DOI: 10.15439/2023F4407
ISSN 2300-5963 ACSIS, Vol. 35

Diffusion Limits for Shortest Remaining Processing Time Queues with Multiple Customer Types

Robert Gieroba
0000-0001-6419-3209
Maria Curie-Skłodowska University
in Lublin
Pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland
Email: robert.gieroba@mail.umcs.pl

Łukasz Kruk
0000-0002-3073-959X
Maria Curie-Skłodowska University
in Lublin
Pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland
Email: lukasz.kruk@mail.umcs.pl

***Abstract*—We consider a single-server queueing system with multiple customer types having bounded processing times in which users are scheduled according to the Shortest Remaining Processing Time (SRPT) discipline, with First In First Out (FIFO) as the tie-breaker. We assume that the processing times of jobs arriving in the system are bounded. We use probabilistic methods to find, under typical heavy traffic assumptions, a suitable approximation of the workload and queue length processes after a long time has passed and show that these processes are divided among the customer classes according to specific proportions, depending on their arrival rates and distributions of initial service times. Our results are confirmed by simulations.**

***Index terms*—Queueing systems, shortest remaining processing time, heavy traffic, diffusion approximations, multiple customer classes.**

I. Introduction

THE SHORTEST Remaining Processing Time (SRPT) service protocol assigns preemptive priority to the task with the smallest residual service time. It started gaining interest after Schrage had proved in [24] that it minimized the number of jobs in a single-server system. SRPT is also well known in queueing theory for minimizing the mean response time (Schrage and Miller [18]). Since then, its properties have been widely studied. Schreiber provides a summary of early work on SRPT in [25]. More recent research of this protocol includes investigating fairness (e.g., Wierman and Harchol-Balter [27]) or tail behavior (Núñez-Queija [19], Nuyens and Zwart [20]). In [12], Gromoll, Kruk and Puha prove, under typical heavy traffic assumptions, a diffusion limit theorem for a measure-valued state descriptor. Another approach, approximating SRPT by the Earliest Deadline First (EDF), leads to the same result (with a slight loss of generality) in [16]. The follow-up research on this topic includes obtaining diffusion limits under nonstandard spatial scaling by Puha [23] and limits for queues with heavy tailed service time distributions by Banerjee, Budhiraja and Puha in [4]. Moreover, Atar, Biswas, Kaspi and Ramanan presented in [3] a unified framework for analyzing single server queueing systems under various service protocols, including SRPT.

Some authors discussed the possibility of implementing the SRPT policy in practice. The main factor impeding it is its unfairness, which could potentially lead to a few tasks having much greater response times (so-called "starving"). However, it has been noted in many works that large jobs are only negligibly penalized at most. For example, Agrawal, Bansal, Harchol-Balter and Schroeder in [1] and [2] propose a method of improving the performance of Web servers by implementing SRPT-based scheduling. Another study by Harchol-Balter and Schroeder ([14]) presents the possibility of a great improvement of the performance of a Web server by changing the traditional fair scheduling policy to SRPT. There are also more recent papers concerning this topic. For example, [6] describes a way of improving the default Linux scheduler by using the existing CFS (Completely Fair Scheduler) and FIFO schedulers to approximate SRPT.

Recent studies suggest that the SRPT protocol performs well in the case of multiserver systems with a single queue. Grosof, Scully and Harchol-Balter proved in [13] that the mean response in the M/G/k queue under the SRPT discipline is asymptotically optimal in the heavy-traffic limit. Dong and Ibrahim ([8]) considered the multiserver M/G/k+G queue with impatient customers with the SRPT protocol and showed that in this setting, SRPT asymptotically maximizes the system throughput among all scheduling disciplines. However, it was also shown that the SRPT protocol can behave suboptimally in multiclass queueing networks to the extent of rendering the queueing system unstable ([7]).

Another related field of research involves studying resource sharing networks under the SRPT service discipline. In [10], [11], a notion of minimality, related to maximizing the corresponding cumulative transmission time with respect to jobs with residual service time not greater than a given threshold, was introduced and it was shown that SRPT is minimal in this sense. Moreover, in [11], another optimality criterion, local edge minimality, was proposed and it was proved that it characterizes a certain subclass of SRPT disciplines, named strong SRPT protocols.

This paper focuses on a single-server queueing system with multiple customer classes. Previous research on this topic includes the work of Peterson ([21]) concerning heavy-traffic limit theorems for queueing networks with multiple customers classes divided into two types: high-priority ones having a preemptive priority over low-priority ones, with customers within each of these types served according to the FIFO policy. In [17], Kruk and Sokołowska establish a fluid limit theorem for a single-server queueing model with K classes of customers, served according to the SRPT protocol, with FIFO used as the tie-breaker.

In this paper, we extend the analysis of [12] and prove a diffusion limit theorem for a multidimensional measure-valued state descriptor $\mathcal{Z}(t)$ defined in Section II, under usual heavy traffic assumptions. More precisely, we first describe a stochastic model for a single-server queuing system with multiple customer classes. We focus on the case of bounded service times of jobs arriving in the system. In order to obtain a diffusion limit, we consider a sequence of such models and apply diffusion scaling as in (1) to follow. We make typical heavy traffic assumptions, as detailed in Section III. The main results of this paper are Theorems 1 and 2. They enable us to easily obtain results for the corresponding workload and queue length processes. This method gives us a way to approximately predict the proportions between the workloads (and queue lengths) of the customer classes in the long run.

The paper is organized as follows. In the second section, we present the mathematical model of the queueing system outlined above and introduce stochastic processes describing its state. In the third section, we introduce a sequence of such models and describe the necessary assumptions. In Sections IV and V, we state and prove the main theorems of this paper. In Section VI, we provide a brief overview of computer simulations illustrating our results.

A. Notation

Let $\mathbb{N}$ denote the set of positive integers, let $\mathbb{R}$ denote the set of real numbers and let $\mathbb{R}_+ = [0, +\infty)$. For $a, b \in \mathbb{R}$, we write $a \vee b$ ($a \wedge b$) for the maximum (minimum) of a and b and $\lfloor a \rfloor$ for the largest integer not greater than a. By convention, a sum over the empty set of indices equals zero. The sets (a, b), $[a, b)$ and $(a, b]$ are empty for $a, b \in [0, \infty]$ with $a \geq b$. The Borel σ-field on $\mathbb{R}_+$ will be denoted by $\mathcal{B}(\mathbb{R}_+)$. For $B \in \mathcal{B}(\mathbb{R}_+)$, we denote the indicator of the set B by $\mathbb{I}_B$. We also define the function $\chi(x) = x,\ x \in \mathbb{R}_+$. For a function $g : \mathbb{R}_+ \to \mathbb{R}$ and $T > 0$, let $\|g\|_T = \sup\{|g(t)| : 0 \leq t \leq T\}$ and $\|g\|_\infty = \sup\{|g(t)| : t \geq 0\}$.

For a vector $a = (a_1, ..., a_K)$, with either real or measure-valued elements, by a_Σ we denote $\sum_{i=1}^K a_i$, unless stated otherwise, where it is a weighted sum. The same notation is used for vector-valued processes.

Let $\mathbf{M}$ denote the set of finite, nonnegative measures on $\mathcal{B}(\mathbb{R}_+)$. When $\mu \in \mathbf{M}$ and $a, b \in \mathbb{R}_+ \cup \{+\infty\}$, we will simply write $\mu(a, b)$, $\mu[a, b)$, $\mu(a, b]$ instead of $\mu((a, b))$, $\mu([a, b))$, $\mu((a, b])$, respectively. Moreover, we will write $\mu(x)$ instead of $\mu(\{x\})$ to denote the measure of a single-element set $\{x\}$. 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 bounded, continuous real functions g on $\mathbb{R}_+$. With this topology, $\mathbf{M}$ is a Polish space ([22]). We denote the zero measure in $\mathbf{M}$ by $\mathbf{0}$ and the measure in $\mathbf{M}$ that puts one unit of mass at a 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. Let $\mathbf{M}_0$ denote the set of those elements of $\mathbf{M}$ that do not charge the origin and have a finite first moment.

We use " $\overset{d}{=}$ " for equality in distribution, " $\xrightarrow{fd}$ " to denote the convergence of finite-dimensional distributions of stochastic processes and " $\Rightarrow$ " to denote convergence in distribution of random elements of a metric space. 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 metric d ([9]). If $\mathcal{S} = \mathbb{R}$, we write $\mathbf{D}[0, \infty)$ instead of $\mathbf{D}([0, \infty), \mathbb{R})$. For $x \in \mathbf{D}([0, \infty), \mathbb{R}^n)$ and $t > 0$, define $x(t-) = \lim\limits_{s \to t^-} x(s)$.

II. Stochastic model for an SRPT queue

The queueing model considered here consists of one server and K job types. Let $\mathcal{K} = \{1, ..., K\}$. The stochastic model involves a random initial condition $(\mathcal{Z}(0), S^x, x > 0)$, describing the system at time zero, together with a measure-valued state descriptor $\mathcal{Z} = (\mathcal{Z}_k, k \in \mathcal{K})$ specifying the time evolution of the system.

A. Initial condition

The initial condition for each class $k \in \mathcal{K}$ consists of the number $Z_k(0)$ of class k jobs in the queue at time zero, the initial service time for each class k job, and the functions describing the order in which the initial jobs with the same initial service time complete service.

Assume that $Z_k(0)$ is a nonnegative integer-valued and finite almost surely random variable for each $k \in \mathcal{K}$. Initial service times for each class k are given by the sequence $\{\tilde{v}_k^j\}_{j \in \mathbb{N}}$ of strictly positive, finite random variables. The initial job with service time $\tilde{v}_k^j, j \leq Z_k(0)$, is called job j for class k. The state of the system will be described by a counting measure with unit masses at the service times of the jobs present in the system. More formally, for $k \in \mathcal{K}$ we define the initial random measure $\mathcal{Z}_k(0) \in \mathbf{M}$ by

$$\mathcal{Z}_k(0) = \sum_{j=1}^{Z_k(0)} \delta_{\tilde{v}_k^j}.$$

Let $\mathcal{Z}(0) = (\mathcal{Z}_1(0), ..., \mathcal{Z}_K(0))$.

For every $x > 0$ the number of initial jobs with initial service time x that are served to completion once t units of time have been devoted to their service is $\lfloor t/x \rfloor \wedge \sum_{k \in \mathcal{K}} \mathcal{Z}_k(0)(x)$. Here we introduce processes $S_k^x,\ k \in \mathcal{K}$, that dictate how much service is allocated across classes. In particular, let $S_k^x(t)$ be the number of initial class k jobs with initial service time x that have left the system by the time that the server has devoted t units of time solely to serving jobs with this initial service time. We assume that the random functions $S_k^x(t)$ satisfy the following consistency conditions:

1) $S_k^x(0) = 0$,

2) $S_k^x(t)$ is nondecreasing and

$$S_k^x(t) = \mathcal{Z}_k(0)(x),\ t \geq x\mathcal{Z}_\Sigma(0)(x),\ k \in \mathcal{K},$$

3) we have

$$\sum_{k\in\mathcal{K}} S_k^x(t) = \lfloor t/x \rfloor \wedge \mathcal{Z}_\Sigma(0)(x).$$

The system $(\mathcal{Z}(0), S^x, x > 0)$, where $S^x = (S_1^x, ..., S_K^x)$, will be called the initial condition.

B. Stochastic primitives

Let E_k be the exogenous arrival process for class $k \in \mathcal{K}$. For $t \geq 0$, $E_k(t)$ is the number of class k arrivals to the system in the time interval $(0, t]$. For each k it is a (possibly delayed) renewal process with rate $\alpha_k > 0$ such that the interarrival times have standard deviation $a_k \geq 0$. Let $E(t) = (E_1(t), ..., E_K(t))$, $\alpha = (\alpha_1, ..., \alpha_K)$ and $a = (a_1, ..., a_K)$. In particular, α_Σ is the total arrival rate. For $t \geq 0$ and $k \in \mathcal{K}$ let $A_k(t) = Z_k(0) + E_k(t)$ and $A(t) = (A_1(t), ..., A_K(t))$. Then job j of class k arrives at time $T_k^j = \inf\{t \geq 0 : A_k(t) \geq j\}$. For $k \in \mathcal{K}$ and $j \in \mathbb{N}$ a random variable v_k^j represents the service time of the $(Z_k(0) + j)$th job of class k. We assume that the random variables $\{v_k^j\}_{j\in\mathbb{N}}$ are strictly positive and form an independent and identically distributed sequence with common distribution ν_k on $\mathbb{R}_+$ for each $k \in \mathcal{K}$. For $k \in \mathcal{K}$ let the sequences $\{v_k^j\}_{j\in\mathbb{N}}$ be mutually independent. Assuming that the mean $\langle \chi, \nu_k \rangle > 0$ and the standard deviation $b_k = \sqrt{\langle \chi^2, \nu_k \rangle - \langle \chi, \nu_k \rangle^2} \geq 0$ we define $\beta_k = \langle \chi, \nu_k \rangle^{-1}$ for each class. We put $\nu = (\nu_1, ..., \nu_K)$. Define $p_k = \alpha_k/\alpha_\Sigma,\ k \in \mathcal{K}$. Then $\nu_\Sigma = \sum_{k\in\mathcal{K}} p_k \nu_k$ is a mixture of service time distributions. It may be thought of as the distribution of the initial service time of a randomly chosen customer. Let $\rho = \alpha_\Sigma \langle \chi, \nu_\Sigma \rangle$ be the traffic intensity.

It will be convenient to combine the stochastic primitives for job classes into measure-valued load processes

$$\mathcal{V}_k(t) = \sum_{j=1}^{E_k(t)} \delta_{v_k^j},\ t \geq 0,\ k \in \mathcal{K},$$

and $\mathcal{V}(t) = (\mathcal{V}_1(t), ..., \mathcal{V}_K(t))$. Then for each $k \in \mathcal{K}$, $\mathcal{V}_k \in \mathbf{D}([0, \infty), \mathbf{M})$, since $E_k \in \mathbf{D}([0, \infty), \mathbb{R}_+)$.

C. Basic performance processes, service protocol

For $j \in \mathbb{N},\ k \in \mathcal{K}$ and $t \geq T_k^j$ let $w_k^j(t)$ denote the residual service time of job j in class k at time t. Thus, w_k^j decreases at rate one when the job j of class k is in service and is constant otherwise. In particular, w_k^j is identically equal to zero after the departure of the job j from the system.

Customers are served using the SRPT service protocol, which gives preemptive priority to the job with the shortest residual service time. In case of a tie, FIFO is used as a tie-breaking rule. When both the service times and the arrival times of two or more jobs are the same, we break the tie in an arbitrary manner.

The state of the system at time t will be described by a K-dimensional vector of counting measures with unit masses at the residual service times of the jobs of each class still present in the system. More formally, for $t \geq 0$ and $k \in \mathcal{K}$, the state descriptor of class k is defined as follows:

$$\mathcal{Z}_k(t) = \sum_{j=1}^{A_k(t)} \delta^+_{w_k^j(t)}.$$

Put $\mathcal{Z}(t) = (\mathcal{Z}_1(t), ..., \mathcal{Z}_K(t))$. For $t \geq 0$ and $k \in \mathcal{K}$, we define the workload of class k by $W_k(t) = \langle \chi, \mathcal{Z}_k(t) \rangle$ and $W(t) = (W_1(t), ..., W_K(t))$. Let $Q_k(t) = \langle 1, \mathcal{Z}_k(t) \rangle$ denote the number of customers of class $k \in \mathcal{K}$ present in the system at time t and define $Q(t) = (Q_1(t), ..., Q_K(t))$.

III. Diffusion limit

We define a sequence of systems to which the diffusion scaling is applied. 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 II. For each $r \in \mathcal{R}$ there is an initial condition $(\mathcal{Z}^r(0), S^{r,x}, x > 0)$, stochastic primitives E_k^r and $\{v_k^{r,j}\}_{j\in\mathbb{N}}$ with parameters $\alpha_k^r, a_k^r, \nu_k^r, \beta_k^r$, b_k^r, p_k^r and ρ^r, for each class $k \in \mathcal{K}$. We also have arrival processes A^r with arrival times $\{T_k^{r,j}\}_{j\in\mathbb{N}},\ k \in \mathcal{K}$ a state descriptor $\mathcal{Z}^r$ and processes W^r, Q^r. The stochastic elements of each model are defined on a probability space $(\Omega^r, \mathcal{F}^r, \mathbf{P}^r)$ with expectation operator $\mathbb{E}^r$.

A diffusion scaling is applied to each model in the $\mathcal{R}$-indexed sequence as follows. For each $r \in \mathcal{R}$ and $t \geq 0$, let

$$\begin{aligned}
\hat{E}^r(t) &= \frac{1}{r}(E^r(r^2t) - r^2t\alpha^r),\\
\hat{\mathcal{Z}}^r(t) &= \frac{1}{r}\mathcal{Z}^r(r^2t),\\
\hat{W}^r(t) &= \frac{1}{r}W^r(r^2t),\\
\hat{Q}^r(t) &= \frac{1}{r}Q^r(r^2t),\\
\hat{\mathcal{V}}^r(t) &= \frac{1}{r}\left(\mathcal{V}^r(r^2t) - r^2t\alpha^r\nu^r\right).
\end{aligned} \tag{1}$$

A fluid scaling is applied to functions $S^{r,x}$. For each $r \in \mathcal{R}$, $x > 0$ and $t \geq 0$, let

$$\bar{S}^{r,x}(t) = \frac{1}{r}S^{r,x}(rt).$$

Let $\alpha = (\alpha_1, ..., \alpha_K) \in (0, +\infty)^K$, $a = (a_1, ..., a_K) \in (0, +\infty)^K$ and define $\alpha(t) = \alpha t,\ t \geq 0$. Let $\nu = (\nu_1, ..., \nu_K)$ be a vector of probability measures on $\mathbb{R}_+$ such that for each $k \in \mathcal{K}$

$$\nu_k(0) = 0, \qquad \langle \chi, \nu_k \rangle = \frac{1}{\alpha_k}, \qquad 0 < \langle \chi^2, \nu_k \rangle < \infty.$$

We make the following asymptotic assumptions for the sequence of stochastic primitives. Assume that as $r \to \infty$,

$$\alpha^r \to \alpha, \qquad a^r \to a, \qquad \hat{E}^r \Rightarrow E^*, \tag{2}$$

where E^* is a K-dimensional Brownian motion starting from zero with drift zero and covariance matrix $\Sigma = [\sigma_{ij}]$ such that

$\sigma_{kk} = a_k^2\alpha_k^3,\ k \in \{1, ..., K\}$. In particular, if the coordinate processes of E^* are independent, then $\sigma_{ij} = 0,\ i \neq j$. Put

$$b_k = \sqrt{\langle \chi^2, \nu_k\rangle - \langle \chi, \nu_k\rangle^2},\ k \in \mathcal{K},$$

and $b = (b_1, ..., b_K)$. In addition, assume the heavy traffic condition that for some $\gamma \in \mathbb{R}$

$$\lim_{r\to\infty} r(1-\rho^r) = \gamma. \tag{3}$$

For the sequence of service time distributions, we assume that $\nu^r = \nu$, i.e., ν^r does not depend on r. Then $\beta^r = \beta$, $p_k^r \to p_k := \frac{\alpha_k}{\alpha_\Sigma}$ as $r \to \infty$, where α is given by (2), $\rho^r \to 1$ as $r \to \infty$ and $b^r = b$. Define

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

We assume that $x^* < \infty$.

For the sequence of diffusion scaled initial conditions $\{(\mathcal{Z}^r(0), S^{r,x}, x > 0)\}_{r>0}$ we assume that as $r \to \infty$

$$\hat{W}^r_\Sigma(0) \Rightarrow W_0^*, \tag{4}$$

where W_0^* is some random variable. It is well known ([15]) that from (2), (3), (4), the fact that service time distributions ν^r do not depend on r and that SRPT is a work conserving discipline, it follows that as $r \to \infty$

$$\hat{W}^r_\Sigma \Rightarrow W^*_\Sigma, \tag{5}$$

where W^*_Σ is a reflected Brownian motion with initial value $W^*_\Sigma(0) \overset{d}{=} W_0^*$ with drift $-\gamma$ and variance $(a_\Sigma^2 + b_\Sigma^2)\alpha_\Sigma$ per unit time. It can be shown ([12]) that if $x^* < \infty$ then $\hat{Q}^r_\Sigma \Rightarrow Q^*_\Sigma := \frac{W^*_\Sigma}{x^*}$.

Before we proceed, we state a standard result, as presented in [12].

Proposition 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 a probability space $(\Omega^r, \mathcal{F}^r, \mathbf{P}^r)$ with finite mean μ^r and standard deviation σ^r, independent of a (possibly delayed) rate $\alpha^r > 0$ renewal process B^r such that the standard deviation of the interarrival times equals $a^r \geq 0$. Assume that $\hat{B}^r \Rightarrow B^*$ as $r \to \infty$, where*

$$\hat{B}^r(t) = \frac{1}{r}\left(B^r(r^2t) - r^2t\alpha^r\right),\ t \geq 0,$$

and B^* is a one-dimensional Brownian motion starting from zero with drift zero and variance $a^2\alpha^3$ per unit time. Suppose that for some finite nonnegative constants μ, σ, and positive α, a we have that $\mu^r \to \mu$, $\sigma^r \to \sigma$, $\alpha^r \to \alpha$ and $a^r \to a$ as $r \to \infty$. Further assume that for each $\varepsilon > 0$,

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

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

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

Then, as $r \to \infty$, $(\hat{B}^r, \hat{X}^r) \Rightarrow (B^*, X^*)$, where X^* is a Brownian motion starting from zero with drift zero and variance σ^2 per unit time, independent of B^*. Furthermore, as $r \to \infty$,

$$\hat{Y}^r(\cdot) \Rightarrow X^*(\alpha(\cdot)) + \mu B^*(\cdot),$$

where for each $r \in \mathcal{R}$ and $t \geq 0$,

$$\hat{Y}^r(t) = \frac{X^r(B^r(r^2t)) - r^2t\alpha^r\mu^r}{r},$$

and $\alpha(t) = \alpha t$.

IV. The case of $\nu_\Sigma(x^*) > 0$

The results in this section require all the assumptions made in Section 3. To simplify the notation, we will write S_k^r and $\bar{S}_k^r$ instead of S_k^{r,x^*} and $\bar{S}_k^{r,x^*}$ and similarly S^r and $\bar{S}^r$ instead of S^{r,x^*} and $\bar{S}^{r,x^*}$.

Theorem 1. *Let $x^* < \infty,\ \nu_\Sigma(x^*) > 0$. Assume that as $r \to \infty$*

$$\left(\hat{\mathcal{Z}}^r(0), \bar{S}^r, \hat{W}^r_\Sigma(0)\right) \Rightarrow (\mathcal{Z}^*(0), D, W_0^*) \tag{6}$$

in $\mathbf{M}^K \times (\mathbf{D}[0,\infty))^K \times \mathbb{R}$, where

$$\mathcal{Z}^*(0) = \left(\frac{p_k\nu_k(x^*)}{\nu_\Sigma(x^*)}\frac{W_0^*}{x^*}\delta_{x^*}, k = 1, ..., K\right) \tag{7}$$

and

$$D(t) = \left(\frac{p_k\nu_k(x^*)}{\nu_\Sigma(x^*)}\frac{t \wedge W_0^*}{x^*}, k = 1, ..., K\right),\ t \geq 0. \tag{8}$$

Then

$$\left(\hat{\mathcal{Z}}_k^r, k = 1, ..., K\right) \Rightarrow \left(p_k\frac{\nu_k(x^*)}{\nu_\Sigma(x^*)}\frac{W^*_\Sigma}{x^*}\delta_{x^*}, k = 1, ..., K\right). \tag{9}$$

A. Proof of Theorem 1

The general idea of the proof of Theorem 1 is to first show that in the diffusion limits all jobs with service times less than x^* are prioritized so that their corresponding workload and queue length processes vanish. This implies that the processes $\hat{\mathcal{Z}}^r_\Sigma$ converge weakly to a multiple of δ_{x^*}. Then we use functional limit theorems and the fact that FIFO is the tie-breaking rule to prove that in the diffusion limit the queue lengths and the workloads of job classes are divided according to the proportions in the statement of the theorem. In the proof, we follow the ideas of Peterson [21], with customers having the residual service times not less than x^* regarded as the low priority (L) ones with necessary modifications. The most important difference is that in [21] the priority of a job cannot change. In our setting however, the priority of a job can change from low to high (H). It happens every time a job with service time x^* receives some service and hence its residual service time is decreased. Consequently, our problem requires a somewhat more careful approach, described below in more detail.

1) Additional notation: For $k \in \mathcal{K}$, $r \in \mathcal{R}$ and $t \geq 0$ denote

$$V_k^r(t) = \langle \chi, \mathcal{V}_k^r(t)\rangle, \quad \hat{V}_k^r(t) = \langle \chi, \hat{\mathcal{V}}_k^r(t)\rangle,$$

$$\begin{aligned} V^r_{k,H}(t) &= \langle \chi \mathbb{I}_{[0,x^*)}, \mathcal{V}_k^r(t)\rangle, \\ Q^r_{k,H}(t) &= \langle \mathbb{I}_{[0,x^*)}, \mathcal{Z}_k^r(t)\rangle, \\ W^r_{k,H}(t) &= \langle \chi \mathbb{I}_{[0,x^*)}, \mathcal{Z}_k^r(t)\rangle, \\ \hat{V}^r_{k,H}(t) &= \langle \chi \mathbb{I}_{[0,x^*)}, \hat{\mathcal{V}}_k^r(t)\rangle, \\ \hat{Q}^r_{k,H}(t) &= \langle \mathbb{I}_{[0,x^*)}, \hat{\mathcal{Z}}_k^r(t)\rangle, \\ \hat{W}^r_{k,H}(t) &= \langle \chi \mathbb{I}_{[0,x^*)}, \hat{\mathcal{Z}}_k^r(t)\rangle \end{aligned}$$

and

$$\begin{aligned} V^r_{k,L}(t) &= V^r_k(t) - V^r_{k,H}(t), \quad W^r_{k,L}(t) = W^r_k(t) - W^r_{k,H}(t), \\ \hat{V}^r_{k,L}(t) &= \hat{V}^r_k(t) - \hat{V}^r_{k,H}(t), \quad \hat{W}^r_{k,L}(t) = \hat{W}^r_k(t) - \hat{W}^r_{k,H}(t). \end{aligned}$$

In general, subscript “H” stands for high priority jobs, i.e. jobs with residual processing times strictly lower than x^* and subscript “L” indicates low priority jobs, i.e. those with residual service times greater than or equal to x^*.

Recall that in our notation

$$\begin{aligned} E^r_\Sigma(t) &= \sum_{k\in\mathcal{K}} E^r_k(t), \quad \hat{E}^r_\Sigma(t) = \sum_{k\in\mathcal{K}} \hat{E}^r_k(t), \\ E^*_\Sigma(t) &= \sum_{k\in\mathcal{K}} E^*_k(t), \quad W^r_\Sigma(t) = \sum_{k\in\mathcal{K}} W^r_k(t), \end{aligned}$$

and similarly,

$$\begin{aligned} Q^r_{\Sigma,H}(t) &= \langle \mathbb{I}_{[0,x^*)}, \mathcal{Z}^r_\Sigma(t)\rangle, \ \hat{Q}^r_{\Sigma,H}(t) = \langle \mathbb{I}_{[0,x^*)}, \hat{\mathcal{Z}}^r_\Sigma(t)\rangle, \\ W^r_{\Sigma,H}(t) &= \langle \chi\mathbb{I}_{[0,x^*)}, \mathcal{Z}^r_\Sigma(t)\rangle, \ \hat{W}^r_{\Sigma,H}(t) = \langle \chi\mathbb{I}_{[0,x^*)}, \hat{\mathcal{Z}}^r_\Sigma(t)\rangle, \\ W^r_{\Sigma,L}(t) &= W^r_\Sigma(t) - W^r_{\Sigma,H}(t), \ \hat{W}^r_{\Sigma,L}(t) = \hat{W}^r_\Sigma(t) - \hat{W}^r_{\Sigma,H}(t). \end{aligned}$$

Let $L^r(t)$ denote the number of units of service dedicated to jobs with initial service times equal to x^* by time $t \geq 0$ in the r-th system. Observe for future reference that $S^r_k(L^r(t))$ is the number of fully served initial class k jobs with initial service time x^* by time t in the r-th system.

2) Concentration of the mass at x^*: We will first show that

$$\hat{Q}^r_{\Sigma,H} \Rightarrow 0, \ r \to \infty. \tag{10}$$

Observe that (10) implies

$$\hat{W}^r_{\Sigma,H} \Rightarrow 0, \ r \to \infty \tag{11}$$

since

$$\hat{W}^r_{\Sigma,H} \leq x^* \hat{Q}^r_{\Sigma,H}. \tag{12}$$

For $r \in \mathcal{R}$ and $t \geq 0$, let

$$\tau^r(t) = \sup\{s \in [0,t] : \hat{Q}^r_{\Sigma,H}(s) = 0\},$$

which equals zero by definition if

$$\{s \in [0,t] : \hat{Q}^r_{\Sigma,H}(s) = 0\} = \varnothing.$$

Then for $r \in \mathcal{R}$ and $t \geq 0$

$$\begin{aligned} \hat{Q}^r_{\Sigma,H}(t) &\leq \hat{Q}^r_{\Sigma,H}(\tau^r(t)) + \frac{1}{r}\left(E^r_\Sigma(r^2 t) - E^r_\Sigma(r^2\tau^r(t)) + 1\right) \\ &= \hat{Q}^r_{\Sigma,H}(\tau^r(t)) + \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)) + \frac{1}{r} + r(t-\tau^r(t))\alpha^r_\Sigma. \end{aligned}$$

Indeed, the inequality above follows from the fact that clearly $\tau^r(t) \leq t$, so we can bound the (diffusion scaled) number of high priority customers in the system at time t from above by an analogous number of customers at time $\tau^r(t)$ increased by the number of external arrivals in this time interval. The addition of 1 on the right-hand side of the inequality is needed because of a possibility that there is no job with service time less than x^* at time $r^2\tau^r(t)$ in the system and a job with service time x^* is chosen for processing at this time, which immediately increases the number of jobs with residual processing times less than x^* by 1.

First we find an upper bound on $\hat{Q}^r_{\Sigma,H}(\tau^r(t))$. Fix $r \in \mathcal{R}$ and $t \geq 0$. If $\tau^r(t) = 0$, then $\hat{Q}^r_{\Sigma,H}(\tau^r(t)) = \hat{Q}^r_{\Sigma,H}(0)$. Otherwise, $\tau^r(t) > 0$. If $\hat{Q}^r_{\Sigma,H}(\tau^r(t)) = 0$, then any nonnegative upper bound suffices, so we can also assume that $\hat{Q}^r_{\Sigma,H}(\tau^r(t)) > 0$. Then $\hat{Q}^r_{\Sigma,H}(\tau^r(t)-) = 0$ and $\hat{Q}^r_{\Sigma,H}(\tau^r(t)) > 0$. Therefore in the rth system at time $\tau^r(t)$ the exogeneous arrival process has a jump. This means that $\hat{Q}^r_{\Sigma,H}(\tau^r(t)) \leq \hat{E}^r_\Sigma(\tau^r(t)) - \hat{E}^r_\Sigma(\tau^r(t)-)$. Combining the bounds for $\tau^r(t) = 0$ or $\tau^r(t) > 0$ gives

$$\hat{Q}^r_{\Sigma,H}(\tau^r(t)) \leq \hat{Q}^r_{\Sigma,H}(0) + \hat{E}^r_\Sigma(\tau^r(t)) - \hat{E}^r_\Sigma(\tau^r(t)-),$$

where by convention $\hat{E}^r_\Sigma(0-) = \hat{E}^r_\Sigma(0) = 0$. Hence

$$\hat{Q}^r_{\Sigma,H}(t) \leq \hat{Q}^r_{\Sigma,H}(0) + \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)-) + \frac{1}{r} + r(t-\tau^r(t))\alpha^r_\Sigma. \tag{13}$$

For $r \in \mathcal{R}$ and $t \geq 0$ let $\theta^r(t) = t - \tau^r(t)$ and $\tilde{\theta}^r(t) = \theta^r(t) + \frac{1}{r^2}$. For now, suppose that

$$r\theta^r \Rightarrow 0. \tag{14}$$

Then, it follows from (14) that as $r \to \infty$,

$$\theta^r \Rightarrow 0 \quad \text{and} \quad \tilde{\theta}^r \Rightarrow 0. \tag{15}$$

Fix $r \in \mathcal{R}$ and $t \geq 0$. By (1), we have

$$\hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)-) = \hat{E}^r_\Sigma(t) - \frac{1}{r}E^r_\Sigma(r^2\tau^r(t)-) + r\tau^r(t)\alpha^r_\Sigma$$

which, together with the fact that the process E^r_Σ is nondecreasing, implies

$$\begin{aligned} \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)) &\leq \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)-) \\ &\leq \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma\left(\left(\tau^r(t) - \frac{1}{r^2}\right)^+\right) + \frac{\alpha^r_\Sigma}{r}. \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(t - \theta^r(t)) &\leq \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)-) \\ &\leq \hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma\left(\left(t - \tilde{\theta}^r(t)\right)^+\right) + \frac{\alpha^r_\Sigma}{r}. \end{aligned}$$

By (2), (15) and the fact that E^*_Σ is continuous almost surely, we obtain (see [5], Section 17) that for $t \geq 0$ as $r \to \infty$

$$\hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(t - \theta^r(t)) \Rightarrow 0$$

and

$$\hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma\left((t - \tilde{\theta}^r(t))^+\right) + \frac{\alpha^r_\Sigma}{r} \Rightarrow 0.$$

Hence, as $r \to \infty$

$$\hat{E}^r_\Sigma(t) - \hat{E}^r_\Sigma(\tau^r(t)-) \Rightarrow 0. \tag{16}$$

Since the space $\mathbf{M}^K \times (\mathbf{D}[0,\infty))^K \times \mathbb{R}$ is separable, we can apply the Skorohod representation theorem ([5], Theorem 6.7) and assume that all the random elements in (6)-(8) are defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$, on which as $r \to \infty$

$$\left(\hat{\mathcal{Z}}^r(0), \bar{S}^r, \hat{W}^r_\Sigma(0)\right)(\omega) \to \left(\mathcal{Z}^*(0), D, W^*_0\right)(\omega) \tag{17}$$

in $\mathbf{M}^K \times (\mathbf{D}[0,\infty))^K \times \mathbb{R}$ for almost every $\omega \in \Omega$. Fix such an ω. In this paragraph, all the random elements under consideration are evaluated at this ω. Observe that from the consistency conditions for functions S^x_k listed in Section II-A it follows that, for a given r, $\bar{S}^r_k(t) = \hat{\mathcal{Z}}_k(0)(x^*)$ for $t \geq x^* \hat{\mathcal{Z}}^r_\Sigma(0)(x^*)$. By (17), there exists a finite constant C such that $\hat{\mathcal{Z}}^r_\Sigma(0)(x^*) \leq C$ for all $r \in \mathcal{R}$. This means that all the functions $\bar{S}^r_k$ as well as the functions D_k are constant on $[Cx^*, \infty)$. In this case, convergence in the Skorohod topology to a continuous limit implies convergence in the uniform topology as well ([5], Section 12). Hence, by (17) as $r \to \infty$

$$\begin{aligned} |\hat{\mathcal{Z}}^r_\Sigma(0)(x^*) - \mathcal{Z}^*_\Sigma(0)(x^*)| &= \left| \left\| \sum_{k \in \mathcal{K}} \bar{S}^r_k \right\|_\infty - \left\| \sum_{k \in \mathcal{K}} D_k \right\|_\infty \right| \qquad (18) \\ &\leq \left\| \sum_{k \in \mathcal{K}} \bar{S}^r_k - \sum_{k \in \mathcal{K}} D_k \right\|_\infty \to 0. \end{aligned}$$

This, together with (6), (7), (17), gives us

$$\hat{Q}^r_{\Sigma,H}(0) = \langle \mathbb{I}_{[0,x^*)}, \hat{\mathcal{Z}}^r_\Sigma(0) \rangle \Rightarrow \langle \mathbb{I}_{[0,x^*)}, \mathcal{Z}^*_\Sigma(0) \rangle = 0, \;\; r \to \infty, \tag{19}$$

which in turn, together with (2), (13), (14), (16) implies (10).

Therefore it remains to prove (14). For each $r \in \mathcal{R}$ and $t \geq 0$, we examine the behavior of $W^r_{\Sigma,H}$ on time intervals of the form $(r^2\tau^r(t), r^2 t]$ to derive an expression that relates $\hat{W}^r_{\Sigma,H}(t)$ and $\theta^r(t)$. In particular, since for each $r \in \mathcal{R}$ and $t \geq 0$, $Q^r_{\Sigma,H}(s) \neq 0$ for all $s \in (r^2\tau^r(t), r^2 t]$ and the service discipline is SRPT, it follows that for each $r \in \mathcal{R}$ and $t \geq 0$,

$$\begin{aligned} W^r_{\Sigma,H}(r^2 t) \leq W^r_{\Sigma,H}(r^2\tau^r(t)) &+ V^r_{\Sigma,H}(r^2 t) - V^r_{\Sigma,H}(r^2\tau^r(t)) \\ &+ x^* - r^2(t - \tau^r(t)). \end{aligned}$$

Again, the addition of x^* on the right-hand side of the inequality is needed because of a possibility that there are only jobs with service time greater than or equal to x^* at time $r^2\tau^r(t)$ in the queue.

By applying diffusion scaling and rearranging, we obtain for $r \in \mathcal{R}$ and $t \geq 0$

$$\begin{aligned} \hat{W}^r_{\Sigma,H}(t) \leq \hat{W}^r_{\Sigma,H}(\tau^r(t)) + \hat{V}^r_{\Sigma,H}(t) - \hat{V}^r_{\Sigma,H}(\tau^r(t)) + \frac{x^*}{r} \\ + (\alpha^r_\Sigma \langle \chi \mathbb{I}_{[0,x^*)}, \nu_\Sigma \rangle - 1) r\theta^r(t). \end{aligned}$$

Using the same line of reasoning that gave rise to (13), for $r \in \mathcal{R}$ and $t \geq 0$,

$$\begin{aligned} \hat{W}^r_{\Sigma,H}(t) \leq \hat{W}^r_{\Sigma,H}(0) + \hat{V}^r_{\Sigma,H}(t) - \hat{V}^r_{\Sigma,H}(\tau^r(t)-) + \frac{x^*}{r} \\ + (\alpha^r_\Sigma \langle \chi \mathbb{I}_{[0,x^*)}, \nu_\Sigma \rangle - 1) r\theta^r(t). \end{aligned}$$

Since $\hat{W}^r_{\Sigma,H}(t) \geq 0$ for all $r \in \mathcal{R}$ and $t \geq 0$, it follows that for such r and t

$$\begin{aligned} (1 - \alpha^r_\Sigma \langle \chi \mathbb{I}_{[0,x^*)}, \nu_\Sigma \rangle) r\theta^r(t) \leq \hat{W}^r_{\Sigma,H}(0) + \hat{V}^r_{\Sigma,H}(t) \\ - \hat{V}^r_{\Sigma,H}(\tau^r(t)-) + \frac{x^*}{r}. \end{aligned} \tag{20}$$

By (2) and the theorem assumption, we have that

$$\lim_{r \to \infty} \left(1 - \alpha^r_\Sigma \langle \chi \mathbb{I}_{[0,x^*)}, \nu_\Sigma \rangle\right) = 1 - \alpha_\Sigma \langle \chi \mathbb{I}_{[0,x^*)}, \nu_\Sigma \rangle > 0. \tag{21}$$

Hence, for r sufficiently large, $\left(1 - \alpha^r_\Sigma \langle \chi \mathbb{I}_{[0,x^*)}, \nu_\Sigma \rangle\right) \theta^r \geq 0$ for all $t \geq 0$. Using Proposition 1, one can prove (see [12]) that

$$\hat{V}^r_{\Sigma,H} \Rightarrow V^*_{\Sigma,H}, \;\; r \to \infty, \tag{22}$$

where $V^*_{\Sigma,H}$ is a Brownian motion starting from zero with zero drift and finite variance per unit time. Then (17) and (20)-(22) together imply that $\theta^r \Rightarrow 0$, $r \to \infty$. By the same line of reasoning that gave rise to (16), we have that for $t \geq 0$

$$\hat{V}^r_{\Sigma,H}(t) - \hat{V}^r_{\Sigma,H}(\tau^r(t)-) \Rightarrow 0.$$

By using this, (6), (7), (12) and (19)-(21) we obtain (14).

3) Proportional breakdown: For $r \in \mathcal{R}$ let $\sigma^r(t)$ be the time of arrival to the rth system of the job with service time x^* which most recently completed service before time t (if none has yet completed we define it as 0). Let $\bar{\sigma}^r(t) = \frac{1}{r^2}\sigma^r(r^2 t)$, $t \geq 0$. We will now prove that $\bar{\sigma}^r \Rightarrow e$, where $e(t) = t$ for $t \geq 0$. For $t \geq 0$ we have

$$\begin{aligned} W^r_{\Sigma,L}(t) \geq V^r_{\Sigma,L}(t) - V^r_{\Sigma,L}(\sigma^r(t)) - u^r(t) \\ + x^* \left(\mathcal{Z}^r_\Sigma(0)(x^*) - S^r_\Sigma(L^r(t))\right), \end{aligned} \tag{23}$$

where $u^r(t)$ denotes the partial service (if any) that has been performed in the time interval $[\sigma^r(t), t)$ on the next job with service time x^*. Recall that $S^r_\Sigma(L^r(t))$ is the number of fully served initial jobs with initial service time x^* by time t in the rth system, so the last term is the workload of initial low priority jobs still present in the system at time t. The fact that (23) holds is a consequence of the FIFO discipline among the jobs with the same residual service time. Notice that $u^r(t) \leq x^*$. On the other hand, we also have that for $t \geq 0$

$$\begin{aligned} W^r_{\Sigma,L}(t) \leq V^r_{\Sigma,L}(t) - V^r_{\Sigma,L}(\sigma^r(t)-) - u^r(t) \\ + x^* \left(\mathcal{Z}^r_\Sigma(0)(x^*) - S^r_\Sigma(L^r(t))\right) + \langle \chi \mathbb{I}_{(x^*,\infty)}, \mathcal{Z}^r_\Sigma(0) \rangle, \end{aligned} \tag{24}$$

where $V^r_{\Sigma,L}(0-) = V^r_{\Sigma,L}(0) = 0$ by convention. The last term in (24) takes into account possible initial jobs with processing times greater than x^*. Under the diffusion scaling $\hat{W}^r_{\Sigma,L}(t) = \frac{1}{r}W^r_{\Sigma,L}(r^2t)$ we have $\hat{W}^r_{\Sigma,L} = \hat{W}^r_{\Sigma} - \hat{W}^r_{\Sigma,H}$ and by (5) and (11),

$$\hat{W}^r_{\Sigma,L} \Rightarrow W^*_{\Sigma}. \tag{25}$$

Under the diffusion scaling, since $\mathcal{Z}^r_\Sigma(0)(x^*) - S^r_\Sigma(L^r(t)) \geq 0$, (23) yields for $t \geq 0$

$$\begin{aligned}\hat{W}^r_{\Sigma,L}(t) \geq \hat{V}^r_{\Sigma,L}(t) - \hat{V}^r_{\Sigma,L}(\bar{\sigma}^r(t)) - \frac{1}{r}u^r(r^2t)\\ + r\alpha^r_\Sigma \nu_\Sigma(x^*)(t - \bar{\sigma}^r(t)).\end{aligned} \tag{26}$$

Since $\sigma^r(t) \leq t$, (26) gives us after rearranging

$$\begin{aligned}0 \leq t - \bar{\sigma}^r(t) \leq \frac{1}{r\alpha^r_\Sigma \nu_\Sigma(x^*)}\Big(\hat{W}^r_{\Sigma,L}(t)\\ -\hat{V}^r_{\Sigma,L}(t) + \hat{V}^r_{\Sigma,L}(\bar{\sigma}^r(t)) + \frac{1}{r}u^r(r^2t)\Big).\end{aligned} \tag{27}$$

Let $T > 0$. Since convergence in the Skorohod topology to a continuous limit implies convergence in the uniform topology, we have $\|\hat{W}^r_{\Sigma,L}\|_T \Rightarrow \|\hat{W}^*_\Sigma\|_T$. For $t \in [0, T]$, we can bound $\hat{V}^r_{\Sigma,L}(\bar{\sigma}^r(t))$ by $\|\hat{V}^r_{\Sigma,L}\|_T$, so (27) after taking norms gives

$$\|t - \bar{\sigma}^r(t)\|_T \leq \frac{1}{r\alpha^r_\Sigma \nu_\Sigma(x^*)}\left(\|\hat{W}^r_{\Sigma,L}\|_T + 2\|\hat{V}^r_{\Sigma,L}\|_T + \frac{x^*}{r}\right).$$

Define the continuous functions $h_r : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ by $h_r(x) = \frac{x}{r}$ and $h(x) = 0, x \in \mathbb{R}$. Then, by Theorem 2.7 of [5], $h_r(\|\hat{W}^r_{\Sigma,L}\|_T + 2\|\hat{V}^r_{\Sigma,L}\|_T + \frac{x^*}{r}) \Rightarrow h(\|\hat{W}^*_\Sigma\|_T + 2\|\hat{V}^*_{\Sigma,L}\|_T) = 0$. Thus the right hand side above converges in probability to zero, which in turn implies that $\bar{\sigma}^r \Rightarrow e$ by Theorem 3.1 of [5].

For $t \geq 0$, let

$$I^r(t) := x^*\left(\hat{\mathcal{Z}}^r_\Sigma(0)(x^*) - \bar{S}^r_\Sigma(\bar{L}^r(rt))\right) + r\alpha^r_\Sigma \nu_\Sigma(x^*)(t - \bar{\sigma}^r(t)),$$

where $\bar{L}^r(t) = \frac{1}{r}L^r(rt)$. From (23)-(24) we have that

$$I^r(t) \leq \hat{W}^r_{\Sigma,L}(t) - \hat{V}^r_{\Sigma,L}(t) + \hat{V}^r_{\Sigma,L}(\bar{\sigma}^r(t)) + \frac{1}{r}u^r(r^2t), \tag{28}$$

$$\begin{aligned}I^r(t) \geq \hat{W}^r_{\Sigma,L}(t) - \hat{V}^r_{\Sigma,L}(t) + \hat{V}^r_{\Sigma,L}(\bar{\sigma}^r(t)-) + \frac{1}{r}u^r(r^2t)\\ - \langle \chi \mathbb{I}_{(x^*,\infty)}, \hat{\mathcal{Z}}^r_\Sigma(0)\rangle.\end{aligned} \tag{29}$$

Notice that

$$\langle \chi \mathbb{I}_{(x^*,\infty)}, \hat{\mathcal{Z}}^r_\Sigma(0)\rangle \Rightarrow 0,\ r \to \infty. \tag{30}$$

Indeed,

$$\hat{W}^r_\Sigma(0) = \langle \chi \mathbb{I}_{[0,x^*]}, \hat{\mathcal{Z}}^r_\Sigma(0)\rangle + \langle \chi \mathbb{I}_{(x^*,\infty)}, \hat{\mathcal{Z}}^r_\Sigma(0)\rangle.$$

By (7) and (18), $\langle \chi \mathbb{I}_{[0,x^*]}, \hat{\mathcal{Z}}^r_\Sigma(0)\rangle \Rightarrow W^*_0$, so, taking (4) into account, we see that (30) must hold.

Using Proposition 1, one can prove (see [12]) that

$$\hat{V}^r_{\Sigma,L} \Rightarrow V^*_{\Sigma,L},\ r \to \infty, \tag{31}$$

where $V^*_{\Sigma,L}$ is a Brownian motion starting from zero with zero drift and finite variance per unit time. From (28)-(29), using (25), (30)-(31), the fact that $\bar{\sigma}^r \Rightarrow e$ and the Random Time Change Theorem ([5], Theorem 14.4) we have that

$$I^r \Rightarrow W^*_\Sigma. \tag{32}$$

We can now obtain the desired breakdown for the workload processes. For $t \geq 0, k \in \mathcal{K}$ we have the following inequalities, analogous to (23)-(24):

$$\begin{aligned}W^r_{k,L}(t) \leq V^r_{k,L}(t) - V^r_{k,L}(\sigma^r(t)-) - u^r_k(t)\\ + x^*\left(\mathcal{Z}^r_k(0)(x^*) - S^r_k(L^r(t))\right) + \langle \chi \mathbb{I}_{(x^*,\infty)}, \mathcal{Z}^r_k(0)\rangle,\end{aligned} \tag{33}$$

$$\begin{aligned}W^r_{k,L}(t) \geq V^r_{k,L}(t) - V^r_{k,L}(\sigma^r(t)) - u^r_k(t)\\ + x^*\left(\mathcal{Z}^r_k(0)(x^*) - S^r_k(L^r(t))\right),\end{aligned} \tag{34}$$

where $u^r_k(t)$ denotes the partial service (if any) that has been performed in the time interval $[\sigma^r(t), t)$ on the next job of class k with service time x^*. Notice that $u^r_k(t) \leq x^*$ for all $k \in \mathcal{K}$.

Under diffusion scaling, (33) yields

$$\begin{aligned}\hat{W}^r_{k,L}(t) \leq \hat{V}^r_{k,L}(t) - \hat{V}^r_{k,L}(\bar{\sigma}^r(t)-) - \frac{1}{r}u^r_k(r^2t)\\ + x^*\left(\hat{\mathcal{Z}}^r_k(0)(x^*) - \bar{S}^r_k(\bar{L}^r(rt))\right)\\ + r\alpha^r_k \nu_k(x^*)(t - \bar{\sigma}^r(t)) + \langle \chi \mathbb{I}_{(x^*,\infty)}, \hat{\mathcal{Z}}^r_k(0)\rangle\end{aligned} \tag{35}$$

and (34) yields

$$\begin{aligned}\hat{W}^r_{k,L}(t) \geq \hat{V}^r_{k,L}(t) - \hat{V}^r_{k,L}(\bar{\sigma}^r(t)) - \frac{1}{r}u^r_k(r^2t)\\ + x^*\left(\hat{\mathcal{Z}}^r_k(0)(x^*) - \bar{S}^r_k(\bar{L}^r(rt))\right)\\ + r\alpha^r_k \nu_k(x^*)(t - \bar{\sigma}^r(t))\end{aligned} \tag{36}$$

for $t \geq 0,\ k \in \mathcal{K}$. Since $\mathcal{Z}^r_k \leq \mathcal{Z}^r_\Sigma$ and χ is nonnegative, from (30) we have that as $r \to \infty$

$$\langle \chi \mathbb{I}_{(x^*,\infty)}, \hat{\mathcal{Z}}^r_k(0)\rangle \Rightarrow 0. \tag{37}$$

Suppose that for each $k \in \mathcal{K}$, $r \in \mathcal{R}$

$$\begin{aligned}I^r_k(t) := x^*\left(\hat{\mathcal{Z}}^r_k(0)(x^*) - \bar{S}^r_k(\bar{L}^r(rt))\right) + r\alpha^r_k \nu_k(x^*)(t - \bar{\sigma}^r(t))\\ \Rightarrow \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} W^*_\Sigma(t).\end{aligned} \tag{38}$$

Using this, (35)-(37), the fact that $\bar{\sigma}^r \Rightarrow e$ and the Random Time Change Theorem, we can write that

$$\hat{W}^r_{k,L} \Rightarrow \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} W^*_\Sigma = p_k \frac{\nu_k(x^*)}{\nu_\Sigma(x^*)} W^*_\Sigma. \tag{39}$$

From Section IV-A2 and (30) it follows that $\hat{\mathcal{Z}}^r_\Sigma \Rightarrow \frac{W^*_\Sigma}{x^*}\delta_{x^*}$. Taking this into account gives us (9). Therefore it remains to prove (38).

Recall from (2) that $\alpha^r \to \alpha$ as $r \to \infty$ and that by the Skorohod representation theorem we may assume that all the random elements in (6)-(8) are defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that (17) holds for almost every $\omega \in \Omega$. In what follows, all the random elements are evaluated at such an ω. Using an analogous line of reasoning as the one that led us to (18) we obtain that, as $r \to \infty$,

$$|\hat{\mathcal{Z}}_k^r(0)(x^*) - \mathcal{Z}_k^*(0)(x^*)| \to 0 \tag{40}$$

and, similarly,

$$\sup_{t\geq 0} |\bar{S}_k^r(t) - D_k(t)| \to 0. \tag{41}$$

Observe that

$$I_k^r - \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} I^r = x^* \left(\hat{\mathcal{Z}}_k^r(0)(x^*) - \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} \hat{\mathcal{Z}}_\Sigma^r(0)(x^*) + \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} \bar{S}_\Sigma^r(\bar{L}^r(rt)) - \bar{S}_k^r(\bar{L}^r(rt)) \right).$$

By (7),

$$\mathcal{Z}_k^*(0)(x^*) = \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} \frac{W_0^*}{x^*} = \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} \mathcal{Z}_\Sigma^*(0)(x^*),$$

so by (40),

$$\left| \hat{\mathcal{Z}}_k^r(0)(x^*) - \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} \hat{\mathcal{Z}}_\Sigma^r(0)(x^*) \right| \to 0$$

as $r \to \infty$. Similarly, by (8) and (41)

$$\sup_{t\geq 0} \left| \bar{S}_k^r(\bar{L}^r(rt)) - \frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} \bar{S}_\Sigma^r(\bar{L}^r(rt)) \right| \to 0$$

as $r \to \infty$, which implies that the limits as $r \to \infty$ of I_k^r and $\frac{\alpha_k \nu_k(x^*)}{\alpha_\Sigma \nu_\Sigma(x^*)} I^r$ coincide and, by (32), proves (38).

V. THE CASE OF $\nu_\Sigma(x^*) = 0$

In this case we take a seemingly different approach. We start from modeling an exogeneous arrival process E_Σ common for all jobs. When a job arrives at the system, we randomly assign it to a class $1, ..., K$ with probabilities $p_1, ..., p_K$ correspondingly[1]. We also assume that there are no customers in the system at time 0. This is described more formally below.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the probability space on which the stochastic elements of the model are defined. Let E_Σ be the exogenous arrival process, i.e. for $t \geq 0$, $E_\Sigma(t)$ is the number of arrivals to the system in the time interval $(0, t]$. It is a delayed renewal process with rate $\alpha_\Sigma > 0$ such that the interarrival times have standard deviation $a_\Sigma \geq 0$. Then job j arrives at time $T^j = \inf\{t \geq 0 : E_\Sigma(t) \geq j\}$.

Let $\{\varphi_i\}_{i\in\mathbb{N}}$ be i.i.d. random vectors, independent of E_Σ such that for each i $\varphi_i = (\varphi_{i,1}, ..., \varphi_{i,K})$, where $\varphi_{i,k} = 1$ if job i belongs to class k and $\varphi_{i,k} = 0$ otherwise. We assume that $\mathbf{P}(\varphi_i = e_k) = p_k$, where e_k is the kth unit vector in $\mathbb{R}^K$ and $p_k \in (0,1)$ for each $k \in \mathcal{K}$, $\sum_{k\in\mathcal{K}} p_k = 1$. Put $p = (p_1, ..., p_K)$. Let $\Phi(n) = (\Phi_1(n), ..., \Phi_K(n)) = \sum_{i=1}^n \varphi_i$. In words, $\Phi_k(n)$ is the number of class k jobs among the first n jobs which arrived in the system. We can now define $E_k(t) = \Phi_k(E_\Sigma(t))$, $k \in \mathcal{K}$, $t \geq 0$ and $E(t) = (E_1(t), ..., E_K(t))$. Put $\alpha_k = p_k \alpha_\Sigma$, $k \in \mathcal{K}$ and $\alpha = (\alpha_1, ..., \alpha_K)$. We define the processes $\mathcal{Z}$, W, Q and $\mathcal{V}$ analogously as in Section II. We assume that $\mathcal{Z}_k(0) = \mathbf{0}$ for each $k \in \mathcal{K}$.

Let v_i represent the initial service time of the ith job in the system. We assume that $(\varphi_i, v_i)_{i\geq 1}$ are i.i.d. random vectors independent of E and that ν_k is the conditional distribution of v_i under the condition of $\varphi_i = e_k$, $k \in \mathcal{K}$. We put $\nu_\Sigma = \sum_{k\in\mathcal{K}} p_k \nu_k$. For $j \in \mathbb{N}$, $k \in \mathcal{K}$ and $t \geq T^j$ let $w^j(t)$ denote the residual service time of job j at time t. As before, the SRPT protocol is used.

We apply diffusion scaling for a sequence of systems similarly as in Section III. Let $\mathcal{R}$ be a sequence of positive numbers increasing to infinity. Consider an $\mathcal{R}$-indexed sequence of stochastic models presented in the previous paragraph. For each $r \in \mathcal{R}$, there are stochastic primitives E_Σ^r, $\{\varphi_i^r\}_{i\in\mathbb{N}}$, Φ_k^r, E^r, $\{v_i^r\}_{i\in\mathbb{N}}$ with parameters $\alpha^r, a_\Sigma^r, p_k^r, \nu_k^r$. The stochastic elements of each model are defined on a probability space $(\Omega^r, \mathcal{F}^r, \mathbf{P}^r)$ with expectation operator $\mathbb{E}^r$ and variance operator Var^r. We also have arrival times $\{T^{r,j}\}_{j\in\mathbb{N}}$, a state descriptor $\mathcal{Z}^r$ and processes $W^r, Q^r, \mathcal{V}^r$.

A diffusion scaling is applied to each model in the $\mathcal{R}$-indexed sequence as in (1). Furthermore, for each $r \in \mathcal{R}$ and $t \geq 0$, let

$$\hat{\Phi}^r(t) = \frac{1}{r}(\Phi^r(\lfloor r^2 t \rfloor) - \lfloor r^2 t p^r \rfloor).$$

Let $\alpha_\Sigma, a_\Sigma \in (0, +\infty)$ and define $\alpha_\Sigma(t) = \alpha_\Sigma t$, $t \geq 0$, $p = (p_1, ..., p_k) \in (0,1)^K$, $\sum_{k\in\mathcal{K}} p_k = 1$. We make the following asymptotic assumptions for the sequence of stochastic primitives. Assume that as $r \to \infty$,

$$\alpha_\Sigma^r \to \alpha_\Sigma, \qquad a_\Sigma^r \to a_\Sigma, \qquad p^r \to p, \qquad \hat{E}_\Sigma^r \Rightarrow E_\Sigma^*, \tag{42}$$

where E_Σ^* is a Brownian motion starting from zero with drift zero and variance $a_\Sigma^2 \alpha_\Sigma^3$ per unit time. Moreover, we assume that $\nu^r = \nu$, i.e., ν^r does not depend on r.

We will now determine the limits of processes $\hat{\Phi}^r$ and $\hat{E}^r$ as $r \to \infty$. We first apply Proposition 1 to the processes $\hat{\Phi}^r = \left(\hat{\Phi}_k^r, r = 1, ..., K\right)$. Fix $k \in \mathcal{K}$. For each $r \in \mathbb{R}$, $\{\varphi_{i,k}^r\}_{i\in\mathbb{N}}$ are independent Bernoulli distributed random variables such that $\mathbf{P}^r(\varphi_{i,k}^r = 1) = p_k^r$. Hence $\mathbb{E}^r \varphi_{i,k}^r = p_k^r$ and $\mathrm{Var}^r\ \varphi_{i,k}^r = p_k^r(1 - p_k^r)$. Using the fact that $p_k^r \to p_k$ and Proposition 1, we obtain

$$\frac{1}{r}\left(\sum_{i=1}^{\lfloor r^2 t \rfloor} \varphi_{i,k} - \lfloor r^2 t \rfloor p_k \right) \Rightarrow \Phi_k^*,\ r \to \infty, \tag{43}$$

where Φ_k^* is a Brownian motion starting from zero with drift zero and variance $p_k(1 - p_k)$ per unit time. By using the Cramer-Wold device, multidimensional central limit theorem and Prohorov theorem, we can generalize this to joint convergence $\hat{\Phi}^r \Rightarrow \Phi^*$, where Φ^* is a K-dimensional Brownian motion starting from 0 with drift 0 and covariance matrix $C = [c_{ij}]_{i,j}$ such that $c_{ii} = p_i(1 - p_i)$, $c_{i,j} = -p_i p_j$ for $i \neq j$ (cf. [26], proof of Theorem 4.3.5). Moreover, this convergence is joint with (42) and Φ^* is independent of E_Σ^*.

[1] As we will see, this is a special case of the approach considered in Sections 1-4, with the exogeneous arrival process E having dependent coordinates.

Now we use Proposition 1 again to obtain the limit of the process $\hat{E}^r = \left(\hat{E}^r_k, r = 1, ..., K\right)$. Fix $k \in \mathcal{K}$. Observe that

$$\hat{E}^r_k(t) = \frac{\Phi^r_k\left(E^r_\Sigma(r^2t)\right) - r^2t\alpha^r_\Sigma p^r_k}{r}.$$

Therefore, by Proposition 1, $\hat{E}^r_k \Rightarrow E^*_k$, where $E^*_k(t) = \Phi^*_k(\alpha_\Sigma t) + p_k E^*_\Sigma(t), t \geq 0$. Note that E^*_k is a Brownian motion starting from 0 with drift 0 and variance per unit time $\alpha_\Sigma p_k(1-p_k) + p_k^2\alpha^3_\Sigma a^2_\Sigma$. Therefore, by an analogous argument as in the previous paragraph, we obtain joint convergence $\hat{E}^r \Rightarrow E^*$, where $E^*(t) = \Phi^*(\alpha_\Sigma t) + pE^*_\Sigma(t), t \geq 0$. Note that E^* is a K-dimensional Brownian motion starting from 0 with drift 0 and covariance matrix $D = [d_{ij}]_{i,j}$ such that $d_{ii} = p_i^2 a^2_\Sigma \alpha^3_\Sigma + \alpha^2_\Sigma p_i(1-p_i)$, $d_{ij} = \mathrm{Cov}(\Phi^*_i(\alpha_\Sigma) + p_i E^*_\Sigma(1), \Phi^*_j(\alpha_\Sigma) + p_j E^*_\Sigma(1)) = p_i p_j a^2_\Sigma \alpha^3_\Sigma - \alpha^2_\Sigma p_i p_j$, $i \neq j$.

Before we proceed, we present a general property of the SRPT protocol, which will be used in the following proofs.

Lemma 1. *Consider a single-server queueing system with customers served according to the SRPT protocol. Let $t \geq 0$ and let i, j be two jobs present in the system at time t with initial processing times v_i, v_j and residual processing times $w_i(t), w_j(t)$ respectively. Then the intervals $(w_i(t), v_i)$ and $(w_j(t), v_j)$ are disjoint.*

Proof. Suppose that $(w_i(t), v_i) \cap (w_j(t), v_j) \neq \varnothing$. First, consider the case when neither of the intervals is a subset of the other. Without loss of generality we may assume that $w_i(t) \leq w_j(t) < v_i \leq v_j$, thus $(w_i(t), v_i) \cap (w_j(t), v_j) = (w_j(t), v_i)$. This means that the job j, even though its initial processing time was not less than v_i, was partially served so that its residual processing time is lower than the initial processing time of the job i. Since the system uses the SRPT protocol, this is possible only when $v_i = v_j$ or when i arrived after j. In the first case, the start of the processing immediately breaks the tie so i cannot be partially processed before j gets fully serviced, thus $w_i(t) = v_i$ and therefore this case is impossible. In the second case, let t_1 be the time of the arrival of the job i. If $w_j(t_1) > v_i$, then for any $t \geq t_1$ the job j could only be chosen for processing after i had been fully serviced, which means that this case is impossible. If $w_j(t_1) \leq v_i$, then for any $t \geq t_1$ the job i could only be chosen for processing after j is fully serviced, which means that this case is impossible as well.

Finally, consider the case when one of the intervals is a subset of the other. Without loss of generality we can assume that $w_j(t) \leq w_i(t) < v_i \leq v_j$. Arguing as above, we can obtain that this is also impossible. This leads to a contradiction, therefore the intervals $(w_i(t), v_i)$ and $(w_j(t), v_j)$ are disjoint. □

We are now ready to formulate the main theorem in the case under consideration.

Theorem 2. *Let $x^* < \infty$, $\nu_\Sigma(x^*) = 0$. Assume that for every $k \in \mathcal{K}$ the limit*

$$\gamma_k = \lim_{x \uparrow x^*} f_k(x), \tag{44}$$

where f_k is the Radon–Nikodym derivative of the measure ν_k with respect to ν_Σ, exists. Under the assumptions of this section, we have that, as $r \to \infty$,

$$\left(\hat{\mathcal{Z}}^r_k, k = 1, ..., K\right) \xrightarrow{fd} \left(p_k\gamma_k\frac{W^*_\Sigma}{x^*}\delta_{x^*}, k = 1, ..., K\right). \tag{45}$$

Proof (somewhat heuritstic). We first show that

$$\hat{\mathcal{Z}}^r_\Sigma \Rightarrow \frac{W^*_\Sigma}{x^*}\delta_{x^*}, \ r \to \infty. \tag{46}$$

and that, for any $x \in (0, x^*)$

$$\left\langle \mathbb{I}_{[0,x)}, \hat{\mathcal{Z}}^r_\Sigma(t)\right\rangle \Rightarrow 0, \quad \left\langle \chi\mathbb{I}_{[0,x)}, \hat{\mathcal{Z}}^r_\Sigma(t)\right\rangle \Rightarrow 0, \ r \to \infty. \tag{47}$$

This is done similarly as in the first part of the proof of Theorem 1.

Fix $t > 0$ and $\varepsilon > 0$. Let $\{i^r_j\}_j$ be the sequence of jobs present in the system r and having residual processing time in $(x^* - \varepsilon, x^*]$ at time t, i.e. such that $w^r_{i_j}(t) > x^* - \varepsilon$. In what follows, we will simply write i_j instead of i^r_j when the system in question can be inferred from the context. Let $n^r(t)$ be the number of such jobs (of all classes) in the rth system at time t. Notice that

$$\sum_{j=1}^{n^r(t)} v^r_{i_j} \geq W^r_\Sigma(t) - \left\langle \chi\mathbb{I}_{[0,x^*-\varepsilon]}, \mathcal{Z}^r_\Sigma(t)\right\rangle. \tag{48}$$

We claim that

$$\sum_{j=1}^{n^r(t)} v^r_{i_j} \leq W^r_\Sigma(t) - \left\langle \chi\mathbb{I}_{[0,x^*-\varepsilon]}, \mathcal{Z}^r_\Sigma(t)\right\rangle + \varepsilon. \tag{49}$$

Indeed, by Lemma 1, the intervals of the form $(w^r_{i_j}(t), v^r_{i_j})$ are pairwise disjoint. All of these intervals are contained in $(x^* - \varepsilon, x^*]$ by the definition of the sequence $\{i^r_j\}_j$, therefore the sum of their lengths cannot exceed ε.

By (5) and (47)-(49) we obtain that as $r \to \infty$[2]

$$\frac{1}{r}\sum_{j=1}^{n^r(r^2t)} v^r_{i_j} \Rightarrow W^*_\Sigma(t). \tag{50}$$

Observe that for $r \in \mathcal{R}$ and almost every (with respect to ν_Σ) $x \in (x^* - \varepsilon, x^*)$

$$\begin{aligned} &\mathbf{P}^r(\varphi^r_i = e_k | v^r_i = x) \\ &= \frac{\mathbf{P}^r(\varphi^r_i = e_k)}{\mathbf{P}^r(v^r_i = x)} \cdot \mathbf{P}^r(v^r_i = x | \varphi^r_i = e_k) \\ &= p^r_k f_k(x). \end{aligned} \tag{51}$$

Consider the sequence $\{v^r_{i_j}\varphi^r_{i_j,k}\}_j$. Obviously $\varphi^r_{i_j,k} = 1$ if the job i^r_j belongs to the class k and $\varphi^r_{i_j,k} = 0$ otherwise. Similarly as before we can show that the limits of $\hat{W}^r_k(t)$ and $\frac{1}{r}\sum_{j=1}^{n^r(r^2t)} v^r_{i_j}\varphi^r_{i_j,k}$ coincide. By (47) and the fact that $Q^r_\Sigma(t) = n^r(t) + \left\langle \mathbb{I}_{[0,x^*-\varepsilon]}, \mathcal{Z}^r_\Sigma(t)\right\rangle$ we obtain that as $r \to \infty$

$$\hat{n}^r(t) := \frac{1}{r}n^r(r^2t) \Rightarrow Q^*_\Sigma(t) = \frac{W^*_\Sigma(t)}{x^*}. \tag{52}$$

[2] In the context of diffusion scaling, we define the sequence $\{i_j\}_j$ for time r^2t instead of t.

If $n^r(r^2t)$ is of the order less than r, then by (47)-(48) and the bound $v_i \leq x^*$ for all i, $\hat{W}^r_\Sigma(t)$ is asymptotically negligible. On the other hand, if $n^r(r^2t)$ is large, using the law of large numbers, we get,

$$\frac{1}{n^r(r^2t)} \sum_{j=1}^{n^r(r^2t)} v^r_{i_j}\varphi^r_{i_j,k} \approx \frac{1}{n^r(r^2t)} \sum_{j=1}^{n^r(r^2t)} \mathbb{E}^r\left(v^r_{i_j}\varphi^r_{i_j,k}\right) \tag{53}$$

and by (51), (44),

$$\begin{aligned}\mathbb{E}^r\left(v^r_{i_j}\varphi^r_{i_j,k}\right) &= \int_{x^*-\varepsilon}^{x^*} u\mathbf{P}^r(\varphi^r_{i_j} = e_k|v^r_i = u)dF^r_j(u) \quad (54)\\ &= p^r_k \int_{x^*-\varepsilon}^{x^*} uf_k(u)dF^r_j(u)\\ &= p^r_k(\gamma_k + o(1))(x^* + O(\varepsilon)),\end{aligned}$$

where F^r_j is the distribution function of $v^r_{i_j}$ and $O(\varepsilon) \in [-\varepsilon, 0]$. Therefore, from (52)-(54) and (42) it follows that, as $r \to \infty$, any weak limiting distribution of a subsequence of the sequence

$$\frac{1}{r} \sum_{j=1}^{n^r(r^2t)} v^r_{i_j}\varphi^r_{i_j,k} = \frac{\hat{n}^r(r^2t)}{n^r(r^2t)} \sum_{j=1}^{n^r(r^2t)} v^r_{i_j}\varphi^r_{i_j,k}$$

is stochastically bounded from below by the distribution of

$$p_k \frac{W^*_\Sigma}{x^*}(\gamma_k(x^* - \varepsilon) + o(1))$$

and stochastically bounded from above by the distribution of

$$p_k W^*_\Sigma(\gamma_k + o(1)).$$

By letting $\varepsilon \downarrow 0$ and taking (46), (50) into account we obtain the desired breakdown and convergence of one-dimensional distributions. It is easy to extend this result to convergence of finite-dimensional distributions.

□

VI. Simulations

In this section we present the results of our computer simulations. We simulated the system described in Section II with two user classes. The times between arrivals of jobs of each class are exponentially distributed with parameters α_1, α_2 correspondingly.

We first consider the case when $\nu_\Sigma(x^*) > 0$. We assume that $\alpha_1 = 0.25$, $\alpha_2 = 1.25$. Hence $\alpha_\Sigma = 1.5$. The initial service times of the first class take the values $0.5, 1, 1.5$ with equal probabilities and the initial service times of the second class take the values $0.2, 0.3, 0.4, 0.6, 1.5$ with equal probabilities. Then we have $x^* = 1.5$, $\rho = 1$, $p_1 = 1/6$, $p_2 = 5/6$, $\nu_\Sigma(1.5) = 2/9$ and $p_1\nu_1(1.5)/\nu_\Sigma(1.5) = 1/4$. We assume that there are no customers in the system at time 0. The results of the simulation in this case are shown in Fig. 1.

Let us add an initial condition consisting of 25 jobs in each of two classes, with the same service time distributions as jobs arriving in the system in the corresponding class. Notice that their workloads are not even approximately distributed according to the asymptotic proportions stated in Theorem 1. The results are shown in Figure 2.

Now, let us change the initial condition. It still consists of 25 jobs in each class, but their processing times are distributed uniformly on $[0, 3]$, which is even further from the assumptions on the initial condition in Theorem 1. The results are shown in Fig. 3.

Now let us consider the case when $\nu_\Sigma(x^*) = 0$. Assume that $\alpha_1 = 1$, $\alpha_2 = 0.6$. Then $\alpha_\Sigma = 1.6$. The initial service times are uniformly distributed in the interval $[0, 1]$ and $[2/3, 1]$ correspondingly. This gives us $x^* = 1$, $\rho = 1$, $p_1 = 0.625$, $p_2 = 0.375$, $\gamma_1 = 4/7$ and $p_1\gamma_1 = 5/14 \approx 0.357$. We assume that there are no customers in the system at time 0. The results are shown in Fig. 4.

We add an initial condition consisting of 25 jobs in each of two classes, with the same service time distributions as jobs arriving in the system in the corresponding class. The results are shown in Fig. 5.

Let us summarize the results. In Fig. 1 we can observe in the left chart that the proportion of workload of class 1 to the total workload in the system stabilizes at $p_1\nu_1\left(\frac{3}{2}\right)/\nu_\Sigma\left(\frac{3}{2}\right) = 1/4$ after a long time has passed, which confirms that Theorem 1 holds true. Moreover, we can notice that the blue graph in the right chart illustrating the predicted workload of class 1 obtained by applying Theorem 1 "lies close" to the red graph presenting the actual workload of class 1. The prediction at a given time is more accurate, if there are more tasks in the system at this time.

Adding an initial condition in the simulation in Fig. 2 does not noticeably change the situation, even though at time 0 a large amount of workload is not distributed between the two classes according to the proportions required by the assumptions in Theorem 1. We can also observe less instability in the left chart, since with such a number of initial tasks the queue lengths remains at higher levels, where the proportions are more stable. However, in Fig. 3 we can see that an initial condition consisting of tasks greatly different from those arriving in the system results in a notably slower convergence. While the decreasing trend is still visible in the left chart, it is very small in magnitude and in the simulated time horizon the system did not manage to stabilize. Therefore, applying Theorem 1 to predict the proportion of class 1 workload to the total workload in the system gives us an inaccurate approximation and the assumptions for the initial conditions cannot be omitted.

In Fig. 4-5 we can see that if $\nu(x^*) = 0$, we also obtain similar results as above and can make analogous observations. This indicates that Theorem 2 (and even its generalized form, without assuming a zero initial condition) holds true.

VII. Conclusion

In this paper, we have proved a diffusion limit theorem for the measure-valued state descriptor for a single-server queuing system with multiple job classes and bounded processing times of arriving jobs. In particular, we have shown that, under suitable assumptions, the workload and the queue length in the

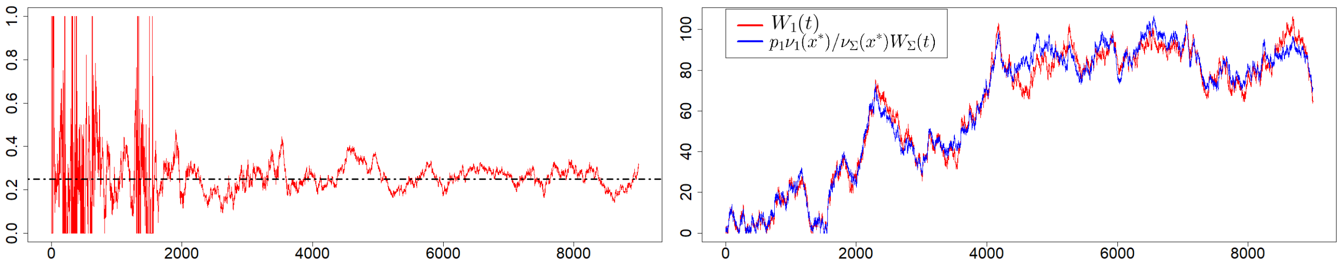

Fig. 1. Simulation in the case of $x^* < \infty, \nu_\Sigma(x^*) > 0$. The left chart illustrates the proportion of workload of class 1 to the total workload in the system as a function of time. The right chart presents the predicted workload of class 1 obtained as a result of applying Theorem 1 (blue) and the actual workload of class 1 (red).

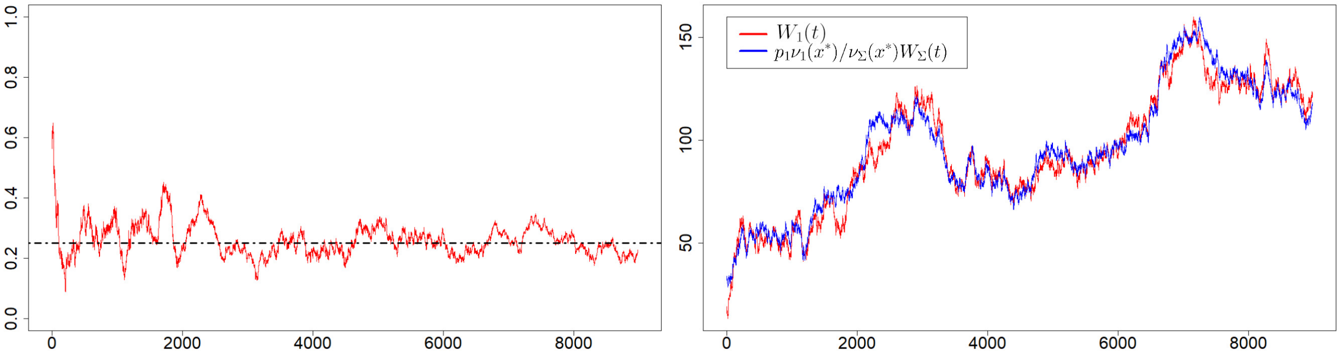

Fig. 2. Simulation in the case of $x^* < \infty, \nu_\Sigma(x^*) > 0$ with a non-zero initial condition. The interpretations of charts is the same as in Fig. 1.

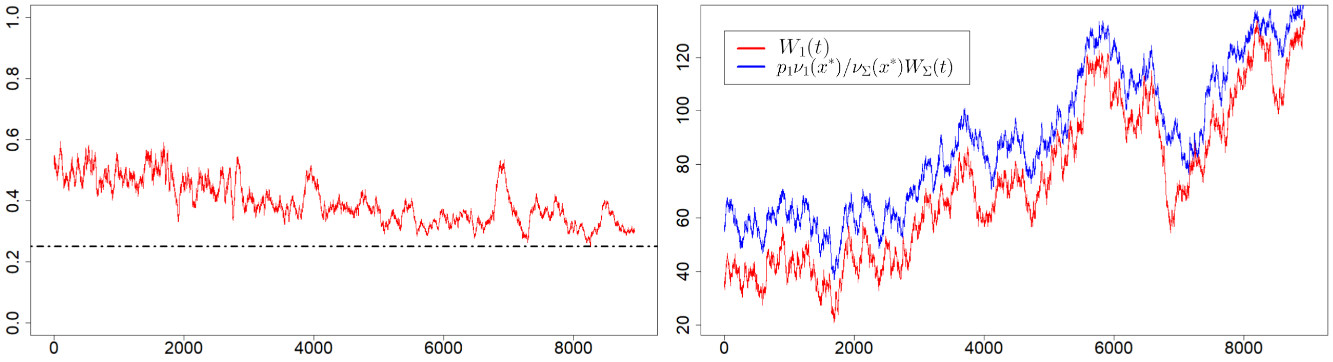

Fig. 3. Simulation in the case of $x^* < \infty, \nu_\Sigma(x^*) > 0$ with a different initial condition.

diffusion limit are divided between these classes according to specific proportions. This result can be applied in practice – it can be used to approximate proportions between workloads of different classes in the long run. The simulations presented in the last section indicate that it should be possible to further generalize Theorem 2 by removing the zero initial condition requirement.

References

[1] M. Agrawal, N. Barsal, M. Harchol-Balter, B. Schroeder, Implementation of SRPT scheduling in Web servers (2001). https://www.cs.cmu.edu/~harchol/Papers/srpt.impl.tech.rept.pdf

[2] M. Agrawal, N. Barsal, M. Harchol-Balter, B. Schroeder, Size-based scheduling to improve Web performance. ACM Transactions on Computer Systems. 21. (2002), https://doi.org/10.1145/762483.762486

[3] R. Atar, A. Biswas, H. Kaspi, K. Ramanan, A Skorokhod map on measure-valued paths with applications to priority queues. Annals of Applied Probability 28:418-481 (2018), https://doi.org/10.1214/17-AAP1309

[4] S. Banerjee, A, Budhiraja, A. L. Puha , Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions, Annals of Applied Probability **32**(4), 2587-2651 (2022), https://dx.doi.org/10.1214/21-AAP1741

[5] P. Billingsley, Convergence of Probability Measures (2nd Edition), John Wiley and Sons, Inc., New York, 1999.

[6] S. Cheng, Y. Cheng, Y. Fu, L. Liu, H. Wang, SFS: Smart OS scheduling for serverless functions, SC '22: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, pp. 1-16 (2022), https://10.1109/SC41404.2022.00047

[7] T. Chojecki, Ł. Kruk, Instability of SRPT, SERPT and SJF queueing networks. Queueing Systems: Theory and Applications 101:57-92 (2022), https://doi.org/10.1007/s11134-021-09733-8

[8] J. Dong, R. Ibrahim, On the SRPT scheduling discipline in many-server

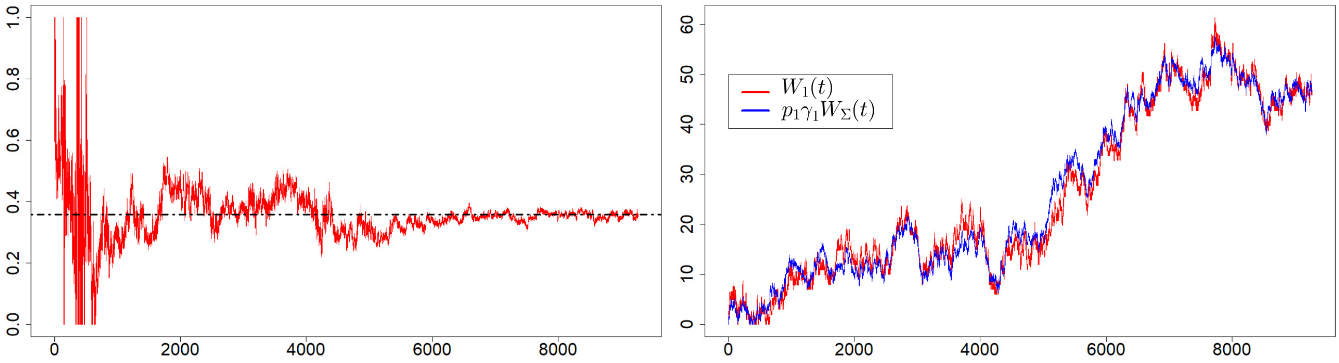

Fig. 4. Simulation in the case of $\nu_\Sigma(x^*) = 0$. As before, the left chart shows the proportion of workload of class 1 to the total workload in the system and the right chart presents the predicted workload of class 1 obtained as a result of applying Theorem 2 (blue) and the actual workload of class 1 (red).

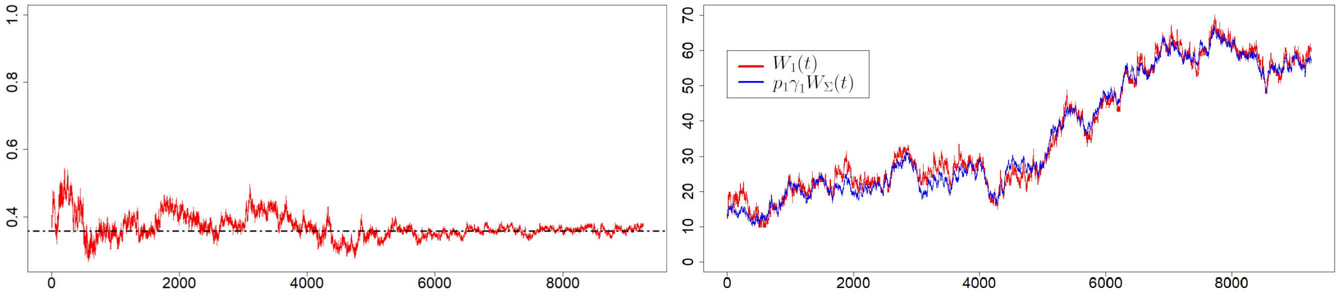

Fig. 5. Simulation in the case of $x^* < \infty, \nu_\Sigma(x^*) = 0$ with a non-zero initial condition.

queues with impatient customers. Management Science 67(12):7708-7718 (2021), https://doi.org/10.1287/mnsc.2021.4110

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

[10] R. Gieroba, Ł. Kruk, Minimality of SRPT networks with resource sharing. WSEAS Transactions on Mathematics 20:74-83 (2021), https://doi.org/10.37394/23206.2021.20.8

[11] R. Gieroba, Ł. Kruk, Local edge minimality of SRPT networks with shared resources, Mathematical Methods of Operations Research, 96:459-492 (2022), https://doi.org/10.1007/s00186-022-00801-0

[12] H. C. Gromoll, Ł. Kruk, A. L. Puha, Diffusion limits for shortest remaining processing time queues, Stochastic Systems 1, 1-16, 2011, https://doi.org/10.1214/10-SSY016.

[13] I. Grosof, Z. Scully, M. Harchol-Balter, SRPT for multiserver systems. Performance Evaluation 127-128:154-175 (2018), https://doi.org/10.1145/3308897.3308902

[14] M. Harchol-Balter, B. Schroeder, Web servers under overload: How scheduling can help. ACM Transactions on Internet Technology 6 (2003), https://doi.org/10.1145/1125274.1125276

[15] D.L. Iglehart, W. Whitt, Multiple channel queues in heavy traffic I, Advances in Applied Probability 2, 150-177, https://doi.org/10.2307/3518347.

[16] Ł. Kruk, Diffusion Limits for SRPT and LRPT Queues via EDF Approximations, QTNA 2019, Lecture Notes in Computer Science, vol. 11688, pp. 263-275. Springer, Cham 2019, https://doi.org/10.1007/978-3-030-27181-7_16

[17] Ł. Kruk, E. Sokołowska, Fluid limits for multiple-input shortest remaining processing time queues, Mathematics of Operation Research 41, 1055-1092, 2016, https://doi.org/10.1287/moor.2015.0768.

[18] L. W. Miller, L. E. Schrage, The queue M/G/1 with the shortest remaining processing time discipline, Operations Research **14**, 670-684 (1966), https://doi.org/10.1287/opre.14.4.670

[19] R. Núñez-Queija: Queues with equally heavy sojourn time and service requirement distributions, Annals of Operations Research **113**, 101-117 (2002), https://doi.org/10.1023/A:1020905810996

[20] M. Nuyens, B. Zwart: A large deviations analysis of the GI/GI/1 SRPT queue, Queueing Systems **54**, 85-97 (2006), https://doi.org/10.1007/s11134-006-8767-1

[21] W. P. Peterson, A heavy traffic limit theorem for networks of Queues with multiple customer types, Mathematics of Operations Research 16, 90-118, 1991, https://doi.org/10.1287/moor.16.1.90

[22] Yu. V. Prohorov, Convergence of random processes and limit theorems in probability theory, Theory of Probability and its Applications 1, 157-214, 1956, https://doi.org/10.1137/1101016.

[23] A. L. Puha, Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling, Annals of Applied Probability **25**(6), 3381–3404 (2015), https://dx.doi.org/10.1214/14-AAP1076

[24] L. E. Schrage, A proof of the optimality of the shortest remaining processing time discipline, Operations Research 16, 687-690 (1968), https://doi.org/10.1287/opre.26.1.197

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

[26] W. Whitt, Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues, Springer-Verlag, New York, 2002.

[27] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/G/1, Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 238-249, 2003, https://doi.org/10.1145/885651.781057