*The Annals of Applied Probability*
2000, Vol. 10, No. 3, 828–876

# STRONG APPROXIMATIONS FOR MULTICLASS FEEDFORWARD QUEUEING NETWORKS

BY HONG CHEN[1] AND XINYANG SHEN[2]

*University of British Columbia*

This paper derives the strong approximation for a multiclass queueing network, where jobs after service completion can only move to a downstream service station. Job classes are partitioned into groups. Within a group, jobs are served in the order of arrival; that is, a first-in first-out (FIFO) discipline is in force, and among groups, jobs are served under a preassigned preemptive priority discipline. We obtain the strong approximation for the network through an inductive application of an input–output analysis for a single-station queue. Specifically, we show that if the input data (i.e., the arrival and the service processes) satisfy an approximation (such as the functional law-of-iterated logarithm approximation or the strong approximation), then the output data (i.e., the departure processes) and the performance measures (such as the queue length, the workload and the sojourn time processes) satisfy a similar approximation. Based on the strong approximation, some procedures are proposed to approximate the stationary distribution of various performance measures of the queueing network. Our work extends and complements the existing work of Peterson and Harrison and Williams on the feedforward queueing network. The numeric examples show that strong approximation provides a better approximation than that suggested by a straightforward interpretation of the heavy traffic limit theorem.

**1. Introduction.** We first consider a single-server station serving multiple classes of jobs, where jobs of different classes may have different arrival and service time distributions. Job classes are partitioned into groups. Within a group, jobs are served in the order of arrival (i.e., a first-in first-out service discipline is in force), and among groups, jobs are served under a preassigned (static) preemptive priority discipline. Our key result is to establish that if the input data (i.e., the arrival and the service processes) satisfy an approximation [such as the functional law-of-iterated logarithm (FLIL) approximation or the strong approximation], then the output data (i.e., the departure processes) and the performance measures (such as the queue length, the workload and the sojourn time processes) satisfy a similar approximation. In an obvious way, this result extends to a feedforward multiclass network, where jobs can

Received March 1999; revised December 1999.

[1]Supported in part by NSERC (Canada) research grant and RGC (Hong Kong) research grant.

[2]Supported in part by NSERC (Canada) research grant and NSERC (Canada) post-graduate Fellowship.

*AMS* 1991 *subject classifications*. Primary 60F17, 60K25, 60G17; secondary 60J70, 90B10, 90B22.

*Key words and phrases*. Multiclass queueing network, diffusion approximations, fluid approximations, heavy traffic, reflected Brownian motion and strong approximation.

only move from a lower numbered station to a higher numbered station. We identify explicitly the strong approximation limit under the network setting.

This paper relates to, extends and complements, some existing works of Peterson (1991), Harrison and Williams (1992) and Chen and Mandelbaum (1994). Peterson (1991) first studied a multiclass feedforward queueing network to derive a heavy traffic limit theorem or a diffusion approximation theorem under a heavy traffic condition. It is shown that the limit can be described by a $J$-dimensional reflected Brownian motion (RBM), where $J$ equals the number of service stations in the network. In particular, a state space collapse phenomenon is observed for higher priority job classes; namely, the limit of the workload or the queue length of high priority jobs is zero. Thus, a direct application of this theorem would yield zero as an approximation to the queue length or the workload of a higher priority class. Clearly this is not satisfactory. Usually heuristics are offered in this case to provide a better approximation. Harrison and Williams (1992) studied the reflected Brownian motion suggested by the heavy traffic limit theorem and obtained a necessary and sufficient for the existence of a product form stationary distribution of the Brownian model. On the other hand, Chen and Mandelbaum (1994) derived a strong approximation for a generalized Jackson network; it was explained that the strong approximation refines the heavy traffic limit theorem in that it provides the rate of convergence and it does not require the network to be under a heavy traffic condition. This paper extends the strong approximation analysis in Chen and Mandelbaum (1994) to the multiclass feedforward network. Strong approximation yields appropriate approximations for the network's workload, queue length and sojourn time processes of all job classes (not just the lowest job class). The numeric examples indicate that the strong approximation does provide a much better approximation than the approximation suggested by the diffusion approximation, for both higher and lower priority classes.

There is a large literature on diffusion approximations, and readers are referred to survey papers by Whitt (1974), Lemoine (1978), Glynn (1990), Harrison and Nguyen (1993) and Williams (1996). Strong approximation is first used by Zhang, Hsu and Wang (1990) for approximating a queueing system. Horváth (1990) and Chen and Mandelbaum (1994) obtain the strong approximation for open generalized Jackson networks, and Zhang (1997) for closed generalized Jackson networks. The strong approximation has also been used to study time-dependent queues [see Mandelbaum and Massey (1995) and Mandelbaum, Massey and Reiman (1998)] and state-dependent queues [see Mandelbaum and Pats (1998)].

The paper is organized as follows. In the next section, we introduce the functional law-of-iterated-logarithm (FLIL) and the functional strong approximation theorem (FSAT) for two fundamental processes, which set a basis for our analysis of the queue. In Section 3, we obtain the FLIL and the FSAT for a single-station queue, which are the key results of the paper. In Section 4, by viewing some of the results in the previous section as an input–output theorem, we extend the strong approximation to the multiclass

feedforward network. Section 5 provides a procedure to approximate by a reflected Brownian motion (RBM) a feedforward queueing network with the renewal (exogenous) arrival process, i.i.d. service times and Markovian routing. This section is almost self-contained to make it convenient for those readers who would only need to obtain a strong approximation. Numeric examples are given in Section 6 to compare the performance measure estimates given by the simulation, the proposed approximations suggested by the strong approximation, and the approximation suggested by the diffusion approximation. In order to simplify the presentation, we assume that the traffic intensity is no greater than 1 through Sections 3–6; this would include almost all cases of practical interest. The more general case, where the traffic intensity may be strictly greater than 1, is summarized in the Appendix for a single-station queue.

We denote by $\Re^K$ the $K$-dimensional Euclidean space, and by $\Re_+^K = \{x \in \Re^K : x \geq 0\}$ its nonnegative orthant. Let $\Re = \Re^1$ and $\Re_+ = \Re_+^1$. All vectors are assumed to be column vectors, and the prime ($'$) is used to denote the transpose of a vector and a matrix. We denote $e^j$ the $j$th unit vector (whose $j$th element equals 1 and all other elements equal zero) and $e$ a vector of 1's (whose elements all equal 1), both in an appropriate dimension from the context. For $x = (x_k)_{k=1}^K \in \Re^K$, define the norm $\|x\| = \max_{1 \leq k \leq K} |x_k|$. Let $\mathscr{D}^K$ be the set of $K$-dimensional functions which are right-continuous and have left limits (RCLL), and let $\mathscr{D}_0^K = \{x \in \mathscr{D}^K : x(0) \geq 0\}$. For $X = (X_k) \in \mathscr{D}^K$, define the norm

$$\|X\|_T = \sup_{0 \leq t \leq T} \|X(t)\| \equiv \sup_{0 \leq t \leq T} \max_{1 \leq k \leq K} |X_k(t)|.$$

Sometimes for convenience, we write $\|X(t)\|_T$ for $\|X\|_T$. The composition $\{x(y(t)), t \geq 0\}$ of $x : \Re_+ \to \Re^K$ with $y : \Re_+ \to \Re_+^K$ is the function from $\Re_+$ to $\Re^K$ whose $k$th coordinate is the real-valued function $\{x_k(y_k(t)), t \geq 0\}$, $k = 1, \ldots, K$.

**2. Preliminaries.** In this section, we consider two fundamental processes. For ease of exposition, we present all results for one-dimensional processes. Since all the results are pathwise on an appropriate probability space, they have obvious generalizations to multidimensional cases, and without explicitly stating so, we shall quote these generalizations in the latter sections. Let $X \in \mathscr{D}$, and let $Y = \{Y(t), t \geq 0\}$ denote its inverse, defined by

$$Y(t) = \sup\{s \geq 0 : X(s) \leq t\}, \qquad 0 \leq t < \infty. \tag{1}$$

One important example of the above pair is that $X$ is a partial sum of a sequence $\xi = \{\xi_i,\ i = 1, 2, \ldots\}$ of nonnegative i.i.d. random variables, namely,

$$X(t) := \sum_{i=1}^{\lfloor t \rfloor} \xi_i, \qquad t \geq 0, \tag{2}$$

[with $X(t) = 0$ for $t < 1$], and its corresponding $Y$ [defined by (1)] is a renewal process.

We have the following result of the functional law-of-iterated-logarithm for the pair.

THEOREM 2.1. *Consider the* $(X, Y)$ *pair as introduced above. Suppose that*

$$\|X(t) - \overline{X}(t)\|_T \equiv \sup_{0\le t\le T} |X(t) - \overline{X}(t)| \overset{a.s.}{=} O(\sqrt{T \log\log T}) \tag{3}$$

*with* $X(t) = mt$ *for* $m > 0$. *Then*

$$\|Y(t) - \overline{Y}(t)\|_T \equiv \sup_{0\le t\le T} |Y(t) - \overline{Y}(t)| \overset{a.s.}{=} O(\sqrt{T \log\log T}) \tag{4}$$

*and* $\overline{Y}(t) = \mu t$ *with* $\mu = 1/m$. *Furthermore, when* $X$ *is a partial sum of nonnegative i.i.d. random variables as given by* (2), *if* $\xi_i$ *has a finite second moment, then* (3) *and* (4) *hold with* $m = E(\xi_i)$.

The proof of this theorem follows almost the same lines of the proof for Theorem 2.1.3 in Csörgő and Horváth (1993). For any stochastic process $X$ having a bound like (3), we say that $X$ has *a FLIL approximation* $\overline{X}$. We note that the FLIL approximation (3) implies a functional strong-law-of-large numbers (FSLLN) limit result. Namely, it implies that almost surely,

$$\overline{X}^n(t) := \frac{1}{n} X(nt) \to \overline{X}(t) \equiv mt,$$

as $n \to \infty$, and the convergence is uniform on any compact set (u.o.c.).

Next, we consider a refined approximation, namely, the functional strong approximation, for processes $X$ and $Y$.

THEOREM 2.2 (FSAT). *Consider the* $(X, Y)$ *pair as introduced above. Suppose that for some* $r > 2$,

$$\|X - \widetilde{X}\|_T \overset{a.s.}{=} \circ(T^{1/r}), \tag{5}$$

*as* $T \to \infty$, *with*

$$\widetilde{X}(t) = mt + \sigma B(t), \qquad t \ge 0,$$

*where* $m$ *and* $\sigma$ *are positive constants, and* $B = \{B(t),\ t \ge 0\}$ *is a standard Brownian motion* (*i.e., a Wiener process*). *Then we have*

$$\|Y - \widetilde{Y}\|_T \overset{a.s.}{=} \circ(T^{1/r'}), \tag{6}$$

*with* $r' = r$ *if* $r < 4$ *and arbitrary* $r' < 4$ *if* $r \ge 4$, *and with*

$$\widetilde{Y}(t) = \mu t - \mu\sigma B(\mu t), \qquad t \ge 0,$$

*where* $\mu = 1/m$. *Furthermore, when* $X$ *is a partial sum of nonnegative i.i.d. random variables as given by* (2), *if* $\xi_i$ *has a finite rth moment with* $r > 2$, *then we can assume that* $X$ *and* $Y$ *are defined on a probability space, on which there exists a standard Brownian motion* $B = \{B(t),\ t \ge 0\}$ *such that both* (5) *and* (6) *hold, with* $m = E(\xi_i)$ *and* $\sigma$ *being the standard deviation of* $\xi_i$.

This theorem follows from Theorems 2.1.1 and 2.1.3 of Csörgő and Horváth (1993), where they actually give a slightly better bound in (6) when (5) holds with $r > 4$. We note that the strong approximation (5) leads to a functional central limit theorem (FCLT) limit. Specifically, let

$$\widehat{X}^n(t) = \frac{1}{\sqrt{n}}[X(nt) - nmt];$$

the approximation (5) implies that

$$\widehat{X}^n \xrightarrow{d} \widehat{X} \quad \text{as } n \to \infty,$$

where $\widehat{X}$ is a driftless Brownian motion with a standard deviation $\sigma$, and "$\xrightarrow{d}$" denotes the weak convergence in $\mathscr{D}$ [refer to, e.g., Billingsley (1968) and Whitt (1980)].

Now we introduce the notion of $r$-strong continuity that will be used extensively in this paper. A function $x \in \mathscr{D}^K$ is said to be *strong continuous* with degree $r$, or $r$-strong continuous, for some $r \in (2, 4)$, if

$$\sup_{\substack{0 \le u,\, v \le T \\ |u-v| \le h(T)}} \|x(u) - x(v)\| = \circ(T^{1/r}) \quad \text{as } T \to \infty, \tag{7}$$

where $h(T) \equiv \sqrt{T \log\log T}$, and it is simply said to be strong continuous if it is $r$-strong continuous for all $r \in (2, 4)$. We note that an $r$-strong continuous function may not be continuous. A stochastic process $X = \{X(t),\ t \ge 0\}$ in $\mathscr{D}^K$ is said to be an $r$-strong continuous process for some $r \in (2, 4)$, if with probability 1, the sample path of this version is $r$-strong continuous. For simplicity, we shall assume throughout this paper that all $r$-strong continuous stochastic processes are defined on such a probability space. A stochastic process is simply said to be strong continuous if it is $r$-strong continuous for all $r \in (2, 4)$. We say a stochastic process $X$ has *a strong approximation* if for some $r \in (2, 4)$, there exists a probability space on which a version of $X$ (for simplicity we still write it as $X$) and an $r$-strong continuous stochastic process $\widehat{X}$ are defined such that

$$\sup_{0 \le t \le T} |X(t) - mt - \widehat{X}(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

where $m$ is a (deterministic) constant. When the above equality holds, we also say that $X$ has a strong approximation $\widetilde{X} = \{\widetilde{X}(t), t \ge 0\}$ with $\widetilde{X}(t) = mt + \widehat{X}(t)$.

LEMMA 2.3.

(i) *A Wiener process* (*i.e., a standard Brownian motion*) *is a strong continuous process.*

(ii) *If a process has a strong approximation, then it must have a* FLIL *approximation. Specifically, if the process $X$ satisfies*

$$\|X(t) - mt - \widehat{X}(t)\|_T \overset{a.s.}{=} \circ(T^{1/r})$$

*with* $\widehat{X} = \{\widehat{X}(t),\ t \geq 0\}$ *begin r-strong continuous, then we have*

$$\|X(t) - mt\|_T \overset{a.s.}{=} O\big(\sqrt{T \log\log T}\big).$$

(iii) *If process* $X = \{X(t),\ t \geq 0\} \in \mathscr{D}$ *has a* FLIL *approximation, then there exist positive* $M$ *and* $T$ *such that with probability* 1,

$$\|X(t)\| \leq Mt \quad \text{for } t \geq T.$$

(iv) *A linear combination of r-strong continuous functions is r-strong continuous, and a (deterministic) linear combination of r-strong continuous processes is r-strong continuous.*

(v) *Let* $X = \{X(t), t \geq 0\}$ *be an r-strong continuous process, and let* $\tau = \{\tau(t), t \geq 0\}$ *be a process with* $\tau(t) \in [0, \infty)$ *for all* $t \geq 0$ *having a* FLIL *approximation:*

$$\|\tau(t) - \alpha t\|_T \overset{a.s.}{=} O(\sqrt{T \log\log T}).$$

*Then we have*

$$\|X(\tau(t)) - X(\alpha t)\|_T \overset{a.s.}{=} \circ(T^{1/r}).$$

(vi) *Let* $\xi = \{\xi_n,\ n \geq 1\}$ *be a sequence of random variables and let*

$$X(t) = \sum_{n=1}^{\lfloor t \rfloor} \xi_n.$$

*Assume that* $X$ *has a strong approximation,*

$$\|X(t) - mt - \widehat{X}(t)\|_T \overset{a.s.}{=} \circ(T^{1/r}).$$

*Suppose that process* $\Xi$ *has a strong approximation,*

$$\|\Xi(t) - \beta t - \widehat{\Xi}(t)\|_T \overset{a.s.}{=} \circ(T^{1/r}).$$

*Let* $Y(t) \equiv X(\Xi(t)),\ t \geq 0$. *Then* $Y = \{Y(t),\ t \geq 0\}$ *has the following strong approximation:*

$$\|Y(t) - m\beta t - m\widehat{\Xi}(t) - \widehat{X}(\beta t)\|_T \overset{a.s.}{=} \circ(T^{1/r}).$$

PROOF. Part (i) of this lemma is a special case of Lemma 3.6.3 in Chen and Mandelbaum (1994), and parts (ii)–(v) clearly follow from the definitions of the strong continuity, the FLIL approximations and the strong approximations. For (vi), we have

$$\begin{aligned}
&\|Y(t) - m\beta t - m\widehat{\Xi}(t) - \widehat{X}(\beta t)\|_T \\
&\quad \leq \|X(\Xi(t)) - m\Xi(t) - \widehat{X}(\Xi(t))\|_T + m\|\Xi(t) - \beta t - \widehat{\Xi}(t)\|_T \\
&\qquad + \|\widehat{X}(\Xi(t)) - \widehat{X}(\beta t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}),
\end{aligned}$$

where the last equality follows from the strong approximation assumptions for $X$ and $\Xi$, part (iii) and (v) of the lemma and the strong continuity of $\widehat{X}$. □

We state an additional property of strong continuity, which relates to the regulator of the one-dimensional reflected mapping. [Refer to Harrison (1985) for the definition and the properties of the one-dimensional reflection mapping.] (The proof of the following proposition can be found in the Appendix, Section A.1.)

PROPOSITION 2.4. *Suppose that $x \in \mathscr{D}_0$ be an $r$-strong continuous function $(2 < r < 4)$. Let*

$$f(t) \equiv \sup_{0 \le s \le t} [-\theta s - x(s)]^+ - [-\theta]^+ t, \tag{8}$$

*where $\theta$ is a real number. Then $f$ is also an $r$-strong continuous function.*

Finally, we state a bound for a special class of reflected Brownian motions (which we shall show may arise as the strong approximation for the feed-forward multiclass queueing network under study in this paper). Let $X$ be a $K$-dimensional Brownian motion starting at $X(0) = x \in \Re_+^K$ with drift $\theta$ and covariance matrix $\Gamma$ (on an appropriate probability space). Let $R$ be a $K \times K$ lower-triangular matrix with positive diagonal elements. Then by inductively applying the one-dimensional reflection mapping, we can show that there exists a unique pair $(Y, Z)$ satisfying

$$Z = X + RY \ge 0,$$

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

$$\int_0^\infty Z_k(t)\, dY_k(t) = 0, \qquad k = 1, \ldots, K.$$

The unique $Z$ is called the reflected Brownian motion and $Y$ the regulator of the reflected Brownian motion, associated with data $(x, \theta, \Gamma, R)$.

THEOREM 2.5. *Let $Z$ be a $K$-dimensional reflected Brownian motion associated with data $(x, \theta, \Gamma, R)$, where $R$ is a lower-triangular matrix with positive diagonal elements. Suppose that $R^{-1}\theta < 0$. Then*

$$\sup_{0 \le t \le T} \|Z(t)\| \overset{a.s.}{=} O(\log T), \tag{9}$$

*which in particular implies that for any $r > 0$,*

$$\sup_{0 \le t \le T} \|Z(t)\| \overset{a.s.}{=} \circ(T^{1/r}).$$

The proof of this theorem is in the Appendix, Section A.1.

**3. A multiclass single server station.** We formally describe the queueing model in Section 3.1, and then establish its FLIL theorem and its strong approximation theorem in Sections 3.3 and 3.4, respectively. In Section 3.5, we provide an improved strong approximation for the sojourn times. We provide a packet queue application in Section 3.6.

3.1. *Queueing model.* The queueing model under consideration is a single-server station serving $K$ classes of jobs. Let $\mathscr{K} = \{1, \ldots, K\}$ denote the set of job class indices. Jobs of all classes arrive exogenously, wait for service and after service completion leave the system. To specify the service discipline, we partition $\mathscr{K}$ into $L$ groups, $1, \ldots, L$, and let $g_\ell$ denote the set of classes belong to group $\ell$. For any $\ell < \ell'$, a job of a class in $g_\ell$ has a higher preemptive-resume priority over any job of any class in $g_{\ell'}$; as a result, the presence of the latter job has no impact on the former job. In this sense, for $\ell < \ell'$, a job of a class in $g_\ell$ *does not see* any job of any class in $g_{\ell'}$. Within each group, jobs are served under FIFO discipline. Let $\pi$ be a mapping from $\mathscr{K}$ (job class index set) to $\mathscr{L} := \{1, \ldots, L\}$ (job group index set); specifically, $k \in g_{\pi(k)}$; that is, class $k$ is in group $\pi(k)$. A job of class $k$ is referred to as a (group) $\ell$ job or it is in $g_\ell$, if $k \in g_\ell$. The station is assumed to have an infinite waiting room. We note that with $L = 1$, the station models a multiclass queue under a (pure) FIFO service discipline, and with $L = K$ (implying each group $\ell$ contains a single class), the station models a multiclass queue under a (pure) priority service discipline.

The queue is described by the following primitive data: $K$ counting processes $A_k = \{A_k(t),\ t \geq 0\}$ $(k \in \mathscr{K})$, and $K$ sequences of nonnegative random variables $v_k = \{v_k(n),\ n \geq 1\}$ $(k \in \mathscr{K})$, all defined on the same probability space. The (integer-valued) quantity, $A_k(t)$, indicates the number of class $k$ jobs that have arrived (exogenously) to the system during $[0, t]$. The random variable $v_k(n)$ is the service time required for the $n$th class $k$ job. We assume that initially there are no jobs in the system.

We introduce some notation. Let $u_k(1)$ be the arrival time of the first class $k$ job and $u_k(n)$, $n > 1$, the interarrival time between the $(n-1)$st and the $n$th class $k$ jobs (corresponding to jump points of $A_k$). We call $u = (u_k)$ with $u_k = \{u_k(n),\ n \geq 1\}$ the interarrival time sequence, and call $v = (v_k)$ the service time sequence. We introduce the summation,

$$V_k(0) = 0, \qquad V_k(n) = \sum_{m=1}^{n} v_k(m), \qquad n \geq 1,\ k \in \mathscr{K},$$

and define its associated counting processes,

$$S_k(t) = \sup\{n \geq 0 \colon V_k(n) \leq t\}, \qquad k \in \mathscr{K}.$$

Let $V_k(t) = V_k(\lfloor t \rfloor)$ for any $t \geq 0$. Let $V = (V_k)$, $A = (A_k)$ and $S = (S_k)$. We call $A$ an exogenous arrival process and $S$ a service process.

The performance measures of interest are the $L$-dimensional aggregated workload process $Z = (Z_\ell)$ with $Z_\ell = \{Z_\ell(t),\ t \geq 0\}$ $(\ell = 1, \ldots, L)$, the $K$-dimensional queue length process $Q = (Q_k)$ with $Q_k = \{Q_k(t),\ t \geq 0\}$ $(k = 1, \ldots, K)$, and the $L$-dimensional cumulative idle time process $Y = (Y_\ell)$ with $Y_\ell = \{Y_\ell(t),\ t \geq 0\}$ $(\ell = 1, \ldots, L)$. All of these processes are nonnegative processes. The quantity, $Z_\ell(t)$, indicates the total amount of current work for the station embodied in jobs that are in groups 1 to $\ell$ and that are queued or in service at the station at time $t$. The quantity, $Q_k(t)$, is integer valued and

indicates the number of class $k$ jobs at the station at time $t$. The quantity, $Y_\ell(t)$, indicates the cumulative amount of time that the server does not serve jobs in groups 1 to $\ell$ during $[0, t]$. It is clear that $Y$ must be nondecreasing with $Y(0) = 0$.

To describe the dynamics of the queue, we need some additional notation:

1. $D_k(t)$ counts the number of departures of class $k$ jobs from the station during $[0, t]$ after their service completions.
2. $W_k(t)$ is the workload process of class $k$ jobs.
3. $T_k(t)$ is the total amount of time that the server at the station has served jobs of class $k$ during $[0, t]$.
4. $\tau_\ell(t)$ is the arrival time of the $g_\ell$ job which has most recently completed service [$\tau_\ell(t)$ is zero if there have been no service completions for group $\ell$].
5. $\nu_k(t)$ is the partial service time (if any) that has been performed on job of type $k$ during $(\tau_\ell(t), t]$, where $k \in g_\ell$.
6. $\mathscr{S}_k(t)$ is the sojourn time of class $k$ jobs at time $t$, denoting the time which will be spent in the queue by the first class $k$ job to arrive at time greater than or equal to $t$.
7. $\eta_k(t)$ is the time at which the first class $k$ job arrives during $[t, \infty)$.
8. $\mathscr{T}_k(t)$ is the time that a class $k$ job would spend at the station if it arrived at time $t$.

From the above definitions, we have the following dynamic relations:

$$Q_k(t) = A_k(t) - D_k(t), \tag{10}$$

$$W_k(t) = V_k(A_k(t)) - V_k(D_k(t)) - \nu_k(t), \tag{11}$$

$$Y_\ell(t) = t - \sum_{i=1}^{\ell} \sum_{j \in g_i} T_j(t), \tag{12}$$

$$Z_\ell(t) = \sum_{i=1}^{\ell} \sum_{j \in g_i} W_j(t) = \sum_{i=1}^{\ell} \sum_{j \in g_i} V_j(A_j(t)) - t + Y_\ell(t), \tag{13}$$

$$D_k(t) = S_k(T_k(t)) = A_k(\tau_\ell(t)) \quad \text{where } \ell = \pi(k), \tag{14}$$

$$0 \le \nu_k(t) \le \max_{1 \le n \le A_k(t)} v_k(n), \tag{15}$$

$$0 \le \eta_k(t) - t \le u_k(A_k(t) + 1), \tag{16}$$

$$\mathscr{S}_k(t) = \mathscr{T}_k(\eta_k(t)). \tag{17}$$

Relations (10) and (11) are flow-balance relations in terms of time and head count of jobs, respectively. The second equality in (14) follows from the FIFO service discipline within jobs in group $\ell$; namely, any group $\ell$ jobs that arrived before the time $\tau_\ell(t)$ must have finished service by time $t$.

The dynamics of $\mathscr{T}_k$ can be described by a recursive relationship,

$$\mathscr{T}_k(t) = Z_{\pi(k)}(t) + \sum_{i=1}^{\pi(k)-1} \sum_{j \in g_i} [V_j(A_j(\mathscr{T}_k(t)+t)) - V_j(A_j(t))] + V_k(A_k(t)) - V_k(A_k(t)-1), \tag{18}$$

where $Z_{\pi(k)}(t)$ is the current workload at time $t$ (contributed by those jobs having priority no less than class $k$) just before this class $k$ job arrives, $V_j(A_j(\mathscr{T}_k(t)+t)) - V_j(A_j(t))$ is the workload of class $j$ jobs that arrive after time $t$ and before the completion of this class $k$ job, and $V_k(A_k(t)) - V_k(A_k(t)-1)$ is the service time required for this job. The summation term in (18) is the total amount work embodied in the jobs which arrive during the sojourn time of the concerned job and which have higher priority than the concerned job (the class $k$ job).

Define the "net-put" process $N_\ell(t)$ by

$$N_\ell(t) \equiv \sum_{i=1}^{\ell} \sum_{j \in g_i} V_j(A_j(t)) - t. \tag{19}$$

The quantity, $N_\ell(t)$, represents the total workload input of all types of jobs in $g_1 \cup \cdots \cup g_\ell$ during $[0, t]$ minus the work that would be finished if the server were never idle. The equality (13) can alternatively be written as

$$Z_\ell(t) = N_\ell(t) + Y_\ell(t) \geq 0. \tag{20}$$

Under the work-conserving (i.e., nonidling) condition, $Y_\ell(\cdot)$ can increase at time $t$ only when $Z_\ell(t) = 0$. Hence, the pair $(Z_\ell, Y_\ell)$ jointly satisfies the one-dimensional reflection mapping theorem in Harrison (1985), which yields

$$Y_\ell(t) = \sup_{0 \leq s \leq t} [-N_\ell(s)]. \tag{21}$$

We assume that there exists a long-run average arrival rate and an average service time; namely,

$$\frac{A(t)}{t} \to \lambda \quad \text{as } t \to \infty,$$
$$\frac{V(n)}{n} \to m \quad \text{as } n \to \infty.$$

We shall call $\lambda_k$, the $k$th coordinate of $\lambda$, the (exogenous) arrival rate of class $k$ job, and call $m_k$, the $k$th coordinate of $m$, the average service time of class $k$ job [alternatively the mean service time of class $k$ job when $v_k(n)$ has the same finite mean for all $n \geq 0$]. We assume that for all $k \in \mathscr{K}$, $\lambda_k > 0$ and $m_k > 0$. Call $\mu_k := 1/m_k$ the service rate of class $k$. Define

$$\beta_i \equiv \sum_{j \in g_i} \lambda_j m_j, \qquad \rho_\ell \equiv \sum_{i=1}^{\ell} \beta_i, \qquad \rho \equiv \rho_L, \tag{22}$$

where $\beta_i$ is the aggregated traffic intensity for job classes in $g_i$, $\rho_\ell$ the aggregated traffic intensity for all classes in $g_1 \cup \cdots \cup g_\ell$ ($\rho_0 \equiv 0$), and $\rho$ is the traffic intensity of the service station.

We shall assume that the traffic intensity $\rho \le 1$, through this section. The discussion for the case when $\rho > 1$ is in the Appendix, Section A.2.

3.2. *Preliminary lemmas.* We state and prove three lemmas that will be used in establishing the main results.

LEMMA 3.1. *Suppose that for $k = 1, \ldots, K$, as $T \to \infty$,*

$$\sup_{0\le t\le T} |A_k(t) - \lambda_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\le t\le T} |V_k(t) - m_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}).$$

*Then we have*

$$\|\nu_k(t)\|_T \overset{a.s.}{=} O(\sqrt{T \log\log T}), \qquad k = 1, \ldots, K,$$

*as $T \to \infty$. If we further assume that*

$$\sup_{0\le t\le T} |V_k(t) - m_k t - \widehat{V}_k(t)| \overset{a.s.}{=} \circ(T^{1/r}),$$

*with $r \in (2, 4)$ and $\widehat{V}_k$ being an $r$-strong continuous process, then we have*

$$\|\nu_k(t)\|_T \overset{a.s.}{=} \circ(T^{1/r}), \qquad k = 1, \ldots, K,$$

*as $T \to \infty$.*

PROOF. We only prove the second half of the lemma; the proof of the first half is almost the same. Use the convention $V_k(-1) = 0$ and $v_k(0) = 0$ in the proof below. From (15), we have for $k = 1, \ldots, K$,

$$\begin{aligned}
\|\nu_k(t)\|_T &\le \sup_{0\le t\le T}\{\max_{1\le n\le A_k(t)} v_k(n)\} \\
&= \sup_{0\le t\le T} v_k(A_k(t)) \\
&= \sup_{0\le t\le T}\{V_k(A_k(t)) - V_k(A_k(t)-1)\} \\
&\le \sup_{0\le t\le T} |V_k(A_k(t)) - m_k A_k(t) + \widehat{V}_k(A_k(t))| \\
&\quad + \sup_{0\le t\le T} |V_k(A_k(t)-1) - m_k(A_k(t)-1) + \widehat{V}_k(A_k(t)-1)| \\
&\quad + \sup_{0\le t\le T} |\widehat{V}_k(A_k(t)) - \widehat{V}_k(A_k(t)-1)| + m_k \\
&\overset{\text{a.s.}}{=} \circ(T^{1/r}),
\end{aligned}$$

where the last equality follows from the strong approximation assumption for $V_k$, the strong continuity of $\widehat{V}_k$, and Lemma 2.3(iii) and (v). □

In view of (16), we can use an argument similar to the one leading to Lemma 3.1 to prove the following lemma.

LEMMA 3.2. *Suppose that for* $k = 1, \dots, K$, *as* $T \to \infty$,

$$\sup_{0 \le t \le T} |A_k(t) - \lambda_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}).$$

*Then*

$$\|\eta_k(t) - t\|_T \overset{a.s.}{=} O(\sqrt{T \log\log T}), \qquad k = 1, \dots, K,$$

*as* $T \to \infty$. *If we further assume that for* $k = 1, \dots, K$, *as* $T \to \infty$,

$$\sup_{0 \le t \le T} |A_k(t) - \lambda_k t - \widehat{A}_k(t)| \overset{a.s.}{=} \circ(T^{1/r}),$$

*with* $r \in (2, 4)$ *and* $\widehat{A}_k$ *being* $r$*-strong continuous, then*

$$\|\eta_k(t) - t\|_T \overset{a.s.}{=} \circ(T^{1/r}), \qquad k = 1, \dots, K,$$

*as* $T \to \infty$.

LEMMA 3.3. *Suppose that the* FLIL *approximations,*

$$\sup_{0 \le t \le T} |A_k(t) - \lambda_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0 \le t \le T} |S_k(t) - \mu_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0 \le t \le T} |V_k(t) - m_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}),$$

*as* $T \to \infty$. *Then there exist* $M_0$ *and* $T_0$ *which are positive such that with probability* 1,

$$\mathscr{T}_k(t) \le M_0 t \quad \text{for } t \ge T_0.$$

PROOF. It follows from the assumptions of the lemma that

$$\|V_k(A_k(t)) - \lambda_k m_k t\|_T \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}).$$

Therefore, there exist positive constants $T_1$ and $a$ such that with probability 1,

$$V_k(A_k(t)) \le \lambda_k m_k t + a\sqrt{t \log\log t} \quad \text{for } t \ge T_1.$$

From (18), we deduce that, with probability 1,

$$\mathscr{T}_k(t) \le Z_{\pi(k)}(t) + \rho_{\pi(k)-1}\mathscr{T}_k(t) + \rho_{\pi(k)-1}t + a\sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i}\sqrt{t\log\log t}$$
$$+V_k(A_k(t)) - V_k(A_k(t)-1) - \sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i} V_j(A_j(t)),$$

for $t \ge T_1$. This implies that with probability 1,

$$(1-\rho_{\pi(k)-1})\mathscr{T}_k(t) \le Z_{\pi(k)}(t) + \rho_{\pi(k)-1}t + a\sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i}\sqrt{t\log\log t}$$
$$+V_k(A_k(t)) - V_k(A_k(t)-1) - \sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i} V_j(A_j(t)),$$

for $t \ge T_1$. Since we assume that $\rho \le 1$, we know that $\rho_{\pi(k)-1} < 1$. The above inequality, together with Lemma 2.3(iii) and (26) in Theorem 3.4, yield the desired result. Even though the proof of Theorem 3.4 (which is to be provided later) makes use of this lemma, the proof of (26) in that theorem does not depend on this lemma and is under the same condition as this lemma. Therefore, the proof is complete. □

3.3. *Functional law-of-iterated-logarithm.* The key result of this section is to show that if the primitive data (the input process) have FLIL approximations, then the departure process (the output process) and the key performance measures of the queue also have FLIL approximations.

To this end, assume that all the primitive data, the exogenous arrival process and the service process, have FLIL approximations: as $T \to \infty$,

$$\sup_{0\le t\le T}|A_k(t) - \lambda_k t| \overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}), \tag{23}$$

$$\sup_{0\le t\le T}|S_k(t) - \mu_k t| \overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}), \tag{24}$$

$$\sup_{0\le t\le T}|V_k(t) - m_k t| \overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}). \tag{25}$$

In fact, it follows from Theorem 2.1 that (25) implies (24). In addition, a sufficient condition for the above approximations is that for each $k \in \mathscr{K}$, the (exogenous) interarrival time sequence $u_k$ and the service time sequence $v_k$ are i.i.d. sequences with finite variances. The main results follow.

THEOREM 3.4. *Suppose that the* FLIL *assumptions* (23)–(25) *hold. Assume that the traffic intensity* $\rho \le 1$. *Then as* $T \to \infty$,

$$\sup_{0\le t\le T} |Z_\ell(t)| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{26}$$

$$\sup_{0\le t\le T} |W_k(t)| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{27}$$

$$\sup_{0\le t\le T} |Q_k(t)| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{28}$$

$$\sup_{0\le t\le T} |\mathcal{S}_k(t)| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{29}$$

$$\sup_{0\le t\le T} |D_k(t) - \lambda_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{30}$$

$$\sup_{0\le t\le T} |Y_\ell(t) - (1-\rho_\ell)t| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{31}$$

$$\sup_{0\le t\le T} |T_k(t) - \beta_k t| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{32}$$

$$\sup_{0\le t\le T} |\tau_\ell(t) - t| \overset{a.s.}{=} O(\sqrt{T \log\log T}), \tag{33}$$

*for all* $\ell \in \mathcal{L}$ *and* $k \in \mathcal{K}$.

REMARK. We note that this theorem holds without assuming the renewal arrival process and the i.i.d. service times. One simple example is to have a compound arrival process (modeling batch arrivals). They hold even without assuming that $A_k$ and $S_k$ are integer-valued, as long as the FLIL approximations (23)–(25) hold. Consider a specific example, where the arrival process $A_k$ takes the form

$$A_k(t) = \int_0^t \alpha_k(s)\, ds$$

and service process $S_k(t) = \mu_k t$. This may represent an ATM communication system, where $\alpha_k(t)$ models the rate at which cells (of class $k$) are generated at time $t$ and $\mu_k$ the rate at which cells (of class $k$) can be processed. [The process $\{\alpha_k(t), t \ge 0\}$ is often modeled by a sum of randomly on–off sources.]

PROOF OF THEOREM 3.4. From the definition of $N_\ell(t)$, we have

$$N_\ell(t) = \sum_{i=1}^{\ell} \sum_{k\in g_i} \{[V_k(A_k(t)) - m_k A_k(t)] + m_k[A_k(t) - \lambda_k t]\} + (\rho_\ell - 1)t.$$

By the FLIL assumptions (23) and (25) and Lemma 2.3(iii), we have

$$\sup_{0\le t\le T} |N_\ell(t) - (\rho_\ell - 1)t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}).$$

Because the pair $(Z_\ell, Y_\ell)$ satisfies the oblique reflection mapping [in view of (20) and (21)], the Lipschitz continuity of the reflection mapping implies (26) and (31).

In view of (11) and (14), we have

$$\begin{aligned}\sum_{k\in g_\ell} W_k(t) &= \sum_{k\in g_\ell} [V_k(A_k(t)) - V_k(D_k(t)) - \nu_k(t)] \\ &= \sum_{k\in g_\ell} \{[V_k(A_k(t)) - m_k A_k(t)] + m_k[A_k(t) - \lambda_k t]\} \\ &\quad - \sum_{k\in g_\ell} \{[V_k(D_k(t)) - m_k D_k(t)] + m_k[A_k(\tau_\ell(t)) - \lambda_k \tau_\ell(t)]\} \\ &\quad + \beta_\ell(t - \tau_\ell(t)) - \sum_{k\in g_\ell} \nu_k(t).\end{aligned}$$

Using Lemma 3.1 and the FLIL assumptions (23) and (25) yields

$$\sup_{0\le t\le T} \left| \sum_{k\in g_\ell} W_k(t) - \beta_\ell(t - \tau_\ell(t)) \right| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

which can be rewritten as

$$\sup_{0\le t\le T} |Z_\ell(t) - Z_{\ell-1}(t) - \beta_\ell(t - \tau_\ell(t))| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}).$$

This combined with (26) implies the FLIL approximation (33) for $\tau_\ell$.

For $k = 1, \ldots, K$, let $\ell = \pi(k)$; the FLIL approximation (30) for the departure process $D_k$ follows from

$$\begin{aligned}\sup_{0\le t\le T} |D_k(t) - \lambda_k t| &= \sup_{0\le t\le T} |A_k(\tau_\ell(t)) - \lambda_k t| \\ &= \sup_{0\le t\le T} |[A_k(\tau_\ell(t)) - \lambda_k \tau_\ell(t)] + \lambda_k(\tau_\ell(t) - t)| \\ &\overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}).\end{aligned}$$

From the relation (14), we can write

$$D_k(t) - \lambda_k t = S(T_k(t)) - \lambda_k t = [S_k(T_k(t)) - \mu_k T_k(t)] + \mu_k[T_k(t) - \lambda_k m_k t].$$

Therefore, the FLIL approximation (32) for $T_k$ can be derived from (30), (24) and the fact that $0 \le T_k(t) \le t$. Similarly, the FLIL approximation (28) for the queue length process $Q_k$ can be proved by observing

$$Q_k(t) = [A_k(t) - \lambda_k t] - [D_k(t) - \lambda_k t].$$

By noting Lemma 3.1, we have that for $k = 1, \ldots, K$,

$$\begin{aligned}\sup_{0\le t\le T}|W_k(t)| &= \sup_{0\le t\le T}|[V_k(A_k(t)) - m_k A_k(t)] + m_k[A_k(t) - \lambda_k t]\\ &\quad -[V_k(D_k(t)) - m_k D_k(t)] + m_k[D_k(t) - \lambda_k t] - \nu_k(t)|\\ &\overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}),\end{aligned}$$

which is the FLIL approximation (27) for the workload process $W_k$.

Finally, we establish the FLIL approximation (29) for the sojourn time process $\mathscr{S}_k$. It follows from (18) and Lemma 3.3 that

$$\begin{aligned}&\|\mathscr{T}_k(t) - \rho_{\pi(k)-1}\mathscr{T}_k(t)\|_T\\ &\quad\le \|Z_{\pi(k)}(t)\|_T + \sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i}\|V_j(A_j(\mathscr{T}_k(t)+t)) - m_j A_j(\mathscr{T}_k(t)+t)\|_T\\ &\qquad + \|V_k(A_k(t)) - V_k(A_k(t)-1)\|_T\\ &\qquad + \sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i} m_j\|A_j(\mathscr{T}_k(t)+t) - \lambda_j(\mathscr{T}_k(t)+t)\|_T\\ &\qquad + \sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i}\|V_j(A_j(t)) - m_j A_j(t)\|_T + \sum_{i=1}^{\pi(k)-1}\sum_{j\in g_i} m_j\|A_j(t) - \lambda_j t\|_T\\ &\quad\overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}).\end{aligned}$$

Therefore, we have

$$\|\mathscr{T}_k(t)\|_T \overset{\text{a.s.}}{=} O\sqrt{T\log\log T}.$$

Since $\mathscr{S}_k(t) = \mathscr{T}_k(\eta_k(t))$, the above combined with Lemma 3.2 yields (29). □

3.4. *Strong approximation.* The key result of this section is to show that if the primitive data of the queue have $r$-strong approximations [for some $r \in (2, 4)$], then the performance measures (such as the workload process, the queue length process and the sojourn time process) and the output process (namely, the departure process) also have $r$-strong approximations.

To this end, we assume that processes $A_k(t), S_k(t), V_k(t)$ are defined on an appropriate probability space such that for some $r \in (2, 4)$ and for $k = 1, \ldots, K$,

$$\sup_{0\le t\le T}|A_k(t) - \lambda_k t - \widehat{A}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{34}$$

$$\sup_{0\le t\le T}|S_k(t) - \mu_k t - \widehat{S}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{35}$$

$$\sup_{0\le t\le T}|V_k(t) - m_k t - \widehat{V}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{36}$$

where $\lambda_k \geq 0,\ \mu_k > 0,\ m_k = 1/\mu_k$, and $\widehat{V}_k(t) = -m_k \widehat{S}_k(m_k t)$. It follows from Theorem 2.2 that (36) implies (35). We shall also assume that $\widehat{A}_k$ and $\widehat{S}_k$ are $r$-strong continuous processes, $k \in \mathscr{K}$. By Theorem 2.2, if we assume that the sequence of service times $v_k$ and the sequence of interarrival times $u_k$ are mutually independent nonnegative i.i.d. sequences having finite $r$th moment with $r \in (2, 4)$, then we can have (34)–(36), with

$$\widehat{A}_k(t) = \lambda_k^{1/2} c_{0,k} B_{0,k}(t), \tag{37}$$

$$\widehat{S}_k(t) = \mu_k^{1/2} c_k B_{1,k}(t), \tag{38}$$

where

$$\begin{aligned} \lambda_k &= 1/E(u_k(n)), \\ \mu_k &= 1/E(v_k(n)), \\ c_{0,k} &= \text{coefficient of variation of } u_k(n), \\ c_k &= \text{coefficient of variation of } v_k(n) \end{aligned}$$

and $B_{0,k}(t)$ and $B_{1,k}(t),\ k \in \mathscr{K}$, are mutually independent standard Brownian motions. (The coefficient of variation of a random variable is its standard deviation divided by its mean.)

THEOREM 3.5. *Suppose that the strong approximation assumptions* (34)–(36) *hold with* $\widehat{A}_k$ *and* $\widehat{S}_k$ *being* $r$*-strong continuous for some* $r \in (2, 4)$. *Assume that the traffic intensity* $\rho \leq 1$. *Then for* $\ell \in \mathscr{L}$ *and* $k \in K$, *as* $T \to \infty$,

$$\sup_{0 \leq t \leq T} |Z_\ell(t) - \widetilde{Z}_\ell(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{39}$$

$$\sup_{0 \leq t \leq T} |\mathscr{S}_k(t) - \widetilde{\mathscr{S}}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{40}$$

$$\sup_{0 \leq t \leq T} |D_k(t) - \widetilde{D}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{41}$$

$$\sup_{0 \leq t \leq T} |Q_k(t) - \widetilde{Q}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{42}$$

$$\sup_{0 \leq t \leq T} |W_k(t) - \widetilde{W}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{43}$$

*where*

$$\widetilde{Z}_\ell(t) = \widetilde{N}_\ell(t) + \widetilde{Y}_\ell(t), \tag{44}$$

$$\widetilde{N}_\ell(t) = (\rho_\ell - 1)t + \sum_{i=1}^{\ell} \sum_{j \in g_i} [m_j \widehat{A}_j(t) - \widehat{V}_j(\lambda_j t)], \tag{45}$$

$$\widetilde{Y}_\ell(t) = \sup_{0 \leq s \leq t} \{-\widetilde{N}_\ell(s)\}^+, \tag{46}$$

$$\widetilde{\mathscr{S}}_k(t) = \frac{\widetilde{Z}_{\pi(k)}(t)}{1-\rho_{\pi(k)-1}}, \tag{47}$$

$$\widetilde{D}_k(t) = \lambda_k t + \widehat{A}_k(t) - \frac{\lambda_k}{\beta_{\pi(k)}}[\widetilde{Z}_{\pi(k)}(t) - \widetilde{Z}_{\pi(k)-1}(t)], \tag{48}$$

$$\widetilde{Q}_k(t) = \frac{\lambda_k}{\beta_{\pi(k)}}[\widetilde{Z}_{\pi(k)}(t) - \widetilde{Z}_{\pi(k)-1}(t)], \tag{49}$$

$$\widetilde{W}_k(t) = \frac{\lambda_k m_k}{\beta_{\pi(k)}}[\widetilde{Z}_{\pi(k)}(t) - \widetilde{Z}_{\pi(k)-1}(t)] = m_k \widetilde{Q}_k(t). \tag{50}$$

REMARKS.

1. By Proposition 2.4, the process $\widetilde{Z}_\ell$ is an $r$-strong continuous process; hence, $\widetilde{\mathscr{S}}_k$, $\widetilde{Q}_k$ and $\widetilde{W}_k$ are $r$-strong continuous. In particular, let

$$\widehat{D}_k(t) = \widehat{A}_k(t) - \frac{\lambda_k}{\beta_{\pi(k)}}[\widetilde{Z}_{\pi(k)}(t) - \widetilde{Z}_{\pi(k)-1}(t)];$$

then $\widehat{D}_k$ is $r$-strong continuous, and the departure process $D_k$ has the strong approximation,

$$\|D_k(t) - \lambda_k t - \widehat{D}_k(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}).$$

This property is essential for the inductive use of the strong approximation of the single-station queue to be extended to a feedforward network case.

2. In Peterson (1991), a state space collapse phenomenon is observed for priority job classes; namely, the workload of high priority group vanishes in the usual heavy traffic normalization. The strong approximation theorem enables us to identify more refined approximation; in this case, it suggests approximating the workload processes of higher priority classes by reflected Brownian motions instead of zero. Indeed, our numerical examples in Section 6 show that the approximations suggested by the strong approximation out performs a straightforward interpretation of the heavy traffic approximation. On the other hand, we can recover the results of heavy traffic limits from the strong approximation by assuming the equalities (37) and (38), and we have

$$Z^*(t) = \widetilde{Z}_L(t), \tag{51}$$

where $Z^*(t)$ is the diffusion approximation of the workload of the station in Peterson (1991). We can also recover the corresponding weak convergence results following a similar approach in Chen and Mandelbaum (1994) that shows how to derive the diffusion limit theorem from the strong approximation theorem for a generalized Jackson network. We note that the strong approximation limit is not unique; it could be any process that differs from the limit in the theorem by an order of magnitude no more than $\circ(T^{1/r})$. Finally, it follows from Theorem 2.5 that when the reflected Brownian

motion with a negative drift ($\rho < e$ case), any constant (including zero) can be the strong approximation limit. Nevertheless, the proposed form of the strong approximation limit is most natural (as can be seen from the proof), and as we shall see through numerical examples, with some refinement, its stationary distribution provides a very good approximation for the stationary distribution of the corresponding quantity in the original queueing network.

3. The second equality in (50) is Little's law for the strong approximation limits of the workload and the queue length process.
4. We would like to point out that, by assuming the equalities (37) and (38), though our approximation for the workload process of the lowest priority group is consistent with the result in Peterson (1991), we have a slightly different approximation for the sojourn time process. For class $k$ jobs which are in the group $g_L$, the diffusion approximation of its sojourn time $\mathscr{S}_k$ in Peterson (1991) is

$$\mathscr{S}_k^*(t) = \frac{Z^*(t)}{\rho_L - \rho_{L-1}}, \tag{52}$$

while in our paper, the strong approximation gives

$$\widetilde{\mathscr{S}}_k(t) = \frac{\widetilde{Z}_L(t)}{1-\rho_L} = \frac{Z^*(t)}{1-\rho_L}. \tag{53}$$

These two formulas are consistent when the traffic intensity at the station is one which is the heavy traffic assumption in the diffusion approximation.

PROOF OF THEOREM 3.5. First, by Lemma 2.3(ii), the strong approximation assumptions (34)–(36) imply the FLIL assumptions (23)–(25); hence, Theorem 3.4 prevails. In the remainder of the proof, we shall repeatedly use Lemma 2.3 without explicitly referring to it.

Next, we rewrite the net-put process as

$$\begin{aligned}
N_\ell(t) &= \sum_{i=1}^{\ell} \sum_{k\in g_i} V_k(A_k(t)) - t \\
&= \sum_{i=1}^{\ell} \sum_{k\in g_i} \Big\{ [V_k(A_k(t)) - m_k A_k(t) + \widehat{V}_k(A_k(t))] \\
&\qquad + m_k[A_k(t) - \lambda_k t - \widehat{A}_k(t)] - [\widehat{V}_k(A_k(t)) - \widehat{V}_k(\lambda_k t)] \\
&\qquad + \lambda_k m_k t + m_k \widehat{A}_k(t) - \widehat{V}_k(\lambda_k t) \Big\} - t \\
&= \sum_{i=1}^{\ell} \sum_{k\in g_i} \Big\{ [V_k(A_k(t)) - m_k A_k(t) + \widehat{V}_k(A_k(t))] \\
&\qquad + m_k[A_k(t) - \lambda_k t - \widehat{A}_k(t)] - [\widehat{V}_k(A_k(t)) - \widehat{V}_k(\lambda_k t)] \Big\} + \widetilde{N}_\ell(t).
\end{aligned}$$

Thus, $N_\ell(t)$ has a strong approximation as

$$\sup_{0\le t\le T}|N_\ell(t)-\widetilde{N}_\ell(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}).$$

Since the pair $(Z_\ell, Y_\ell)$ satisfies the reflection mapping, by the Lipschitz continuity of the reflection mapping, we have

$$\sup_{0\le t\le T}|Z_\ell(t)-\widetilde{Z}_\ell(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

$$\sup_{0\le t\le T}|Y_\ell(t)-\widetilde{Y}_\ell(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

where $\widetilde{Z}_\ell$ and $\widetilde{Y}_\ell$ are defined in (44) and (46), respectively. This proves the strong approximation (39) for the workload process $Z_\ell$.

In view of (11) and (14), we have

$$\begin{aligned}\sum_{k\in g_\ell} W_k(t) = \sum_{k\in g_\ell}\Big\{&[V_k(A_k(t))-m_kA_k(t)-\widehat{V}_k(A_k(t))]+m_k[A_k(t)-\lambda_kt-\widehat{A}_k(t)]\\ &+[\widehat{V}_k(A_k(t))-\widehat{V}_k(\lambda_kt)]-[V_k(D_k(t))-m_kD_k(t)-\widehat{V}_k(D_k(t))]\\ &-m_k[A_k(\tau_\ell(t))-\lambda_k\tau_\ell(t)-\widehat{A}_k(\tau_\ell(t))]-m_k[\widehat{A}_k(\tau_\ell(t))-\widehat{A}_k(t)]\\ &-[\widehat{V}_k(D_k(t))-\widehat{V}_k(\lambda_kt)]+\lambda_km_k[t-\tau_\ell(t)]-\nu_k(t)\Big\}.\end{aligned}$$

Thus, by Lemma 3.1, it follows that

$$\sup_{0\le t\le T}\left|\sum_{k\in g_\ell} W_k(t)-\beta_\ell[t-\tau_\ell(t)]\right| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

which, in view of (39), implies

$$\sup_{0\le t\le T}\left|\widetilde{Z}_\ell(t)-\widetilde{Z}_{\ell-1}(t)-\beta_\ell[t-\tau_\ell(t)]\right| \overset{\text{a.s.}}{=} \circ(T^{1/r}). \tag{54}$$

Now fix $k\in\mathscr{K}$ and $\ell=\pi(k)$. Rewrite $D_k$ as

$$\begin{aligned}D_k(t) &= [A_k(\tau_\ell(t))-\lambda_k\tau_\ell(t)-\widehat{A}_k(\tau_\ell(t))]+[\widehat{A}_k(\tau_\ell(t))-\widehat{A}_k(t)]\\ &\quad+\lambda_kt+\lambda_k[\tau_\ell(t)-t]+\widehat{A}_k(t)\\ &= [A_k(\tau_\ell(t))-\lambda_k\tau_\ell(t)-\widehat{A}_k(\tau_\ell(t))]+[\widehat{A}_k(\tau_\ell(t))-\widehat{A}_k(t)]\\ &\quad+\frac{\lambda_k}{\beta_\ell}\Big\{\beta_\ell(\tau_\ell(t)-t)+[\widetilde{Z}_\ell(t)-\widetilde{Z}_{\ell-1}(t)]\Big\}\\ &\quad-\frac{\lambda_k}{\beta_\ell}[\widetilde{Z}_\ell(t)-\widetilde{Z}_{\ell-1}(t)]+\widehat{A}_k(t)+\lambda_kt.\end{aligned}$$

This, together with (54), yields the strong approximation (41) for the departure process $D_k$.

Similarly, we prove the strong approximation (42) for the queue length process $Q_k$ by observing

$$\sup_{0\le t\le T}|Q_k(t)-\widetilde{Q}_k(t)| = \sup_{0\le t\le T}|[A_k(t)-\lambda_k t=\widehat{A}_k(t)]-[D_k(t)-\widetilde{D}_k(t)]|$$
$$\stackrel{\text{a.s.}}{=}\circ(T^{1/r}) \quad \text{as } T\to\infty,$$

and the strong approximation (43) for the workload process $W_k$ by observing

$$\begin{aligned}
&W_k(t)-\widetilde{W}_k(t)\\
&\quad=[V_k(A_k(t))-m_kA_k(t)-\widehat{V}_k(A_k(t))]+m_k[A_k(t)-\lambda_k t-\widetilde{A}_k(t)]\\
&\qquad+[\widehat{V}_k(A_k(t))-\widehat{V}_k(\lambda_k t)]-[V_k(D_k(t))-m_kD_k(t)-\widehat{V}_k(D_k(t))]\\
&\qquad-m_k[D_k(t)-\widetilde{D}_k(t)]-[\widehat{V}_k(D_k(t))-\widehat{V}_k(\lambda_k t)]-\nu_k(t).
\end{aligned}$$

Finally, we establish the strong approximation (40) for the sojourn time process $\mathscr{S}_k$. Note that $\mathscr{S}_k(t)=\mathscr{T}_k(\eta_k(t))$; we first rewrite (18),

$$\begin{aligned}
\mathscr{T}_k(t)=&\widetilde{Z}_{\pi(k)}(t)+\rho_{\pi(k)-1}\mathscr{T}_k(t)+[Z_{\pi(k)}(t)-\widetilde{Z}_{\pi(k)}(t)]\\
&+\sum_{i=1}^{\pi(k)-1}\sum_{j\in\mathscr{I}_i}\Big\{[V_j(A_j(\mathscr{T}_k(t)+t))\\
&\qquad -m_jA_j(\mathscr{T}_k(t)+t)-\widehat{V}_j(A_j(\mathscr{T}_k(t)+t))]\\
&\qquad +m_j[A_j(\mathscr{T}_k(t)+t)-\lambda_j(\mathscr{T}_k(t)+t)-\widehat{A}_j(\mathscr{T}_k(t)+t)]\\
&\qquad +m_j[\widehat{A}_j(\mathscr{T}_k(t)+t)-\widehat{A}_j(t)]\\
&\qquad +[\widehat{V}_j(A_j(\mathscr{T}_k(t)+t))-\widehat{V}_j(\lambda_j t)]\\
&\qquad -[V_j(A_j(t))-\lambda_jm_jt-m_j\widehat{A}_j(t)-\widehat{V}_j(\lambda_jt)]\Big\}\\
&+[V_k(A_k(t))-\lambda_km_kt-m_k\widehat{A}_k(t)-\widehat{V}_k(\lambda_kt)]\\
&-[V_k(A_k(t)-1)-\lambda_km_kt-m_k\widehat{A}_k(t)-\widehat{V}_k(\lambda_kt)].
\end{aligned}$$

Note that $\|\mathscr{T}_k(t)\|=O(\sqrt{T\log\log T})$ a.s. as shown in the proof of Theorem 3.4; in view of Lemma 2.3 and the strong approximations for $Z_{\pi(k)}$, $V_j$ and $A_j$, we obtain from the above equality,

$$\left\|\mathscr{T}_k(t)-\rho_{\pi(k)-1}\mathscr{T}_k(t)-\widetilde{Z}_{\pi(k)}(t)\right\|_T\stackrel{\text{a.s.}}{=}\circ(T^{1/r}) \quad \text{as } T\to\infty,$$

or equivalently,

$$\left\|\mathscr{T}_k(t)-\frac{1}{1-\rho_{\pi(k)-1}}\widetilde{Z}_{\pi(k)}(t)\right\|_T\stackrel{\text{a.s.}}{=}\circ(T^{1/r}) \quad \text{as } T\to\infty.$$

The latter, together with Lemmas 3.2 and (17), implies (40). □

3.5. *Fine tuning the strong approximation for sojourn time.* A more detailed study reveals that the strong approximation of sojourn time that we obtained in Theorem 3.5 should be better interpreted as the strong approximation to the *sojourn queue time* which is the time between the arrival of the job and the time that it just begins service. The reason is that, in the proof of FSAT approximation for the sojourn time process we approximate $V_k(A_k(t))-V_k(A_k(t)-1)$, the service time, by zero. If we replace $V_k(A_k(t)) - V_k(A_k(t) - 1)$ by zero in (18), then the new $\mathscr{T}_k$ is exactly the sojourn queue time.

Thus, the strong approximation for the sojourn queue time is same as that for the sojourn time given by Theorem 3.5. As we know, the sojourn time should be larger than the sojourn queue time. For a single-class single-station queue, the sojourn time of a job equals its sojourn queue time plus its service time. For preemptive priority queueing networks, even if a job is in service, it may well be interrupted by another arriving job with higher priority. Hence, in general, the sojourn time of a job should be longer than or equal to its sojourn queue time plus its service time. Our numeric examples suggest that approximating the service time by its mean would yield an improved strong approximation for the sojourn time (47),

$$\widetilde{\mathscr{S}}_k(t) = \frac{1}{1-\rho_{\pi(k)-1}}[\widetilde{Z}_{\pi(k)}(t) + m_k]. \tag{55}$$

[We note that we might obtain a better approximation by using a true service time random variable in place of $m_k$ in the above; this will be discussed in Shen (2000).]

Because of the nature of strong approximation, we could replace $m_k$ in (55) by any constant and the strong approximation still holds. However, there is strong reason to believe that $m_k$ is the best constant to put in (55). Our numerical examples in Section 6 show that (55) gives fairly good approximation to the steady-state average sojourn time. For $M/G/1$ preemptive priority queue, the approximated steady-state mean sojourn time is

$$E\mathscr{S}_k = \frac{1}{1-\rho_{\pi(k)-1}}\left(\frac{\sum_{l=1}^{\pi(k)}\sum_{i\in\mathscr{I}_l} m_i^2(1+b_i^2)}{2(1-\rho_{\pi(k)})} + m_k\right), \tag{56}$$

which is the same as the exact mean sojourn time; see (3.39) in Kleinrock (1976).

3.6. *A packet queue application.* Our strong approximation in Theorem 3.5 provides a framework to evaluate the performance of single-station multiclass queues. It holds without assuming renewal arrival processes or renewal service times. Here, we provide as an example a batch renewal model which was used by Fendick, Saksena and Whitt (1989) to study the dependence in packet communication networks. As we know, a batch renewal process is a renewal process if and only if it is a batch-Poisson process and the batch size is geometrically distributed on the positive integers. Typically, the superposition process

of batch renewal processes is not renewal. The strong approximation theorem can be applied to them to get performance measures easily.

Let $u_k = \{u_k(n), n \geq 1\}$ be i.i.d. batch interarrival times, with mean $\lambda_k$ and squared coefficient of variation (SCV) $c_{0k}^2$; let $v_k = \{v_k(n), n \leq 1\}$ be i.i.d. service times, with mean $m_k$ and squared coefficient of variation $c_k^2$; let $\eta_k = \{\eta_k(n),\ n \geq 1\}$ be i.i.d. batch sizes with $\eta_k(n)$ being the $n$th batch size of class $k$ packets and with mean $b_k$ and squared coefficient of variation $c_{bk}^2$. Furthermore, we assume that all three of them have finite $r$th moments with $r > 2$ and are mutually independent. All classes belong to one priority group, so service discipline is strictly FIFO.

Let $A_k(t)$ denote the arrival process of class $k$ packets,

$$X_k(t) = \sum_{i=1}^{\lfloor t \rfloor} u_k(i), \qquad t \geq 0,$$

$$Y_k(t) = \sup\{s \geq 0\colon X_k(s) \leq t\}, \qquad 0 \leq t < \infty,$$

$$U_k(t) = \sum_{i=1}^{\lfloor t \rfloor} v_k(i), \qquad t \geq 0, \qquad V_k(t) = \sum_{i=1}^{\lfloor t \rfloor} \eta_k(i), \qquad t \geq 0,$$

with $X(T) = 0, U(t) = 0$ and $V_k(t) = 0$ for $t < 1$. Thus

$$A(t) = U(Y(t)).$$

By Theorem 2.2 and Lemma 2.3, there exist three independent standard Brownian motions $B_k^X$, $B_k^U$ and $B_k^V$ such that

$$\left\| U_k(t) - b_k t - b_k^{1/2} c_{bk} B_k^U(b_k t) \right\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r'}),$$

$$\left\| V_k(t) - m_k t - m_k^{1/2} c_k B_k^V(m_k t) \right\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r'}),$$

$$\left\| Y_k(t) - \lambda_k t - \lambda_k^{1/2} c_{0k} B_k^X(t) \right\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r'}),$$

$$\left\| A_k(t) - \lambda_k b_k t - \lambda_k^{1/2} b_k c_{0k} B_k^X(t) - b_k^{1/2} c_{bk} B_k^U(\lambda_k b_k t) \right\| \overset{\text{a.s.}}{=} \circ(T^{1/r'}),$$

where $r' = r$ if $r < 4$, and $r' < 4$ if $r \geq 4$.

Therefore, by Theorem 3.5, the strong approximation for the total workload process at the station is

$$\widetilde{Z}(t) = \widetilde{N}(t) + \widetilde{Y}(t),$$

$$\widetilde{N}(t) = (\rho - 1)t + \sum_{k=1}^{K} \Big\{ m_k \lambda_k^{1/2} b_k c_{0k} B_k^X(t) + m_k b_k^{1/2} c_{bk} B_k^U(\lambda_k b_k t) - m_k^{1/2} c_k B_k^V(m_k \lambda_k b_k t)] \Big\},$$

$$\widetilde{Y}_\ell(t) = \sup_{0 \leq s \leq t} \left\{ - \widetilde{N}_\ell(s) \right\}^+,$$

where $\rho = \Sigma_{k=1}^{K} \lambda_k b_k m_k$. This is a one-dimensional reflected Brownian motion. In particular, we can get approximated steady-state mean work load by

$$EZ = \frac{\sum_{k=1}^{K} \lambda_k m_k^2 [b_k c_k^2 + b_k^2 (c_{0k}^2 + c_{bk}^2)]}{2(1-\rho)}.$$

Fendick, Saksena and Whitt (1989) obtained the same result by using a heavy traffic limit theorem.

**4. Multiclass feedforward networks.** In this section, we shall assume the FLIL approximation without much discussion and focus on the strong approximation. The reasons are that the derivation of the FLIL approximation for the network case is quite similar to that for the single-station case and that the strong approximation yields more useful approximation.

4.1. *Queueing network model.* We first describe the primitive data and then the performance measures and their dynamics.

4.1.1. *Primitive data and assumptions.* The queueing network consists of a set of $J$ service stations, indexed by $j = 1, \ldots, J$, serving $K$ classes of jobs, indexed by $k = 1, \ldots, K$. There are $L$ priority groups, indexed by $\ell$, $\ell = 1, \ldots, L$, and $g_\ell$ is the set of all job classes belong to group $\ell$. Let $\pi(\cdot)$ be a many-to-one mapping from class indices to group indices; specifically, job class $k$ belongs to the priority group $\pi(k)$. Jobs from group $\ell$ ($\ell = 1, \ldots, L$) are served exclusively at station $j = \sigma(\ell)$, where $\sigma(\cdot)$ is a many-to-one mapping from group indices to station indices. While each group is served at one station exclusively, each station may serve more than one group. For simplicity, we define $\sigma(0) \equiv 0$. Note that the composition $\sigma \circ \pi$ is a many-to-one mapping from class indices to station indices. If $\ell < m$, then jobs in group $\ell$ are assumed to have a preemptive priority over jobs in group $m$($\ell, m = 1, \ldots, L$). Within a group, jobs of all classes are served in the order of arrival, that is, first-in first-out (FIFO). The network is a feedforward queueing network in the sense that any job at station $i$ can turn into another class at station $j$ only if $j > i$ ($i, j = 1, \ldots, J$). To illustrate our notation, consider the network given by Figure 1, which has $J = 2$ stations serving $K = 6$ classes of jobs with $L = 4$ priority groups. Job class 1 belongs to priority group 1; job classes 2 and 3 belong to priority group 2; job class 4 belongs to priority group 3 and job classes 5 and 6 belong to priority group 4. Priority groups 1 and 2 reside at station 1 and all the other groups reside at station 2. Then, $\pi$, $g$ and $\sigma$ defined above can be written as

$$\begin{aligned}
&\pi(1) = 1, \qquad \pi(2) = \pi(3) = 2,\\
&\pi(4) = 3, \qquad \pi(5) = \pi(6) = 4;\\
&g_1 = \{1\}, \qquad g_2 = \{2, 3\}, g_3 = \{4\}, g_4 = \{5, 6\};\\
&\sigma(1) = \sigma(2) = 1, \sigma(3) = \sigma(4) = 2.
\end{aligned}$$

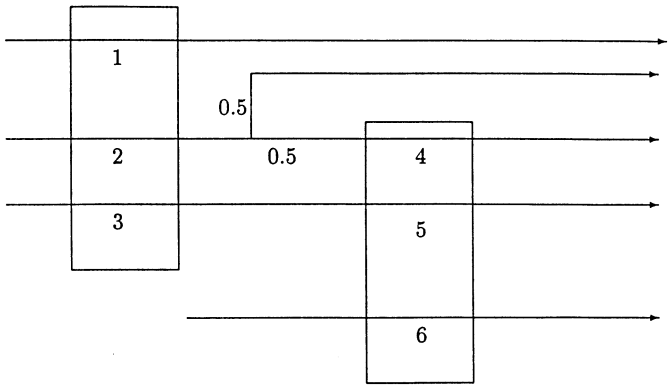


FIG. 1. *A multiclass feedforward queue network.*

To facilitate our analysis, we make some assumptions on the way of indexing priority groups and job classes. As we will see later, this is critical to obtaining a compact presentation of the main results.

1. Every station has at least one priority group, and every priority group has at least one job class; hence, necessarily, $K \geq L \geq J$.
2. For $k, j = 1, \ldots, K$, $\pi(k) \leq \pi(j)$ if $k < j$. Therefore, jobs of class 1 must be in group 1 and jobs of class $K$ must be in group $L$.
3. For $\ell, m = 1, \ldots, L$ $\sigma(\ell) \leq \sigma(m)$ if $\ell < m$. Thus, jobs from group 1 must be served in station 1, and jobs from group $L$ must be served in station J.

The queueing network is described by the following primitive data: $2K$ sequences of nonnegative random variables $u_k = \{u_k(n), n \geq 1\}$ and $v_k = \{v_k(n), n \geq 1\}(k = 1, \ldots, K)$, and $K$ sequences of $K$-dimensional vector $\phi^k = \{\phi^k(n), n \geq 1\}(k = 1, \ldots, K)$, all defined on the same probability space.

We assume that there are no jobs in the network at time $t = 0$. The random variable $u_k(1)$ is the time of the first exogenously arrived class $k$ job, and $u_k(n), n > 1$, is the time between the $(n-1)$st and $n$th exogenous arrived class $k$ jobs. The random variable $v_k(n)$ is the service time required for the $n$th class $k$ jobs. The random variable $\phi^k$ describes the routing mechanism for class $k$ jobs: the $n$th class $k$ job after service completion turns into a class $j$ job if $\phi^k(n) = e^j$ and leave the network if $\phi^k(n) = 0$. By the feedforward structure and our numbering convention, it follows that, for all $n \geq 1$ and $k, i = 1, \ldots, K$, the class transitions must satisfy $\phi^k(n) \neq e^i$ if $\sigma(\pi(i)) \leq \sigma(\pi(k))$.

We introduce the summations,

$$U_k(0) = 0, \; U_k(n) = \sum_{m=1}^{n} u_k(m), n \geq 1, \qquad k = 1, \ldots, K,$$

$$V_k(0) = 0, \; V_k(n) = \sum_{m=1}^{n} v_k(m), n \geq 1, \qquad k = 1, \ldots, K,$$

$$\Phi^k(0) = 0,\ \Phi^k(n) = \sum_{m=1}^{n} \phi^k(m), n \geq 1, \qquad k = 1, \ldots, K.$$

Define their associated counting processes

$$E_k(t) = \sup\{n \geq 0\colon U_k(n) \leq t\}, \qquad k = 1, \ldots, K,$$
$$S_k(t) = \sup\{n \geq 0\colon V_k(n) \leq t\}, \qquad k = 1, \ldots, K.$$

Let $U = (U_k)$, $V = (V_k)$, $\Phi = (\Phi^1, \ldots, \Phi^K)$, $E = (E_k)$ and $S = (S_k)$. We call $E$ an exogenous arrival process, $S$ a service process, and $\Phi$ a routing sequence. Note that we do not assume that the arrival process $E$, the service process $S$ and the routing processes are renewal processes.

Similar to the single-station queueing model above, we assume that there exist a long-run average arrival rate, an average service time and a long-run average transition (routing) rate; namely,

$$\frac{E(t)}{t} \to \alpha \quad \text{as } t \to \infty,$$
$$\frac{V(n)}{n} \to m \quad \text{as } n \to \infty,$$
$$\frac{\Phi^k(n)}{n} \to P_k' \quad \text{as } n \to \infty,$$

where $P_k'$ is the $k$th row of a $K \times K$ matrix $P = (p_{kj})$. We shall call $\alpha_k$, the $k$th coordinate of $\alpha$, the (exogenous) arrival rate of class $k$ job and call $m_k$, the $k$th coordinate of $m$, the average service time of class $k$ job [alternatively the mean service time of class $k$ job when $v_k(n)$ has the same finite mean for all $n \geq 1$]. Call $p_{kj}$, the $j$th coordinate of $P_k'$ [and the $(k, j)$th element of $P$], the average transition rate that a class $k$ job turns into a class $j$ job after completing its service. When $\Phi^K$ is an i.i.d. summation, $p_{kj}$ is the probability that a class $k$ job turns into a class $j$ job after its service completion. We assume that for all $1 \leq k \leq K$, $m_k > 0$ and call $\mu_k := 1/m_k$ the service rate of class $k$. By our assumption on the routing sequence, it follows that matrix $P$ is a strictly upper triangular matrix. For the network shown in Figure 1, the routing matrix takes the form

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where it is assumed that upon service completion, any class 3 job will turn into a class 5 job, a class 2 job will either leave the network or turn into a class 4 job with equal probability and jobs of all other classes will leave the network.

We assume that the primitive processes $(V, E, \Phi^1, \ldots, \Phi^K)$ have strong approximations; namely, we assume that they are defined on a probability space such that there exist a $K(K+2)$-dimensional $r$-continuous process $(\widehat{V}, \widehat{E}, \widehat{\Phi}^1, \ldots, \widehat{\Phi}^K)$ satisfying

$$\|V(t) - mt - \widehat{V}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{57}$$

$$\|E(t) - \alpha t - \widehat{E}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{58}$$

$$\|\Phi^k(t) - P_k' t - \widehat{\Phi}^k(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \qquad k = 1, \ldots, K, \tag{59}$$

as $T \to \infty$, for some $r \in (2, 4)$. We note, in particular, that we neither assume that $(\widehat{V}, \widehat{E}, \widehat{\Phi}^1, \ldots, \widehat{\Phi}^K)$ is a Brownian motion, nor assume its components $\widehat{V}$, $\widehat{E}$ and $\widehat{\Phi}^1, \ldots, \widehat{\Phi}^K$ are mutually independent.

However, if $u_k$, $v_k$ and $\phi^k$, $k = 1, \ldots, K$, are mutually independent i.i.d. sequences, and $u_k$ and $v_k$ have finite moments of order $r \in (2, 4)$, then by a multidimensional generalization of Theorem 2.2, the strong approximation assumptions (57)–(59) hold with $m_k$ (the $k$th component of $m$), $1/\alpha_k$ (where $\alpha_k$ is the $k$th component of $\alpha$) being the means of random variables $v_k(1)$ and $u_k(1)$, respectively, with $p_{kj} = \boldsymbol{P}\{\phi^k(n) = e^j\}$ (the $j$th component of $P_k'$) and with $\widehat{V}$, $\widehat{E}$, $\widehat{\Phi}^1, \ldots, \widehat{\Phi}^K$ being mutually independent driftless Brownian motions. The covariance matrices of these Brownian motions are, respectively,

$$\begin{aligned}
(\Gamma_E)_{i\ell} &= \delta_{i\ell}\alpha_\ell c_{0,\ell}^2, \\
(\Gamma_\Phi^k)_{i\ell} &= p_{ki}(\delta_{i\ell} - p_{k\ell}), \qquad k = 1, \ldots, K, \\
(\Gamma_V)_{i\ell} &= \delta_{i\ell} m_\ell^2 c_\ell^2,
\end{aligned}$$

where $c_{0,k}$ and $c_k$ are the coefficients of variations, of random variables, $u_k(1)$ and $v_k(1)$, respectively, and $\delta_{i\ell} = 1$ if $i = \ell$ and $\delta_{i\ell} = 0$ otherwise.

Our model has a slightly more general structure than the one described in Peterson (1991). In particular, we allow the routing sequences to include Markovian routing, while Peterson (1991) only considers the deterministic routing.

4.1.2. *Performance measures and their dynamics.* The performance measures of interest are the $L$-dimensional (*aggregated*) workload process $Z = (Z_\ell)$ with $Z_\ell = \{Z_\ell(t),\ t \geq 0\}$ $(\ell = 1, \ldots, L)$, the $K$-dimensional workload process $W = (W_k)$ with $W_k = \{W_k(t),\ t \geq 0\}$ $(k = 1, \ldots, K)$, the $K$-dimensional queue length process $Q = (Q_k)$ with $Q_k = \{Q_k(t),\ t \geq 0\}$ $(k = 1, \ldots, K)$, and the $L$-dimensional cumulative idle time process $Y = (Y_\ell)$ with $Y_\ell = \{Y_\ell(t),\ t \geq 0\}$ $(l = 1, \ldots, L)$. The process $Z$ is nonnegative with $Z_\ell(t)$ indicating the total amount of immediate work for station $\sigma(\ell)$ embodied in jobs that are in groups 1 to $\ell$ and that are either queued or in service at station $\sigma(\ell)$ at time $t$. The quantity $W_k(t)$ indicates the amount of work embodied in all class $k$ jobs that are either queued or in service at time $t$. The quantity $Q_k(t)$ indicates the number of class $k$ jobs in the network at time $t$. We assume that $Q(0) = 0$ and thus $Z(0) = 0$. The quantity $Y_\ell(t)$ indicates the cumulative

amount of time that the server at station $\pi(\ell)$ does not process jobs in groups 1 to $\ell$ during $[0, t]$. It is clear that $Y$ must be nondecreasing and $Y(0) = 0$.

We introduce some additional notation.

1. $A_k(t)$ is the total number of class $k$ jobs arrived to station $\sigma(\pi(k))$ during $[0, t]$ either exogenously or from other stations.
2. $D_k(t)$ is the total number of service completions of class $k$ jobs at station $\sigma(\pi(k))$ during $[0, t]$.
3. $T_k(t)$ is the total amount of time that server $\sigma(\pi(k))$ has served jobs of class $k$ during $[0, t]$.
4. $\tau_\ell(t)$ is the arrival time of the $g_\ell$ job which has most recently completed service at station $\sigma(\ell)$ ($\tau_\ell(t)$ is zero if there have been no service completions for group $\ell$).
5. $\mathscr{S}_k(t)$ is the sojourn time of class $k$ jobs at time $t$ at station $\sigma(\pi(k))$, denoting the time which will be spent at station $\sigma(\pi(k))$ by the first class $k$ job who arrives at time greater than or equal to $t$.
6. $\eta_k(t)$ is the time at which the first class $k$ job arrives during $[t, \infty)$.
7. $\mathscr{T}_k(t)$ is the time that a class $k$ job would spend at station $\sigma(\pi(k))$, if it arrived at time $t$.

Define two $L \times K$ matrices: a higher priority group constituent matrix $C$ and a group constituent matrix $C_1$. The $(\ell, k)$th component of $C$, $C_{\ell k} = 1$ if $\sigma(\ell) = \sigma(\pi(k))$ and $\pi(k) \le \ell$, and $C_{\ell k} = 0$ otherwise. The $(\ell, k)$th component of $C_1$, $C_{1\ell k} = 1$ if $\pi(k) = \ell$, and $C_{1\ell k} = 0$ otherwise. Define a $K \times K$ (strictly) higher priority class constituent matrix $C_2 = (C_{2ij})$ with $C_{2ij} = 1$ if $\sigma(\pi(i)) = \sigma(\pi(j))$ and $\pi(i) > \pi(j)$, and $C_{2ij} = 0$ otherwise. Consider the example shown in Figure 1; under the priority group specification given earlier, the higher priority group constituent matrix $C$, the group constituent matrix $C_1$, and the (strictly) higher priority class constituent matrix $C_2$, respectively, take the form

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}, \qquad C_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

$$C_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Let $\lambda = (I - P')^{-1}\alpha$ and call its $k$th component, $\lambda_k$, the long-run average arrival rate of class $k$ jobs, $k = 1, \ldots, K$. Let $M = \mathrm{diag}(m)$ and $\Lambda = \mathrm{diag}(\lambda)$ be $K \times K$ diagonal matrices with $k$th diagonal elements $m_k$ and $\lambda_k$, respectively. Let $\rho = CM\lambda$. Note that $\rho$ is of dimension $L$; if $g_\ell$ is the lowest priority group at its station $\sigma(\ell)$, then $\rho_\ell$ is the traffic intensity at that station. We shall

assume that the traffic intensity at all stations are no greater than one and hence, $\rho \le e$. Let $\delta_k = (1 - \rho_{\pi(k-1)})^{-1}$ if $\sigma(\pi(k-1)) = \sigma(\pi(k))$ and $\delta_k = 1$ if $\sigma(\pi(k-1)) < \sigma(\pi(k))$, and let $\Delta = \mathrm{diag}(\delta)$ be a $K \times K$ diagonal matrix whose $k$th diagonal element is $\delta_k$.

It is from the above definitions that we have the following dynamic relations:

$$Q(t) = A(t) - D(t), \tag{60}$$

$$W(t) = V(A(t)) - V(D(t)) - \nu(t), \tag{61}$$

$$Y(t) = et - CT(t), \tag{62}$$

$$Z(t) = CW(t), \tag{63}$$

$$Z(t) = N(t) + Y(t), \tag{64}$$

$$N(t) = CV(A(t)) - et,$$

$$D(t) = A(C_1'\tau(t)), \tag{65}$$

$$A(t) = E(t) + \sum_{k=1}^{K} \Phi^k(D_k(t)), \tag{66}$$

$$\mathscr{S}(t) = \mathscr{T}(\eta(t)), \tag{67}$$

$$\begin{aligned}\mathscr{T}_k(t) = {} & Z_{\pi(k)}(t) + C_{2k}'[V(A(\mathscr{T}_k(t) + t)) - V(A(t))] \\ & + V_k(A_k(t)) - V_k(A_k(t) - 1),\end{aligned} \tag{68}$$

where $\nu_k(t)$, the $k$th component of $\nu(t)$, is the partial service time (if any) that has been performed on the class $k$ job during $(\tau_{\pi(k)}, t]$, which is dominated by an inequality similar to (15). In (68), $C_{2k}'$ is the $k$th row of matrix $C_2$, and from the context, we hope it will not be confused with the $(2, k)$th element of matrix $C$. For understanding the above relations, it is helpful to compare them with the relations (10)–(18) for the single-station case. In particular, relation (64) is a workload flow balance relation (in terms of time) for service stations, and relation (60) is a job flow balance relation (in terms of number of jobs) for job classes. We shall assume that the work-conserving condition is in force. Hence, the pair $(Z, Y)$ satisfies the reflection mapping relation, which implies that

$$Y(t) = \sup_{0 \le s \le t} [-N(s)] = \sup_{0 \le s \le t} [es - CV(A(s))]. \tag{69}$$

4.2. *Main result.* For the queueing network model described in Section 4.1, jobs can route from station $i$ to station $j$ only if $j > i$. Now we argue that given the strong approximations (57)–(59) for the primitive data, we could inductively apply Theorem 3.5 (the strong approximation theorem for a single station) to the network from stations 1 to station $J$. First, by Theorem 3.5 and Remark 1 after it, the departure process of each job class from station 1 has a strong approximation; this, the assumption (59) (that the routing sequence

has a strong approximation) and Lemma 2.3(vi) imply that the arrival processes to station 2 from station 1 have strong approximations. Since jobs arrive at station 2 either exogenously or from station 1, the total arrival process to station 2 for each job class must have a strong approximation as well. Hence, by applying Theorem 3.5 to station 2, we know in particular that the departure process of each job class from station 2 satisfies a strong approximation. Inductively, we can show that the departure process and the arrival process for each class in the network must have some strong approximations. Therefore, we could apply Theorems 3.5 to each station to obtain the strong approximations for all the performance measures, especially the workload process of each job class, the aggregated workload processes, the queue length processes and the sojourn time processes. The following theorem presents the strong approximations in a compact form.

THEOREM 4.1. *Suppose that the strong approximations* (57)–(59) *hold. Let*

$$H = \Lambda C_1'(CM\Lambda C_1')^{-1},$$
$$G = CM(I - P')^{-1}P'H$$

*and*

$$R = (I + G)^{-1}.$$

*Then, as* $T \to \infty$,

$$\|Z(t) - \widetilde{Z}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{70}$$

$$\|Y(t) - \widetilde{Y}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{71}$$

$$\|Q(t) - \widetilde{Q}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{72}$$

$$\|W(t) - \widetilde{W}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{73}$$

$$\|\mathscr{S}(t) - \widetilde{\mathscr{S}}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{74}$$

*where*

$$\widetilde{Q}(t) = H\widetilde{Z}(t) \quad \textit{for } t \geq 0, \tag{75}$$

$$\widetilde{W}(t) = MH\widetilde{Z}(t) = M\widetilde{Q}(t), \tag{76}$$

$$\widetilde{\mathscr{S}}(t) = \Delta\big[C_1'\widetilde{Z}(t) + m\big]; \tag{77}$$

*and* $(\widetilde{Z}, \widetilde{Y})$ *are defined as follows:*

$$\widetilde{Z}(t) = \theta t + \widetilde{X}(t) + R\widetilde{Y}(t) \geq 0 \quad \textit{for } t \geq 0, \tag{78}$$

$$\theta = R(\rho - e), \tag{79}$$

$$\widetilde{X}(t) = RC\left[\widehat{V}(\lambda t) + M(I - P')^{-1}\big[\widehat{E}(t) + \sum_{k=1}^{K}\widehat{\Phi}^k(\lambda_k t)\big]\right]. \tag{80}$$

(81) $\widetilde{Y}(t)$ *is continuous and nondecreasing with* $\widetilde{Y}(0) = 0$,

(82) $$\int_0^\infty \widetilde{Z}_\ell(t)d\widetilde{Y}_\ell(t) = 0 \quad \textit{for } l = 1, \ldots, L,$$

REMARKS.

1. It is shown in the Appendix (Lemma A.1) that the matrix $H$ is well defined. In fact, let $\beta = C_1 M\lambda$; then the $(k, \ell)$th element of the $K \times L$ matrix $H$ is given by
$$H_{k\ell} = \begin{cases} \dfrac{\lambda_k}{\beta_{\pi(k)}}, & \text{if } \ell = \pi(k), \\ -\dfrac{\lambda_k}{\beta_{\pi(k)}}, & \text{if } \ell = \pi(k) - 1 \text{ and } \sigma(l) = \sigma(\pi(k)), \\ 0, & \text{otherwise.} \end{cases}$$
[The $\ell$th component of $\beta$, $\beta_\ell$, is the traffic intensity of priority group $\ell$ at station $\sigma(\ell)$.] It is also shown in the Appendix that the matrix $G$ is well defined and is strictly lower triangular. Hence, matrix $R$ is also well defined and is lower triangular.
2. Since matrix $R$ is triangular, by inductively applying the one-dimensional reflection mapping, it is clear that for the given $\theta$ in (79) and $\widetilde{X}$ in (80), relations (78), (81) and (82) uniquely determine the process $\widetilde{Z}$ and $\widetilde{Y}$. In particular, when the interarrival sequences $u_k$ $(k = 1, \ldots, K)$, the service sequences $v_k$ $(k = 1, \ldots, K)$ and the routing sequences $\phi^k$ $(k = 1, \ldots, K)$ are mutually independent i.i.d. sequences, the process $\widetilde{X}$ is a Brownian motion and the process $\widetilde{Z}$ is a reflected Brownian motion with reflection mapping $R$. The covariance matrix of the Brownian motion $\widetilde{X}$ in this case is given by
$$\Gamma = RC\left[\Gamma_V\Lambda + M(I - P')^{-1}\left[\Gamma_E + \sum_{k=1}^K \lambda_k\Gamma_\Phi^k\right](I - P)^{-1}M\right]C'R', \tag{83}$$
where $\Gamma_E$, $\Gamma_V$ and $\Gamma_\Phi$ are as given toward the end of Section 4.1.1. When the sequences $u_k$, $v_k$ and $\phi^k$, $k = 1, \ldots, K$, are i.i.d. but not mutually independent, $\widetilde{X}$ given in (80) is still a Brownian motion but with its covariance matrix computed differently.
3. The second equality in (76) is Little's law for strong approximation limits.

PROOF OF THEOREM 4.1. In view of the previous discussion, we only need to show that the strong approximation limits in (70)–(74) are given by (75)–(82). Specifically, the starting points of our proof are the following results: first the FLIL approximations hold, in particular,

$$\|A(t) - \lambda t\|_T \overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}), \tag{84}$$

$$\|D(t) - \lambda t\|_T \overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}), \tag{85}$$

$$\|\tau(t) - et\|_T \overset{\text{a.s.}}{=} O(\sqrt{T\log\log T}), \tag{86}$$

$$\|\mathscr{T}(t)\|_T \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}). \tag{87}$$

Second, the strong approximations (57)–(59), the strong approximations (70)–(74) and

$$\|A(t) - \lambda t - \widehat{A}(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{88}$$

$$\|\nu(t)\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{89}$$

$$\|\eta(t) - et\|_T \overset{\text{a.s.}}{=} \circ(T^{1/r}), \tag{90}$$

hold. That the FLIL approximations (84)–(87) hold can be proved under the FLIL approximation assumptions for the primitive data which are implied by the strong approximation assumptions (57)–(59); the proof is similar to that for Theorem 3.4 and will not be provided. That the strong approximations (70)–(74) and (88) hold for some limits $\widetilde{Z}$, $\widetilde{Y}$, $\widetilde{Q}$, $\widetilde{W}$, $\widetilde{\mathscr{S}}$, and $\widehat{A}$ follows from an induction proof as outlined before the theorem. That (89) and (90) hold is an extension of Lemmas 3.1 and 3.2, respectively. What remains now is to identify limits $\widetilde{Z}$, $\widetilde{Y}$, $\widetilde{Q}$, $\widetilde{W}$, $\widetilde{\mathscr{S}}$ and $\widehat{A}$ and show that they satisfy (75)–(82).

For any two processes $X$ and $Y$, if $\|X - Y\|_T = \circ(T^{1/r})$ a.s., we shall write "$X(t) \approx Y(t)$" to simplify the presentation. With this notation, if $Y(t) \approx \widetilde{X}(t)$, then $Y$ is a strong approximation for $X$ implies that $\widetilde{X}$ is also a strong approximation for $X$. In the following proof, we shall repeatedly use without explicitly referring to the FLIL approximations (84)–(87), the strong approximations (57)–(59) and Lemma 2.3.

First, note that $\lambda = \Lambda C_1' e$; from (61), (65) and (89), we have

$$\begin{aligned} W(t) &= V(A(t)) - V(D(t)) - \nu(t) \\ &\approx [MA(t) + \widehat{V}(A(t))] - [MD(t) + \widehat{V}(D(t))] \\ &\approx [M\lambda t + M\widehat{A}(t) + \widehat{V}(\lambda t)] - [MA(C_1'\tau(t)) + \widehat{V}(\lambda t)] \\ &\approx [M\lambda t + M\widehat{A}(t)] - [M\Lambda C_1'\tau(t) + M\widehat{A}(C_1'\tau(t))] \\ &\approx M\Lambda C_1'(et - \tau(t)); \end{aligned}$$

this, combined with (73), yields

$$\widetilde{W}(t) \approx M\Lambda C_1'(et - \tau(t)). \tag{91}$$

From the above and (63), we have

$$\widetilde{Z}(t) = C\widetilde{W}(t) \approx CM\Lambda C_1'(et - \tau(t)); \tag{92}$$

substituting the above into (91) leads to

$$\widetilde{W}(t) \approx M\Lambda C_1'(CM\Lambda C_1')^{-1}\widetilde{Z}(t) = MH\widetilde{Z}(t);$$

this establishes the first equality in (76). We also note that the above, together with (91), implies

$$\Lambda C_1'(et - \tau(t)) \approx M^{-1}\widetilde{W}(t) \approx H\widetilde{Z}(t). \tag{93}$$

In view of (60), (65), (92) and (93), the relation (75) and the second equation in (76) follow from

$$\begin{aligned} Q(t) &= A(t) - D(t) \\ &\approx [\lambda t + \widehat{A}(t)] - A(C_1'\tau(t)) \\ &\approx [\lambda t + \widehat{A}(t)] - [\Lambda C_1'\tau(t) + \widehat{A}(C_1'\tau(t))] \\ &\approx \Lambda C_1'(et - \tau(t)) \\ &\approx H\widetilde{Z}(t). \end{aligned}$$

Next, in view of (65) and (66), we have

$$\begin{aligned} A(t) &= E(t) + \sum_{k=1}^{K} \Phi^k(D_k(t)) \\ &\approx \big[\alpha t + \widehat{E}(t)\big] + \sum_{k=1}^{K} \big[P_k' D_k(t) + \widehat{\Phi}_k(D_k(t))\big] \\ &\approx \alpha t + \widehat{E}(t) + P' A(C_1'\tau(t)) + \sum_{k=1}^{K} \widehat{\Phi}^k(\lambda_k t) \\ &\approx \alpha t + \widehat{E}(t) + \sum_{k=1}^{K} \widehat{\Phi}^k(\lambda_k t) + P'\Lambda C'\tau(t) + P'\widehat{A}(t). \end{aligned}$$

Note that $A(t) \approx \lambda t + \widehat{A}(t)$ and $\lambda = \alpha + P'\lambda = \alpha + P'\Lambda C' e$; the above leads to

$$\begin{aligned} \widehat{A}(t) &\approx (I - P')^{-1}\left[\widehat{E}(t) + \sum_{k=1}^{K} \widehat{\Phi}^k(\lambda_k t) - P'\Lambda C_1'(et - \tau(t))\right] \\ &\approx (I - P')^{-1}\left[\widehat{E}(t) + \sum_{k=1}^{K} \widehat{\Phi}^k(\lambda_k t) - P' H\widetilde{Z}(t)\right], \end{aligned}$$

where the last approximation follows from (93). Using the above approximation, we can rewrite (64) to obtain

$$\begin{aligned} Z(t) &= CV(A(t)) - et + Y(t) \\ &\approx C\big[MA(t) + \widehat{V}(A(t))\big] - et + \widehat{Y}(t) \\ &\approx CM\lambda t + CM\widehat{A}(t) + \widehat{V}(\lambda t) - et + \widehat{Y}(t) \\ &\approx (\rho - e)t + C\left[\widehat{V}(\lambda t) + M(I - P')^{-1}\big[\widehat{E}(t) + \sum_{k=1}^{K} \widehat{\Phi}^k(\lambda_k t)\big]\right] \\ &\quad - CM(I - P')^{-1}P' H\widetilde{Z}(t) + \widetilde{Y}(t); \end{aligned}$$

this, together with (70), implies that

$$\widetilde{Z}(t) \approx \theta(t) + \widetilde{X}(t) + R\widetilde{Y}(t),$$

with $\theta$ and $\widetilde{X}$ as defined by (79) and (80), respectively. This establishes the relation (78). The relations (81) and (82) follow from the corresponding properties for the original processes $Y$ and $Z$ and the Lipschitz continuity of the reflection mapping; specifically; the first relation corresponds to the nondecreasing property of $Y$ and the second relation corresponds to the work-conserving condition as stated by (69).

Finally, we establish (77). In view of (68) and (87), we have

$$\begin{aligned}
\mathscr{T}_k(t) &= Z_{\pi(k)}(t) + C'_{2k}\big[V(A(\mathscr{T}_k(t)+t)) - V(A(t))\big] \\
&\quad + V_k(A_k(t)) - V_k(A_k(t)-1), \\
&\approx \widetilde{Z}_{\pi(k)}(t) + C'_{2k}\big[MA(\mathscr{T}_k(t)+t) \\
&\quad + \widehat{V}(A(\mathscr{T}_k(t)+t)) - MA(t) - \widehat{V}(A(t))\big] + m_k \\
&\approx \widetilde{Z}_{\pi(k)}(t) + C'_{2k}\big[M\lambda(\mathscr{T}_k(t)+t) + M\widehat{A}(\mathscr{T}_k(t)+t) + \widehat{V}(\lambda(\mathscr{T}_k(t)+t)) \\
&\quad - M\lambda t - M\widehat{A}(t) - \widehat{V}(\lambda t)\big] + m_k \\
&\approx \widetilde{Z}_{\pi(k)}(t) + C'_{2k}\big[M\lambda\mathscr{T}_k(t) + M\widehat{A}(t) + \widehat{V}(\lambda t) - M\widehat{A}(t) - \widehat{V}(\lambda t))\big] + m_k \\
&= \widetilde{Z}_{\pi(k)}(t) + C'_{2k}M\lambda\mathscr{T}_k(t) + m_k;
\end{aligned}$$

this establishes the relation

$$\mathscr{T}(t) \approx \Delta(C'_1\widetilde{Z}(t) + m).$$

Therefore, combined with (67) and (90), we can conclude (77). □

**5. Performance analysis procedure.** Based on the strong approximation theorem in Section 4, we outline a procedure to approximate various performance measures of queueing networks. Specifically, we consider the case where the interarrival time, the service time and the routing sequences are mutually independent i.i.d. sequences. In this case, we can approximate the aggregated workload process $Z$ by an RBM $\widetilde{Z}$ with drift $\theta$, covariance matrix $\Gamma$ and reflection matrix $R$, which are described by

$$\begin{aligned}
R &= (I + CM(I-P')^{-1}P'H)^{-1}, \\
\theta &= -R(e-\rho), \\
\Gamma &= RC\left[\Gamma_V\Lambda + M(I-P')^{-1}\left[T_E + \sum_{k=1}^{K}\lambda_k\Gamma_\Phi^k\right](I-P)^{-1}M\right]C'R'.
\end{aligned}$$

Readers are referred to Section 4.1.1 for the definitions of the vectors and matrices used in the above equalities. In particular, we note that all of them are from the service disciplines, the routing probability, and the mean and the

variance of the interarrival and service times. Following from Lemma A.1 in the Appendix and Lemma 3.2 of Chen (1996), when $\rho < 1$, $\widetilde{Z}$ has a unique stationary distribution. Under certain conditions (which will be elaborated below), the stationary distribution has an explicit product form. In the more general case, the stationary distribution can be computed numerically. Dai and Harrison (1992) first develop such an algorithm, known as *QNET* (which approximates the stationary distribution by polynomial basis functions). Recently, Shen (2000) developed a new numerical algorithm *QNET*+ (which approximates the stationary distribution by finite element basis functions). All numerical examples in this paper are calculated by this new algorithm. Though this new algorithm can compute the stationary distribution function, we focus on approximating the mean of the stationary distribution only; approximating the distribution function will be dealt with in Shen (2000).

Given the estimate for the stationary distribution of the aggregated workload process, we could obtain estimates for some other performance measures of queueing networks. Let $\boldsymbol{E}(\widetilde{Z}_\ell)$ $(\ell = 1, \ldots, L)$ be the stationary mean for the aggregated workload. We shall describe two alternative methods to obtain the estimates of the stationary mean queue length and mean sojourn time.

The first method is to approximate the mean queue length by (75) in Section 4, and we have

$$\boldsymbol{E}(Q) = H\boldsymbol{E}(\widetilde{Z}). \tag{94}$$

Then, we use Little's law to obtain mean sojourn time as

$$\boldsymbol{E}(\mathscr{S}_k) = \frac{1}{\lambda_k}\boldsymbol{E}(Q_k). \tag{95}$$

The second method is to approximate the mean sojourn time via (77) by

$$\boldsymbol{E}(\mathscr{S}_k) = \delta_k[\boldsymbol{E}(\widetilde{Z}_{\pi(k)}) + m_k], \tag{96}$$

and then obtain the mean queue length by Little's law

$$\boldsymbol{E}(Q_k) = \lambda_k\boldsymbol{E}(\mathscr{S}_k). \tag{97}$$

ALGORITHM 1. *Computing steady-state average queue length and sojourn time*:

$$\boldsymbol{E}(Q) = H\boldsymbol{E}(\widetilde{Z}),$$
$$\boldsymbol{E}(\mathscr{S}_k) = \frac{1}{\lambda_k}\boldsymbol{E}(Q_k).$$

ALGORITHM 2. *Computing steady-state average queue length and sojourn time*:

$$\boldsymbol{E}(\mathscr{S}_k) = \delta_k[\boldsymbol{E}(\widetilde{Z}_{\pi(k)}) + m_k],$$
$$\boldsymbol{E}(Q_k) = \lambda_k\boldsymbol{E}(\mathscr{S}_k).$$

These two methods are summarized as Algorithms 1 and 2. They usually give different approximations. From our numerical experiments in Section 6, Algorithm 2 seems to provide much more accurate estimation than Algorithm 1. Therefore, Algorithm 2 is recommended to obtain approximations for the mean stationary queue length and the mean stationary sojourn time. The numeric evidence also suggests that both algorithms are doing well and are asymptotically identical for a class $k$ if $\beta_{\pi(k)}$ [the traffic intensity of priority group $\pi(k)$] is close to 1. Intuitively, when $\beta_{\pi(k)}$ is close to 1 and $\rho_{\pi(k)} < 1$, $\delta_k$ is close to or equal to 1 and the workload of all other priority groups at that station should almost be zero. Thus, both algorithms give

$$\boldsymbol{E}(\mathscr{S}_k) \approx \boldsymbol{E}(\widetilde{Z}_k).$$

5.1. *Product form solution.* Harrison and Williams (1992) showed that $\widetilde{Z}$ has a product form stationary distribution if and only if

$$\Gamma_{ij} = \tfrac{1}{2} R_{ji} \Gamma_{ii} \quad \text{for all } 1 \le i < j \le L, \tag{98}$$

in which case the solution is

$$p(x) = \prod_{\ell=1}^{L} \kappa_\ell \exp(-\kappa_\ell x_\ell), \qquad x \ge 0, \tag{99}$$

where $\kappa_1, \ldots, \kappa_L$ are the positive constants defined as

$$\kappa_\ell = \frac{2(1-\rho_\ell)R_{ll}}{\Gamma_{\ell\ell}} \quad \text{for } \ell = 1, \ldots, L. \tag{100}$$

The product form condition is rarely satisfied. Peterson (1991) pointed out a special case where the product form condition is satisfied; this special case requires Kelly network structures; namely, all jobs at each station have the same service time distributions and are served under FIFO service disciplines, and jobs follow deterministic routing.

**6. Numeric examples.** This section is devoted to analyzing two examples, both of which are feedforward queueing networks as described in Section 4. We apply our strong approximation to these models to obtain RBM models. Then, we compare the performance estimates from our RBM approximations with the estimates obtained from the RBM approximations obtained by using diffusion approximation in Peterson (1991) and with simulation results. To calculate the steady-state performance measures from RBM models, we use a newly developed numerical algorithm *QNET* + from Shen (2000).

6.1. *Single station with two job classes.* Consider the single-station network pictured in Figure 2. There are two job classes. Class 1 jobs have higher preemptive priority over class 2 jobs. We consider four versions of systems:

1. All interarrival and service times are taken to be Erlang of order 4 (SCV = 0.25).

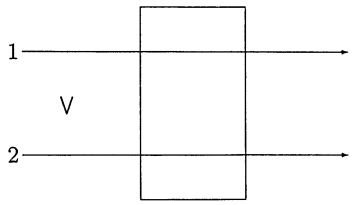


FIG. 2. *Single station with two job classes.*

2. All interarrival and service times are taken to be exponential (SCV = 1).
3. All interarrival and service times are taken to be Gamma distributed with SCV = 2.
4. All interarrival times are taken to be exponential and all service times are taken to be Erlang of order 4.

The arrival rates of both classes are 1. For each system, we examine five cases of the mean service times:

1. $m_1 = 0.7$, $m_2 = 0.1$;
2. $m_1 = 0.5$, $m_2 = 0.3$;
3. $m_1 = 0.3$, $m_2 = 0.5$;
4. $m_1 = 0.1$, $m_2 = 0.7$;
5. $m_1 = 0.2$, $m_2 = 0.2$.

Although there is no product form solution for the joint stationary distribution of $(\widetilde{Z}_1, \widetilde{Z}_2)$, the marginal distributions of $\widetilde{Z}_1$ and $\widetilde{Z}_2$ are exponentially distributed with means

$$E(\widetilde{Z}_1) = \frac{m_1^2(c_{01}^2 + c_1^2)}{2(1-\rho_1)},$$

$$E(\widetilde{Z}_2) = \frac{\sum_{i=1}^2 m_1^2(c_{0i}^2 + c_i^2)}{2(1-\rho_2)},$$

respectively ($\rho_2 = \lambda_1 m_2 + \lambda_2 m_2$ and $\rho_1 = \lambda_1 m_1$). The following three analytical methods are used to obtain approximations of the mean queue lengths and mean sojourn times:

1. By Algorithm 1 in Section 5,

$$E(\mathscr{S}_1) = EQ_1 = \frac{m_1(c_{01}^2 + c_1^2)}{2(1-\rho_1)},$$

$$E(\mathscr{S}_2) = EQ_2 = \frac{1}{m_2}\left(\frac{\sum_{i=1}^2 m_i^2(c_{0i}^2 + c_i^2)}{2(1-\rho_2)} - \frac{m_1^2(c_{01}^2 + c_1^2)}{2(1-\rho_1)}\right).$$

2. By Algorithm 2 in Section 5,

$$E\mathscr{S}_1 = EQ_1 = \frac{m_1^2(c_{01}^2 + c_1^2)}{2(1-\rho_1)} + m_1,$$

$$E\mathscr{S}_2 = EQ_2 = \frac{1}{1-\rho_1}\left(\frac{\sum_{i=1}^2 m_i^2(c_{0i}^2 + c_i^2)}{2(1-\rho_2)} + m_2\right).$$

3. By the diffusion approximation in Peterson (1991),

$$E\mathscr{S}_1 = EQ_1 = m_1,$$

$$E\mathscr{S}_2 = EQ_2 = \frac{1}{m_2}\frac{\sum_{i=1}^2 m_i^2(c_{0i}^2 + c_i^2)}{2(1-\rho_2)}.$$

(Note that a straightforward interpretation of the diffusion approximation would yield $E\mathscr{S}_1 = EQ_1 = 0$; the suggested approximation above represents a slight improvement.)

Table 1 summarizes the mean queue length estimates of each job class by using strong approximations, diffusion approximations and simulation. The columns "Strong (1)" and "Strong (2)" in Table 1 correspond to the approximations obtained by Algorithm 1 and Algorithm 2, respectively. The numbers in parentheses after the simulation figures show 95% confidence intervals as percentages of the simulation numbers. The numbers in parentheses following other figures are the percentage errors (in absolute value) as compared to simulation numbers. This convention will also be used in the subsequent tables. First we note that as indicated in Section 3.5 [cf. (56)], Algorithm 2 gives the exact mean queue lengths for the cases (systems 1 and 2) of Poisson arrivals. (In the corresponding rows, we could have reported the percentage errors relative to the exact mean, but we report the percentage errors relative to simulation results for consistency with other rows.) It seems that in almost all other cases, Algorithm 2 of the strong approximation also gives the best approximations and its estimates are quite close to simulation results. When SCVs of interarrival and service times are 1, Algorithms 1 and 2 coincide. This is true even when the arrival rates are not equal to 1. Note that the strong approximation Algorithm (2) also performs well when the station is lightly loaded (with $\rho = 0.4$). It should be expected that the diffusion approximation would not give a good estimate for the mean queue length of the higher priority class, but it is quite surprising that it also does poorly in estimating the mean queue length of the lower priority class. However, we note that the approximation for the mean queue length of the lower priority class does improve as the traffic intensity for the lower priority class increases (relative to the traffic intensity for the higher priority class); this corresponds to the case when $m_2$ increases from 0.1 to 0.7 and $m_1$ decreases from 0.7 to 0.1 in Table 1.

6.2. *Two-station tandem queue.* Pictured in Figure 3 is a two-station tandem queueing network. Each station has two different job classes. We assume

TABLE 1
*Average queue length in a single-station network*

| System | $m_1$ | $m_2$ | Class | Strong (1) | Strong (2) | Diffusion | Simulation |
|---|---|---|---|---|---|---|---|
| 1 | 0.7 | 0.1 | $Q_1$ | 0.58 (42.6%) | 1.11 (9.9%) | 0.70 (30.7%) | 1.01 (0.5%) |
| | | | $Q_2$ | 2.17 (10.0%) | 2.42 (0.4%) | 6.25 (159.3%) | 2.41 (1.9%) |
| | 0.5 | 0.3 | $Q_1$ | 0.25 (56.1%) | 0.63 (10.5%) | 0.50 (12.3%) | 0.57 (0.3%) |
| | | | $Q_2$ | 1.00 (27.0%) | 1.45 (5.8%) | 1.42 (3.6%) | 1.37 (1.0%) |
| | 0.3 | 0.5 | $Q_1$ | 0.11 (65.5%) | 0.33 (6.5%) | 0.30 (3.2%) | 0.31 (0.2%) |
| | | | $Q_2$ | 0.79 (35.1%) | 1.32 (9.1%) | 0.85 (29.8%) | 1.21 (0.7%) |
| | 0.1 | 0.7 | $Q_1$ | 0.03 (72.0%) | 0.10 (0.0%) | 0.10 (0.0%) | 0.10 (0.1%) |
| | | | $Q_2$ | 0.88 (34.8%) | 1.47 (8.9%) | 0.89 (34.1%) | 1.35 (0.6%) |
| | 0.2 | 0.2 | $Q_1$ | 0.06 (70.0%) | 0.21 (6.3%) | 0.20 (0.0%) | 0.20 (0.1%) |
| | | | $Q_2$ | 0.10 (64.3%) | 0.29 (3.6%) | 0.17 (39.3%) | 0.28 (0.3%) |
| 2 | 0.7 | 0.1 | $Q_1$ | 2.33 (1.7%) | 2.33 (1.7%) | 0.70 (70.5%) | 2.37 (2.2%) |
| | | | $Q_2$ | 8.67 (2.9%) | 8.67 (2.9%) | 25.0 (180.0%) | 8.93 (4.9%) |
| | 0.5 | 0.3 | $Q_1$ | 1.00 (0.0%) | 1.00 (0.0%) | 0.50 (50.0%) | 1.00 (1.1%) |
| | | | $Q_2$ | 4.00 (2.6%) | 4.00 (2.6%) | 5.67 (45.4%) | 3.90 (1.8%) |
| | 0.3 | 0.5 | $Q_1$ | 0.43 (0.0%) | 0.43 (0.0%) | 0.30 (30.2%) | 0.43 (0.7%) |
| | | | $Q_2$ | 3.14 (0.6%) | 3.14 (0.6%) | 3.40 (7.6%) | 3.16 (1.9%) |
| | 0.1 | 0.7 | $Q_1$ | 0.11 (0.0%) | 0.11 (0.0%) | 0.10 (9.1%) | 0.11 (0.5%) |
| | | | $Q_2$ | 3.56 (0.8%) | 3.56 (0.8%) | 3.57 (1.1%) | 3.53 (2.5%) |
| | 0.2 | 0.2 | $Q_1$ | 0.25 (0.0%) | 0.25 (0.0%) | 0.20 (20.0%) | 0.25 (0.5%) |
| | | | $Q_2$ | 0.42 (6.8%) | 0.42 (6.8%) | 0.66 (70.9%) | 0.39 (0.6%) |
| 3 | 0.7 | 0.1 | $Q_1$ | 4.67 (13.3%) | 3.97 (3.6%) | 0.70 (83.0%) | 4.12 (3.7%) |
| | | | $Q_2$ | 17.31 (0.1%) | 17.00 (3.6%) | 50.0 (190.2%) | 17.23 (8.8%) |
| | 0.5 | 0.3 | $Q_1$ | 2.00 (25.0%) | 1.50 (6.3%) | 0.50 (68.8%) | 1.60 (2.0%) |
| | | | $Q_2$ | 8.00 (4.3%) | 7.40 (3.5%) | 11.33 (47.7%) | 7.67 (5.0%) |
| | 0.3 | 0.5 | $Q_1$ | 0.87 (42.6%) | 0.56 (8.2%) | 0.30 (50.8%) | 0.61 (1.2%) |
| | | | $Q_2$ | 6.29 (10.7%) | 5.57 (1.9%) | 6.8 (19.7%) | 5.68 (3.8%) |
| | 0.1 | 0.7 | $Q_1$ | 0.22 (70.8%) | 0.12 (7.7%) | 0.10 (23.1%) | 0.13 (0.6%) |
| | | | $Q_2$ | 7.12 (7.7%) | 6.34 (4.1%) | 7.14 (8.0%) | 6.61 (3.8%) |
| | 0.2 | 0.2 | $Q_1$ | 0.50 (51.0%) | 0.30 (9.0%) | 0.20 (39.4%) | 0.33 (0.9%) |
| | | | $Q_2$ | 0.83 (57.2%) | 0.58 (10.0%) | 1.33 (151.6%) | 0.53 (1.0%) |
| 4 | 0.7 | 0.1 | $Q_1$ | 1.46 (15.1%) | 1.72 (0.0%) | 0.70 (59.3%) | 1.72 (1.5%) |
| | | | $Q_2$ | 5.42 (2.9%) | 5.54 (0.7%) | 7.81 (40.0%) | 5.58 (3.9%) |
| | 0.5 | 0.3 | $Q_1$ | 0.63 (22.2%) | 0.81 (0.0%) | 0.50 (38.3%) | 0.81 (0.7%) |
| | | | $Q_2$ | 2.50(8.8%) | 2.73 (0.4%) | 3.54 (29.2%) | 2.74 (2.2%) |
| | 0.3 | 0.5 | $Q_1$ | 0.27 (28.9%) | 0.38 (0.0%) | 0.30 (21.1%) | 0.38 (0.4%) |
| | | | $Q_2$ | 1.96 (12.1%) | 2.23 (0.0%) | 2.13 (4.5%) | 2.23 (2.0%) |
| | 0.1 | 0.7 | $Q_1$ | 0.07 (36.4%) | 0.11 (0.0%) | 0.10 (9.1%) | 0.11 (0.3%) |
| | | | $Q_2$ | 2.22 (11.9%) | 2.51 (0.4%) | 2.23 (11.5%) | 2.52 (2.0%) |
| | 0.2 | 0.2 | $Q_1$ | 0.16 (30.4%) | 0.24 (4.3%) | 0.20 (13.0%) | 0.23 (0.3%) |
| | | | $Q_2$ | 0.26 (25.6%) | 0.35 (0.0%) | 0.42 (20.0%) | 0.35 (0.5%) |

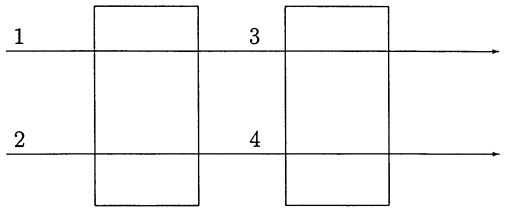


FIG. 3. *Two-station tandem queue.*

that all exogenous arrival processes and service processes are mutually independent renewal processes.

We will estimate the performance of this network under two different service disciplines:

1. The service discipline at station 1 is preemptive priority and the service discipline at station 2 is FCFS. Class 1 jobs have a higher priority over class 2 jobs at station 1.
2. The service discipline at station 1 is FCFS and the service discipline at station 2 is preemptive priority. Class 3 jobs have a higher priority over class 4 jobs at station 2.

For each different service discipline type, we will compute three versions of systems with different service and interarrival time distributions. We list parameters of three systems in Table 2. All service and interarrival time distributions are taken to be Erlang of order 4 (SCV = 0.25) in the first system, all are taken to be exponential (SCV = 1) in the second system and all are taken to be Gamma distribution with SCV = 2 in the third system.

Tables 3 and 4 present the simulation estimates and $QNET+$ estimates of the mean queue length for each job class for each different queueing system configuration. We use $QNET+$ to get the mean aggregated workload numerically and then use Algorithm 2 in Section 5 to obtain the mean queue lengths. (In this case, both Algorithm 1 and the estimate based on the diffusion approximation provide inferior estimates as well, so they are not presented.) The $QNET+$ estimates of the mean queue lengths for this two-station network are quite impressive compared with the simulation estimates, except the estimates for the job class 4 in the queueing networks with first service discipline type, in which the $QNET+$ significantly underestimates the queue length for job

TABLE 2
*System specifications of two-station tandem queue*

| System | Distribution | $\alpha_1$ | $\alpha_2$ | $m_1$ | $m_2$ | $m_3$ | $m_4$ |
|---|---|---|---|---|---|---|---|
| 1 | $E_4$ | 1.0 | 3.0 | 0.5 | 0.1 | 0.3 | 0.2 |
| 2 | $M$ | 1.0 | 3.0 | 0.5 | 0.1 | 0.3 | 0.2 |
| 3 | Gamma | 1.0 | 3.0 | 0.5 | 0.1 | 0.3 | 0.2 |

TABLE 3
*Average queue length of network* 1

| System No. | Approximation Method | $Q_1$ | $Q_2$ | $Q_3$ | $Q_4$ |
|---|---|---|---|---|---|
| 1 | Simulation | 0.57 (0.3%) | 2.71 (1.1%) | 0.99 (1.3%) | 4.00 (1.1%) |
| | *QNET*+ | 0.63 (10.5%) | 2.70 (0.4%) | 1.04 (5.1%) | 2.83 (29.3%) |
| 2 | Simulation | 1.00 (0.9%) | 9.02 (2.6%) | 3.32 (2.5%) | 12.50 (1.9%) |
| | *QNET*+ | 1.01 (1.0%) | 8.99 (3.3%) | 3.2 (3.6%) | 9.31 (25.5%) |
| 3 | Simulation | 1.59 (1.6%) | 17.54 (4.0%) | 6.59 (4.6%) | 24.41 (3.9%) |
| | *QNET*+ | 1.54 (3.1%) | 17.15 (2.2%) | 6.25 (5.2%) | 18.45 (24.4%) |

TABLE 4
*Average queue length of network* 2

| System No. | Approximation Method | $Q_1$ | $Q_2$ | $Q_3$ | $Q_4$ |
|---|---|---|---|---|---|
| 1 | Simulation | 0.78 (0.6%) | 1.51 (0.8%) | 0.31 (0.2%) | 3.79 (1.2%) |
| | *QNET*+ | 0.85 (9.0%) | 1.34 (11.3%) | 0.33 (6.5%) | 3.58 (5.5%) |
| 2 | Simulation | 1.90 (2.1%) | 4.50 (2.2%) | 0.40 (0.7%) | 12.17 (2.9%) |
| | *QNET*+ | 1.89 (0.5%) | 4.47 (0.6%) | 0.41 (3.5%) | 11.80 (3.0%) |
| 3 | Simulation | 3.42 (3.1%) | 8.53 (3.5%) | 0.55 (1.0%) | 23.86 (4.4%) |
| | *QNET*+ | 3.30 (3.5%) | 8.69 (1.9%) | 0.51 (7.3%) | 22.65 (5.1%) |

class 4. We have no theoretical explanation for it at the moment, though we feel that the large errors might be due to the large variations in the interarrival times of this class (which correspond to the departure times of class 2, the lower priority class at station 1).

## APPENDIX

### A.1. Proofs and an elementary lemma.

PROOF OF PROPOSITION 2.4. Without loss of generality, we assume that $x(0) = 0$. If $u \geq v$, we have

$$
\begin{aligned}
f(u)-f(v) &= \sup_{0\leq s\leq u}\{-\theta s-x(s)\}-\sup_{0\leq s\leq v}\{-\theta s-x(s)\}-[-\theta]^+(u-v)\\
&= \left[\sup_{v\leq s\leq u}\{-\theta s-x(s)\}-\sup_{0\leq s\leq v}\{-\theta s-x(s)\}\right]^+-[-\theta]^+(u-v)\\
&= \left[\sup_{0\leq s\leq u-v}\{-\theta(v+s)-x(v+s)\}-\sup_{0\leq s\leq v}\{-\theta s-x(s)\}\right]^+-[-\theta]^+(u-v)\\
&= \left[\sup_{0\leq s\leq u-v}\{-\theta s-x(v+s)+x(v)\}-\theta v-x(v)-\sup_{0\leq s\leq v}\{-\theta s-x(s)\}\right]^+\\
&\quad -[-\theta]^+(u-v)
\end{aligned}
$$

$$\leq \sup_{0\leq s\leq u-v}\{-\theta s - x(v+s)+x(v)\}-[-\theta]^+(u-v)$$
$$\leq \sup_{0\leq s\leq u-v}\{-\theta s\}+\sup_{0\leq s\leq u-v}\{-x(v+s)+x(v)\}-[-\theta]^+(u-v)$$
$$\leq \sup_{0\leq s\leq u-v}|x(v+s)-x(v)|.$$

If $\theta \geq 0$, we have

$$f(u)-f(v)=\sup_{0\leq s\leq u}\{-\theta s - x(s)\}-\sup_{0\leq s\leq v}\{-\theta s - x(s)\}\geq 0.$$

If $\theta < 0$, we have

$$\begin{aligned}
f(u)-f(v) &= \sup_{0\leq s\leq u}\{-\theta s - x(s)\}-\sup_{0\leq s\leq v}\{-\theta s - x(s)\}+\theta(u-v)\\
&\geq \sup_{u-v\leq s\leq u}\{-\theta s - x(s)\}-\sup_{0\leq s\leq v}\{-\theta s - x(s)\}+\theta(u-v)\\
&= \sup_{0\leq s\leq v}\{-\theta(s+u-v)-x(s+u-v)\}\\
&\quad -\sup_{0\leq s\leq v}\{-\theta s - x(s)\}+\theta(u-v)\\
&= \sup_{0\leq s\leq v}\{-\theta s - x(s+u-v)\}-\sup_{0\leq s\leq v}\{-\theta s - x(s)\}\\
&\geq -\sup_{0\leq s\leq v}\{x(s)-x(s+u-v)\}\\
&\geq -\sup_{0\leq s\leq u-v}|x(s+u-v)-x(s)|.
\end{aligned}$$

Let $h(T)=\sqrt{T\log\log T}$, for $\forall\theta$, we have

$$\begin{aligned}
&\sup_{\substack{0\leq u,\,v\leq T\\ |u-v|\leq h(T)}}|f(u)-f(v)|\\
&\quad\leq \sup_{\substack{0\leq u,\,v\leq T\\ |u-v|\leq h(T)}}\left\{\sup_{0\leq s\leq v}|x(s+u-v)-x(s)|\vee \sup_{0\leq s\leq u-v}|x(v+s)-x(v)|\right\}\\
&\quad\leq \sup_{\substack{0\leq u,\,v\leq T\\ |u-v|\leq h(T)}}\{|x(u)-x(v)|\}\\
&\quad= \circ(T^{1/r})\quad \text{as } T\to\infty.
\end{aligned}$$

□

PROOF OF THEOREM 2.5. We provide an inductive proof on dimension $K$. First, consider $K=1$. Then $R$ is a positive scalar. In this case, the one-dimensional RBM $Z$ can be expressed as

$$Z(t)=X(t)+\sup_{0\leq s\leq t}\{-X(s)\}^+,$$

where $X$ is a Brownian motion with a negative drift $\theta$ and a standard deviation $\sigma$. If $X(0) = x = 0$, then following from a standard argument for the reflected Brownian motion and the maximum of the Brownian motion [see, for example, Harrison (1985)], we have

$$\begin{aligned} P(||Z||_T \geq z) &= P\Big( \sup_{0\leq t\leq T} \sup_{0\leq s\leq t} [X(t) - X(s)] \geq z \Big) \\ &= P\Big( \sup_{0\leq s\leq T} X(s) \geq z \Big) \\ &\leq P\Big( \sup_{0\leq s\leq \infty} X(s) \geq z \Big) \\ &= e^{-\nu z} \end{aligned}$$

for any $z \geq 0$, where $\nu = 2|\theta|/\sigma^2$. Taking $z = (2 \log T)/\nu$ in the above yields

$$P\big(||Z||_T/\log T \geq 2/\nu\big) \leq \frac{1}{T^2};$$

then by the Borel–Cantelli lemma, we establish (9). If $X(0) = x \neq 0$, define $X^1(t) = X(t) - X(0)$. Let $\Phi$ denote the one-dimensional reflection mapping; that is, $Z = \Phi(X)$. Define $Z^1 = \Phi(X^1)$. Since $X^1$ is a Brownian motion with a negative drift starting from the origin, the above proof establishes that the bound (9) holds for $Z^1$; then the Lipschitz continuity of the reflection mapping $\Phi$ establishes the bound (9) for $Z$.

Next, suppose that the theorem holds for dimension $d-1$; we show it also holds for dimension $K = d$. Since $R$ is a lower triangular matrix, it follows from the induction hypothesis that

$$\sup_{0\leq t\leq T} |Z_k(t)| \overset{\text{a.s.}}{=} O(\log T) \tag{101}$$

holds for all $k = 1, \ldots, d-1$. Note that $R^{-1}Z = R^{-1}X + Y$; we have

$$(R^{-1})_{dd} Z_d(t) = \sum_{k=1}^{d} (R^{-1})_{dk} X_k(t) - \sum_{k=1}^{d-1} (R^{-1})_{dk} Z_k(t) + Y_d(t),$$

with $(R^{-1})_{dd} > 0$. Let

$$X^d(t) + \sum_{k=1}^{d} (R^{-1})_{dk} X_k(t)/(R^{-1})_{dd}$$

and $Z^d = \Phi(X^d)$. Note that $X^d$ is a one-dimensional process (actually a Brownian motion), $Z^d$ is its reflected process. Then, in view of (101) and the Lipschitz continuity of the one-dimensional reflection mapping $\Phi$, we have

$$\sup_{0\leq t\leq T} |Z_d(t) - Z^d(t)| \overset{\text{a.s.}}{=} O(\log T). \tag{102}$$

Furthermore, note that $X^d$ is a (one-dimensional) Brownian motion with a negative drift,

$$\sum_{k=1}^{d}(R^{-1})_{dk}\theta_k/(R^{-1})_{dd},$$

since $R^{-1}\theta < 0$ and $(R^{-1})_{dd} > 0$. Applying the proved result for the one-dimensional case yields

$$\sup_{0\le t\le T}|Z^d(t)| \stackrel{\text{a.s.}}{=} O(\log T);$$

this, together with the bound (102), implies

$$\sup_{0\le t\le T}|Z_d(t)| \stackrel{\text{a.s.}}{=} O(\log T).$$

The above and (101) prove (9) for $K = d$. □

LEMMA A.1. *Both matrices* $H = \Lambda C_1'(CM\Lambda C_1')^{-1}$ *and* $G = CM(I-P')^{-1}P'H$ *are well-defined, and matrix* $G$ *is strictly lower triangular.*

PROOF. Let $k_i$ denote the number of job classes at station $i$, and $\ell_i$ denote the number of priority groups at station $i$, $i = 1, \ldots, J$. Because the queueing network is feedforward, we can express the matrix $P'$ as

$$P' = \begin{Vmatrix} P_{11} & 0 & 0 & \ldots & 0 \\ P_{21} & P_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{J1} & P_{J2} & P_{J3} & \ldots & P_{JJ} \end{Vmatrix}, \tag{103}$$

where $P_{ij}\,(1 \le i,\ j \le J)$ is a $k_i \times k_j$ submatrix. Since there is no self-feedback, all elements of $P_{ii}$, $i = 1, \ldots, J$, must be zero. Matrix $(P^2)'$ has a structure similar to (103) where the diagonal submatrices are zero. Since

$$(I - P')^{-1} = (I + P + P^2 + \cdots)',$$

and $M = \text{diag}\{m_i\}$, to prove the result of this lemma, it suffices to prove that $H$ is well defined and $CP'H$ is a lower triangular matrix.

Matrix $C$ can be represented by

$$C = \begin{Vmatrix} C_{11} & 0 & 0 & \ldots & 0 \\ 0 & C_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & C_{JJ} \end{Vmatrix},$$

where $C_{ii}\,(1 \le i \le J)$ is an $\ell_i \times k_i$ submatrix. This implies that $D \equiv CM\Lambda C_1'$ is lower triangular. Furthermore, $D$ can be written as

$$D = \begin{Vmatrix} D_{11} & 0 & 0 & \ldots & 0 \\ 0 & D_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & D_{JJ} \end{Vmatrix},$$

where $D_{ii}\,(1 \le i \le J)$ is an $\ell_i \times \ell_i$ submatrix and has the form of

$$D_{ii} = \begin{Vmatrix} \beta_{\ell+1} & 0 & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{\ell+1} & \beta_{\ell+2} & \dots & \beta_{\ell+\ell_i} \end{Vmatrix},$$

where $\beta = C_1 M\lambda$, and group $\ell$ has the highest priority at station $i$. We know that the inverse of $D_{ii}$ has the form

$$D_{ii}^{-1} = \begin{Vmatrix} \dfrac{1}{\beta_{\ell+1}} & 0 & \dots & 0 \\ -\dfrac{1}{\beta_{\ell+2}} & \dfrac{1}{\beta_{\ell+2}} & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \dots & -\dfrac{1}{\beta_{\ell+\ell_i}} & \dfrac{1}{\beta_{\ell+\ell_i}} \end{Vmatrix}.$$

Thus, it can be verified easily that

$$H_{kl} = \begin{cases} \dfrac{\lambda_k}{\beta_{\pi(k)}}, & \text{if } \ell = \pi(k), \\ -\dfrac{\lambda_k}{\beta_{\pi(k)}}, & \text{if } \ell = \pi(k) - 1 \text{ and } \sigma(\ell) = \sigma(\pi(k)), \\ 0, & \text{otherwise} \end{cases}$$

It is obvious that $H$ has the following representation:

$$H = \begin{Vmatrix} H_{11} & 0 & 0 & \dots & 0 \\ 0 & H_{22} & 0 & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \dots & H_{JJ} \end{Vmatrix},$$

where $H_{ii}(1 \le i \le J)$ is a $k_i \times \ell_i$ submatrix. By these representations, it is clear that $CP'H$ is lower triangular. This completes the proof. □

**A.2. The general traffic intensity case.** In this section, we show that strong approximation can also be applied for approximation even if the traffic intensity at a station is larger than 1. For simplicity, we only consider single-station queueing networks. This is an extension of the discussion in Section 3. We are satisfied not to provide a proof here since the proof is analogous to that in Section 3.

Define a partition of $\mathcal{K}$ by

$$\begin{aligned} \mathcal{K}_n &= \{k \in \mathcal{K} : \rho_{\pi(k)} \le 1\}, \\ \mathcal{K}_b &= \{k \in \mathcal{K} : \rho_{\pi(k)-1} < 1 < \rho_{\pi(k)}\}, \\ \mathcal{K}_s &= \{k \in \mathcal{K} : \rho_{\pi(k)} > 1 \text{ and } \rho_{\pi(k)-1} > 1\}. \end{aligned}$$

Note that both $\mathcal{K}_b$ and $\mathcal{K}_s$ are empty sets, if $\rho \le 1$; also note that the set $\mathcal{K}_b$ would be an empty set if $\rho_{\pi(k)} = 1$ for some $k \in \mathcal{K}_n$. Also note that all classes in $\mathcal{K}_n$ have higher priorities than classes in $\mathcal{K}_b$ and $\mathcal{K}_s$, and all classes in $\mathcal{K}_b$ have higher priorities than classes in $\mathcal{K}_s$. Hence, jobs of classes in $\mathcal{K}_n$ do not see jobs of classes in $\mathcal{K}_b$ and $\mathcal{K}_s$, and jobs of classes in $\mathcal{K}_b$ do not see jobs of classes in $\mathcal{K}_s$. On the other hand, jobs of classes in $\mathcal{K}_s$ see all jobs of classes in $\mathcal{K}_n$ and $\mathcal{K}_b$ in front of them in the queue, and jobs of classes in $\mathcal{K}_b$ see

all jobs of classes in $\mathscr{K}_n$ in front of them in the queue. Based on what they see, jobs of classes in $\mathscr{K}_n$ observe the queue with a traffic intensity less than or equal to 1, or the queue is nonbottleneck or balanced bottleneck; jobs of classes in $\mathscr{K}_s$ observe the queue with a traffic intensity strictly greater than 1, or the queue is strictly bottleneck. When $\mathscr{K}_b \neq \emptyset$, all jobs in $\mathscr{K}_n$ see the queue with a traffic intensity strictly less than 1. In this case, if a job in $\mathscr{K}_b$ were given the highest (preemptive) priority over the other jobs in $\mathscr{K}_b$, then it would observe that the queue is nonbottleneck, and if a job in $\mathscr{K}_b$ were given the lowest (preemptive) priority to the other jobs in $\mathscr{K}_b$, then it would observe that the queue is strictly bottleneck; hence, overall, jobs in $\mathscr{K}_b$ hold a balance or fall between nonbottlenecks and strict bottlenecks. Actually, when $\mathscr{K}_b \neq \emptyset$, $\mathscr{K}_b = g_{\ell_b}$ for some $1 \leq \ell_b \leq L$, and hence, all jobs in $\mathscr{K}_b$ are served in the order of their arrival (FIFO).

THEOREM A.2. *Suppose that the FLIL asumptions* (23)–(25) *hold. Then as* $T \to \infty$,

$$\sup_{0\leq t\leq T} |Z_\ell(t) - \overline{Z}_\ell(t)| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |W_k(t) - (\lambda_k - \lambda_k^*)m_k t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |Q_k(t) - (\lambda_k - \lambda_k^*)t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |\mathscr{S}_k(t) - \overline{\mathscr{S}}_k(t)| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |D_k(t) - \lambda_k^* t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |Y_\ell(t) - (1-\rho_\ell)^+ t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |T_k(t) - \lambda_k^* m_k t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

$$\sup_{0\leq t\leq T} |\tau_\ell(t) - \overline{\tau}_\ell t| \overset{\text{a.s.}}{=} O(\sqrt{T \log\log T}),$$

*where* $\ell = 1, \ldots, L$, $k, \ldots, K$ *and*

$$\overline{Z}_\ell(t) = (\rho_\ell - 1)^+ t,$$

$$\overline{\mathscr{S}}_k(t) = \begin{cases} 0, & \text{for } \pi(k) \in \mathscr{K}_n, \\ \dfrac{(\rho_{\pi(k)} - 1)}{1 - \rho_{\pi(k)-1}} t, & \text{for } k \in \mathscr{K}_n \cup \mathscr{K}_b \end{cases},$$

$$\lambda_k^* = \begin{cases} \lambda_k, & \text{for } k \in \mathscr{K}_n, \\ \dfrac{\lambda_k(1 - \rho_{\pi(k)-1})}{\rho_{\pi(k)}}, & \text{for } k \in \mathscr{K}_b, \\ 0, & \text{for } k \in \mathscr{K}_s, \end{cases}$$

$$\overline{\tau}_\ell = \begin{cases} 1, & \text{for } \ell \subset \mathscr{K}_n, \\ \dfrac{1 - \rho_{\ell-1}}{\beta_\ell}, & \text{for } \ell \subset \mathscr{K}_b, \\ 0, & \text{for } \ell \subset \mathscr{K}_s. \end{cases}$$

THEOREM A.3. *Suppose that assumptions* (34)–(36) *hold with* $\widehat{A}_k$ *and* $\widehat{S}_k$ *being* $r$*-strong continuous for some* $r \in (2,4)$. *Then for* $\ell = 1, \ldots, L$, $k = 1, \ldots, K$, *as* $T \to \infty$,

$$\sup_{0\le t\le T} |Z_\ell(t) - \widetilde{Z}_\ell(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

$$\sup_{0\le t\le T} |\mathscr{S}_k(t) - \widetilde{\mathscr{S}}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

$$\sup_{0\le t\le T} |D_k(t) - \lambda_k^* t - \widetilde{W}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

$$\sup_{0\le t\le T} |Q_k(t) - (\lambda_k - \lambda_k^*)t - \widehat{A}_k(t) + \widetilde{W}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

$$\sup_{0\le t\le T} |W_k(t) - (\lambda_k - \lambda_k^*)m_k t - \widehat{W}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

*where*

$$\widetilde{Z}_\ell(t) = \widetilde{N}_\ell(t) + \widetilde{Y}_\ell(t),$$

$$\widetilde{N}_\ell(t) = (\rho_\ell - 1)t + \sum_{i=1}^{l} \sum_{j\in g_i} [m_j \widehat{A}_j(t) - \widehat{V}_j(\lambda_j t)],$$

$$\widetilde{Y}_\ell(t) = \sup_{0\le s\le t} \{-\widetilde{N}_\ell(s)\}^+,$$

$$\widetilde{\mathscr{S}}_k(t) = \frac{1}{1-\rho_{\pi(k)-1}} \Bigg( \widetilde{Z}_{\pi(k)}(t) + \sum_{i=1}^{\pi(k)} \sum_{j\in g_i} m_j [\widehat{A}_j(t) + \widehat{V}_j(\lambda_j t) - \widehat{A}_j(\overline{Z}_{\pi(k)}(t) + t) - \widehat{V}_j(\lambda_j(\overline{Z}_{\pi(k)}(t) + t))] \Bigg),$$

$$\widetilde{W}_k(t) = \widehat{A}_k(\overline{\tau}_{\pi(k)} t) - \frac{\lambda_k}{\beta_{\pi(k)}} \sum_{j\in g_\ell} [m_j \widehat{A}_j(\overline{\tau}_{\pi(k)} t) - \widehat{V}_j(\lambda_j^* t)] - \frac{\lambda_k}{\beta_{\pi(k)}} [\widetilde{Y}_{\pi(k)}(t) - (1-\rho_{\pi(k)})^+ t] + \frac{\lambda_k}{\beta_{\pi(k)}} [\widetilde{Y}_{\pi(k)-1}(t) - (1-\rho_{\pi(k)-1})^+ t],$$

$$\widehat{W}_k(t) = -m_k \widetilde{W}_k(t) + m_k \widehat{A}_k(t) - \widehat{V}_k(\lambda_k t) + \widehat{V}_k(\lambda_k^* t),$$

*and* $\widetilde{W}_k$ *is* $r$*-strong continuous.*

In fact, we can get an equivalent but easier to understand strong approximation form for the workload and queue length of individual job classes.

COROLLARY A.4. *Suppose that the assumptions* (34)–(36) *hold with* $2 < r < 4$. *Then for* $k = 1, \ldots, K$, *as* $T \to \infty$,

$$\sup_{0 \le t \le T} |Q_k(t) - \widetilde{Q}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

$$\sup_{0 \le t \le T} |W_k(t) - \widetilde{W}_k(t)| \overset{\text{a.s.}}{=} \circ(T^{1/r}),$$

*where*

$$\widetilde{W}_k(t) = \frac{\lambda_k m_k}{\beta_{\pi(k)}}[\widetilde{Z}_{\pi(k)} - \widetilde{Z}_{\pi(k)-1}(t) - \widehat{W}^{\pi(k)}(t)] + m_k[\widehat{A}_k(t) - \widehat{A}_k(\gamma^{\pi(k)}t)]$$
$$+[\widehat{V}_k(\lambda_k t) - \widehat{V}_k(\lambda_k^* t)],$$

$$\widetilde{Q}_k(t) = \frac{\lambda_k}{\beta_{\pi(k)}}[\widetilde{Z}_{\pi(k)}(t) - \widetilde{Z}_{\pi(k)-1}(t) - \widehat{W}^{\pi(k)}(t)] + [\widehat{A}_k(t) - \widehat{A}_k(\gamma^{\pi(k)}t)],$$

$$\widehat{W}^{\ell}(t) = \sum_{i \in g_\ell} \{m_i \widehat{A}_i(t) - \widehat{V}_i(\lambda_i t) - m_i \widehat{A}_i(\gamma^{\ell} t) + \widehat{V}_i(\lambda_i^* t)\}.$$

## REFERENCES


BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

CHEN, H. (1996). A sufficient condition for the positive recurrence of a semimartingale reflecting Brownian motion in an orthant. *Ann. Appl. Probab*. **6** 758–765.

CHEN, H. and MANDELBAUM, A. (1994). Hierarchical modeling of stochastic networks II. Strong approximations. In *Stochastic Modeling and Analysis of Manufacturing Systems* (D. D. Yao, ed.) 107–131. Springer, Berlin.

CSÖRGŐ, M. and HORVÁTH, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley, New York.

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.

FENDICK, K. W., SAKSENA, V. R. and WHITT, W. (1989). Dependence in packet queues. *IEEE Trans. Communications* **37** 1173–1183.

GLYNN, P. W. (1990). Diffusion approximations. In *Handbooks in Operations Research and Management Science* **2**. *Stochastic Models* (D. P. Heyman and M. J. Sobel, eds.) 145–198. North-Holland, Amsterdam.

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

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.

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.

HORVÁTH, L. (1990). Strong approximations of open queueing networks. *Math Oper. Res*. **17** 487–508.

KLEINROCK, L. (1976). *Queueing Systems* **2**. *Computer Applications*. Wiley, New York.

LEMOINE, A. J. (1978). Network of queues—a survey of weak convergence results. *Management Sci.* **24** 1175–1193.

MANDELBAUM, A. and MASSEY, W. A. (1995). Strong approximation for time-dependent queues. *Math. Oper. Res*. **20** 33–64.

MANDELBAUM, A., MASSEY, W. A. and REIMAN, M. (1998). Strong approximation for Markov service networks. *Queueing Systems Theory Appl*. **30** 149–201.

MANDELBAUM, A. and PATS, G. (1998). Stochastic networks I. Approximations and applications with continuous diffusion limits. *Ann. Appl. Probab*. **8** 569–646.

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

SHEN, X. (2000). Performance evaluation of multiclass queueing networks via Brownian motions. Ph.D. thesis, Univ. British Columbia.

WHITT, W. (1974). Heavy traffic theorems for queues: a survey. In *Mathematical Methods in Queueing Theory* (A. B. Clarke, ed.) 307–350. Springer, New York.

WHITT, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res*. **5** 67–85.

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

ZHANG, H. (1997). Strong approximations for irreducible closed queueing networks. *Adv. in Appl. Probab*. **29** 498–522.

ZHANG, H., HSU, G. and WANG, R. (1990). Strong approximations for multiple channel queues in heavy traffic. *J. Appl. Probab*. **28** 658–670.

FACULTY OF COMMERCE
AND BUSINESS ADMINISTRATION
UNIVERSITY OF BRITISH COLUMBIA
2053 MAIN MALL
VANCOUVER, BRITISH COLUMBIA
CANADA V6T 1Z2
E-MAIL: chen@hong.commerce.ubc.ca
xshen@hong.commerce.ubc.ca