The Annals of Applied Probability
2003, Vol. 13, No. 1, 100–139

FLUID AND HEAVY TRAFFIC DIFFUSION LIMITS FOR A GENERALIZED PROCESSOR SHARING MODEL

BY KAVITA RAMANAN AND MARTIN I. REIMAN

Bell Laboratories

Under fairly general assumptions on the arrival and service time processes, we prove fluid and heavy traffic limit theorems for the unfinished work, queue length, sojourn time and waiting time processes associated with a single station multiclass generalized processor sharing model. The fluid limit of the unfinished work process is characterized by the Skorokhod map associated with a Skorokhod problem formulation of the generalized processor sharing model, while the heavy traffic diffusion limit is characterized using the corresponding extended Skorokhod map. An interesting feature of the diffusion limits is that they may fail to be semimartingales.

1. Introduction. The efficient sharing of a single processing or transmission resource among traffic from several sources is a recurrent issue in high speed networks. One scheme that has been proposed is known as generalized processor sharing (GPS). GPS has several desirable characteristics that make it a reasonable candidate for a resource sharing policy. First, it provides a minimal guaranteed throughput rate to each source, regardless of the behavior of other sources. In addition, it allows for discrimination between the various sources in the sense that the minimal guaranteed throughputs can be different for each source. Furthermore, it ensures that the resource is fully utilized as long as there is a backlog for any of the sources.

Given a single server that can process one unit of work per unit of time and that is being shared by J $(1 < J < \infty)$ sources, the information needed to implement the GPS policy is contained in the weight vector $\alpha = (\alpha_1, \ldots, \alpha_J)$. When all sources have a backlog of work, source j is allotted a fraction α_j of the total capacity of the server. When some sources achieve no backlog by using less than their allotted capacity, the remaining service capacity of the server is split among the other sources in proportion to their α_i's. A more precise description of the model is given in Section 2, where a slight generalization is also considered. The GPS policy was introduced in [17] and [18], where worst case bounds on the delay and unfinished work for deterministically shaped traffic were also calculated. Estimates of the unfinished work for stochastically bounded traffic were derived in [29], and large deviation results were presented in [2], [8] and [16].

Received September 2001; revised May 2002.

AMS 2000 subject classifications. Primary 60F05, 60F17; secondary 60K25, 90B22, 68M20.

Key words and phrases. Diffusion approximations, heavy traffic, fluid limits, generalized processor sharing, queueing networks, Skorokhod problem, Skorokhod map, extended Skorokhod problem, semimartingales.

The analysis in [8] was based on a Skorokhod problem formulation of the GPS model, in which the Skorokhod map associated with the Skorokhod problem was shown to take the input processes into the buffer content process. The Skorokhod problem provides a convenient representation in many queueing problems. Skorokhod problems associated with most queueing applications previously considered in the literature can be equivalently formulated as dynamic complementarity problems associated with certain reflection matrices [14, 15]. The Skorokhod maps associated with these Skorokhod problems are usually referred to as reflection maps, and are well defined (though possibly multivalued) on the space of right-continuous paths with left limits if and only if the associated reflection matrices satisfy what is known as the completely-$\mathcal{S}$ condition [1, 14, 15, 22]. The latter condition implies in particular that the so-called $\mathcal{V}$-set associated with the Skorokhod problem ([20], Definition 3.4) is empty. For some classes of queueing networks, such as single-class open queueing networks, the corresponding reflection matrices satisfy additional regularity conditions along with the completely-$\mathcal{S}$ condition, which guarantee that the corresponding reflection maps are Lipschitz continuous and consequently single-valued [10, 14]. This enables the use of continuous mapping techniques to characterize fluid and heavy traffic limits of the associated queueing processes as images under the reflection map of corresponding unconstrained diffusions that arise as functional limits of the primitive input processes [6, 21].

In this paper we also use a representation based on the Skorokhod problem to prove fluid and diffusion limits for GPS, with the latter valid under a heavy traffic condition. The Skorokhod problem associated with the GPS model cannot be posed as a standard dynamic complementarity problem and hence does not fall into the reflection mapping setting ([9], Section 2.4). Nevertheless, using techniques developed in [9] for more general Skorokhod problems, it was shown in [10] that the GPS Skorokhod map is well defined and Lipschitz continuous on the subspace of bounded variation functions. Thus a continuous mapping approach can be used to obtain the a.s. fluid limit for the associated unfinished work process and to represent it as the image of an affine function under the GPS Skorokhod map (see Theorem 4.3). However, the Skorokhod problem appears to be inadequate to characterize the diffusion limit of the GPS model. In particular, since the GPS Skorokhod problem has a nonempty $\mathcal{V}$-set ([20], Lemma 4.10) and the associated Skorokhod map is defined only on a strict subset of right-continuous paths that does not include certain paths of unbounded variation ([20], Theorem 3.11), the Skorokhod map associated with the GPS Skorokhod problem cannot in general be applied to diffusions. Instead we use a generalization of the Skorokhod problem, called the extended Skorokhod problem, that was formulated in [20], Definition 2.2. This generalization allows one to obtain a Lipschitz continuous mapping (the GPS extended Skorokhod map) that is single-valued and well defined on all continuous paths and that can therefore be applied to

Brownian motion, yielding the heavy traffic diffusion limit for the unfinished work process (see Theorem 4.14). The advantage of the continuous mapping technique is that it yields not just weak convergence, but almost sure convergence of the scaled processes to the diffusion limit. The limit process is a special case of a more general class of reflected diffusion processes obtained by applying extended Skorokhod maps to unconstrained diffusions. This class of reflected diffusion processes is introduced and studied in [20]. Under suitable regularity conditions (that are satisfied in particular by the heavy traffic limit obtained herein), it is shown in [20] that these reflected processes are continuous strong Markov processes. However, a unique feature of these processes is that they may fail to be semimartingales ([20], Section 1.1B). In particular it follows from [27], Theorem 2, that the diffusion limit for the two-dimensional GPS model is not a semimartingale. This is in contrast to more conventional heavy traffic limit theorems obtained previously (under similar assumptions on the input processes), where the diffusion limits are semimartingale reflected Brownian motion [14, 22].

The Skorokhod and extended Skorokhod map representations of limits of the unfinished work process also yield fluid and diffusion limit theorems for the busy time processes. The fluid and diffusion limits for the queue length, sojourn time and waiting time processes are then obtained by representing them as suitable functionals of the unfinished work and busy time processes, and once again invoking continuous mapping theorems.

There are two other disciplines which have names that contain the term "processor sharing" and for which heavy traffic diffusion limits have been proved. The head-of-the-line proportional processor sharing discipline (HLPPS) was introduced in [4, 28]. In HLPPS, only the customers at the head of the line in each class are served, and for each such customer the service rate is proportional to the number of customers of that class in the system. The generalized head-of-the-line proportional processor sharing (GHLPPS) discipline operates similarly to HLPPS except that each class has an associated (positive) weight and the service rate allocated to each class is proportional to the number of customers of that class in the system multiplied by the weight for that class. [With a single class, both HLPPS and GHLPPS reduce to first-in-first-out (FIFO).] Both of these disciplines exhibit state space collapse in the heavy traffic limit in the sense that the stochastic behavior of the limit process associated with each discipline is captured by a one-dimensional reflected Brownian motion. In contrast, the limit that we obtain here for the GPS discipline does not exhibit state space collapse. Indeed, the limit process that we obtain is a J-dimensional reflected diffusion.

This paper is organized as follows. In Section 2 we introduce the GPS model in greater detail and characterize the associated unfinished work, queue length, sojourn time and waiting time processes. The Skorokhod problem, Skorokhod

map, extended Skorokhod problem and extended Skorokhod map are described in Section 3, where we also show that the unfinished work process can be represented as the Skorokhod map applied to a simpler unconstrained process. The fluid and heavy traffic diffusion limit theorems are proved in Section 4.

We now collect together some of the notational conventions used in this paper. The sets of nonnegative reals, nonnegative integers and positive integers are denoted by $\mathbb{R}_+$, $\mathbb{Z}_+$ and $\mathbb{N}$, respectively. Given $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minimum of a and b. Vector inequalities are to be interpreted componentwise. The standard orthonormal basis in $\mathbb{R}^J$ is represented by $\{e_i,\ i = 1, \ldots, J\}$, and the J-dimensional nonnegative orthant $\mathbb{R}_+^J$ is equal to $\{x \in \mathbb{R}^J : x \geq 0\}$. Let $\mathcal{J}$ denote the set $\{1, \ldots, J\}$. Given $E \subset \mathbb{R}^J$, $\mathcal{D}([0, \infty) : E)$ represents the space of E-valued right-continuous functions with left limits. Unless indicated otherwise, we will assume that $\mathcal{D}([0, \infty) : E)$ is equipped with the topology of uniform convergence on compact sets (frequently abbreviated to u.o.c.). We use $1_{\{E\}}$ to represent the indicator function of the set E, which is equal to 1 on E and is 0 elsewhere. For $f \in \mathcal{D}([0, \infty) : E)$, as usual $f(t-) = \lim_{s \uparrow t} f(s)$. For $t \in [0, \infty)$, $|f|(t)$ denotes the total variation of f on $[0, t]$ with respect to the Euclidean norm $|\cdot|$ on $\mathbb{R}^J$. The composition of two functions f and g is as usual denoted by $f \circ g$. We sometimes consider functions that are not defined on all of $[0, \infty)$. In this situation $\mathrm{dom}(f)$ is used to denote the domain of the function, that is, the set of points for which the function is defined. The identity function $\iota : [0, \infty) \to [0, \infty)$ is such that $\iota(t) = t$ for all $t \in [0, \infty)$. Given a set $A \subset \mathbb{R}^J$, $\overline{\mathrm{co}}[A]$ denotes the closure of the convex hull of A, $\mathrm{cone}[A]$ is the cone generated by A and A° is the interior of A. Finally, given a matrix D we use D' to denote its transpose.

2. Model description. We consider a single server queueing system with J customer classes, where $1 < J < \infty$. Each customer arriving into the system brings in a certain amount of work that is measured in terms of the amount of time required to process it using the server's total processing capacity, which is assumed without loss of generality to be 1. The server processes the incoming work using the GPS scheduling discipline, which is described in Section 2.1. The work of class i customers is stored in the class i buffer, which is assumed to be infinite. We study four processes associated with this model—the unfinished work U, the queue length Q, the sojourn time V and the waiting time W. For $i \in \mathcal{J}$ and $t \in [0, \infty)$, $U_i(t)$ is defined to be the amount of work of class i that is in the system at time t and $Q_i(t)$ is the number of class i customers in the system at time t. For $i \in \mathcal{J}$ and $k \in \mathbb{N}$, $V_i(k)$ is the sojourn time (i.e., the time elapsed from entry to exit) and $W_i(k)$ is the waiting time (i.e., the time elapsed from entry to the beginning of service) of the kth class i customer that arrived into the system after time 0. In Section 2.2 we characterize these four processes in terms of primitives such as the cumulative workload arrival process H and the cumulative customer arrival process A, and state our basic assumptions.

2.1. *The GPS discipline.* For $E \subseteq \mathcal{I}$ we define α_i^E to be the fraction of the capacity of the server that is given to class i when the set of empty buffers is equal to E. We assume that the processor is work-conserving, so that $\sum_{i \notin E} \alpha_i^E = 1$ when $E \neq \mathcal{I}$. In this paper, we focus on the case when the fractions α_i^E are determined in the following manner by two weight vectors $\alpha \in [0, 1]^J$ and $\beta \in (0, 1]^J$ that satisfy $\sum_{i \in \mathcal{I}} \alpha_i = \sum_{i \in \mathcal{I}} \beta_i = 1$. Given the weight vectors, for $E = \varnothing$ (i.e., when no queue is empty) we define $\alpha_i^{\varnothing} = \alpha_i$ and for $E \subseteq \mathcal{I}$ we let

$$\alpha_i^E \doteq \begin{cases} \alpha_i + \dfrac{\beta_i}{\sum_{j \notin E} \beta_j} \left(\sum_{i \in E} \alpha_i \right), & \text{for } i \notin E, \\ 0, & \text{otherwise.} \end{cases}$$

For all $E \subset \mathcal{I}$, $\alpha_i^E \geq \alpha_i$ for $i \in \mathcal{I} \setminus E$ and $\sum_{i \notin E} \alpha_i^E = \sum_{i \in \mathcal{I}} \alpha_i = 1$. Thus α_i represents the minimum guaranteed rate assigned to class i. Any excess capacity is split among the remaining classes in proportion to the corresponding components of the vector β. The condition $\beta_i > 0$ for each $i \in \mathcal{I}$ is required to ensure that the processor is work-conserving. On the other hand, we allow $\alpha_i = 0$ for some $i \in \mathcal{I}$. This represents the case when the ith class is of relatively low priority and only receives service when one of the high priority classes (with $\alpha_j > 0$) does not require all of its assigned capacity. Note, however, that if there is more than one high priority class, then a low priority class could receive service even when a high priority class is backlogged. In particular this situation would arise if one high priority class is empty (causing its nominal capacity to be reassigned to all the other nonempty classes in proportion to their β values), while another high priority class is still backlogged. Also note that since β must lie in $(0, 1]^J$ for the discipline to be work-conserving, our model cannot capture more than one level of priority. Multiple levels of priority (i.e., beyond just high vs. low) may be captured by a more general definition of the α_i^E, $E \subset \mathcal{I}$, but this is not considered here. When $\alpha = \beta$ this model reduces to the well known GPS model (see, e.g., [8, 17, 18]). In this case, the requirement that the discipline be work-conserving imposes the restriction that $\alpha = \beta \in (0, 1]^J$ and hence it is not possible to incorporate even one level of priority into that model. By some abuse of terminology we will also refer to the above more general model as the GPS model.

2.2. *Characterization of the processes.* In Section 2.2.1 we introduce the primitives of the GPS model such as the cumulative workload, arrival and service time processes, as well as the residual service time vector. In Sections 2.2.2 and 2.2.3 we state the characterizing equations for the unfinished work and queue length processes, respectively. In Section 2.2.4 we define the sojourn and waiting time processes. We assume that all processes are measurable functions defined on the probability space $(\Omega, \mathcal{F}, P)$.

2.2.1. *The primitive input processes.* Let A be the $\mathcal{D}([0,\infty):\mathbb{Z}_+^J)$-valued measurable function on the probability space $(\Omega,\mathcal{F},P)$ such that $A_i(t)$ represents the cumulative number of class i customers that have come into the system in the interval $[0,t]$ and let H be the $\mathcal{D}([0,\infty):\mathbb{R}_+^J)$-valued process such that $H_i(t)$ represents the cumulative workload brought into the system by class i in the interval $[0,t]$. Consider the sequence of random variables $\{S_i(n),\ n\in\mathbb{Z}_+\}$, $i\in\mathcal{I}$, where $S_i(0)=0$ and for $n\in\mathbb{N}$, $S_i(n)$ is the cumulative amount of time required to process the first n class i customers in the system (including those already in the system at time 0) if the processor were entirely devoted to class i. Also consider the $\mathcal{D}([0,\infty):\mathbb{Z}_+^J)$-valued process L, where $L_i(t)$ is the number of class i customers (including those in the queue at time 0) that are fully served after the processor has devoted t units of cumulative service effort to class i. For simplicity, we assume throughout that P a.s. no customer has zero service time.

The processes A, H, S and L are related in the following manner. First note that if $Q_i(0)$ is the initial number of customers in the class i buffer at time 0, then the unfinished work of class i that is in the system at time 0 can be expressed as $U_i(0)=S_i(Q_i(0))$. The cumulative class i workload can then be expressed as

$$H_i(t)=S_i\big(A_i(t)+Q_i(0)\big)-S_i\big(Q_i(0)\big). \tag{2.1}$$

Conversely, given any piecewise-constant cumulative workload process H and residual service time vector $r\in\mathbb{R}_+^J$, under the assumption that P a.s. no arrivals take place simultaneously, since no customer has a zero service time, the processes A and S can be recovered as follows. For $i\in\mathcal{I}$, P a.s. $A_i(t)$ is equal to the number of jump points of H_i in the interval $[0,t]$, $S_i(0)=0$ and for $n\in\mathbb{N}$,

$$S_i(n)\doteq\begin{cases}\displaystyle\sum_{k=1}^{n}s_i(k), & \text{if } Q_i(0)=0,\\[2ex] \displaystyle\sum_{k=-Q_i(0)+1}^{n-Q_i(0)}s_i(k)+r_i, & \text{if } Q_i(0)>0,\end{cases} \tag{2.2}$$

where $s_i(0)=0$, for $k\in\mathbb{N}$, $s_i(k)>0$ is the size of the kth jump of H_i, for $k=1,\dots,Q_i(0)-1$, $s_i(-k)>0$ is the amount of work that was brought in by the $(Q_i(0)-k)$th customer in the queue at time 0 and $r_i>0$ denotes the residual time required to complete the processing of the head-of-the-line customer in the queue at time 0. The processes S and L clearly satisfy the relationships

$$L_i(t)=\sup\big\{n\geq 0: S_i(n)\leq t\big\} \tag{2.3}$$

for $n\in\mathbb{N}$,

$$S_i(n)=\inf\big\{t>0: L_i(t)\geq n\big\}, \tag{2.4}$$

and since no customer has zero service time, P a.s.,

$$L_i\big(S_i(n)\big)=n. \tag{2.5}$$

Finally, suppose $\tau_i(k)$ is the time interval between the $(k-1)$th and kth jumps of A_i and for $n \in \mathbb{N}$ define

$$C_i(n) \doteq \sum_{k=1}^{n} \tau_i(k).$$

Then it is clear that

$$A_i(t) = \sup\{n \geq 0 : C_i(n) \leq t\}. \tag{2.6}$$

The ultimate assumptions that we require on the primitive processes are quite weak: A, S, L and H must satisfy functional strong laws of large numbers and functional central limit theorems. The simplest concrete example where this happens is where the J component processes of H, A and L form compound renewal processes (see, e.g., [21], Lemma 2). In the next three sections we describe the equations satisfied by the unfinished work, queue length, sojourn time and waiting time processes. For the description of the unfinished work process U in Section 2.2.2 and the related results in Sections 4.1.1 and 4.2.1 we do not make any assumptions about how the server's effort is apportioned among the customers of each class in the system, as long as the total effort devoted to the class is as dictated by the GPS discipline. (Recall that under our assumption that $\beta \in (0, 1]^J$ the GPS discipline is work-conserving.) However, the description of the queue length, sojourn time and waiting time given in Sections 2.2.3 and 2.2.4, and the related results in Sections 4.1.2, 4.2.2, 4.1.3 and 4.2.3 are based on the assumption that customers within each class are served in a FIFO order. Under this FIFO restriction within each class, the GPS discipline considered here with $\alpha = \beta$ has also been referred to by various authors as the generalized head-of-the-line processor sharing discipline or as the head-of-the-line generalized processor sharing discipline [5, 8, 11].

2.2.2. *The unfinished work process.* In the definition and analysis of the unfinished work process recall that no assumption is made as to how the service allocated to a class is divided among the customers present in that class. For $E \subseteq \mathcal{I}$, the cumulative idle time $I_E(t)$ is the amount of time in $[0, t]$ that the set of empty buffers is equal to E. We characterize the unfinished work process U, and idle time processes I_E, $E \subseteq \mathcal{I}$, as solutions to the following equations. For $i \in \mathcal{I}$,

$$\begin{aligned} U_i(t) &= U_i(0) + H_i(t) - \sum_{E \subset \mathcal{I}: i \notin E} \alpha_i^E I_E(t), \\ I_E(t) &\doteq \int_0^t 1_{\{E(s)=E\}}\, ds \qquad \text{for } E \subseteq \mathcal{I}, \\ E(s) &\doteq \{i \in \mathcal{I} : U_i(s) = 0\}. \end{aligned} \tag{2.7}$$

The busy time process T_i defined by

$$T_i \doteq \sum_{E \subset \mathcal{I}: i \notin E} \alpha_i^E I_E \tag{2.8}$$

represents the cumulative amount of service given to class i.

We suppose that the cumulative workload arrival process H and initial conditions satisfy the following assumptions.

ASSUMPTION 2.1. 1. $U(0) \in \mathbb{R}_+^J$.

2. $H \in \mathcal{D}([0,\infty) : \mathbb{R}_+^J)$ is nondecreasing and piecewise constant with $H(0) = 0$.

3. H has a finite number of jump points in every finite interval, almost surely.

We now show that if Assumption 2.1 holds, then representation (2.7) uniquely characterizes the set of processes $(U, I_E, E \subseteq \mathcal{I})$.

LEMMA 2.2. *Suppose Assumption* 2.1 *holds. Then there exists a unique set of processes* $(U, I_E,\ E \subseteq \mathcal{I})$ *that satisfies* (2.7).

PROOF. Define $I_E(0) = 0$ for all $E \subseteq \mathcal{I}$. Since $H_i(0) = 0$, (2.7) is satisfied by U for $t = 0$. We now proceed by induction. Let $0 < t_1 < t_2 < \cdots$ denote the set of jump points of H, let $t_0 = 0$ and let $\mathcal{J} = \{t_n,\ n \geq 0\}$. Suppose there exists a unique solution $(U, I_E,\ E \subseteq \mathcal{I})$ to (2.7) on the interval $[0, t_n]$ for some $t_n \in \mathcal{J}$. For $t \in [t_n, \infty)$, let

$$\tilde{U}_i^0(t) \doteq U_i(t_n) - \alpha_i^{E(t_n)}(t - t_n) + H_i(t) - H_i(t_n).$$

Let $\kappa^0(t) \doteq \{i \in \mathcal{I} : \tilde{U}_i^0(t) = 0\}$, and note that $\kappa^0(t_n) = E(t_n)$. Define $\sigma_1 \doteq t_{n+1} \wedge \inf\{t \geq t_n : \kappa^0(t) \neq E(t_n)\}$. Observe that for $t \in [t_n, t_{n+1})$, $H_i(t) = H_i(t_n)$. If $E(t_n) = \mathcal{I}$, then by definition $U_i(t_n) = 0$ and $\alpha_i^{E(t_n)} = 0$ for every $i \in \mathcal{I}$, and thus $\tilde{U}_i^0(t) = 0$ for $t \in [t_n, t_{n+1})$ and $\sigma_1 = t_{n+1} \wedge \infty = t_{n+1}$. On the other hand, if $E(t_n) \neq \mathcal{I}$, then since $\alpha_i^{E(t_n)} = 0$ for $i \in E(t_n)$ and $\sum_{i \notin E(t_n)} \alpha_i^{E(t_n)} = 1 > 0$, $\tilde{U}_i^0(t) = U_i(t_n) = 0$ for $i \in E(t_n)$ and $\sum_{i \notin E(t_n)} \tilde{U}_i^0(t)$ is monotonically decreasing for $t \in [t_n, t_{n+1})$. Moreover, since $t_{n+1} > t_n$, $\tilde{U}_i^0(t_n) > 0$ for $i \notin \kappa^0(t_n)$ and $\alpha_i^E \leq 1$ for every $E \subseteq \mathcal{I}$, clearly $\sigma_1 > t_n$. Let $\sigma_0 \doteq t_n$ and for $k \geq 1$, define

$$\tilde{U}_i^k(t) \doteq \tilde{U}_i^{k-1}(\sigma_k) - \alpha_i^{E(\sigma_k)}(t - \sigma_k) + H_i(t) - H_i(\sigma_k), \tag{2.9}$$

$\kappa^k(t) \doteq \{i \in \mathcal{I} : \tilde{U}_i^k(t) = 0\}$ and $\sigma_{k+1} \doteq t_{n+1} \wedge \inf\{t \geq \sigma_k : \kappa^k(t) \neq E(\sigma_k)\}$. Let $k^* = \max\{k \in Z_+ : \sigma_k < t_{n+1}\}$. From the monotonicity properties of $\tilde{U}_i^k$ on $[t_n, t_{n+1})$, it is clear that $k^* \leq J$ and $\sigma_{k^*+1} = t_{n+1}$. If $t_{n+1} < \infty$, we extend the definition of (U, I_E) to $[0, t_{n+1}]$ as follows. For $k = 0, 1, \ldots, k^*$ and $t \in [\sigma_k, \sigma_{k+1}]$, let $U(t) \doteq \tilde{U}^k(t)$ and for $E \subseteq \mathcal{I}$, define

$$I_E(t) \doteq \begin{cases} I_E(\sigma_k) + t - \sigma_k, & \text{for } E = E(\sigma_k), \\ I_E(\sigma_k), & \text{otherwise.} \end{cases}$$

It is easy to verify that (U, I_E) defined above satisfy (2.7) on $[0, t_{n+1}]$. If $t_{n+1} = \infty$ the definition above, with the obvious restriction that it is valid only for $t \in [\sigma_k, \sigma_{k+1})$ when $k = k^*$, can be used to extend (U, I_E) to $[0, \infty)$. Since either $t_{n+1} = \infty$ or $t_{n+1} \in \mathcal{J}$, and $t_n \to \infty$ as $n \to \infty$ (because H has only a finite number of jumps in any finite interval), the existence of (U, I_E) satisfying (2.7) follows by induction. Uniqueness is a simple consequence of the construction. □

2.2.3. *The queue length process.* Recall that in this section and the next, it is assumed that the service is FIFO within each class, in the sense that the service allocated to each class by the GPS discipline is given entirely to the head-of-the-line customer in that class. Let A and L be the arrival and service processes described in Section 2.2.1. Then the queue length process $Q \in \mathcal{D}([0, \infty) : \mathbb{Z}_+^J)$ satisfies the equation

$$Q_i(t) = Q_i(0) + A_i(t) - L_i\big(T_i(t)\big), \tag{2.10}$$

where T_i is defined by (2.8). Note that since for every $s \in [0, \infty)$, $U_i(s) = 0$ if and only if $Q_i(s) = 0$, one can equivalently set

$$\begin{aligned} I_E(t) &\doteq \int_0^t 1_{\{E(s)=E\}}\, ds \qquad \text{for } E \subseteq \mathcal{I}, \\ E(s) &\doteq \big\{i \in \mathcal{I} : Q_i(s) = 0\big\}, \end{aligned} \tag{2.11}$$

in the definition (2.8) of T_i. Thus T is determined completely by Q and so equations (2.8), (2.10) and (2.11) are self-contained coupled equations.

We now introduce assumptions on the cumulative arrival and service processes that are sufficient for the existence of a unique solution (Q, T) to the set of equations (2.8), (2.10) and (2.11).

ASSUMPTION 2.3. 1. $Q(0) \in Z_+^J$.

2. $A, L \in \mathcal{D}([0, \infty) : Z_+^J)$ are nondecreasing and piecewise constant with $A(0) = L(0) = 0$.

3. For $i \in \mathcal{I}$ and $t \in [0, \infty)$, $L_i(t) - L_i(t-) \le 1$.

LEMMA 2.4. *Suppose Assumption* 2.3 *holds. Then there exists a unique set of processes* (Q, T) *that satisfy* (2.10).

PROOF. The proof of this lemma is similar to the proof of Lemma 2.2, and thus we provide only a rough sketch. Let $t_0 = 0$. Since $A_i(0) = L_i(0) = 0$, it is clear that (2.10) and (2.11) are satisfied for $t = t_0$. Suppose there exist $t_n \ge 0$ and (Q, T) that satisfy (2.10) and (2.11) for $t \in [0, t_n]$. Let $E^* \doteq E(t_n)$ and for

$t \in [t_n, \infty)$, define

$$\tilde{T}_i(t) \doteq \begin{cases} \alpha_i^{E^*}(t - t_n) + T_i(t_n), & \text{if } i \notin E^*, \\ T_i(t_n), & \text{if } i \in E^*, \end{cases}$$

$$\tilde{Q}_i(t) \doteq Q_i(0) + A_i(t) - L_i\big(\tilde{T}_i(t)\big),$$

$$\kappa(t) \doteq \big\{i \in \mathcal{I} : \tilde{Q}_i(t) = 0\big\},$$

and let $t_{n+1} \doteq \inf\{t \geq t_n : \kappa(t) \neq E^*\}$. For $t \in [t_n, t_{n+1}]$, let $T(t) \doteq \tilde{T}(t)$ and $Q(t) \doteq \tilde{Q}(t)$. Then $t_{n+1} > t_n$ and it is easy to see that (Q, T) satisfy (2.10) and (2.11) on $[0, t_{n+1}]$. As in Lemma 2.2, this can be extended to $[0, \infty)$ by induction. $\square$

2.2.4. *The sojourn time and waiting time processes.* Recall from Section 2.2.1 that $C_i(k)$ is the kth jump point of A_i and represents the arrival time of the kth customer after time 0. Under the assumption made in the last section that the service is FIFO within each class, for $k \in \mathbb{N}$ and $i \in \mathcal{I}$, the sojourn time of the kth class i customer to arrive after 0 can be written as

$$V_i(k) = \inf\big\{s \geq 0 : T_i(s) - T_i\big(C_i(k)\big) \geq U_i\big(C_i(k)\big)\big\} - C_i(k)$$

and likewise, the kth class i customer's waiting time can be expressed as

$$W_i(k) = \inf\big\{s \geq 0 : T_i(s) - T_i\big(C_i(k)\big) \geq U_i\big(C_i(k) - \big)\big\} - C_i(k).$$

To represent the sojourn time and waiting time processes more succinctly it will be convenient to introduce the mapping F defined below. Given a nondecreasing function $f \in \mathcal{D}([0, \infty) : \mathbb{R})$ with $f(0) = 0$, $F[f]$ is defined by

$$F[f](t) \doteq \inf\big\{s \geq 0 : f(s) \geq t\big\} \tag{2.12}$$

for $t \in \operatorname{dom}(F[f])$, where

$$\begin{aligned} \operatorname{dom}(F[f]) &= \big\{t \in [0, \infty) : f(s) \geq t \text{ for some } s \in [0, \infty)\big\} \\ &= \bigcup_{M < \infty} \Big[0, \sup_{s \in [0, M]} f(s)\Big]. \end{aligned} \tag{2.13}$$

Thus we can rewrite

$$V_i(k) = F[T_i] \circ (T_i + U_i) \circ C_i(k) - C_i(k) \tag{2.14}$$

and

$$W_i(k) = F[T_i] \circ \big(T_i \circ C_i(k) + U_i\big(C_i(k) - \big)\big) - C_i(k). \tag{2.15}$$

Then V_i is defined on the set

$$\operatorname{dom}(V_i) = \big\{k \in \mathbb{Z}_+ : (T_i + U_i) \circ C_i(k) \leq T_i(s) \text{ for some } s \in [0, \infty)\big\},$$

and an analogous statement holds for W_i.

If $f \in \mathcal{D}([0,\infty) : \mathbb{R})$ is strictly increasing and has range $[0,\infty)$, then it is easy to see that $F[f] = f^{-1}$ [in particular this follows from (4.26) in the proof of Lemma 4.10], where the inverse function $f^{-1} \in \mathcal{D}([0,\infty) : \mathbb{R})$ is given by

$$f^{-1}(t) \doteq \inf\{s \geq 0 : f(s) > t\} \qquad \text{for } t \in [0,\infty). \tag{2.16}$$

However, for general nondecreasing $f \in \mathcal{D}([0,\infty) : \mathbb{R})$, $F[f]$ may not be defined on the whole of $[0,\infty)$. Indeed, if there exists $t < \infty$ such that $f(s) = M$ for all $t \geq s$, then $\mathrm{dom}(F(f)) = [0, M]$. On the other hand, if $\lim_{s\uparrow\infty} f(s) \uparrow M$ for some $M \in [0,\infty]$, but $f(s) \neq M$ for any $s \in [0,\infty)$, then $\mathrm{dom}(F(f)) = [0, M)$. In particular, if f is unbounded, then $M = \infty$ and the domain is $[0,\infty)$. Another problem when f is not strictly increasing is that $F[f]$ need not be right-continuous. However, as shown in Section 4.1.3, the scaled sojourn and waiting times have a representation in terms of the composition of $F[f]$ with g, where g is a piecewise-constant nondecreasing function. In this case $F[f] \circ g$ is right-continuous (e.g., see the remark in [19], page 950). Thus the processes of interest to us that have a representation in terms of $F[f]$ are right-continuous with left limits.

3. The GPS Skorokhod and extended Skorokhod problems. As mentioned in the Introduction, the Skorokhod problem and its generalization, the extended Skorokhod problem, provide a convenient representation for many constrained processes. The GPS model for $\alpha = \beta$ with stochastic fluid inputs was analyzed in [8]. It was shown there that the mapping taking the inputs to the buffer content can be represented in terms of a Skorokhod problem. Here too, in Section 3.3, we derive a similar representation for the unfinished work process U associated with the slight generalization of the GPS model. In Section 3.1 we define the Skorokhod problem and extended Skorokhod problem associated with the GPS model. In Section 3.2 we summarize some useful properties of the extended Skorokhod map associated with the GPS extended Skorokhod problem, which are then used in Section 4 to establish fluid and heavy traffic limit theorems.

3.1. *Definition of the Skorokhod and extended Skorokhod problems.* Roughly speaking, given the closure G of a domain in $\mathbb{R}^J$, directions of constraint $d(\cdot)$ on ∂G and a path ψ, the solutions to both the associated Skorokhod and extended Skorokhod problems define constrained versions ϕ of ψ that lie in G. For the Skorokhod problem the constraint mechanism $\phi - \psi$ must be of bounded variation and must act along the direction $d(\phi(s))$ using the "least effort" required to keep ϕ in G. The solution to the extended Skorokhod problem relaxes the requirement on the constraint mechanism, imposing only that the increments of the constraining process $\phi - \psi$ in every interval $[s,t]$ lie in the convex hull of the directions of

constraint $d(\phi(u))$ for $u \in (s,t]$. Skorokhod or extended Skorokhod problems on polyhedral domains with a constant direction of constraint on each face can be represented by a finite set of triplets $\{(d_i, n_i, c_i),\ i = 1, \ldots, K\}$. The domain of such a Skorokhod problem is defined to be

$$G \doteq \bigcap_{i=1}^{K} \{x \in \mathbb{R}^J : \langle x, n_i \rangle \geq c_i\}$$

and the set of constraining directions for any point x on the boundary ∂G is given by

$$d(x) \doteq \left\{ \sum_{i \in I(x)} a_i d_i : a_i \geq 0 \right\}, \tag{3.1}$$

where

$$I(x) \doteq \{i = 1, \ldots, K : \langle x, n_i \rangle = c_i\}.$$

See [7, 20] for definitions of the Skorokhod and extended Skorokhod problems on more general domains. For conciseness we define $d^1(x) \doteq d(x) \cap \{x \in \mathbb{R}^J : |x| = 1\}$. Below we specialize to the case of the GPS model.

The GPS model described in Section 2 is characterized by two weight vectors, $\alpha \in [0,1]^J$ and $\beta \in (0,1]^J$, that satisfy $\sum_{i=1}^J \alpha_i = \sum_{i=1}^J \beta_i = 1$, where α_i is the minimum fraction of the total processing capacity guaranteed to the ith class and any excess unused capacity is redistributed among the nonempty classes in proportion to their β_i values. For $i = 1, \ldots, J$, let $n_i \doteq e_i$ and

$$d_i \doteq e_i - \sum_{j \neq i} \frac{\beta_j e_j}{(1 - \beta_i)}. \tag{3.2}$$

Moreover, let $n_{J+1} \doteq d_{J+1} \doteq \sum_{k=1}^J e_k/\sqrt{J}$ and $c_i = 0$ for $i = 1, \ldots, J+1$. We will refer to the Skorokhod and extended Skorokhod problems that have the representation $\{(d_i, n_i, c_i),\ i = 1, \ldots, J+1\}$ as the GPS Skorokhod and extended Skorokhod problems, respectively, associated with the weight vector β. Note that the GPS Skorokhod problem associated with the weight vector β is specified by $K = J + 1$ directions of constraint, has domain $G = \mathbb{R}_+^J$ and the directions of constraint on the boundary are defined by (3.2) and (3.1). The set of directions of constraint on the boundary describes how service is reallocated when one or more buffers is empty. Since this reallocation is determined solely by the weight vector β, the description of the GPS Skorokhod problem only depends on β (and not on α). Thus the GPS Skorokhod problem defined here corresponds exactly to the GPS Skorokhod problem defined in [8, 10], which considers the GPS model for the case $\alpha = \beta$ (and denotes this common weight vector by ρ), and the results derived for the GPS Skorokhod problem in [10] can be directly applied here. On the other hand, our model is more general than the one considered in [8, 10] due to the fact

that the nominal service allocation to each class (which equals the actual allocation when all buffers are nonempty) is determined by another weight vector α, in general different from β. Thus α determines the drift of the unconstrained process in the GPS Skorokhod problem representation of the unfinished work process U [see (3.5) and Lemma 3.4], but does not influence the directions of constraint of the Skorokhod problem.

We now give the formal definitions of the Skorokhod problem and extended Skorokhod problem associated with the GPS model. Recall that for $\eta \in \mathcal{D}([0,\infty) : \mathbb{R}^J)$, $|\eta|(T)$ denotes the total variation of η on $[0,T]$ with respect to the Euclidean norm on $\mathbb{R}^J$.

DEFINITION 3.1 (Skorokhod problem). Let $\psi \in \mathcal{D}([0,\infty) : \mathbb{R}^J)$ with $\psi(0) \in \mathbb{R}^J_+$ be given. Then (ϕ, η) solves the GPS Skorokhod problem (SP) for ψ if $\phi(0) = \psi(0)$ and if, for all $t \in [0,\infty)$, the following five properties hold:

1. $\phi(t) = \psi(t) + \eta(t)$.
2. $\phi(t) \in \mathbb{R}^J_+$.
3. $|\eta|(t) < \infty$.
4. $|\eta|(t) = \int_{[0,t]} I_{\{\phi(s) \in \partial\mathbb{R}^J_+\}}\, d|\eta|(s)$.
5. There exists a measurable $\gamma : [0,\infty) \to \mathbb{R}^J$ such that $\gamma(t) \in d^1(\phi(t))$ ($d|\eta|$ almost everywhere) and

$$\eta(t) = \int_{[0,t]} \gamma(s)\, d|\eta|(s).$$

Note that ϕ is constrained to remain within $\mathbb{R}^J_+$ and that η changes only when ϕ is on the boundary $\partial\mathbb{R}^J_+$, in which case the change points in one of the directions of $d(\phi)$. If $(\phi, \phi - \psi)$ solve the SP for ψ, then we denote $\phi = \Gamma(\psi)$ and refer to Γ as the GPS Skorokhod map. It was shown in [10], Theorem 3.8, that Γ is unique on its domain of definition, which ensures that $\Gamma(\psi)$ is uniquely defined. The values of $\psi \in \mathcal{D}([0,\infty) : \mathbb{R}^J)$ for which there exists $\phi \in \mathcal{D}([0,\infty) : \mathbb{R}^J_+)$ such that $\phi = \Gamma(\psi)$, is called the domain of Γ [dom(Γ)].

We now introduce the definition of the GPS extended Skorokhod problem. Recall that for $A \subset \mathbb{R}^J$, $\overline{\text{co}}[A]$ represents the closure of the convex hull of the set A.

DEFINITION 3.2 (extended Skorokhod problem). Let $\psi \in \mathcal{D}([0,\infty) : \mathbb{R}^J)$ with $\psi(0) \in \mathbb{R}^J_+$ be given. Then (ϕ, η) solves the GPS extended Skorokhod problem (ESP) for ψ if $\phi(0) = \psi(0)$ and if, for all $t \in [0,\infty)$:

1. $\phi(t) = \psi(t) + \eta(t)$.
2. $\phi(t) \in \mathbb{R}^J_+$.

3. For every $s \in [0,t]$,

$$\eta(t) - \eta(s) \in \overline{\text{co}}\Bigg[\bigcup_{u\in(s,t]} d(\phi(u))\Bigg]. \tag{3.3}$$

4. $\eta(t) - \eta(t-) \in \overline{\text{co}}[d(\phi(t))]$.

Analogous to the GPS Skorokhod map, if $(\phi, \phi - \psi)$ solve the GPS extended Skorokhod problem for ψ, then we denote $\phi = \overline{\Gamma}(\psi)$ and refer to $\overline{\Gamma}$ as the GPS extended Skorokhod map. Also, the domain of the extended Skorokhod map is denoted by $\text{dom}(\overline{\Gamma})$. In Theorem 3.3 we show that there is a unique (ϕ, η) that satisfies the GPS ESP for any $\psi \in \mathcal{D}([0,\infty) : \mathbb{R}^J)$, and thus that $\overline{\Gamma}(\psi)$ is uniquely defined.

3.2. *Properties of the GPS extended Skorokhod map.* In this section we summarize some properties of the GPS extended Skorokhod problem that are required to prove the limit theorems in Section 4. We first recall that a map $\Gamma : \mathcal{D}([0,\infty) : \mathbb{R}^J) \to \mathcal{D}([0,\infty) : \mathbb{R}^J)$ is said to be Lipschitz continuous on its domain (with respect to the topology of uniform convergence on compact sets) if for every $T < \infty$, there exists $K_T < \infty$ such that given any $\psi_1, \psi_2 \in \text{dom}(\Gamma)$ and $\phi_1 = \Gamma(\psi_1)$, $\phi_2 = \Gamma(\psi_2)$,

$$\sup_{t\in[0,T]} \|\phi_1(t) - \phi_2(t)\| \le K_T \sup_{t\in[0,T]} \|\psi_1(t) - \psi_2(t)\|. \tag{3.4}$$

THEOREM 3.3. *For any $\beta \in (0,1]^J$ the GPS Skorokhod map Γ associated with the weight vector β is defined on a strict subset of $\mathcal{D}([0,\infty) : \mathbb{R}^J)$ that includes all paths of bounded variation and is Lipschitz continuous on its domain. The corresponding GPS extended Skorokhod map $\overline{\Gamma}$ on $\mathcal{C}([0,\infty) : \mathbb{R}^J)$ is equal to the unique Lipschitz continuous extension of Γ to $\mathcal{C}([0,\infty) : \mathbb{R}^J)$.*

PROOF. The example provided in [20], Theorem 3.12, shows that the domain of the GPS Skorokhod map is a strict subset of $\mathcal{D}([0,\infty) : \mathbb{R}^J)$. It was shown in [10], Theorem 3.8, that the GPS Skorokhod map is Lipschitz continuous (with the Lipschitz constant $K_T = K$ in fact independent of T) and, therefore, single-valued on a domain that includes all paths of bounded variation. Since this set of paths is dense in $\mathcal{D}([0,\infty) : \mathbb{R}^J)$, Γ consequently has a unique Lipschitz continuous extension to $\mathcal{D}([0,\infty) : \mathbb{R}^J)$ ([23], page 149). From [20], Theorem 3.11, it then follows that $\overline{\Gamma}$ coincides with the Lipschitz continuous extension of Γ to $\mathcal{C}([0,\infty) : \mathbb{R}^J)$. □

Note that although the Skorokhod and extended Skorokhod maps could in general be multivalued, we need not be concerned with that situation here since in this paper we only consider the GPS Skorokhod and extended Skorokhod maps, which are Lipschitz continuous (and therefore single-valued) on their domains of definition.

3.3. *Skorokhod problem representation for the unfinished work.* In this section we derive a convenient representation for the unfinished work process U associated with the GPS model with the two weight vectors $\alpha \in [0,1]^J$ and $\beta \in (0,1]^J$ in terms of the GPS Skorokhod map Γ associated with the weight vector β. Define

$$X_i(t) \doteq U_i(0) + H_i(t) - \alpha_i t \tag{3.5}$$

and

$$Y_i(t) \doteq \alpha_i t - T_i(t) \tag{3.6}$$

where, as in (2.8),

$$T_i(t) = \sum_{E \subset \mathcal{I}: i \notin E} \alpha_i^E I_E(t).$$

Note that by (2.7) $U = X + Y$.

LEMMA 3.4. *Suppose Assumption* 2.1 *holds. Let X be as defined in* (3.5) *and let Γ be the GPS Skorokhod map associated with the weight vector $\beta \in (0,1]^J$. If $(U, I_E, E \subseteq \mathcal{I})$ satisfy* (2.7), *then $U = \Gamma(X)$.*

PROOF. The fact that $U(0)$ is nonnegative and $H(0) = 0$ implies that $U(0) = \Gamma(X)(0)$. Equations (2.7) imply that $U \in \mathbb{R}_+^J$. If Y is as defined in (3.6), then $Y = U - X$ and since U and X are of bounded variation, it follows that Y is also of bounded variation. To prove the lemma, it remains to show that Y satisfies properties 4 and 5 of Definition 3.1.

From the construction of U given in the proof of Lemma 2.2, it is clear that there exists a partition of $[0,\infty)$ given by $s_0 = 0 < s_1 < s_2 < \cdots < s_n < \cdots$ such that $E(\cdot)$ is constant on each interval $[s_{j-1}, s_j)$. Now suppose that for some $j \geq 1$, $U(t) = \Gamma(X)(t)$ for all $t \in [0, s_{j-1}]$ and let $E^* = E(s_{j-1}) = \{i \in \mathcal{I} : U_i(s_{j-1}) = 0\}$. By construction, we know that for all $t \in [s_{j-1}, s_j]$, $I_{E^*}(t) = I_{E^*}(s_{j-1}) + t - s_{j-1}$ and $I_E(t) = I_E(s_{j-1})$ for $E \subseteq \mathcal{I}, E \neq E^*$. Thus for $t \in [s_{j-1}, s_j]$,

$$Y_i(t) - Y_i(s_{j-1}) = \begin{cases} \alpha_i(t - s_{j-1}), & \text{for } i \in E^*, \\ -\dfrac{\beta_i}{\sum_{j \notin E^*} \beta_j} \left[\sum_{j \in E^*} \alpha_j \right] (t - s_{j-1}), & \text{for } i \notin E^*. \end{cases} \tag{3.7}$$

From the above expression it is clear that if $E^* = \varnothing$, then Y_i is constant on $[s_{j-1}, s_j]$ for $i \in \mathcal{I}$. This establishes property 4.

We now verify property 5. If $E^* \neq \varnothing$, for $j \in E^*$ define

$$\theta_j \doteq (1 - \beta_j) \left[\alpha_j + \frac{\beta_j [\sum_{k \in E^*} \alpha_k]}{\sum_{k \notin E^*} \beta_k} \right]$$

and note that since $\sum_{j\in\mathcal{I}}\beta_j = 1$,

$$\sum_{j\in E^*}\frac{\theta_j}{1-\beta_j} = \frac{[\sum_{j\in E^*}\alpha_j][\sum_{k\notin E^*}\beta_k] + [\sum_{j\in E^*}\beta_j][\sum_{k\in E^*}\alpha_k]}{\sum_{k\notin E^*}\beta_k} = \frac{\sum_{k\in E^*}\alpha_k}{\sum_{k\notin E^*}\beta_k}.$$

Moreover, observe that $\theta_j \geq 0$ and

$$\begin{aligned}\sum_{j\in E^*}\theta_j d_j &= \sum_{j\in E^*}\theta_j e_j - \sum_{j\in E^*}\theta_j\Bigg(\sum_{k\neq j}\frac{\beta_k}{1-\beta_j}e_k\Bigg)\\ &= \sum_{j\in E^*}\theta_j e_j + \sum_{j\in E^*}\frac{\theta_j\beta_j}{1-\beta_j}e_j - \sum_{j\in E^*}\sum_{k=1}^{J}\frac{\theta_j\beta_k}{1-\beta_j}e_k\\ &= \sum_{j\in E^*}\frac{\theta_j}{1-\beta_j}e_j - \sum_{k=1}^{J}\Bigg(\sum_{j\in E^*}\frac{\theta_j}{1-\beta_j}\Bigg)\beta_k e_k.\end{aligned}$$

Using the last three displays, we note that for $i \in E^*$,

$$\begin{aligned}\Bigg(\sum_{j\in E^*}\theta_j d_j\Bigg)_i &= \frac{\theta_i}{1-\beta_i} - \beta_i\sum_{j\in E^*}\frac{\theta_j}{1-\beta_j}\\ &= \alpha_i + \frac{\beta_i[\sum_{k\in E^*}\alpha_k]}{\sum_{k\notin E^*}\beta_k} - \frac{\beta_i[\sum_{k\in E^*}\alpha_k]}{\sum_{k\notin E^*}\beta_k} = \alpha_i,\end{aligned} \tag{3.8}$$

and for $i \notin E^*$,

$$\Bigg(\sum_{j\in E^*}\theta_j d_j\Bigg)_i = -\beta_i\sum_{j\in E^*}\frac{\theta_j}{1-\beta_j} = -\frac{\beta_i[\sum_{k\in E^*}\alpha_k]}{\sum_{k\notin E^*}\beta_k}. \tag{3.9}$$

Thus for $t \in [s_{j-1}, s_j)$, (3.7), (3.8) and (3.9) together imply

$$Y(t) - Y(s_{j-1}) = (t - s_{j-1})\Bigg(\sum_{j\in E^*}\theta_j d_j\Bigg) \in \operatorname{cone}(d_j,\, j \in E^*) = d\big(U(t)\big),$$

where the last equality follows from the definitions of $d(\cdot)$ and E^*. Finally, $Y(s_j) - Y(s_j-) = 0 \in d(U(s_j))$, which completes the verification of property 5. □

REMARK. In the network setting, the mapping that takes the input processes to the buffer content would assume a more complicated form, which includes routing between different nodes. In contrast with the single-node case, in the network setting this mapping may not always be sufficiently regular, even when restricted to paths of bounded variation. Indeed, in [11] a particular four-class two-node GPS network was considered and it was shown there that the fluid limit has a representation in terms of a Skorokhod map, but that the Skorokhod map is not Lipschitz continuous for all parameter values of the network. Therefore a key element to extending the results of this paper to the network setting is the identification of the class of GPS networks in heavy traffic for which the associated Skorokhod maps are Lipschitz continuous.

4. Limit theorems. In this section we consider a sequence of GPS servers defined on $(\Omega, \mathcal{F}, P)$ by a sequence $\{H^n\}$ of cumulative workload arrival processes that satisfy Assumption 2.1, and sequences $\{A^n\}$ and $\{L^n\}$ of cumulative arrival and service counting processes that satisfy Assumption 2.3. Let U^n, T^n and Q^n be the associated unfinished work, busy time and queue length processes uniquely characterized by equations (2.7), (2.8) and (2.10), and let $\{V^n\}$ and $\{W^n\}$ be the corresponding sojourn time and waiting time sequences defined by (2.14) and (2.15), respectively. We also consider the associated sequences $\{X^n\}$ and $\{Y^n\}$, where X^n and Y^n are defined by the relationships (3.5) and (3.6), respectively, with $U_i(0)$, H_i and T_i replaced by $U_i^n(0)$, H_i^n and T_i^n. Assume $\mathcal{F}$ is complete with respect to P and for $n \in \mathbb{N}$, let $\{\mathcal{F}_t^n\}$ be complete filtrations such that $H^n(t)$ and $A^n(t)$ are adapted to $\{\mathcal{F}_t^n\}$.

In Sections 4.1 and 4.2 we establish fluid and diffusion limit theorems, respectively, for the unfinished work, busy time, queue length, sojourn time and waiting time processes associated with the GPS model. The limit theorems rely on the Skorokhod problem representation for the unfinished work process derived in Lemma 3.4 and the continuity properties of the associated Skorokhod map stated in Theorem 3.3.

4.1. *Fluid limits.* Given a sequence $\{f^n\} \subset \mathcal{D}([0,\infty) : \mathbb{R}^J)$, we define the associated fluid scaled sequence $\{\overline{f}^n\} \subset \mathcal{D}([0,\infty) : \mathbb{R}^J)$ by

$$\overline{f}^n(t) \doteq \frac{f^n(nt)}{n} \qquad \text{for } t \in [0,\infty). \tag{4.1}$$

For the sojourn and waiting times, we also introduce the sequences $\{\widetilde{V}^n\}$ and $\{\widetilde{W}^n\}$ defined by

$$\begin{aligned} \widetilde{V}_i^n(t) &\doteq \frac{V_i^n(A_i^n(nt))}{n} = \overline{V}_i^n \circ \overline{A}_i^n(t), \\ \widetilde{W}_i^n(t) &= \frac{W_i^n(A_i^n(nt))}{n} = \overline{W}_i^n \circ \overline{A}_i^n(t). \end{aligned} \tag{4.2}$$

Likewise we define the scaled version of the jump times of A_i^n as

$$\widetilde{C}_i^n(t) \doteq \frac{C_i^n(A_i^n(nt))}{n} = \overline{C}_i^n \circ \overline{A}_i^n(t). \tag{4.3}$$

In the following lemma we state an elementary property that will be useful in the sequel.

LEMMA 4.1. *Suppose f, g are nondecreasing functions in $\mathcal{D}([0,\infty) : \mathbb{R})$ such that*

$$\lim_{m\to\infty} \frac{f \circ g(mt)}{m} = t, \tag{4.4}$$

where the convergence is uniform on any compact subset of $[0,\infty)$. *If there exists* $\theta \in \mathbb{R} \setminus \{0\}$ *such that*

$$\lim_{m\to\infty} \frac{f(mt)}{m} = \theta\iota \qquad u.o.c., \tag{4.5}$$

then

$$\lim_{m\to\infty} \frac{g(mt)}{m} = \frac{1}{\theta}\iota \qquad u.o.c. \tag{4.6}$$

PROOF. Define

$$\overline{f}^m(t) \doteq \frac{f(mt)}{m} \quad \text{and} \quad \overline{g}^m(t) \doteq \frac{g(mt)}{m}.$$

Suppose (4.5) holds. If (4.6) does not hold, then there exist $t \in [0,\infty)$, $\varepsilon > 0$ and a subsequence $\{m_k\}$ such that

$$\liminf_{k\to\infty} |\theta\overline{g}^{m_k}(t) - t| > \varepsilon.$$

Since $\theta \neq 0$ and $\overline{f}^m$ is nondecreasing and satisfies (4.5), this implies that either

$$\liminf_{k\to\infty} \overline{f}^{m_k} \circ \overline{g}^{m_k}(t) \geq \liminf_{k\to\infty} \overline{f}^{m_k}\big((t+\varepsilon)/\theta\big) = t + \varepsilon$$

or

$$\limsup_{k\to\infty} \overline{f}^{m_k} \circ \overline{g}^{m_k}(t) \leq \limsup_{k\to\infty} \overline{f}^{m_k}\big((t-\varepsilon)/\theta\big) = t - \varepsilon,$$

which contradicts (4.4) and hence establishes that (4.5) implies (4.6). □

4.1.1. *The fluid limit of the unfinished work process.* From Definition 3.1 and the fact that the GPS Skorokhod map is Lipschitz continuous on its domain (see Theorem 3.3), it is easy to verify that the GPS Skorokhod map Γ is nonanticipatory in the sense that $\Gamma(X)(t)$ depends only on $\{X(s), s \leq t\}$ (e.g., see [5]). Since H^n is adapted to the filtration $\{\mathcal{F}_t^n\}$, U^n is right-continuous and $U^n = \Gamma(X^n)$ by Lemma 3.4, this implies that U^n is progressively measurable with respect to the filtration $\{\mathcal{F}_t^n\}$. We now assume that the primitive processes satisfy a functional strong law of large numbers. Recall that the abbreviation u.o.c. represents uniform convergence on compact time intervals.

ASSUMPTION 4.2. 1. There exists $\overline{u} \in \mathbb{R}_+^J$ such that a.s.

$$\lim_{n\to\infty} \frac{U^n(0)}{n} = \overline{u}.$$

2. For each $n \in \mathbb{N}$ there exists $\gamma^n \in \mathbb{R}_+^J$ such that a.s.

$$\lim_{m\to\infty} \frac{H^n(mt)}{m} = \gamma^n t,$$

where the convergence is u.o.c.

3. There exists $\gamma \in \mathbb{R}_+^J$ such that

$$\lim_{n\to\infty} \gamma^n = \gamma.$$

Recall that $\iota : \mathbb{R}_+ \to \mathbb{R}_+$ is the identity map. Define

$$\nu \doteq \gamma - \alpha \tag{4.7}$$

and let

$$\overline{X} \doteq \overline{u} + \nu\iota, \qquad \overline{U} \doteq \Gamma(\overline{X}), \qquad \overline{Y} \doteq \overline{U} - \overline{X} \tag{4.8}$$

and

$$\overline{T} \doteq \alpha\iota - \overline{Y} = \gamma\iota + \overline{u} - \overline{U}. \tag{4.9}$$

THEOREM 4.3. *Suppose Assumptions* 2.1 *and* 4.2 *hold, and let* $\overline{U}$ *and* $\overline{Y}$ *be given by* (4.8). *Then* P *a.s. as* $n \to \infty$, $\overline{U}^n \to \overline{U}$, $\overline{Y}^n \to \overline{Y}$ *and* $\overline{T}^n \to \overline{T}$ *u.o.c.*

PROOF. From the representation proved in Lemma 3.4, one obtains $U^n = \Gamma(X^n)$ for $n \in \mathbb{N}$, where Γ is the GPS Skorokhod map associated with the weight vector β. As stated in Theorem 3.3, Γ is defined on a domain that includes all functions of bounded variation. Since the domain $\mathbb{R}_+^J$ of the GPS Skorokhod problem is conical with vertex at the origin and the directions of constraint are radially homogeneous [i.e., $d(x) = d(cx)$ for $c \in (0, \infty)$], it is easy to see that the GPS Skorokhod map is homogeneous in both time and space. In other words, $\Gamma(\lambda\psi) = \lambda\Gamma(\psi)$ for $\lambda \in \mathbb{R}$ and $\psi \in \mathrm{dom}(\Gamma)$. Also, $\Gamma(\psi^c)(t) = \Gamma(\psi)(ct)$, where $\psi^c(t) \doteq \psi(ct)$. Thus the scaled processes also have the representation $\overline{U}^n(t) = \Gamma(\overline{X}^n)(t)$ for $n \in \mathbb{N}$.

By Assumption 4.2 and the expresssion (3.5) for X^n, it is evident that P a.s. $\overline{X}^n \to \overline{X}$ u.o.c. as $n \to \infty$. The fact that Γ is Lipschitz continuous on its domain (which follows from Theorem 3.3 and the fact that $\beta > 0$) then yields P a.s. $\overline{U}^n \to \overline{U}$ u.o.c. Since $\overline{Y}^n = \overline{U}^n - \overline{X}^n$ and addition is continuous in the u.o.c. topology, P a.s. $\overline{Y}^n \to \overline{Y}$ u.o.c. The P a.s. convergence $\overline{T}^n \to \overline{T}$ u.o.c. then follows directly from the fact that $\overline{T}^n = \alpha\iota - \overline{Y}^n$ and $\overline{T} = \alpha\iota - \overline{Y}$. □

The next lemma describes the structure of the paths of the fluid limit $\overline{U}$ of the unfinished work process. As can be seen from the definition given in (4.8), $\overline{U}$ is the image of an affine trajectory under the GPS Skorokhod map. The structure of images of affine trajectories under the Skorokhod map on polyhedral domains has been studied in [1, 5, 12]. Stability properties of such trajectories for conical polyhedral domains and Lipschitz continuous Skorokhod maps were analyzed in [5]. For the reflection mapping case, [1] showed that these trajectories were affine if the reflection matrix was completely-$\mathcal{S}$ and [12] showed that they comprised a finite number of linear pieces if the reflection matrix satisfied the

so-called Harrison–Reiman condition. As mentioned in the introduction, the GPS Skorokhod problem does not fall into the reflection mapping setting nor does it satisfy the generalization of the completely-$\mathcal{S}$ condition to general Skorokhod problems (namely that the associated $\mathcal{V}$-set be empty [20]). Nevertheless, we show in Lemma 4.4 that analogous results continue to hold even for the GPS Skorokhod problem. In the proof, we use the existence of a unique Lipschitz continuous discrete projection operator $\pi : \mathbb{R}^J \to \mathbb{R}^J_+$ associated with the GPS Skorokhod problem, which satisfies $\pi(x) = x$ if $x \in \mathbb{R}^J_+$, while $\pi(x) \in \partial\mathbb{R}^J_+$ and $\pi(x) - x \in d(\pi(x))$ if $x \notin \mathbb{R}^J_+$ ([10], Theorem 3.8).

LEMMA 4.4. *Let* $\overline{U}$ *be as defined in* (4.8) *and define* $\kappa \doteq \pi(\nu)$. *Then the following properties are satisfied*:

1. $\overline{U}$ *is piecewise affine with at most* J *changes of slope*.
2. *If* $\overline{u} = 0$, *then* $\overline{U} = \kappa\iota$ *and* $\overline{Y} = (\kappa - \nu)\iota$. *Moreover*, $\kappa = 0$ *if and only if* $\sum_{i=1}^J \nu_i \le 0$.
3. *If* $\sum_{i=1}^J \nu_i < 0$, *then there exists* $\sigma < \infty$ *such that* $\overline{U}(t) = 0$ *for all* $t \ge \sigma$.

PROOF. To prove the first property, we define $\sigma \doteq \inf\{t > 0 : \overline{U}(t) = 0\}$. Let D be the $J \times J$ matrix whose ith column is d_i and for any $\mathcal{K} \subseteq \mathcal{I}$, let K be the cardinality of $\mathcal{K}$, let $D_{\mathcal{K}}$ be the submatrix that consists of all entries D_{ij} with $i, j \in \mathcal{K}$ and let I_K be the $K \times K$ identity matrix. Given any submatrix $D_{\mathcal{K}}$ such that $\mathcal{K} \ne \mathcal{I}$, using the definition of d_i given in (3.2) it follows that for any $k \in \mathcal{K}$,

$$\sum_{i\in\mathcal{K}} (I_K - D_{\mathcal{K}})_{ik} = \frac{\sum_{i\in\mathcal{K}, i\ne k} \beta_i}{1 - \beta_k} = \frac{1 - \beta_k - \sum_{j\notin\mathcal{K}} \beta_j}{1 - \beta_k} < 1,$$

where the last inequality follows from the fact that $\beta > 0$ and $\mathcal{K} \ne \mathcal{I}$. Thus $I_K - D_{\mathcal{K}}$ is nonnegative and the spectral radius of $|I_K - D_{\mathcal{K}}|$ is less than 1 [14]. In particular, this implies that $D_{\mathcal{K}}$ satisfies the completely-$\mathcal{S}$ condition. Indeed if y is the K-dimensional vector of all 1's, then the last display shows that $D'_{\mathcal{L}} y > 0$ for every principal submatrix $D_{\mathcal{L}}$ of $D_{\mathcal{K}}$, which implies that $D'_{\mathcal{K}}$ is completely-$\mathcal{S}$. Since the class of completely-$\mathcal{S}$ matrices is closed under transposition ([22], Lemma 3; [15]), this implies $D_{\mathcal{K}}$ satisfies the completely-$\mathcal{S}$ condition. From [1], Remarks 2 and 4, one can infer that $\overline{U}$ must be piecewise affine on $[0, \sigma)$. In other words, given any $t \in [0, \sigma)$, there exists $\varepsilon > 0$ and $\mathcal{K} \subset \mathcal{I}$ such that for $s \in [0, \varepsilon)$,

$$\mathcal{K} = \{i \in \mathcal{I} : \overline{U}_i(t + s) = 0\} \quad \text{and} \quad \overline{U}(s) = \overline{U}(t) + sw,$$

where for some $\theta_i \ge 0$, $i \in \mathcal{K}$,

$$w = \nu + \sum_{i\in\mathcal{K}} \theta_i d_i.$$

It is easy to see that for every $t < \sigma$ there exists $\varepsilon > 0$ such that on the interval $[t, t+\varepsilon]$, $\overline{U}$ can be identified with the image of $\nu\iota$ under the reflection map associated with the matrix $D_{\mathcal{K}}$, where $\mathcal{K} = \{i \in \mathcal{I} : \overline{U}_i(t) = 0\}$. The argument used in the proof of Theorem 4.4(1) of [12] then shows that if $\mathcal{K} = \{i \in \mathcal{I} : \overline{U}_i(t) = 0\}$ for $t \in (0, \sigma)$, then $\mathcal{K} \subseteq \{i \in \mathcal{I} : \overline{U}_i(s) = 0\}$ for all $s \in [t, \sigma)$ and that if $\sigma < \infty$, then $\overline{U}(s) = \dot{\overline{U}}(s) = 0$ for all $s \in [\sigma, \infty)$. This implies that $\overline{U}$ must move to a lower-dimensional face every time it changes slope, and so $\overline{U}$ can have at most $J-1$ changes of slope on $[0, \sigma)$, and at most J changes of slope on $[0, \infty)$.

If $\overline{u} = 0$, then $\overline{U} = \Gamma(\nu\iota)$. Thus to prove that $\overline{U} = \kappa\iota$ we need to show that $\kappa\iota$ solves the GPS Skorokhod problem for $\nu\iota$. Properties 1, 2 and 3 of the Skorokhod problem follow automatically from the definition of π and the conical structure of the domain. If $\nu \in \mathbb{R}^J_+$, then $\pi(\nu) = \nu$, and it is clear that $\pi(\nu)\iota = \nu\iota$ solves the GPS Skorokhod problem for $\nu\iota$. If $\nu \notin \mathbb{R}^J_+$, then $\pi(\nu) \in \partial\mathbb{R}^J_+$, and by the definition of π and the fact that d is a cone, $(\pi(\nu) - \nu)t \in d(\pi(\nu))$ for all $t \geq 0$. Moreover, due to the fact that $\mathbb{R}^J_+$ is a cone and the fact that d is constant along faces of $\partial\mathbb{R}^J_+$, $d(\pi(\nu)t) = d(\pi(\nu))$ for all $t > 0$. This verifies the remaining two properties of the GPS Skorokhod problem stated in Definition 3.1, and hence shows that $(\kappa\iota, (\kappa - \nu)\iota)$ solve the GPS Skorokhod problem for $\nu\iota$. Since the GPS Skorokhod map is single-valued, we conclude that $\overline{U} = \kappa\iota$. From the definition of d_i it is easy to see that $d(0) = \{x \in \mathbb{R}^J : \sum_{i=1}^J x_i \geq 0\}$. If $\sum_{i=1}^J \nu_i \leq 0$, then $-\nu \in d(0)$ and so the definition of π ensures that $\kappa = \pi(\nu) = 0$. This establishes the second property.

For the third property, note that if ν satisfies $\sum_{i=1}^J \nu_i < 0$, then, as stated above, ν lies in the interior of $-d(0)$. Thus it follows from [5] that there exists $\sigma < \infty$ such that $\overline{U}(t) = 0$ for all $t \geq \sigma$. $\square$

4.1.2. *The fluid limit of the queue length process.* We first state the assumptions on the primitive processes that yield a fluid limit for the queue length process. Recall that $(\mathbb{R}^J_+)^\circ$ denotes the interior of the J-dimensional nonnegative orthant.

ASSUMPTION 4.5. 1. There exists $\overline{q} \in \mathbb{R}^J_+$ such that a.s.

$$\lim_{n\to\infty} \frac{Q^n(0)}{n} = \overline{q}.$$

2. For every $n \in \mathbb{N}$ there exists $\lambda^n \in \mathbb{R}^J_+$ such that P a.s.

$$\lim_{m\to\infty} \frac{A^n(mt)}{m} = \lambda^n t \qquad \text{u.o.c.}$$

3. For every $n \in \mathbb{N}$ there exists $\mu^n \in (\mathbb{R}^J_+)^\circ$ such that P a.s.

$$\lim_{m\to\infty} \frac{L^n(mt)}{m} = \mu^n t \qquad \text{u.o.c.}$$

4. There exist $\lambda \in \mathbb{R}_+^J$ and $\mu \in (\mathbb{R}_+^J)^\circ$ such that

$$\lim_{n\to\infty} \lambda^n = \lambda \quad \text{and} \quad \lim_{n\to\infty} \mu^n = \mu.$$

Moreover $\gamma, \gamma^n \in \mathbb{R}_+^J$ are defined by

$$\gamma_i \doteq \lambda_i/\mu_i \quad \text{and} \quad \gamma_i^n \doteq \lambda_i^n/\mu_i^n \qquad \text{for } i \in \mathcal{I},\ n \in \mathbb{N}.$$

REMARK 4.6. Given sequences $\{Q^n(0)\}$, $\{A^n\}$ and $\{L^n\}$ that satisfy Assumption 2.3, the corresponding sequence of cumulative workload processes $\{H^n\}$ defined via relationships (2.1) and (2.4) clearly satisfies Assumption 2.1. Moreover, it is straightforward to check that if Assumption 4.5 holds, then Assumption 4.2 is satisfied with $\overline{u}_i = \overline{q}_i/\mu_i$ and $\gamma_i^n = \lambda_i^n/\mu_i^n$ for $n \in \mathbb{N}$ and $\gamma_i = \lambda_i/\mu_i$ for $i \in \mathcal{I}$.

We now introduce the process $\overline{Q}$ given by

$$\overline{Q}_i \doteq \mu_i \overline{U}_i \qquad \text{for } i \in \mathcal{I}, \tag{4.10}$$

where $\overline{U}$ is defined by (4.8) with $\overline{u}_i = \overline{q}_i/\mu_i$.

THEOREM 4.7. *Suppose Assumptions* 2.3 *and* 4.5 *hold and let* $\overline{Q}$ *be defined by* (4.10). *Then* P *a.s. as* $n \to \infty$, $\overline{Q}^n \to \overline{Q}$ *u.o.c.*

PROOF. Applying the fluid scaling (4.1), note that for each $i \in \mathcal{I}$ and $t \in [0, \infty)$,

$$\frac{L_i^n(T_i^n(nt))}{n} = \frac{L_i^n(n\overline{T}_i^n(t))}{n} = \overline{L}_i^n\big(\overline{T}_i^n(t)\big).$$

Substituting this in (2.10), we infer that

$$\overline{Q}_i^n(t) = \overline{Q}_i^n(0) + \overline{A}_i^n(t) - \overline{L}_i^n\big(\overline{T}_i^n(t)\big). \tag{4.11}$$

Along with Remark 4.6, Theorem 4.3 implies that P a.s. $\lim_{n\to\infty} \overline{T}^n \to \overline{T}$ u.o.c., where $\overline{T}$ is given by (4.9) with $\gamma_i = \lambda_i/\mu_i$ and $\overline{u}_i = \overline{q}_i/\mu_i$. Thus Assumption 4.5 and (4.9) yield

$$\lim_{n\to\infty} \overline{Q}_i^n(t) = \overline{q}_i + \lambda_i t - \mu_i \overline{T}_i(t) = \mu_i[\overline{u}_i + \gamma_i t - \alpha_i t + \overline{Y}_i(t)].$$

From (4.7) and (4.8) it follows that $\overline{U}_i(t) = \overline{u}_i + (\gamma_i - \alpha_i)t + \overline{Y}_i(t)$. Substituting this into the last display yields the theorem. □

In [3] a general class of head-of-the-line disciplines is considered, which in particular includes our (generalized) GPS discipline under the assumption that the service is FIFO within each class. A fluid limit for the queue length was proved in [3], Section 4, under the additional condition (translated to the GPS discipline and our notation) that $\alpha_j > \gamma_j$ for $1 \le j \le J$, and was used to prove stability of the original queueing system. There is thus some overlap between the result in Theorem 4.7 of this paper and that in [3], Section 4, but neither result contains the other.

4.1.3. *The fluid limit of the sojourn time and waiting time processes.* Recall the definitions of $\overline{T}$ and the mapping F given in (4.9) and (2.12), respectively, and for $i \in \mathcal{I}$ let

$$\overline{V}_i = F[\overline{T}_i] \circ (\overline{T}_i + \overline{U}_i) - \iota \quad \text{and} \quad \overline{W} \doteq \overline{V}. \tag{4.12}$$

Then

$$\text{dom}(\overline{V}_i) = \left\{ t \geq 0 : \overline{T}_i(t) + \overline{U}_i(t) \leq \sup_{t \in [0,T]} \overline{T}_i(s) \text{ for some } T < \infty \right\}. \tag{4.13}$$

Also recall from Section 2.2.4 that

$$V_i^n(k) \doteq F[T_i^n] \circ (T_i^n + U_i^n) \circ C_i^n(k) - C_i^n(k) \tag{4.14}$$

and

$$W_i^n(k) \doteq F[T_i^n] \circ \left(T_i^n \circ C_i^n(k) + U_i^n\left(C_i^n(k) - \right)\right) - C_i^n(k). \tag{4.15}$$

As shown in Lemma 4.9, the following condition guarantees that each of the functions V_i^n, W_i^n, $\overline{V}_i$ and $\overline{W}_i$ has $[0, \infty)$ as its domain.

CONDITION 4.8. At least one of the following two properties holds.

1. $\gamma_k < \alpha_k$ for some $k \in \mathcal{I}$.
2. $\sum_{i=1}^J \gamma_i = 1$ and $\overline{u}_i = 0$ for all i such that $\alpha_i = 0$.

LEMMA 4.9. *Suppose Assumptions* 2.3 *and* 4.5 *hold, and let* $\overline{T}$ *and* $\overline{V}$ *be defined by* (4.9) *and* (4.12), *respectively. Then the following properties are satisfied*:

1. *If Condition* 4.8 *holds, then* $\text{dom}(\overline{V}) = [0, \infty)$.
2. *Suppose* $\gamma_i > 0$. *If Condition* 4.8 *is satisfied, then there exist* $\delta > 0$ *and* $\tau \in [0, \infty)$ *such that* $\overline{T}_i(t) = 0$ *for* $t \in [0, \tau]$ *and* $\dot{\overline{T}}_i(t) > \delta$ *for a.e.* $t > \tau$, *while if* $\alpha_i > 0$, *then the above statement holds with* $\tau = 0$.
3. *If* $\alpha_i > 0$, *then* $\text{dom}(\overline{V}_i) = [0, \infty)$.
4. *If Condition* 4.8 *is satisfied or* $\alpha_i > 0$, *then* $\gamma_i > 0$ *implies that given any* $M > 0$ *there exists* $N < \infty$ *such that* $[0, M] \subset \text{dom}(\overline{V}_i^n)$ *for all* $n \geq N$.

PROOF. We first establish some general properties that will be useful in proving the lemma.

General properties. For $i \in \mathcal{I}$ let $\tau_i \doteq \inf\{t > 0 : \overline{U}_i(t) = 0\}$ and let $\sigma = \inf\{t > 0 : \overline{U}(t) = 0\}$. As argued in the proof of Lemma 4.4, if $\tau_i < \infty$, then $\overline{U}_i(t) = 0$ for all $t \geq \tau_i$. For any set $\mathcal{J} \subset \mathcal{I}$, the vectors d_j, $j \in \mathcal{J}$, are linearly independent ([10], Lemma 3.1). The definition of the SP and the fact that $\{i : \overline{U}_i(t) = 0\}$ is monotonically nondecreasing in t dictates that for Lebesgue

a.e. $t \in [0, \sigma)$ there exists a unique decomposition of $\eta(t)$ in terms of the vectors $\{d_j : \overline{U}_j(t) = 0\}$, and so there exist unique $\theta_j(t) \in [0, \infty)$, $j \in \mathcal{I}$, with $\theta_j(t) > 0$ only if $\overline{U}_j(t) = 0$ and $\dot{\overline{X}}_j(t) = \gamma_j - \alpha_j < 0$, such that

$$\dot{\overline{Y}}_i(t) = \sum_{j \in \mathcal{I}} \theta_j(t)(d_j)_i \qquad \text{for } i \in \mathcal{I}. \tag{4.16}$$

In particular, since $(d_j)_i < 0$ for $i \neq j$ and for a.e. $t < \tau_i$, $\overline{U}_i(t) > 0$ and $\theta_i(t) = 0$, it follows that

$$\dot{\overline{Y}}_i(t) = \sum_{j \in \mathcal{I}} \theta_j(t)(d_j)_i = \sum_{j \in \mathcal{I} \setminus \{i\}} \theta_j(t)(d_j)_i \leq 0. \tag{4.17}$$

Thus by (4.8) and (4.9) for a.e. $t < \tau_i$,

$$\dot{\overline{U}}_i(t) = (\gamma_i - \alpha_i) + \dot{\overline{Y}}_i(t) \leq \gamma_i - \alpha_i \tag{4.18}$$

and

$$\dot{\overline{T}}_i(t) = \alpha_i - \dot{\overline{Y}}_i(t) \geq \alpha_i. \tag{4.19}$$

Moreover, for i such that $\tau_i < \infty$, using (4.9) along with the fact that for $t \geq \tau_i$, $\overline{U}_i(t) = \overline{U}_i(\tau_i) = 0$, one obtains $\overline{T}_i(t) = \overline{T}_i(\tau_i) + \gamma_i(t - \tau_i)$, and so

$$\dot{\overline{T}}_i = \gamma_i \qquad \text{for } t \geq \tau_i. \tag{4.20}$$

If $\gamma_i = 0$, then $\overline{X}_i(t) = \overline{u}_i$ and $\overline{U}_i(t) + \overline{T}_i(t) = \overline{u}_i$ for all $t \in [0, \infty)$. Thus if $\tau_i < \infty$, then

$$\overline{T}_i(\tau_i) = \overline{u}_i = \sup_{t \in [0, \infty)} \overline{T}_i(t) = \sup_{t \in [0, \infty)} [\overline{T}_i(t) + \overline{U}_i(t)]. \tag{4.21}$$

Thus for any $i \in \mathcal{I}$ such that $\tau_i < \infty$, if $\gamma_i > 0$, (4.20) implies that $\overline{T}_i$ is unbounded and so $\mathrm{dom}(\overline{V}_i) = [0, \infty)$, while if $\gamma_i = 0$, then (4.13) and (4.21) imply $\mathrm{dom}(\overline{V}_i) = [0, \infty)$. Let $\mathcal{K} \doteq \{i \in \mathcal{I} : \tau_i < \infty\}$. The above argument shows that

$$\mathrm{dom}(\overline{V}_i) = [0, \infty) \qquad \text{for } i \in \mathcal{K}. \tag{4.22}$$

If $\mathcal{K} = \mathcal{I}$ (which implies $\sigma < \infty$), then the lemma follows as a direct consequence of the result given above, and so henceforth we assume that $\mathcal{K} \neq \mathcal{I}$.

Property 1. If $\gamma_k < \alpha_k$ for some $k \in \mathcal{I}$, then (4.18) shows that $\dot{\overline{U}}_k \leq \gamma_k - \alpha_k < 0$ for $t < \tau_k$ and thus $\tau_k < \infty$ or equivalently $k \in \mathcal{K}$. Then (4.16) and (4.18) and the fact that $\overline{U}_i(t) > 0$ for $i \notin \mathcal{K}$ show that for a.e. $t \in [0, \infty)$ and $\theta_j(t) \in [0, \infty)$, $j \in \mathcal{K}$,

$$\dot{\overline{U}}(t) = \gamma - \alpha + \theta_k(t) d_k + \sum_{j \in \mathcal{K} \setminus \{k\}} \theta_j(t) d_j.$$

For $t > \tau_k$, taking the kth component of the above display and using the fact that $\dot{\overline{U}}_k(t) = 0$, $(d_k)_k = 1$ and $(d_j)_k < 0$ for $j \neq k$, one obtains $\theta_k(t) \geq \alpha_k - \gamma_k$. Hence for $i \notin \mathcal{K}$ and a.e. $t \in [\tau_k, \infty)$ substituting the definition of d_k yields

$$\dot{\overline{U}}_i(t) = \gamma_i - \alpha_i + \theta_k(t)(d_k)_i + \sum_{j \in \mathcal{K} \setminus \{k\}} \theta_j(t)(d_j)_i \leq \gamma_i - \alpha_i - \frac{(\alpha_k - \gamma_k)\beta_i}{1 - \beta_k},$$

which implies that for $i \notin \mathcal{K}$,

$$\dot{\overline{T}}_i(t) \geq \alpha_i + \frac{(\alpha_k - \gamma_k)\beta_i}{1 - \beta_k} > 0, \tag{4.23}$$

where the last inequality uses the fact that $\beta_i > 0$ and $\alpha_k > \gamma_k$. Thus $\overline{T}_i$ is unbounded and $\mathrm{dom}(\overline{V}_i) = [0, \infty)$ for $i \notin \mathcal{K}$. Along with (4.22) this establishes property 1 when Condition 4.8(1) is satisfied.

Now suppose Condition 4.8(1) is not satisfied, but Condition 4.8(2) is satisfied. Then it must be that $\gamma = \alpha$. If $\alpha_i > 0$, then $\gamma_i > 0$, and (4.19) and (4.20) show that $\dot{\overline{T}}(t) \geq \alpha_i \wedge \gamma_i > 0$ for a.e. $t \in [0, \infty)$ and hence $\overline{T}_i(t)$ is unbounded and $\mathrm{dom}(\overline{V}_i) = [0, \infty)$. On the other hand, if $\alpha_i = 0$, then $\gamma_i = 0$ and $\overline{u}_i = 0$; hence $\tau_i = 0 < \infty$ and (4.21) shows that $\mathrm{dom}(\overline{V}_i) = [0, \infty)$. This completes the proof of property 1.

Property 2. Suppose $\gamma_i > 0$. If $\alpha_i > 0$, then (4.19) and (4.20) imply that $\dot{\overline{T}}_i(t) \geq \alpha_i \wedge \gamma_i > 0$ for all $t \in [0, \infty)$. Now suppose $\alpha_i = 0$ and Condition 4.8 holds. Then we show below that Condition 4.8(1) must hold. Indeed, define $\tau \doteq \min_{j \in \mathcal{I}: \gamma_j < \alpha_j} \tau_j$ and $\delta \doteq \gamma_i \wedge \min_{j:\alpha_j = 0} \beta_j \min_{j:\gamma_j < \alpha_j} (\alpha_j - \gamma_j)$. Clearly $\delta > 0$ because $\beta, \gamma_i > 0$ and $\tau < \infty$ because (as shown in the proof of property 1) $\gamma_k < \alpha_k$ implies $k \in \mathcal{K}$. Observe that $\sum_{j \in \mathcal{I}} \dot{\overline{T}}_j \leq 1$ and, in addition, that from (4.19) it follows that for $t \in (0, \tau)$, $\sum_{j \in \mathcal{I}: \alpha_j > 0} \dot{\overline{T}}_j(t) \geq \sum_{j \in \mathcal{I}: \alpha_j > 0} \alpha_j = \sum_{j \in \mathcal{I}} \alpha_j = 1$. Consequently $\dot{\overline{T}}_i(t) = 0$ for $t \in (0, \tau)$. From (4.23) it is clear that $\dot{\overline{T}}_i(t) \geq \beta_i \min_{j:\gamma_j < \alpha_j} (\alpha_j - \gamma_j)$ for $t \in (\tau, \tau_i)$ and if $\tau_i < \infty$, then for $t \in (\tau_i, \infty)$, (4.20) shows that $\dot{\overline{T}}_i(t) \geq \gamma_i$. This concludes the proof.

Property 3. If $\alpha_i > 0$ and $\gamma_i > 0$, property 3 follows from property 2, while if $\alpha_i > 0$ and $\gamma_i = 0$, then (4.18) implies $\tau_i < \infty$ and hence (4.13) and (4.21) show that $\mathrm{dom}(\overline{V}_i) = [0, \infty)$.

Property 4. If Condition 4.8 or $\alpha_i > 0$ holds, then the proofs of the above properties show that $\gamma_i > 0$ implies that given $M < \infty$, there exists $M' < \infty$ such that

$$\sup_{t \in [0, M]} [\overline{U}_i(t) + \overline{T}_i(t)] < \sup_{t \in [0, M']} \overline{T}_i(t).$$

The fact that P a.s. $\overline{T}_i^n \to \overline{T}_i$ and $\overline{U}_i^n \to \overline{U}_i$ u.o.c. by Theorem 4.3 then implies that there exists $N < \infty$ such that for all $n \geq N$,

$$\sup_{t\in[0,M]} [\overline{U}_i^n(t) + \overline{T}_i^n(t)] \leq \sup_{t\in[0,M']} \overline{T}_i^n(t),$$

which establishes the last property of the lemma.

REMARK. From property 3 of the lemma it follows that if $\alpha > 0$, then $\mathrm{dom}(\overline{V}) = [0, \infty)$. Note that $\overline{V}_i$ may not be right-continuous at zero if $\alpha_i = 0$. Also, if $\gamma \geq \alpha$ and $\gamma_i = \bar{u}_i = 0$ whenever $\alpha_i = 0$, a trivial extension of the lemma will show that $\mathrm{dom}(\overline{V}) = [0, \infty)$. In any other case, one does not expect $\overline{V}$ to be well defined. In particular, if $\gamma \geq \alpha$ and both $\alpha_i = 0$ and $\overline{u}_i > 0$ for some $i \in \mathcal{I}$, then (in the fluid limit) each class uses all its guaranteed capacity and so source i never gets any service. Thus any new class i arrival will remain in the system forever, which corresponds to an infinite sojourn time.

Given a sequence of functions $\{f^n\}$ converging u.o.c. to a function f, the following lemma establishes conditions under which $F[f^n]$ tends u.o.c. to f^{-1}.

LEMMA 4.10. *Suppose there exist $\tau \in [0, \infty)$, $\theta > 0$ and a continuous nondecreasing function f such that $f(t) = 0$ for $t \in [0, \tau]$ and $\dot{f}(t) \geq \theta$ for a.e. $t \in [\tau, \infty)$, and let f^{-1} and $F[f]$ be defined by* (2.16) *and* (2.12), *respectively. Then f^{-1} is a strictly increasing continuous function and $F[f](t) = f^{-1}(t)$ for $t \in (0, \infty)$. Moreover if $\{f^n\}$ is a sequence of right-continuous nondecreasing functions on $[0, \infty)$ with $f^n(0) = 0$ for $n \in \mathbb{N}$ and such that $f^n \to f$ u.o.c., then*

$$f^{-1} - F[f^n] \to 0 \qquad \text{as } n \to \infty, \tag{4.24}$$

where the convergence is uniform on compact subsets of $(0, \infty)$ (and of $[0, \infty)$ if $\tau = 0$).

PROOF. Recall from (2.16) and (2.12) that for $t > 0$,

$$g^{-1}(t) = \inf\{s \geq 0 : g(s) > t\} \quad \text{and} \quad F[g](t) = \inf\{s \geq 0 : g(s) \geq t\}.$$

We first establish some general properties of the functions $F[g]$ and g^{-1} when g is right-continuous and nondecreasing. Let $R = \sup\{g(s), s \in [0, \infty)\}$. Then g^{-1} and $F[g]$ are well defined on $[0, R)$. It follows directly from the definitions that $F[g](t) \leq g^{-1}(t)$ for $t \in [0, R)$. From the right continuity of g it follows that $g(g^{-1}(t)) \geq t$ and $g(F[g](t)) \geq t$. Furthermore, as shown below, if $g^{-1}(t) \in (0, \infty)$ is a point of continuity for g, then

$$g\big(g^{-1}(t)\big) = t. \tag{4.25}$$

Indeed, suppose $g(g^{-1}(t)) > t$ for some $t \in [0, R)$ such that $g^{-1}(t) \in (0, \infty)$ is a point of continuity for g. Then by continuity there exists $s < g^{-1}(t)$ such that

$g(s) > t$, which contradicts the definition of $g^{-1}(t)$ and so establishes (4.25). Now consider any $t \in [0, R)$ such that $g(g^{-1}(t)) = t$. Then since g is nondecreasing it is easy to see that $g(F[g](t)) = t$ and that g must be constant and equal to t on the interval $[F[g](t), g^{-1}(t)]$. On the other hand, for $t \in [0, R)$ such that $g(g^{-1}(t)) > t$, clearly $g^{-1}(t)$ must be a point of discontinuity of g and the definition of g^{-1} dictates that $g(g^{-1}(t)-) \le t$. For such t, if $g(s) < t$ for all $s < g^{-1}(t)$, then clearly $F[g](t) = g^{-1}(t) \ne t$. Otherwise [i.e., if there exists $s < g^{-1}(t)$ such that $g(s) = t$], g must be constant and equal to t on the interval $[F[g](t), g^{-1}(t))$. Since $g(g^{-1}(t)) \ge t$ for all $t \in [0, R)$, the above discussion shows that for any right-continuous nondecreasing function g and $t \in [0, \sup_{s\in[0,\infty)} g(s))$,

$$0 \le g^{-1}(t) - F[g](t) \le \sup_{0\le u\le v\le g^{-1}(t)} \{v - u : g(v) = g(u) = t\}, \tag{4.26}$$

where we adopt the convention here that the supremum of the empty set is equal to zero.

Let f be the function satisfying the conditions stated in the lemma. Then $\sup_{s\in[0,\infty)} f(s) = \infty$, and so f^{-1} and $F[f]$ are well defined and finite on $[0, \infty)$. It is straightforward to check that f^{-1} is nondecreasing, $f^{-1}(0) = \tau$, $F[f](0) = 0$ and $f(u) = t$ for some $t \in (0, \infty)$ implies $u \in (\tau, \infty)$. We now show that f^{-1} is in fact continuous and strictly increasing. Suppose $f^{-1}(s) = f^{-1}(t)$ for $s, t \in (0, \infty)$. Then the continuity of f and (4.25) dictate that $s = f(f^{-1}(s)) = f(f^{-1}(t)) = t$, which shows that f^{-1} is strictly increasing. To see that f^{-1} is also continuous, note that if there exists $t \in (0, \infty)$ such that $f^{-1}(t) > f^{-1}(t-)$, then $f^{-1}(t) \in (\tau, \infty)$ and the fact that f is strictly increasing on (τ, ∞) implies that $t = f(f^{-1}(t)) > f(f^{-1}(t-)) = f(f^{-1}(t)-) = t$, which leads to a contradiction. In addition, note that since f is strictly increasing on (τ, ∞) and since for $t \in (0, \infty)$, $f(v) = t$ implies $v > \tau$, it is clear that for $t \in (0, \infty)$ the right-hand side in (4.26) holds with f replaced by g and 0 replaced by τ. However, since f is strictly increasing on (τ, ∞), the right-hand side in (4.26) must be zero for $t \in (0, \infty)$, which establishes the fact that $f^{-1}(t) = F[f](t)$ for $t \in (0, \infty)$.

Now consider the sequence $\{f^n\}$ of right-continuous nondecreasing functions. We claim that given any $M \in [\tau, \infty)$,

$$\lim_{n\to\infty} \sup_{\tau\le u\le v\le M} \{v - u : f^n(v) = f^n(u)\} = 0. \tag{4.27}$$

We defer the proof of the claim to the end and instead first show that (4.24) follows from the claim. Let C be a compact subset of $(0, \infty)$ and choose $M \in [\tau, \infty)$ such that $M/2 \ge \sup_{t\in C} f^{-1}(t)$. From the fact that f^{-1} is continuous on $(0, \infty)$ and [26], Lemma 7.2, one deduces that

$$(f^n)^{-1} \to f^{-1}, \tag{4.28}$$

where the convergence is uniform for compact subsets of $(0, \infty)$ (and of $[0, \infty)$ if $\tau = 0$). Since $f^n \to f$ u.o.c., $f(t) = 0$ for $t \in [0, \tau]$, C is compact and

$0 \notin C$, there exists $N < \infty$ such that for all $n \geq N$ $\sup_{u\in[0,\tau]} f^n(u) \leq \min\{t : t \in C\}$. Moreover due to (4.28) one can assume that N is large enough so that $\sup_{n\geq N} \sup_{t\in C} (f^n)^{-1}(t) \leq M$. This implies that if there exists $n \geq N$ and $t \in C$ such that $f^n(u) = t$, then $u \in [\tau, M]$. Combining this with the fact that (4.26) holds with g replaced by f^n one obtains for $n \geq N$,

$$\sup_{t\in C} |(f^n)^{-1}(t) - F[f^n](t)| \leq \sup_{\tau\leq u\leq v\leq M} \{v - u : f^n(v) = f^n(u)\},$$

which, together with (4.27), yields

$$\lim_{n\to\infty} \sup_{t\in C} |(f^n)^{-1}(t) - F[f^n](t)| = 0. \tag{4.29}$$

Since for any $n \in \mathbb{N}$ and compact $C \subset (0, \infty)$,

$$\sup_{t\in C} |f^{-1}(t) - F[f^n](t)| \leq \sup_{t\in C} |f^{-1}(t) - (f^n)^{-1}(t)| + \sup_{t\in C} |(f^n)^{-1}(t) - F[f^n](t)|$$

combining (4.28) with (4.29) one obtains

$$\lim_{n\to\infty} \sup_{t\in C} |f^{-1}(t) - F[f^n](t)| = 0.$$

It is easy to see from the proof that C can be taken to be a compact subset of $[0, \infty)$ if $\tau = 0$.

To complete the proof, it only remains to establish the claim (4.27). Fix $M \in [\tau, \infty)$ and recall that $\dot{f}(t) \geq \theta > 0$ for a.e. $t \in [\tau, M]$. Moreover, since $f^n \to f$ u.o.c., given $\varepsilon > 0$ there exists $N < \infty$ such that for $n \geq N$,

$$\sup_{t\in[\tau,M]} |f^n(t) - f(t)| < \theta\varepsilon/2.$$

Given $\tau \leq s \leq t \leq M$, it follows that for $n \geq N$,

$$\begin{aligned} \big||f(t) - f(s)| - |f^n(t) - f^n(s)|\big| &\leq |f(t) - f^n(t) + f^n(s) - f(s)| \\ &\leq |f(t) - f^n(t)| + |f(s) - f^n(s)| \\ &\leq \theta\varepsilon. \end{aligned}$$

Consequently for any $\tau \leq s \leq t \leq M$ such that $t - s > \varepsilon$, for $n \geq N$,

$$f^n(t) - f^n(s) \geq f(t) - f(s) - \theta\varepsilon \geq \theta(t - s) - \theta\varepsilon > 0,$$

which implies that given any $M \in [\tau, \infty)$ and $\varepsilon > 0$ there exists $N < \infty$ such that for $n \geq N$,

$$\sup_{\tau\leq s\leq t\leq M} \{t - s : f^n(t) = f^n(s)\} \leq \varepsilon.$$

Sending $n \uparrow \infty$ and then $\varepsilon \downarrow 0$ yields (4.27). $\square$

THEOREM 4.11. *Suppose Assumptions* 2.3 *and* 4.5 *hold with* $\gamma > 0$, *and let* $\overline{V}$ *and* $\overline{W}$ *be defined by* (4.12). *If Condition* 4.8 *is satisfied, then a.s.* $\widetilde{V}^n \to \overline{V}$ *and* $\widetilde{W}^n \to \overline{W}$ *uniformly on every compact subset of* $(0, \infty)$ *as* $n \to \infty$. *Moreover, if* $\alpha_i > 0$, *then the convergence* $\widetilde{V}_i^n \to \overline{V}_i$ *and* $\widetilde{W}_i^n \to \overline{W}_i$ *is uniform on compact subsets of* $[0, \infty)$.

PROOF. The definition of $V^n(k)$ stated in (2.14), and the fluid scalings (4.1)–(4.3) show that for $t \in \mathrm{dom}(\widetilde{V}_i^n)$,

$$\begin{aligned}\widetilde{V}_i^n(t) &= \frac{1}{n}\big[\inf\big\{s \geq 0 : T_i^n(s) \geq T_i^n\big(C_i^n\big(A_i^n(nt)\big)\big) + U_i^n\big(C_i^n\big(A_i^n(nt)\big)\big)\big\} \\ &\qquad - C_i^n\big(A_i^n(nt)\big)\big] \\ &= \inf\big\{s \geq 0 : T_i^n(ns) \geq T_i^n\big(n\widetilde{C}_i^n(t)\big) + U_i^n\big(n\widetilde{C}_i^n(t)\big)\big\} - \widetilde{C}_i^n(t) \\ &= \inf\big\{s \geq 0 : \overline{T}_i^n(s) \geq \overline{T}_i^n\big(\widetilde{C}_i^n(t)\big) + \overline{U}_i^n\big(\widetilde{C}_i^n(t)\big)\big\} - \widetilde{C}_i^n(t).\end{aligned}$$

Using the mapping F defined in (2.12) one obtains the representation

$$\widetilde{V}_i^n = F[\overline{T}_i^n] \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n - \widetilde{C}_i^n. \tag{4.30}$$

By Assumption 2.1 for $n \in \mathbb{N}$, $\overline{T}_i^n + \overline{U}_i^n = \overline{U}_i^n(0) + \overline{H}_i^n$ is nondecreasing and piecewise constant, and hence $(\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n$ is nondecreasing and piecewise constant. Hence (see the discussion in Section 2.2.4) $\widetilde{V}_i^n$ is a nondecreasing right-continuous function with left limits on its domain of definition. Moreover, Lemmas 4.9(2) and 4.10 and the fact that $\bar{T}_i$ is continuous imply that for $i \in \mathcal{I}$ and $t \in (0, \infty)$,

$$F[\overline{T}_i](t) = (\overline{T}_i)^{-1}(t) \tag{4.31}$$

with the above equality also holding for $t = 0$ if $\alpha_i > 0$.

Consider the mapping $R : \mathcal{D}([0, \infty) : \mathbb{R}^J)^3 \to \mathcal{D}([0, \infty) : \mathbb{R}^J)$ defined by

$$R(f_1, f_2, f_3) = (f_1)^{-1} \circ (f_1 + f_2) \circ f_3 - f_3.$$

The last two displays and (4.12) show that

$$\overline{V}_i(t) = R(\overline{T}_i, \overline{U}_i, \iota)(t) \qquad \text{for } t \in (0, \infty) \tag{4.32}$$

and also for $t = 0$ if $\alpha_i > 0$. Theorem 4.3 shows that for $i \in \mathcal{I}$, P a.s.

$$\lim_{n\to\infty} \overline{T}_i^n = \overline{T}_i \quad \text{and} \quad \lim_{n\to\infty} \overline{U}_i^n = \overline{U}_i \qquad \text{u.o.c.}$$

Assumption 4.5 implies that P a.s. $\widetilde{C}_i^n \to \iota$ u.o.c. as $n \to \infty$. Thus P a.s.

$$(\overline{T}_i^n, \overline{U}_i^n, \widetilde{C}_i^n) \to (\overline{T}_i, \overline{U}_i, \iota) \qquad \text{u.o.c.} \tag{4.33}$$

Since $\overline{T}_i^n + \overline{U}_i^n \to \overline{T}_i + \overline{U}_i$ u.o.c. as $n \to \infty$, $\overline{T}_i + \overline{U}_i$ is nondecreasing and $\widetilde{C}_i^n \le \iota$, given any $M < \infty$ there exist $K, N < \infty$ such that for all $n \ge N$,

$$\sup_{t\in[0,M]} [(\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n] \le K.$$

From Lemmas 4.9(2) and 4.10 and (4.33) one concludes that for any $0 < \varepsilon \le M < \infty$ (and for $0 \le \varepsilon \le M < \infty$ if $\alpha_i > 0$),

$$\lim_{n\to\infty} \sup_{t\in[\varepsilon,M]} \left|F[\overline{T}_i^n] \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n(t) - (\overline{T}_i)^{-1} \circ (\overline{T}_i + \overline{U}_i)(t)\right| = 0. \tag{4.34}$$

Along with (4.31), (4.32) and the fact that $\widetilde{C}_i^n \to \iota$ u.o.c. this implies that

$$\lim_{n\to\infty} \sup_{t\in[\varepsilon,M]} |\widetilde{V}_i^n(t) - \overline{V}_i(t)| = 0. \tag{4.35}$$

The proof for the limit of the waiting times is similar. As above, using Theorem 4.3, Assumption 4.5 and the fact that $\overline{U}$ is continuous, it follows that for $i \in \mathcal{I}$, P a.s.

$$\lim_{n\to\infty} \big(\overline{T}_i^n(t), \overline{T}_i^n \circ \widetilde{C}_i^n(t) + \overline{U}_i^n\big(\widetilde{C}_i^n(t)-\big)\big) = \big(\overline{T}_i(t), \overline{T}_i(t) + \overline{U}_i(t)\big),$$

where the convergence is uniform for t in compact sets of $[0,\infty)$. Once again Lemmas 4.9(2) and 4.10 show that

$$\begin{aligned}\lim_{n\to\infty} \Big|&F[\overline{T}_i^n] \circ \big(\overline{T}_i^n \circ \widetilde{C}_i^n(t) + \overline{U}_i^n\big(\widetilde{C}_i^n(t)-\big)\big)\\ &- (\overline{T}_i^n)^{-1} \circ \big(\overline{T}_i^n \circ \widetilde{C}_i^n(t) + \overline{U}_i^n\big(\widetilde{C}_i^n(t)-\big)\big)\Big| = 0,\end{aligned}$$

where the convergence is uniform for t in compact subsets of $(0,\infty)$ (or of $[0,\infty)$ if $\alpha_i > 0$). The result then follows from (4.15), (4.32) and the fact that $\widetilde{C}_i^n \to \iota$ u.o.c. □

4.2. *Heavy traffic diffusion approximations.* As in the previous section, we consider a sequence of networks with associated processes H^n, A^n, L^n and so forth that satisfy Assumptions 2.1 and 2.3. In addition, we also assume that they satisfy the functional strong laws stated in Assumptions 4.2 and 4.5. Recall from Remark 4.6 that then $\gamma_i = \lambda_i/\mu_i$ for $i \in \mathcal{I}$. Also recall the defining equations (4.8), (4.9), (4.10) and (4.12) for $\overline{U}$, $\overline{X}$, $\overline{Y}$, $\overline{T}$, $\overline{Q}$, $\overline{V}$ and $\overline{W}$, and consider the diffusion scalings

$$\begin{aligned}&\hat{H}^n \doteq \sqrt{n}[\overline{H}^n - \gamma^n\iota], && \hat{A}^n \doteq \sqrt{n}[\overline{A}^n - \lambda^n\iota],\\ &\hat{U}^n \doteq \sqrt{n}[\overline{U}^n - \overline{U}], && \hat{X}^n \doteq \sqrt{n}[\overline{X}^n - \overline{X}],\\ &\hat{Y}^n \doteq \sqrt{n}[\overline{Y}^n - \overline{Y}], && \hat{T}^n \doteq \sqrt{n}[\overline{T}^n - \overline{T}],\\ &\hat{Q}^n \doteq \sqrt{n}[\overline{Q}^n - \overline{Q}],\end{aligned} \tag{4.36}$$

$$\hat{V}^n \doteq \sqrt{n}[\widetilde{V}^n - \overline{V}] \quad \text{and} \quad \hat{W}^n \doteq \sqrt{n}[\widetilde{W}^n - \overline{W}], \tag{4.37}$$

and for $i \in \mathcal{I}$,

$$\hat{L}_i^n(t) \doteq \sqrt{n}[\overline{L}_i^n(\gamma_i^n t) - \lambda_i^n t] \quad \text{and} \quad \hat{S}_i^n(t) \doteq \sqrt{n}[\overline{S}_i^n(\lambda_i^n t) - \gamma_i^n t]. \tag{4.38}$$

In the following sections we impose a heavy traffic condition that in particular implies that $\overline{U} = \overline{X} = \overline{Y} = \overline{Q} = \overline{V} = \overline{W} = 0$ (see Lemma 4.13) and so the diffusion scaling above simplifies considerably. Observe that the diffusion scalings for the primitive processes $\hat{H}^n$, $\hat{A}^n$, $\hat{L}^n$ and $\hat{S}^n$ have been chosen in such a way that in the heavy traffic limit, the scaled processes tend to driftless diffusions. For the case of $\hat{L}^n$ and $\hat{S}^n$, we also introduce a time change so that roughly speaking, for large enough $n \in \mathbb{N}$, $\hat{L}^n(t)$ represents the fluctuations around the mean of the number of customers served (in the scaled system) in the interval $[0, t]$ and $\hat{S}^n(t)$ captures the fluctuations in the amount of total service required for all customers that arrived into the (scaled) system in the interval $[0, t]$.

4.2.1. *The diffusion limit of the unfinished work process.* To prove the heavy traffic limit theorem for the unfinished work process we first assume that the primitive sequence $\{H^n\}$ satisfies a functional central limit theorem, in addition to the functional strong law imposed by Assumption 4.2.

ASSUMPTION 4.12. 1. There exists a random variable $\hat{u}$ taking values in $\mathbb{R}_+^J$ such that P a.s.

$$\lim_{n\to\infty} \frac{U^n(0)}{\sqrt{n}} = \hat{u}.$$

2. Almost surely,

$$\lim_{n\to\infty} \hat{H}^n = B \qquad \text{u.o.c.},$$

where B is a driftless J-dimensional Brownian motion with covariance matrix M^H.

3. There exists $\hat{c} \in \mathbb{R}^J$ such that

$$\lim_{n\to\infty} \sqrt{n}(\gamma^n - \alpha) = \hat{c}.$$

LEMMA 4.13. *Suppose Assumptions* 2.1, 2.3, 4.2, 4.5 *and* 4.12 *hold. Then* $\gamma = \alpha$, $\overline{U} = \overline{X} = \overline{Y} = \overline{Q} = \overline{V} = \overline{W} = 0$ *and* $\overline{T} = \alpha\iota$.

PROOF. Note that Assumption 4.12(1) implies that Assumption 4.2(1) holds with $\overline{u} = 0$. In addition, Assumption 4.12(3) implies that $\gamma = \alpha$, where as before $\gamma = \lim_{n\to\infty} \gamma^n$. From (4.7) this implies that $\nu = 0$, and by (4.8) it follows that $\overline{X} = 0$. Theorem 4.3 then shows that $\overline{U} = 0$, $\overline{Y} = 0$ and $\overline{T} = \alpha\iota$. From

definitions (4.10) and (4.12) it then follows that $\overline{Q} = \overline{V} = \overline{W} = 0$, which completes the proof. $\square$

Define

$$\hat{X} \doteq \hat{u} + B + \hat{c}\iota, \qquad \hat{U} \doteq \overline{\Gamma}(\hat{X}), \qquad \hat{Y} \doteq \hat{U} - \hat{X}, \tag{4.39}$$

where we recall that $\overline{\Gamma}$ is the GPS extended Skorokhod map.

THEOREM 4.14. *Suppose Assumptions* 2.1, 4.2 *and* 4.12 *are satisfied, and let* $\hat{U}$, $\hat{Y}$ *and* $\hat{X}$ *be defined as in* (4.39). *Then a.s.* $\hat{U}^n \to \hat{U}$, $\hat{Y}^n \to \hat{Y}$, $\hat{X}^n \to \hat{X}$ *and* $\hat{T}^n \to -\hat{Y}$ *u.o.c.*

PROOF. As in the proof of Theorem 4.3, the representation proved in Lemma 3.4 and homogeneity properties of the Skorokhod map imply that $\hat{U}^n = \Gamma(\hat{X}^n)$. We can also write $\hat{U}^n = \overline{\Gamma}(\hat{X}^n)$, since $\hat{X}^n$ is of bounded variation, and $\overline{\Gamma}$ and Γ coincide on paths of bounded variation by Theorem 3.3. Using the definitions of $\hat{X}^n$ and $\hat{H}^n$ we obtain

$$\hat{X}^n(t) = \hat{U}^n(0) + \hat{H}^n(t) + \sqrt{n}(\gamma^n - \alpha)t. \tag{4.40}$$

Hence by Assumption 4.12, P a.s. $\hat{X}^n \to \hat{X}$ u.o.c. Due to the Lipschitz continuity of $\overline{\Gamma}$, an application of the continuous mapping theorem yields a.s. $\hat{U}^n \to \overline{\Gamma}(\hat{X}) = \hat{U}$ u.o.c. Since $\hat{Y}^n = \hat{U}^n - \hat{X}^n$, the almost sure convergence $\hat{Y}^n \to \hat{Y}$ u.o.c. also follows. Finally, observe from (3.6) that

$$\hat{Y}^n = \sqrt{n}[\alpha\iota - T^n] = -\hat{T}^n.$$

Taking limits as $n \to \infty$ one obtains $\hat{T}^n \to -\hat{Y}$ u.o.c. $\square$

From the above theorem we see that the heavy traffic diffusion limit $\hat{U}$ of the unfinished work process has a representation in terms of the extended Skorokhod map applied to Brownian motion with drift. It follows from [20], Theorem 4.2, that $\hat{U}$ is a J-dimensional reflected diffusion process and from [20], Theorem 4.14, that $\hat{U}(\cdot \wedge \tau)$ is an $\mathcal{F}_t$-semimartingale, where τ is the first time to hit the origin. However, an interesting feature of this diffusion limit is that $\hat{U}$ need not in general be a semimartingale (the proof for the two-dimensional case follows from [27], Theorem 2).

The case when $\alpha_i = 0$ for one or more i, $1 \le i \le J$, is worthy of some discussion. Note that if $\alpha_i = 0$, then by Lemma 4.13 we must have $\gamma_i = 0$ as well. It is easy to deduce that then $M^H_{ii} = 0$ and B_i is identically zero. By Assumption 4.12(3), $\sqrt{n}\gamma^n_i \to \hat{c}_i \ge 0$ as $n \to \infty$. If $\hat{c}_i = 0$, then class i simply disappears from the limit process. Specifically, if $\hat{u}_i = 0$, then $\hat{U}_i(t) = 0$ for all $t \ge 0$, while if $\hat{u}_i > 0$, then $\hat{U}_i(t) = 0$ for $t \ge \tau$, where $\tau \doteq \inf\{t > 0 : \hat{U}_i(t) = 0\}$. On the other hand, if $\hat{c}_i > 0$, then class i is present in the limit, but the sample

paths of $\hat{U}_i(\cdot)$ are a.s. of bounded variation. Focusing on the case $J = 2$ with $\alpha_2 = 0$ (so that class 1 is high priority), we can compare our results with those for preemptive (resume) priority queues in [13] and [25]. In these papers it was assumed (translating to our notation) that $\gamma_i > 0$ for $i = 1, 2$. Thus in [13, 25] the system with only high priority customers is not in heavy traffic. This leads to state space collapse, where $\hat{U}_1(t) = 0$ for $t > 0$. In our setting, with $\gamma_2 = 0$, the system is in heavy traffic even without the low priority class and there is no state space collapse.

In the next lemma we show that the work-conserving nature of the GPS discipline yields simple limits for the total unfinished work process. Recall that the classical one-dimensional Skorokhod or reflection mapping [24] $\Gamma_1 : \mathcal{D}([0, \infty) : \mathbb{R}) \to \mathcal{D}([0, \infty) : \mathbb{R}_+)$ is defined by

$$\Gamma_1(f)(t) \doteq f(t) - \left[\inf_{s\in[0,t]} f(s)\right] \wedge 0 \qquad \text{for } t \in [0, \infty).$$

Let $X_T^n(t) \doteq \sum_{i=1}^J X_i^n(t)$ and $U_T^n(t) \doteq \sum_{i=1}^J U_i^n(t)$, and let $\overline{X}_T^n$, $\overline{X}_T$, $\hat{X}_T^n$, $\hat{X}_T$, $\overline{U}_T^n$ and $\hat{U}_T^n$ be defined analogously. Also let

$$\overline{U}_T \doteq \Gamma_1(\overline{X}_T) \quad \text{and} \quad \hat{U}_T \doteq \Gamma_1(\hat{X}_T).$$

LEMMA 4.15. *Suppose Assumption* 2.1 *is satisfied, suppose U and X are defined by* (2.7) *and* (3.5), *respectively, and let* $\overline{U}_T^n$, $\hat{U}_T^n$, $\overline{U}_T$ *and* $\hat{U}_T$ *be defined as above. If Assumption* 4.2 *is satisfied, then P a.s.* $\lim_{n\to\infty} \overline{U}_T^n = \overline{U}_T$ *u.o.c. If in addition Assumption* 4.12 *is also satisfied, then* $\overline{U}_T = 0$ *and a.s.* $\lim_{n\to\infty} \hat{U}_T^n = \hat{U}_T$ *u.o.c.*

PROOF. We first show that $U_T^n = \Gamma_1(X_T^n)$. Note that from the definition of X_i given in (3.5) and the fact that $\sum_{i=1}^J \alpha_i = 1$, it follows that

$$X_T^n(t) = \sum_{i=1}^J U_i^n(0) + \sum_{i=1}^J H_i^n(t) - t$$

and using (2.7) we can write

$$U_T^n(t) = \sum_{i=1}^J U_i^n(0) + \sum_{i=1}^J H_i^n(t) - \sum_{i=1}^J \sum_{E\subset\mathcal{I},\, i\notin E} \alpha_i^E I_E^n(t). \tag{4.41}$$

By the definitions of α_i^E, $i \in \mathcal{I}$, $E \subset \mathcal{I}$, $\sum_{i:i\notin E} \alpha_i^E = 1$ for every $E \subset \mathcal{I}$, $E \neq \mathcal{I}$. Indeed, this is the work-conserving property of the GPS discipline. Combining the last two displays and noting the fact that $t = \sum_{E\subset\mathcal{I}} I_E^n(t) + I_{\mathcal{I}}^n(t)$ for every $n \in \mathbb{N}$, (4.41) can be rewritten as

$$U_T^n(t) = X_T^n(t) + I_{\mathcal{I}}^n(t).$$

Since each $U_i^n(t) \geq 0$, it is clear that $U_T^n(t) \in \mathbb{R}_+$ for every $t \in [0, \infty)$. Moreover, $I_{\mathbb{1}}^n$ is clearly nondecreasing and increases only at points t where $E^n(t) = \mathbb{1}$ or equivalently where $U_T^n(t) = 0$. It is well known that these properties uniquely characterize the one-dimensional Skorokhod map [24] and therefore we conclude that $U_T^n = \Gamma_1(X_T^n)$.

From the explicit expression for Γ_1 it is easy to see that the one-dimensional Skorokhod map is Lipschitz continuous (with constant 2). Moreover, due to the scaling properties of Γ_1 we have $\overline{U}_T^n = \Gamma_1(\overline{X}_T^n)$. Hence by the continuous mapping theorem and Theorem 4.3, it follows that P a.s.

$$\lim_{n\to\infty} \overline{U}_T^n = \lim_{n\to\infty} \Gamma_1(\overline{X}_T^n) = \Gamma_1(\overline{X}_T) = \overline{U}_T \qquad \text{u.o.c.}$$

If Assumption 4.12 is also satisfied, then by Lemma 4.13, $\overline{X}_T = 0$, and so $\overline{U}_T = 0$ and

$$\hat{U}_T^n = \sqrt{n}[U_T^n - \overline{U}_T] = \sqrt{n}U_T^n = \sqrt{n}\Gamma_1(\overline{X}_T^n) = \Gamma_1(\hat{X}_T^n),$$

where we used the radial homogeneity of the one-dimensional Skorokhod map in the last equality. Taking limits, using the continuous mapping theorem again along with Theorem 4.14, we conclude that P a.s.

$$\lim_{n\to\infty} \hat{U}_T^n = \Gamma_1(\hat{X}_T) \qquad \text{u.o.c.} \qquad \square$$

4.2.2. *The diffusion limit of the queue length process.* In this section, we introduce some more detailed assumptions that imply Assumption 4.12. When Assumption 4.5 and Assumption 4.12(3) hold, from Remark 4.6 it follows that

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\lambda_i^n}{\mu_i^n} - \alpha_i\right) = \lim_{n\to\infty} \sqrt{n}(\gamma_i^n - \alpha_i) = \hat{c}_i. \tag{4.42}$$

As stated in Lemma 4.13, this corresponds to a heavy traffic condition that ensures that the fluid limits $\overline{U}$, $\overline{Q}$, $\overline{W}$ and $\overline{V}$ are trivial.

Let $\hat{A}$ and $\hat{L}$ denote driftless J-dimensional Brownian motions with joint covariance matrix $\mathcal{M}$, where

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}^{AA} & \mathcal{M}^{AL} \\ \mathcal{M}^{LA} & \mathcal{M}^{LL} \end{pmatrix}$$

with

$$\mathcal{M}_{ij}^{AA} = \operatorname{cov}(\hat{A}_i, \hat{A}_j),$$

$$\mathcal{M}_{ij}^{AL} = \operatorname{cov}(\hat{A}_i, \hat{L}_j),$$

$$\mathcal{M}_{ij}^{LA} = \mathcal{M}_{ji}^{AL},$$

$$\mathcal{M}_{ij}^{LL} = \operatorname{cov}(\hat{L}_i, \hat{L}_j).$$

ASSUMPTION 4.16. 1. There exists a random variable $\hat{q}$ taking values in $\mathbb{R}_+^J$ such that a.s.

$$\lim_{n\to\infty} \frac{Q^n(0)}{\sqrt{n}} = \hat{q}.$$

2. Almost surely,

$$\lim_{n\to\infty} \hat{A}^n = \hat{A} \qquad \text{u.o.c.}$$

3. Almost surely,

$$\lim_{n\to\infty} \hat{L}^n = \hat{L} \qquad \text{u.o.c.}$$

In analogy with the fluid scaling relationship $\gamma_i = \lambda_i/\mu_i$ that holds when both Assumptions 4.2 and 4.5 are satisfied (see Remark 4.6), the following lemma establishes the consistency conditions required for both Assumptions 4.12 and 4.16 to be satisfied.

LEMMA 4.17. *Suppose Assumptions* 4.12 *and* 4.16 *hold. Then*

$$M_{ij}^H = \frac{1}{\mu_i\mu_j}\Big[\mathcal{M}_{ij}^{AA} - \mathcal{M}_{ij}^{AL} - \mathcal{M}_{ji}^{AL} + \mathcal{M}_{ij}^{LL}\Big]. \tag{4.43}$$

PROOF. From the definitions of S_i^n and L_i^n given in Section 2.2.1 and the relationship (2.5) it follows that for every $n, m \in \mathbb{N}$ and $t \in [0,\infty)$, $L_i^n(S_i^n\lfloor mt\rfloor) = \lfloor mt\rfloor$, and thus

$$\lim_{m\to\infty} \frac{L_i^n(S_i^n(\lfloor mt\rfloor))}{m} = t,$$

where the convergence is uniform on compact sets. Assumption 4.5(3) and Lemma 4.1 then imply that for every $n \in \mathbb{N}$,

$$\lim_{m\to\infty} \frac{S_i^n(\lfloor mt\rfloor)}{m} = \frac{t}{\mu_i^n} \qquad \text{u.o.c.}$$

uniformly on compacts, which along with Assumption 4.5(4) implies

$$\lim_{n\to\infty} \overline{S}_i^n = \frac{\iota}{\mu_i} \qquad \text{u.o.c.} \tag{4.44}$$

Recall the diffusion scalings given in (4.38) and observe that

$$\overline{L}_i^n\big(\overline{S}_i^n(\lambda_i^n t)\big) = \frac{\lfloor \lambda_i^n nt\rfloor}{n} + O\Big(\frac{1}{n}\Big),$$

which implies that

$$\begin{aligned}\sqrt{n}[\,\overline{L}_i^n(\overline{S}_i^n(\lambda_i^n t)) - \mu_i^n \overline{S}_i^n(\lambda_i^n t)]\\ = -\mu_i^n\sqrt{n}[\,\overline{S}_i^n(\lambda_i^n t) - \gamma_i^n t] - \sqrt{n}\Big(\lambda_i^n t - \frac{\lfloor \lambda_i^n n t\rfloor}{n}\Big) + O\Big(\frac{1}{\sqrt{n}}\Big),\end{aligned}$$

which in turn implies that

$$\hat{L}_i^n\Big(\frac{\overline{S}_i^n(\lambda_i^n t)}{\gamma_i^n}\Big) = -\mu_i^n \hat{S}_i^n(t) + O\Big(\frac{1}{\sqrt{n}}\Big).$$

Taking limits as $n \to \infty$ in the above display, and using (4.42), (4.44) and the continuity of $\hat{L}$ one concludes that

$$\hat{S}_i \doteq \lim_{n\to\infty} \hat{S}_i^n = -\frac{1}{\mu_i}\hat{L}_i \qquad \text{u.o.c.} \tag{4.45}$$

The above identity along with (2.1) and the fact that $\lambda_i^n/\mu_i^n = \gamma_i^n$ shows that

$$\begin{aligned}\hat{H}_i^n(t) &= \sqrt{n}[\,\overline{S}_i^n(\overline{A}_i^n(t) + \overline{Q}_i^n(0)) - \gamma_i^n t - \overline{S}_i^n(\overline{Q}_i^n(0))]\\ &= \hat{S}_i^n\Big(\frac{(\overline{A}_i^n(t) + \overline{Q}_i^n(0))}{\lambda_i^n}\Big) + \frac{\sqrt{n}}{\mu_i^n}[\overline{A}_i^n(t) - \lambda_i^n t]\\ &\quad + \sqrt{n}\Big[\frac{\overline{Q}_i^n(0)}{\mu_i^n} - \overline{S}_i^n(\overline{Q}_i^n(0))\Big].\end{aligned}$$

Sending $n \to \infty$ in the above display, Remark 4.6, (4.45) and Assumptions 4.5 and 4.16 yield

$$B_i(t) \doteq \lim_{n\to\infty} \hat{H}_i^n(t) = \hat{S}_i(t) + \frac{1}{\mu_i}\hat{A}_i(t) = \frac{1}{\mu_i}[\hat{A}_i(t) - \hat{L}_i(t)], \tag{4.46}$$

from which (4.43) follows. □

THEOREM 4.18. *Suppose Assumptions* 2.3, 4.5, 4.12 *and* 4.16 *are satisfied, and let* $\hat{U}$ *be as in* (4.39). *Then P a.s.* $\hat{Q}^n \to \hat{Q}$ *u.o.c., where* $\hat{Q}_i \doteq \mu_i \hat{U}_i$ *for* $i \in \mathcal{I}$.

PROOF. It is easy to see that if Assumption 2.3 is satisfied, then Assumption 2.1 is also satisfied. Moreover, by Remark 4.6, we know that if Assumption 4.5 is satisfied, then Assumption 4.2 is satisfied with $\gamma_i^n = \lambda_i^n/\mu_i^n$. Under the heavy traffic condition Assumption 4.12(3), $\gamma = \alpha$ and by Lemma 4.13, $\overline{Q} = 0$ and $\overline{Y} = 0$. Therefore, from (4.11) and (4.36) it follows that

$$\begin{aligned}\hat{Q}_i^n(t) &= \sqrt{n}\,\overline{Q}_i^n(t)\\ &= \sqrt{n}\,\overline{Q}_i^n(0) + \sqrt{n}\,\overline{A}_i^n(t) - \sqrt{n}\,\overline{L}_i^n(\overline{T}_i^n(t))\end{aligned}$$

$$= \sqrt{n}\,\overline{Q}_i^n(0) + \hat{A}_i^n(t) - \hat{L}_i^n\left(\frac{\overline{T}_i^n(t)}{\gamma_i^n}\right) + \sqrt{n}(\lambda_i^n - \mu_i^n\alpha_i)t$$
$$+ \sqrt{n}\mu_i^n\left(\alpha_i t - \overline{T}_i^n(t)\right)$$
$$= \mu_i^n\left[\frac{\sqrt{n}\overline{Q}_i^n(0)}{\mu_i^n} + (\mu_i^n)^{-1}\left[\hat{A}_i^n(t) - \hat{L}_i^n\left(\frac{\overline{T}_i^n(t)}{\gamma_i^n}\right)\right] + \sqrt{n}(\gamma_i^n - \alpha_i)t\right.$$
$$\left. + \sqrt{n}\left(\overline{Y}_i^n(t)\right)\right].$$

Since $\gamma^n \to \gamma$, and since Theorem 4.3 and Assumption 4.12 show that $\gamma = \alpha$ and a.s. $\overline{T}^n \to \alpha\iota$ u.o.c., one obtains

$$\lim_{n\to\infty} \hat{Q}_i^n(t) = \mu_i[\hat{u}_i + \mu_i^{-1}[\hat{A}_i(t) - \hat{L}_i(t)] + \hat{c}_i t + \hat{Y}_i(t)].$$

Substituting (4.46) into the last display and using (4.39), we deduce that

$$\hat{Q}_i = \lim_{n\to\infty} \hat{Q}_i^n = \mu_i[\hat{X}_i + \hat{Y}_i] = \mu_i\hat{U}_i$$

u.o.c., which concludes the proof of the theorem. □

4.2.3. *The diffusion limits of the sojourn and waiting time processes.* As in the case of fluid limits, the proof of heavy traffic limits for the sojourn and waiting times again relies on their characterization via the maps $F[f]$ and f^{-1} introduced in Section 2.2.4. In this section we assume that $\alpha \in (0,1]^J$. We define the process $\hat{V}$ as

$$\hat{V}_i \doteq \frac{\hat{U}_i}{\alpha_i} \qquad \text{for } i \in \mathcal{I}, \tag{4.47}$$

where $\hat{U}$ is as defined in (4.39).

THEOREM 4.19. *Suppose Assumptions* 4.2, 4.5, 4.12 *and* 4.16 *hold*, $\alpha \in (0,1]^J$ *and* $\hat{V}$ *is defined by* (4.47). *Then a.s.* $\hat{V}^n \to \hat{V}$ *and* $\hat{W}^n \to \hat{V}$ *u.o.c.*

PROOF. We first observe that by Lemma 4.13, $\overline{T} = \alpha\iota$, $\overline{U} = 0$ and $\overline{W} = \overline{V} = 0$. Let $(T_i^n)^{-1}$ be the inverse of T_i^n as defined by (2.16). It is clear that $(\overline{T}_i)^{-1} = \iota/\alpha_i$. Since $\overline{T}_i^n \to \overline{T}_i$ u.o.c. by Theorem 4.3 and $(\bar{T}_i)^{-1}$ is strictly increasing, [26], Theorem 7.2, implies that P a.s.

$$\overline{T_i^{-1}} \doteq \lim_{n\to\infty} \overline{(T_i^n)^{-1}} = \frac{\iota}{\alpha_i} \qquad \text{u.o.c. for } i \in \mathcal{I}. \tag{4.48}$$

Now consider the diffusion scaling

$$\widehat{(T_i^n)^{-1}} \doteq \sqrt{n}\left[\overline{(T_i^n)^{-1}} - \frac{\iota}{\alpha_i}\right] \qquad \text{for } i \in \mathcal{I}.$$

Employing the same reasoning that was used in the proof of (4.43) in Lemma 4.17, one can deduce that P a.s. and u.o.c.

$$\widehat{T_i^{-1}} \doteq \lim_{n\to\infty} \widehat{(T_i^n)^{-1}} = -\alpha_i^{-3/2}\hat{T}_i = \alpha_i^{-3/2}\hat{Y}_i, \tag{4.49}$$

where the last equality follows by Theorem 4.14.

We now argue by contradiction to show that P a.s. for every $M < \infty$,

$$\lim_{n\to\infty} \sqrt{n} \max_{t\in[0,M]} \sup_{0\le u\le v<\infty} \{v-u : \overline{T}_i^n(v) = \overline{T}_i^n(u) = t\} = 0. \tag{4.50}$$

Suppose (4.50) did not hold. Then there exists $t \in [0, M]$ and $\varepsilon > 0$ such that for every $n \in \mathbb{N}$ there exists an interval $(u_n, v_n) \in [0,\infty)$ with $|v_n - u_n| > \varepsilon/\sqrt{n}$ such that $\overline{T}_i^n(v_n) = \overline{T}_i^n(u_n) = t$. This implies that $(\overline{T}_i^n)^{-1}(\underline{t}) \le u_n$ for $\underline{t} < t$ and $(\overline{T}_i^n)^{-1}(\bar{t}) > v_n$ for $\bar{t} > t$, which implies that

$$\sqrt{n}\big[(\overline{T}_i^n)^{-1}(\bar{t}) - (\overline{T}_i^n)^{-1}(\underline{t})\big] > \varepsilon.$$

Letting $\bar{t} \downarrow t$ and $\underline{t} \uparrow t$, the last display along with (4.49) implies that $\hat{Y}_i$ has a jump at t, which contradicts the fact that $\hat{Y}_i$ is continuous and hence establishes (4.50).

Since

$$0 \le \sqrt{n}\big[(\overline{T}_i^n)^{-1} - F[\overline{T}_i^n]\big] \le \sqrt{n} \max_{t\in[0,M]} \sup_{0\le u\le v<\infty} \{v-u : \overline{T}_i^n(v) = \overline{T}_i^n(u) = t\},$$

(4.50) implies that $\lim_{n\to\infty} g_n = 0$ u.o.c., where

$$g_n \doteq \sqrt{n}\big[F[\overline{T}_i^n] \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n - (\overline{T}_i^n)^{-1} \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n\big]. \tag{4.51}$$

Using the representation for $\widetilde{V}^n$ given in (4.30) and the definition (4.37) of $\hat{V}^n$ we obtain

$$\begin{aligned}
\hat{V}_i^n &= \sqrt{n}\,\big[F[\overline{T}_i^n] \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n - \widetilde{C}_i^n\big] \\
&= \sqrt{n}\big[\widehat{(T_i^n)^{-1}} \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n - \widetilde{C}_i^n\big] + g_n \\
&= \widehat{(T_i^n)^{-1}} \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n + \sqrt{n}\Big[\frac{1}{\alpha_i}\overline{T}_i^n \circ \widetilde{C}_i^n - \widetilde{C}_i^n\Big] + \frac{1}{\alpha_i}\overline{U}_i^n \circ \widetilde{C}_i^n + g_n \\
&= \widehat{(T_i^n)^{-1}} \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n + \frac{1}{\alpha_i}\hat{T}_i^n \circ \widetilde{C}_i^n + \frac{1}{\alpha_i}\hat{U}_i^n \circ \widetilde{C}_i^n + g_n.
\end{aligned}$$

The mapping that takes

$$\begin{aligned}
&\big(\widehat{(T_i^n)^{-1}}, \hat{T}_i^n, \overline{T}_i^n, \overline{U}_i^n, \widetilde{C}_i^n\big) \\
&\qquad \to \widehat{(T_i^n)^{-1}} \circ (\overline{T}_i^n + \overline{U}_i^n) \circ \widetilde{C}_i^n + \frac{1}{\alpha_i}\hat{T}_i^n \circ \widetilde{C}_i^n + \frac{1}{\alpha_i}\hat{U}_i^n \circ \widetilde{C}_i^n
\end{aligned}$$

is clearly continuous in the u.o.c. topology at the point $(\widehat{T_i^{-1}}, \hat{T}_i, \alpha_i \iota, 0, \iota)$. Thus by the continuous mapping theorem, the fact that $g_n \to 0$ u.o.c. and the fact that the limit is continuous, one concludes that

$$\lim_{n\to\infty} \hat{V}_i^n = \hat{T}_i^{-1} \circ \alpha_i \iota + \frac{1}{\alpha_i}\hat{T}_i + \frac{\hat{U}_i}{\alpha_i} = \hat{V}_i \qquad \text{u.o.c.},$$

where the last equality follows using (4.49) and the fact that $\hat{T}_i^{-1} \circ \alpha_i \iota = \sqrt{\alpha_i}\hat{T}_i^{-1}$.

The proof for the waiting time is very similar and is thus omitted. □

REMARK. If the convergence in the various assumptions of the paper hold only in distribution rather than a.s., then using the Skorokhod representation theorem one can still obtain all the results of the paper, with the convergence in distribution rather than almost surely.

REFERENCES

[1] BERNARD, A. and EL KHARROUBI, A. (1991). Régulation de processus dans le premier orthant de $\mathbb{R}^n$. *Stochastics Stochastics Rep.* **34** 149–167.

[2] BERTSIMAS, D., PASCHALIDIS, I. CH. and TSITSIKLIS, J. N. (1999). Large deviations analysis of the generalized processor sharing policy. *Queueing Systems* **32** 319–349.

[3] BRAMSON, M. (1998). Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems Theory Appl.* **28** 7–31.

[4] BRAMSON, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems Theory Appl.* **30** 89–148.

[5] BUDHIRAJA, A. and DUPUIS, P. (1999). Simple necessary and sufficient conditions for the stability of constrained processes. *SIAM J. Appl. Math.* **59** 1686–1700.

[6] CHEN, H. and MANDELBAUM, A. (1991). Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann. Probab.* **19** 1436–1519.

[7] DUPUIS, P. and ISHII, H. (1991). On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics* **35** 31–62.

[8] DUPUIS, P. and RAMANAN, K. (1998). A Skorokhod problem formulation and large deviation analysis of a processor sharing model. *Queueing Systems Theory Appl.* **28** 109–124.

[9] DUPUIS, P. and RAMANAN, K. (1999). Convex duality and the Skorokhod problem—I. *Probab. Theory Related Fields* **115** 153–195.

[10] DUPUIS, P. and RAMANAN, K. (1999). Convex duality and the Skorokhod problem—II. *Probab. Theory Related Fields* **115** 197–236.

[11] DUPUIS, P. and RAMANAN, K. (2000). A multiclass feedback queueing network with a regular Skorokhod problem. *Queueing Systems Theory Appl.* **36** 327–349.

[12] DUPUIS, P. and RAMANAN, K. (2002). A time-reversed representation for the tail probabilities of stationary reflected Brownian motion. *Stochastic Process. Appl.* **98** 253–287.

[13] HARRISON, J. M. (1973). A limit theorem for priority queues in heavy traffic. *J. Appl. Probab.* **10** 907–912.

[14] HARRISON, J. M. and REIMAN, M. I. (1981). Reflected Brownian motion on an orthant. *Ann. Probab.* **9** 302–308.

[15] MANDELBAUM, A. and VAN DER HEYDEN, L. (1987). Complementarity and reflection. Unpublished manuscript.

[16] MASSOULIÉ, L. (1999). Large deviations for polling and weighted fair queueing service systems. *Advances in Performance Analysis* **2** 103–128.
[17] PAREKH, A. K. and GALLAGER, R. G. (1993). A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Transactions on Networking* **1** 344–357.
[18] PAREKH, A. K. and GALLAGER, R. G. (1994). A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking* **2** 137–150.
[19] PUHALSKII, A. (1994). On the invariance principle for the first passage time. *Math. Oper. Res.* **19** 946–954.
[20] RAMANAN, K. (2001). Reflected diffusions defined via the extended Skorokhod map. Preprint.
[21] REIMAN, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* **9** 441–458.
[22] REIMAN, M. I. and WILLIAMS, R. J. (1988). A boundary property of semimartingale reflecting Brownian motions. *Probab. Theory Related Fields* **77** 87–97.
[23] ROYDEN, H. L. (1989). *Real Analysis*. Macmillan, New York.
[24] SKOROKHOD, A. V. (1961). Stochastic equations for diffusions in a bounded region. *Theory Probab. Appl.* **6** 264–274.
[25] WHITT, W. (1971). Weak convergence theorems for priority queues: Preemptive-resume discipline. *J. Appl. Probab.* **8** 78–94.
[26] WHITT, W. (1980). Some useful functions for functional central limit theorems. *Math. Oper. Res.* **5** 67–85.
[27] WILLIAMS, R. J. (1985). Reflected Brownian motion in a wedge: Semimartingale property. *Probab. Theory Related Fields* **69** 161–176.
[28] WILLIAMS, R. J. (1998). Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems Theory Appl.* **30** 27–88.
[29] ZHANG, Z.-L., TOWSLEY, D. and KUROSE, J. (1995). Statistical analysis of the generalized processor sharing discipline. *IEEE Journal on Selected Areas in Communications* **13** 1071–1080.

LUCENT TECHNOLOGIES
BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974
E-MAIL: kavita@lucent.com
marty@lucent.com