MATHEMATICS OF OPERATIONS RESEARCH
Vol. 20, No. 3, August 1995
Printed in U.S.A.

A MULTICLASS STATION WITH MARKOVIAN FEEDBACK IN HEAVY TRAFFIC

J. G. DAI AND THOMAS G. KURTZ

This paper proves a heavy traffic limit theorem for a multiclass service station with Markovian feedback. This result generalizes the one proved by Reiman (1988). Our approach also significantly simplifies Reiman's original proof. Numerical examples are presented to illustrate the effectiveness of the QNET method which is rooted in the theorem.

1. Introduction. We consider a multiclass single server station. There are c classes of customers, and each class k has its own exogenous arrival process $E_k = \{E_k(t),\ t \geqslant 0\}$ (possibly null), where $E_k(t)$ is the number of class k customers who arrive at the network by time t. For each customer class $k = 1, \ldots, c$, it is assumed that $E_k(0) = 0$. We denote by E the c-dimensional process with components $E_1, \ldots, E_c$. (All vectors are envisioned as column vectors.) The service time for the ith class k customer is $v_k(i)$. Let $V_k(n) = v_k(1) + \cdots + v_k(n)$. We call $\{V_k(n),\ n \geqslant 1\}$ the class k service process. Upon completion of service at the station, 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. To be more precise for the last statement, let $\phi^k(n)$ be the routing vector for the nth class k customer who finishes service at the station. The lth component of $\phi^k(n)$ is one if this customer becomes a class l customer and zero otherwise. Therefore, $\phi^k(n)$ is a c-dimensional "Bernoulli random variable" with parameter P_k', where P_k denotes the kth row of $P = (P_{kl})$ and prime denotes transpose. We assume that $\phi^k = \{\phi^k(n),\ n \geqslant 1\}$ is i.i.d., and $\phi^1, \ldots, \phi^c$ are independent and are independent of the arrival processes and service processes. Such a routing mechanism is often called Bernoulli routing. The transition matrix $P = (P_{kl})$ is taken to be transient. That is,

$$I + P + P^2 + \cdots \text{ is convergent.} \tag{1.1}$$

This condition implies that all customers eventually leave the system. Hence the systems we are considering are open queueing systems. We assume that the waiting buffer at the station has infinite capacity, and that customers are served on a first-in-first-out (FIFO) basis. Hereafter, we will refer to such a system as a *multiclass station.* Our multiclass station model is very general by conventional standards. Besides that inter-arrival time and service time distributions for each customer class can be general, arrival processes and service processes among different classes can be *dependent.* Furthermore, the arrival processes can also be dependent on the service processes. The main assumption we will make is that the $2c$-dimensional vector of

Received October 12, 1993; revised April 6, 1994.
AMS 1991 subject classification. Primary: 60K25; Secondary: 60F17, 60G17, 60J15, 90B22, 68M20.
OR/MS Index 1978 subject classification. Primary: 697 Queues/Networks.
Key words. Multiclass queueing network, heavy traffic, diffusion approximation, reflecting Brownian motion, performance analysis.

0364-765X/95/2003/0721/$01.25

arrival processes and service processes jointly satisfies a *functional central limit theorem*, see (2.6). This assumption is quite mild.

The main result of this paper is to prove a *heavy traffic limit theorem* (Theorem 2.1) which says that the workload process, properly normalized, will converge to a one-dimensional reflecting Brownian motion (RBM) under the heavy traffic condition. Reiman proved a similar heavy traffic limit theorem in Reiman (1988) under the additional assumption that the number of visits to the station by each customer is bounded by some prespecified bound. There are two major contributions in this paper. First, the model considered here is more general than the one considered by Reiman (1988). The major added generality is that we allow *Markovian feedback* among different customer classes. Markovian switching among classes can be used to represent the sort of probabilistic routing that arises from rework, spoilage, and the like. Because there is no prespecified bound on the number of visits to the station by each customer, this generality cannot be handled by Reiman's original proof. Next, the approach used in our proof is new. In many ways it is more systematic. It has the potential to be generalized to certain *multiclass networks* with feedback.

It is known that to prove a heavy traffic limit theorem for a general multiclass network is difficult, see Dai and Wang (1993), Whitt (1993) and Dai and Nguyen (1994) for more elaborate discussions. Indeed, Bramson (1994) has recently shown that the fundamental question of stability for FIFO queueing discipline in a nondeterministic multiclass queueing network has not been resolved; in particular, the traditional definition of heavy traffic in terms of nominal traffic intensities being close to one at each station is not appropriate for all multiclass networks. It is a challenging open problem to determine a suitable class of multiclass networks with feedback for which the heavy traffic limit theorem prevails. Known heavy traffic limit theorems of the type discussed in this paper for open networks were proved by Iglehart and Whitt (1970a, b), Reiman (1984), Johnson (1983), Peterson (1991), Reiman (1988) and Chen and Shanthikumar (1994).

The primary motivation in proving heavy traffic limit theorems is for performance analysis of queueing networks. In the case that a heavy traffic limit theorem prevails for a queueing network, a Brownian system can be used to approximate the queueing network. Such approximations, known as the QNET methods, were proposed in Harrison and Nguyen (1993). Even under the assumption that all arrival processes are independent Poisson processes and class k service times are independent identically distributed (i.i.d.) exponential random variables and some additional independence assumptions, the multiclass station discussed in this paper is not subject to exact mathematical analysis. This is the nature of many multiclass networks with *class dependent* service times. The heavy traffic limit theorem proved in this paper provides a rigorous justification for using a Brownian model to approximate the queueing system. Hence the Brownian model provides a practical tool for performance analysis of a multiclass station. In §5, we present two numerical examples to illustrate the effectiveness of the QNET method for performance analysis of the multiclass station.

Let us introduce notation that will be used in this paper. As mentioned earlier, all vectors are envisioned as column vectors. Prime denotes transpose. We use e to denote the vector of ones, whose dimension will be determined from the context. Unless otherwise stated, vector operations and relations are interpreted componentwise. Therefore for $x = (x_1, \ldots, x_c)'$, $|x| = (|x_1|, \ldots, |x_c|)'$. To rigorously state our convergence result, we need to introduce the path space $D^c[0, \infty)$, which is the space all functions f: $[0, \infty) \to \mathbb{R}^c$ which 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 §3.5 of Ethier and Kurtz (1986). 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 distribu-

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

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

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

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

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 $t \geqslant 0$, $\|f^n - f\|_t \to 0$ as $n \to \infty$. (The symbol zero denotes either the zero vector or the zero scalar, depending on the context.) For a sequence $\{X^n\}$ of $D^c[0,\infty)$-valued stochastic processes and $X \in D^c[0,\infty)$, we write $X^n(\cdot) \to X(\cdot)$ u.o.c. if almost surely, X^n converges to X uniformly on compact sets.

The paper is organized as follows. The main theorem is stated in §2 with preliminaries presented in §3. Section 4 proves the main theorem. The paper concludes with §5, where two numerical examples are presented.

2. The heavy traffic limit theorem. Recall that $\{\phi^k(1), \phi^k(2), \ldots\}$ is 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

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

Let $W(t)$ be the *immediate workload* at time t. It is the amount of time that the server has to work to empty out everyone at the station provided that no more external and internal arrivals to the station are allowed. In order to rigorously state a heavy traffic limit theorem, we need to consider a "sequence of systems" indexed by n. Our setup here follows closely that of Harrison and Nguyen (1993). Let α^n and m^n be two sequences of nonnegative vectors. We interpret α_k^n the external arrival rate for class k customers associated with the nth system. Similarly, $1/m_k^n$ will be the service rate for class k customers. We assume, however, that the routing vector does not depend on n. Because P is transient, $(I - P')$ is invertible and

$$(I - P')^{-1} = \sum_{i=0}^{\infty} (P')^i. \tag{2.2}$$

Let

$$\lambda^n = (I - P')^{-1} \alpha^n. \tag{2.3}$$

We define the traffic intensity for the nth system ρ^n to be

$$\rho^n = \sum_{k=1}^{c} \lambda_k^n m_k^n. \tag{2.4}$$

Before we state the main theorem, we need to define some scaled processes. For

each $t \geqslant 0$ and $n \geqslant 1$, set

$$\tilde{E}^n(t) = \frac{1}{\sqrt{n}}(E^n(nt) - \alpha^n nt),$$

$$\tilde{V}^n(t) = \frac{1}{\sqrt{n}}(V^n([nt]) - m^n nt),$$

$$\tilde{\Phi}_k^{i,n}(t) = \frac{1}{\sqrt{n}}\left(\Phi_k^i([nt]) - P_{ik}nt\right), \qquad i = 1,\ldots,c, k = 1,\ldots,c,$$

$$\tilde{W}^n(t) = \frac{1}{\sqrt{n}}W^n(nt),$$

where $[x]$ is the integer part of x, and $E^n(\cdot)$, $V^n(\cdot)$ and $W^n(\cdot)$ are corresponding processes associated with the nth system. (Again, note that the processes Φ^i do not change with n.) It follows from the classical functional central limit theorem (Donsker's Theorem) that as $n \to \infty$,

$$\tilde{\Phi}^{i,n} \Rightarrow \xi^i, \qquad i = 1,\ldots,c. \tag{2.5}$$

where ξ^i is a c-dimensional zero-drift Brownian motion with covariance matrix Γ^i $(i = 1,\ldots,c)$, and $\xi^1,\ldots,\xi^c$ are independent. It can be verified that

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

We now assume that

(2.6) $\tilde{E}^n$ and $\tilde{V}^n$ satisfy a joint functional central limit theorem.

Because the Brownian motion has continuous sample paths and the Skorohod representation theorem holds, we may assume throughout this paper that the sequences $\{E^n(t), t \geqslant 0\}$ and $\{V^n([t]), t \geqslant 0\}$, and $\{\tilde{\Phi}^{i,n}([t]), t \geqslant 0\}$ can be constructed on a *common* probability space such that as $n \to \infty$,

$$\tilde{E}^n \to \xi^a, \quad \text{u.o.c.} \tag{2.7}$$

$$\tilde{V}^n \to \xi^s, \quad \text{u.o.c.} \tag{2.8}$$

$$\tilde{\Phi}^{i,n} \to \xi^i, \quad \text{u.o.c. } i = 1,\ldots,c, \tag{2.9}$$

where (ξ^a, ξ^s) is a $2c$-dimensional zero-drift Brownian motion with covariance matrix Γ, and it is independent of $(\xi^1,\ldots,\xi^c)$.

THEOREM 2.1. *Assume that* (2.6)–(2.9) *hold. Assume further that as* $n \to \infty$,

$$\lambda^n = (I - P')^{-1}\alpha^n \to \lambda > 0, \tag{2.10}$$

$$m^n \to m > 0, \tag{2.11}$$

$$\sqrt{n}(\rho^n - 1) \to \theta. \tag{2.12}$$

Then the sequence of normalized workload processes

$$\tilde{W}^n(t) \to W^*(t) \equiv \frac{1}{\beta} Z^*(t), \quad u.o.c., \text{ as } n \to \infty, \tag{2.13}$$

where

$$\beta = \left(1 + \sum_{i=1}^{\infty} m'(P')^i \lambda\right), \tag{2.14}$$

$$Z^*(t) = \xi^*(t) + \theta t - \inf_{0 \leqslant s \leqslant t} (\xi^*(s) + \theta s), \tag{2.15}$$

$$\xi^*(t) = e'\xi^s(\lambda t) + m'(I - P')^{-1}\left(\xi^a(t) + \sum_{l=1}^{c} \xi^l(\lambda_l t)\right). \tag{2.16}$$

REMARK 1. Assumptions (2.6)–(2.9) are quite mild. First, if arrival process E^n is a vector of independent renewal processes, V^n is a vector of independent random walks and it is independent of E^n, then by conventional Donsker type functional central limit theorem $(\tilde{E}^n, \tilde{V}^n, \tilde{\Phi}^{1,n}, \ldots, \tilde{\Phi}^{c,n})$ converges weakly to a multidimensional Brownian motion under some additional moment assumptions on E^n and V^n. The Skorohod representation theorem allows all these processes be constructed in one probability space such that the convergence takes place as in (2.7)–(2.9).

REMARK 2. Assumptions (2.10) and (2.11) are quite natural. The key assumption is condition (2.12). This is the so-called *heavy traffic condition*. It not only requires that $\rho^n \to 1$, but also that ρ^n converges to one at the specified rate.

REMARK 3. It can be checked that $\xi^*(t)$ is a one-dimensional Brownian motion with zero drift and variance

$$\begin{aligned} \sigma^2 \equiv & \left(m'(I - P')^{-1}, \sqrt{\lambda'}\right)\Gamma\left(m'(I - P')^{-1}, \sqrt{\lambda'}\right)' \\ & + m'(I - P')^{-1}\left(\sum_{l=1}^{c} \lambda_l \Gamma^l\right)(I - P)^{-1} m, \end{aligned} \tag{2.17}$$

where $\sqrt{\lambda'} = (\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_c})$. It follows that Z^* is a one-dimensional reflecting Brownian motion with drift θ, and variance σ^2, see §5.6 of Harrison (1985). Therefore, W^* is an RBM with drift θ/β and variance σ^2/β^2. As we will see in later sections, Z^* is the heavy traffic limit of the *total workload process* of the multiclass station. This fact was also proved by Iglehart and Whitt (1970b). While it is straightforward to obtain this limit, it is difficult to get the kind of convergence in our theorem.

REMARK 4. The limit $\{W^*(t), t \geqslant 0\}$ of immediate workload processes can be used to estimate performance measures of the multiclass station (see §5). From this theorem, one can also easily establish that normalized class level queue length processes and class level workload processes converge to a constant times $W^*(t)$; see Corollary 4.1. This result was also proved by Reiman (1988). Peterson (1991) proved an analogous result for feedforward networks.

3. Preliminaries. Let us for the moment fix a system in the sequence and temporarily drop the superscript n for notational convenience. First let $Z(t)$ denote the sum of *all* future service times at the station for customers who are present at the station at time t, plus the remaining service time of any customer who may be in

service at the station at time t. If there were no new external arrivals to the station after time t, the $Z(t)$ would represent the total amount of work required from the server to empty out the system. Thus, it is also called the *total workload* at the station. Let $N_k(t)$ be the total number of visits to class k made by those customers who enter the system before time t (regardless of the entering customers original class designation). Assuming initially that the system is empty, we have

$$N_k(t) = E_k(t) + \sum_{l=1}^{c} \Phi_k^l(N_l(t)),$$

or in vector form

$$N(t) = E(t) + \sum_{l=1}^{c} \Phi^l(N_l(t)). \tag{3.1}$$

Let $Y(t)$ be the cumulative idleness time for the server by time t. Then $t - Y(t)$ is the cumulative time that the server has been busy by time t. Hence, we have

$$Z(t) = \sum_{k=1}^{c} V_k(N_k(t)) - t + Y(t),$$

$$Z(t) \geqslant 0,$$

$$Y(\cdot) \text{ is nondecreasing and } Y(0) = 0,$$

$$Y(\cdot) \text{ increases only when } Z(t) = 0.$$

The last assertion holds because FIFO is a *work-conserving* queueing discipline. Define the following centered processes

$$\hat{E}_k(t) = E_k(t) - \alpha_k t, \tag{3.2}$$

$$\hat{V}_k(r) = V_k(r) - m_k r, \tag{3.3}$$

$$\hat{\Phi}_k^i(r) = \Phi_k^i(r) - P_{ik} r, \tag{3.4}$$

$$\hat{N}_k(t) = N_k(t) - \lambda_k t. \tag{3.5}$$

It follows from (3.1) that

$$\hat{N}(t) = \hat{E}(t) + \sum_{l=1}^{c} \hat{\Phi}^l(N_l(t)) + P'\hat{N}(t),$$

or

$$\hat{N}(t) = (I - P')^{-1}\left(\hat{E}(t) + \sum_{l=1}^{c} \hat{\Phi}^l(N_l(t))\right). \tag{3.6}$$

One can rewrite Z as

$$Z(t) = \sum_{k=1}^{c} \left(\hat{V}_k(N_k(t)) + m_k \hat{N}_k(t)\right) + (\rho - 1)t + Y(t)$$

$$\equiv \zeta(t) + (\rho - 1)t + Y(t),$$

where

$$\zeta(t) = \sum_{k=1}^{c} \left(\hat{V}_k(N_k(t)) + m_k \hat{N}_k(t)\right). \tag{3.7}$$

It follows by the usual reflection argument, see §2.2 of Harrison (1985), that

$$Z(t) = \zeta(t) + (\rho - 1)t - \inf_{0 \leqslant s \leqslant t} (\zeta(s) + (\rho - 1)s). \tag{3.8}$$

Recall that $W(t)$ is the immediate workload at the station at time t. Obviously, we have

$$W(t) \leqslant Z(t) \quad \text{for } t \geqslant 0. \tag{3.9}$$

It turns out that proving the heavy traffic limit for Z is relatively easy. However, to prove the heavy traffic limit for W is difficult. In order to make the connection between W and Z, we derive a set of equations for W. Let $A_k(t)$ be the total number of customer visits to class k by time t and

$$\hat{A}_k(t) = A_k(t) - \lambda_k t.$$

Then,

$$W(t) = \sum_{k=1}^{c} V_k(A_k(t)) - t + Y(t) \tag{3.10}$$

$$= \sum_{k=1}^{c} \left(\hat{V}_k(A_k(t)) + m_k \hat{A}_k(t)\right) + (\rho - 1)t + Y(t),$$

$$W(t) \geqslant 0,$$

$$Y(\cdot) \text{ is nondecreasing and } Y(0) = 0,$$

$$Y(\cdot) \text{ increases only when } W(t) = 0.$$

For each $t \geqslant 0$, define $\tau(t)$ to be the arrival time of the customer who has the most recent service completion or the beginning of the most recent idle period, whichever is later. Note that this definition makes $\tau(t)$ monotone. This definition of τ was first introduced by Reiman (1988) and later used by Peterson (1991) and Dai and Nguyen (1994). It is a key quantity linking workload for different classes to the immediate workload under the FIFO queueing discipline. One can first check that $A_k(\tau(t))$ is

the total number of customer departures from class k by time t. One can next check that

$$A_k(t) = E_k(t) + \sum_{i=1}^{c} \Phi_k^i(A_i(\tau(t))). \tag{3.11}$$

Let

$$\hat{\tau}(t) = t - \tau(t). \tag{3.12}$$

Using (2.3) and the vector form of (3.11), we have

$$\hat{A}(t) = \hat{E}(t) + \sum_{l=1}^{c} \hat{\Phi}^l(A_l(\tau(t))) + P'\hat{A}(\tau(t)) - P'\lambda\hat{\tau}(t). \tag{3.13}$$

Iterating (3.13) $k-1$ times, we have

$$\begin{aligned}\hat{A}(t) = {} & \sum_{i=1}^{k} (P')^{i-1}\hat{E}(\tau^{i-1}(t)) + \sum_{i=1}^{k} (P')^{i-1}\left(\sum_{l=1}^{c} \hat{\Phi}^l(A_l(\tau^i(t)))\right) \\ & - \sum_{i=1}^{k} (P')^i \lambda\hat{\tau}(\tau^{i-1}(t)) + (P')^k \hat{A}(\tau^k(t)),\end{aligned} \tag{3.14}$$

where

$$\tau^i(t) = \tau^{i-1}(\tau(t)) \quad \text{for } i \geqslant 1 \text{ and } \tau^0(t) = t. \tag{3.15}$$

Because P is transient and $\tau^k(t) \leqslant t$ for all $k \geqslant 0$, we have

$$\begin{aligned}\hat{A}(t) = {} & \sum_{i=1}^{\infty} (P')^{i-1}\hat{E}(\tau^{i-1}(t)) + \sum_{i=1}^{\infty} (P')^{i-1}\left(\sum_{l=1}^{c} \hat{\Phi}^l(A_l(\tau^i(t)))\right) \\ & - \sum_{i=1}^{\infty} (P')^i \lambda\hat{\tau}(\tau^{i-1}(t)).\end{aligned} \tag{3.16}$$

In the remainder of this section, we derive an alternative expression for $\hat{\tau}(t)$. Let

$$\epsilon_1(t) \equiv \hat{\tau}(t) - W(\tau(t)). \tag{3.17}$$

Because we assume the FIFO discipline, $\epsilon_1(t)$ has the following interpretations. If the server is idle at time t, $\epsilon_1(t)$ is the idle time the server has experienced in the current idle period. When the server is currently serving the first customer in the current busy period, $\epsilon_1(t)$ is equal to the amount of service that the customer has received plus the last idle period. When the server is currently serving a customer who is not the first customer in the current busy period, $\epsilon_1(t)$ is the amount of service that the customer has received. Let

(3.18)

$\epsilon_2(t)$ be the amount of service that the current customer has received or zero if the server is currently in idle.

It follows that we have the following alternative expression for W:

$$\begin{aligned} W(t) &= \sum_{k=1}^{c} \left(V_k(A_k(t)) - V_k(A_k(\tau(t)))\right) - \epsilon_2(t) \\ &= e'\left(\hat{V}(A(t)) - \hat{V}(A(\tau(t))) + m'(A(t) - A(\tau(t))\right) - \epsilon_2(t) \\ &= e'\left(\hat{V}(A(t)) - \hat{V}(A(\tau(t))) + m'\left(\hat{A}(t) - \hat{A}(\tau(t))\right) + \rho\hat{\tau}(t) - \epsilon_2(t)\right. \\ &= \xi(t) - \xi(\tau(t)) + \sum_{i=0}^{\infty} m'(P')^i(I - P')\lambda\hat{\tau}\left(\tau^i(t)\right) - \epsilon_2(t) \end{aligned}$$

where the last equality is based on (3.16) and

$$\xi(t) \equiv e'\hat{V}(A(t)) + \sum_{i=1}^{\infty} m'(P')^{i-1}\hat{E}\left(\tau^{i-1}(t)\right) + \sum_{i=1}^{\infty} m'(P')^{i-1}\left(\sum_{l=1}^{c} \hat{\Phi}^l\left(A_l(\tau^i(t))\right)\right). \tag{3.19}$$

Noting that $(I - P')\lambda = \alpha$, we have

$$W(\tau(t)) = \xi(\tau(t)) - \xi(\tau^2(t)) - \epsilon_2(\tau(t)) + \sum_{i=0}^{\infty} m'(P')^i\alpha\hat{\tau}\left(\tau^{i+1}(t)\right).$$

Using (3.17), we have

$$\hat{\tau}(t) = \epsilon_1(t) + \xi(\tau(t)) - \xi\left(\tau^2(t)\right) - \epsilon_2(\tau(t)) + \sum_{i=0}^{\infty} m'(P')^i\alpha\hat{\tau}\left(\tau^{i+1}(t)\right). \tag{3.20}$$

Put

$$\beta_i = m'(P')^i\alpha, \qquad i = 0, 1, \ldots, \tag{3.21}$$

$$\eta(t) = \left(\hat{\tau}(t), \hat{\tau}(\tau(t)), \hat{\tau}\left(\tau^2(t)\right), \ldots\right)', \tag{3.22}$$

$$e_1 = (1, 0, 0, \ldots)', \tag{3.23}$$

$$\epsilon_3(t) \equiv \epsilon_1(t) + \xi(\tau(t)) - \xi\left(\tau^2(t)\right) - \epsilon_2(\tau(t)) \tag{3.24}$$

and

$$Q = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_i & \cdots \\ 1 & 0 & \cdots & 0 & \cdots \\ 0 & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \cdots \\ \vdots & \vdots & \cdots & \vdots & \ddots \end{pmatrix}. \tag{3.25}$$

Then we can rewrite (3.20) in matrix form,

$$\eta(t) = e_1\epsilon_3(t) + Q\eta(\tau(t)).$$

Iterating this identity k times, we have

$$\eta(t) = \sum_{i=0}^{k} Q^i e_1 \epsilon_3(\tau^i(t)) + Q^{k+1}\eta(\tau^{k+1}(t)). \tag{3.26}$$

4. Proof of the heavy traffic limit theorem. Recall that we are considering a sequence of systems indexed by n. Hence all processes introduced in §3 will have a superscript n. The only exception is that we use $\tau_n(t)$ to denote $\tau(t)$ in the nth system because we have used $\tau^i(t)$ to denote the ith iteration of $\tau(t)$. Let

$$\bar{N}^n(t) \equiv \frac{1}{n}N^n(nt), \qquad \bar{A}^n(t) \equiv \frac{1}{n}A^n(nt), \tag{4.1}$$

$$\bar{\tau}_n(t) \equiv \frac{1}{n}\tau_n(nt), \qquad \tilde{\tau}_n(t) = \frac{1}{\sqrt{n}}\hat{\tau}_n(nt), \tag{4.2}$$

$$\tilde{\xi}^n(t) = \frac{1}{\sqrt{n}}\xi^n(nt), \qquad \tilde{Z}^n(t) = \frac{1}{\sqrt{n}}Z^n(nt), \tag{4.3}$$

$$\tilde{\epsilon}_i^n(t) = \frac{1}{\sqrt{n}}\epsilon_i^n(nt), \qquad \tilde{N}^n(t) = \frac{1}{\sqrt{n}}\hat{N}^n(nt). \tag{4.4}$$

LEMMA 4.1. *For each* $k = 1, \ldots, c$,

$$\bar{N}_k^n(t) \to \lambda_k t \quad u.o.c., \quad as\ n \to \infty.$$

PROOF. Let $\zeta_l^k(i)$ be the number of visits to class l by the ith external arrival to class k. Then for each pair (k, l), $\{\zeta_l^k(i), i \geqslant 1\}$ is i.i.d., and for each l,

$$N_l^n(nt) = \sum_{k=1}^{c} \sum_{i=1}^{E_k^n(nt)} \zeta_l^k(i).$$

Therefore,

$$\lim_{n\to\infty} \frac{1}{n}N_l^n(nt) = \lim_{n\to\infty} \sum_{k=1}^{c} \frac{E_k^n(nt)}{n} \frac{1}{E_k^n(nt)} \sum_{i=1}^{E_k^n(nt)} \zeta_l^k(i) = \sum_{k=1}^{c} \alpha_k t E\left[\zeta_l^k(1)\right] \quad \text{a.s.}$$

Because

$$E\left[\zeta_l^k(1)\right] = \delta_{kl} + \sum_{m=1}^{\infty} P_{kl}^m = \left[(I - P)^{-1}\right]_{kl},$$

we have

$$\lim_{n\to\infty} \frac{1}{n}N_l^n(nt) = \lambda_l t \quad \text{a.s.},$$

where $\delta_{kl} = 1$ if $k = l$ and zero otherwise, and P^m is the mth power of P. Because

$\bar{N}^n(t)$ is a nondecreasing function of t for each n, and λt is a continuous function of t, one can show that $\bar{N}^n(t) \to \lambda t$ u.o.c., see Lemma 4.1 of Dai (1994). □

LEMMA 4.2. *As* $n \to \infty$,

$$\tilde{N}^n(t) \to (I - P')^{-1}\left(\xi^a(t) + \sum_{l=1}^{c} \xi^l(\lambda_l t)\right) \quad u.o.c.$$

PROOF. It follows from Lemma 4.1 that

$$\bar{N}_k^n(t) - \lambda_k^n t \to 0 \quad u.o.c.$$

The lemma then follows from (3.6), (2.7), (2.9) and the continuity of Brownian motions ξ^a and ξ^i, $i = 1, \ldots, c$. □

THEOREM 4.1. *As* $n \to \infty$, $\tilde{Y}^n(t) \to Y^*(t)$ *u.o.c. and* $\tilde{Z}^n$ *converges to the one-dimensional reflecting Brownian motion* Z^*, *where*

$$Y^*(t) = -\inf_{0 \leqslant s \leqslant t} (\xi^*(s) + \theta s) \quad \text{and} \quad Z^*(t) = \xi^*(t) + \theta t + Y^*(t)$$

and $\xi^*(t)$ *is defined in* (2.16).

PROOF. By Lemma 4.2, (3.7) and (2.8), one can check that $\tilde{\zeta}^n(t) = \zeta^n(nt)/\sqrt{n} \to \xi^*(t)$ u.o.c. The proof follows from (2.12) and the continuity of the mapping (3.8). □

LEMMA 4.3. *For each* $t \geqslant 0$ *and for* $i = 1, 2$,

$$\left\| \frac{1}{\sqrt{n}} \epsilon_i^n(n \cdot) \right\|_t \to 0 \quad \text{as } n \to \infty,$$

where ϵ_1 *and* ϵ_2 *are defined in* (3.17) *and* (3.18).

PROOF. The lemma follows from Lemma 3.3 of Iglehart and Whitt (1970a). □

LEMMA 4.4. *There exists a random variable* κ *independent of* n *such that for each sample path and each* $t \geqslant 0$,

$$\|\tilde{\tau}^n(\cdot)\|_t \leqslant \kappa \quad \text{for all } n.$$

PROOF. From (3.17), we have

$$\begin{aligned} \tilde{\tau}_n(s) &= \tilde{W}^n(\bar{\tau}^n(s)) + \tilde{\epsilon}_1^n(s) \\ &\leqslant \tilde{Z}^n(\bar{\tau}_n(s)) + \tilde{\epsilon}_1^n(s). \end{aligned} \tag{4.5}$$

The lemma follows from (4.5), Lemma 4.3 and Theorem 4.1. □

LEMMA 4.5. *For each* $k = 1, \ldots, c$, *as* $n \to \infty$,

$$\bar{A}_k^n(t) \to \lambda_k t, \quad u.o.c.$$

PROOF. First $\bar{A}_k^n(t) - \lambda_k^n t = \hat{A}_k^n(nt)/n$ can be written as the summation of three terms as in (3.16). Note that $\tau_n(t) \leqslant t$. We have the following estimate for the term corresponding to the first term in (3.16). (Recall that for a vector $x = (x_1, \ldots, x_c)'$ we

use $|x|$ to denote $(|x_1|, \ldots, |x_c|)'$.)

$$\sup_{0\leqslant s\leqslant t}\left|\sum_{i=1}^{\infty}(P')^{i-1}\frac{1}{n}\hat{E}^n\big(\tau_n^{i-1}(ns)\big)\right| \leqslant \sup_{0\leqslant s\leqslant t}\sum_{i=1}^{\infty}(P')^{i-1}\frac{1}{n}\left|\hat{E}^n\big(\tau_n^{i-1}(ns)\big)\right|$$

$$\leqslant \sum_{i=1}^{\infty}(P')^{i-1}\frac{1}{n}\left\|\hat{E}^n(n\cdot)\right\|_t = (I-P')^{-1}\frac{1}{n}\left\|\hat{E}^n(n\cdot)\right\|_t.$$

The last expression converges to the zero vector because of (2.7). Similarly, we can prove that the term corresponding to the middle term in (3.16) converges to zero u.o.c. The term corresponding to the last term in (3.16) converges to zero u.o.c. by Lemma 4.4. It follows that $\bar{A}_k^n(t) - \lambda_k^n t \to 0$ u.o.c. □

LEMMA 4.6. *As* $n \to \infty$,

$$\tilde{\xi}^n(t) \to \xi^*(t), \quad u.o.c.,$$

where ξ^* *is a one-dimensional Brownian motion defined in* (2.16), *with drift zero and variance* σ^2 *defined in* (2.17).

PROOF. First from the definition of ξ in (3.19) and (2.2), it follows that for any $s \geqslant 0$,

$$\begin{aligned}\left|e'\big(\tilde{\xi}^n(s) - \xi^*(s)\big)\right| \leqslant{}& \left|e'\big(\tilde{V}^n\big(\bar{A}^n(s)\big) - \xi^s(\lambda s)\big)\right| \\ &+ \sum_{i=1}^{\infty}\left|(m^n)'(P')^{i-1}\tilde{E}^n\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big) - m'(P')^{i-1}\xi^a(s)\right| \\ &+ \sum_{i=1}^{\infty}\left|(m^n)'(P')^{i-1}\left(\sum_{l=1}^{c}\tilde{\Phi}^{l,n}\Big(\bar{A}_l^n\Big(\frac{1}{n}\tau_n^i(ns)\Big)\Big)\right)\right. \\ &\left. - m'(P')^{i-1}\left(\sum_{l=1}^{c}\xi^l(\lambda_l s)\right)\right|.\end{aligned} \tag{4.6}$$

For any $0 \leqslant s \leqslant t$, one has

$$\begin{aligned}&\left|(m^n)'(P')^{i-1}\tilde{E}^n\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big) - m'(P')^{i-1}\xi^a(s)\right| \\ &\leqslant (m^n)'(P')^{i-1}\left|\tilde{E}^n\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big) - \xi^a(s)\right| \\ &\quad + |(m^n - m)'|(P')^{i-1}|\,\xi^a(s)| \\ &\leqslant (m^n)'(P')^{i-1}\left|\tilde{E}^n\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big) - \xi^a\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big)\right| \\ &\quad + (m^n)'(P')^{i-1}\left|\xi^a\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big) - \xi^a(s)\right| + |(m^n - m)'|(P')^{i-1}|\,\xi^a(s)| \\ &\leqslant (m^n)'(P')^{i-1}\left\|\tilde{E}^n(\cdot) - \xi^a(\cdot)\right\|_t \\ &\quad + (m^n)'(P')^{i-1}\left|\xi^a\Big(\frac{1}{n}\tau_n^{i-1}(ns)\Big) - \xi^a(s)\right| \\ &\quad + |(m^n - m)'|(P')^{i-1}|\,\xi^a(s)|.\end{aligned}$$

Therefore, for any $0 \leqslant s \leqslant t$ the second expression in (4.6) is

$$\sum_{i=1}^{\infty} \left| (m^n)'(P')^{i-1} \tilde{E}^n\left(\frac{1}{n}\tau_n^{i-1}(ns)\right) - m'(P')^{i-1}\xi^a(s) \right|$$

$$\leqslant \sum_{i=1}^{\infty} (m^n)'(P')^{i-1} \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t$$

$$+ \sum_{i=1}^{\infty} (m^n)'(P')^{i-1} \left| \xi^a\left(\frac{1}{n}\tau_n^{i-1}(ns)\right) - \xi^a(s) \right|$$

$$+ \sum_{i=1}^{\infty} |(m^n - m)'|(P')^{i-1} \|\xi^a(\cdot)\|_t$$

$$= (m^n)'(I - P')^{-1} \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t$$

$$+ \sum_{i=1}^{k} (m^n)'(P')^{i-1} \left| \xi^a\left(\frac{1}{n}\tau_n^{i-1}(ns)\right) - \xi^a(s) \right|$$

$$+ \sum_{i=k+1}^{\infty} (m^n)'(P')^{i-1} \left| \xi^a\left(\frac{1}{n}\tau_n^{i-1}(ns)\right) - \xi^a(s) \right|$$

$$+ |(m^n - m)'|(I - P')^{-1} \|\xi^a(\cdot)\|_t$$

$$\leqslant (m^n)'(I - P')^{-1} \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t$$

$$+ \sum_{i=1}^{k} (m^n)'(P')^{i-1} \left| \xi^a\left(\frac{1}{n}\tau_n^{i-1}(ns)\right) - \xi^a(s) \right|$$

$$+ (m^n)'(I - P')^{-1}(P')^k 2\|\xi^a(\cdot)\|_t + |(m^n - m)'|(I - P')^{-1} \|\xi^a(\cdot)\|_t.$$

Because $(P')^k \to 0$ as $k \to \infty$, $\xi^a(\cdot)$ is continuous, (2.11), (2.7) and Lemma 4.4 hold, we have

$$\sum_{i=1}^{\infty} \left\| (m^n)'(P')^{i-1} \tilde{E}^n\left(\frac{1}{n}\tau_n^{i-1}(n\cdot)\right) - m'(P')^{i-1}\xi^a(\cdot) \right\|_t \to 0.$$

Similarly, we can prove the first term and the third term converge to zero u.o.c. Hence we have $\tilde{\xi}^n(t) \to \xi^*(t)$ u.o.c. □The following lemma is the most critical to the proof of our main theorem.

LEMMA 4.7. *For each* $t \geqslant 0$,

$$\left\| \frac{1}{\sqrt{n}} \hat{\tau}_n(n\cdot) - \frac{1}{\sqrt{n}} \hat{\tau}_n(\tau_n(n\cdot)) \right\|_t \to 0, \quad \text{as } n \to \infty.$$

PROOF. Let $Q^{(n)}$ be defined as in (3.25) and η^n be defined as in (3.22) for the nth system. From (3.26), we have

$$\frac{1}{\sqrt{n}}\eta^n(nt) = \sum_{i=0}^{k} (Q^{(n)})^i e_1 \frac{1}{\sqrt{n}}\epsilon_3^n(\tau_n^i(nt)) + (Q^{(n)})^{k+1}\frac{1}{\sqrt{n}}\eta^n(\tau_n^{k+1}(nt)). \tag{4.7}$$

Let $\hat{e} = (1, -1, 0, 0, \ldots)'$ and $\tilde{\eta}^n(t) = \eta^n(nt)/\sqrt{n}$. Premultiplying both sides of (4.7) by $\hat{e}'$, we have for each s such that $0 \leqslant s \leqslant t$,

$$|\hat{e}'\tilde{\eta}^n(s)| \leqslant \sum_{i=0}^{k} |\hat{e}'(Q^{(n)})^i|e_1|\,\tilde{\epsilon}_3^n(\tau_n^i(ns)/n)| + |\hat{e}'(Q^{(n)})^{k+1}|\frac{1}{\sqrt{n}}\eta^n(\tau_n^{k+1}(ns)).$$

It follows from Lemmas 4.3, 4.4 and 4.6 that

$$\left\|\frac{1}{\sqrt{n}}\epsilon_3^n(n\cdot)\right\|_t \to 0, \quad \text{as } n \to \infty,$$

where $\epsilon_3^n(t)$ is defined in (3.24). Also, for each $i \geqslant 1$ we have $|\hat{e}(Q^{(n)})^i| \to |\hat{e}Q^i|$ as $n \to \infty$. Therefore,

$$\begin{aligned}
\limsup_{n\to\infty}\left\|\frac{1}{\sqrt{n}}\hat{\tau}_n(n\cdot) - \frac{1}{\sqrt{n}}\hat{\tau}_n(\tau_n(n\cdot))\right\|_t &= \limsup_{n\to\infty}\sup_{0\leqslant s\leqslant t}|\hat{e}'\tilde{\eta}^n(s)| \\
&\leq \sum_{i=0}^{k}|\hat{e}'Q^i|e_1 \limsup_{n\to\infty}\|\tilde{\epsilon}_3^n(\cdot)\|_t + |\hat{e}'Q^{k+1}|\limsup_{n\to\infty}\|\tilde{\eta}^n(\cdot)\|_t \\
&\leq \sum_{i=0}^{k}|\hat{e}'Q^i|e_1 \limsup_{n\to\infty}\|\tilde{\epsilon}_3^n(\cdot)\|_t + |\hat{e}'Q^{k+1}|e\limsup_{n\to\infty}\|\tilde{\tau}^n(\cdot)\|_t \\
&= |\hat{e}'Q^{k+1}|e\limsup_{n\to\infty}\|\tilde{\tau}^n(\cdot)\|_t \\
&= \sum_{i=1}^{\infty}|Q_{1,i}^{k+1} - Q_{2,i}^{k+1}|\limsup_{n\to\infty}\|\tilde{\tau}^n(\cdot)\|_t,
\end{aligned}$$

where e is the column vector of ones. Because k is arbitrary, the next lemma implies that

$$\limsup_{n\to\infty}\left\|\frac{1}{\sqrt{n}}\hat{\tau}_n(n\cdot) - \frac{1}{\sqrt{n}}\hat{\tau}_n(\tau_n(n\cdot))\right\|_t = 0,$$

and hence the lemma is proved. □

LEMMA 4.8. *For the matrix Q defined in* (3.25),

$$\lim_{k\to\infty}\sum_{i=1}^{\infty}|Q_{1,i}^{k+1} - Q_{2,i}^{k+1}| = 0.$$

PROOF. Recall the definition of $\beta_i = m'(P')^i\alpha$. One can check that $\sum_{i=0}^{\infty}\beta_i = 1$ because of (2.12). Hence the corresponding matrix Q is stochastic. Because each component of m is strictly positive by assumption (2.11), one has $\beta_{i+1} = 0$ if $\beta_i = 0$.

Let p be the largest index i such that $\beta_i \neq 0$ (or $= \infty$ if none of the β_i is zero). Assume that $p = \infty$. (For the case that $p < \infty$, we can consider the $p \times p$ stochastic submatrix of Q. The corresponding proof is analogous.) Because $\beta_i > 0$ for each i, the matrix Q is irreducible. Therefore, Q can be considered to be a transition matrix of an irreducible, aperiodic discrete time Markov chain on state space $\{0, 1, \ldots, \}$. Because P is transient, one can readily verify from (1.1) that

$$\sum_{i=1}^{\infty} i\beta_i < \infty.$$

Thus, the expected return time for the Markov chain to state zero is finite. Hence, the Markov chain is positive recurrent. Because the Markov chain is aperiodic, it follows that for each starting state l,

$$\lim_{k\to\infty} Q^k_{l,i} = \pi_i, \qquad i = 0, 1, \ldots,$$

where $\pi = (\pi_1, \pi_2, \ldots)'$ is the stationary distribution for the Markov chain. Because Q is row stochastic and π is a probability distribution, it follows from Chung (1974, Theorem 4.5.4) that

$$\lim_{k\to\infty} \sum_{i=1}^{\infty} |Q^k_{l,i} - \pi_i| = 0.$$

Hence, we have

$$\lim_{k\to\infty} \sum_{i=1}^{\infty} \left|Q^{k+1}_{1,i} - Q^{k+1}_{2,i}\right| = 0. \quad \square$$

LEMMA 4.9. *For each $t \geqslant 0$, as $n \to \infty$,*

$$\sup_{0\leqslant s\leqslant t} \left| \sum_{i=1}^{\infty} (m^n)'(P')^i \lambda^n \left(\tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau^i_n(ns)\right)\right)\right| \to 0.$$

PROOF. Because $m^n \to m$ and $\lambda^n \to \lambda$, it can be checked that

$$\limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \left| \sum_{i=1}^{\infty} (m^n)'(P')^i \lambda^n \left(\tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau^i_n(ns)\right)\right)\right|$$

$$\leqslant \limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \left| \sum_{i=1}^{k} (m^n)'(P')^i \lambda^n \left(\tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau^i_n(ns)\right)\right)\right|$$

$$+ \limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \left| \sum_{i=k+1}^{\infty} (m^n)'(P')^i \lambda^n \left(\tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau^i_n(ns)\right)\right)\right|$$

$$\leqslant \sum_{i=1}^{k} (m)'(P')^i \lambda \limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \left| \tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau^i_n(ns)\right)\right|$$

$$+ m'(P')^{k+1}(I - P')^{-1}\lambda 2 \limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} |\tilde{\tau}_n(s)|.$$

Because $(P')^{k+1} \to 0$ and $\limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} |\tilde{\tau}_n(s)|$ is finite, it suffices to show that for each i,

$$\limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \left|\tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau_n^i(ns)\right)\right| = 0.$$

This follows from Lemma 4.7 and

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{0\leqslant s\leqslant t} \left|\tilde{\tau}_n(s) - \tilde{\tau}_n\left(\frac{1}{n}\tau_n^i(ns)\right)\right| \\
&\leqslant \limsup_{n\to\infty} \sum_{l=1}^{i} \left\| \frac{1}{\sqrt{n}}\hat{\tau}_n\left(\tau_n^{l-1}(n\cdot)\right) - \frac{1}{\sqrt{n}}\hat{\tau}_n\left(\tau_n^{l}(n\cdot)\right)\right\|_t \\
&\leqslant \limsup_{n\to\infty} \left\| \frac{1}{\sqrt{n}}\hat{\tau}_n(n\cdot) - \frac{1}{\sqrt{n}}\hat{\tau}_n(\tau_n(n\cdot))\right\|_t. \quad \square
\end{aligned}$$

PROOF OF THEOREM 2.1. By (3.10) and (3.16), we have

$$\begin{aligned}
W(t) &= e'\hat{V}(A(t)) + m'\hat{A}(t) + (\rho - 1)t + Y(t) \\
&= \xi(t) - \sum_{i=1}^{\infty} m'(P')^i \lambda \hat{\tau}(\tau^{i-1}(t)) + (\rho - 1)t + Y(t)
\end{aligned} \tag{4.8}$$

where e is the c-dimensional vector of ones, ρ is defined as in (2.4) and ξ is defined in (3.19). Hence

$$W(\tau(t)) = \xi(\tau(t)) - \sum_{i=1}^{\infty} m'(P')^i \lambda \hat{\tau}(\tau^i(t)) + (\rho - 1)\tau(t) + Y(\tau(t)).$$

Therefore, by (3.17),

$$\begin{aligned}
&\left(1 + \sum_{i=1}^{\infty} m'(P')^i \lambda\right)\hat{\tau}(t) \\
&\quad = \xi(\tau(t)) + (\rho - 1)\tau(t) + Y(\tau(t)) + \epsilon_1(t) \\
&\qquad + \sum_{i=1}^{\infty} m'(P')^i \lambda\left(\hat{\tau}(t) - \hat{\tau}(\tau^i(t))\right).
\end{aligned} \tag{4.9}$$

Because $\tau_n(nt)/n \to t$ u.o.c., $\xi^n(\tau_n(nt))/\sqrt{n} \to \xi^*(t)$ u.o.c., $\sqrt{n}(\rho^n - 1) \to \theta$, $Y^n(\tau_n(nt))/\sqrt{n} \to Y^*(t)$ u.o.c. and Lemma 4.9 holds, we have

$$\tilde{\tau}_n(t) \to \frac{1}{\beta}(\xi^*(t) + \theta t + Y^*(t)) = \frac{1}{\beta}Z^*(t) = W^*(t) \quad \text{u.o.c.},$$

where β is defined in (2.14). Thus for each $i \geqslant 1$,

$$\tilde{\tau}_n(\tau_n^{i-1}(nt)/n) \to \frac{1}{\beta}Z^*(t) \quad \text{u.o.c.}$$

Also, similar to the proof of Lemma 4.9, one can show that

$$\sup_{0\leqslant s\leqslant t}\left|\sum_{i=1}^{\infty}(m^n)'(P')^i\lambda^n\left(\tilde{\tau}_n(\tau_n^{i-1}(ns)/n)-\frac{1}{\beta}Z^*(s)\right)\right|\to 0.$$

It follows from (4.8) that

$$\tilde{W}^n(t)\to\xi^*(t)+\theta t+Y^*(t)-\sum_{i=1}^{\infty}m'(P')^i\lambda\frac{1}{\beta}Z^*(t)$$

$$=Z^*(t)-\sum_{i=1}^{\infty}m'(P')^i\lambda\frac{1}{\beta}Z^*(t)=\frac{1}{\beta}Z^*(t)\quad\text{u.o.c.}\quad\square$$

COROLLARY 4.1. *Let $W_k^n(t)$ and $Q_k^n(t)$ be the workload process and the queue length process for class k customers in the nth system. Under the condition of Theorem* 2.1, *as* $n\to\infty$,

$$\frac{1}{\sqrt{n}}Q_k^n(nt)\to\lambda_k W^*(t)\quad\text{u.o.c.},\tag{4.10}$$

$$\frac{1}{\sqrt{n}}W_k^n(nt)\to\lambda_k m_k W^*(t)\quad\text{u.o.c.}\tag{4.11}$$

PROOF. Recall the definition of $\tau_n(t)$ defined in §3. The number of class k customers in the system at time t is

$$Q_k^n(t)=A_k^n(t)-A_k^n(\tau_n(t)).$$

Let $\tilde{A}_k^n(t)=\hat{A}_k^n(nt)/\sqrt{n}$. Then

$$\frac{1}{\sqrt{n}}Q_k^n(nt)=\tilde{A}_k^n(t)-\tilde{A}_k^n(\tilde{\tau}_n(t))+\lambda_k^n\tilde{\tau}_n(t).$$

Following the proof of Theorem 2.1 and (3.16), we have

$$\tilde{A}^n(t)\to(I-P')^{-1}\left(\xi^a(t)+\sum_{k=1}^{c}\xi^k(\lambda_k t)-P'\lambda W^*(t)\right),\quad\text{u.o.c.}$$

Therefore,

$$\frac{1}{\sqrt{n}}Q_k^n(nt)\to\lambda_k W^*(t).$$

Similarly, we can show that

$$\frac{1}{\sqrt{n}}W_k^n(nt)=\frac{1}{\sqrt{n}}(V_k^n(A_k^n(nt)))-V_k^n(A_k^n(\tau_n(nt)))\to\lambda_k m_k W^*(t),\quad\text{u.o.c.}\quad\square$$

5. QNET analysis. In this section we apply Harrison and Nguyen's QNET method to the performance analysis of the queueing system introduced in §1. The QNET method was proposed by Harrison and Nguyen (1990, 1993) for the performance analysis of general multiclass queueing networks. The heavy traffic limit theorem proved in this paper justifies the QNET method for our system.

Based on Theorem 2.1, the workload process $W = \{W(t), t \geqslant 0\}$ can be approximated by a one-dimensional reflecting Brownian motion (RBM) with variance $(1 + g)^{-2}\sigma^2$ and drift $(1 + g)^{-1}(\rho - 1)$, where σ^2 is definend in (2.17), ρ is the traffic intensity defined in (2.4) and

$$g = \sum_{i=1}^{\infty} m'(P')^i \lambda = m'(I - P')^{-1}\lambda - \rho. \tag{5.1}$$

Note that from the definition in (2.14), $1 + g$ is simply β. The introduction of the extra symbol g makes our analysis completely analogous to the one given by Harrison and Nguyen (1993). (They used matrix G instead of g in their paper.) Assume that $\rho < 1$. Then the limiting RBM has the exponential stationary distribution with mean $\sigma^2/(2(1 - \rho)(1 + g))$ (see, for example, §5.6 of Harrison (1985)). Therefore, the QNET estimate of the average waiting time per visit to the station is

$$E(W(\infty)) \approx \frac{\sigma^2}{2(1 + g)(1 - \rho)}. \tag{5.2}$$

Harrison and Nguyen (1993, §6) proposed a refined QNET method, which coincides with their original QNET method proposed in Harrison and Nguyen (1990). Specializing to our case, their refined QNET estimate can be obtained by replacing g in (5.1) by $\hat{g} \equiv g/\rho$. The resulting average waiting time formula is the same as in (5.2) except that g is replaced by $\hat{g}$. When the system is in heavy traffic, these two estimates are close. However, when the traffic intensity is moderate, say, less than 70%, the difference between these two estimates can be significant. It is expected that this refined QNET should perform better most of the time. Readers are referred to §6 of Harrison and Nguyen (1993) for an informal defense of the refinement. In the remainder of this section, we present two network examples, where the refined QNET estimates are compared with simulation estimates.

5.1. *A feedback station with one type of customer.* Pictured in Figure 1 is a multiclass station, where customers arrive at the station according to a renewal process with rate 1 and interarrival time squared coefficient of variation (SCV) c_a^2. (SCV of a positive random variable is defined as variance divided by mean squared.) Each customer visits the station exactly twice and then exits. It is assumed that service times $\{v_k(i), i \geqslant 1\}$ in the kth visit are i.i.d. random variables with mean m_k and SCV $c_{s,k}^2$, $k = 1, 2$. We further assume that $\{v_1(i), i \geqslant 1\}$ and $\{v_2(i), i \geqslant 1\}$ are mutually independent and they are independent of the arrival process. One can check that for this model $\alpha = (1, 0)'$, $m = (m_1, m_2)'$,

$$P = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \Gamma = \operatorname{diag}\left(c_a^2, 0, m_1^2 c_{s,1}^2, m_2^2 c_{s,2}^2\right).$$

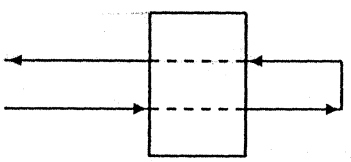

FIGURE 1. A multiclass station with feedback.

Therefore, we have $g = m_2$, $\sigma^2 = (m_1 + m_2)^2 c_a^2 + m_1^2 c_{s,1}^2 + m_2^2 c_{s,2}^2$ and

$$E(W(\infty)) \approx \frac{(m_1 + m_2)^2 c_a^2 + m_1^2 c_{s,1}^2 + m_2^2 c_{s,2}^2}{2(1 + m_2/\rho)(1 - \rho)}.$$

In Harrison and Nguyen (1993), the authors also proposed the QNET estimates of the *total* mean sojourn time in a network. Specializing to our case, the QNET estimates of the total mean sojourn time is

$$2E(W(\infty)) + m_1 + m_2 = 2E(W(\infty)) + \rho.$$

We consider four versions of this queueing system. Each version corresponds to a different triad of SCV's $(c_a^2, c_{s,1}^2, c_{s,2}^2)$ chosen from the set: (1, 1, 1), (2, 0.25, 2), (2, 2, 0.25) and (0.25, 2, 2). We label these four versions as systems A, B, C and D. In each system we consider three cases: ($m_1 = 0.8$, $m_2 = 0.1$), ($m_1 = 0.1$, $m_2 = 0.8$) and ($m_1 = 0.45$, $m_2 = 0.45$). Cases 1 and 2 have very different service requirements for the two visits. Whereas for Case 3, each customer's two mean service requirements are the same. Table 1 gives the simulation estimates and QNET estimates of the *total* mean sojourn time in the system and the mean waiting time for *each* visit to the station. In simulation, service times and interarrival times are fitted with an Erlang distribution, exponential distributions, or gamma distributions depending on the SCV being less than one, equal to one, or larger than one, respectively. For example, when service times have mean m and SCV $c^2 = 2$, we use a gamma distribution with density function

$$f(x) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta}, \qquad x > 0,$$

to fit the service time distribution, where the shape parameter $\alpha = 1/c^2 = 1/2$ and the scale parameter $\beta = mc^2$. The simulations were performed using SIMAN IV. In all cases 10 replications were run. In each run we let 10,000 customers leave the system. In this table as in all subsequent tables, the numbers in parentheses after the simulation results represent the half-width of 95% confidence intervals, expressed as

TABLE 1. Simulation estimates and QNET estimates for the *total* mean sojourn time and mean waiting time for *each* visit.

Sys / Case		*Mean Waiting Time*				*Mean Sojourn Time*			
		SIM		*QNET*		*SIM*		*QNET*	
A	1	6.63	(8.87%)	6.57	(−0.90%)	14.10	(8.37%)	14.04	(−0.43%)
	2	3.78	(8.65%)	3.87	(2.38%)	8.46	(7.72%)	8.63	(2.01%)
	3	4.01	(8.20%)	4.05	(1.00%)	8.93	(7.37%)	10.00	(11.98%)
B	1	8.56	(8.47%)	8.10	(−5.37%)	18.00	(8.06%)	17.10	(−5.00%)
	2	7.36	(7.55%)	7.68	(4.39%)	15.60	(7.12%)	16.27	(4.29%)
	3	7.11	(10.60%)	6.92	(−2.67%)	15.10	(10.00%)	14.74	(−2.38%)
C	1	12.50	(9.28%)	13.06	(4.48%)	25.90	(8.96%)	27.02	(4.32%)
	2	5.05	(13.25%)	4.77	(−5.54%)	11.00	(12.18%)	10.44	(−5.09%)
	3	7.08	(11.53%)	6.92	(−2.26%)	15.00	(10.87%)	14.74	(−1.73%)
D	1	5.85	(12.44%)	6.76	(15.56%)	12.60	(11.59%)	14.42	(14.44%)
	2	3.84	(9.79%)	3.98	(3.65%)	8.58	(8.85%)	8.86	(3.26%)
	3	3.10	(8.90%)	3.38	(9.03%)	7.09	(7.88%)	7.66	(8.04%)
Average absolute percentage error		9.80%		4.77%		9.08%		5.25%	

a percentage of the simulation average. The number in parentheses after the QNET estimates represent percentage errors from the simulation average. This format, suggested by Reiman (1990), makes it easy to determine the statistical significance of the errors. As we can see from the table, the QNET estimates (except in Case D/1) are almost always within the 95% confidence intervals of the simulation results. It is worth noting that having the shorter service time in the first visit will significantly reduce mean waiting time as well as the total mean sojourn time. For example, in System C, both the QNET and simulation predict that Case 2 has a reduction of waiting time by a factor larger than 2 compared with Case 1. Reduction ratio can be estimated from our QNET analysis. For example, when $c_{s,1}^2 = c_{s,2}^2$, the ratio of the mean waiting times in Case 1 and Case 2 is

$$\frac{1 + m_2^{\text{case 2}}/\rho}{1 + m_2^{\text{case 1}}/\rho} = 1.7.$$

5.2. *A system with two types of customers.* Consider a multiclass station with two types of customers. Type 1 customers visit the station five times and then exit and type 2 customers visit the station twice and then exit. We assume that type k customers arrive at the station according to a Poisson process with rate 1, $k = 1, 2$ and service times in each stage are i.i.d. random variables. We assume that two arrival processes are mutually independent, all service time sequences are mutually independent and are independent of the arrival processes. The service times during the ith visit for type 1 customers have mean m_i and SCV $c_{s,i}^2$, $i = 1, 2, 3, 4, 5$. The service times during the first and second visit for type 2 customers have mean m_6, SCV $c_{s,6}^2$ and m_7, SCV $c_{s,7}^2$, respectively. Therefore the station traffic intensity is $\rho = \sum_{i=1}^7 m_i$. We consider four versions of the system. Systems A, B, C and D corresponds to

$$\left(c_{s,1}^2, c_{s,2}^2, c_{s,3}^2, c_{s,4}^2, c_{s,5}^2, c_{s,6}^2, c_{s,7}^2\right)$$

being equal to (0.25, 2, 1, 0.25, 2, 2, 0.25), (2, 0.25, 0.25, 1, 1, 2, 2), (1, 0.25, 2, 2, 0.25, 1, 0.25) and (2, 2, 1, 2, 2, 0.25, 0.25). We assume $\rho = 0.90$ and we consider four cases for each system. Case 1, 2, 3 and 4 have

$$(m_1, m_2, m_3, m_4, m_5, m_6, m_7)$$

chosen from (0.1, 0.1, 0.05, 0.1, 0.1, 0.2, 0.25), (0.05, 0.05, 0.05, 0.2, 0.1, 0.2, 0.25), (0.2, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1) and (0.05, 0.05, 0.05, 0.025, 0.025, 0.6, 0.1). For Cases 1 and 2, we have $\sum_{i=1}^5 m_i = m_6 + m_7 = 0.45$, which indicates that type 1 customers and type 2 customers have the same average offered load. For Case 3 and 4, we have $\sum_{i=1}^5 m_i = 0.7$, $m_6 + m_7 = 0.2$ and $\sum_{i=1}^5 m_i = 0.2$, $m_6 + m_7 = 0.7$, respecitvely. Table 2 gives the simulation estimates and QNET estimates of mean waiting time for each visit and mean sojourn time for each customer type. The QNET estimates of the mean waiting time and sojourn time for type 2 customers are quite impressive compared with the simulation estimates. Note that both QNET and simulation predict that Case 4 always causes much longer delay than the other three cases for all four systems. It is interesting, however, to observe that the QNET always significantly underestimates the sojourn time for type 1 customers. We have no theoretical explanation for it at the moment.

TABLE 2. Simulation estimates and QNET estimates of mean sojourn times and mean waiting time for the multiclass station with two types of customers

	Waiting Time For Each Visit				*Sojourn Time For Type 1 Customers*				*Sojourn Time For Type 2 Customers*			
	SIM		*QNET*		*SIM*		*QNET*		*SIM*		*QNET*	
A1	1.15	(7.97%)	1.20	(4.63%)	6.10	(7.51%)	5.26	(−13.77%)	2.82	(6.67%)	2.86	(1.42%)
2	1.01	(7.59%)	1.05	(4.37%)	5.42	(7.07%)	4.67	(−13.84%)	2.57	(6.11%)	2.56	(−0.39%)
3	1.19	(0.90%)	1.22	(2.35%)	6.68	(8.10%)	5.57	(−16.62%)	2.56	(8.40%)	2.64	(3.13%)
4	3.64	(10.30%)	4.29	(17.76%)	17.50	(10.98%)	17.35	(−0.86%)	8.83	(8.88%)	9.27	(4.98%)
B1	1.34	(7.46%)	1.43	(6.99%)	7.03	(7.40%)	6.18	(−12.09%)	3.22	(6.52%)	3.32	(3.11%)
2	1.23	(7.32%)	1.30	(5.98%)	6.47	(7.11%)	5.66	(−12.52%)	3.02	(6.29%)	3.06	(1.32%)
3	1.35	(6.74%)	1.38	(2.18%)	7.46	(6.26%)	6.22	(−16.62%)	2.88	(6.15%)	2.96	(2.78%)
4	3.80	(11.11%)	4.34	(14.18%)	18.20	(11.60%)	17.56	(−3.52%)	9.15	(9.73%)	9.38	(2.51%)
C1	1.04	(5.55%)	1.10	(5.67%)	5.59	(5.24%)	4.85	(−13.24%)	2.59	(4.40%)	2.65	(2.32%)
2	1.01	(5.65%)	1.08	(6.79%)	5.40	(5.30%)	4.76	(−11.85%)	2.56	(4.57%)	2.61	(1.95%)
3	1.30	(9.00%)	1.35	(3.47%)	7.21	(8.22%)	6.08	(−15.67%)	2.78	(8.27%)	2.89	(3.96%)
4	2.78	(8.27%)	3.06	(10.21%)	13.40	(8.59%)	12.45	(−7.09%)	6.90	(7.00%)	6.83	(−1.01%)
D1	1.07	(6.03%)	1.13	(5.27%)	5.72	(5.79%)	4.96	(−13.29%)	2.65	(4.60%)	2.70	(1.89%)
2	0.99	(5.07%)	1.06	(7.01%)	5.32	(4.79%)	4.70	(−11.65%)	2.53	(3.94%)	2.58	(1.98%)
3	1.41	(8.87%)	1.46	(3.38%)	7.74	(8.26%)	6.53	(−15.63%)	3.01	(7.97%)	3.12	(3.65%)
4	1.41	(8.87%)	1.46	(3.38%)	7.74	(8.26%)	6.53	(−15.63%)	3.01	(7.97%)	3.12	(3.65%)
Average absolute error	7.09%		6.82%		7.36%		11.65%		6.51%		2.50%	

Acknowledgement. We thank Hong Chen and Ruth Williams for helpful comments on an earlier draft of this paper. We are grateful to an anonymous referee for pointing out an error in an earlier proof of Lemma 4.1. Research of the first author supported in part by two grants from Texas Instruments Corporation and by NSF Grants DMS-9209586 and DDM-9215233. Research of the second author supported in part by NSF Grant DMS-8901464.

References

Bramson, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4** 414–431.

Chen, H., J. G. Shanthikumar (1994). Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. *J. Discrete Event Dynamic Systems: Theory Appl.* **4** 269–291.

Chung, K. L. (1974). *A Course in Probability.* Wiley, New York.

Dai, J. G. (1994). On a positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

_____,V. Nguyen (1994). On the convergence of multiclass queueing networks in heavy traffic. *Ann. Appl. Probab.* **4** 26–42.

_____, Wang, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems Theory Appl.* **13** 41–46.

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

Harrison, J. M. (1985). *Brownian Motion and Stochastic Flow Systems.* Wiley, New York.

_____,Nguyen, V. (1990). The QNET method for two-moment analysis of open queueing networks. *Queueing Systems Theory Appl.* **6** 1–32.

_____,_____ (1993). Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems Theory Appl.* **13**, 5–40.

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

_____, _____ (1970b). Multiple channel queues in heavy traffic II. *Adv. Appl. Probab.* **2** 355–364.

Johnson, D. P. (1983). *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks.* PhD Thesis, University of Wisconsin.

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

Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* **9** 441–458.
_____ (1988). A multiclass feedback queue in heavy traffic. *Adv. Appl. Probab.* **20** 179–207.
_____ (1990). Asymptotically exact decomposition approximations for queueing networks. *Oper. Res. Lett.* **9** 363–370.
Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Sci.* **39** 1020–1028.

J. G. Dai: School of Industrial and Systems Engineering and School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332-0205; email: dai@isye.gatech.edu
T. G. Kurtz: Department of Mathematics and Statistics, University of Wisconsin, Madison, Wisconsin 53706; email: kurtz@math.wisc.edu