DIFFUSION LIMITS FOR SHORTEST REMAINING PROCESSING TIME QUEUES

H. Christian Gromoll*, Łukasz Kruk, Amber L. Puha

University of Virginia, Maria Curie-Skłodowska University,

California State University San Marcos

May 7, 2010

Abstract

We present a heavy traffic analysis for a single server queue with renewal arrivals and generally distributed i.i.d. service times, in which the server employs the Shortest Remaining Processing Time (SRPT) policy. Under typical heavy traffic assumptions, we prove a diffusion limit theorem for a measure-valued state descriptor, from which we conclude a similar theorem for the queue length process. These results allow us to make some observations on the queue length optimality of SRPT. In particular, they provide the sharpest illustration of the well-known tension between queue length optimality and quality of service for this policy.

AMS 2010 subject classifications. Primary 60K25, 60F17; secondary 60G57, 68M20, 90B22.
Key words. Heavy traffic, queueing, shortest remaining processing time, diffusion limit.

1 Introduction

In a single server queue employing the Shortest Remaining Processing Time (SRPT) policy, preemptive priority is given to the job that can be completed first, that is, the job with the shortest remaining processing time. More

*Research supported in part by NSF grant DMS 0707111

precisely, consider a single server queue with renewal arrivals and i.i.d. service times, and let $\mathcal{I}(t)$ index those jobs that are in the queue at time t. For $i \in \mathcal{I}(t)$, let $w_i(t)$ denote the *residual service time* at time t of job i. This is the remaining amount of processing time required to complete this job. If $j \in \mathcal{I}(t)$ is the smallest index such that $w_j(t) \leq w_i(t)$ for all $i \in \mathcal{I}(t)$, then under SRPT, $\frac{d}{dt}w_j(t) = -1$ and $\frac{d}{dt}w_i(t) = 0$ for all $i \in \mathcal{I}(t) \setminus j$.

Interest in the SRPT policy goes back to the first optimality result of Schrage [15], who showed that SRPT minimizes the number of jobs in the system, or queue length, at each point in time (see also Smith [18]). More explicitly, given fixed arrival and service processes, if $Z(t)$ is the queue length at time t under SRPT and $Q(t)$ is the queue length at t under an arbitrary work conserving policy, then almost surely,

$$Z(t) \leq Q(t), \qquad \text{for all } t \geq 0. \tag{1.1}$$

This holds with no distributional assumptions on the underlying arrival and service processes.

Expressions for the mean response time for an M/G/1 SRPT queue were developed earlier by Schrage and Miller [16], and extended later in Schassberger [14] and Perera [12] (see Schreiber [17] for a survey of the same time period). Another notable contribution was made by Pavlov [10] and Pechinkin [11], who characterized the heavy traffic limit of the steady state distributions for the queue length of an M/G/1 SRPT queue.

Recently, there has been renewed interest in the SRPT policy, mainly in computer science. For example, Bansal and Harchol-Balter [1] study fairness for SRPT ([1] is also a good source for a more extended list of prior work on SRPT). More recent work seeks to provide a framework for comparing policies in the M/G/1 setting, see for example Wierman and Harchol-Balter [20].

There has also been a recent body of work on the tail behavior of single server queues under SRPT; see for example Núñez Queija [8] and Nuyens and Zwart [9]. They discuss the advisability of implementing SRPT using large deviations techniques.

In [5], Down and Wu employ diffusion limits to show certain optimality properties of a multi-layered round robin routing policy for a system of parallel servers, each operating under SRPT. This was done under the assumption of a finitely supported service time distribution, mainly due to the absence at the time of diffusion limits for more general service time distributions. In

the case of a general service time distribution, Down, Gromoll, and Puha [4] developed fluid limits for SRPT queues, and used these to obtain a formula for state-dependent response times (on fluid scale) of jobs entering the system (see also [3]).

In this paper, we prove a diffusion limit theorem that holds for a general service time distribution, under usual heavy traffic assumptions. We do this for a measure-valued state descriptor, so that diffusion limits for various other performance measures may be obtained as corollaries; see Theorem 3.1. In particular, we obtain a diffusion limit theorem for the queue length process. This result reveals just how optimal SRPT is, in the sense of (1.1), and is explained below.

Let $\widehat{Z}^r(t) = r^{-1}Z^r(r^2t)$, $t \geq 0$, be the rth diffusion scaled queue length process from an r-indexed sequence of SRPT models, as detailed in Section 3. In particular, we assume the fairly standard heavy traffic assumptions (3.4), (3.5), (3.6), (3.7), (3.8), and (3.10). We use $W^*(\cdot)$ to denote the limit in distribution of the corresponding sequence of diffusion scaled workload processes (see (3.9)). As noted there, $W^*(\cdot)$ is the same for all work conserving policies and is a reflected Brownian motion in $\mathbb{R}_+$ [7]. We use ν to denote the limiting service time distribution (see (3.5)) and x^* to denote the supremum of the support of ν. Informally, x^* is the largest possible job size. Then,

Theorem 1.1 *As $r \to \infty$, the processes $\widehat{Z}^r(\cdot)$ converge in distribution to*

$$Z^*(\cdot) \stackrel{d}{=} \begin{cases} \frac{W^*(\cdot)}{x^*}, & \text{if } x^* < \infty, \\ 0, & \text{if } x^* = \infty. \end{cases}$$

This result follows from Theorem 3.1 by the continuous mapping theorem.

Theorem 1.1 makes a striking statement about the queue length optimality of SRPT. Consider the following simple lower bound, valid for any work conserving policy and service time distribution ν. Assume for the moment that $x^* < \infty$. Let $Q(t)$, $t \geq 0$, be the queue length process under an arbitrary work conserving policy. Then at each time $t \geq 0$, the workload $W(t)$ is bounded above by $Q(t)x^*$, because it is the sum of $Q(t)$ residual service times, each of which is bounded above by x^*. So almost surely,

$$Q(t) \geq \frac{W(t)}{x^*}, \qquad \text{for all } t \geq 0. \tag{1.2}$$

Note that (1.2) makes sense when $x^* = \infty$ as well, as the right side is interpreted as zero.

Unlike (1.1), which gives a universal lower bound (over all work conserving policies) in terms of the queue length process of one such policy, (1.2) gives a universal lower bound in terms of the common workload process of all such policies. In particular, we may combine these bounds and have, almost surely,

$$\frac{W(t)}{x^*} \leq Z(t) \leq Q(t), \qquad \text{for all } t \geq 0.$$

The bound (1.2) is intuitively appealing because it results from the hypothetical configuration of residual service times that minimizes the queue length at time t, given the workload at t. At each $t \geq 0$, the queue length minimizing configuration is the one that puts as many residual service times as possible at x^*, such that they sum to $W(t)$. (To be precise, all of them if x^* divides $W(t)$ and all but one of them otherwise). Additionally, since the workload process is a much simpler object than the queue length process under SRPT, (1.2) may be easier to work with in practice, when $x^* < \infty$, than (1.1).

Of course, this bound is hypothetical because no work conserving policy, including SRPT, can achieve such optimal configurations for all $t \geq 0$, although many may achieve it for some t (including for example all times t for which $W(t) = 0$). The interesting fact contained in Theorem 1.1 is that, on diffusion scale in heavy traffic, SRPT actually achieves the hypothetical lower bound asymptotically, almost surely for all $t \geq 0$.

So SRPT is not only better than any other work conserving policy in the sense of (1.1), it is in fact as optimal as possible in the heavy traffic limit. Of course, this optimality is from the point of view of the server, who one imagines wants to minimize queue length. As is well known, SRPT performs poorly from the point of view of large jobs (see e.g. [4]), who wish to minimize their time in queue, but tend to wait for long periods as they are preempted by smaller jobs. Indeed the queue length optimality of SRPT comes at the expense of long sojourn times for large jobs, and this tension is made explicit by Thereom 3.1, which gives the measure-valued diffusion limit. From this result, we see that in the heavy traffic limit, all mass is concentrated at x^*. So asymptotically for all $t \geq 0$, the queue consists entirely of jobs of the largest possible size, whereas smaller jobs are flushed out instantly. That is, the diffusion limit in Theorem 3.1 puts the contrast between queue length optimality and poor performance for large jobs in the sharpest light.

In the remainder of the paper, we give a precise definition of the stochastic model for an SRPT queue (Section 2), state our assumptions and main result

(Section 3), and provide the proofs (Section 4).

1.1 Notation

The following notation will be used throughout the paper. Let $\mathbb{N}$ denote the set of positive integers and let $\mathbb{R}$ denote the set of real numbers. For $a, b \in \mathbb{R}$, we write $a \vee b$ for the maximum of a and b, and $\lfloor a \rfloor$ for the largest integer less than or equal to a. The nonnegative real numbers $[0, \infty)$ will be denoted by $\mathbb{R}_+$. By convention, a sum of the form $\sum_{i=n}^{m}$ with $n > m$, or a sum over an empty set of indices equals zero. The sets (a, b), $[a, b)$, and $(a, b]$ are empty for $a, b \in [0, \infty]$ with $a \geq b$. For a Borel set $B \subset \mathbb{R}_+$, we denote the indicator of the set B by 1_B. We also define the real valued function $\chi(x) = x$, for $x \in \mathbb{R}_+$.

Let $\mathbf{M}$ denote the set of finite, nonnegative Borel measures on $\mathbb{R}_+$. For $\xi \in \mathbf{M}$ and a Borel measurable function $g : \mathbb{R}_+ \to \mathbb{R}$ that is integrable with respect to ξ, define $\langle g, \xi \rangle = \int_{\mathbb{R}_+} g(x)\xi(dx)$. The set $\mathbf{M}$ is endowed with the weak topology. That is, for $\xi_n, \xi \in \mathbf{M}$, we have $\xi_n \xrightarrow{w} \xi$ if and only if $\langle g, \xi_n \rangle \to \langle g, \xi \rangle$ as $n \to \infty$, for all $g : \mathbb{R}_+ \to \mathbb{R}$ that are bounded and continuous. With this topology, $\mathbf{M}$ is a Polish space [13]. We denote the zero measure in $\mathbf{M}$ by $\mathbf{0}$ and the measure in $\mathbf{M}$ that puts one unit of mass at the point $x \in \mathbb{R}_+$ by δ_x. For $x \in \mathbb{R}_+$, the measure δ_x^+ is δ_x if $x > 0$ and $\mathbf{0}$ otherwise. For $\xi \in \mathbf{M}$, we say that $x \in \mathbb{R}_+$ is a ξ-continuity point if $\langle 1_{\{x\}}, \xi \rangle = 0$. Let $\mathbf{M}_a$ denote those elements of $\mathbf{M}$ that do not charge the origin. We say that a measure $\xi \in \mathbf{M}$ has a finite first moment if $\langle \chi, \xi \rangle < \infty$. Let $\mathbf{M}_\chi$ denote the set of all such measures and let $\mathbf{M}_0 = \mathbf{M}_\chi \cap \mathbf{M}_a$.

We use "$\stackrel{d}{=}$" for equality in distribution and "$\Rightarrow$" to denote convergence in distribution of random elements of a metric space. Unless otherwise specified, all stochastic processes used in this paper are assumed to have paths that are right continuous with finite left limits (r.c.l.l.). For a Polish space $\mathcal{S}$, we denote by $\mathbf{D}([0, \infty), \mathcal{S})$ the space of r.c.l.l. functions from $[0, \infty)$ into $\mathcal{S}$, endowed with the Skorohod J_1-topology [6].

2 Stochastic Model for an SRPT Queue

Our stochastic model of an SRPT queue consists of the following: a random initial condition $\mathcal{Z}(0) \in \mathbf{M}$ specifying the state of the system at time zero, stochastic primitives $E(\cdot)$ and $\{v_k\}_{k \in \mathbb{N}}$ describing the arrival of jobs to the

queue and their service times, and a measure valued state descriptor $\mathcal{Z}(\cdot)$ describing the time evolution of the system. These are defined below.

Initial condition. The initial condition specifies the number $Z(0)$ of jobs in the queue at time zero, as well as the initial service time of each job. Assume that $Z(0)$ is a nonnegative integer valued random variable that is finite almost surely. The initial service times are the first $Z(0)$ elements of a sequence $\{\tilde{v}_j\}_{j\in\mathbb{N}}$ of strictly positive, finite random variables. The initial job with service time $\tilde{v}_j$, $j \le Z(0)$, is called job j.

A convenient way to express the initial condition is to define an initial random measure $\mathcal{Z}(0) \in \mathbf{M}$ by

$$\mathcal{Z}(0) = \sum_{j=1}^{Z(0)} \delta_{\tilde{v}_j},$$

which equals $\mathbf{0}$ if $Z(0) = 0$. Our assumptions imply that $\mathcal{Z}(0)$ satisfies

$$\mathbf{P}(\langle 1, \mathcal{Z}(0)\rangle \vee \langle \chi, \mathcal{Z}(0)\rangle < \infty) = 1. \tag{2.1}$$

In particular, the number of initial jobs and the initial workload are finite almost surely, and so $\mathcal{Z}(0) \in \mathbf{M}_0$ almost surely.

Stochastic primitives. The stochastic primitives consist of an exogenous arrival process $E(\cdot)$ and a sequence of initial service times $\{v_k\}_{k\in\mathbb{N}}$. The arrival process $E(\cdot)$ is a rate $\alpha \in (0,\infty)$ delayed renewal process such that the interarrival times have standard deviation $a \in [0,\infty)$. For $t \in [0,\infty)$, $E(t)$ represents the number of jobs that arrive to the queue during the time interval $(0,t]$. Jobs arriving after time 0 are indexed by integers $j > Z(0)$. For $t \in [0,\infty)$, let

$$A(t) = Z(0) + E(t). \tag{2.2}$$

Then job $j \in \mathbb{N}$ arrives at time $T_j = \inf\{t \in [0,\infty) : A(t) \ge j\}$. Hence, for $i < j$, $T_i \le T_j$ and we say that job i arrives before job j.

For each $k \in \mathbb{N}$, the random variable v_k represents the initial service time of the $(Z(0)+k)$th job. That is, job $j > Z(0)$ has initial service time $v_{j-Z(0)}$. Assume that the random variables $\{v_k\}_{k\in\mathbb{N}}$ are strictly positive and form an independent and identically distributed sequence with common Borel distribution ν on $\mathbb{R}_+$. Assume that the mean $\langle \chi, \nu\rangle \in (0,\infty)$ and standard

deviation $b = \sqrt{\langle \chi^2, \nu \rangle - \langle \chi, \nu \rangle^2} \in [0, \infty)$. Let $\beta = \langle \chi, \nu \rangle^{-1}$. Define the traffic intensity $\rho = \alpha/\beta$.

It will be convenient to combine the stochastic primitives into a single, measure valued load process.

Definition 2.1 *The load process is given by*

$$\mathcal{V}(t) = \sum_{k=1}^{E(t)} \delta_{v_k}, \quad \textit{for } t \in [0, \infty).$$

Then $\mathcal{V}(\cdot) \in \mathbf{D}([0, \infty), \mathbf{M})$.

Evolution of the residual service times. In an SRPT queue, the smallest nonzero residual service time decreases at rate one until either it becomes zero or a job arrives that has a smaller initial service time, at which time the rate changes to zero and the new smallest nonzero residual service time begins decreasing at rate one. We adopt the convention that in case of a tie, the residual service time of the job that arrived first (that is, the job with smaller index) begins decreasing at rate one.

For $j \in \mathbb{N}$ and $t \in [0, \infty)$, let $w_j(t)$ denote the residual service time of job j. By convention, for $j \in \mathbb{N}$ and $t \in [0, T_j]$,

$$w_j(t) = \begin{cases} \tilde{v}_j, & 1 \le j \le Z(0), \\ v_{j-Z(0)}, & j > Z(0). \end{cases}$$

Furthermore, for $j \in \mathbb{N}$, if D_j denotes the time at which job j completes service and departs the system, then $w_j(t) = 0$ for all $t \ge D_j$. On (T_j, D_j), $w_j(\cdot)$ is nonincreasing. In particular, $w_j(\cdot)$ decreases at rate one when job j is in service, and is constant when job j is not in service. See [4] for a detailed definition of the residual service times.

Measure-valued state descriptor. For $t \in [0, \infty)$, define the state descriptor by

$$\mathcal{Z}(t) = \sum_{j=1}^{A(t)} \delta^+_{w_j(t)}. \tag{2.3}$$

3 Diffusion Limit Theorem

We first define a sequence of systems over which the limit is taken. Let $\mathcal{R}$ be a sequence of positive real numbers increasing to infinity. Consider an $\mathcal{R}$-indexed sequence of stochastic models, each defined as in Section 2. For each $r \in \mathcal{R}$, there is an initial condition $\mathcal{Z}^r(0)$; there are stochastic primitives $E^r(\cdot)$ and $\{v_k^r\}_{k\in\mathbb{N}}$ with parameters α^r, a^r, ν^r, β^r, b^r, and ρ^r, and an arrival process $A^r(\cdot)$ with arrival times $\{T_j^r\}_{j\in\mathbb{N}}$; there is a corresponding measure valued load process $\mathcal{V}^r(\cdot)$; there is a state descriptor $\mathcal{Z}^r(\cdot)$. The stochastic elements of each model are defined on a probability space $(\Omega^r, \mathcal{F}^r, \mathbf{P}^r)$ with expectation operator $\mathbf{E}^r$. A diffusion scaling (or central limit theorem scaling) is applied to each model in the $\mathcal{R}$-indexed sequence as follows. For each $r \in \mathcal{R}$ and $t \in [0, \infty)$, let

$$\widehat{E}^r(t) = \frac{1}{r}\left(E^r(r^2t) - r^2t\alpha^r\right). \tag{3.1}$$

Also, for each $r \in \mathcal{R}$ and $t \in [0, \infty)$, let

$$\widehat{\mathcal{Z}}^r(t) = \frac{1}{r}\mathcal{Z}^r(r^2t) \qquad \text{and} \qquad \widehat{W}^r(t) = \langle \chi, \widehat{\mathcal{Z}}^r(t)\rangle. \tag{3.2}$$

Let $\alpha, a \in (0, \infty)$ and define $\alpha(t) = \alpha t$ for all $t \in [0, \infty)$. Let ν be a probability measure such that

$$\langle 1_{\{0\}}, \nu\rangle = 0, \qquad \langle \chi, \nu\rangle = 1/\alpha, \qquad \text{and} \qquad 0 < \langle \chi^2, \nu\rangle < \infty. \tag{3.3}$$

Set $b = \sqrt{\langle \chi^2, \nu\rangle - \langle \chi, \nu\rangle^2}$ and

$$x^* = \sup\{x \in \mathbb{R}_+ : \langle 1_{[0,x]}, \nu\rangle < 1\}.$$

For the sequence of stochastic primitives we make the following asymptotic assumptions. For the exogenous arrival processes, assume that as $r \to \infty$,

$$\alpha^r \to \alpha, \qquad a^r \to a, \qquad \text{and} \qquad \widehat{E}^r(\cdot) \Rightarrow E^*(\cdot), \tag{3.4}$$

where $E^*(\cdot)$ is a Brownian motion starting from zero with drift zero and variance $a^2\alpha^3$ per unit time. This implies a functional weak law of large numbers for the exogenous arrival processes. In particular, it implies that as $r \to \infty$,

$$\bar{E}^r(\cdot) \Rightarrow \alpha(\cdot),$$

where $\bar{E}^r(t) = E^r(r^2 t)/r^2$ for all $t \in [0,\infty)$ and $r \in \mathcal{R}$. For the sequence of service time distributions, assume that as $r \to \infty$,

$$\nu^r \xrightarrow{w} \nu \qquad \text{and} \qquad \langle \chi^2, \nu^r \rangle \to \langle \chi^2, \nu \rangle. \tag{3.5}$$

Then $\beta^r \to \alpha$, $\rho^r \to 1$, and $b^r \to b$ as $r \to \infty$. Furthermore, $\{\nu^r, r \in \mathcal{R}\}$ satisfies a Lindeberg-Feller condition, i.e., for all $\varepsilon > 0$,

$$\lim_{r\to\infty} \langle (\chi - \langle \chi, \nu^r \rangle)^2 \left(1_{[0,\langle \chi,\nu^r\rangle - \varepsilon r)} + 1_{(\langle \chi,\nu^r\rangle + \varepsilon r,\infty)}\right), \nu^r \rangle = 0.$$

In addition, assume the heavy traffic condition that for some $\gamma \in \mathbb{R}$,

$$\lim_{r\to\infty} r(1-\rho^r) = \gamma. \tag{3.6}$$

Finally, if $x^* < \infty$, also assume that for all $x > x^*$,

$$\lim_{r\to\infty} r\langle \chi 1_{(x,\infty)}, \nu^r \rangle = 0. \tag{3.7}$$

For the sequence of fluid scaled initial conditions $\{\widehat{\mathcal{Z}}^r(0) : r > 0\}$, assume that as $r \to \infty$,

$$\widehat{W}^r(0) \Rightarrow W_0^*, \tag{3.8}$$

for some random variable W_0^*. Then from (3.4), (3.5), (3.6), (3.8), and the fact that SRPT is a work conserving discipline, it follows that, as $r \to \infty$,

$$\widehat{W}^r(\cdot) \Rightarrow W^*(\cdot), \tag{3.9}$$

where $W^*(\cdot)$ is a reflected Brownian motion with initial value $W^*(0) \stackrel{d}{=} W_0^*$, variance $(a^2+b^2)\alpha$ per unit time, and drift $-\gamma$ (see [7]). Further assume that, as $r \to \infty$,

$$\widehat{\mathcal{Z}}^r(0) \Rightarrow \begin{cases} \frac{W_0^*}{x^*}\delta_{x^*}, & \text{if } x^* < \infty, \\ \mathbf{0}, & \text{if } x^* = \infty. \end{cases} \tag{3.10}$$

Theorem 3.1 *Under the asymptotic assumptions* (3.4), (3.5), (3.6), (3.7), (3.8), *and* (3.10), *the sequence* $\{\widehat{\mathcal{Z}}^r(\cdot) : r \in \mathcal{R}\}$ *converges in distribution on* $\mathbf{D}([0,\infty), \mathbf{M})$ *to a measure valued process* $\mathcal{Z}^*(\cdot)$ *such that*

$$\mathcal{Z}^*(\cdot) \stackrel{d}{=} \begin{cases} \frac{W^*(\cdot)}{x^*}\delta_{x^*}, & \textit{if } x^* < \infty, \\ \mathbf{0}, & \textit{if } x^* = \infty. \end{cases}$$

This result, in the first case when $x^* < \infty$, is a continuous analog of the diffusion limit result for a multi-class static buffer priority queue, where in the diffusion limit work only resides in the lowest priority class [19]. In an SRPT queue, those jobs with larger service times receive lower priority. Hence, an informal restatement of the first case is that in the diffusion limit the work concentrates in jobs with the largest possible service time, i.e., the lowest priority. The case when $x^* = \infty$ is the natural extension of this result when there is no largest possible service time. Indeed, for the work to get pushed out to infinity on diffusion scale while the diffusion scaled workload process converges, the queue length must necessarily tend to zero.

4 Proofs

Throughout this section we assume that (3.4), (3.5), (3.6), (3.7), (3.8), and (3.10) hold. In Section 4.1, we state a well known result and use it to derive three diffusion limit results to be used in the sequel. In Section 4.2, Theorem 3.1 is proved.

4.1 Diffusion Limits for Load Related Processes

The following result is well known and follows from [13, Theorem 3.1] used to extend [2, Section 17.3].

Proposition 4.1 *For each $r \in \mathcal{R}$, let $\{x_k^r\}_{k=1}^{\infty}$ be an independent and identically distributed sequence of nonnegative random variables on $(\Omega^r, \mathcal{F}^r, \mathbf{P}^r)$ with finite mean μ^r and standard deviation σ^r, that is independent of $E^r(\cdot)$. Suppose that for some finite nonnegative constants μ and σ, $\mu^r \to \mu$ and $\sigma^r \to \sigma$ as $r \to \infty$. Further assume that for each $\varepsilon > 0$,*

$$\lim_{r\to\infty} \mathbf{E}^r \left[(x_1^r - \mu^r)^2; |x_1^r - \mu^r| > r\varepsilon \right] = 0.$$

For $r \in \mathcal{R}$, $n \in \mathbb{N}$, and $t \in [0, \infty)$, let

$$X^r(n) = \sum_{k=1}^{n} x_k^r. \qquad \text{and} \qquad \widehat{X}^r(t) = \frac{X^r(\lfloor r^2 t \rfloor) - \lfloor r^2 t \rfloor \mu^r}{r}.$$

Then as $r \to \infty$, $(\widehat{E}^r(\cdot), \widehat{X}^r(\cdot)) \Rightarrow (E^*(\cdot), X^*(\cdot))$, where $E^*(\cdot)$ is given by (3.4) and $X^*(\cdot)$ is a Brownian motion starting from zero with zero drift and

variance σ^2 per unit time, that is independent of $E^*(\cdot)$. Furthermore, as $r \to \infty$,

$$\frac{X^r(r^2\bar{E}^r(\cdot)) - r^2\alpha^r(\cdot)\mu^r}{r} \Rightarrow X^*(\alpha(\cdot)) + \mu E^*(\cdot),$$

where for each $r \in \mathcal{R}$ and $t \in [0,\infty)$, $\alpha^r(t) = \alpha^r t$.

Note that the limiting process $X^*(\alpha(\cdot)) + \mu E^*(\cdot)$ in Proposition 4.1 is a Brownian motion starting from zero with zero drift and variance $\alpha\sigma^2+\mu^2\alpha^3a^2$ per unit time. We apply this proposition to three processes of interest here, that we respectively refer to as the total load, the truncated load, and the tail load processes. For $r \in \mathcal{R}$ and $t \in [0,\infty)$, let

$$\widehat{\mathcal{V}}^r(t) = \frac{1}{r}\left(\mathcal{V}^r(r^2t) - r^2\alpha^r(\cdot)\nu^r\right).$$

Then, for $r \in \mathcal{R}$, let the total load and scaled total load processes be given respectively by

$$V^r(\cdot) = \langle \chi, \mathcal{V}^r(\cdot)\rangle \qquad \text{and} \qquad \widehat{V}^r(\cdot) = \langle \chi, \widehat{\mathcal{V}}^r(\cdot)\rangle.$$

Then, for $r \in \mathcal{R}$,

$$\widehat{V}^r(\cdot) = \frac{\sum_{k=1}^{r^2\bar{E}^r(\cdot)} v_k^r - r^2\alpha^r(\cdot)\langle \chi, \nu^r\rangle}{r}.$$

From (3.5) and Proposition 4.1, it follows that as $r \to \infty$,

$$\widehat{V}^r(\cdot) \Rightarrow V^*(\cdot),$$

where $V^*(\cdot)$ is a Brownian motion starting from zero with zero drift and variance $\alpha(a^2+b^2)$ per unit time.

Next we consider the truncated load process. For $r \in \mathcal{R}$ and $x \in \mathbb{R}_+$, let

$$V_x^r(\cdot) = \langle \chi 1_{[0,x]}, \mathcal{V}^r(\cdot)\rangle \qquad \text{and} \qquad \widehat{V}_x^r(\cdot) = \langle \chi 1_{[0,x]}, \widehat{\mathcal{V}}^r(\cdot)\rangle.$$

Then, for $r \in \mathcal{R}$ and $x \in \mathbb{R}_+$,

$$\widehat{V}_x^r(\cdot) = \frac{\sum_{k=1}^{r^2\bar{E}^r(\cdot)} v_k^r 1_{\{v_k^r \le x\}} - r^2\alpha^r(\cdot)\langle \chi 1_{[0,x]}, \nu^r\rangle}{r}.$$

Note that (3.5) implies that for any ν-continuity point $x \in \mathbb{R}_+$, as $r \to \infty$,

$$\langle \chi^2 1_{[0,x]}, \nu^r\rangle \to \langle \chi^2 1_{[0,x]}, \nu\rangle. \tag{4.1}$$

Hence (3.5) and Proposition 4.1 imply that for any ν-continuity point $x \in \mathbb{R}_+$, as $r \to \infty$,

$$\widehat{V}_x^r(\cdot) \Rightarrow V_x^*(\cdot), \tag{4.2}$$

where $V_x^*(\cdot)$ is a Brownian motion starting from zero with drift zero and finite variance per unit time.

Finally, for each $r \in \mathcal{R}$ and $x \in \mathbb{R}_+$, we consider the tail load process $V^r(\cdot) - V_x^r(\cdot)$. Then, for $r \in \mathcal{R}$ and $x \in \mathbb{R}_+$,

$$\widehat{V}^r(\cdot) - \widehat{V}_x^r(\cdot) = \frac{\sum_{k=1}^{r^2 \bar{E}^r(\cdot)} v_k^r 1_{\{v_k^r > x\}} - r^2 \alpha^r(\cdot) \langle \chi 1_{(x,\infty)}, \nu^r \rangle}{r}.$$

Note that (3.5) (which implies (4.1)) also implies that for any ν-continuity point $x \in \mathbb{R}_+$, as $r \to \infty$,

$$\langle \chi^2 1_{(x,\infty)}, \nu^r \rangle \to \langle \chi^2 1_{(x,\infty)}, \nu \rangle.$$

Hence, (3.5) and Proposition 4.1 imply that for any ν-continuity point $x \in \mathbb{R}_+$, as $r \to \infty$,

$$\widehat{V}^r(\cdot) - \widehat{V}_x^r(\cdot) \Rightarrow T_x^*(\cdot), \tag{4.3}$$

where $T_x^*(\cdot)$ is a Brownian motion starting from zero with drift zero and variance s_x^2 per unit time. Here,

$$s_x^2 = \alpha(\langle \chi^2 1_{(x,\infty)}, \nu \rangle - \langle \chi 1_{(x,\infty)}, \nu \rangle^2) + \langle \chi 1_{(x,\infty)}, \nu \rangle^2 \alpha^3 a^2. \tag{4.4}$$

Notice that if $x^* < \infty$ and $x > x^*$, then x is a ν-continuity point and $\langle 1_{(x,\infty)}, \nu \rangle = 0$. Hence, if $x > x^*$, then in (4.4), $s_x^2 = 0$, i.e.,

$$T_x^*(\cdot) \equiv 0. \tag{4.5}$$

4.2 Proof of the Main Theorem

Here we use the diffusion limits for the load related processes derived in Section 4.1 to prove the main result. We use the result about the scaled truncated load process to prove that, on diffusion scale, the truncated queue length tends to zero when the truncation is below x^*, the supremum of the support of the limiting service time distribution. Then we use the result about the scaled tail load processes to prove that, on diffusion scale, the queue length above x tends to zero when x is above x^*. Then these two

results are put together to show that in the diffusion limit, the queue mass concentrates at x^*.

For $r \in \mathcal{R}$ and $x \in \mathbb{R}_+$, let

$$\begin{aligned} Z_x^r(\cdot) &= \langle 1_{[0,x]}, \mathcal{Z}^r(\cdot)\rangle \qquad \text{and} \qquad W_x^r(\cdot) = \langle \chi 1_{[0,x]}, \mathcal{Z}^r(\cdot)\rangle, \\ \widehat{Z}_x^r(\cdot) &= \langle 1_{[0,x]}, \widehat{\mathcal{Z}}^r(\cdot)\rangle \qquad \text{and} \qquad \widehat{W}_x^r(\cdot) = \langle \chi 1_{[0,x]}, \widehat{\mathcal{Z}}^r(\cdot)\rangle. \end{aligned}$$

Lemma 4.2 *For any $x \in (0, x^*)$, as $r \to \infty$,*

$$\widehat{Z}_x^r(\cdot) \Rightarrow 0. \tag{4.6}$$

Proof. Since $\widehat{Z}_y^r(\cdot) \le \widehat{Z}_x^r(\cdot)$ for each $0 < y \le x < x^*$, it suffices to verify (4.6) for $x \in (0, x^*)$ that are ν-continuity points. Fix such an x. For $r \in \mathcal{R}$ and $t \in [0, \infty)$, let

$$\tau_x^r(t) = \sup\{s \in [0,t] : \widehat{Z}_x^r(s) = 0\},$$

which is taken to be zero if $\{s \in [0,t] : \widehat{Z}_x^r(s) = 0\} = \emptyset$. Then, for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\begin{aligned} \widehat{Z}_x^r(t) &\le \widehat{Z}_x^r(\tau_x^r(t)) + \frac{E^r(r^2 t) - E^r(r^2 \tau_x^r(t))}{r} \\ &= \widehat{Z}_x^r(\tau_x^r(t)) + \widehat{E}^r(t) - \widehat{E}^r(\tau_x^r(t)) + r\,(t - \tau_x^r(t))\,\alpha^r. \end{aligned} \tag{4.7}$$

We wish to obtain an upper bound on $\widehat{Z}_x^r(\tau_x^r(\cdot))$. Fix $r \in \mathcal{R}$ and $t \in [0, \infty)$. Either $\tau_x^r(t) = 0$ or $\tau_x^r(t) > 0$. If $\tau_x^r(t) = 0$, then $\widehat{Z}_x^r(\tau_x^r(t)) = \widehat{Z}_x^r(0)$. Otherwise, $\tau_x^r(t) > 0$. If $\widehat{Z}_x^r(\tau_x^r(t)) = 0$, then any nonnegative upper bound suffices. Hence, without loss of generality, we also assume that $\widehat{Z}_x^r(\tau_x^r(t)) > 0$. Then $\widehat{Z}_x^r(\tau_x^r(t)-) = 0$ and $\widehat{Z}_x^r(\tau_x^r(t)) > 0$ results from the following two possibilities. In the rth system at time $r^2\tau_x^r(t)$, the exogenous arrival process jumps and at least one of the entering jobs has an initial service time in $[0, x]$, and/or the residual service time of the job in service just before time $r^2\tau_x^r(t)$ decreases to x. Hence, $\widehat{Z}_x^r(\tau_x^r(t)) \le \widehat{E}^r(\tau_x^r(t)) - \widehat{E}^r(\tau_x^r(t)-) + \frac{1}{r}$. Combining the bounds for $\tau_x^r(t) = 0$ or $\tau_x^r(t) > 0$ gives

$$\widehat{Z}_x^r(\tau_x^r(t)) \le \widehat{Z}_x^r(0) + \widehat{E}^r(\tau_x^r(t)) - \widehat{E}^r(\tau_x^r(t)-) + \frac{1}{r},$$

where we adopt the convention $\widehat{E}^r(0-) = \widehat{E}^r(0) = 0$. Hence, for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\widehat{Z}_x^r(t) \leq \widehat{Z}_x^r(0) + \widehat{E}^r(t) - \widehat{E}^r(\tau_x^r(t)-) + \frac{1}{r} + r\left(t - \tau_x^r(t)\right)\alpha^r. \tag{4.8}$$

For $r \in \mathcal{R}$ and $t \in [0, \infty)$, let $\theta_x^r(t) = t - \tau_x^r(t)$. In order to show that the upper bound in (4.8) tends to zero and thereby prove (4.6), it suffices to show that as $r \to \infty$,

$$r\theta_x^r(\cdot) \Rightarrow 0. \tag{4.9}$$

To see this, for $r \in \mathcal{R}$ and $t \in [0, \infty)$, let

$$\tilde{\theta}_x^r(t) = \theta_x^r(t) + \frac{1}{r^2}.$$

Then (4.9) implies that as $r \to \infty$

$$\theta_x^r(\cdot) \Rightarrow 0 \qquad \text{and} \qquad \tilde{\theta}_x^r(\cdot) \Rightarrow 0. \tag{4.10}$$

Also note that for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\widehat{E}^r(t) - \widehat{E}^r(\tau_x^r(t)-) = \widehat{E}^r(t) - \frac{1}{r}E^r(r^2\tau_x^r(t)-) + r\tau_x^r(t)\alpha^r.$$

Hence, for each $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\widehat{E}^r(t) - \widehat{E}^r\left(\tau_x^r(t)\right) \leq \widehat{E}^r(t) - \widehat{E}^r(\tau_x^r(t)-) \leq \widehat{E}^r(t) - \widehat{E}^r\left(\tau_x^r(t) - \frac{1}{r^2}\right) + \frac{\alpha^r}{r},$$

where we adopt the convention that $E^r(t) = E^r(0)$ if $t < 0$. Therefore, for each $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\widehat{E}^r(t) - \widehat{E}^r\left(t - \theta_x^r(t)\right) \leq \widehat{E}^r(t) - \widehat{E}^r(\tau_x^r(t)-) \leq \widehat{E}^r(t) - \widehat{E}^r\left(t - \tilde{\theta}_x^r(t)\right) + \frac{\alpha^r}{r}.$$

By (3.4), the fact that $E^*(\cdot)$ is continuous almost surely, and (4.10), it follows that, as $r \to \infty$,

$$\widehat{E}^r(\cdot) - \widehat{E}^r\left(\cdot - \theta_x^r(\cdot)\right) \Rightarrow 0 \qquad \text{and} \qquad \widehat{E}^r(\cdot) - \widehat{E}^r\left(\cdot - \tilde{\theta}_x^r(\cdot)\right) + \frac{\alpha^r}{r} \Rightarrow 0.$$

(see [2, Section 17]). Hence, as $r \to \infty$,

$$\widehat{E}^r(\cdot) - \widehat{E}^r(\tau_x^r(\cdot)-) \Rightarrow 0. \tag{4.11}$$

Then (4.8), (3.10), (4.11), (3.4), and (4.9) together imply (4.6).

Hence all that remains is to prove (4.9). For this, we exploit the behavior of $\widehat{W}_x^r(\cdot)$ on $(\tau_x^r(\cdot), \cdot]$ to derive an expression that involves $\theta_x^r(\cdot)$. In particular, since the service discipline is SRPT, for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$W_x^r(r^2 t) = W_x^r(r^2 \tau_x^r(t)) + V_x^r(r^2 t) - V_x^r(r^2 \tau_x^r(t)) - r^2(t - \tau_x^r(t)).$$

Then, for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\widehat{W}_x^r(t) = \widehat{W}_x^r(\tau_x^r(t)) + \widehat{V}_x^r(t) - \widehat{V}_x^r(\tau_x^r(t)) + \left(\alpha^r \langle \chi 1_{[0,x]}, \nu^r \rangle - 1\right) r\theta_x^r(t).$$

Using the same line of reasoning that gave rise to (4.8), for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\widehat{W}_x^r(t) \leq \widehat{W}_x^r(0) + \widehat{V}_x^r(t) - \widehat{V}_x^r(\tau_x^r(t)-) + \frac{x}{r} + \left(\alpha^r \langle \chi 1_{[0,x]}, \nu^r \rangle - 1\right) r\theta_x^r(t).$$

Since $\widehat{W}_x^r(t) \geq 0$ for all $r \in \mathcal{R}$ and $t \in [0, \infty)$, it follows that for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\left(1 - \alpha^r \langle \chi 1_{[0,x]}, \nu^r \rangle\right) \theta_x^r(t) \leq \frac{\widehat{W}_x^r(0)}{r} + \frac{\widehat{V}_x^r(t)}{r} - \frac{\widehat{V}_x^r(\tau_x^r(t)-)}{r} + \frac{x}{r^2}. \quad (4.12)$$

By (3.5) and the fact that x is a ν-continuity point, we have that

$$\lim_{r \to \infty} 1 - \alpha^r \langle \chi 1_{[0,x]}, \nu^r \rangle = 1 - \alpha \langle \chi 1_{[0,x]}, \nu \rangle > 0. \quad (4.13)$$

Hence, for r sufficiently large, $\left(1 - \alpha^r \langle \chi 1_{[0,x]}, \nu^r \rangle\right) \theta_x^r(\cdot) \geq 0$. Then (4.12), (3.10), (4.2), and (4.13) together imply that as $r \to \infty$,

$$\theta_x^r(\cdot) \Rightarrow 0.$$

Hence, by (4.2) and the same line of reasoning that gave rise to (4.11), as $r \to \infty$,

$$\widehat{V}_x^r(\cdot) - \widehat{V}_x^r(\tau_x^r(\cdot)-) \Rightarrow 0.$$

Therefore, if one multiplies (4.12) by r and uses this and (3.10), (4.9) follows. □

We are ready to use Lemma 4.2, (4.3), and (4.5) to prove the main theorem.

Proof of Theorem 3.1. First suppose that $x^* = \infty$. Then it suffices to show that as $r \to \infty$,

$$\widehat{Z}^r(\cdot) \Rightarrow 0. \tag{4.14}$$

For $r \in \mathcal{R}$, $x \in \mathbb{R}_+$ and $t \in [0, \infty)$, we have

$$\begin{aligned}\widehat{Z}^r(t) &= \widehat{Z}^r_x(t) + \langle 1_{(x,\infty)}, \widehat{\mathcal{Z}}^r(t)\rangle \\ &\le \widehat{Z}^r_x(t) + \frac{1}{x}\langle \chi 1_{(x,\infty)}, \widehat{\mathcal{Z}}^r(t)\rangle \\ &\le \widehat{Z}^r_x(t) + \frac{1}{x}\widehat{W}^r(t).\end{aligned}$$

Hence (4.14) follows from Lemma 4.2, (3.9), and the fact that x is arbitrary.

Next suppose that $x^* < \infty$ and let $\varepsilon > 0$ be such that $x^* - \varepsilon$ is a ν-continuity point. Then by Lemma 4.2, as $r \to \infty$,

$$\widehat{Z}^r_{x^*-\varepsilon}(\cdot) \Rightarrow 0. \tag{4.15}$$

For $r \in \mathcal{R}$ and $t \in [0, \infty)$, we have

$$\langle 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(t)\rangle \le \frac{1}{x^* + \varepsilon}\langle \chi 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(t)\rangle.$$

But, for $r \in \mathcal{R}$ and $t \in [0, \infty)$,

$$\begin{aligned}\langle \chi 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(t)\rangle &\le \langle \chi 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(0)\rangle + \frac{V^r(r^2t) - V^r_{x^*+\varepsilon}(r^2t)}{r} \\ &\le \langle \chi 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(0)\rangle + \widehat{V}^r(t) - \widehat{V}^r_{x^*+\varepsilon}(t) \\ &\quad + rt\alpha^r\langle \chi 1_{(x^*+\varepsilon,\infty)}, \nu^r\rangle.\end{aligned}$$

Hence, by (3.8), (3.10), (4.3), (4.5), and (3.7), as $r \to \infty$,

$$\langle \chi 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(\cdot)\rangle \Rightarrow 0. \tag{4.16}$$

Therefore,

$$\langle 1_{(x^*+\varepsilon,\infty)}, \widehat{\mathcal{Z}}^r(\cdot)\rangle \Rightarrow 0. \tag{4.17}$$

In addition, (4.16) together with (4.15) and (3.9) implies that as $r \to \infty$,

$$\langle \chi 1_{(x^*-\varepsilon,x^*+\varepsilon]}, \widehat{\mathcal{Z}}^r(\cdot)\rangle \Rightarrow W^*(\cdot). \tag{4.18}$$

Since for $r \in \mathcal{R}$,

$$\frac{1}{x^* + \varepsilon}\langle \chi 1_{(x^*-\varepsilon,x^*+\varepsilon]}, \widehat{\mathcal{Z}}^r(\cdot)\rangle \le \langle 1_{(x^*-\varepsilon,x^*+\varepsilon]}, \widehat{\mathcal{Z}}^r(\cdot)\rangle \le \frac{1}{x^* - \varepsilon}\langle \chi 1_{(x^*-\varepsilon,x^*+\varepsilon]}, \widehat{\mathcal{Z}}^r(\cdot)\rangle,$$

(4.18), (4.15), (4.17), and the fact that $\varepsilon > 0$ can be made arbitrarily small completes the proof. □

References

[1] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: investigating unfairness. In *ACM SIGMETRICS Performance Evaluation Review*, 29:279–290, 2001.

[2] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, Inc., New York, 1968.

[3] D. G. Down, H. C. Gromoll, and A. L. Puha. State-dependent response times via fluid limits for shortest remaining processing time queues. In *ACM SIGMETRICS Performance Evaluation and Review*, 37:75–76, 2009.

[4] D. G. Down, H. C. Gromoll, and A. L. Puha. Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research*, 34:880–911, 2009.

[5] D. G. Down and R. Wu. Multi-layered round robin routing for parallel servers. *Queueing Systems*, 53:177–188, 2006.

[6] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, Inc., New York, 1986.

[7] D. Iglehart and W. Whitt. Multiple channel queues in heavy traffic I. *Advances in Applied Probability*, 2:150–177, 1970.

[8] R. Núñez Queija. Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research*, 113:101–117, 2002.

[9] M. Nuyens and B. Zwart. A large deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems*, 54:85–97, 2006.

[10] A. V. Pavlov. A system with Schrage servicing discipline in the case of a high load. *Engrg. Cybernetics* 21:114–121, 1984; translated from *Izv. Akad. Nauk SSSR Tekhn. Kibernet*, 6:59–66, 1983 (Russian).

[11] A. V. Pechinkin. Heavy traffic in a system with a discipline of priority servicing for the job with the shortest remaining length with interruption (Russian). *Math. Issled. No. 89, Veroyatn. Anal.*, 97:85–93, 1986.

[12] R. Perera. The variance of delay time in queueing system M/G/1 with optimal strategy SRPT. *Archiv für Elektronik und Übertragungstechnik*, 47:110–114, 1993.

[13] Yu. V. Prohorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability and its Applications*, 1:157–214, 1956.

[14] R. Schassberger. The steady-state appearance of the M/G/1 queue under the discipline of shortest remaining processing time. *Advances in Applied Probability*, 22:456–479, 1990.

[15] L. E. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:687–690, 1968.

[16] L. E. Schrage and L. W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14:670–684, 1966.

[17] F. Schreiber. Properties and applications of the optimal queueing strategy SRPT: a survey. *Archiv für Elektronik und Übertragungstechnik*, 47:372–378, 1993.

[18] D. R. Smith. A new proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 26:197–199, 1976.

[19] W. Whitt. Weak convergence theorems for priority queues: Preemptive-resume discipline. *Journal of Applied Probability*, 8:74–94, 1970.

[20] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/G/1. In *Proceedings of the 2003 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, 238–249, 2003.

Department of Mathematics
University of Virginia
Charlottesville, VA 22903
E-mail: gromoll@virginia.edu

Maria Curie-Skłodowska University
Department of Mathematics
Lublin, Poland
E-mail: lkruk@hektor.umcs.lublin.pl

Department of Mathematics
California State University, San Marcos
San Marcos CA 92096
E-mail: apuha@csusm.edu