider the limited processor sharing (LPS) queue which is a generalization of the processor sharing (PS) queue. As inferred by the name, we limit the number of jobs that can share the server at any time by $K \geq 1$, instead of letting all the jobs share the server. The server is shared equally by those jobs in service, i.e. at any time each job in service is processed at a rate that is the reciprocal of the number of jobs in service. An arriving job will immediately enter the server and start receiving service if there are less than K jobs in the server when it arrives; otherwise it will wait in the

This research is supported in part by National Science Foundation grant CNS-0718701.

J. Zhang (✉) · B. Zwart
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: jrz@gatech.edu

B. Zwart
e-mail: bertzwart@gatech.edu

buffer. When the number of jobs in the server drops from K to $K-1$, the server will immediately admit the longest waiting job from the buffer if there is any. A job will leave the system immediately after the server has fulfilled its service requirement. This is quite a general model since letting $K=\infty$ makes the system a PS queue and taking $K=1$ reduces the system to a first-come-first-serve (FCFS) queue.

The PS model has been widely used in the analysis of computer systems, network servers and data transmission over the Internet. The PS discipline can be viewed as an idealization of time-sharing protocols in computer systems, as described in [17] and [20]. The advantage is that a big job will not block the whole system as in a FCFS queue. However, allowing too many jobs to time-share at once can lead to significant overhead (due to switching), hence reduce overall performance. This point has already been observed in early papers on operating systems [4, 6], as well as in more recent studies on Web server design [7, 16], and databases [15, 21]. So in several applications, a sharing limit is normally imposed, which results in the LPS model.

Despite its wide range of applications, there are only a few studies on the LPS queue. Avi-Itzhak and Halfin [2] propose an approximation for the mean response time assuming Poisson arrivals. A computational analysis based on matrix geometric methods is performed in Zhang and Lipsky [23, 24]. Some stochastic ordering results are derived in Nuyens and van der Weij [19]. Recently, Zhang, Dai and Zwart [25, 26] studied the stochastic processes underlying the LPS queue in the heavy traffic regime, an asymptotic regime where the traffic intensity converges to 1. The study was carried in a general setting, allowing the inter-arrival and service times to have general distributions. Since the exact performance analysis of the LPS queue seems not tractable, we are interested in obtaining approximations for the performance characteristics of the system.

To give a proper description of the system dynamics, it is necessary to use a *measure-valued state descriptor*: we need finite Borel measures on $\mathbb{R}_+=(0,\infty)$ to describe the system. At any time $t\geq 0$, we record all the remaining service times using a measure $\mathcal{Z}(t)$. For any Borel set $B\subset\mathbb{R}_+$, $\mathcal{Z}(t)(B)$ indicates the number of jobs in server with remaining service time belonging to B at that time. Similarly, we use a measure $\mathcal{Q}(t)$ to describe the state at the buffer. At time $t\geq 0$, $\mathcal{Q}(t)(B)$ indicates the number of jobs in buffer with job size belonging to B. The descriptor $(\mathcal{Q}(\cdot),\mathcal{Z}(\cdot))$ contains a wealth of information; almost all the usual performance processes (such as workload and queue length) can be recovered from it. More details will be discussed when we give a detailed model description in Sect. 2. In fact, the framework of using measure-valued process has been successfully applied to study models where multiple jobs are processed at the same time at the process level. Existing works include Gromoll, Puha and Williams [12], Gromoll and Kruk [11] and Gromoll, Robert and Zwart [13] etc. Zhang, Dai and Zwart [25] use the framework to obtain the diffusion limit of the LPS queue, i.e. the limit of a sequence of diffusion scaled processes $(\hat{\mathcal{Q}}^r(\cdot),\hat{\mathcal{Z}}^r(\cdot))$ (the scaling is defined in (3.1)) as $r\to\infty$.

In this paper, we use the measure-valued framework to study the steady state limit of the LPS queue. The measure-valued process is still regenerative, and will converge weakly as $t\to\infty$ to a steady state. However, no explicit solution for the stationary distribution seems available. On the other hand, the diffusion limit established in

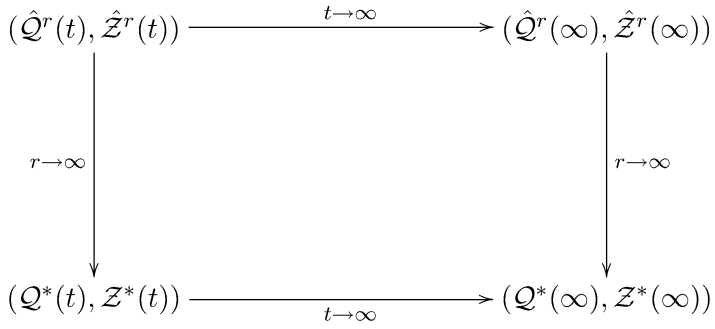

Fig. 1 The interchange of heavy traffic and steady state limits

Zhang, Dai and Zwart [25], which is denoted by $(\mathcal{Q}^*(\cdot), \mathcal{Z}^*(\cdot))$ is tractable and its steady state limit $(\mathcal{Q}^*(\infty), \mathcal{Z}^*(\infty))$ is, as we show in this paper, rather explicit, cf. (4.2).

The first contribution of the current paper is to establish the interchange of heavy traffic and steady state limit as depicted in Fig. 1. The main idea is to couple the measure-valued process for the LPS queue with its corresponding stationary version. This helps to obtain a version of the classical coupling inequality, cf. (4.6). The interchange can be established after we prove the uniform convergence of the upper bound for the coupling inequality in Lemma 4.1. It should be pointed out that the framework of using coupling to establish interchange of the heavy traffic and steady state limit works for all single buffer single server systems in work conserving disciplines. In particular, for the classical PS queue our technique works as well, and allows one to recover Grishechkin's [9] steady state approximation from Gromoll's [10] process limit theorem. The main point is that, although we are dealing with a measure-valued process, the limit interchange problem is relatively tractable since the workload process is explicitly known. For networks, the interchange is more involved, cf. Budhiraja and Lee [5] and Gamarnik and Zeevi [8]. The idea of interchanging limits is also applied in other models, cf. Maulik and Zwart [18].

The validity of the interchange, established in Theorem 4.1, provides the necessary theoretical support for using the tractable limit $(\mathcal{Q}^*(\infty), \mathcal{Z}^*(\infty))$ as approximation of the steady state of a given LPS queue. Section 5 in this paper demonstrates how to analyze performance quantities such as queue size, delay probability and response times. As implication of Theorem 4.1, Corollaries 5.2 and 5.3 analyze the performance quantities such as queue size, delay probability and response time. Along the way, Proposition 5.1 states a process limit for the time-dependent virtual response time process. Note that the analysis of (virtual) response times in heavy traffic is non-trivial, as in the standard PS queue [11].

From a practical perspective, the main insights of this paper are the approximation formulas (6.1)–(6.4) for queue size, delay probability and response times. In particular, our results show that the following two-moment approximation of the queue size $E[X]$ is accurate in heavy traffic:

$$\mathbb{E}[X] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{\rho}{1-\rho}(1 - d_p) + \frac{c_a^2 + c_s^2}{2} \frac{\rho}{1-\rho} d_p, \tag{1.1}$$

$$d_p \approx \rho^{\frac{1+c_s^2}{c_a^2+c_s^2}K}. \tag{1.2}$$

In the above display, c_a^2 and c_s^2 are squared coefficients of variation for the inter-arrival time and service time, ρ is the traffic intensity and d_p is the probability that a customer cannot enter service immediately upon arrival. Interestingly, our approximation is consistent with that of Avi-Itzhak and Halfin [2] if the arrival process is Poisson.

The paper is organized as follows: In Sect. 2, we present the mathematical model and study the steady state limit of the LPS queue. A brief review of heavy traffic limit theorems that will be used in this paper is given in Sect. 3. Section 4 establishes the validity of interchanging the heavy traffic and steady state limit of the LPS queue. The interchange provides the foundation for performance analysis in Sect. 5. Finally, based on the previous analysis, we obtain some approximation formulas for various performance quantities in Sect. 6, where some simulation results are also presented to show the quality of the approximation formulas.

2 The LPS queue and its steady state limit

We consider a $G/G/1$ queue operating under the LPS policy, with the sharing limit equal to K. Let $Q(t)$, $Z(t)$, and $X(t)$ denote the number of jobs in the buffer, number of jobs in service, and the total number of jobs in the system at time t, respectively. Thus,

$$X(t) = Q(t) + Z(t), \quad t \geq 0.$$

The system is allowed to be non-empty initially, i.e. $X(0) > 0$. We index jobs by $i = -X(0)+1, -X(0)+2, \ldots, 0, 1, \ldots$. The first $X(0)$ jobs are initially in the system, with jobs $i = -X(0)+1, \ldots, -Q(0)$ in service and jobs $i = -Q(0)+1, \ldots, 0$ waiting in the buffer. Jobs arrived after time 0 are indexed by $i = 1, 2, \ldots$. Let $E(t)$ denote the number of jobs that arrive to the system during time interval $(0, t]$, for all $t \geq 0$. According to the policy, a job may have to wait for a certain amount of time after arrival to get service. Let w_i denote the waiting time, and U_i denote the arrival time of the ith job for all $i > -X(0)$. By convention, $U_i = 0$ for $i \leq 0$, and $w_i = 0$ for $i \leq -Q(0)$. The quantity

$$\tau_i = U_i + w_i, \quad i > -X(0),$$

can be viewed as the time that the ith job starts service. We use v_i to denote the job size of the ith job for all $i > -Q(0)$. We assume that $E(\cdot)$ is a renewal process and $\{v_i\}_{i=-\infty}^{\infty}$ is a sequence of i.i.d. random variables with distribution F, and the sequence is independent of the arrival process $E(\cdot)$. For jobs with index $-X(0) < i \leq -Q(0)$, i.e. the first $Z(0)$ jobs that are initially in service, we use $\tilde{v}_i$ to denote the job sizes of these jobs. They are not assumed to be i.i.d. We call $\{E(\cdot), \{v_i\}_{i=1}^{\infty}\}$ the stochastic primitives of the system, and $\{Z(0), Q(0), \{v_i\}_{i=-\infty}^{0}, \{\tilde{v}_i\}_{i=-\infty}^{0}\}$ the initial conditions of the system.

As in [25, 26], we introduce a measure-valued state descriptor $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$, which is rich enough to describe the evolution of the system with given initial conditions and

stochastic primitives. For any Borel set $A \subset \mathbb{R}_+$, $\mathcal{Q}(t)(A)$ denotes the total number of jobs in buffer whose job size belongs to A; and $\mathcal{Z}(t)(A)$ denotes the total number of jobs in service whose residual job size belongs to set A. Denote $\mathbf{M}$ the space of all non-negative finite Borel measures on $[0, \infty)$. For any $\nu_1, \nu_2 \in \mathbf{M}$, let $\mathbf{d}[\nu_1, \nu_2]$ denote the Prohorov metric between the two measures (cf. p. 72, [3]). For any Borel measurable function $f : \mathbb{R}_+ \to \mathbb{R}$, the integration of this function with respect to any measure $\nu_0 \in \mathbf{M}$ is denoted by $\langle f, \nu_0 \rangle$. It is clear that we have the following relationship:

$$Q(t) = \big\langle 1, \mathcal{Q}(t) \big\rangle, \qquad Z(t) = \big\langle 1, \mathcal{Z}(t) \big\rangle.$$

Let λ be the rate of the arrival process $E(\cdot)$. Denote ν the probability measure of the service time (i.e. ν is the probability measure associated with the distribution F, it is clear that $\nu \in \mathbf{M}$). The traffic intensity of the LPS queue is defined as $\rho = \lambda \langle \chi, \nu \rangle$.

Define the *cumulative service amount* up to time t by

$$S(t) = \int_0^t \psi\big(Z(\tau)\big)\, d\tau, \tag{2.1}$$

where $\psi(x) = 1/x$ if $x > 0$ and $\psi(x) = 0$ if $x = 0$. A job will have received a cumulative amount of processing time

$$S(s, t) = \int_s^t \psi\big(Z(\tau)\big)\, d\tau$$

during time interval $[s, t]$ if it is in service in this time period. Let

$$B(t) = E(t) - Q(t). \tag{2.2}$$

Note that at time $t \geq 0$, $B(t)$ is the index of the last job who has entered into service by time t. Thus

$$B(s, t) = B(t) - B(s) \tag{2.3}$$

represents the number of jobs which have left the buffer and entered the server during time interval $(s, t]$. Using the notation introduced in this section, the state descriptor can be written as

$$\mathcal{Q}(t)(A) = \sum_{i=B(t)+1}^{E(t)} \delta_{v_i}(A), \tag{2.4}$$

$$\mathcal{Z}(t)(A) = \sum_{i=-X(0)+1}^{-Q(0)} \delta_{\tilde{v}_i}\big(A + S(t)\big) + \sum_{i=-Q(0)+1}^{B(t)} \delta_{v_i}\big(A + S(\tau_i, t)\big), \tag{2.5}$$

for any Borel set $A \subset \mathbb{R}_+$, where $\delta_a(A)$ denotes the Dirac measure of point a on $\mathbb{R}$ and $A + y = \{a + y : a \in A\}$. Due to the LPS policy, the sharing limit K must be enforced at any time t,

$$Q(t) = \big(X(t) - K\big)^+, \tag{2.6}$$

$$Z(t) = \big(X(t) \wedge K\big). \tag{2.7}$$

We call (2.4) and (2.5) the *stochastic dynamic equations* and (2.6) and (2.7) the policy constraints.

It is clear that the measure-valued process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$ is a regenerative process. Using the measure-valued descriptor, we can recover the workload $W(t)$ at time $t > 0$ by

$$W(t) = \big\langle \chi, \mathcal{Q}(t) + \mathcal{Z}(t) \big\rangle, \tag{2.8}$$

where χ denotes the identity function on $\mathbb{R}$. Let $R_0 = 0$ and define the regenerative points R_n, $n \geq 1$, as the following:

$$R_n = \inf\big\{t > R_{n-1} : W(t-) = 0 \text{ and } W(t) > 0\big\}. \tag{2.9}$$

The regeneration points are those time epochs that the workload jumps from 0. It is clear that the jump happens because of the new arrival. By (2.8), $(\mathcal{Q}(t), \mathcal{Z}(t)) = (\mathbf{0}, \mathbf{0})$ if and only if $W(t) = 0$ for any $t \geq 0$. So the process starts from empty at time R_n with a new job just arriving at R_n. Thus, the evolution of the process from time R_n onwards does not depend on any information of the process before that time.

Note that the workload process of a single buffer single server system is the same for all non-idling policies. It is well known that the workload process (for any non-idling policy) is a delayed regenerative process if $W(0) > 0$, and the above definition of R_n is one way to define the regenerative points. By Proposition 3.1 in Chap. X of [1], the mean of the regenerative cycles Y_i's ($Y_i = R_i - R_{i-1}$) with $i > 1$ is finite if $\rho < 1$. By Proposition 3.2 in Chap. X of [1], the distribution of them is non-lattice if the service time distribution F is non-lattice.

In summary, the process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$ can be modeled as a delayed regenerative process. Denote $\mathbb{E}_0(\cdot) = \mathbb{E}(\cdot|(\mathcal{Q}(0), \mathcal{Z}(0)) = (\mathbf{0}, \delta_{v_1}), U_1 = 0)$, that is, the expectation operator given that the first job arrives to an empty system at time 0. We write $Y = Y_1$ for the length of the first cycle. Let $\mathbf{M} \times \mathbf{M}$ denote the Cartesian product. There are a number of ways to define the metric on the product space. For convenience, we define the metric to be the maximum of the Prohorov metric between each component, so that the product space is still a Polish space. Now, we define a distribution π on the product space by

$$\pi(A) = \frac{1}{\mathbb{E}_0 Y} \mathbb{E}_0 \int_0^Y 1_{\{(\mathcal{Q}(s), \mathcal{Z}(s)) \in A\}}\, ds,$$

for any Borel set $A \in \mathbf{M} \times \mathbf{M}$. The following result about the steady state distribution of the LPS queue follows directly from Theorem 1.2 in Chap. X of [1].

Proposition 2.1 (Stochastic stability of LPS) *Suppose that the traffic intensity $\rho < 1$ and the service time distribution F is non-lattice. The above defined distribution π is the unique stationary distribution for the measure-valued process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$. The distribution of $(\mathcal{Q}(t), \mathcal{Z}(t))$ converges to π as $t \to \infty$.*

The above theorem establishes the convergence of the regenerative process to the steady state limit, which has the stationary distribution π. In fact, there exists

a stationary version $(\mathcal{Q}_\pi(\cdot), \mathcal{Z}_\pi(\cdot))$ of the regenerative process (see [22]) such that the marginal distribution at any time $t \geq 0$ is π. The stationarity of the process $(\mathcal{Q}_\pi(\cdot), \mathcal{Z}_\pi(\cdot))$ will help to obtain a coupling inequality in Sect. 4.

3 Process level limit theorems

Process level limit theorems for the LPS queue are developed in [25]. Some of the results there will be used throughout this paper. For completeness and reader's convenience, we briefly summarize the necessary definitions, results and notations in this section.

We consider a family of limited processor sharing queues indexed by r, where r increases to ∞ through a sequence in $(0, \infty)$. Each queueing model is defined in the same way as in Sect. 2, and each of them is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. To distinguish models with different indices, quantities of the rth model are accompanied by superscript r.

Our results concern the asymptotic behavior of the descriptor under the *diffusion* scaling, which is defined by

$$\left(\hat{\mathcal{Q}}^r(t), \hat{\mathcal{Z}}^r(t)\right) = \left(\frac{1}{r}\mathcal{Q}^r\left(r^2 t\right), \frac{1}{r}\mathcal{Z}^r\left(r^2 t\right)\right) \quad \text{for all } t \geq 0, \tag{3.1}$$

in the *heavy traffic* regime. Similarly as in Sect. 2, define the traffic intensity of the rth system by $\rho^r = \lambda^r \langle \chi, \nu^r \rangle$, where λ^r and ν^r denote the arrival rate and service time measure of the rth system, respectively. The heavy traffic regime is a parameter regime specified by the following conditions:

$$\lim_{r\to\infty} r\left(1 - \rho^r\right) = \theta, \tag{3.2}$$

$$\lim_{r\to\infty} K^r / r = K, \tag{3.3}$$

for some positive constants θ, K. To establish the heavy traffic limit, we also need the following regularity conditions, which are quite general and standard. We assume that the arrival processes satisfy

$$\frac{E^r(r^2\cdot) - \lambda^r r^2 \cdot}{r} \Rightarrow E^*(\cdot) \quad \text{as } r \to \infty, \tag{3.4}$$

where

$$\lim_{r\to\infty} \lambda^r = \lambda > 0, \tag{3.5}$$

and $E^*(\cdot)$ is a Brownian motion with drift 0 and variance λc_a^2. And the measures of job sizes satisfy that, as $r \to \infty$,

$$\mathbf{d}\left[\nu^r, \nu\right] \to 0, \tag{3.6}$$

$$\left\langle \chi^{2+2p}, \nu^r \right\rangle \to \left\langle \chi^{2+2p}, \nu \right\rangle \quad \text{for some } p > 0, \tag{3.7}$$

where the probability measure ν satisfies

$$\nu \text{ has no atoms.} \tag{3.8}$$

Let $\beta = \langle \chi, \nu \rangle$ be the mean and $c_s^2 = \frac{\langle \chi^2, \nu \rangle - \beta^2}{\beta^2}$ be the squared coefficient of variation (SCV) of the job size distribution ν. Also, the following initial condition will be assumed:

$$\big(\hat{\mathcal{Q}}^r(0), \hat{\mathcal{Z}}^r(0)\big) \Rightarrow \big(\xi^*, \mu^*\big), \tag{3.9}$$

$$\big\langle \chi^{1+p}, \hat{\mathcal{Q}}^r(0) + \hat{\mathcal{Z}}^r(0) \big\rangle \Rightarrow \big\langle \chi^{1+p}, \xi^* + \mu^* \big\rangle, \tag{3.10}$$

as $r \to \infty$, where p is the same as in (3.7), $(\xi^*, \mu^*) \in \mathcal{I}$ and

$$\mu^* \text{ has no atoms,} \tag{3.11}$$

where $\mathcal{I}$ denotes the set of all $(\xi, \mu) \in \mathbf{M} \times \mathbf{M}$ that satisfies

$$\xi = \big(\langle 1, \xi + \mu \rangle - K\big)^+ \nu,$$
$$\langle 1, \mu \rangle = \langle 1, \xi + \mu \rangle \wedge K.$$

Roughly speaking, all initial states must be consistent with the limited sharing policy and their initial waiting jobs have the same service time distribution as arriving jobs.

The following proposition is a well-known heavy traffic approximation for the workload process of a single queue operated under a non-idling policy, including the LPS policy. Readers are referred to [10] for a proof.

Proposition 3.1 *Assume* (3.2), (3.4)–(3.7), (3.9) *and* (3.10). *The sequence of diffusion scaled workload processes*

$$\hat{W}^r(\cdot) \Rightarrow W^*(\cdot) \quad \text{as } r \to \infty,$$

where $W^*(\cdot)$ *is a reflected Brownian motion with drift* $-\theta$, *variance* $\beta(c_a^2 + c_s^2)$ *and initial value* $w^* = \langle \chi, \xi^* + \mu^* \rangle$.

For the heavy traffic limit of the measure-valued process, we need the following definition of the lifting map $\Delta_{K,\nu} : \mathbb{R}_+ \to \mathbf{M} \times \mathbf{M}$.

Definition 3.1 Denote $\beta_e = \langle \chi, \nu_e \rangle$, where ν_e is the *equilibrium* measure of ν, i.e. $\nu_e([0, x]) = \frac{1}{\beta} \int_0^x \nu((y, \infty))\, dy$ for all $x \geq 0$. Let $\Delta_{K,\nu} : \mathbb{R}_+ \to \mathbf{M} \times \mathbf{M}$ be the lifting map (associated with the probability measure ν and constant K) given by

$$\Delta_{K,\nu} w = \left(\frac{(w - K\beta_e)^+}{\beta} \nu, \frac{w \wedge K\beta_e}{\beta_e} \nu_e \right) \quad \text{for } w \in \mathbb{R}_+.$$

Theorem 3.1 *Assume* (3.2)–(3.11). *If the limit* (ξ^*, μ^*) *in* (3.9) *satisfies*

$$\big(\xi^*, \mu^*\big) = \Delta_{K,\nu} w^*, \tag{3.12}$$

then the sequence of diffusion scaled state descriptors

$$\big(\hat{\mathcal{Q}}^r(\cdot), \hat{\mathcal{Z}}^r(\cdot)\big) \Rightarrow \Delta_{K,\nu} W^*(\cdot) \quad \text{as } r \to \infty,$$

where $W^*(\cdot)$ *is the reflected Brownian motion in Proposition* 3.1.

4 Validity of heavy traffic steady state approximations

As we have seen in Theorem 3.1, the heavy traffic limiting process $(\mathcal{Q}^*(\cdot), \mathcal{Z}^*(\cdot))$ is the image of the workload process $W^*(\cdot)$ under the continuous mapping $\Delta_{K,\nu}$. The limit is in the sense of weak convergence of probability measures, so the limiting process may not be in the same probability space where each process with index r is defined. Denote $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ the probability space where the weak limit is defined. It is well known that the marginal distribution $W^*(t)$ of the reflected Brownian motion $W^*(\cdot)$ converges weakly to that of the steady state random variable $W^*(\infty)$, which has the stationary distribution

$$\mathbb{P}^*\big(W^*(\infty) > x\big) = \exp\left(\frac{-2\theta x}{\beta(c_a^2 + c_s^2)}\right). \tag{4.1}$$

By the continuous mapping theorem, the measure-valued process $(\mathcal{Q}^*(\cdot), \mathcal{Z}^*(\cdot))$ converges weakly to $\Delta_{K,\nu} W^*(\infty)$. Denote the distribution of $\Delta_{K,\nu} W^*(\infty)$ by π^*. For any open set $B \in \mathbf{M} \times \mathbf{M}$,

$$\pi^*(B) = \mathbb{P}^*\big(\Delta_{K,\nu} W^*(\infty) \in B\big) = \mathbb{P}^*\big(W^*(\infty) \in \Delta_{K,\nu}^{-1} B\big). \tag{4.2}$$

On the other hand, for each r, since the traffic intensity $\rho^r < 1$ and the service time distribution is non-lattice, the diffusion scaled process $(\hat{\mathcal{Q}}^r(\cdot), \hat{\mathcal{Z}}^r(\cdot))$ is a regenerative process. By Proposition 2.1, $(\hat{\mathcal{Q}}^r(t), \hat{\mathcal{Z}}^r(t))$ converges to the steady state $(\hat{\mathcal{Q}}^r(\infty), \hat{\mathcal{Z}}^r(\infty))$ which has distribution $\hat{\pi}^r$ as $t \to \infty$.

Now, the question is: Does the stationary distribution $\hat{\pi}^r$ converge to π^*, which is obtained by first taking heavy traffic limit and then steady state limit? We have the following theorem, which validates the interchange of steady state limit and heavy traffic limit.

Theorem 4.1 *Assume* (3.2)–(3.12). *The sequence* $\{\hat{\pi}^r\}$ *converges weakly to* π^*.

The major steps of proving the above theorem are, first, obtaining inequality (4.6) via coupling, and second, establishing the uniform convergence on the right-hand side of (4.6) (i.e. the uniform bound of the coupling time) in Lemma 4.1. The proof of Theorem 4.1 will be presented at the end of this section.

Following the discussion in Sect. 2, we can construct a stationary version of the regenerative process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ such that at any time $t \geq 0$, it has distribution $\hat{\pi}^r$, i.e.

$$\mathbb{P}\Big(\big(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(t), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(t)\big) \in B\Big) = \hat{\pi}^r(B), \tag{4.3}$$

for any open set $B \in \mathbf{M} \times \mathbf{M}$. Let $\hat{W}^r_{\hat{\pi}^r}(\cdot) = \langle \chi, \hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot) + \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot) \rangle$ denote the corresponding workload process.

Let us now couple the stationary process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ with the corresponding process $(\hat{\mathcal{Q}}^r_0(\cdot), \hat{\mathcal{Z}}^r_0(\cdot))$ which starts with a zero initial condition. In other words, both $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ and $(\hat{\mathcal{Q}}^r_0(\cdot), \hat{\mathcal{Z}}^r_0(\cdot))$ are driven by the same stochastic primitives $(\hat{E}^r(\cdot), \{v^r_i\}_{i\geq 1})$. The only difference is the initial condition. Note that the stationarity assumption forces the renewal arrival process to be a stationary delayed renewal process. Define

$$\hat{t}^r_c = \inf\bigl\{t \geq 0 : \bigl(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(t), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(t)\bigr) = (\mathbf{0}, \mathbf{0})\bigr\}. \tag{4.4}$$

Note that the workload of $(\hat{\mathcal{Q}}^r_0(\cdot), \hat{\mathcal{Z}}^r_0(\cdot))$ starts at 0, which is less than or equal to $\hat{W}^r_{\hat{\pi}^r}(0)$. Since the LPS policy is work conserving, and both processes have the same stochastic primitives, for any $t \geq 0$

$$\bigl(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(t), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(t)\bigr) = (\mathbf{0}, \mathbf{0}) \quad \text{implies} \quad \bigl(\hat{\mathcal{Q}}^r_0(t), \hat{\mathcal{Z}}^r_0(t)\bigr) = (\mathbf{0}, \mathbf{0}). \tag{4.5}$$

Since both systems are driven by the same arrival process, $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(t), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(t))$ and $(\hat{\mathcal{Q}}^r_0(t), \hat{\mathcal{Z}}^r_0(t))$ are identical for all $t \geq t_c$. It then follows from Corollary 2.2 in Chap. VII of [1] that

$$\bigl|\mathbb{P}\bigl(\bigl(\hat{\mathcal{Q}}^r_0(t), \hat{\mathcal{Z}}^r_0(t)\bigr) \in B\bigr) - \hat{\pi}^r(B)\bigr| \leq \mathbb{P}\bigl(\hat{t}^r_c > t\bigr) \quad \text{for all } t \geq 0. \tag{4.6}$$

We now show that the probability $\mathbb{P}(\hat{t}^r_c > t)$ converges to 0 as $t \to \infty$ uniformly in r.

Lemma 4.1 *If* (3.2)–(3.8) *hold, then*

$$\sup_r \mathbb{P}\bigl(\hat{t}^r_c > t\bigr) \to 0 \quad \text{as } t \to \infty. \tag{4.7}$$

Proof Let $\hat{C}^r(t) = \frac{1}{r}\sum_{i=1}^{E^r(r^2 t)} v^r_i - rt$ for all $t \geq 0$. The summation in the above denotes the total amount of arrived work (under diffusion scaling) by time t, the second term $-rt$ denotes the amount of work the server has finished by time t without idling. So the first time the process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ reaches zero is the first time that $\hat{W}^r_{\hat{\pi}^r}(0) + \hat{C}^r(t) = 0$. By the definition of $\hat{t}^r_c$ in (4.4),

$$\hat{t}^r_c = \inf\bigl\{t \geq 0 : \hat{C}^r(t) = -\hat{W}^r_{\hat{\pi}^r}(0)\bigr\}.$$

So for any $M > 0$,

$$\begin{aligned}
\mathbb{P}\bigl(\hat{t}^r_c > t\bigr) &= \mathbb{P}\bigl(\hat{C}^r(s) > -\hat{W}^r_{\hat{\pi}^r}(0), \text{ for all } s \leq t\bigr) \\
&\leq \mathbb{P}\bigl(\hat{C}^r(t) > -\hat{W}^r_{\hat{\pi}^r}(0)\bigr) \\
&\leq \mathbb{P}\bigl(\hat{C}^r(t) > -\hat{W}^r_{\hat{\pi}^r}(0), -\hat{W}^r_{\hat{\pi}^r}(0) \geq -M\bigr) + \mathbb{P}\bigl(-\hat{W}^r_{\hat{\pi}^r}(0) < -M\bigr) \\
&\leq \mathbb{P}\bigl(\hat{C}^r(t) > -M\bigr) + \mathbb{P}\bigl(\hat{W}^r_{\hat{\pi}^r}(0) > M\bigr).
\end{aligned}$$

Since the regenerative process $(\hat{\mathcal{Q}}^r_{\hat{\pi}^r}(\cdot), \hat{\mathcal{Z}}^r_{\hat{\pi}^r}(\cdot))$ is stationary, the corresponding workload process $\hat{W}^r_{\hat{\pi}^r}(\cdot)$ is also stationary. By Corollary 7.5 in Chap. X of [1], the stationary distribution of the workload converges weakly to an exponential distribution as $r \to \infty$. This implies that for any $\epsilon > 0$, there exists $M' > 0$ such that

$$\sup_r \mathbb{P}\big(\hat{W}^r_{\hat{\pi}^r}(0) > M\big) < \epsilon \quad \text{for all } M \geq M'. \tag{4.8}$$

From now on, we fix a constant $M \in [M', \infty)$. Note that the process $\hat{C}^r(\cdot)$ converges weakly to a Brownian motion $B(\cdot)$ with drift $-\theta$ and variance $\beta(c_a^2 + c_s^2)$. This implies that

$$\lim_r \mathbb{P}\big(\hat{C}^r(t) > -M\big) = \mathbb{P}\big(B(t) > -M\big),$$

which goes to zero as $t \to \infty$. So for any constant M, there exists $t(M)$ such that

$$\limsup_r \mathbb{P}\big(\hat{C}^r(t) > -M\big) < \epsilon/2, \quad \text{for all } t \geq t(M).$$

This means that there exists $r_0 > 0$ such that for all $r \geq r_0$,

$$\mathbb{P}\big(\hat{C}^r(t) > -M\big) < \epsilon, \quad \text{for all } t \geq t(M).$$

For each $r < r_0$, we can choose $t_r(M)$ large enough (depending on r) such that

$$\mathbb{P}\big(\hat{C}^r(t) > -M\big) < \epsilon, \quad \text{for all } t \geq t_r(M).$$

Since there are only finitely many of those r's that are less than r_0, let $t_0(M) = \max_{r<r_0} t_r(M)$. We now have that

$$\sup_r \mathbb{P}\big(\hat{C}^r(t) > -M\big) < \epsilon, \quad \text{for all } t \geq \max\big(t(M), t_0(M)\big). \tag{4.9}$$

The lemma follows immediately from (4.8) and (4.9). □

Proof of Theorem 4.1 For any closed set $B \in \mathbf{M} \times \mathbf{M}$, we have

$$\begin{aligned}\hat{\pi}^r(B) - \pi^*(B) \leq{}& \big|\hat{\pi}^r(B) - \mathbb{P}\big(\big(\hat{\mathcal{Q}}^r_0(t), \hat{\mathcal{Z}}^r_0(t)\big) \in B\big)\big| \\ &+ \mathbb{P}\big(\big(\hat{\mathcal{Q}}^r_0(t), \hat{\mathcal{Z}}^r_0(t)\big) \in B\big) - \mathbb{P}^*\big(\big(\mathcal{Q}^*(t), \mathcal{Z}^*(t)\big) \in B\big) \\ &+ \mathbb{P}^*\big(\big(\mathcal{Q}^*(t), \mathcal{Z}^*(t)\big) \in B\big) - \pi^*(B). \end{aligned} \tag{4.10}$$

By the coupling inequality, the first term on the right-hand side of (4.10) is bounded by

$$\sup_r \mathbb{P}\big(\hat{t}^r_c > t\big),$$

which vanishes as $t \to \infty$, as proved in Lemma 4.1. According to the definition of π^* and Portmanteau Theorem (cf. Theorem 2.1 in [3]), the lim sup of the third term

on the right-hand side of (4.10) equals 0 as $t \to \infty$. So for any $\epsilon > 0$, there exists a $t_1 > 0$ (may be very large, but still finite) such that

$$\sup_r \mathbb{P}\big(\hat{t}_c^r > t_1\big) < \epsilon,$$

$$\mathbb{P}^*\big(\big(\mathcal{Q}^*(t_1), \mathcal{Z}^*(t_1)\big) \in B\big) - \pi^*(B) < \epsilon.$$

For this fixed t_1, by Theorem 3.1 and Portmanteau Theorem, we have that

$$\limsup_r \mathbb{P}\big(\big(\hat{\mathcal{Q}}_0^r(t_1), \hat{\mathcal{Z}}_0^r(t_1)\big) \in B\big) \le \mathbb{P}^*\big(\big(\mathcal{Q}^*(t_1), \mathcal{Z}^*(t_1)\big) \in B\big).$$

So there exists r_0 such that when $r \ge r_0$,

$$\mathbb{P}\big(\big(\hat{\mathcal{Q}}_0^r(t_1), \hat{\mathcal{Z}}_0^r(t_1)\big) \in B\big) - \mathbb{P}^*\big(\big(\mathcal{Q}^*(t_1), \mathcal{Z}^*(t_1)\big) \in B\big) < \epsilon.$$

So we have that for any $\epsilon > 0$, there exists r_0 such that when $r \ge r_0$,

$$\hat{\pi}^r(B) - \pi^*(B) < 3\epsilon.$$

This implies that $\limsup_r \hat{\pi}^r(B) \le \pi^*(B)$ for any closed set B. The result of the theorem follows from Portmanteau Theorem. □

5 Performance evaluation

So far, we have obtained results for the measure-valued description of the LPS queue. We now establish some more concrete results on the queue size, delay probability and response time.

5.1 Queue length and delay probability

The following result on the diffusion limit for the queue size process follows immediately from Theorem 3.1. The proof can be found in [25].

Corollary 5.1 (Piecewise reflected Brownian motion) *Assume* (3.2)–(3.10). *The sequence of diffusion scaled total job size processes* $\hat{X}^r(\cdot) = \langle 1, \hat{\mathcal{Q}}^r(\cdot) + \hat{\mathcal{Z}}^r(\cdot)\rangle$ *converges in distribution as* $r \to \infty$ *to* $X^*(\cdot)$, *where*

$$X^*(t) = \frac{(W^*(t) - K\beta_e)^+}{\beta} + \frac{W^*(t) \wedge K\beta_e}{\beta_e} \quad \text{for } t \ge 0,$$

and $W^*(\cdot)$ *is the reflected Brownian motion as in Proposition* 3.1.

In other words, $X^*(\cdot)$ is a reflected Brownian motion with drift $\frac{-\theta}{\beta}$ and variance $\frac{c_a^2+c_s^2}{\beta}$ when it is above K, and with drift $\frac{-\theta}{\beta_e}$ and variance $\frac{\beta(c_a^2+c_s^2)}{\beta_e^2}$ when it is below K.

Define the map $f:\mathbb{R}_+\to\mathbb{R}_+$ by $f(w)=\frac{1}{\beta}(w-K\beta_e)^+ + \frac{1}{\beta_e}(w\wedge K\beta_e)$ for all $w\in\mathbb{R}_+$. It is clear that $f(\cdot)$ is a continuous mapping with inverse

$$f^{-1}(x)=\begin{cases}\beta_e x & x\le K,\\ \beta_e K+\beta(x-K) & x>K.\end{cases}$$

By the continuous mapping theorem, $X^*(t)$ converges weakly to the steady state $X^*(\infty)=f(W^*(\infty))$ as $t\to\infty$, and $\mathbb{P}(X^*(\infty)>x)=\mathbb{P}(W^*(\infty)>f^{-1}(x))$. By the definition of the equilibrium distribution, it is easy to see that $\frac{\beta_e}{\beta}=\frac{1+c_s^2}{2}$. Since the stationary distribution of the reflected Brownian motion $W^*(\cdot)$ is explicitly known as in (4.1), it is easy to compute the distribution of $X^*(\infty)$,

$$\mathbb{P}\big(X^*(\infty)>x\big)=\begin{cases}\exp(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}x) & x\le K,\\ \exp(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}K-\frac{2\theta}{c_a^2+c_s^2}(x-K)) & x>K.\end{cases}\tag{5.1}$$

Let

$$d_p^*(\infty)=\mathbb{P}\big(X^*(\infty)>K\big)=\exp\left(-\frac{1+c_s^2}{c_a^2+c_s^2}\theta K\right)\tag{5.2}$$

be the steady state probability that the limiting queue size $X^*(\cdot)$ is above the sharing level K. As we see, the steady state limit of the heavy traffic limit is so tractable that the stationary distribution can be explicitly written down. Since we have established the interchange of steady state limit and heavy traffic limit, the following result is a direct implication of Theorem 4.1 and the continuous mapping theorem.

Corollary 5.2 *If* (3.2)–(3.12) *hold, then*

$$\hat{X}^r(\infty)\Rightarrow X^*(\infty),$$
$$\hat{d}_p^r(\infty)\to d_p^*(\infty),$$

as $r\to\infty$, *where* $\hat{d}_p^r(\infty)=\mathbb{P}(\hat{X}^r(\infty)>K^r/r)$ *is the steady state delay probability for the* r*th system.*

5.2 Response time

Let $R(t,v)$ denote the total time (including both waiting and service times) a job will stay in the system if it arrives at time t and with job size v. Since at a time t, there may not be an arrival or the arrival may not have job size v, the quantity $R(t,v)$ is often referred as the *virtual* response time. It contains two parts,

$$R(t,v)=R_B(t)+R_Z(t,v),\tag{5.3}$$

where $R_B(t)$ is the time that this virtual job spends on waiting in buffer (which does not depend on its job size) and $R_Z(t,v)$ is the service time of this virtual job. Let $W_B(\cdot)=\langle\chi,\mathcal{Q}(\cdot)\rangle$ and $W_Z(\cdot)=\langle\chi,\mathcal{Z}(\cdot)\rangle$ denote the workload in buffer and the

workload in server, respectively. From the time t in which this virtual job enters the system until it is about to enter service at $t + R_B(t)$, the server never idles. So the workload the server processes during this time period is equal to $R_B(t)$. Since the LPS policy is work conserving, we must have that

$$W_B(t) + W_Z(t) = R_B(t) + W_Z\big(t + R_B(t)\big). \tag{5.4}$$

It is clear that the service time of this virtual job should satisfy

$$S\big(t + R_B(t), t + R_B(t) + R_Z(t, v)\big) = v. \tag{5.5}$$

We now study the heavy traffic limit of the diffusion scaled virtual response time $\hat{R}^r(t, v) = \frac{1}{r} R^r(r^2 t, v)$.

Proposition 5.1 (Heavy traffic limit for virtual response time process) *Assume* (3.2)–(3.12). *For any fixed* $v \geq 0$, *the diffusion scaled virtual waiting time* $(\hat{R}^r_B(\cdot), \hat{R}^r_Z(\cdot, v))$ *converges weakly to* $(R^*_B(\cdot), R^*_Z(\cdot, v))$, *where*

$$R^*_B(t) = \beta\big(X^*(t) - K\big)^+, \quad R^*_Z(t, v) = v\big(X^*(t) \wedge K\big), \quad t \geq 0. \tag{5.6}$$

Proof Since $\hat{W}^r_B(\cdot) = \langle \chi, \hat{\mathcal{Q}}^r(\cdot) \rangle$ and $\hat{W}^r_Z(\cdot) = \langle \chi, \hat{\mathcal{Z}}^r(\cdot) \rangle$, by Theorem 3.1 and the continuous mapping theorem,

$$\big(\hat{W}^r_B(\cdot), \hat{W}^r_Z(\cdot)\big) \Rightarrow \big(\big(W^*(\cdot) - K\beta_e\big)^+, \big(W^*(\cdot) \wedge K\beta_e\big)\big), \tag{5.7}$$

as $r \to \infty$. The diffusion scaled version of (5.4) can be written as

$$\hat{W}^r_B(t) + \hat{W}^r_Z(t) = \hat{R}^r_B(t) + \hat{W}^r_Z\left(t + \frac{1}{r}\hat{R}^r_B(t)\right). \tag{5.8}$$

It is clear that $\hat{R}^r_B(t) \leq \hat{W}^r(t)$, which converges to a reflected Brownian motion as $r \to \infty$. So on any finite interval $[0, T]$, for any $\epsilon > 0$ there exists $M > 0$ such that

$$\limsup_{r\to\infty} \mathbb{P}\Big(\sup_{t\in[0,T]} \hat{R}^r_B(t) > M\Big) < \epsilon.$$

Since $\hat{W}^r_Z(\cdot)$ converges to $W^*(\cdot) \wedge K\beta_e$, which is almost surely continuous, we have that

$$\sup_{t\in[0,T]} \left|\hat{W}^r_Z\left(t + \frac{1}{r}M\right) - \hat{W}^r_Z(t)\right| \Rightarrow 0 \quad \text{as } r \to \infty.$$

So for any $\epsilon > 0$,

$$\limsup_{r\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \left|\hat{W}^r_Z\left(t + \frac{1}{r}\hat{R}^r_B(t)\right) - \hat{W}^r_Z(t)\right| > \epsilon\right) < \epsilon.$$

It then follows from (5.8) that

$$\sup_{t\in[0,T]} \big|\hat{R}^r_B(t) - \hat{W}^r_B(t)\big| \Rightarrow 0 \quad \text{as } r \to \infty. \tag{5.9}$$

The diffusion scaled version of (5.5) can be written as

$$S^r\big(r^2t + r\hat{R}^r_B(t), r^2t + r\hat{R}^r_B(t) + r\hat{R}^r_Z(t,v)\big) = v. \tag{5.10}$$

Due to the sharing level K^r, the processing time of a job with size v has bound

$$\hat{R}^r_Z(t,v) \le \frac{K^r}{r}v,$$

which is less than $Kv+1$ for all large enough r. By Theorem 3.1, the server size $\hat{Z}^r(\cdot)$ converges weakly to $(X^*(\cdot) \wedge K\beta_e)$ as $r \to \infty$. Again, the limiting process is almost surely continuous. So

$$\sup_{t\in[0,T]} \sup_{s\le(Kv+1)/r} \left|\hat{Z}^r\left(t + \frac{1}{r}\hat{R}^r_B(t) + s\right) - \hat{Z}^r(t)\right| \Rightarrow 0 \quad \text{as } r \to \infty.$$

In other words, the server size will not oscillate much during the whole service time. Thus

$$\limsup_{r\to\infty} \mathbb{P}\Big(\sup_{t\in[0,T]} \sup_{x\le Kv+1} \big|S^r\big(r^2t + r\hat{R}^r_B(t), r^2t + r\hat{R}^r_B(t) + rx\big)\hat{Z}^r(t) - x\big| > \epsilon\Big) < \epsilon.$$

It then follows from (5.10) that, for any $v \ge 0$,

$$\sup_{t\in[0,T]} \big|\hat{R}^r_Z(t,v) - \hat{Z}^r(t)v\big| \Rightarrow 0 \quad \text{as } r \to \infty. \tag{5.11}$$

By Corollary 5.1, as $r \to \infty$,

$$\hat{Z}^r(\cdot) \Rightarrow \big(X^*(\cdot) \wedge K\big),$$

where $X^*(\cdot) = \frac{(W^*(\cdot)-K\beta_e)^+}{\beta} + \frac{W^*(\cdot)\wedge K\beta_e}{\beta_e}$. In fact, this convergence is also a direct application of Theorem 3.1 and the continuous mapping theorem. So the convergence of $\hat{Z}^r(\cdot)$ holds jointly with the convergence in (5.7). In particular,

$$\big(\hat{W}^r_B(\cdot), \hat{Z}^r(\cdot)\big) \Rightarrow \big(\beta\big(X^*(\cdot) - K\big)^+, \big(X^*(\cdot) \wedge K\big)\big) \quad \text{as } r \to \infty. \tag{5.12}$$

So the joint convergence of $\hat{R}^r_B(\cdot)$ and $\hat{R}^r_Z(\cdot, v)$ follows immediately from (5.9), (5.11) and the above convergence. □

From this proposition, we see that the limiting response times are piecewise linear and continuous functions of the limiting queue size process. It follows from (5.1) and the continuous mapping theorem that the steady state distributions of the response times are

$$\mathbb{P}\big(R^*_B(\infty) > x\big) = \exp\left(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}K - \frac{2\theta}{c_a^2+c_s^2}\frac{x}{\beta}\right), \quad x \ge 0, \tag{5.13}$$

$$\mathbb{P}\big(R^*_Z(\infty, v) > x\big) = \exp\left(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}\left(\frac{x}{v} \wedge K\right)\right), \quad x \ge 0, \tag{5.14}$$

$$\mathbb{P}\big(R^*(\infty, v) > x\big) = \begin{cases} \exp(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}\frac{x}{v}), & 0 \le x \le Kv, \\ \exp(-\frac{(1+c_s^2)\theta}{c_a^2+c_s^2}K - \frac{2\theta}{c_a^2+c_s^2}\frac{x-Kv}{\beta}), & x \ge Kv. \end{cases} \tag{5.15}$$

Similarly as in Sect. 5.1, we can obtain the result below as a corollary of Theorem 4.1. The difference is that the linear and continuous relationship (5.6) only holds for the heavy traffic limit, not for each rth system, so we cannot apply the continuous mapping theorem. However, the coupling inequality (4.6) holds for the response times as well as the measure-valued process. (The reason is that if two queues are the same, then the virtual response times will also be the same.) So the following result can be proved following the same approach as in the proof of Theorem 4.1. We omit the proof for brevity.

Corollary 5.3 *Assume* (3.2)–(3.12). *For any* $v \ge 0$,

$$\big(\hat{R}_B^r(\infty), \hat{R}_Z^r(\infty, v)\big) \Rightarrow \big(R_B^*(\infty), R_Z^*(\infty, v)\big),$$

as $r \to \infty$.

6 Approximations

In this section, we apply our limit theorems to obtain approximations for the steady state queue length and response time.

6.1 Queue size

Since we have validated the heavy traffic steady state approximation, we can use the steady state random variable $X^*(\infty)$ to approximate the steady state of the diffusion scaled rth system $\hat{X}^r(\infty)$. It follows from (5.1) that

$$\mathbb{E}\big(X^*(\infty)\big) = \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{1}{\theta}\big(1 - d_p^*(\infty)\big) + \frac{c_a^2 + c_s^2}{2}\frac{1}{\theta}d_p^*(\infty),$$

where $d_p^*(\infty)$ is given in (5.2). According to the heavy traffic conditions (3.2) and (3.3), $\frac{1}{r}$ can be approximately written as $\frac{1-\rho^r}{\rho^r\theta}$ and θK can be approximately written as $\frac{1-\rho^r}{\rho^r}K^r$. So we obtain the following approximation for $\mathbb{E}(X^r(\infty))$:

$$\mathbb{E}\big(X^r(\infty)\big) \approx \frac{c_a^2 + c_s^2}{1 + c_s^2}\frac{\rho^r}{1-\rho^r}\big(1 - d_p^r(\infty)\big) + \frac{c_a^2 + c_s^2}{2}\frac{\rho^r}{1-\rho^r}d_p^r(\infty),$$

where the delay probability $d_p^r(\infty)$ could be taken as $\exp(-\frac{1+c_s^2}{c_a^2+c_s^2}\frac{1-\rho^r}{\rho^r}K^r)$. Since $\frac{1-\rho^r}{\rho^r} \sim -\ln\rho^r$, we prefer to use the asymptotically equivalent description $d_p^r(\infty) = (\rho^r)^{\frac{1+c_s^2}{c_a^2+c_s^2}K^r}$.

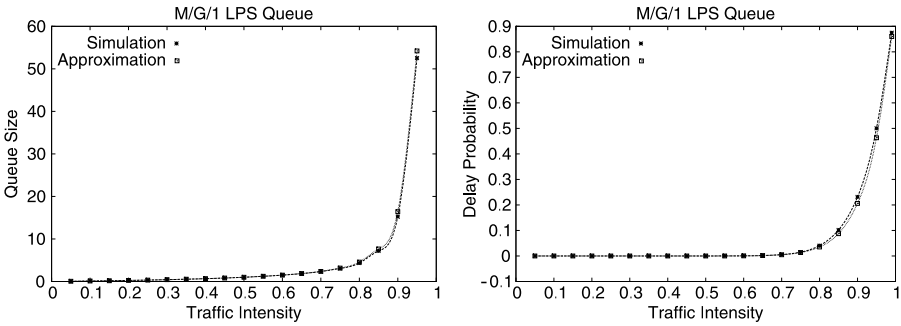

Fig. 2 A comparison of the approximation formulas with simulation estimates of steady state response times of the $M/G/1$ LPS queue. The sharing level $K = 15$, service time distribution is log-normal with $c_s^2 = 9$ and traffic intensities range from 0.05 to 0.95

In practice, only one system with certain sharing level K and traffic intensity $\rho < 1$ will be given. So we can drop the index r and obtain the following approximation formula:

$$\mathbb{E}[X] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{\rho}{1-\rho}(1 - d_p) + \frac{c_a^2 + c_s^2}{2} \frac{\rho}{1-\rho} d_p, \tag{6.1}$$

$$d_p \approx \rho^{\frac{1+c_s^2}{c_a^2+c_s^2}K}. \tag{6.2}$$

The resulting approximation (6.1) reduces to Kingman's formula for the FCFS queue when the sharing level $K = 1$, and the formula in [9] for the PS queue when the sharing level $K = \infty$. Although the approximation formulas are derived from heavy traffic theorems, they are actually explicit if the arrival process is Poisson and either $K = 1$ or $K = \infty$. In addition, the quality of the approximations is actually reasonable for all traffic intensities, cf. Fig. 2.

The approximation formulas are derived in the context of the $G/G/1$ LPS queue, and use the first two moments of inter-arrival time and service time distributions. Table 1 demonstrates the quality of our approximations for various combinations of inter-arrival time and service time distributions. As suggested by the formulas, no matter how we change the combination of distributions, the approximation will be the same as long as the coefficients of variation c_a^2 and c_s^2 (for inter-arrival and service times, respectively) are fixed. This is also reflected by the numerical results in Table 1.

6.2 Response time

As for the response time, we can use the steady state $R_B^*(\infty)$ and $R_Z^*(\infty, v)$ to approximate the steady state of the diffusion scaled response time of the rth system, i.e. $\hat{R}_B^r(\infty)$ and $\hat{R}_Z^r(\infty, v)$. By (5.13) and (5.14), the following expectations can be easily computed:

$$\mathbb{E}\big[R_B^*(\infty)\big] = \frac{c_a^2 + c_s^2}{2} \frac{1}{\theta} d_p^*(\infty)\beta,$$

Table 1 $G/G/1$ LPS queue. The sharing limit $K = 20$, traffic intensity $\rho = 0.9$. The squared coefficient of variation of inter-arrival time and service time distribution is fixed at $c_a^2 = 4$ and $c_s^2 = 8$ respectively. Here and later on, we choose to look at the 95% confidence interval. $a \pm \delta$ means the confidence interval is $(a - \delta, a + \delta)$

Arr. dist.	Serv. dist.	$\mathbb{E}[X]$	d_p
HyperExp2p	HyperExp2p	20.5378 ± 0.3148	0.2519 ± 0.0022
	Log-normal	20.6243 ± 0.2753	0.2798 ± 0.0019
	Hyper2star	20.5642 ± 0.1619	0.2066 ± 0.0014
Log-normal	HyperExp2p	19.6028 ± 0.2957	0.2334 ± 0.0020
	Log-normal	19.3615 ± 0.1894	0.2562 ± 0.0017
	Hyper2star	19.5913 ± 0.2435	0.1981 ± 0.0018
Hyper2star	HyperExp2p	20.9429 ± 0.3561	0.2676 ± 0.0027
	Log-normal	21.1600 ± 0.2136	0.2972 ± 0.0015
	Hyper2star	20.8725 ± 0.2505	0.2089 ± 0.0018
Approximation formulas		20.6474	0.2059

HyperExp2p is the hyper-exponential distribution with 2 phases. A Hyper2star random variable has probability p to be 0 and probability $(1 - p)$ to be an exponential distribution

$$\mathbb{E}\big[R_Z^*(\infty, v)\big] = \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{1}{\theta} \big(1 - d_p^*(\infty)\big) v.$$

Now, we approximate $\frac{1}{r}$ using $\frac{1-\rho^r}{\theta}$. This way of approximating $\frac{1}{r}$ is equivalent in the limit to using $\frac{1-\rho^r}{\rho^r \theta}$ based on the heavy traffic condition (3.2). The main reason for the difference is to make the approximations of waiting time and buffer queue size consistent with Little's law. So we obtain the following approximation formulas for a given system:

$$\mathbb{E}[R_B] \approx \frac{c_a^2 + c_s^2}{2} \frac{1}{1 - \rho} d_p \beta, \tag{6.3}$$

$$\mathbb{E}\big[R_Z(v)\big] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{1}{1 - \rho} (1 - d_p) v, \tag{6.4}$$

where d_p is the same as in (6.2).

Table 2 shows the quality of approximations (6.3) and (6.4) for the $M/G/1$ LPS queue with various service time distributions. Again, our approximations use up to the second moment of the service time distribution, so the simulation gives the similar performance for different distributions with the same squared coefficient of variation c_s^2.

Let R_Z be the unconditional steady state service time, then

$$\mathbb{E}[R_Z] \approx \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{1}{1 - \rho} (1 - d_p) \beta.$$

Table 2 $M/G/1$ LPS queue. The sharing limit $K = 30$, traffic intensity $\rho = 0.95$. The squared coefficient of variation of service time is fixed at $c_s^2 = 19$

Distribution	$\mathbb{E}[R_B]$	$\mathbb{E}[R_Z(v)/v]$
HyperExp2p	42.1670 ± 2.0085	15.7579 ± 0.0871
Log-normal	37.2947 ± 1.6338	15.9390 ± 0.0942
Hyper2Star	41.0397 ± 1.6116	15.6039 ± 0.0851
Bimodal	41.8724 ± 1.3550	15.6162 ± 0.0958
Approximations	42.9278	15.7072

Table 3 $G/M/1$ and $M/G/1$ LPS queues. The sharing limit $K = 10$, traffic intensity $\rho = 0.9$

Perf. meas.	$M/G/1$		$G/M/1$	
	Simulation	Approx.	Simulation	Approx.
d_p	0.3437 ± 0.0025	0.3478	0.2436 ± 0.0026	0.2580
$E(X)$	8.2459 ± 0.0569	8.2155	6.8441 ± 0.0555	7.0000
$E(R_B)$	2.6591 ± 0.0466	2.6151	1.8629 ± 0.0426	2.0070
$E(R_Z)$	6.4958 ± 0.0146	6.5132	5.6779 ± 0.0171	5.7708

The service time distribution for $M/G/1$ is Erlang with 2 phases (E_2), mean is 1 and $c_s^2 = 1/2$. The inter-arrival time distribution for $G/M/1$ is E_2 with mean $1/0.9$ and $c_a^2 = 1/(2 \times 0.9)$

So we obtain an approximation of the unconditional response time

$$\mathbb{E}[R] \approx \frac{c_a^2 + c_s^2}{2} \frac{\beta}{1-\rho} d_p + \frac{c_a^2 + c_s^2}{1 + c_s^2} \frac{\beta}{1-\rho} (1 - d_p). \tag{6.5}$$

Finally, we show in Table 3 a comparison of all our performance approximations with simulations of the $M/G/1$ and the $G/M/1$ LPS queues. All the numerical results show that the two-moment approximations are reasonably fit, with the exception of log-normal service times (in Table 2). This is in accordance with other numerical studies on the quality of two-moment approximations, see for example [14].

References

1. Asmussen, S.: Applied Probability and Queues, 2nd edn. Applications of Mathematics (New York), vol. 51. Springer, New York (2003)
2. Avi-Itzhak, B., Halfin, S.: Expected response times in a non-symmetric time sharing queue with a limited number of service positions. In: Proceedings of the 12th International Teletraffic Congress. Torino (1988)
3. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley, New York (1999)
4. Blake, R.: Optimal control of thrashing. In: Proceedings of the 1982 ACM SIGMETRICS Conference on Measurements and Modeling of Computer Systems. Seattle, WA (1982)
5. Budhiraja, A., Lee, C.: Stationary distribution convergence for generalized Jackson networks in heavy traffic. Technical Report, University of North Carolina at Chapel Hill (2008). http://www.unc.edu/%7Echlee/jackson.pdf
6. Denning, P.J., Kahn, K.C., Leroudier, J., Potier, D., Suri, R.: Optimal multiprogramming. Acta Inform. **7**, 197–216 (1976)

7. Elnikety, S., Nahum, E., Tracy, J., Zwaenepoel, W.: A method for transparent admission control and request scheduling in e-commerce web sites. In: World-Wide-Web Conference (2004)
8. Gamarnik, D., Zeevi, A.: Validity of heavy traffic steady-state approximation in generalized Jackson networks. Ann. Appl. Probab. **16**(1), 56–90 (2006)
9. Grishechkin, S.: $GI/G/1$ processor sharing queue in heavy traffic. Adv. Appl. Probab. **26**(2), 539–555 (1994)
10. Gromoll, H.C.: Diffusion approximation for a processor sharing queue in heavy traffic. Ann. Appl. Probab. **14**(2), 555–611 (2004)
11. Gromoll, H.C., Kruk, Ł.: Heavy traffic limit for a processor sharing queue with soft deadlines. Ann. Appl. Probab. **17**(3), 1049–1101 (2007)
12. Gromoll, H.C., Puha, A.L., Williams, R.J.: The fluid limit of a heavily loaded processor sharing queue. Ann. Appl. Probab. **12**(3), 797–859 (2002)
13. Gromoll, H.C., Robert, P., Zwart, B.: Fluid limits for processor sharing queues with impatience. Math. Oper. Res. **33**(2), 375–402 (2008)
14. Gupta, V., Dai, J.G., Harchol-Balter, M., Zwart, B.: On the inapproximability of $M/G/K$: Why two moments of job size distribution are not enough. Technical report, Carnegie Mellon University (2007)
15. Heiss, H.-U., Wagner, R.: Adaptive load control in transaction processing systems. In: Proceedings of the 17th International Conference on Large Data Bases (1991)
16. Kamra, A., Misra, V., Nahum, E.M.: Yaksha: A self-tuning controller for managing the performance of 3-tiered web sites. In: Twelfth IEEE International Workshop on Quality of Service (2004)
17. Kleinrock, L.: Queueing Systems. Computer Applications, vol. II. Wiley, New York (1976)
18. Maulik, K., Zwart, B.: An extension of the square root law of TCP. Ann. Oper. Res. (2008, to appear)
19. Nuyens, M., van der Weij, W.: The limited processor sharing queue. Technical report, CWI, Amsterdam (2007)
20. Ritchie, D.M., Thompson, K.: The Unix time-sharing system. J. Assoc. Comput. Mach. **17**(7), 365–375 (1974)
21. Schroeder, B., Harchol-Balter, M., Iyengar, A., Nahum, E., Wierman, A.: How to determine a good multi-programming level for external scheduling. In: Proceedings of the 22nd International Conference on Data Engineering. Atlanta, GA (2006)
22. Sigman, K., Wolff, R.W.: A review of regenerative processes. SIAM Rev. **35**(2), 269–288 (1993)
23. Zhang, F., Lipsky, L.: Modelling restricted processor sharing. In: Proc. of the 2006 Int'l Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA06) (2006)
24. Zhang, F., Lipsky, L.: An analytical model for computer systems with non-exponential service times and memory thrashing overhead. In: Proc. of the 2007 Int'l Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA07) (2007)
25. Zhang, J., Dai, J.G., Zwart, B.: Diffusion limits of limited processor sharing queues. Technical report, Georgia Institute of Technology (2007). http://www.isye.gatech.edu/~jzhang/research/lps-ht.pdf
26. Zhang, J., Dai, J.G., Zwart, B.: Law of large number limits of limited processor sharing queues. Technical report, Georgia Institute of Technology (2007). http://www.isye.gatech.edu/~jzhang/research/fl-lps.pdf