Brownian Models of Closed Queueing Networks: Explicit Solutions for Balanced Three-Station Systems

Elizabeth Schwerer, Jan A. Van Mieghem

The Annals of Applied Probability, Volume 4, Issue 2 (May, 1994), 448-477.

The Annals of Applied Probability
1994, Vol. 4, No. 2, 448–477

BROWNIAN MODELS OF CLOSED QUEUEING NETWORKS: EXPLICIT SOLUTIONS FOR BALANCED THREE-STATION SYSTEMS

BY ELIZABETH SCHWERER AND JAN A. VAN MIEGHEM

Stanford University

We study a closed, three-station queueing network with general service time distributions and balanced workloads (that is, each station has the same relative traffic intensity). If the customer population is large, then the queue length process of such a network can be approximated by driftless reflected Brownian motion (RBM) in a simplex. Building on earlier work by Harrison, Landau and Shepp, we develop explicit formulas for various quantities associated with the stationary distribution of RBM in a general triangle and use them to derive approximate performance measures for the closed queueing network. In particular, we develop approximations for the throughput rate and for moments and tail fractiles of the throughput time distribution. Also, crude bounds on the throughput rate and mean throughput time are proposed. Finally, we present three examples that test the accuracy of both the Brownian approximation and our performance estimates.

1. Introduction. Recently Dai and Harrison [2] proposed a general scheme for approximating multiclass closed queueing networks by what they call Brownian system models. Their scheme approximates the queue length process of a J-station closed network by a regulated or reflected Brownian motion (RBM) whose state space is a J-dimensional simplex. The drift vector, covariance matrix and boundary data of the approximating RBM are determined by the routing data, service discipline and first and second moments of the service time distributions of the original queueing network. Dai and Harrison conjecture that their approximation can be rigorously justified by a heavy traffic limit theorem in which the number of customers in the closed network becomes large. Their method culminates in the numerical solution of a partial differential equation that determines the stationary distribution of the approximating RBM.

In this paper we consider the very special case of a three-station closed network in which all stations have the same relative traffic intensity. We refer to such a system as a *balanced* three-station closed network. For such networks the approximating RBM proposed by Dai and Harrison is driftless, and its state space is the unit simplex in $\mathbb{R}^3$. Because the RBM is

Received May 1993; revised September 1993.

AMS 1991 *subject classifications.* 60K25, 68M20, 90B22.

Key words and phrases. Closed queueing networks, heavy traffic analysis, reflected Brownian motion, sojourn time analysis, performance analysis.

driftless and its state space is effectively two-dimensional, we can apply the results of Harrison, Landau and Shepp (HLS) [7] to derive a formula for its stationary distribution. Actually, the HLS formula is not entirely explicit, but from it we will derive explicit formulas for the throughput rate and for moments and tail fractiles of the throughput time distribution associated with the Brownian system model.

Few other families of Brownian approximations of queueing models have proved amenable to exact analysis. By offering a comprehensive analysis that derives explicit formulas for network performance measures, we hope to provide both insight into the mathematical content of Brownian system theory and a benchmark for testing numerical methods. In particular, our formulas can be used to evaluate the numerical methods developed by Dai and Harrison [2] as part of their QNET method for the analysis of closed networks. In addition, they offer qualitative insights into the stationary density that cannot be obtained numerically and that might illuminate studies of more general networks. Finally, the systems we analyze serve in their own right as Brownian approximations for certain special classes of manufacturing and computer systems.

A closed queueing network is one in which the number of jobs remains constant: This is often interpreted to mean that the completion of one job triggers the start of a new one. Solberg [23] first proposed the use of closed queueing networks to model machine shops that are physically restricted to a fixed number of jobs by the number of pallets available for transporting work among stations. Spearman, Woodruff and Hopp [24] discuss the advantages of pull-based production control systems—such as kanban and their own CONWIP (CONstant Work In Process) system—over push systems such as Materials Requirements Planning. They argue that push systems can be modeled as open queueing networks where throughput is scheduled and WIP is determined endogenously. Pull systems, however, are best modeled as closed queueing networks where WIP is fixed and throughput is determined endogenously. Dai and Harrison [2] discuss the general applicability of closed queueing systems as models of manufacturing operations. Sauer and Chandy [22] suggest modeling time-shared computer systems as closed queueing networks. Even though the number of customers using a time-shared system will change over time, they argue, these changes will be slow enough that the system can be modeled as a sequence of closed systems in equilibrium. Lazowska, Zahorjan, Graham and Sevcik [13] claim that closed queueing networks make good models of the memory subsystems of computers with memory constraints. If jobs impose fairly uniform demands on the memory system, then it is reasonable to model a memory constraint as a constraint on the number of jobs in the system.

In order to preserve analytic tractability, we restrict ourselves to balanced three-station networks in heavy traffic. In other words, we assume that each of the three stations in the network receives on average the same amount of work (measured in time content) and that this amount is close to the station's full capacity (i.e., the utilization of each station approaches 1). Although manufacturing operations often try to balance capacity, this is nonetheless a very restrictive assumption. We allow our networks to be multiclass—that is, we allow transitions among stations to depend on jobs' past routes—but following Dai and Harrison, we require that all classes served at a station share a common service distribution. This restriction ensures that the Brownian model is well posed.

The heavy traffic theory justifying Brownian approximations to queueing networks is still incomplete. Reiman [20] proves that the heavy traffic limit of an open, generalized Jackson network is RBM in an orthant, and Chen and Mandelbaum [1] prove an analogous limit theorem for closed networks of the generalized Jackson type, obtaining RBM in a simplex as their heavy traffic limit. Their work has not yet been fully extended to multiclass networks. Peterson [18] proves a heavy traffic limit theorem for open feedforward networks, and Dai and Nguyen [3] show for general open networks that if a heavy traffic limit exists, then it is reflected Brownian motion in an orthant. However, Dai and Wang [4] and Whitt [27] have found examples of multiclass networks with feedback for which the approximating Brownian model is not well posed. It is to avoid such problems that we restrict ourselves to a single service distribution for all classes at a station. For such networks, Dai and Harrison show that the Brownian approximation is well posed, and they make an informal argument that it is the correct two-moment approximation to use, even though it has yet to be justified by a heavy traffic limit theorem.

To our knowledge, this paper is the first to "solve" a Brownian model of a non-product-form closed queueing network. The first Brownian models to be analyzed were single-station systems with and without buffer limitations; see Harrison [6]. In his work on open generalized Jackson networks, Reiman [21] observes that certain of those networks have product-form stationary distributions which can be written out explicitly. Harrison and Williams [11] spell out exactly which open networks have product form solutions, and Harrison, Williams and Chen [12] provide the analogous condition for closed networks of the generalized Jackson type. Among networks that do not satisfy the product-form condition, Harrison [5] computes the stationary distribution for the Brownian model of a certain special tandem queue, and Harrison and Shepp [10] do the same for a balanced tandem queue with one finite storage buffer. Trefethen and Williams [25] study the same balanced tandem queue assuming *two* finite buffers, using the same method of HLS that we apply in this paper. The balanced tandem queue with two finite buffers leads to a Brownian system model with a rectangular state space, which is more complicated than the triangular case treated here. To handle the rectangular case, Trefethen and Williams must first specify a Schwarz–Christoffel transform appearing in the HLS formula, and they do this numerically, using a software package called SCPACK. Specification of the Schwarz–Christoffel transform is trivial in our triangular case, so we do not need the sophisticated numerical methods embodied in SCPACK.

The paper is organized as follows. In Section 2, we describe the network model and its Brownian approximation. Our notation and development follow

Dai and Harrison [2], who in turn follow Harrison and Nguyen [8]. The queueing network model differs from that of Dai and Harrison in two ways: The mean service times τ_i at the three stations are constrained by the restriction to balanced networks, and we consider only Markovian routing among classes. In Section 3, we develop explicit formulas for some steady state quantities associated with the Brownian system model and discuss the functional form of the stationary density. These formulas are used in Section 4 to derive performance measures for the original queueing network, and in Section 5 we calculate the performance measures explicitly for some sample networks.

2. A balanced three-station closed network model.

Given a balanced closed network with three stations and n customers, we are interested in the evolution of its three-dimensional state vector $N(t)$, which records the number of customers at each station. We will approximate this state vector by a reflected Brownian motion in the three-dimensional simplex $S = \{N \geq 0: e'N = n\}$, where e is a 3-vector of ones. This section describes how the network primitives determine the approximating RBM.

We specialize Dai and Harrison's model to $J = 3$ single-server stations. The network has K customer classes, each class k being served at a unique station $s(k)$. The many-to-one relationship between customer classes and servers is recorded in the $3 \times K$ constituency matrix C: namely, $C_{jk} = 1$ if $s(k) = j$. Customers change class in a Markovian fashion: After completing service at station $s(k)$, a customer of class k turns into a customer of class l with probability P_{kl}, independent of all previous history. This setup allows a customer's future route to depend not only on his current location in the network, but also on his past processing history, insofar as past history can be captured in the customer's class designation. For example, the deterministic routing pictured in Figure 1 can be captured by defining six customer classes. We assume that the stochastic matrix $P = (P_{kl})$ is irreducible. Dai and Harrison consider a somewhat more general class of closed networks,

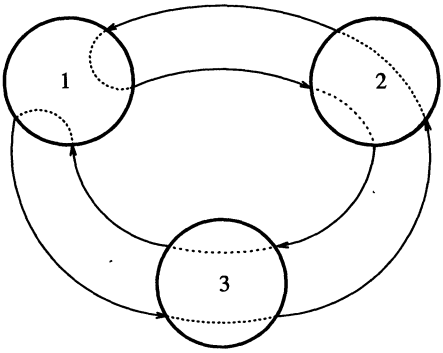

FIG. 1. *A closed three-station network with deterministic routing.*

arguing that such generality may be desirable when there are natural "input" classes. Although the analysis in Sections 3–5 extends to the case covered by Dai and Harrison (indeed, one of the examples of Section 5 exploits their greater generality), we develop the RBM model for this simpler case in order to keep the exposition uncluttered.

To conform to the setup of Dai and Harrison, we artificially divide the routing history of each customer into a succession of "cycles," each of which begins with an entry into class 1. That is, each customer cycle begins with a visit to class 1, and the transition structure within a cycle is given by the $K \times K$ substochastic matrix $\tilde{P}$, where $\tilde{P}_{kl} = P_{kl}$ if $l > 1$ and $\tilde{P}_{k1} = 0$. Let λ_k be the expected number of visits that a customer makes to class k during one cycle. Then the row vector $\lambda' = (\lambda_1, \dots, \lambda_K)$ satisfies $\lambda' P = \lambda'$ with the auxiliary condition that $\lambda_1 = 1$. In other words, λ is the stationary distribution of the original transition matrix P, except for a rescaling. For future reference, let $Q = (I - \tilde{P}')^{-1}$. Because P is an irreducible transition matrix, $\tilde{P}$ has spectral radius less than 1, the fundamental matrix Q is well defined and λ is uniquely determined by the conditions previously given.

To complete a two-moment model of the network, we assume that each service station $i = 1, 2, 3$ has an associated service time distribution with mean τ_i and squared coefficient of variation (SCV) b_i^2. Let $\gamma = C\lambda$ so that γ_i represents the expected number of visits to station i that a customer makes in a cycle, and define $\rho_i = \gamma_i \tau_i$ for each $i = 1, 2, 3$. We call ρ_i the *relative traffic intensity*, or relative utilization, for station i. It represents the average amount of work that a customer delivers to station i during one cycle. A balanced network is one in which $\rho_1 = \rho_2 = \rho_3$. Assuming hereafter that the network is balanced, we can choose our unit of time so that $\rho_i = 1$ for all i, or equivalently

$$\tau_i = 1/\gamma_i \quad \text{for } i = 1, 2, 3. \tag{1}$$

A central objective of our analysis is to determine the throughput rate a, defined as the long-run average number of customer cycles completed (or begun) per unit time. The choice of time unit implicit in (1) allows a to be interpreted alternatively as the long-run utilization rate at any of the network's three stations. Following the development of Dai and Harrison [2], a initially will be treated as if it were an input to the model, as it would be in the case of an open network, but eventually it will be determined through a "closure approximation."

To describe the approximating Brownian system model, we will need the following notation: let $B = \operatorname{diag}(b_1^2, b_2^2, b_3^2)$, $T = \operatorname{diag}(\tau_{s(1)}, \dots, \tau_{s(K)})$, $\Lambda = \operatorname{diag}(\lambda)$,

$$\hat{R} = \operatorname{diag}(\gamma)(CTQ\Lambda C')^{-1}, \tag{2}$$

$$R = \hat{R} - \hat{R}ee'\hat{R}/(e'\hat{R}e), \tag{3}$$

$$\Gamma = CTQ\left[\operatorname{diag}(\tilde{P}'\lambda) - \tilde{P}'\Lambda\tilde{P}\right](CTQ)' + CTB\Lambda(CT)', \tag{4}$$

$$\hat{\Omega} = R\Gamma R'. \tag{5}$$

In Dai and Harrison's model, $\hat{R}$ is a "reflection matrix" that relates immediate workload to a long-run ("total workload") process that in turn can be shown to approach Brownian motion in the heavy traffic limit. The matrix R is the projection of $\hat{R}$ into the $(J-1)$-dimensional space where the process lives. The matrices Γ and $\hat{\Omega}$ are covariance matrices derived from the primitive routing and service processes. Thus they are positive definite. Given a value for $a > 0$, Dai and Harrison show that there exists a pair of three-dimensional processes N and I, unique in distribution, satisfying the following five conditions:

$$N(t) = N(0) + X(t) + RI(t), \tag{6}$$

where

$$X(t) \text{ is a Brownian motion with zero drift and covariance matrix } \Omega = a\hat{\Omega}, \tag{7}$$

$$N(t) \in S, \tag{8}$$

$$I(\cdot) \text{ is continuous and increasing with } I(0) = 0, \tag{9}$$

$$I_j(\cdot) \text{ increases only at times } t \text{ when } N_j(t) = 0. \tag{10}$$

In the Brownian system model, the increasing process $I_j(\cdot)$ represents cumulative server idleness at station j, and $N(\cdot)$ is the queue length process described earlier. Conditions (6)–(10) identify N as an RBM with state space S, zero drift, covariance matrix Ω and reflection matrix R. From (6) we see that an increase in $I_j(\cdot)$ displaces $N(\cdot)$ in direction R^j (the jth column of R), and (10) says that $I_j(\cdot)$ only increases when $N_j(\cdot) = 0$. Given these two facts, R^j is commonly described as the "direction of reflection" associated with the boundary surface $\{N_j = 0\}$, but Harrison [6] has suggested the alternative term "direction of control." Given the stationary density function p of this RBM, we can calculate the throughput rate a and other performance measures of interest. Hereafter N is described as driftless RBM(Ω, R) in the simplex S.

3. Driftless RBM in a triangle. In this section we study the stationary behavior of the Brownian system model described by (6)–(10) and present explicit formulas that will be used in Section 4 to compute various performance measures of the network. Although this derivation focuses on the simplex S, the results are valid for driftless RBM in a *general triangular* state space.

As a preliminary, define the differential operators

$$\mathscr{G} = \frac{1}{2}\nabla'\Omega\nabla = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\Omega_{ij}\frac{\partial^2}{\partial x_i\,\partial x_j} \tag{11}$$

and

$$\mathscr{D}_j = (R^j)'\nabla = \sum_{i=1}^{3} R_{ij}\frac{\partial}{\partial x_i}. \tag{12}$$

Because Ω is positive definite, $\mathscr{G}$ is an elliptic operator associated with the underlying Brownian motion X in (7). The operator $\mathscr{D}_j$ is proportional to the directional derivative in the "direction of reflection" associated with boundary surface

$$(13) \qquad F_j = \{x \in S \colon x_j = 0\}, \qquad j = 1,2,3.$$

Building on earlier work on RBM, Dai and Harrison [2] establish that there exist "associated boundary densities p_j ($j = 1,2,3$)" on F_j such that for each bounded Borel measurable function f on F_j,

$$(14) \qquad \lim_{t \to \infty} \frac{1}{t} E\left[\int_0^t f(N(s))\, dI_j(s)\right] = \int_{F_j} f \cdot p_j \, d\sigma_j,$$

where $d\sigma_j$ is surface Lebesgue measure on F_j. Furthermore, the stationary interior and boundary density functions (p, p_1, p_2, p_3) satisfy the following "basic adjoint relationship":

$$(15) \qquad \int_S \mathscr{G} f \cdot p \, dx + \sum_j \int_{F_j} \mathscr{D}_j f \cdot p_j \, d\sigma_j = 0 \quad \text{for all } f \in C^2(S),$$

where dx is Lebesgue measure on S and $C^2(S)$ is the set of functions that, together with their first and second order derivatives, are continuous and bounded on S.

Using (14) and (15), we can express many interesting functionals of N in terms of the moments of the boundary densities p_j. For example, the expected value of $h(N)$ for a particular function $h\colon S \to \mathbb{R}$ can be calculated using the boundary densities as follows. Find a function f such that $\mathscr{G} f = h$ on S and use (15) to obtain

$$(16) \qquad E(h(N)) = -\sum_j \int_{F_j} \mathscr{D}_j f \cdot p_j \, d\sigma_j \quad \text{where } \mathscr{G} f = h \text{ on } S.$$

We will use this relationship in Sections 4 and 5 to calculate the moments of the throughput time. Also, setting $f = 1$ in (14) yields an expression for δ_j, the *long-run average idleness rate for station j*:

$$(17) \qquad \delta_j \equiv \lim_{t \to \infty} \frac{1}{t} E[I_j(t)] = \int_{F_j} p_j \, d\sigma_j.$$

The idleness rate δ_j corresponds to the amount of control exerted at boundary face F_j, i.e., to the local time spent on F_j. In light of the pivotal role of the line integrals of the boundary densities, we focus on computing the moments of the boundary densities in closed form.

3.1. *Moments of the boundary density.* To facilitate the exposition we introduce the functions I_s:

$$(18) \quad I_s(a,b,c,d) = \frac{1}{B(a,b)} \int_0^1 \left(\frac{B_x(a,b)}{B(a,b)} \right)^s x^{a+c-1} (1-x)^{b+d-1} \, dx,$$

for positive $\mathrm{Re}(a)$, $\mathrm{Re}(b)$, $\mathrm{Re}((s+1)a+c)$ and $\mathrm{Re}((s+1)b+d)$, where $B(a,b)$ and $B_x(a,b)$ denote the beta and incomplete beta functions of a and b. Denoting the two-dimensional restrictions

$$(19) \qquad \tilde{\Omega} = (\Omega_{ij})_{1\le i,j\le 2} \quad \text{and} \quad \tilde{R} = (R_{ij})_{1\le i\le 2,\, 1\le j\le 3},$$

the moments of the boundary density can be expressed in terms of the functions I_s as follows.

PROPOSITION 1. *The moments of the boundary density of driftless $RBM(\Omega, R)$ in the simplex $S = \{N \ge 0\colon e'N = n\}$ are*

$$(20) \qquad \int_{F_i} \sigma_i^k p_i \, d\sigma_i = cn^{k-1} \frac{\Omega_{ii}}{R_{ii}} I_k(F_i) \cos \theta_i,$$

where $I_k(F_i)$ is an abbreviation for $I_k(\xi_{i+1}, \xi_{i+2}, \alpha_{i+1}, \alpha_{i+2})$,

$$(21) \qquad \sin \pi\xi_i = \sqrt{\frac{|\tilde{\Omega}|}{\Omega_{i+1,i+1}\Omega_{i+2,i+2}}},$$

$$(22) \qquad \sin \theta_i = \frac{(R^i \times \Omega^i)_1}{\sqrt{\Omega_{ii} |\tilde{\Omega}| (\tilde{R}^i)' \tilde{\Omega}^{-1} \tilde{R}^i}},$$

$$(23) \qquad \pi\alpha_i = \theta_{i+2} - \theta_{i+1}$$

and

$$(24) \quad c = -\frac{1}{2}\left(\frac{R_{21}\Omega_{11}}{R_{11}\Omega_{22}} (I_0(F_1) - I_1(F_1)) \cos \theta_1 + \frac{R_{23}\Omega_{33}}{R_{33}\Omega_{22}} I_1(F_3) \cos \theta_3 \right)^{-1}.$$

Since all components of the vector product $R^i \times \Omega^i$ are identical, any of them could be used in (22) to define θ_i.

3.2. *Proof of Proposition* 1. Because $e'N = n$, it suffices to determine the density of the two-dimensional projection (denoted by tildes) $\tilde{N} = (N_1, N_2)$ in the solid simplex $\tilde{S} = \{\tilde{N} \ge 0\colon e'\tilde{N} \le n\}$. The process $\tilde{N}$ is driftless RBM with covariance matrix $\tilde{\Omega}$ and reflection matrix $\tilde{R}$ defined in (19). First we will transform $\tilde{N}$ into standard RBM. Then we can apply the work of HLS to calculate the density and its boundary moments. Finally, we transform the results back into original quantities.

3.2.1. *Transformation of $RBM(\tilde{\Omega}, \tilde{R})$ into standard $RBM(I, R^*)$.* Applying a nonsingular linear transform to the $(\tilde{\Omega}, \tilde{R})$ RBM $\tilde{N}$ in the solid simplex $\tilde{S}$ yields the RBM N^*, which has a general triangular state space S^*. If N^* is to be standard RBM, that is, if $\Omega^* = I$, then the linear transform must be a square root of $\tilde{\Omega}^{-1}$. This condition determines the transform up to a rotation. From $N^* = \tilde{\Omega}^{-1/2}\tilde{X} + \tilde{\Omega}^{-1/2}\tilde{R}\tilde{Y}$, it follows that the transformed reflection matrix is $R^* = \tilde{\Omega}^{-1/2}\tilde{R}$ and that the idleness process Y remains unchanged. Denote the length of the jth side (or face F_j^*) of the triangle S^* by $\|F_j^*\|$. It is readily verified that

$$\|F_j^*\| = n\sqrt{\frac{\Omega_{jj}}{|\tilde{\Omega}|}}\,. \tag{25}$$

Let G^* denote the triangle S^* without its three vertices (the unsmooth part of ∂S^*). Harrison, Williams and Chen ([12], Section 8, Lemma 8) establish that if the stationary distribution $p^* \in C^2(G^*)$, the associated boundary density p_j $(j = 1, 2, 3)$ is a scaled restriction of the density p to the face F_j:

$$p_j = \frac{\Omega_{jj}}{2R_{jj}}p \quad \text{on } F_j,\ j = 1, 2, 3. \tag{26}$$

We will show in Section 3.2.2 that p^* is harmonic on G^* and therefore in $C^2(G^*)$. Using these relations, the basic adjoint relationship in the transformed domain becomes

$$\int_{S^*} \Delta f^* \cdot p^*\, dz^* + \sum_j \frac{\Omega_{jj}^{1/2}}{R_{jj}} \int_{F_j^*} (R^{*j})'\, \nabla f^* \cdot p^*\, d\sigma_j^* = 0 \tag{27}$$

$$\text{for all } f^* \in C^2(S^*).$$

Also, the moments of the boundary density are transformed into

$$\int_{F_i} \sigma_i^k p_i\, d\sigma_i = \frac{|\tilde{\Omega}|^{(1+k)/2}\Omega_{ii}^{(1-k)/2}}{2R_{ii}} \int_{F_i^*} (\sigma_i^*)^k p^*\, d\sigma_i^*. \tag{28}$$

3.2.2. *Explicit solution of standard RBM(I, R^*).* HLS compute the stationary distribution of standard RBM in a general polygonal state space S with angles of reflection at the faces. Figure 2 shows the transformed triangle S^* and the relevant quantities: v_k and $\pi\xi_k$ $(k = 1, 2, 3)$ represent the vertex

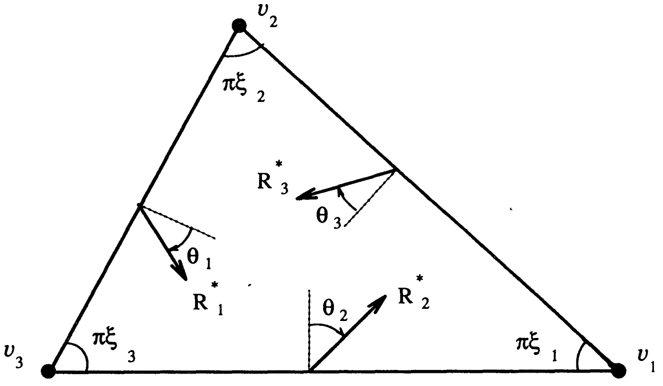

FIG. 2. *The transformed state space S^*.*

and the interior angle opposite face F_k^*, and θ_k is the angle between the reflection vector R_k^* and the inward normal to face F_k^*, measured clockwise. It will be convenient to set $\theta_0 = \theta_3$ and $\xi_4 = \xi_1$. We change the labeling convention of [7] slightly (θ_k in this paper corresponds to θ_{k+1} in [7]) in order to highlight some of the symmetries of the triangular case.

HLS calculate the stationary distribution $p^*(z)$ by mapping the general polygonal state space S onto the upper half-plane by means of a conformal mapping g such that $g(\sigma_0) = \infty$ for some nonsingular point $\sigma_0 \in \partial S$. When S is a polygon, g is the inverse of the Schwarz–Christoffel map. For the triangular state space S^*, they derive the stationary distribution

$$(29) \qquad p^*(z) = c^* \operatorname{Re}\left\{\exp(-i\theta_3)\prod_{k=1}^{3}\left[g(z) - g(v_k)\right]^{\alpha_k}\right\}, \qquad z \in S^*,$$

where $\pi\alpha_k = \theta_{k+2} - \theta_{k+1}$ and c^* is a normalization constant ensuring that p^* integrates to 1. For a general polygonal state space the values of the "prevertices" $g(v_k)$ are not known a priori and need to be computed numerically. Trefethen and Williams [25] discuss the numerical computation of $p^*(z)$ for this general case. The Riemann mapping theorem ([15], page 174) shows that one can choose three prevertices. Therefore, for the triangular domain, the determination of the values $g(v_k)$ is trivial and it is possible to derive analytic formulas for the moments of the boundary density.

A simpler form of the stationary density $p^*(z)$ can be derived by mapping one vertex, say v_l, to infinity. The three possible choices for v_l provide three expressions for p^* as a product of two terms. Then Theorem 3.13 in reference 7 becomes (with the HLS labeling of reflection angles) the following theorem.

THEOREM 1. *The stationary density is*

$$(30) \qquad p^*(z) = c^* \operatorname{Re}\left\{\exp(i\theta_{l-1})\prod_{\substack{k=1\\k\neq l}}^{3}\left[g(z) - g(v_k)\right]^{\alpha_k}\right\},$$

where $z \in S^*$, $l = 1, 2, 3$, and g is the inverse Schwarz–Christoffel mapping with $g(v_l) = \infty$.

The proof in [7] extends easily to this form [by redefining the HLS function L as $L = i\theta_{l-1} + \log(\prod_{k=1,\,k\neq l}^{K}[g(z) - g(v_k)]^{\alpha_k})$]. This simplified expression also holds for a general polygonal state space and is in agreement with the results of Trefethen and Williams [25].

This simplification of p^* plays a crucial role in obtaining simple symmetric results for the triangular case. The Riemann mapping theorem allows us to choose g such that $g(v_l) = \infty$, $g(v_{l+1}) = 0$, and $g(v_{l+2}) = 1$. Then the density becomes

$$(31) \qquad p^*(z) = c^* \operatorname{Re}\left\{e^{i\theta_l} g(z)^{\alpha_{l+1}}\left[1 - g(z)\right]^{\alpha_{l+2}}\right\},$$

where $z \in S^*$, $l = 1, 2, 3$, the inverse mapping $g^{-1}(w)$ is given by

$$\begin{aligned} g^{-1}(w) &= \frac{\|F_l^*\|}{B(\xi_{l+1}, \xi_{l+2})} \int_0^w t^{\xi_{l+1}-1}(1-t)^{\xi_{l+2}-1}\, dt \\ &= \frac{\|F_l^*\|}{B(\xi_{l+1}, \xi_{l+2})} B_w(\xi_{l+1}, \xi_{l+2}). \end{aligned} \tag{32}$$

This gives us three equivalent expressions for the density p^*, each involving a different Schwarz–Christoffel map g^{-1}. When $l = k$ we say that p^* is expressed with reference to face F_k^*. This will be the natural representation of p^* for analyzing the behavior of $\tilde{N}$ on face F_k^*. Because g is real on the boundary, p^* restricted to face F_k^* reduces to

$$p^*(\sigma) = c^* g(\sigma)^{\alpha_{k+1}}[1 - g(\sigma)]^{\alpha_{k+2}} \cos\theta_k, \qquad \sigma \in F_k^*, \tag{33}$$

where p^* is expressed with reference to face F_k^*. The mth moment of p^* on face F_k^* is given by

$$\int_{F_k^*} \sigma^m p^*\, d\sigma_k = c^* \cos\theta_k \int_{F_k^*} \sigma_k^m g(\sigma_k)^{\alpha_{k+1}}[1 - g(\sigma_k)]^{\alpha_{k+2}}\, d\sigma_k, \tag{34}$$

where $d\sigma_k$ is surface Lebesgue measure on F_k^*. The change of variable $x = g(\sigma)$ yields the boundary moments formula

$$\int_{F_k^*} \sigma_k^m p^*\, d\sigma_k = c^* \cos\theta_k \|F_k^*\|^{m+1} I_m(\xi_{k+1}, \xi_{k+2}, \alpha_{k+1}, \alpha_{k+2}). \tag{35}$$

Observe that

$$I_m(a, b, c, d) = \frac{1}{[B(a,b)]^{m+1}} \int_0^1 [B_x(a,b)]^m\, dB_x(a+c, b+d) \tag{36}$$

and hence,

$$I_0(a, b, c, d) = \frac{B(a+c, b+d)}{B(a,b)} \quad \text{and} \quad I_m(a, b, 0, 0) = \frac{1}{m+1}. \tag{37}$$

3.2.3. *Transformation back into original quantities.* Substituting (35) into (28) and using (25) yields the boundary moments formula (20) in the original domain S, where $c = c^* n^2/2$, a scale independent quantity. Using the transformation matrix $\tilde{\Omega}^{-1/2}$, one can express ξ_k and θ_k in terms of original quantities, yielding the expressions (21) and (22).

The normalization constant c is calculated by using $f_i = x_i^2/\Omega_{ii}$ in the basic adjoint relationship (15) for any i. To preserve symmetry at the expense of parsimony, one could choose a "symmetric" function such as $f = x' \operatorname{diag}(\Omega)^{-1} x$. Using f_2 we have that

$$\sum_j \int_{F_j} 2x_2 \frac{R_{2j}}{\Omega_{22}} p_j\, d\sigma_j = -1. \tag{38}$$

Observing that x_2 takes the value $n - \sigma_1$ on F_1, 0 on F_2 and σ_3 on F_3 and using the boundary moments formula (20) yields (24), which concludes the proof of Proposition 1. □

3.3. *Three different cases for the density p.* Three different shapes of the density p can arise corresponding to different combinations of poles and zeros at the vertices. First, since p is harmonic, it cannot have maxima on the interior of S. Singularities can occur only at the vertices, and we find that $n\mathbf{e}_k$, where $\mathbf{e}_k$ is the unit vector in the kth direction, is a pole, a positive real number or a zero if α_k is, respectively, negative, zero or positive. If the diffusion process exists, as it does for the networks that we are considering, then $\alpha_k > -2\pi\xi_k$ for all k ([7], (1.6)). Therefore, any system with a pole has at least one zero, and vice versa. Further, because the harmonic density $p^*(z) = \mathrm{Re}\{h(g(z))\}$, where $h(x) = c\exp(i\theta_l)x^{\alpha_{l+1}}(1-x)^{\alpha_{l+2}}$ with $l = 1, 2, 3$, saddle points of p^* must satisfy

$$\frac{dh(g(z))}{dz} = 0. \tag{39}$$

Except at vertices, we have that $g'(z) \neq 0$ for the Schwarz–Christoffel map g. Thus, excluding the vertices, the condition for saddle points reduces to

$$\alpha_{l+1}(1 - g(z)) + \alpha_{l+2}g(z) = 0. \tag{40}$$

There is no saddle point if $\alpha_{l+1} = \alpha_{l+2}$, and otherwise there is a unique saddle point located on the boundary point with

$$z = g^{-1}\left(\frac{\alpha_{l+1}}{\alpha_{l+1} - \alpha_{l+2}}\right). \tag{41}$$

Using the transformation of Section 3.2.1, one can identify the corresponding boundary point of the simplex S. In conclusion, we have three possible cases:

1. $\alpha_1 = \alpha_2 = \alpha_3 = 0$. The density p is a constant function.
2. One vertex, say $n\mathbf{e}_1$, has $\alpha_1 = 0$. The density has a finite positive limit at $n\mathbf{e}_1$, and one pole and one zero at the other vertices. There is no saddle point.
3. α_1 positive (negative), α_2, α_3 negative (positive). The density has two zeros (poles) at $n\mathbf{e}_2, n\mathbf{e}_3$ and one pole (zero) at $n\mathbf{e}_1$. There is one saddle point on the side opposite $n\mathbf{e}_1$.

Newell [16] reported a similar situation for a rectangular state space. In the next section we will use these relationships to compute network performance measures.

4. Network performance characteristics.

4.1. *Throughput rate.* Let $A(t)$ be the total number of cycles completed in the network up to time t. The *throughput rate* a is the long-run average number of cycles completed per unit time, so $A(t) \sim at$ almost surely as

$t \to \infty$. In a stable open network the throughput rate equals the input rate (given as data), whereas in a closed network it is an endogenous quantity determined by the number of customers n, the service time distributions and the routing structure. Together with the throughput times for the various customer routes, the throughput rate is one of the most important network performance characteristics.

Balanced networks have the same utilization rate at each station and thus the idleness rate at each station i is equal to a common constant δ. Recall that our time scaling convention sets the throughput rate a equal to the common utilization rate, so

$$a + \delta = 1. \tag{42}$$

Equation (17) in Section 3 says that

$$\delta = \int_{F_j} p_j \, d\sigma_j, \tag{43}$$

where $F_j = \{x \in S\colon\ x_j = 0\}$ and p_j is the boundary density defined in Section 3.

Dai and Harrison ([2], Section 8) develop the following iterative scheme to determine the throughput rate. Assume an initial estimate a_1, and observe that the covariance matrix in the Brownian system model becomes $a_1 \hat{\Omega}$. Using (42) and (43), compute $\delta(a_1)$ and a new estimate $a_2 = 1 - \delta(a_1)$. The limiting value a of this iterative algorithm, if it exists, will be a fixed point of the relationship $a + \delta(a) = 1$. Dai and Harrison call a the *refined QNET estimate* of the throughput rate.

They further show that for balanced networks, a time rescaling argument shows that $\delta(a) = a\delta(1)$ for all $a > 0$. Thus the (unique) fixed point a referred to previously can be computed without iteration via

$$a = \frac{1}{1 + \delta(1)} \quad \text{and hence} \quad \delta = 1 - a = \frac{\delta(1)}{1 + \delta(1)}. \tag{44}$$

From the boundary moments formula,

$$\delta(1) = c \frac{\Omega_{ii}}{R_{ii}} I_0(F_i) \cos \theta_i \quad \text{for } i = 1, 2, 3, \tag{45}$$

where c is the normalization constant given in Proposition 1.

4.2. *Throughput times.* The *throughput time* (or *cycle time* or *sojourn time*) T of a customer in an open network is the random variable measuring the total elapsed time between that customer's arrival at the network and eventual departure. For closed networks, T is the time it takes a customer to complete one cycle. Clearly, a given customer's throughput time depends on the locations of the other customers in the network. Section 3 showed how to find the *time*-indexed stationary distribution of customer configurations N. This section investigates the *customer*-indexed stationary distribution of T. In other words, it relates the discrete-time stochastic process T, which is evaluated at customer arrivals, to the continuous-time stochastic process N. We restrict discussion to the stationary distribution of T conditioned on a customer route vector r giving the number of visits the customer makes to each station. We follow the approach of Harrison and Nguyen [9], but we incorporate Little's relation between time averages and customer averages.

The Brownian approximation of throughput time as reviewed by Harrison and Nguyen [9] relies heavily on Reiman's [19] "snapshot principle," stating that the time a customer spends in a heavily loaded system is negligible compared to the time scale on which the state evolves. One can take a "snapshot" of the system when the customer arrives and regard the workloads at the various stations as relatively constant during that customer's entire cycle.

Let ν_{ji} be the number of customers in queue j ahead of a particular "marked" customer upon that customer's ith arrival," let w_{ji} be the total workload these customers represent (measured in time content), and let s_{ji} be the "marked" customer's own service requirement. Then

$$T = \sum_{j=1}^{3} \sum_{i=1}^{f_j} (w_{ji} + s_{ji}). \tag{46}$$

Notice that w_{ji} is the sum of $(\nu_{ji} - 1)^+$ complete service times plus the remaining time of the customer being served when the marked customer arrives. We do not know the exact distributions of ν_{ji} and w_{ji}, but Reiman's snapshot principle implies that they will be virtually constant over the customer cycle, so we will drop the subscripts i. The best approximation to $E(T)$ is the one that Dai and Harrison [2] obtain from Little's law and the Brownian approximation to δ, namely, $E(T) = n/(1 - \delta)$. This gives the correct first moment for an approximating distribution for w.

For open networks, Harrison and Nguyen [9] propose approximating w_j by a random variable distributed according to the time-indexed stationary distribution of the workload process. This approximation, which corresponds to the queue length approximation $\nu_j \simeq_d N_j$, would be exact for an open Kelly network, that is, for an open multiclass network of the sort that Kelly investigated in his classic book [14]. An exact relationship between the distribution of ν_j and N_j is also available in the case of closed Kelly networks, but it is not as simple as the open case. The networks in our setting that fall into Kelly's class are those with exponential service times at each station. Their stationary distributions are uniform over the state space, so it is simply a matter of counting states to calculate

$$\Pr(\nu_j = m) = \left(\frac{n - m}{n + 1 - m} \cdot \frac{n + 2}{n} \right) \Pr(N_j = m), \tag{47}$$

which provides the following relationships for the first and second moments:

$$\begin{gathered} E(\nu_j) = E(N_j) - \frac{1}{3}, \qquad \operatorname{Var}[\nu_j] = \operatorname{Var}[N_j] - \frac{n+1}{9}, \\ \operatorname{Cov}[\nu_j, \nu_k] = \operatorname{Cov}[N_j, N_k] + \frac{n-1}{18}. \end{gathered} \tag{48}$$

Equivalently, ν_j is distributed like the corresponding time-indexed stationary queue length in a Kelly network with $n - 1$ customers. We denote this quantity $N_j^{(n-1)}$. As $n \to \infty$, the two distributions N_j and $N_j^{(n-1)} = \nu_j$ become indistinguishable. In addition, the residual service time of the customer in service, being exponential, is distributed like a full service time. Following the spirit of the discussion in Harrison and Nguyen [9], we propose a first *raw* QNET approximation for T based on the Kelly relationship $\nu_j =_d N_j^{(n-1)}$ and relating queue length to workload by replacing the random service times U_j with their means τ_j:

$$T_{\text{raw}} \approx_d \sum_{j=1}^{3} f_j\left(N_j^{(n-1)} + 1\right)\tau_j. \tag{49}$$

In general, the distributions of ν_j and w_j will vary depending on the service variability parameters b_i^2. Little's law tells how the mean of w_j differs from $E(N_j\tau_j)$, but there is no corresponding relationship for higher moments. In particular,

$$E(N_j) = \frac{1 - \delta_j}{\tau_j} \cdot E[w_j + s_j], \tag{50}$$

and summing over j and canceling τ_j gives

$$\begin{aligned} \sum_{j=1}^{3} E(N_j) = n &= \frac{1}{1 + \delta(1)} E\left[\sum_{j=1}^{3} \frac{w_j}{\tau_j} + \frac{s_j}{\tau_j}\right] \\ &= \frac{1}{1 + \delta(1)} E\left[\sum_{j=1}^{3}\left[(\nu_j - 1)^+ + 1 + \frac{\tilde{U}_j}{\tau_j}\right]\right], \end{aligned} \tag{51}$$

where $\tilde{U}_j$ is the residual service time of the job in service at a customer arrival and the last equality follows from the definition of w and s: w is the waiting time of the $(\nu_j - 1)^+$ customers plus the residual service time of the customer in service. The *refined* QNET approximation will take $(\nu_j - 1)^+ + \tilde{U}_j/\tau_j$ to be distributed like $N_j^{(n-1)}$ shifted deterministically to satisfy (51):

$$(\nu_j - 1)^+ + \tilde{U}_j/\tau_j \approx_d N_j^{(n-1)} + \varepsilon - 1 \quad \text{where } \varepsilon = \frac{n\delta(1) + 1}{3} \tag{52}$$

and the possibly fractional ε is interpreted as the proportion of a normal service time remaining for the customer in service. Thus, the *refined* QNET approximation to T is

$$T_{\text{ref}} \approx_d \sum_{j=1}^{3}\left(\sum_{i=1}^{r_j(N_j^{(n-1)} - 1)} U_{ij} + r_j\tilde{U}_j\right) + \mathscr{S}, \tag{53}$$

where $\tilde{U}_j =_d \varepsilon U$ and $\mathscr{S}$ is the sum of all the services that the customer requires from the network. That is, to obtain a throughput time estimate, we add the customer's service requirements to the service requirements of the $N^{(n-1)}$ other customers. We take these $N^{(n-1)}$ customers to be distributed according to the Brownian model, but we adjust the service time requirement of the one in service according to Little's law. Because r, N and U are mutually independent, the mean throughput time is

$$E\left[T_{\mathrm{ref}}\middle| r\right] = \sum_{j=1}^{3} r_j\left(E\left(N_j^{(n-1)}\right) + \varepsilon\right)\tau_j \tag{54}$$

and the second moment of the throughput time is

$$\begin{aligned} E\left[T_{\mathrm{ref}}^2\middle| r\right] &= \sum_{j=1}^{3} r_j\left(E\left(N_j^{(n-1)}\right) + \varepsilon\right)b_j^2\tau_j^2 \\ &\quad + \sum_{i,j=1}^{3} (r_i r_j)E\left[\left(N_i^{(n-1)} + \varepsilon\right)\left(N_j^{(n-1)} + \varepsilon\right)\right]\tau_i\tau_j. \end{aligned} \tag{55}$$

Because $N_j \to \infty$ with probability 1 as $n \to \infty$, the first term of $E(T^2)$ is asymptotically negligible. Dropping it amounts to replacing U_{ij} with its mean τ_j in (53). We keep it in our second moment estimates because it seems to provide significantly better results, even for fairly large customer populations. However, we will exploit the simpler representation to find estimates of the tail of T in Section 4.3.

Raising (53) to some power yields the corresponding moment of T in terms of moments of N and network primitives. Moments of N are most easily determined by working in the transformed domain S^*. Recall that the basic adjoint relationship (27) allows us to express the moments of N^* in terms of moments of the boundary density by using appropriate test functions f^*. For example, suppose we want to calculate $E(h(N))$. Letting $h(N_1, N_2, n - N_1 - N_2) = \tilde{h}(\tilde{N})$, we have that $E(h(N)) = E(\tilde{h}(\tilde{N})) = \int_{\tilde{S}} \tilde{h}(z)p(z)\,dz$. Applying the linear transform $z^* = \tilde{\Omega}^{-1/2}z$ gives

$$E(h(N)) = \int_{S^*} h^*(z^*)p^*(z^*)\,dz^*, \tag{56}$$

where $h^*(z^*) = \tilde{h}(\tilde{\Omega}^{1/2}z^*)$. To apply the basic adjoint relationship (27) we need to find a function f^* such that $\Delta f^* = h^*$ on S^*. For example, $f^* = x^{m+2}y^n/((m + 1)(m + 2))$ solves $\Delta f^* = x^m y^n$ for $m \in N$, $n = 0, 1$, supplying us with the first and second moments of N (we do not know whether there is always a polynomial solution f^*). The basic adjoint relationship (27) yields

$$E(h(N)) = -\sum_j \frac{\Omega_{jj}^{1/2}}{R_{jj}} \int_{F_j^*} R^{*j'}\,\nabla f^*(\sigma^*)\cdot p^*(\sigma^*)\,d\sigma^*, \tag{57}$$

which can be evaluated in terms of the functions $I_n(F_i^*)$.

4.3. *Approximate tail behavior of the throughput time density.* In many applications the 95th percentile of the throughput time is another important performance measure. Indeed, organizations often take more interest in the quality of customer service they can guarantee with high likelihood than in their mean performance. Although we cannot find an exact closed form expression for the 95th percentile, we can approximate it using the "linearized" version of the QNET approximation obtained by replacing U_{ij} with τ_j in (53):

$$T_{\text{lin}} \simeq \sum_j f_j \tau_j \left(N_j^{(n-1)} + \varepsilon \right). \tag{58}$$

In particular, if the network has deterministic routing, then $T = n + n\delta(1)$ is a constant. Thus the linearized QNET approximation predicts deterministic throughput times for networks with deterministic routing. For more general networks, the approximating distribution $P(T \le t)$ of T involves the integral of p over the polygonal intersection of the simplex S with the half-space $T \le t$. This integral cannot be done in closed analytic form because the inverse Schwarz–Christoffel map is not known in closed form. Nevertheless, we can estimate the tail fractiles of T.

Denote the $100(1-\eta)$th percentile of T by T_η: $P(T \ge T_\eta) = \eta$. Given a specific route (i.e., given f_j), the plane $T = T_\eta$ maps onto a line in the triangle S^*. For small η, this line will be close to one vertex, say v_k, and can be expressed in a coordinate system centered at this vertex as $c_1 x^* + c_2 y^* = c_3 - T$. (The proposed approximation scheme does not apply to the unlikely case where the line is parallel to a side.) Using a polar coordinate system (r, θ) centered at v_k and setting $\beta_k = (\theta_{k-1} - \theta_{k-2})/\pi\xi_k$, the density has the following approximation in the vicinity of v_k:

$$p^*(z) = A r^{\beta_k} \cos(\beta_k \theta + \theta_{k-1}) + o(r^{\beta_k}), \tag{59}$$

where

$$A = 2c \left(\frac{\sqrt{\det(\Omega)}}{n} \right)^{2+\beta_k} \left(\frac{\xi_k B(\xi_k, \xi_{k+1})}{\sqrt{\Omega_{k-1,k-1}}} \right)^{\beta_k} \tag{60}$$

and all quantities are defined in Proposition 1. Using this approximation, we estimate

$$P\left(T \ge T_\eta \middle| f, \mathscr{S}\right) \simeq B\left(c_3 - T_\eta\right)^{\beta_k + 2}, \tag{61}$$

where

$$B = \frac{A}{\beta_k + 2} \int_0^{\pi\xi_k} \frac{\cos(\beta_k \theta + \theta_{k-1})}{(c_1 \cos\theta + c_2 \sin\theta)^{\beta_k + 2}} \, d\theta. \tag{62}$$

For a given route, T_η can be solved explicitly from (61) as

$$T_\eta \simeq c_3 - \left(\frac{\eta}{B} \right)^{1/(\beta_k + 2)} \tag{63}$$

This relationship is linear in n (because both terms are). For more general routing, T_η can be calculated from $P(T \ge T_\eta) = \eta$ by unconditioning (61) using the joint distribution of the routing structure $(\mathscr{S}, f)$.

The next section shows how to use the Brownian system model together with the relationships of this section to calculate performance measures for some specific networks.

5. Examples. In this section we compute the throughput rate, the first two moments of the throughput time distribution and the 95th percentile of the throughput time distribution for three illustrative models and compare them to results obtained from simulation. This allows us to test the quality of the Brownian approximation for different sizes of the customer population and to quantify the improvement of our refinement to the raw QNET approximation. In addition, we propose an approximate expression for the idleness rate and expected throughput time based only on network primitives (i.e., not requiring the Brownian calculations of Section 3) and test its accuracy. Finally, we explore the effects of randomness in routing.

We ran each simulation experiment on SIMAN 3.5 [17] for 300,000 time units unless noted otherwise. We used lognormal distributions at each station. In accordance with the convention established in [21], the number reported in parentheses after each simulation result is the half-width of the 95% confidence interval expressed as a percentage of the simulation value. The number in parentheses after the Brownian estimate is the percentage deviation from the simulation average. We show no percentage deviation when the two numbers agree.

The appendix to [2] provides a condition under which a general closed network's queue length vector has a "product form" stationary distribution. In this special case the distribution is a multidimensional exponential (with exponents related to the differences between station utilizations), and the density p is particularly easy to compute in closed form. Both our raw and refined QNET approximations are based on intuition derived from the product form case. Here we use the product form case to develop a cruder approximation that does not require computing the stationary distribution of the approximating RBM.

For three-station balanced networks, the product form condition reduces to

$$\Omega_{11}/R_{11} = \Omega_{22}/R_{22} = \Omega_{33}/R_{33}, \tag{64}$$

and in that case the associated interior distribution is uniform over the state space. The throughput rate and idleness depend not on the interior density, but on the boundary densities. They can be found for product form networks by means of a simple comparison to conventional queueing results, as the next paragraph explains. We define the vector $\beta = (\Omega_{11}/R_{11}, \Omega_{22}/R_{22}, \Omega_{33}/R_{33})$ to simplify future formulas.

In the case when the service time distributions are exponential, we call the network a "Kelly network" because it falls into the class studied in Chapter 3 of [14]. A balanced three-station Kelly network in heavy or light traffic has $b_1^2 = b_2^2 = b_3^2 = 1$ and $\beta_i = \Omega_{ii}/R_{ii} = 2$. Its distribution is uniform over its entire state space (both interior and boundary), and its throughput rate can be shown to be

$$a = \frac{n}{n+2} \quad \text{with } \delta = \frac{2}{n+2}. \tag{65}$$

This result is independent of routing and is exact for any customer population size n. If the service time SCV's of a *Brownian* network are scaled in such a way that $\beta_i^* = \lambda\beta_i$ while the routing parameters remain unchanged, then $\delta^*(1) = \lambda\delta(1)$. Thus, for networks satisfying the product form condition, we can compute the Brownian estimates of δ and a using the result for Kelly networks:

$$\delta = \frac{\beta_i}{n + \beta_i} = \frac{\delta(1)}{1 + \delta(1)}, \qquad a = \frac{n}{n + \beta_i}. \tag{66}$$

In light of this, we propose the following conjecture.

CONJECTURE 1. *For a general non-product-form network,*

$$\min_i \beta_i \le n\delta(1) \le \max_i \beta_i, \text{ where } \beta_i = \frac{\Omega_{ii}}{R_{ii}} \tag{67}$$

and in particular,

$$\begin{gathered}\min_i \frac{\beta_i}{n + \beta_i} \le \delta \le \max_i \frac{\beta_i}{n + \beta_i} \quad \text{and} \\ \min_i (n + \beta_i) \le \mathrm{E}T \le \max_i (n + \beta_i).\end{gathered} \tag{68}$$

We will see in the examples in Sections 5.1 and 5.2 that the average of the β_i (denoted $\bar{\beta}$) is a good first approximation for $n\delta(1)$. In effect, it replaces each station's variability (as measured by β_i) by the average variability of the three stations when they are treated as isolated $GI/G/1$ queues. This *product form* approximation gives idleness and throughput time estimates of

$$\delta \simeq \frac{\bar{\beta}}{n + \bar{\beta}} \quad \text{and} \quad E(T) \simeq n + \bar{\beta}. \tag{69}$$

In fact, the examples suggest that δ is often an upper bound on idleness, which is achieved only in the case of product form networks.

5.1. *Cyclic queue.* The simplest three-station closed network is a cyclic queue like the one shown in Figure 3. We use this case to illustrate the calculations of Sections 2–4. Because the network is balanced and because this is a two-moment approximation, the model of a cyclic queue is determined by the squared coefficients of variation of the service time distributions $b = (b_1^2, b_2^2, b_3^2)$. The routing and constituency matrices for the cyclic queue are

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{70}$$

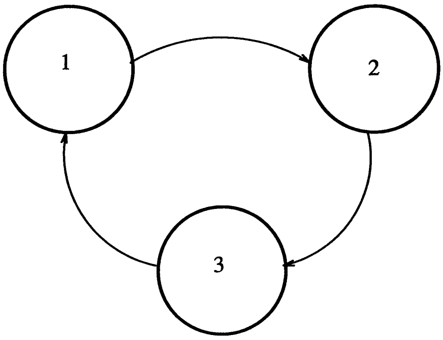

FIG. 3. *The cyclic queueing network.*

These lead to particularly simple reflection and covariance matrices R and $\hat{\Omega}$,

$$R = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \quad \text{and}$$
$$\hat{\Omega} = \begin{pmatrix} b_1^2 + b_3^2 & -b_1^2 & -b_3^2 \\ -b_1^2 & b_1^2 + b_2^2 & -b_2^2 \\ -b_3^2 & -b_2^2 & b_2^2 + b_3^2 \end{pmatrix}, \tag{71}$$

and to $\beta = (b_1^2 + b_3^2,\ b_1^2 + b_2^2,\ b_2^2 + b_3^2)$. Thus the product form condition is $b_1^2 = b_2^2 = b_3^2$, in which case the functions $I_n(F_i)$ take the form $I_n(F_i) = 1/(n+1)$. In the product form case, Proposition 1 yields

$$c = \frac{1}{\cos\theta_1} \quad \text{and} \quad \int_{F_i} \sigma^k p_i(\sigma)\, d\sigma = \frac{2n^{k-1}}{k+1}, \tag{72}$$

from which we have that

$$\delta(1) = \frac{2b_1^2}{n} \quad \text{and} \quad a = \frac{1}{1 + 2b_1^2/n}, \tag{73}$$

in agreement with the general product form expressions for δ and a given in (66). The functions $I_n(F_i)$ cannot be calculated in closed form for non-product-form networks, so we rely on the boundary moments formula (20).

Because routing is deterministic, the raw QNET approximation predicts that the throughput time should be constant and equal to n/a. In the notation of Section 4.2, $f_k = 1$ and $\tau_k = 1$ for all k. The raw and refined QNET approximations are

$$E(T_{\mathrm{raw}}) = n + 2, \tag{74}$$

$$E(T_{\mathrm{raw}}^2) = n^2 + 4n + 4, \tag{75}$$

$$E(T_{\mathrm{ref}}) = n(1 + \delta(1)), \tag{76}$$

$$E(T_{\mathrm{ref}}^2) = n^2(1 + \delta(1))^2 + b'E(N) + \frac{n\delta(1)}{3}e'b. \tag{77}$$

Table 1 compares the two sets of Brownian estimates for the cyclic queue to the results of simulation for some particular parameter values. In each case, only δ and $E(T_{\mathrm{ref}})$ are exact measures for the Brownian system; $E(T_{\mathrm{raw}})$, $E(T^2)$ and $T_{0.95}$ are approximations. As we approach the Brownian regime, the refined estimate is clearly superior (by more than an order of magnitude). Although the raw QNET approximation discards all service variability data when estimating $E(T)$ and $E(T^2)$, it significantly outperforms the refined approximation in the high variability case with $n = 10$. While the refined Brownian estimates of mean throughput time range from 12 to 25 for the 10-customer networks, the simulated values remain relatively constant between 12 and 14. This can best be interpreted as compensating errors—the approximating Brownian network is far from the original network, and the raw QNET estimates are far from the Brownian network. We conjecture that a cyclic queue with a small customer population enjoys so

TABLE 1
Simulation and Brownian estimates of performance measures of the cyclic queue

(b_1^2, b_2^2, b_3^2)	n		δ		$E[T]$		$(E[T^2])^{1/2}$	
(1, 0.64, 0.81)	10	SIM	0.140	(1.1%)	11.65	(0.3%)	12.08	(0.3%)
		RAW	0.140		12.00	(3.0%)	12.00	(0.7%)
		REF	0.140		11.63	(0.2%)	12.03	(0.4%)
	100	SIM	0.017	(10.0%)	101.7	(0.3%)	102.2	(0.3%)
		RAW	0.016	(5.9%)	102.0	(0.3%)	102.0	(0.2%)
		REF	0.016	(5.9%)	101.6	(0.1%)	102.0	(0.2%)
(4, 2, 5)	10	SIM	0.293	(1.0%)	14.2	(0.7%)	16.5	(1.3%)
		RAW	0.422	(44%)	12.0	(15.5%)	12.0	(27.3%)
		REF	0.422	(44%)	17.3	(21.8%)	19.1	(15.8%)
	100*	SIM	0.064	(2.0%)	106.4	(0.2%)	108.7	(0.2%)
		RAW	0.068	(6.3%)	102.0	(4.1%)	102.0	(6.2%)
		REF	0.068	(6.3%)	107.3	(0.8%)	109.2	(0.5%)
	1000*	SIM	0.0072	(12.5%)	1004	(0.2%)	1007	(0.2%)
		RAW	0.0073	(1.4%)	1002	(0.2%)	1002	(0.5%)
		REF	0.0073	(1.4%)	1007	(0.3%)	1009	(0.2%)
(0.25, 1, 25)	10	SIM	0.289	(1.4%)	14.1	(1.3%)	20	(5.4%)
		RAW	0.601	(108.0%)	12.0	(14.9%)	12	(40%)
		REF	0.601	(108.0%)	25.0	(77.3%)	30	(50%)
	100*	SIM	0.082	(2.6%)	108.6	(0.4%)	118	(1.4%)
		RAW	0.131	(59.8%)	102.0	(6.1%)	102	(13.6%)
		REF	0,131	(59.8%)	115.0	(5.9%)	121	(2.5%)
	1000*	SIM	0.013	(12.9%)	1011	(0.6%)	1021	(0.5%)
		RAW	0.015	(15.4%)	1002	(0.9%)	1002	(1.9%)
		REF	0.015	(15.4%)	1015	(0.4%)	1021	

*Run for 3,000,000 cycles.

much dependency between its state transitions that it is not susceptible to much performance degradation due to worsening service SCV's.

Table 1 shows that as the service distributions become more variable, the network converges more slowly to its Brownian limit, both in terms of the tightness of the confidence intervals and in terms of the number of customers in the network. The latter phenomenon is illustrated in Figure 4, which shows how the relative error of the Brownian approximation decreases as the number of customers grows. It is approximately $O(n^{-1})$ or C/n, where the constant C depends on the station SCV's. Because the linearized approximation (58) for T used in the 95th percentile estimate is degenerate in the cyclic case, expression (61) cannot be used to find an estimate for $T_{0.95}$. In fact, since T_{lin} is a constant, its 95th percentile equals its mean. In other words, all of the throughput time variability in a cyclic Brownian network is due to service variability, but the estimate $T_{0.95}$ is based on discarding service variability in order to explore queue length variability. Examples in Sections 5.2 and 5.3 explore the quality of the 95th percentile approximation.

Figure 5 compares the throughput time densities obtained from simulation and from the Brownian approximation for the cyclic network with $b = (4, 2, 5)$ and 100 customers. (The Brownian approximation is taken to be a normal distribution with first and second moments supplied by the refined QNET estimate.) The greater width of the simulation distribution suggests that the

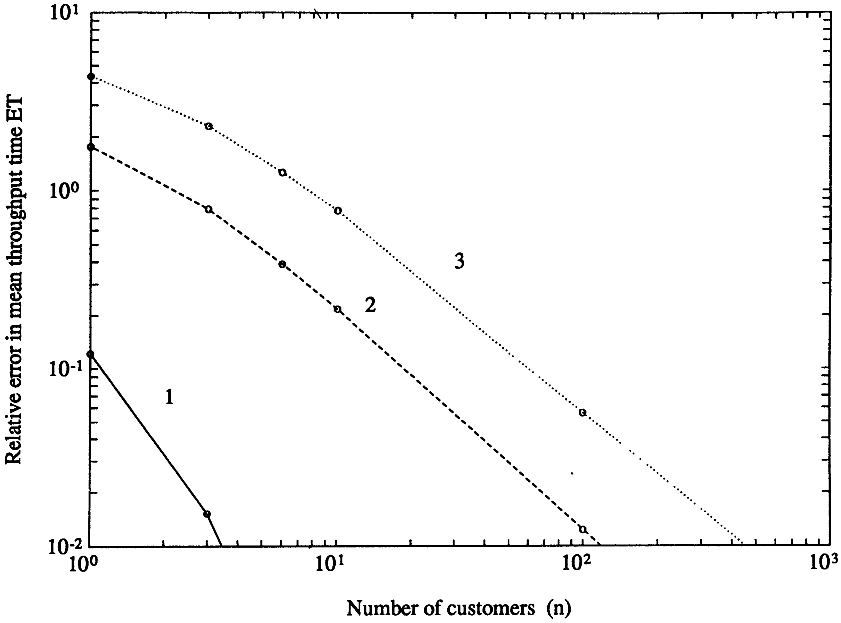

FIG. 4. *Convergence to the Brownian limit as $n \to \infty$ for the cyclic network with case* 1 [b = (1, 0.64, 0.81)], *case* 2 [b = (4, 2, 5)] *and case* 3 [b = (0.25, 1, 25)].

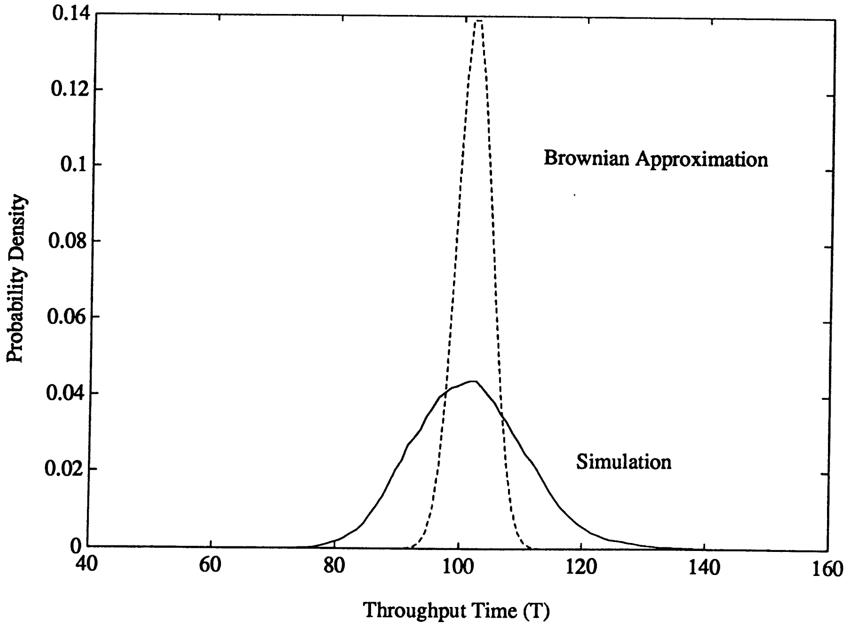

FIG. 5. *The throughput time density of the Brownian approximation compared to the (wider) simulation result for the cyclic network with* $b = (4, 2, 5)$ *and* 100 *customers.*

snapshot principle does not apply at this population size: The state must be evolving on a time scale comparable to customer throughput times.

Figure 6 shows the idleness proxy $\delta(1)$ as a function of b_1^2 and b_2^2, where $0 \le b_1^2,\ b_2^2 \le 2$ and $b_3^2 = 1$. We conclude from the contour plot that (for all practical purposes) $n\delta(1)$ is only a function of the sum of the coefficients of variation, $b_1^2 + b_2^2$. In Figure 7, we plot $n\delta(1)$ as a function of b_1^2 for three fixed values of b_3^2 with $b_2^2 = (3 - b_1^2 - b_3^2)$. That is, the sum $\sum b_i^2$ is held constant while the relative variabilities of the three stations are allowed to vary. Each of the curves in Figure 7 shows the value of $n\delta(1)$ along a diagonal of a plot like the one in Figure 6. We see that $n\delta(1)$ is close to $\frac{2}{3}(b_1^2 + b_2^2 + b_3^2) = 2$, the "product form" approximation proposed in (69), and is only slowly varying in the given range. For the ranges shown, approximating the network by one with identical servers introduces errors of less than 5%.

5.2. *A model of a CPU and two I/O's.* Consider a 10-customer computer system composed of a central processing unit and two input/output (I/O) devices with different service variabilities as shown in Figure 8. The CPU randomly splits messages between I/O_1 and I/O_2, sending a proportion ϕ to I/O_1. This might represent a small office computer system, for example. (Of course, mean service times at the two I/O devices are adjusted in all cases to maintain perfect balance.) The CPU, I/O_1 and I/O_2 are labeled stations 1, 2

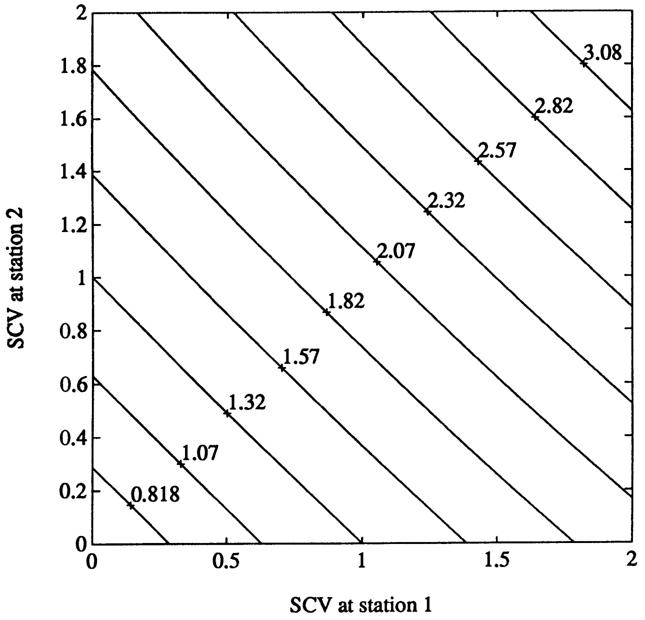

FIG. 6. *Contour plot of the cyclic network's* $n\delta(1)$ *as a function of* b_1^2 *and* b_2^2 *for* $0 \leq b_1^2, b_2^2 \leq 2$ *and* $b_3^2 = 1$.

and 3 respectively. Routing is Markovian among stations, so this is a generalized Jackson network, and we can take the constituency matrix to be $C = I$. Then

$$P = \begin{pmatrix} 0 & \phi & 1-\phi \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{78}$$

Product form networks have $b_1^2 = 1$, $b_2^2 = b_3^2$, with associated idleness rate

$$\delta = \frac{b_2^2 + 1}{n + b_2^2 + 1} \quad \text{independent of } \phi. \tag{79}$$

Table 2 compares the results of simulation to the refined Brownian estimates. A comparison to the deterministically routed cyclic network reveals that the random splitting in this network tends to temper service SCV's greater than 1 and to amplify service SCV's less than 1. Thus this network converges faster to its Brownian limit than a cyclic network with the same parameters. From a managerial perspective, we learn that the random routing of the computer system enhances the throughput rate of highly variable networks relative to deterministically routed systems with the same service variabilities and degrades it for less variable ones.

In addition, it is interesting to pose the question: given two I/O technologies with known variabilities, what is the optimal allocation of capacity

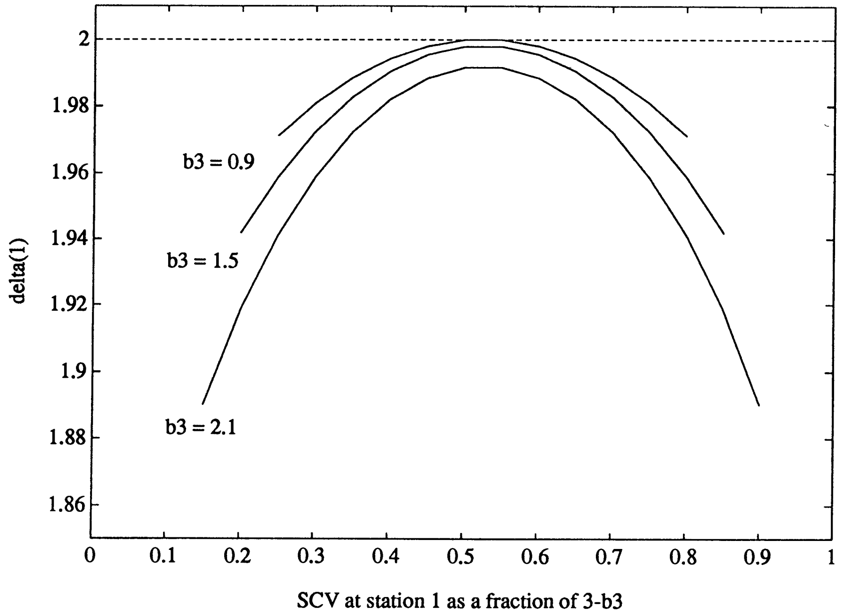

FIG. 7. *Comparison of* $n\delta(1)$ *to its product form upper bound of* 2 *as a function of* b_1^2, *where* $\sum b_i^2 = 3$.

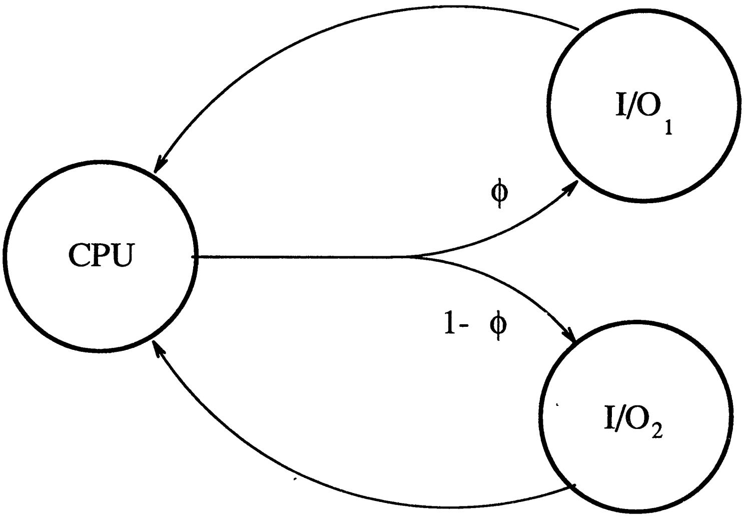

FIG. 8. *The computer network.*

TABLE 2
Simulation, Brownian and product form (P.F.) *estimates for the computer model, with* $n = 10$ *customers*

(b_1^2, b_2^2, b_3^2)	f		δ		$E[T]$		$(E[T^2])^{1/2}$		$T_{0.95}$	
(1, 0.64, 0.81)	0.1	SIM	0.147	(3.4%)	11.74	(0.7%)	18.4	(2.5%)	36	(50%)
		REF	0.147		11.73	(0.1%)	17.8	(3.3%)	38	(5.6%)
		P.F.	0.149	(1.4%)	11.75	(0.1%)				
	1/3	SIM	0.147	(2.7%)	11.73	(0.4%)	13.5	(0.7%)	24	(14.4%)
		REF	0.147		11.73		13.3	(1.5%)	22	(8.3%)
		P.F.	0.148	(0.7%)	11.73					
	2/3	SIM	0.147	(2.0%)	11.70	(0.4%)	13.7	(0.7%)	25	(15.6%)
		REF	0.147		11.73	(0.3%)	13.4	(2.2%)	23	(4.0%)
		P.F.	0.146	(0.7%)	11.72	(0.2%)				
	0.9	SIM	0.146	(2.7%)	11.70	(0.7%)	19.0	(2.7%)	38	(52.0%)
		REF	0.147	(0.7%)	11.73	(0.3%)	18.4	(3.2%)	40	(5.3%)
		P.F.	0.145	(0.7%)	11.70					
(4, 2, 5)	0.1	SIM	0.264	(2.3%)	13.5	(1.2%)	23	(4.6%)	39	(38.5%)
		REF	0.375	(42.0%)	16.0	(18.5%)	23		42	(7.7%)
		P.F.	0.408	(54.5%)	16.9	(25.2%)				
	1/3	SIM	0.265	(1.5%)	13.6	(0.9%)	17.7	(1.6%)	32	(20.5%)
		REF	0.375	(41.5%)	16.0	(17.6%)	19.1	(7.9%)	24	(25.0%)
		P.F.	0.400	(50.9%)	16.7	(22.8%)				
	2/3	SIM	0.268	(1.5%)	13.7	(1.0%)	18.9	(2.5%)	34	(25.5%)
		REF	0.375	(38.9%)	16.0	(16.8%)	20.2	(6.9%)	29	(14.7%)
		P.F.	0.388	(44.8%)	16.3	(19.1%)				
	0.9	SIM	0.266	(3.0%)	13.7	(1.9%)	28	(12.2%)	36	(56.6%)
		REF	0.375	(38.9%)	16.0	(16.8%)	28		55	(52.8%)
		P.F.	0.379	(42.5%)	16.1	(17.5%)				

between them? The approximation for δ in (69) is minimized by choosing ϕ to minimize $\bar{\beta}$. On the other hand, from the insights of Figure 7, we suspect that performance will improve when the differences $|\beta_i - \beta_j|$ are large. In this CPU example, maximizing the sum of these differences amounts to maximizing

$$(80) \qquad (1-\phi)|\varepsilon_2 + \varepsilon_1| + \phi|\varepsilon_1 - \varepsilon_2| + \left|(2\phi - 1)\varepsilon_1 - \varepsilon_2\right|,$$

where $\varepsilon_1 = b_1^2 - 1$ and $\varepsilon_2 = b_3^2 - b_2^2$. In the two simulation cases shown in Table 2, $\varepsilon_2 > \varepsilon_1 \geq 0$, and a small amount of algebra shows that $\bar{\beta}$ is minimized for $\phi = 1$, while the sum of the differences is maximized for $\phi = 0$. Thus we cannot predict heuristically which ϕ will provide the best performance, and from our simulation and RBM results, we see that the two considerations cancel: δ and $E(T)$ are independent of ϕ. Although the routing variability does not affect δ and $E(T)$, it has a great impact on the variability of the throughput time. Both the second moment and the 95th percentile increase for asymmetric splitting. The minimal second moment $E(T^2)$ is achieved at some intermediate value, which the refined QNET approximation predicts is 0.486 for $[b_1^2, b_2^2, b_3^2] = [1, 0.64, 0.81]$ and 0.411 for $[4, 2, 5]$. The tail approximation seems to perform very well, yielding results that are all except one in the confidence interval. It gives the best results for service SCV's near 1, in which case the errors are around 5%.

We show the product form approximations for δ and $E(T)$ in Table 2 and find that they give errors of roughly the same magnitude as those shown in Figure 7.

5.3. *A multiclass manufacturing example.* As a final example, consider Dai and Harrison's model of a three-station manufacturing system inspired by Solberg. Figure 9 illustrates this system. It consists of a mill, a drill and an inspector that produce two types of products, A and B. Product A requires milling, drilling and milling operations, in that order, plus a 50% chance of inspection after the drilling. Product B requires one visit to the drill followed by one visit to the mill. In contrast to the two previous examples, this one has multiple classes at two of the nodes, so it allows us to test the quality of the Brownian approximation for a case in which heavy traffic limits have not been proved. In addition, because there are two natural customer types, it provides a test of the performance approximations for more general "cycles." Although the route of each product satisfies the assumption of Markovian class switching, Dai and Harrison point out the desirability of allowing the

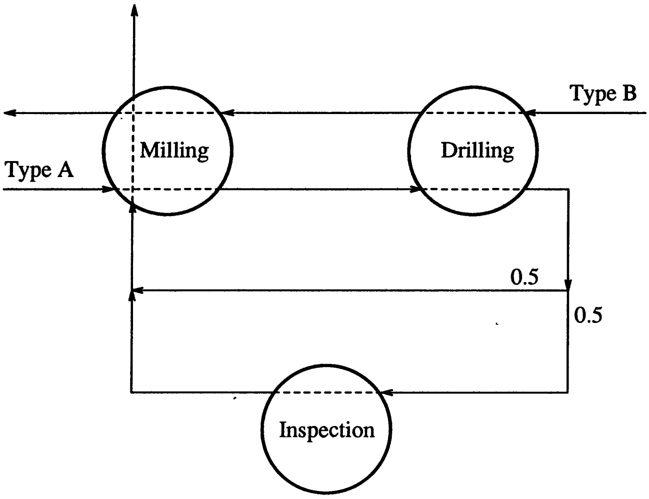

FIG. 9. *A multiclass manufacturing network.*

replacement of completed jobs with new ones to be more regular than a Markov process. We take advantage of the greater generality of Dai and Harrison's formulation to let the transition from output to input classes alternate deterministically between types A and B. Labeling the mill, drill and inspector as stations 1, 2 and 3, respectively, we have that

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

and new jobs alternate between classes 1 and 5. Dai and Harrison [2] discuss how to compute the parameters of the approximating Brownian network in this greater generality. (In their notation, the vector of relative input rates is $\alpha = (0{,}5, 0, 0, 0, 0.5, 0)$ with associated covariance matrix $\Delta = \mathbf{0}$.) We take $n = 100$ customers as representative of a large manufacturing shop.

The product form condition for this network is $b_1^2 = \frac{1}{2}b_2^2 + \frac{1}{2} = \frac{2}{5}b_3^2 + \frac{3}{5}$. The results for this network are shown in Table 3. By deriving separate performance predictions for the two natural subpopulations, we lose some accuracy; throughput times of type A are overestimated and those for type B are underestimated relative to the simulation results. Nevertheless, even the 95th percentile estimates remain very accurate.

TABLE 3

Simulation and Brownian estimates for the multiclass example, with n = 100 customers. All simulations run for 3,000,000 cycles

(b_1^2, b_2^2, b_3^2)	Type		δ		$E[T]$		$(E[T^2])^{1/2}$		$T_{0.95}$	
(1, 0.64, 0.81)	A	SIM	0.015	(9.0%)	144.5	(0.9%)	167	(1.6%)	308	(8.4%)
		REF	0.015		147.1	(1.8%)	170	(1.8%)	302	(1.9%)
	B	SIM			58.0	(2.0%)	62	(1.6%)	90	(4.4%)
		REF			56.1	(3.3%)	60	(3.2%)	86	(4.4%)
$(\frac{4}{3}, \frac{4}{3}, \frac{3}{2})$	A	SIM	0.024	(7.1%)	147	(0.9%)	171	(1.5%)	318	(8.8%)
		REF	0.024		148	(0.7%)	171		300	(5.7%)
	B	SIM			58	(2.1%)	62	(1.5%)	94	(5.4%)
		REF			57	(1.7%)	61	(1.6%)	88	(6.4%)
(4, 2, 5)	A	SIM	0.052	(6.3%)	152	(0.9%)	182	(1.5%)	348	(11.2%)
		REF	0.058	(11.5%)	155	(1.9%)	183	(0.5%)	318	(8.6%)
	B	SIM			58	(1.9%)	63	(1.3%)	99	(6.6%)
		REF			57	(1.7%)	62	(1.6%)	88	(11.1%)

Acknowledgments. We gratefully acknowledge Professor J. Michael Harrison for suggesting this research project and for his helpful advice and active supervision. We also thank Professors Viên Nguyen, Larry Wein and Ruth Williams for their thoughtful comments on an earlier draft. Piet Van Mieghem proposed formula (36).

REFERENCES

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

[2] Dai, J. and Harrison, J. M. (1993). The QNET method for two-moment analysis of closed manufacturing systems. *Ann. Appl. Probab.* **3** 968–1012.

[3] Dai, J. G. and Nguyen, V. (1994). On the convergence of multiclass queueing networks in heavy traffic. Preprint. *Ann. Appl. Probab.* To appear.

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

[5] Harrison, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Adv. in Appl. Probab.* **10** 886–905.

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

[7] Harrison, J. M., Landau, H. J. and Shepp, L. A. (1985). The stationary distribution of reflected Brownian motion in a planar region. *Ann. Probab.* **13** 744–757.

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

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

[10] Harrison, J. M. and Shepp, L. (1984). A tandem storage system and its diffusion limit. *Stochastic Process. Appl.* **16** 257–274.

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

[12] Harrison, J. M., Williams, R. J. and Chen, H. (1990). Brownian models of closed queueing networks with homogeneous customer populations. *Stochastics* **29** 37–74.

[13] Lazowska, E. D., Zahorjan, J., Graham, G. S. and Sevcik, K. C. (1984). *Quantitative System Performance.* Prentice-Hall, Englewood Cliffs, NJ.

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

[15] Nehari, Z. (1975). *Conformal Mapping.* Dover, New York.

[16] Newell, G. F. (1979). *Approximate Behavior of Tandem Queues. Lecture Notes in Econon. and Math. Systems* **171**. Springer, Berlin.

[17] Pegden, C. D. (1989). *Introduction to SIMAN.* Systems Modeling Corporation, Sewickley, PA.

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

[19] Reiman, M. I. (1982). The heavy traffic diffusion approximation for sojourn times in Jackson networks. In *Applied Probability-Computer Science: The Interface* (R. L. Disney and T. J. Ott, eds.) **2** 409–422. Birkhäuser, Boston.

[20] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Operat. Res.* **9** 441–458.

[21] Reiman, M. I. (1990). Asymptotically exact decomposition approximations for queueing networks. *Operat. Res. Lett.* **9** 363–370.

[22] Sauer, C. H. and Chandy, K. M. (1981). *Computer Systems Performance Modeling.* Prentice-Hall, Englewood Cliffs, NJ.

[23] SOLBERG, J. J. (1981). Capacity planning with a stochastic workflow model. *AIIE Trans.* **13** 116–122.

[24] SPEARMAN, M. L., WOODRUFF, D. L. and HOPP, W. J. (1990). CONWIP: A pull alternative to kanban. *International Journal of Production Research* **28** 879–894.

[25] TREFETHEN, L. N. and WILLIAMS, R. J. (1986). Conformal mapping solution of Laplace's equation on a polygon with oblique derivative boundary conditions. *J. Comput. Appl. Math.* **14** 227–249.

[26] WHITT, W. (1983). The queueing network analyzer and Performance of the queueing network analyzer. *Bell Syst. Tech. J.* **62** 2779–2843.

[27] WHITT, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Sci.* **39** 1020–1027.

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-5015