Annals of Operations Research (2022) 310:411–429
https://doi.org/10.1007/s10479-021-03929-0

S.I. : QUEUESTOCHMOD2019

Heavy traffic analysis for single-server SRPT and LRPT queues via EDF diffusion limits

Łukasz Kruk[1]

Accepted: 5 January 2021 / Published online: 19 January 2021

Abstract
Extending the results of Kruk (Queueing theory and network applications. QTNA 2019. Lecture notes in computer science, vol 11688. Springer, Cham, pp 263–275, 2019), we derive heavy traffic limit theorems for a single server, single customer class queue in which the server uses the Shortest Remaining Processing Time (SRPT) policy from heavy traffic limits for the corresponding Earliest Deadline First queueing systems. Our analysis allows for correlated customer inter-arrival and service times and heavy-tailed inter-arrival and service time distributions, as long as the corresponding stochastic primitive processes converge weakly to continuous limits under heavy traffic scaling. Our approach yields simple, concise justifications and new insights for SRPT heavy traffic limit theorems of Gromoll et al. (Stoch Syst 1(1):1–16, 2011). Corresponding results for the longest remaining processing time policy are also provided.

Keywords Heavy traffic · Queueing · Shortest remaining processing time · Earliest deadline first · Heavy traffic limit

Mathematics Subject Classification 60K25 · 60G57 · 68M20 · 90B22 · 90B36

1 Introduction

1.1 Summary of prior results

Under the Shortest Remaining Processing Time (SRPT) service protocol, preemptive priority is given to a job with the shortest residual service time. This discipline has long been known to possess several attractive features, in particular to minimize the mean response time (see Schrage and Miller 1966) and the queue length at any point of time in a single-server system (Schrage 1968). However, SRPT is not often encountered in practical applications, because it is believed to unfairly penalize big tasks (see, e.g., Bender et al. 1998), although this objection has largely been dismissed, e.g., by Bansal and Harchol-Balter (2001), Wier-

✉ Łukasz Kruk
lkruk@hektor.umcs.lublin.pl

[1] Department of Mathematics, Maria Curie-Skłodowska University, Pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland

man and Harchol-Balter (2003). In particular, Bansal and Harchol-Balter (2001) showed that for an M/G/1 queue with a heavy-tailed service time distribution, at least 99% of the jobs had significantly better expected response times under SRPT than under processor sharing (PS), widely regarded as a fair policy. Wierman and Harchol-Balter (2003) observed that for an M/G/1 queue, SRPT treated fairly all jobs under light loads and, moreover, for highers loads, a load increase resulted in an increase of jobs being treated fairly. Schrage and Miller (1966) provided formulae for the mean response time in an M/G/1 SRPT queue. Their work was later extended by Schassberger (1990) and Perera (1993). Another notable contribution was made by Pavlov (1983) and Pechinkin (1986), who obtained the heavy traffic limit of the invariant distributions for the queue length of an M/G/1 SRPT queue. The tail behavior of single server queues under the SRPT protocol was investigated, e.g., by Núñez-Queija (2002) and Nuyens and Zwart (2006). Núñez-Queija (2002) has shown that in the case of heavy-tailed service time distributions, the tail of the sojourn time distribution and the tail of the corresponding service time distribution coincide up to a constant, which is the best possible behavior. For light-tailed service time distributions, using large-deviations techniques, Nuyens and Zwart (2006) have shown that the decay rate of the sojourn time distribution under SRPT is suboptimal (in fact, in most cases the worst possible). They have concluded that, from a large-deviations point of view, it is not advisable to switch from the First-In, First-Out (FIFO) service discipline, known for the maximal sojourn time distribution decay rate (see Ramanan and Stolyar 2001), to SRPT.

Several functional limit theorems for SRPT queueing systems have been established. Down et al. (2009) defined a fluid model and obtained fluid limits for G/G/1 SRPT queues. Gromoll and Keutel (2012) obtained the same fluid limits in the case of the Shortest Job First (SJF) protocol, a non-preemptive version of SRPT. Kruk and Sokołowska (2016) generalized the results of Down et al. (2009) to SRPT queues with multiple inputs. Down and Wu (2006) used diffusion limits to obtain some optimality properties of a multi-layered round robin routing policy in a system of parallel servers operating under SRPT, with a finitely supported service time distribution. Gromoll et al. (2011) established diffusion limits for G/G/1 SRPT queues with general service time distributions. Recently, Puha (2015) and Banerjee et al. (2020) obtained diffusion limits for a G/G/1 SRPT system under nonstandard spatial scaling. Kruk (2019) observed that, under suitable assumptions, the results of Gromoll et al. (2011) could be obtained from heavy traffic limits of Kruk (2007) for preemptive G/G/1 Earliest Deadline First (EDF) queues with job service times correlated with their initial lead times. Recall that under the EDF service discipline, priority is given to the task with the shortest lead time. The main idea of the arguments in Kruk (2019) was to compare the SRPT system with queueing systems having the same stochastic primitives, but operating under EDF, with initial job lead times set to be large multiples of their service times. Since, by the theorem of Schrage (1968), SRPT minimizes the number of customers in the system, the queue length in the EDF system is an upper bound for the queue length in the corresponding SRPT system. Applying this observation, together with heavy traffic limits of Kruk (2007) and elementary lower bounds, the required results may be obtained. The idea of comparing the performance of a SRPT queue with that of the corresponding EDF system, already present in Bender et al. (1998), was previously used to "regularize" the SRPT protocol in order to make it more fair to big tasks. As it has been observed in Kruk (2019), after suitable adjustments, our approach carries over to the Longest Remaining Processing Time (LRPT) discipline, giving preemptive priority to the task with the longest remaining processing time. The latter protocol appears in some applications; see Kittsteiner and Moldovanu (2005), where both SRPT and LRPT queue disciplines arise in equilibria for some priority auctions.

1.2 Our work

In this paper, we observe that, after suitable modifications, the arguments of Kruk (2019) work as long as functional central limit theorems for the customer arrival and service times hold (see (8) and (10), to follow), with the limiting processes having continuous sample paths. These assumptions are satisfied in a much more general setting than the i.i.d. case with finite second moments, analyzed in Kruk (2019). For example, all the stochastic primitives may be correlated and may exhibit short- or long-range dependence. In the latter case, we can consider heavy-tailed interarrival and/or service time distributions. Accordingly, the (appropriately scaled) stochastic primitive processes may converge, e.g., to a fractional Brownian motion or a linear fractional stable motion, in addition to the classic Brownian motion case considered in Kruk (2019). For more information on these issues, see Sections 4.4, 4.6 and 4.7 of Whitt (2002).

The extension we present here is obtained without adding notable technical complications to the proofs. Its main idea is simply truncation of the lead times in the corresponding auxiliary EDF systems (i.e., large multiples of the service times) at suitably large levels. This allows for an uniform application of heavy traffic limit theory for EDF queues with bounded lead times, which is notably simpler than its counterpart for the unbounded lead times case (used in Kruk 2019 to analyze SRPT systems with unbounded service times) and requires less technical assumptions. See Kruk (2007) and our Sect. 7, to follow. On the other hand, it turns out that the resulting EDF systems approximate their SRPT counterparts well enough to provide suitable upper bounds for the queue lengths in the latter systems.

Let us also note that in Kruk (2019), the right endpoints of supports of the customer service time distributions in the pre-limit systems were assumed to coincide with the right endpoint of support of their weak limit. This technical condition was necessary in order to use the heavy traffic limits of Kruk (2007) for EDF queues, where an analogous assumption on the customer lead times was made. In order to relax it, in the appendix we provide Theorem 7, a variant of Theorem A.2 from Kruk (2007), allowing the right endpoints of the rescaled initial lead time distributions in the pre-limit EDF queues to vary in a controlled way. Theorem 7 is the basis of our current work.

As a consequence, even in the classical G/G/1 case with finite second moments, our results extend their counterparts from Kruk (2019). Indeed, in Sect. 3.2, to follow, we give Theorems 3 and 4 , providing diffusion limits for this case under weaker assumptions than the ones required in their counterparts from Kruk (2019). For example, we no longer require that $[0, \infty]$, equipped with a special semimetric, defined in terms of the service time distribution, is a totally bounded semimetric space.

This paper is organized as follows. Section 2 presents the model, notation and assumptions. In Sect. 3, we state our limit theorems for SRPT queues and we show that they extend analogous results for G/G/1 SRPT systems given in Kruk (2019). In Sect. 4, we define a sequence of auxiliary EDF systems and we characterize their asymptotic behavior in heavy traffic. In Sect. 5, we prove our main results stated in Sect. 3, following the lines of the arguments from Kruk (2019). In Sect. 6, we provide analogous results for single server, single customer class LRPT queueing systems. Section 7 is an appendix in which a variant of Theorem A.2 in Kruk (2007) [extending the results of Doytchinov et al. (2001)], suitable for the applications presented in the previous sections, is stated. Section 8 concludes.

2 The model, assumptions and notation

2.1 Notation

We will use the following notation. Let $\mathbb{N} = \{1, 2, \ldots\}$, let $\mathbb{R}$ denote the set of real numbers and let $\mathbb{R}_+ = [0, \infty)$. Let $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ and $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ be equipped with the obvious topologies. For $a, b \in \mathbb{R}$, we write $a \wedge b$ for the minimum of a and b, a^+ for the positive part of a and $\lfloor a \rfloor$ for the largest integer less than or equal to a. Denote by e the identity map on $\mathbb{R}$, i.e., $e(t) = t$, $t \in \mathbb{R}$. For functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}$, we denote the composition of f and g by $g \circ f$, i.e., $(g \circ f)(t) = g(f(t)), t \in \mathbb{R}$. For a function $f : \mathbb{R} \to \mathbb{R}$ and $t \in \mathbb{R}$, define $f(t-) = \lim_{s \uparrow t} f(s)$.

Denote by $\mathcal{B}(\mathbb{R})$ the Borel σ-field on $\mathbb{R}$. Let $\mathcal{M}$ denote the set of all finite, nonnegative measures on $\mathcal{B}(\mathbb{R})$. Under the weak topology, $\mathcal{M}$ is a Polish space (see Prokhorov 1956). We denote the zero measure in $\mathcal{M}$ by $\mathbf{0}$ and the Dirac delta measure with unit mass at $x \in \mathbb{R}$ by δ_x. For $x \in \mathbb{R}_+$, let δ_x^+ be δ_x if $x > 0$ and $\mathbf{0}$ otherwise. For $\xi \in \mathcal{M}$ and a Borel measurable function $f : \mathbb{R} \to \mathbb{R}$ integrable with respect to ξ, we write $\langle f, \xi \rangle$ to denote $\int_R f(x)\xi(dx)$.

The symbol $\Rightarrow$ will be used to denote weak convergence of measures, either on $\mathbb{R}$ (in this case, we use the same symbol for convergence of the corresponding cumulative distribution functions (c.d.f.s)), or on the space $D_S[0, \infty)$ of right-continuous functions with left-hand limits (RCLL functions) from $[0, \infty)$ to a Polish space S with the Skorokhod J_1 topology. Note that $D_S[0, \infty)$ is itself a Polish space. See Ethier and Kurtz (1985) for details. While considering $D_S[0, \infty)$, we take $S = \mathbb{R}$ or $\mathbb{R}^d$, with appropriate dimension d for vector-valued functions, unless explicitly stated otherwise. In the case of $S = \mathbb{R}$, the lower index S in $D_S[0, \infty)$ will usually be skipped.

2.2 The basic model

Consider a sequence of single-server queueing systems, indexed by superscript n, each with a single customer class. The customer inter-arrival times are $\{u_j^n\}_{j=1}^{\infty}$, a sequence of strictly positive, identically distributed random variables (r.v.s) with mean $1/\lambda_n$. The corresponding service times are $\{v_j^n\}_{j=1}^{\infty}$, a sequence of strictly positive, identically distributed r.v.s with distribution function G^n and mean $1/\mu_n$. Each system is empty at time zero and

$$\lim_{n\to\infty} \lambda_n = \lim_{n\to\infty} \mu_n = \lambda > 0. \tag{1}$$

Moreover, we have

$$G^n \Rightarrow G \tag{2}$$

for some c.d.f. G and

$$G_v^n(y) \stackrel{\Delta}{=} \mathbb{E}\left[v_j^n \mathbb{I}_{\{v_j^n \le y\}}\right] \;\Rightarrow\; G_v(y), \tag{3}$$

where G_v is a c.d.f. of a finite positive measure on $\mathbb{R}_+$. By (1), G_v has total mass $1/\lambda$.

We define the customer arrival times

$$S_0^n \stackrel{\Delta}{=} 0, \qquad S_k^n \stackrel{\Delta}{=} \sum_{i=1}^{k} u_i^n, \quad k \ge 1, \tag{4}$$

the customer arrival process $A^n(t) \triangleq \max\{k : S_k^n \le t\}$, $t \ge 0$, and the work arrival process

$$V^n(t) \triangleq \sum_{j=1}^{\lfloor t \rfloor} v_j^n, \quad t \ge 0. \tag{5}$$

Let $W^n(t)$, $t \ge 0$, be the workload process which records the amount of work in the queue. The above processes do not depend of the queue service discipline, provided that the server is never idle when there are unfinished tasks in the system. However, the queue length process $Q^n(t)$, $t \ge 0$, depends on the underlying service protocol.

2.3 Heavy traffic assumptions

Let c_n be a sequence of constants such that $c_n \to \infty$, $n/c_n \to \infty$. We make the *heavy traffic assumption*

$$\lim_{n\to\infty} n(1-\rho_n)/c_n = \gamma \tag{6}$$

for some $\gamma \in \mathbb{R}$, where $\rho_n \triangleq \lambda_n/\mu_n$ is the nth system's *traffic intensity*. For $n \ge 1$ and $t \ge 0$, let

$$\widehat{S}^n(t) \triangleq c_n^{-1}\left[S^n_{\lfloor nt \rfloor} - \lambda_n^{-1} nt\right], \qquad \widehat{V}^n(t) \triangleq c_n^{-1}\left[V^n(nt) - \mu_n^{-1} nt\right], \tag{7}$$
$$\widehat{W}^n(t) \triangleq c_n^{-1} W^n(nt), \qquad \widehat{Q}^n(t) \triangleq c_n^{-1} Q^n(nt).$$

We assume that

$$\left(\widehat{S}^n, \widehat{V}^n\right) \Rightarrow (S^*, V^*) \tag{8}$$

in $D_{\mathbb{R}^2}[0,\infty)$, where S^* and V^* have continuous sample paths. Thus, by Theorem 9.3.4 in Whitt (2002),

$$\widehat{W}^n(t) \Rightarrow W^*(t) \tag{9}$$

in $D[0,\infty)$, where, for any $t \ge 0$, $W^*(t) = N^*(t) - \inf_{0\le s\le t} N^*(s)$ and $N^* = (V^* - S^*) \circ \lambda e - \gamma e$. Finally, we assume that there exists a dense subset C of $\mathbb{R}$ such that for every $a < b$, $a, b \in C$, we have

$$\tilde{V}^n_{a,b}(t) \triangleq \frac{1}{c_n} \sum_{j=1}^{\lfloor nt \rfloor} \left(v_j^n \mathbb{I}_{\{a < v_j^n \le b\}} - (G_v^n(b) - G_v^n(a)) \right) \Rightarrow V^*_{a,b}(t) \tag{10}$$

in $D[0,\infty)$, where $V^*_{a,b}$ has continuous sample paths. Since $\mathbb{E}[v_j^n \mathbb{I}_{\{a<v_j^n\le b\}}] = G_v^n(b) - G_v^n(a)$, the process $\tilde{V}^n_{a,b}$ is an analog of $\widehat{V}^n$, with the customer service times truncated above b and below a. Hence, the assumption (10) is of the same nature as (8) and it holds under similar conditions. Thanks to (10), the assumption (64) in Theorem 7, to follow, which is the cornerstone of our analysis, is satisfied in the cases presented in the next section.

3 Main results

3.1 Results

In this section and in Sects. 4–5 we assume that customers are served using the SRPT queue discipline. In our analysis, we will consider two cases, corresponding to bounded and unbounded service times, respectively. In the first one, we additionally assume that as $n \to \infty$,

$$v_n^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G^n(y) = 1\} \ \to \ v^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G(y) = 1\} < \infty. \tag{11}$$

In the second case,

$$v^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G(y) = 1\} = \infty. \tag{12}$$

The following theorems are the main results of this paper.

Theorem 1 *Assume that* (1)–(3), (6), (8), (10) *for all* $a < b$, $a, b \in C$, *and* (11) *hold. Then* $\widehat{Q}^n \Rightarrow \frac{1}{v^*} W^*$ *in* $D[0, \infty)$ *jointly with* (9) *as* $n \to \infty$.

Theorem 2 *Assume that* (1)–(3), (6), (8), (10) *for all* $a < b$, $a, b \in C$, *and* (12) *hold. Then* $\widehat{Q}^n \Rightarrow 0$ *in* $D[0, \infty)$ *as* $n \to \infty$.

Theorems 1 and 2 can be easily refined to limit theorems for the corresponding measure-valued state descriptors. Let $w_j^n(t)$ be the residual service time at time t of the jth customer to appear in the nth SRPT queueing system. For $n \in \mathbb{N}$, $t \geq 0$ and $B \in \mathcal{B}(\mathbb{R})$, let

$$\mathcal{Q}^n(t) = \sum_{j=1}^{A^n(t)} \delta^+_{w_j^n(t)}, \qquad \widehat{\mathcal{Q}}^n(t)(B) \stackrel{\Delta}{=} c_n^{-1} \mathcal{Q}^n(nt)(B). \tag{13}$$

Note that $\langle 1, \mathcal{Q}^n(t)\rangle = Q^n(t)$ and $\langle e, \mathcal{Q}^n(t)\rangle = W^n(t)$, so $\langle 1, \widehat{\mathcal{Q}}^n(t)\rangle = \widehat{Q}^n(t)$ and $\langle e, \widehat{\mathcal{Q}}^n(t)\rangle = \widehat{W}^n(t)$. Theorem 1 implies

Corollary 1 *Under the assumptions of Theorem 1 we have* $\widehat{\mathcal{Q}}^n \Rightarrow \frac{1}{v^*} W^* \delta_{v^*}$ *in* $D_{\mathcal{M}}[0, \infty)$ *as* $n \to \infty$.

Theorem 2 obviously implies

Corollary 2 *Under the assumptions of Theorem 2 we have* $\widehat{\mathcal{Q}}^n \Rightarrow \mathbf{0}$ *in* $D_{\mathcal{M}}[0, \infty)$ *as* $n \to \infty$.

The proofs of Theorems 1, 2 and Corollary 1 will be given in Sect. 5.

The assumptions of Theorems 1 and 2 can be simplified if the service time distribution $G^n \equiv G$ does not depend on n. In this case, we have

Corollary 3 *Asume that* $G^n \equiv G$ *for every* $n \in \mathbb{N}$,

$$\lim_{n\to\infty} \lambda_n = \lambda \stackrel{\Delta}{=} \frac{1}{\mathbb{E}v_j^n}, \tag{14}$$

and that (6), (8), (10) *for* $a < b$, $a, b \in C$, *hold. If* $v^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G(y) = 1\} < \infty$, *then* $\widehat{\mathcal{Q}}^n \Rightarrow \frac{1}{v^*} W^* \delta_{v^*}$ *in* $D_{\mathcal{M}}[0, \infty)$ *and* $\widehat{Q}^n \Rightarrow \frac{1}{v^*} W^*$ *in* $D[0, \infty)$ *jointly with* (9) *as* $n \to \infty$. *In the opposite case,* $\widehat{\mathcal{Q}}^n \Rightarrow \mathbf{0}$ *in* $D_{\mathcal{M}}[0, \infty)$ *as* $n \to \infty$.

Corollary 3 follows directly from Theorem 1 and Corollaries 1, 2.

3.2 Special case: independent stochastic primitives with second moments

In this subsection we show that the results of the previous subsection extend the corresponding theorems and corollaries from Kruk (2019).

As in Kruk (2019), in this subsection we make the following assumptions. The customer interarrival times form a sequence of strictly positive, independent, identically distributed (i.i.d.) r.v.s with mean $1/\lambda_n$ and standard deviation α_n. The service times $\{v_j^n\}_{j=1}^{\infty}$ are also strictly positive, i.i.d., with distribution function G^n, mean $1/\mu_n$ and standard deviation β_n. For every n, the sequences $\{u_j^n\}_{j=1}^{\infty}$ and $\{v_j^n\}_{j=1}^{\infty}$ are mutually independent, each system is empty at time zero and (1)–(3) hold. We assume that

$$\lim_{n\to\infty} \alpha_n = \alpha > 0, \quad \lim_{n\to\infty} \beta_n = \beta > 0, \tag{15}$$

and that (6) holds with

$$c_n = \sqrt{n}, \qquad n \in \mathbb{N}. \tag{16}$$

We also impose the Lindeberg condition on the inter-arrival times:

$$\lim_{n\to\infty} \mathbb{E}\left[\left(u_j^n - (\lambda_n)^{-1}\right)^2 \mathbb{I}_{\left\{\left|u_j^n-(\lambda_n)^{-1}\right|>c\sqrt{n}\right\}}\right] = 0 \qquad \forall c > 0. \tag{17}$$

One may check that the Lindeberg condition on the customer service times follows from (1)–(2) and (15).

Under the above assumptions, Theorem 3.1 of Prokhorov (1956) implies (8), where S^* and V^* are independent, driftless Brownian motions with variances α^2 and β^2 per unit time, respectively. Accordingly, by the (already mentioned) Theorem 9.3.4 in Whitt (2002) (or by the results of Billingsley 1999, Section 17.3), we have (9), where W^* is a Brownian motion with drift $-\gamma$ and variance $\lambda(\alpha^2+\beta^2)$ per unit time, instantaneously reflected at the origin.

We will show that Theorems 1, 2 imply

Theorem 3 *Under the assumptions of this subsection, suppose that* (11) *holds. Then* $\widehat{Q}^n \Rightarrow \frac{1}{v^*} W^*$ *in* $D[0,\infty)$ *jointly with* (9) *as* $n \to \infty$.

Theorem 4 *Under the assumptions of this subsection, suppose that* (12) *holds. Then* $\widehat{Q}^n \Rightarrow 0$ *in* $D[0,\infty)$ *as* $n \to \infty$.

Theorem 3 extends Theorem 1 of Kruk (2019), because our assumption (11) is less restrictive than the corresponding assumption (4) of Kruk (2019), stating that $v_n^* = v^*$ for each $n \in \mathbb{N}$. Also, Theorem 4 extends Theorem 2 of Kruk (2019), because some of the assumptions of the latter theorem [namely (5)–(7) in Kruk (2019) and the assertion that $(\overline{\mathbb{R}}_+, \rho)$ is a totally bounded semimetric space, where the semimetric ρ was defined on p. 266 of Kruk 2019] are not necessary for our Theorem 4 to hold.

In order to apply Theorems 1, 2 in the present context, we have to justify the assumption (10) for a, b belonging to a suitable set C dense in $\mathbb{R}$. Let

$$G_{v^2}^n(y) = \mathbb{E}\left[(v_j^n)^2 \mathbb{I}_{\{v_j^n \le y\}}\right], \qquad y \in \mathbb{R}, \quad n \in \mathbb{N}.$$

By (1) and (15),

$$\sup_{y\in\mathbb{R}} G_{v^2}^n(y) = \lim_{y\to\infty} G_{v^2}^n(y) = \mathbb{E}(v_j^n)^2 = \beta_n^2 + \frac{1}{\mu_n^2} \to \beta^2 + \frac{1}{\lambda^2}, \qquad n \to \infty,$$

and hence the sequence $\{G^n_{v^2}\}_{n=1}^{\infty}$ is uniformly bounded. By Helley's selection principle (see, e.g., Billingsley 1986, Theorem 25.9), for each subsequence of the sequence of the distribution functions $\{G^n_{v^2}\}$, we may extract a further subsequence (also indexed by $n \in \mathbb{N}$ for notational simplicity) such that along this subsequence,

$$G^n_{v^2}(y) \Rightarrow G_{v^2}(y), \tag{18}$$

where G_{v^2} is a c.d.f of a finite positive measure on $\mathbb{R}_+$ (possibly dependent on this subsequence). Let

$$C = \{y \in \mathbb{R} : \ G_v(y-) = G_v(y), \ G_{v^2}(y-) = G_{v^2}(y)\}$$

be the set points of continuity of both G_v and G_{v^2}. Clearly, C is dense in $\mathbb{R}$, since $\mathbb{R}\backslash C$ is at most countable. For the corresponding subsequence of indices, along which (18) holds, and for every $a < b$, $a, b \in C$, Theorem 3.1 of Prokhorov (1956) implies that (10) holds with c_n given by (16), where the limiting process $V^*_{a,b}$ is a driftless Brownian motion with variance $G_{v^2}(b) - G_{v^2}(a) - (G_v(b) - G_v(a))^2$. Therefore, assuming (11) (resp., (12)–(34)), we have that *along the subsequence satisfying* (18), by Theorem 1 (resp., Theorem 2), Theorem 3 (resp., Theorem 4) holds. However, the resulting limiting distribution of the processes $(\widehat{Q}^n, \widehat{W}^n)$ does not depend on the subsequence chosen. Moreover, by Theorem 1.11 of Prokhorov (1956), weak convergence in $D[0, \infty)$ (and hence in $(D[0, \infty))^2$) is metrizable. Consequently, it is not hard to see that actually weak convergence of $(\widehat{Q}^n, \widehat{W}^n)$ holds *for the entire sequence* $(\widehat{Q}^n, \widehat{W}^n)_{n=1}^{\infty}$, as claimed in Theorem 3 (resp., Theorem 4).

Similarly as in the more general setup of Sect. 3.1, Theorems 3 and 4 easily imply their counterparts for the measure-valued state descriptors, namely Corollaries 1 and 2 . We also get the following immediate corollary.

Corollary 4 (Corollary 3 of Kruk 2019) *Assume that* $\{u^n_j\}_{j=1}^{\infty}$ *and* $\{v^n_j\}_{j=1}^{\infty}$ *form mutually independent i.i.d. sequences. Moreover, suppose that* $G^n \equiv G$ *for every* $n \in \mathbb{N}$ *and that* (6), (14)–(17) *hold. If* $v^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G(y) = 1\} < \infty$, *then* $\widehat{\mathcal{Q}}^n \Rightarrow \frac{1}{v^*} W^* \delta_{v^*}$ *in* $D_{\mathcal{M}}[0, \infty)$ *and* $\widehat{Q}^n \Rightarrow \frac{1}{v^*} W^*$ *in* $D[0, \infty)$ *jointly with* (9) *as* $n \to \infty$, *where* W^* *is a reflected Brownian motion with drift* $-\gamma$ *and variance* $\lambda(\alpha^2 + \beta^2)$. *In the opposite case,* $\widehat{\mathcal{Q}}^n \Rightarrow \mathbf{0}$ *in* $D_{\mathcal{M}}[0, \infty)$ *as* $n \to \infty$.

4 Approximating EDF systems and their asymptotics

This section is a preparation for the proofs of our main results, which will be provided in Sect. 5. Here we define a sequence of auxiliary EDF systems and we characterize their asymptotic behavior in heavy traffic.

4.1 Approximating EDF systems

Without loss of generality we can take $\lambda = 1$, because it is only a convenient rescaling. Fix $M \in \mathbb{N}$. Consider a sequence of auxiliary EDF queueing systems, indexed by superscript n. The inter-arrival times for the nth system are $\{u^n_j\}_{j=1}^{\infty}$ and the service times are $\{v^n_j\}_{j=1}^{\infty}$. The jth customer arrives at the nth system with an initial lead time (i.e., the time between the arrival and the deadline for completion of service for that customer)

$$L^n_j = Mc_n(v^n_j \wedge M). \tag{19}$$

Note that

$$G_M^n(y) \stackrel{\Delta}{=} \mathbb{P}\{L_j^n \le c_n y\} = \begin{cases} G^n(y/M), & y < M^2, \\ 1 & y \ge M^2, \end{cases} \tag{20}$$

$$G_{v,M}^n(y) \stackrel{\Delta}{=} \mathbb{E}\left[v_j^n \mathbb{I}_{\{L_j^n \le c_n y\}}\right] = \begin{cases} G_v^n(y/M), & y < M^2, \\ 1/\mu_n & y \ge M^2. \end{cases} \tag{21}$$

In the case of $v^* < \infty$, (11) implies that $v_j^n \le v^{**} \stackrel{\Delta}{=} \sup_{n\in\mathbb{N}} v_n^*$ almost surely (a.s.) and hence, for $M \ge v^{**}$ (which will be assumed in the corresponding proof), we have

$$L_j^n = M c_n v_j^n. \tag{22}$$

Denote by $Q_M^n(t)$ the queue length process in the nth EDF system and let

$$\widehat{Q}_M^n(t) = c_n^{-1} Q_M^n(nt).$$

By the above-mentioned SRPT optimality result, $Q^n(t) \le Q_M^n(t)$ for each $t \ge 0$, and thus

$$\widehat{Q}^n(t) \le \widehat{Q}_M^n(t), \qquad t \ge 0. \tag{23}$$

4.2 Auxiliary functions

It is easy to check that the assumption (3) implies uniform integrability of the sequence $\{v_1^n\}_{n=1}^\infty$. This, together with (1)–(2), yields

$$\int_0^\infty \big(1 - G(\eta)\big)\, d\eta = \lim_{n\to\infty} \int_0^\infty \big(1 - G^n(\eta)\big)\, d\eta = \lim_{n\to\infty} \frac{1}{\mu_n} = \frac{1}{\lambda} = 1. \tag{24}$$

For any $y \in \overline{\mathbb{R}}$, define

$$H(y) \stackrel{\Delta}{=} \int_y^\infty \big(1 - G(\eta)\big)\, d\eta.$$

By (24), the function H is finite and, moreover, it is the complementary c.d.f. of the residual lifetime distribution

$$G_e(x) = \int_0^x \big(1 - G(\eta)\big)\, d\eta, \qquad x \ge 0,$$

corresponding to the limiting service time distribution G.

For $y \in \mathbb{R}$, let

$$G_M(y) = \begin{cases} G(y/M), & y < M^2, \\ 1 & y \ge M^2, \end{cases} \qquad G_{v,M}(y) = \begin{cases} G_v(y/M), & y < M^2, \\ 1/\lambda & y \ge M^2. \end{cases}$$

The assumptions (1)–(3) and the formulae (20)–(21) imply that

$$G_M^n \Rightarrow G_M, \qquad G_{v,M}^n \Rightarrow G_{v,M}. \tag{25}$$

For $y \in \overline{\mathbb{R}}$, define

$$\begin{aligned} H^M(y) &= \int_y^\infty (1 - G_M(\eta))\, d\eta = \int_{y\wedge M^2}^{M^2} (1 - G(\eta/M))\, d\eta \\ &= M \int_{\frac{y}{M}\wedge M}^{M} (1 - G(\eta))\, d\eta = M\left[H(y/M) - H(M)\right]^+. \end{aligned} \tag{26}$$

Note that in the case of $v^* < \infty$ and $M \geq v^*$, the formula (26) simplifies to

$$H^M(y) = MH(y/M). \tag{27}$$

Similarly, let

$$H_v^M(y) = \int_y^\infty (1 - G_{v,M}(\eta))\, d\eta = M \int_{\frac{y}{M} \wedge M}^M (1 - G_v(\eta))\, d\eta. \tag{28}$$

If for some (and hence all) $y \in \mathbb{R}$, we have

$$H_v(y) \stackrel{\Delta}{=} \int_y^\infty (1 - G_v(\eta))\ d\eta < \infty, \tag{29}$$

then the function H_v^M may be defined by the formula

$$H_v^M(y) = M\left[H_v(y/M) - H_v(M)\right]^+,$$

analogous to (26). In particular, if $v^* < \infty$ and $M \geq v^*$, (28) simplifies to

$$H_v^M(y) = MH_v(y/M). \tag{30}$$

In general, the function H_v may be identically equal to ∞, for example, if the service time distribution $G^n \equiv G$ does not vary with n and it has sufficiently heavy tails, implying $\mathbb{E}(v_j^n)^2 = \infty$. In any case, however, under the assumptions of either Theorem 1, or Theorem 2, the function H_v^M defined by (28) maps $(-\infty, v^* \wedge M^2]$ onto $\mathbb{R}_+$ and is strictly decreasing and continuous. Therefore, there exists a continuous, strictly decreasing inverse function $(H_v^M)^{-1}$ mapping $\mathbb{R}_+$ onto $(-\infty, v^* \wedge M^2]$.

4.3 Heavy traffic limits for the EDF systems

Let $y_{n,M}^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G_M^n(y) = 1\}$, $y_M^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G_M(y) = 1\}$. Under the assumptions of Theorem 1, by (20), (22) and (11), as $n \to \infty$,

$$y_{n,M}^* = Mv_n^* \to Mv^* = y_M^*, \tag{31}$$

while under the assumptions of Theorem 2, by (20), (2) and (12), for n large enough,

$$y_{n,M}^* = M^2 = y_M^*. \tag{32}$$

We claim that under the assumptions of Theorem 1,

$$y_M^* = y_{v,M}^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G_{v,M}(y) = 1/\lambda\}. \tag{33}$$

Indeed, let y, z be the points of continuity of both G_M and $G_{v,M}$ such that $y_M^* \leq y < z$. Then, by (20)–(22) and (25),

$$\begin{aligned} 0 \leq G_{v,M}^n(z) - G_{v,M}^n(y) = \mathbb{E}\left[v_j^n \mathbb{I}_{\{y < Mv_j^n \leq z\}}\right] &\leq \frac{z}{M}\mathbb{P}[y < Mv_j^n \leq z] \\ &= \frac{z}{M}(G_M^n(z) - G_M^n(y)). \end{aligned}$$

Letting $n \to \infty$, we get $0 \leq G_{v,M}(z) - G_{v,M}(y) \leq (z/M)(G_M(z) - G_M(y)) = 0$. Consequently, $G_{v,M} \equiv const$ on $[y_M^*, \infty)$, which implies that $y_{v,M}^* \leq y_M^*$. In order to get

the opposite inequality, let y, z be the points of continuity of both G_M and $G_{v,M}$ such that $y^*_{v,M} \le y < z$. Then, by (20)–(22) and (25),

$$0 \le G^n_M(z) - G^n_M(y) = \mathbb{P}[y < Mv^n_j \le z] \le \frac{M}{y} E\left[v^n_j \mathbb{I}_{\{y<Mv^n_j \le z\}}\right] = \frac{M}{y}(G^n_{v,M}(z) - G^n_{v,M}(y)).$$

Letting $n \to \infty$, we obtain $G_M \equiv const$ on $[y^*_{v,M}, \infty)$, which implies that $y^*_{v,M} \ge y^*_M$, so (33) follows. By a similar argument, (12) implies that

$$G_v(y) < 1/\lambda, \qquad y \in \mathbb{R}, \tag{34}$$

and hence $y^*_{v,M} = M^2$, so (33) holds by (32). Hence, (31)–(33) imply that under the assumptions of either Theorem 1 or Theorem 2, we have

$$\lim_{n\to\infty} y^*_{n,M} = y^*_{v,M}. \tag{35}$$

Let

$$f_M = H^M \circ (H^M_v)^{-1}. \tag{36}$$

By definition, f_M is a continuous, strictly increasing mapping of $\mathbb{R}_+$ onto $\mathbb{R}_+$. Under the assumptions of either Theorem 1 or Theorem 2, (35) and Corollary 6 imply that

$$\widehat{Q}^n_M \Rightarrow Q^*_M \stackrel{\Delta}{=} f_M(W^*), \tag{37}$$

in $D[0, \infty)$ as $n \to \infty$, jointly with (9), where W^* is as in (9).

5 Proofs of the main results

5.1 Proof of Theorem 1

Since $v^* < \infty$ by assumption, for $M \ge v^*$, the formulae (27), (29)–(30) hold, and hence $f_M(x) = MH\left((H_v)^{-1}(x/M)\right)$ for $x \ge 0$. Consequently, our Theorem 1 can be proved like Theorem 1 in Kruk (2019). We recall the corresponding argument below.

First, we will check that for every $x \ge 0$,

$$\lim_{M\to\infty} f_M(x) = \frac{x}{v^*}. \tag{38}$$

Because $f_M(0) = 0$ for every M, we only have to show (38) for $x > 0$. Fix $x > 0$ and let $y_M = (H^M_v)^{-1}(x)$. Then

$$\int_{y_M/M}^{v^*} (1 - G_v(\eta))d\eta = H_v(y_M/M) = \frac{x}{M} \to 0, \qquad M \to \infty,$$

so $y_M/M < v^*$ and

$$\lim_{M\to\infty} \frac{y_M}{M} = v^*. \tag{39}$$

For $z < v^*$ being a point of continuity for both G and G_v we have, by (1)–(3), (11), the fact that $\lambda = 1$ and the Markov inequality,

$$\frac{1 - G_v(z)}{1 - G(z)} = \lim_{n\to\infty} \frac{\frac{1}{\mu_n} - G_v^n(z)}{1 - G^n(z)} = \lim_{n\to\infty} \frac{\mathbb{E}[v_1^n \mathbb{I}_{[z<v_1^n \le v_n^*]}]}{\mathbb{P}[z < v_1^n \le v_n^*]}. \tag{40}$$

However, for each n,

$$\frac{\mathbb{E}[v_1^n \mathbb{I}_{[z<v_1^n \le v_n^*]}]}{\mathbb{P}[z < v_1^n \le v_n^*]} \in [z, v_n^*],$$

so (11) and (40) yield

$$\frac{1 - G_v(z)}{1 - G(z)} \in [z, v^*]. \tag{41}$$

If $z < v^*$ is a point of discontinuity of G or G_v, there exist points $z_n < v^*$, $z_n \downarrow z$ such that both G and G_v are continuous at z_n. Hence, (41) holds at z_n and right-continuity of G, G_v implies

$$z \le \frac{1 - G_v(z)}{1 - G(z)} \le v^*, \qquad 0 < z < v^*. \tag{42}$$

Therefore, by (39) and (42), as $M \to \infty$, we have

$$\frac{H^M(y_M)}{H_v^M(y_M)} = \frac{\int_{y_M/M}^{v^*}(1 - G(\eta))d\eta}{\int_{y_M/M}^{v^*}(1 - G_v(\eta))d\eta} \to \frac{1}{v^*}.$$

Consequently, as $M \to \infty$,

$$f_M(x) = H^M(y_M) = H_v^M(y_M)\frac{H^M(y_M)}{H_v^M(y_M)} = x\frac{H^M(y_M)}{H_v^M(y_M)} \to \frac{x}{v^*}.$$

We have justified (38). The functions $f_M(x)$, x/v^* are continuous and increasing, so it is easy to verify that the convergence (38) is actually uniform on compact subsets of $\mathbb{R}_+$ (see, e.g., the proof of Proposition 3.4 in Doytchinov et al. (2001) for a similar reasoning).

Let us choose $T > 0$ and $\epsilon > 0$. Take N large enough to assure that

$$\mathbb{P}\left[\max_{0\le t\le T} W^*(t) \le N\right] \ge 1 - \frac{\epsilon}{2}. \tag{43}$$

Next, choose M so large that $\sup_{0\le x\le N} |f_M(x) - x/v^*| \le \epsilon/2$. Consequently, by (43), we have

$$\mathbb{P}\left[\max_{0\le t\le T} \left| f_M(W^*(t)) - \frac{1}{v^*}W^*(t)\right| \le \frac{\epsilon}{2}\right] \ge 1 - \frac{\epsilon}{2}. \tag{44}$$

Using (9), (37) and the Skorokhod representation theorem (see, e.g., Billingsley 1999, Theorem 6.7), we can construct the model primitives u_j^n and v_j^n for $j \in \mathbb{N}$, $n \in \mathbb{N}$, on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the sequences of processes $\widehat{W}^n$, $\widehat{Q}_M^n$, $n \in \mathbb{N}$, and the process W^* are defined on this space and

$$\widehat{W}^n \to W^*, \qquad \widehat{Q}_M^n \to f_M(W^*) \tag{45}$$

almost surely (a.s.). Here each a.s. convergence is in the J_1 topology on $D[0,\infty)$ and since the limits are continuous, this is equivalent to uniform convergence on compact intervals. Hence, for n large enough, we have

$$\mathbb{P}\left[\max_{0\le t\le T}\left|\widehat{Q}_M^n(t)-f_M(W^*(t))\right|\le\frac{\epsilon}{2}\right]\ge 1-\frac{\epsilon}{2}.$$

This, together with (44), yields

$$\mathbb{P}\left[\max_{0\le t\le T}\left|\widehat{Q}_M^n(t)-\frac{1}{v^*}W^*(t)\right|\le\epsilon\right]\ge 1-\epsilon \tag{46}$$

for n large enough. From (23) and (46), we get

$$\mathbb{P}\left[\widehat{Q}^n(t)\le\frac{1}{v^*}W^*(t)+\epsilon\quad\forall t\in[0,T]\right]\ge 1-\epsilon. \tag{47}$$

On the other hand, (11) implies that for each $n\in\mathbb{N}$, we have $W^n(t)\le v_n^*Q^n(t)$ (and hence $\widehat{W}^n(t)\le v_n^*\widehat{Q}^n(t)$) for all $t\ge 0$ almost surely. Thus, by (45), for n large enough,

$$\mathbb{P}\left[\widehat{Q}^n(t)\ge\frac{1}{v_n^*}W^*(t)-\epsilon\quad\forall t\in[0,T]\right]\ge 1-\epsilon. \tag{48}$$

Finally, the relations (11) and (47)–(48) imply Theorem 1.

5.2 Proof of Theorem 2

We follow the lines of the proof of Theorem 2 in Kruk (2019).

First, we will check that for every $x\ge 0$,

$$\lim_{M\to\infty}f_M(x)=0. \tag{49}$$

By (26), (28) and (36), we have $f_M(0)=H^M((H_v^M)^{-1}(0))=H^M(M^2)=0$ for each M, so it is sufficient to show (49) for $x>0$. Fix $x>0$ and let $y_M=(H_v^M)^{-1}(x)$. By (28), we have

$$\int_{y_M/M}^{M}(1-G_v(\eta))d\eta=\frac{x}{M}\to 0.\qquad M\to\infty. \tag{50}$$

We claim that

$$\lim_{M\to\infty}\frac{y_M}{M}=\infty. \tag{51}$$

Indeed, suppose, to the converse, that for some sequence $M_k\to\infty$ as $k\to\infty$ and some constant $C<\infty$, we have $y_{M_k}/M_k\le C$ for all k. Then for k sufficiently large, (34), (50) and the assumption $\lambda=1$ imply that

$$\frac{x}{M_k}\ge\int_C^{M_k}(1-G_v(\eta))d\eta\ge\int_C^{2C}(1-G_v(\eta))d\eta>0,$$

which contradicts the convergence $M_k\to\infty$. We have proved (51).

For z being a point of continuity for both G and G_v, by (2)–(3), the fact that $\lambda=1$ and the Markov inequality, we have

$$\frac{1-G_v(z)}{1-G(z)}=\lim_{n\to\infty}\frac{\frac{1}{\mu_n}-G_v^n(z)}{1-G^n(z)}=\lim_{n\to\infty}\frac{\mathbb{E}[v_1^n\mathbb{I}_{[v_1^n>z]}]}{\mathbb{P}[v_1^n>z]}\ge z. \tag{52}$$

If z is a point of discontinuity of G or G_v, there exist points $z_n \downarrow z$ such that both G and G_v are continuous at z_n. Consequently, (52) holds at z_n and right-continuity of G, G_v implies

$$\frac{1-G_v(z)}{1-G(z)} \geq z, \qquad z > 0. \tag{53}$$

Thus, by (51) and (53), as $M \to \infty$, we have

$$\begin{aligned}\frac{H^M(y_M)}{H_v^M(y_M)} &= \frac{\int_{y_M/M}^M (1-G(\eta))d\eta}{\int_{y_M/M}^M (1-G_v(\eta))d\eta} \leq \frac{\int_{y_M/M}^M (1-G(\eta))d\eta}{\int_{y_M/M}^M \eta(1-G(\eta))d\eta} \\ &\leq \frac{\int_{y_M/M}^M (1-G(\eta))d\eta}{\frac{y_M}{M}\int_{y_M/M}^M (1-G(\eta))d\eta} = \frac{M}{y_M} \to 0,\end{aligned}$$

and hence as $M \to \infty$,

$$f_M(x) = H^M(y_M) = H_v^M(y_M)\frac{H^M(y_M)}{H_v^M(y_M)} = x\frac{H^M(y_M)}{H_v^M(y_M)} \to 0.$$

We have proved (49).

Let $T > 0$ and $\epsilon > 0$. Take N large enough to assure (43). By (49), for M sufficiently large, we have $f_M(N) \leq \epsilon/2$. Hence, by (37), (43) and monotonicity of the function f_M, we get

$$\mathbb{P}\left[\max_{0\leq t\leq T} \widehat{Q}_M^n(t) \leq \epsilon\right] \geq 1-\epsilon$$

for large n. This, together with (23) and nonnegativity of the process $\widehat{Q}^n$, proves Theorem 2.

5.3 Proof of Corollary 1

We have $w_j^n(t) \leq v_j^n$, so (11) implies that $\widehat{\mathcal{Q}}^n(t)(v_n^*, \infty) = 0$ for every $n \in \mathbb{N}$, $t \geq 0$. Hence, again by (11), for any $x > v^*$ and n large enough, we have $\widehat{\mathcal{Q}}^n(t)(x, \infty) = 0$, $t \geq 0$. By Theorem 1, the total mass $\widehat{Q}^n(t)$ of the random measure $\widehat{\mathcal{Q}}^n(t)$ is convergent to $\frac{1}{v^*}W^*(t)$ in $D[0,\infty)$. Consequently, in order to show Corollary 1, it suffices to check that for each $x \in (0, v^*)$,

$$\widehat{\mathcal{Q}}^n(t)[0,x) \Rightarrow 0, \qquad n \to \infty. \tag{54}$$

Let $x \in (0, v^*)$. Then

$$\begin{aligned}\widehat{W}^n(t) &= \langle e, \widehat{\mathcal{Q}}^n(t)\rangle \ \leq x\ \widehat{\mathcal{Q}}^n(t)[0,x) + v_n^*\widehat{\mathcal{Q}}^n(t)[x, v_n^*] \\ &= v_n^*\widehat{Q}^n(t) - (v_n^* - x)\widehat{\mathcal{Q}}^n(t)[0,x).\end{aligned}$$

From this, we get,

$$\widehat{\mathcal{Q}}^n(t)[0,x) \leq \frac{1}{v_n^* - x}\left(v_n^*\widehat{Q}^n(t) - \widehat{W}^n(t)\right). \tag{55}$$

The right-hand side of (55) converges weakly to zero by (11) and Theorem 1, proving (54).

6 Limiting distributions for LRPT

In this section we assume that jobs are served according to the LRPT protocol. We also assume (1)–(3), (6), (8) and the following variant of (10). There exists a dense subset C of $\overline{\mathbb{R}}$, containing ∞, such that for all $a < b$, $a, b \in C$, we have

$$\tilde{V}^n_{a-,b-}(t) \stackrel{\Delta}{=} \frac{1}{c_n} \sum_{j=1}^{\lfloor nt \rfloor} \left(v_j^n \mathbb{I}_{\{a \le v_j^n < b\}} - (G_v^n(b-) - G_v^n(a-)) \right) \Rightarrow V^*_{a-,b-}(t) \tag{56}$$

in $D[0,\infty)$, where $V^*_{a-,b-}$ has continuous sample paths. Let $v_* \stackrel{\Delta}{=} \inf\{y \in \mathbb{R} : G(y) > 0\}$. If $v_* > 0$, then we additionally assume that as $n \to \infty$,

$$v^n_* \stackrel{\Delta}{=} \inf\{y \in \mathbb{R} : G^n(y) > 0\} \to v_*. \tag{57}$$

In the case of LRPT, we have the following two theorems.

Theorem 5 *If $v_* > 0$ and (57) holds, then $\widehat{Q}^n \Rightarrow \frac{1}{v_*} W^*$ in $D[0,\infty)$ jointly with (9) as $n \to \infty$.*

Theorem 6 *If $v_* = 0$, then for every fixed $t > 0$ $\widehat{Q}^n(t) \Rightarrow \infty$ as $n \to \infty$.*

Let us remark that it is not possible to generalize Theorem 6 to convergence in $D_{\overline{\mathbb{R}}}[0,\infty)$. Indeed, for each n we have $Q^n(0) = 0$ a.s.. Furthermore, if $\rho_n < 1$, $0 < c < T$ and if n is large, then a "typical" sample path of the process Q^n hits zero at some time $t \in [nc, nT]$. Hence, for any $0 < c < T$, the convergence $\widehat{Q}^n \Rightarrow \infty$, $n \to \infty$, cannot hold in $D_{\overline{\mathbb{R}}}[c,T]$.

Let the measure-valued processes $\mathcal{Q}^n$ and $\widehat{\mathcal{Q}}^n$ be defined by (13). Theorem 5 has the following

Corollary 5 *Under the assumptions of Theorem 5 we have $\widehat{\mathcal{Q}}^n \Rightarrow \frac{1}{v_*} W^* \delta_{v_*}$ in $D_{\mathcal{M}}[0,\infty)$ as $n \to \infty$.*

The proofs of Theorems 5, 6 and Corollary 5 are similar to the proofs of Theorems 1, 2 and Corollary 1. The most important difference is that in the case of LRPT, we use approximating single server, single customer class EDF systems with initial lead times equal to $-Mc_n(v_j^n \wedge M)$ (or just $-Mc_n v_j^n$) instead of $Mc_n(v_j^n \wedge M)$. We omit the details.

As in the case of SRPT, if the service time distribution $G^n \equiv G$ does not depend on n, then the assumptions for the above results can be simplified. In fact, in this case only (6), (8), (14) and (56) for $a < b$, $a, b \in C$, need to be assumed for Theorems 5, 6 and Corollary 5 to hold.

The results of this section extend their counterparts from Section 5 of Kruk (2019), just like the results of Sect. 3.1 extend the main results of Kruk (2019) (see our Sect. 3.2). Indeed, as we have already seen, the heavy traffic regime assumed in Kruk (2019) is a special case of the one considered in this paper. Moreover, our assumption (57) is less restrictive than the assumption (9) from Kruk (2019), stating that (in the notation of (57)) $v^n_* = v_*$ for each $n \in \mathbb{N}$.

Acknowledgements The author thanks the referees for their helpful remarks and suggestions for improving the paper.

7 Appendix: Heavy traffic limits for EDF queues

In this Appendix, we state Theorem 7, a variant of Theorem A.2 in Kruk (2007) (extending the results of Doytchinov et al. 2001), which is suitable for the applications described in this paper. The point here is to allow the right endpoint of the rescaled lead time distributions G^n to vary with n in a controlled way, see (58)–(59), (63) and (66), to follow. Note that in Doytchinov et al. (2001), it is assumed that the initial lead time distribution $G^n \equiv G$ does not vary with n, while in Kruk (2007), initial lead time distributions G^n depending on n and satisfying (2) are allowed, but their right endpoints of support y_n^*, $n \in \mathbb{N}$, are assumed to coincide with y^*, the right endpoint of support of the limiting distribution G.

Consider a sequence of single-station queueing systems, indexed by superscript n, each with one customer class. Assume that $\{u_j^n\}_{j=1}^{\infty}$, the customer inter-arrival times, are strictly positive, identically distributed r.v.s with mean $1/\lambda_n$ and $\{v_j^n\}_{j=1}^{\infty}$, the customer service times, are strictly positive r.v.s with mean $1/\mu_n$. We assume that each queue is empty at time zero and (1) holds. Let c_n be a sequence of constants such that $c_n \to \infty$, $n/c_n \to \infty$. Let L_j^n denote the customer initial lead times, with distribution given by

$$\mathbb{P}\{L_j^n \le c_n y\} = G^n(y), \qquad y \in \mathbb{R}, \qquad j, n \ge 1. \tag{58}$$

Let

$$y_n^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G^n(y) = 1\}, \qquad n \ge 1. \tag{59}$$

We assume that

$$\sup_{n\in\mathbb{N}} y_n^* < \infty, \tag{60}$$

and that (2) holds for some c.d.f. G. By (2) and (60), we have

$$y^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G(y) = 1\} \le \liminf_{n\to\infty} y_n^* \le \sup_{n\in\mathbb{N}} y_n^* < \infty. \tag{61}$$

Next, we assume that the random vectors $\left\{\left(v_j^n, L_j^n\right)\right\}_{j=1}^{\infty}$ are identically distributed and that

$$G_v^n(y) \stackrel{\Delta}{=} \mathbb{E}\left[v_j^n \mathbb{I}_{\{L_j^n \le c_n y\}}\right] \Rightarrow G_v(y), \tag{62}$$

where G_v is a c.d.f. of a finite, positive measure on $\mathbb{R}$ with total mass $1/\lambda$. Let

$$y_v^* \stackrel{\Delta}{=} \min\{y \in \mathbb{R} : G_v(y) = 1/\lambda\}. \tag{63}$$

By (2) and (61)–(62), we have $y_v^* \le y^*$.

We make the heavy traffic assumption (6) for some $\gamma \in \mathbb{R}$, where $\rho_n \stackrel{\Delta}{=} \lambda_n/\mu_n$. Let S^n, V^n be defined by (4)–(5). We assume that (8) holds in $D_{\mathbb{R}^2}[0,\infty)$, where the processes $\widehat{S}^n$, $\widehat{V}^n$ are defined by (7) and their limits S^*, V^* have continuous sample paths. Let $W^n(t)$, $t \ge 0$, be the workload process in the nth system. By Theorem 9.3.4 in Whitt (2002), $\widehat{W}^n(t) \stackrel{\Delta}{=}$

$c_n^{-1} W^n(nt) \Rightarrow W^*(t)$ in $D[0,\infty)$, where, for any $t \geq 0$, $W^*(t) = N^*(t) - \inf_{0\leq s\leq t} N^*(s)$ and $N^* = (V^* - S^*) \circ \lambda e - \gamma e$. Finally, we assume that there exists a dense subset C of $\mathbb{R} \cup \{-\infty\}$, containing $-\infty$, such that for every $a < b$, $a, b \in C$, we have

$$\tilde{V}^n_{a,b}(t) \stackrel{\Delta}{=} \frac{1}{c_n} \sum_{j=1}^{\lfloor nt \rfloor} \left(v_j^n \mathbb{I}_{\{c_n a < L_j^n \leq c_n b\}} - (G_v^n(b) - G_v^n(a)) \right) \Rightarrow V^*_{a,b}(t) \tag{64}$$

in $D[0,\infty)$, where $V^*_{a,b}$ has continuous sample paths.

For $y \in \mathbb{R}$, let $H_v(y) \stackrel{\Delta}{=} \int_y^{y_v^*} (1-\lambda G_v(\eta)) d\eta = \int_y^{\infty} (1-\lambda G_v(\eta)) d\eta$. The function H_v maps $(-\infty, y_v^*]$ onto $\mathbb{R}_+$ and is continuous and strictly decreasing on $(-\infty, y_v^*]$. Thus, there exists a continuous inverse function H_v^{-1} that maps $\mathbb{R}_+$ onto $(-\infty, y_v^*]$. Let $F^*(t) \stackrel{\Delta}{=} H_v^{-1}(W^*(t))$, $t \geq 0$. For any Borel set $B \subseteq \mathbb{R}$, let

$$\mathcal{Q}^n(t)(B) \stackrel{\Delta}{=} \left\{ \begin{array}{l} \text{Number of customers in the queue at time } t \\ \text{having lead times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\},$$

$$\mathcal{W}^n(t)(B) \stackrel{\Delta}{=} \left\{ \begin{array}{l} \text{Work in the queue at time } t \text{ associated with customers} \\ \text{in this queue having lead times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\},$$

$$\widehat{\mathcal{Q}}^n(t)(B) \stackrel{\Delta}{=} c_n^{-1} \mathcal{Q}^n(nt)(c_n B), \qquad \widehat{\mathcal{W}}^n(t)(B) \stackrel{\Delta}{=} c_n^{-1} \mathcal{W}^n(nt)(c_n B).$$

Let $\mathcal{W}^*$ and $\mathcal{Q}^*$ be measure-valued processes defined by

$$\mathcal{W}^*(t)(B) \stackrel{\Delta}{=} \int_{B \cap [F^*(t),\infty)} \left(1 - \lambda\, G_v(\eta)\right) d\eta,$$

$$\mathcal{Q}^*(t)(B) \stackrel{\Delta}{=} \lambda \int_{B \cap [F^*(t),\infty)} \left(1 - G(\eta)\right) d\eta,$$

for all Borel sets $B \subseteq \mathbb{R}$.

Theorem 7 *Under the preemptive EDF service protocol, subject to the assumptions made above, we have*

$$\widehat{\mathcal{W}}^n \Rightarrow \mathcal{W}^* \tag{65}$$

in $D_{\mathcal{M}}[0,\infty)$ as $n \to \infty$. If, additionally,

$$\lim_{n\to\infty} y_n^* = y_v^*, \tag{66}$$

then $\widehat{\mathcal{Q}}^n \Rightarrow \mathcal{Q}^*$ in $D_{\mathcal{M}}[0,\infty)$ as $n \to \infty$, jointly with (65).

Theorem 7 can be proved by a suitable generalization of the arguments of Doytchinov et al. (2001).

For $y \in \mathbb{R}$, define $H(y) \stackrel{\Delta}{=} \lambda \int_y^{y^*} (1 - G(\eta)) d\eta = \lambda \int_y^{\infty} (1 - G(\eta)) d\eta$. Let $Q^n(t) = \mathcal{Q}^n(t)(\mathbb{R})$ denote the queue length at time t in the nth system and let $\widehat{Q}^n(t) \stackrel{\Delta}{=} c_n^{-1} Q^n(nt) = \widehat{\mathcal{Q}}^n(t)(\mathbb{R})$. From Theorem 7, we immediately have

Corollary 6 *Under the preemptive EDF service protocol, subject to all the assumptions of this section, including (66), $\widehat{Q}^n \Rightarrow Q^* = H(F^*) = H(H_v^{-1}(W^*(t)))$ in $D[0,\infty)$ as $n \to \infty$, jointly with (9).*

It is plausible that the condition (66) in Theorem 7 and Corollary 6 may be replaced by a suitable assumption on the rate of convergence of $G^n(y_v^*)$ to 1 (implying, in particular, that $y_v^* = y^*$), or on the rate of convergence of $G_v^n(y^*)$ to $1/\lambda$. However, we have not attempted to obtain such a generalization.

8 Conclusion

We have provided a derivation of heavy traffic limit theorems for single server, single customer class SRPT and LRPT queues from the corresponding heavy traffic limits for EDF queueing systems. A crucial assumption in our analysis, namely weak convergence of the rescaled stochastic primitive processes to continuous limits, is satisfied in many cases of interest, including the G/G/1 case with finite second moments considered in Kruk (2019) and processes exhibiting short- or long-range dependence. We have also presented a variant of a heavy traffic limit theorem for single server EDF queues, suitable for the applications considered in this paper.

References

Banerjee, S., Budhiraja, A., & Puha, A. L. (2020). Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions. arXiv: 2003.03655v1.

Bansal, N., & Harchol-Balter, M. (2001). Analysis of SRPT scheduling: Investigating unfairness. *ACM Sigmetrics Performance Evaluation Review*, *29*, 279–290.

Bender, M., Chakrabarti, S., & Muthukrishnan, S. (1998). Flow and stretch metrics for scheduling continuous job streams. In *proceedings of the 9th annual ACM-SIAM symposium on discrete algorithms*.

Billingsley, P. (1999). *Convergence of probability measures* (2nd ed.). New York: Wiley.

Billingsley, P. (1986). *Probability and measure* (2nd ed.). New York: Wiley.

Down, D. G., Gromoll, H. C., & Puha, A. L. (2009). Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research*, *34*, 880–911.

Down, D. G., & Wu, R. (2006). Multi-layered round robin routing for parallel servers. *Queueing Systems*, *53*, 177–188.

Doytchinov, B., Lehoczky, J. P., & Shreve, S. E. (2001). Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Annals of Applied Probability*, *11*, 332–378.

Gromoll, H. C., & Keutel, M. (2012). Invariance of fluid limits for the shortest remaining processing time and shortest job first policies. *Queueing Systems*, *70*, 145–164.

Gromoll, H. C., Kruk, Ł., & Puha, A. L. (2011). Diffusion limits for shortest remaining processing time queues. *Stochastic Systems*, *1*(1), 1–16.

Ethier, S. N., & Kurtz, T. G. (1985). *Markov processes: Characterization and convergence*. New York: Wiley.

Kittsteiner, T., & Moldovanu, B. (2005). Priority auctions and queue disciplines that depend on processing time. *Management Science*, *51*, 236–248.

Kruk, Ł. (2007). Diffusion approximation for a G/G/1 EDF queue with unbounded lead times. *Annales UMCS Mathematica A*, *61*, 51–90.

Kruk, Ł. (2019). Diffusion limits for SRPT and LRPT queues via EDF approximations. In T. Phung-Duc, S. Kasahara, S. Wittevrongel (Eds.), *Queueing theory and network applications. QTNA 2019. Lecture notes in computer science* (Vol. 11688, pp. 263–275). Cham: Springer.

Kruk, Ł., & Sokołowska, E. (2016). Fluid limits for multiple-input shortest remaining processing time queues. *Mathematics of Operations Research*, *41*(3), 1055–1092.

Núñez-Queija, R. (2002). Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research*, *113*, 101–117.

Nuyens, M., & Zwart, B. (2006). A large deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems*, *54*, 85–97.

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

Pechinkin, A. V. (1986). 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.

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

Prokhorov, Yu. (1956). Convergence of random processes and limit theorems in probability. *Theory of Probability and Applications*, *1*, 157–214.

Puha, A. L. (2015). Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *Annals of Applied Probability*, *25*(6), 3301–3404.

Ramanan, K., & Stolyar, A. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Annals of Applied Probability*, *11*, 1–48.

Schassberger, R. (1990). 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.

Schrage, L. E. (1968). A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, *16*, 687–690.

Schrage, L. E., & Miller, L. W. (1966). The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, *14*, 670–684.

Whitt, W. (2002). *Stochastic-process limits*. New York: Springer.

Wierman, A., & Harchol-Balter, M. (2003). 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* (pp. 238–249).