Heavy Traffic Limits for Some Queueing Networks

Maury Bramson[1]
School of Mathematics, University of Minnesota, Minneapolis, MN 55455
bramson@math.umn.edu

J. G. Dai[2]
School of Industrial and Systems Engineering and School of Mathematics
Georgia Institute of Technology, Atlanta, GA 30332-0205
dai@isye.gatech.edu

August 30, 1999

Abstract

Using a slight modification of the framework in Bramson [7] and Williams [52], we prove heavy traffic limit theorems for six families of multiclass queueing networks. The first three families are single station systems operating under first-in first-out (FIFO), generalized head-of-the-line proportional processor sharing (GHLPPS) and static buffer priority (SBP) service disciplines. The next two families are re-entrant lines operating under first-buffer-first-serve (FBFS) and last-buffer-first-serve (LBFS) service disciplines; the last family consists of certain 2-station, 5-class networks operating under an SBP service discipline. Some of these heavy traffic limits have appeared earlier in the literature; our new proofs demonstrate the significant simplifications that can be achieved in the present setting.

AMS 1991 subject classification. Primary 60K25, 60F17, 60G17; secondary 60J15, 90B22, 68M20.

Key words and phrases. Multiclass queueing network, Brownian model, heavy traffic, reflecting Brownian motion, diffusion approximation.

1 Introduction

Queueing networks have been extensively used to model computer systems, telecommunications networks, and manufacturing systems (see, e.g., Bertsekas and Gallager [2], and Yao [54]). Classical queueing network theory imposes restrictive assumptions on the distributions of the interarrival and service times, and on the service disciplines employed in a queueing network (Jackson [35], Baskett et. al. [1] and Kelly [37]). These restrictions exclude the use of such theory for many practical systems. Brownian model approximations have been employed as an alternative tool for more general queueing networks (see, e.g., Harrison and Nguyen [27]). They share two distinctive features: (a) the analysis of a Brownian model is mathematically more tractable than that of the corresponding queueing network, since a complicated Markov chain is replaced by a diffusion process and (b) the Brownian model uses just the first two moments of the interarrival and service times, and of the routing vectors associated with the queueing network.

In formulating the Brownian model for a queueing network, one replaces the workload process of the queueing network by a multi-dimensional semimartingale reflecting Brownian (SRBM). Many quantities for the SRBM, including the stationary distribution, can be computed either exactly or numerically (Harrison and Williams [30], Dai and Harrison [16]). Ideally, these should provide

[1]Research supported in part by NSF grant DMS-9971248

[2]Research supported in part by NSF grants DMI-9457336 and DMI-9813345, and by MaPhySto—Centre for Mathematical Physics and Stochastics, funded by The Danish National Research Foundation

estimates for the corresponding queueing networks. Unfortunately, this is not always the case (Dai and Wang [21]); it is thus essential to determine when a Brownian model can be used for the analysis of a queueing network. This task is often carried out by establishing a *heavy traffic limit* for a sequence of related queueing networks, which justifies the comparison with a Brownian model when each server is heavily utilized. Such an assumption is reasonable in many systems, including semiconductor wafer production lines, where extremely high capital cost of equipment demands high utilization of machines.

The order in which jobs at a station are executed is an important component in heavy traffic limits. When each station has a single class of jobs, a queueing network is referred to as a *single-class* network; when at least one station has more than one job class, it is a *multiclass* network. In the latter setting, a policy dictating the order in which jobs at each station are served is called a *service discipline*. Examples of service disciplines include first-in first-out (FIFO), generalized head-of-the-line proportional processor sharing (GHLPPS) and static buffer priority (SBP) disciplines, each of which will be defined in Section 2. When the routing is deterministic and only one class has external arrivals, the network is called a *re-entrant line*.

In a typical setup for heavy traffic limits, one considers a sequence of queueing networks, indexed by r. The basic network topology remains fixed across the entire sequence, with, however, the arrival and service rates, and the corresponding distributions varying over r. As $r \to \infty$, the traffic intensity at each station is assumed to converge to 1, i.e., each service station is *critically loaded* in the limit. In this setting, one expects the queue length and workload processes to typically grow without bound as $r \to \infty$. For standard heavy traffic limits, under diffusive scaling, the workload processes converge to a limit which is an SRBM, with dimension equal to the number of stations in each network. The corresponding limit of the queue lengths, with dimension equal to the number of classes in each network, will be a constant multiple of the workload limit. This last property is an example of *state space collapse*, a term first used in Reiman [44], although such phenomena were observed earlier in Whitt [48] and Foschini and Salz [26]. The nature of the queue length limit will be strongly influenced by the service discipline for the networks in the sequence.

The study of heavy traffic limits of queueing systems has a long history, which dates back to Kingman [38, 39], Prohorov [42], Borovkov [3, 4] and Iglehart [32]. Heavy traffic limits, in the form of functional central limit theorems, were first studied by Iglehart and Whitt [33, 34]; a survey can be found in Whitt [49]. Reiman [43] proved a heavy traffic limit theorem for single-class networks; his proof was simplified by Johnson [36] by studying the corresponding fluid models. (Readers are referred to Chen and Mandelbaum [8] for a survey on single-class networks.) There have been a number of heavy traffic limits for multiclass queueing networks in Whitt [48], Peterson [41], Reiman [45], Dai and Kurtz [17], Chen and Zhang [11, 12, 10], Bramson [7] and Williams [52]; Williams [51] provides a survey. Examples of non-existence of heavy traffic limits were given in Whitt [50], and Dai and Nguyen [18]. Unconventional heavy traffic limits were obtained in Harrison and Williams [31], and Coffman et. al. [14].

In this paper, we establish heavy traffic limits for six families of multiclass queueing networks. The first three families are single station systems operating under FIFO, GHLPPS and SBP service disciplines. The next two families are re-entrant lines operating under first-buffer-first-serve (FBFS) and last-buffer-first-serve (LBFS) service disciplines. Last, we prove that, under an additional, unconventional heavy traffic condition, the heavy traffic limit holds for a given family of 2-station, 5-class networks operating under an SBP service discipline. Our proofs of the heavy traffic limit theorems are based on a slight modification of the framework given in Bramson [7] and Williams [52], in which state space collapse and fluid limits play a central role. Our criteria will consist of showing that (a) the reflecting matrix R, corresponding to the sequence of queueing networks, is completely-

$\mathcal{S}$ and (b) the critically loaded fluid model, corresponding to the queueing networks, is uniformly convergent.

Some of our heavy traffic results are not new. Heavy traffic limits for sequences of FIFO single station systems were established in Reiman [43] and in Dai and Kurtz [17]. Chen and Zhang [11] proved a heavy traffic limit for a family of FBFS re-entrant lines. These known results are included here to show how one may significantly simplify their proofs by using the framework of Bramson [7] and Williams [52]. Presumably, this framework can be employed for further heavy traffic limit theorems. It has been reported to us that, in a contemporaneous independent work, Chen and Ye [9] have shown a heavy traffic limit theorem for LBFS re-entrant lines by using a related framework from Chen and Zhang [13].

The paper is organized as follows. Multiclass networks are introduced in Section 2. In Section 3, we present the background for and state our main results on heavy traffic limits; the remainder of the paper is devoted to demonstrating these results. The equations of the queueing networks we consider and the corresponding fluid model equations are given in Section 4. The framework of Bramson [7] and Williams [52] is applied to our setting in Section 5. The proofs of our heavy traffic limit theorems are given in Sections 6-8. Such limits are demonstrated for single station systems (Theorems 3.1-3.3) in Section 6, and for FBFS and LBFS re-entrant lines in Section 7 (Theorems 3.4-3.5). In Section 8, heavy traffic limits are demonstrated for a particular family of 2-station, 5-class re-entrant lines (Theorem 3.6).

2 Open multiclass queueing networks

Multiclass queueing networks were introduced in Section 1; in this section, we give a more detailed description of these networks. Each station is assumed to have a single server, with unlimited waiting space. When a job arrives from outside the network, it receives service at a finite number of stations, after which it leaves the network. At any given time during its lifetime in the network, the job belongs to one of the job *classes*. The job changes classes as it moves through the network, changing classes each time a service is completed; all jobs within a class are served at a unique station. Since the network is multiclass, more than one class might be served at a station. Each job is assumed to eventually leave the network. The ordered sequence of classes that a job visits in the network is called its *route*; if all jobs follow the same route, the network is called a *re-entrant line*.

We use J to denote the number of service stations and K to denote the number of job classes in the network. Stations are labelled $j = 1, \ldots, J$, and classes by $k = 1, \ldots, K$. We use $\mathcal{C}(j)$ to denote the set of classes belonging to station j, and $s(k)$ to denote the station to which class k belongs; when j and k appear together, we implicitly set $j = s(k)$. Associated with each class k of a queueing network, there are two i.i.d. sequences of random variables, $u_k = \{u_k(i),\ i \geq 1\}$ and $v_k = \{v_k(i),\ i \geq 1\}$, an i.i.d. sequence of K-dimensional random vectors, $\phi^k = \{\phi^k(i),\ i \geq 1\}$, and two real numbers, $\alpha_k \geq 0$ and $m_k > 0$. We assume that $3K$ sequences

$$u_1, \ldots, u_K, v_1, \ldots, v_K, \phi^1, \ldots, \phi^K \tag{2.1}$$

are mutually independent. We refer to them as the *primitive increments* of the network. We set $a_k = \mathrm{var}(u_k(1))$ and $b_k = \mathrm{var}(v_k(1))$, and assume that $a_k < \infty$ and $b_k < \infty$, and that u_k and v_k are *unitized*, i.e., $\mathbb{E}[u_k(1)] = 1$ and $\mathbb{E}[v_k(1)] = 1$. For each i, $u_k(i)/\alpha_k$ will denote the interarrival time between the $(i-1)$th and the ith *externally* arriving job at class k, $m_k v_k(i)$ will denote the service time for the ith class k job, and $\phi^k(i)$ will denote the routing vector of the ith class k job.

It follows that, for each class k, m_k is the mean service time for class k jobs, α_k is the external arrival rate to class k, and a_k and b_k are the squared coefficients of variation for interarrival and service times. (The squared coefficient of variation of a positive random variable is defined to be the variance divided by the squared mean.) We allow $\alpha_k = 0$ for some classes k, and we set $\mathcal{E} = \{k : \alpha_k \neq 0\}$. We assume that the routing vector $\phi^k(i)$ takes values in $\{e_0, e_1, \ldots, e_K\}$, where e_0 is the K-dimensional vector of all 0's and, for $\ell = 1, \ldots, K$, e_ℓ is the K-dimensional vector with ℓth component 1 and other components 0. When $\phi^k(i) = e_\ell$, the ith job departing class k becomes a class ℓ job. We let $P_{k\ell} = \mathbb{P}\{\phi^k(i) = e_\ell\}$ be the probability that a job departing class k becomes a class ℓ job. The $K \times K$ matrix $P = (P_{k\ell})$ is the *routing matrix* of the network. We assume our networks are *open*, that is, the matrix

$$Q \stackrel{\text{def}}{=} I + P' + (P')^2 + \cdots$$

is finite, which is equivalent to $(I - P')$ being invertible, with $Q = (I - P')^{-1}$. (The symbol $'$ on a vector or a matrix denotes the transpose.)

We define the cumulative arrival, cumulative service and cumulative routing processes by the sums

$$U_k(n) = \sum_{i=1}^{n} u_k(i), \quad V_k(n) = \sum_{i=1}^{n} v_k(i), \quad \Phi^k(n) = \sum_{i=1}^{n} \phi^k(i),$$

where $n = 1, 2, \ldots$ and $k = 1, \ldots, K$. For each class k, $m_k V_k(n)$ is the total amount of service required for the first n class k jobs. Also, for each k and $t \geq 0$, let $E_k = \{E_k(t),\ t \geq 0\}$ denote the renewal process associated with the i.i.d. sequence $\{u_k(i),\ i \geq 1\}$, i.e.,

$$E_k(t) = \max\{n : U_k(n) \leq t\}.$$

For $t \geq 0$, $E_k(\alpha_k t)$ counts the number of external arrivals to class k in $(0, t]$. We also write $V_k = \{V_k(n),\ n \geq 1\}$ and $\Phi^k = \{\Phi^k(n),\ n \geq 1\}$. The processes

$$E_1, \ldots, E_k, V_1, \ldots, V_K, \Phi^1, \ldots, \Phi^K \tag{2.2}$$

are referred to as the *primitive processes*. They contain the same information as the primitive increments in (2.1).

A service discipline dictates the order in which jobs are served at each station. A service discipline is said to be *non-idling* if a server is always active when there are jobs waiting to be served at its station. In this paper, we restrict our disciplines to three families of disciplines, first-in first-out (FIFO), generalized head-of-the-line proportional processor sharing (GHLPPS) and static buffer priority (SBP), which are defined below.

Under the FIFO discipline, jobs at each station are served on a first-in first-out basis, regardless of their class designations. Under a GHLPPS service discipline with *weight vector* $\beta = (\beta_1, \ldots, \beta_K)$, with $\beta_k > 0$ for all k, the server at each station simultaneously serves the leading job of each (non-empty) class. The server allocates effort to each class k in proportion to the number of jobs in that class, weighted by β_k. Such disciplines are mathematical idealizations of certain round-robin processor sharing disciplines, which are common in telecommunication networks. When the weight vector $\beta = (1, \ldots, 1)$, the GHLPPS discipline becomes the head-of-the-line proportional processor sharing (HLPPS) service discipline in Bramson [5, 7] and Williams [52].

Under an SBP discipline, the classes at each station are assigned a fixed ranking. When the server switches from one job to another, the new job will be taken from the leading (or longest waiting) job at the highest ranking non-empty class at the server's station. We assume that the

ranking is strict, i.e., there is no tie in the ranking. We also assume that the service discipline is *preemptive-resume*. That is, when a job, with a higher rank than the one currently being served, arrives at the server's station, the service of the current job is interrupted. When service of all jobs with higher ranks is completed, the interrupted service continues from where it left off. Two SBP disciplines for re-entrant lines that have been studied in the literature are first-buffer first-served (FBFS) and last-buffer first-served (LBFS). Under the FBFS discipline, earlier classes along the route are assigned higher priorities. Under the LBFS discipline, later classes along the route are assigned higher priorities.

All of these disciplines are examples of *head-of-the-line* (HL) disciplines, that is, only the leading job from each class may receive service at any given time. It is assumed that the discipline is non-idling, and that service within each class is on a FIFO 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. 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. Readers are referred to Bramson [5] for precise definition of such disciplines. (Williams [52] gives a slightly more general definition.)

In this paper, we focus our study on six network models. The first three consist of the multiclass single server stations (i.e., $J = 1$) under the FIFO, GHLPPS and SBP service disciplines. The other three models are the family of re-entrant lines under FBFS and LBFS disciplines, and the 2-station, 5-class re-entrant line pictured in Figure 1, in Section 3.

3 Heavy traffic limit results

In order to state our heavy traffic results, we require additional terminology. This is provided in Sections 3.1-3.5, where performance processes, traffic equations, initial conditions, scaling and heavy traffic conditions, and the definition of reflecting Brownian motion are discussed. Our heavy traffic results are then presented in Section 3.6.

3.1 Performance processes

The following processes Z, D, W, and Y will be used to measure the performance of our queueing network. The processes $Z = \{Z(t),\ t \geq 0\}$ and $D = \{D(t),\ t \geq 0\}$ are both K-dimensional, with $Z_k(t)$ denoting the number of class k jobs at time t, and $D_k(t)$ denoting the cumulative number of departures from class k over $[0, t]$. They are called the *queue length process* and *departure process*, respectively. The other two processes, $W = \{W(t),\ t \geq 0\}$ and $Y = \{Y(t),\ t \geq 0\}$, are both J-dimensional. For each station j, $W_j(t)$ denotes the amount of work for server j (measured in units of remaining service time) embodied in those jobs who are at station j at time t. If no more arrivals (either external and internal) are allowed at station j after time t, server j needs to work $W_j(t)$ additional units of time before the station is empty. The process W is called the (immediate) *workload process*. For each station j, $Y_j(t)$ denotes the total amount of time that the server at station j has been idle over $[0, t]$. Y 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.

3.2 Traffic equations

To investigate open multiclass queueing networks, one employs the solution λ_ℓ, $\ell = 1, \ldots, K$, of the *traffic equations*

$$\lambda_\ell = \alpha_\ell + \sum_{k=1}^{K} \lambda_k P_{kl}, \tag{3.1}$$

or equivalently, in vector form, of $\lambda = \alpha + P'\lambda$. (All vectors in this paper are to be interpreted as column vectors unless explicitly stated otherwise.) Since the network corresponding to P is open, the unique solution in (3.1) of λ is $\lambda = Q\alpha$. The term λ_k is referred to as the *nominal total arrival rate* at class k; it depends on both external and internal arrivals. If, for each class k, there is a long-run average rate of flow into the class which is equal to the long-run average rate out of that class, this rate will equal λ_k.

Employing m and λ, one defines the *traffic intensity* ρ_j for the jth server as

$$\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k. \tag{3.2}$$

In vector form, ρ is given by $\rho = CM\lambda$, where $M = \mathrm{diag}(m)$ and C is the *constituency matrix*

$$C_{jk} = \begin{cases} 1 & \text{if } k \in \mathcal{C}(j), \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

(For a d-dimensional vector x, $\mathrm{diag}(x)$ denotes the $d \times d$ matrix whose diagonal entries are given by the components of x and all other entries are 0.) When $\rho_j \leq 1$, ρ_j is also referred to as the nominal fraction of time that server j is busy. In this paper, we are interested in networks in which ρ_j is close to one for each station j. Such networks are said to be "heavily loaded."

3.3 Initial conditions

Heavy traffic limit theorems have frequently required the corresponding networks to be empty initially [10, 17, 41, 45]. Here, we allow each class to have a positive number of jobs at time 0; we assume that the probabilistic behavior of these jobs is the same as for jobs arriving at the class after time 0 (in terms of service requirement and routing). This assumption is less general than those in [7, 52], where the initial jobs are allowed to have different service time distributions. We restrict ourselves to the current framework to keep the exposition simple.

Depending on the discipline used, the amount of information encoded in the initial state can differ. The initial state should have enough information so that, under the service discipline, the evolution of the queueing network is completely determined by the initial state and the primitive processes in (2.2). For a GHLPPS or an SBP discipline, we take $Z(0)$ to be the initial state. (Recall that the kth component $Z_k(0)$ is the number of jobs initially in class k.) For a FIFO service discipline, however, one needs to specify the order of the initial jobs at each station or, equivalently, the order in which the initial jobs depart after their service completions. The initial state for a network with the FIFO discipline is given by

$$\{D_k(s), \text{ for } 0 \leq s \leq W_j(0)\} \quad \text{for each } k \in \mathcal{C}(j).$$

3.4 Scaling and heavy traffic conditions

We will use α^r and m^r to denote the vectors of the external arrival rates and mean service times for a family of networks indexed by r, where r tends to infinity through a strictly increasing family of values in $(0, \infty)$. (With some abuse of notation, we refer to such networks as a sequence of networks.) Let $\lambda^r = Q\alpha^r$ and $\rho^r = CM^r\lambda^r$, with $M^r = \text{diag}(m^r)$. We assume that the set $\mathcal{E} = \{k : \alpha_k^r \neq 0\}$ and the routing matrix P do not depend on r. We assume further that α^r and m^r are so chosen that, as $r \to \infty$,

$$\alpha_k^r \to \alpha_k > 0 \quad \text{for } k \in \mathcal{E}, \qquad m_k^r \to m_k > 0 \quad \text{for } k = 1, \ldots, K, \tag{3.4}$$

and that $\rho^r \to e$ at the rate

$$r(\rho^r - e) \to \gamma, \tag{3.5}$$

where e is the J-dimensional vector of all 1's and γ is some J-dimensional vector. Note that (3.4)-(3.5) imply that

$$\rho = CM\lambda = e, \tag{3.6}$$

namely, each station is *critically loaded* in the limit. The interarrival times at class k are given by $\{u_k(i)/\alpha_k^r,\ i = 1, \ldots\}$ and the service times by $\{m_k^r v_k(i) : i = 1, \ldots\}$. Therefore, the squared coefficients of variation of the interarrival times and service times for class k, a_k and b_k, and do not depend on the index r.

Conditions (3.4)-(3.5) are referred to as heavy traffic conditions; they will be employed in Section 3.6. Readers who are not familiar with this setting may be puzzled by our reason for introducing a *sequence* of networks. As motivation, one can consider the following situation. In a production system, it is up to the manager to decide how quickly jobs are to be released into the system. In particular, one needs to decide how heavily the system should be loaded in order to effectively use its resources. Ideally, one would like to choose each ρ_j close to 1. A sequence corresponding to such a network arises by varying the load condition imposed by the manager; one envisions the network as a member of the sequence, with r chosen large since ρ is close to e. The heavy traffic limit corresponding to this sequence of networks should then provide insight on the behavior of the original network. For the re-entrant line pictured in Figure 1 of Section 3.6, with m constant and satisfying

$$m_1 + m_3 + m_5 = m_2 + m_4,$$

let $\alpha_1^r = 1/(m_1 + m_3 + m_5) - 1/r$, with $r > 0$. Then, $\rho_1 = \rho_2 = 1 - (m_1 + m_3 + m_5)/r$, and so the heavy traffic conditions (3.4)-(3.5) are satisfied with $\gamma_1 = \gamma_2 = -(m_1 + m_3 + m_5)$.

When $\rho^r \to e$ as $r \to \infty$, we expect the queue length, workload and idletime processes to grow. With functional central limit theorems in mind, we define the scaled queue length process $\tilde{Z}^r(t) = (\tilde{Z}_1^r(t), \ldots, \tilde{Z}_K^r(t))'$, by

$$\tilde{Z}_k^r(t) = r^{-1} Z_k^r(r^2 t).$$

For $0 \leq t \leq 1$, the scaled process $\tilde{Z}^r(t)$ records the queue lengths over $[0, r^2]$ at resolution $1/r$ (each job is assigned weight $1/r$). As r increases, the scaled process employs longer and longer time intervals at coarser and coarser resolutions. We similarly define $\tilde{W}^r(t)$ and $\tilde{Y}^r(t)$ by

$$\tilde{W}_k^r(t) = r^{-1} W_k^r(r^2 t) \quad \text{and} \quad \tilde{Y}_k^r(t) = r^{-1} Y_k^r(r^2 t).$$

3.5 Reflecting Brownian motion

In this section, we recall the definition of semimartingale reflecting Brownian motion (SRBM). Such processes will be the limits for our heavy traffic limit theorems. Throughout this section, $\mathcal{B}$ denotes the σ-algebra of Borel subsets of $\mathbb{R}_+^J$, θ is a vector in $\mathbb{R}^J$, Γ is a $J \times J$ symmetric and strictly positive definite matrix, R is a $J \times J$ matrix, and ν is a probability measure on $(\mathbb{R}_+^J, \mathcal{B})$.

The following definition of an SRBM is taken from Williams [53, Section 6].

Definition 3.1 (SRBM). An SRBM associated with the data $(\mathbb{R}_+^J, \theta, \Gamma, R, \nu)$ 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 $\mathbb{P}$-a.s.:

(i) W has continuous paths with $W(t) \in \mathbb{R}_+^J$ for $t \geq 0$, and

(ii) $W = X + RY$ for appropriate J-dimensional processes X and Y.

The processes X and Y satisfy the following properties. Under $\mathbb{P}$,

(iii) X is a Brownian motion with drift vector θ and covariance matrix Γ, such that $X(0)$ has distribution ν, and

(iv) $\{X(t) - X(0) - \theta t,\ \mathcal{F}_t,\ t \geq 0\}$ is a martingale.

The process Y is an $\{\mathcal{F}_t\}$-adapted, J-dimensional process such that $\mathbb{P}$-a.s., for each $j = 1, \ldots, J$,

(v) $Y_j(0) = 0$,

(vi) Y_j is continuous and nondecreasing,

(vii) Y_j can increase only at times t where $W_j(t) = 0$.

In (vii), we mean that, for each $t > 0$, $W_j(t) > 0$ implies $Y_j(t - \delta) = Y_j(t + \delta)$ for some $\delta > 0$. This is equivalent to $\int_0^\infty W_j(s) dY_j(s) = 0$ for all j. Loosely speaking, an SRBM behaves like a Brownian motion with drift vector θ and covariance matrix Γ in the interior of the orthant $\mathbb{R}_+^J$, with the processes being confined to the orthant by instantaneous "reflection" (or "pushing") at the boundary, where the direction of "reflection" on the jth face, $F_j \equiv \{x \in \mathbb{R}_+^J : x_j = 0\}$, is given by the jth column of R. The parameters θ, Γ and R are called the *drift vector*, *covariance matrix* and *reflection matrix* of the SRBM, respectively. Results of Reiman and Williams [46] and Taylor and Williams [47] show that a necessary and sufficient condition for the existence and uniqueness (in distribution) of the SRBM associated with $(\mathbb{R}_+^J, \theta, \Gamma, R, \nu)$, for each initial distribution ν on $(\mathbb{R}_+^J, \mathcal{B})$, is that the reflection matrix R be completely-$\mathcal{S}$, which is defined as follows. For a $J \times J$ matrix R and a subset $\mathcal{J} \subset \{1, \ldots, J\}$, the principal submatrix associated with $\mathcal{J}$ is the $|\mathcal{J}| \times |\mathcal{J}|$ matrix obtained from R by deleting the rows and columns that are not in $\mathcal{J}$, where $|\mathcal{J}|$ is the cardinality of $\mathcal{J}$. A $J \times J$ matrix R is an $\mathcal{S}$ matrix if there exists $u \geq 0$ such that $Ru > 0$. (Vector inequalities are to be interpreted componentwise.) The matrix R is *completely-$\mathcal{S}$* if each principal submatrix of R is an $\mathcal{S}$ matrix.

In Definition 3.1, the SRBM W has the semimartingale decomposition (ii) with respect to a Brownian motion X defined on *some* probability space. In the stochastic differential equation literature, such a W is called a *weak solution* of (i)-(vii). If, for a Brownian motion X defined on a given probability space, one can find W and Y that are defined on the same probability space, are adapted to X, and satisfy conditions (i)-(vii) of Definition 3.1 with $\{\mathcal{F}_t\}$ being the filtration generated by X, then W is called a *strong solution* of (i)-(vii). Note that, in the strong solution

setting, condition (iv) is redundant because Brownian motion minus the drift is always a martingale with respect to its own filtration. If the reflection matrix R satisfies an appropriate spectral radius condition, such as in Harrison and Reiman [29], the strong solution always exists and is unique.

3.6 Heavy traffic limit theorems

We state here the heavy traffic limit theorems 3.1-3.6, which are the main results of the paper. For these results, we will need some general assumptions. Recall that α and m are the limits in (3.4) and that $\lambda = Q\alpha$. We will henceforth assume that (3.4) holds, and that $\lambda_k > 0$ for all k.

Let H^k be the $K \times K$ matrix given by

$$H^k_{\ell\ell'} = \begin{cases} P_{k\ell}(1 - P_{k\ell}) & \text{for } \ell = \ell', \\ -P_{k\ell}P_{k\ell'} & \text{for } \ell \neq \ell', \end{cases} \tag{3.7}$$

with $\ell, \ell' = 1, \ldots, K$. One can check that H^k is the covariance matrix of the routing vector $\phi^k(1)$. Thus, it is symmetric and nonnegative definite. Set

$$\Sigma = C\left(\mathrm{diag}(\lambda_1 b_1, \ldots, \lambda_K b_K) + MQ\left(\mathrm{diag}(\alpha_1^3 a_1, \ldots, \alpha_K^3 a_K) + \sum_{k=1}^{K} \lambda_k H^k\right) Q'M\right) C'. \tag{3.8}$$

Since $\sum_{k=1}^K \lambda_k H^k$ and the two diagonal matrices in (3.8) are each symmetric and nonnegative definite, Σ is symmetric and nonnegative definite. The role of the diagonal matrices $\mathrm{diag}(\alpha_1^3 a_1, \ldots, \alpha_K^3 a_K)$ and $\mathrm{diag}(\lambda_1 b_1, \ldots, \lambda_K b_K)$ is to quantify the randomness of the interarrival and service times. Similarly, the matrix $\sum_{k=1}^K \lambda_k H^k$ quantifies the randomness of the routing vectors. Thus, the matrix Σ can be thought of as measuring the randomness in the queueing network due to the above three quantities. We will always assume that Σ is positive definite. Since Σ is always nonnegative definite, this is equivalent to the determinant of Σ being positive. This condition is needed for the uniqueness of the SRBM discussed after Definition 3.1.

To properly talk about the convergence of the stochastic processes under discussion, we employ the path spaces $\mathbb{D}^d[0,\infty)$, with $d \in \mathbb{Z}_+$. Each path $x \in \mathbb{D}^d[0,\infty)$ is a function $x : [0,\infty) \to \mathbb{R}^d$ that is right continuous in $[0,\infty)$, and has left limits on $(0,\infty)$. We endow the path space with the usual Skorohod J_1-topology (see, e.g., Ethier and Kurtz [25]). We note that when a limit point is a continuous path, convergence in the Skorohod topology is equivalent to uniform convergence on compact intervals. For a sequence of stochastic processes $\{\xi^r,\ r > 0\}$ taking values in $\mathbb{D}^d[0,\infty)$ for some $d \in \mathbb{Z}_+$, we use $\xi^r \Rightarrow \xi^*$ to denote the convergence of ξ^r to ξ^* in distribution.

In the following definition, we assume that (3.4) and (3.5) hold, and employ the following notation. Let Δ denote a $K \times J$ nonnegative matrix. Also, set

$$G = CMQP'\Delta, \tag{3.9}$$

$$R = (I + G)^{-1}, \tag{3.10}$$

$$\theta = R\gamma, \tag{3.11}$$

$$\Gamma = R\Sigma R'. \tag{3.12}$$

In defining R in (3.10), one implicitly assumes that $I + G$ is invertible.

Definition 3.2. Let Δ denote a $K \times J$ nonnegative matrix. For a sequence of networks indexed by r that satisfy (3.4) and (3.5), set $\tilde{X}^r = \tilde{W}^r - R\tilde{Y}^r$. Assume that

$$(\tilde{W}^r, \tilde{X}^r, \tilde{Y}^r, \tilde{Z}^r) \Longrightarrow (W^*, X^*, Y^*, Z^*) \quad \text{as } r \to \infty, \tag{3.13}$$

for some W^*, X^*, Y^* and Z^*, where $W^* = X^* + RY^*$ is an $(\mathbb{R}^J_+, \theta, \Gamma, R, \nu)$-SRBM. Also, assume that

$$Z^* = \Delta W^*. \tag{3.14}$$

Then, (3.13) is said to be a *heavy traffic limit* with *lifting matrix* Δ.

The condition (3.14) is an example of state space collapse. It says that the K-dimensional process Z^*, corresponding to the classes of the networks, is deterministically given by the J-dimensional process W^*, corresponding to the stations. Note that by the discussion following Definition 3.1, for a heavy traffic limit theorem to hold, the reflection matrix R needs to be completely-$\mathcal{S}$.

In order for (3.13) to hold, the initial data must satisfy

$$\tilde{W}^r(0) \Longrightarrow W^*(0) \quad \text{as } r \to \infty \tag{3.15}$$

for some nonnegative random vector $W^*(0)$. State space collapse in (3.14) implies that $Z^*(0) = \Delta W^*(0)$, and hence

$$|\tilde{Z}^r(0) - \Delta \tilde{W}^r(0)| \to 0 \quad \text{in probability as } r \to \infty\,. \tag{3.16}$$

In general (3.15) is not needed for (3.16), since it is possible for $\tilde{Z}^r(0)$ and $\tilde{W}^r(0)$ to be in a fixed proportion, but for both to diverge as $r \to \infty$.

We note that the matrix Δ will typically depend on the discipline and other features of the networks in the above sequence. When the service discipline is FIFO, we will set

$$\Delta w = (\lambda_1 w_{s(1)}, \ldots, \lambda_K w_{s(K)}), \quad \text{for } w \in \mathbb{R}^J.$$

One then needs the following stronger condition on the initial data in order to show (3.13):

$$r^{-1} \sum_{0 \le s \le W^r_j(0)} |D^r_k(s) - \lambda_k s| \to 0 \quad \text{in probability as } r \to \infty, \tag{3.17}$$

for $k \in \mathcal{C}(j)$, $j = 1, \ldots, J$. One can check that (3.16) follows from (3.17). Note that when

$$\tilde{W}^r(0) \Longrightarrow 0 \quad \text{as } r \to \infty, \tag{3.18}$$

(3.16) and (3.17) are both automatically satisfied. In order to understand the following theorems, the reader may find it useful to substitute (3.18) for (3.16) and (3.17).

We now state the main results of the paper, which consist of the following six heavy traffic limit theorems. The first three theorems are for multiclass single station systems. The last three theorems are for re-entrant lines.

Theorem 3.1 (FIFO single station). *Assume that the service discipline is FIFO, and that* $J = 1$. *Assume* (3.4)*, and let the* $K \times 1$ *matrix* $\Delta = (\Delta_1, \ldots, \Delta_K)'$ *be given by* $\Delta_k = \lambda_k$. *Assume further that* (3.5)*,* (3.15) *and* (3.17) *all hold. Then, the heavy traffic limit* (3.13) *holds with lifting matrix* Δ.

A heavy traffic limit for FIFO single station systems was first proved by Reiman [45], for $W^*(0) = 0$, under the additional assumption that jobs can make at most a pre-specified number of visits to each class before leaving the station. This assumption does not allow the feedback to

be Markovian. Dai and Kurtz [17] provided a simpler proof in the more general Markov setting. They, too, considered zero initial data.

It is known that the network version of Theorem 3.1, with more than one station, does not hold; see Dai and Wang [21], Whitt [50] and Dai and Nguyen [18]. A sequence of networks satisfying (3.4) is said to be asymptotically of Kelly type if $m_k = m_\ell$ whenever $s(k) = s(\ell)$. Bramson [7] and Williams [52] showed that the heavy traffic limit theorem holds for FIFO networks which are asymptotically of Kelly type.

Theorem 3.2 (GHLPPS single station). *Assume that the service discipline is GHLPPS with weight vector $\beta = (\beta_1, \ldots, \beta_K)$, and that $J = 1$. Assume* (3.4)*, and let the $K \times 1$ matrix $\Delta = (\Delta_1, \ldots, \Delta_K)'$ be given by*

$$\Delta_k = \frac{(\lambda_k m_k/\beta_k)}{\sum_{\ell=1}^K (\lambda_\ell m_\ell^2/\beta_\ell)} .$$

Assume further that (3.5)*,* (3.15) *and* (3.16) *all hold. Then, the heavy traffic limit* (3.13) *holds with lifting matrix Δ.*

When the weight vector $\beta = (1, \ldots, 1)$, the GHLPPS discipline reduces to the HLPPS discipline. Bramson [7] and Williams [52] showed that the heavy traffic limit holds for HLPPS networks. For a general weight vector β, it is not difficult to show that the network version of Theorem 3.2 does not hold.

Theorem 3.3 (Static buffer priority single station). *Assume that the service discipline is SBP, and that $J = 1$. Assume* (3.4)*, and let the $K \times 1$ matrix $\Delta = (\Delta_1, \ldots, \Delta_K)'$ be given by $\Delta_k = 1/m_k$ if k is the lowest priority class at the station and 0 otherwise. Assume further that* (3.5)*,* (3.15) *and* (3.16) *all hold. Then, the heavy traffic limit* (3.13) *holds with lifting matrix Δ.*

Whitt [48] showed the limit theorem when the station has no feedback, i.e., every job visits the station exactly once before leaving the system.

Our next two theorems are for FBFS and LBFS re-entrant lines.

Theorem 3.4 (FBFS re-entrant line). *Consider a re-entrant line with the FBFS service discipline. Assume* (3.4)*, and let the $K \times J$ matrix Δ be given by $\Delta_{kj} = 1/m_k$ if k is the lowest priority class at station j and 0 otherwise. Assume further that* (3.5)*,* (3.15) *and* (3.16) *all hold. Then, the heavy traffic limit* (3.13) *holds with lifting matrix Δ.*

Theorem 3.5 (LBFS re-entrant line). *Consider a re-entrant line with the LBFS service discipline. Assume* (3.4)*, and let the $K \times J$ matrix Δ be given by $\Delta_{kj} = 1/m_k$ if k is the lowest priority class at station j and 0 otherwise. Assume further that* (3.5)*,* (3.15) *and* (3.16) *all hold. Then, the heavy traffic limit* (3.13) *holds with lifting matrix Δ.*

When the service discipline is FBFS, Chen and Zhang [11] proved the heavy traffic limit theorem under (3.18). Therefore, Theorem 3.4 is not new (except for the more general initial data). Nevertheless, it is a good illustration of the framework developed in Bramson [7] and Williams [52] for proving heavy traffic limit theorems; it is, in particular, much shorter than a proof "from scratch."

Finally, consider the 2-station, 5-class re-entrant line pictured in Figure 1. We assume that the service discipline there is the SBP discipline

$$\{(5, 3, 1), (2, 4)\}, \tag{3.19}$$

that gives the highest priority to class 5, the next priority to class 3 and the lowest priority to class 1 at station 1; and the highest priority to class 2, and the lowest priority to class 4 at station 2.

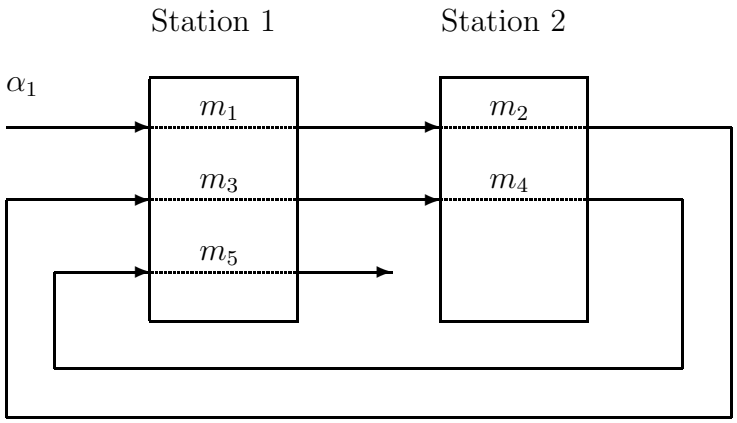

Figure 1: A 2-station, 5-class priority network

Theorem 3.6 (A 2-station, 5-class priority network). *Consider the* 2*-station,* 5*-class priority network in* Figure 1*, with priority ranking given in* (3.19)*. Assume* (3.4)*, and let the* 5×2 *matrix* Δ *be given by*

$$\Delta = \begin{pmatrix} 1/m_1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1/m_4 \\ 0 & 0 \end{pmatrix} . \tag{3.20}$$

Assume further that (3.5)*,* (3.15) *and* (3.16) *all hold, and that*

$$\alpha_1(m_2 + m_5) < 1 . \tag{3.21}$$

Then, the heavy traffic limit (3.13) *holds with lifting matrix* Δ*.*

The above network has certain interesting properties which will be discussed in Section 8. In addition to the proof of Theorem 3.6, a partial converse, Theorem 8.1, will be given there.

A number of assumptions in the preceding theorems can be relaxed. The i.i.d. assumptions on the primitive increments u, v, and ϕ in (2.1), that were used in Theorems 3.1- 3.6, can be replaced, in Theorems 3.1-3.4 and 3.6, by the assumption that the functional central limit theorem holds for each of the primitive processes E, V, and Φ. The i.i.d. assumption employed in our theorems allows us to quote results from Williams [52]. In proving heavy traffic convergence, Williams employed this assumption to show that each limit process is an SRBM having the martingale property given in (iv) of Definition 3.1, and hence that the process is unique in distribution. When the uniqueness is guaranteed through other means, e.g., the reflection matrix R in Definition 3.1 is of the type given in Harrison and Reiman [29], the i.i.d. assumption can be relaxed to a functional central limit theorem assumption for E, V and Φ. When $J = 1$, R is of the type given in Harrison and Reiman [29]. For a FBFS re-entrant line, R is of upper triangular form, and so is also of Harrison and Reiman type. For a LBFS re-entrant line, R is not of Harrison and Reiman type, and the i.i.d. assumption is needed for the proof in this case. (For the last observation, see the appendix of Dai, Yeh and Zhou [23].)

Bramson [6] and Williams [53] considered more general initial data than that assumed in Section 3.3. They allowed the service times and routing vectors for the initial jobs to have distributions

that are different from those for the jobs arriving at the network after time 0. They also allowed the residual external interarrival time for the first job arriving at each class k after time 0 and the residual service time for the first job in class k to depend on each other, and on the other parts of the initial data. To keep the exposition simple, we employ our more restrictive assumptions.

For our sequence of queueing networks in (3.13), we employed the same primitive increments u, v and ϕ, for each r, to construct the interarrival times, service times and routing vectors. In a more general setting, these three variables are given by *triangular* arrays of random variables, where the underlying u, v and ϕ may vary. Heavy traffic limit theorems under this more general setup show that the approximations given by (3.13) are robust under perturbations of the interarrival, service and routing vectors. The purpose of the present setup is to keep the notation simple. Since Bramson [6] and Williams [53] use the framework of triangular arrays, all of the theorems in this paper can be generalized straightforwardly to that setting.

4 Queueing network and fluid model equations

In this section, we write down systems of equations for the queueing networks of interest to us. We also introduce fluid models, which are the continuous, deterministic analogs of queueing networks; their fluid model equations are the analogs of the queueing network equations.

4.1 Queueing network equations

We consider a sequence of queueing networks indexed by r, with performance processes Z^r, D^r, W^r, Y^r defined as in Section 3.1. To describe the dynamics of the queueing network, we introduce two additional K-dimensional processes, $A^r = \{A^r(t),\ t \geq 0\}$ and $T^r = \{T^r(t),\ t \geq 0\}$, where $A^r_k(t)$ denotes the total number of arrivals, over $[0,t]$, at class k (including both external and internal arrivals), and $T^r_k(t)$ denotes the amount of time that server $s(k)$ has spent serving class k jobs over $[0,t]$. One can check that A^r, D^r, T^r, W^r, Y^r, and Z^r satisfy the *queueing network equations*

$$A^r(t) = E^r(t) + \sum_k \Phi^k\left(D^r_k(t)\right), \tag{4.1}$$

$$Z^r(t) = Z^r(0) + A^r(t) - D^r(t), \tag{4.2}$$

$$W^r(t) = CV^r(A^r(t) + Z^r(0)) - CT^r(t), \tag{4.3}$$

$$CT^r(t) + Y^r(t) = et, \tag{4.4}$$

$$Y^r_j(t) \text{ can increase only at times } t \text{ where } W^r_j(t) = 0, \quad j = 1, \ldots, J, \tag{4.5}$$

for all $t \geq 0$. Here, C is the constituency matrix defined in (3.3), e denotes the J-vector of all 1's, $E^r_k(t) = E_k(\alpha^r_k t)$ and $V^r_k(n) = m^r_k V_k(n)$. We note that T^r and Y^r are continuous in t, and that A^r, D^r, W^r, and Z^r are right continuous with left limits. All of the variables are nonnegative in each component, with A^r, D^r, T^r and Y^r being nondecreasing. By assumption, one has

$$A^r(0) = D^r(0) = T^r(0) = 0 \quad \text{and} \quad Y^r(0) = 0. \tag{4.6}$$

In (4.5), we mean that $Y^r_j(t_2) > Y^r_j(t_1)$ implies $W^r_j(t) = 0$ for some $t \in [t_1, t_2]$, which reflects the non-idling property. Since Y^r is continuous, this can also be written as

$$\int_0^\infty W^r_j(t) dY^r_j(t) = 0, \quad j = 1, \ldots, J. \tag{4.7}$$

All queueing networks that we will be working with are HL networks, which were introduced in Section 2. For such networks, one has

$$V^r(D(t)) \leq T^r(t) \leq V^r(D^r(t) + e) \tag{4.8}$$

in addition to (4.1)-(4.5), where the inequalities are componentwise and e denotes the K-vector of all 1's.

From our perspective, the 6-tuple

$$\mathbb{X}^r(t) = (A^r(t), D^r(t), T^r(t), W^r(t), Y^r(t), Z^r(t)), \quad t \geq 0, \tag{4.9}$$

will contain all of the essential information on the evolution of the system. As in Bramson [7], we refer to $\mathbb{X}^r$ as the *queueing network process* for the queueing network, or, in the HL setting, as the *HL queueing network process*. The above equations do not specify the discipline of the queueing network. Below, we give the appropriate equations for the FIFO, GHLPPS and SBP disciplines.

FIFO queueing networks

We recall that for FIFO queueing networks, jobs are served in the order of their arrival at each station. This property can be written as

$$D_k^r(t + W_j^r(t)) = Z_k^r(0) + A_k^r(t), \quad k = 1, \ldots, K, \tag{4.10}$$

for all $t \geq 0$. Together, (4.1)-(4.5), (4.8) and (4.10) form the *FIFO queueing network equations*; the corresponding 6-tuple $\mathbb{X}^r$ will be referred to as a *FIFO queueing network process*. One can check that these equations, together with the values taken by $(E, V, \Phi, \alpha^r, m^r)$ and

$$\left\{D_k^r(t), \text{ for } t \leq W_j^r(0), \quad k = 1, \ldots, K\right\}, \tag{4.11}$$

determine $\mathbb{X}^r(t)$, for all $t \geq 0$. (One also needs to specify an ordering among classes to take care of possible ties among arrivals of customers at different classes.) Thus, the quantity in (4.11) serves the role of the *initial data* for these equations.

GHLPPS networks

Under a GHLPPS discipline with weight vector $\beta = (\beta_\ell)$, all nonempty classes present at a station are served simultaneously, with the fraction of time spent serving a class, say k, being proportional to β_k times the number of jobs in the class. All service goes into the first job of each class to arrive at the station, with the job departing from the station when the service requirement is attained.

The GHLPPS property can be written as

$$T^r(t) = \int_0^t Z^{r,\beta}(s)ds \tag{4.12}$$

for all $t \geq 0$, where

$$Z_k^{r,\beta}(s) = \begin{cases} \beta_k Z_k^r(s) / \sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell^r(s) & \text{if } \sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell^r(s) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

The term $Z_k^{r,\beta}(s)$ is the proportion of effort devoted by the server $s(k)$ to the class k at time s. Together, (4.1)-(4.5), (4.8) and (4.12) form the *GHLPPS queueing network equations*; the corresponding 6-tuple $\mathbb{X}^r$ will be referred to as a *GHLPPS queueing network process*. The equations, together with the values taken by $(E, V, \Phi, \alpha^r, m^r)$ and $Z^r(0)$, determine $\mathbb{X}^r(t)$ for all $t \geq 0$; $Z^r(0)$ serves the role of the *initial data* for these equations.

SBP networks

Under an SBP discipline, classes at each station are assigned a fixed ranking, with jobs from higher ranking classes being served first. For each class k, we denote by $Z_k^{r,+}(t)$ the total number of jobs present in classes whose priorities are at least as great as k, and by $T_k^{r,+}(t)$ the cumulative time that server $s(k)$ has spent on classes whose priorities are at least as great as k. Since the discipline is assumed to be preemptive resume, the SBP property is given by

$$t - T_k^{r,+}(t) \text{ can increase only at times } t \text{ where } Z_k^{r,+}(t) = 0, \quad k = 1, \ldots, K, \tag{4.14}$$

for all $t \geq 0$. (In this setting, (4.5) is redundant, since it is equivalent to (4.14) when k is the lowest ranked class at its station.) As in (4.7), one can instead write this as

$$\int_0^\infty Z_k^{r,+}(t) d\left(t - T_k^{r,+}(t)\right) = 0, \quad k = 1, \ldots, K. \tag{4.15}$$

Together, (4.1)-(4.5), (4.8) and (4.15) form the *SBP queueing network equations*; the corresponding 6-tuple $\mathbb{X}^r$ will be referred to as an *SBP queueing network process*. These equations, together with the values taken by $(E, V, \Phi, \alpha^r, m^r)$ and $Z^r(0)$, determine $\mathbb{X}^r(t)$ for all $t \geq 0$; $Z^r(0)$ therefore serves the role of the *initial data* for these equations.

4.2 Fluid model equations

The formal deterministic analog of a queueing network process has components which satisfy the equations

$$A(t) = \alpha t + P' D(t), \tag{4.16}$$
$$Z(t) = Z(0) + A(t) - D(t), \tag{4.17}$$
$$W(t) = CM(A(t) + Z(0)) - CT(t), \tag{4.18}$$
$$CT(t) + Y(t) = et, \tag{4.19}$$
$$Y_j(t) \text{ can increase only at times } t \text{ where } W_j(t) = 0, \quad j = 1, \ldots, J, \tag{4.20}$$

for all $t \geq 0$. The analog of (4.8) is given by

$$T(t) = MD(t). \tag{4.21}$$

Here, $\alpha = (\alpha_1, \ldots, \alpha_K)'$ is assumed to have nonnegative components, $M = \text{diag}(m)$, where $m = (m_1, \ldots, , m_K)'$ has positive components, and P is a subprobability transition matrix.

The equations in the displays (4.16)-(4.20) are known as *fluid model equations*; their solutions, written as

$$\mathbb{X}(t) = (A(t), D(t), T(t), W(t), Y(t), Z(t)), \quad t \geq 0,$$

will be referred to as *fluid model solutions*. When (4.21) is included with (4.16)-(4.20), we refer to the corresponding quantities as *HL fluid model equations* and *HL fluid model solutions*. When convenient, we will employ the same vocabulary for the fluid model analogs of queueing network quantities, such as the workload W.

We will be interested in HL fluid model solutions for which $\alpha = \lim_{r\to\infty} \alpha^r$ and $m = \lim_{r\to\infty} m^r$, where α^r, m^r and P are the means of sequences of queueing network processes as in (4.1)-(4.5) and (4.8). One formally obtains (4.16)-(4.21) from (4.1)-(4.5) and (4.8) by scaling both time and the

weight of the individual jobs by r, and applying the law of large numbers to $E^r(\cdot)$, $V^r(\cdot)$ and $\Phi(\cdot)$ in (4.1), (4.3) and (4.8).

We will assume that all of the components of $\mathbb{X}$ are continuous and nonnegative, with A, D, T, and Y being nondecreasing. One can check that

$$A(0) = D(0) = T(0) = 0 \quad \text{and} \quad Y(0) = 0$$

all follow from (4.16)-(4.20), and that

$$W(t) = CMZ(t), \quad \text{for all } t \geq 0, \tag{4.22}$$

follows from (4.17), (4.18) and (4.21). Using (4.16)-(4.21), it is easy to show that each component of $\mathbb{X}$ is Lipschitz continuous. That is, for some $N > 0$ (depending on (α, m, P)),

$$|f(t_2) - f(t_1)| \leq N|t_2 - t_1| \quad \text{for all } t_1, t_2 \geq 0,$$

if f is any of the above functions. (When dealing with vectors, we always employ the max norm, although this is a matter of convenience.) In particular, each component of $\mathbb{X}$ is absolutely continuous, and hence differentiable almost everywhere with respect to Lebesgue measure on $[0, \infty)$. A time $t > 0$ is said to be a *regular point* for the fluid model solution $\mathbb{X}$ if $\mathbb{X}$ is differentiable at this time. Whenever we employ the derivative of a component of $\mathbb{X}$ at a time t, we will implicitly assume that t is a regular point. We use $\dot{f}(t)$ to denote the derivative of a function f at t.

For each service discipline, there are additional equations for $\mathbb{X}$ to satisfy. Such equations will be similar to those specifying the discipline of the corresponding queueing network process. Fluid model solutions need *not* be unique, even though their queueing network counterparts determine the evolution of the corresponding queueing network uniquely. This is, for example, the case for the fluid model that corresponds to the well known Lu-Kumar network in [40]. (Dai and Weiss [22, Section 5] presented a divergent fluid solution with $Z(0) = 0$; another solution is given by $Z(\cdot) \equiv 0$.)

The *FIFO fluid model equations* consist of (4.16)-(4.21), together with

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t), \quad k = 1, \ldots, K, \tag{4.23}$$

for all $t \geq 0$. The *initial data* are given by

$$\{D_k(t), \text{ for } t \leq W_j(0), \quad k = 1, \ldots, K\}. \tag{4.24}$$

These last two conditions are the analogs of (4.10) and (4.11). By (4.19)-(4.21),

$$\sum_{k \in \mathcal{C}(j)} m_k D_k(t) = 1 \quad \text{for } t \leq W_j(0),$$

which serves as a consistency condition on the initial data.

The *GHLPPS fluid model equations* consist of (4.16)-(4.21), together with

$$\dot{T}_k(t) = Z_k^{\beta}(t) \quad \text{for} \quad \sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(t) > 0, \quad k = 1, \ldots, K, \tag{4.25}$$

where

$$Z_k^{\beta}(t) = \frac{\beta_k Z_k(t)}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(t)} \quad \text{for} \quad \sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(t) > 0.$$

The equality (4.25) states that when a station j is nonempty, the server allocation rates $\dot{T}_k(t)$ exist and are proportional to the weighted fluid level of each class k. (When a station is empty, $\dot{T}_k(t)$ may still be positive, and so (4.12) need not hold for the fluid model.) Here, $Z(0)$ serves the role of the *initial data* for the GHLPPS fluid model equations.

The *SBP fluid model equations* consist of (4.16)-(4.21), together with

$$\dot{T}_k^+(t) = 1 \quad \text{when } Z_k^+(t) > 0, \quad k = 1, \ldots, K, \tag{4.26}$$

for all regular values of t. (In this setting, (4.20) is redundant, since it is equivalent to (4.26) when k is the lowest ranked class at its station.) The corresponding 6-tuples $\mathbb{X}$ are the *SBP fluid model solutions*. Here, $Z(0)$ serves the role of the *initial data* for these equations.

5 Heavy traffic limits and uniform convergence of fluid models

Consider a sequence of queueing networks that satisfies (3.4) and (3.5), and has FIFO, GHLPPS or SBP service discipline. One can then define the corresponding fluid model, with parameters $\alpha = \lim_{r\to\infty} \alpha^r$, $m = \lim_{r\to\infty} m^r$ and P, as in Section 4.2. Each fluid model solution $\mathbb{X} = (A, D, T, W, Y, Z)$ satisfies the fluid equations (4.16)-(4.21), and additional equations that are specific to the service discipline.

In this section, we provide criteria under which heavy traffic limits hold for such sequences of networks, based on the behavior of the corresponding fluid models and their reflection matrices. These results are modifications of results in Bramson [7] and Williams [52]. To state the conditions on the fluid models succinctly, we introduce the following terminology.

Definition 5.1. Let Δ be a $K \times J$ nonnegative matrix. A fluid model is said to be *uniformly convergent with lifting matrix* Δ if there exists a function $h : \mathbb{R}_+ \to \mathbb{R}_+$, with $h(t) \to 0$ as $t \to \infty$, such that for each fluid model solution $\mathbb{X}$ with $|Z(0)| = 1$,

$$|Z(t) - Z(\infty)| \le h(t) \quad \text{for all } t \ge 0, \tag{5.1}$$

for some $Z(\infty) \in \mathbb{R}_+^K$ satisfying

$$Z(\infty) = \Delta w \quad \text{for some } w \in \mathbb{R}_+^J. \tag{5.2}$$

Condition (5.1) requires that all fluid model solutions, with $|Z(0)| = 1$, converge uniformly quickly to limits satisfying (5.2). (Recall that fluid model solutions need not be unique.) The next two lemmas state that for critical FIFO, GHLPPS and SBP networks, two additional properties automatically follow. These results will be used for Theorems 5.1-5.3. They may be skipped by readers not concerned with the proofs of the theorems. As in Section 3.6, the following $K \times J$ lifting matrices Δ are assigned to each of the disciplines: for FIFO,

$$\Delta_{kj} = \begin{cases} \lambda_k & \text{if } j = s(k), \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

for GHLPPS,

$$\Delta_{kj} = \frac{\lambda_k m_k / \beta_k}{\sum_{\ell \in \mathcal{C}(j)} \lambda_\ell m_\ell^2 / \beta_\ell}, \tag{5.4}$$

and for SBP,

$$\Delta_{kj} = \begin{cases} 1/m_k & \text{if } k \text{ is the lowest priority class at station } j, \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

Lemma 5.1. *Assume that a fluid model operates under a FIFO, GHLPPS or SBP discipline, and is uniformly convergent. Then,* (5.1) *and* (5.2) *also hold for each fluid model solution with* $|Z(0)| \leq 1$.

Lemma 5.2. *Assume that a fluid model operates under a GHLPPS or SBP discipline, and is uniformly convergent. For each fluid model solution with* $Z(0) = \Delta w$ *for some* $w \in \mathbb{R}_+^J$, *one has* $Z(t) = Z(0)$ *for* $t \geq 0$. *Assume that a fluid model operates under the FIFO discipline and is uniformly convergent. For each fluid model solution with* $D_k(t) = \lambda_k t$ *for* $0 \leq t \leq W_j(0)$, *one has* $Z(t) = Z(0)$ *for all* $t \geq 0$.

Lemma 5.1 states that (5.1)-(5.2) remain valid under $|Z(0)| \leq 1$, if they are satisfied under $|Z(0)| = 1$. Lemma 5.2 states that, for the GHLPPS and SBP disciplines $Z(t) = Z(0)$ for all t if $Z(0) = \Delta w$; for FIFO, the same conclusion holds if one instead assumes that $D_k(t) = \lambda_k t$ for $0 \leq t \leq W_j(0)$. This shows that bifurcation from these initial states cannot occur. Lemmas 5.1 and 5.2 are proved at the end of the section. The reasoning is elementary in each case.

In this paper, we will use the following heavy traffic limit results for sequences of networks with the FIFO, GHLPPS and SBP disciplines. In each case, the two main conditions are that the fluid model corresponding to sequences of queueing networks, with the limits (3.4), be uniformly convergent, and that the reflection matrix R given in (3.10) exist and be completely-$\mathcal{S}$.

Theorem 5.1 (FIFO networks). *Assume that the service discipline of a sequence of queueing networks is FIFO, and that* Δ *is given by* (5.3). *Assume that* (3.4), (3.5), (3.15) *and* (3.17) *hold. If (i) the corresponding FIFO fluid model is uniformly convergent with lifting matrix* Δ *and (ii) the matrix* R *in* (3.10) *is completely-*$\mathcal{S}$, *then the heavy traffic limit holds with lifting matrix* Δ.

Theorem 5.2 (GHLPPS networks). *Assume that the service discipline of a sequence of queueing networks is GHLPPS with weight vector* $\beta = (\beta_1, \ldots, \beta_K)$, *and that* Δ *is given by* (5.4). *Assume that* (3.4), (3.5), (3.15) *and* (3.16) *hold. If (i) the corresponding GHLPPS fluid model is uniformly convergent with lifting matrix* Δ *and (ii) the matrix* R *in* (3.10) *is completely-*$\mathcal{S}$, *then the heavy traffic limit holds with lifting matrix* Δ.

Theorem 5.3 (SBP networks). *Assume that the service discipline of a sequence of queueing networks is SBP, and that* Δ *is given by* (5.5). *Assume that* (3.4), (3.5), (3.15) *and* (3.16) *all hold. If (i) the corresponding SBP fluid model is uniformly convergent with lifting matrix* Δ *and (ii) the matrix* R *in* (3.10) *is completely-*$\mathcal{S}$, *then the heavy traffic limit holds with lifting matrix* Δ.

The proofs of Theorems 5.1-5.3 are similar, and all follow from the reasoning employed in Bramson [7] and Williams [52] for related networks. We provide a brief summary here.

Since the FIFO, GHLPPS and SBP disciplines are all HL, it suffices to check that the conditions of Theorem 7.1 in Williams [52] are satisfied in each case. Most of the assumptions in Theorem 7.1 are automatically satisfied, because our construction of each sequence of networks is in terms of the same primitive increments u and v, which have finite second moments, and because of the initial conditions given in the first paragraph of Section 3.3 and the assumptions (3.4), (3.5) and (3.15) on α^r, m^r and $\hat{W}^r(0)$. Two further conditions in Theorem 7.1 remain to be verified for each discipline,

namely that (a) the matrix R in (3.10) exists and is completely-$\mathcal{S}$, and that (b) *multiplicative state space collapse* (MSSC) occurs for the sequence $\mathbb{X}^r$. The latter condition means that, for each $t \geq 0$,

$$\frac{\|\hat{Z}^r(\cdot) - \Delta\hat{W}^r(\cdot)\|_t}{\|\hat{W}^r(\cdot)\|_t \vee 1} \to 1 \quad \text{in probability} \tag{5.6}$$

as $r \to \infty$, where $\|\cdot\|_t$ denotes the sup norm on $[0,t]$ and $a \vee b = \max(a,b)$. We explicitly assume (a) in part (ii) of each of Theorems 5.1-5.3. So, in order to demonstrate the heavy traffic limits in each of these theorems, it remains to demonstrate MSSC in each case. The condition, (3.16) is needed for the GHLPPS and SBP disciplines, and (3.17) is needed for FIFO.

MSSC for SBP networks, in Theorem 5.3, follows immediately from Theorem 4 of Bramson [7], where the results are phrased slightly differently. There, uniform convergence of the fluid model and the resulting properties in Lemmas 5.1 and 5.2 are all assumed. The current approach is more efficient, since the properties given in the lemmas follow automatically because of the discipline.

In order to demonstrate MSSC for the FIFO and GHLPPS networks given in Theorems 5.1 and 5.2, one needs to modify slightly the proofs of Theorems 1 and 1′ of Bramson [7]. In Theorem 1, the FIFO networks are assumed to satisfy the additional condition that $m_k = m_\ell$ for $s(k) = s(\ell)$ (that is, the sequence $\mathbb{X}^r$ is asymptotically of Kelly type), in place of uniform convergence. For these networks, the conclusion of Proposition 6.3, in Bramson [7], contains a stronger version of uniform convergence and the conclusions of Lemmas 5.1 and 5.2. As mentioned on page 134 of Bramson [7], Proposition 6.3 is the only place in the proof of Theorem 1 where this condition is used. Assuming uniform convergence instead, one can show MSSC by closely following the same steps. The only other difference, in the two arguments, is that *multiplicative strong state space collapse*, in addition to MSSC, is demonstrated in Theorem 1. In order to demonstrate this stronger variant of MSSC, one employs a suitable norm on $\mathbb{X}^r$ rather than working directly with Z^r. As a result, the proof simplifies slightly, when instead demonstrating MSSC for Theorem 5.1.

MSSC for HLPPS networks is demonstrated in Theorem 1′. The analog of Proposition 6.3, Proposition 7.2, holds for such networks, and is applied in the same manner. Upon replacing the proposition with the assumption of uniform convergence, the same proof shows that MSSC holds for the sequences of GHLPPS networks given in Theorem 5.2. Since the proof of Theorem 1′ already deals with Z^r rather than $\mathbb{X}^r$, the changes required in the preceding paragraph, for replacing $\mathbb{X}^r$ by Z^r for FIFO networks, are not needed here. The summary provided at the beginning of Section 7 of Bramson [7], for adapting the proof of Theorem 1 to that of Theorem 1′, can be used as a guide for demonstrating MSSC for GHLPPS networks.

Checking the uniform convergence of a fluid model may involve entropy arguments (for FIFO networks of Kelly type and HLPPS networks), comparisons with Markov chains (for single station FIFO and GHLPPS networks) and piecewise linear Lyapunov functions (for SBP networks). The completely-$\mathcal{S}$ property can always be checked, at least numerically, because of the linear algebra involved. For single station networks, the completely-$\mathcal{S}$ property becomes trivial, since R reduces to a positive scalar.

We now return to the proofs of Lemmas 5.1 and 5.2. Instead of demonstrating Lemma 5.1, it is natural to work in a more general setting. For a given fluid model solution $\mathbb{X}(\cdot)$ and $c > 0$, we set

$$\mathbb{X}^c(t) = (A(t+c) - A(c), D(t+c) - D(c), T(t+c) - T(c), W(t+c), Y(t+c) - Y(c), Z(t+c))\,,$$

for $t \geq 0$; $\mathbb{X}^c(\cdot)$ corresponds to restarting $\mathbb{X}(\cdot)$ at time c.

Definition 5.2. A fluid model is said to be *shift invariant* if for each fluid model solution $\mathbb{X}(\cdot)$, $\mathbb{X}^c(\cdot)$ is also a fluid model solution for each $c > 0$. A fluid model is said to be *scale invariant* if for each fluid model solution $\mathbb{X}(\cdot)$, $c^{-1}\mathbb{X}(c\cdot)$ is also a fluid model solution for each $c > 0$.

Plugging into the fluid model equations (4.16)-(4.21), (4.23), (4.25) and (4.26), it is not difficult to check that FIFO, GHLPPS and SBP fluid models are all shift and scale invariant. So, Lemma 5.1 is included in the following result.

Lemma 5.3. *Assume that a fluid model is shift and scale invariant, and assume that there exists a function $h : \mathbb{R}_+ \to \mathbb{R}_+$, with $h(t) \to 0$ as $t \to \infty$, such that* (5.1) *and* (5.2) *hold for each fluid model solution $\mathbb{X}(\cdot)$ with $|Z(0)| = 1$. Then* (5.1) *and* (5.2) *also hold for each fluid model solution $\mathbb{X}(\cdot)$ with $|Z(0)| \leq 1$.*

Proof. We may assume without loss of generality that h is bounded and nonincreasing. For example, letting $\tilde{h}$ denote the original choice of h, one may set $h(t) = \sup_{t' \geq t} \tilde{h}(t') \wedge M$ for large enough M, since $Z(\cdot)$ is Lipschitz, where $a \wedge b = \min(a, b)$.

We first show that, for any fluid model solution $\mathbb{X}(\cdot)$,

$$Z(0) = 0 \text{ implies } Z(t) = 0 \quad \text{for all } t \geq 0. \tag{5.7}$$

Assume, on the contrary, that $|Z(t_1)| = c > 0$ for some $t_1 > 0$. Then, by the continuity of $Z(\cdot)$, for any $M > 1$, there exists $t_0 \in (0, t_1)$, so that $|Z(t_0)| = c/M$. Set $\tilde{\mathbb{X}}(t) = \frac{M}{c} \mathbb{X}^{t_0}\left(\frac{c}{M} t\right)$, $t \geq 0$. By the shift and scale invariance of the fluid model, $\tilde{\mathbb{X}}(\cdot)$ is also a solution; it clearly satisfies $|\tilde{Z}(0)| = 1$. But, $\left|\tilde{Z}\left(\frac{M}{c}(t_1 - t_0)\right)\right| = \frac{M}{c}|Z(t_1)| = M$; since M is arbitrary, this contradicts (5.1). So, (5.7) holds.

Because of (5.7), it suffices to consider $|Z(0)| \in (0, 1]$, in order to demonstrate the lemma. Setting $c = |Z(0)|$, it follows that $\tilde{X}(\cdot) = c^{-1}\mathbb{X}(c\cdot)$ is a fluid model solution with $|\tilde{Z}(0)| = 1$. So, by assumption, $|\tilde{Z}(t) - \tilde{Z}(\infty)| \leq h(t)$ for $t \geq 0$, where $\tilde{Z}(\infty) = \Delta\tilde{w}$ for some $\tilde{w} \in \mathbb{R}_+^J$. Equivalently, $|Z(ct) - Z(\infty)| \leq ch(t)$ for $t \geq 0$, where $Z(\infty) = \Delta(c\tilde{w})$. Therefore,

$$|Z(t) - Z(\infty)| \leq ch(c^{-1}t) \leq h(t),$$

as desired. □

The demonstration of Lemma 5.2 also employs the shift invariance of the FIFO, GHLPPS and SBP disciplines.

Proof of Lemma 5.2. It suffices to show in all three cases, that for each $c > 0$, there exists a fluid model solution $\tilde{\mathbb{X}}(\cdot)$, with $|\tilde{Z}(0)| \leq 1$ and $\mathbb{X}(\cdot) = \tilde{\mathbb{X}}^c(\cdot)$. Then, by Lemma 5.1,

$$|Z(t) - \tilde{Z}(\infty)| = |\tilde{Z}(t + c) - \tilde{Z}(\infty)| \leq h(c)$$

for appropriate $\tilde{Z}(\infty)$ and any $t \geq 0$, where $h(c) \to 0$ as $c \to \infty$, and $h(\cdot)$ does not depend on $\tilde{Z}(\cdot)$. Consequently, $Z(t)$ is constant.

We construct $\tilde{\mathbb{X}}(\cdot)$ by shifting $\mathbb{X}(\cdot)$ by $-c$, and defining $\tilde{\mathbb{X}}(\cdot)$ over $[0, c]$ to be $\hat{\mathbb{X}}(\cdot)$, the invariant fluid model solution with $\hat{Z}(t) = Z(0)$ and $\hat{D}(t) = \lambda t$ for all t. Over $t \geq c$, this means that

$$\begin{aligned}
\tilde{A}(t) &= \hat{A}(c) + A(t - c), \quad & \tilde{D}(t) &= \hat{D}(c) + D(t - c),\\
\tilde{T}(t) &= \hat{T}(c) + T(t - c), \quad & \tilde{Y}(t) &= \hat{Y}(c) + Y(t - c),\\
\tilde{W}(t) &= W(t - c), & \tilde{Z}(t) &= Z(t - c).
\end{aligned}$$

It is not difficult to see that $\tilde{\mathbb{X}}(\cdot)$ always satisfies the fluid model equations (4.16)-(4.21), and, when the discipline is GHLPPS or SBP, either (4.25) or (4.26) holds. In particular, $\tilde{\mathbb{X}}(\cdot)$ is a fluid model solution for the GHLPPS and SBP disciplines. Using the assumption that $D_k(t) = \lambda_k t$ for $0 \leq t \leq W_j(0)$, one can check that $\tilde{\mathbb{X}}(\cdot)$ also satisfies (4.23) when $\mathbb{X}(\cdot)$ and $\hat{\mathbb{X}}(\cdot)$ are FIFO. (The behavior of $\tilde{\mathbb{X}}(\cdot)$ on the time interval $[c, c + W_j(c)]$ requires a little work.) So, $\tilde{X}(\cdot)$ is, in this case, a fluid model solution for the FIFO discipline. □

6 Proofs of Theorems 3.1, 3.2 and 3.3

In this section, we prove Theorems 3.1-3.3, which are the heavy traffic limits for single station systems operating under the FIFO, GHLPPS and SBP disciplines. Since $J = 1$, the matrix R defined in (3.10) reduces to a scalar. One can check that it is always positive, and hence that R is completely-$\mathcal{S}$. Therefore, by Theorems 5.1-5.3, to prove Theorems 3.1-3.3, it is enough to show that each fluid model is uniformly convergent with the corresponding lifting matrix Δ, which is specified, in (5.3)-(5.5), for each discipline. In our proofs, we drop the station index j, since there is only one station in the system.

Before proceeding with the proofs of the theorems, we point out, in Lemma 6.1, that for all critically loaded one-station fluid models, the *total workload* $CMQZ(t)$ is invariant. This observation will be needed in the proofs of Theorems 3.1 and 3.3.

Lemma 6.1. *For a one-station, critically loaded fluid model,*

$$CMQ(Z(t) - Z(0)) = 0 \tag{6.1}$$

holds for all solutions and all t.

Proof. By (4.16), (4.17) and the definition of Q,

$$CMQ(Z(t) - Z(0)) = tCMQ\alpha - CMD(t).$$

On account of (4.21) and the equalities, $\lambda = Q\alpha$ and $CM\lambda = 1$, this equals $t - CT(t)$. By (4.19), $t \geq CT(t)$ always holds, and so $CMQZ(t)$ is nondecreasing.

Suppose now that $CMQZ(t_1) > 0$ for some t_1. Then, by the preceding paragraph, $CMQZ(t) > 0$ for all $t \geq t_1$. Since $Z(t) \neq 0$ implies that $W(t) > 0$, it follows from (4.19) and (4.20), that $t - CT(t)$ remains constant on $[t_1, \infty)$. So, $CMQZ(t)$ is constant on $[t_1, \infty)$. Since $CMQZ(t)$ is continuous in t and is always nonnegative, (6.1) follows from this. □

The following corollary of Lemma 6.1 will be used in the proofs of Theorems 3.1 and 3.2. It follows immediately from (6.1) and the inequalities

$$|Z(t_1)| \min_k\{m_k\} \leq CMQZ(t_1) = CMQZ(t_2) \leq K|(I - P')^{-1}|\,|Z(t_2)| \max_k\{m_k\}$$

for $t_1, t_2 \geq 0$, where $|\cdot|$ denotes the max norm for both vectors and matrices.

Corollary 6.1. *For a one-station, critically loaded fluid model,*

$$|Z(0)|/a \leq |Z(t)| \leq a|Z(0)|, \quad t \geq 0 \tag{6.2}$$

for appropriate $a > 0$, depending only on M and P.

6.1 Proof of Theorem 3.1

In order to show that a critical one-station FIFO fluid model is uniformly convergent, we need to show that for $|Z(0)| = 1$, $Z(t)$ converges uniformly to a scalar multiple of λ. The reasoning consists of four main steps. Let $\tau(s) = s + W(s)$, $\bar{Z}(t) = \Lambda^{-1}Z(t)$ and $B = \Lambda^{-1}(\alpha CM + P')\Lambda$, where $\Lambda = \mathrm{diag}(\lambda)$. We first show, in Step 1, that

$$\bar{Z}(\tau(s)) = B\bar{Z}(s) \quad \text{for } s \geq 0. \tag{6.3}$$

In Step 2, we show that B is the transition matrix of an irreducible aperiodic K-state Markov chain. Letting $\tau^n(s)$ denote the n-fold iterate of $\tau(s)$, it will then follow, as in Step 3, that $Z(\tau^n(s))$ converges to a multiple of λ as $n \to \infty$, where the convergence is uniform in s. In Step 4, we conclude from this, that $Z(t)$ converges to a multiple of λ as $t \to \infty$. This is the desired result.

Step 1. (6.3) holds.

Proof. Combining (4.16) and (4.17), we have

$$Z(s) = Z(0) + \alpha s - (I - P')D(s), \quad s \geq 0.$$

Thus, for $s \geq 0$,

$$\begin{aligned} Z(s + W(s)) &= Z(0) + \alpha(s + W(s)) - (I - P')D(s + W(s)) && (6.4)\\ &= Z(0) + \alpha(s + W(s)) - (I - P')(Z(0) + A(s)) && (6.5)\\ &= \alpha W(s) + P'(Z(0) + A(s)) + \alpha s - A(s) \\ &= \alpha W(s) + P'(Z(0) + A(s) - D(s)) && (6.6)\\ &= (\alpha CM + P')Z(s), && (6.7) \end{aligned}$$

where we have used the FIFO equation (4.23) in getting (6.5), (4.16) in getting (6.6), and (4.17) and (4.22) in getting (6.7). Substitution of $\bar{Z} = \Lambda^{-1}Z$ and $B = \Lambda^{-1}(\alpha CM + P')\Lambda$ above implies (6.3). □

Step 2. B is the transition matrix of an irreducible aperiodic K-state Markov chain.

Proof. Observe that each entry of B is nonnegative, and that the kth row of B is given by

$$\alpha_k \lambda_k^{-1}(\lambda_1 m_1, \ldots, \lambda_K m_K) + \lambda_k^{-1}(\lambda_1 P_{1k}, \ldots, \lambda_K P_{Kk}).$$

Therefore, the sum of the entries of the kth row of B is

$$\alpha_k \lambda_k^{-1} \sum_{\ell=1}^{K} \lambda_\ell m_\ell + \lambda_k^{-1} \sum_{\ell=1}^{K} \lambda_\ell P_{\ell k} = \alpha_k \lambda_k^{-1} + \lambda_k^{-1}(\lambda_k - \alpha_k) = 1,$$

where we have used equations (3.1) and (3.6). It follows that B is a stochastic matrix. We claim that B is irreducible. Note that for any $n \geq 1$, $B^n = \Lambda^{-1}(\alpha CM + P')^n \Lambda$. Fix k. Since $\lambda_k > 0$, it follows from $\lambda = Q\alpha$ that the kth component of $(P')^{n-1}\alpha$, which we denote by c_k, is positive for some $n \geq 1$. Also,

$$(\alpha CM + P')^n \geq (P')^{n-1} \alpha CM,$$

with the kth row of $(P')^{n-1}\alpha CM$ being given by $c_k(m_1, \ldots, m_K)$, each component of which is strictly positive. It follows that $B^n_{k\ell} > 0$ for appropriate n and all ℓ. Consequently, B is irreducible.

We still need to show that B is aperiodic. This is a simple consequence of the observation that $\alpha_k > 0$ for some k, and that the kth diagonal entry of αCM is given by $\alpha_k m_k$, which is therefore positive for this k. □

By Step 2 and discrete time Markov chain theory,

$$B^n \to \Pi \quad \text{as } n \to \infty \tag{6.8}$$

for some matrix Π, where all rows of Π are identical, and the entries $\pi_1, \ldots, \pi_K$ are positive and sum to 1. Set $\pi(z) = \sum_k \pi_k z_k$.

Step 3. $Z(\tau^n(s)) \to \lambda \pi(\bar{Z}(0))$ as $n \to \infty$, where convergence is uniform over all s and initial data $|Z(0)| = 1$.

Proof. By Step 1 and (6.8),

$$Z(\tau^n(s)) = \Lambda B^n \bar{Z}(s) \to \Lambda\Pi\bar{Z}(s) \quad \text{as } n \to \infty. \tag{6.9}$$

To see the uniformity of the convergence in (6.9) over all s and initial data satisfying $|Z(0)| = 1$, note that

$$|\Lambda B^n \bar{Z}(s) - \Lambda\Pi\bar{Z}(s)| \leq K|\Lambda|\,|B^n - \Pi|\,|\bar{Z}(s)|. \tag{6.10}$$

By Corollary 6.1, the right side of (6.10) is at most $aK|\Lambda|\,|\Lambda^{-1}|\,|B^n - \Pi|$, which does not depend on s or on $Z(0)$, since $|Z(0)| = 1$ is assumed.

One can write $\Lambda\Pi\bar{Z}(s)$, on the right side of (6.9), as $\lambda\pi(\bar{Z}(s))$, which is a scalar multiple of λ. By Lemma 6.1, $CMQZ(\tau^n(s))$ is constant, and so the limit, as $n \to \infty$, does not depend on s. It follows that $\pi(\bar{Z}(s)) = \pi(\bar{Z}(0))$ must always hold. This implies the claim. □

Step 4. $Z(t) \to \lambda\pi(\bar{Z}(0))$ as $t \to \infty$, where convergence is uniform over initial data satisfying $|Z(0)| = 1$.

Proof. By the upper bound in (6.2), $|Z(s)|$, and hence $\tau(s) - s = W(s)$, remains bounded for all s. It follows that

$$\tau^{n-1}(s) < cn \tag{6.11}$$

for appropriate c, and all $n \geq 1$ and $s \in [0, \tau(0))$. Also, note that since $W(s)$ is continuous in s, so is $\tau(s)$. Hence, for every $t \in [\tau^{n'}(0), \tau^{n'+1}(0))$, there exists an $s \in [0, \tau(0))$ with $t = \tau^{n'}(s)$. Setting $\tau^\infty(0) = \lim_{n\to\infty} \tau^n(0)$, one has, in particular, that for $t \in [cn, \tau^\infty(0))$, $t = \tau^{n'}(s)$ for some $s \in [0, \tau(0))$ and $n' \geq n$.

The last observation, together with Step 3, implies that

$$\sup_{t\in[cn,\tau^\infty(0))} |Z(t) - \lambda\pi(\bar{Z}(0))| \to 0 \quad \text{as } n \to \infty \tag{6.12}$$

uniformly over initial data satisfying $|Z(0)| = 1$. It follows from the lower bound in (6.2), that $\tau^n(0) \geq n/c$, for appropriate $c > 0$. So $\tau^\infty(0) = \infty$. Together with (6.12), this implies the claim. □

This concludes the proof of Theorem 3.1.

6.2 Proof of Theorem 3.2

The proof of Theorem 3.2 is similar to that of Theorem 3.1, but simpler, since one can use $\dot{Z}(t)$, rather than iterate $Z(s + W(s))$. In order to show that a critical one-station GHLPPS fluid model is uniformly convergent, we need to show that for $|Z(0)| = 1$, $Z(t)$ converges uniformly to a scalar multiple of $(\lambda m_1/\beta_1, \ldots, \lambda_K m_K/\beta_K)$.

By the lower bound in (6.2), $Z(t) \neq 0$ for all t. Applying (4.16) and (4.17), and then (4.21) and (4.25), one obtains

$$\begin{aligned}\dot{Z}_\ell(t) &= \alpha_\ell + \sum_{k=1}^K P_{k\ell}\dot{D}_k(t) - \dot{D}_\ell(t) \\ &= \frac{1}{|Z(t)|_\beta}\left(\alpha_\ell|Z(t)|_\beta + \sum_{k=1}^K P_{k\ell}\beta_k Z_k(t)/m_k - \beta_\ell Z_\ell(t)/m_\ell\right),\end{aligned} \tag{6.13}$$

where $|Z(t)|_\beta = \sum_k \beta_k Z_k(t)$. Let $\bar{Z}_k(t) = \beta_k Z_k(t)/(\lambda_k m_k)$. Substituting into (6.13), one can check that

$$\dot{\bar{Z}}_\ell(t) = \frac{\beta_\ell}{m_\ell |Z(t)|_\beta} \left(\sum_{k=1}^{K} \left(\alpha_\ell \lambda_k m_k \lambda_\ell^{-1} + \lambda_k P_{k\ell} \lambda_\ell^{-1} \right) \bar{Z}_k(t) - \bar{Z}_\ell(t) \right). \tag{6.14}$$

Therefore,

$$\dot{\bar{Z}}(t) = \frac{1}{|Z(t)|_\beta} G \bar{Z}(t), \tag{6.15}$$

where

$$G = \mathrm{diag}(\beta_1 \mu_1, \ldots, \beta_K \mu_K)(B - I) \quad \text{and} \quad B = \Lambda^{-1}(\alpha C M + P')\Lambda\,.$$

Solving (6.15) gives

$$\bar{Z}(t) = \exp\left(G \int_0^t \frac{1}{|Z(s)|_\beta}\, ds \right) \bar{Z}(0). \tag{6.16}$$

(Here, $\exp(H) = I + \cdots + H^n/n! + \ldots$.)

The matrix B was employed in the proof of Theorem 3.1. In Step 2, it was shown that B is the transition matrix of an irreducible Markov chain. Consequently, G is the infinitesimal generator of a continuous time Markov chain. On the other hand, by the lower bound in (6.2), $|Z(t)|_\beta$ is bounded away from 0, uniformly for all fluid model solutions with $|Z(0)| = 1$. Hence, $\int_0^t 1/|Z(s)|_\beta\, ds \to \infty$ uniformly as $t \to \infty$. It follows from continuous time Markov chain theory, that

$$\exp\left(G \int_0^t \frac{1}{|Z(s)|_\beta}\, ds \right) \to \Pi \quad \text{as } t \to \infty, \tag{6.17}$$

where all rows of Π are identical, and convergence is uniform over $|Z(0)| = 1$.

Let π be a row of Π, and $\pi(z) = \sum_k \pi_k z_k$ for $z \in \mathbb{R}^K$. From (6.16) and (6.17), we have

$$\bar{Z}(t) \to e\pi(\bar{Z}(0)) \quad \text{as } t \to \infty,$$

where e the K-vector of all 1's. Therefore,

$$Z(t) \to (\lambda_1 m_1/\beta_1, \ldots, \lambda_K m_K/\beta_K)' \pi(\bar{Z}(0)) \quad \text{as } t \to \infty,$$

with convergence being uniform over all fluid model solutions $\mathbb{X}$ satisfying $|Z(0)| = 1$. This demonstrates Theorem 3.2.

6.3 Proof of Theorem 3.3

In order to show that a critical one-station priority fluid model is uniformly convergent, we need to show that for $|Z(0)| = 1$, $Z(t)$ converges uniformly to a scalar multiple of Δ, with $\Delta_k = 1/m_k$ if k is the lowest priority class at the station and $\Delta_k = 0$ otherwise. The argument differs from those of Theorems 3.1 and 3.2. We will show that fluid levels in non-lowest priority classes reach zero in a finite time. We will then use Lemma 6.1 to conclude that the fluid level in the lowest priority class remains constant after that time.

To show that the fluid levels of non-lowest priority classes reach zero, we construct a Lyapunov function of the fluid levels for such classes. We introduce the following notation. Denote by K

the class with lowest priority and by $\mathcal{H} = \{1, \ldots, K-1\}$ the set of higher priority classes. Set $M_{\mathcal{H}}$ equal to the $(K-1) \times (K-1)$ diagonal matrix, with diagonal entries m_k, $k = 1, \ldots, K-1$. Partition the transition matrix P according to

$$P = \begin{pmatrix} P_{\mathcal{H}\mathcal{H}} & P_{\mathcal{H}K} \\ P_{K\mathcal{H}} & P_{KK} \end{pmatrix},$$

where $P_{\mathcal{H}\mathcal{H}}$ is the $(K-1) \times (K-1)$ submatrix of P, with (k,ℓ)th entry $P_{k\ell}$, $P_{\mathcal{H}K}$ is the $(K-1)$ column vector, with kth component P_{kK}, and $P_{K\mathcal{H}}$ is the $(K-1)$ row vector with kth component P_{Kk}. For a vector y, $(y_{\mathcal{H}}, y_K)'$ denotes the corresponding partition. Also, set $Q_{\mathcal{H}} = (I - P'_{\mathcal{H}\mathcal{H}})^{-1}$ and e the $(K-1)$-vector of all 1's, and define

$$f(t) = eM_{\mathcal{H}}Q_{\mathcal{H}}Z_{\mathcal{H}}(t), \tag{6.18}$$

which is the total workload for the modified network, which is obtained by removing fluid upon arrival at K.

We wish to show that $f(t) = 0$ for $t \geq \delta$ and appropriate $\delta \geq 0$. Using (4.16) and (4.17), one has

$$Z_{\mathcal{H}}(t) = Z_{\mathcal{H}}(0) + \alpha_{\mathcal{H}}t + P'_{K\mathcal{H}}D_K(t) - (I - P'_{\mathcal{H}\mathcal{H}})D_{\mathcal{H}}(t). \tag{6.19}$$

Substitution of (6.19) into (6.18) implies that

$$\dot{f}(t) = eM_{\mathcal{H}}Q_{\mathcal{H}}\alpha_{\mathcal{H}} + eM_{\mathcal{H}}Q_{\mathcal{H}}P'_{K\mathcal{H}}\dot{D}_K(t) - \sum_{k\in\mathcal{H}} \dot{T}_k(t). \tag{6.20}$$

By the traffic equation (3.1), $\lambda_{\mathcal{H}} = \alpha_{\mathcal{H}} + P'_{\mathcal{H},\mathcal{H}}\lambda_{\mathcal{H}} + P'_{K\mathcal{H}}\lambda_K$, and so

$$Q_{\mathcal{H}}\alpha_{\mathcal{H}} = \lambda_H - Q_{\mathcal{H}}P'_{K\mathcal{H}}\lambda_K \leq \lambda_H.$$

Therefore,

$$eM_{\mathcal{H}}Q_H\alpha_{\mathcal{H}} \leq eM_{\mathcal{H}}\lambda_{\mathcal{H}} = \sum_{k\in\mathcal{H}} \lambda_k m_k = 1 - \lambda_K m_K. \tag{6.21}$$

When $f(t) > 0$, then $\sum_{k\in\mathcal{H}} Z_k(t) > 0$, and hence $\dot{D}_K(t) = 0$ and $\sum_{k\in\mathcal{H}} \dot{T}_k(t) = 1$. Substitution of this and the bound in (6.21) into (6.20) implies that $\dot{f}(t) \leq -\lambda_K m_K$ whenever $f(t) > 0$. It follows without difficulty, since $f(\cdot)$ is absolutely continuous, that $f(t) = 0$, and hence $Z_k(t) = 0$, for $k \in \mathcal{H}$ and $t \geq \delta' \equiv f(0)/(\lambda_K m_K)$. By Lemma 6.1, the total workload for the entire network is constant, and so $Z_K(t) = Z_K(\delta')$ for $t \geq \delta'$. Set $Z(\infty) = (0, \ldots, 0, Z_K(\delta'))'$. Clearly, $Z(t) = Z(\infty)$ for $t \geq \delta'$. By (6.18), with $t = 0$, $\delta \equiv \sup_{|Z(0)|=1} \delta'$ is finite. This demonstrates Theorem 3.3.

7 Proofs of Theorems 3.4 and 3.5

In this section, we prove Theorems 3.4 and 3.5, which are the heavy traffic limits for re-entrant lines with FBFS and LBFS static buffer priorities. By Theorems 3.1 and 3.2 of Dai, Yeh and Zhou [23], under either discipline, the matrix R given by (3.10) is completely-$\mathcal{S}$. Using Theorem 5.3, it therefore suffices to prove that the corresponding fluid models, in each case, are uniformly convergent, with the lifting matrices given in Theorems 3.4 and 3.5. The proof of this for FBFS is quite quick; the proof for LBFS is more involved.

By (3.6), $\rho = e$ automatically holds in both settings. We can assume without loss of generality that $\alpha_1 = 1$; it then follows that $\sum_{k \in \mathcal{C}(j)} m_k = 1$ for each j. In the proofs, we enumerate the classes according to the order of their appearance along the route of the re-entrant line. For a fluid model solution $\mathbb{X}$, we will find it convenient to set $d_k(t) = \dot{D}_k(t)$, for the departure rate from a class k. As mentioned in Section 4.2, because of the absolute continuity of $\mathbb{X}$, we need only consider $d_k(t)$ at t which are regular points of $\mathbb{X}$.

Proof of Theorem 3.4. We use induction to prove that, for each $k = 1, \ldots, K$, there exists a $t_k \geq 0$ such that, for any fluid model solution $\mathbb{X}$ with $|Z(0)| = 1$, $Z_\ell(t)$ is constant on $[t_k, \infty)$, for $\ell = 1, \ldots, k$. Furthermore, this value is zero if ℓ is not a lowest priority class, i.e., the last class to be visited at some station. For the induction step, we assume that $Z_\ell(t)$ is constant on $[t_{k-1}, \infty)$, for $\ell = 1, \ldots, k-1$. We break the argument into two cases, depending on whether or not the class k has lowest priority.

We note that for $t \geq t_{k-1}$, $d_{k-1}(t) = d_{k-2}(t) = \cdots = d_1(t) = \alpha_1 = 1$. Set $\mathcal{H}_k = \{\ell \leq k : s(\ell) = s(k)\}$. If $Z_k(t) > 0$, then, by the FBFS property, $\sum_{\ell \in \mathcal{H}_k} m_\ell d_\ell(t) = 1$. So, for $Z_k(t) > 0$ and $t \geq t_{k-1}$,

$$d_k(t) = \mu_k \Bigg(1 - \sum_{\ell \in \mathcal{H}_k \setminus \{k\}} m_\ell \Bigg). \tag{7.1}$$

Case (i). Class k is not a lowest priority class. In this case, $\sum_{\ell \in \mathcal{H}_k} m_\ell < 1$, and so

$$\dot{Z}_k(t) = d_{k-1}(t) - d_k(t) = 1 - \mu_k \Bigg(1 - \sum_{\ell \in \mathcal{H}_k \setminus \{k\}} m_\ell \Bigg) < 0$$

whenever $Z_k(t) > 0$ and $t \geq t_{k-1}$. Setting

$$t'_k = t_{k-1} + \frac{Z_k(t_{k-1})}{\mu_k \left(1 - \sum_{\ell \in \mathcal{H}_k \setminus \{k\}} m_\ell \right) - 1},$$

it follows that $Z_k(t) = 0$ for $t \geq t'_k$. Since $|Z(0)| = 1$ and $\alpha_1 = 1$, one has $Z_k(t) \leq |Z(t)| \leq t + 1$ for all t. So, we can choose $t_k \geq t_{k-1}$ independently of the fluid solution $\mathbb{X}$, with $Z_k(t) = 0$ for $t \geq t_k$.

Case (ii). Class k is a lowest priority class. Since $\sum_{\ell \in \mathcal{H}_k} m_\ell = 1$, it follows from (7.1) that $d_k(t) = 1$, whenever $Z_k(t) > 0$ and $t \geq t_{k-1}$. From this, it follows that $\dot{Z}_k(t) = d_{k-1}(t) - d_k(t) = 0$. When $Z_k(t) = 0$, $\dot{Z}_k(t) = 0$ holds. (Recall that t is a regular point of $\mathbb{X}$.) It follows that $Z_k(t)$ is constant on $[t_{k-1}, \infty)$. So, we simply choose $t_k = t_{k-1}$ in this case. □

Proof of Theorem 3.5. We use induction to prove that, for each $k = 1, \ldots, K$, there exists a $t_k \geq 0$ such that, for any fluid model solution $\mathbb{X}$ with $|Z(0)| = 1$, $Z_\ell(t)$ is constant on $[t_k, \infty)$ for $\ell = k, k+1, \ldots, K$. Furthermore, this value is zero if ℓ is not a lowest priority class, i.e., the first class to be visited at some station. For the induction step, we assume that $Z_\ell(t)$ is constant on $[t_{k+1}, \infty)$, for $\ell = k+1, k+2, \ldots, K$. As before, we break the argument into two cases, depending on whether or not the class k has lowest priority.

We first present some preliminaries. It will be convenient, for bookkeeping purposes, to append an extra one-class station, denoted by $k = 0$, to the beginning of the network, with $m_0 = 1$ and $Z_0(0) = 1$. Then, $Z_0(t) = 1$ will always hold, and the evolution of $\mathbb{X}$ proceeds as before in the remainder of the network. (For the new network, $|Z(0)| = 2$.) Set $\mathcal{G}_k = \{\ell \geq k : s(\ell) = s(k)\}$ and

$\mathcal{G}_{h,k} = \{h \le \ell < k : s(\ell) = s(h)\}$, for $h, k = 0, 1 \ldots, K$ and $h < k$, and set $\bar{m}_k = \sum_{\ell \in \mathcal{G}_k} m_\ell$. Since $Z_\ell(t)$ is assumed to be constant on $[t_{k+1}, \infty)$, $d_k(t) = d_{k+1}(t) = \cdots = d_K(t)$ for $t \ge t_{k+1}$. By the fluid model equations (4.19) and (4.21), $\sum_{\ell \in \mathcal{C}(j)} m_\ell d_\ell(t) \le 1$ always holds for each j. Consequently,

$$d_k(t) \le 1/\bar{m}_k \quad \text{for } t \ge t_{k+1}, \tag{7.2}$$

and by the LBFS discipline,

$$d_k(t) = 1/\bar{m}_k \quad \text{for } t \ge t_{k+1}, \quad \text{whenever } Z_k(t) > 0. \tag{7.3}$$

We will employ the functions $f_k(\cdot)$ and $g_{h,k}(\cdot)$, where

$$f_k(t) = \sum_{\ell \le k} Z_\ell(t), \quad g_{h,k}(t) = \sum_{\ell \in \mathcal{G}_{h,k}} m_\ell \sum_{\ell < \ell' \le k} Z_{\ell'}(t). \tag{7.4}$$

One can check that

$$f_k(t) = f_k(0) + t - D_k(t), \quad g_{h,k}(t) = g_{h,k}(0) + \sum_{\ell \in \mathcal{G}_{h,k}} m_\ell (D_\ell(t) - D_k(t)). \tag{7.5}$$

The meaning of $f_k(\cdot)$ is clear. The functions $g_{h,k}(\cdot)$ are more difficult to motivate; the main point is that, by (7.6), $\dot{g}_{h,k}(\cdot)$ will have a fixed sign over the intervals of interest to us, in Cases (i) and (ii) of the proof.

Lemma 7.1. *Let $\mathbb{X}$ be a solution of the fluid model equations* (4.16)-(4.21) *and* (4.26) *for a reentrant line whose discipline is LBFS, with $\rho = e$. Assume that for given classes h and k, with $h < k$, $Z_\ell(t)$ is constant on (u, v), for $\ell = k+1, \ldots, K$, and $Z_h(t) > 0$ on (u, v). Then,*

$$\dot{g}_{h,k}(t) = 1 - \bar{m}_h d_k(t) \quad \text{on } (u, v). \tag{7.6}$$

Proof. By (7.5),

$$\dot{g}_{h,k}(t) = \sum_{\ell \in \mathcal{G}_{h,k}} m_\ell (d_\ell(t) - d_k(t)).$$

Since $Z_\ell(t)$, $\ell \ge k+1$, is constant on (u, v), this equals

$$\sum_{\ell \in \mathcal{G}_h} m_\ell (d_\ell(t) - d_k(t)) = \left(\sum_{\ell \in \mathcal{G}_h} m_\ell d_\ell(t) \right) - \bar{m}_h d_k(t)$$

on (u, v). On (u, v), $Z_h(t) > 0$, and hence $\sum_{\ell \in \mathcal{G}_h} m_\ell d_\ell(t) = 1$. This implies (7.6). □

We first consider the case where k is not a lowest priority class at a station. One then has $\bar{m}_k < 1$.

Case (i). Class k is not a lowest priority class. We wish to show that for appropriate t_k, not depending on $\mathbb{X}$, $Z_k(t) = 0$ on $[t_k, \infty)$. We do this in two steps. In Step 1, we show that for appropriate t'_k satisfying (7.7), $Z_k(t'_k) = 0$. In Step 2, we show that $Z_k(t) = 0$ on $[t'_k, \infty)$. Since the bound in (7.7) is uniform over all $\mathbb{X}$, t_k can be chosen not to depend on $\mathbb{X}$.

Step 1. For each fluid model solution $\mathbb{X}$, there is a t'_k satisfying $Z_k(t'_k) = 0$ and

$$t'_k - t_{k+1} \in (0, (t_{k+1} + 2)/(1/\bar{m}_k - 1)]. \tag{7.7}$$

Proof. By (7.5), $\dot{f}_k(t) = 1 - d_k(t)$ for $t \geq 0$. So, by (7.3), whenever $t \geq t_{k+1}$ and $Z_k(t) > 0$, $\dot{f}_k(t) = 1 - 1/\bar{m}_k < 0$. Consequently, there exists a t'_k satisfying $f_k(t'_k) = 0$, and hence $Z_k(t'_k) = 0$, with

$$t'_k - t_{k+1} \in (0, f_k(t_{k+1})/(1/\bar{m}_k - 1)]. \tag{7.8}$$

Since $|Z(0)| = 2$ and $\alpha_1 = 1$, $f_k(t) \leq |Z(t)| \leq t + 2$ for all t. Together with (7.8), this implies (7.7). □

Step 2. Choose t'_k as in Step 1. Then, $Z_k(t) = 0$ for $t \in [t'_k, \infty)$.

Proof. We assume the claim is not true, and that there exists an interval $[a, b] \subset [t'_k, \infty)$ such that $Z_k(a) = 0$ and $Z_k(t) > 0$ for $t \in (a, b]$. We will show that, depending on the sign of $\dot{g}_{h,k}(t)$ close to a, this will result in a contradiction for either $t < a$ or $t > a$.

Let h be the first class before k that is nonempty at time a, that is,

$$h = \max\{\ell < k : Z_\ell(a) > 0\}. \tag{7.9}$$

(The class $k = 0$ ensures that the set is not empty.) By the continuity of $Z(\cdot)$, there is an interval $(u, v) \subset [t_{k+1}, b]$, containing a, on which $Z_h(t) > 0$. One always has $g_{h,k}(t) \geq 0$. By the choice of h and $[a, b]$, $g_{h,k}(a) = 0$ and $g_{h,k}(t) > 0$, for $t \in (a, v] \subset (a, b]$, also hold.

Suppose now that $\bar{m}_k \leq \bar{m}_h$. On (a, v), $Z_k(t) > 0$, and so, by (7.3), $d_k(t) = 1/\bar{m}_k$. Therefore, by Lemma 7.1, $\dot{g}_{h,k}(t) = 1 - \bar{m}_h/\bar{m}_k \leq 0$ on (a, v). This contradicts the last sentence of the previous paragraph. Suppose instead that $\bar{m}_k > \bar{m}_h$. By (7.2), $d_k(t) \leq 1/\bar{m}_k$, and so by Lemma 7.1, $\dot{g}_{h,k}(t) \geq 1 - \bar{m}_h/\bar{m}_k > 0$ on $(u, a) \subset (u, v)$. This contradicts $g_{h,k}(a) = 0$ and the nonnegativity of $g_{h,k}(\cdot)$. So, there exists no point a as specified at the beginning of the proof. This implies that $Z_k(t) = 0$ for $t \in [t'_k, \infty)$. □

We now consider the case where k is the lowest priority class at a station. One then has $\bar{m}_k = 1$.

Case (ii). Class k is a lowest priority class. We wish to show that for appropriate t_k, not depending on $\mathbb{X}$, $Z_k(t)$ is constant on $[t_k, \infty)$. This requires three steps. In Step 1, we show that $Z_k(t)$ is nondecreasing on $[t_{k+1}, \infty)$. Set

$$\epsilon = \min\{1 - \bar{m}_\ell : \ell \text{ is not a lowest priority class}\};$$

then, $\epsilon > 0$. In Step 2, we show that for appropriate t'_k satisfying (7.11), $\dot{Z}_k(t'_k) < \epsilon$. Then, in Step 3, we show that $Z_k(t)$ is constant on $[t'_k, \infty)$. Since the bound in (7.11) is uniform over all $\mathbb{X}$, t_k can be chosen not to depend on $\mathbb{X}$. Steps 2 and 3 are the analogs of Steps 1 and 2 in Case (i). The reasoning here is similar, except that there are more alternatives to be considered in Step 3.

Step 1. For each fluid model solution $\mathbb{X}$, $Z_k(t)$ is nondecreasing on $[t_{k+1}, \infty)$.

Proof. For a given $t \in [t_{k+1}, \infty)$, set $h = \max\{\ell < k : Z_\ell(t) > 0\}$. Since $Z_\ell(t)$, $\ell \geq k+1$, is constant on $[t_{k+1}, \infty)$, one has $d_\ell = d_k$ for $\ell \geq k$. Also, since $Z_\ell(t) = 0$, and hence $\dot{Z}_\ell(t) = 0$, for $h < \ell < k$, one has $d_\ell(t) = d_{k-1}(t)$ for $h \leq \ell \leq k - 1$. It therefore follows from (4.21) and the LBFS property, that

$$d_{k-1}(t) \sum_{\ell \in \mathcal{G}_{h,k}} m_\ell + d_k(t) \sum_{\ell \in \mathcal{G}_h \setminus \mathcal{G}_{h,k}} m_\ell = \sum_{\ell \in \mathcal{G}_h} d_\ell(t) m_\ell = 1. \tag{7.10}$$

Also, $\sum_{\ell \in \mathcal{G}_h} m_\ell \leq 1$, and, by (7.2), $d_k(t) \leq 1$. So, by (7.10), $d_{k-1}(t) \geq 1$, and hence $\dot{Z}_k(t) = d_{k-1}(t) - d_k(t) \geq 0$. Since $t \in [t_{k+1}, \infty)$ is arbitrary, this implies that $Z_k(t)$ is nondecreasing on the set. □

Step 2. For each fluid model solution $\mathbb{X}$, there is a regular point t'_k, with $\dot{Z}_k(t'_k) < \epsilon$ and

$$t'_k - t_{k+1} \in (0, (t_{k+1} + 3)/\epsilon]. \tag{7.11}$$

Proof. Assume that, on the contrary, for a given $w > t_{k+1}$, $\dot{Z}_k(t) \geq \epsilon$ for all regular $t \in (t_{k+1}, w)$. Then, $f_k(t)$ is constant over $[t_{k+1}, w]$, where $f_k(t)$ is given by (7.4). To see this, note that since $Z_k(t) > 0$ for $t \in (t_{k+1}, w)$, it follows from (7.3) that $d_k(t) = 1/\bar{m}_k = 1$. Together with (7.5), this implies $f_k(t)$ is constant on $[t_{k+1}, w]$.

It follows that $Z_k(w) \leq f_k(w) = f_k(t_{k+1})$. Since $Z_k(w) - Z_k(t_{k+1}) \geq \epsilon(w - t_{k+1})$ also holds, we have

$$w - t_{k+1} \leq Z_k(w)/\epsilon \leq f_k(t_{k+1})/\epsilon.$$

Consequently, there exists a regular t'_k, with $\dot{Z}_k(t'_k) < \epsilon$ and

$$t'_k - t_{k+1} \in (0, f_k(t_{k+1})/\epsilon + 1].$$

Since $f_k(t_{k+1}) \leq |Z(t_{k+1})| \leq t_{k+1} + 2$, this implies (7.11). □

Step 3. Choose t'_k as in Step 2. Then $Z_k(t)$ is constant over $[t'_k, \infty)$.

Proof. We assume the claim is not true. By the monotonicity of $Z_k(t)$ in Step 1, there exists an interval $[a, b] \subset [t'_k, \infty)$ such that $Z_k(t) = Z_k(t'_k)$ for $t \in [t'_k, a]$ and $Z_k(t) > Z_k(t'_k)$ for $t \in (a, b]$. As in Step 2 of Case (i), we will show that, depending on the value of $\dot{g}_{h,k}(t)$ close to a, this will result in a contradiction for either $t > a$ or $t < a$. In the first case, we use the monotonicity of $Z_k(t)$ in Step 1, and in the second case, we use Steps 1 and 2. The reasoning will be different, depending on whether $a = t'_k$ or $a > t'_k$.

We consider first the case where $a > t'_k$. Choose h as in (7.9). There is an interval $(u, v) \subset [t'_k, b]$, containing a, on which $Z_h(t) > 0$. On $[a, v)$, $t = a$ is a strict minimum for $g_{h,k}(t)$, and on $(u, a]$, it is a minimum. This is because $t = a$ is a strict minimum (respectively, minimum) for $Z_k(t)$ on $[a, v)$ (respectively, $(u, a]$), and $Z_\ell(a) = 0$ for $h < \ell < k$.

The reasoning now proceeds as in Step 2 in Case (i). First, suppose that $\bar{m}_k \leq \bar{m}_h$ (i.e., $\bar{m}_h = 1$). On (a, v), $d_k(t) = 1/\bar{m}_k$, and so by Lemma 7.1, $\dot{g}_{h,k}(t) = 1 - \bar{m}_h/\bar{m}_k \leq 0$ on (a, v). This contradicts the conclusion, in the previous paragraph, that a is a strict minimum for $g_{h,k}(t)$ on $[a, v)$. So, suppose instead that $\bar{m}_k > \bar{m}_h$ (i.e., $\bar{m}_h < 1$). On (u, a), $d_k(t) \leq 1/\bar{m}_k$, and so by Lemma 7.1, $\dot{g}_{h,k}(t) \geq 1 - \bar{m}_h/\bar{m}_k > 0$. This contradicts the conclusion, in the previous paragraph, that a is a minimum for $g_{h,k}(t)$ on $(u, a]$. So, $a > t'_k$ is not possible.

We turn now to the case where $a = t'_k$. The possibility that $\bar{m}_k \leq \bar{m}_h$ is excluded by exactly the same reasoning, as in the previous paragraph, by analyzing the behavior of $\dot{g}_{h,k}(t)$ over (a, v), with v chosen close to a. So, suppose instead that $\bar{m}_k > \bar{m}_h$. By Step 2, $\dot{D}(t'_k)$ and $\dot{Z}(t'_k)$ exist. Since $Z_\ell(t'_k) = 0$ for $h < \ell < k$, it follows that $d_h(t'_k) = d_{k-1}(t'_k)$. Consequently,

$$\dot{Z}_k(t'_k) = d_{k-1}(t'_k) - d_k(t'_k) = d_h(t'_k) - d_k(t'_k). \tag{7.12}$$

Also, $d_h(t'_k) = \cdots = d_{k-1}(t'_k) \geq d_k(t'_k) = \cdots = d_K(t'_k)$, with the inequality following from Step 1. It therefore follows, as in (7.3), that $d_h(t'_k) \geq 1/\bar{m}_h$. Together with (7.2), this implies that (7.12) is at least

$$1/\bar{m}_h - 1 = \bar{m}_h^{-1}(1 - \bar{m}_h) \geq \epsilon.$$

This contradicts Step 2, where $\dot{Z}_k(t'_k) < \epsilon$ is given. So, $a = t'_k$ is also not possible. Hence, there exists no point a as specified at the beginning of the proof. This implies that $Z_k(t)$ is constant over $[t'_k, \infty)$. □

Together, Cases (i) and and (ii) imply that $Z_k(t)$ is constant on $[t_k, \infty)$, for appropriate t_k, with $Z_k(t) = 0$ if k is not a lowest priority class. This completes the proof of Theorem 3.5. □

8 The 2-station, 5-class priority network

In this section, we analyze the behavior of the family of 2-station, 5-class SBP networks that was introduced in Section 3. We first demonstrate a heavy traffic result, Theorem 3.6, for these networks. In addition to the standard assumptions (3.4) and (3.5) on α^r, m^r and ρ^r, this result requires (3.21), i.e., that $\alpha_1(m_2 + m_5) < 1$. At the end of the section, we show that this condition is, in fact, needed.

This example shows that it is sometimes possible to apply standard heavy traffic limit results, such as Theorem 5.3, in unconventional situations. One can also show analogous results for the more elementary Lu-Kumar network in Lu and Kumar [40]. (There, a heavy traffic limit holds exactly when $\alpha_1(m_2 + m_4) < 1$.) We prefer to investigate the networks given in Theorem 3.6 since they are without immediate feedback; the somewhat more sophisticated arguments used here also give a better idea of the type of reasoning that is often needed to verify the conditions of Theorem 5.3. (The networks in Theorem 3.6 were first examined in Dai and Vande Vate [19].)

Throughout the section, we assume that $\alpha_1 = 1$, except when specified otherwise. Also, as in the previous section, we set $d_k(t) = \dot{D}_k(t)$ for the departure rate from a class k.

Proof of Theorem 3.6. Since the discipline is an SBP re-entrant line, we can use Theorem 5.3 to demonstrate the theorem. We first show that (a) the matrix R in (3.10) is well-defined and is completely-$\mathcal{S}$. We then show that (b) the corresponding fluid model is uniformly convergent with the lifting matrix Δ in (3.20). Checking (a) is quite straightforward. We use a piecewise linear Lyapunov function to check (b).

To show (a), recall the definition of the matrix G in (3.9). One can check that for this network,

$$I + G = \begin{pmatrix} m_1 + m_3 + m_5 & m_5 \\ m_2 + m_4 & m_4 \end{pmatrix} [\mathrm{diag}(m_1, m_4)]^{-1}.$$

It follows from (3.6), that

$$m_1 + m_3 + m_5 = 1, \quad m_2 + m_4 = 1. \tag{8.1}$$

Thus,

$$I + G = \begin{pmatrix} 1 & m_5 \\ 1 & m_4 \end{pmatrix} [\mathrm{diag}(m_1, m_4)]^{-1}.$$

By (8.1) and (3.21), $m_5 < m_4$. The determinant of $I + G$ is therefore positive; the inverse R is given by

$$R = (m_4 - m_5)^{-1} \,\mathrm{diag}(m_1, m_4) \begin{pmatrix} m_4 & -m_5 \\ -1 & 1 \end{pmatrix}.$$

To check that R is completely-$\mathcal{S}$, note that its diagonal elements are positive. Also, choose $a > 1$ so that $am_5 < m_4$ and $u = (1, a)'$; then, $Ru > 0$. This shows that R is completely-$\mathcal{S}$, and completes the proof of (a).

To prove (b), we show that there exists a $\delta > 0$ such that for any fluid model solution $\mathbb{X}$ with $|Z(0)| = 1$, $Z_k(t)$ is constant on $[\delta, \infty)$, for $k = 1, \ldots, 5$, and is zero, for $k = 2$, 3 and 5. We separate the proof into two steps. Let

$$f(t) = Z_2(t) + Z_3(t) + Z_5(t).$$

In Step 1, we show that there exists a $\delta > 0$, depending only on m, such that $f(t) = 0$ for $t \geq \delta$. In Step 2, we show that $Z_k(t)$ is constant on $[\delta, \infty)$ for $k = 1$ and 4.

Step 1. There exists a $\delta > 0$, such that for each fluid model solution $\mathbb{X}$ with $|Z(0)| = 1$, $f(t) = 0$ for $t \geq \delta$.

Proof. We claim that there exists an $\epsilon > 0$ such that $\dot{f}(t) \leq -\epsilon$ whenever $f(t) > 0$. It then follows that $f(t) = 0$ for $t \geq f(0)/\epsilon$. Since $f(0) \leq |Z(0)| = 1$, setting $\delta = 1/\epsilon$ implies that $f(t) = 0$ for $t \geq \delta$.

To prove the claim, we consider different cases, depending on whether or not the individual components $Z_k(t)$ are positive. For this, we recall the priority structure of the re-entrant line, which, in descending order, is given by $(5, 3, 1)$ at station 1 and $(2, 4)$ at station 2. We will also repeatedly use the observation that $Z_k(t) = 0$ implies $\dot{Z}_k(t) = 0$ at regular points, and so $d_{k-1}(t) = d_k(t)$ there.

We first assume that $Z_5(t) > 0$. This implies $d_1(t) = d_3(t) = 0$. If, in addition, either $Z_2(t) > 0$ or $Z_4(t) = 0$, then $d_4(t) = 0$. It follows that $\dot{f}(t) = -\mu_5$. If, on the other hand, $Z_2(t) = 0$ and $Z_4(t) > 0$, then $d_4(t) = \mu_4$, and so $\dot{f}(t) = \mu_4 - \mu_5$, which is also negative, since $\mu_5 > \mu_4$.

For the other cases, we assume that $Z_5(t) = 0$. If, in addition, both $Z_2(t) > 0$ and $Z_3(t) > 0$, then $d_1(t) = d_4(t) = d_5(t) = 0$. It follows that $\dot{f}(t) = -\mu_3$. Assume instead, then, that $Z_2(t) > 0$ but $Z_3(t) = 0$. Then,

$$\dot{f}(t) = \dot{Z}_2(t) = d_1(t) - d_2(t) = d_1(t) - \mu_2.$$

Two subcases arise, depending on whether or not $Z_1(t) = 0$. Under $Z_1(t) = 0$, one has $d_1(t) = 1$, and so $\dot{f}(t) = 1 - \mu_2 < 0$. Under $Z_1(t) > 0$ one has $m_1 d_1(t) + m_3 d_3(t) = 1$. Since $Z_3(t) = 0$ implies that $d_3(t) = d_2(t) = \mu_2$, solving the above equation produces $d_1(t) = \mu_1(1 - \mu_2 m_3)$. Consequently, $\dot{f}(t) = \mu_1(1 - \mu_2 m_3) - \mu_2$, which we wish to show is < 0. This is equivalent to $m_2 < m_1 + m_3$, which is implied by (3.21) and the first equation in (8.1).

We still need to consider the case where $Z_2(t) = Z_5(t) = 0$ and $Z_3(t) > 0$. Then,

$$\dot{f}(t) = \dot{Z}_3(t) = d_2(t) - d_3(t) = -d_3(t),$$

with the last equality holding since there is no service at class 1 when class 3 is occupied. There are two subcases, depending on whether or not $Z_4(t) = 0$. Under $Z_4(t) = 0$, one has $d_3(t) = d_5(t)$. Together with $m_3 d_3(t) + m_5 d_5(t) = 1$, this implies that $d_3(t) = 1/(m_3 + m_5)$, and so $\dot{f}(t) = -1/(m_3 + m_5)$. Suppose, instead, that $Z_4(t) > 0$. Since $Z_2(t) = 0$ and $Z_3(t) > 0$, there is no service at class 2, and so $d_4(t) = \mu_4$. Since $d_4(t) = d_5(t)$, it follows from this and $m_3 d_3(t) + m_5 d_5(t) = 1$, that $d_3(t) = \mu_3(1 - \mu_4 m_5)$. Consequently, $\dot{f}(t) = -\mu_3(1 - \mu_4 m_5)$, which is also negative since $m_5 < m_4$.

Together, the values of $\dot{f}(t)$ that are computed above show that, for appropriate $\epsilon > 0$, $\dot{f}(t) \leq -\epsilon$ whenever $f(t) > 0$. This completes the argument for Step 1. □

Step 2. Let δ be chosen as in Step 1. Then, $Z_1(t)$ and $Z_4(t)$ are constant on $[\delta, \infty)$.

Proof. By Step 1, $Z_k(t)$ is constant on $[\delta, \infty)$, for $k = 2$, 3 and 5, and so

$$d_1(t) = d_2(t) = d_3(t), \quad d_4(t) = d_5(t). \tag{8.2}$$

In order to show $Z_1(t)$ and $Z_4(t)$ are constant on $[\delta, \infty)$, it therefore suffices to show

$$d_2(t) = d_4(t) = 1 \tag{8.3}$$

for such t.

There are four cases to show, depending on whether or not $Z_1(t)$ and $Z_4(t)$ are positive. Assume, first, that $Z_1(t) > 0$ and $Z_4(t) > 0$. Then, $\sum_{k\in\mathcal{C}(j)} m_k d_k(t) = 1$ for $j = 1, 2$. Employing (8.2) at the first station, this can be rewritten as

$$\begin{aligned}(m_1 + m_3)d_2(t) + m_5 d_4(t) = 1,\\ m_2 d_2(t) + m_4 d_4(t) = 1.\end{aligned} \tag{8.4}$$

By (8.1), $d_2(t) = d_4(t) = 1$ solves (8.4). Using (8.1) and (3.21), it is easy to check that the system is nonsingular, and so this solution is unique.

Assume next that $Z_1(t) = 0$ and $Z_4(t) > 0$. Then, $d_1(t) = 1$ and $m_2 d_2(t) + m_4 d_4(t) = 1$. Together with (8.1) and (8.2), this implies (8.3). The reasoning for $Z_1(t) > 0$ and $Z_4(t) = 0$ is similar. Here, one instead employs $d_3(t) = d_4(t)$ and $m_1 d_1(t) + m_3 d_3(t) + m_5 d_5(t) = 1$, together with (8.1) and (8.2). For $Z_1(t) = Z_4(t) = 0$, one has $d_1(t) = 1$ and $d_3(t) = d_4(t)$. Together with (8.2), this implies (8.3). □

This completes the proof of Theorem 3.6. □

We conclude this section by providing a partial converse to Theorem 3.6, where the assumption (3.21) is replaced by (8.5). Some motivation for this is provided by Lemma 8.1 and the following discussion.

Theorem 8.1. *Consider a sequence of* 2*-station,* 5*-class SBP networks in Figure* 1*, with priority ranking given by* (3.19)*. Assume* (3.4), (3.5)*, and that* $Z^r(0) = 0$ *for each* $r > 0$*. If*

$$\alpha_1(m_2 + m_5) > 1, \tag{8.5}$$

then, with probability one, $|\tilde{Z}^r(t)| \to \infty$ *for each* $t > 0$*, as* $r \to \infty$*.*

As before, one can, without loss of generality, assume that $\alpha_1 = 1$. We also note that $|\tilde{W}^r(t)| \to \infty$ as $r \to \infty$ follows from the above assumptions, with the argument being the same as that for $|\tilde{Z}^r(t)| \to \infty$.

A key ingredient in the proof of the theorem is the following elementary lemma.

Lemma 8.1. *Consider a sequence of* 2*-station,* 5*-class SBP networks as in Figure* 1*, with priority ranking given by* (3.19)*, and with* $Z^r(0) = 0$ *for each* $r > 0$*. Then,*

$$Z_2^r(t) Z_5^r(t) = 0, \quad t \geq 0. \tag{8.6}$$

Consequently,

$$T_2^r(t) + T_5^r(t) \leq t, \quad t \geq 0. \tag{8.7}$$

The first part of Lemma 8.1 states that at any given time, either class 2 or class 5 must be empty. This condition holds at $t = 0$; it persists at later times because a job cannot move from class 1 to class 2 as long as class 5 is occupied, and cannot move from class 4 to class 5 as long as class 2 is occupied. The second part of the lemma is an immediate consequence of the inability of classes 2 and 5 to simultaneously receive service. Because of this behavior, classes 2 and 5 are said to form a *virtual station*; (3.21) is thus a *virtual station condition*. This type of behavior was first observed in Harrison and Nguyen [28], and in Dumas [24] for certain networks; it has been systematically employed in Dai and Vande Vate [19, 20].

We provide an abbreviated argument for Theorem 8.1. This is the only proof in the paper where we need to work with random quantities; on account of Theorems 5.1-5.3, it sufficed to work with fluid models for the proofs in Sections 6 and 7. We therefore provide only a quick account of the machinery, referring the reader elsewhere for more detail.

Proof of Theorem 8.1. By the strong law of large numbers,

$$\lim_{t\to\infty} \frac{E_k(t)}{t} = 1, \quad \lim_{n\to\infty} \frac{V_k(n)}{n} = 1, \quad \lim_{n\to\infty} \frac{\Phi_\ell^k(n)}{n} = P_{k\ell}, \tag{8.8}$$

with probability 1. Choose a sample path ω such that (8.8) holds and consider the sequence $\bar{\mathbb{X}}^r(\cdot,\omega)$, $r>0$, where

$$\bar{\mathbb{X}}^r(t,\omega) = r^{-2}\mathbb{X}^r(r^2t,\omega).$$

Since $|T^r(t,\omega) - T^r(s,\omega)| \le t-s$ for any $0 \le s \le t$ and $r>0$, by the Azela-Ascoli lemma, $\{\bar{T}^r(\cdot,\omega),\ r>0\}$ is precompact in the topology of uniform convergence on compact intervals. It follows from queueing network equations (4.1)-(4.5) and (4.8), and the strong law of large numbers in (8.8), that $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$ is precompact, as $r\to\infty$, in $\mathbb{D}^{4K+2J}[0,\infty)$, under the topology of uniform convergence on compact intervals. One can show that each limit point $\bar{\mathbb{X}}$, of $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$, is a fluid model solution to (4.16)-(4.21). (The reasoning is now quite standard, see, for example, Dai [15] for an analogous argument.)

We claim that the queue length $|\bar{Z}(t)|$ grows linearly in t. For this, we observe that

$$\begin{aligned} m_2(\bar{Z}_1(t)+\bar{Z}_2(t)) + m_5(\bar{Z}_1(t)+\cdots+\bar{Z}_5(t)) &= m_2(t-\bar{D}_2(t)) + m_5(t-\bar{D}_5(t)) \\ &= (m_2+m_5)t - (\bar{T}_2(t)+\bar{T}_5(t)). \end{aligned} \tag{8.9}$$

The equalities follow from (4.16)-(4.21); they rephrase the total workload at the virtual station, at a given time, in terms of the arrival and departure of fluid. On account of (8.7), $\bar{T}_2(t)+\bar{T}_5(t) \le t$ for $t>0$. Together with (8.5), this implies that (8.9) is at least ct, for some $c>0$. So, for appropriate $c'>0$, $|\bar{Z}(t)| \ge c't$ for $t>0$. It follows that

$$\liminf_{r\to\infty} |\bar{Z}^r(t,\omega)| \ge c't.$$

Since $\tilde{Z}^r(t,\omega) = r\bar{Z}^r(t,\omega)$, this implies that

$$\lim_{r\to\infty} |\tilde{Z}^r(t,\omega)| = \infty$$

for each $t>0$, as desired. □

Theorems 3.6 and 8.1 analyze the behavior of sequences of queueing networks, as in Figure 1, where $m_2+m_5<1$ and $m_2+m_5>1$, respectively. One can, naturally, ask what happens in the borderline case, where $m_2+m_5=1$. It is not difficult to check that, in this case, $I+G$ is not invertible, and so R, in (3.10), is not defined. One can also construct a fluid model solution $\mathbb{X}$, with $Z(0)=(1,0,0,0,0)$, that is periodic (see Dai and Vande Vate [20, Section 8]). Therefore, the fluid model is not uniformly convergent. Moreover, $\bar{Z}^r$ is not tight as $r\to\infty$; we omit the rather long argument.

Acknowledgment. The second author appreciates the hospitality of the University of Aarhus during his stay there in Fall, 1998.

References

[1] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260 (1975).

[2] Bertsekas, D. and Gallager, R. *Data Networks*. Prentice-Hall, Englewood Cliffs, NJ, 1992.

[3] Borovkov, A. Some limit theorems in the theory of mass service, I. *Theory of Probability and its Applications* **9**, 550–565 (1964).

[4] Borovkov, A. Some limit theorems in the theory of mass service, II. *Theory of Probability and its Applications* **10**, 375–400 (1965).

[5] Bramson, M. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems: Theory and Applications* **23**, 1–26 (1997).

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

[7] Bramson, M. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory and Applications* **30**, 89–148 (1998).

[8] Chen, H. and Mandelbaum, A. Hierarchical modeling of stochastic networks II: strong approximations. In Yao, D. D., editor, *Probability Models in Manufacturing Systems*, chapter 3, pages 107–132. Springer, New York, 1994.

[9] Chen, H. and Ye, H. Private communication.

[10] Chen, H. and Zhang, H. Diffusion approximations for multiclass FIFO queueing networks. (1997). Preprint.

[11] Chen, H. and Zhang, H. Diffusion approximations for re-entrant lines with a first-buffer-first-served priority discipline. *Queueing Systems: Theory and Applications* **23**, 177–195 (1997).

[12] Chen, H. and Zhang, H. Diffusion approximations for Kumar-Seidman network under a priority service discipline. *Operations Research Letters* (1998). To appear.

[13] Chen, H. and Zhang, H. A sufficient condition for the diffusion approximations of multiclass queueing networks under priority service disciplines. Technical report, Faculty of Commerce and Business Administration, UBC, 1998.

[14] Coffman Jr., E. G., Puhalskii, A. A., and Reiman, M. I. Polling systems with zero switchover times: a heavy traffic averaging principle. *Annals of Applied Probability* **5**(3), 681–719 (1995).

[15] Dai, J. G. A fluid-limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability* **6**, 751–757 (1996).

[16] Dai, J. G. and Harrison, J. M. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Annals of Applied Probability* **2**, 65–86 (1992).

[17] Dai, J. G. and Kurtz, T. G. A multiclass station with Markovian feedback in heavy traffic. *Mathematics of Operations Research* **20**, 721–742 (1995).

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

[19] Dai, J. G. and VandeVate, J. Global stability of two-station queueing networks. In Paul Glasserman, K. S. and Yao, D., editors, *Proceedings of Workshop on Stochastic Networks: Stability and Rare Events*, volume 117 of *Lecture Notes in Statistics*, pages 1–26. Springer, New York, 1996.

[20] Dai, J. G. and VandeVate, J. The stability of two-station multi-type fluid networks. *Operations Research* (1998). To appear.

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

[22] Dai, J. G. and Weiss, G. Stability and instability of fluid models for re-entrant lines. *Mathematics of Operations Research* **21**, 115–134 (1996).

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

[24] Dumas, V. A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems: Theory and Applications* **25**, 1–43 (1997).

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

[26] Foschini, G. J. and Salz, J. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications* **26**, 320–327 (1978).

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

[28] Harrison, J. M. and Nguyen, V. Some badly behaved closed queueing networks. In Kelly, F. P. and Williams, R. J., editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 117–124. Springer, New York, 1995.

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

[30] Harrison, J. M. and Williams, R. J. Multidimensional reflected Brownian motions having exponential stationary distributions. *Annals of Probability* **15**, 115–137 (1987).

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

[32] Iglehart, D. L. Limit theorems for queues with traffic intensity one. *Ann. Math. Statist.* **36**, 1437–1449 (1965).

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

[34] Iglehart, D. L. and Whitt, W. Multiple channel queues in heavy traffic II: Sequences, networks, and batches. *Adv. Appl. Probab.* **2**, 355–369 (1970).

[35] Jackson, J. R. Networks of waiting lines. *Operations Research* **5**, 518–521 (1957).

[36] Johnson, D. P. *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks*. PhD thesis, University of Wisconsin, 1983.

[37] Kelly, F. P. Networks of queues with customers of different types. *J. Appl. Probab.* **12**, 542–554 (1975).

[38] Kingman, J. F. C. On queues in heavy traffic. *Proc. Camb. Phil. Soc.* **57**, 902–904 (1961).

[39] Kingman, J. F. C. The heavy traffic approximation in the theory of queues. In Smith, W. L. and *et al.*, editors, *Proc. Symp. on Congestion Theory*, pages 137–159. University of North Carolina Press, 1965.

[40] Lu, S. H. and Kumar, P. R. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control* **36**, 1406–1416 (1991).

[41] Peterson, W. P. A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research* **16**, 90–118 (1991).

[42] Prohorov, Y. Transient phenomena in processes of mass service. *Litovsk. Mat. Sb.* **3**, 199–205 (1963). In Russian.

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

[44] Reiman, M. I. Some diffusion approximations with state space collapse. In Baccelli, F. and Fayolle, G., editors, *Modeling and Performance Evaluation Methodology*, pages 209–240. Springer, Berlin, 1984.

[45] Reiman, M. I. A multiclass feedback queue in heavy traffic. *Advances in Applied Probability* **20**, 179–207 (1988).

[46] Reiman, M. I. and Williams, R. J. A boundary property of semimartingale reflecting Brownian motions. *Probability Theory and Related Fields* **77**, 87–97 (1988). Correction: **80**, 633 (1989).

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

[48] Whitt, W. Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Probab.* **8**, 74–94 (1971).

[49] Whitt, W. Heavy traffic theorems for queues: a survey. In Clarke, A. B., editor, *Mathematical Methods in Queueing Theory*, pages 307–350. Springer, New York, 1974.

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

[51] Williams, R. J. On the approximation of queueing networks in heavy traffic. In Kelly, F. P., Zachary, S., and Ziedins, I., editors, *Stochastic Networks: Theory and Applications*. Royal Statistical Society, Oxford University Press, 1996.

[52] Williams, R. J. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems: Theory and Applications* **30**, 27–88 (1998).

[53] Williams, R. J. An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems: Theory and Applications* **30**, 5–25 (1998).

[54] Yao, D. D. *Probability Models in Manufacturing Systems.* Springer Series in Operations Research. Springer, New York, 1994.