The Annals of Applied Probability
1994, Vol. 4, No. 1, 26–42

ON THE CONVERGENCE OF MULTICLASS QUEUEING NETWORKS IN HEAVY TRAFFIC

BY J. G. DAI[1] AND VIÊN NGUYEN

Georgia Institute of Technology and M.I.T.

The subject of this paper is the heavy traffic behavior of a general class of queueing networks with first-in–first-out (FIFO) service discipline. For special cases that require various assumptions on the network structure, several authors have proved heavy traffic limit theorems to justify the approximation of queueing networks by reflecting Brownian motions. Based on these theorems, some have conjectured that the Brownian approximation may in fact be valid for a more general class of queueing networks.

In this paper, we prove that the Brownian approximation does *not* hold for such a general class of networks. Our findings suggest that it may be fruitful to consider a more general class of approximating processes.

1. Introduction. The past few years have witnessed a surge of research activities in the area of Brownian approximations of queueing networks. The thrust of these activities is to establish a theoretical framework from which one can approximate queueing networks by reflecting Brownian motions (RBM's). These approximations are justified by so-called heavy traffic limit theorems and such limit theorems have been proved for a number of special cases that require (often restrictive) assumptions on the network structure. Nevertheless, some authors have proposed that one may apply Brownian approximations to a more general class of networks that operate under the first-in–first-out service discipline [13, 14].

In this paper, we prove that Brownian approximations are *not* valid for the general class of networks described by Harrison and Nguyen [13, 14]. We do so by first proving a "pseudo" heavy traffic limit theorem, which states that if the processes associated with the queueing network converge to a continuous limit, then that limit must be the Brownian system specified in [13] and [14]. We then present a queueing network example developed by Dai and Wang [9] for which the specified Brownian approximation is not well defined. Our findings suggest that it may be fruitful to consider a more general class of approximating processes. In addition, other service disciplines may yield more tractable structures.

Received October 1992; revised June 1993.

[1] Supported in part by NSF Grants DMS-92-03524, DDM 92-15233 and two grants from Texas Instruments Corporation.

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

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

We consider a network composed of d single server stations, which we index by $j = 1, \dots, d$. The network is populated by c classes of customers, and each class k has its own exogenous renewal arrival process $E_k = \{E_k(t), t \geq 0\}$ (possibly null), where $E_k(t)$ is the number of class k customers who have arrived at the network by time t. For each customer class $k = 1, \dots, c$, it is assumed that $E_k(0) = 0$ and customer interarrival times have mean $1/\alpha_k$ with squared coefficient of variation $c^2_{a,k}$. (The squared coefficient of variation, henceforth SCV, of a random variable is defined as its variance divided by the square of its mean.) We denote by E the c-dimensional process with components $E_1, \dots, E_c$. (All vectors are envisioned as column vectors.) We assume that arrival processes $E_1, \dots, E_c$ are independent and $\alpha_k > 0$ for at least one k. For each k, α_k is interpreted as the *long-run average arrival rate* for class k customers. These customers require service at station $s(k)$, and their service times are independent and identically distributed (i.i.d.) with mean m_k and SCV $c^2_{s,k}$. The service time sequences for the various customer classes are independent of one another and are also independent of the arrival processes. Upon completion of service at station $s(k)$, a class k customer becomes a customer of class l with probability P_{kl} and exits the network with probability $1 - \Sigma_l P_{kl}$, independent of all previous history. The transition matrix $P = (P_{kl})$ is taken to be transient, which simply means that all customers eventually leave the network. Hence the networks we are considering are open queueing networks. We assume that the waiting buffer at each station has infinite capacity, and that customers are served at each station on a first-in–first-out (FIFO) basis. Hereafter, we will refer to such a network as a *multiclass open queuing network*.

Such a description of a multiclass network is now quite standard, as in Harrison and Nguyen [13], [14]. (The class of queueing networks described here is, in fact, an important special case of the setup in [13 and 14].) Figure 1 shows an example of such a network, which Dai and Wang [9] have studied. Customers arrive at station 1 according to a Poisson process with rate $\alpha_1 = 1$. Each customer follows a deterministic route whose sequence of station visitations is $1, 1, 2, 2, 1$, after which the customer departs from the network. Hence, each customer makes five stops before exiting the network, and we designate those customers in their kth stop as class k customers.

In his pioneering paper on queueing networks, Jackson [20] assumed that customers visiting or occupying any given station are essentially indistinguishable from one another, and that a customer completing service at

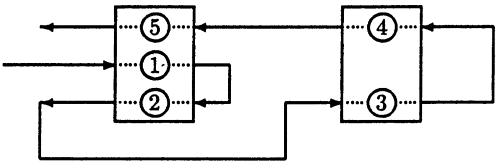

FIG. 1. *A two-station network with self-feedback.*

station i will move next to station j with some fixed probability P_{ij}, independent of all previous history. Thus in Jackson's networks, each station serves a single customer class; hence, these networks have been called *single-class* networks. Jackson's model was extended by Baskett, Chandy, Muntz and Palacios [1] and Kelly [22] to networks populated by multiple types of customers, each type following a deterministic route. The routing mechanism described in this paper subsumes those considered in [1] and [22]. Readers are referred to [12], [13] and [14] for further discussion.

For each $j = 1, \dots, d$ and each $t \geq 0$, let $W_j(t)$ denote the sum of the impending service times for all customers who are queued at station j at time t, plus the remaining service time for any customer who may be in service there at time t. If a new customer arrives to station j at time t, that customer must wait $W_j(t)$ time units before gaining access to the server, so one can also describe $W_j(t)$ as the virtual waiting time process for station j. Set $W(t)$ to be the d-vector with components $W_1(t), \dots, W_d(t)$. Define the process $W_j = \{W_j(t), t \geq 0\}$ and let W be the corresponding d-dimensional work-load process defined in the natural way.

Intuitively, when the system is heavily loaded, the work load $W(t)$ at time t will be large for large t. Let ρ_j be the *traffic intensity* at station j (this term will be defined in Section 2). As an example, because the arrival rate is set to be 1, the traffic intensities for the network pictured in Figure 1 are given by $\rho_1 = m_1 + m_2 + m_5$ and $\rho_2 = m_3 + m_4$. To facilitate our explanation of the underlying concepts, let us for the moment assume that

$$\rho_j = 1, \qquad j = 1, \dots, d. \tag{1.1}$$

Condition (1.1) is a special form of the *heavy traffic condition* as described in Section 2. For fixed $t \geq 0$, we are interested in how fast $W(nt)$ goes to infinity as $n \to \infty$. It has been widely believed that

$$\tilde{W}^n(\cdot) \equiv \frac{1}{\sqrt{n}} W(n\,\cdot) \Rightarrow W^*(\cdot) \quad \text{as } n \to \infty, \tag{1.2}$$

where $W^* = \{W^*(t), t \geq 0\}$ is a d-dimensional semimartingale reflecting Brownian motion (SRBM) and the symbol $\Rightarrow$ denotes weak convergence. We will clarify the notion of weak convergence in Section 2. For the definition of an SRBM, we refer the reader to Definition 1.1 of [27]. Our main contribution in this paper is the proof that conjecture (1.2) does not hold in general.

In cases where (1.2) holds, the corresponding theorem is called a heavy traffic limit theorem. There now exists a variety of heavy traffic limit theorems for networks with certain special structures. The first heavy traffic limit theorem for networks of queues is due to Iglehart and Whitt [18, 19], who studied single-class queues in series. For single-class networks with routing mechanisms similar to that of Jackson's networks, but whose interarrival times and service times may have general distributions, Reiman [24] proved that under the heavy traffic condition, the normalized queue length process converges weakly to a reflecting Brownian motion (RBM) as defined and constructed in Harrison and Reiman [15]. Reiman's proof was later simplified by Johnson [21]. Reiman's result was extended by Chen and Shanthikumar [5] to networks in which stations may have multiple servers. Peterson [23] proved an analogous heavy traffic limit theorem for multiclass feedforward networks. The term "feedforward" denotes a routing structure in which stations can be numbered so that customers always travel from lower numbered stations to higher numbered ones. Reiman [25] proved a heavy traffic limit theorem for a multiclass one station feedback queue. Dai and Kurtz [8] have greatly simplified Reiman's proof. A heavy traffic limit theorem for single-class closed networks was proved by Chen and Mandelbaum [3]. Strong approximations for single class networks were discussed in Glynn and Whitt [11] and Chen and Mandelbaum [4].

Until recently, it was believed that a heavy limit theorem should hold for multiclass open queueing networks of the type described in this section. Based on existing heavy traffic limit theorems, Harrison and Nguyen [13, 14] proposed Brownian models to approximate these networks. Unfortunately, Dai and Wang [9] have found two- and three-station networks for which Harrison and Nguyen's Brownian models fail to exist. (A more explicit interesting example showing no convergence was given by Whitt [28], who demonstrated "chaotic" behavior for certain multiclass open queueing networks.) Building on Dai and Wang's example, we prove in this paper the following general result: *There exist multiclass open queueing networks for which the sequence of normalized work-load processes* $\{\tilde{W}_n,\ n \geq 1\}$ *is not tight in the Skorohod topology.*

With networks for which corresponding heavy traffic limit theorems prevail, the Brownian approximation models proposed by Harrison and Nguyen [13, 14] are asymptotically justified. Their approximation scheme, known as the QNET method, provides a promising new tool for performance analysis of queueing networks that are not amenable to the exact mathematical analysis. The first step in a QNET analysis is to replace one's "exact" queueing model by an approximating Brownian system; see [13, 14]. The second step involves steady-state analysis of the approximating Brownian system; see [16], [6] and [7]. For a queueing network with d stations, this analysis requires that one determine the stationary distribution of a d-dimensional reflecting Brownian motion. Finally, summary statistics derived from that stationary distribution are used to obtain approximate steady-state performance measures for the original system. Unlike previous approximations, the Brownian approximations culminate in estimates of complete distributions; readers can find examples of Brownian estimates for complete sojourn time distributions in [14].

The remainder of this paper is organized as follows. We introduce some additional notation and definitions in Section 2. In Section 3, we state the heavy traffic conjecture and the main theorem of this paper. We prove our theorem by way of a "pseudo heavy traffic limit theorem," which we state and justify in Section 4. The proof of our main theorem is in Section 5. Finally, we discuss some open problems in Section 6.

2. Preliminaries.

We now define several processes that will be used in later sections. Let $\{\phi^k(1), \phi^k(2), \dots\}$ be a sequence of i.i.d. *routing vectors* for class k customers. The lth component of the vector $\phi^k(i)$ equals 1 if the ith class k customer next goes to class l, and all other components are zero. Also, define the c-dimensional cumulative sum processes

$$\Phi^k(r) = \phi^k(1) + \cdots + \phi^k(r). \tag{2.1}$$

Finally, let $\mathscr{C}(j)$ be the set of all customer classes k that receive service at station j, that is, $\mathscr{C}(j) = \{k\colon s(k) = j\}$. This set is called the *constituency* of server j in Harrison [12]. We require that $\mathscr{C}(j)$ be nonempty for each $j = 1, \dots, d$.

Next, set C to be the $d \times c$ incidence matrix with components

$$C_{jk} = \begin{cases} 1, & \text{if } k \in \mathscr{C}(j), \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Recall that $E_k(t)$ is the *external* arrival process for class k. Denote by $A_k(t)$ the *total* number of customer visits to class k by time t and by $D_k(t)$ the total number of customer departures from class k by time t. One has as a matter of definition

$$A_k(t) = E_k(t) + \sum_{i=1}^{c} \Phi_k^i(D_i(t)). \tag{2.3}$$

Let $\{v_k(1), v_k(2), \dots\}$ be a sequence of i.i.d. service times associated with class k customers and let $V_k(r)$ be the cumulative sum process defined by

$$V_k(r) = v_k(1) + \cdots + v_k(r).$$

We denote by $V(A(t))$ the c-dimensional process whose kth component is given by $V_k(A_k(t))$, and we set

$$L(t) = CV(A(t)). \tag{2.4}$$

Note that $L = \{L(t),\ t \geq 0\}$ is a d-dimensional process; one interprets $L_j(t)$ as the amount of work for server j brought by all those customers who have arrived at station j by time t. The process L_j was referred to as the *immediate work-load input* process for station j by Harrison and Nguyen [14].

Let $Y_j(t)$ be the amount of cumulative idleness experienced by server j up to time t and let $Y(t) = (Y_1(t), \dots, Y_d(t))'$ be the corresponding vector process. (Prime denotes transpose.) We can express the d-dimensional work-load process $W = \{W(t),\ t \geq 0\}$ as follows:

$$W(t) = L(t) - te + Y(t), \tag{2.5}$$

where e is the d-dimensional vector of ones. Clearly, the idleness process $Y_j(\cdot)$ may increase only at times t such that $W_j(t) = 0$; hence,

$$Y_j(t) = -\inf_{0 \le s \le t} \{L_j(s) - s\}, \qquad j = 1, \dots, d. \tag{2.6}$$

To finish our description of the network model, let us define the fundamental matrix

$$Q = (I - P')^{-1} = (I + P + P^2 + \cdots)'. \tag{2.7}$$

The (k, l)th element of Q represents the expected number of visits to class k made by a customer who starts in class l. Let $\alpha = (\alpha_1, \dots, \alpha_c)'$ and define $\lambda = (\lambda_1, \dots, \lambda_c)'$ via

$$\lambda = Q\alpha. \tag{2.8}$$

One interprets λ_k as the long-run average number of customer visits to class k per unit time resulting from external arrivals as well as internal transitions. The total traffic intensity at station j is then defined by

$$\rho_j = \sum_{k \in \mathscr{C}(j)} \lambda_k m_k. \tag{2.9}$$

Let ρ be the vector of traffic intensities at stations $1, \dots, d$. One can express the vector of traffic intensities in matrix form via

$$\rho = CM\lambda, \tag{2.10}$$

where M is the $c \times c$ diagonal matrix with diagonal elements $m_1, \dots, m_c$.

To state our convergence result rigorously, we need to introduce the path space $D^c[0, \infty)$, which is the space all functions f: $[0, \infty) \to \mathbb{R}^c$ that are right continuous on $[0, \infty)$ and have finite left limits on $(0, \infty)$. The path space $D^c[0, \infty)$ is endowed with the Skorohod topology; see Billingsley [2]. For a sequence $\{X^n\}$ of $D^c[0, \infty)$-valued stochastic processes and $X \in D^c[0, \infty)$, we write $X^n(\cdot) \Rightarrow X(\cdot)$ if X^n converges to X in distribution.

For a function f: $[0, \infty) \to \mathbb{R}$ and $t \ge 0$, put

$$\|f\|_t \equiv \sup_{0 \le s \le t} |f(s)|,$$

and for a vector of functions $f = (f_1, \dots, f_k)'$: $[0, \infty) \to \mathbb{R}^k$ and $t \ge 0$, put

$$\|f\|_t = (\|f_1\|_t, \dots, \|f_k\|_t)'.$$

A sequence $\{f^n\}$ of functions f^n: $[0, \infty) \to \mathbb{R}^k$ is said to converge *uniformly on compact sets* (u.o.c.) to f: $[0, \infty) \to \mathbb{R}^k$ if for each $t \ge 0$, $\|f^n - f\|_t \to 0$ as $n \to \infty$. For a sequence $\{X^n\}$ of $D^c[0, \infty)$-valued stochastic processes and $X \in D^c[0, \infty)$ defined on a probability space, we write $X^n(\cdot) \to X(\cdot)$ u.o.c. if almost surely, X^n converges to X uniformly on compact sets.

3. Conjecture and the main theorem. In order to rigorously state a heavy traffic limit theorem, we need to consider a "sequence of networks" indexed by n. Our setup here follows closely that of Harrison and Nguyen [14]. Let α^n and m^n be vectors of interarrival rates and mean service times, respectively, associated with the nth network in the sequence. We may assume without loss of generality, however, that the routing matrix and the squared coefficients of variations for interarrival times and service times remain fixed across the sequence of networks. Let ρ^n be the vector of traffic intensities for the nth network defined similarly to (2.9). We are interested in a sequence of networks such that for some vectors α, m and β, as $n \to \infty$,

$$\alpha^n \to \alpha, \qquad m^n \to m > 0 \tag{3.1}$$

and

$$\sqrt{n}\,(\rho^n - e) \to \beta, \tag{3.2}$$

where e, as before, is the d-dimensional vector of ones. Condition (3.2) requires that $\rho_j^n \to 1$ at an appropriate rate and is known as the *heavy traffic condition.* As $n \to \infty$, we are interested in the limit of the normalized work-load process $\tilde{W}^n$ defined by

$$\tilde{W}^n(t) = \frac{1}{\sqrt{n}} W^n(nt), \qquad t \geq 0. \tag{3.3}$$

Before we state the conjecture, let us define some more normalized processes. For each $t \geq 0$ and $n \geq 1$, set

$$\begin{aligned}
\tilde{E}^n(t) &= \frac{1}{\sqrt{n}}(E^n(nt) - \alpha^n nt),\\
\tilde{V}^n(t) &= \frac{1}{\sqrt{n}}(V^n([nt]) - m^n nt),\\
\tilde{\Phi}_k^{i,n}(t) &= \frac{1}{\sqrt{n}}(\Phi_k^i([nt]) - P_{ik}nt), \qquad i = 1,\dots,c,\ k = 1,\dots,c,
\end{aligned}$$

where $[x]$ is the integer part of x. (Again, note that the processes Φ^i do not change with n.) It follows from the classical Donsker theorem that as $n \to \infty$,

$$\tilde{E}^n \Rightarrow \xi^a, \tag{3.4}$$

$$\tilde{V}^n \Rightarrow \xi^s, \tag{3.5}$$

$$\tilde{\Phi}^{i,n} \Rightarrow \xi^i, \qquad i = 1,\dots,c, \tag{3.6}$$

where ξ^a, ξ^s and ξ^i $(i = 1,\dots,c)$ are $(c+2)$ independent c-dimensional zero-drift Brownian motions with covariance matrices Γ^a, Γ^s and Γ^i $(i = 1,\dots,c)$, respectively. It is easily verified that $\Gamma^a = \operatorname{diag}(\alpha_1 c_{a,1}^2,\dots,\alpha_c c_{a,c}^2)$, $\Gamma^s = \operatorname{diag}(m_1^2 c_{s,1}^2,\dots,m_c^2 c_{s,c}^2)$ and Γ^i is a matrix defined by

$$\Gamma_{kl}^i = \begin{cases} P_{ik}(1 - P_{ik}), & \text{if } k = l,\\ -P_{ik}P_{il}, & \text{if } k \neq l.\end{cases}$$

CONJECTURE 1. Under the heavy traffic conditions (3.1)–(3.2), the sequence of normalized work-load processes $\{\tilde{W}^n,\ n \geq 1\}$ defined in (3.3) converges in distribution to a continuous process $W^* = \{W^*(t), t \geq 0\}$ as $n \to \infty$. That is,

$$(3.7) \qquad \tilde{W}^n(\cdot) \equiv \frac{1}{\sqrt{n}} W^n(n\,\cdot) \Rightarrow W^*(\cdot) \quad \text{as } n \to \infty.$$

THEOREM 3.1. *There exist multiclass open queueing networks for which the sequence of normalized work-load processes* $\{\tilde{W}^n, n \geq 1\}$ *does not converge in distribution to any continuous limit. In particular*, *Conjecture* 1 *is false.*

The key to the proof of Theorem 3.1 is the "pseudo" heavy traffic result stated in Theorem 4.1 and proved in the next section together with the Dai–Wang example [9]. We leave the proof of Theorem 3.1 to Section 5.

COROLLARY 3.1. *There exist multiclass open queueing networks for which the sequence of normalized work-load processes* $\{\tilde{W}^n, n \geq 1\}$ *is not D-tight.*

The definition of tightness is given, for example, in Section 3.2 of Ethier and Kurtz [10]. The proof of the corollary is given at the end of Section 5.

4. A psuedo heavy traffic limit theorem.

Set

$$G = CMQP'\Lambda C',$$

where $M = \operatorname{diag}(m_1, \dots, m_c)$, $\Lambda = \operatorname{diag}(\lambda)$ and $\operatorname{diag}(\lambda)$ is the diagonal matrix with diagonal elements $\lambda_1, \dots, \lambda_c$. Recall that $Y_j^n(t)$ is the cumulative idleness of server j by time t for the nth system and $Y^n(t)$ is the d-dimensional vector with components $Y_1^n, \dots, Y_d^n$. Set $\tilde{Y}^n(t) = n^{-1/2} Y^n(nt)$.

Because Brownian motions are continuous and $\tilde{E}^n$, $\tilde{V}^n$ and $\tilde{\Phi}^{i,n}$ ($i = 1, \dots, c$) are independent, we can and will assume by the Skorohod representation theorem that the convergence in (3.4)–(3.6) holds u.o.c. Similarly, henceforth whenever we invoke Conjecture 1, we will also assume by the Skorohod representation theorem that the convergence in (3.7) holds u.o.c. With this approach, our exposition becomes considerably cleaner.

THEOREM 4.1. *Assume Conjecture* 1 *is true*, *namely*, *that the convergence in* (3.7) *holds. Then the sequence of normalized idleness processes* $\{\tilde{Y}^n, n \geq 1\}$ *converges to* Y^* *u.o.c. and the limiting processes* (W^*, Y^*) *must satisfy the following statements*:

$$(4.1) \qquad (I + G)W^*(t) = C\xi^s(\lambda t) + CMQ\left(\xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t)\right) + \beta t + Y^*(t),$$

$$(4.2) \qquad W^*(t) \geq 0,$$

(4.3) $\quad Y^*(0) = 0$, Y^* *is continuous and nondecreasing*,

(4.4) $\quad Y_j^*(\cdot)$ *increases only at times* t *such that* $W_j^*(t) = 0$, $j = 1, \dots, d$.

REMARK. Theorem 4.1 states that if (3.7) is true, then the Brownian model proposed by Harrison and Nguyen [13, 14] is the correct model. In fact, Harrison and Nguyen summarily referred to this result (namely, Theorem 4.1) in Section 5 of [14]. We offer a complete proof in this paper and consequently use this theorem to prove Theorem 3.1 in Section 5.

Turning to the proof of Theorem 4.1, we begin by introducing some important notation. For $j = 1, \dots, d$ and $t \geq 0$, define $\tau_j^n(t)$ to be the arrival time at station j of the customer currently being serviced there if $W_j^n(t) > 0$, and to be t if $W_j^n(t) = 0$. Let $\tau^n(t)$ be the d-dimensional vector defined in the obvious manner. This definition of $\tau^n(t)$, which is slightly different from what was given in Peterson ([23], page 103), enables us to give a concise proof of Lemma 4.2. With $\tau_j^n(t)$, one can verify that the number of class k customers who have departed from station $j = s(k)$ by time t is given by

$$D_k^n(t) = \begin{cases} A_k^n(\tau_j^n(t)) - 1, & \text{if server } j \text{ is currently serving a class } k \text{ customer,} \\ A_k^n(\tau_j^n(t)), & \text{otherwise.} \end{cases} \tag{4.5}$$

We will use $A^n(\tau^n(t))$ to denote the c-dimensional process whose kth component is $A_k^n(\tau_{s(k)}^n(t))$. For each $t \geq 0$ and $n \geq 1$, define

$$\bar{\tau}^n(t) = \frac{1}{n}\tau^n(nt), \qquad \tilde{\tau}^n(t) = \frac{1}{\sqrt{n}}(ent - \tau^n(nt))$$

and

$$\bar{A}^n(t) = \frac{1}{n}A^n(nt), \qquad \tilde{A}^n(t) = \frac{1}{\sqrt{n}}(A^n(nt) - \lambda^n nt),$$

where λ^n is defined similarly to (2.8). The following Lemmas 4.1 and 4.2 hold in general without the assumption of convergence in (3.7).

LEMMA 4.1. *For almost every sample path ω and each $t \geq 0$, there exists $\kappa = \kappa(\omega, t)$ independent of n such that*

$$\|\bar{A}_k^n(\omega, \cdot)\|_t \leq \kappa, \qquad k = 1, \dots, c, \ n \geq 1.$$

PROOF. Let $S_k^n = \{S_k^n(t),\ t \geq 0\}$ be the renewal process associated with class k service times. That is, for any $t \geq 0$,

$$S_k^n(t) = \max\{l \geq 0 : V_k^n(l) \leq t\}.$$

Let $T_k^n(t)$ be the cumulative time that server $s(k)$ has devoted to class k customers in the interval $[0, t]$. Then, the number of class k customers who have departed from station $s(k)$ by time t is $D_k^n(t) = S_k^n(T_k^n(t)) \leq S_k^n(t)$. Therefore, from (2.3), we have

$$A^n(t) = E^n(t) + \sum_{k=1}^{c} \Phi^k(S_k^n(T_k^n(t))) \leq E^n(t) + \sum_{k=1}^{c} \Phi^k(S_k^n(t)).$$

(Inequalities between vectors here and in the sequel are inequalities between corresponding coordinates.) The lemma then follows from the functional strong law of large numbers for random walks and renewal processes. □

From the definition of $\tau_j^n(t)$, it follows that

$$t = \tau_j^n(t) + W_j^n\big(\tau_j^n(t)\big) - \epsilon_j^n(t), \tag{4.6}$$

where $\epsilon_j^n(t)$ is 0 if $W_j^n(t) = 0$, and otherwise is equal to the remaining service time of the customer currently occupying server j. Define $\hat{\tau}_j^n(t) = t - \tau_j^n(t)$ and note that

$$W_j^n\big(\tau_j^n(t)\big) - \epsilon_j^n(t) = \hat{\tau}_j^n(t) \le W_j^n\big(\tau_j^n(t)\big).$$

The next three results show that under the heavy traffic scaling, the processes $\tilde{\tau}^n(t)$ and $\tilde{W}^n(t)$ are close for large n. We begin with the following lemma, which proves that $\epsilon_j^n(t)$ is negligible under the heavy traffic scaling.

Lemma 4.2. *For $j = 1, \dots, d$,*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \epsilon_j^n(nt) \to 0, \quad u.o.c. \ as\ n \to \infty.$$

Proof. It follows from the definition of $\epsilon_j^n(t)$ that

$$0 \le \epsilon_j^n(t) \le \max_{k \in \mathscr{C}(j)} \max_{1 \le i \le A_k^n(t)} v_k^n(i),$$

where $\{v_k^n(1), v_k^n(2), \dots\}$ is the sequence of i.i.d. service times for class k customers. An application of Lemma 3.3 from Iglehart and Whitt [18] yields

$$\left\| \frac{1}{\sqrt{n}} \epsilon_j^n(n\,\cdot) \right\|_t \to 0, \quad \text{a.s. as } n \to \infty. \qquad \square$$

Lemma 4.3. *Suppose the convergence in* (3.7) *holds. Then*

$$\bar{\tau}^n(t) \to et \quad u.o.c. \ as\ n \to \infty,$$

where e is the d-dimensional vector of ones.

Proof. Let $\overline{W}_j^n(t) = (1/n)W^n(nt)$. Then,

$$\overline{W}_j^n\big(\bar{\tau}_j^n(t)\big) - \frac{1}{n}\epsilon_j^n(nt) = \frac{1}{n}\hat{\tau}_j^n(nt) \le \overline{W}_j^n\big(\bar{\tau}_j^n(t)\big).$$

Because $\bar{\tau}^n(s) \le s$ for $s \ge 0$,

$$\left\| \frac{1}{n}\hat{\tau}_j^n(n\,\cdot) \right\|_t \le \frac{1}{\sqrt{n}} \left\| \frac{1}{\sqrt{n}} W_j^n(n\,\cdot) \right\|_t + \frac{1}{\sqrt{n}} \left\| \frac{1}{\sqrt{n}} \epsilon_j^n(n\,\cdot) \right\|_t.$$

With the assumption of (3.7) and Lemma 4.2, the lemma is proved. □

Lemma 4.4. *Suppose the convergence in* (3.7) *holds. Then,*

$$\tilde{\tau}^n(t) \to W^*(t) \quad u.o.c. \ as\ n \to \infty.$$

PROOF. Because

$$W_j^n\big(\tau_j^n(t)\big) - \epsilon_j^n(t) = t - \tau_j^n(t) \le W_j^n\big(\tau_j^n(t)\big),$$

we have

$$\tilde{W}_j^n\big(\bar{\tau}_j^n(t)\big) - \frac{1}{\sqrt{n}}\epsilon_j^n(nt) = \tilde{\tau}_j^n(t) \le \tilde{W}_j^n\big(\bar{\tau}_j^n(t)\big).$$

The lemma follows immediately from assumption (3.7) and Lemmas 4.2 and 4.3. □

LEMMA 4.5. *Suppose the convergence in* (3.7) *holds. Then*

$$\bar{A}^n(t) \to \lambda t \quad \textit{u.o.c. and} \quad \bar{D}^n(t) \to \lambda t \quad \textit{u.o.c. as } n \to \infty,$$

where $\bar{D}^n(t) = (1/n)D^n(nt)$.

PROOF. It follows from (2.3) that

$$\bar{A}^n(t) = \bar{E}^n(t) + \sum_{k=1}^{c} \bar{\Phi}^{k,n}\big(\bar{D}_k^n(t)\big),$$

where $\bar{E}^n(t) = (1/n)E^n(nt)$ and $\bar{\Phi}^{k,n}(t) = (1/n)\Phi^k([nt])$ for $k = 1,\dots,c$. Therefore,

$$\begin{aligned}\bar{A}^n(t) - \lambda t = \bar{E}^n(t) - \alpha t &+ \sum_{k=1}^{c}\Big(\bar{\Phi}^{k,n}\big(\bar{D}_k^n(t)\big) - P_k'\bar{D}_k^n(t)\Big)\\ &+ P'\big(\bar{D}^n(t) - \Lambda C'\bar{\tau}^n(t)\big) - P'\Lambda C'\big(te - \bar{\tau}^n(t)\big),\end{aligned} \tag{4.7}$$

where we have used the fact that

$$\lambda = \alpha + P'\lambda$$

and P_k denotes the kth row of P. Using (4.5), we have

$$|\bar{A}^n(\bar{\tau}^n(t)) - \bar{D}^n(t)| \le \frac{1}{n}.$$

Therefore, we can replace $\bar{D}^n(t)$ in the third expression on the right side of (4.7) by $\bar{A}^n(\bar{\tau}^n(t))$ when n is large. Hence, by Lemmas 4.1 and 4.3 and a functional strong law of large numbers, we have almost surely,

$$\begin{aligned}\limsup_{n\to\infty}\|\bar{A}^n(\cdot) - \lambda\cdot\|_t &\le \limsup_{n\to\infty} P'\|\bar{A}^n(\bar{\tau}^n(\cdot)) - \Lambda C'\bar{\tau}^n(\cdot)\|_t\\ &\le P'\limsup_{n\to\infty}\|\bar{A}^n(\cdot) - \lambda\cdot\|_t.\end{aligned}$$

Because all entries of $(I - P')^{-1}$ are negative,

$$\limsup_{n\to\infty}\|\bar{A}^n(\cdot) - \lambda\cdot\|_t \le 0 \quad \text{a.s.}$$

and hence

$$\lim_{n\to\infty}\|\bar{A}^n(\cdot) - \lambda\cdot\|_t = 0 \quad \text{a.s.}$$

The second part of this lemma follows from the first part, (4.5) and Lemma 4.3. □

LEMMA 4.6. *Suppose the convergence in* (3.7) *holds. Define for each* $t \geq 0$,

$$\eta(t) = Q\left(\xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t) - P'\Lambda C'W^*(t) \right).$$

Then

$$\tilde{A}^n(t) \to \eta(t) \quad u.o.c. \text{ as } n \to \infty.$$

PROOF. Let $\Lambda^n = \operatorname{diag}(\lambda^n)$. First, note that

$$\tilde{A}^n(t) = \tilde{E}^n(t) + \sum_{k=1}^{c} \tilde{\Phi}^{k,n}\big(\bar{D}_k^n(t)\big) + P'\tilde{A}^n(\bar{\tau}^n(t))$$
$$- P'\Lambda^n C'\tilde{\tau}^n(t) + P'\sqrt{n}\,\big(\bar{D}^n(t) - \bar{A}^n(\bar{\tau}^n(t))\big)$$

and

$$\eta(t) = \xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t) + P'\eta(t) - P'\Lambda C'W^*(t).$$

Thus,

$$\tilde{A}^n(t) - \eta(t) = \tilde{E}^n(t) - \xi^a(t) + \sum_{k=1}^{c} \big(\tilde{\Phi}^{k,n}\big(\bar{D}_k^n(t)\big) - \xi^k(\lambda_k t)\big)$$
$$+ P'\big(\tilde{A}^n(\bar{\tau}^n(t)) - \eta(\bar{\tau}^n(t))\big) + P'(\eta(\bar{\tau}^n(t)) - \eta(t))$$
$$- P'(\Lambda^n C'\tilde{\tau}^n(t) - \Lambda C'W^*(t)) + P'\sqrt{n}\,\big(\bar{D}^n(t) - \bar{A}^n(\bar{\tau}^n(t))\big).$$

Hence

$$\begin{aligned} \|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t &\leq \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t + \sum_{k=1}^{c} \|\tilde{\Phi}^{k,n}\big(\bar{D}_k^n(\cdot)\big) - \xi^k(\lambda_k\,\cdot)\|_t \\ &\quad + P'\|\tilde{A}^n(\bar{\tau}^n(\cdot)) - \eta(\bar{\tau}^n(\cdot))\|_t + P'\|\eta(\bar{\tau}^n(\cdot)) - \eta(\cdot)\|_t \\ &\quad + P'\|\Lambda^n C'\tilde{\tau}^n(\cdot) - \Lambda C'W^*(\cdot)\|_t + P'\tilde{e}\frac{1}{\sqrt{n}}, \end{aligned} \tag{4.8}$$

where $\tilde{e}$ is the c-dimensional vector of ones. Because $\bar{\tau}_j^n(s) \leq s$ for all $s \geq 0$ and $j = 1, \dots, d$, we have

$$\|\tilde{A}^n(\bar{\tau}^n(\cdot)) - \eta(\bar{\tau}^n(\cdot))\|_t \leq \|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t.$$

Therefore, it follows from (4.8) that

$$\begin{aligned} &(I - P')\|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t \\ &\quad \leq \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t + \sum_{k=1}^{c} \|\tilde{\Phi}^{k,n}\big(\bar{D}_k^n(\cdot)\big) - \xi^k(\lambda_k t)\|_t \\ &\qquad + P'\|\eta(\bar{\tau}^n(\cdot)) - \eta(\cdot)\|_t + P'\|\Lambda^n C'\tilde{\tau}^n(\cdot) - \Lambda C'W^*(\cdot)\|_t + P'\tilde{e}\frac{1}{\sqrt{n}} \\ &\quad \equiv \zeta^n(t). \end{aligned} \tag{4.9}$$

Again, note that Q. Premultiplying both sides of (4.9) by Q, we have

$$\|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t \le Q\zeta^n(t).$$

By (3.4) and (3.6), Lemmas 4.4 and 4.5, the continuity of processes ξ^k and η, $\zeta^n(t) \to 0$ a.s. as $n \to \infty$. Hence we have proved Lemma 4.6. □

Proof of Theorem 4.1. To prove Theorem 4.1, observe that from (2.4) and (2.5),

$$\tilde{W}^n(t) = C\tilde{V}^n(\bar{A}^n(t)) + CM^n\tilde{A}^n(t) + \sqrt{n}(\rho^n - e)t + \tilde{Y}^n(t).$$

By Lemmas 4.5 and 4.6 and assumptions (3.5) and (3.2), we have

$$\begin{aligned} &C\tilde{V}^n(\bar{A}^n(t)) + CM^n\tilde{A}^n(t) + \sqrt{n}(\rho^n - e)t \\ &\quad \to C\xi^s(\lambda t) + CMQ\left(\xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t) - P'\Lambda C'W^*(t)\right) + \beta t \end{aligned}$$

u.o.c. as $n \to \infty$. Because the mapping defined in (2.6) is continuous, Theorem 4.1 follows immediately from the continuous mapping theorem. □

Remark. Because a Brownian motion is almost surely not a process with bounded variation, the matrix $I + G$ must be nonsingular in order for a solution of the system (4.1)–(4.4) to exist. (See the argument given in the next section.) Multiplying both sides of (4.1) by $R \equiv (I + G)^{-1}$, one has

$$W^*(t) = RC\xi^s(\lambda t) + RCMQ\left(\xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t)\right) + R\beta t + RY^*(t).$$

For each $t \ge 0$, set

$$X^*(t) = RC\xi^s(\lambda t) + RCMQ\left(\xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t)\right) + R\beta t.$$

Then X^* is a Brownian motion with drift vector $\theta \equiv R\beta$ and covariance matrix

$$\Gamma^* = RC\left[\Gamma^s\Lambda + MQ\left(\Gamma^a + \sum_{k=1}^{c} \lambda_k\Gamma^k\right)Q'M'\right]C'R'.$$

If one can show that $\{X^*(t) - \theta t,\ t \ge 0\}$ is a martingale with respect to the filtration generated by (W^*, Y^*), one recognizes that W^* is a semimartingale reflecting Brownian motion (SRBM) starting from zero, defined by

$$W^*(t) = X^*(t) + RY^*(t), \qquad t \ge 0, \tag{4.10}$$

and (4.2)–(4.4) with covariance matrix Γ^*, drift vector θ and reflection matrix R; see [27] for the definition of an SRBM. Reiman and Williams [26] proved that if an SRBM exists, the matrix R must be completely-$\mathscr{S}$. In particular, the diagonal elements of R are positive. Conversely, Taylor and Williams [27] proved that if R is a completely-$\mathscr{S}$ matrix, then the corresponding SRBM exists and is unique in law.

5. Proof of Theorem 3.1. Now we present the example of Dai and Wang [9] to show the limiting process W^* in Theorem 4.1 does not exist for certain networks.

Consider the two-station network pictured in Figure 1. Customers arrive at station 1 according to a Poisson process with rate α_1^n. (The index n indicates the nth system.) Each customer makes five stops before departing from the network, and the stations visited are in the following order: $1, 1, 2, 2, 1$. As explained in Section 1, we designate those customers in their kth stop as class k customers. The service times for class k customers are assumed to be exponentially distributed with mean m_k ($k = 1, \dots, 5$), independent of n.

Choose $m = (1/10, 1/10, 22/27, 5/27, 8/10)'$ and

$$\alpha_1^n = \left(1 - \frac{1}{\sqrt{n}}\right).$$

Then $\alpha_1^n \to 1$, $\rho^n < e$, for each n and

$$\lim_{n \to \infty} \sqrt{n}\,(\rho^n - e) = (-1, -1)'.$$

For the foregoing specific data, one can check that $\det(I + G) = 0$; thus $I + G$ is singular. Therefore, there exists a vector $u \neq 0$ such that $u'(I + G) = 0$.

Now, assume that conjecture (3.7) is true. By Theorem 4.1, the normalized work-load process and idleness process $(\tilde{W}^n, \tilde{Y}^n)$ converge to the limiting processes (W^*, Y^*) u.o.c., where

$$(I + G)W^*(t) = \xi(t) + \beta t + Y^*(t) \tag{5.1}$$

and

$$\xi(t) = C\xi^s(\lambda t) + CMQ\left(\xi^a(t) + \sum_{k=1}^{c} \xi^k(\lambda_k t)\right)$$

is a Brownian motion with zero drift and covariance matrix

$$\Gamma \equiv C\Gamma^s\Lambda C' + CMQ\left(\Gamma^a + \sum_{k=1}^{c} \lambda_k \Gamma^k\right)Q'MC'.$$

Multiplying both sides of (5.1) by u', we get

$$u'\xi(t) = -u'\beta t - u'Y(t) \quad \text{for all } t \geq 0. \tag{5.2}$$

It is easy to check that Γ is a positive definite matrix and hence $u'\xi$ is a zero drift Brownian motion with variance $u'\Gamma u > 0$. Note that the right side of (5.2) is a process of bounded variation, while a Brownian motion is almost surely not a process of bounded variation. Therefore, the conjecture cannot possibly hold, and Theorem 3.1 is proved. □

REMARK. If one takes

$$m = (1/10, 1/10, 23/27, 4/27, 8/10)',$$

the reflection matrix $R = (I + G)^{-1}$ becomes

$$R = \begin{pmatrix} -310/27 & 16 \\ 20 & -27 \end{pmatrix}.$$

Because the diagonal elements of R are negative, R is not a completely-$\mathscr{S}$ matrix. Hence by [26], there is no SRBM W^* associated with the corresponding reflection matrix R. Therefore, $\tilde{W}^n$ cannot converge to an SRBM in this case. □

PROOF OF COROLLARY 3.1. For $x \in D_{\mathbb{R}^d}[0, \infty)$, define

$$J(x) = \int_0^\infty e^{-u}[J(x, u) \wedge 1]\, du,$$

where

$$J(x, u) = \sup_{0 \le t \le u} \sum_{i=1}^{d} |x_i(t) - x_i(t-)|.$$

It follows from Lemma 4.2 that

$$J(\tilde{W}^n) \to 0,$$

almost surely as $n \to \infty$. Therefore, by Theorem 3.10.2 of [10], any limit $W^* = \{W^*(t), t \ge 0\}$ of a convergent subsequence of $\{\tilde{W}^n(\cdot), n \ge 1\}$ (under the Skorohod topology) is continuous. From the proof of Theorem 3.1, we know that such a process W^* does not exist. Therefore, $\{\tilde{W}^n(\cdot), n \ge 1\}$ cannot be D-tight. □

6. Concluding remarks and open problems. In this paper, we have proved that conventional heavy traffic limit theorems do not hold for general multiclass open queueing networks. To identify a maximal subset of multiclass networks such that the corresponding heavy traffic limit theorems prevail seems to be a formidable task for the moment. We conjecture that when there is a single service time distribution associated with each server, the convergence in (3.7) holds.

In his example, Whitt [28] demonstrated that the nonconvergence of the normalized work-load process may be caused by large fluctuations of the work load. In [28], these large fluctuations occur because batches of customers with short service times build up in the queues. One way to avoid such fluctuation is to employ some kind of processor sharing discipline (like head-of-the-line processor sharing) among the customer classes at each station. It is worthwhile to investigate the heavy traffic behavior for multiclass queueing networks under non-FIFO queueing disciplines. Research in this direction is just beginning; see the Appendix of [17].

Acknowledgment. Ruth Williams' comment on previous drafts have greatly improved the clarity and content of this paper. We gratefully acknowledge her assistance in our work.

REFERENCES

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

[2] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[3] Chen, H. and Mandelbaum, A. (1991). Stochastic discrete flow networks: Diffusion approximation and bottlenecks. *Ann. Probab.* **19** 1463–1519.

[4] Chen, H. and Mandelbaum, A. (1992). Hierarchical modeling of stochastic networks II: Strong approximations. Preprint.

[5] Chen, H. and Shanthikumar, J. G. (1992). Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. Preprint.

[6] Dai, J. G. and Harrison, J. M. (1992). Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis. *Ann. Appl. Probab.* **2** 65–86.

[7] Dai, J. G. and Kurtz, T. G. (1992). Characterization of the stationary distribution for a semimartingale reflecting Brownian motion in a convex polyhedron. Preprint.

[8] Dai, J. G. and Kurtz, T. G. (1993). A multiclass station with Markovian feedback in heavy traffic. Preprint.

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

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

[11] Glynn, P. W. and Whitt, W. (1991). Departures from many queues in series. *Ann. Appl. Probab.* **1** 546–572.

[12] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Proceedings of the IMA Workshop on Stochastic Differential Systems*. Springer, New York.

[13] Harrison, J. M. and Nguyen, V. (1990). The QNET method for two-moment analysis of open queueing networks. *Queueing Systems Theory Appl.* **6** 1–32.

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

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

[16] Harrison, J. M. and Williams, R. J. (1987). Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* **22** 77–115.

[17] Harrison, J. M. and Williams, R. J. (1992). Brownian models of feedforward queueing networks: Quasireversibility and product form solutions. *Ann. Appl. Probab.* **2** 263–293.

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

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

[20] Jackson, J. R. (1957). Networks of waiting lines. *Oper. Res.* **5** 518–521.

[21] Johnson, D. P. (1983). Diffusion approximations for optimal filtering of jump processes and for queueing networks. Ph.D. thesis, Univ. Wisconsin.

[22] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, New York.

[23] Peterson, W. P. (1991). A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Oper. Res.* **16** 90–118.

[24] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* **8** 441–458.

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

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

[27] Taylor, L. M. and Williams, R.J. (1993). Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probab. Theory Related Fields.* To appear.

[28] Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Manag. Sci.* To appear.

School of Mathematics and
Industrial / Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139