Applied Mathematics and Mechanics (English Edition), 2007, **28**(10):1331–1342

Diffusion approximations for multiclass queueing networks under preemptive priority service discipline *

DAI Wan-yang (戴万阳)

(Department of Mathematics, Nanjing University, Nanjing 210093, P. R. China)

(Communicated by GUO Mao-zheng)

Abstract We prove a heavy traffic limit theorem to justify diffusion approximations for multiclass queueing networks under preemptive priority service discipline and provide effective stochastic dynamical models for the systems. Such queueing networks appear typically in high-speed integrated services packet networks about telecommunication system. In the network, there is a number of packet traffic types. Each type needs a number of job classes (stages) of processing and each type of jobs is assigned the same priority rank at every station where it possibly receives service. Moreover, there is no inter-routing among different traffic types throughout the entire network.

Key words queueing network, preemptive priority, heavy traffic, semimartingale reflecting Brownian motion, fluid model, diffusion approximation, Lyapunov function

Chinese Library Classification O211, O226
2000 Mathematics Subject Classification 60F17, 60J60, 60K25, 90B15
Digital Object Identifier(DOI) 10.1007/s10483-007-1006-x

Introduction

Motivated from high-speed integrated services packet networks in telecommunication system, we study a type of multiclass queueing networks and establish associated approximating stochastic dynamical models in guidance of system performance evaluation and prediction. An important feature of such queueing networks is that each station in the network can process more than one class of jobs and have complicated feedback structure. Heavily loaded (close to service capacity) networks, where congestion and blocking are compelling problems, are of particular interest. Frequently, exact analysis of such networks is unavailable and it is natural to seek tractable approximations. In connection with this, it has been shown that a certain class of diffusion processes, known as semimartingale reflecting Brownian motions (SRBMs), approximate normalized versions of the queue length or workload processes in many single class queueing networks and some multiclass queueing networks under conditions of heavy traffic (e.g., Ref. [1–11]). The approximating Brownian models (SRMBs) use only the first two moments of the interarrival times, service times, and routing vectors associated with the networks. Moreover, for the Brownian models, many quantities including the stationary distribution can be computed either exactly or numerically[2,12–14]. Therefore, one can obtain performance estimates for the networks, like average queue lengths and average queueing delays, from their Brownian counterparts. Unfortunately, it is known that not all multiclass networks with feedback can be approximated by SRBMs in heavy traffic[15]. In fact, one of the challenges in contemporary research on queueing networks is to identify broad categories of networks which can be so approximated and to prove a heavy traffic limit theorem justifying the approximation. Hence, in this paper, we prove a heavy traffic limit theorem to show that SRBMs can be used

* Received Oct. 3, 2005; Revised Jul. 11, 2007
Project supported by the National Natural Science Foundation of China (No. 10371053)
Corresponding author DAI Wan-yang, Professor, Doctor, E-mail: nan5lu8@netra.nju.edu.cn

to approximate preemptive priority multiclass queueing networks with general interarrival time and service time distributions.

The traditional methods for demonstrating the approximations rely on the existence and uniqueness of solutions to the Skorohod problems associated with corresponding queueing networks[4]. However, uniqueness does not always hold, for example, in the case of networks with finite buffers[1−3], and in the case of multiclass networks with feedback[15]. In these cases, traditional methods cannot be readily extended. To overcome the difficulty, the weak convergence method was initially established by authors, such as W. Dai[1,2], J. G. Dai and W. Dai[3], for finite buffer networks, then along the similar line by authors like Williams[7,8,16] and Bramson[5,6] for certain family of multiclass networks. By the method, the keys to prove a heavy traffic limit theorem for multiclass networks with feedback are to show a state space collapse property and a completely-$\mathcal{S}$ property for some reflection matrix. Bramson and J. G. Dai[9] further summarized the results in Refs. [5,7,16] and claimed that as long as one can show the uniform convergence for associated fluid model then one can prove the state space collapse property. By this way, they established a heavy traffic limit theorem in Ref. [9] for a single station multiclass queueing system under preemtive priority service discipline. Nevertheless, it is not trivial to establish the uniform convergence and the completely-$\mathcal{S}$ property for more general network models (e.g., as presented above). Hence the proofs of these two properties are the main parts in our justification.

The rest of this paper is organized as follows. The network model is described in Section 1. Our main theorem is stated in Section 2 and its proof is given in Section 3.

1 Queueing network model

The queueing network under consideration has J single server stations and each station has an infinite capacity waiting buffer. In the network, there are h traffic types and each type consists of J job classes distributed at different stations. Therefore, the network is populated by $K(=Jh)$ job classes. Stations are labeled by $j = 1, \cdots, J$, and classes by $k = 1, \cdots, K$. When a job of a type arrives from outside the network, it may only receive service for part of J job classes and may visit a particular class more than once, then 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 and becomes a new job of a new class after a change, changing classes happens only at 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 leave the network eventually and there is no inter-routing among different job types throughout the entire network.

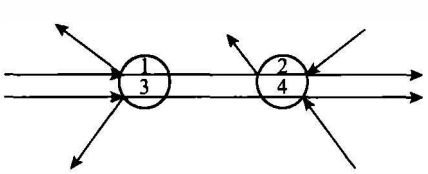

Fig. 1 A two-station, two-type and four-class network

Concerning the service discipline that controls the order in which jobs are served at each station, each type of jobs is assigned the same priority rank at every station where it possibly receives service (see, for example in Fig. 1, where type 1 traffic possibly requires class 1 and class 2 services, type 2 traffic possibly requires class 3 and class 4 service, job classes in type 1 have the lower priority at their corresponding stations). 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 ranked non-empty class at the server station. Moreover, we 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 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. Finally, we assume our policy is non-idling, namely, a server is never idle when there are

jobs waiting to be served at its station.

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 independent and identically distributed (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 the $3K$ sequences

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

are mutually independent. We set $a_k = \text{var}(u_k(1))$ and $b_k = \text{var}(v_k(1))$, and assume that $a_k < \infty$ and $b_k < \infty$, and that u_k and v_k are unitized, i.e., $E[u_k(1)] = 1$ and $E[v_k(1)] = 1$. For each i, $u_k(i)/\alpha_k$ denote the interarrival time between the $(i-1)$th and the ith externally arriving job at class k, $m_k v_k(i)$ denote the service time for the ith class k job, and $\phi^k(i)$ 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 set $\mathcal{E} = \{k : \alpha_k \neq 0\}$. We suppose that the routing vector $\phi^k(i)$ takes values in $\{e_0, e_1, \cdots, e_K\}$, where e_0 is the K-dimensional vector with components 0 and, for $l = 1, \cdots, K$, e_l is the K-dimensional vector with lth component 1 and other components 0. When $\phi^k(i) = e_l$, the ith job departing class k becomes a class l job. We let $P_{kl} = P\{\phi^k(i) = e_l\}$ be the probability that a job departing class k becomes a class l job (of the same type). The $K \times K$ matrix $P = (P_{kl})$ is the routing matrix of the network. From our assumption that every job will leave the system eventually, our network is open, namely, the matrix

$$Q = I + P' + (P')^2 + \cdots \tag{2}$$

is finite, which is equivalent to $(I - P')$ being invertible, with $Q = (I - P')^{-1}$, where I denotes the identity matrix and the prime symbol on P denotes its transpose.

To study open multiclass queueing networks, one employs the solution λ_l, $l = 1, \cdots, K$, of the traffic equations

$$\lambda_l = \alpha_l + \sum_{k=1}^{K} \lambda_k P_{kl}, \tag{3}$$

or equivalently, in vector form, of $\lambda = \alpha + P'\lambda$. Since the network corresponding to P is open, the unique solution λ of Eq. (3) 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 equals to the long-run average rate out of that class, this rate will equal λ_k. Employing vectors $m = (m_1, \cdots, m_K)'$ and λ, one defines the traffic intensity ρ_j for the jth server as

$$\rho_j = \sum_{k \in \mathcal{C}(j)} \lambda_k m_k. \tag{4}$$

In vector form, ρ is given by $\rho = CM\lambda$, where $M = \text{diag}(m)$ is the $K \times K$ diagonal matrix whose diagonal entries are given by the components of m and all other entries are 0, moreover, C is the constituency matrix,

$$C_{jk} = \begin{cases} 1, & \text{if } k \in \mathcal{C}(j), \text{ that is, class } k \text{ is served at station } j; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

When $\rho_j \leq 1$, ρ_j is also referred to as the nominal fraction of time server j is busy. In this paper, we study the networks in which ρ_j is close to one for each station j. Such networks are said to be "heavily loaded".

2 Heavy traffic limit theorem

Before we state our heavy traffic limit theorem, we require additional terminology including scaling, heavy traffic conditions and some definitions.

2.1 Scaling and heavy traffic conditions

Let α^r and m^r denote the vectors of the external arrival rates and mean service times for a sequence of networks indexed by $r \in \{1,2,\cdots\}$. Let $\lambda^r = Q\alpha^r$ and $\rho^r = CM^r\lambda^r$ with $M^r = \text{diag}(m^r)$. It is supposed that the set $\mathcal{E} = \{k : \alpha_k^r \neq 0\}$ and the routing matrix P do not depend on r. It is further assumed that α^r and m^r are so chosen that the below heavy traffic conditions are satisfied as $r \to \infty$,

$$\alpha_k^r \to \alpha_k > 0 \text{ for } k \in \mathcal{E}, \quad m_k^r \to m_k > 0 \text{ for } k = 1,\cdots,K, \tag{6}$$

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

$$r(\rho^r - e) \to \gamma, \tag{7}$$

where e is the J-dimensional vector with components 0 and γ is some J-dimensional vector. Notice that from Eqs. (6) and (7), we have

$$\rho = CM\lambda = e, \tag{8}$$

that is, each station is critically loaded in the limit. The interarrival times for class k are given by $\{u_k(i)/\alpha_k^r, i = 1,2,\cdots\}$ and the service times by $\{m_k^r v_k(i), i = 1,2,\cdots\}$. Therefore the squared coefficients of variation (SCV) of the interarrival times and service times for class k, a_k and b_k, do not depend on the index r.

For the rth network, the below processes Z^r, W^r and Y^r will be employed to measure its performance. The process $Z^r = \{Z^r(t), t \geq 1\}$ is K-dimensional with $Z_k^r(t)$ representing the number of class k jobs at time t. It is called the queue length process. The other two processes, $W^r = \{W^r(t), t \geq 1\}$ and $Y^r = \{Y^r(t), t \geq 1\}$, are both J-dimensional. For each station j, $W_j^r(t)$ denotes the amount of work for server j (measured in units of remaining service time) embodied in those jobs which are at station j at time t. The process W^r is called the (immediate) workload process. For each station j, $Y_j^r(t)$ denotes the total amount of time that the server at station j has been idle over time interval $[0,t]$. Y^r is called the (cumulative) idle-time process. The queue length and workload processes measure congestion and delay in the network; the idle-time process measures utilization of the resources (servers) in the network. The queue length, workload and idle-time processes are expected to grow when $\rho^r \to e$ as $r \to \infty$. Considering functional central limit theorems, we define the scaled queue length processes $\tilde{Z}^r(t) = (\tilde{Z}_1^r(t),\cdots,\tilde{Z}_K^r(t))'$, the scaled workload processes $\tilde{W}^r(t) = (\tilde{W}_1^r(t),\cdots,\tilde{W}_K^r(t))'$ and the scaled idle-time processes $\tilde{Y}^r(t) = (\tilde{Y}_1^r(t),\cdots,\tilde{Y}_K^r(t))'$ as follows:

$$\tilde{Z}_k^r = \frac{1}{r}Z_k^r(r^2t), \quad \tilde{W}_k^r(t) = \frac{1}{r}W_k^r(r^2t), \quad \tilde{Y}_k^r(t) = \frac{1}{r}Y_k^r(r^2t).$$

2.2 Some definitions

Throughout this section, $\mathcal{B}$ denotes the σ-algebra of Borel subsets of $\mathcal{R}_+^J$ with $\mathcal{R}_+ = [0,\infty)$, θ is a vector in $\mathcal{R}^J$ that denotes the J-dimensional Euclidean space, 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 $(\mathcal{R}_+^J, \mathcal{B})$. The following definition of an SRBM is adapted from Ref. [16].

Definition 1 (SRBM) *An SRBM associated with the data* $(\mathcal{R}_+^J, \theta, \Gamma, R, \nu)$ *is an* $\{\mathcal{F}_t\}$*-adapted,* J*-dimensional process* W*, defined on some filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathcal{P})$*, such that* $\mathcal{P}$*-a.s.:*

1. W *has continuous paths with* $W(t) \in \mathcal{R}_+^J$ *for* $t \geq 1$,

2. $W = X + RY$ *for appropriate* J*-dimensional processes* X *and* Y.

3. *the processes* X *and* Y *satisfy the following properties, under* $\mathcal{P}$,

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

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

4. *the process* Y *is an* $\{\mathcal{F}_t\}$*-adapted* J*-dimensional process such that* $\mathcal{P}$*-a.s., for each* $j = 1, \cdots, J$,

(a) $Y_j(0) = 0$,

(b) Y_j *is continuous and nondecreasing,*

(c) Y_j *can increase only at time* t *where* $W_j(t) = 0$.

Definition 2 (Completely-$\mathcal{S}$) *A* $J \times J$ *matrix is called completely-*$\mathcal{S}$ *if and only if for each principal submatrix* $\tilde{R}$ *of* R *there is* $x > 0$ *such that* $\tilde{R}x > 0$*, where vector inequalities are interpreted componentwise.*

2.3 Main theorem

To state our heavy traffic limit theorem, we need some general assumptions. Recall that α and m are the limits in Eq. (6) and $\lambda = Q\alpha$. We will suppose that Eq. (6) holds, and $\lambda_k > 0$ for all k. Let Λ, Σ and Π denote the diagonal matrices with entries λ_k, b_k and $\alpha_k^3 a_k$ for $k = 1, \cdots, K$ along the main diagonal, and let Γ^k be the $K \times K$ matrix given by

$$\Gamma_{ll'}^k = \begin{cases} P_{kl}(1 - P_{kl}) & \text{if } l = l', \\ -P_{kl}P_{kl'} & \text{if } l \neq l', \end{cases}$$

with $l, l' = 1, \cdots, K$. One can check that Γ^k is the covariance matrix of the routing vector $\phi^k(1)$. Thus it is symmetric and nonnegative definitive. Set

$$H = C\left(\Lambda\Sigma + MQ\left(\Pi + \sum_{k=1}^{K} \lambda_k \Gamma^k\right) Q'M\right) C'. \tag{9}$$

From Eq. (9), it is easy to see that H is symmetric and nonnegative definitive since $\sum_{k=1}^{K} \lambda_k \Gamma^k$ and the two diagonal matrices are each symmetric and nonnegative definitive. In the sequel, we will always assume that H is strictly positive definitive to guarantee the randomness of our network. Let Δ denote a $K \times J$ nonnegative matrix with entries given by

$$\Delta_{kj} = \begin{cases} \dfrac{1}{m_k}, & \text{if } k \text{ is the lowest priority class at station } j; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

For a vector $a = (a_1, \cdots, a_d)'$, define $\|a\| = \max_{i=1}^{d} |a_i|$ and let $\Rightarrow$ denote the weak convergence in Skorohod topological space[17]. Then we have the following theorem.

Theorem 1 *Assume that Eqs.* (6) *and* (7) *are true, the initial data satisfy*

$$\tilde{W}^r(0) \Rightarrow W^*(0) \quad \text{as } r \to \infty \tag{11}$$

for some nonnegative random vector $W^*(0)$*, and*

$$\|\tilde{Z}^r(0) - \Delta\tilde{W}^r(0)\| \to 0 \quad \text{in probability as } r \to \infty. \tag{12}$$

Let $\tilde{X}^r = \tilde{W}^r - R\tilde{Y}^r$. *Then, as* $r \to \infty$,

$$(\tilde{W} \quad , \quad , \quad , \quad , \quad , \quad , \tag{13}$$

for some W^*, X^*, Y^* *and* Z^*, *where* $W^* = X^* + RY^*$ *is a* $(\mathcal{R}_+^J, \theta, \Gamma, R, \nu)$*-SRBM, and the limits have state space collapse property*

$$Z^* = \Delta W^*. \tag{14}$$

Moreover, R *is a completely-*$\mathcal{S}$ *matrix given by*

$$R = (I + CMQP'\Delta)^{-1}, \tag{15}$$

and the parameters θ *and* Γ *are as follows*:

$$\theta = R\gamma, \quad \Gamma = RHR'. \tag{16}$$

3 Demonstrating main theorem

Lemma 1 *Under the conditions of Theorem* 2.3, *the matrix* $I + CMQP'\Delta$ *is invertible, and its inverse matrix* R *given by Eq.* (15) *is completely-*$\mathcal{S}$.

Proof Without loss of generality, we use consecutive numbers to index the job classes that have the same priority rank at stations 1 to J, i.e., the lowest priority job classes for station 1 to J are indexed by $\mathcal{I}_1 = \{1, \cdots, J\}$, the second lowest priority classes are indexed by $\mathcal{I}_2 = \{J+1, \cdots, 2J\}, \cdots$, and the highest priority classes are indexed by $\mathcal{I}_h = \{K-J+1, \cdots, K\}$, where $h = K/J$ is the number of job types. There is no inter-routing among job classes $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_h$. Thus, from the index method, we know that the routing matrix P is a block diagonal matrix with h $J \times J$ submatrices $P_{\mathcal{I}_i\mathcal{I}_i}$ for $i \in \{1, \cdots, h\}$ along the main diagonal, where the matrix $P_{\mathcal{I}_i\mathcal{I}_i}$ is corresponding to the routing probabilities of the job classes that have priority rank i at their stations. Then P can be denoted by

$$P = \mathrm{diag}(P_{\mathcal{I}_1\mathcal{I}_1}, \cdots, P_{\mathcal{I}_h\mathcal{I}_h}). \tag{17}$$

Therefore the matrix Q is also a block diagonal matrix and it can be written as

$$Q = (I - P')^{-1} = \mathrm{diag}\left((I - P'_{\mathcal{I}_1\mathcal{I}_1})^{-1}, \cdots, (I - P'_{\mathcal{I}_h\mathcal{I}_h})^{-1}\right). \tag{18}$$

The matrix CM can be blocked as

$$CM = [\mathrm{diag}(m_1, \cdots, m_J), \cdots, \mathrm{diag}(m_{K-J+1}, \cdots, m_K)]. \tag{19}$$

The matrix Δ is the following blocked matrix consisting of h $J \times J$ submatrices

$$\Delta = [\mathrm{diag}(m_1, \cdots, m_J)^{-1}, 0, \cdots, 0]', \tag{20}$$

where 0 is the $J \times J$ zero matrix. From Eqs. (18), (19) and (20), we have

$$\begin{aligned} I + G &= I + CM(I - P')^{-1}\Delta - CM\Delta \\ &= \mathrm{diag}(m_1, \cdots, m_J)(I - P'_{\mathcal{I}_1\mathcal{I}_1})^{-1} [\mathrm{diag}(m_1, \cdots, m_J)]^{-1}. \end{aligned} \tag{21}$$

From Eq. (21), we see that $I + G$ is invertible and the inverse R is given by

$$R = \mathrm{diag}(m_1, \cdots, m_J)(I - P'_{\mathcal{I}_1\mathcal{I}_1})[\mathrm{diag}(m_1, \cdots, m_J)]^{-1}. \tag{22}$$

Since the network is open, we know that the matrix $(I - P'_{\mathcal{I}_1\mathcal{I}_1})$ is completely-$\mathcal{S}$ (see, for example, Bernard and Kharroubi[18], Harrison and Reiman[19]). Then it is easy to check that R is completely-$\mathcal{S}$.

The remaining proof of our main theorem heavily depends on the below fluid model which is the analog of queueing network process as explained in Ref. [9],

$$A(t) = \alpha t + P'D(t), \tag{23}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{24}$$

$$W(t) = CM(A(t) + Z(0)) - CT(t), \tag{25}$$

$$CT(t) + Y(t) = \mathrm{e}^J t, \tag{26}$$

$$Y_j(t) \text{ can increase only at time } t \text{ where } W_j(t) = 0, \text{ for } j = 1, \cdots, J, \tag{27}$$

for all $t \geq 0$, where $\alpha = (\alpha_1, \cdots, \alpha_K)'$ is supposed to have nonnegative components and $m = (m_1, \cdots, m_K)'$ has positive components, $M = \mathrm{diag}(m)$, and P is a flow transition matrix. Moreover, we have

$$T(t) = MD(t). \tag{28}$$

We will assume that all of the fluid process are continuous and nonnegative with $A(\cdot)$, $D(\cdot)$, $T(\cdot)$ and $Y(\cdot)$ being nondecreasing. Moreover, one can check that

$$A(0) = D(0) = T(0) = Y(0) = 0 \tag{29}$$

follow from Eqs. (23)–(26). And that

$$W(t) = CMZ(t), \quad \text{for all } t \geq 0, \tag{30}$$

follows from Eqs. (24), (25) and (28).

Employing Eqs. (23)–(28), one can show that each of these fluid process is Lipschitz continuous, that is, for some $N > 0$ (depending only 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 fluid processes. In particular, they are 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 (A, D, T, W, Z, Y) satisfying Eqs. (23)–(30) if they are all differentiable at this time. Whenever we employ the derivative of a fluid process at a time t, we implicitly assume that t is a regular point. The notation, $\dot{f}(t)$, will be used to denote the derivative of a function f at the time t. Hence, it follows from our service discipline that

$$\dot{T}_k^+(t) = 1 \quad \text{when } Z_k^+(t) > 0, \quad \text{for all regular values of } t, \quad k = 1, \cdots, K, \tag{31}$$

where $Z_k^+(\cdot)$ corresponds to the total number of jobs presenting in classes whose priorities are at least as great as k and $T_k^+(\cdot)$ is corresponding to the cumulative amount of time that server $s(k)$ has spent on classes whose priorities are at least as great as k.

Definition 3 *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 g: $\mathcal{R}_+ \to \mathcal{R}_+$ with $g(t) \to 0$ as $t \to \infty$, such that for each fluid model solution $(A(\cdot), D(\cdot), T(\cdot), W(\cdot), Y(\cdot), Z(\cdot))$ with $\|Z(0)\| = 1$,*

$$\|Z(t) - Z(\infty)\| \leq g(t) \quad \textit{for all } t \geq 0, \tag{32}$$

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

$$Z(\infty) = \Delta w \quad \textit{for some } w \in \mathcal{R}_+^K. \tag{33}$$

Proposition 2 *Assume that the matrix* Δ *is given by Eq.* (10) *and the conditions* (6), (7), (11) *and* (12) *all hold. Then the fluid models* (23)–(31) *is uniformly convergent with the lifting matrix* Δ.

Proof With the notations in the proof of Lemma 1, we know that $\mathcal{I}_1$ is the index set of all classes with the lowest priority at each station, and $\mathcal{I}_i$ for $i \in \{2, \cdots, h\}$ denotes the index set of all classes with priority rank i at each station. We will show that the fluid level $Z_k(t)$ for $k \in \mathcal{I}_i$ with $i \in \{2, \cdots, h\}$ reaches zero in a finite time, and that the fluid level $Z_k(t)$ for $k \in \mathcal{I}_1$ remains constant after that time with the initial data $\|Z(0)\| = 1$.

Step 1 We show that the fluid level $Z_k(t)$ for $k \in \mathcal{I}_i$ with $i \in \{2, \cdots, h\}$ reaches zero in a finite time. For a K-dimensional vector x, we use $(x_{\mathcal{I}_1}, \cdots, x_{\mathcal{I}_h})'$ to denote the corresponding partition. Let $M_{\mathcal{I}_i}$ for $i \in \{2, \cdots, h\}$ be the $J \times J$ diagonal matrix with entries m_k, $k \in \mathcal{I}_i$, along the main diagonal, and let

$$Q_{\mathcal{I}_i} = (I - P'_{\mathcal{I}_i\mathcal{I}_i})^{-1}.$$

Then we can construct a Lyapunov function (total workload) of the fluid levels of all classes that have priority rank i as follows:

$$f_i(t) = e' M_{\mathcal{I}_i} Q_{\mathcal{I}_i} Z_{\mathcal{I}_i}(t), \tag{34}$$

where e denotes the J-dimensional vector with components 0. Noticing Eqs. (17) and (18) in the proof of Lemma 1, we can construct a Lyapunov function of the fluid levels of all higher priority classes

$$f(t) = \sum_{i=2}^{h} f_i(t). \tag{35}$$

Now we try to show that there exists an appropriate $\delta \geq 0$ such that $f(t) = 0$ for all $t \geq \delta$. In doing so, we use the induction method in terms of $i \in \{2, \cdots, h\}$.

Before going to the induction procedure, we need to derive some equations associated with the set $\mathcal{I}_i$ for $i \in \{1, 2, \cdots, h\}$. From Eqs. (23), (24) and the structure of the routing matrix P, we have

$$Z_{\mathcal{I}_i}(t) = Z_{\mathcal{I}_i}(0) + \alpha_{\mathcal{I}_i} t - (I - P'_{\mathcal{I}_i\mathcal{I}_i}) D_{\mathcal{I}_i}(t). \tag{36}$$

In the sequel, we need to consider the cases where $Z_k(t) = 0$ and $Z_k(t) \neq 0$ for $k \in \mathcal{I}_i$ and a fixed $t \geq 0$. Therefore we introduce the following notations:

$$\mathcal{I}_i^o = \{k : Z_k(t) = 0,\ k \in \mathcal{I}_i\}, \quad \mathcal{I}_i^n = \{k : Z_k(t) \neq 0,\ k \in \mathcal{I}_i\}. \tag{37}$$

Notice that the network is open, then we can solve Eq. (36) and obtain

$$D_{\mathcal{I}_i^o}(t) = Q_{\mathcal{I}_i^o} \left(Z_{\mathcal{I}_i^o}(0) + \alpha_{\mathcal{I}_i^o} t + P'_{\mathcal{I}_i^n \mathcal{I}_i^o} D_{\mathcal{I}_i^n}(t) - Z_{\mathcal{I}_i^o}(t) \right), \tag{38}$$

where $Q_{\mathcal{I}_i^o} = (I - P'_{\mathcal{I}_i^o \mathcal{I}_i^o})^{-1}$ and $Z_{\mathcal{I}_i^o}(t) = \mathbf{0}$ by Eq. (37). From the traffic equation (3) and the structure of the routing matrix P, we have

$$\lambda_{\mathcal{I}_i} = \alpha_{\mathcal{I}_i} + P'_{\mathcal{I}_i\mathcal{I}_i} \lambda_{\mathcal{I}_i}. \tag{39}$$

Solving Eq. (39), we get

$$\lambda_{\mathcal{I}_i^o} = Q_{\mathcal{I}_i^o} \left(\alpha_{\mathcal{I}_i^o} + P'_{\mathcal{I}_i^n \mathcal{I}_i^o} \lambda_{\mathcal{I}_i^n} \right). \tag{40}$$

We are now ready to go into the induction justification procedure. In the first step of the induction, we consider the case $i = h$ and show that the fluid level $Z_k(t)$ for $k \in \mathcal{I}_h$ reaches zero in a finite time, where $\mathcal{I}_h$ is the index set of all classes that have the highest priority at each station. Instead, we demonstrate that there exists an appropriate $\delta_h \geq 0$ such that $f_h(t) = 0$ for all $t \geq \delta_h$, and hence $Z_k(t) = 0$ for $t \geq \delta_h$ and $k \in \mathcal{I}_h$. In fact, when $f_h(t) > 0$ at a point t, then there is at least one $k \in \mathcal{I}_h$ such that $Z_k(t) > 0$ and hence $\mathcal{I}_h^n$ is nonempty. Thus from Eqs. (34), (36) and (39) and for each regular point t, we have

$$\begin{aligned}
\dot{f}_h(t) &= e' \\
&= e'_{\mathcal{I}_h^o} M_{\mathcal{I}_h^o} \left(\lambda_{\mathcal{I}_h^o} - \dot{D}_{\mathcal{I}_h^o}(t)\right) + e'_{\mathcal{I}_h^n} M_{\mathcal{I}_h^n} \left(\lambda_{\mathcal{I}_h^n} - \dot{D}_{\mathcal{I}_h^n}(t)\right) \\
&= e'_{\mathcal{I}_h^o} M_{\mathcal{I}_h^o} Q_{\mathcal{I}_h^o} P'_{\mathcal{I}_h^n \mathcal{I}_h^o} \left(\lambda_{\mathcal{I}_h^n} - \dot{D}_{\mathcal{I}_h^n}(t)\right) + e'_{\mathcal{I}_h^n} M_{\mathcal{I}_h^n} \left(\lambda_{\mathcal{I}_h^n} - \dot{D}_{\mathcal{I}_h^n}(t)\right) \\
&= \left(e'_{\mathcal{I}_h^o} M_{\mathcal{I}_h^o} Q_{\mathcal{I}_h^o} P'_{\mathcal{I}_h^n \mathcal{I}_h^o} M_{\mathcal{I}_h^n}^{-1} + e'_{\mathcal{I}_h^n}\right) \left(M_{\mathcal{I}_h^n} \lambda_{\mathcal{I}_h^n} - M_{\mathcal{I}_h^n} \dot{D}_{\mathcal{I}_h^n}(t)\right) \\
&= \sum_{k \in \mathcal{I}_h^n} \left(1 + \left(e'_{\mathcal{I}_h^o} M_{\mathcal{I}_h^o} Q_{\mathcal{I}_h^o} P'_{\mathcal{I}_h^n \mathcal{I}_h^o} M_{\mathcal{I}_h^n}^{-1}\right)_k\right) \left(\lambda_k m_k - m_k \dot{D}_k(t)\right) \\
&= -\sum_{k \in \mathcal{I}_h^n} \left(1 + \left(e'_{\mathcal{I}_h^o} M_{\mathcal{I}_h^o} Q_{\mathcal{I}_h^o} P'_{\mathcal{I}_h^n \mathcal{I}_h^o} M_{\mathcal{I}_h^n}^{-1}\right)_k\right) \left(\sum_{l \in \mathcal{C}(s(k)) \setminus \{k\}} \lambda_l m_l\right), \qquad (41)
\end{aligned}$$

where $e_{\mathcal{I}_h^o}$ and $e_{\mathcal{I}_h^n}$ are vectors with components ones with dimensions corresponding to the numbers of indices in $\mathcal{I}_h^o$ and $\mathcal{I}_h^n$, respectively. In the third equality of Eq. (41), we used Eqs. (38), (40) and the fact $Z_{\mathcal{I}_i^o}(t) = 0$. In the fifth equality of Eq. (41), $(\cdot)_k$ denotes the kth component of the corresponding vector. In the last equality of Eq. (41), we used the heavy traffic condition (8), and the fluid model property (31) and (28) to conclude that $\dot{D}_l(t) = 0$ for $l \in \mathcal{C}(s(k)) \setminus \{k\}$ and $m_k \dot{D}_k(t) = 1$, where k is the index of the highest priority class at station $s(k)$ and $\mathcal{C}(s(k)) \setminus \{k\}$ is the index set of all classes but class k at station $s(k)$. Now notice that $e'_{\mathcal{I}_h^o} M_{\quad h \quad h \quad h \quad h \quad h}$

$$\delta'_h \equiv \min_{\mathcal{I}_h^n \in \mathcal{G}_h} \left\{ \sum_{k \in \mathcal{I}_h^n} \left(1 + \left(e'_{\mathcal{I}_h^o} M_{\mathcal{I}_h^o} Q_{\mathcal{I}_h^o} P'_{\mathcal{I}_h^n \mathcal{I}_h^o} M_{\mathcal{I}_h^n}^{-1}\right)_k\right) \left(\sum_{l \in \mathcal{C}(s(k)) \setminus \{k\}} \lambda_l m_l\right) \right\} > 0, \qquad (42)$$

where $\mathcal{G}_h$ is the set consisting of sets of all possible nonempty combinations of class indices in $\mathcal{I}_h$. Then from Eqs. (41) and (42), we know that $\dot{f}_h(t) \leq -\delta'_h$ whenever $f_h(t) > 0$. Define

$$\delta_h \equiv f_h(0)/\delta'_h. \qquad (43)$$

Then, it is easy to show that $f_h(t) = 0$ for $t \geq \delta_h$ since $f_h(\cdot)$ is absolutely continuous. Moreover, $Z_k(t) = 0$ for all $t \geq \delta_h$ and $k \in \mathcal{I}_h$.

In the second step of the induction, we suppose that, for a fixed $i \in \{2, \cdots, h-1\}$, the claim is true for $i+1 \leq j \leq h$, that is, there exists a nonnegative number δ_{i+1} such that $Z_k(t) = 0$ for $t \geq \delta_{i+1}$ and $k \in \cup_{j=i+1}^h \mathcal{I}_j$ (union of sets $\mathcal{I}_j$). Furthermore, we can conclude that $\dot{Z}_k(t) = 0$ at regular points for all $t \geq \delta_{i+1}$ and $k \in \cup_{j=i+1}^h \mathcal{I}_j$. Thus from Eqs. (36) and (39) and for $i+1 \leq j \leq h$, we have

$$\dot{D}_{\mathcal{I}_j}(t) = (I - P'_{\mathcal{I}_j \mathcal{I}_j})^{-1} \alpha_{\mathcal{I}_j} = \lambda_{\mathcal{I}_j}. \qquad (44)$$

In the third step of the induction, we prove that, for the fixed i in the second step, the claim is true for $i \leq j \leq h$, that is, there exists a nonnegative number δ_i such that $Z_k(t) = 0$ for all $t \geq \delta_i$ and $k \in \cup_{j=i}^h \mathcal{I}_j$. Instead, we demonstrate that there exists an appropriate $\delta_i \geq 0$ such

that $f_j(t) = 0$ for all $t \geq \delta_i$ and $i \leq j \leq h$. To achieve this objective, we consider $t \geq \delta_{i+1}$. From the induction assumption in the second step, we know that $f_j(t) = 0$ for $i+1 \leq j \leq h$. Then we only need to discuss $f_i(t)$ for $t \geq \delta_{i+1}$. Notice that there is at least one $k \in \mathcal{I}_i$ such that $Z_k(t) > 0$ and hence $\mathcal{I}_i^n$ is nonempty when $f_i(t) > 0$ at a point t. Thus from Eqs. (34), (36), (39) and for each regular point t, we have

$$\begin{aligned}
\dot{f}_i(t) &= e' M_{\mathcal{I}_i} \left(\lambda_{\mathcal{I}_i} - \dot{D}_{\mathcal{I}_i}(t)\right) \\
&= e'_{\mathcal{I}_i^o} M_{\mathcal{I}_i^o} \left(\lambda_{\mathcal{I}_i^o} - \dot{D}_{\mathcal{I}_i^o}(t)\right) + e'_{\mathcal{I}_i^n} M_{\mathcal{I}_i^n} \left(\lambda_{\mathcal{I}_i^n} - \dot{D}_{\mathcal{I}_i^n}(t)\right) \\
&= e'_{\mathcal{I}_i^o} M_{\mathcal{I}_i^o} Q_{\mathcal{I}_i^o} P'_{\mathcal{I}_i^n \mathcal{I}_i^o} \left(\lambda_{\mathcal{I}_i^n} - \dot{D}_{\mathcal{I}_i^n}(t)\right) + e'_{\mathcal{I}_i^n} M_{\mathcal{I}_i^n} \left(\lambda_{\mathcal{I}_i^n} - \dot{D}_{\mathcal{I}_i^n}(t)\right) \\
&= \left(e'_{\mathcal{I}_i^o} M_{\mathcal{I}_i^o} Q_{\mathcal{I}_i^o} P'_{\mathcal{I}_i^n \mathcal{I}_i^o} M_{\mathcal{I}_i^n}^{-1} + e'_{\mathcal{I}_i^n}\right) \left(M_{\mathcal{I}_i^n} \lambda_{\mathcal{I}_i^n} - M_{\mathcal{I}_i^n} \dot{D}_{\mathcal{I}_i^n}(t)\right) \\
&= \sum_{k \in \mathcal{I}_i^n} \left(1 + \left(e'_{\mathcal{I}_i^o} M_{\mathcal{I}_i^o} Q_{\mathcal{I}_i^o} P'_{\mathcal{I}_i^n \mathcal{I}_i^o} M_{\mathcal{I}_i^n}^{-1}\right)_k\right) \left(\lambda_k m_k - m_k \dot{D}_k(t)\right) \\
&= -\sum_{k \in \mathcal{I}_i^n} \left(1 + \left(e'_{\mathcal{I}_i^o} M_{\mathcal{I}_i^o} Q_{\mathcal{I}_i^o} P'_{\mathcal{I}_i^n \mathcal{I}_i^o} M_{\mathcal{I}_i^n}^{-1}\right)_k\right) \left(\sum_{l \in \mathcal{C}(s(k)) \setminus \mathcal{H}_i(s(k))} \lambda_l m_l\right). \qquad (45)
\end{aligned}$$

Here, we need give some interpretations about the last equality of Eq. (45). $\mathcal{H}_i(s(k))$ is the index set of classes whose priority ranks are at least i at station $s(k)$, and $\mathcal{C}(s(k)) \setminus \mathcal{H}_i(s(k))$ denotes the index set of classes with priority ranks lower than i at station $s(k)$. Notice that the class k at station $s(k)$ has priority rank i and is nonempty. Thus, from Eq. (28) and the SBP fluid model property (31), we know that $\dot{D}_l(t) = 0$ for $l \in \mathcal{C}(s(k)) \setminus \mathcal{H}_i(s(k))$ and

$$\sum_{l \in \mathcal{H}_i(s(k))} m_l \dot{D}_l(t) = 1. \qquad (46)$$

Thus, from the heavy traffic condition (8), the induction assumption (44) in the second step, and Eq. (46), we have, for $k \in \mathcal{I}_i^n$,

$$\lambda_k m_k - m_k \dot{D}_k(t) = \sum_{l \in \mathcal{H}_i(s(k))} \left(\lambda_l m_l - m_l \dot{D}_l(t)\right) = -\sum_{l \in \mathcal{C}(s(k)) \setminus \mathcal{H}_i(s(k))} \lambda_l m_l.$$

Therefore we provide a proof for Eq. (45). Next notice that $e'_{\mathcal{I}_i^o}$ we have

$$\delta'_i \equiv \min_{\mathcal{I}_i^n \in \mathcal{G}_i} \left\{ \sum_{k \in \mathcal{I}_i^n} \left(1 + \left(e'_{\mathcal{I}_i^o} M_{\mathcal{I}_i^o} Q_{\mathcal{I}_i^o} P'_{\mathcal{I}_i^n \mathcal{I}_i^o} M_{\mathcal{I}_i^n}^{-1}\right)_k\right) \left(\sum_{l \in \mathcal{C}(s(k)) \setminus \mathcal{H}_i(s(k))} \lambda_l m_l\right) \right\} > 0, \qquad (47)$$

where $\mathcal{G}_i$ is the set consisting of sets of all possible nonempty combinations of class indices in $\mathcal{I}_i$. Then from Eqs. (45) and (47), we know that, for $t \in [\delta_{i+1}, \infty)$, $\dot{f}_i(t) \leq -\delta'_i$ whenever $f_i(t) > 0$. Define

$$\delta_i \equiv \delta_{i+1} + f_i(\delta_{i+1})/\delta'_i. \qquad (48)$$

Then, it is easy to show that $f_i(t) = 0$ for $t \geq \delta_i$ since $f_i(\cdot)$ is absolutely continuous, and hence $Z_k(t) = 0$ for all $t \geq \delta_i$ and $k \in \cup_{j=i}^h \mathcal{I}_j$.

In the end of this step, we take $\delta = \delta_2$ obtained from the above induction procedure. Then, we have that $f(t) = 0$ for $t \geq \delta$ and hence $Z_k(t) = 0$ for all $t \geq \delta$ and $k \in \cup_{j=2}^h \mathcal{I}_j$, where $f(t)$ is defined in Eq. (35).

Step 2 We demonstrate that the fluid level $Z_k(t)$ for $k \in \mathcal{I}_1$ remains constant after the finite time δ obtained in Step 1. In order to prove the claim, we restart the fluid model solution at time δ as follows:

$$\begin{aligned}&(A^\delta(t), D^\delta(t), T^\delta(t), W^\delta(t), Y^\delta(t), Z^\delta(t))\\ &= (A(t+\delta) - A(\delta), D(t+\delta) - D(\delta), T(t+\delta) - T(\delta), W(t+\delta), Y(t+\delta) - Y(\delta), Z(t+\delta))\end{aligned} \tag{49}$$

for $t \geq 0$. It is easy to check that the left-hand side of Eq. (49) is still a fluid model solution with the initial data $Z^\delta(0) = Z(\delta)$. From Eqs. (36), (40) and $Z^\delta_{\mathcal{I}_i}(t) = 0$ for $i \in \{2, \cdots, h\}$, we have

$$D^\delta_{\mathcal{I}_i}(t) = (I - P'_{\mathcal{I}_i\mathcal{I}_i})^{-1} Z^\delta_{\mathcal{I}_i}(0) + \lambda_{\mathcal{I}_i} t. \tag{50}$$

Hence, we have

$$\begin{aligned}Z^\delta_{\mathcal{I}_1}(t) &= Z^\delta_{\mathcal{I}_1}(0) + \alpha_{\mathcal{I}_1} t - (I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1} \left(e_{\mathcal{I}_1} t - \sum_{i=2}^{h} T^\delta_{\mathcal{I}_i}(t) - Y^\delta(t) \right).\\ &= X^\delta_{\mathcal{I}_1}(t) + \left(\alpha_{\mathcal{I}_1} - (I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1} (e_{\mathcal{I}_1} - \sum_{i=2}^{h} M_{\mathcal{I}_i} \lambda_{\mathcal{I}_i}) \right) t + (I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1} Y^\delta(t)\\ &= X^\delta_{\mathcal{I}_1}(t) + (I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1} Y^\delta(t),\end{aligned} \tag{51}$$

where

$$X^\delta_{\mathcal{I}_1}(t) \equiv Z^\delta_{\mathcal{I}_1}(0) + (I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1} \sum_{i=2}^{h} M_{\mathcal{I}_i} (I - P'_{\mathcal{I}_i\mathcal{I}_i})^{-1} Z^\delta_{\mathcal{I}_i}(0).$$

In the first equality of Eq. (52), we used Eqs. (36), (28) and (26). In the second equality, Eqs. (28) and (51) are employed. The third equality is obtained from the heavy traffic condition (8) and the traffic equation (39). Moreover, we have

$$\int_0^\infty Z^\delta_k(t) dY^\delta_k(t) = 0 \quad \text{for } k \in \mathcal{I}_1, \tag{52}$$

$$Y^\delta_k(\cdot) \text{ is non-decreasing and } Y^\delta_k(0) = 0 \text{ for } k \in \mathcal{I}_1. \tag{53}$$

Therefore, Eq. (52) to Eq. (54) form a so-called continuous deterministic regulation (Skorohod) problem. Since the network is open, we know that the matrix $(I - P'_{\mathcal{I}_1\mathcal{I}_1})$ is completely-$\mathcal{S}$ (See, for example, Bernard and Kharroubi[18], Harrison and Reiman[19]). Then it is easy to check that the reflection matrix $(I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1}$ in Eq. (52) is completely-$\mathcal{S}$. Thus, following an oscillation inequality in Bernard and Kharroubi[18] and Eq. (52), we have

$$\begin{aligned}\text{Osc}(Z^\delta_{\mathcal{I}_1}(\cdot), [t_1, t_2]) &\equiv \sup \left\{ \| Z^\delta_{\mathcal{I}_1}(t) - Z^\delta_{\mathcal{I}_1}(s) \| : t_1 \leq s < t \leq t_2 \right\}\\ &\leq \kappa \, \text{Osc}(X^\delta_{\mathcal{I}_1}(\cdot), [t_1, t_2]) = 0\end{aligned} \tag{54}$$

for any $0 \leq t_1 < t_2 < \infty$, where κ is a positive constant depending only on the reflection matrix $(I - P'_{\mathcal{I}_1\mathcal{I}_1}) M^{-1}_{\mathcal{I}_1}$ in Eq. (52). Then we have $Z_{\mathcal{I}_1}(\delta + t) = Z_{\mathcal{I}_1}(\delta)$ for all $t \geq 0$. Thus we finish the proof of the second step.

Finally, set $Z_{\mathcal{I}_1}(\infty) = Z_{\mathcal{I}_1}(\delta)$ and $Z_{\mathcal{I}_i}(\infty) = 0$ for $i \in \{2, \cdots, h\}$, then we have $Z(t) = Z(\infty)$ for $t \geq \delta$. If we can show $\sup_{\|Z(0)\|=1} \delta < \infty$, Proposition 2 is proved. As a matter of fact, this claim can be demonstrated by induction. First, for $j = h$, we have that $\sup_{\|Z(0)\|=1} \delta_h < \infty$ by

Eq. (43) and (34). Secondly, suppose that $\sup_{\|Z(0)\|=1} \delta_{i+1} < \infty$ for a fixed $i \in \{2, \cdots, h-1\}$. Thirdly, we consider the case $j = i$. Since $Z(\cdot)$ is Lipschitz continuous, we have

$$\|Z(\delta_{i+1})\| \leq \|Z(\delta_{i+1}) - Z(0)\| + \|Z(0)\| = N\delta_{i+1} + 1, \tag{55}$$

where N is a constant depending only on (α, M, P). From Eqs. (48), (34) and the induction assumption, we know that $\sup_{\|Z(0)\|=1} \delta_i < \infty$.

Proof of Theorem 2.3 Due to Lemma 1 and Proposition 2, it follows from Bramson and J. G. Dai[9] that our main theorem is true.

References

[1] Dai W. A heavy traffic limit theorem for queueing networks with finite capacity[C]. In: *INFORMS Applied Probability Conference*, Atlanta, U.S.A., June 14–16, 1995.

[2] Dai W. Brownian approximations for queueing networks with finite buffers: modeling, heavy traffic analysis and numerical implementations[D]. Ph D Dissertation, School of Mathematics, Georgia Institute of Technology, 1996. Aslo published in UMI Dissertation Services, A Bell & Howell Company, 300 N.Zeeb Road, Ann Arbor, Michican 48106, U.S.A. 1997.

[3] Dai J G, Dai W. A heavy traffic limit theorem for a class of open queueing networks with finite buffers[J]. *Queueing Systems*, 1999, **32**(1/3):5–40.

[4] Reiman M I. Open queueing networks in heavy traffic[J]. *Mathematics of Operations Research*, 1984, **9**(3):441–458.

[5] Bramson M. State space collapse with application to heavy traffic limits for multiclass queueing networks[J]. *Queueing Systems*, 1998, **30**(1/2):89–148.

[6] Bramson M. State space collapse for queueing networks[C]. In: *Proceedings of the International Congress of Mathematicians*, 1998, Vol III, 213–222.

[7] Williams R J. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse[J]. *Queueing Systems: Theory and Applications*, 1998, **30**(1/2):27–88.

[8] Williams R J. Reflecting diffusions and queueing networks[C]. In: *Proceedings of the International Congress of Mathematicians*, 1998, Vol III, 321–330.

[9] Bramson M, Dai J G. Heavy traffic limits for some queueing networks[J]. *Annals of Applied Probability*, 2001, **11**(1):49–90.

[10] Chen H, Zhang H. A sufficient condition and a necessary condition for the diffusion approximations of multiclass queueing networks under priority service displines[J]. *Queueing Systems*, 2000, **34**(1/4):237–268.

[11] Chen H, Zhang H. Diffusion approximations for some multiclass queueing networks with FIFO service disciplines[J]. *Mathematics of Operations Research*, 2000, **25**(4):679–707.

[12] Harrison J M, Williams R J. Multidimensional reflected Brownian motions having exponential stationary distributions[J]. *Annals of Probability*, 1987, **15**(1):115–137.

[13] Dai J G, Harrison J M. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis[J]. *Annals of Applied Probability*, 1992, **2**(1):65–86.

[14] Shen X, Chen H, Dai J G, Dai W. The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks[J]. *Queueing Systems*, 2002, **42**(1):33–62.

[15] Dai J G, Wang Y. Nonexistence of Brownian models of certain multicalss queueing networks[J]. *Queueing Systems*, 1993, **13**(1/3):41–46.

[16] Williams R J. An invariance principle for semimartingale reflecting Brownian motions in an orthant[J]. *Queueing Systems*, 1998, **30**(1/2):5–25.

[17] Ethier S N, Kurtz T G. Markov processes: charaterization and convergence[M]. New York: Wiley, 1986.

[18] Bernard A, Kharroubi A El. Régulation déterministes et stochastiques dans le premier "orthant" de R^n[J]. *Stochastics Stochastics Rep*, 1991, **34**(3/4):149–167.

[19] Harrison J M, Reiman M I. Reflected Brownian motion on an orthant[J]. *Annals of Probability*, 1981, **9**(2):302–308.