Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse[1]

R. J. Williams
Department of Mathematics
University of California, San Diego
La Jolla CA 92093-0112

ABSTRACT

Certain diffusion processes known as semimartingale reflecting Brownian motions (SRBMs) have been shown to approximate many single class and some multiclass open queueing networks under conditions of heavy traffic. While it is known that not all multiclass networks with feedback can be approximated in heavy traffic by SRBMs, one of the outstanding challenges in contemporary research on queueing networks is to identify broad categories of networks that can be so approximated and to prove a heavy traffic limit theorem justifying the approximation. In this paper, general sufficient conditions are given under which a heavy traffic limit theorem holds for open multiclass queueing networks with head-of-the-line (HL) service disciplines, which in particular require that service *within* each class is on a first-in-first-out (FIFO) basis. The two main conditions that need to be verified are that (a) the reflection matrix for the SRBM is well defined and completely-$\mathcal{S}$, and (b) a form of state space collapse holds. A result of Dai and Harrison shows that condition (a) holds for FIFO networks of Kelly type and their proof is extended here to cover networks with the HLPPS (head-of-the-line proportional processor sharing) service discipline. In a companion work, Bramson shows that a multiplicative form of state space collapse holds for these two families of networks. These results, when combined with the main theorem of this paper, yield new heavy traffic limit theorems for FIFO networks of Kelly type and networks with the HLPPS service discipline.

Keywords: Multiclass queueing networks, heavy traffic, FIFO Kelly type, head-of-the-line-proportional processor sharing, semimartingale reflecting Brownian motions, diffusions, completely-$\mathcal{S}$.

AMS 1991 Subject Classification: 60F17, 60K25, 60J70, 68M20, 90B22.

NOTE TO THE COPYEDITOR AND TYPESETTER: *The table of contents is essential to this paper and should not be deleted. In addition, the numbering of Sections, Definitions, Assumptions, Lemmas, Propositions and Theorems needs to be kept as it is since these items are referenced by number in this paper and in another paper submitted to the same volume of QUESTA.*

[1]Research supported in part by NSF grants GER 9023335 and DMS 9703891.

Contents

1 Introduction

Queueing network models are of current interest for analyzing congestion and delay in computer systems, communication networks and complex manufacturing systems (see e.g., [3, 54]). Many of these systems have stations that can process more than one class of customer or job (so-called multiclass networks) and/or have complex feedback structures. Heavily loaded networks, where congestion is a compelling problem, are of particular interest. Frequently, exact analysis of such networks is unavailable and it is natural to seek tractable approximations. In connection with this, certain diffusion processes, known as semimartingale reflecting Brownian motions (SRBMs), have been proposed as approximations for normalized versions of the queue length or workload processes in heavily loaded networks [28, 30]. Heavy traffic limit theorems justifying these approximations have been proved for many single class and some multiclass queueing networks (see [52] for a summary of such results). However, it is known that not all multiclass networks with feedback can be approximated in heavy traffic by SRBMs (cf. [22, 51, 21, 33]), and one of the outstanding challenges in contemporary research on queueing networks is to identify broad categories of networks that can be so approximated and to prove a heavy traffic limit theorem justifying the approximation.

In this paper, general sufficient conditions are given under which a heavy traffic limit theorem holds for open multiclass queueing networks with head-of-the-line (HL) service disciplines, which in particular require that customers *within* each class are served on a first-in-first-out (FIFO) basis. For open networks, the associated SRBMs live in the positive orthant of a Euclidean space. Besides the usual functional central limit theorem assumptions and independence assumptions on the primitive processes for our queueing network model, the two main conditions of our heavy traffic limit theorem are that

(a) the reflection matrix for the limit SRBM is well defined and completely-$\mathcal{S}$,

(b) a form of state space collapse holds.

The definition of a completely-$\mathcal{S}$ reflection matrix is given in Section 6. This condition is known to be necessary and sufficient for the existence and uniqueness of an SRBM. State space collapse is defined in Section 4 and its key role in the proof of our heavy limit theorem is explained below. For certain service disciplines, e.g., FIFO (across all classes at a station), state space collapse is necessary for a heavy traffic limit theorem to hold (see Appendix B).

The two key features of our proof of the heavy traffic limit theorem are

(i) the assumption of *state space collapse* is used to approximate the normalized high dimensional queue length process at each station by a constant vector times the normalized workload process for the station,

(ii) an *invariance principle* (or perturbation result) for SRBMs and condition (a) are used to turn the independence and functional central limit theorem assumptions for the normalized external arrival, service time, and routing processes, into weak convergence of the normalized workload processes to an SRBM.

State space collapse is critical to our treatment of *multiclass* networks. This *phenomenon* was first established by Whitt [50] for the case of a single multiclass station with a static priority discipline, whereas the *term* was first used in the later work of Reiman [45, 46], which was also concerned with a single station. Verification of state space collapse is a key feature of the heavy traffic limit theorem of Peterson [42] which applies to feedforward multiclass queueing networks with a preemptive resume static priority service discipline. In our limit theorem, state space collapse is an assumption, which would need to be verified in any particular application. In connection with this, in a companion work, Bramson [11] has shown that a multiplicative form of state space collapse holds for two interesting families of multiclass networks. In Proposition 8.1 of this paper, it is shown that this multiplicative form implies regular state space collapse under our heavy traffic conditions and condition (a), both of which are assumptions for our heavy traffic limit theorem.

Most prior proofs of heavy traffic limit theorems for open queueing networks, such as those of Reiman [44] for single class FIFO networks, of Peterson [42] for feedforward multiclass networks with preemptive resume static priorities, and of Chen and Zhang [15] for re-entrant lines with a first-buffer-first-served priority service discipline, used a stronger and more restrictive result than the invariance principle referred to in (ii) above. Those proofs relied on the existence and uniqueness of a continuous path-to-path mapping which could be applied to obtain an SRBM path as the continuous image of a Brownian motion path (cf. [31, 26]). A major difficulty in generalizing those proofs to multiclass networks with feedback is that uniqueness does not always hold for such a path-to-path mapping [2, 41], even when existence and uniqueness *in law* of an SRBM is known. Indeed, Dai, Wang and Wang [23] have given an example of a FIFO network of Kelly type for which the continuous mapping argument used in [42, 44] cannot be readily extended. (The details of this example are given in Appendix A of this paper.)

To overcome this difficulty, in [53], this author established an invariance principle (or perturbation result) for SRBMs which could take the place of the continuous mapping argument used by Reiman [44], Peterson [42], and Chen and Zhang [15]. This result is based on a distributional characterization of an SRBM which involves a certain martingale property. Accordingly, in establishing our heavy traffic limit theorem, a martingale property of the queueing network model needs to be established, and this in turn requires that a stopping time property be verified. For our proof of the martingale property (cf. Lemma 8.4), independence assumptions are imposed between and within the (external) interarrival time, service time and routing sequences (cf. Section 3.1.6). In particular, this independence is stronger than what is needed for a conventional continuous mapping argument, where the main condition is that a functional central limit theorem holds for these primitive processes. Though it is possible that this independence assumption could be relaxed, this would require a more general approach than that used here. The stopping time property (cf. Lemma 8.3) is expected to hold for a wide variety of non-anticipating service disciplines. However, to give a rigorous proof of this property, one needs to formalize the notion of a service discipline. The stopping time property is verified here for so-called *head-of-the-line* (HL) service disciplines (our notion of an HL service discipline is a little more general than that used by Bramson in [10, 11]). An HL service discipline requires that service within each class is on a FIFO basis (that is, service for each class is concentrated on the customer at the head-of-the-line for the class) and each class receives a proportion (possibly zero) of the associated server's time, where this proportion may be random but it is kept constant between changes in the arrival and departure processes. Furthermore, these proportions should depend in a measurable way on the "state" of the queueing network, and they should not anticipate (external) interarrival times, service times or routing vectors for future arrivals (our measurability and dependence assumptions are made precise in Section 3.1.5). Common service disciplines which satisfy the HL requirement are FIFO (across all classes at a station), preemptive resume static priorities and HLPPS (head-of-the-line proportional processor sharing). Static priority policies without preemption could be included in an extension of our model. However, to reduce the complexity of the stopping time argument, that case has not been treated here (see Section 3.1.5 for more discussion).

We note here that in recent work Chen and Zhang [16] have given a heavy traffic limit theorem for a multiclass FIFO network with a restrictive spectral radius condition on the workload contents matrix (denoted by G here). They do not use a continuous mapping argument and in a sense they implicitly verify (i) and (ii). Their approach to proving the stopping time property is different from that used here. Furthermore, they only consider a FIFO service discipline. The main result stated in Chen and Zhang [16] is neither subsumed by nor subsumes the results indicated below for FIFO networks of Kelly type. It remains an open problem to identify collections of FIFO networks that naturally satisfy their spectral radius condition.

As applications of the main theorem of this paper, the results of Bramson [11] on multiplicative state space collapse are combined with a verification of condition (a) (see Dai and Harrison [20] for FIFO networks of Kelly type and Section 9 for an extension to HLPPS networks), to yield new heavy traffic limit theorems for FIFO networks of Kelly type and HLPPS networks. In a FIFO network, customers are served in the order of their arrival at each station, without regard to their class designation. A *Kelly type* network is such that the mean service time is the same for all classes at a given station. (In fact, Bramson works with a mild generalization of Kelly type networks, namely those that are asymptotically of Kelly type. Our heavy traffic limit theorem also holds for such networks with a FIFO service discipline. However, to simplify terminology in the text, we shall simply use the term Kelly type until we come to the precise statement of the relevant heavy traffic limit theorem or to proofs of related results, see e.g., Lemma 9.2 and Theorem 10.1.) In an HLPPS network, all non-empty classes present at a station are served simultaneously, where the fraction of time spent on a given class is proportional to the number of customers present in that class, and all of this service goes to the customer at the head of the line of the class. When the service times are exponentially distributed, the queue length process for an HLPPS network is equivalent in distribution to that for a network having the usual processor sharing discipline in which all customers at a station receive an equal proportion of the server's time. Thus, HLPPS networks have some similarities with processor sharing and FIFO networks, but they turn out to be easier to work with for the purpose of verifying state space collapse.

This paper is organized as follows. Some preliminary notation is established in Section 2. Our model for an open queueing network is described in Section 3. Although much of the setup described there applies for arbitrary non-idling service disciplines, our heavy traffic limit theorem is only proved for head-of-the-line (HL) service disciplines. Thus, from Section 3.1.5 onwards we restrict to such service disciplines which include FIFO (across all classes at a station), preemptive resume static priority and HLPPS policies. Fluid and diffusion scaling for the primitive network processes and performance processes are defined and applied to equations satisfied by our network model in Section 4. Assuming state space collapse and invertibility of a certain data matrix, these equations are then reduced to the form given by equations (74)–(77). Our heavy traffic assumptions, in particular, conditions which imply a functional central limit theorem for the primitive external arrival, service time, and routing processes, are specified in Section 5. The precise definition of a semimartingale reflecting Brownian motion in an orthant (SRBM) is given in Section 6. The main result of this paper, namely, the statement of general sufficient conditions which imply a heavy traffic limit theorem for open multiclass queueing networks with HL service disciplines, is given in Section 7. This result is proved in Section 8. In Section 9, Assumption 7.1 (or equivalently, condition (a) above) on the data matrices for the queueing networks is verified to hold for FIFO networks of Kelly type and HLPPS networks. This was originally shown by Dai and Harrison [20] for FIFO networks of Kelly type. Their proof is repeated in Section 9 for completeness and to facilitate the extension made there to HLPPS networks. In Section 10, the results of this paper are combined with those of Bramson [11] to yield new heavy traffic limit theorems for FIFO networks of Kelly type and HLPPS networks.

As a possible roadmap for reading this paper, we suggest the reader start with Section 3, to become familiar with our queueing network model and the associated performance processes. After reviewing the definitions of the diffusion scaled processes in Section 4, the reader may then turn to Section 10 where the main theorem of this paper is applied, in combination with the results of Bramson [11], Dai-Harrison [20] and Section 9, to yield new heavy traffic limit theorems for two families of open multiclass networks. In order to fully appreciate the meaning of the results stated there, the reader will need to consult Section 2 for notation related to weak convergence, the heavy traffic assumptions of Section 5, and the definition of an SRBM given in Section 6. However, knowing the statements of the results should guide the reader in focussing on what is essential for their understanding. The reader will note that the proofs of these two results in Section 10 are quite short. This is no accident. On consulting the main theorem of the paper, Theorem 7.1, the reader will see that this theorem is set up in such a way that a heavy traffic limit theorem is immediate once Assumptions 7.1 and 7.2 (i.e., (a) and (b) above) have been verified. For the two applications given in Section 10, Assumption 7.1 on the data matrices follows from Dai-Harrison [20] and Section 9, and Assumption 7.2 on multiplicative state space collapse is shown to hold in Bramson [11]. Finally, the reader will want to turn to Sections 7 and 8 where the main theorem of this paper is stated and proved. At this stage the reader will need to consult Section 4 in more detail.

2 Notation

The set of non-negative integers will be denoted by $\mathbb{N}$. For each positive integer n, the n-dimensional Euclidean space will be denoted by $\mathbb{R}^n$ and the n-dimensional positive orthant will be denoted by $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x_i \geq 0,\ i = 1, \ldots, n\}$. When $n = 1$, the superscript will be suppressed on $\mathbb{R}^1$ and $\mathbb{R}^1_+$. Vectors will be column vectors unless indicated otherwise. Inequalities between vectors in $\mathbb{R}^n$ should be interpreted componentwise. For $a, b \in \mathbb{R}^n$, we shall use $a \vee b$ to denote the vector whose i^{th} component is the maximum of a_i and b_i for $i = 1, \ldots, n$. Similarly, $a \wedge b$ will denote the componentwise minimum of a and b. The superscript $'$ will be used to denote the operation of taking the transpose of a vector or matrix. For $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$, we will use the norm $|x| = \max_{i=1}^n |x_i|$, and for the norm of an $n \times l$ matrix A we will use $|A| = \max_{i=1}^n \sum_{j=1}^l |A_{ij}|$. For a vector $x \in \mathbb{R}^n$, the $n \times n$ diagonal matrix whose diagonal entries are given by the components of x will be denoted by $\text{diag}(x)$.

The notation $I\!P$ will be used to denote a probability measure and $I\!E$ will be used to denote the expectation relative to $I\!P$. A triple $(\Omega, \mathcal{F}, \{\mathcal{F}_t, t \geq 0\})$ will be called a filtered space if Ω is a set, $\mathcal{F}$ is a σ-field of subsets of Ω, and $\{\mathcal{F}_t, t \geq 0\}$ is an increasing family of sub-σ-fields of $\mathcal{F}$, i.e., a filtration. If in addition, $I\!P$ is a probability measure on $(\Omega, \mathcal{F})$, then $(\Omega, \mathcal{F}, \{\mathcal{F}_t, t \geq 0\}, I\!P)$ is called a filtered probability space. A filtration $\{\mathcal{F}_t, t \geq 0\}$ will often be simply written as $\{\mathcal{F}_t\}$. Given a fixed filtration, an n-dimensional process will be called a martingale (relative to the filtration) if and only if each component is a martingale (relative to the filtration).

For each positive integer n, let $\mathbf{D}^n$ be the space of "Skorokhod paths" in $\mathbb{R}^n$ having time domain $[0, \infty)$. That is, $\mathbf{D}^n$ is the set of all functions $w : [0, \infty) \to \mathbb{R}^n$ that are right continuous on $[0, \infty)$ and have finite left limits on $(0, \infty)$. We shall use the abbreviation r.c.l.l. for "right continuous with finite left limits". The identically zero path in $\mathbf{D}^n$ will be denoted by $\mathbf{0}$. For $w \in \mathbf{D}^n$ and $T \geq 0$, we let

$$\|w\|_T = \sup_{t \in [0,T]} |w(t)|. \tag{1}$$

Consider $\mathbf{D}^n$ to be endowed with the usual Skorokhod $\mathbf{J}_1$-topology (see Skorokhod [48] and Ethier and Kurtz [27]). Let $\mathcal{M}^n$ denote the Borel σ-algebra on $\mathbf{D}^n$ associated with this

Skorokhod topology. This is the same as the σ-algebra generated by the coordinate maps, i.e., $\mathcal{M}^n = \sigma\{w(s) : 0 \leq s < \infty\}$. Each continuous-time (stochastic) process in this paper will be assumed to have paths in $\mathbf{D}^n$ for a suitable value of n. That is, such a process is a measurable function from some probability space $(\Omega, \mathcal{F}, I\!\!P)$ into $(\mathbf{D}^n, \mathcal{M}^n)$. To indicate the dependence on n, we shall call such a process an n-dimensional process. Consider n-dimensional processes $W^1, W^2, \ldots,$ and W. The filtration generated by W is taken to be $\{\mathcal{F}_t, t \geq 0\}$ where $\mathcal{F}_t = \sigma\{W(s) : 0 \leq s \leq t\}$ for all $t \geq 0$. The sequence $\{W^i\}_{i=1}^{\infty}$ is said to be tight if and only if the probability measures induced by the W^i on $(\mathbf{D}^n, \mathcal{M}^n)$ form a tight sequence. The notation "$W^i \Rightarrow W$" will mean that the probability measures induced by the W^i on $(\mathbf{D}^n, \mathcal{M}^n)$ converge weakly to the probability measure induced on $(\mathbf{D}^n, \mathcal{M}^n)$ by W as $i \to \infty$; this same state of affairs may be expressed by the statement "W^i converges in distribution to W as $i \to \infty$". For more on tightness and convergence in distribution of processes taking values in $\mathbf{D}^n$, see Chapter 3 of Ethier and Kurtz [27].

3 Open multiclass queueing networks: the model

In this section, our model for an open multiclass queueing network is defined. This model is a variant of that used by Harrison and Nguyen [30] for open multiclass networks with a first-in-first-out (FIFO) service discipline. In particular, different service disciplines besides FIFO are allowed here and model elements incorporate information about residual service times, and order and age of customers present in the network at time zero. The latter are important for our treatment of networks that are initially non-empty. (For simplicity, prior limit theorems have frequently assumed that the networks are initially empty, see e.g., [44, 42]. For multiclass networks, this is a non-trivial assumption, see [29] for related discussion.) Also, rather than considering a single collection of primitive processes with accelerated long run average arrival rates, as is done in Peterson [42] and Harrison and Nguyen [30], we revert to the more general framework of Iglehart and Whitt [34, 35] and Reiman [44], where a family of networks is considered, each member of which may have a different collection of primitive processes. In short, in the terminology of central limit theorems, a triangular array of primitive (external) interarrival time, service time and routing sequences is considered. We have chosen this more general setup because a heavy traffic limit theorem in this context implies a certain uniformity or robustness of the diffusion approximation to small perturbations in the distributions of the (external) interarrival times, service times and routing sequences. Our description of the initial data is a little different from that used in Bramson [11], where the focus is on establishing state space collapse. The correspondence between our setup and that used by Bramson is described at the end of this section.

Here a queueing network will be taken to consist of a fixed finite set of (service) stations through which customers (or jobs) are processed. A customer visits stations in a sequential manner and receives service at each station visited. In an open network, each customer arrives from outside the network, visits a finite number of stations and then exits the network. The route of a customer is the ordered sequence of stations that the customer visits during its sojourn in the network. It is assumed that for each station there is an infinite buffer to store customers awaiting service there. Each customer at a station belongs to one of a finite number of classes. The mapping from classes to stations is many-to-one and customers change class as they go from one station to the next. Customers in different classes may be distinguished for example by their (external) interarrival time or service time distributions, routing protocols or treatment under the service discipline.

3.1 Model primitives

Our description of the primitive elements for our open multiclass queueing network model is broken down into assumptions concerning network structure, external arrivals, service times, routing, service discipline, initial conditions, and independence. Given these model primitives, the evolution of the queueing network is determined as that of a discrete event system. In particular, the descriptive processes Z, W, Y, A, D, introduced in Section 3.2 below, are determined from this.

3.1.1 Network structure

We consider an open multiclass queueing network consisting of J ($1 \leq J < \infty$) stations with a single perfectly reliable server at each station and K ($J \leq K < \infty$) customer classes for the entire network. For convenience, let $\mathcal{J} = \{1, \ldots, J\}$ and $\mathcal{K} = \{1, \ldots, K\}$. The many-to-one mapping from customer classes to stations is described by a $J \times K$ *constituency matrix* C where for $j \in \mathcal{J}$, $k \in \mathcal{K}$,

$$C_{jk} = \begin{cases} 1 & \text{if class } k \text{ is served at station } j, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

For $j \in \mathcal{J}$, let $\mathcal{C}(j)$ denote the *constituency of server* j, i.e., $\mathcal{C}(j) = \{k \in \mathcal{K} : C_{jk} = 1\}$. We assume that $\mathcal{C}(j) \neq \emptyset$ for each j. For $k \in \mathcal{K}$, let $s(k)$ denote the station at which class k is served, i.e., $s(k)$ is the unique $j \in \mathcal{J}$ such that $C_{jk} = 1$.

3.1.2 External arrivals

We assume that there is a K-dimensional *external arrival process* $E = \{E(t), t \geq 0\}$ such that for each class $k \in \mathcal{K}$, $E_k(t)$ is the number of arrivals to class k from outside the network that have occurred by time t. We may have $E_k(\cdot) \equiv \mathbf{0}$ for some k. Let $\mathcal{A} = \{k \in \mathcal{K} : E_k(\cdot) \not\equiv \mathbf{0}\}$, the set of classes that have some external arrivals. We assume that $\mathcal{A}$ is non-empty. For each $k \in \mathcal{A}$, E_k is assumed to be defined from a sequence of independent random variables $\{u_k(i), i = 1, 2, \ldots\}$, where for $i = 2, 3, \ldots$, $u_k(i)$ denotes the time between the $(i-1)^{\text{st}}$ and the i^{th} external arrival of a class k customer, and $u_k(1)$ denotes the first (residual) interarrival time that is equal to the time measured from zero until the first external arrival to class k. It is assumed that $\{u_k(i), i = 2, 3, \ldots\}$ is a sequence of strictly positive i.i.d. random variables, each with mean $1/\alpha_k \in (0, \infty)$ and variance $a_k \in [0, \infty)$. The residual interarrival time $u_k(1)$ is assumed to be independent of $\{u_k(i), i = 2, 3, \ldots\}$ and to be strictly positive, but is otherwise arbitrary. For $k \in \mathcal{A}$, setting $U_k(0) = 0$ and

$$U_k(n) = \sum_{i=1}^{n} u_k(i), \quad \text{for } n = 1, 2, \ldots, \tag{3}$$

we have

$$E_k(t) = \max\{n \geq 0 : U_k(n) \leq t\} \quad \text{for all } t \geq 0. \tag{4}$$

Thus, E_k is a (delayed) renewal process. We interpret α_k as the long run average external arrival rate for class k. For convenience, we define $\alpha_k = 0, a_k = 0$ for $k \notin \mathcal{A}$, and we let $E_{\mathcal{A}}, U_{\mathcal{A}}$ denote the $|\mathcal{A}|$-dimensional processes whose components are given by E_k, U_k, respectively, for $k \in \mathcal{A}$.

3.1.3 Service times

For each $k \in \mathcal{K}$, we assume that there are two sequences of service times for the class k customers. A first sequence of random variables $\{v_k^o(i), i = 1, 2, \ldots\}$ gives the (residual) service times for those class k customers who are *originally* at station $s(k)$ at time zero. (There are more elements in this infinite sequence than needed, but there is no loss of generality in this and having an infinite sequence makes for a more convenient mathematical setup.) These "original" customers are assumed to have an ordering corresponding to (fictitious) arrival times occurring at or before time zero. For instance the first "original" class k customer is the one who arrived the longest time ago, the second customer in class k is the next original class k customer to have arrived, and so on. If there is an i^{th} customer of class k in the network at time zero, then $v_k^o(i)$ is interpreted as the (residual) service time for that customer, i.e., it is the amount of service time (as measured from time zero) that the customer still needs to receive before it can depart station $s(k)$. We define an *original cumulative service time process* for class k by

$$V_k^o(n) = \sum_{i=1}^{n} v_k^o(i), \quad n = 1, 2, \ldots, \tag{5}$$

and for convenience we define $V_k^o(0) = 0$. If Z_k^o is a random variable denoting the number of customers in class k at time zero, then we assume that $v_k^o(i) > 0$ for $1 \leq i \leq Z_k^o$. Further essential distributional assumptions for the sequences $\{v_k^o(i), i = 1, 2, \ldots\}$, $k \in \mathcal{K}$, will be embodied in the independence assumptions of Section 3.1.6 and the heavy traffic assumptions of Section 5.

A second sequence of random variables $\{v_k^s(i), i = 1, 2, \ldots\}$ gives the service times for the *subsequent* customers arriving into class k after time zero. The time $v_k^s(i)$ is the amount of service time at station $s(k)$ required by the i^{th} customer who arrives into class k after time zero. Thus, these service times are ordered according to the times of arrival of the associated class k customers. It is assumed that $\{v_k^s(i), i = 1, 2, \ldots\}$ is a sequence of strictly positive i.i.d. random variables, each with finite mean $m_k > 0$ and variance $b_k \in [0, \infty)$. We let M denote the $K \times K$ diagonal matrix with m_k as the k^{th} diagonal element. Let $\mu_k = \frac{1}{m_k}$, which is the long run average rate at which class k customers would be served if the server $s(k)$ were never idle and the server devoted all of its attention to class k. We define the *subsequent cumulative service time process* for class k by

$$V_k^s(n) = \sum_{i=1}^{n} v_k^s(i), \quad n = 1, 2, \ldots, \tag{6}$$

and for convenience we define $V_k^s(0) = 0$.

For a multi-index $n = (n_1, \ldots, n_K)$ where each n_k, $k \in \mathcal{K}$, is a non-negative integer, we let

$$V^o(n) = (V_1^o(n_1), \ldots, V_K^o(n_K))', \tag{7}$$

and we define $V^s(n)$ similarly.

3.1.4 Routing

We assume Markovian routing which can be described as follows.

Let $e_1, \ldots, e_K$ be the unit basis vectors parallel to the K coordinates axes in $\mathbb{R}^K$ and let e_0 be the K-dimensional vector of all zeros. For each class $k \in \mathcal{K}$, $\{\phi^k(i), i = 1, 2, \ldots\}$ is a sequence of i.i.d. *routing vectors* where $\phi^k(i)$ takes values in the set $\{e_0, e_1, \ldots, e_K\}$. The interpretation of the routing vectors $\phi^k(i)$ is that the i^{th} class k customer to depart from station $s(k)$ is next routed to class l if $\phi^k(i) = e_l$ for some $l \in \mathcal{K}$, or it leaves the network if $\phi^k(i) = e_0$.

Let $P_{kl} = \mathbb{P}(\phi^k(i) = e_l)$, $k \in \mathcal{K}$, $l \in \mathcal{K}$. The $K \times K$ matrix P with components P_{kl} is called the *routing matrix.* To simplify later notation involving other superscripts, we let $\tilde{P}$ denote P', the transpose of the matrix P. To satisfy our *open* queueing network assumption, the matrix P, or equivalently $\tilde{P}$, is assumed to have spectral radius strictly less than one. Hence

$$Q \equiv (I - \tilde{P})^{-1} = I + \tilde{P} + (\tilde{P})^2 + (\tilde{P})^3 + \ldots \tag{8}$$

is well defined, where $(\tilde{P})^n$ denotes the n^{th} power of $\tilde{P}$. The matrix entry Q_{lk} is interpreted as the average number of visits to class l made by a customer who starts in class k.

Note that for $k \in \mathcal{K}$,

$$\mathbb{E}[\phi^k(i)] = \tilde{P}^k \quad \text{and} \quad Cov[\phi^k(i)] = \Upsilon^k, \tag{9}$$

where $\tilde{P}^k$ denotes the k^{th} column of $\tilde{P}$ and Υ^k is the $K \times K$ matrix defined by

$$\Upsilon_{lm}^k = \begin{cases} P_{kl}(1 - P_{kl}) & \text{if } l = m, \\ -P_{kl}P_{km} & \text{if } l \neq m. \end{cases} \tag{10}$$

(The reader should not confuse $\tilde{P}^k$ with the k^{th} power of $\tilde{P}$, for we shall always write the latter as $(\tilde{P})^k$.) For each $k \in \mathcal{K}$, we define a K-dimensional *cumulative routing process* for class k by

$$\Phi^k(n) = \sum_{i=1}^{n} \phi^k(i), \quad n = 1, 2, \ldots, \tag{11}$$

and for convenience we define $\Phi^k(0) = 0$. We let $\Phi = (\Phi^1, \ldots, \Phi^K)$.

3.1.5 Service discipline and initial conditions

A service discipline is a *non-idling* policy if a server is never idle when there are customers waiting to be served at its station. (Some authors refer to this as a work conserving policy.) We only consider non-idling policies. Furthermore, while the general model described so far could apply to non-idling policies such as last-in-first-out (LIFO) and processor sharing (PS), our heavy traffic limit theorem is only proved for *head-of-the-line* (HL) service disciplines. Although LIFO and PS are excluded, the collection of HL disciplines does include such common service disciplines as FIFO (across classes), preemptive resume static priorities and head-of-the-line processor sharing, in particular, it includes the HLPPS (head-of-the-line proportional processor sharing) service discipline.

HL service disciplines. The rough description of an HL service discipline given in the introduction will be made more precise here. An HL service discipline is a non-idling policy and service within each class is on a FIFO (first-in-first-out) basis. Each class receives a proportion (possibly zero) of the associated server's time, where this proportion may be random but it is kept constant between changes in the arrival or departure processes. To further describe an HL service discipline, we need to specify the information about the state of the queueing network that may be used to update the proportions of the server's time allocated to the various classes when a new arrival or departure occurs in the network.

For this, we define a strictly increasing sequence $\{\sigma_\ell\}_{\ell=0}^{\infty}$ that specifies the successive times at which an arrival occurs to, or a departure occurs from, some class in the network. (The sequence $\{\sigma_\ell\}_{\ell=0}^{\infty}$ is strictly increasing and so more than one arrival or departure may occur at a given σ_ℓ.) More precisely, we define $\sigma_0 = 0$ and σ_ℓ (for $\ell \geq 1$) as the time of the ℓ^{th} change in the number of arrivals to, or the number of departures from, some class in the network. (For a constructive approach, one should define the σ_ℓ inductively along with the proportions r^ℓ associated with the description of the service discipline below. In this, for each ℓ, σ_ℓ would be defined first and then r^ℓ would be defined. While it is clear that this could be done (cf. Lemma 8.3), here we take a descriptive viewpoint and leave a formal inductive construction to the interested reader.) Using the fact that at least one class has a non-zero external arrival process, it can be shown that $\sigma_\ell < \infty$ a.s. for each ℓ, and furthermore, using the a priori bound (192) on the arrival process it can be shown that $\sigma_\ell \to \infty$ a.s. as $\ell \to \infty$. To avoid the need to always mention the exceptional null sets where these properties may not hold, we assume that such exceptional null sets have been removed from the probability space a priori. There is no loss of generality in this, since our main result, Theorem 7.1, is a statement about convergence in distribution.

For each $\ell \in \mathbb{N}$, $l \in \mathcal{A}$, and $k \in \mathcal{K}$, let Z_k^ℓ denote the number of class k customers that are in queue or being served at the time σ_ℓ, let u_l^ℓ denote the residual interarrival time for class l as measured from σ_ℓ, (i.e., u_l^ℓ is the amount of time remaining after σ_ℓ until the next external arrival to class l), and let $v_k^\ell(i)$ for $1 \le i \le Z_k^\ell$ denote the residual service time for the i^{th} customer present in class k at the time σ_ℓ (i.e., the amount of server $s(k)$'s time still required by the i^{th} customer present in class k at σ_ℓ). Here customers in class k are ordered according to their times of arrival into the class, with the first customer being the one who arrived the longest time ago, etc. A tie breaking rule (see below) can be invoked for simultaneous arrivals to a class. For $\ell = 0$, recalling the definitions in Sections 3.1.2 and 3.1.3 of the original queuelength, residual interarrival times and residual service times, we have $Z_k^0 = Z_k^o$ is the number of customers initially in class k, $u_l^0 = u_l(1)$, and $v_k^0(i) = v_k^o(i)$, for $i = 1, \ldots, Z_k^0$. For convenience, we extend the identity $v_k^0(i) = v_k^o(i)$ to $i > Z_k^0$, so that this equality holds for all i. (The times $v_k^0(i), v_k^o(i)$ for $i > Z_k^0$, are not used in specifying the queueing network model, they are simply included for the mathematical convenience of avoiding sequences of random length.) Similarly and even more simply, we set $v_k^\ell(i) = 0$ for all $i > Z_k^\ell$ and $\ell \ge 1$. (We did not make this simplistic assumption of a zero value for the case $\ell = 0$, because the more general structure of our sequence of original residual service times allows us to easily explain the correspondence with Bramson's model, which differs slightly in its treatment of the initial data.) For certain service disciplines, such as FIFO and preemptive resume static priorities, additional information is needed to specify the action of the service discipline. For uniformity, we include this with the information for all HL service disciplines. For each station $j \in \mathcal{J}$, let $\mathcal{Z}_j^\ell = \sum_{k \in \mathcal{C}(j)} Z_k^\ell$. Now, consider the $\mathcal{Z}_j^\ell$ customers that are at station j at σ_ℓ to be ordered according to the times at which they arrived at the station, where the first customer is the one who arrived the longest time ago, etc. A tie breaking rule (see below) is invoked to determine the ordering of customers who arrive at a station at the same time. For each $\ell \in \mathbb{N}$ and $j \in \mathcal{J}$, the composition of the customer population at station j at time σ_ℓ is recorded in a sequence $\{O_j^\ell(i)\}_{i=1}^\infty$ of (class, age) pairs. More precisely, $\{O_j^\ell(i)\}_{i=1}^\infty$ is a sequence of pairs of random variables where each pair takes values in $\mathcal{K}_0 \times \mathbb{R}_+$ (for $\mathcal{K}_0 \equiv \{0\} \cup \mathcal{K}$) such that for $1 \le i \le \mathcal{Z}_j^\ell$, $O_j^\ell(i)$ is associated with the i^{th} customer at station j at the time σ_ℓ. The first coordinate of this ordered pair is the class designator for the customer (an element of $\mathcal{K}$), and the second coordinate is the "age" of that customer, i.e., the amount of time that has passed since the customer arrived to its current class at station j. For $i \ge \mathcal{Z}_j^\ell + 1$, $O_j^\ell(i) \equiv (0,0)$. For example, provided $\mathcal{Z}_j^\ell > 0$, $O_j^\ell(1)$ consists of the class designator and the age of the customer that is at station j at σ_ℓ and who arrived there the longest time ago. For $\ell = 0$, the customers present at each station at time zero are assumed to have (fictitious) arrival times that occurred before time zero. (These fictitious times may also come into the computation of O_j^ℓ for $\ell > 0$ through customers who are initially at station j and who are still there at some later times σ_ℓ for $\ell > 0$.) For many purposes the additional information about the order of arrival of the customers to station j that is carried by the first coordinate of the sequence $\{O_j^\ell(i), i=1,2,\ldots\}$ suffices for the specification of the service discipline at the station. However, the additional information about ages of customers carried by the second coordinate of $\{O_j^\ell(i), i=1,2,\ldots\}$ has been included to allow for a wide variety of HL disciplines, including those that might provide service based on the specific age of the customers at the station.

For each $\ell \in \mathbb{N}$, let $Z^\ell = (Z_1^\ell, \ldots, Z_K^\ell)$, $u^\ell = (u_l^\ell : l \in \mathcal{A})$, $v^\ell = \{(v_1^\ell(i), \ldots, v_K^\ell(i))\}_{i=1}^\infty$, $O^\ell = \{(O_1^\ell(i), \ldots, O_J^\ell(i))\}_{i=1}^\infty$ and

$$\mathcal{X}^\ell = (Z^\ell, u^\ell, v^\ell, O^\ell). \tag{12}$$

Then, $\mathcal{X}^\ell$ is the information that can be used to determine the proportions of server effort to be allocated to the various classes for the period $[\sigma_\ell, \sigma_{\ell+1})$. Define $\mathcal{X}(t) = \sum_{\ell=0}^\infty \mathcal{X}^\ell 1_{\{\sigma_\ell \le t < \sigma_{\ell+1}\}}$ for all $t \ge 0$. We regard $\mathcal{X}(0) = \mathcal{X}^0$ as specifying the *initial conditions* for our queueing network model.

An HL service discipline is such that at the time σ_ℓ, for each $j \in \mathcal{J}$, proportions r_k^ℓ (of server j's effort) are assigned to the classes $k \in \mathcal{C}(j)$ at station j, where $r_k^\ell = 0$ if $Z_k^\ell = 0$ and $\sum_{k \in \mathcal{C}(j)} r_k^\ell = 1$ if $\mathcal{Z}_j^\ell \neq 0$. These proportions are specified by a measurable function $\Psi : \mathbb{N}^K \times \mathbb{R}_+^{|\mathcal{A}|} \times (\mathbb{R}_+^K)^{\mathbb{N}} \times ((\mathcal{K}_0 \times \mathbb{R}_+)^J)^{\mathbb{N}} \to [0,1]^K$ where $r^\ell = \Psi(\mathcal{X}^\ell)$. (Here, the measurable structures on the domain and range are specified in the natural way — product σ-algebras are used, where $\mathbb{N}$ and $\mathcal{K}_0$ have their discrete σ-algebras and $[0,1]$, $\mathbb{R}_+$, $\mathbb{R}$ are endowed with the usual Borel σ-algebras.) Over the interval $[\sigma_\ell, \sigma_{\ell+1})$, for each class $k \in \mathcal{K}$, service time is dispensed by server $s(k)$ to class k at a constant rate of r_k^ℓ. Then r_k^ℓ is the proportion of server $s(k)$'s time devoted to class k over the interval $[\sigma_\ell, \sigma_{\ell+1})$. It is also the rate at which the workload (measured in units of required service time) is being depleted for class k over this time interval. More formally, if for each $t \ge 0$, $T_k(t)$ denotes the amount of time server $s(k)$ has spent serving class k up to time t, then for $t \in [\sigma_\ell, \sigma_{\ell+1})$, $\dot{T}_k(t) = r_k^\ell$, where the dot notation denotes the time derivative from the right. An HL service discipline is a service discipline satisfying all of the above assumptions. From now on we restrict to such service disciplines. We describe several common HL service disciplines in more detail below.

FIFO. Under a FIFO service discipline, customers at each station are served on a first-in-first-out basis. Thus, for this discipline, for $j \in \mathcal{J}$ and $k \in \mathcal{C}(j)$, $r_k^\ell = 1$ if the first coordinate of $O_j^\ell(1)$ equals k and $r_k^\ell = 0$ otherwise. A FIFO network of *Kelly type* is a FIFO network such that for each station $j \in \mathcal{J}$, the mean service time m_k is the same for all customer classes $k \in \mathcal{C}(j)$ served at station j.

Static Priorities (Preemptive Resume). Under a static priority discipline, the classes at each station have a fixed ranking. When the server switches attention from one customer to another, the new customer is taken from the head-of-the-line of the highest ranking non-empty class at the server's station. If two or more classes are tied at the same rank, then all customers with the same rank are served on a FIFO basis, without regard to class designation.

We consider preemptive resume static priority service. For this, when a customer arrives at the server's station with a higher rank than the one currently being served, the service of the current customer is interrupted until service of all customers with higher ranks is completed, at which time the interrupted service continues from where it left off (preemptive resume). For this discipline, for $j \in \mathcal{J}$, if $k \in \mathcal{C}(j)$ is the single highest ranking class that appears in the first coordinate of the sequence $O_j^\ell = \{O_j^\ell(i)\}_{i=1}^\infty$ then $r_k^\ell = 1$ and $r_l^\ell = 0$ for $l \in \mathcal{C}(j)$, $l \neq k$. If there is more than one class of highest rank appearing in O_j^ℓ, then $r_k^\ell = 1$ for the first such class k appearing in the sequence O_j^ℓ and $r_l^\ell = 0$ for all $l \in \mathcal{C}(j) \setminus \{k\}$. One could handle non-preemptive priority service by adding an additional component to $\mathcal{X}$ to keep track of the customer currently being served at each station. For the sake of simplicity (in verifying the stopping time property in Lemma 8.3), we have not included this case here. Indeed, based on extant limit theorems and the fact that a single customer is infinitesimal in the heavy traffic limit, it is reasonable to conjecture that the heavy traffic behavior will be the same whether preemptive or non-preemptive service is used. For future reference, for each $j \in \mathcal{J}$, we let $\mathcal{L}(j)$ denote the index (or indices) for that class (or those classes) that have lowest priority at station j.

Head-of-the-line processor sharing (HLPS). An HLPS service discipline is an HL discipline where the server simultaneously serves the customers at the head-of-the-line of each (non-empty) class, i.e., $r_k^\ell > 0$ for each k such that $Z_k^\ell > 0$. If the server allocates effort to each class in proportion to the number of customers in that class, the discipline is called the head-of-the-line proportional processor sharing discipline (HLPPS). For this discipline, for $j \in \mathcal{J}$, $k \in \mathcal{C}(j)$, $r_k^\ell = Z_k^\ell / \mathcal{Z}_j^\ell$ if $\mathcal{Z}_j^\ell \neq 0$, or $r_k^\ell = 0$ if $\mathcal{Z}_j^\ell = 0$. Alternatively, another type of HLPS discipline is obtained if the server allocates its effort equally among the non-empty classes.

Remark. Bramson's [10, 11] definition of an HL service discipline is a little more restrictive than that described here in that his proportions r^ℓ depend on a modified $\mathcal{X}^\ell$, where the sequence $v^\ell = \{v^\ell(i)\}_{i=1}^\infty$ is replaced by the vector $v^\ell(1)$, consisting simply of the first residual service time for each class. Since the information in v^ℓ is not used in specifying the FIFO, preemptive resume priority or HLPPS service disciplines, Bramson's results apply for these disciplines.

Breaking ties. In addition to the above, if simultaneous arrivals are allowed to a station, then a rule is needed for determining the (FIFO) ordering of these arrivals. For concreteness and for the purpose of rigorous proof, we describe below a specific tie breaking rule to be used in such cases. We call this a deterministic tie breaking rule because it does not require the use of additional randomness beyond that already present in the system in order to break the tie.

For simultaneous arrivals to different classes at a given station, the FIFO ordering of the classes is determined by considering arrivals to higher numbered classes to have occurred immediately ahead of arrivals to lower numbered classes. For simultaneous arrivals to a given class, their FIFO ordering within the class is based on the classes from which the customers just came, with customers who just departed from higher numbered classes being placed ahead of customers who just departed from lower numbered classes, and external arrivals being considered to have just departed from a lowest numbered class labelled "zero". If simultaneous arrivals occur both to the same class and to different classes at a given station, the classes are first ordered and then the customers are ordered within each class, according to the above rules.

The reader should note that, in the presence of simultaneous arrivals to different classes at a station, a tie breaking rule is needed to update the order sequence $\{O^\ell\}_{\ell=0}^\infty$. However, depending on the service discipline, this rule may not affect the evolution of our measures of performance, namely, the queue length and workload processes. For example, although the order of arrival to different classes is used in executing the FIFO (across classes) service discipline, this information is not needed for a strict priority or HLPPS discipline. Furthermore, for simultaneous arrivals to the same class, the service times and routing vectors for the class are i.i.d. and the information inserted in the order sequence $\{O^\ell\}_{\ell=0}^\infty$ for these arrivals is the same. Consequently, the specific tie breaking rule used within a class will not affect our heavy traffic limit results. We have given an explicit tie breaking rule for this case so that the reader who wishes to do so will have something concrete to focus on and will not be concerned that this is an undefined quantity.

3.1.6 Independence

We assume that $\{u_l(i), i = 2, 3, \ldots\}$, $\{v_k^s(i), i = 1, 2, \ldots\}$ and $\{\phi^k(i), i = 1, 2, \ldots\}$, for $l \in \mathcal{A}$ and $k \in \mathcal{K}$ are mutually independent sequences of i.i.d. random variables (or vectors in the case of the $\phi^k(i)$) and that collectively these are independent of $\mathcal{X}(0)$. These properties are used to verify the functional central limit theorem in Section 5 and a certain martingale property in Section 8. While the independence assumption could be relaxed somewhat and the functional central limit theorem would still hold, the proof of the martingale property uses the independence in a crucial way and generalizing the proof would require a different approach than adopted here.

We shall refer to the basic stochastic processes E, V^o, V^s, Φ as the primitive processes for our multiclass open queueing network model. They are primitives in the sense that given these processes, the service discipline, initial conditions $\mathcal{X}(0)$ and the rule for "breaking ties", one can construct path-by-path the queue length, workload, cumulative idletime, arrival and departure processes.

3.2 Performance processes and model equations

The following descriptive processes Z, W, Y will be used to measure the performance of our queueing network. For each class $k \in \mathcal{K}$, let $Z_k(t)$ denote the number of class k customers that are in queue or being served at station $s(k)$ at time t. The K-dimensional process

$$Z = \{Z(t), t \geq 0\} \tag{13}$$

is called the *queue length process.* For each station $j \in \mathcal{J}$, let $W_j(t)$ denote the amount of work (measured in units of remaining service time) embodied in those customers who are at station j at time t. The J-dimensional process

$$W = \{W(t), t \geq 0\} \tag{14}$$

is called the (immediate) *workload process.* For each station $j \in \mathcal{J}$, let $Y_j(t)$ denote the total amount of time that the server at station j has been idle in the time interval $[0,t]$. The J-dimensional process

$$Y = \{Y(t), t \geq 0\} \tag{15}$$

is called the *cumulative idletime process.* The queue length and workload processes measure congestion and delay in the network. The idletime process measures utilization of the resources (servers) in the network.

We further define the K-dimensional descriptive processes $A = \{A(t), t \geq 0\}$ and $D = \{D(t), t \geq 0\}$ where for each $k \in \mathcal{K}$ and $t \geq 0$, $A_k(t)$ denotes the number of arrivals (both internal and external) to class k that have occurred in the time interval $[0,t]$ and $D_k(t)$ denotes the number of departures from class k that have occurred in the interval $[0,t]$. We assume that $A(0) = 0$ and $D(0) = 0$. Let e denote the J-dimensional vector, each component of which is equal to one. Then the processes A, D, W, Y, Z are related by the following *model equations* (cf. [30]). Here new processes L and F are introduced to elucidate the system structure. These are defined in equations (17) and (22), respectively. We emphasize here that the model equations (16)–(22) are simply relations satisfied by the descriptive processes associated with our queueing network model. These equations do not give a complete description of the behavior of our queueing network (see the Remark at the end of (this) Section 3.2).

For each $t \geq 0$,

$$\begin{aligned}
A(t) &= E(t) + F(t) && (16)\\
L(t) &= CV^s(A(t)) && (17)\\
W(0) &= CV^o(Z(0)) && (18)\\
W(t) &= W(0) + L(t) - et + Y(t) && (19)\\
Y(t) &= \sup_{0 \leq s \leq t} (W(0) + L(s) - es)^- && (20)\\
Z(t) &= Z(0) + A(t) - D(t) && (21)\\
F(t) &= \sum_{k=1}^{K} \Phi^k(D_k(t)) && (22)
\end{aligned}$$

where $x^- = \max(-x, 0)$ for $x \in \mathbb{R}^J$ and the maximum (as well as the supremum in (20)) is taken componentwise.

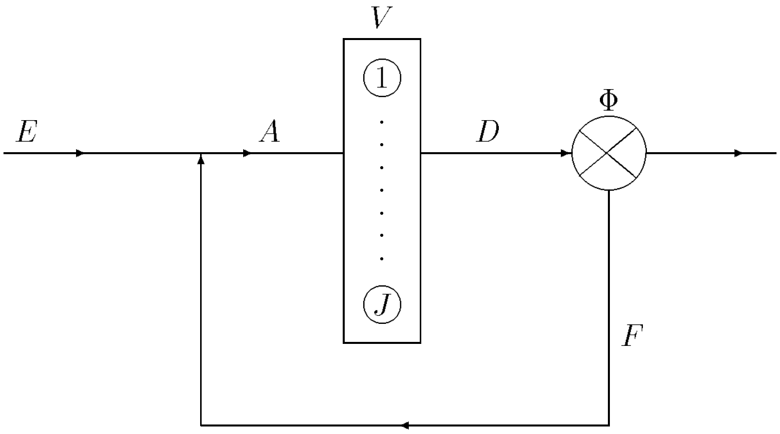

Figure 1: Schematic for a multiclass open queueing network

These equations can be interpreted as follows. The arrivals to class l in $[0,t]$ consist of $E_l(t)$ *external* arrivals plus $F_l(t) = \sum_{k=1}^{K} \Phi_l^k(D_k(t))$ arrivals from *feedback* due to the fraction of the $D_k(t)$ departures from class k that are routed next to class l, summed over all classes $k \in \mathcal{K}$. The total amount of work that has arrived at station j by time t consists of the initial workload $W_j(0) = \sum_{k\in\mathcal{C}(j)} V_k^o(Z_k(0))$ due to the $\sum_{k\in\mathcal{C}(j)} Z_k(0)$ customers who are originally at station j at time zero, plus the amount of work that has come in due to arrivals in $[0,t]$ to the classes $k \in \mathcal{C}(j)$. The i^{th} such arrival to class k brings with it an amount of work $v_k^s(i)$ and hence the total amount of work that has come to station j in $[0,t]$ due to such arrivals is the cumulative *load* $L_j(t) = \sum_{k\in\mathcal{C}(j)} V_k^s(A_k(t))$. The amount of work $W_j(t)$ remaining at station j at time t is then $W_j(0) + L_j(t)$ minus the amount of time $t - Y_j(t)$ that the server has been *busy* up to time t. This leads to the expression for W given by equations (17)–(19). Now for each $j \in \mathcal{J}$, W_j is a process with r.c.l.l. paths that take values in $[0,\infty)$, and Y_j is continuous, non-decreasing and starts from zero. Furthermore, we have the integral relation

$$\int_{[0,\infty)} W_j(t)\, dY_j(t) = 0 \quad \text{for all } j \in \mathcal{J}, \tag{23}$$

which embodies the non-idling property that server j can only be idle when there is no work to be done at station j. It then follows from uniqueness of the solution of the one-dimensional Skorokhod problem (cf. [13]) that Y_j is given by equation (20). The equation (21) simply expresses the fact that the queue length $Z_k(t)$ for class k at time t is equal to the initial queue length $Z_k(0)$ plus the number of arrivals $A_k(t)$ that have occurred to class k in $[0,t]$ minus the number of departures $D_k(t)$ from class k that have occurred in that time interval.

Remark. The fact that equations (16)–(22) do not give a complete description of the behavior of our queueing network model is deliberate. These are generic equations satisfied by the descriptive processes associated with any queueing network model satisfying the assumptions of Section 3.1. In particular, these equations do not include information about the service discipline. In certain situations (e.g., for a single class FIFO network as considered by Reiman [44]), one can add one or more equations to uniquely specify the evolution of the processes Z, W, Y, A, D. However, for a general HL service discipline in a multiclass network, one would at least need to add equations that are equivalent to the description given in Section 3.1.5. This has not been done here because the queueing network model described in Section 3.1 is taken as primitive and the model equations, plus a $K \times J$ matrix Δ to be introduced later as a distillation of the service discipline, are the main derivatives of this model needed to develop the diffusion approximation to our open multiclass queueing network.

3.3 Relationship to Bramson's model

In a companion work, Bramson [11] considers the problem of proving state space collapse for open multiclass queueing networks. In particular, he verifies multiplicative state space collapse for FIFO networks of Kelly type and for networks with an HLPPS service discipline. Since Bramson focusses on proving state space collapse, rather than the full diffusion approximation, he is able to use a slightly simpler setup for the service times than used here. Rather than introducing two sequences of service times for the original and subsequent class k customers as we do, Bramson uses a single sequence of random variables $\{v_k(i), i = 1, 2, \ldots\}$, where $v_k(1)$ has a distribution corresponding to that of the first customer in class k (equal to that of our $v_k^o(1)$), and the $\{v_k(i), i = 2, 3, \ldots\}$ are i.i.d. and have the same distribution as our $\{v_k^s(i), i = 1, 2, \ldots\}$. He further assumes that the sequences $\{v_k(i), i = 2, 3, \ldots\}$ for $k \in \mathcal{K}$ are mutually independent, and as a collection are independent of the other model primitives E, Φ and the initial variables $Z(0)$, $\{v_k(1),\ k \in \mathcal{K}\}$ and O^0.

Indeed, with the addition of some assumptions concerning our $\{v_k^o(i), i = 1, 2, \ldots\}$, $k \in \mathcal{K}$, one can think of constructing Bramson's sequences from ours as follows. In addition to the assumptions already made, assume that for each $k \in \mathcal{K}$, $\{v_k^o(i), i = 2, 3, \ldots\}$ is a sequence of i.i.d. random variables that has the same distribution as our $\{v_k^s(i), i = 1, 2, \ldots\}$. Further suppose that the sequences $\{v_k^o(i), i = 2, 3, \ldots\}$, $k \in \mathcal{K}$, are mutually independent and as a collection are independent of the initial queue length vector $Z(0)$, the initial residual times $v_k^o(1), k \in \mathcal{K}$, and $u_l, l \in \mathcal{A}$, and the initial sequence of (class, age) pairs O^0. Then one can define Bramson's sequence from ours as follows. Let $v_k(i) = v_k^o(i)$ for $1 \le i \le Z_k(0)$ and $v_k(i) = v_k^s(i - Z_k(0))$ for $i > Z_k(0)$, $k \in \mathcal{K}$. In terms of the model equations, the correspondence between Bramson's model and ours is then given by

$$V_k(Z_k(0) + A_k(t)) = V_k^o(Z_k(0)) + V_k^s(A_k(t)), \tag{24}$$

where $V_k(n) = \sum_{i=1}^{n} v_k(i)$, $n = 0, 1, 2, \ldots$.

There is one other distinction between Bramson's setup and that used here. In the next section, we consider a *sequence* of open multiclass queueing networks with the same basic structure as described in this section. On the other hand, Bramson [11] considers an uncountable family of such networks indexed by all of the positive real numbers $r \in (0, \infty)$.

One can embed our sequence in a family of the type considered by Bramson simply by defining a family that is piecewise constant, i.e., it is constant between successive values taken by our sequence parameter. Furthermore, one can easily check that the assumptions required by Bramson for multiplicative state space collapse follow for this family from the heavy traffic assumptions that we will assume in Section 5. Although we shall not need it, in fact, one can go back and forth between our results for sequences and Bramson's result for families. We have chosen to keep to sequences here for compatibility with previous heavy traffic limit theorems.

In summary, by combining the above correspondences, it follows that the conclusions obtained by Bramson concerning state space collapse apply to our situation when we make the additional assumptions on the original service times described in the second paragraph of (this) Section 3.3.

4 Scaling

We now consider a sequence of open multiclass queueing networks indexed by r, where r tends to infinity through a strictly increasing sequence of values in $(0, \infty)$. (As in Bramson [11], we use r rather than n to allow for the possibility of scaling by non-integer r.) Each of these networks is to have the same basic structure as described in the previous section. The numbers of stations and classes and the mapping C from classes to stations will remain fixed for all r. Furthermore, the collection of classes $\mathcal{K} \setminus \mathcal{A}$ that have no external arrivals will not change with r. On the other hand, the primitive stochastic processes E, V^o, V^s, Φ and the initial conditions embodied in $\mathcal{X}(0)$ are allowed to change with r. We also allow the HL service discipline (as specified by the mapping Ψ in Section 3.1.5) to vary with r. However, in practice, some continuity or limiting behavior of the HL service disciplines is needed as $r \to \infty$ in order for the critical property of state space collapse to hold (cf. Definitions 4.1 and 7.1, and Assumption 7.2). Indeed, although it is no more complicated to treat the general case, to fix ideas the reader may wish to first consider the case where the service discipline (i.e., Ψ) is fixed for the sequence of networks. This still includes the common service disciplines given as examples here, namely, FIFO, preemptive resume priority and the HLPPS service discipline. To indicate the position in the sequence of networks, a superscript r will be appended to the network parameters and processes. When there is already a superscript, the r will be appended to the right of the current superscript, e.g., as in $\Phi^{k,r}$.

We first define λ^r to be the unique K-dimensional vector solution to the traffic equation:

$$\lambda^r = \alpha^r + \tilde{P}^r \lambda^r, \tag{25}$$

i.e., using the definition (8) of $Q^r \equiv (I - \tilde{P}^r)^{-1}$,

$$\lambda^r = Q^r \alpha^r. \tag{26}$$

We also define

$$\rho_j^r = \sum_{k \in \mathcal{C}(j)} \lambda_k^r m_k^r \quad \text{for all } j \in \mathcal{J}, \tag{27}$$

or equivalently, $\rho^r = CM^r\lambda^r$. For the r^{th} network, λ_k^r is interpreted as the *nominal arrival rate* to class k due to external and internal arrivals, and ρ_j^r is the *nominal traffic intensity for station* j. The qualifier *nominal* is used here because the *traffic equation* (25) implicitly assumes that for each class k there is a long run average rate λ_k^r of flow into and out of that class and that this does not exceed the maximal mean service rate $\mu_k^r = 1/m_k^r$ for class k. The issue of whether the nominal arrival rate is actually a long run average arrival rate is related to stability of the network. Rather than elaborating further on this, as a start, we refer the interested reader to the articles in [37] by Harrison [29], Kumar [38], Dai [18], and Bramson [7]. For our purposes, we simply regard (26) and (27) as useful definitions.

We define the following fluid and diffusion scaled processes. Fluid (or law of large numbers) scaling is indicated by placing a bar over a process. Diffusion (or central limit theorem) scaling is indicated by placing a hat over a process. Here $[x]$ denotes the integer part of a non-negative real number x and $\tilde{P}^{k,r}$ denotes the k^{th} column of $\tilde{P}^r$. We extend the definition of the scaled cumulative service time and routing processes to all non-negative times by making them piecewise constant.

Fluid scaling.

$$\bar{E}^r(t) = r^{-2}E^r(r^2t) \tag{28}$$

$$\bar{V}^{o,r}(t) = r^{-2}V^{o,r}([r^2t]) \tag{29}$$

$$\bar{V}^{s,r}(t) = r^{-2}V^{s,r}([r^2t]) \tag{30}$$

$$\bar{\Phi}^{k,r}(t) = r^{-2}\Phi^{k,r}([r^2t]), \tag{31}$$

and define $\bar{A}^r, \bar{D}^r, \bar{F}^r, \bar{L}^r, \bar{W}^r, \bar{Y}^r, \bar{Z}^r$ from $A^r, D^r, F^r, L^r, W^r, Y^r, Z^r$, respectively, in the same manner as $\bar{E}^r$ is defined from E^r, namely, accelerate time by a factor of r^2 and divide space by a factor of r^2.

Diffusion scaling.

$$\hat{E}^r(t) = r^{-1}(E^r(r^2t) - \alpha^r r^2 t) \tag{32}$$

$$\hat{A}^r(t) = r^{-1}(A^r(r^2t) - \lambda^r r^2 t) \tag{33}$$

$$\hat{D}^r(t) = r^{-1}(D^r(r^2t) - \lambda^r r^2 t) \tag{34}$$

$$\hat{F}^r(t) = r^{-1}(F^r(r^2t) - \tilde{P}^r\lambda^r r^2 t) \tag{35}$$

$$\hat{V}^{s,r}(t) = r^{-1}(V^{s,r}([r^2t]) - m^r[r^2t]) \tag{36}$$

$$\hat{\Phi}^{k,r}(t) = r^{-1}(\Phi^{k,r}([r^2t]) - \tilde{P}^{k,r}[r^2t]) \tag{37}$$

$$\hat{L}^r(t) = r^{-1}(L^r(r^2t) - \rho^r r^2 t) \tag{38}$$

$$\hat{W}^r(t) = r^{-1}W^r(r^2t) \tag{39}$$

$$\hat{Z}^r(t) = r^{-1}Z^r(r^2t) \tag{40}$$

$$\hat{Y}^r(t) = r^{-1}Y^r(r^2t). \tag{41}$$

The reader who consults the related work of Bramson [11] is cautioned that he employs a different fluid scaling than is used here. Though both definitions incorporate the notion of a law of large numbers scaling, we accelerate time by a factor of r^2 whereas Bramson only accelerates time by a factor of r in his fluid scaling. To obtain results on state space collapse relevant to our work, Bramson looks at his fluid scaled processes over long periods of time of order r in size and thus is able to prove results relevant to our time scaling by r^2. We shall need one process with the same kind of fluid scaling as Bramson, namely

$$\tilde{V}^{o,r}(t) = r^{-1} V^{o,r}([rt]). \tag{42}$$

Although the spatial scaling in (42) is the same as that for our diffusion scaling, the scaling in time is only of order r. The scaling for $V^{o,r}$ is different from that for $V^{s,r}$ because the argument of $V^{o,r}$ in the model equations is $Z^r(0)$ which will have diffusion scaling applied to it (cf. (40)), and so to compensate for division of this spatial variable by r, we need only accelerate the argument of $V^{o,r}$ by r.

Substituting these scaled processes into the model equations (16)–(22) yields the following fluid and diffusion scaled equations.

Fluid scaled equations.

$$\begin{aligned}
\bar{A}^r(t) &= \bar{E}^r(t) + \bar{F}^r(t) && (43)\\
\bar{L}^r(t) &= C\bar{V}^{s,r}(\bar{A}^r(t)) && (44)\\
\bar{W}^r(0) &= C\bar{V}^{o,r}(\bar{Z}^r(0)) && (45)\\
\bar{W}^r(t) &= \bar{W}^r(0) + \bar{L}^r(t) - et + \bar{Y}^r(t) && (46)\\
\bar{Y}^r(t) &= \sup_{0\le s\le t} (\bar{W}^r(0) + \bar{L}^r(s) - es)^- && (47)\\
\bar{Z}^r(t) &= \bar{Z}^r(0) + \bar{A}^r(t) - \bar{D}^r(t). && (48)\\
\bar{F}^r(t) &= \sum_{k=1}^{K} \bar{\Phi}^{k,r}(\bar{D}^r_k(t)) && (49)
\end{aligned}$$

Note that for equation (51) below we use equation (25) and for equation (54) we define

$$\hat{\gamma}^r = (\rho^r - e)r \equiv (CM^r\lambda^r - e)r. \tag{50}$$

Diffusion scaled equations.

$$\begin{aligned}
\hat{A}^r(t) &= \hat{E}^r(t) + \hat{F}^r(t) && (51)\\
\hat{L}^r(t) &= C\left(\hat{V}^{s,r}(\bar{A}^r(t)) + M^r\hat{A}^r(t)\right) && (52)\\
\hat{W}^r(0) &= C\tilde{V}^{o,r}(\hat{Z}^r(0)) && (53)\\
\hat{W}^r(t) &= \hat{W}^r(0) + \hat{L}^r(t) + \hat{\gamma}^r t + \hat{Y}^r(t) && (54)\\
\hat{Y}^r(t) &= \sup_{0\le s\le t} (\hat{W}^r(0) + \hat{L}^r(s) + \hat{\gamma}^r s)^- && (55)
\end{aligned}$$

$$\hat{Z}^r(t) = \hat{Z}^r(0) + \hat{A}^r(t) - \hat{D}^r(t) \tag{56}$$

$$\hat{F}^r(t) = \sum_{k=1}^{K} \hat{\Phi}^{k,r}(\bar{D}_k^r(t)) + \tilde{P}^r \hat{D}^r(t) \tag{57}$$

The fluid scaled equations will be used later to determine the behavior of $\bar{A}^r$ and $\bar{D}^r$ as $r \to \infty$. On the other hand, the diffusion scaled equations will used to determine the behavior of the normalized workload $\hat{W}^r$ and queue length $\hat{Z}^r$ processes as $r \to \infty$. Before proceeding with this, we reduce the system of equations (51)–(57) as follows.

By substituting $\hat{D}^r$ from (56) into (57) and in turn substituting $\hat{F}^r$ from (57) into (51) and solving for $\hat{A}^r$ we obtain

$$\hat{A}^r(t) = Q^r \left(\hat{E}^r(t) + \sum_{k=1}^{K} \hat{\Phi}^{k,r}(\bar{D}_k^r(t)) - \tilde{P}^r(\hat{Z}^r(t) - \hat{Z}^r(0)) \right). \tag{58}$$

Substituting this into (52) and then substituting the result into (54) yields

$$\hat{W}^r(t) = \hat{W}^r(0) + \hat{\xi}^r(t) - CM^r Q^r \tilde{P}^r(\hat{Z}^r(t) - \hat{Z}^r(0)) + \hat{Y}^r(t), \tag{59}$$

where

$$\hat{\xi}^r(t) = C\hat{V}^{s,r}(\bar{A}^r(t)) + CM^r Q^r \left(\hat{E}^r(t) + \sum_{k=1}^{K} \hat{\Phi}^{k,r}(\bar{D}_k^r(t)) \right) + \hat{\gamma}^r t. \tag{60}$$

Note that the only stochastic processes appearing in the expression for $\hat{\xi}^r$ are the diffusion scaled primitive processes $\hat{E}^r, \hat{V}^{s,r}, \hat{\Phi}^{k,r}$ for $k \in \mathcal{K}$, and the fluid scaled arrival and departure processes $\bar{A}^r, \bar{D}^r$.

We now introduce a $K \times J$ matrix Δ^r:

$$\Delta_{kj}^r = \begin{cases} \delta_k^r & \text{if } k \in \mathcal{C}(j) \\ 0 & \text{otherwise,} \end{cases} \tag{61}$$

where the non-negative K-dimensional vector δ^r depends on the service discipline and parameters for the r^{th} network. We consider the specification of Δ^r to be part of the description of the r^{th} network. Some examples of forms we propose for Δ^r for some common service disciplines are given below. In fact, the key desired property of Δ^r is what we call *state space collapse,* which is defined below. When we have state space collapse, it allows us to approximate the k^{th} component of the higher dimensional rescaled queue length process $\hat{Z}^r$ by a multiple δ_k^r of the j^{th} component of the lower dimensional rescaled workload process $\hat{W}^r$ where $k \in \mathcal{C}(j)$.

Definition 4.1 (State space collapse)
State space collapse is said to hold for our sequence of open multiclass queueing networks if for each $T \geq 0$,

$$\|\hat{Z}^r(\cdot) - \Delta^r \hat{W}^r(\cdot)\|_T \to 0 \quad \textit{in probability as } r \to \infty. \tag{62}$$

Remark. Bramson [11] has a slightly different definition of state space collapse. Under our heavy traffic condition (80) (which implies we are in the critical case in Bramson's terminology) and the assumption that Δ^r converges (componentwise) to a $K \times J$ matrix Δ as $r \to \infty$, these two definitions are equivalent. Later, in Section 7, we shall define a weaker notion called *multiplicative state space collapse.* In fact, it is this condition that Bramson [11] verifies for FIFO networks of Kelly type and HLPPS networks. His work illustrates that from the point of view of verification, multiplicative state space collapse can be a more natural condition to work with. On the other hand, from the point of view of proving a heavy traffic limit theorem, state space collapse as defined above is a more natural condition. Fortunately, as shown in Proposition 8.1, the two conditions are equivalent under the assumptions given in Section 5 and Assumption 7.1, both of which are used as conditions for our heavy traffic limit theorem (Theorem 7.1).

From extant limit theorems, we extrapolate to propose the following forms for the δ^r (or equivalently, Δ^r) for some common service disciplines. The forms for the FIFO and preemptive resume static priority service disciplines may be found in the Appendix to [32] and are based on limit theorems of Reiman [44, 46] and Peterson [42]. The form for the HLPPS case is based on the work of Bramson [11]. We emphasize here that when referring to these common service disciplines, we assume that the discipline is held fixed for the entire sequence of networks. Accordingly, we shall refer to a sequence of networks with a FIFO (respectively, static priority or HLPPS) service discipline as a sequence of FIFO (respectively, static priority or HLPPS) networks. In the following $j \in \mathcal{J}$ and $k \in \mathcal{C}(j)$. Here we assume the denominators are strictly positive (this can be achieved for instance by making the mild assumption that λ^r has strictly positive entries).

FIFO.

$$\delta_k^r = \frac{\lambda_k^r}{\sum\limits_{l \in \mathcal{C}(j)} \lambda_l^r m_l^r}. \tag{63}$$

Static Priorities (Preemptive Resume). The priority ranking is fixed for the entire sequence of networks and hence the list $\mathcal{L}(j)$, consisting of those classes served at station j that have lowest priority, does not depend on r.

$$\delta_k^r = \begin{cases} \lambda_k^r / (\sum_{l \in \mathcal{L}(j)} \lambda_l^r m_l^r) & \text{for } k \in \mathcal{L}(j) \\ 0 & \text{for } k \in \mathcal{C}(j) \setminus \mathcal{L}(j). \end{cases} \tag{64}$$

Head-of-the-line proportional processor sharing (HLPPS).

$$\delta_k^r = \frac{\lambda_k^r m_k^r}{\sum_{l \in \mathcal{C}(j)} \lambda_l^r (m_l^r)^2}. \tag{65}$$

Note that in all cases,

$$C M^r \Delta^r = I. \tag{66}$$

For the aforementioned service disciplines, we take Δ^r to have the form indicated. For other HL service disciplines, we assume that a Δ^r of the form (61) has been given. (A more specific form for Δ^r is not known in general.)

Now, for each $t \geq 0$ define

$$\hat{\varepsilon}^r(t) = \hat{Z}^r(t) - \Delta^r \hat{W}^r(t), \tag{67}$$

$$\hat{\eta}^r(t) = CM^r Q^r \tilde{P}^r(\hat{\varepsilon}^r(0) - \hat{\varepsilon}^r(t)), \tag{68}$$

$$G^r = CM^r Q^r \tilde{P}^r \Delta^r. \tag{69}$$

Note that the state space collapse condition (62) can be rewritten as

$$\|\hat{\varepsilon}^r\|_T \to 0 \quad \text{in probability as } r \to \infty. \tag{70}$$

Substituting the above definitions into (59) yields

$$\hat{W}^r(t) = \hat{W}^r(0) + \hat{\xi}^r(t) - G^r(\hat{W}^r(t) - \hat{W}^r(0)) + \hat{\eta}^r(t) + \hat{Y}^r(t) \tag{71}$$

and hence

$$(I + G^r)(\hat{W}^r(t) - \hat{W}^r(0)) = \hat{\xi}^r(t) + \hat{\eta}^r(t) + \hat{Y}^r(t), \tag{72}$$

where, since $Q^r = (I - \tilde{P}^r)^{-1}$ and equation (66) holds, we have

$$I + G^r = CM^r Q^r \Delta^r. \tag{73}$$

To proceed further we need the following Condition 4.1 to be satisfied by $I + G^r$. This is known to hold for certain service disciplines and network structures, provided r is sufficiently large. In particular, in Dai-Harrison [20] this was shown for FIFO networks of Kelly type, it is shown for HLPPS networks in Section 9, and in Dai-Yeh-Zhou [25] it is shown for re-entrant lines with a first-buffer-first-served or last-buffer-first-served static priority discipline.

Condition 4.1 *$I + G^r$ is invertible.*

Assuming that Condition 4.1 holds, on multiplying (72) by $R^r \equiv (I + G^r)^{-1}$, we obtain the following reduced form of the diffusion scaled equations:

$$\hat{W}^r(t) = \hat{X}^r(t) + R^r \hat{Y}^r(t), \quad t \geq 0, \tag{74}$$

where

$$\hat{X}^r(t) = \hat{W}^r(0) + R^r \left(\hat{\xi}^r(t) + \hat{\eta}^r(t)\right), \tag{75}$$

and $\hat{\xi}^r$, $\hat{\eta}^r$ are given by equations (60), (68), respectively. Note that from the basic properties of the workload and cumulative idletime processes, W^r, Y^r, respectively, we have

$$\hat{W}^r(t) \in \mathbb{R}_+^J \text{ for all } t \geq 0, \tag{76}$$

and for each $j \in \mathcal{J}$,

$$\hat{Y}_j^r(0) = 0, \quad \hat{Y}_j^r \text{ is non-decreasing} \quad \text{and} \quad \int_0^\infty \hat{W}_j^r(t)\, d\hat{Y}_j^r(t) = 0. \tag{77}$$

5 Heavy traffic assumptions

To obtain a heavy traffic limit theorem for our sequence of open multiclass queueing networks, we need to impose some additional conditions on the behavior of the primitive processes as $r \to \infty$. In particular, we assume that α^r, m^r, P^r, which represent the long run average external arrival rates, mean service times, and transition probabilities for the customer classes, respectively, satisfy

$$\alpha^r \to \alpha, \quad m^r \to m, \quad P^r \to P \tag{78}$$

as $r \to \infty$, where $\alpha \geq 0$ and $m > 0$ (inequalities are to be interpreted componentwise) are finite vectors, and P is a $K \times K$ substochastic matrix with spectral radius strictly less than one. Recalling the definition (27) of the traffic intensity $\rho^r = CM^r\lambda^r$, we see that the above implies that $\rho^r \to \rho = CM\lambda$ as $r \to \infty$, where $M = \mathrm{diag}(m)$, $Q = (I - \tilde{P})^{-1}$, $\tilde{P} = P'$, and

$$\lambda = Q\alpha. \tag{79}$$

Our assumption of *heavy traffic* is expressed by the requirements that $\rho = e$ and

$$\hat{\gamma}^r \equiv (\rho^r - e)r \to \gamma \quad \text{as } r \to \infty, \tag{80}$$

for some $\gamma \in \mathbb{R}^J$.

Remark. We note that these assumptions (78)–(80) are consistent with the usual heavy traffic assumptions. For instance, setting $r = \sqrt{n}$ for $n \in \mathbb{N}$, one obtains scaling and assumptions like those in Iglehart and Whitt [34, 35], Reiman [44] and Chen and Zhang [15, 16]. Setting $r = 1/\epsilon_n$, where $\epsilon_n \to 0$, yields the assumptions of Peterson [42].

Next we make the following assumptions which imply a functional central limit theorem for the diffusion scaled primitive processes. Recall that for each r the sequences of external interarrival times $\{u_l^r(i), i = 2, 3, \ldots\}$, service times $\{v_k^{s,r}(i), i = 1, 2, \ldots\}$, and routing vectors $\{\phi^{k,r}(i), i = 1, 2, \ldots\}$, for $l \in \mathcal{A}$ and $k \in \mathcal{K}$, are mutually independent sequences of i.i.d. random variables/vectors and that these are independent of the initial conditions $\mathcal{X}^r(0)$. We suppose that in addition to (78), the variances of the external interarrival times and service times satisfy

$$a_l^r \to a_l \in [0, \infty) \ \text{ for each } l \in \mathcal{A}, \quad b_k^r \to b_k \in [0, \infty) \ \text{ for each } k \in \mathcal{K}, \tag{81}$$

the original *residual* interarrival times satisfy

$$r^{-1}u_l^r(1) \to 0 \quad \text{in probability as } r \to \infty \text{ for each } l \in \mathcal{A}, \tag{82}$$

the original *residual* service times for the customers at the head-of-the-line of each class satisfy

$$r^{-1}v_k^{o,r}(1) \to 0 \quad \text{in probability as } r \to \infty \text{ for each } k \in \mathcal{K}, \tag{83}$$

and the following conditions (which imply those of Lindeberg type) hold:

$$\sup_{l\in\mathcal{A}} \sup_r I\!\!E[(u_l^r(i))^2; u_l^r(i) > n] \quad\to\quad 0 \quad \text{as } n\to\infty \text{ for } i = 2,3,\ldots \tag{84}$$

$$\sup_{k\in\mathcal{K}} \sup_r I\!\!E[(v_k^{s,r}(i))^2; v_k^{s,r}(i) > n] \quad\to\quad 0 \quad \text{as } n\to\infty \text{ for } i = 1,2,\ldots. \tag{85}$$

Note that since the $\{u_l^r(i), i = 2,3,\ldots\}$ are i.i.d., if condition (84) holds for $i = 2$, then it holds for $i = 2,3,\ldots$. Similarly, condition (85) holds for all i if it holds for $i = 1$. Condition (83) will not be needed immediately. It is included here since it is only a mild condition on the initial residual service times and it will be needed in order to apply Bramson's results on state space collapse. We further assume that the normalized initial workload vector $\hat{W}^r(0)$ converges in distribution to a random variable $W(0)$ as $r\to\infty$.

From the above, the independence described in Section 3.1.6, and renewal and ordinary functional central limit theorems for triangular arrays (cf. [34, 35, 36, 43, 27]), we have the following. Let $\Pi = \text{diag}(\alpha_1^3 a_1, \ldots, \alpha_K^3 a_K)$, $\Sigma = \text{diag}(b_1,\ldots,b_K)$ and Υ^k be given by (10) for $k\in\mathcal{K}$. There are $K+2$ independent driftless K-dimensional Brownian motions $E, V^s, \Phi^1,\ldots,\Phi^K$, that are jointly independent of $W(0)$, that each start from the origin and have covariance matrices $\Pi,\Sigma,\Upsilon^1,\ldots,\Upsilon^K$, respectively, such that

$$(\hat{W}^r(0), \hat{E}^r, \hat{V}^{s,r}, \hat{\Phi}^r) \Rightarrow (W(0), E, V^s, \Phi) \quad \text{as } r\to\infty, \tag{86}$$

where $\hat{\Phi}^r = (\hat{\Phi}^{1,r},\ldots,\hat{\Phi}^{K,r})$ and $\Phi = (\Phi^1,\ldots,\Phi^K)$. (Note that we have reused symbols from Section 3.2 for a different purpose here, in particular, E, V^s, Φ now denote limiting diffusion processes. Since all primitive queueing network processes now have a superscript of r to indicate their position in the sequence of networks, this should not cause confusion for the reader.)

Observe that (78) together with (86) imply the following functional weak law of large numbers:

$$\left(\bar{W}^r(0), \bar{E}^r(\cdot), \bar{V}^{s,r}(\cdot), \bar{\Phi}^{1,r}(\cdot),\ldots,\bar{\Phi}^{K,r}(\cdot)\right) \to \left(0, \alpha(\cdot), m(\cdot), \tilde{P}^1(\cdot),\ldots,\tilde{P}^K(\cdot)\right) \tag{87}$$

in probability as $r\to\infty$, where $\alpha(t) = \alpha t$, $m(t) = mt$, $\tilde{P}^k(t) = \tilde{P}^k t$ for all $t\ge 0$, $k\in\mathcal{K}$, and $\tilde{P}^k$ denotes the k^{th} column of the matrix $\tilde{P} = P'$.

Let $\Lambda = \text{diag}(\lambda)$ and

$$H \equiv C\left(\Lambda\Sigma + MQ\left(\Pi + \sum_{k=1}^K \lambda_k \Upsilon^k\right) Q'M\right) C'. \tag{88}$$

We assume that H is non-degenerate, i.e., H is strictly positive definite. (Note that this can be achieved by assuming that $\lambda_k > 0$ for all k and that the diagonal entries of Σ or Π are all strictly positive. Alternatively, this could be achieved by imposing suitable restrictions on the λ_k and the Υ^k for each $k\in\mathcal{K}$. Rather than making such assumptions explicit and thereby restricting our model a priori, we simply require that H be non-degenerate.)

The assumptions stated in this section will be assumed to hold for the remainder of this paper.

6 Definition of an SRBM

Throughout this section, $S \equiv \mathbb{R}_+^J$ (the positive J-dimensional orthant), $\mathcal{B}$ denotes the σ-algebra of Borel subsets of S, θ is a constant vector in $\mathbb{R}^J$, Γ is a $J \times J$ non-degenerate covariance matrix (symmetric and strictly positive definite), and R is a $J \times J$ matrix.

The following definition of an SRBM is the same as that in Williams [53]. It is a slight generalization (allowing an arbitrary initial distribution ν) of that used by Taylor and Williams [49] in their work on the existence and uniqueness in law of such diffusion processes.

Definition 6.1 (SRBM) *Given a probability measure ν on $(S, \mathcal{B})$, a semimartingale reflecting Brownian motion (abbreviated as SRBM) associated with the data $(S, \theta, \Gamma, R, \nu)$ is an $\{\mathcal{F}_t\}$-adapted, J-dimensional process W defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ such that*

(i) $W = X + RY$, *$\mathbb{P}$-a.s.,*

(ii) *$\mathbb{P}$-a.s., W has continuous paths and $W(t) \in S$ for all $t \geq 0$,*

(iii) *under $\mathbb{P}$,*

- **(a)** *X is a J-dimensional Brownian motion with drift vector θ, covariance matrix Γ and $X(0)$ has distribution ν,*
- **(b)** $\{X(t) - X(0) - \theta t, \mathcal{F}_t, t \geq 0\}$ *is a martingale,*

(iv) *Y is an $\{\mathcal{F}_t\}$-adapted, J-dimensional process such that $\mathbb{P}$-a.s. for each $j \in \mathcal{J}$,*

- **(a)** $Y_j(0) = 0$,
- **(b)** *Y_j is continuous and non-decreasing,*
- **(c)** *Y_j can increase only when W is on the face $F_j \equiv \{x \in S : x_j = 0\}$, i.e., $\int_0^\infty 1_{(0,\infty)}(W_j(s))\, dY_j(s) = 0$.*

Loosely speaking, an SRBM behaves like a Brownian motion in the interior of the orthant S and it is confined to the orthant by instantaneous "reflection" (or "pushing") at the boundary, where the direction of "reflection" on the i^{th} face F_i is given by the i^{th} column of the *reflection matrix* R. The results of Reiman and Williams [47], Taylor and Williams [49] and Williams [53] show that a necessary and sufficient condition for the existence and uniqueness (in law) of an SRBM associated with $(S, \theta, \Gamma, R, \nu)$ for each initial distribution ν on $(S, \mathcal{B})$ is that the reflection matrix R be completely-$\mathcal{S}$, as defined below.

Definition 6.2 (Completely-$\mathcal{S}$) *For a $J \times J$ matrix R, a principal submatrix of R is any square matrix obtained by deleting all of the rows and columns of R with indices in some strict subset $\mathcal{I} \subset \mathcal{J}$ where $\mathcal{I}$ may be empty. A $J \times J$ matrix R is called completely-$\mathcal{S}$ if and only if for each principal submatrix $\tilde{R}$ of R there is $\tilde{y} > 0$ such that $\tilde{R}\tilde{y} > 0$. Here vector inequalities are to be interpreted componentwise.*

Remark. The following is a geometric interpretation of the completely-$\mathcal{S}$ condition. For each point x on the boundary of S, let $\mathcal{V}(x)$ denote the collection of those columns of R with indices j such that $x_j = 0$, i.e., the j^{th} column of R is included in $\mathcal{V}(x)$ if and only if $x_j = 0$. Intuitively, $\mathcal{V}(x)$ is the set of "push" or "control" directions that may be used at the boundary point x. The completely-$\mathcal{S}$ condition requires that for each point x on the boundary of S, there is a positive linear combination of the vectors in $\mathcal{V}(x)$ that points into the interior of S from x.

7 Main theorem

For the statement of our heavy traffic limit theorem, in addition to the general hypotheses of Section 5 on our sequence of open multiclass queueing networks, we need the following Assumptions 7.1 and 7.2 whose verification depends on the service discipline used in each network. The first assumption involves the data matrices.

Assumption 7.1 (Data matrices) *$\Delta^r \to \Delta$ as $r \to \infty$ and for $G \equiv CMQ\tilde{P}\Delta$, $I + G$ is invertible and $R = (I + G)^{-1}$ is completely-$\mathcal{S}$.*

Under a FIFO service discipline, assuming $\lambda > 0$, for all r sufficiently large Δ^r is given by (61) and (63), and assumption (78) implies the convergence of Δ^r to Δ. However, it has been known since an example of Dai and Wang [22] that the condition that R is completely-$\mathcal{S}$ can fail to hold for some multiclass FIFO queueing network models with feedback. Indeed, examples of Bramson [5, 6] have shown that not all FIFO queueing networks are stable (i.e., are positive recurrent when viewed as a Markov process) when the traffic intensity is less than one at each station. On the other hand, Bramson [8, 9] has shown that stability does hold under this traffic intensity condition for FIFO queueing networks of Kelly type and for queueing networks with the HLPPS service discipline. A proof of Dai and Harrison [20] shows that under the conditions of Section 5 and the assumption that $\lambda > 0$, Assumption 7.1 holds for any sequence of FIFO queueing networks of Kelly type. In Section 9 their proof is extended to cover any sequence of HLPPS queueing networks.

The next assumption is a form of state space collapse which relates the diffusion scaled workload and queue length processes. Here we use the notion of multiplicative state space collapse, which under suitable initial conditions and the assumptions of Section 5 has been verified by Bramson [11] to hold for FIFO networks of Kelly type and HLPPS networks. Clearly, multiplicative state space collapse is implied by state space collapse (see Definition 4.1). On the other hand, it is shown in Proposition 8.1 that under the conditions of Section 5 and Assumption 7.1, state space collapse is implied by multiplicative state space collapse, and hence in these circumstances the two notions are equivalent.

Definition 7.1 (Multiplicative state space collapse) *Multiplicative state space collapse is said to hold for our sequence of open multiclass queueing networks if for each $T \geq 0$,*

$$\frac{\|\hat{Z}^r(\cdot) - \Delta^r \hat{W}^r(\cdot)\|_T}{\|\hat{W}^r(\cdot)\|_T \vee 1} \to 0 \quad \textit{in probability as } r \to \infty. \tag{89}$$

Remark. Assuming the heavy traffic (critical case) conditions (80) and $\Delta^r \to \Delta$, Bramson's [11] definition of multiplicative state space collapse is the same as ours except that he has Δ in place of Δ^r in (89). Under the assumption that $\Delta^r \to \Delta$, it is easy to see that the two definitions are equivalent.

Assumption 7.2 *Multiplicative state space collapse holds for our sequence of open multiclass queueing networks.*

For the next theorem, which is the main result of this paper, we recall the following notation and definitions associated with our sequence of open multiclass queueing networks as described in Sections 3 and 4. Recall the definition of an HL service discipline from Section 3 and the fact that for each member of our sequence of networks an HL service discipline is used. From Section 4, recall the definitions of the diffusion scaled workload $\hat{W}^r$, cumulative idletime $\hat{Y}^r$ and queue length $\hat{Z}^r$ processes, and the matrix Δ^r. In particular, note that under Assumption 7.1, for r sufficiently large, Condition 4.1 holds and hence $\hat{W}^r$ has the representation (74) involving $\hat{X}^r$ defined by (75), (60), (67), (68). Associated with the heavy traffic assumptions of Section 5 we have the asymptotic parameters γ and H, defined in (80) and (88), respectively, and the limit $W(0)$ in distribution of $\hat{W}^r(0)$ as $r \to \infty$. The definition of an SRBM is given in Section 6.

Theorem 7.1 (Heavy traffic limit theorem) *Consider a sequence of open multiclass queueing networks as defined in Sections 3 and 4 that satisfies the heavy traffic conditions of Section 5 and Assumptions 7.1, 7.2. Let $S = \mathbb{R}_+^J$, $\theta = R\gamma$, $\Gamma = RHR'$ and let ν denote the distribution of $W(0)$. Then*

$$(\hat{W}^r, \hat{X}^r, \hat{Y}^r, \hat{Z}^r) \Rightarrow (W, X, Y, Z) \quad \textit{as } r \to \infty, \tag{90}$$

where $Z = \Delta W$ and (W, X, Y) satisfy the conditions of Definition 6.1, in particular, $W = X + RY$ a.s. and W is an SRBM associated with the data $(S, \theta, \Gamma, R, \nu)$.

8 Proof of the main theorem

Throughout this section we assume that both the conditions of Section 5 and Assumption 7.1 hold for our sequence of open multiclass queueing networks. In particular, by (78) and Assumption 7.1, we have that $G^r = CM^rQ^r\tilde{P}^r\Delta^r \to G$ as $r \to \infty$ where $I + G$ is invertible and $R = (I+G)^{-1}$ is completely-$\mathcal{S}$. Since the properties of invertibility and being completely-$\mathcal{S}$ are preserved under small perturbations, it follows that for all r sufficiently large, $I + G^r$ will be invertible and $R^r = (I + G^r)^{-1}$ will be completely-$\mathcal{S}$. For the purposes of proving Theorem 7.1, we may and do assume henceforth that r is large enough that these conditions hold.

We first state four preliminary results which are used in the proof of our main Theorem 7.1. So as not to disrupt the flow of the proof of that main theorem, we defer the technical proofs of these four results to the end of this section.

Proposition 8.1 *Under the conditions of Section 5 and Assumption 7.1, multiplicative state space collapse implies state space collapse.*

The next result shows that the limits of the fluid scaled arrival and departure processes are deterministic linear processes which move in the k^{th} coordinate direction at the limiting arrival rate λ_k, $k \in \mathcal{K}$.

Lemma 8.2 *Under the conditions of Section 5 and Assumptions 7.1 and 7.2, we have that for each $T \geq 0$,*

$$\|\bar{A}^r(\cdot) - \lambda(\cdot)\|_T \to 0 \quad \textit{and} \quad \|\bar{D}^r(\cdot) - \lambda(\cdot)\|_T \to 0, \tag{91}$$

in probability as $r \to \infty$, where $\lambda(t) = \lambda t$ for all $t \geq 0$.

The next two lemmas are key to the proof that any weak limit point of the sequence of triples $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r)\}$ has the martingale property (iii)(b) of Definition 6.1, which enables us to identify such a limit point as being associated with an SRBM. The first lemma establishes a mild stopping time property for any HL service discipline.

First we need to define the notion of a multiparameter stopping time. Since the process L defined in Section 3 does not appear in the sequel, we shall use the symbol L for another purpose in this and subsequent sections. Recall from Section 3 that $E^r_{\mathcal{A}}$, $U^r_{\mathcal{A}}$ denote the $L \equiv |\mathcal{A}|$-dimensional processes whose components are given by E^r_l, U^r_l, respectively, for $l \in \mathcal{A}$, where $\mathcal{A}$ is the collection of classes for which the associated external arrival processes are non-zero. For notational convenience, by relabelling the classes if necessary, in the following we may and do assume that $\mathcal{A} = \{1, \ldots, L\}$. We let $e_{\mathcal{A}}$ denote the vector in $\mathbb{R}^L$ whose components are all ones. Consider $\mathbb{N}^L \times \mathbb{N}^K$ to be partially ordered such that for $n^1, n^2 \in \mathbb{N}^L \times \mathbb{N}^K$, $n^1 \leq n^2$ if and only if this inequality holds component by component. For each $p \in \mathbb{N}^L$, $q \in \mathbb{N}^K$, let

$$\mathcal{G}^r_{pq} = \sigma\{U^r_{\mathcal{A}}(\,\cdot \wedge (p + e_{\mathcal{A}})), V^{s,r}(\,\cdot \wedge q), \Phi^r(\,\cdot \wedge (Z^r(0) + q)), \mathcal{X}^r(0)\}, \tag{92}$$

where $U^r_{\mathcal{A}}(\,\cdot \wedge (p+e_{\mathcal{A}})) = (U^r_l(\,\cdot \wedge (p_l+1)) : l \in \mathcal{A})$, $V^{s,r}(\,\cdot \wedge q) = (V^{s,r}_k(\,\cdot \wedge q_k) : k \in \mathcal{K})$, and $\Phi^r(\,\cdot \wedge (Z^r(0) + q)) = (\Phi^{k,r}(\,\cdot \wedge (Z^r_k(0) + q_k)) : k \in \mathcal{K})$ and $\mathcal{X}^r(0)$ is defined in Section 3.1.5 (i.e., by (12) with $\ell = 0$, where Z^0, u^0, v^0, O^0 depend on r). Then $\{\mathcal{G}^r_{pq} : (p, q) \in \mathbb{N}^L \times \mathbb{N}^K\}$ is a multiparameter filtration (cf. [27], Section 2.8).

Definition 8.1 *A (multiparameter) stopping time relative to* $\{\mathcal{G}^r_{pq} : (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}$ *is a random variable* τ *taking values in* $\mathbb{N}^L \times \mathbb{N}^K$ *such that*

$$\{\tau = (p,q)\} \in \mathcal{G}^r_{pq} \tag{93}$$

for all $(p,q) \in \mathbb{N}^L \times \mathbb{N}^K$.

Lemma 8.3 *For any HL service discipline used in the* r^{th} *network, we have the following stopping time property. For* $t \geq 0$,

$$\tau^r(t) \equiv (E^r_{\mathcal{A}}(t), A^r(t)) \tag{94}$$

is a (multiparameter) stopping time relative to $\{\mathcal{G}^r_{pq} : (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}$.

Remark. The reader will note that it is easy to verify that $E^r_{\mathcal{A}}(t)$ is a stopping time relative to the filtration $\{\mathcal{G}^r_{p,0} : p \in \mathbb{N}^L\}$. Thus the main emphasis is on proving the property for $A^r(t)$. We include $E^r_{\mathcal{A}}(t)$ in the statement of the stopping time property because this formulation facilitates the proof. The reader will note that $\mathcal{G}^r_{pq}$ has an unsymmetric appearance, in that $e_{\mathcal{A}}$ is added to the argument of $U_{\mathcal{A}}(\cdot)$, but not to the arguments of the service time or routing processes in $\mathcal{G}^r_{pq}$. This occurs because we need to know the first $p_l + 1$ external interarrival times for class l before we can determine whether $E^r_l(t) = p_l$ or not, whereas for $A^r(t)$, once the number of external arrivals up to time t is known for each class, the number of additional arrivals that have occurred up to time t is determined by service times and routing vectors for customers that have already arrived.

Lemma 8.4 *Suppose that the conditions of Section 5 and Assumptions 7.1 and 7.2 hold. In addition, suppose that*

$$\sup_r \sup_{k \in \mathcal{K}} \mathbb{E}[\bar{Z}^r_k(0)] \;<\; \infty. \tag{95}$$

Then, for all r *sufficiently large,* $\hat{X}^r$ *as given by (75) has a decomposition:*

$$\hat{X}^r(\cdot) = \check{X}^r(\cdot) + \check{\epsilon}^r(\cdot), \tag{96}$$

where $\check{X}^r, \check{\epsilon}^r$ *are* J*-dimensional processes satisfying the following conditions (i)–(iii).*

(i) *There is a sequence of constants* $\{\theta^r\}$ *such that* $\theta^r \to \theta$ *as* $r \to \infty$ *and for each* r, $\{\check{X}^r(t) - \check{X}^r(0) - \theta^r t, t \geq 0\}$ *is a martingale with respect to the filtration generated by* $(\hat{W}^r, \check{X}^r, \hat{Y}^r)$.

(ii) *For each* $T \geq 0$, $\|\check{\epsilon}^r(\cdot)\|_T \to 0$ *in probability as* $r \to \infty$.

(iii) *For each* $t \geq 0$*, the collection* $\{\check{X}^r(t) - \check{X}^r(0)\}$ *as indexed by* r *is uniformly integrable.*

Proof of Theorem 7.1. We first prove the Theorem under the L^1-boundedness assumption (95) on $|\bar{Z}^r(0)|$. An outline of the proof for this case is as follows. Using the heavy traffic assumptions of Section 5 (especially the functional central limit theorems for the primitive processes $E^r, V^{s,r}, \Phi^r$), together with the fluid limit result of Lemma 8.2, we show that the process $\hat{\xi}^r$ (cf. (60)) converges in distribution to a Brownian motion as $r \to \infty$. This is then combined with state space collapse (which holds by Assumption 7.2 and Proposition 8.1), to show that $\hat{X}^r$ converges in distribution to a Brownian motion as $r \to \infty$. In addition, from Lemma 8.4 we have that a small perturbation $\check{X}^r$ of $\hat{X}^r$ has a martingale property relative to the filtration generated by $\hat{W}^r, \check{X}^r, \hat{Y}^r$. Using the weak convergence of $\hat{X}^r$ and the aforementioned martingale property, the invariance principle for SRBMs given in Williams [53] can then be invoked to conclude that the sequence $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r)\}$ converges in distribution as $r \to \infty$ to a triple (W, X, Y) which satisfies the conditions of Definition 6.1, and in particular, W is an SRBM. Finally, state space collapse is used to deduce that along with this we have convergence in distribution of $\hat{Z}^r$ to ΔW. We now proceed with the details of the proof assuming that (95) holds.

Note that by Proposition 8.1, state space collapse holds, i.e., $\|\hat{\varepsilon}^r\|_T \to 0$ in probability as $r \to \infty$. Combining this with the functional central limit theorem assumption (86) of Section 5, and the fluid limit result Lemma 8.2, we have by Theorem 4.4 of Billingsley [4] that as $r \to \infty$,

$$(97)\ \left(\hat{W}^r(0), \hat{E}^r(\cdot), \hat{V}^{s,r}(\cdot), \hat{\Phi}^r(\cdot), \bar{A}^r(\cdot), \bar{D}^r(\cdot), \hat{\varepsilon}^r(\cdot)\right) \Rightarrow (W(0), E(\cdot), V^s(\cdot), \Phi(\cdot), \lambda(\cdot), \lambda(\cdot), \mathbf{0}).$$

Since the processes in the right member above have continuous paths, it follows from the random time change theorem (cf. (17.9) of [4]) and the continuous mapping theorem (cf. Theorem 5.1 of [4]), together with the assumed convergence of the deterministic parameters in (78), (80) and Assumption 7.1, that as $r \to \infty$,

$$(98)\quad \hat{X}^r(\cdot) \Rightarrow X(\cdot) \equiv W(0) + R\left(CV^s(\lambda(\cdot)) + CMQ\left(E(\cdot) + \sum_{k=1}^{K} \Phi^k(\lambda_k(\cdot))\right) + \gamma(\cdot)\right),$$

where $\gamma(t) = \gamma t$ for $t \geq 0$, and X is a J-dimensional Brownian motion with drift $\theta = R\gamma$, non-degenerate covariance matrix $\Gamma = RHR'$ and initial distribution ν.

Combining the above with (74), (76) and (77), we see that $\hat{W}^r, \hat{X}^r, \hat{Y}^r$ satisfy conditions (i)–(v) of the SRBM invariance principle given in Theorem 4.1 of [53] with $J, r, \hat{W}^r, \hat{X}^r, \hat{Y}^r$ in place of d, n, W^n, X^n, Y^n and with $\alpha^n = \gamma^n = \mathbf{0}$, $\delta^n = 0$ there. From that theorem (or more precisely, the oscillation inequality on which the first part of the theorem is based), we can conclude that the sequence $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r)\}$ inherits tightness from the sequence $\{\hat{X}^r\}$. In order to identify any weak limit point (W, X, Y) of this sequence as being associated with an SRBM W for the data $(S, \theta, \Gamma, R, \nu)$, it suffices to verify that any such limit point satisfies the martingale property (iii)(b) of Definition 6.1. By Proposition 4.2(III) of [53], a sufficient condition for this is that the conclusion of Lemma 8.4 holds. But, under the hypotheses of Theorem 7.1 and (95), the conclusion indeed holds. It follows (cf. Corollary 4.3 of [53]) that

$(\hat{W}^r, \hat{X}^r, \hat{Y}^r) \Rightarrow (W, X, Y)$ as $r \to \infty$ where (W, X, Y) satisfies the conditions of Definition 6.1 and hence $W = X + RY$ a.s. and W is an SRBM associated with the data $(S, \theta, \Gamma, R, \nu)$. Furthermore, from the state space collapse condition (62) and the convergence of Δ^r to Δ, it follows that joint with the convergence of $(\hat{W}^r, \hat{X}^r, \hat{Y}^r)$ we have $\hat{Z}^r \equiv \Delta^r \hat{W}^r + \hat{\varepsilon}^r \Rightarrow Z = \Delta W$. Thus, Theorem 7.1 has been proved under the additional assumption that (95) holds.

Finally, we reduce the general case to that where (95) holds. For this, we modify the initial queue length in the network indexed by r by defining the new initial queue length to be given by $\tilde{Z}^r(0) = Z^r(0) 1_{\{|Z^r(0)| < r^{3/2}\}}$ and by modifying the initial conditions to be given by

$$\tilde{\mathcal{X}}^r(0) = \mathcal{X}^r(0) 1_{\{|Z^r(0)| < r^{3/2}\}} + (0, u^r(1), 0, 0) 1_{\{|Z^r(0)| \geq r^{3/2}\}}. \tag{99}$$

(Here the zeros in the last term denote zero entries in the appropriate spaces, corresponding to the initial queue length, residual service times, and ordering and age of customers, respectively.) The sequence of networks resulting from this modification of the initial conditions still satisfies all of the model assumptions of Section 3. Moreover, for each r, the performance processes W^r, Y^r, Z^r and the processes A^r, D^r for the modified sequence of networks agree with those for the original sequence on $\{|Z^r(0)| < r^{3/2}\}$ and we have an initial workload of zero on $\{|Z^r(0)| \geq r^{3/2}\}$. Now, by Assumptions 7.1, 7.2, Proposition 8.1, and the assumption that $\hat{W}^r(0) \Rightarrow W(0)$ as $r \to \infty$, it follows that $\hat{Z}^r(0) \Rightarrow \Delta W(0)$ as $r \to \infty$ and hence

$$I\!P(|Z^r(0)| \geq r^{3/2}) \to 0 \quad \text{as } r \to \infty. \tag{100}$$

Using the above and Assumption 7.2, we see that multiplicative state space collapse still holds for the modified sequence of networks and that the diffusion scaled modified initial workload process still converges in distribution to $W(0)$. Furthermore, it is clear that with the modification, condition (95) will be satisfied with $\tilde{Z}^r(0)$ in place of $Z^r(0)$. It then follows from the first part of the proof of this Theorem that the quadruple $(\hat{W}^r, \hat{X}^r, \hat{Y}^r, \hat{Z}^r)$ associated with the modified sequence of networks converges in distribution to (W, X, Y, Z) where $Z = \Delta W$, $W = X + RY$ a.s. and W is an SRBM associated with $(S, \theta, \Gamma, R, \nu)$. However, since the set $\{|Z^r(0)| \geq r^{3/2}\}$ on which this quadruple may differ from the original one has probability going to zero as $r \to \infty$, it follows that convergence in distribution to the same limit also holds for the original sequence $\{(\hat{W}^r, \hat{X}^r, \hat{Y}^r, \hat{Z}^r)\}$. □

Proof of Proposition 8.1. Recall that r is assumed to be sufficiently large that $I + G^r$ is invertible and $R^r = (I + G^r)^{-1}$ is completely-$\mathcal{S}$.

The idea of the proof is to show that for each $T \geq 0$, $\|\hat{W}^r\|_T$ is bounded in probability as r varies, so that state space collapse can be deduced from multiplicative state space collapse. For this, an oscillation inequality from [53] is used in conjunction with multiplicative state space collapse to estimate $\|\hat{W}^r\|_T$ in terms of $|\hat{W}^r(0)|$ and $\|\hat{\xi}^r\|_T$. We estimate the latter by overestimating the arrival and departure processes appearing in (60). This estimation of $\hat{\xi}^r$ will also be used in the proof of Lemma 8.2. Now we proceed with the details of the proof.

For each $T \geq 0$, let

$$\hat{\zeta}_T^r(t) = \frac{\hat{Z}^r(t) - \Delta^r \hat{W}^r(t)}{\|\hat{W}^r\|_T \vee 1} \quad \text{for all } t \in [0, T]. \tag{101}$$

From the assumption of multiplicative state space collapse we have that for each $T \geq 0$,

$$\|\hat{\zeta}_T^r\|_T \to 0 \quad \text{in probability as } r \to \infty. \tag{102}$$

Note that by (87),

$$\bar{W}^r(0) = r^{-1}\hat{W}^r(0) \to 0 \quad \text{in probability as } r \to \infty. \tag{103}$$

From the definition of $\hat{\varepsilon}^r(\cdot)$, we have

$$\hat{\varepsilon}^r(t) = \hat{\zeta}_T^r(t)(\|\hat{W}^r\|_T \vee 1) \quad \text{for all } 0 \leq t \leq T, \tag{104}$$

and

$$\bar{Z}^r(0) = \Delta^r \bar{W}^r(0) + r^{-1}\hat{\varepsilon}^r(0) = \Delta^r \bar{W}^r(0) + \hat{\zeta}_0^r(0)(|\bar{W}^r(0)| \vee r^{-1}). \tag{105}$$

On combining this with (102), (103), and the convergence of Δ^r to Δ, we obtain

$$\bar{Z}^r(0) \to 0 \quad \text{in probability as } r \to \infty. \tag{106}$$

Now we shall estimate $\bar{A}^r$. From the fluid scaled equations (43) and (49), we have

$$\bar{A}^r(t) = \bar{E}^r(t) + \sum_{k=1}^{K} \bar{\Phi}^{k,r}(\bar{D}_k^r(t)). \tag{107}$$

Now $D_k^r(t)$ (prior to fluid scaling) is the number of class k customers served up to time t and this is less than or equal to the initial number of customers in class k (namely $Z_k^r(0)$) plus the number (denoted by $S_k^r(t)$) of subsequent customers that could be served if the server at station $s(k)$ devoted all of its attention to subsequent class k customers and was never idle in $[0, t]$. Hence (107) yields

$$\bar{A}^r(t) \leq \bar{E}^r(t) + \sum_{k=1}^{K} \bar{\Phi}^{k,r}(\bar{Z}_k^r(0) + \bar{S}_k^r(t)) \equiv \bar{B}^r(t), \tag{108}$$

where $\bar{S}_k^r(t) = r^{-2}S_k^r(r^2 t)$. Now, under any service discipline that processes customers within each class in FIFO order we have

$$S_k^r(t) = \max\{n \geq 0 : V_k^{s,r}(n) \leq t\} \quad \text{for each } k \in \mathcal{K}. \tag{109}$$

It then follows (cf. Iglehart and Whitt [36]) from the functional weak law of large numbers (87) for $\bar{V}^{s,r}$ that for any $T \geq 0$, the fluid scaled renewal process $\bar{S}^r(\,\cdot\,) \equiv r^{-2}S^r(r^2\,\cdot\,)$ satisfies

$$\|\bar{S}^r(\cdot) - \mu(\cdot)\|_T \to 0 \quad \text{in probability as } r \to \infty, \tag{110}$$

where $\mu_k(t) \equiv m_k^{-1} t$ for all $t \geq 0$ and $k \in \mathcal{K}$. Combining the above with the functional weak law of large numbers (87), (106), and Billingsley's random time change result (see (17.9) of [4]), we have for each $T \geq 0$,

$$\left\| \bar{B}^r(\cdot) - \alpha(\cdot) - \sum_{k=1}^{K} \tilde{P}^k \mu_k(\cdot) \right\|_T \to 0 \text{ in probability as } r \to \infty, \tag{111}$$

where $\alpha(t) = \alpha t$ for all $t \geq 0$.

Now we shall use the above and the functional central limit theorem assumptions in (86) to estimate $\hat{\xi}^r$. By (60), (108) and the fact that $\bar{D}_k^r(t) \leq \bar{Z}_k^r(0) + \bar{A}_k^r(t)$ for all $t \geq 0$, on setting $\bar{\beta}^r(t) = \max_{k=1}^K (\bar{Z}_k^r(0) + \bar{B}_k^r(t))$ for all $t \geq 0$, we have for any $T \geq 0$,

$$\begin{aligned} \|\hat{\xi}^r\|_T \quad &\leq \quad |C| \|\hat{V}^{s,r}\|_{\bar{\beta}^r(T)} + |CM^r Q^r| K \max_{k=1}^K \|\hat{\Phi}^{k,r}\|_{\bar{\beta}^r(T)} \\ &\quad + |CM^r Q^r| \, \|\hat{E}^r\|_T + |\hat{\gamma}^r| T. \end{aligned} \tag{112}$$

Now, by (78), (80), the continuous mapping theorem, (86), (106), (111), and the random time change theorem, we have that the right member of (112) converges in distribution as $r \to \infty$ to

$$|C| \|V^s\|_{\beta(T)} + |CMQ| K \max_{k=1}^K \|\Phi^k\|_{\beta(T)} + |CMQ| \, \|E\|_T + |\gamma| T, \tag{113}$$

where $\beta(T) = \max_{l=1}^K |\alpha_l + \sum_{k=1}^K P_{kl} \mu_k| T$.

Now from (67), (68), (74), (75) and (104), we have for each $T \geq 0$ and all $t \in [0, T]$,

$$\hat{W}^r(t) = \hat{X}^r(t) + R^r \hat{Y}^r(t) \tag{114}$$

where

$$\hat{X}^r(t) = \hat{W}^r(0) + R^r \hat{\xi}^r(t) + R^r C M^r Q^r \tilde{P}^r (\hat{\zeta}_T^r(0) - \hat{\zeta}_T^r(t))(\|\hat{W}^r\|_T \vee 1). \tag{115}$$

We may rewrite (114) as

$$\hat{W}^r(t) = \hat{X}^r(t) + (R^r - R)\hat{Y}^r(t) + R\hat{Y}^r(t), \tag{116}$$

where $\hat{W}^r, \hat{Y}^r$ satisfy (76), (77). Since R is completely-$\mathcal{S}$, it then follows from the oscillation inequality given in Theorem 5.1 of [53] that there is a constant C_R that depends only on R such that for each $T \geq 0$,

$$\mathrm{Osc}(\hat{Y}^r, [0,T]) \leq C_R \, \mathrm{Osc}(\hat{X}^r(\cdot) + (R^r - R)\hat{Y}^r(\cdot), [0,T]) \tag{117}$$

$$\mathrm{Osc}(\hat{W}^r, [0,T]) \leq C_R \, \mathrm{Osc}(\hat{X}^r(\cdot) + (R^r - R)\hat{Y}^r(\cdot), [0,T]) \tag{118}$$

where $\mathrm{Osc}(w, [0,T]) \equiv \sup\{|w(t) - w(s)| : 0 \leq s < t \leq T\}$ for any $w \in \mathbf{D}^J$. It then follows as in Section 6 of [53], that for all r such that $C_R |R^r - R| < \frac{1}{2}$,

$$\mathrm{Osc}(\hat{Y}^r, [0,T]) \leq 2C_R \mathrm{Osc}(\hat{X}^r, [0,T]) \tag{119}$$

$$\mathrm{Osc}(\hat{W}^r, [0,T]) \leq 2C_R \mathrm{Osc}(\hat{X}^r, [0,T]). \tag{120}$$

On substituting the expression for $\hat{X}^r$ from (115) in the above, we obtain

$$\begin{aligned}\mathrm{Osc}(\hat{W}^r,[0,T]) &\le 2C_R\,\mathrm{Osc}(R^r\hat{\xi}^r(\cdot)+R^rCM^rQ^r\tilde{P}^r(\hat{\zeta}^r_T(0)-\hat{\zeta}^r_T(\cdot))(\|\hat{W}^r\|_T\vee 1),[0,T]) \\ &\le 2C_R|R^r|\mathrm{Osc}(\hat{\xi}^r,[0,T]) \\ &\quad +2C_R|R^rCM^rQ^r\tilde{P}^r|(\|\hat{W}^r\|_T\vee 1)\mathrm{Osc}(\hat{\zeta}^r_T,[0,T]).\end{aligned}$$

Hence

$$\begin{aligned}\|\hat{W}^r\|_T &\le |\hat{W}^r(0)|+\mathrm{Osc}(\hat{W}^r,[0,T]) \\ &\le |\hat{W}^r(0)|+\ 4C_R|R^r|\|\hat{\xi}^r\|_T \\ &\quad +4C_R|R^rCM^rQ^r\tilde{P}^r|\,(\|\hat{W}^r\|_T\vee 1)\,\|\hat{\zeta}^r_T\|_T.\end{aligned}\tag{121}$$

Fix $T\ge 0$ and $\varepsilon>0$. Then given the convergence of $R^r, M^r, Q^r, \tilde{P}^r$ and $\|\hat{\zeta}^r_T\|_T$, we can find an r_0 such that for all $r\ge r_0$, (121) holds and

$$I\!P\left(4C_R|R^rCM^rQ^r\tilde{P}^r|\,\|\hat{\zeta}^r_T\|_T>\frac{1}{2}\right)\le\frac{\varepsilon}{6}.\tag{122}$$

Then it follows from (121) that for each $r\ge r_0$,

$$I\!P(\|\hat{W}^r\|_T\vee 1\le 2(|\hat{W}^r(0)|+4C_R|R^r|\|\hat{\xi}^r\|_T+1))\ge 1-\frac{\varepsilon}{6}.\tag{123}$$

Now by the properties of Brownian motion, there is $\kappa_\varepsilon>0$ such that (113) is bounded by κ_ε with probability at least $1-\frac{\varepsilon}{12}$ and since the right member of (112) converges in distribution to (113), it follows that there is $r_1\ge r_0$ such that for all $r\ge r_1$,

$$I\!P(\|\hat{\xi}^r\|_T\le 2\kappa_\varepsilon)\ge 1-\frac{\varepsilon}{6}.\tag{124}$$

Finally, since $\hat{W}^r(0)$ converges in distribution, there is $\hat{\kappa}_\varepsilon\ge 2\kappa_\varepsilon$ and $r_2\ge r_1$ such that for all $r\ge r_2$,

$$I\!P(|\hat{W}^r(0)|\le\hat{\kappa}_\varepsilon)\ge 1-\frac{\varepsilon}{6}.\tag{125}$$

Combining the above we see that for each $r\ge r_2$,

$$I\!P(\|\hat{W}^r\|_T\vee 1\le C_{R,\varepsilon})\ge 1-\frac{\varepsilon}{2},\tag{126}$$

where $C_{R,\varepsilon}=2(\hat{\kappa}_\varepsilon+8C_R\sup_{r\ge r_2}|R^r|\hat{\kappa}_\varepsilon+1)$. Now by (102) there is $r_3\ge r_2$ such that for all $r\ge r_3$,

$$I\!P(\|\hat{\zeta}^r_T\|_T\ge\varepsilon/C_{R,\varepsilon})\le\frac{\varepsilon}{2}.\tag{127}$$

It then follows that for all $r\ge r_3$,

$$I\!P(\|\hat{Z}^r(\cdot)-\Delta^r\hat{W}^r(\cdot)\|_T\ge\varepsilon)\ \le\ I\!P(\|\hat{\zeta}^r_T\|_T\ge\varepsilon/C_{R,\varepsilon})+I\!P(\|\hat{W}^r\|_T\vee 1>C_{R,\varepsilon})\le\varepsilon.$$

Since $\varepsilon>0$ and $T\ge 0$ were arbitrary, this completes the proof that state space collapse holds. □

Proof of Lemma 8.2. The method of proof for this lemma follows an idea suggested by Maury Bramson.

Our proof borrows heavily from the proof of Proposition 8.1. Briefly the idea of our proof is as follows. Using the oscillation inequality (120) together with the estimate (112) and state space collapse, we show that $\|\bar{W}^r\|_T \to 0$ in probability as $r \to \infty$. Then, by combining this with state space collapse and (48), we show that the difference between the fluid scaled arrival and departure processes goes to zero (uniformly on compact time intervals) in probability as $r \to \infty$. Upon substituting this into the expression (107), we obtain an implicit equation for $\bar{A}^r$ which can be solved to yield the asymptotic behavior of $\bar{A}^r$ as $r \to \infty$. We now provide the details of the proof.

We suppose that r is large enough that the oscillation inequality (120) holds. Upon dividing this inequality through by r and recalling the definition of $\hat{X}^r$ from (75) we have that for any $T \geq 0$,

$$\mathrm{Osc}(\bar{W}^r, [0,T]) \leq 2C_R \mathrm{Osc}(\bar{X}^r, [0,T]), \tag{128}$$

where

$$\bar{X}^r(t) = \bar{W}^r(0) + R^r(\bar{\xi}^r(t) + \bar{\eta}^r(t)), \tag{129}$$

$\bar{\xi}^r(\cdot) = r^{-1}\hat{\xi}^r(\cdot)$, $\bar{\eta}^r(\cdot) = r^{-1}\hat{\eta}^r(\cdot)$, and $\hat{\xi}^r, \hat{\eta}^r$ are defined by (60), (68), respectively. From (87), we have that $\bar{W}^r(0) \to 0$ in probability as $r \to \infty$. By the a priori estimate (112) on $\hat{\xi}^r$ and the convergence in distribution of the right member of (112) to (113), it follows that for each $T \geq 0$,

$$\|\bar{\xi}^r(\cdot)\|_T \to 0 \text{ in probability as } r \to \infty. \tag{130}$$

Furthermore, by Proposition 8.1 we have that state space collapse holds, i.e., for each $T \geq 0$,

$$\|\hat{\varepsilon}^r\|_T \to 0 \quad \text{in probability as } r \to \infty. \tag{131}$$

On combining this with the definition of $\hat{\eta}^r$ and the convergence of the model parameters given in (78), we have that $\|\hat{\eta}^r\|_T \to 0$ in probability as $r \to \infty$. Then, by combining the above with (129) and the convergence of R^r to R, we conclude that for each $T \geq 0$,

$$\|\bar{X}^r\|_T \to 0 \quad \text{in probability as } r \to \infty. \tag{132}$$

It then follows from (128) that for each $T \geq 0$,

$$\mathrm{Osc}(\bar{W}^r, [0,T]) \to 0 \quad \text{in probability as } r \to \infty. \tag{133}$$

Then by (48) and (67), we have for each $T \geq 0$,

$$\begin{aligned}
\|\bar{A}^r(\cdot) - \bar{D}^r(\cdot)\|_T &= \|\bar{Z}^r(\cdot) - \bar{Z}^r(0)\|_T && (134)\\
&\leq |\Delta^r| \|\bar{W}^r(\cdot) - \bar{W}^r(0)\|_T + r^{-1}\|\hat{\varepsilon}^r(\cdot) - \hat{\varepsilon}(0)\|_T && (135)\\
&\leq |\Delta^r| \mathrm{Osc}(\bar{W}^r, [0,T]) + r^{-1}\|\hat{\varepsilon}^r(\cdot) - \hat{\varepsilon}^r(0)\|_T, && (136)
\end{aligned}$$

where the last line goes to zero in probability as $r \to \infty$, by Assumption 7.1, (133) and (131). Hence,

$$\|\bar{A}^r(\cdot) - \bar{D}^r(\cdot)\|_T \to 0 \text{ in probability as } r \to \infty. \tag{137}$$

From (107) and the definitions of $\bar{\Phi}^{k,r}$ and $\hat{\Phi}^{k,r}$ we have

$$\begin{aligned} \bar{A}^r(t) &= \bar{E}^r(t) + \sum_{k=1}^K \left(\tilde{P}^{k,r}\bar{D}^r_k(t) + r^{-1}\hat{\Phi}^{k,r}(\bar{D}^r_k(t))\right) \\ &= \bar{E}^r(t) + \sum_{k=1}^K \tilde{P}^{k,r}\bar{A}^r_k(t) + \bar{\zeta}^r(t), \end{aligned} \tag{138}$$

where

$$\bar{\zeta}^r(t) = \sum_{k=1}^K \left(\tilde{P}^{k,r}\left(\bar{D}^r_k(t) - \bar{A}^r_k(t)\right) + r^{-1}\hat{\Phi}^{k,r}(\bar{D}^r_k(t))\right). \tag{139}$$

Now, $\hat{\Phi}^r(\cdot)$ converges in distribution to $\Phi(\cdot)$ as $r \to \infty$, and $\bar{D}^r(t) \le \bar{Z}^r(0) + \bar{B}^r(t)$ for all $t \ge 0$, where $\bar{B}^r$ is defined in (108). It then follows by the same kind of reasoning as that which led to (130) that for each $T \ge 0$ and $k \in \mathcal{K}$,

$$\|r^{-1}\hat{\Phi}^{k,r}(\bar{D}^r_k(\cdot))\|_T \to 0 \quad \text{in probability as } r \to \infty. \tag{140}$$

Combining this with (78) and (137), it follows that for each $T \ge 0$,

$$\|\bar{\zeta}^r\|_T \to 0 \quad \text{in probability as } r \to \infty. \tag{141}$$

Solving for $\bar{A}^r$ in (138), we have

$$\bar{A}^r(t) = Q^r\left(\bar{E}^r(t) + \bar{\zeta}^r(t)\right). \tag{142}$$

From the known convergence (cf. (78), (79), (87) and (141)) of the entities in the right member above we have for each $T \ge 0$,

$$\|\bar{A}^r(\cdot) - Q\alpha(\cdot)\|_T \to 0 \text{ in probability as } r \to \infty. \tag{143}$$

Since $\lambda(\cdot) = Q\alpha(\cdot)$, this establishes the desired result for $\bar{A}^r$ and the result for $\bar{D}^r$ follows from this by (137). ▫

Proof of Lemma 8.3. This proof is for fixed r and so the superscript r will be suppressed in the following. Also, for this proof only, m will denote a member of $\mathbb{N}^L$, rather than the vector of mean service times.

As in Section 3.1.5, we define a strictly increasing sequence of real-valued random times $\{\sigma_\ell\}_{\ell=0}^\infty$ for the (discrete event) queueing network model such that $\sigma_0 \equiv 0$ and for $\ell = 1, 2, \ldots$, σ_ℓ is the time of the ℓ^{th} change in the status of the arrival-departure process pair (A, D), i.e., σ_ℓ is the ℓ^{th} time of occurrence of an arrival to, or a departure from, some class. Note that at a given σ_ℓ, simultaneous departures and arrivals are allowed (where for simultaneous arrivals to a given station the tie breaking rule described in Section 3.1.5 is invoked). We have $\sigma_\ell < \infty$ for each ℓ, and $\sigma_\ell \to \infty$ as $\ell \to \infty$. (We remind the reader that, without loss of generality for the proof of Theorem 7.1, it was assumed in Section 3.1.5 that any exceptional null sets where these properties may not hold have been removed a priori from the probability space.) Then, for each $t \geq 0$, $p \in \mathbb{N}^L$, $q \in \mathbb{N}^K$,

$$\{E_{\mathcal{A}}(t) = p, A(t) = q\} = \bigcup_{i=0}^{\infty} \bigcap_{\ell=i}^{\infty} \{E_{\mathcal{A}}(t \wedge \sigma_\ell) = p, A(t \wedge \sigma_\ell) = q\}. \tag{144}$$

Thus it suffices to show that for each $t \geq 0$, $\ell \geq 0$, $p \in \mathbb{N}^L$ and $q \in \mathbb{N}^K$,

$$B^\ell_{pq} \equiv \{E_{\mathcal{A}}(t \wedge \sigma_\ell) = p, A(t \wedge \sigma_\ell) = q\} \in \mathcal{G}_{pq}. \tag{145}$$

For this, fix $t \geq 0$. It will be shown by induction that for each $\ell \geq 0$, the following two properties hold for all $p \in \mathbb{N}^L$, $q \in \mathbb{N}^K$:

(i) $B^\ell_{pq} \in \mathcal{G}_{pq}$,

(ii) for $\mathcal{I}^\ell \equiv (t \wedge \sigma_\ell, E(\cdot \wedge t \wedge \sigma_\ell), A(\cdot \wedge t \wedge \sigma_\ell), D(\cdot \wedge t \wedge \sigma_\ell), \mathcal{X}(\cdot \wedge t \wedge \sigma_\ell))$, we have $1_{B^\ell_{pq}} \mathcal{I}^\ell \in \mathcal{G}_{pq}$.

Recall, from Section 3.1.5, the definition of $\mathcal{X}(\cdot) = \{\mathcal{X}(t), t \geq 0\}$ where

$$\mathcal{X}(t) = \sum_{\ell=0}^{\infty} 1_{\{\sigma_\ell \leq t < \sigma_{\ell+1}\}} (Z^\ell, u^\ell, v^\ell, O^\ell) \quad \text{for all } t \geq 0. \tag{146}$$

We now proceed with the induction proof. For $\ell = 0$, one has $\sigma_0 = 0$, $E_{\mathcal{A}}(0) = 0$ and $A(0) = D(0) = 0$, by definition. Moreover, $\mathcal{X}(0) \in \mathcal{G}_{pq}$ for all p, q. Then (i) and (ii) are easily verified to hold for $\ell = 0$.

For the induction step, assume that (i)–(ii) hold for all $p \in \mathbb{N}^L$, $q \in \mathbb{N}^K$, for some $\ell \geq 0$. Now,

$$B^{\ell+1}_{pq} = \bigcup_{m,n} \left(B^{\ell+1}_{pq} \cap B^\ell_{mn} \right), \tag{147}$$

where the union is over all $(m, n) \in \mathbb{N}^L \times \mathbb{N}^K$ such that $m \leq p$ and $n \leq q$. By the induction assumption, for fixed $(p, q) \in \mathbb{N}^L \times \mathbb{N}^K$ and any $(m, n) \in \mathbb{N}^L \times \mathbb{N}^K$ such that $m \leq p, n \leq q$, we have

$$B^\ell_{mn} \in \mathcal{G}_{mn}, \qquad 1_{B^\ell_{mn}} \mathcal{I}^\ell \in \mathcal{G}_{mn} \tag{148}$$

and hence $B^\ell_{mn} \cap \{\sigma_\ell \geq t\} \in \mathcal{G}_{mn}$ and $B^\ell_{mn} \cap \{\sigma_\ell < t\} \in \mathcal{G}_{mn}$.

Now, on $B^\ell_{mn} \cap \{\sigma_\ell \geq t\}$, $\sigma_{\ell+1} \wedge t = \sigma_\ell \wedge t$, $E_{\mathcal{A}}(t \wedge \sigma_{\ell+1}) = E_{\mathcal{A}}(t \wedge \sigma_\ell) = m$, $A(t \wedge \sigma_{\ell+1}) = A(t \wedge \sigma_\ell) = n$ and $\mathcal{I}^{\ell+1} = \mathcal{I}^\ell$. Thus, if $(m, n) = (p, q)$ we have

$$B^{\ell+1}_{pq} \cap B^\ell_{mn} \cap \{\sigma_\ell \geq t\} = B^\ell_{mn} \cap \{\sigma_\ell \geq t\} \in \mathcal{G}_{mn}, \tag{149}$$

or if $(m,n) \neq (p,q)$, then the left member of (149) is the empty set which is still in $\mathcal{G}_{mn}$. Thus, combining the above with the induction assumption (148) we obtain

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell \geq t\}} \mathcal{I}^{\ell+1} = 1_{\{(m,n)=(p,q)\}} 1_{B_{mn}^{\ell} \cap \{\sigma_\ell \geq t\}} \mathcal{I}^{\ell} \in \mathcal{G}_{mn}. \tag{150}$$

On the other hand, on $B_{mn}^{\ell} \cap \{\sigma_\ell < t\}$, $E_{\mathcal{A}}(\sigma_\ell) = E_{\mathcal{A}}(t \wedge \sigma_\ell) = m$ and the first time after σ_ℓ that a new external arrival occurs is $\sigma = \min_{l \in \mathcal{A}} U_l(m_l + 1)$. Furthermore, on the set $B_{mn}^{\ell} \cap \{\sigma_\ell < t\}$, we have

$$\mathcal{I}^{\ell} = (\sigma_\ell, E(\cdot \wedge \sigma_\ell), A(\cdot \wedge \sigma_\ell), D(\cdot \wedge \sigma_\ell), \mathcal{X}(\cdot \wedge \sigma_\ell)). \tag{151}$$

Recall that $\mathcal{X}(\sigma_\ell) = \mathcal{X}^{\ell} = (Z^{\ell}, u^{\ell}, v^{\ell}, O^{\ell})$ and the proportions of service time for each of the classes over $[\sigma_\ell, \sigma_{\ell+1})$ are given by $r^{\ell} = \Psi(\mathcal{X}^{\ell})$ where Ψ is a measurable function (cf. Section 3.1.5). It follows that if we define

$$\eta = \sigma_\ell + \inf\{s \geq 0 : v_k^{\ell}(1) - r_k^{\ell} s = 0 \quad \text{for some } k \text{ such that } r_k^{\ell} > 0\}, \tag{152}$$

then on $B_{mn}^{\ell} \cap \{\sigma_\ell < t\}$, $\sigma_{\ell+1} = \sigma \wedge \eta$ where the latter is a measurable function of $(U_{\mathcal{A}}(\cdot \wedge (m + e_{\mathcal{A}})), \sigma_\ell, \mathcal{X}^{\ell})$, and hence by the induction assumption (148), (151), and the definition of $\mathcal{G}_{mn}$ we have $1_{B_{mn}^{\ell} \cap \{\sigma_\ell < t\}} \sigma_{\ell+1} \in \mathcal{G}_{mn}$. Moreover, on $B_{mn}^{\ell} \cap \{\sigma_\ell < t\}$, since no external interarrival or service time is zero, we have

$$D(\sigma_{\ell+1}) \leq Z(0) + A(\sigma_\ell) = Z(0) + n, \tag{153}$$

and on this set we can express $E(\sigma_{\ell+1})$, $D(\sigma_{\ell+1})$, $A(\sigma_{\ell+1})$, $Z(\sigma_{\ell+1})$ as measurable functions of σ_ℓ, $\sigma_{\ell+1}$, $E(\sigma_\ell)$, $D(\sigma_\ell)$, $U_{\mathcal{A}}(m + e_{\mathcal{A}})$, $\Phi(\cdot \wedge (Z(0) + n))$, $\mathcal{X}^{\ell}$, $Z(0)$, as follows. For $l \in \mathcal{A}$, $k \in \mathcal{K}$,

$$\begin{aligned}
E_l(\sigma_{\ell+1}) &= E_l(\sigma_\ell) + 1_{\{U_l(m_l+1) = \sigma_{\ell+1}\}}, \\
D_k(\sigma_{\ell+1}) &= D_k(\sigma_\ell) + 1_{\{v_k^{\ell}(1) = r_k^{\ell}(\sigma_{\ell+1} - \sigma_\ell), r_k^{\ell} > 0\}}, \\
A_k(\sigma_{\ell+1}) &= E_k(\sigma_{\ell+1}) + \sum_{i=1}^{K} \Phi_k^i(D_i(\sigma_{\ell+1})), \\
Z_k(\sigma_{\ell+1}) &= Z_k(0) + A_k(\sigma_{\ell+1}) - D_k(\sigma_{\ell+1}).
\end{aligned}$$

(Note that $E_k \equiv 0$ if $k \notin \mathcal{A}$.) Since E, D, A, Z are constant on $[\sigma_\ell, \sigma_{\ell+1})$, on combining the above with the induction assumption (148) and (151), we have that

$$1_{B_{mn}^{\ell} \cap \{\sigma_\ell < t\}} (\sigma_{\ell+1}, E(\cdot \wedge \sigma_{\ell+1}), A(\cdot \wedge \sigma_{\ell+1}), D(\cdot \wedge \sigma_{\ell+1}), Z(\sigma_{\ell+1})) \in \mathcal{G}_{mn}. \tag{154}$$

In particular,

$$1_{B_{mn}^{\ell} \cap \{\sigma_\ell < t\}} (E_{\mathcal{A}}(t \wedge \sigma_{\ell+1}), A(t \wedge \sigma_{\ell+1})) \in \mathcal{G}_{mn} \tag{155}$$

and hence

$$B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell < t\} \in \mathcal{G}_{mn}. \tag{156}$$

On combining this with (149) we see that $B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \in \mathcal{G}_{mn} \subset \mathcal{G}_{pq}$ and hence by (147),

$$B_{pq}^{\ell+1} \in \mathcal{G}_{pq}. \tag{157}$$

Thus, (i) holds with $\ell+1$ in place of ℓ. Similarly,

$$B_{pq}^{\ell+1} \cap \{\sigma_{\ell+1} \le t\} = \bigcup_{m,n} \left(B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell < t\} \cap \{\sigma_{\ell+1} \le t\}\right) \in \mathcal{G}_{pq}, \tag{158}$$

where the union is over all $(m,n) \in \mathbb{N}^L \times \mathbb{N}^K$ such that $m \le p, n \le q$.

It remains to verify (ii) with $\ell+1$ in place of ℓ. Now, by (147), (150), (154), (156), and the definition of $\mathcal{I}^{\ell+1}$, it suffices to verify that

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell < t\}} \mathcal{X}(\cdot \wedge t \wedge \sigma_{\ell+1}) \in \mathcal{G}_{pq}. \tag{159}$$

Furthermore, by (148), (151), (154) and (156), we have for $m \le p, n \le q$,

$$\begin{aligned} &1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell < t < \sigma_{\ell+1}\}} \mathcal{X}(\cdot \wedge t \wedge \sigma_{\ell+1}) \\ &= 1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell < t < \sigma_{\ell+1}\}} \mathcal{X}(\cdot \wedge t \wedge \sigma_\ell) \in \mathcal{G}_{mn} \subset \mathcal{G}_{pq}. \end{aligned} \tag{160}$$

Thus, it remains to show for each pair $(m,n) \le (p,q)$ that

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_\ell < t\} \cap \{\sigma_{\ell+1} \le t\}} \mathcal{X}(\cdot \wedge \sigma_{\ell+1}) \in \mathcal{G}_{pq}. \tag{161}$$

Indeed, in view of (148), (151), (154) and (156), and the fact that $\mathcal{X}(\cdot)$ is constant on $[\sigma_\ell, \sigma_{\ell+1})$ and $\sigma_\ell < \sigma_{\ell+1}$, for this it suffices to verify that

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_{\ell+1} \le t\}} (Z^{\ell+1}, u^{\ell+1}, v^{\ell+1}, O^{\ell+1}) \in \mathcal{G}_{pq}. \tag{162}$$

This is done below. In fact, the result for the first component (that involving $Z^{\ell+1}$) follows from (154). Although the idea for the other components is simple enough, namely, to write $u^{\ell+1}$, $v^{\ell+1}$ and $O^{\ell+1}$ as measurable functions of $Z^\ell, v^\ell, O^\ell, r^\ell, \sigma_\ell, \sigma_{\ell+1}, U_{\mathcal{A}}(\cdot \wedge (E_{\mathcal{A}}(\sigma_{\ell+1}) + e_{\mathcal{A}})), A(\cdot \wedge \sigma_{\ell+1}), D(\cdot \wedge \sigma_{\ell+1}), V^s(\cdot \wedge A(\sigma_{\ell+1}))$, a variety of details are needed to take account of the various changes in the queue lengths that might occur at $\sigma_{\ell+1}$.

Fix $(m,n) \le (p,q)$. Now,

$$u^{\ell+1} = U_{\mathcal{A}}(E_{\mathcal{A}}(\sigma_{\ell+1}) + e_{\mathcal{A}}) - \sigma_{\ell+1} e_{\mathcal{A}}, \tag{163}$$

and on $B_{pq}^{\ell+1} \cap \{\sigma^{\ell+1} \le t\}$,

$$u^{\ell+1} = U_{\mathcal{A}}(p + e_{\mathcal{A}}) - \sigma_{\ell+1} e_{\mathcal{A}}. \tag{164}$$

Then by the definition of $\mathcal{G}_{pq}$, (154) and (158), it follows that

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^{\ell} \cap \{\sigma_{\ell+1} \le t\}} u^{\ell+1} \in \mathcal{G}_{pq}. \tag{165}$$

Now, for $v^{\ell+1}$, note that a customer will depart from class k at the time $\sigma_{\ell+1}$ if and only if $v_k^\ell(1) - r_k^\ell(\sigma_{\ell+1} - \sigma_\ell) = 0$ and $r_k^\ell > 0$. On the other hand, at least one new customer enters class k at the time $\sigma_{\ell+1}$ if and only if $A_k(\sigma_{\ell+1}) > A_k(\sigma_\ell)$. We use these observations to express $v_k^{\ell+1}$ as a measurable function of $Z^\ell = Z(\sigma_\ell)$, v^ℓ, r^ℓ, σ_ℓ, $\sigma_{\ell+1}$, $A(\cdot \wedge \sigma_{\ell+1})$, $V^s(\cdot \wedge A(\sigma_{\ell+1}))$, as follows. For $k \in \mathcal{K}$,

$$v_k^{\ell+1}(i) = \begin{cases} v_k^\ell(i) & \text{for } 1 \le i \le Z_k^\ell, \text{ if } r_k^\ell = 0, \\ v_k^\ell(1) - r_k^\ell(\sigma_{\ell+1} - \sigma_\ell) & \text{for } i = 1, \text{ if } r_k^\ell > 0 \text{ and } v_k^\ell(1) - r_k^\ell(\sigma_{\ell+1} - \sigma_\ell) > 0, \\ v_k^\ell(i) & \text{for } 2 \le i \le Z_k^\ell, \text{ if } r_k^\ell > 0 \text{ and } v_k^\ell(1) - r_k^\ell(\sigma_{\ell+1} - \sigma_\ell) > 0, \\ v_k^\ell(i+1) & \text{for } 1 \le i \le Z_k^\ell - 1, \text{ if } r_k^\ell > 0 \text{ and } v_k^\ell(1) = r_k^\ell(\sigma_{\ell+1} - \sigma_\ell), \\ v_k^s(A_k(\sigma_\ell) + j) & \text{for } i = Z_k^\ell + j - 1_{\{r_k^\ell > 0,\ v_k^\ell(1) = r_k^\ell(\sigma_{\ell+1} - \sigma_\ell)\}} \text{ and} \\ & \qquad 0 < j \le A_k(\sigma_{\ell+1}) - A_k(\sigma_\ell), \\ 0 & \text{for } i > Z_k^\ell + A_k(\sigma_{\ell+1}) - A_k(\sigma_\ell) - 1_{\{r_k^\ell > 0, v_k^\ell(1) = r_k^\ell(\sigma_{\ell+1} - \sigma_\ell)\}}. \end{cases}$$

The first four options above deal with potential departures from class k. The first option corresponds to the situation where class k receives no service in $[\sigma_\ell, \sigma_{\ell+1})$, the second and third options correspond to the situation where the customer at the head of the line for class k receives service in $[\sigma_\ell, \sigma_{\ell+1})$ but the service is not completed by the end of this time interval, and the fourth option corresponds to the situation where a class k service completion occurs at the time $\sigma_{\ell+1}$. The fifth option takes account of new arrivals to class k occurring at $\sigma_{\ell+1}$ and the sixth option is simply a convention that $v_k^{\ell+1}(i)$ is defined to be zero when i exceeds the number of customers in class k at the time $\sigma_{\ell+1}$. Recall that $v^{\ell+1} = \{(v_1^{\ell+1}(i), \ldots, v_K^{\ell+1}(i))\}_{i=1}^\infty$. Using (148), (151), (154), (157) and the facts that $A_k(\sigma_{\ell+1}) = q$ on $B_{pq}^{\ell+1} \cap \{\sigma_{\ell+1} \le t\}$ and $V^s(\cdot \wedge q) \in \mathcal{G}_{pq}$, it follows from the above that

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^\ell \cap \{\sigma_{\ell+1} \le t\}} v^{\ell+1} \in \mathcal{G}_{pq}. \tag{166}$$

It remains to verify that

$$1_{B_{pq}^{\ell+1} \cap B_{mn}^\ell \cap \{\sigma_{\ell+1} \le t\}} O^{\ell+1} \in \mathcal{G}_{pq}. \tag{167}$$

For this we claim that for each $j \in \mathcal{J}$,

$$O_j^{\ell+1} = \Theta_j(O_j^\ell, \sigma_\ell, \sigma_{\ell+1}, Z^\ell, D(\sigma_{\ell+1}) - D(\sigma_\ell), A(\sigma_{\ell+1}) - A(\sigma_\ell)) \tag{168}$$

where Θ_j is a measurable function from $(\mathcal{K}_0 \times \mathbb{R}_+)^{\mathbb{N}} \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{N}^K \times \mathbb{N}^K \times \mathbb{N}^K$ into $(\mathcal{K}_0 \times \mathbb{R}_+)^{\mathbb{N}}$, where $\mathcal{K}_0 = \{0, 1, \ldots, K\}$. Indeed, the action of Θ_j may be described as follows. Although we have not invented notation to describe the deletion, update and insertion operations below, it is evident from this description that Θ_j can be defined as a measurable function of the form indicated above. For each $k \in \mathcal{C}(j)$, if $D_k(\sigma_{\ell+1}) - D_k(\sigma_\ell) > 0$ (in which case the left member must equal one since simultaneous departures from a class cannot occur), delete the first pair in the sequence O_j^ℓ that has the value k as its first coordinate. Denote the new sequence obtained in this way by $\tilde{O}_j^\ell$. Then, $\tilde{O}_j^\ell$ indicates the order and age at the time σ_ℓ of those customers who are left at station j after the departures have occurred at $\sigma_{\ell+1}$, but before the new arrivals at $\sigma_{\ell+1}$ have been taken into account. There are

$$\zeta_j^{\ell+1} \equiv \sum_{k\in\mathcal{C}(j)} \left(Z_k^\ell - (D_k(\sigma_{\ell+1}) - D_k(\sigma_\ell)) \right) \tag{169}$$

of these "remaining" customers at station j. We update the ages of these remaining customers by adding $\sigma_{\ell+1} - \sigma_\ell$ to the second coordinate of the first $\zeta_j^{\ell+1}$ pairs in $\tilde{O}_j^\ell$. To include information about the new arrivals and so obtain $O_j^{\ell+1}$, for each $k \in \mathcal{C}(j)$, insert $A_k(\sigma_{\ell+1}) - A_k(\sigma_\ell)$ copies of the class designator, age pair $(k, 0)$ into $\tilde{O}_j^\ell$ just after the first $\zeta_j^{\ell+1}$ pairs in this sequence. The order in which these new pairs are inserted is given by the tie breaking rule for treating simultaneous arrivals. Note that simultaneous arrivals to the same class are indistinguishable in terms of the information to be inserted in $\tilde{O}_j^\ell$ and so only the deterministic rule for breaking ties between classes is needed here. It now follows from (148), (151), (154) and (168) that

$$1_{B_{mn}^\ell \cap \{\sigma_{\ell+1} \le t\}} O^{\ell+1} \in \mathcal{G}_{mn} \tag{170}$$

and hence on combining this with (157) we obtain (167). Thus (162) has been verified and this completes the induction step. ▫

Proof of Lemma 8.4. In this proof only, $\mathcal{M}^r$ will denote a martingale rather than the σ-algebra on $\mathbf{D}^r$.

An outline of our proof is as follows. The idea of the proof of part (i) is that apart from small error terms associated with state space collapse and residual interarrival times, by suitably centering and scaling the primitive processes $(E^r, V^{s,r}, \Phi^{k,r})$, we can reexpress $\hat{X}^r$, as given by (60), (68) and (75), in terms of a multiparameter martingale evaluated at a stopping time. Indeed, we use the i.i.d. and independence assumptions on the primitive sequences $\{u_l^r(i), i = 2, 3, \ldots\}$, $\{v_k^{s,r}(i), i = 1, 2, \ldots\}$, $\{\phi^{r,k}(i), i = 1, 2, \ldots\}$, $l \in \mathcal{A}$, $k \in \mathcal{K}$, as described in Section 3.1.6, to establish the multiparameter martingale property (182) below. This implies that the process $\{\mathcal{T}^r(p,q), \mathcal{G}_{pq}^r, (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}$ as defined in (183) is a multiparameter martingale. In order to conclude that the stopped process in (199) is also a martingale, we establish L^2-bounds on the multiparameter martingale $\mathcal{T}^r$ and bounds on the mean of the stopping time $\tau^r(t) = (E_{\mathcal{A}}^r(t), A^r(t))$. The martingale property in part (i) of the lemma follows from this stopped martingale property and the fact that Y^r, W^r are adapted to $\mathcal{G}_{\tau^r(t)}^r$. State space collapse and the asymptotic negligibility of error terms associated with the martingale property of the renewal process $E_{\mathcal{A}}^r$ are used to show part (ii). Finally, the uniform integrability property in part (iii) follows from L^2 bounds used in obtaining the stopped martingale property mentioned above. Now we provide the details of the proof.

For the moment, let r be fixed. Now, $\{\mathcal{G}^r_{pq}\} \equiv \{\mathcal{G}^r_{pq} : (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}$ defined by (92) is a (multiparameter) filtration and for each $t \geq 0$, by Lemma 8.3,

$$\tau^r(t) \equiv (E^r_{\mathcal{A}}(t), A^r(t)) \tag{171}$$

is a (multiparameter) stopping time relative to this filtration. If $(\Omega^r, \mathcal{F}^r)$ is the measurable space on which all of the processes indexed by r are defined, then for each $t \geq 0$ we can define a σ-algebra associated with $\tau^r(t)$ as follows:

$$\mathcal{G}^r_{\tau^r(t)} = \{B \in \mathcal{F}^r : B \cap \{\tau^r(t) \leq (p,q)\} \in \mathcal{G}^r_{pq} \text{ for all } (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}. \tag{172}$$

Then $\{\mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$ is a filtration in the usual one-parameter sense.

We first verify that Y^r, W^r are adapted to the filtration $\{\mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$. Since $\mathcal{X}^r(0)$ contains $(V^{o,r}(\cdot), Z^r(0))$, it follows that $W^r(0) \equiv CV^{o,r}(Z^r(0))$ is $\mathcal{G}^r_{00}$-measurable. Furthermore, the multiparameter process $\{(U^r_{\mathcal{A}}(p+e_{\mathcal{A}}), V^{s,r}(q)) : (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}$ is adapted to the filtration $\{\mathcal{G}^r_{pq}\}$ and hence by the discrete nature of $\mathbb{N}^L \times \mathbb{N}^K$ it is progressively measurable. Then by Proposition 2.8.5 of [27] and the stopping time property of $\tau^r(t)$ we have that for each $t \geq 0$,

$$(U^r_{\mathcal{A}}(E^r_{\mathcal{A}}(t) + e_{\mathcal{A}}), V^{s,r}(A^r(t))) \in \mathcal{G}^r_{\tau^r(t)}. \tag{173}$$

It then follows (using approximation of the supremum at a countable set of times) that (cf. (20)),

$$Y^r(t) \equiv \sup_{0 \leq s \leq t} (W^r(0) + CV^{s,r}(A^r(s)) - es)^- \in \mathcal{G}^r_{\tau^r(t)} \tag{174}$$

and

$$W^r(t) \equiv W^r(0) + CV^{s,r}(A^r(t)) - et + Y^r(t) \in \mathcal{G}^r_{\tau^r(t)}. \tag{175}$$

We now introduce the fundamental multiparameter martingales $\mathcal{M}^r, \mathcal{O}^r, \mathcal{Q}^r$, and martingales associated with squares of their components. For each $p \in \mathbb{N}^L, q \in \mathbb{N}^K$, let

$$\begin{aligned}
\mathcal{M}^r_l(p_l) &= (\alpha^r_l (U^r_l(p_l+1) - u^r_l(1)) - p_l) && (176)\\
\mathcal{N}^r_l(p_l) &= \left((\mathcal{M}^r_l(p_l))^2 - p_l(\alpha^r_l)^2 a^r_l\right) && (177)\\
\mathcal{O}^r_k(q_k) &= V^{s,r}_k(q_k) - q_k m^r_k && (178)\\
\mathcal{P}^r_k(q_k) &= (\mathcal{O}^r_k(q_k))^2 - q_k b^r_k && (179)\\
\mathcal{Q}^{k,r}(q_k) &= \Phi^{k,r}(Z^r_k(0) + q_k) - \Phi^{k,r}(Z^r_k(0)) - \tilde{P}^{k,r} q_k && (180)\\
\mathcal{R}^{k,r}_{ij}(q_k) &= \mathcal{Q}^{k,r}_i(q_k)\mathcal{Q}^{k,r}_j(q_k) - q_k \Upsilon^{k,r}_{ij} && (181)
\end{aligned}$$

for all $l \in \mathcal{A}$, $i,j,k \in \mathcal{K}$. Let $\mathcal{M}^r(p) = (\mathcal{M}^r_l(p_l) : l \in \mathcal{A})$, $\mathcal{N}^r(p) = (\mathcal{N}^r_l(p_l) : l \in \mathcal{A})$, $\mathcal{O}^r(q) = (\mathcal{O}^r_k(q_k) : k \in \mathcal{K})$, $\mathcal{P}^r(q) = (\mathcal{P}^r_k(q_k) : k \in \mathcal{K})$, $\mathcal{Q}^r(q) = (\mathcal{Q}^{1,r}(q_1), \ldots, \mathcal{Q}^{K,r}(q_K))$, $\mathcal{R}^r(q) = (\mathcal{R}^{1,r}(q_1), \ldots, \mathcal{R}^{K,r}(q_K))$. Because of the independence, i.i.d. and integrability assumptions of Section 3, we have that the $(2L + 2K + K^2 + K^3)$-dimensional process:

$$\{\mathcal{S}^r(p,q) \equiv (\mathcal{M}^r(p), \mathcal{N}^r(p), \mathcal{O}^r(q), \mathcal{P}^r(q), \mathcal{Q}^r(q), \mathcal{R}^r(q)) : (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\}, \tag{182}$$

is a multiparameter martingale relative to $\{\mathcal{G}^r_{pq}\}$.

For each $(p,q) \in \mathbb{N}^L \times \mathbb{N}^K$, let

$$\mathcal{T}^r(p,q) = (\mathcal{M}^r(p), \mathcal{O}^r(q), \mathcal{Q}^r(q)). \tag{183}$$

We aim to show that $\{\mathcal{T}^r(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$ is a martingale. However, we cannot immediately deduce this from the martingale property of $\mathcal{S}^r$, since $\tau^r(t)$ is a possibly unbounded stopping time. So we first truncate time, apply the multiparameter stopping theorem and then pass to the limit in the truncation using uniform integrability to deduce the desired result. The bounds obtained for the uniform integrability will also prove useful in verifying part (iii) of the lemma. For $n \in \mathbb{N}$, let n^L (respectively, n^K) denote the L-dimensional (respectively K-dimensional) vector whose components all have value n. Then for the minima $p \wedge n^L$, $q \wedge n^K$, defined componentwise, we can verify (in a similar manner to that for $\mathcal{S}^r$) that

$$\{\mathcal{S}^{r,n}(p,q) \equiv \mathcal{S}^r(p \wedge n^L, q \wedge n^K) : (p,q) \in \mathbb{N}^L \times \mathbb{N}^K\} \tag{184}$$

is a multiparameter martingale relative to $\{\mathcal{G}^r_{pq}\}$. Then by the multiparameter optional stopping theorem (see [27], Theorem 2.8.7) we have that

$$\{\mathcal{S}^{r,n}(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\} \tag{185}$$

is a martingale for each $n \in \mathbb{N}$. Now, for $(p,q) \in \mathbb{N}^L \times \mathbb{N}^K$ and $n \in \mathbb{N}$, let

$$\mathcal{T}^{r,n}(p,q) = (\mathcal{M}^r(p \wedge n^L), \mathcal{O}^r(q \wedge n^K), \mathcal{Q}^r(q \wedge n^K)). \tag{186}$$

For each $n \in \mathbb{N}$, it follows from the martingale property for $\{\mathcal{S}^{r,n}(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\}$ that

$$\{\mathcal{T}^{r,n}(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\} \tag{187}$$

is a martingale. We aim to prove that the same is true with $\mathcal{T}^r$ in place of $\mathcal{T}^{r,n}$. For $t \geq 0$ fixed, $\mathcal{T}^{r,n}(\tau^r(t)) \to \mathcal{T}^r(\tau^r(t))$ pointwise as $n \to \infty$, and so it suffices to show that $\{\mathcal{T}^{r,n}(\tau^r(t))\}_{n=1}^\infty$ is L^2-bounded for each $t \geq 0$, since this implies that it is uniformly integrable. By the martingale properties of the $\mathcal{N}^r$, $\mathcal{P}^r$ and $\mathcal{R}^r$ elements of $\mathcal{S}^{r,n}(\tau^r(\cdot))$ we have for all $l \in \mathcal{A}$, $i, k \in \mathcal{K}$, $n \geq 1$:

$$\mathbb{E}[(\mathcal{M}^r_l(E^r_l(t) \wedge n))^2 - (E^r_l(t) \wedge n)(\alpha^r_l)^2 a^r_l] = 0, \tag{188}$$

$$\mathbb{E}[(\mathcal{O}^r_k(A^r_k(t) \wedge n))^2 - (A^r_k(t) \wedge n) b^r_k] = 0 \tag{189}$$

$$\mathbb{E}[(\mathcal{Q}^{k,r}_i(A^r_k(t) \wedge n))^2 - (A^r_k(t) \wedge n)\Upsilon^{k,r}_{ii}] = 0. \tag{190}$$

From Lorden's inequality for renewal processes (cf. Lindvall [39], pp. 77–78; Carlsson-Nerman [12]) and by setting the first external interarrival time to zero, we obtain the following upper bound for all $l \in \mathcal{A}$,

$$\mathbb{E}[E^r_l(t)] \leq \alpha^r_l t + (\alpha^r_l)^2 a^r_l + 2. \tag{191}$$

Indeed, since $E^r_l \equiv 0$, $\alpha^r_l = a^r_l = 0$ for $l \in \mathcal{K} \setminus \mathcal{A}$, the above holds for all $l \in \mathcal{K}$. Furthermore, since we have assumed FIFO service within each class, as in the proof of Lemma 8.2 (cf. (108)) we have

$$A^r(t) \leq E^r(t) + \sum_{l=1}^{K} \Phi^{l,r}(Z^r_l(0) + S^r_l(t)) \equiv B^r(t) \tag{192}$$

where $S^r_l(t)$ is given by (109) with l in place of k there. Now, for each $k, l \in \mathcal{K}$,

$$\Phi^{l,r}_k(Z^r_l(0) + S^r_l(t)) \leq Z^r_l(0) + S^r_l(t) \tag{193}$$

and by Lorden's inequality again,

$$\mathbb{E}[S^r_l(t)] \leq \mu^r_l t + (\mu^r_l)^2 b^r_l + 2. \tag{194}$$

Combining (191)–(194) we have for each $k \in \mathcal{K}$,

$$\mathbb{E}[A^r_k(t)] \leq (\alpha^r_k t + (\alpha^r_k)^2 a^r_k + 2) + \sum_{l=1}^{K} (\mathbb{E}[Z^r_l(0)] + \mu^r_l t + (\mu^r_l)^2 b^r_l + 2) \equiv c^r_k(t), \tag{195}$$

where $c^r_k(t)$ is finite by (95). It then follows from (188)–(190), (191) and (195) that for all $n \geq 1$, $l \in \mathcal{A}$, $i, k \in \mathcal{K}$,

$$\mathbb{E}[(\mathcal{M}^r_l(E^r_l(t) \wedge n))^2] \leq (\alpha^r_l)^2 a^r_l c^r_l(t) \tag{196}$$

$$\mathbb{E}[(\mathcal{O}^r_k(A^r_k(t) \wedge n))^2] \leq b^r_k c^r_k(t) \tag{197}$$

$$\mathbb{E}[(\mathcal{Q}^{k,r}_i(A^r_k(t) \wedge n))^2] \leq \Upsilon^{k,r}_{ii} c^r_k(t). \tag{198}$$

This establishes the desired L^2-boundedness and hence

$$\{\mathcal{T}^r(\tau^r(t)), \mathcal{G}^r_{\tau^r(t)}, t \geq 0\} \tag{199}$$

is a martingale for each r.

We now apply the above martingale properties to establish part (i) of the lemma. First extend the definition of $\mathcal{M}^r$ to coordinates $k \notin \mathcal{A}$ by defining $\mathcal{M}^r_k \equiv 0$ for all $k \in \mathcal{K} \setminus \mathcal{A}$. Then, rewriting $\hat{\xi}^r$, as defined by (60), in terms of some elements of $\mathcal{S}^r$, we have

$$\hat{\xi}^r(t) = \check{\xi}^r(t) + CM^rQ^r(r^{-1}\zeta^{r,1}(t) + \zeta^{r,2}(t)), \tag{200}$$

where

$$\begin{aligned}
\check{\xi}^r(t) &= r^{-1}C\left(\mathcal{O}^r(A^r(r^2t)) + M^rQ^r\left(-\mathcal{M}^r(E^r(r^2t)) + \sum_{k=1}^{K}\mathcal{Q}^{k,r}(A^r_k(r^2t))\right)\right) + \hat{\gamma}^r t,\\
\zeta^{r,1}_k(t) &= \begin{cases} 0 & \text{if } k \notin \mathcal{A},\\ \alpha^r_k(U^r_k(E^r_k(r^2t)+1) - u^r_k(1) - r^2t) & \text{if } k \in \mathcal{A},\end{cases}\\
\zeta^{r,2}(t) &= \sum_{k=1}^{K}\left(\hat{\Phi}^{k,r}(\bar{D}^r_k(t)) - \hat{\Phi}^{k,r}(\bar{Z}^r_k(0) + \bar{A}^r_k(t)) + \hat{\Phi}^{k,r}(\bar{Z}^r_k(0))\right).
\end{aligned}$$

Since $\mathcal{T}^r(\tau^r(r^2t)) = (\mathcal{M}^r_{\mathcal{A}}(E^r_{\mathcal{A}}(r^2t)), \mathcal{O}^r(A^r(r^2t)), \mathcal{Q}^r(A^r(r^2t)))$, it follows from the martingale property of $\mathcal{T}^r(\tau^r(\cdot))$ that

$$\{\check{\xi}^r(t) - \hat{\gamma}^r t, \mathcal{G}^r_{\tau^r(r^2t)}, t \geq 0\} \tag{201}$$

is a martingale. Now, from (75) and (200) we have

$$\hat{X}^r(t) = \check{X}^r(t) + \check{\epsilon}^r(t) \tag{202}$$

where

$$\begin{aligned} \check{X}^r(t) &= \hat{W}^r(0) + R^r\check{\xi}^r(t) \\ \check{\epsilon}^r(t) &= R^r\left(\hat{\eta}^r(t) + CM^rQ^r(r^{-1}\zeta^{r,1}(t) + \zeta^{r,2}(t))\right). \end{aligned}$$

By the martingale property for $\check{\xi}^r$, for $\theta^r = R^r\hat{\gamma}^r$,

$$\{\check{X}^r(t) - \check{X}^r(0) - \theta^r t, \mathcal{G}^r_{\tau^r(r^2t)}, t \geq 0\} \tag{203}$$

is a martingale. Note that $\hat{Y}^r$, $\hat{W}^r$ are adapted to $\{\mathcal{G}^r_{\tau^r(r^2t)}, t \geq 0\}$, by (174)–(175) and (39), (41). Hence, $\{\check{X}^r(t) - \check{X}^r(0) - \theta^r t, t \geq 0\}$ is a martingale relative to the filtration generated by $(\hat{W}^r, \check{X}^r, \hat{Y}^r)$.

In order to show part (ii) of the lemma, note that by (68), (78), and the fact that state space collapse holds by Assumption 7.2 and Proposition 8.1, we have for each $T \geq 0$, $\|\hat{\eta}^r(\cdot)\|_T \to 0$ in probability as $r \to \infty$. Furthermore, by the definition of E^r_k from U^r_k for $k \in \mathcal{A}$, for each $T \geq 0$,

$$\|r^{-1}\zeta^{r,1}(\cdot)\|_T \leq 2\max_{k\in\mathcal{A}} |\alpha^r_k|\, \|r^{-1}u^r_k(E^r_k(r^2\,\cdot\,) + 1)\|_T, \tag{204}$$

where, as a consequence of our functional central limit theorem assumptions in Section 5, the right member goes to zero in probability as $r \to \infty$ (see the proof of Lemma 6 in Iglehart and Whitt [36]). By (87), Assumption 7.1 and state space collapse, we have that $\bar{Z}^r(0) = \Delta^r\bar{W}^r(0) + r^{-1}\hat{\varepsilon}^r(0) \to 0$ in probability as $r \to \infty$. Now,

$$\|\zeta^{r,2}(\cdot)\|_T \leq \sum_{k=1}^{K} \|\hat{\Phi}^{k,r}(\bar{D}^r_k(\cdot)) - \hat{\Phi}^{k,r}(\bar{Z}^r_k(0) + \bar{A}^r_k(\cdot)) + \hat{\Phi}^{k,r}(\bar{Z}^r_k(0))\|_T, \tag{205}$$

where, by a similar argument to that used in deducing (98) from (97), using (86) and Lemma 8.2 we have that the right member above converges to zero in distribution (or equivalently, in probability) as $r \to \infty$. Combining the above with the convergence of CM^rQ^r and R^r, we see that for each $T \geq 0$, $\|\check{\epsilon}^r(\cdot)\|_T \to 0$ in probability as $r \to \infty$.

It remains to show part (iii) of the lemma. Since $\check{X}^r(0) = \hat{W}^r(0)$ and $R^r \to R$, for this it suffices to show that $\{\check{\xi}^r(t)\}$ as r varies is uniformly integrable for each fixed $t \geq 0$. Now by Fatou's lemma, (196)–(198) hold with the n's removed. Fix $t \geq 0$. By (78), (95) and (195), we have

$$\sup_r \max_{k\in\mathcal{K}} \left(\frac{c^r_k(r^2t)}{r^2}\right) < \infty. \tag{206}$$

Replacing t by $r^2 t$ in (196)–(198), and combining the above with the assumed convergence of $\alpha_k^r, \mu_k^r = \frac{1}{m_k^r}$, a_k^r, b_k^r, $\Upsilon^{k,r}$ (cf. (78), (81) and (10)), we see that

$$\{r^{-1}(\mathcal{M}^r(E^r(r^2t)), \mathcal{O}^r(A^r(r^2t)), \mathcal{Q}^r(A^r(r^2t)))\}$$

as a collection indexed by r is L^2-bounded, and hence uniformly integrable. Combining this with the convergence of M^r, Q^r, and $\hat{\gamma}^r$, we see that $\{\check{\xi}^r(t)\}$ is uniformly integrable, as desired. ▫

9 Verification of Assumption 7.1 for FIFO networks of Kelly type and HLPPS networks

For the statements of the definitions and lemmas in this section, consider a sequence of open multiclass queueing networks as defined in Sections 3 and 4 that satisfies the heavy traffic assumptions of Section 5. In this section, we refer to this as "our sequence of queueing networks". For this section only, A and D will have different meanings than elsewhere (they will denote matrices here).

The following matrix theoretic result, which is taken from Dai and Harrison [20], plays a key role in the verification of Assumption 7.1 for FIFO networks of Kelly type and HLPPS networks.

Definition 9.1 *A $J \times J$ matrix A is said to be admissible if there is a $J \times J$ diagonal matrix D with strictly positive diagonal entries such that*

(207) $$AD + DA'$$

is (strictly) positive definite.

Lemma 9.1 *If a $J \times J$ matrix A is admissible, then A is invertible, A^{-1} is admissible, and A and A^{-1} are completely-$\mathcal{S}$.*

Proof. See Dai and Harrison [20]. ▫

Definition 9.2 *Our sequence of queueing networks is said to be asymptotically of Kelly type if m as defined in (78) is such that for each $j \in \mathcal{J}$, m_k has the same value for all $k \in \mathcal{C}(j)$.*

Lemma 9.2 *Suppose that our sequence of queueing networks has FIFO service, is asymptotically of Kelly type, and $\lambda > 0$, i.e., $\lambda_k > 0$ for each $k \in \mathcal{K}$. Then Assumption 7.1 holds.*

Remark. The proof of this lemma is given in Section 3 of Dai and Harrison [20]. (Although it is not emphasized there, it is implicit in the proof that $\lambda > 0$, i.e., $\Lambda = \text{diag}(\lambda)$ has strictly positive diagonal entries.) The proof is repeated below for the convenience of the reader and because our verification of Assumption 7.1 for HLPPS networks uses elements from this proof.

Proof. From the FIFO assumption, by (63), for all r sufficiently large that $\lambda^r > 0$, we have

$$\delta_k^r = \frac{\lambda_k^r}{\rho_j^r} \quad \text{for all } k \in \mathcal{C}(j),\ j \in \mathcal{J}, \tag{208}$$

and so as $r \to \infty$, since $\rho_j^r \to 1$ for each $j \in \mathcal{J}$ and $\lambda^r \to \lambda$, we have

$$\Delta^r \to \Delta \equiv \Lambda C'. \tag{209}$$

Furthermore, from (73) and the convergence of M^r to M, Q^r to Q, and Δ^r to Δ, we have

$$I + G = CMQ\Delta. \tag{210}$$

From the assumption that the sequence of networks is asymptotically of Kelly type we have that

$$CM = \bar{M}C, \tag{211}$$

where $\bar{M}$ is a $J \times J$ diagonal matrix whose j^{th} diagonal entry is the common value of m_k for all $k \in \mathcal{C}(j)$. Then by (209)–(211), we have

$$I + G = CMQ\Lambda C' = \bar{M}CQ\Lambda C'. \tag{212}$$

We now show that $I + G$ is admissible. An application of Lemma 9.1 then completes the proof that Assumption 7.1 holds.

Using $\Lambda = \text{diag}(\alpha) + \sum_{k=1}^K \lambda_k \text{diag}(\tilde{P}^k)$, where $\tilde{P}^k$ denotes the k^{th} column of $\tilde{P} = P'$, we can verify (cf. [20], (3.19)–(3.20)) that

$$\Lambda(I - P) + (I - P')\Lambda = \text{diag}(\alpha) + \Upsilon + (I - P')\Lambda(I - P), \tag{213}$$

where $\Upsilon = \sum_{k=1}^K \lambda_k \Upsilon^k$ and Υ^k is given by (10). Note that the right member here is positive definite because the first two terms in that member are non-negative definite and the last term is positive definite by the assumption that Λ has strictly positive diagonal entries and the fact that $I - P$ is invertible. On multiplying (213) on the left by $Q = (I - P')^{-1}$ and on the right by Q' we then have that $Q\Lambda + \Lambda Q'$ is positive definite. Then, since the constituency matrix C has full rank and so the null space of C' is trivial, it follows that $CQ\Lambda C' + C\Lambda Q'C'$ is positive definite and hence so too is

$$(I + G)\bar{M} + \bar{M}(I + G') = \bar{M}CQ\Lambda C'\bar{M} + \bar{M}C\Lambda Q'C'\bar{M}. \tag{214}$$

Since $\bar{M}$ is a diagonal matrix with strictly positive diagonal entries, this completes the proof that $I + G$ is admissible. □

Lemma 9.3 *Suppose that our sequence of queueing networks has the HLPPS service discipline and $\lambda > 0$. Then Assumption 7.1 holds.*

Proof. This proof is similar to that for Lemma 9.2. The main difference is that the matrix Δ has a slightly different form. Indeed, for the HLPPS service discipline (see (65)), for all r sufficiently large we have

$$\delta_k^r = \frac{\lambda_k^r m_k^r}{\sum_{l \in \mathcal{C}(j)} \lambda_l^r (m_l^r)^2} \quad \text{for all } k \in \mathcal{C}(j),\ j \in \mathcal{J}, \tag{215}$$

and so as $r \to \infty$, since $\lambda^r \to \lambda > 0$ and $m^r \to m > 0$, we have

$$\Delta^r \to \Delta \equiv M \Lambda C' (C M^2 \Lambda C')^{-1}. \tag{216}$$

Here $CM^2 \Lambda C'$ is a $J \times J$ diagonal matrix whose j^{th} diagonal entry is $\sum_{k \in \mathcal{C}(j)} \lambda_k m_k^2$ which is strictly positive since $\lambda_k > 0$ and $m_k > 0$ for all $k \in \mathcal{K}$. Then since (210) holds, we have

$$I + G = CMQ\Delta = CMQM\Lambda C'(CM^2 \Lambda C')^{-1}. \tag{217}$$

Hence, using the fact that $M\Lambda = \Lambda M$, we have

$$I + G = CMQ\Lambda MC'(CM^2 \Lambda C')^{-1}. \tag{218}$$

The proof given in Lemma 9.2 that $Q\Lambda + \Lambda Q'$ is positive definite does not depend on the specific service discipline and so it also applies here. Since M is a diagonal matrix with strictly positive diagonal entries and C has full rank, it follows that for $D = CM^2 \Lambda C'$,

$$(I + G)D + D(I + G') \equiv CMQ\Lambda MC' + CM\Lambda Q'MC' \tag{219}$$

is positive definite. Since D is a diagonal matrix with strictly positive diagonal entries, $I + G$ is admissible and hence by Lemma 9.1 it is invertible and its inverse is completely-$\mathcal{S}$. This completes the verification of Assumption 7.1. ▫

10 Heavy traffic limit theorems for FIFO networks of Kelly type and HLPPS networks

For the statements of the two theorems in this section, we consider a sequence of open multiclass queueing networks as defined in Sections 3 and 4. In order that our model be consistent with that of Bramson, we also assume the additional conditions described in the second paragraph of Section 3.3, i.e., for each r and each $k \in \mathcal{K}$, $\{v_k^{o,r}(i), i = 2, 3, \ldots\}$ is a sequence of i.i.d. random variables that has the same distribution as our $\{v_k^{s,r}(i), i = 1, 2, \ldots\}$, and for each r, the sequences $\{v_k^{o,r}(i), i = 2, 3, \ldots\}$, $k \in \mathcal{K}$, are mutually independent, and, as a collection, are independent of the initial queue length $Z^r(0)$, the initial residual times

$\{v_k^{o,r}(1), k \in \mathcal{K}\}$ and $\{u_l^r, l \in \mathcal{A}\}$, and the initial sequence of (class, age) pairs $O^{0,r}$. Recall from Section 4 the definitions of the diffusion scaled processes associated with workload $\hat{W}^r$, cumulative idletime $\hat{Y}^r$ and queue length $\hat{Z}^r$, and the non-negative entries δ_k^r in the matrix Δ^r. Associated with the heavy traffic assumptions of Section 5 we have the asymptotic parameters γ and H, defined in (80) and (88), respectively, and the limit $W(0)$ in distribution of $\hat{W}^r(0)$ as $r \to \infty$. The law of $W(0)$ is denoted by ν. The definition of an SRBM is given in Section 6.

In order to obtain heavy traffic approximations using Theorem 7.1 (and in particular, to apply the results of Bramson [11] on multiplicative state space collapse), we need to assume that state space collapse holds initially. In Bramson's [11] terminology, the initial data must be asymptotically close to a certain invariant manifold. In fact, for FIFO networks of Kelly type, this involves assumptions on the initial distribution of the workload amongst the different classes and on the ordering of the customers initially at each station, or equivalently on the departure process for a small interval of time. This requirement is stronger than simply requiring that (62) hold with $T = 0$ and is captured in condition (220) of Theorem 10.1. On the other hand, for HLPPS networks, since service is distributed across the classes in proportion to the length of the queue for each class, in this case it suffices to have (62) for $T = 0$ (cf. (224)).

Theorem 10.1 (Heavy traffic limit theorem for FIFO networks of Kelly type) *Consider a sequence of open multiclass queueing networks with the FIFO service discipline that satisfies the heavy traffic conditions of Section 5. Suppose that this sequence is asymptotically of Kelly type (cf. Definition (9.2)) and that $\lambda > 0$, i.e., $\lambda_k > 0$ for all $k \in \mathcal{K}$. Further suppose that for each $j \in \mathcal{J}$ and $k \in \mathcal{C}(j)$,*

$$(220) \qquad \sup_{t \in [0, W_j^r(0)]} r^{-1} |D_k^r(t) - \delta_k^r t| \to 0 \quad \textit{in probability as } r \to \infty,$$

where δ_k^r is given by (63). Then

$$(221) \qquad (\hat{W}^r, \hat{X}^r, \hat{Y}^r, \hat{Z}^r) \Rightarrow (W, X, Y, Z) \quad \textit{as } r \to \infty$$

where $Z = \Delta W$ for $\Delta = \Lambda C'$, and (W, X, Y) satisfy the conditions of Definition 6.1, in particular, $W = X + RY$ a.s. and W is an SRBM associated with the data $(S, \theta, \Gamma, R, \nu)$ for $\theta = R\gamma$, $\Gamma = RHR'$.

Remark. Condition (220) is used to describe the required initial behavior of our networks with a FIFO service discipline. The assumption of a FIFO service discipline implies that

$$(222) \qquad D_k^r(W_j^r(0)) = Z_k^r(0),$$

for all $k \in \mathcal{K}$. This can be interpreted as follows: $W_j^r(0)$ is the first time that all of the work initially held at station j has been completed, and at this time, because the service is on a FIFO basis, the number of departures that have occurred from any class $k \in \mathcal{C}(j)$ up to that time is equal to the number of customers $Z_k^r(0)$ initially in that class. One can combine (222) with (220), and use the definitions of fluid and diffusion scaling, and of Δ^r (see (61)), to conclude that for each $j \in \mathcal{J}$ and $k \in \mathcal{C}(j)$,

$$(223)\quad |\hat{Z}_k^r(0) - \Delta_{kj}^r \hat{W}_j^r(0)| = r^{-1}|D_k^r(W_j^r(0)) - \delta_k^r W_j^r(0)| \to 0 \quad \text{in probability as } r \to \infty.$$

However, (220) requires more than this, it requires that the ordering of the customers initially at station j and the distribution of the initial workload amongst these customers be such that up to an error that is $o(r)$, in the initial interval $[0, W_j^r(0)]$, server j will expend effort in the proportions given by δ_k^r to the classes $k \in \mathcal{C}(j)$. We note in particular that if $\hat{W}^r(0)$ and $\hat{Z}^r(0)$ converge to zero in probability as $r \to \infty$, then (220) is automatically satisfied.

Proof of Theorem 10.1. By Theorem 7.1, we simply need to verify that Assumptions 7.1 and 7.2 hold. By Lemma 9.2, Assumption 7.1 on the data matrices holds. For the multiplicative state space collapse Assumption 7.2, note from the assumptions that $\Delta^r \to \Delta$, $\hat{W}^r(0)$ converges in distribution, and the initial condition (220), that the latter will still hold if for each k, δ_k^r there is replaced by $\delta_k = \lim_{r\to\infty} \delta_k^r$, which equals λ_k for a FIFO service discipline. With this replacement, condition (220) is the same as condition (3.6) (or equivalently, (3.9)) of Bramson [11]. (We remind the reader, as noted in Section 3.3, that Bramson [11] considers an uncountable family of networks indexed by all $r \in (0, \infty)$. However, we can embed our sequence in such a family, by a (right continuous) piecewise constant extension, and it is easy to check that Bramson's conditions (3.1)–(3.6) hold for the resulting family.) It then follows from Theorem 1 of Bramson [11] and the remark following Definition 7.1 that the Assumption 7.2 of multiplicative state space collapse is satisfied. Our desired result then follows from Theorem 7.1. ▫

Theorem 10.2 (Heavy traffic limit theorem for HLPPS networks) *Consider a sequence of open multiclass queueing networks with the HLPPS service discipline that satisfies the heavy traffic conditions of Section 5. Suppose that $\lambda > 0$ and*

$$(224)\qquad |\hat{Z}^r(0) - \Delta^r \hat{W}^r(0)| \to 0 \quad \textit{in probability as } r \to \infty,$$

where Δ^r is given by (61), (65). Then

$$(225)\qquad (\hat{W}^r, \hat{X}^r, \hat{Y}^r, \hat{Z}^r) \Rightarrow (W, X, Y, Z) \quad \textit{as } r \to \infty,$$

where $Z = \Delta W$ for $\Delta = M\Lambda C'(CM^2\Lambda C')^{-1}$, and (W, X, Y) satisfy the conditions of Definition 6.1, in particular, $W = X + RY$ a.s. and W is an SRBM associated with the data $(S, \theta, \Gamma, R, \nu)$ for $\theta = R\gamma$, $\Gamma = RHR'$.

Proof. The proof is the same as for Theorem 10.1, except that Lemma 9.3 is used in place of Lemma 9.2 and Theorem 1′ of Bramson [11] is used in place of his Theorem 1. ▫

Remark. Versions of Theorems 10.1 and 10.2 above also appear as Theorems 2 and 2′, respectively, in Bramson [11].

11 Directions for further research

A compelling problem left open by this paper is that of identifying other families of networks (besides those discussed in Section 10) which satisfy the conditions of our heavy traffic limit theorem (Theorem 7.1). More precisely, which open multiclass networks with HL service disciplines are such that (a) the reflection matrix R is well defined and completely-$\mathcal{S}$ (see Assumption 7.1), and (b) (multiplicative) state space collapse holds (see Assumption 7.2)? Natural networks to examine in this light are those with (preemptive resume) static priority service disciplines. For re-entrant lines with a first-buffer-first-served or last-buffer-first-served static priority discipline, Dai, Yeh and Zhou [25] (see their appendix) have shown that condition (a) holds. In [11], Bramson gives sufficient conditions for (b) to hold for networks having a fixed static priority discipline. His conditions are phrased in terms of required behavior of associated balanced fluid models. It is a challenging open problem to identify static priority networks for which the sufficient conditions given by Bramson are satisfied and for which condition (a) holds. A related problem is to further investigate when state space collapse is necessary for a heavy traffic limit theorem of the type contained in Section 7 (cf. Appendix B).

The assumption of FIFO service within a class is used in the proofs of Proposition 8.1 and Lemmas 8.2 and 8.4. In particular, this assumption is used to obtain (108) and (192), where estimates on the behavior of S_k^r can be obtained directly from the behavior of the primitive process $V_k^{s,r}$. While it seems likely that the results in Proposition 8.1, and Lemmas 8.2 and 8.4, could be extended to other service disciplines (e.g., LIFO and PS), suitable replacements would need to be found for the aforementioned estimates. In a similar vein, it is desirable to extend the stopping time property of Lemma 8.3 to other service disciplines besides those of the HL type considered here. It seems likely that this property holds for a wide variety of non-idling, non-anticipating service disciplines.

Other generalizations that might be attempted are to relax the independence assumptions of Section 3.1.6 between and within the external interarrival time, service time and routing sequences, and to allow external interarrival and service times that may take the value zero with positive probability. Relaxation of the independence assumption would substantially affect verification of the martingale property in Lemma 8.4. Allowing zero external interarrival and service times would at least further complicate the verification of the stopping time property (cf. Lemma 8.3).

Finally, it is natural to try to extend the results of this paper to closed and capacitated networks. For such networks, the associated SRBMs live in convex polyhedrons. Sufficient conditions for existence and uniqueness of such SRBMs have been given in Dai and Williams [24] and an invariance principle for those SRBMs has been established by Dai and Dai [19].

Acknowledgements. J. Michael Harrison first suggested to the author that state space collapse and an invariance principle for SRBMs are essential keys to the proof of a heavy traffic limit theorem for multiclass networks. In connection with this, Elizabeth Schwerer (a research assistant to Michael Harrison at the time) was the one who emphasized the pivotal role played by state space collapse. The author expresses her thanks to these two researchers for generously sharing their insights. Further thanks are extended to Jim Dai, Gang Wang and Yang Wang for permission to include their illuminating Example A.1 here, to Jim Dai for consultation concerning the verification of Assumption 7.1 for FIFO networks of Kelly type and static priority networks, to Maury Bramson for suggesting an idea for the proof of Lemma 8.2, to Tom Kurtz for helpful discussions on Lemmas 8.3 and 8.4, and to the referees for their helpful comments. Finally, the author thanks J. Michael Harrison and Maury Bramson for extensive discussions on the research reported here and for helpful comments on a preliminary version of this paper.

A Appendix: FIFO network of Kelly type for which the continuous mapping argument fails

We first describe the continuous mapping approach to construction of SRBMs. In the following, $\mathbf{C}^J = \{x : [0,\infty) \to \mathbb{R}^J,\ x \text{ is continuous}\}$ and $\mathbf{C}^J_+ = \{x \in \mathbf{C}^J : x(0) \in S\}$. We endow $\mathbf{C}^J$ (and hence $\mathbf{C}^J_+$) with the topology of uniform convergence on compact time intervals. Consider the following so-called deterministic Skorokhod problem for a given $J \times J$ matrix R.

Definition A.1 (Skorokhod problem) *Let $x \in \mathbf{C}^J_+$. Then $(w,y) \in \mathbf{C}^J \times \mathbf{C}^J$ solves the Skorokhod problem (SP) for x (with respect to R) if*

(i) $w(t) = x(t) + Ry(t) \in S$ *for all* $t \geq 0$,

(ii) y *is such that for* $j = 1, \ldots, J$,

- **(a)** $y_j(0) = 0$,
- **(b)** y_j *is non-decreasing,*
- **(c)** y_j *can increase only when* w_j *is zero, i.e.,* $\int_0^\infty w_j(t)\, dy_j(t) = 0$.

From the work of Harrison and Reiman [31] and Dupuis and Ishii [26] it is known that if R is of the form $I - Q$ where $|Q|$ (the matrix obtained by replacing each entry in Q by its absolute value) has spectral radius strictly less than one, then for each $x \in \mathbf{C}^J_+$ there is a unique pair (w,y) that solves the Skorokhod problem for x and moreover the mapping $\psi : \mathbf{C}^J_+ \to \mathbf{C}^J \times \mathbf{C}^J$ given by $\psi(x) = (w,y)$ is continuous and adapted (i.e., for each $t \geq 0$, the mapping $x \to (w(t), y(t))$ is $\mathcal{M}_t$-measurable where $\mathcal{M}_t = \sigma\{x(s) : 0 \leq s \leq t\}$). The continuous mapping ψ can then be used to construct a SRBM W with data $(S, \theta, \Gamma, R, \nu)$, as defined in Section 6, from a given (θ, Γ) Brownian motion X with initial distribution ν by defining $(W,Y) = \psi(X)$. Furthermore, under the assumptions mentioned at the beginning of this paragraph, ψ can be extended continuously to a mapping from $\mathbf{D}^J_+ = \{x \in \mathbf{D}^J : x(0) \in S\}$ into $\mathbf{D}^J \times \mathbf{D}^J$. One can then use this continuous mapping to turn approximations of X (obtained by functional central limit theorems) into approximations of W. This is a key element of the heavy traffic limit theorem proofs of Reiman [44], Peterson [42] and Chen and Zhang [15].

Unfortunately, uniqueness of solutions to the Skorokhod problem need not always hold, even when R is completely-$\mathcal{S}$. For example, Mandelbaum [40] has shown that for $J = 2$ there can be non-uniqueness of solutions of the Skorokhod problem for certain reflection matrices. In particular, one can adapt his idea to show non-uniqueness of such solutions for the reflection matrix

$$\hat{R} = \begin{pmatrix} 1 & \frac{4}{5} \\ -\frac{8}{5} & 1 \end{pmatrix}.$$

In such cases of non-uniqueness we do not know that there is a continuous path-to-path mapping that can be applied to yield an SRBM path from a Brownian motion path and so we cannot use this approach to prove a heavy traffic limit theorem. The following example illustrates that reflection matrices of this type can arise as SRBM data associated with a FIFO network of Kelly type. This example was provided to the author by Jim Dai in 1997 and is based on work of Jim Dai, Gang Wang and Yang Wang [23] done in 1992.

Example A.1 Consider a FIFO network of Kelly type as described in Section 3. The network has three stations, six customer classes, constituency matrix

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix},$$

and deterministic routing such that the only non-zero entries in P are $P_{k,k+1} = 1$ for $k = 1, 2, \ldots, 5$. External arrivals are assumed to only occur to class 1 and the (long run average) arrival rate is assumed to be $\alpha > 0$. The network is assumed to be of Kelly type and so there is only one mean service time per station. We let $\bar{m}_j$ denote the mean service time for customers at station j, $j = 1, 2, 3$. Service is on a FIFO basis at each station. The traffic intensity parameter for station j is given by $\rho_j = 2\alpha\bar{m}_j$, $j = 1, 2, \ldots$. If we suppose that α and $\bar{m}_j$ are fixed such that $\rho_j = 1$, then under the heavy traffic scaling the matrix G remains fixed and from (212) we have

$$I + G = \mathrm{diag}(\alpha\bar{m}_1, \alpha\bar{m}_2, \alpha\bar{m}_3) C (I - P')^{-1} C' = \frac{1}{2} \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 0 \\ 2 & 4 & 3 \end{pmatrix}.$$

Hence,

$$R \;=\; (I + G)^{-1} = \frac{2}{19} \begin{pmatrix} 1 & \frac{2}{5} & -\frac{6}{5} \\ -\frac{2}{3} & 1 & \frac{4}{5} \\ \frac{2}{9} & -\frac{8}{5} & 1 \end{pmatrix} \mathrm{diag}(9, 5, 5).$$

Note that, apart from normalization by a diagonal matrix, the matrix R contains the 2×2 submatrix $\hat{R}$ considered by Mandelbaum [40]. The non-uniqueness exhibited by Mandelbaum can be parlayed into non-uniqueness of solutions for the Skorokhod problem associated with R. Thus the continuous mapping approach cannot be used to prove a heavy traffic limit theorem for this FIFO network of Kelly type. Consequently, the approach presented in this paper, using the invariance principle for SRBMs, is currently the only viable approach for FIFO networks of Kelly type.

B Appendix: State space collapse is necessary for a heavy traffic limit theorem for FIFO networks

Proposition B.1 *Consider a sequence of open multiclass queueing networks with FIFO service that satisfies the heavy traffic assumptions of Section 5. Suppose that jointly with the convergence assumed in (86), we have*

$$(226) \qquad (\hat{W}^r, \hat{Z}^r) \Rightarrow (W, Z) \quad \text{as } r \to \infty$$

where W, Z are continuous processes. Then for Δ^r given by (61), (63), we have that state space collapse holds, i.e., (62) holds.

Proof. From (226), it follows that

$$(227) \qquad (\bar{W}^r, \bar{Z}^r, \hat{W}^r, \hat{Z}^r) \Rightarrow (\mathbf{0}, \mathbf{0}, W, Z) \quad \text{as } r \to \infty.$$

For a FIFO service discipline we have the additional model equation (cf. Bramson [11]):

$$(228) \qquad D_k^r(t + W_j^r(t)) - D_k^r(t) = Z_k^r(t) \quad \text{for all } t \geq 0, k \in \mathcal{C}(j), j \in \mathcal{J}.$$

By a similar proof to that for Lemma 8.2, we have that (91) holds. To see this, note that from (48) and (227), we have for each $T \geq 0$,

$$(229) \qquad \|\bar{A}^r(\cdot) - \bar{D}^r(\cdot)\|_T = \|\bar{Z}^r(\cdot) - \bar{Z}^r(0)\|_T \to 0 \quad \text{in probability as } r \to \infty.$$

Then one can proceed precisely as in the proof of Lemma 8.2 (from (138) onwards) to conclude that (91) holds. Applying (91), (226) and the heavy traffic assumptions of Section 5, we see from (58) that

$$(230) \qquad \hat{A}^r(\cdot) \Rightarrow Q\left(E(\cdot) + \sum_{k=1}^K \Phi^k(\lambda(\cdot)) - \tilde{P}(Z(\cdot) - Z(0))\right) \equiv A(\cdot) \quad \text{as } r \to \infty,$$

and this convergence is joint with (226) and (86). Combining this with (56) we have

$$(231) \qquad \hat{D}^r(\cdot) \Rightarrow A(\cdot) + (Z(0) - Z(\cdot)) \equiv D(\cdot) \quad \text{as } r \to \infty,$$

jointly with (230), (226) and (86). Now, by (228), for $j \in \mathcal{J}$, $k \in \mathcal{C}(j)$ and $t \geq 0$,

$$\begin{aligned}\hat{Z}_k^r(t) &= \frac{D_k^r(r^2(t+\bar{W}_j^r(t))) - D_k^r(r^2t)}{r} \\ &= \hat{D}_k^r(t+\bar{W}_j^r(t)) - \hat{D}_k^r(t) + \lambda_k^r \hat{W}_j^r(t).\end{aligned}$$

It follows from the fact that $\|\bar{W}^r(\cdot)\|_T \to 0$ in probability for each $T \geq 0$, (226) and (231) that for $j \in \mathcal{J}$, $k \in \mathcal{C}(j)$,

(232) $$\hat{Z}_k^r(\cdot) \Rightarrow D_k(\cdot) - D_k(\cdot) + \lambda_k W_j(\cdot) = \Delta_{kj} W_j(\cdot) \quad \text{as } r \to \infty,$$

jointly with (226), where for $k \in \mathcal{K}$,

(233) $$\Delta_{kj} = \begin{cases} \lambda_k & \text{if } k \in \mathcal{C}(j), \\ 0 & \text{if } k \notin \mathcal{C}(j). \end{cases}$$

It then follows from the fact that $\Delta^r \to \Delta$, (232) and (226) that

$$\begin{aligned}\hat{Z}^r(\cdot) - \Delta^r \hat{W}^r(\cdot) &= \hat{Z}^r(\cdot) - \Delta W(\cdot) + \Delta(W(\cdot) - \hat{W}^r(\cdot)) + (\Delta - \Delta^r)\hat{W}^r(\cdot) \\ &\Rightarrow \mathbf{0} \quad \text{as } r \to \infty,\end{aligned}$$

and hence (62) holds. □

Remark. For the case when the networks are initially empty, i.e., $\hat{W}^r(0) = 0$ and $\hat{Z}^r(0) = 0$ for all r, the above result can be improved using elements of the proofs in Dai-Nguyen [21]. In this case, using an alternative FIFO model equation to (228) that holds when a network is initially empty, one can show that the conclusion of Proposition B.1 still holds when (226) is replaced by the simpler assumption that $\hat{W}^r \Rightarrow W$ as $r \to \infty$, where W is a continuous process. While it is likely that this result can be extended to networks that are initially non-empty, a generalization of the model equation used by Dai and Nguyen would have to be found. We leave such a generalization for further research.

References

[1] Berman, A., and Plemmons, R. J. (1994). *Nonnegative Matrices in the Mathematical Sciences.* SIAM, Philadelphia, PA.

[2] Bernard, A., and El Kharroubi, A. (1991). Régulation de processus dans le premier orthant de $\mathbb{R}^n$. *Stochastics and Stochastics Reports,* **34,** 149–167.

[3] Bertsekas, D., and Gallagher, R. (1992). *Data Networks.* Prentice-Hall, Englewood Cliffs, N.J.

[4] Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley and Sons, New York.

[5] Bramson, M. (1994). Instability of FIFO queueing networks. *Annals of Applied Probability*, **4**, 414–431.

[6] Bramson, M. (1994). Instability of FIFO queueing networks with quick service times. *Annals of Applied Probability*, **4**, 693–718.

[7] Bramson, M. (1995). Two badly behaved queueing networks. In *Stochastic Networks,* IMA Volumes in Mathematics and Its Applications, Kelly, F. P., and Williams, R. J. (eds.), **71,** Springer-Verlag, New York, 105–116.

[8] Bramson, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems: Theory and Applications,* **22,** 5–45.

[9] Bramson, M. (1996). Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems: Theory and Applications,* **23,** 1–26.

[10] Bramson, M. (1998). Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems: Theory and Applications,* **28,** 7–31.

[11] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. To appear in *Queueing Systems: Theory and Applications.*

[12] Carlsson, H., and Nerman, O. (1986). An alternative proof of Lorden's renewal inequality. *Adv. Appl. Prob.* **18,** 1015–1016.

[13] Chen, H., and Mandelbaum, A. (1991). Leontief Systems, RBV's and RBM's. In *Applied Stochastic Analysis,* M. H. A. Davis and R. J. Elliott (eds.), Gordon and Breach Science Publishers, New York, pp. 1–43.

[14] Chen, H., and Whitt, W. (1993). Diffusion approximations for open queueing networks with service interruptions. *Queueing Systems: Theory and Applications,* **13,** 335–359.

[15] Chen, H., and Zhang, H. (1996). Diffusion approximations for re-entrant lines with a first-buffer-first-served priority discipline. *Queueing Systems: Theory and Applications,* **23,** 177–195.

[16] Chen, H., and Zhang, H. (1997). Diffusion approximations for some multiclass queueing networks with FIFO service disciplines. Preprint.

[17] Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability*, **5**, 49–77.

[18] Dai, J. G. (1995). Stability of open multiclass queueing networks via fluid models. In *Stochastic Networks,* IMA Volumes in Mathematics and Its Applications, Kelly, F. P., and Williams, R. J. (eds.), **71,** Springer-Verlag, New York, 71–90.

[19] Dai, J. G., and Dai, W. (1997). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. Submitted to *Queueing Systems: Theory and Applications.*

[20] Dai, J. G., and Harrison, J. M. (1993). The QNET method for two-moment analysis of closed manufacturing systems. *Annals of Applied Probability,* **3,** 968–1012.

[21] Dai, J. G., and Nguyen, V. (1994). On the convergence of multiclass queueing networks in heavy traffic. *Annals of Applied Probability*, **4**, 26–42.

[22] Dai, J. G., and Wang, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems: Theory and Applications*, **13**, 41–46.

[23] Dai, J. G., Wang, G., and Wang, Y. (1992). Private communication.

[24] Dai, J. G., and Williams, R. J. (1995). Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Theory of Probability and Its Applications*, **40,** 1–40.

[25] Dai, J. G., Yeh, D. H., and Zhou, C. (1997). The QNET method for re-entrant queueing networks with priority disciplines. *Operations Research*, **45,** 610–623.

[26] Dupuis, P., and Ishii, H. (1991). On the Lipschitz continuity of the solution mapping to the Skorokhod problem. *Stochastics and Stochastics Reports,* **35,** 31–62.

[27] Ethier, S. N., and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence.* Wiley, New York.

[28] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Volumes in Mathematics and Its Applications, W. Fleming and P.-L. Lions (eds.), Springer-Verlag, New York, 147–186.

[29] Harrison, J. M. (1995). Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. In *Stochastic Networks*, IMA Volumes in Mathematics and Its Applications, F. P. Kelly and R. J. Williams (eds.), **71**, Springer-Verlag, New York, 1–20.

[30] Harrison, J. M., and Nguyen, V. (1993). Brownian models of multiclass queueing networks: current status and open problems. *Queueing Systems: Theory and Applications*, **13**, 5–40.

[31] Harrison, J. M., and Reiman, M. I. (1981). Reflected Brownian motion on an orthant. *Annals of Probability,* **9,** 302–308.

[32] Harrison, J. M., and Williams, R. J. (1992). Brownian models of feedforward queueing networks: quasireversibility and product form solutions. *Annals of Applied Probability*, **2,** 263–293.

[33] Harrison, J. W., and Williams, R. J. (1996). A multiclass closed queueing network with unconventional heavy traffic behavior. *Annals of Applied Probability*, **6,** 1–47.

[34] Iglehart, D. L., and Whitt, W. (1970). Multiple channel queues in heavy traffic I. *Adv. Appl. Prob.*, **2**, 150–177.

[35] Iglehart, D. L., and Whitt, W. (1970). Multiple channel queues in heavy traffic II. *Adv. Appl. Prob.*, **2**, 355–364.

[36] Iglehart, D. L., and Whitt, W. (1971). The equivalence of functional central limit theorems for counting processes and associated partial sums. *Annals of Mathematical Statistics,* **42,** 1372–1378.

[37] Kelly, F. P., and Williams, R. J. (eds.) (1995). *Stochastic Networks,* IMA Volumes in Mathematics and Its Applications, **71** Springer-Verlag, New York.

[38] Kumar, P. R. (1995). Scheduling queueing networks: stability, performance analysis and design. In *Stochastic Networks,* IMA Volumes in Mathematics and Its Applications, Kelly, F. P., and Williams, R. J. (eds.), **71,** Springer-Verlag, New York, pp. 21–70.

[39] Lindvall, T. (1992). *Lectures on the Coupling Method.* Wiley, New York.

[40] Mandelbaum, A. (1992). The dynamic complementarity problem. Preprint.

[41] Mandelbaum, A., and Van der Heyden, L. (1987, unpublished work). Complementarity and reflection.

[42] Peterson, W. P. (1991). Diffusion approximations for networks of queues with multiple customer types. *Mathematics of Operations Research*, **9**, 90–118.

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

[44] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Mathematics of Operations Research* **9**, 441–458.

[45] Reiman, M. I. (1984). Some diffusion approximations with state space collapse. In *Proceedings International Seminar on Modeling and Performance Evaluation Methodology,* Lecture Notes in Control and Information Sciences, F. Baccelli and G. Fayolle (eds.), Springer-Verlag, New York, 209–240.

[46] Reiman, M. I. (1988). A multiclass feedback queue in heavy traffic. *Adv. Appl. Prob.*, **20**, 179–207.

[47] Reiman, M. I., and Williams, R. J. (1988–89). A boundary property of semimartingale reflecting Brownian motions. *Probability Theory and Related Fields,* **77,** 87–97, and **80,** 633.

[48] Skorokhod, A. V. (1956). Limit Theorems for Stochastic Processes. *Theory of Probability and Its Applications*, **1**, 261–290.

[49] Taylor, L. M., and Williams, R. J. (1993). Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probability Theory and Related Fields*, **96**, 283–317.

[50] Whitt, W. (1971). Weak convergence theorems for priority queues: preemptive resume discipline. *J. Applied Probability*, **8**, 74–94.

[51] Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Science*, **39,** 1020–1028.

[52] Williams, R. J. (1996). On the approximation of queueing networks in heavy traffic. In *Stochastic Networks: Theory and Applications,* F. P. Kelly, S. Zachary, and I. Ziedins (eds.), Oxford University Press, Oxford, pp. 35–56.

[53] Williams, R. J. (1998). An invariance principle for semimartingale reflecting Brownian motions in an orthant. To appear in *Queueing Systems: Theory and Applications.*

[54] Yao, D. D. (ed.) (1994). *Stochastic Modeling and Analysis of Manufacturing Systems.* Springer-Verlag, New York.