THE QNET METHOD FOR RE-ENTRANT QUEUEING NETWORKS WITH PRIORITY DISCIPLINES

J. G. DAI

Georgia Institute of Technology, Atlanta, Georgia

D. H. YEH

Kao Shiung Institute of Technology, Taiwan

C. ZHOU

Georgia Institute of Technology, Atlanta, Georgia

(Received October 1994; revisions received September 1995; accepted October 1995)

This paper is concerned with the estimation of performance measures of two priority disciplines in a d-station re-entrant queueing network. Such networks arise from complex manufacturing systems such as wafer fabrication facilities. The priority disciplines considered are First-Buffer-First-Served (FBFS) and Last-Buffer-First-Served (LBFS). An analytical method is developed to estimate the long-run average workload at each station and the mean sojourn time in the network. When the first-buffer-first-served discipline is used, a refined estimate of the mean sojourn time is also developed. The workload estimation has two steps. In the first step, following Harrison and Williams (1992), we use a d-dimensional reflecting Brownian motion (RBM) to model the workload process. We prove that the RBM exists and is unique in distribution and that it has a unique stationary distribution. We then use an algorithm of Dai and Harrison (1992) to compute the stationary distribution of the RBM. Our method uses both the first and second moment information, and it is rooted in heavy traffic theory. It is closely related to the QNET method of Harrison and Nguyen (1993) for two-moment analysis of First-In-First-Out (FIFO) discipline. Our performance estimates of several example problems are compared to the simulation estimates to illustrate the effectiveness of our method.

This paper is concerned with queueing network models of job-shop or batch manufacturing systems. For our purposes, a manufacturing system is a collection of workstations, or simply stations, each of which has one server working at the station. A server may represent either a machine or an operator. The entities that are processed at the workstations will be called jobs or customers. Depending on the particular manufacturing context, what we call a job might actually be referred to as a part, a work order, a production lot, or a production batch. In the models considered here, each job that enters the system requires a particular sequence of operations, each of which must be performed at a particular station. The route of a job is the ordered sequence of stations that it visits. The time required to perform any given operation is called a service time.

In this paper, we restrict our attention to what Kumar (1993) has called a *re-entrant line*, which is a special type of d-station queueing network in which all jobs follow a deterministic route of K stages, and the jobs may visit some stations multiple times. An example of two station re-entrant line with $d = 2$ and $K = 3$ is shown in Figure 1. A distinctive feature of a re-entrant line is that jobs at different stages may be processed at the same station. For example, a machining cell is used to perform cutting operations of a family of complex components using computer numerically controlled machines. In order to satisfy the final tolerances and finish, various features on a part are processed in several setups based on different data systems and motion requirements to reduce the deformations caused by thermal expansion and forces in fixturing and cutting. Another example of a re-entrant line can be found in semiconductor wafer fabrication facilities. The manufacturing process there consists of building up layers of implanted material according to a sequence of masks. Some layers are implanted on a common machine at different processing stages. As an example, one wafer process (after major simplifications and aggregations) can be described as a re-entrant line with $d = 12$ and $K = 60$, with some stations being revisited 14 times.

When a station in a re-entrant line becomes available and there are several jobs at different stages waiting in their respective buffers at the station, it is necessary to decide which job to process first. Here, the buffer for a stage does not have to be physically present. The simplest queueing discipline (or dispatching rule) is first-in-first-out (FIFO). Alternately, an operator can assign priorities based on the stages of a job. Obviously, the queueing disciplines impact the number of jobs waiting at various stations and the sojourn (or flow) times. The queueing disciplines considered in this paper are two buffer priority disciplines: first-buffer-first-served (FBFS) and last-buffer-first-served (LBFS). Under the FBFS discipline, priority is given to the waiting job that is at the earliest stage in its route. Under

Subject classifications: Queues, networks, diffusion models: Brownian models, heavy traffic, performance measures, re-entrant lines, priority disciplines.
Area of review: Stochastic Processes & Their Applications.

0030-364X/97/4504-0610 $05.00

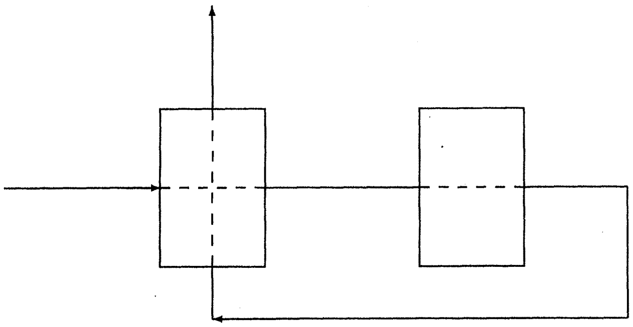

Figure 1. A three-buffer, two-station re-entrant line.

the LBFS discipline, priority is given to the waiting job that is at the latest stage in its route. We assume preemptive priority disciplines, though our simulations show that the performance difference between a preemptive discipline and a nonpreemptive discipline is not significant.

In this paper, we present an analytical method to predict the performance of the two queueing disciplines. The method, called the *QNET method* following Harrison and Nguyen (1990, 1993), has two steps. We first use a d-dimensional reflecting Brownian motion (RBM) to model the d-dimensional workload process. This amounts to specializing the Brownian model developed by Harrison and Williams (1992) for a multiclass queueing network to the re-entrant line with the FBFS and LBFS disciplines. We simplify their formulas that convert queueing network data to Brownian model data. The Brownian data are calculated explicitly from the first and second moment information of the network. Furthermore, we prove that the proposed RBM exists, and is unique in distribution, and that it has a unique stationary distribution when the traffic intensity at each station is less than one. In the second step, we use an algorithm developed by Dai and Harrison (1992) to compute the stationary distribution of the RBM. We then use the steady-state mean of the RBM to estimate the long-run average workload at each station and the mean sojourn time in the network under each of the two priority disciplines. When the FBFS discipline is used, the steady-state mean of the RBM is used to develop a refined estimate of the mean sojourn time in the network. We present some numerical results to compare the estimates obtained from Brownian models with the estimates obtained from simulations. These results show that our Brownian estimation performs well for average workload, and very well for mean sojourn time under the FBFS discipline.

Performance analysis of queueing networks has mostly been restricted to the FIFO discipline. When all interarrival and service time distributions are exponential, Jackson (1957), Baskett et al. (1975), and Kelly (1975) derived exact formulas for performance measures. (The latter two papers also dealt with some non-FIFO disciplines.) If general distributions are allowed, one does not have exact formulas. Whitt (1983) developed a Queueing Network Analyzer to approximately predict the performance of Jackson-type networks. That method was generalized to multiclass networks by Bitran and Tirupati (1988) and Segal and Whitt (1989). In all these approximations, a network is decomposed into many single-station systems, and each system is analyzed separately. Shalev-Oren et al. (1985) developed a generalization of mean value analysis for closed queueing networks that allows static priorities at each station. Harrison and Nguyen (1990, 1993) adopted a completely different approach called the QNET method. They proposed to use d-dimensional RBM to model the d-dimensional workload process in an open multiclass network under the FIFO discipline. They also used an algorithm of Dai and Harrison (1992) to compute the stationary distribution of the RBM. In most cases, their performance estimates are quite accurate. Recently, there is a growing effort to study Brownian models of queueing networks with non-FIFO queueing disciplines. Coffman et al. (1995, 1996) developed Brownian models for single-station polling systems. Reiman and Wein (1996) developed Brownian models for polling systems in tandem.

Harrison and Williams (1992) developed Brownian models of multiclass queueing networks under some general queueing disciplines, which include FIFO discipline, processor sharing discipline, and static buffer priority disciplines. However, they had not attempted to establish the existence of their Brownian models, nor had they assessed the accuracy in using their Brownian models to predict performance measures of queueing networks. There are two major reasons for us to study Brownian models of the FBFS and LBFS disciplines. First, these disciplines have been proven to be stable as long as the traffic intensity at each station is less than one. (See Kumar 1993, Dai and Weiss 1996, and Kumar and Kumar 1996.) Recent work on the stability of multiclass queueing networks shows that such stability property is *not* shared by all priority disciplines or by the widely used FIFO discipline. (See Kumar and Seidman 1990, Lu and Kumar 1991, Rybko and Stolyar 1992, Bramson 1994, and Seidman 1994.) The stability of the FBFS and LBFS disciplines indicates that these disciplines are reasonably good, at least when they are compared with the FIFO discipline; therefore, it is important to understand their performance. The other motivation of our work is to present a pilot study for Brownian models of priority disciplines. It is unclear at this point which type of priority disciplines Brownian models can be applied for.

Our Brownian model is based on a heuristic that is rooted in heavy traffic theory. A heavy traffic limit theorem would assert that as the traffic intensity at each station approaches the critical value of one, the normalized queue length process in high-priority classes converges to zero, and the normalized d-dimensional workload process converges to an RBM. In particular, the theorem would assert that at each station only customers of the lowest priority are in the queue in heavy traffic. As of this writing, there is no such heavy traffic limit theorem for a re-entrant line under a general priority discipline; however, Chen (1994) has recently shown that for some two-station re-entrant lines under the FBFS and LBFS disciplines, the heavy traffic limit theorem indeed prevails. It is believed that such a limit theorem holds in general under the FBFS and LBFS disciplines.

Multidimensional RBMs were first introduced by Harrison and Reiman (1981) to approximate Jackson-type networks. These approximations were justified by the so-called heavy traffic limit theorems in Reiman (1984), Johnson (1983), and Chen and Mandelbaum (1991). There has been much recent progress in the study of RBMs. For example, see Harrison and Williams (1987), Reiman and Williams (1988), Dai and Harrison (1992), Taylor and Williams (1993), Dupuis and Williams (1994), and Dai and Kurtz (1994). With the recent work of Taylor and Williams (1993), RBMs can be used to model *multiclass* networks under FIFO discipline; however, the corresponding heavy traffic limit theorems are limited either to multiclass single-station systems or to feedforward networks. (See Reiman 1988, Dai and Kurtz 1995, and Peterson 1991.) Readers are referred to Harrison and Nguyen (1993) for the current status of heavy traffic limit theorems under FIFO discipline.

The following conventions will be used in this paper. All vectors are envisioned as column vectors unless stated otherwise. Vector inequalities are interpreted componentwise. A prime on a vector or a matrix means transpose. The symbol $\Rightarrow$ denotes weak convergence of stochastic processes (cf. Ethier and Kurtz 1986). We end this introduction by outlining the rest of the paper. The re-entrant multiclass queueing network model is introduced in Section 1. The background materials on RBM are discussed in Section 2. The derivation of the Brownian model for our queueing network is given in Section 3. The performance analysis procedure is summarized in Section 4. In Sections 5 through 7 we present three network examples, where Brownian estimates are compared with simulation estimates. The article concludes in Section 8 with some open problems. Readers can first read Section 1 to get the model description, and proceed directly to Section 4, and then to the examples to get an overview of our method.

1. MODEL DESCRIPTIONS AND PRELIMINARIES

Consider a d-station re-entrant queueing network, or a re-entrant line. Customers arrive at station 1 from the outside according to a general process $E_1 = \{E_1(t), t \geq 0\}$, where $E_1(t)$ is the cumulative number of arrivals by time t. Each customer follows a deterministic route: $\sigma(1) = 1$, $\sigma(2), \ldots, \sigma(K)$, where $\sigma(k)$ is the station number that the customer visits during the kth stage of its service. We designate customers in their kth visit as *class k* customers, and they wait in buffer k for service at station $\sigma(k)$. We assume that each station has a single server and that the buffer size for each class is infinite. In the example shown in Figure 1, there are two stations ($d = 2$) and three customer classes ($K = 3$). Customers of classes 1, 2, and 3 visit stations $\sigma(1) = 1$, $\sigma(2) = 2$, and $\sigma(3) = 1$, respectively. Two classes of customers are served at station 1, competing services from server 1.

Let $v_k(i)$ be the service time for the ith class k customer. Define:

$$S_k(t) = \max\{n : n \geq 0,\ v_k(1) + \cdots + v_k(n) \leq t\}, \qquad t \geq 0, \quad k = 1, \ldots, K. \tag{1.1}$$

We interpret $S_k(t)$ as the number of class k services completed in the first t units of time that are devoted by server $\sigma(k)$ to the service of class k. We call $S_k = \{S_k(t), t \geq 0\}$ class k service process. We assume that the arrival process $E_1 = \{E_1(t), t \geq 0\}$ is independent of the service processes $S_1, \ldots, S_K$, and that as $t \to \infty$, almost surely:

$$\frac{E_1(t)}{t} \to \alpha_1, \tag{1.2}$$

$$\frac{S_k(t)}{t} \to \mu_k, \quad k = 1, \ldots, K, \tag{1.3}$$

for some constants $\alpha_1 > 0$, $\mu_k > 0$ ($k = 1, \ldots, K$). We interpret α_1 the external arrival rate and μ_k the service rate for class k jobs if server $\sigma(k)$ devotes its full efforts to class k. We further assume that the arrival process and service processes satisfy a *functional central limit theorem*. That is, there exist a constant $c_a^2 > 0$ and a $K \times K$ positive definite matrix Γ such that as $n \to \infty$:

$$\frac{1}{\sqrt{n}}(E_1(n\cdot) - \alpha_1 n\cdot) \Rightarrow \xi^a(\cdot), \tag{1.4}$$

$$\frac{1}{\sqrt{n}}(S_1(n\cdot) - \mu_1 n\cdot, \ldots, S_K(n\cdot) - \mu_K n\cdot)' \Rightarrow \xi^s(\cdot) = (\xi_1^s(\cdot), \ldots, \xi_K^s(\cdot))', \tag{1.5}$$

where $\xi^a(\cdot)$ is a one-dimensional, driftless Brownian motion with variance $\alpha_1 c_a^2$, ξ^s is a K-dimensional, driftless Brownian motion with covariance matrix Γ and is independent of ξ^a.

We call a re-entrant queueing network satisfying the *standard assumptions* if interarrival times are iid with mean $1/\alpha_1$ and variance c_a^2/α_1^2, service processes $S_1, \ldots, S_K$ are independent, and class k service times are iid with mean m_k and variance $m_k^2 c_{s,k}^2$ ($k = 1, \ldots, K$). Such re-entrant line is called a *standard network*. The variability parameters c_a^2 and $c_{s,k}^2$ are the squared coefficient of variation (SCV) of the interarrival times and class k service times. (The squared coefficient of variation SCV of a random variable is defined to be variance divided by squared mean.) For a standard network, conditions (1.2) through (1.5) hold with service rate $\mu_k = 1/m_k$ and the covariance matrix of the Brownian motion in (1.5) being:

$$\Gamma = \mathrm{diag}(c_{s,1}^2/m_1, \ldots, c_{s,K}^2/m_K).$$

Our assumption (1.5) holds under much weaker assumptions. For example, (1.5) holds when the service time sequence $\{(v_1(i), \ldots, v_K(i))', i \geq 1\}$ is iid with finite second moment, and for fixed i the service times $(v_1(i), \ldots,$

$v_K(i))'$ at different stages of service have arbitrary dependencies. The latter feature is useful for certain applications, notably in computer communications and manufacturing systems. There, the length of a computer message or the size of a manufacturing lot may be random; however, the service times in general are proportional to the message length or the lot size, and therefore are positively correlated.

Without loss of generality, we assume that $\alpha_1 = 1$ in the rest of this paper. Let $m_k = 1/\mu_k$ be the mean service time for class k customers. We define the nominal workload per unit of time at station i:

$$\rho_i = \sum_{k \in \mathscr{C}_i} m_k, \quad i = 1, \ldots, d, \tag{1.6}$$

where $\mathscr{C}_i = \{k : 1 \leqslant k \leqslant K, \sigma(k) = i\}$ is the constituency of station i. Throughout this paper, we assume that:

$$\rho_i < 1, \quad i = 1, \ldots, d. \tag{1.7}$$

For future use, we define the $d \times K$ incidence matrix $C = (C_{ik})$:

$$C_{ik} = \begin{cases} 1 & \text{if } \sigma(k) = i, \\ 0 & \text{otherwise.} \end{cases} \tag{1.8}$$

Let $Q_k(t)$ be the number of class k customers in station $\sigma(k)$ at time t, including possibly the one being served. For each station i, define the *workload process* at the station:

$$Z_i(t) = \sum_{k \in \mathscr{C}_i} m_k Q_k(t), \quad i = 1, \ldots, d.$$

Intuitively, $Z_i(t)$ is the *average* amount of work for server i if no more arrivals are allowed to station i after time t. Let $Q(t) = (Q_1(t), \ldots, Q_K(t))'$ and $Z(t) = (Z_1(t), \ldots, Z_d(t))'$. In vector form, we have:

$$Z(t) = CMQ(t),$$

where $M = \operatorname{diag}(m_1, \ldots, m_K)$. We call $Z = \{Z(t), t \geqslant 0\}$ the station-level workload process, or simply the workload process. Our definition of workload process is slightly different from the traditional one as defined in Harrison and Nguyen (1990, 1993). More will be said on this at the end of Section 2. The workload process is the key process that we will study in this paper. In particular, we will estimate a primary performance measure:

$$\bar{Z}_i \equiv \lim_{t \to \infty} \frac{1}{t} \int_0^t Z_i(s)\, ds, \quad (i = 1, \ldots, d), \tag{1.9}$$

which measures the long-run average workload at each station. When interarrival and service times are iid with finite second moments, the long run average queue length:

$$\bar{Q}_k \equiv \lim_{t \to \infty} \frac{1}{t} \int_0^t Q_k(s)\, ds, \quad (k = 1, \ldots, K)$$

exists and is finite under the FBFS and LBFS disciplines. (See Dai and Weiss 1996 and Dai and Meyn 1995.) Thus, the long-run average workload defined in (1.9) exists and is equal to:

$$\bar{Z}_i = \sum_{k \in \mathscr{C}_i} m_k \bar{Q}_k.$$

Another performance measure commonly used is the mean sojourn (flow) time $\bar{F}$ of each job defined to be:

$$\bar{F} \equiv \lim_{n \to \infty} \frac{F_1 + \cdots + F_n}{n},$$

where F_n is the sojourn time in the network for the nth job. By Little's law, the mean sojourn time $\bar{F}$ exists when the long-run average queue lengths are finite. Furthermore:

$$\bar{F} = \bar{Q}_1 + \cdots + \bar{Q}_K, \tag{1.10}$$

when the external arrival rate $\alpha_1 = 1$.

Even when all distributions are exponential, there is no analytical method to estimate $\bar{Q}_k$ $(k = 1, \ldots, K)$ in a re-entrant line. We are going to devise a method to estimate the long-run average workload by using an RBM described in the next section.

2. REFLECTED BROWNIAN MOTION

Let Σ be a $d \times d$ positive definite matrix, θ be a d-dimensional vector, and R be a $d \times d$ matrix. A d-dimensional continuous process $Z^* = \{Z^*(t), t \geqslant 0\}$ is said to be a (θ, Σ, R)-RBM (see Taylor and Williams 1993 for a more precise definition) if:

$$Z^*(t) = X(t) + RY(t) = X(t) + \sum_{i=1}^{d} v_i Y_i, \tag{2.1}$$

$$Z^*(t) \geqslant 0, \tag{2.2}$$

$$Y(\cdot) \text{ is nondecreasing and continuous with } Y(0) = 0, \tag{2.3}$$

$$Y_i(\cdot) \text{ increases only at times } t \text{ when } Z_i^*(t) = 0, \quad i = 1, \ldots, d, \tag{2.4}$$

where $X = \{X(t), t \geqslant 0\}$ is a d-dimensional Brownian motion with drift θ and covariance matrix Σ, and v_i is the ith column of the *reflection matrix* R. Note that the RBM Z^* is confined to the orthant $\Re_+^d$. Heuristically, the behavior of an RBM Z^* may be described as follows. The process Z^* behaves like a Brownian motion in the interior of the orthant, and it is confined to the orthant by instantaneous "reflection" (or "pushing") by $Y_i(\cdot)$'s at the boundary, where the direction of reflection on the ith face $\{x \in \Re_+^d : x_i = 0\}$ is given by v_i.

In order for an RBM to exist, the directions of reflection must point to the interior of the orthant. To fully describe the condition on the directions of reflection, we need the following definition.

Definition 2.1. A square matrix $A = (v_1, \ldots, v_d)$ is an $\mathscr{S}$ matrix if there exist $a_i > 0$ $(i = 1, \ldots, d)$ such that $\sum_{i=1}^d a_i v_i > 0$, where v_i is the ith column of A. A square matrix A is a *completely $\mathscr{S}$ matrix* if each principal submatrix of A is an $\mathscr{S}$ matrix.

Reiman and Williams (1988) proved that a necessary condition for existence of process Z^* is that R is a *completely* $\mathscr{S}$ matrix. Conversely, Taylor and Williams (1993) proved the following foundational result: if R is *completely* $\mathscr{S}$, then a (θ, Σ, R)-RBM Z^* exists and Z^* is unique in distribution.

We have intentionally used symbol Z^* in our description of the RBM. The resemblance to the queueing network analog is to emphasize the connection of this process with the workload process in our queueing model. Specifically, as we will see in later sections, the workload process Z defined in Section 1 will be replaced by an RBM Z^*. We hope such usage does not cause any confusion.

3. REDUCTION TO AN RBM

In this section we present an RBM that approximates the workload process in the queueing network. In the Appendix of Harrison and Williams (1992), the authors presented Brownian models for multiclass queueing networks under several queueing disciplines. Their multiclass queueing networks include re-entrant lines considered in this paper. Among the queueing disciplines that they considered are static buffer priority disciplines, which include the FBFS and LBFS disciplines. Harrison and Nguyen (1990, 1993) also studied Brownian models for multiclass queueing networks, but their queueing discipline primarily focused on FIFO.

According to Harrison and Williams (1992), the workload process $Z = \{Z(t), t \geqslant 0\}$ can be modeled by a (θ, Σ, R)-RBM $Z^* = \{Z^*(t), t \geqslant 0\}$ as defined in Section 2. The Brownian data (θ, Σ, R) were given on page 288 of Harrison and Williams (1992). Specifically, the reflection matrix R was given by:

$$R = (I + G)^{-1}, \tag{3.1}$$

where

$$G = CM(I - P')^{-1}P'\Delta, \tag{3.2}$$

P is a $K \times K$ matrix whose entries are zero, except that $P_{k,k+1} = 1$ for $k = 1, \ldots, K - 1$, and Δ is a $K \times d$ matrix to be explained shortly. The drift vector θ was given as:

$$\theta = -R(e - \rho),$$

where e is a d-dimensional vector of ones, and ρ is the vector of traffic intensities as defined in (1.6). The covariance matrix Σ was given as $R\tilde{\Gamma}R'$, where $\tilde{\Gamma}$ was defined through expressions in (A.1), (A.12), (A.17), (A.29), and (A.51) of Harrison and Williams (1992). Specializing to our model, we have:

$$\Sigma = RC[\Gamma + \alpha_1 c_a^2 Mee'M]C'R'.$$

It is expected that the Brownian data should depend on a particular queueing discipline used. It turns out that the matrix Δ used in (3.2) does depend on the queueing discipline used. The determination of Δ is often based on a *heavy traffic* theory, although the resultant Δ dominates in any traffic conditions. The system is in heavy traffic if ρ_i is less than one, but is close to one at each station i. For some queueing networks, Reiman (1984b) discovered what he called the *state space collapse* phenomenon. That is, the queue lengths among different classes at a station are a fixed proportion of the workload at the station under a heavy traffic scaling. For a given queueing discipline, if there is a state space collapse, namely:

$$Q_k(t) \approx \delta_k Z_i(t) \quad \text{for each } k \in \mathscr{C}_i \tag{3.3}$$

in heavy traffic, then Harrison and Williams (1992) suggested to take:

$$\Delta_{ki} = \begin{cases} \delta_k, & \text{if } k \in \mathscr{C}_i, \\ 0, & \text{otherwise.} \end{cases}$$

The state space collapse is usually a key to proving a heavy traffic limit theorem, which is often used to justify Brownian approximations of the type proposed by Harrison and Williams (1992) and Harrison and Nguyen (1993). Under the FBFS and LBFS disciplines, when the system is in heavy traffic, we expect that customers in the lowest priority class experience most of the waiting. In fact, Johnson (1983) and Peterson (1991) proved that for certain multiclass queueing networks, under heavy traffic scaling, only the lowest priority class has nonempty queue in heavy traffic. This suggests that:

$$Z_i(t) \approx m_{\tau(i)} Q_{\tau(i)},$$

where $\tau(i)$ is the lowest priority class at station i. Therefore:

$$\Delta_{ki} = \begin{cases} 1/m_k, & \text{if } k \text{ is the lowest priority class in } \mathscr{C}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Because $CM\Delta = I$, one can check that:

$$I + G = CM(I - P')^{-1}\Delta.$$

Because of the special structure of P, $(I - P')^{-1}$ is a lower triangular matrix with each entry in the lower triangular part being equal to 1. Therefore, the (i, j)th entry of $I + G$ is:

$$\left(\sum_{k \in \mathscr{C}_i, k \geqslant \ell(j)} m_k \right) \frac{1}{m_{\ell(j)}}. \tag{3.5}$$

The following two theorems are proved in the appendix.

Theorem 3.1. *Under the FBFS and LBFS disciplines,* $(I + G)$ *is invertible and* $R = (I + G)^{-1}$ *is a completely* $\mathscr{S}$ *matrix as defined in Definition* 2.1. *Therefore, the* (θ, Σ, R)*-RBM* Z^* *defined in* (2.1) *although* (2.4) *exists and is unique in distribution.*

Theorem 3.2. *Under the FBFS and LBFS disciplines, the RBM* Z^* *has a stationary distribution when the traffic condition* (1.7) *is satisfied.*

It was shown by Dai and Kurtz (1994) that the stationary distribution of Z^* was unique. Furthermore, it was characterized by a basic adjoint relationship. (See Harrison and

Williams 1987.) The stationary distribution of the RBM can be computed by a numerical algorithm devised by Dai and Harrison (1992). When the reflection matrix R and covariance matrix Σ satisfy a special condition, the stationary distribution of the RBM is of an exponential form that can be determined analytically. The condition, called the skew symmetric condition for historical reasons as in Harrison and Williams (1987), takes the form:

$$2\Sigma = RD^{-1}\Lambda + \Lambda D^{-1}R', \tag{3.6}$$

where $D = \text{diag}(R)$ and $\Lambda = \text{diag}(\Sigma)$. When (3.6) is satisfied, the stationary density has a *product-form* given by:

$$\kappa \exp(-\eta' x), \quad x = (x_1, \ldots, x_d)' \in \Re^d_+ \tag{3.7}$$

with $\eta = -2\Lambda^{-1}\, D\, R^{-1}\theta$ and some normalizing constant $\kappa > 0$.

In Harrison and Williams (1992), the RBM Z^* was used to model what they called the *immediate workload process* $W(t) = (W_1(t), \ldots, W_d(t))'$, where $W_i(t)$ is the sum of the impending service times of customers who are queued at station i at time t, plus the remaining service times of those customers (if any) who are being serviced there at time t. Readers might have noticed that their definition of $W_i(t)$ is slightly different from the definition of average workload $Z_i(t)$ defined in this paper. If all service times at station i are deterministic, and server i has just initiated a new service at time t, then $W_i(t)$ is equal to $Z_i(t)$. In general, when the system is in heavy traffic, there will be a lot of customers waiting in buffer $k = \ell(i)$ at each station i. By the strong law of large numbers (1.5), the amount of immediate workload contributed by this class is roughly $m_k Q_k(t)$. Because workload contribution from other classes are negligible, we have:

$$Z(t) \approx W(t).$$

Therefore, Z and W should yield the same Brownian model Z^*.

4. SUMMARY OF THE PERFORMANCE ANALYSIS PROCEDURE

Recall that the input data for the re-entrant queueing network contain external arrival rate $\alpha_1 = 1$, SCV c_a^2 for interarrival times, service rate μ_k $(k = 1, \ldots, K)$, covariance matrix Γ that captures the variability of service times for each class, as well as the correlation of service times among different classes. The queueing discipline used is either FBFS or LBFS. For a given priority discipline, let $\ell(i)$ be the lowest priority class at station i.

In Section 3 we have shown that the workload process Z can be replaced by an RBM Z^* with drift vector θ, covariance matrix Σ, and reflection matrix R, or simply a (θ, Σ, R)-RBM. To repeat, the Brownian data (θ, Σ, R) are given by:

$$R = \text{diag}(m_{\ell(1)}, \ldots, m_{\ell(d)}) \Bigg(\sum_{k \in \mathscr{C}_i, k \geqslant \ell(j)} m_k \Bigg)^{-1}, \tag{4.1}$$

$$\theta = -R(e - \rho), \tag{4.2}$$

$$\Sigma = RC[\Gamma + \alpha_1 c_a^2 Mee'M]C'R', \tag{4.3}$$

where, as before, $m_k = 1/\mu_k$, $\mathscr{C}_i$ is the constituency of station i, e is a d-dimensional vector of ones, ρ is the vector of traffic intensities, $M = \text{diag}(m_1, \ldots, m_K)$, and C is the constituency matrix. Notice that the second moment information is contained in Σ only. Also, the reflection matrix R, and hence the drift vector θ, covariance matrix Σ of the RBM Z^* depend on the discipline used.

It has been proven in Theorems 3.1 and 3.2 that the (θ, Σ, R)-RBM Z^* exists, that it is unique in distribution, and when the traffic condition (1.7) is satisfied, Z^* has a unique stationary distribution. The stationary distribution of the RBM can be computed by a numerical algorithm, which has been implemented in a QNET software package that can be run on virtually any type of computer platform. In particular, the long-run average position:

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t Z_i^*(s)\, ds \quad (i = 1, \ldots, d)$$

of the RBM can be computed. We use these numbers to estimate the long-run average workloads $\bar{Z}_i$ $(i = 1, \ldots, d)$ in the network. By the heavy traffic state space collapse heuristic alluded in Section 3, we can further obtain rough estimates of average queue lengths:

$$\bar{Q}_k \approx \begin{cases} \dfrac{1}{m_k} \bar{Z}_i, & \text{if } k \text{ is the lowest priority class at station } i, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

The mean sojourn time $\bar{F}$ can be estimated based on mean queue length $(\bar{Q}_1, \ldots, \bar{Q}_K)'$ and Little's law as in (1.10).

When the queueing discipline is FBFS, we propose the following recursive procedure to obtain a *refined* estimate of $(\bar{Q}_1, \ldots, \bar{Q}_K)'$ from the estimate of mean workload. First, observe that in analyzing the first L $(L \leqslant K)$, buffers $1, \ldots, L$, the rest of the buffers $L + 1, \ldots, K$ can be ignored because of the FBFS discipline. We call the subnetwork consisting of buffers $1, \ldots, L$ the L-truncated network. We choose the largest $K_1 \leqslant K$ such that there is exactly one class at each station in the K_1-truncated network. For $k = 1, \ldots, K_1$, $\bar{Q}_k$ can be calculated via $\bar{Q}_k = \bar{Z}_i^{(1)}/m_k$, where $\bar{Z}_i^{(1)}$ is the mean workload at station $i = \sigma(k)$ in the K_1-truncated network. If $K_1 = K$, we are done; otherwise, choose the largest K_2 $(K_1 < K_2 \leqslant K)$ such that in the K_2-truncated network each station has, at most, one job class that has an unknown mean queue length. For any station i in the K_2-truncated network, there are, at most, two classes visiting the station. If station i has one job class k, then $\bar{Q}_k$ can be computed as before via $\bar{Q}_k = \bar{Z}_i^{(2)}/m_k$; otherwise, there are two classes, k and ℓ, with $k \leqslant K_1$ and $K_1 < \ell \leqslant K_2$. In this case, $\bar{Q}_k$ has been computed in the K_1-truncated network. Furthermore, $\bar{Q}_\ell$ can be computed via:

$$\bar{Q}_\ell = (\bar{Z}_i^{(2)} - m_k \bar{Q}_k)/m_\ell,$$

where $\bar{Z}_i^{(2)}$ is the mean workload at station i in the K_2-truncated network. Thus, we have obtained estimates for

$\bar{Q}_k$ ($k = 1, \ldots, K_2$). Continuing in this way, we will eventually have $K_1 < K_2 < \cdots < K_\ell = K$ and an estimate of $\bar{Q}_k$ for $k = 1, \ldots, K$.

In the next three sections, we present numerical studies for three re-entrant lines.

5. A SIMPLE RE-ENTRANT LINE

Let us come back to the two-station network shown in Figure 1. Assume that the standard assumptions in Section 1 hold, and the exogenous arrival process to class 1 is Poisson with rate 1. For this network, the workloads $Z(t) = (Z_1(t), Z_2(t))'$ at time t at both stations are:

$$Z_1(t) = m_1 Q_1(t) + m_3 Q_3(t),$$
$$Z_2(t) = m_2 Q_2(t).$$

Under the FBFS discipline, the workload process Z can be replaced by a (θ, Σ, R)-RBM Z^* with:

$$R = \begin{pmatrix} 1 & -m_3/m_2 \\ 0 & 1 \end{pmatrix},$$
$$\theta = (m_1 + m_3/m_2 - 1, m_2 - 1)',$$
$$\Sigma = \begin{pmatrix} m_1^2(c_a^2 + c_{s,1}^2) + m_3^2(c_{s,2}^2 + c_{s,3}^2) & m_1 m_2 c_a^2 - m_2 m_3 c_{s,2}^2 \\ m_1 m_2 c_a^2 - m_2 m_3 c_{s,2}^2 & m_2^2(c_a^2 + c_{s,2}^2 \end{pmatrix}.$$

Because of the form of the reflection matrix, the second component $Z_2^*(t)$ of Z^* is a one-dimensional RBM. Therefore, its stationary distribution is exponential with mean (for example, see Harrison 1985):

$$\frac{\Sigma_{22}}{2(1 - m_2)} = m_2 \left(\frac{m_2}{1 - m_2}\right)\left(\frac{1 + c_{s,2}^2}{2}\right),$$

which is identical to the familiar Pollaczek-Kinchine formula for a $M/G/1$ queue. (See, for example, Gross and Harris 1985.) This result is not surprising, because under our heuristics customers entering system experience no waiting at station 1 and proceed directly to station 2. Therefore, station 2 acts like an $M/G/1$ queue, for which the Pollaczek-Kinchine formula prevails. For the RBM, condition (3.6) is equivalent to:

$$-m_2 m_3 (c_a^2 + c_{s,2}^2) = 2 m_1 m_2 c_a^2 - 2 m_2 m_3 c_{s,2}^2,$$

which reduces to:

$$(2m_1 + m_3) c_a^2 = m_3 c_{s,2}^2.$$

Therefore, in our case where the arrival process is Poisson, the corresponding RBM has product form stationary distribution if, and only if, $c_{s,2}^2 = (2m_1 + m_3)/m_3$.

Under the LBFS discipline, the workload process Z can be replaced by a (θ, Σ, R)-RBM Z^* with:

$$R = \begin{pmatrix} 1 & -m_3/m_2 \\ -m_2/m_1 & 1 + m_3/m_1 \end{pmatrix},$$
$$\theta = (m_1 + m_3/m_2 - 1, \quad m_2/m_1 - m_3/m_1 - 1)',$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where

$$\Sigma_{11} = m_1^2(c_a^2 + c_{s,1}^2) + m_3^2(c_{s,2}^2 + c_{s,3}^2),$$
$$\Sigma_{12} = \Sigma_{21} = -m_1 m_2 c_{s,1}^2 - m_2 m_3 (1 + m_3/m_1) c_{s,2}^2 - (m_2/n_1) m_3^2 c_{s,3}^2,$$
$$\Sigma_{22} = m_2^2 c_{s,1}^2 + (1 + m_3/m_1)^2 m_2^2 c_{s,2}^2 + (m_2/m_1)^2 m_3^2 c_{s,3}^2.$$

Unlike the FBFS case, neither of the marginal distributions of Z^* is a one-dimensional RBM. Condition (3.6) reduces to:

$$\tilde{R}_{12}\Sigma_{22} + \tilde{R}_{21}\Sigma_{11} = 2\Sigma_{12},$$

where:

$$\tilde{R} = RD^{-1} = \begin{pmatrix} 1 & -m_1 m_3/[m_2(m_1 + m_3)] \\ -m_2/m_1 & 1 \end{pmatrix},$$

which can be further reduced to:

$$m_1(m_1 + m_3) c_a^2 = m_1^2 c_{s,1}^1 + m_3(m_1 + m_3) c_{s,2}^2 + m_3^2 c_{s,3}^2. \tag{5.1}$$

Notice that condition (5.1) does not involve mean service time at station 2. When service times at station 1 are deterministic ($c_{s,1}^2 = c_{s,3}^2 = 0$), and service times at station 2 are exponentially distributed ($c_{s,2}^2 = 1$), condition (5.1) is satisfied; hence, the RBM has a product form stationary distribution.

The proposed method in this paper gives estimates of the mean workloads $\bar{Z}_1 = m_1\bar{Q}_1 + m_3\bar{Q}_3$ and $\bar{Z}_2 = m_2\bar{Q}_2$. According to the heavy traffic state space collapse heuristic, under the FBFS discipline, we have:

$$\bar{Q}_1 = 0, \quad \bar{Q}_2 = \frac{1}{m_2}\bar{Z}_2, \quad \bar{Q}_3 = \frac{1}{m_3}\bar{Z}_1. \tag{5.2}$$

Under the LBFS discipline, we have:

$$\bar{Q}_1 = \frac{1}{m_1}\bar{Z}_1, \quad \bar{Q}_2 = \frac{1}{m_2}\bar{Z}_2, \quad \bar{Q}_3 = 0. \tag{5.3}$$

Finally, the mean sojourn time $\bar{F}$ under either of the queueing disciplines can also be estimated via $\bar{F} = \bar{Q}_1 + \bar{Q}_2 + \bar{Q}_3$.

Under the FBFS discipline, one can obtain a refined estimate of $(\bar{Q}_1, \bar{Q}_2, \bar{Q}_3)$ from $(\bar{Z}_1, \bar{Z}_2)$, as discussed in Section 4. Because the FBFS discipline is used, buffer 1 can be analyzed in isolation, ignoring the existence of other buffers. Thus the 1-truncated network is an $M/G/1$ queue, and our Brownian estimate of the mean workload is:

$$m_1\left(\frac{m_1}{1 - m_1}\right)\left(\frac{1 + c_{s,1}^2}{2}\right).$$

Table I
Estimates of Mean Workloads and Mean Sojourn Times in a Two-station Network

		FBFS		LBFS	
Case Number	Station Number	SIMAN	Refined QNET	SIMAN	QNET
A-1	1	4.74 (5.6%)	4.83 (1.9%)	3.53 (4.6%)	3.58 (1.4%)
	2	7.93 (3.0%)	8.10 (2.1%)	9.46 (3.1%)	9.99 (5.6%)
	Sojourn	19.4 (3.1%)	19.73 (1.7%)	18.3 (2.7%)	19.1 (4.1%)
A-2	1	5.95 (7.2%)	5.92 (0.5%)	5.33 (6.3%)	5.25 (1.5%)
	2	7.91 (2.9%)	8.10 (2.4%)	8.11 (3.3%)	8.36 (3.1%)
	Sojourn	32.2 (5.6%)	32.77 (1.8%)	17.2 (3.3%)	16.8 (2.4%)
A-3	1	5.59 (4.9%)	5.85 (4.7%)	3.44 (3.1%)	3.48 (1.2%)
	2	8.16 (3.0%)	8.10 (0.7%)	15.4 (3.5%)	18.63 (21.0%)
	Sojourn	17.7 (3.4%)	17.54 (0.9%)	28.5 (3.3%)	37.8 (32.6%)
B-1	1	5.62 (4.8%)	6.18 (10.0%)	4.19 (2.9%)	4.51 (7.6%)
	2	8.59 (3.2%)	8.10 (5.7%)	10.7 (3.6%)	11.3 (5.7%)
	Sojourn	22.0 (3.2%)	22.73 (3.3%)	21.1 (2.3%)	22.6 (7%)
B-2	1	9.37 (5.9%)	10.8 (15.2%)	8.41 (4.2%)	9.41 (11.9%)
	2	10.1 (2.9%)	8.10 (19.8%)	11.3 (3.8%)	11.8 (4.4%)
	Sojourn	50.0 (4.2%)	51.33 (2.6%)	25.8 (3.5%)	26.6 (2.9%)
B-3	1	4.32 (3.3%)	4.33 (0.2%)	2.83 (2.3%)	2.63 (7.1%)
	2	8.05 (2.9%)	8.10 (0.6%)	12.9 (3.3%)	15.2 (18.2%)
	Sojourn	15.5 (3.9%)	15.54 (0.3%)	23.7 (2.7%)	30.0 (26.7%)
C-1	1	5.88 (5.7%)	6.76 (15.0%)	4.18 (4.5%)	4.60 (10.0%)
	2	6.14 (2.9%)	5.06 (17.6%)	7.93 (3.9%)	7.93 (0.0%)
	Sojourn	20.3 (3.9%)	20.64 (1.7%)	18.1 (3.4%)	19.0 (5.2%)
C-2	1	9.36 (5.8%)	11.0 (17.4%)	8.40 (5.5%)	9.41 (12.0%)
	2	7.79 (3.1%)	5.06 (35.0%)	8.71 (4.0%)	8.72 (0.1%)
	Sojourn	47.3 (6.1%)	48.96 (3.5%)	22.5 (3.5%)	23.1 (2.6%)
C-3	1	4.97 (6.1%)	5.36 (7.8%)	2.67 (3.4%)	2.68 (0.4%)
	2	5.63 (3.1%)	5.06 (10.1%)	10.6 (3.0%)	11.94 (12.6%)
	Sojourn	14.3 (2.8%)	13.64 (4.8%)	21.2 (3.1%)	26.7 (25.8%)

The Brownian estimate of the mean queue length $\bar{Q}_1$ is:

$$\left(\frac{m_1}{1-m_1}\right)\left(\frac{1+c_{s,1}^2}{2}\right), \tag{5.4}$$

which differs from the exact value for $M/G/1$ queue:

$$m_1 + m_1\left(\frac{m_1}{1-m_1}\right)\left(\frac{1+c_{s,1}^2}{2}\right). \tag{5.5}$$

The relative error of our Brownian estimate is:

$$(1-m_1)\,\frac{|c_{s,1}^2-1|}{1+c_{s,1}^2}.$$

Once we have an estimate of $\bar{Q}_1$, we have the estimates of $\bar{Q}_2$ and $\bar{Q}_3$ given by:

$$\bar{Q}_2 = \bar{Z}_2/m_2,$$

$$\bar{Q}_3 = (\bar{Z}_1 - m_1\bar{Q}_1)/m_3.$$

Now we present some numerical experiments for this network. We consider three systems with different combinations of service time varabilities among classes. For systems A, B, and C, the corresponding service time SCVs ($c_{s,1}^2$, $c_{s,2}^2$, $c_{s,3}^2$) are given as (1, 1, 1), (3, 1, 0.25), and (3, 0.25, 1), respectively. For System A, all service times follow exponential distributions. Although such a system can be modeled by a continuous-time, discrete state space Markov chain, its stationary distribution is beyond the domain of exact analysis. In System B, class 3 has low variability service times, whereas in system C, class 2 has low variability service times. We should expect that the mean workload at station 2 be smaller in system C than in other systems, regardless of whether the FBFS or LBFS discipline is employed.

For every system, we fix the traffic intensity at both stations to be 0.90, which is reasonably heavy. For each system we consider three different cases. Each case corresponds to a different work allocation for server 1. In case 1, server 1 evenly splits its efforts between class 1 and class 3 customers; therefore, $m_1 = m_3 = 0.45$. In case 2, $m_1 = 0.7$ and $m_3 = 0.2$; therefore, server 1 devotes significantly more time on class 1 customers. For case 3, we let $m_1 = 0.2$ and $m_3 = 0.7$. Table I gives the simulation and QNET estimates of the long-run average workload at each station and the mean sojourn time in the network under the FBFS and LBFS disciplines. When the FBFS discipline is used, QNET estimates of the mean sojourn time are calculated from the refined procedure proposed in Section 4. The following conventions apply to this table, as well as to all subsequent tables. The column SIMAN contains the estimates obtained by simulations, and the numbers in

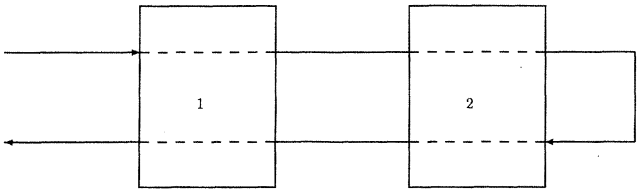

Figure 2. The Lu-Kumar network.

parentheses after the simulation estimates represent the half-width of 95-percent confidence intervals, which is expressed as a percentage of the simulation average. The estimates obtained by the proposed method are contained in the QNET column, and the numbers in the parentheses after the QNET estimates represent percentage errors from the simulation average. The simulations are performed using SIMAN 3.5. In all cases, simulation estimates are collected from 10 replications, and the simulation time of each replication is set to be 10^5, at which time the systems seem to have been in steady state for a long time. Since service time distributions are allowed to be general, in simulations we use Erlang distributions, exponential distributions, and hyperexponential distributions to fit service time distributions, with SCV being less than one, equal to one, and larger than one, respectively. A random variable is said to have hyperexponential distribution with mean m and SCV c^2 if its probability density function is given by:

$$f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \quad x \geq 0, \tag{5.6}$$

where $p = ½ + ½\sqrt{(c^2-1)/(c^2+1)}$, $\lambda_1 = 2p/m$ and $\lambda_2 = 2(1-p)/m$.

Notice that, as we discussed earlier, our QNET method cannot tell the difference among the three cases for long-run average workload at station 2 under the FBFS discipline, whereas simulation indicates that they are significantly different between Systems B and C. We expect this discrepancy to become smaller if the traffic intensities at both stations get higher. Also, under the FBFS discipline, in case 3 ($m_1 = 0.2$, $m_3 = 0.7$) of all systems the QNET estimates are more accurate. We attribute this accuracy to the small mean service time for class 1, which allows class 1 customers to pass through station 1 quickly. The quick passage of class 1 customers is consistent with the heavy traffic conjecture. Notice that the refined QNET estimates of the mean sojourn times under the FBFS discipline are extremely accurate compared with the simulation results. In calculating the QNET estimates of mean sojourn times, we used the Brownian estimates for mean queue length in formula (5.4) instead of the exact formula (5.5). Our calculations show that the relative difference of mean sojourn times based on these two estimates are insignificant (within 1%).

6. THE LU-KUMAR NETWORK

Shown in Figure 2 is the Lu-Kumar two-station re-entrant line. Customers enter the network from outside and follow a deterministic routing sequence given by stations 1, 2, 2, 1. Lu and Kumar (1991) show that if classes 2 and 4 receive higher priorities, the network may not be stable, even if (1.7) is satisfied.

Under the standard assumptions in Section 1, and the assumption that the exogenous arrival process is a Poisson process with rate 1, the FBFS and LBFS disciplines are stable, as proved by Dai and Weiss (1996), for any re-entrant lines. The workload process $Z(t)$ is given by:

$$Z_1(t) = m_1 Q_1(t) + m_4 Q_4(t),$$
$$Z_2(t) = m_2 Q_2(t) + m_3 Q_3(t).$$

Under the FBFS discipline, the workload process Z can be replaced by a (θ, Σ, R)-RBM Z^* with:

$$\theta = ((m_1 m_3 - m_2 m_4)/m_3 + m_4/m_3 - 1,$$
$$(m_2 + m_3) - 1)',$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$
$$R = \begin{pmatrix} 1 & -m_4/m_3 \\ 0 & 1 \end{pmatrix},$$

where:

$$\Sigma_{11} = (m_1 - m_2 m_4/m_3)^2 c_a^2 + m_1^2 c_{s,1}^2$$
$$+ (m_4/m_3)^2 m_2^2 c_{s,2}^2$$
$$+ m_4^2(c_{s,3}^2 + c_{s,4}^2),$$
$$\Sigma_{12} = \Sigma_{21} = (m_2 + m_3)(m_1 - m_2 m_4/m_3)c_a^2$$
$$- (m_4/m_3)m_2^2 c_{s,2}^2 - m_3 m_4 c_{s,3}^2,$$
$$\Sigma_{22} = (m_2 + m_3)^2 c_a^2 + m_2^2 c_{s,2}^2 + m_3^2 c_{s,3}^2.$$

The RBM Z^* exists and is unique in the pathwise sense because R is of upper-triangular form. Again, as in the FBFS case in Section 5, the Brownian estimates of the workload at station 2 can be analytically calculated without using the QNET software. In this network, the product-form condition (3.6) is equivalent to:

$$-m_4/m_3 \Sigma_{22} = 2\Sigma_{12}.$$

After a lengthy calculation, the condition is further reduced to:

$$(2m_1 m_2 + 2m_1 m_3 + m_3 m_4)c_a^2$$
$$= \frac{m_2^2 m_4}{m_3}(c_a^2 + c_{s,2}^2) + m_3 m_4 c_{s,3}^2.$$

Notice that $c_{s,1}^2$ does not play any role in the product-form condition.

Under the LBFS discipline, the workload process can be replaced by a (θ, Σ, R)-RBM with:

$$\theta = (m_1 + m_4/(m_2 + m_3) - 1, m_2/m_1$$
$$- m_2(m_1 + m_4)/[m_1(m_2 + m_3)])',$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$
$$R = \begin{pmatrix} 1 & -m_4/(m_2 + m_3) \\ -m_2/m_1 & m_2(m_1 + m_4)/[m_1(m_2 + m_3)] \end{pmatrix},$$

where:

Table II
Estimates of Mean Workloads and Mean Sojourn Times in the Lu-Kumar Network

		FBFS		LBFS	
Case Number	Station Number	SIMAN	Refined QNET	SIMAN	QNET
A-1	1	4.84 (1.8%)	4.84 (0.0%)	3.36 (1.9%)	3.28 (2.4%)
	2	5.46 (3.7%)	6.08 (11.4%)	3.64 (2.0%)	3.56 (2.2%)
	Sojourn	22.9 (2.6%)	24.3 (5.97%)	15.5 (1.94%)	15.2 (1.9%)
A-2	1	6.04 (3.0%)	5.87 (2.9%)	5.21 (2.3%)	5.22 (0.1%)
	2	6.02 (3.6%)	6.70 (11.3%)	5.13 (2.2%)	5.26 (2.5%)
	Sojourn	48.0 (4.2%)	51.2 (6.63%)	15.8 (2.53%)	15.0 (5.2%)
A-3	1	5.67 (2.0%)	5.87 (3.5%)	3.20 (1.6%)	3.00 (6.3%)
	2	6.25 (4.3%)	6.70 (7.2%)	3.36 (2.4%)	3.36 (0.0%)
	Sojourn	18.3 (2.7%)	18.3 (0.08%)	25.0 (2.4%)	31.8 (27.2%)
B-1	1	5.98 (4.4%)	6.20 (3.7%)	3.96 (3.8%)	4.20 (6.1%)
	2	6.27 (4.2%)	6.08 (3.0%)	4.31 (4.0%)	4.25 (1.4%)
	Sojourn	26.9 (2.6%)	27.3 (1.5%)	18.5 (3.9%)	18.8 (1.5%)
B-2	1	10.3 (4.6%)	10.8 (4.6%)	8.55 (4.2%)	9.35 (9.4%)
	2	8.48 (4.2%)	6.70 (21.0%)	7.93 (4.1%)	8.03 (1.3%)
	Sojourn	73.8 (5.2%)	66.5 (9.9%)	25.1 (4.0%)	24.8 (1.1%)
B-3	1	4.38 (3.7%)	4.36 (0.5%)	2.43 (2.8%)	2.14 (11.9%)
	2	6.17 (5.2%)	6.70 (8.6%)	2.78 (4.0%)	2.61 (6.1%)
	Sojourn	16.2 (3.1%)	16.4 (1.2%)	20.2 (3.5%)	23.8 (17.6%)

$$\Sigma_{11} = m_1^2(c_a^2 + c_{s,1}^2) + \left(\frac{m_4}{m_2 + m_3}\right)^2 (m_2^2 c_{s,2}^2 + m_3^2 c_{s,3}^2) + m_4^2 c_{s,4}^2,$$

$$\Sigma_{12} = \Sigma_{21} = -m_1 m_2 c_{s,1}^2 - \frac{m_2 m_4 (m_1 + m_4)}{m_1 (m_2 + m_3)^2} \cdot (m_2^2 c_{s,2}^2 + m_3^2 c_{s,3}^2) - (m_2/m_1) m_4^2 c_{s,4}^2,$$

$$\Sigma_{22} = m_2^2 c_{s,1}^2 + \left(\frac{m_2(m_1 + m_4)}{m_1(m_2 + m_3}\right)^2 (m_2^2 c_{s,2}^2 + m_3^2 c_{s,3}^2) + (m_2/m_1)^2 m_4^2 c_{s,4}^2.$$

For the RBM, condition (3.6) reduces to:

$$\frac{m_1^2}{m_1 + m_4} c_{s,1}^2 + m_1 c_a^2 + \frac{m_4}{(m_2 + m_3)^2} \cdot \left(-2m_2^2 - m_2^2 \frac{m_4}{m_1} + 1 + \frac{m_4}{m_1}\right) c_{s,2}^2 = \frac{m_3^2 m_4}{(m_2 + m_3)^2} c_{s,2}^2.$$

We consider two systems of this network. In System *A*, all service times are exponentially distributed. In System *B*, service time SCVs are (3, 1, 1, 0.25). Again, as in the first example, we fix the traffic intensities at both stations to be 0.90. For each system, we consider three cases, each case corresponds to a different combination of mean service times. In Case 1, all mean service times are the same, equal to 0.45. In Case 2, classes 1 and 2 have long mean service times (equal to 0.7), and classes 3 and 4 have short mean service times (equal to 0.2). In Case 3, classes 1 and 2 have short mean service times (equal to 0.2), and classes 3 and 4 have long mean service times (equal to 0.7). The numerical results are summarized in Table II. The conclusion is similar to the first example. In particular, the QNET estimates of the mean sojourn times under the FBFS discipline are again quite accurate.

7. A SIX-STATION RE-ENTRANT LINE

Consider the six-station queueing network depicted in Figure 3. In the network, customers enter from outside and follow a deterministic routing sequence given by stations 1, 2, 3, 2, 1, 4, 5, 6, 5, 4, 6. We assume the standard assumptions as in Section 1, and the exogenous arrival process is a Poisson process with rate 1.

In this example we consider two systems. In System A, all service time distributions are exponential. In System B, the service time SCVs are given:

(1, 2.25, 0.25, 1, 1, 0.25, 1, 2.25, 1, 1, 1).

In both systems, traffic intensities are 0.9 at all stations. Class 3 has mean service time 0.9, and the rest of the classes have mean service time 0.45. The numerical results are summarized in Table III, which shows that the QNET estimates are still encouraging. This example shows that

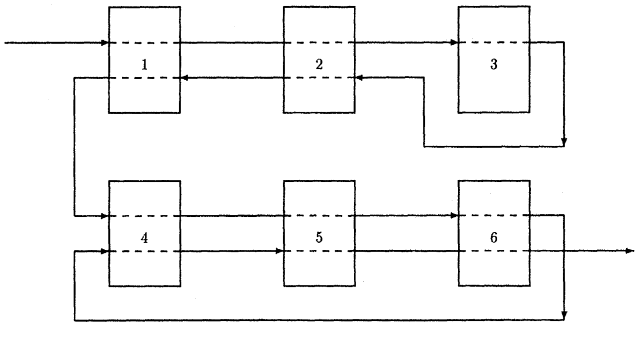

Figure 3. A six-station network.

Table III
Estimates of Mean Workloads and Mean Sojourn Times in a Six-station Network

		FBFS		LBFS	
System	Station Number	SIMAN	Refined QNET	SIMAN	QNET
A	1	4.66 (4.1%)	4.54 (4.54%)	3.70 (5.5%)	3.65 (1.4%)
	2	4.82 (4.5%)	4.82 (0.0%)	3.93 (4.4%)	3.73 (5.1%)
	3	8.36 (4.3%)	8.10 (3.1%)	10.2 (3.7%)	10.5 (2.9%)
	4	6.23 (5.3%)	5.32 (14.6%)	3.41 (3.1%)	3.20 (6.2%)
	5	9.81 (4.5%)	10.2 (4.0%)	3.80 (3.3%)	3.53 (7.1%)
	6	5.50 (4.3%)	4.69 (14.7%)	3.70 (2.6%)	3.63 (1.9%)
	Sojourn	74.9 (3.1%)	74.71 (0.25%)	53.3 (2.4%)	51.1 (4.1%)
B	1	5.29 (2.9%)	5.37 (1.7%)	3.60 (4.6%)	3.27 (9.2%)
	2	5.52 (5.9%)	5.88 (6.5%)	4.60 (2.6%)	4.55 (1.1%)
	3	5.89 (3.0%)	5.06 (14.1%)	7.30 (3.5%)	7.27 (0.4%)
	4	5.70 (5.0%)	4.43 (22.3%)	2.84 (2.5%)	2.31 (18.7%)
	5	10.4 (4.3%)	10.91 (4.9%)	3.63 (3.3%)	3.04 (16.3%)
	6	5.21 (3.5%)	5.50 (5.6%)	4.78 (3.1%)	4.39 (8.2%)
	Sojourn	79.8 (4.0%)	76.9 (3.6%)	49.3 (2.4%)	47.1 (4.4%)

the proposed method is quite robust in analyzing complicated networks. When the FBFS discipline is used, the refined QNET estimates of the mean sojourn times are calculated as follows. First, as discussed in Section 4, we let $K_1 = 3$. We obtain the QNET estimates of $\bar{Q}_1$, $\bar{Q}_2$ and $\bar{Q}_3$. Next, we let $K_2 = 8$ and obtain the QNET estimates of $\bar{Q}_4$, $\bar{Q}_5$, $\bar{Q}_6$, $\bar{Q}_7$ and $\bar{Q}_8$. Finally, letting $K_3 = K = 11$, we obtain QNET estimates of $\bar{Q}_9$, $\bar{Q}_{10}$ and $\bar{Q}_{11}$. The mean sojourn time in the network is $\bar{Q}_1 + \cdots + \bar{Q}_{11}$ by the Little's law.

8. CONCLUSIONS AND OPEN PROBLEMS

In this paper, we have presented a Brownian system model that can be used to predict the long-run average workload level at each station and the mean sojourn time in a re-entrant line under the FBFS and LBFS disciplines. When the discipline is FBFS, our method also yields a refined estimate of the mean sojourn (or flow) time in the system. The Brownian model was first proposed by Harrison and Williams (1992), in which a d-dimensional reflected Brownian motion was used to model the workload process in the re-entrant line. We show that the reflected Brownian motion exists, and that it is unique in distribution, and that it has a unique stationary distribution when the usual traffic condition (1.7) is satisfied. We also present three network examples in which performance estimates based on the Brownian model are shown to be reasonably accurate.

We have not attempted to prove a heavy traffic limit theorem that would justify the approximation procedure presented here. We conjecture that a properly normalized sequence of workload processes converges to the RBM in Section 4 under a heavy traffic condition when the FBFS or LBFS discipline is used. In fact, our method can readily be applied to any buffer priority disciplines for which a conventional heavy traffic limit theorem holds. Finally, it is desirable that more numerical studies, which represent all traffic intensities, be conducted to test the accuracy of our method, which is rooted in heavy traffic theory.

APPENDIX

PROOFS FOR THEOREMS 3.1 AND 3.2

In this appendix, we present proofs for Theorems 3.1 and 3.2. Recall that $\ell(i)$ is the lowest priority class at station i. We make the convention that stations are numbered such that $\ell(1) < \ell(2) < \cdots < \ell(d)$. Also, recall the definition of the reflection matrix $R = (I + G)^{-1}$ in (3.1). It follows from (3.5) that:

$$R = \operatorname{diag}(m_{\ell(1)}, \ldots, m_{\ell(d)}) A^{-1}, \tag{A.1}$$

where A is the $d \times d$ matrix:

$$A = \left(\sum_{k \in \mathscr{C}_i, k \geq \ell(j)} m_k \right). \tag{A.2}$$

Proof of Theorem 3.1. (a) Assume that the FBFS discipline is used. Then $\ell(i)$ is the last class that visits station i; hence, for $j > i$, $\Sigma_{k \in \mathscr{C}_i, k \geq \ell(j)} m_k = 0$. Therefore, A, and hence R, is a lower triangular matrix with positive diagonal entries. Thus, R is a completely $\mathscr{S}$ matrix.

(b) Assume that the LBFS discipline is used. It will be a consequence of the proof in part (b) of Theorem 3.2 that the determinant of A is positive. Let L be any subset of $\{1, \ldots, d\}$, and let $|L|$ be the cardinality of L. The $|L| \times |L|$ submatrix:

$$\left(\sum_{k \in \mathscr{C}_i, k \geq \ell(j)} m_k \right)_{i,j \in L},$$

is a *principal* submatrix of A. By the same consequence of the proof in part (b) of Theorem 3.2, the determinant of this submatrix is again positive. Therefore, A is a $\mathscr{P}$-matrix as defined in Berman and Plemmons (1979). It follows from Berman and Plemmons (1979) that A is invertible and A^{-1} is also a $\mathscr{P}$-matrix. The latter fact implies that A^{-1} is a completely $\mathscr{S}$ matrix. Therefore, $R = \operatorname{diag}(m_{\ell(1)}, \ldots, m_{\ell(d)}) A^{-1}$ is a completely $\mathscr{S}$ matrix. □

Before proving Theorem 3.2, let us state a lemma that was proved by Dupuis and Williams (1994, Theorem 2.6).

Lemma A.1. *Let R be a completely $\mathcal{S}$ matrix. A (θ, Σ, R)-RBM has a stationary distribution if every solution $z(\cdot)$ to the following deterministic Skorohod problem eventually reaches zero for any $z(0) \geqslant 0$.*

The Skorohod Problem:

$$z(t) = z(0) + \theta t + Ry(t), \tag{A.3}$$

$$z(t) \geqslant 0, \tag{A.4}$$

$$y(0) = 0 \text{ and each component of } y(\cdot) \text{ is nondecreasing,} \tag{A.5}$$

$$\int_0^\infty z_i(t)\, dy_i(t) = 0,\ i = 1, \ldots, d. \tag{A.6}$$

Proof of Theorem 3.2. (a) Assume that the FBFS discipline is used. In this case, R is a lower triangular matrix, as argued in part (a) of the proof for Theorem 3.1. Recall that $R = (I + G)^{-1}$ where G is a lower triangular matrix defined in (3.2). Let $z(\cdot)$ be any solution to the Skorohod problem with $z(0) \geqslant 0$. It follows from (A.3) and (4.2) that:

$$\begin{aligned}(I + G)z(t) &= R^{-1}z(t)\\ &= (I + G)z(0) + R^{-1}\theta t + y(t)\\ &= (I + G)z(0) - (e - \rho)t + y(t).\end{aligned} \tag{A.7}$$

Because G is a lower triangular matrix with positive diagonal entries, we have:

$$(1 + G_{11})z_1(t) = (1 + G_{11})z_1(0) - (1 - \rho_1)t + y_1(t).$$

It follows from Harrison (1985, Chapter 2) that $z_1(t) \equiv 0$ for $t \geqslant t_1 \equiv (1 + G_{11})z_1(0)/(1 - \rho_1)$. Now for $t \geqslant t_1$, from (A.7):

$$\begin{aligned}(1 + G_{22})z_2(t) = {}&(1 + G_{22})z_2(t_1) - (1 - \rho_2)(t - t_1)\\ &+ y_2(t) - y_2(t_1).\end{aligned}$$

Therefore, for $t \geqslant t_1 + t_2$, $z_2(t) = 0$, where $t_2 = (1 + G_{22})z_2(t_1)/(1 - \rho_2)$. By the same argument, one can show that for $i = 1, \ldots, d$, $z_i(t) \equiv 0$ for $t \geqslant t_1 + \cdots + t_i$, where $t_i = (1 + G_{ii})z_i(t_{i-1})/(1 - \rho_i)$ (assume that $t_0 = 0$). By Lemma A.1, any (θ, Σ, R)-RBM has a stationary distribution.

(b) Assume that the LBFS discipline is used. In light of (A.1), we first find the inverse of the $d \times d$ matrix A by using the traditional Gaussian elimination method. To begin with, we form a $2d \times d$ matrix on the left side of (A.8) and transform it into a matrix on the right side of (A.8) by performing column transformations:

$$\begin{pmatrix} A \\ I \end{pmatrix} \to \begin{pmatrix} I \\ A^{-1} \end{pmatrix}. \tag{A.8}$$

For $j = 2, \ldots, d$, subtracting column j from column $j - 1$, we obtain a matrix on the left side of (A.9):

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1d} \\ 0 & a_{22} & a_{23} & \cdots & a_{2d} \\ 0 & 0 & a_{33} & \cdots & a_{3d} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{dd} \\ 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \to \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{dd} \\ 1 & b_{12}^{(1)} & b_{13}^{(2)} & \cdots & b_{1d}^{(d-1)} \\ -1 & b_{22}^{(1)} & b_{23}^{(2)} & \cdots & b_{2d}^{(d-1)} \\ 0 & -1 & b_{33}^{(2)} & \cdots & b_{3d}^{(d-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{dd}^{(d-1)} \end{pmatrix}, \tag{A.9}$$

where $a_{ij} = \Sigma_{k \in \mathscr{C}_i, \ell(j) \leqslant k < \ell(j+1)}\, m_k$ with the convention $\ell(d + 1) = K + 1$. From this matrix, we have that the determinant of A is $a_{11}a_{22} \ldots a_{dd} > 0$ (this fact has been used in the proof of Theorem 3.1). Continuing the column operations on the matrix on the left side of (A.9), we obtain the matrix on the right of (A.9), where for $r = 1, \ldots, d - 1$:

$$b_{r+1,r+1}^{(r)} = 1 + \frac{a_{r,r+1}}{a_{rr}}, \tag{A.10}$$

$$b_{r+1,j}^{(r)} = \frac{a_{rj}}{a_{rr}}, \quad j = r + 2, \ldots, d, \tag{A.11}$$

$$b_{k,j}^{(r)} = b_{k,j}^{(r-1)} - \frac{a_{rj}}{a_{rr}} b_{kr}^{(r-1)}, \quad k = 1, \ldots, r, \quad j = r + 1, \ldots, d. \tag{A.12}$$

Therefore, $A^{-1} = [\mathrm{diag}(a_{11}, \ldots, a_{dd})]^{-1} B$, where:

$$B = \begin{pmatrix} 1 & b_{12}^{(1)} & b_{13}^{(2)} & \cdots & b_{1d}^{(d-1)} \\ -1 & b_{22}^{(1)} & b_{23}^{(2)} & \cdots & b_{2d}^{(d-1)} \\ 0 & -1 & b_{33}^{(2)} & \cdots & b_{3d}^{(d-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{dd}^{(d-1)} \end{pmatrix}.$$

Hence the reflection matrix R has the following representation:

$$R = \mathrm{diag}(m_{\ell(1)}/a_{11}, \ldots, m_{\ell(d)}/a_{dd})B.$$

Let $z(\cdot)$ be a solution to the Skorohod problem as defined in (A.3) through (A.6). Because $\theta = -R(e - \rho)$ as in (4.2), we have:

$$z(t) = z(0) + \theta t + Ry(t) = z(0) + Rx(t),$$

where:

$$x(t) = (\rho - e)t + y(t).$$

Let $w_i(t) = a_{ii}z_i(t)/m_{\ell(i)}$ and $w(t) = (w_1(t), \ldots, w_d(t))'$. It is easy to see that $w(\cdot)$ solves the Skorohod problem defined by:

$$w(t) = w(0) + B(\rho - e)t + By(t), \tag{A.13}$$

plus the analogous equations (A.4)–(A.6) with reflection matrix B. Let

$$g(t) = e'w(t) = w_1(t) + \ \ldots\ w_d(t).$$

It is a linear function of $w(t)$, and hence of $z(t)$. It is often called a linear Laypunov function. We are going to show that $g(\cdot)$ decreases to zero at a certain rate. A vector-valued function $x(\cdot)$ is said to be *regular* at time t if it is differentiable at t. We use $\dot{x}(t)$ to denote the derivative of $x(\cdot)$ at a regular point t. Let t be a regular point for $w(\cdot)$. Assume that $g(t) > 0$. There exists an r such that:

$$w_r(t) > 0,\ w_{r+1}(t) = 0, \ldots, w_d(t) = 0.$$

We would like to show that for $i = r, r + 1, \ldots, d$:

$$\dot{x}_i(t) = \left(\frac{1}{a_{ii}} \sum_{j=i}^{d} a_{ij}\right) \dot{x}_d(t). \tag{A.14}$$

Suppose that (A.14) holds for $i = r + 1, \ldots, d$, we show that (A.14) holds for $i = r$. To see this, notice that $w_{r+1}(t) = 0$ implies $\dot{w}_{r+1}(t) = 0$; hence, from (A.13), we have:

$$\begin{aligned}
\dot{x}_r(t) &= \sum_{j=r+1}^{d} b_{r+1,j}^{(j-1)} \dot{x}_j(t) \\
&= \left[\sum_{j=r+1}^{d} b_{r+1,j}^{(j-1)} \frac{1}{a_{jj}} \sum_{k=j}^{d} a_{jk}\right] \dot{x}_d(t) \\
&= \left[\frac{1}{a_{rr}} \sum_{k=r}^{d} a_{rk}\right] \dot{x}_d(t),
\end{aligned}$$

where the last equality follows from Lemma A.2 below. Therefore, (A.14) holds by an induction argument. It is clear from (A.10) and (A.11) that for $r = 1, \ldots, d$:

$$\sum_{j=r+1}^{d} b_{r+1,j}^{(r)} = \frac{1}{a_{rr}} \sum_{j=r}^{d} a_{rj}. \tag{A.15}$$

Thus, one can check that $e'B = (0, \ldots, 0, 1)$, and hence:

$$g(t) = x_d(t).$$

Because $w_r(t) > 0$, by (A.6), $\dot{y}_r(t) = 0$, and hence:

$$\dot{x}_r(t) = \rho_r - 1 < 0.$$

Therefore, we have:

$$\begin{aligned}
\dot{g}(t) = \dot{x}_d(t) &= \left[\frac{\sum_{k=r}^{d} a_{rk}}{a_{rr}}\right] \dot{x}_r(t) \\
&= \left[\frac{\sum_{k=r}^{d} a_{rk}}{a_{rr}}\right] (\rho_r - 1) \leqslant -\epsilon,
\end{aligned}$$

where:

$$\epsilon = \min_{1 \leqslant r \leqslant d} \left[\frac{\sum_{k=r}^{d} a_{rk}}{a_{rr}}\right] (1 - \rho_r) > 0.$$

It follows from Lemma 5.2 of Dai (1995) that $g(t) = 0$ for $t \geqslant g(0)/\epsilon$. □

Lemma A.2. *For each* $r = 1, \ldots, d,$

$$\sum_{i=r+1}^{d} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i} a_{ik}\right) = \frac{1}{a_{rr}} \sum_{i=r}^{d} a_{ri}.$$

Proof. We prove by induction that for each $j = d, \ldots, r$:

$$\sum_{i=r+1}^{d} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i} a_{ik}\right) = \sum_{i=r+1}^{j} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i} a_{ik}\right) + \sum_{k=j+1}^{d} b_{r+1,k}^{(j)}. \tag{A.16}$$

(The empty summation is interpreted as zero.) If (A.16) holds, the lemma is proved by taking $j = r$. It is obvious that (A.16) holds for $j = d$. Suppose that (A.16) holds for j, we would like to show that it holds for $j - 1$ as well. Now:

$$\begin{aligned}
&\sum_{i=r+1}^{d} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i}^{d} a_{ik}\right) \\
&= \sum_{i=r+1}^{j} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i}^{d} a_{ik}\right) + \sum_{k=j+1}^{d} b_{r+1,k}^{(j)} \\
&= \sum_{i=r+1}^{j-1} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i}^{d} a_{ik}\right) + b_{r+1,j}^{(j-1)} \left(\frac{1}{a_{jj}} \sum_{k=j}^{d} a_{jk}\right) \\
&\quad + \sum_{k=j+1}^{d} b_{r+1,k}^{(j)} \\
&= \sum_{i=r+1}^{j-1} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i}^{d} a_{ik}\right) + b_{r+1,j}^{(j-1)} \left(\frac{1}{a_{jj}} \sum_{k=j}^{d} a_{jk}\right) \\
&\quad + \sum_{k=j+1}^{d} \left[b_{r+1,k}^{(j-1)} - \frac{a_{jk}}{a_{jj}} b_{r+1,j}^{(j-1)}\right] \\
&= \sum_{i=r+1}^{j-1} b_{r+1,i}^{(i-1)} \left(\frac{1}{a_{ii}} \sum_{k=i}^{d} a_{ik}\right) + \sum_{k=j}^{d} b_{r+1,k}^{(j)}.
\end{aligned}$$

By induction (A.16) holds, and therefore the lemma is proved. □

ACKNOWLEDGMENTS

Mike Harrison pointed out that the Brownian model we derived in an earlier version of this paper was identical to the one in Harrison and Williams (1992). Following his insight we have been able to prove Theorems 3.1 and 3.2, which were stated as conjectures. We would also like to thank Mike and an anonymous referee for suggestions that have substantially improved the exposition of the paper. This research was in part supported by National Science Foundation grants DMS-9203524, DDM-9215233, DMI-9457336, and a grant from Texas Instruments Corporation.

REFERENCES

Baskett, F., K. M. Chandy, R. R. Muntz, and F. G. Palacios. 1975. Open, Closed and Mixed Networks of Queues with Different Classes of Customers. *J. ACM.* **22,** 248–260.

Berman, A. and R. J. Plemmons. 1979. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York.

Bitran, G. R. and D. Tirupati. 1988. Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference. *Mgmt. Sci.* **34,** 75–100.

Bramson, M. 1994. Instability of FIFO Queueing Networks. *Anns. Appl. Prob.* **4,** 414–431.

Chen, H. Personal Communications. 1994.

Chen, H. and A. Mandelbaum. 1991. Stochastic Discrete Flow Networks: Diffusion Approximation and Bottlenecks. *Anns. Appl. Prob.* **19,** 1463–1519.

Chen, H. and A. Mandelbaum. 1994. Hierarchical Modeling of Stochastic Networks II: Strong Approximations. In *Probability Models in Manufacturing Systems*, D. D. Yao (ed.), 107–132. Springer, New York.

Coffman, E. G., Jr., A. A. Puhalskii, and M. I. Reiman. 1995. Polling Systems with Zero Switchover Times: a Heavy Traffic Averaging Principle. *Anns. Appl. Prob.* **5,** 681–719.

Coffman, E. G., Jr., A. A. Puhalskii, and M. I. Reiman. 1996. Polling Systems in Heavy Traffic: a Bessel Process Limit. Preprint.

Dai, J. G. 1995. On Positive Harris Recurrence of Multiclass Queueing Networks: a Unified Approach via Fluid Limit Models. *Anns. Appl. Prob.* **5,** 49–77.

Dai, J. G. and J. M. Harrison. 1992. Reflected Brownian Motion in an Orthant: Numerical Methods for Steady-State Analysis. *Anns. Appl. Prob.* **2,** 65–86.

Dai, J. G. and T. G. Kurtz. 1994. Characterization of the Stationary Distribution for a Semimartingale Reflecting Brownian Motion in a Convex Polyhedron. Preprint.

Dai, J. G. and T. G. Kurtz. 1995. A Multiclass Station with Markovian Feedback in Heavy Traffic. *Math. O. R.* **20,** 721–742.

Dai, J. G. and S. P. Meyn. 1995. Stability and Convergence of Moments for Multiclass Queueing Networks via Fluid Limit Models. *IEEE Trans.* **40,** 1889–1904.

Dai, J. G. and G. Weiss. 1996. Stability and Instability of Fluid Models for Certain Re-entrant Lines. *Math. O. R.* **21,** 115–134.

Dupuis, P. and R. J. Williams. 1994. Lyapunov Functions for Semimartingale Reflecting Brownian Motions. *Anns. Appl. Prob.* **22,** 680–702.

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

Gross, D. and C. M. Harris. 1985. *Fundamentals of Queueing Theory*. Wiley, New York.

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

Harrison, J. M. and V. Nguyen. 1990. The QNET Method for Two-Moment Analysis of Open Queueing Networks. *Queueing Systems: Theory and Applications,* **6,** 1–32.

Harrison, J. M. and V. Nguyen. 1993. Brownian Models of Multiclass Queueing Networks: Current Status and Open Problems. *Queueing Systems: Theory and Applications,* **13,** 5–40.

Harrison, J. M. and M. I. Reiman. 1981. Reflected Brownian Motion on an Orthant. *Anns. Appl. Prob.* **9,** 302–308.

Harrison, J. M. and R. J. Williams. 1987. Brownian Models of Open Queueing Networks with Homogeneous Customer Populations. *Stochastics*, **22,** 77–115.

Harrison, J. M. and R. J. Williams. 1992. Brownian Models of Feedforward Queueing Networks: Quasireversibility and Product Form Solutions. *Anns. Appl. Prob.* **2,** 263–293.

Jackson, J. R. 1957. Networks of Waiting Lines. *Opns. Res.* **5,** 518–521.

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

Kelly, F. P. 1975. Networks of Queues with Customers of Different Types. *J. Appl. Prob.* **12,** 542–554.

Kumar, P. R. 1993. Re-entrant Lines. *Queueing Systems: Theory and Applications.* **13,** 87–110.

Kumar, S. and P. R. Kumar. 1996. Fluctuation Smoothing Policies Are Stable for Stochastic Reentrant Lines. *Discrete Event Dynamic Systems,* **6,** 361–370.

Kumar, P. R. and T. I. Seidman. 1990. Dynamic Instabilities and Stabilization Methods in Distributed Real-Time Scheduling of Manufacturing Systems. *IEEE Trans. Automat. Control.* **AC-35,** 289–298.

Lu, S. H. and P. R. Kumar. 1991. Distributed Scheduling Based on Due Dates and Buffer Priorities. *IEEE Trans. Automat. Control.* **36,** 1406–1416.

Peterson, W. P. 1991. A Heavy Traffic Limit Theorem for Networks of Queues with Multiple Customer Types. *Math. O. R.* **16,** 90–118.

Reiman, M. I. 1984. Open Queueing Networks in Heavy Traffic. *Math. O. R.* **9,** 441–458.

Reiman, M. I. 1984b. Some Diffusion Approximations with State Space Collapse. In *Modeling and Performance Evaluation Methodology*. F. Baccelli, and G. Fayolle (eds.), 209–240, Berlin, Springer.

Reiman, M. I. 1988. A Multiclass Feedback Queue in Heavy Traffic. *Adv. in Appl. Probab.* **20,** 179–207.

Reiman, M. I. and Wein, L. 1996. Heavy Traffic Analysis of Polling Systems in Tandem. Preprint.

Reiman, M. I. and R. J. Williams. 1988 and 1989. A Boundary Property of Semimartingale Reflecting Brownian Motions. *Probability Theory and Related Fields,* **77,** 87–97, and **80,** 633.

Rybko, A. N. and A. L. Stolyar. 1992. Ergodicity of Stochastic Processes Describing the Operation of Open Queueing Networks. *Problems of Inform. Transmission,* **28,** 199–220.

Segal, M. and W. Whitt. 1989. A Queueing Network Analyzer for Manufacturing. In Bonatti, M. (ed.). *TELETRAFFIC SCIENCE for New Cost Effective Systems, Networks and Services*, ITC-12, Elsevier, North Holland, 1146–1152.

Seidman, T. I. 1994. First come, first served Can be Unstable! *IEEE Trans. Automat. Control*, **39,** 2166–2171.

Shalev-Oren, S., A. Seidman, and P. J. Schweitzer. 1985. Analysis of Flexible Manufacturing Systems with Priority Scheduling: PMVA. *Anns. Opns. Res.* **3,** 115–139.

Taylor, L. M. and R. J. Williams. 1993. Existence and Uniqueness of Semimartingale Reflecting Brownian Motions in an Orthant. *Probability Theory and Related Fields,* **96,** 283–317.

Whitt, W. 1983. The Queueing Network Analyzer. *Bell Sys. Tech. J.* **62,** 2779–2815.