A heavy traffic limit theorem for a class of open queueing networks with finite buffers *

J.G. Dai [a] and W. Dai [b,**]

[a] *School of Industrial and Systems Engineering, and School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA*
E-mail: dai@isye.gatech.edu
[b] *School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA*

Submitted 1 February 1998; accepted 1 December 1998

We consider a queueing network of d single server stations. Each station has a finite capacity waiting buffer, and all customers served at a station are homogeneous in terms of service requirements and routing. The routing is assumed to be deterministic and hence feedforward. A server stops working when the downstream buffer is full. We show that a properly normalized d-dimensional queue length process converges in distribution to a d-dimensional semimartingale reflecting Brownian motion (RBM) in a d-dimensional box under a heavy traffic condition. The conventional continuous mapping approach does not apply here because the solution to our Skorohod problem may not be unique. Our proof relies heavily on a uniform oscillation result for solutions to a family of Skorohod problems. The oscillation result is proved in a general form that may be of independent interest. It has the potential to be used as an important ingredient in establishing heavy traffic limit theorems for general finite buffer networks.

Keywords: finite capacity network, blocking probabilities, loss network, semimartingale reflecting Brownian motion, RBM, heavy traffic, limit theorems, oscillation estimates

1. Introduction

This paper proves a heavy traffic limit theorem for an open queueing network with finite buffers. The queueing network has d single server stations. Each station has a finite capacity waiting buffer, and all customers served at a station are *homogeneous* in terms of service requirements and routing. The routing is assumed to be deterministic and hence feedforward. Since there is a single customer class associated with each station, our network is a single class queueing network as opposed to the multiclass queueing networks widely discussed in the literature in recent years (see, e.g., Harrison [24]).

* Research supported in part by NSF Grants DMI-9457336 and DMI-9813345, US–Israel Binational Science Foundation Grant 94-00196 and a grant from Harris Semiconductor.
** Current address: End-to-End Network Architecture Department, Lucent Technologies, Warren, NJ 07059, USA. E-mail: wdai@lucent.com.

Queueing networks have been used to model telecommunication networks and manufacturing systems. All these networks have finite buffers in practice. See, e.g., Bertsekas and Gallager [3], Buzacott [9], Mitra and Mitrani [31], Perros and Altiok [32], and Yao [40]. In some applications, notably in some manufacturing systems like existing wafer fabrication facilities, buffer constraints have not been a major problem. Therefore, it is safe to ignore buffer constraints in the analysis of these networks. However, in telecommunication networks, more recently in asynchronous transfer mode (ATM) networks, buffer constraints have a major impact on system performances (see, e.g., Elwalid and Mitra [20] or Kroner et al. [29]). Thus, it is imperative to model the finiteness of the buffer sizes in these networks.

In our network the interarrival times and service times at each station are assumed to be independent, identically distributed (iid) sequences with finite first two moments. We show that the normalized d-dimensional queue length process converges in distribution to a d-dimensional reflecting Brownian motion (RBM) under a heavy traffic condition. The RBM lives in a d-dimensional box. The Brownian data, including the drift vector, covariance matrix and reflection matrix, can be calculated explicitly from the moments, network topology and the blocking mechanism employed. There are algorithms to numerically compute the stationary distribution of the RBM. Therefore, one can obtain performance estimates for the queueing network, like blocking probabilities and average queue lengths, from their Brownian counterparts [15].

The normalization of the queue length involves a scaling in time by a factor n and a scaling in space by a factor $1/\sqrt{n}$ for large n. Thus the heavy traffic limit theorem provides qualitative insight for the queueing network when it is operated for a long period of time, and each individual customer's movement is not of primary concern. The heavy traffic condition assumes that the traffic intensity ρ_i at each station i is close to 1 so that $1-\rho_i$ is of order $1/\sqrt{n}$. In addition, it requires the buffer size at a station is of order $\sqrt{n}$. The limit theorem suggests that this is the magnitude of the buffer size for the network to experience a "moderate level" of blocking.

Although many blocking mechanisms can be employed for a finite buffer network, we will focus on the "block-and-hold-0" mechanism. Under such a blocking mechanism, a server will stop working whenever an immediate downstream buffer is full. Therefore, the number of blocked customers that have completed services is 0. Readers are referred to Cheng and Yao [13] or Cheng [12] for the definition of the general "block-and-hold-k" mechanism. We note that the terms "manufacturing blocking" and "communication blocking" may not have a standard meaning in the literature; see, e.g., Cheng [12], and Konstantopoulos and Walrand [28]. A loss mechanism will be briefly discussed in section 9.

Due to the finiteness of the buffer sizes and the blocking mechanism used, the Skorohod problem associated with the queueing network may not have a unique solution (see the example at the end of section 5). Therefore the conventional continuous mapping approach, as used in Iglehart and Whitt [25,26] for feedforward single class networks, in Reiman [34] for single class networks with feedback and in Peterson [33] for feedforward multiclass queueing networks, does not apply here, although some

authors, such as Bardhan and Mithal [1], attempted such an extension. Instead we establish a uniform oscillation result for solutions to a sequence of Skorohod problems. Using this result, one can establish that the sequence of normalized queue length processes is precompact in the space of right continuous paths with left limits. Each limit point of the sequence is shown to be an RBM. Care has been taken to show that the limit satisfies a martingale property which is a defining property of the RBM. (Lemma 7.1 of this paper plays a key role in proving this martingale property. The proof of this lemma is adapted from Williams [38].) Finally, the heavy traffic limit theorem follows from the uniqueness (in distribution) of the RBM [18].

Almost all prior proofs of heavy traffic limit theorems for open networks assume the buffer sizes are infinite. For multiclass queueing networks, the mapping associated with the Skorohod problem is not well defined in general, as illustrated by an example of Dai et al. [17] which is included as appendix A of Williams [38]. The nonuniqueness excludes the usage of the continuous mapping theorem used in Iglehart and Whitt [25, 26], Reiman [34], Johnson [27], Peterson [33], and Chen and Zhang [10] to prove heavy traffic limit theorems. Reiman [35] proved a heavy traffic limit theorem for a multiclass station; see Dai and Kurtz [16] for an alternative proof and extension. Chen and Zhang [11] showed a heavy traffic limit theorem for a multiclass FIFO network with a restrictive spectral radius condition on a certain matrix. Although these three works went beyond the conventional continuous mapping paradigm, until very recently, we have not seen a viable approach to the proof of general heavy traffic limit theorems. The contemporaneous, independent works of Bramson [7] and Williams [38,39] provided sufficient conditions for a heavy traffic limit theorem for multiclass queueing networks under many conventional queueing disciplines, including the FIFO discipline, static buffer priority discipline, and head-of-the-line proportional processor sharing (HLPPS) discipline. These results represent a major breakthrough for proving heavy traffic limit theorems for infinite buffer multiclass queueing networks. In fact, using the sufficient conditions and Bramson [5,6], they established new heavy traffic limit theorems for FIFO networks of Kelly type and open multiclass queueing networks under the HLPPS discipline. The two key ingredients in establishing their heavy traffic limit theorems are oscillation result [38] and "state space collapse" [7].

Although the oscillation result in this paper looks similar to the oscillation result in [38], neither one implies the other. Our oscillation result deals with the Skorohod problem in a general state space and requires some control on the jump sizes of the pushing process, whereas Williams' result deals with a more general family of perturbed Skorohod problems in an orthant. Our oscillation result, which is proved in a much more general setting than needed in this paper, has the potential to be used as an important ingredient to prove a heavy traffic limit theorem for a general *finite* buffer queueing network, although other important ingredients, like deadlock in feedback networks and "state space collapse" in multiclass networks, have to be dealt with separately.

We now introduce the notation to be used in the paper. The number of stations in the network is assumed to be $d \geqslant 1$. Let $\boldsymbol{I} = \{1, \ldots, d\}$. The set of nonnegative

integers is denoted by $\mathbb{Z}_+$, and the k-dimensional nonnegative lattice is denoted by $\mathbb{Z}_+^k$. We use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space. Let $\mathbb{R}_+ = [0,\infty)$. Unless stated otherwise, all vectors are envisioned as column vectors. The prime symbol on a vector or a matrix denotes transpose. For $a = (a_1,\ldots,a_k)' \in \mathbb{R}^k$, $|a| = \max_{i=1}^k |a_i|$. For an $n \times k$ matrix A, $||A|| = \max_{i=1}^n \sum_{j=1}^k |A_{ij}|$. For a vector $a \in \mathbb{R}^k$, we use diag(a) to denote the $k \times k$ diagonal matrix whose diagonal entries are given by the components of a. Vector inequalities are interpreted componentwise. We use e to denote the d-dimensional vector of ones.

For $k \geqslant 1$, the k-dimensional path space $D([0,\infty),\mathbb{R}^k)$ is the set of functions $x:[0,\infty) \to \mathbb{R}^k$ that are right continuous on $[0,\infty)$ and have finite left limits on $(0,\infty)$. For a path $x \in D([0,\infty),\mathbb{R}^k)$, we sometimes use $x(\cdot)$ to denote the path. For a vector $a \in \mathbb{R}^k$ and a path $x \in D([0,\infty),\mathbb{R}^k)$, $x(a\cdot)$ is the path with $x(at) = (x_1(a_1t),\ldots,x_k(a_kt))'$. More generally, for an $h \in D([0,\infty),\mathbb{R}_+^k)$, $x(h(\cdot))$ is the path with $x(h(t)) = (x_1(h_1(t)),\ldots,x_d(h_d(t)))'$. A path $x \in D([0,\infty),\mathbb{R}^k)$ is nondecreasing if each component is. We use $x(s-)$ to denote the left limit at $s > 0$. The space $D([0,\infty),\mathbb{R}^k)$ is endowed with the Skorohod J_1-topology (see, e.g., Ethier and Kurtz [21]). For a sequence of paths $\{f^n\}$, for each $n \geqslant 1$ the paths $\bar{f}^n$ and $\tilde{f}^n$ are defined by

$$\bar{f}^n(\cdot) = \frac{1}{n} f^n(n\cdot) \quad \text{and} \quad \tilde{f}^n(\cdot) = \frac{1}{\sqrt{n}} f^n(n\cdot).$$

The sequence $\{f^n\}$ is said to converge to f uniformly on compact sets if for each $T > 0$

$$\sup_{0 \leqslant t \leqslant T} |f^n(t) - f(t)| \to 0$$

as $n \to \infty$. We denote such converge by $f^n \to f$ u.o.c.

In section 2, the queueing network model is introduced. The heavy traffic limit theorem is stated in section 3. The Skorohod problem is stated in section 4, where a general oscillation result is established. In section 5, we represent the queue length process as a solution to a Skorohod problem. In section 6 we prove a fluid limit theorem which will be used in the proof of the heavy traffic limit theorem. In section 7 we prove a stopping time property that is needed to prove a martingale property. The proof of the heavy traffic limit theorem is completed in section 8. Extensions will be discussed in section 9.

2. The queueing network model

The queueing network under consideration has d single server stations indexed by $i \in \boldsymbol{I} \equiv \{1,\ldots,d\}$. Customers visiting station i are *homogeneous* in terms of service time distribution and routing. We assume that routing is deterministic. That is, customers leaving station i all go next to station $\sigma(i) \in \boldsymbol{I}$ or leave the system. In the latter case we let $\sigma(i) = 0$. Because all customers leaving station i are deterministically

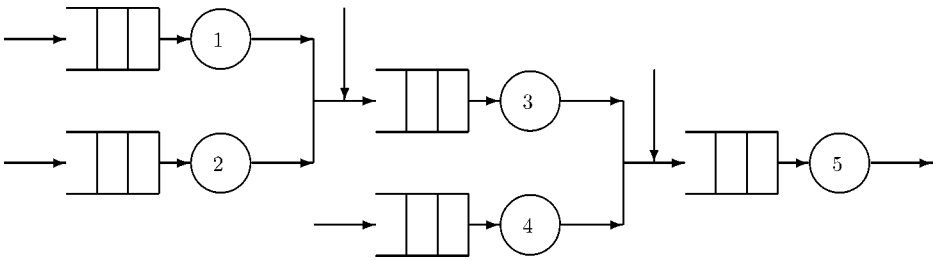

Figure 1. A five station network.

routed to a station, the routing must be feedforward. The network is sometimes called an *in-tree* network. This routing assumption is quite restrictive by conventional standards. An example of such a network is pictured in figure 1. (Other routing assumptions will be discussed in section 9.) We assume that the size b_i of the buffer associated with each station i is *finite*, $i \in \boldsymbol{I}$. Therefore, at each station i there are at most b_i customers, including the one possibly being served. We assume that the network is open. That is, all customers eventually leave the network.

Associated with each station i, there are two sequences of iid positive random variables $\{u_{ik},\ k \geqslant 1\}$ and $\{v_{ik},\ k \geqslant 1\}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that

$$\begin{aligned} \mathbb{E}(u_{i1}) &= 1, \quad \text{Var}(u_{i1}) = c_i^a < \infty, \quad i \in \boldsymbol{I}, \\ \mathbb{E}(v_{i1}) &= 1, \quad \text{Var}(v_{i1}) = c_i^s < \infty, \quad i \in \boldsymbol{I}. \end{aligned}$$

Also associated with each station i, there are two numbers: $\alpha_i \geqslant 0$ and $m_i > 0$. The iid random variables $\{v_{ik},\ k \geqslant 1\}$ are the *normalized* service times and the iid random variables $\{u_{ik}, k \geqslant 1\}$ are the *normalized* interarrival times. The actual service times for the kth customer at station i is $m_i v_{ik}$. If $\alpha_i = 0$, there are no external customer arrivals to station i. If $\alpha_i > 0$, the interarrival between the kth and the $(k-1)$th customer is u_{ik}/α_i. Although it is not necessary, for notational convenience, we assume that $\alpha_i > 0$ for each $i \in \boldsymbol{I}$.

An important feature in the network is that the sizes of buffers are *finite*. When the buffer at a downstream station $\sigma(i)$ is full, server i stops working although a customer may still occupy station i. This phenomenon is called the "block-and-hold-0" blocking; see Cheng and Yao [13] for a discussion of general blocking mechanisms. One can envision that when the kth customer enters service at station i, a service time clock (stopwatch) is set to $m_i v_{ik}$. The service is completed when the clock reading reaches zero. During the service period, the clock is turned off or on depending on whether the server is blocked or not. Our blocking mechanism applies to arrivals too. Upon the kth external arrival to station i, an arrival clock at station i is set to $u_{i,k+1}/\alpha_i$. When the clock reading reaches zero, the $k+1$ customer arrives at station i. During this interarrival period, the arrival clock is turned off or on depending on whether buffer i is full or not.

We admit that our blocking mechanism for external arrivals is restrictive for some applications. However, in many manufacturing applications, external arrivals can be controlled. Our blocking mechanism represents one way of modeling arrival processes. In section 9, we will discuss other blocking mechanisms, including loss networks. In heavy traffic analysis, the blocking in our network introduces complications that do not exist in networks with infinite buffers.

For $i \in \boldsymbol{I}$, let $Z_i(t)$ be the number of customers at station i at time t, including possibly the one being served. Note that $Z_i(0)$ is the initial number of customers at station i at time 0. It represents part of an initial network configuration. Let $Y_i(t)$ be the amount of time that server i has been idle while server i is not blocked in time interval $[0, t]$, and let $Y_{i+d}(t)$ be the amount of time that buffer i has been full in time interval $[0, t]$. That is,

$$Y_i(t) = \int_0^t 1_{\{Z_i(s)=0, Z_{\sigma(i)}(s)<b_{\sigma(i)}\}} \, \mathrm{d}s, \qquad Y_{i+d}(t) = \int_0^t 1_{\{Z_i(s)=b_i\}} \, \mathrm{d}s. \tag{2.1}$$

Hereafter, whenever $\sigma(i) = 0$ condition $\{a_{\sigma(i)} < b_{\sigma(i)}\}$ always holds for any $a, b \in \mathbb{R}^d$. Let $Z(t) = (Z_1(t), \ldots, Z_d(t))'$ and $Y(t) = (Y_1(t), \ldots, Y_{2d})'$. The process $Z = \{Z(t),\ t \geqslant 0\}$ is called the queue length process and the process $Y = \{Y(t),\ t \geqslant 0\}$ is called the allocation process. Clearly, Y is a nondecreasing, continuous process. Given the iid interarrival time sequences and service time sequences, one can uniquely construct the queue length process and the allocation process. Such detailed construction, though not attempted here, is implicitly assumed in section 7.

For each $i \in \boldsymbol{I}$ and $t \geqslant 0$, let

$$F_i(t) = t - Y_{i+d}(t), \qquad B_i(t) = t - Y_i(t) - Y_{\sigma(i)+d}(t). \tag{2.2}$$

Hereafter, whenever $\sigma(i) = 0$, $Y_{\sigma(i)+d}(t)$ is understood to be 0.

It is clear that $B_i(t)$ is the cumulative amount of time that server i has been busy in $[0, t]$ and $F_i(t)$ is the cumulative amount of time that buffer i has not been full in $[0, t]$. That is,

$$F_i(t) = \int_0^t 1_{\{Z_i(s)<b_i\}} \, \mathrm{d}s, \qquad B_i(t) = \int_0^t 1_{\{Z_i(s)>0, Z_{\sigma(i)}(s)<b_{\sigma(i)}\}} \, \mathrm{d}s.$$

3. A heavy traffic limit theorem

To state a heavy traffic limit theorem, we need to consider a sequence of networks indexed by n. The network depends on the index n through the external arrival rates α^n, mean service times m^n and buffer sizes b^n, where

$$\alpha^n = \left(\alpha_1^n, \ldots, \alpha_d^n\right)', \qquad m^n = \left(m_1^n, \ldots, m_d^n\right)', \qquad b^n = \left(b_1^n, \ldots, b_d^n\right)'.$$

We let $\mu_i^n = 1/m_i^n$ be the mean service rate at station i. The normalized interarrival and service times, and the routing do not depend on n. Let $Z^n = \left\{Z^n(t),\ t \geqslant 0\right\}$ be the queue length process and $Y^n = \left\{Y^n(t),\ t \geqslant 0\right\}$ be the allocation process

associated with the nth network. In the following theorem, P is the $d \times d$ routing matrix, i.e., $P_{ij} = 1$ if station i customers go next to station j and $P_{ij} = 0$, otherwise.

Theorem 3.1. Assume that as $n \to \infty$,

$$\alpha^n \to \alpha > 0 \quad \text{and} \quad m^n \to m > 0, \tag{3.1}$$

$$\frac{b^n}{\sqrt{n}} \to b > 0, \tag{3.2}$$

$$\sqrt{n}\left(\alpha^n - \left(I - P'\right)\mu^n\right) \to \theta. \tag{3.3}$$

Assume that for each n, $Z^n(0)$ is defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $Z^n(0)$ is independent of the interarrival and service time sequences such that

$$\frac{1}{\sqrt{n}} Z^n(0) \Longrightarrow \xi, \quad n \to \infty. \tag{3.4}$$

Assume further that

$$\Gamma = \operatorname{diag}\left(\alpha_1 c_1^a, \ldots, \alpha_d c_d^a\right) + \left(I - P'\right) \operatorname{diag}\left(\mu_1 c_1^s, \ldots, \mu_d c_d^s\right)(I - P) \tag{3.5}$$

is (strictly) positive definite. Then

$$\left(\frac{1}{\sqrt{n}} Z^n(n\cdot), \frac{1}{\sqrt{n}} Y^n(n\cdot)\right) \Longrightarrow \left(Z^*(\cdot), Y^*(\cdot)\right), \quad \text{as } n \to \infty, \tag{3.6}$$

where Z^*, together with Y^*, is a semimartingale reflecting Brownian motion (RBM) defined on a filtered probability space $(\Omega^*, \{\mathcal{F}_t^*\}, \mathcal{F}^*, \mathbb{P}^*)$. The process (Z^*, Y^*) is uniquely determined in distribution from the following equations:

$$\mathbb{P}^*\text{-a.s.}, \quad Z^*(t) = Z^*(0) + X^*(t) + RY^*(t) \quad \text{for all } t \geqslant 0, \tag{3.7}$$

$$\mathbb{P}^*\text{-a.s.}, \quad 0 \leqslant Z^*(t) \leqslant b \quad \text{for all } t \geqslant 0, \tag{3.8}$$

$$Z^*(0) \text{ has the same distribution as } \xi, \tag{3.9}$$

$$Z^*(\cdot) \text{ and } Y^*(\cdot) \text{ are} \{\mathcal{F}_t^*\}\text{-adapted}, \tag{3.10}$$

$$\mathbb{P}^*\text{-a.s.}, \quad Y^*(0) = 0,\ Y^*(\cdot) \text{ is continuous and nondecreasing}, \tag{3.11}$$

$$\mathbb{P}^*\text{-a.s.}, \quad \text{for } i \in \boldsymbol{I},\ Y_i^*(\cdot) \text{ increases only at times } t \text{ when } Z_i^*(t) = 0, \tag{3.12}$$

$$\mathbb{P}^*\text{-a.s.}, \quad \text{for } i \in \boldsymbol{I},\ Y_{i+d}^*(\cdot) \text{ increases only at times } t \text{ when } Z_i^*(t) = b_i, \tag{3.13}$$

$$X^* \text{ is a Brownian motion with drift } \theta \text{ and covariance matrix } \Gamma, \tag{3.14}$$

$$\left\{X^*(t) - \theta t\right\} \text{ is an } \left\{\mathcal{F}_t^*\right\}\text{-martingale}, \tag{3.15}$$

where θ is defined in (3.3), Γ is defined in (3.5) and

$$R = \left(\left(I - P'\right) \operatorname{diag}(\mu),\ \left[\left(I - P'\right) \operatorname{diag}(\mu)\right]_\sigma - \operatorname{diag}(\alpha)\right). \tag{3.16}$$

For a $d \times d$ matrix A and a vector $x \in \mathbb{R}^d$,

$$x_\sigma = (x_{\sigma(1)}, \ldots, x_{\sigma(d)})' \quad \text{and} \quad A_\sigma x = A x_\sigma. \tag{3.17}$$

The theorem will be proved in section 8. The vector θ, the matrix Γ and the $d \times 2d$ matrix R are called the drift, the covariance matrix and the reflection matrix of the RBM Z^*, respectively. For $i \in \boldsymbol{I}$, the ith column of R is the direction of reflection used when $Z_i^*(t) = 0$, and the $(i+d)$th column of R is the direction of reflection used when $Z_i^*(t) = b_i$. Because of (3.8), the RBM Z^* lives in the d-dimensional box $\boldsymbol{S}$ defined by

$$\boldsymbol{S} \equiv \big\{x = (x_1, \ldots, x_d)' \in \mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i \text{ for } i \in \boldsymbol{I}\big\}. \tag{3.18}$$

Therefore, the RBM Z^* in the theorem has state space $\boldsymbol{S}$. From now on, we call the RBM Z^* a $(\Gamma, \theta, R, \boldsymbol{S})$-RBM. The process Y^* is the pushing processes associated with the RBM Z^*. In the stochastic differential equation terminology, the process (Z^*, Y^*) is a *weak* solution to (3.7)–(3.14). Because the corresponding Skorohod problem may not have a unique solution (see the example at the end of section 5), it is not known whether a (strong) solution exists for *each* Brownian motion X^* defined on a probability space. The uniqueness of (Z^*, Y^*) (in distribution) follows from Dai and Williams [18] that generalized an earlier result of Taylor and Williams [37] for RBM's in an orthant.

4. The Skorohod problem and an oscillation theorem

In this section, we define the Skorohod problem and establish an oscillation result for solutions to a family of Skorohod problems. We choose to prove our results in a general polyhedral state space $\boldsymbol{S}$, instead of the d-dimensional box introduced in (3.18). We believe our oscillation result in a general state space is of independent interest.

In this section we follow most of the notation introduced in section 1 of Dai and Williams [18]. Symbols m and F are reused in this section. In the subsequent sections, they retain the original meaning. The polyhedron is defined in terms of m ($m \geqslant 1$) d-dimensional unit vectors $\{n_i,\ i \in \boldsymbol{J}\}$, $\boldsymbol{J} \equiv \{1, \ldots, m\}$, and an m-dimensional vector $a = (a_1, \ldots, a_m)'$. The state space $\boldsymbol{S}$ is defined by

$$\boldsymbol{S} \equiv \big\{x \in \mathbb{R}^d\colon\ n_i \cdot x \geqslant a_i \text{ for all } i \in \boldsymbol{J}\big\}, \tag{4.1}$$

where $n_i \cdot x = n_i' x$ denotes the inner product of the vectors n_i and x. It is assumed that the interior of $\boldsymbol{S}$ is non-empty and that the set $\{(n_1, a_1), \ldots, (n_m, a_m)\}$ is minimal in the sense that no proper subset defines $\boldsymbol{S}$. That is, for any strict subset $\boldsymbol{K} \subset \boldsymbol{J}$, the set $\{x \in \mathbb{R}^d\colon\ n_i \cdot x \geqslant a_i\ \forall i \in \boldsymbol{K}\}$ is strictly larger than $\boldsymbol{S}$. This is equivalent to the assumption that each of the faces

$$F_i \equiv \{x \in \boldsymbol{S}\colon\ n_i \cdot x = a_i\}, \quad i \in \boldsymbol{J}, \tag{4.2}$$

has dimension $d-1$ (cf. [8, theorem 8.2]). As a consequence, n_i is the unit normal to F_i that points into the interior of $\boldsymbol{S}$. Let N denote the $m \times d$ matrix whose ith row is given by the row vector n_i' for each $i \in \boldsymbol{J}$.

For each face F_i, $i \in \boldsymbol{J}$, we associate a d-dimensional vector v_i with it. We use R to denote the $d \times m$ matrix whose ith column v_i. Let us first define the Skorohod problem associated with the data $(\boldsymbol{S}, R)$. The matrix R is called the reflection matrix.

In the following, for a Borel set $U \subset \mathbb{R}^k$, $k \geqslant 1$, we define $D([0,T],U) = \{w : [0,T] \to U,\ w \text{ is right continuous in } [0,T) \text{ having left limits in } (0,T]\}$.

Definition 4.1 (The Skorohod problem). Given $T > 0$ and $x \in D([0,T],\mathbb{R}^d)$ with $x(0) \in \boldsymbol{S}$, an $(\boldsymbol{S}, R)$-regulation of x over $[0,T]$ is a pair $(z,y) \in D([0,T], \boldsymbol{S}) \times D([0,T], \mathbb{R}^m_+)$ such that

(i) $z(t) = x(t) + Ry(t)$ for all $t \in [0,T]$,

(ii) $z(t) \in \boldsymbol{S}$ for all $t \in [0,T]$,

(iii) for each $i \in \boldsymbol{J}$,

(a) $y_i(0) = 0$,

(b) y_i is nondecreasing,

(c) $\int_{(0,T]} (n_i \cdot z(t) - a_i)\, \mathrm{d}y_i(t) = 0$.

Remarks. (a) Although in the rest of this paper, y is known to be continuous, we allow y to have jumps in the definition of the Skorohod problem.

(b) The integral $\int_{(0,T]} (n_i \cdot z(t) - a_i)\, \mathrm{d}y_i(t)$ is well defined as a Lebesgue–Stieltjes integral, because any path $z \in D([0,T],\mathbb{R}^d)$ is bounded in $[0,T]$. Loosely speaking, condition (iii)(c) says that y_i can increase only at times $t \in [0,T]$ for which $z(t) \in F_i$. (See lemma 4.4 for a more precise statement.)

The existence and uniqueness of an $(\boldsymbol{S}, R)$-regulation heavily depends on the reflection matrix R.

Definition 4.2. For each $\emptyset \neq \boldsymbol{K} \subset \boldsymbol{J}$, define $F_{\boldsymbol{K}} = \bigcap_{i \in \boldsymbol{K}} F_i$. Let $F_\emptyset = \boldsymbol{S}$. A set $\boldsymbol{K} \subset \boldsymbol{J}$ is *maximal* if $\boldsymbol{K} \neq \emptyset$, $F_{\boldsymbol{K}} \neq \emptyset$, and $F_{\boldsymbol{K}} \neq F_{\widetilde{\boldsymbol{K}}}$ for any $\widetilde{\boldsymbol{K}} \supset \boldsymbol{K}$ such that $\widetilde{\boldsymbol{K}} \neq \boldsymbol{K}$.

Now we introduce an assumption on N and R.

Completely-$\mathcal{S}$ assumption. For each maximal $\boldsymbol{K} \subset \boldsymbol{J}$,

(S.a) there is a positive linear combination $v = \sum_{i \in \boldsymbol{K}} c_i v_i$ ($c_i > 0\ \forall i \in \boldsymbol{K}$) of the $\{v_i,\ i \in \boldsymbol{K}\}$ such that $n_i \cdot v > 0$ for all $i \in \boldsymbol{K}$;

(S.b) there is a positive linear combination $\eta = \sum_{i \in \boldsymbol{K}} c_i n_i$ ($c_i > 0\ \forall i \in \boldsymbol{K}$) of the $\{n_i,\ i \in \boldsymbol{K}\}$ such that $\eta \cdot v_i > 0$ for all $i \in \boldsymbol{K}$.

The labels (S.a) and (S.b) stand for $\mathcal{S}$-condition (a) and (b), respectively. The origin of these labels becomes apparent when the conditions are written in matrix form

as below. For a vector $x \in \mathbb{R}^k$, the notation $x > 0$ indicates that all coordinates of x are strictly positive, and the notation $x \geqslant 0$ indicates that all coordinates of x are nonnegative.

Definition 4.3. A matrix A is called an $\mathcal{S}$-matrix if there is a vector $x \geqslant 0$ such that $Ax > 0$.

For an $m \times m$ matrix A and $\boldsymbol{K} \subset \boldsymbol{J}$, let $A_{\boldsymbol{K}}$ denote the $|\boldsymbol{K}| \times |\boldsymbol{K}|$ matrix obtained from A by deleting those rows and columns with indices in $\boldsymbol{J} \backslash \boldsymbol{K}$.

Conditions (S.a) and (S.b) are equivalent to the following:

(S.a) the matrix $(NR)_{\boldsymbol{K}}$ is an $\mathcal{S}$ matrix;

(S.b) the matrix $(NR)'_{\boldsymbol{K}}$ is an $\mathcal{S}$ matrix.

Definition 4.4. The convex polyhedron $\boldsymbol{S}$ is *simple* if for each $\boldsymbol{K} \subset \boldsymbol{J}$ such that $\boldsymbol{K} \neq \emptyset$ and $F_{\boldsymbol{K}} \neq \emptyset$, exactly $|\boldsymbol{K}|$ distinct faces contain $F_{\boldsymbol{K}}$.

The convex polyhedron $\boldsymbol{S}$ is *simple* if and only if for each $\boldsymbol{K} \subset \boldsymbol{J}$, $F_{\boldsymbol{K}} \neq \emptyset$ implies that $\boldsymbol{K}$ is maximal. One can check that the d-dimensional box in (3.18) is a simple polyhedron. The following proposition was proved in Dai and Williams [18, proposition 1.1]. It is a straightforward generalization of Reiman and Williams [36, lemma 3].

Proposition 1. Suppose that $\boldsymbol{S}$ is simple. Then (S.a) holds for all maximal $\boldsymbol{K} \subset \boldsymbol{J}$ if and only if (S.b) holds for all maximal $\boldsymbol{K} \subset \boldsymbol{J}$.

The following oscillation result is concerned with paths in a family of $(\boldsymbol{S}^r, R^r)$-regulations indexed by $r > 0$. In the case that $\boldsymbol{S} = \mathbb{R}^d_+$, and all paths are continuous and from a single $(\boldsymbol{S}, R)$-regulation, this result was proved previously by Bernard and El Kharroubi [2]. Dai and Williams [18] generalized the result to a general polyhedral state space. Our proof here is adapted from [18].

For any $f \in D([t_1, t_2], \mathbb{R}^k)$ with some $k \geqslant 1$, let

$$\begin{aligned}
\text{Osc}\big(f, [t_1, t_2]\big) &= \sup_{t_1 \leqslant s \leqslant t \leqslant t_2} \big|f(t) - f(s)\big|, \\
\text{Osc}\big(f, [t_1, t_2)\big) &= \sup_{t_1 \leqslant s \leqslant t < t_2} \big|f(t) - f(s)\big|, \\
||\Delta f||_{(t_1, t_2]} &= \sup_{t_1 < s \leqslant t_2} \big|\Delta f(s)\big|,
\end{aligned}$$

where, as before, $\Delta f(s) = f(s) - f(s-)$ and $f(s-)$ is the left limit at s. Note that when f is left continuous at t_2, $\text{Osc}(f, [t_1, t_2]) = \text{Osc}(f, [t_1, t_2))$.

We consider a sequence of state spaces $\boldsymbol{S}^r$ indexed by $r > 0$. The shape of the space state does not change with r. That is, the normal vectors $\{n_i,\ i \in \boldsymbol{J}\}$ do not depend on r. However, the size $(a_1^r, \ldots, a_m^r)'$ of the state space depends on r. Hence,

$$\boldsymbol{S}^r \equiv \big\{x \in \mathbb{R}^d\colon\ n_i \cdot x \geqslant a_i^r \text{ for all } i \in \boldsymbol{J}\big\}.$$

The reflection matrix associated with each state space $\boldsymbol{S}^r$ is R^r, whose ith column is denoted by v_i^r. Recall that N is a matrix whose ith row is given by n_i'.

Theorem 4.2. Assume that $R^r \to R$ as $r \to \infty$ and (N, R) satisfies the *Completely-$\mathcal{S}$ assumption*. There exist constants $\kappa > 0$ and $\hat{r} > 0$ that depend only on (N, R) such that for any $T > 0,\ r \geqslant \hat{r},\ x \in D([0,T], \mathbb{R}^d)$ with $x(0) \in \boldsymbol{S}^r$, and an $(\boldsymbol{S}^r, R^r)$-regulation (y, z) of x over $[0, T]$, the following holds for each interval $[t_1, t_2] \subset [0, T]$:

$$\begin{aligned}
\mathrm{Osc}\big(y, [t_1, t_2]\big) &\leqslant \kappa\big(\mathrm{Osc}\big(x, [t_1, t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big),\\
\mathrm{Osc}\big(z, [t_1, t_2]\big) &\leqslant \kappa\big(\mathrm{Osc}\big(x, [t_1, t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big).
\end{aligned}$$

We leave the lengthy proof to the end of this section. To prepare for the proof, we need a few lemmas.

Lemma 4.3. Let $f \in D([0,\infty), \mathbb{R})$. Suppose f is of bounded variation on each finite time interval, and assume that $f(0) = 0$. Then for each $t \geqslant 0$:

$$f^2(t) + \sum_{0<s\leqslant t} \big[\Delta f(s)\big]^2 = 2\int_{(0,t]} f(s)\,\mathrm{d}f(s).$$

Proof. The result is quite standard. See, for example, Last and Brandt [30, theorem A.4.6]. □

Let $g \in D([0,\infty), \mathbb{R})$ be a nondecreasing function. The function g is said to increase at time $t > 0$ if there exists a $\delta > 0$ such that $g(u) < g(v)$ for each $t - \delta < u < t < v < t + \delta$. The following lemma should also be standard. For completeness, we provide a direct proof.

Lemma 4.4. Let $g \in D([0,\infty, \mathbb{R})$ be a nondecreasing function and $f \in D([0,\infty), \mathbb{R})$ be a nonnegative function. For $t > 0$, if $\int_{(0,t]} f(s)\,\mathrm{d}g(s) = 0$ and $f(s) > 0$ for $s \in [0, t)$, then $g(s) = g(0)$ for $s \in [0, t)$.

Proof. Suppose that there is an $s \in (0, t)$ such that $g(s) > g(0)$. If g has jump at a point $t' \in (0, t)$, then

$$\int_{(0,t]} f(s)\,\mathrm{d}g(s) \geqslant f\big(t'\big)\Delta g\big(t'\big) > 0,$$

contradicting the fact that $\int_{(0,t]} f(s)\,dg(s) = 0$. Thus g must be continuous on $(0,t)$. Let

$$t' = \inf\big\{s \in (0,t):\ g(s) > g(0)\big\}.$$

By the continuity of g, $g(t') = g(0)$. By the definition of t', for any $s > t'$, $g(s) > g(t')$. Because $f(t') > 0$ and f is right continuous, there is a $\delta > 0$ such that $\inf_{t' \leqslant s \leqslant t'+\delta} f(s) > 0$. Now,

$$\int_{(0,t]} f(s)\,\mathrm{d}g(s) \geqslant \int_{(t',t'+\delta]} f(s)\,\mathrm{d}g(s) \geqslant \inf_{t' \leqslant s \leqslant t'+\delta} f(s)\big(g\big(t'+\delta\big) - g\big(t'\big)\big) > 0,$$

contradicting the fact that $\int_{(0,t]} f(s)\,\mathrm{d}g(s) = 0$. Therefore, $g(s) = g(0)$ for $0 \leqslant s < t$. □

Lemma 4.5. Let $\boldsymbol{S} = [0,\infty)$ and $R = 1$. Then the $(\boldsymbol{S}, R)$-regulation of x with $x(0) \geqslant 0$ has a unique solution (z,y) given by

$$\begin{aligned} y(t) &= \sup_{0 \leqslant s \leqslant t} x^-(s) \quad \text{for } 0 \leqslant t \leqslant T,\\ z(t) &= x(t) + y(t), \end{aligned}$$

where $x^-(t) = \max\{-x(t), 0\}$.

Proof. We first show the uniqueness. Suppose there are two solutions (z,y) and $(\hat{z},\hat{y})$ to the $(\boldsymbol{S}, R)$-regulation of x. Then $z - \hat{z} = y - \hat{y}$. Now let $f = y - \hat{y}$. By lemma 4.3, we have for each $t \geqslant 0$

$$\begin{aligned} 0 \leqslant f^2(t) + \sum_{0<s\leqslant t} \big[\Delta f(s)\big]^2 &= 2\int_{(0,t]} f(s)\,\mathrm{d}f(s)\\ &= 2\int_{(0,t]} \big(z(s) - \hat{z}(s)\big)\,\mathrm{d}\big(y(s) - \hat{y}(s)\big)\\ &= -2\int_{(0,t]} \hat{z}(s)\,\mathrm{d}y(s) - 2\int_{(0,t]} z(s)\,\mathrm{d}\hat{y}(s) \leqslant 0. \end{aligned}$$

Hence, $f(t) = 0$, thus proving uniqueness.

For existence, let $y(t) = \sup_{0\leqslant s\leqslant t} x^-(s)$. Since $x(0) \geqslant 0$, $x^-(0) = 0$ and so $y(0) = 0$. Clearly,

$$z(t) \equiv x(t) + y(t) \geqslant x(t) + x^-(t) \geqslant 0 \quad \text{for all } t \geqslant 0,$$

y is nondecreasing, and, hence, it has left limits on $(0,T]$. Since $x(\cdot)$ is right continuous, y is right continuous. It remains to be verified that y satisfies property (iii)(c) in the definition of the Skorohod problem. Suppose y has a jump at time t. Because

$$y\big(t^-\big) = \sup_{0\leqslant s<t} x^-(s) \quad \text{and} \quad y(t) = \max\big\{y\big(t^-\big), x^-(t)\big\} > y\big(t^-\big),$$

we have $y(t) = x^-(t) = -x(t)$. Thus, $z(t) = x(t) + y(t) = x(t) + x^-(t) = 0$. Therefore, without loss of generality, we assume that y is continuous. If y increases

at time t, it follows from the proof of lemma 8.1 in Chung and Williams [14] that $z(t) = 0$. Therefore, by Graves [22, p. 269],

$$\int_0^t z(s)\,\mathrm{d}y(s) = \lim_{n\to\infty} \sum_{k=1}^{2^n t} \Big(\inf_{s\in[(k-1)t/2^n,\, kt/2^n]} z(s)\Big)\left(y\left(\frac{kt}{2^n}\right) - y\left(\frac{(k-1)t}{2^n}\right)\right) = 0.$$

□

Let C be the constant determined in Dai and Williams [18, lemma B.1]. It depends on $\{n_i,\ i \in J\}$ only, not on $(a_1^r, \ldots, a_m^r)'$. For each $\varepsilon \geqslant 0$ and $\boldsymbol{K} \subset \boldsymbol{J}$ (including the empty set), define

$$\begin{aligned} F_{\boldsymbol{K}}^{r,\varepsilon} = \big\{ x \in \mathbb{R}^d\colon\ & 0 \leqslant n_i x - a_i^r \leqslant C_\varepsilon \text{ for all } i \in \boldsymbol{K} \\ & \text{and } n_i x - a_i^r > \varepsilon \text{ for all } i \in \boldsymbol{J}\backslash\boldsymbol{K} \big\}, \end{aligned} \tag{4.3}$$

where $C_\varepsilon = Cm\varepsilon$. The following lemma, which was proved in [18, lemma 4.1], plays a key role in the proof of the oscillation theorem.

Lemma 4.6. For each $\varepsilon \geqslant 0$,

$$\boldsymbol{S}^r = \bigcup_{\boldsymbol{K}\in\mathcal{C}} F_{\boldsymbol{K}}^{r,\varepsilon}, \tag{4.4}$$

where $\mathcal{C}$ denotes the collection of subsets of $\boldsymbol{J}$ consisting of all maximal sets in $\boldsymbol{J}$ together with the empty set.

Proof of theorem 4.2. Our proof is adapted from that of lemma 4.3 in Dai and Williams [18] who generalized lemma 1 of Bernard and El Kharroubi [2]. We proceed via an induction on the size of $\boldsymbol{J}$, the index set for the faces of $\boldsymbol{S}$. Throughout this proof, T, x, y, z, t_1, t_2 will be as in the statement of the theorem. In general, z and y depend on the index r, but we suppress the dependence in the proof.

First consider the case $|\boldsymbol{J}| = 1$. Then $R^r = v_1^r$ is a vector in $\mathbb{R}^d$ and $v_1^r \to v_1$ as $r \to \infty$. By (S.a), $n_1 \cdot v_1 > 0$. Take r_0 such that

$$n_1 \cdot v_1^r \geqslant \frac{1}{2}(n_1 \cdot v_1) \quad \text{and} \quad \frac{||v_1^r||}{(n_1 \cdot v_1^r)} \leqslant \frac{2||v_1||}{n_1 \cdot v_1}$$

for $r \geqslant r_0$. Fix $r \geqslant r_0$. In this case, y is uniquely given by the one-dimensional regulator mapping for $n_1 \cdot x - a_1^r$ in lemma 4.5:

$$y(t) = \Big(-\min_{0\leqslant s\leqslant t} \big(n_1 \cdot x - a_1^r\big)(s)\Big)^+ \big/ \big(n_1 \cdot v_1^r\big) \quad \text{for all } t \in [0, T]. \tag{4.5}$$

Together with

$$n_1 \cdot z(t) = n_1 \cdot x(t) + n_1 \cdot v_1^r y(t) \quad \text{for all } t \in [0, T],$$

this defines a $([a_1^r,\infty), n_1\cdot v_1^r)$-regulation of $n_1\cdot x$ over $[0,T]$. The oscillation estimates in the theorem then follow easily from (4.5) and the fact that $z = x + v_1^r y$. That is, for $r \geqslant r_0$,

$$\begin{aligned}
&\mathrm{Osc}\big(y,[t_1,t_2]\big) \leqslant \frac{1}{n_1\cdot v_1^r}\mathrm{Osc}\big(x,[t_1,t_2]\big) \leqslant \frac{2}{n_1\cdot v_1}\mathrm{Osc}\big(x,[t_1,t_2]\big),\\
&\mathrm{Osc}\big(z,[t_1,t_2]\big) \leqslant 1+\frac{||v_1^r||}{(n_1\cdot v_1^r)}\mathrm{Osc}\big(x,[t_1,t_2]\big) \leqslant 1+\frac{2||v_1||}{(n_1\cdot v_1)}\mathrm{Osc}\big(x,[t_1,t_2]\big).
\end{aligned}$$

Thus the theorem holds for $|\boldsymbol{J}| = 1$ with $\hat{r} = r_0$ and

$$\kappa = \max\left\{\left(1+\frac{2||v_1||}{n_1\cdot v_1}\right),\ \frac{2}{n_1\cdot v_1}\right\}.$$

For the induction step, suppose that the theorem is true for $1 \leqslant |\boldsymbol{J}| < m$. Now consider a state space $\boldsymbol{S}$ with $|\boldsymbol{J}| = m$. Our proof of the induction step is separated into several parts.

Part (a). We claim that there exists a constant C_1 that depends only on (N,R) and a constant $r_0 > 0$ such that for $r \geqslant r_0$ and each $\boldsymbol{K} \in \mathcal{C}\backslash\{\boldsymbol{J}\}$ (see lemma 4.6 for the definition of $\mathcal{C}$), if $y_{\boldsymbol{J}\backslash\boldsymbol{K}}$ does not increase on $[t_1,t_2)$, then one has:

$$\mathrm{Osc}\big(y,[t_1,t_2]\big) \leqslant C_1\big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big), \tag{4.6}$$

$$\mathrm{Osc}\big(z,[t_1,t_2]\big) \leqslant C_1\big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big). \tag{4.7}$$

To see this, note that under the assumptions of the claim, for $t \in [0, t_2 - t_1)$,

$$z(t+t_1) = z(t_1) + x(t+t_1) - x(t_1) + \sum_{i\in\boldsymbol{K}} v_i^r\big(y_i(t+t_1) - y_i(t_1)\big). \tag{4.8}$$

For any t_2' such that $t_1 \leqslant t_2' < t_2$, it follows that $(z(\cdot+t_1), y_{\boldsymbol{K}}(\cdot+t_1) - y_{\boldsymbol{K}}(t_1))$ is an $(\boldsymbol{S}_{\boldsymbol{K}}^r, R_{\boldsymbol{K}}^r)$-regulation of $z(t_1) + x(\cdot+t_1) - x(t_1)$ over $[0, t_2' - t_1]$. If $\boldsymbol{K} = \emptyset$, then y does not increase on $[t_1, t_2']$ and the oscillation estimate trivially holds with $C_1 = 1$. If $\boldsymbol{K} \neq \emptyset$, then $\boldsymbol{K}$ is maximal and so by Dai and Williams [18, lemma 4.2], (S.a) and (S.b) hold for $(N_{\boldsymbol{K}}, R_{\boldsymbol{K}})$. Then, by the induction assumption, since $|\boldsymbol{K}| < m$, we have that there exist constants $C_{\boldsymbol{K}} \geqslant 1$ and $r_{0,\boldsymbol{K}} > 0$ that depend only on $(N_{\boldsymbol{K}}, R_{\boldsymbol{K}})$, such that for $r \geqslant r_{0,\boldsymbol{K}}$

$$\begin{aligned}
\mathrm{Osc}\big(y,[t_1,t_2']\big) &= \mathrm{Osc}\big(y_{\boldsymbol{K}}(\cdot+t_1),[0,t_2'-t_1]\big)\\
&\leqslant C_{\boldsymbol{K}}\big(\mathrm{Osc}\big(x(\cdot+t_1)-x(t_1),\big[0,t_2'-t_1\big]\big) + \sup_{t_1<s\leqslant t_2'}\big|\Delta y_{\boldsymbol{K}}(s)\big|\big)\\
&\leqslant C_{\boldsymbol{K}}\big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + \sup_{t_1<s\leqslant t_2}\big|\Delta y(s)\big|\big).
\end{aligned}$$

Letting $t_2' \uparrow t_2$,

$$\mathrm{Osc}\big(y,[t_1,t_2)\big) \leqslant C_{\boldsymbol{K}}\big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + \sup_{t_1<s\leqslant t_2}|\Delta y(s)|\big).$$

Therefore,

$$\begin{aligned}\mathrm{Osc}\big(y,[t_1,t_2]\big) &\leqslant \mathrm{Osc}\big(y,[t_1,t_2)\big) + \big|\Delta y(t_2)\big| \\ &\leqslant 2C_{\boldsymbol{K}}\big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + \sup_{t_1<s\leqslant t_2}\big|\Delta y(s)\big|\big).\end{aligned}$$

It follows from $z(t) = x(t) + R^r y(t)$ that

$$\begin{aligned}\mathrm{Osc}\big(z,[t_1,t_2]\big) &\leqslant \mathrm{Osc}\big(x,[t_1,t_2]\big) + ||R^r||\mathrm{Osc}\big(y,[t_1,t_2]\big) \\ &\leqslant \big(1+\|R^r\|2C_{\boldsymbol{K}}\big)\mathrm{Osc}\big(x,[t_1,t_2]\big).\end{aligned}$$

Because $R^r \to R$ as $r \to \infty$, we can choose r_0 such that r_0 is at least the maximum of the $r_{0,\boldsymbol{K}}$'s for $\boldsymbol{K}$ running through $\mathcal{C}\backslash\{\boldsymbol{J}\}$ and $||R^r|| \leqslant ||R|| + 1$ for $r \geqslant r_0$. Let C_1 be the maximum of $1 + (||R|| + 1)2C_{\boldsymbol{K}}$ for $\boldsymbol{K}$ running through $\mathcal{C}\backslash\{\boldsymbol{J}\}$. Then inequalities (4.6) and (4.7) follow.

For parts (b) and (c), we let

$$\varepsilon = \big(\mathrm{Osc}\big(x,[t_1,t_2]\big) + ||\Delta y||_{(t_1,t_2]}\big).$$

Without loss of generality we assume that $\varepsilon > 0$. By lemma 4.6, $z(t_1) \in F_{\boldsymbol{K}}^{C_1\varepsilon}$ for some $\boldsymbol{K} \in \mathcal{C}$.

Part (b). Suppose that the $\boldsymbol{K}$ found above is not $\boldsymbol{J}$. Then, for all $i \in \boldsymbol{J}\backslash\boldsymbol{K}$,

$$d\big(z(t_1), F_i\big) \geqslant n_i \cdot z(t_1) - a_i^r > C_1\varepsilon,$$

where $d(x, F)$ is the distance from a point x to a set F. We claim that $n_i \cdot z(s) - a_i^r > 0$ for $s \in [t_1, t_2]$ and $i \in \boldsymbol{J}\backslash\boldsymbol{K}$. Assume, on the contrary, that there exist $i \in \boldsymbol{J}\backslash\boldsymbol{K}$ and $s \in [t_1, t_2]$ such that $n_i \cdot z(s) - a_i^r = 0$. Let

$$t_2' = \inf\big\{s \in [t_1, t_2]\colon\ n_i \cdot z(s) - a_i^r = 0\big\}.$$

By the right continuity of z, $n_i \cdot z(t_2') - a_i^r = 0$. From the definition of t_2', $n_i \cdot z(s) - a_i^r > 0$ for $s \in [t_1, t_2')$, and hence y_i does not increase on $[t_1, t_2')$ by lemma 4.4. By part (a), we have

$$\begin{aligned}n_i \cdot z\big(t_2'\big) - a_i^r &= n_i \cdot \big(z\big(t_2'\big) - z(t_1)\big) + n_i \cdot z(t_1) - a_i^r \\ &> -C_1\big(\mathrm{Osc}\big(x,\big[t_1,t_2'\big]\big) + \sup_{t_1<s\leqslant t_2'}\big|\Delta y(s)\big|\big) + C_1\varepsilon \geqslant 0\end{aligned}$$

contradicting $n_i \cdot z(t_2') - a_i^r = 0$. Thus, z does not reach F_i^r for any $i \in \boldsymbol{J}\backslash\boldsymbol{K}$ during the interval $[t_1, t_2]$ and therefore $y_{\boldsymbol{J}\backslash\boldsymbol{K}}$ does not increase on $[t_1, t_2]$.

Then part (a) implies that (4.6) holds in this case.

Part (c). Suppose that the $\boldsymbol{K}$ described before part (b) is equal to $\boldsymbol{J}$. Since $z(t_1) \in F_{\boldsymbol{J}}^{C_1\varepsilon}$, by [18, lemma B.1], $d(z(t_1), F_i) \leqslant C_2\varepsilon$, where $C_2 = C_1Cm$. Now one of the following two situations holds.

(i) For every $i \in \boldsymbol{J}$, $d(z(t), F_i) \leqslant 2C_2\,\varepsilon$ for all $t \in [t_1, t_2]$. Then for each $i \in \boldsymbol{J}$,

$$0 \leqslant n_i \cdot z(t) - a_i^r \leqslant d\big(z(t), F_i\big) \leqslant 2C_2\varepsilon \quad \text{for all } t \in [t_1, t_2], \tag{4.9}$$

and so

$$\mathrm{Osc}\big(n_i \cdot z, [t_1, t_2]\big) \leqslant 2C_2\varepsilon. \tag{4.10}$$

Now, since $\boldsymbol{K} = \boldsymbol{J}$ is maximal, there is an $x_0 \in F_{\boldsymbol{J}}$ and by (S.b) there exists a positive linear combination $\eta = \sum_{i\in\boldsymbol{J}} \gamma_i n_i$ ($\gamma_i > 0$ for all i) of the $\{n_i,\ i \in \boldsymbol{J}\}$ such that $\eta \cdot v_i > 0$ for all $i \in \boldsymbol{J}$. Then

$$\eta \cdot \big(z(t) - x_0\big) = \eta \cdot \big(x(t) - x_0\big) + \sum_{i\in\boldsymbol{J}} \big(\eta \cdot v_i^r\big) y_i(t) \quad \text{for all } t \in [0, T]. \tag{4.11}$$

Thus,

$$\begin{aligned}
&\min_{i\in\boldsymbol{J}} \big(\eta \cdot v_i^r\big) \mathrm{Osc}\big(y_1 + \cdots + y_m, [t_1, t_2]\big) \\
&\quad \leqslant \mathrm{Osc}\big(\eta \cdot z, [t_1, t_2]\big) + \mathrm{Osc}\big(\eta \cdot x, [t_1, t_2]\big) \\
&\quad \leqslant \sum_{i\in\boldsymbol{J}} \gamma_i \big(\mathrm{Osc}\big(n_i \cdot z, [t_1, t_2]\big) + \mathrm{Osc}\big(n_i \cdot x, [t_1, t_2]\big)\big).
\end{aligned} \tag{4.12}$$

Since

$$\min_{i\in J} \big(\eta \cdot v_i^r\big) \to \min_{i\in J} (\eta \cdot v_i) > 0$$

as $r \to \infty$, using (4.10) and $z = x + R^r y$, we see that one can choose a constant C_3 depending only on (N, R) and an $r_1 > r_0$ such that

$$\mathrm{Osc}\big(y, [t_1, t_2]\big) \leqslant C_3\varepsilon, \qquad \mathrm{Osc}\big(z, [t_1, t_2]\big) \leqslant C_3\varepsilon.$$

(ii) There is an $i \in \boldsymbol{J}$ and $t_3 \in [t_1, t_2]$ such that $d(z(t_3), F_i) > 2C_2\varepsilon$. Define

$$t_1' = \inf\big\{t > t_1\colon\ d\big(z(t), F_i\big) > 2C_2\varepsilon \text{ for some } i \in \boldsymbol{J}\big\}.$$

By the definition of t_1', for any $\delta > 0$, over $[t_1, t_1' - \delta]$ we have the situation in part (c)(i) above. That is,

$$\mathrm{Osc}\big(y, \big[t_1, t_1' - \delta\big]\big) \leqslant C_3\varepsilon, \qquad \mathrm{Osc}\big(z, [t_1, t_1' - \delta]\big) \leqslant C_3\varepsilon.$$

Letting $\delta \to 0^+$, we have

$$\mathrm{Osc}\big(y, \big[t_1, t_1'\big)\big) \leqslant C_3\varepsilon, \qquad \mathrm{Osc}\big(z, \big[t_1, t_1'\big)\big) \leqslant C_3\varepsilon.$$

Over $[t_1', t_2]$, by lemma 4.6, we have $z(t_1') \in F_{\boldsymbol{K}}^{C_1\varepsilon}$ for some $\boldsymbol{K} \in \mathcal{C}\backslash\{\boldsymbol{J}\}$, and then we have the situation in part (b). Thus,

$$\mathrm{Osc}\big(y, \big[t_1', t_2\big]\big) \leqslant C_1\varepsilon, \qquad \mathrm{Osc}\big(z, \big[t_1', t_2\big]\big) \leqslant C_1\varepsilon.$$

Therefore,

$$\mathrm{Osc}\big(y,[t_1,t_2]\big) \leqslant \mathrm{Osc}\big(y,\big[t_1,t_1'\big)\big) + \big|\Delta y\big(t_1'\big)\big| + \mathrm{Osc}\big(y,\big[t_1',t_2\big]\big) \leqslant (1+C_1+C_3)\varepsilon.$$

Hence, there is a constant C_4 depending only on $(N,\ R)$ such that

$$\mathrm{Osc}\big(y,[t_1,t_2]\big) \leqslant C_4\varepsilon, \qquad \mathrm{Osc}\big(z,[t_1,t_2]\big) \leqslant C_4\varepsilon.$$

Thus, the theorem holds for $\kappa = \max\{C_1, C_3, C_4\}$ and $\hat{r} = r_1$. □

5. Network dynamics and preliminaries

For each $t \geqslant 0$, $i \in \boldsymbol{I}$ and $j > 0$ let $U_i(0) = V_i(0) = 0$,

$$\begin{aligned} &U_i(j) = u_{i1} + \cdots + u_{ij}, \qquad V_i(j) = v_{i1} + \cdots + v_{ij},\\ &E_i(t) = \max\{k \geqslant 0:\ u_{i1} + \cdots + u_{ik} \leqslant t\},\\ &S_i(t) = \max\{k \geqslant 0:\ v_{i1} + \cdots + v_{ik} \leqslant t\}. \end{aligned}$$

Let

$$\widehat{E}_i(t) = E_i(t) - t \quad \text{and} \quad \widehat{S}_i(t) = S_i(t) - t \quad \text{for } i \in \boldsymbol{I}.$$

Let

$$\widehat{E}(t) = \big(\widehat{E}_1(t), \ldots, \widehat{E}_d(t)\big)' \quad \text{and} \quad \widehat{S}(t) = \big(\widehat{S}_1(t), \ldots, \widehat{S}_d(t)\big)'.$$

The two d-dimensional processes $\{\widehat{E}(t),\ t \geqslant 0\}$ and $\{\widehat{S}(t),\ t \geqslant 0\}$ contain all the randomness in the queueing network. It is known that they satisfy the Functional Strong Law of Large Numbers [23, lemma V.2.1]: $\mathbb{P}$-a.s, as $r \to \infty$,

$$\frac{1}{r}\widehat{E}(r\cdot) \to 0 \text{ u.o.c.}, \qquad \frac{1}{r}\widehat{S}(r\cdot) \to 0 \text{ u.o.c.} \tag{5.1}$$

and the Functional Central Limit Theorem [4, section 17]: as $r \to \infty$,

$$\left(\frac{1}{\sqrt{r}}\widehat{E}(r\cdot), \frac{1}{\sqrt{r}}\widehat{S}(r\cdot)\right) \Longrightarrow \big(E^*, S^*\big), \tag{5.2}$$

where E^* and S^* are two independent, d-dimensional Brownian motions with drift zero and covariance matrices $\mathrm{diag}(c_1^a, \ldots, c_d^a)$ and $\mathrm{diag}(c_1^s, \ldots, c_d^s)$, respectively.

Recall that we are considering a sequence of networks indexed by n. In particular, α_i^n and μ_i^n are the external arrival rate to station i and the service rate of server i for the nth network. Let

$$E_i^n(t) = E_i\big(\alpha_i^n t\big), \qquad S_i^n(t) = S_i\big(\mu_i^n t\big).$$

If server i has been busy all the time in $[0, t]$, $S_i^n(t)$ is the number of services completed by time t at station i. Similarly, if buffer i has never been full in $[0, t]$, $E_i^n(t)$ is the number of arrivals by time t to station i. Recall that $F_i^n(t)$ is the cumulative amount of time that buffer i is not full by time t. From our model assumption, $E_i^n(F_i^n(t))$

is the number of external arrivals to station i by time t in the nth network. Also, $B_i^n(t)$ is the cumulative amount of time that server i has been working by time t and $S_i^n(B_i^n(t))$ is the number of departures from station i by time t in the nth network. Now we can write down the main equation that governs the dynamics of the queue length process. Namely,

$$Z_i^n(t) = Z_i^n(0) + E_i^n\big(F_i^n(t)\big) + \sum_{j\in \boldsymbol{I},\sigma(j)=i} S_j^n\big(B_j^n(t)\big) - S_i^n\big(B_i^n(t)\big), \quad i \in \boldsymbol{I}, \tag{5.3}$$

where $Z_i^n(0)$ is the initial queue length at station i. Let

$$E^n\big(F^n(t)\big) = \big(E_1^n\big(F_1^n(t)\big), \ldots, E_d^n\big(F_d^n(t)\big)\big)'$$

and

$$S^n\big(B^n(t)\big) = \big(S_1^n\big(B_1^n(t)\big), \ldots, S_d^n\big(B_d^n(t)\big)\big)'.$$

Recall that the routing matrix is defined as

$$P_{ij} = \begin{cases} 1 & \text{if station } i \text{ customers go to station } i, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have the vector form of (5.3):

$$Z^n(t) = Z^n(0) + E^n\big(F^n(t)\big) - (I - P')S^n\big(B^n(t)\big). \tag{5.4}$$

Following Harrison [24], we introduce the centered processes

$$\widehat{E}^n(t) = \big(\widehat{E}_1^n(t), \ldots, \widehat{E}_d^n(t)\big)' \quad \text{and} \quad \widehat{S}^n(t) = \big(\widehat{S}_1^n(t), \ldots, \widehat{S}_d^n(t)\big)',$$

where

$$\widehat{E}_i^n(t) = E_i^n(t) - \alpha_i^n t = \widehat{E}_i\big(\alpha_i^n t\big) \quad \text{and} \quad \widehat{S}_i^n(t) = S_i^n(t) - \mu_i^n t = \widehat{S}_i\big(\mu_i^n t\big), \quad i \in I. \tag{5.5}$$

It follows from (5.4) that

$$\begin{aligned} Z^n(t) = Z^n(0) &+ \widehat{E}^n\big(F^n(t)\big) - \big(I - P'\big)\widehat{S}^n\big(B^n(t)\big) \\ &+ \operatorname{diag}\big(\alpha^n\big)F^n(t) - \big(I - P'\big)\operatorname{diag}\big(\mu^n\big)B^n(t). \end{aligned} \tag{5.6}$$

It follows from (5.6) and (2.2) that

$$Z^n(t) = Z^n(0) + X^n(t) + R^n Y^n(t), \tag{5.7}$$

where

$$X^n(t) = \widehat{E}^n\big(F^n(t)\big) - \big(I - P'\big)\widehat{S}^n\big(B^n(t)\big) + \big(\alpha^n - \big(I - P'\big)\mu^n\big)t, \tag{5.8}$$

R^n is the $d \times 2d$ matrix given by

$$R^n = \big(\big(I - P'\big)\operatorname{diag}\big(\mu^n\big), \big[\big(I - P'\big)\operatorname{diag}\big(\mu^n\big)\big]_\sigma - \operatorname{diag}\big(\alpha^n\big)\big)$$

and for a matrix A, A_σ is defined in (3.17). Let $\boldsymbol{S}^n$ be the d-dimensional box defined by

$$\boldsymbol{S}^n = \{x \in \mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i^n\ \forall i \in \boldsymbol{I}\}.$$

One can check that for each sample path:

(i) $Z^n(t) = Z^n(0) + X^n(t) + R^n Y^n(t)$ for all $t \geqslant 0$,

(ii) $Z^n(t) \in \boldsymbol{S}^n$ for all $t \geqslant 0$,

(iii) for each $i = 1, \ldots, 2d$,

(a) $Y_i^n(0) = 0$,

(b) Y_i^n is nondecreasing and continuous,

(c) for $i = 1, \ldots, d$, Y_i increases only when $Z_i^n(t) = 0$ and for $i = d+1, \ldots, 2d$, Y_i^n increases only when $Z_i^n(t) = b_i^n$.

It follows that for each sample path, the pair $(Z^n(\cdot), Y^n(\cdot))$ is an $(\boldsymbol{S}^n, R^n)$-regulation of $Z^n(0) + X^n(\cdot)$.

Using the notion in section 4, for each boundary face of $\boldsymbol{S}^n$, there is a unit vector n_i that is normal to the face. (We number faces such that the ith face is $\{x \in S^n\colon\ x_i = 0\}$ for $i = 1, \ldots, d$ and $\{x \in S^n\colon\ x_i = b_i^n\}$ for $i = d+1, \ldots, 2d$.) Recall that N is a $2d \times d$ matrix whose ith row is the row vector n_i'. It is easy to check that

$$N = (I, -I)',$$

where I is the $d \times d$ identity matrix. Under the assumptions (3.1), as $n \to \infty$, $R^n \to R$ as defined in (3.16).

Lemma 5.1. The completely-$\mathcal{S}$ assumption in theorem 4.2 holds for (N, R).

Proof. Because the state space S^n is simple, by proposition 1, it is enough to show that for each maximal $\boldsymbol{K} \subset \boldsymbol{J} \equiv \{1, \ldots, 2d\}$, $(NR)_{\boldsymbol{K}}$ is an $\mathcal{S}$-matrix. Let R_0 and R_b be two $d \times d$ submatrices of R such that $R = (R_0, R_b)$. It is easy to check that

$$NR = \begin{pmatrix} R_0 & R_b \\ -R_0 & -R_b \end{pmatrix}.$$

A $\boldsymbol{K} \subset \boldsymbol{J}$ is maximal if $\bigcap_{i \in \boldsymbol{K}} F_i^n$ is non-empty. Because F_i^n and F_{i+d}^n are parallel to each other, a non-empty $\boldsymbol{K}$ is maximal if and only if for each $i \in \boldsymbol{K}$, $i + d \notin \boldsymbol{K}$. Let $M = (NR)_{\boldsymbol{K}}$. Then M has the following form:

$$M = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} = \begin{pmatrix} M_1 & 0 \\ 0 & M_4 \end{pmatrix} + \begin{pmatrix} 0 & M_2 \\ M_3 & 0 \end{pmatrix},$$

where M_1 is a principal submatrix of R_0, M_4 is a principal submatrix of $-R_b$, M_2 is a submatrix of R_b and M_3 is a submatrix of $-R_0$. Because $\boldsymbol{K}$ is maximal, M_3

does not contain any diagonal elements of $-R_0$. Hence, M_3 is a nonnegative matrix. Similarly, M_2 is a nonnegative matrix. Because R_0 is a completely-$\mathcal{S}$ matrix, hence, M_1 is an $\mathcal{S}$-matrix. Because $-R_b$ is an upper triangular matrix with positive diagonal elements, M_4 is an $\mathcal{S}$-matrix. Thus, M is an $\mathcal{S}$-matrix. □

We end this section by presenting an example in which the associated Skorohod problem does not have a unique solution. Consider, for example, a network of two stations in tandem. The routing matrix

$$P = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Assume that $\alpha^n = (1,0)'$ and $\mu^n = (1,1)'$ for each n. Then the corresponding reflection matrix

$$R = \begin{pmatrix} 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \end{pmatrix}.$$

We claim that the $(\boldsymbol{S}, R)$-regulation of $x(\cdot)$ is *not* unique for some path $x(\cdot)$. Here the state space $\boldsymbol{S}$ is a box, say, $\{x \in \mathbb{R}^2\colon\ 0 \leqslant x_i \leqslant 1 \text{ for } i = 1,2\}$. Note that the directions of reflection that correspond to the corner $(0,1)'$ are *parallel*, both being $(1,-1)'$. Let $x_1(t) = -t$ and $x_2(t) = 1 + t$ for $t \geqslant 0$. One can check that both (z, y) and $(\hat{z}, \hat{y})$ are $(\boldsymbol{S}, R)$-regulations of $x(\cdot)$, where, for $t \geqslant 0$,

$$z_1(t) = 0, \quad z_2(t) = 1, \quad y_1(t) = t, \quad y_2(t) = y_3(t) = y_4(t) = 0,$$

$$\hat{z}_1(t) = 0, \quad \hat{z}_2(t) = 1, \quad \hat{y}_1(t) = \hat{y}_2(t) = \hat{y}_3(t) = 0, \quad \text{and} \quad \hat{y}_4(t) = t.$$

Thus, the conventional continuous mapping approach for proving heavy traffic limit theorems cannot be applied to the network.

6. Fluid limits

Theorem 6.1 (Fluid limit theorem). Assume that (3.1)–(3.3) in theorem 3.1 hold. Then, for each sample path such that (5.1) holds,

$$\frac{1}{n} F^n(n\cdot) \longrightarrow et \text{ u.o.c.}, \quad \text{and} \quad \frac{1}{n} B^n(n\cdot) \longrightarrow et \text{ u.o.c.} \tag{6.1}$$

Proof. Fix a sample path such that (5.1) holds. For a sequence of paths f^n, recall that

$$\bar{f}^n(t) = \frac{1}{n} f^n(nt).$$

By (5.1),

$$\frac{\widehat{E}^n(nt)}{n} = \frac{\widehat{E}(\alpha^n t)}{n} \to 0 \text{ u.o.c.} \quad \text{and} \quad \frac{\widehat{S}^n(nt)}{n} = \frac{\widehat{S}(\mu^n t)}{n} \to 0 \text{ u.o.c.} \tag{6.2}$$

Let ω be a fixed sample path such that (6.2) holds. We claim that as $n \to \infty$,

$$\overline{X}^n(t) \to 0 \text{ u.o.c.}$$

In fact, for $s > 0$,

$$\big|\overline{X}^n(s)\big| \leqslant \frac{1}{n}\big|\widehat{E}^n\big(F^n(ns)\big)\big| + \big\|\big(I-P'\big)\big\|\frac{1}{n}\big|\widehat{S}^n\big(B^n(ns)\big)\big| + \big|\alpha^n - \big(I-P'\big)\mu^n\big|s,$$

where for a matrix A, $||A|| = \max_i \sum_j |A_{ij}|$. Thus,

$$\sup_{0\leqslant s\leqslant t}\big|\overline{X}^n(s)\big| \leqslant \frac{1}{n}\sup_{0\leqslant s\leqslant t}\big|\widehat{E}^n(ns)\big| + \big\|\big(I-P'\big)\big\|\frac{1}{n}\sup_{0\leqslant s\leqslant t}\big|\widehat{S}^n(ns)\big| + \big|\alpha^n - \big(I-P'\big)\mu^n\big|t$$

and $\sup_{0\leqslant s\leqslant t}|\overline{X}^n(s)| \to 0$ as $n \to \infty$.

It is easy to check that $(\overline{Z}^n(\cdot), \overline{Y}^n(\cdot))$ is an $(\overline{S}^n, R^n)$-regulation of $\overline{Z}^n(0) + \overline{X}^n(\cdot)$, where

$$\overline{S}^n = \big\{x \in \mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i^n/n\ \forall i \in \boldsymbol{I}\big\}.$$

By lemma 5.1, theorem 4.2 and the fact that $\overline{Y}^n(\cdot)$ is continuous, there exist constants $\kappa > 0$ and $n_0 > 0$ such that for each $t_1 < t_2$ and $n \geqslant n_0$,

$$\text{Osc}\big(\overline{Y}^n(\cdot,\omega), [t_1,t_2]\big) \leqslant \kappa \text{Osc}\big(\overline{X}^n(\cdot), [t_1,t_2]\big). \tag{6.3}$$

Since $|\overline{Y}^n(t) - \overline{Y}^n(s)| \leqslant 2d(t-s)$ for all n and $t > s > 0$, the sequence $\{\overline{Y}^n(\cdot)\}$ is precompact in $C([0,\infty), \mathbb{R}^{2d})$. Let $\overline{Y}$ be a limit of this sequence. Because $\overline{X}^n(\cdot,\omega) \to 0$ u.o.c. as $n \to \infty$, it follows from (6.3) that $\text{Osc}(\overline{Y}, [t_1,t_2]) = 0$ for any $0 \leqslant t_1 < t_2$. Since $\overline{Y}(0) = 0$, $\overline{Y}(t) = 0$ for all $t \geqslant 0$. Because each limit point $\overline{Y}$ is identically zero, $\overline{Y}^n(\cdot) \to 0$ u.o.c. as $n \to \infty$. The lemma then follows from (2.2). □

7. A stopping time property

In this section we prove a stopping time property that is essential to the proof of the main theorem. Let $p, q \in \mathbb{Z}_+^d$ be d-dimensional indexes. We use $U_i(\cdot \wedge j)$ to denote process $\{U_i(k \wedge j),\ k \geqslant 0\}$. For a d-dimensional index p, let $U(\cdot \wedge p) = (U_1(\cdot \wedge p_1), \ldots, U_d(\cdot \wedge p_d))$. For any $p, q \in \mathbb{Z}_+^d$, let

$$\mathcal{G}_{p,q}^n = \sigma\big\{U\big(\cdot \wedge (p+e)\big), V\big(\cdot \wedge (q+e)\big), Z^n(0)\big\}, \tag{7.1}$$

where e is the d-dimensional vector of ones. We assume that $\mathcal{G}_{p,q}^n$ has been augmented with all $\mathbb{P}$-null sets. Recall that $E_i^n(F_i^n(t))$ is the number of external arrivals to station i by time t and $S_i^n(B_i^n(t))$ is the number of departures from station i by time t.

Lemma 7.1 (Stopping time property). For any $p, q \in \mathbb{Z}_+^d$ and $t \geqslant 0$,

$$\big\{E^n\big(F^n(t)\big) = p,\ S^n\big(B^n(t)\big) = q\big\} \in \mathcal{G}_{p,q}^n. \tag{7.2}$$

Proof. Since we are going to prove (7.2) is true for each $n \geqslant 1$, we drop the superscript n in this proof. Let

$$A(t) = E\big(F(t)\big) \quad \text{and} \quad D(t) = S\big(B(t)\big).$$

When the event $A_i(t) = p_i$ occurs, the dynamics of the network in $[0, t]$ does *not* depend on the interarrival times $u_{i\ell}$ for $\ell > p_i+1$. Similarly, when the event $D_i(t) = q_i$ occurs, the dynamics of the network in $[0, t]$ does not depend on the service times $v_{i\ell}$ for $\ell > q_i + 1$. Thus, the lemma is intuitively obvious. However, a rigorous proof is needed to show that $A(t)$ and $D(t)$ *measurably* depend on interarrival and service times. The proof essentially requires us to go through the detailed construction of $A(t)$ and $D(t)$ from the primitive interarrival and service times.

We mimic the proof in Williams [38], where open multiclass queueing networks with *unlimited* buffer size were considered. An event time is the instant when a service completion or an arrival has just occurred. Let $e_0 = 0$ and e_l be the lth event time. Because the mean interarrival times and mean service times are positive, with probability one, $e_l \to \infty$ as $l \to \infty$. Thus, we have, with probability one,

$$\big\{A(t) = p,\ D(t) = q\big\} = \bigcup_{k\geqslant 1} \bigcap_{l\geqslant k} \big\{A(t \wedge e_l) = p,\ D(t \wedge e_l) = q\big\}.$$

Therefore, to show (7.2), it is enough to show

$$\big\{A(t \wedge e_l) = p,\ D(t \wedge e_l) = q\big\} \in \mathcal{G}_{p,q}.$$

for each $t \geqslant 0,\ l \geqslant 0$ and $p, q \in \mathbb{Z}_+^d$. (Here we used the fact that each $\mathcal{G}_{p,q}$ has been augmented with all $\mathbb{P}$-null sets.) For each $t \geqslant 0$ and $i \in \boldsymbol{I}$, let $R_i^a(t)$ be the remaining time (from time t) for the next external arrival to station i to occur if the arrival will never be turned off. Similarly, let $R_i^s(t)$ be the remaining time for the next service at station i to complete if the service will never be interrupted. If there is no customer in service at time t, $R_i^s(t) = \infty$. We adopt the convention that $\infty - a = \infty$ and $\min\{\infty, a\} = a$ for any constant a. We want to use induction to show that for each $l \geqslant 0$

$$C_{l,p,q} \equiv \big\{A(t \wedge e_l) = p,\ D(t \wedge e_l) = q\big\} \in \mathcal{G}_{p,q}, \tag{7.3}$$

$$1_{\{A(t\wedge e_l)=p, D(t\wedge e_l)=q\}}\xi_l \in \mathcal{G}_{p,q} \tag{7.4}$$

hold for each $t \geqslant 0$ and $p, q \in \mathbb{Z}_+^d$, where

$$\xi_l = \big(Z(t \wedge e_l),\ R^a(t \wedge e_l),\ R^s(t \wedge e_l),\ t \wedge e_l\big).$$

From our model assumption, $A_i(0) = 0$ and $D_i(0) = 0$, $R_i^a(0) = u_{i1}/\alpha_i$ and

$$R_i^s(0) = \begin{cases} m_i v_{i1} & \text{if } Z_i(0) > 0, \\ \infty & \text{if } Z_i(0) = 0. \end{cases}$$

Thus, $\xi_0 = (Z(0), R^a(0), R^s(0), 0) \in \mathcal{G}_{0,0}$. For any $(p, q) \neq (0, 0)$,

$$1_{\{A(t\wedge e_0)=p, D(t\wedge e_0)=q\}}\xi_0 = 0 \in \mathcal{G}_{p,q}.$$

Therefore, (7.3) and (7.4) hold for $l = 0$.

We now make the induction assumption that $C_{l,p,q} \in \mathcal{G}_{p,q}$ and $1_{C_{l,p,q}}\xi_l \in \mathcal{G}_{p,q}$ for all $p, q \in \mathbb{Z}_+^d$ and $t \geqslant 0$. We would like to show that $C_{l+1,p,q} \in \mathcal{G}_{p,q}$ and $1_{C_{l+1,p,q}}\xi_{l+1} \in \mathcal{G}_{p,q}$ for all $p, q \in \mathbb{Z}_+^d$ and $t \geqslant 0$. We first show that $1_{C_{l+1,p,q}}\xi_{l+1} \in \mathcal{G}_{p,q}$. Note that

$$1_{C_{l+1,p,q}}\xi_{l+1} = 1_{C_{l+1,p,q}}\xi_{l+1}1_{\{t\leqslant e_l\}} + 1_{C_{l+1,p,q}}\xi_{l+1}1_{\{e_l<t\}}.$$

It is clear that

$$1_{C_{l+1,p,q}}\xi_{l+1}1_{\{t\leqslant e_l\}} = 1_{C_{l,p,q}}\xi_l 1_{\{t=t\wedge e_l\}} \in \mathcal{G}_{p,q}$$

by the induction assumption. It remains to be shown that

$$1_{C_{l+1,p,q}}\xi_{l+1}1_{\{e_l<t\}} \in \mathcal{G}_{p,q}.$$

On $\{t > e_l\}$,

$$e_{l+1} = t \wedge e_l + \min_{i\in \boldsymbol{I}\setminus\boldsymbol{F},\, j\in\boldsymbol{I}\setminus\boldsymbol{B}} \big\{R_i^a(t\wedge e_l),\ R_j^s(t\wedge e_l)\big\}, \tag{7.5}$$

where $\boldsymbol{F} \subset \boldsymbol{I}$ is the set of buffers that are full at time $t \wedge e_l$, i.e.,

$$\boldsymbol{F} \equiv \big\{i \in \boldsymbol{I}\colon\ Z_i(t\wedge e_l) = b_i\big\},$$

and $\boldsymbol{B} \subset \boldsymbol{I}$ is the set of stations are blocked at time $t \wedge e_l$, i.e.,

$$\boldsymbol{B} \equiv \big\{i \in \boldsymbol{I}\colon\ Z_{\sigma(i)}(t\wedge e_l) = b_{\sigma(i)}\big\}.$$

It follows from (7.5) and the induction assumption that

$$1_{C_{l,m,n}}1_{\{t>e_l\}}e_{l+1} \in \mathcal{G}_{m,n} \tag{7.6}$$

for all $m, n \in \mathbb{Z}_+^d$.

Now,

$$1_{C_{l+1,p,q}}\xi_{l+1}1_{\{t>e_l\}} = \sum_{(\boldsymbol{a},\boldsymbol{s})} 1_{C_{l,\tilde{p},\tilde{q}}}\xi_{l+1}1_{\{t>e_l\}}1_{B_{\boldsymbol{a},\boldsymbol{s}}},$$

where

$$\begin{aligned} B_{\boldsymbol{a},\boldsymbol{s}} = {} & \bigcap_{i\in\boldsymbol{a}} \big\{R_i^a(t\wedge e_l) = e_{l+1} - t\wedge e_l\big\} \\ & \times \bigcap_{i\notin\boldsymbol{a}} \big(\big\{R_i^a(t\wedge e_l) > e_{l+1} - t\wedge e_l\big\} \cup \big\{Z_i(t\wedge e_l) = b_i\big\}\big) \\ & \times \bigcap_{i\in\boldsymbol{s}} \big\{R_i^s(t\wedge e_l) = e_{l+1} - t\wedge e_l\big\} \\ & \times \bigcap_{i\notin\boldsymbol{s}} \big(\big\{R_i^s(t\wedge e_l) > e_{l+1} - t\wedge e_l\big\} \cup \big\{Z_{\sigma(i)}(t\wedge e_l) = b_{\sigma(i)}\big\}\big), \end{aligned}$$

$$\tilde{p}_i = \begin{cases} p_i - 1 & \text{if } i \in \boldsymbol{a}, \\ p_i & \text{if } i \notin \boldsymbol{a}, \end{cases} \qquad \tilde{q}_i = \begin{cases} q_i - 1 & \text{if } i \in \boldsymbol{s}, \\ q_i & \text{if } i \notin \boldsymbol{s}, \end{cases}$$

$$Z_i(t \wedge e_{l+1}) = \begin{cases} Z_i(t \wedge e_l) + 1 & \text{if } i \in \boldsymbol{a} \setminus \boldsymbol{s}, \\ Z_i(t \wedge e_l) - 1 & \text{if } i \in \boldsymbol{s} \setminus \boldsymbol{a}, \\ Z_i(t \wedge e_l) & \text{otherwise,} \end{cases}$$

$$R_i^a(t \wedge e_{l+1}) = \begin{cases} u_{i,p_i} & \text{if } i \in \boldsymbol{a}, \\ R_i^a(t \wedge e_l) - (t \wedge e_{l+1} - t \wedge e_l) & \text{if } i \in \boldsymbol{I} \setminus \boldsymbol{a}, \end{cases}$$

$$R_i^s(t \wedge e_{l+1}) = \begin{cases} v_{i,q_i} & \text{if } i \in \boldsymbol{s}, \\ R_i^s(t \wedge e_l) - (t \wedge e_{l+1} - t \wedge e_l) & \text{if } i \in \boldsymbol{I} \setminus \boldsymbol{s}, \end{cases}$$

and the summation is over all pairs $(\boldsymbol{a}, \boldsymbol{s})$ with $\boldsymbol{a} \subset \boldsymbol{I}$ and $\boldsymbol{s} \subset \boldsymbol{I}$. The set $\boldsymbol{a} \cup \boldsymbol{s}$ is the set of indexes whose clocks "expire" exactly at e_{l+1}. If $\boldsymbol{a} \cup \boldsymbol{s} = \emptyset$, then the $(l+1)$th event has not yet happened by time t. It follows from (7.6) and the induction assumption that $1_{C_{l+1,p,q}} \xi_{l+1}$ is $\mathcal{G}_{p,q}$ measurable. Similarly, we can show that $1_{C_{l+1,p,q}} \in \mathcal{G}_{p,q}$. □

8. Proof of the heavy traffic limit theorem

For a sequence of functions f^n, recall that

$$\tilde{f}^n(t) = \frac{1}{\sqrt{n}} f^n(nt).$$

Lemma 8.1. Under the assumptions (3.1)–(3.3) in theorem 3.1, as $n \to \infty$,

$$\big(\widetilde{E}^n, \widetilde{S}^n, \widetilde{X}^n\big) \Longrightarrow \big(E^*, S^*, X^*\big),$$

where E^* and S^* are independent Brownian motions as in (5.2),

$$X^*(t) = E^*(\alpha t) - \big(I - P'\big) S^*(\mu t) + \theta t, \tag{8.1}$$

and X^* is a Brownian motion with drift θ and covariance matrix Γ given in (3.16).

Proof. Let $\widetilde{E}^n(t) = (1/\sqrt{n})\widehat{E}(\alpha^n nt)$ and $\widetilde{S}^n(t) = (1/\sqrt{n})\widehat{S}(\mu^n nt)$. It follows from (5.2), (3.1), (6.1) and the Random Change of Time Theorem [4, section 17] that

$$\big(\widetilde{E}^n\big(\overline{F}^{\,n}(\cdot)\big), \widetilde{S}^n\big(\overline{B}^{\,n}(\cdot)\big)\big) \Longrightarrow \big(E^*(\alpha\cdot), S^*(\mu\cdot)\big).$$

By the continuous mapping theorem,

$$\begin{aligned} \widetilde{X}^n(\cdot) \;&=\; \widetilde{E}^n\big(\overline{F}^{\,n}(\cdot)\big) - \big(I - P'\big)\widetilde{S}^n\big(\overline{B}^{\,n}(\cdot)\big) - \big(I - P'\big)\mu^n \cdot \\ &\Longrightarrow E^*(\alpha\cdot) - \big(I - P'\big) S^*(\mu\cdot) + \theta \cdot . \end{aligned}$$

It is easy to check that X^* is a Brownian motion with drift θ and covariance matrix Γ given in (3.16). □

A sequence of stochastic processes $\{X^n\}$ in $D([0,\infty),\mathbb{R}^k)$ is said to be relatively compact if for every sequence $\{n_k\}$, there is a subsequence $\{n_{k_j}\}$ such that $X^{n_{k_j}}$ converges in distribution.

Lemma 8.2. Under the assumptions (3.1)–(3.4) in theorem 3.1, the sequence $\{\widetilde{X}^n, \widetilde{Z}^n, \widetilde{Y}^n\}$ is relatively compact.

Proof. To prove the lemma it suffices to verify conditions (a) and (b) in corollary 7.4 in chapter 3 of Ethier and Kurtz [21]. To state the conditions, we need to define the modulus of continuity of a path $x(\cdot)$. For $T > 0$ and $\delta > 0$, let

$$w\big(x(\cdot),\delta,T\big) = \inf_{t_i} \max_i \operatorname{Osc}\big(x(\cdot),[t_{i-1},t_i)\big), \tag{8.2}$$

where the infimum extends over the finite sets $\{t_i\}$ of points satisfying $0 = t_0 < t_1 < \cdots < t_r = T$ and $t_j - t_{j-1} > \delta$ for $j = 1,\ldots,r$.

(a) For every $\eta > 0$ and rational $t \geqslant 0$, there exists a constant $c(\eta,t) > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\big\{\big|\big(\widetilde{X}^n(t),\widetilde{Z}^n(t),\widetilde{Y}^n(t)\big)\big| \leqslant c(\eta,t)\big\} \geqslant 1-\eta.$$

(b) For every $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\big\{w\big(\big(\widetilde{X}^n,\widetilde{Z}^n,\widetilde{Y}^n\big),\delta,T\big) \geqslant \eta\big\} \leqslant \eta.$$

To verify condition (a), by lemma 8.1, $\widetilde{X}^n$ converges in distribution. Hence, it follows from remark 7.3 in chapter 3 of Ethier and Kurtz [21] that $\{\widetilde{X}^n\}$ satisfies the following compact containment condition: for every $\eta > 0$ and $T > 0$, there is a constant $M_1 > 0$ such that

$$\inf_n \mathbb{P}\big\{\big|\widetilde{X}^n(t)\big| \leqslant M_1,\ 0 \leqslant t \leqslant T\big\} \geqslant 1-\eta/2.$$

By assumption (3.4), there exists a constant $M_2 > 0$ such that $\sup_n \mathbb{P}\{|\widetilde{Z}^n(0)| > M_2\} \leqslant \eta/2$. It is easy to check that for each sample path, $(\widetilde{Z}^n,\widetilde{Y}^n)$ is an $(\widetilde{S}^n,R^n)$-regulation of $\widetilde{Z}^n(0)+\widetilde{X}^n$, where

$$\widetilde{\boldsymbol{S}}^n = \big\{x\in\mathbb{R}^d\colon\ 0 \leqslant x_i \leqslant b_i^n/\sqrt{n}\ \forall i \in \boldsymbol{I}\big\}.$$

Therefore by theorem 4.2 and the continuity of $\widetilde{Y}^n$, there exist constants $\kappa > 0$ and $n_0 > 0$ such that for all $0 \leqslant t_1 < t_2$ and all $n \geqslant n_0$,

$$\operatorname{Osc}\big(\big(\widetilde{X}^n,\widetilde{Z}^n,\widetilde{Y}^n\big),[t_1,t_2]\big) \leqslant \kappa \operatorname{Osc}\big(\widetilde{X}^n,[t_1,t_2]\big). \tag{8.3}$$

Thus, we have, for $n \geqslant n_0$,

$$\begin{aligned}
&\big|\big(\widetilde{X}^n(t), \widetilde{Z}^n(t), \widetilde{Y}^n(t)\big)\big| \\
&\quad \leqslant \big|\big(\widetilde{X}^n(0), \widetilde{Z}^n(0), \widetilde{Y}^n(0)\big)\big| + \big|\big(\widetilde{X}^n(t), \widetilde{Z}^n(t), \widetilde{Y}^n(t)\big) - \big(\widetilde{X}^n(0), \widetilde{Z}^n(0), \widetilde{Y}^n(0)\big)\big| \\
&\quad \leqslant \big|\widetilde{Z}^n(0)\big| + \operatorname{Osc}\big(\big(\widetilde{X}^n, \widetilde{Z}^n, \widetilde{Y}^n\big), [0,t]\big) \\
&\quad \leqslant \big|\widetilde{Z}^n(0)\big| + \kappa \operatorname{Osc}\big(\widetilde{X}^n, [0,t]\big) \leqslant \big|\widetilde{Z}^n(0)\big| + \kappa \sup_{0\leqslant t\leqslant T} \big|\widetilde{X}^n(t)\big|.
\end{aligned}$$

Hence, for $n \geqslant n_0$

$$\begin{aligned}
&\mathbb{P}\big\{\big|\big(\widetilde{X}^n(t), \widetilde{Z}^n(t), \widetilde{Y}^n(t)\big)\big| > M_2 + \kappa M_1 \text{ for some } t \in [0,T]\big\} \\
&\quad \leqslant \mathbb{P}\big\{\big|\widetilde{Z}^n(0)\big| > M_2\big\} + \mathbb{P}\big\{\big|\widetilde{X}^n(t)\big| > M_1 \text{ for some } t \in [0,T]\big\} \leqslant \eta.
\end{aligned}$$

Therefore, $\{(\widetilde{X}^n, \widetilde{Z}^n, \widetilde{Y}^n)\}$ satisfies the containment condition. Thus, condition (a) in corollary 7.4 holds.

To verify condition (b) in corollary 7.4, because $\{\widetilde{X}^n\}$ is relatively compact, for each $\eta > 0$ and $T > 0$, there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left\{w\big(\widetilde{X}_n, \delta, T\big) \geqslant \frac{\eta}{\kappa+1}\right\} \leqslant \frac{\eta}{\kappa+1}.$$

From (8.3), for $n \geqslant n_0$,

$$w\big(\big(\widetilde{X}_n, \widetilde{Z}^n, \widetilde{Y}^n\big), \delta, T\big) \leqslant \kappa w\big(\widetilde{X}_n, \delta, T\big).$$

Therefore, for $n \geqslant n_0$,

$$\begin{aligned}
\mathbb{P}\big\{w\big(\big(\big(\widetilde{X}_n, \widetilde{Z}^n, \widetilde{Y}^n\big), \delta, T\big) \geqslant \eta\big\} &\leqslant \mathbb{P}\big\{\kappa w\big(\widetilde{X}_n, \delta, T\big) \geqslant \eta\big\} \\
&\leqslant \mathbb{P}\left\{w\big(\widetilde{X}_n, \delta, T\big) \geqslant \frac{\eta}{\kappa+1}\right\} \leqslant \frac{\eta}{\kappa+1} \leqslant \eta.
\end{aligned}$$

Thus, condition (b) in corollary 7.4 holds. □

Lemma 8.3. Suppose z^n converges to z in $D([0,\infty), \mathbb{R}^d)$, y^n converges to y in $D([0,\infty), \mathbb{R}_+)$ and y is continuous. Assume that for each n, $y^n(\cdot)$ is nondecreasing. Then, for any $f \in C_b(\mathbb{R}^d)$, we have

$$\int_0^t f\big(z^n(s)\big)\, \mathrm{d}y^n(s) \to \int_0^t f\big(z(s)\big)\, \mathrm{d}y(s) \quad \text{as } n \to \infty \tag{8.4}$$

uniformly for t in any compact subset of $[0,\infty)$.

Proof. Noting that $z^n \to z$ in $D([0,\infty), \mathbb{R}^d)$, by proposition 3.5.3 and remark 3.5.4 in Ethier and Kurtz [21] or Billingsley [4, p. 112], there exists a sequence $\{\gamma_n\}$

of continuous, strictly increasing functions from $[0,\infty)$ onto $[0,\infty)$ such that, as $n \to \infty$,

$$z^n\big(\gamma_n(t)\big) \to z(t) \text{ u.o.c.} \quad \text{and} \quad \gamma_n(\cdot) \to t \text{ u.o.c.} \tag{8.5}$$

Now, fix $t > 0$ and observe that for each $u \in [0,t]$,

$$\begin{aligned}
&\int_0^u f\big(z^n(s)\big)\,\mathrm{d}y^n(s) - \int_0^u f\big(z(s)\big)\,\mathrm{d}y(s) \\
&\quad = \int_0^{\gamma_n^{-1}(u)} \big(f\big(z^n\big(\gamma_n(s)\big)\big) - f\big(z(s)\big)\big)\,\mathrm{d}y^n\big(\gamma_n(s)\big) \\
&\qquad + \int_u^{\gamma_n^{-1}(u)} f\big(z(s)\big)\,\mathrm{d}y^n\big(\gamma_n(s)\big) + \int_0^u f\big(z(s)\big)\,\mathrm{d}\big(y^n(\gamma_n) - y\big)(s).
\end{aligned} \tag{8.6}$$

The first term on the right side of (8.6) is bounded by

$$\max_{0\leqslant s\leqslant \gamma_n^{-1}(t)} \big|f\big(z^n\big(\gamma_n(s)\big)\big) - f\big(z(s)\big)\big|\, y^n(t),$$

which converges to zero as $n \to \infty$ uniformly on $u \in [0,t]$ because $f \in C_b(\mathbb{R}^d)$, $y(t)$ is continuous, and $y^n(t) \to y(t)$.

The second term on the right side of (8.6) is dominated by

$$\begin{aligned}
&\|f\|_\infty \sup_{0\leqslant u\leqslant t} \big|y^n(u) - y^n\big(\gamma_n(u)\big)\big| \\
&\quad \leqslant \|f\|_\infty \Big(\sup_{0\leqslant u\leqslant t} \big|y^n(u) - y(u)\big| + \sup_{0\leqslant u\leqslant t} \big|y(u) - y\big(\gamma_n(u)\big)\big| \\
&\qquad + \sup_{0\leqslant u\leqslant t} \big|y\big(\gamma_n(u)\big) - y^n\big(\gamma_n(u)\big)\big| \Big),
\end{aligned}$$

which converges to zero because $y(t)$ is continuous, and $y^n(t) \to y(t)$ u.o.c.

Finally, we claim that the third term on the right side of (8.6) converges to zero. In fact, since $f(z(\cdot)) \in D([0,\infty),\mathbb{R})$, by theorem 3.5.6, proposition 3.5.3 and remark 3.5.4 of Ethier and Kurtz [21], there is a sequence of step functions $\{g^k(\cdot)\}_{k=1}^\infty$ of the form

$$g^k(\cdot) = \sum_{i=1}^{l_k} g^k\big(t_i^k\big) I_{[t_i^k, t_{i+1}^k)}(\cdot), \tag{8.7}$$

where $0 = t_1^k < t_2^k < \cdots < t_{l_{k+1}}^k < \infty$, $I_{[s,t)}$ is the indicator function on $[s,t)$, and

$$\sup_{0\leqslant s\leqslant t} \big|f\big(z(s)\big) - g^k(s)\big| \to 0 \quad \text{as } k \to \infty.$$

Notice that

$$\begin{aligned}
&\left| \int_0^u f\big(z(s)\big)\, \mathrm{d}\big(y^n(\gamma_n) - y\big)(s) \right| \\
&\quad \leqslant \left| \int_0^u \big(f\big(z(s)\big) - g^k(s)\big)\, \mathrm{d}\big(y^n(\gamma_n) - y\big)(s) \right| + \left| \int_0^u g^k(s)\, \mathrm{d}\big(y^n(\gamma_n) - y\big)(s) \right| \\
&\quad \leqslant \sup_{0\leqslant s\leqslant t} \big|f\big(z(s)\big) - g^k(s)\big| \big(y^n(\gamma_n)(t) + y(t)\big) \\
&\qquad + \sup_{0\leqslant u\leqslant t} \sum_{i=1}^{l_k} \big|g^k\big(t_i^k \wedge u\big)\big| \\
&\qquad \times \big|\big(y^n(\gamma_n) - y\big)\big(t_{i+1}^k \wedge u\big) - \big(y^n(\gamma_n) - y\big)\big(t_i^k \wedge u\big)\big|. \qquad (8.8)
\end{aligned}$$

Because $y^n(\cdot) \to y(\cdot)$ u.o.c. and y is continuous, for each $t > 0$, there exists $M > 0$ such that

$$\limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \big|y^n(s)\big| \leqslant M.$$

Letting $n \to \infty$ in (8.8), noticing that for fixed k, the last term of (8.8) converges to zero, we have

$$\limsup_{n\to\infty} \sup_{0\leqslant u\leqslant t} \left| \int_0^u f\big(z(s)\big) \mathrm{d}\big(y^n(\gamma_n) - y\big)(s) \right| \leqslant 2M \sup_{0\leqslant s\leqslant t} \big|f\big(z(s)\big) - g^k(s)\big|. \qquad (8.9)$$

Let $k \to \infty$, we have

$$\limsup_{n\to\infty} \sup_{0\leqslant u\leqslant t} \left| \int_0^u f\big(z(s)\big) \mathrm{d}\big(y^n(\gamma_n) - y\big)(s) \right| = 0, \qquad (8.10)$$

thus proving the lemma. □

Lemma 8.4. For $i \in \boldsymbol{I}$ and any $t \geqslant 0$,
(a)

$$\mathbb{E}\left[\frac{1}{\sqrt{n}} \max_{1\leqslant j\leqslant E_i(nt)+1} u_{ij} \right] \to 0 \quad \text{as } n \to \infty.$$

(b)

$$\left\{ \frac{1}{\sqrt{n}} \sup_{0\leqslant s\leqslant 1} \big|E_i(ns) - ns\big| \colon\ n \geqslant 1 \right\}$$

is uniformly integrable.

Proof. Noting that $E_i(t) + 1$ is a stopping time for the discrete filtration $\{\mathcal{G}_j\}$ with

$$\mathcal{G}_j = \sigma\{u_{i1}, \ldots, u_{ij}\},$$

we can write

$$\frac{E_i(nt)-nt}{\sqrt{n}} = \frac{E_i(nt)+1-U_i(E_i(nt)+1)}{\sqrt{n}} - \frac{1}{\sqrt{n}} + \frac{U_i(E_i(nt)+1)-t}{\sqrt{n}}. \quad (8.11)$$

The first term on the right, denoted by $M_i^n(t)$, is a square integrable martingale with

$$\mathbb{E}\big[M_i^n(t)^2\big] = c_i^a \frac{\mathbb{E}[E_i(nt)+1]}{n}. \quad (8.12)$$

Since the right-hand side of (8.12) is bounded in n [23, theorem II.5.1], by [21, corollary 2.2.17] the sequence $\mathbb{E}[\sup_{0\leqslant t\leqslant 1}|M_i^n(t)|]^2$ is bounded, hence,

$$\Big\{ \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big|,\ n\geqslant 1\Big\}$$

is uniformly integrable. Using the fact that

$$\sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)-M_i^n(t-)\big| \leqslant 2 \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big|,$$

for $0\leqslant t\leqslant 1$, the last term on the right of (8.11) (the overshoot of the renewal process) is bounded by

$$\max_{0\leqslant j\leqslant E_i(n)+1} \frac{u_{ij}}{\sqrt{n}} \leqslant 2 \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big| + \frac{1}{\sqrt{n}}.$$

We then have

$$\sup_{0\leqslant t\leqslant 1} \left|\frac{E_i(nt)-nt}{\sqrt{n}}\right| \leqslant 3 \sup_{0\leqslant t\leqslant 1} \big|M_i^n(t)\big| + \frac{2}{\sqrt{n}},$$

and (a) and (b) follow from the uniform integrability of $\{\sup_{0\leqslant t\leqslant 1}|M_i^n(t)|,\ n\geqslant 1\}$. □

Proof of theorem 3.1. By lemma 8.2 the sequence

$$\big\{\big(\widetilde{Z}^n,\widetilde{X}^n,\widetilde{Y}^n\big),\ n\geqslant n_0\big\}$$

is precompact. Therefore,

$$\big\{\big(\widetilde{E}^n,\widetilde{S}^n,\widetilde{Z}^n,\widetilde{X}^n,\widetilde{Y}^n\big),\ n\geqslant n_0\big\}$$

is precompact. Let (E^*,S^*,Z^*,X^*,Y^*) be a weak limit defined on a probability space $(\Omega^*,\mathcal{F}^*,\mathbb{P}^*)$. That is, there is a sequence $\{n_k\}$ such that as $n_k\to\infty$

$$\big(\widetilde{E}^{n_k},\widetilde{S}^{n_k},\widetilde{Z}^{n_k},\widetilde{X}^{n_k},\widetilde{Y}^{n_k}\big) \Longrightarrow \big(E^*,S^*,Z^*,X^*,Y^*\big).$$

By lemma 8.1,

$$X^*(t) = E^*(\alpha t) - \big(I-P'\big)S^*(\mu t) \quad (8.13)$$

is a d-dimensional Brownian motion with drift θ and covariance matrix Γ. We will show that Z^*, together with Y^*, is an RBM associated with the Brownian motion X^*.

Because the $(\Gamma, \theta, R, \boldsymbol{S})$-RBM with initial distribution $\mathbb{P}^* Z^*(0)^{-1}$ is unique in distribution (see Dai and Williams [18]), we have

$$\big(\widetilde{Z}^n, \widetilde{X}^n, \widetilde{Y}^n\big) \Longrightarrow \big(Z^*, X^*, Y^*\big),$$

as $n \to \infty$, thus proving the theorem.

To show Z^* is an RBM, notice that

(i) $\widetilde{Z}^n(t) = \widetilde{Z}^n(0) + \widetilde{X}^n(t) + R^n \widetilde{Y}^n(t)$ for all $t \geqslant 0$,

(ii) $0 \leqslant \widetilde{Z}_i^n(t) \leqslant b_i^n/\sqrt{n}$ for all $t \geqslant 0$ and $i = 1, \ldots, d$,

(iii) for each $i = 1, \ldots, 2d$,

(a) $\widetilde{Y}_i^n(0) = 0$,

(b) $\widetilde{Y}_i^n$ is nondecreasing,

(c) for $i = 1, \ldots, d$, $\widetilde{Y}_i^n$ increases only when $\widetilde{Z}_i^n(t) = 0$ and for $i = d+1, \ldots, 2d$, $\widetilde{Y}_i^n$ increases only when $\widetilde{Z}_i^n(t) = b_i^n/\sqrt{n}$.

To show that the limit process (Z^*, X^*, Y^*) satisfies (3.7)–(3.15), we invoke the Skorohod representation theorem [21, theorem 3.1.8]. Therefore, we assume that $\{(\widetilde{Z}^{n_k}, \widetilde{X}^{n_k}, \widetilde{Y}^{n_k}),\ n \geqslant n_0\}$ and (Z^*, X^*, Y^*) are defined on the same probability space $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ such that $\mathbb{P}^*$-a.s., (i)–(iii) hold and

$$\big(\widetilde{Z}^{n_k}, \widetilde{X}^{n_k}, \widetilde{Y}^{n_k}\big) \to \big(Z^*, X^*, Y^*\big) \text{ u.o.c.} \quad \text{as } n_k \to \infty. \tag{8.14}$$

It follows from (3.4) that $Z^*(0)$ has the same distribution as ξ. Clearly, (3.10) is satisfied with

$$\mathcal{F}_t^* \equiv \sigma\big\{\big(Z^*(s), X^*(s), Y^*(s)\big),\ 0 \leqslant s \leqslant t\big\}.$$

It easy to check that (3.7), (3.8), and (3.11) follow from (i), (ii), (iii)(a) and (iii)(b). Because $\widetilde{Y}_{i+d}^n(t)$ increases only at times t such that $\widetilde{Z}_i^n(t) = b_i^n/\sqrt{n}$, we have for each $T > 0$

$$\int_0^T \left(\frac{b_i^n}{\sqrt{n}} - \widetilde{Z}_i^n(t)\right) \wedge 1 \, \mathrm{d}\widetilde{Y}_{i+d}^n(t) = 0. \tag{8.15}$$

Let

$$f : (b, z) \in \mathbb{R}^2 \to f(b, z) = (b - z) \wedge 1.$$

Clearly, $f \in C_b(\mathbb{R}^2)$. By lemma 8.3 and (8.14),

$$\int_0^T \big(b_i - Z_i^*(t)\big) \wedge 1 \, \mathrm{d}\widetilde{Y}_{i+d}^*(t) = 0, \quad \text{for all } T > 0.$$

Therefore, $Y_{i+d}^*(\cdot)$ increases only at times t such that $Z_i^*(t) = b_i$, showing (3.13). Similarly, we can show that $Y_i^*(\cdot)$ increases only at times t when $Z_i^*(t) = 0$, i.e., (3.12) holds.

It remains to prove (3.15), i.e., $\{X^*(t) - \theta t,\ t \geqslant 0\}$ is an $\{\mathcal{F}_t^*\}$-martingale. It is enough to show that for each $i \in \boldsymbol{I}$, each $r \geqslant 1$, any $0 \leqslant s_1 < s_2 < \cdots < s_r \leqslant s < t$, and any $f_k, g_k, h_k \in C_b(\mathbb{R}^d)$,

$$\mathbb{E}^*\left[\big(X_i^*(t+s) - X_i^*(s) - \theta_i t\big)\prod_{k=1}^{r} f_j\big(X^*(s_j)\big)g_j\big(Y^*(s_j)\big)h_j\big(Z^*(s_j)\big)\right] = 0. \quad (8.16)$$

Let $p, q \in \mathbb{Z}_+^d$ be d-dimensional indexes. Let

$$\widehat{U}_i^n(p_i) = \sum_{k=2}^{p_i+1} \frac{u_{ik} - 1}{\alpha_i^n}, \qquad \widehat{V}_i^n(q_i) = \sum_{k=2}^{q_i+1} \frac{v_{ik} - 1}{\mu_i^n}.$$

Recall the definition of $\mathcal{G}_{p,q}^n$ in (7.1). Because $Z^n(0)$ is assumed to be independent of the interarrival time and service time sequences, it easy to check that

$$\big\{\big(\widehat{U}^n(p), \widehat{V}^n(q)\big),\ \mathcal{G}_{p,q}^n,\ (p,q) \in \mathbb{Z}_+^d \times \mathbb{Z}_+^d\big\}$$

is a multiparameter martingale (see [21, section 2.8] for the definition). Let

$$A^n(t) = E^n\big(F^n(t)\big), \qquad D^n(t) = S^n\big(B^n(t)\big), \qquad \tau^n(t) = \big(A^n(t), D^n(t)\big).$$

By lemma 7.1, for each fixed t, $\tau^n(t)$ is a multidimensional stopping time with respect to the filtration $\{\mathcal{G}_{p,q}^n\}$. Define

$$\mathcal{G}_{\tau^n(t)}^n \equiv \big\{B \in \mathcal{F},\ B \cap \big\{\tau^n(t) \leqslant (p,q)\big\} \in \mathcal{G}_{p,q}^n \text{ for all } (p,q) \in \mathbb{Z}_+^d \times \mathbb{Z}_+^d\big\}.$$

It is clear that $\tau^n(t) \in \mathcal{F}_{\tau^n(t)}^n$. Because $Z^n(0) \in \mathcal{G}_{0,0}^n$, it follows from (5.4) that $Z^n(t) \in \mathcal{G}_{\tau^n(t)}^n$. From (2.1) $Y^n(t) \in \mathcal{G}_{\tau^n(t)}^n$ and from (5.7) $X^n(t) \in \mathcal{G}_{\tau^n(t)}^n$. Let

$$\widehat{U}^{n,k}(p) = \big(\widehat{U}_1^n(p_1 \wedge k), \ldots, \widehat{U}_d^n(p_d \wedge k)\big),$$
$$\widehat{V}^{n,k}(q) = \big(\widehat{V}_1^n(q_1 \wedge k), \ldots, \widehat{V}_d^n(q_d \wedge k)\big).$$

By the multiparameter optional stopping theorem [21, theorem 2.8.7] we have that for each $n \geqslant n_0$ and $k \geqslant 1$,

$$\big\{\big(\widehat{U}^{n,k}\big(A^n(t)\big), \widehat{V}^{n,k}\big(D^n(t)\big)\big), \mathcal{G}_{\tau^n(t)}^n,\ t \geqslant 0\big\}$$

is a martingale, or

$$\left\{\left(\frac{1}{\sqrt{n}}\widehat{U}^{n,k}\big(A^n(nt)\big), \frac{1}{\sqrt{n}}\widehat{V}^{n,k}\big(D^n(nt)\big)\right), \mathcal{G}_{\tau^n(nt)}^n,\ t \geqslant 0\right\}$$

is a martingale. Therefore, for each $n \geqslant n_0$ and $k \geqslant 1$,

$$\mathbb{E}\Bigg[\bigg(\frac{1}{\sqrt{n}}\widehat{U}_i^{n,k}\big(A_i^n\big(n(t+s)\big)\big) - \frac{1}{\sqrt{n}}\widehat{U}_i^{n,k}\big(A_i^n(ns)\big)\bigg) \times \prod_{j=1}^{r} f_j\big(\widetilde{X}^n(s_j)\big) g_j\big(\widetilde{Y}^n(s_j)\big) h_j\big(\widetilde{Z}^n(s_j)\big)\Bigg] = 0. \tag{8.17}$$

For a fixed n and for each $k \geqslant 1$,

$$\begin{aligned}\big|\widehat{U}_i^{n,k}\big(A_i^n(ns)\big)\big| &\leqslant \sum_{j=2}^{(A_i^n(ns)\wedge k)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{A_i^n(ns)\wedge k}{\alpha_i^n} \\ &\leqslant \sum_{j=1}^{A_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{A_i^n(ns)}{\alpha_i^n} \leqslant \sum_{j=1}^{E_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{E_i^n(ns)}{\alpha_i^n}.\end{aligned}$$

Letting $k \to \infty$ in (8.17), by [23, theorem III.3.1],

$$\mathbb{E}\Bigg[\sum_{j=1}^{E_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} + \frac{E_i^n(ns)}{\alpha_i^n}\Bigg] < \infty,$$

it follows from the dominated convergence theorem that for each $n \geqslant 1$,

$$\mathbb{E}\Bigg[\bigg(\frac{1}{\sqrt{n}}\widehat{U}_i^{n}\big(A_i^n\big(n(t+s)\big)\big) - \frac{1}{\sqrt{n}}\widehat{U}_i^{n}\big(A_i^n(ns)\big)\bigg) \times \prod_{j=1}^{r} f_j\big(\widetilde{X}^n(s_j)\big) g_j\big(\widetilde{Y}^n(s_j)\big) h_j\big(\widetilde{Z}^n(s_j)\big)\Bigg] = 0. \tag{8.18}$$

$$\begin{aligned}\widehat{U}_i^n\big(A_i^n(ns)\big) &= \sum_{j=2}^{A_i^n(ns)+1} \frac{u_{ij}}{\alpha_i^n} - \frac{A_i^n(ns)}{\alpha_i^n} = \sum_{j=2}^{E_i^n(F_i^n(ns))+1} \frac{u_{ij}}{\alpha_i^n} - \frac{E_i^n(F_i^n(ns))}{\alpha_i^n} \\ &= \sum_{j=2}^{E_i^n(F_i^n(ns))+1} \frac{u_{ij}}{\alpha_i^n} - F_i^n(ns) + F_i^n(ns) - \frac{E_i^n(F_i^n(ns))}{\alpha_i^n} \\ &= \varepsilon^n - \frac{\widehat{E}_i^n(F_i^n(ns))}{\alpha_i^n},\end{aligned}$$

where

$$\varepsilon^n = \sum_{j=2}^{E_i^n(F_i^n(ns))+1} \frac{u_{ij}}{\alpha_i^n} - F_i^n(ns).$$

Because

$$\begin{aligned}\left|\varepsilon^n\right| &\leqslant \frac{u_{i1}}{\alpha_i^n} + \max_{1\leqslant j\leqslant E_i^n(F_i^n(ns))+1} \frac{u_{i,j}}{\alpha_i^n} \leqslant \frac{u_{i1}}{\alpha_i^n} + \max_{1\leqslant j\leqslant E_i^n(ns)+1} \frac{u_{i,j}}{\alpha_i^n} \\ &\leqslant \frac{u_{i1}}{\alpha_i^n} + \max_{1\leqslant j\leqslant E_i(n\alpha_i^n s)+1} \frac{u_{i,j}}{\alpha_i^n},\end{aligned}$$

it follows from part (a) of lemma 8.4 that as $n \to \infty$,

$$\mathbb{E}\left[\frac{1}{\sqrt{n}}\left|\varepsilon^n\right|\right] \to 0.$$

Because $\alpha_i^n \to \alpha_i$, by part (b) of lemma 8.4,

$$\left\{\frac{1}{\sqrt{n}} \sup_{0\leqslant t\leqslant s} \left|\widehat{E}_i\big(\alpha_i^n nt\big)\right|,\ n \geqslant 1\right\}$$

is uniformly integrable. Notice that

$$\left|\widehat{E}_i^n\big(F_i^n(ns)\big)\right| \leqslant \sup_{0\leqslant t\leqslant s} \left|\widehat{E}_i^n(nt)\right|$$

and, therefore,

$$\left\{\frac{1}{\sqrt{n}}\left|\widehat{E}_i^n\big(F_i^n(ns)\big)\right|,\ n \geqslant 1\right\}$$

is uniformly integrable. Because

$$\left(\frac{1}{\sqrt{n_k}}\widehat{E}_i^{n_k}\big(F_i^{n_k}(n_k s)\big), \widetilde{X}^{n_k}(\cdot), \widetilde{Z}^{n_k}(\cdot), \widetilde{Y}^{n_k}(\cdot)\right) \Longrightarrow \big(E_i^*(\alpha_i s), X^*(\cdot), Z^*(\cdot), Y^*(\cdot)\big),$$

we have

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\big(A_i^{n_k}(n_k s)\big)\right)\prod_{j=1}^r f_j\big(\widetilde{X}^{n_k}(s_j)\big)g_j\big(\widetilde{Y}^{n_k}(s_j)\big)h_j\big(\widetilde{Y}^{n_k}(s_j)\big)\right] \\ &\quad\to \mathbb{E}^*\left[-\frac{E_i^*(\alpha_i s)}{\alpha_i}\prod_{j=1}^r f_j\big(X^*(s_j)\big)g_j\big(Y^*(s_j)\big)h_j\big(Z^*(s_j)\big)\right].\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\big(A_i^{n_k}\big(n_k(t+s)\big)\big)\right)\prod_{j=1}^r f_j\big(\widetilde{X}^{n_k}(s_j)\big)g_j\big(\widetilde{Y}^{n_k}(s_j)\big)h_j\big(\widetilde{Y}^{n_k}(s_j)\big)\right] \\ &\quad\to \mathbb{E}^*\left[-\frac{E_i^*\big(\alpha_i(t+s)\big)}{\alpha_i}\prod_{j=1}^r f_j\big(X^*(s_j)\big)g_j\big(Y^*(s_j)\big)h_j\big(Z^*(s_j)\big)\right].\end{aligned}$$

Therefore, we have

$$\mathbb{E}\Bigg[\bigg(\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\big(A_i^{n_k}\big(n_k(t+s)\big)\big)-\frac{1}{\sqrt{n_k}}\widehat{U}_i^{n_k}\big(A_i^{n_k}(n_k s)\big)\bigg)\times\prod_{j=1}^{r}f_j\big(\widetilde{X}^{n_k}(s_j)\big)g_j\big(\widetilde{Y}^{n_k}(s_j)\big)h_j\big(\widetilde{Y}^{n_k}(s_j)\big)\Bigg]$$
$$\to-\frac{1}{\alpha_i}\mathbb{E}^*\Bigg[\big(E_i^*(\alpha_i(t+s)-E_i^*(\alpha_i s)\big)\prod_{j=1}^{r}f_j\big(X^*(s_j)\big)g_j\big(Y^*(s_j)\big)h_j\big(Z^*(s_j)\big)\Bigg].$$

From (8.18), we have

$$\mathbb{E}^*\Bigg[\big(E_i^*\big(\alpha_i(t+s)\big)-E_i^*(\alpha_i s)\big)\prod_{j=1}^{r}f_j\big(X^*(s_j)\big)g_j\big(Y^*(s_j)\big)h_j\big(Z^*(s_j)\big)\Bigg]=0.$$

By the exact same proof, we have

$$\mathbb{E}^*\Bigg[\big(S_i^*\big(\mu_i(t+s)\big)-S_i^*(\mu_i s)\big)\prod_{j=1}^{r}f_j\big(X^*(s_j)\big)g_j\big(Y^*(s_j)\big)h_j\big(Z^*(s_j)\big)\Bigg]=0.$$

Therefore, (8.16) follows from the fact that

$$X_i^*(t+s)-X_i^*(s)-\theta_i t=E_i^*\big(\alpha_i(t+s)\big)-E_i^*(\alpha_i s)-\big(S_i^*\big(\mu_i(t+s)\big)-S_i^*(\mu_i s)\big).\ \square$$

9. Extensions

Consider the queueing network described in section 2, except that probabilistic routing is allowed. Assume that a customer leaving station $i\in\boldsymbol{I}$ goes to station $j\in\boldsymbol{I}$ with probability P_{ij} or exits the network with probability $1-\sum_{j\in\boldsymbol{I}}P_{ij}$, independent of all previous history. Assume the network is feedforward, i.e., the stations can be numbered so that $P_{ij}=0$ for $j\leqslant i$. Furthermore, we assume that each station has at most one predecessor. That is,

$$\sigma(i)\cap\sigma(j)=\emptyset\quad\text{for any } i\neq j,$$

where $\sigma(i)=\{j\in\boldsymbol{I}\colon\ P_{ij}>0\}$.

For this network, using the techniques developed in this paper, we can show that the heavy traffic limit theorem in theorem 3.1 holds with Γ replaced by the formula

$$\Gamma=\operatorname{diag}\big(\alpha_1c_1^a,\ldots,\alpha_dc_d^a\big)+\big(I-P'\big)\operatorname{diag}\big(\mu_1c_1^s,\ldots,\mu_dc_d^s\big)(I-P)+\sum_{j\in\boldsymbol{I}}\mu_j\Gamma^j,\quad(9.1)$$

where

$$\Gamma^j_{lk}=\begin{cases}P_{jl}(1-P_{jl}) & \text{if } l=k,\\ -P_{jl}P_{jk} & \text{if } l\neq k.\end{cases}$$

See [19, sections 2.2, 4.3] for more discussion on this network.

Consider another modification to the network in section 2, where general probabilistic routing is allowed, but a customer arriving at a full buffer is lost. Therefore, the network is a generalized Jackson network [40] except that a customer arriving at a full buffer station is lost. It can be shown that the heavy traffic limit theorem in theorem 3.1 holds with

$$R = \big(I - P', -I\big),$$

and Γ given in (9.1).

Acknowledgements

We thank Tom Kurtz for providing the proof of lemma 8.4. We thank John Hasenbein, Takis Konstantopoulos, Ruth Williams and two referees for suggesting various improvements to the paper.

References

[1] I. Bardhan and S. Mithal, Heavy-traffic limits for an open network of finite-buffer overflow queues: The single class case, preprint (1993).
[2] A. Bernard and A. El Kharroubi, Régulations déterministes et stochastiques dans le premier "orthant" de $\mathbb{R}^n$, Stochastics Stochastics Rep. 34 (1991) 149–167.
[3] D. Bertsekas and R. Gallagher, *Data Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1992).
[4] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).
[5] M. Bramson, Convergence to equilibria for fluid models of FIFO queueing networks, Queueing Systems 22 (1996) 5–45.
[6] M. Bramson, Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks, Queueing Systems 23 (1997) 1–26.
[7] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks, Queueing Systems 30 (1998) 89–148.
[8] A. Brondstred, *An Introduction to Convex Polytopes* (Springer, New York, 1983).
[9] J.A. Buzacott, Automatic transfer lines with buffer stocks, Internat. J. Prod. Res. 5 (1967) 183–200.
[10] H. Chen and H. Zhang, Stability of multiclass queueing networks under FIFO service discipline, Math. Oper. Res. 22 (1997) 691–725.
[11] H. Chen and H. Zhang, Diffusion approximations for multiclass FIFO queueing networks, preprint.
[12] D.W. Cheng, Second order properties in a tandem queue with general blocking, Oper. Res. Lett. 12 (1992) 139–144.
[13] D. Cheng and D.D. Yao, Tandem queues with general blocking: A unified model and comparison results, Discrete Event Dyn. Systems 2 (1993) 207–234.
[14] K.L. Chung and J.R. Williams, *Introduction to Stochastic Integration* (Birkhäuser, Boston, 1983).
[15] J.G. Dai and J.M. Harrison, Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application, Ann. Appl. Probab. 1 (1991) 16–35.
[16] J.G. Dai and T.G. Kurtz, A multiclass station with Markovian feedback in heavy traffic, Math. Oper. Res. 20 (1995) 721–742.
[17] J.G. Dai, G. Wang and Y. Wang, Nonuniqueness of the Skorohod problem arising from FIFO Kelly type network, private communication (1992).

[18] J.G. Dai and R.J. Williams, Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons, Theory Probab. Appl. 40 (1995) 3–53.
[19] W. Dai, Brownian approximations for queueing networks with finite buffers: modeling, heavy traffic analysis and numerical implementations, Ph.D. thesis, School of Mathematics, Georgia Institute of Technology (1996).
[20] A.I. Elwalid and D. Mitra, Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulation, Queueing Systems 9 (1991) 29–64.
[21] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence* (Wiley, New York, 1986).
[22] L.M. Graves, *The Theory of Functions of Real Variables* (McGraw-Hill, New York, 1956).
[23] A. Gut, *Stopped Random Walks: Limit Theorems and Applications* (Springer, Berlin, 1988).
[24] J.M. Harrison, Brownian models of queueing networks with heterogeneous customer populations, in: *Proc. of the IMA Workshop on Stochastic Differential Systems* (Springer, Berlin, 1988).
[25] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic I, Adv. in Appl. Probab. 2 (1970) 150–177.
[26] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic II, Adv. in Appl. Probab. 2 (1970) 355–364.
[27] D.P. Johnson, Diffusion approximations for optimal filtering of jump processes and for queueing networks, Ph.D. thesis, University of Wisconsin (1983).
[28] T. Konstantopoulos and J. Walrand, On the ergodicity of networks of $\cdot/GI/1/N$ queues, Adv. in Appl. Probab. 22 (1990) 263–267.
[29] H. Kroner, M. Eberspacher, T.H. Theimer, P.J. Kuhn and U. Briem, Approximate analysis of the end to end delay in ATM networks, in: *Proc. of the IEEE INFOCOM '92*, Florence, Italy (1992) pp. 978–986.
[30] G. Last and A. Brandt, *Marked Point Processes on the Real Line: The Dynamic Approach* (Springer, New York, 1995).
[31] D. Mitra and I. Mitrani, Analysis of a Kanban discipline for cell coordination in production lines: I, Managm. Sci. 36 (1990) 1458–1566.
[32] H. Perros and T. Altiok, Queueing networks with blocking: A bibliography, Performance Evaluation Rev.: ACM Sigmetrics 12 (1984) 8–12.
[33] W.P. Peterson, A heavy traffic limit theorem for networks of queues with multiple customer types, Math. Oper. Res. 16 (1991) 90–118.
[34] M.I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (1984) 441–458.
[35] M.I. Reiman, A multiclass feedback queue in heavy traffic, Adv. in Appl. Probab. 20 (1988) 179–207.
[36] M.I. Reiman and R.J. Williams, A boundary property of semimartingale reflecting Brownian motions, Probab. Theory Related Fields 77 (1988) 87–97 and 80 (1989) 633.
[37] L.M. Taylor and R.J. Williams, Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant, Probab. Theory Related Fields 96 (1993) 283–317.
[38] R.J. Williams, An invariance principle for semimartingale reflecting Brownian motions in an orthant, Queueing Systems 30 (1998) 5–25.
[39] R.J. Williams, Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse, Queueing Systems 30 (1998) 27–88.
[40] D.D. Yao, *Probability Models in Manufacturing Systems*, Springer Series in Operations Research (Springer, Berlin, 1994).