The Annals of Applied Probability
1993, Vol. 3, No. 4, 968–1012

THE QNET METHOD FOR TWO-MOMENT ANALYSIS OF CLOSED MANUFACTURING SYSTEMS[1]

BY J. G. DAI AND J. M. HARRISON

Georgia Institute of Technology and Stanford University

Consider a job-shop or batch-flow manufacturing system in which new jobs are introduced only as old ones depart, either because of physical constraints or as a matter of management policy. Assuming that there is never a shortage of new work to be done, the number of active jobs remains constant over time, and the system can be modeled as a kind of closed queueing network. With manufacturing applications in mind, we formulate a general closed network model and develop a mathematical method to estimate its steady-state performance characteristics. A restrictive feature of our network model is that all the job classes that are served at any given node or station share a common service time distribution.

Our analytical method, which is based on an algorithm for computing the stationary distribution of an approximating Brownian model, is motivated by heavy traffic theory; it is precisely analogous to a method developed earlier for analysis of open queueing networks. The required inputs include not only first-moment information, such as average product mix and average processing rates, but also second-moment data that serve as quantitative measures of variability in the processing environment. We present numerical examples that show that system performance is very much affected by changes in second-moment data. In these few numerical examples, our estimates of average throughput rates and average throughput times for different product families are generally accurate when compared against simulation results.

Contents

Received August 1992; revised April 1993.

[1] Supported in part by Semiconductor Research Corporation Grant 84-01-046, Texas Instruments Corporation Grant 90456-034 and NSF Grants DMS-92-03524 and DDM-92-15233.

AMS 1991 *subject classifications.* Primary 60K25, 60J70; secondary 90B22, 90B30.

Key words and phrases. Performance analysis, closed multiclass queueing network, Brownian approximation, heavy traffic.

1. Introduction. This paper is concerned with queueing network models of job-shop or batch-flow manufacturing systems. For our purposes a manufacturing system is a collection of "workstations," or simply "stations," each of which has one or more interchangeable "servers" working in parallel. (The "servers" may represent either machines or operators.) The entities that are processed at the workstations will be called "jobs," although the standard term in queueing theory is "customer." Depending on the particular manufacturing context, what we call a job might actually be referred to as a "work order," a "production lot," a "manufacturing order" or a "production batch." In the models considered here, there are J distinct workstations, and 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, and the time required to perform any given operation is called a "service time." Both the route and the service times of a given job may be viewed a priori as uncertain, or stochastic.

The original work of Jackson [25, 26] on queueing network models was motivated by manufacturing applications, but Jackson's model contained a simplifying assumption that many found objectionable: Jobs at any given workstation i were treated as essentially identical, and the probability that a job leaving station i would go next to station j was assumed to be some number P_{ij} that did not depend on the job's previous processing history. Later research [1, 27] showed that Jackson's analysis could be extended to networks with much more general routing: In the general scheme, jobs at a given station may be of various *classes*, and the future routing of a job (that is, the switching probabilities P_{ij}) may depend on the job's class designation. One can represent virtually any kind of job routing within such a model structure (see Section 2).

The extension of Jackson's analysis to multiclass networks is crucially important in every area of application, but unfortunately, conventional queueing network theory is still limited in its predictive power by very restrictive distributional assumptions. To be specific, it is assumed in conventional theory that all input processes are Poisson and that each station is characterized by either (a) a common exponential service time distribution for the job classes served at the station or else (b) a special type of service discipline, such as processor sharing, that causes the station to behave essentially as if it had a common exponential service time distribution. The net effect of these special assumptions is that estimates or predictions of steady-state performance depend only on first-moment data; that is, on average arrival rates, average service times and average switching proportions. In other words, conventional queueing formulas do not show how system performance changes as one increases or decreases the amount of *variability* in the operating environment, although it is statistical variability that causes the congestion effects that are the object of study in performance analysis.

This paper continues the work of Harrison, Nguyen and Dai [18, 19, 10] on two-moment analysis of multiclass queueing networks. Because of their multiclass structure, our models are very general in the job routing allowed, and they use both first-moment information and second-moment information (that is, both average-rate data and variability data) in estimating system performance. What distinguishes this paper from its predecessors is that we treat *closed systems* rather than open systems, or closed network models rather than open network models. In an open network model, the arrivals of new jobs are viewed as uncontrollable and (perhaps) random events, and the number of active jobs fluctuates over time. At first glance this may seem like a reasonable representation of manufacturing reality, but as Solberg [36, 37] observed, it is clearly wrong for the kind of automated machining facility that is increasingly common in the metal cutting industry. In such a system, one typically has a fixed number of pallets on which workpieces are mounted, and when a workpiece has finished its route, it is removed from the pallet and another workpiece immediately replaces it. Thus, as Solberg pointed out, it is really pallets that play the role of "jobs," and the correct queueing model is a closed network; that is, a network with a fixed set of "jobs" that circulate perpetually through the workcenters, with no arrivals and no departures.

Solberg's proposal of a closed network model was based on physical restrictions, but a much larger number of manufacturing facilities are operated with "closed loop input control," either exact or approximate, which leads to the same type of model for performance analysis purposes. That is, in order to reduce physical clutter or simplify the operating environment, management may set a target value for the number of active manufacturing orders, decreeing that new orders may enter the factory floor only as old ones leave it. Such iron-fisted restrictions on work-in-process are one way in which managers seek to implement the "just-in-time" philosophy of material flow control. Chen, Harrison, Mandelbaum, van Ackere and Wein [6] describe a Hewlett–Packard wafer fabrication facility operating under just such a discipline, but the same sort of input control is exercised in many other plants in a variety of industries.

In adopting a closed network model of the factory, of course, we are assuming that there is always enough demand or enough internally generated manufacturing orders to keep the factory full. That assumption is not as stringent as it may first appear: If management maintains a constant number of active jobs by restricting input *and* reducing the factory's operating hours in times of slack demand, one is led to the same closed network model, but with "elapsed time" interpreted to mean cumulative operating hours. On a different front, one may object to a closed network model of make-to-order manufacturing because it excludes consideration of delays suffered by customer orders before they are released to the factory floor. That is quite true, but it is still important to understand the performance of the factory itself, and one may expand the analysis later in an attempt to understand overall response times seen by customers.

Our method for steady-state analysis of closed networks is precisely analogous to one described in [18] and [10], called *the QNET method*, for two-moment analysis of *open* networks. That same name will be used here, so it becomes necessary to distinguish between QNET analysis of open systems and QNET analysis of closed systems. In both settings, the QNET method is based on an algorithm for computing the stationary distribution of an approximate Brownian system model, and the motivation for the Brownian approximation comes from heavy traffic theory. In the case of a closed network, this means that one expects the QNET method to work best when the job population is large, but the few numerical examples presented in this paper suggest that good results can also be expected with moderate population sizes.

In their original description of the QNET method for open networks, Harrison and Nguyen [18] allowed the various job classes served at a given node or station to have different service time distributions. In a later paper, however, Dai and Wang [12] showed that the Brownian approximation proposed in [18] makes no sense for certain multiclass networks with feedback routing. This perplexing state of affairs was discussed in the recent survey by Harrison and Hguyen [19], and Whitt [39] has thrown further light on "pathologies" that can occur with feedback routing, but much remains to be understood at the time of this writing. To avoid foundational problems in the current work, we will impose the following critical restriction: Each node or station of the network is characterized by a single service time distribution, which is common to the various job classes served there. This restriction does rule out some interesting applications, such as metal-cutting operations where jobs following different routes may have very different processing times at a particular machining center, but our model is still very general by conventional standards.

For a closed network with J stations, our approximating Brownian model is essentially a "reflected" or "regulated" Brownian motion (RBM) whose state space is a J-dimensional simplex. Earlier work by Chen and Mandelbaum [7] and by Harrison, Williams and Chen [24] also identified RBM in a simplex as the appropriate Brownian model of a closed queueing network, but those papers considered only "generalized Jackson networks," in which a single job class is served at each station. By expanding the scope of the analysis to include multiclass networks, which is necessary to get a realistic representation of job routing, one encounters a more general kind of RBM.

Also, in the earlier work referred to previously, no means was given for deriving numerical performance estimates from the Brownian approximation, whereas we develop an algorithm for practical performance analysis. On the other side of the ledger, Chen and Mandelbaum [7] were able to justify their Brownian approximation by a rigorous heavy traffic limit theorem, but that limit theory has not been extended yet to the class of networks treated here.

System performance analysis, whether based on queueing models or on computer simulation, is more highly developed and more broadly accepted in computer systems engineering than in manufacturing, and closed network models are generally considered to be more important in the computer systems domain than open network models [28, 29, 35]. Closed network models of computer systems are typically justified by the fact that user populations are literally fixed, and if one has several fixed and distinct user populations, all competing for common processing resources, the appropriate queueing network model is a so-called *multichain closed network*. As we have explained earlier, closed network models arise in manufacturing primarily because of management policies that maintain a fixed population size, not because "customers" literally circulate forever. This suggests a type of closed network model that is different, and in certain ways simpler, than those commonly used in computer performance analysis. In other words, our definition of a closed queueing network is chosen with manufacturing applications in mind, and our objective is to develop a method for two-moment performance analysis of that particular network class.

The paper is organized as follows. Section 2 describes and motivates the foregoing multiclass closed network model. An approximating Brownian system model is developed by stages in Sections 3–6, assuming that readers are familiar with the corresponding analysis of open networks in [18]. Sections 7 and 8 explain how numerical performance estimates are derived from the approximating Brownian model, with technical details relegated to an Appendix. The complete procedure for practical performance analysis is summarized in Section 9, and to clarify connections with previous work, we explain in Section 10 how our method applies to network models of the generalized Jackson type.

Three numerical examples are discussed in Sections 11–13. To be more precise, three *families* of examples are examined in those sections, where we discuss not only the accuracy of our approximation procedures, but also qualitative phenomena. Section 14 explains how QNET analysis can be extended to accommodate multiserver stations, unreliable servers and non-FIFO service disciplines, all of which are important in manufacturing applications. Finally, Section 15 contains a description of an open problem and some miscellaneous concluding remarks.

Readers who wish to get a quick overview may first read Section 2, then jump to Section 9 for a summary of QNET mechanics, and then peruse the examples in Sections 11–13.

2. A multiclass closed network model.

Our basic closed network model, which is closely related to the "structured open network" described in Section 5 of [18], is portrayed schematically in Figure 1. The notation and terminology introduced here are mostly repeated from Section 5 of [18].

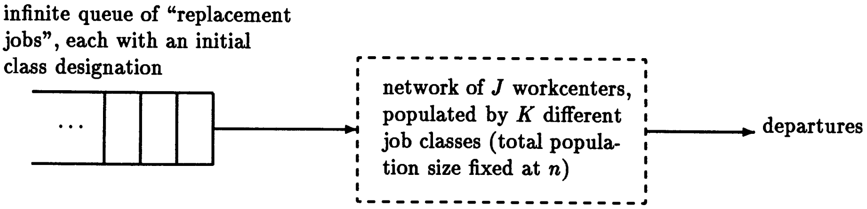
infinite queue of "replacement jobs", each with an initial class designation
...
network of J workcenters, populated by K different job classes (total population size fixed at n)
departures

FIG. 1. *Schematic representation of the closed network model.*

Workstations are indexed by $i, j = 1, \ldots, J$ and, as explained earlier, there is a single service time distribution at each station. That is, service times at station i are independent and identically distributed with a given general distribution, and the service time sequences at different stations are mutually independent as well. Let τ_i and b_i be the mean and the coefficient of variation, respectively, of service times at station i. (The coefficient of variation for a random variable is its standard deviation divided by its mean.) For the time being, each workstation is assumed to have a single server and a first-in–first-out (FIFO) service discipline. Also, servers are assumed to be perfectly reliable, but the relaxation of these restrictions will be discussed in Section 14.

Job classes are indexed by $k, l = 1, \ldots, K$. One should think of a job's class designation as a complete summary of that job's identity and past processing history. In a manufacturing system where jobs represent production lots, for example, the class designation might tell what product is being made, how many units are in the lot, which operations have already been performed and how many and what types of rework operations have been required thus far. This example suggests that the required number K of job classes is huge, which may be true in principle, but it is not actually necessary to list all potential classes, because the data that ultimately describe our approximating Brownian system model are a relatively small number of aggregate quantities.

Jobs of class k require service at some particular station $s(k)$. The probability that a class k job, upon completion of service at station $s(k)$, will turn next into a class l job is P_{kl}, and the probability that a class k job will exit the system upon completion of service is $1 - \Sigma_{l=1}^{K} P_{kl}$, independent of all past processing history. The $K \times K$ Markov switching matrix $P = (P_{kl})$ is assumed to be transient, which means simply that all jobs eventually leave the system. In our earlier example involving production lots at various stages of completion, the great majority of the switching probabilities P_{kl} would be zeros, the remainder would be mostly ones and a few intermediate values (that is, true probabilistic switching) would occur to represent the possibility of rework, scrapping or breakage of units within a lot, and so forth. Because the number of classes can be arbitrarily large, with a many-to-one mapping of classes to workstations, one can easily accommodate multiproduct manufacturing and complex product routes within the structure of our multiclass network. See Section 5 of [18] for more discussion of that point.

We define a $J \times K$ *constituency matrix* $C = (C_{jk})$ by setting $C_{jk} = 1$ if $s(k) = j$ and $C_{jk} = 0$ otherwise. That is, $C_{jk} = 1$ if job class k is served at station j, and $C_{jk} = 0$ otherwise. As in Section 5 of [18], define a $K \times K$ matrix (here and later, primes denote transposes):

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

Also, specializing the notation in Section 5 of [18], let T be the $K \times K$ diagonal matrix with

$$T_{kk} = \tau_j \quad \text{for all } k \text{ such that } s(k) = j. \tag{2.1}$$

Finally, for each $m = 1, \dots, K$, define a $K \times K$ covariance matrix $H^m = (H^m_{kl})$ via

$$H^m_{kl} = \begin{cases} P_{mk}(1 - P_{mk}), & \text{if } k = l, \\ -P_{mk}P_{ml}, & \text{if } k \neq l. \end{cases} \tag{2.2}$$

It remains to specify the input mechanism for our closed network model. As Figure 1 suggests, we imagine that there is an ordered queue of "replacement jobs" outside the processing system itself, and that the job at the head of this line enters the system each time an active job completes its route. Each replacement job has an initial class designation that specifies, for example, the quantity of product to be made and the required sequence of operations. Taken as primitive in our model is the joint distribution of initial class designations for jobs in the replacement queue. That is, the characteristics of replacement jobs may be deterministically prespecified, or they may depend on chance factors exogenous to the processing system itself, but they may not depend on events internal to the processing system. This means, for example, that we exclude from consideration systems where the choice of job to be released next depends on which workstations are currently least congested. In other words, the *sequence* of jobs to enter the system is exogenously specified, either deterministically or probabilistically, but the timing of those releases is endogenously determined by our closed-loop input control rule.

To capture this notion mathematically, we take as given a K-dimensional stochastic process $\zeta = \{\zeta(r),\ r = 1, 2, \dots\}$ whose kth component process is

$$\zeta_k(r) = \text{the number of the first } r \text{ replacement jobs that are of class } k.$$

To simplify notation in the future, we extend ζ to a continuous time parameter in a piecewise constant, right-continuous fashion, which means that $\zeta(t) = \zeta(r)$ if $r \le t < r + 1$. It is assumed that there exists a K-vector α and a $K \times K$ covariance matrix Δ such that

$$E[\zeta(t)] \sim \alpha t \quad \text{and} \quad \mathrm{Cov}[\zeta(t)] \sim \Delta t \quad \text{as } t \to \infty. \tag{2.3}$$

One interprets α_k as the long-run fraction of new jobs that are of class k, and the covariance matrix Δ describes the statistical variability in our cumulative input process ζ. The requirement that there exist such an asymptotic mean vector α and asymptotic covariance matrix Δ is a very weak stability assumption. Two input schemes of particular interest are deterministic, cyclic input and independent, identically distributed (iid) input. In the former case, the class designations of successive "replacement jobs" repeat in a deterministic cycle, such as $1, 2, 1, 3, 1, 2, 1, 3$ and so forth. The input fractions α_k are computed from cycle data in the obvious way (in our example, $\alpha_1 = 0.50$, $\alpha_2 = 0.25$, $\alpha_3 = 0.25$), and the asymptotic covariance matrix is $\Delta = 0$. In the iid case, the class designations of successive replacement jobs are independent and identically distributed according to a given probability vector α. Then $\zeta(r)$ has a multinomial distribution for each fixed r, implying that $\mathrm{Cov}[\zeta(r)] = \Delta r$, where

$$(2.4) \qquad \Delta_{kl} = \begin{cases} \alpha_k(1-\alpha_k), & \text{if } k = l, \\ -\alpha_k \alpha_l, & \text{if } k \neq l. \end{cases}$$

To justify our approximating Brownian system model, we must also assume the following functional central limit theorem (FCLT) for ζ: The centered and scaled processes $\{r^{-1/2}[\zeta(rt) - \alpha rt],\ t \geq 0\}$, indexed by $r = 1, 2, \dots,$ converge weakly as $r \to \infty$ to a Brownian motion with zero drift and covariance matrix Δ. By conventional standards this is also a very weak assumption. It is satisfied, for example, in the case of deterministic, cyclic input and the case of iid input (examples of both types are presented in Sections 11–13).

The cumulative input process ζ is assumed to be independent of the service times and class switches that occur within the processing system. Finally, as shown in Figure 1, we denote by n the fixed number of jobs within the processing system; that is, the constant population size for our closed network model.

Our formulation is general enough to include systems where jobs or customers circulate forever, but only under certain restrictions. Consider, for example, a closed network with a fixed population and K user classes. The switching matrix P is stochastic in this case; that is, each row sums to 1. Let us suppose that P is also irreducible, which means that all user classes "communicate" with one another, and hence all users eventually visit all class designations. Such a network can be mapped into our formulation by the following artifice. Choose one class k arbitrarily and designate this the "exit class." A user who completes service in class k is deemed to have finished a "route," and a new route begins with whatever class the user may visit next. Because of the irreducibility assumption, all routes terminate after a finite number of operations, and the probability that a new route begins with a visit to class l is $\alpha_l = P_{kl}$. With this particular setup, the class designations of successive replacement jobs are iid.

Unfortunately, the same trick does not work when P is reducible. In that case the job population can be divided into two or more subpopulations that never intermingle. That is, each user stays within its initial subpopulation, and one obtains a multichain closed network of the type alluded to in Section 1. Such a model cannot be mapped into our formulation, because if one designates an "exit class" for each subpopulation, one finds that the probability distribution of a replacement job's initial class designation depends on what has been happening in the network, and that does not conform to our setup.

Although they are important in the computer systems domain, we believe that multichain closed networks are relatively unimportant for manufacturing applications. However, they are not completely irrelevant, as the following example demonstrates. Consider an automated machining facility that produces both transmission housings and clutch housings. Products in the two families are made from castings of different sizes, and hence they are mounted on different types of pallets while on the factory floor. Thus a transmission housing must be replaced by another transmission housing when it finishes its route, and similarly for clutch housings, so one is led to a multichain closed network model, which unfortunately cannot be accommodated within our formulation. Consider again a traditional closed network where jobs circulate perpetually and switch class according to an irreducible stochastic matrix P. Our closed network model, with its infinite queue of replacement jobs, is more general, because it allows for such phenomena as deterministic, cyclic input schedules in multiproduct manufacturing. The example discussed in Section 11 will show that random input and deterministic, cyclic input produce substantially different system performance.

3. Brownian models for a family of open networks. To facilitate development of our approximating Brownian system model, let us consider first an *open* network that is identical to the system described in Section 2, except that new jobs from the "replacement queue" enter at fixed intervals of length $1/a$, where $a > 0$. The letter a is mnemonic for "arrival rate," and this parameter indexes a *family* of open network models of the type considered in [18]. In that earlier paper we focused on the J-dimensional *total workload process* $Z(t)$ and the J-dimensional *immediate workload process* $W(t)$, the definitions of which are as follows. For each $j = 1, \dots, J$ and each $t \geq 0$, let $W_j(t)$ denote the sum of the impending service times for jobs that are queued at station j at time t, plus the remaining service time for any job that may be in service there at time t. If there were a new arrival to station j at time t, that job would have to wait $W_j(t)$ time units before gaining access to the server, so one could also describe $W_j(t)$ as the *virtual waiting time process* for station j. Next, let $Z_j(t)$ denote the sum of all future service times at station j for jobs that are present anywhere in the network at time t, plus the remaining service time of any job that may be in service at station j at time t. If there were no new arrivals to the network after time t, then $Z_j(t)$ would represent the total amount of work required from server j to empty the system. In the current context, it is more natural to focus on $Z(t)$ and a J-dimensional *jobcount process* $N(t)$ whose jth component is

$$N_j(t) = \text{the total number of jobs at station } j \text{ at time } t, \text{ regardless of class.}$$

In these verbal definitions, of course, we are imagining that all jobs belonging to a class k in the constituency of server j [that is, all jobs of a class k such that $s(k) = j$] physically "occupy" station j as they await service.

If one is concerned only with Brownian approximations, the substitution of the process $N(t)$ for $W(t)$ in one's system description hardly changes anything, because in a Brownian system model, jobcount and immediate workload are deterministically linked by the relationship

$$W_j(t) = \tau_j N_j(t). \tag{3.1}$$

(Recall that τ_j is the mean service time at station j.) Now let us define a K-vector λ and a J-vector ρ via

$$\lambda = Q\alpha \quad \text{and} \quad \rho = CT\lambda. \tag{3.2}$$

Recalling that $Q = (I - P')^{-1}$, one sees that λ satisfies the "traffic equation" $\lambda = \alpha + P'\lambda$, so λ_k is the average rate at which class k jobs "arrive" at station $s(k)$ when the external arrival rate is $a = 1$. Thus the vector of class-level arrival rates for a general value of a is $a\lambda$. Similarly, ρ_j represents the total hours of work for server j entering the system per hour if $a = 1$, and for a general value of a, the vector of overall "traffic intensity parameters" (one component for each workstation) is $a\rho$. Hereafter, attention is restricted to a values such that $a\rho_j < 1$ for each station j. If we now define

$$\gamma = C\lambda, \tag{3.3}$$

then γ_j is the total arrival rate to station j, regardless of job class, when $a = 1$.

Imitating the notation used in [18], let us denote by $L_j(r)$ the sum of all the service times at station j required by the first r jobs in the replacement queue. Also, define a J-dimensional vector process $L(r)$ in the obvious way, and extend $L(\cdot)$ to a continuous-time-parameter process $\{L(t),\ t \geq 0\}$ in a right-continuous, piecewise constant fashion. If the external arrival rate is $a = 1$, then $L_j(t)$ represents the total amount of work for server j entering the system before time t, and by specializing the results developed in Section 5 of [18], one has

$$E[L(t)] \sim \rho t \quad \text{and} \quad \operatorname{Cov}[L(t)] \sim \Gamma t \quad \text{as } t \to \infty, \tag{3.4}$$

where $\rho = CT\lambda$ as before and

$$\Gamma = (CTQ)G(CTQ)', \tag{3.5}$$

$$G = \Delta + H + (I - P')D(I - P), \tag{3.6}$$

$$H = \sum_{m=1}^{K} \lambda_m H^m, \tag{3.7}$$

$$D = \operatorname{diag}\left(\lambda_1 b_{s(1)}^2, \ldots, \lambda_K b_{s(K)}^2\right). \tag{3.8}$$

Actually, in [18] we assumed that the exogenous inputs to the various classes were independent renewal processes, which gave Δ a certain special structure, but the argument is virtually unchanged with a general input process. With that generalization, our formula (3.6) for G is identical to formula (52) of [18], but readers should be warned that formula (40) of [18], which was said to give an equivalent specification of G, is only valid if one assumes $P_{kk} = 0$ for all k, and we do not wish to impose that restriction in this paper.

To avoid trivial complications, we will assume as in [18] that Γ has full rank J. That is, we assume that there is some randomness in the workload input to each station, either because of service time variability or because of routing variability, or because of variability in the initial job class designations. Moreover, given the FCLT that we have assumed for ζ, plus our various independence assumptions, it is easy to establish the following FCLT for L. Defining

$$\hat{L}(t) = L(t) - \rho t, \qquad t > 0, \tag{3.9}$$

one can show that the scaled processes $\{r^{-1/2}\hat{L}(rt),\ t \geq 0\}$, indexed by $r = 1, 2, \ldots$, converge weakly as $r \to \infty$ to a Brownian motion with drift zero and covariance Γ. As a final preliminary, let us define $J \times J$ matrices F and M via

$$F = C(TQ\Lambda)C' \quad \text{and} \quad M_{ij} = F_{ij}/\gamma_j, \tag{3.10}$$

where $\Lambda = \operatorname{diag}(\lambda)$. Our definition of F is exactly as in [18], but that of M is slightly different, for reasons to be explained shortly.

In [18] we denoted by $\xi(t)$ the *total workload netflow process* for an open network. In the current context, with a general arrival rate a, this is given by $\xi(t) = L(at) - et$, where e is the J-vector of ones. That expression can be rewritten as

$$\xi(t) = \hat{L}(at) - (e - a\rho)t \tag{3.11}$$

and, given the FCLT for $\hat{L}$ stated previously, it is natural to approximate ξ as in (3.13). Let $I_j(t)$ be the cumulative idleness suffered by server j up to time t, and define a J-dimensional vector process $I(t)$ in the obvious way. The Brownian approximation for the open network that was developed in [18] is defined by the following six relationships:

(3.12) $Z(t) = Z(0) + \xi(t) + I(t)$,

(3.13) ξ is a Brownian motion with drift vector $-(e - a\rho)$ and covariance matrix $a\Gamma$,

(3.14) $I(\cdot)$ is continuous and increasing with $I(0) = 0$,

(3.15) $I_j(\cdot)$ only increases at times t when $N_j(t) = 0$,

(3.16) $Z(t) = MN(t)$,

(3.17) $N(t) \geq 0$.

Actually, the Brownian system model proposed in [18] differed from (3.12)–(3.17) in two regards: $W(t)$ appeared in place of $N(t)$ in the analog of (3.16), and the matrix M was defined via $M_{ij} = F_{ij}/\rho_j$, whereas (3.10) specifies $M_{ij} = F_{ij}/\gamma_j$. From (3.1) and (3.3), one sees that the two formulations are in fact equivalent.

As explained in [19], the approximating Brownian model for our multiclass open network is uniquely determined by (3.12)–(3.17) if and only if M is invertible and its inverse is *completely-$\mathscr{S}$*, whose definition follows. For future reference, we also define a class of "admissible" matrices and prove that an admissible matrix must be a completely-$\mathscr{S}$ matrix.

DEFINITION 3.1. A $J \times J$ matrix A is said to be an $\mathscr{S}$ matrix if there exists a J-dimensional vector $u \geq 0$ such that $Au > 0$, and to be a completely-$\mathscr{S}$ matrix if each of its principal submatrices is an $\mathscr{S}$ matrix.

DEFINITION 3.2. A $J \times J$ matrix A is said to be admissible if there is a positive diagonal matrix D such that

$$\Gamma^* \equiv AD + DA' \tag{3.18}$$

is positive definite.

The following lemma is used several times in this paper.

LEMMA 3.1. *If A is admissible, then A is completely-$\mathscr{S}$.*

PROOF. This lemma can be proved by quoting existing known results. First, by Theorem 2.3 of Berman and Plemmons ([3], page 134), an admissible matrix must be a $\mathscr{P}$ matrix. Next, it follows from the discussion in Section 2 of Mandelbaum [30] that a $\mathscr{P}$ matrix is a completely-$\mathscr{S}$ matrix. Here we provide a direct proof that seems to be new.

Assume that A is admissible. We first show that A is an $\mathscr{S}$ matrix. Let D and Γ^* be the matrices defined in (3.18). It suffices to prove that

$$R \equiv AD$$

is an $\mathscr{S}$ matrix.

Let $S^0 = \{x \in \mathbb{R}^j \colon x_i > 0 \ (i = 1, \dots, J)\}$ and $S^* = RS^0$. Then both S^0 and S^* are convex open cones in $\mathbb{R}^J$. If $S^0 \cap S^* \neq \emptyset$, then it is obvious that R is an $\mathscr{S}$ matrix. Assume on the contrary that $S^0 \cap S^* = \emptyset$. Because S^0 and S^* are two disjoint convex sets in $\mathbb{R}^J$, there is a hyperplane K separating S^0 and S^*. Let $\{x \colon n \cdot x = 0\}$ be the equation determining the hyperplane K, where n is the normal direction of K. Because S^0 lies entirely on one side of K, the inner product $\langle n, y \rangle$ takes one sign for all $y \in S^0$. Without loss of generality, we assume that $\langle n, y \rangle \geq 0$ for all $y \in S^0$ and, therefore, by the same argument, $\langle n, y \rangle \leq 0$ for all $y \in S^*$. By the continuity of the inner product function $y \to \langle n, y \rangle$, $\langle n, y \rangle \geq 0$ for all $y \in \overline{S}$, where $\overline{S}$ is the closure of S^0. Similarly, $\langle n, y \rangle \leq 0$ for all y in $R\overline{S}$, the closure of S^*. Because $\langle n, y \rangle \geq 0$ for all $y \in \overline{S}$, we have $n \in \overline{S}$ and hence Rn is in $R\overline{S}$. Therefore, $\langle n, Rn \rangle \leq 0$, contradicting $n'Rn = \frac{1}{2}n'\Gamma^* n > 0$. Thus R is an $\mathscr{S}$ matrix. Applying the same reasoning to any principal submatrix of A, we can prove that A is completely-$\mathscr{S}$. □

PROPOSITION 3.1. *Let matrix M be defined as in* (3.10). *Then M is invertible and M^{-1} is completely-$\mathscr{S}$.*

PROOF. First suppose that M is admissible. Then for any $x \neq 0$, $2x'MDx = x'\Gamma^* x > 0$, where D and Γ^* are defined in (3.18). Thus MD is invertible and, therefore, M is invertible. It is easy to check that the property of being admissible is closed under matrix inversion. Thus M^{-1} is admissible and, therefore, from Lemma 3.1 M^{-1} is completely-$\mathscr{S}$.

The remainder of this proof is devoted to showing that M is admissible. First recall that τ_i is the mean service time at station i $(i = 1, \dots, J)$, and $T = \operatorname{diag}(\tau_{s(1)}, \dots, \tau_{s(K)})$ is the diagonal matrix defined in (2.1). Let $\overline{T} = \operatorname{diag}(\tau_1, \dots, \tau_J)$. One can check that $CT = \overline{T}C$ and, therefore, $F = \overline{T}CQ\Lambda C'$. Next we show that

$$\Lambda(I - P) + (I - P')\Lambda = G^*, \tag{3.19}$$

where

$$G^* = \operatorname{diag}(\alpha) + H + (I - P')\Lambda(I - P),$$

with H being defined as in (3.7), and α_i's being external arrival rates. [This matrix G^* is what one would get from (3.6) if ζ were a vector of independent Poisson inputs and $b_i = 1$ for all $i = 1, \dots, J$.] In fact, it is easy to check by some simple algebra that (3.19) is equivalent to

$$\Lambda = H + \operatorname{diag}(\alpha) + P'\Lambda P. \tag{3.20}$$

To prove (3.20), we note that

$$H^k = \operatorname{diag}(P_k') - P_k' P_k,$$

where P_k, as before, is the kth row of P. Therefore,

$$\begin{aligned} H &= \sum_{k=1}^{K} \lambda_k H^k \\ &= \sum_{k=1}^{K} \lambda_k \operatorname{diag}(P_k') - \sum_{k=1}^{K} \lambda_k P_k' P_k \\ &= \operatorname{diag}\left(\sum_{k=1}^{K} \lambda_k P_k' \right) - P'\Lambda P \\ &= \operatorname{diag}(\lambda - \alpha) - P'\Lambda P \\ &= \operatorname{diag}(\lambda) - \operatorname{diag}(\alpha) - P'\Lambda P, \end{aligned}$$

where, in the next-to-last equation, we have used the traffic equation (3.2). It follows that (3.20) and hence (3.19) holds.

It follows from (3.19) that G^* is a positive definite matrix. Premultiplying by CQ and postmultiplying by $Q'C'$ on the both sides of (3.19), we have that

$$C(Q\Lambda + \Lambda Q')C' = CQ\Lambda C' + C\Lambda Q'C' = CQG^* Q'C'.$$

It follows that

$$\overline{T}CQ\Lambda C'\overline{T} + \overline{T}C\Lambda Q'C'\overline{T} = \overline{T}CQG^* Q'C'\overline{T}.$$

Because $M = F[\mathrm{diag}(\gamma)]^{-1} = \overline{T}CQ\Lambda C'[\mathrm{diag}(\gamma)]^{-1}$, we have

$$M\,\mathrm{diag}(\gamma)\overline{T} + \overline{T}\,\mathrm{diag}(\gamma)M' = \overline{T}CQG^{*}Q'C'\overline{T} \equiv \Gamma^{*}. \tag{3.21}$$

Because the constituency matrix C has the full rank J, we have $CQG^{*}Q'C'$ positive definite, thus concluding that $\Gamma^{*} = \overline{T}CQG^{*}Q'C'\overline{T}$ is positive definite. □

Defining

$$R = M^{-1}, \tag{3.22}$$

one can now proceed exactly as in Section 4 of [18] to eliminate $Z(t)$ from (3.12)–(3.17), eventually arriving at the following more compact representation of the Brownian model:

(3.23) $N(t) = N(0) + X(t) + RI(t)$,

(3.24) $X = R\xi$ is a Brownian motion with drift vector $\mu = -R(e - a\rho)$ and covariance matrix $a\Omega$, where $\Omega = R\Gamma R'$,

(3.25) $N(t) \ge 0$,

(3.26) $I(\cdot)$ is contionuous and increasing with $I(0) = 0$,

(3.27) $I_j(\cdot)$ increases only at times t when $N_j(t) = 0$.

4. The Brownian model of a multiclass closed network. Returning to the multiclass closed network described in Section 2, we assume that the initial population size is $e'N(0) = n$ (again e denotes the J-vector of ones) and that new jobs are injected so as to keep the total population size fixed. Denoting by $A(t)$ the total number of replacement jobs injected up to time t, there presumably exists a long-run average *throughput rate* $a > 0$ such that

$$E[A(t)] \sim at \quad \text{as } t \to \infty. \tag{4.1}$$

A fundamental purpose of system performance analysis is to determine the throughput rate a, but for the moment let us treat it as a known constant and define

$$Y(t) = A(t) - at, \qquad t \ge 0. \tag{4.2}$$

Thus $Y(\cdot)$ is a nonmonotone, one-dimensional process that describes fluctuations of the cumulative input $A(\cdot)$ around its central tendency. The workload netflow process seen by the various workstations under closed-loop input control can then be expressed as (here and later, tildes are used to signify processes associated with the closed network model)

$$\begin{aligned}\tilde{\xi}(t) &= L(A(t)) - et = \hat{L}(A(t)) + \rho A(t) - et\\ &= \hat{L}(A(t)) - (e - a\rho)t + \rho Y(t).\end{aligned} \tag{4.3}$$

Given a so-called functional version of (4.1), plus the FCLT for $\hat{L}(\cdot)$ stated earlier, one can show that the scaled processes $\{r^{-1/2}\hat{L}(A(rt)),\ t \ge 0\}$, indexed by $r = 1, 2, \dots$, converge weakly as $r \to \infty$ to a Brownian motion with zero drift and covariance $a\Gamma$, and thus we are led to approximate the first two terms on the right side of (4.3) by a Brownian motion ξ as in (4.5). As analogs of (3.12)–(3.17), one then arrives at (4.4)–(4.9), and (4.10) is the distinguishing additional feature of a closed network model:

(4.4) $\tilde{Z}(t) = \tilde{Z}(0) + \xi(t) + \tilde{I}(t) + \rho\tilde{Y}(t)$,

(4.5) ξ is a Brownian motion with drift vector $-(e - a\rho)$ and covariance matrix $a\Gamma$,

(4.6) $\tilde{I}(\cdot)$ is continuous and increasing with $\tilde{I}(0) = 0$,

(4.7) $\tilde{I}_j(\cdot)$ increases only at times t when $\tilde{N}_j(t) = 0$,

(4.8) $\tilde{Z}(t) = M\tilde{N}(t)$,

(4.9) $\tilde{N}(t) \geq 0$,

(4.10) $e'\tilde{N}(t) = n$.

It is not obvious, of course, that (4.4)–(4.10) uniquely determine the processes $\tilde{Z}$, $\tilde{I}$, $\tilde{N}$ and $\tilde{Y}$, but in the remainder of this section and in Section 5 we will show that to be the case.

Proceeding exactly as in Section 3, we can now simplify the Brownian model (4.4)–(4.10) by eliminating $\tilde{Z}$. Defining $R = M^{-1}$ as before and

$$u = R\rho, \tag{4.11}$$

one can premultiply (4.4) by R and then equate $R\tilde{Z}(t)$ to $\tilde{N}(t)$ because of (4.8), to obtain

$$\tilde{N}(t) = \tilde{N}(0) + X(t) + R\tilde{I}(t) + u\tilde{Y}(t), \tag{4.12}$$

where

(4.13) $X(t) = R\xi(t)$ is a Brownian motion with drift vector $-R(e - a\rho)$ and covariance matrix $a\Omega$, where $\Omega = R\Gamma R'$.

Equation (4.12) is analogous to (3.23), and (4.13) is identical to our earlier definition (3.24) of the Brownian motion X. To summarize, the Brownian approximation for our closed network model is defined by (4.12), (4.13), (4.6), (4.7), (4.9) and (4.10). Equations (4.4) and (4.8) have been incorporated into (4.12) and thus will not be needed hereafter.

5. Reduction to RBM in a simplex.

We continue to treat the throughput rate a, which appears in our Brownian system model through (4.13), as if it were a known constant. To further simplify the Brownian model, premultiply (4.12) by e' and then set $e'\tilde{N}(t) = e'\tilde{N}(0) = n$ because of (4.10), to obtain

$$e'X(t) + e'R\tilde{I}(t) + c\tilde{Y}(t) = 0, \tag{5.1}$$

where

$$c = e'u = e'R\rho. \tag{5.2}$$

One can show that $c > 0$. In fact, noting that $\mathrm{diag}(\gamma)\overline{T} = \mathrm{diag}(\rho)$ and (3.21), one has

$$R\,\mathrm{diag}(\rho) + \mathrm{diag}(\rho)R' = R\Gamma^* R'. \tag{5.3}$$

Premultiplying (5.3) by e' and postmultiplying by e, we have

$$2c = 2e'R\rho = 2e'R\,\mathrm{diag}(\rho)e = e'R\Gamma^* R'e > 0,$$

because $R\Gamma^* R'$ is positive definite. Of course, one can use (5.1) to solve for $\tilde{Y}(t)$:

$$\tilde{Y}(t) = -\frac{1}{c}\big[e'X(t) + e'R\tilde{I}(t)\big]. \tag{5.4}$$

When (5.4) is substituted into (4.12), one obtains yet another simplification of the Brownian system model. To express the result compactly, let us define the $J \times J$ matrices

$$B = I - \frac{1}{c}ue' \quad \text{and} \quad \tilde{R} = BR. \tag{5.5}$$

From (5.5) and the definition (5.2) of c, it follows that

$$B \text{ is of rank } J - 1 \quad \text{and} \quad e'B = 0. \tag{5.6}$$

Recalling that R is invertible, one then has from (5.5) and (5.6) that

$$\tilde{R} \text{ is of rank } J - 1 \quad \text{and} \quad e'\tilde{R} = 0. \tag{5.7}$$

Also, using (4.11), (5.2) and (5.5), readers may easily verify that

$$\tilde{R}\rho = 0. \tag{5.8}$$

By substituting (5.4) into (4.12) and then using (5.5), one obtains

$$\tilde{N}(t) = \tilde{N}(0) + \tilde{X}(t) + \tilde{R}\tilde{I}(t), \tag{5.9}$$

where $\tilde{X} = BX = \tilde{R}\xi$ is a Brownian motion with drift vector $-\tilde{R}(e - a\rho)$ and covariance matrix $a\tilde{R}\Gamma\tilde{R}'$. However, (5.8) says that $\tilde{R}\rho = 0$, and thus

$$\begin{gathered}\tilde{X} \text{ is a Brownian motion with drift vector } \tilde{\mu} = -\tilde{R}e \text{ and} \\ \text{covariance matrix } a\tilde{\Omega}, \text{ where } \tilde{\Omega} = \tilde{R}\Gamma\tilde{R}'.\end{gathered} \tag{5.10}$$

To completely specify our Brownian approximation for the closed network model, one still needs conditions (4.6), (4.7) and (4.9), and for ease of reference we repeat those conditions:

(5.11) $\tilde{N}(t) \geq 0$,

(5.12) $\tilde{I}(\cdot)$ is continuous and increasing with $\tilde{I}(0) = 0$,

(5.13) $\tilde{I}_j(\cdot)$ increases only at times t when $\tilde{N}_j(t) = 0$.

To review, our original seven conditions (4.4)–(4.10) that defined the Brownian system model have been reduced to the equivalent five conditions (5.9)–(5.13), first by using (4.8) to eliminate $\tilde{Z}(t)$ from the system of equations and then by using (4.10) to eliminate $\tilde{Y}(t)$. It will be proved in the next section that $\tilde{N}$ is uniquely determined by (5.9)–(5.13), which defines $\tilde{N}$ as a reflected or regulated Brownian motion (RBM) whose state space is the simplex

$$\tilde{S}(n) = \{x \in \mathbb{R}^J\colon x \geq 0 \text{ and } e'x = n\}.$$

Let us consider its sample path behavior. Recall from (5.7) that $e'\tilde{R} = 0$, or equivalently, each column of $\tilde{R}$ lies in the hyperplane $H = \{x \in \mathbb{R}^J\colon e'x = 0\}$. It then follows from (5.10) that the one-dimensional Brownian motion $e'\tilde{X}$ has zero drift and zero variance, which is to say that sample paths of $\tilde{X}$ lie in the hyperplane H. Thus, because we assume that $e'\tilde{N}(0) = n$, the sample path of $\tilde{N}$ remains in the simplex $\tilde{S}(n)$, at least until the boundary of $\tilde{S}(n)$ is hit. According to (5.13), the cumulative idleness process $\tilde{I}_j$ increases only when the boundary surface $\{\tilde{N}_j = 0\}$ is hit, and according to (5.9), each such increase "pushes" $\tilde{N}(\cdot)$ in a direction given by the jth column of $\tilde{R}$; the magnitude of that "push" is the minimal amount required to assure that $\tilde{N}_j(t) \geq 0$ for all t. Because each column of $\tilde{R}$ lies in the hyperplane H, it follows that $\tilde{N}$ remains always within the simplex $\tilde{S}(n)$.

In this section and its predecessor, we have developed a Brownian system model that approximates the multiclass closed network described in Section 2, treating the throughput rate α as if it were a known constant. In Section 7, again taking α to be given, we will explain how one can compute steady-state quantities associated with the jobcount process $\tilde{N}$ defined by (5.9)–(5.13). Finally, in Section 8, a general method will be described for approximate steady-state analysis of the original multiclass closed queueing network, where α is actually an output from the analysis rather than an input to it.

6. RBM in a simplex. As discussed in the previous section, the jobcount process $\tilde{N}$ will be approximated by an RBM in the simplex $\tilde{S} = \tilde{S}(n)$. In its most compact form, our Brownian model of a multiclass closed network is specified by the five equations, definitions and auxiliary conditions numbered (5.9)–(5.13). One must show that the Brownian system model is well posed, which amounts to showing that the process $\tilde{N}$ exists and is unique in distribution. Our formal definition of an RBM is analogous to the one given in Dai and Williams [13]. Because relations (5.9)–(5.13) capture the main ingredients in the definition, we do not repeat their formal mathematical definition here. In order to establish existence and uniqueness, we first reduce $\tilde{N}$ to an RBM in the *solid* simplex

$$S = S(n) = \{z \in \mathbb{R}_+^{J-1} : e'z \leq n\}, \tag{6.1}$$

where e is the $(J-1)$-dimensional vector of ones. In this section, we use e to denote both the J-dimensional and the $(J-1)$-dimensional vector of ones. Denoting the projection of $\tilde{N}$ onto $\mathbb{R}_+^{J-1}$ by N, we now argue that N and $\tilde{N}$ can be constructed from each other and, therefore, the existence and uniqueness of $\tilde{N}$ is equivalent to that of N. (The letters N, I, X, R, μ and Ω were given one meaning in Section 3, and they will be reused with a different meaning here and in the Appendix, but that should cause no confusion.) Set

$$\begin{gathered} \mu = (\tilde{\mu}_i)_{1 \leq i \leq J-1}, \qquad \Omega = \left(\tilde{\Omega}_{ij}\right)_{1 \leq i, j \leq J-1} \quad \text{and} \\ R = \left(\tilde{R}_{ij}\right)_{1 \leq i \leq J-1, 1 \leq j \leq J}. \end{gathered} \tag{6.2}$$

Notice that Ω is positive definite by definitions (3.5), (5.5) and (5.10). Recall the definitions of $\tilde{X}$ and $\tilde{I}$ in (5.9)–(5.13). Let $N_j(t) = \tilde{N}_j(t)$, $X_j(t) = \tilde{X}_j(t)$ $(j = 1, \dots, J-1)$ and $I_j(t) = \tilde{I}_j(t)$ $(j = 1, \dots, J)$ for each $t \geq 0$. Then we have

$$N(t) = N(0) + X(t) + RI(t), \tag{6.3}$$

$$X(t) \text{ is a Brownian motion with drift } \mu \text{ and covariance } \Omega, \tag{6.4}$$

$$N(t) \in S, \tag{6.5}$$

$$I(\cdot) \text{ is continuous and increasing with } I(0) = 0, \tag{6.6}$$

$$\begin{gathered} I_j(\cdot) \text{ increases only at times } t \text{ when } N_j(t) = 0 \; (j = 1, \dots, J-1); \\ I_J(\cdot) \text{ increases only at times } t \text{ when } e'N(t) = n. \end{gathered} \tag{6.7}$$

Equations (6.3)–(6.7) define N as an (S, μ, Ω, R)-RBM. Conversely, suppose that there is given an RBM N, together with X and I, satisfying (6.3)–(6.7). By defining $\tilde{N}_j(t) = N_j(t)$ and $\tilde{X}_j(t) = X_j(t)$ $(j = 1, \dots, J-1)$, $\tilde{N}_J(t) = n - (N_1(t) + \cdots + N_{J-1}(t))$ and $\tilde{X}_J(t) = -(X_1(t) + \cdots + X_{J-1}(t))$ and $\tilde{I}(t) = I(t)$, then

$$\begin{aligned} \tilde{N}_J(t) &= n - e'N(t) = n - e'N(0) - e'X(t) - e'RI(t) \\ &= \tilde{N}_J(0) + \tilde{X}_J(t) + \tilde{R}_J\tilde{I}(t), \end{aligned}$$

where $\tilde{R}_J$ is the Jth row of $\tilde{R}$. Readers can easily check that $\tilde{N}$, $\tilde{X}$ and $\tilde{I}$ satisfy (5.9)–(5.13), with the minor change that $\tilde{X}$ is a Brownian motion with drift $\tilde{\mu}$ and covariance matrix $\tilde{\Omega}$. That is, $\tilde{N}$ is an $(\tilde{S}, \tilde{\mu}, \tilde{\Omega}, \tilde{R})$-RBM. The foregoing arguments show that an RBM $\tilde{N}$ in the simplex $\tilde{S}$ with data $(\tilde{S}, \tilde{\mu}, \tilde{\Omega}, \tilde{R})$ is equivalent to an RBM N in the *solid* simplex S with data (S, μ, Ω, R).

Having shown that $\tilde{N}$ is equivalent to an RBM N whose state space is the $(J-1)$-dimensional solid simplex $S(n)$, one needs to prove the following foundational result:

(6.8) the process, putatively defined by (6.3)–(6.7), exists and is unique in distribution.

Dai and Williams [13] have recently shown this to be true if and only if the "reflection matrix" R appearing in (6.3) satisfies a natural generalization of the "completely-$\mathscr{S}$" condition that appeared in the foundational work by Reiman and Williams [34] and by Taylor and Williams [38] on RBM in an orthant.

To check that the conditions in Dai and Williams [13] hold, define

$$\begin{aligned} F_j &= \{(z_1,\dots,z_{J-1}) \in S\colon z_j = 0\}, \qquad j = 1,\dots,J-1, \\ F_J &= \{z \in S\colon e'z = n\}. \end{aligned} \tag{6.9}$$

It is clear that F_j is the jth boundary piece of S. For $j = 1,\dots,J-1$, e_j is an inward normal to F_j and $-e$ is an inward normal to F_J. There are J vertices for the state space S given by $\bigcap_{j\neq i} F_j$ for $i = 1,\dots,J$. For each $j = 1,\dots,J$, let $\tilde{R}_{|j}$ be the $(J-1)\times(J-1)$ principal submatrix of $\tilde{R}$ obtained by deleting the jth row and the jth column of $\tilde{R}$. To show that the RBM associated with (S,μ,Ω,R) exists and is unique in law, we first prove the following lemma.

Lemma 6.1. *For each $j = 1,\dots,J$, $\tilde{R}_{|j}$ is admissible and hence is completely-$\mathscr{S}$.*

Proof. We first show that

$$\tilde{R}\operatorname{diag}(\rho) + \operatorname{diag}(\rho)\tilde{R}' = \tilde{R}\Gamma^*\tilde{R}', \tag{6.10}$$

where Γ^* is defined by (3.21). From (5.3) we know that

$$R\operatorname{diag}(\rho) + \operatorname{diag}(\rho)R' = R\Gamma^*R'.$$

It follows that

$$BR\operatorname{diag}(\rho)B' + B\operatorname{diag}(\rho)R'B' = BR\Gamma^*R'B',$$

or equivalently,

$$\tilde{R}\operatorname{diag}(\rho)B' + B'\operatorname{diag}(\rho)\tilde{R}' = \tilde{R}\Gamma^*\tilde{R}'. \tag{6.11}$$

Because $B = I - (1/c)ue'$,

$$\begin{aligned} \tilde{R}\operatorname{diag}(\rho)B' &= \tilde{R}\operatorname{diag}(\rho) - \frac{1}{c}\tilde{R}\operatorname{diag}(\rho)eu' \\ &= \tilde{R}\operatorname{diag}(\rho) - \frac{1}{c}\tilde{R}\rho u' \\ &= \tilde{R}\operatorname{diag}(\rho), \end{aligned}$$

where in the last equality we have used equation (5.8). Equation (6.10) follows from (6.11) immediately. Because the rank of $\tilde{R}$ is $J-1$, each $(J-1)\times(J-1)$ principal submatrix of $\tilde{R}\Gamma^*\tilde{R}'$ is positive definite. Therefore, it follows from (6.10) that $\tilde{R}_{|j}$ is admissible and hence is completely-$\mathscr{S}$ by Lemma 3.1. □

We end this section by proving that the RBM associated with data (S,μ,Ω,R) exists and is unique in law. Readers are referred to Dai and Williams [13] for the precise meaning of the statement. Let v_i be the ith column vector of R, $i=1,\dots,J$. If for each $i=1,\dots,J$, corresponding to the vertex $\bigcap_{j\neq i}F_j$, one can show that

$$\text{(6.12)} \qquad \begin{pmatrix} n'_1 \\ \vdots \\ \hat{n}_i \\ \vdots \\ n'_J \end{pmatrix} (v_1,\dots,\hat{v}_i,\dots,v_J)$$

is completely-$\mathscr{S}$, where n_i is an inward normal to F_i and the symbol $\hat{v}$ means the deletion operation for vector v, then by Theorem 1.1 of Dai and Williams [13] the RBM associated with data (S,μ,Ω,R) exists and is unique in law.

For $i=J$, expression (6.12) is equal to $\tilde{R}_{|J}$, which is completely-$\mathscr{S}$ by Lemma 6.1. For $i=1,\dots,J-1$, the matrix in (6.12) reduces to

$$\begin{pmatrix} e'_1 \\ \vdots \\ \hat{e}_i \\ \vdots \\ e'_{J-1} \\ -e' \end{pmatrix} (v_1,\dots,\hat{v}_i,\dots,v_{J-1},v_J),$$

which is equal to $\tilde{R}_{|i}$ because $e'\tilde{R}=0$. Thus by Lemma 6.1 and [13], Theorem 1.1, the RBM exists and is unique in law. That is, (6.8) has been proved.

7. Stead-state analysis of RBM in a simplex.

The process $\tilde{N}=\{\tilde{N}(t), t\geq 0\}$ defined by (5.9)–(5.13) is a reflected Brownian motion in the J-dimensional simplex $\tilde{S}(n)=\{x\in\mathbb{R}^J_+ : e'x=n\}$, with drift vector $\tilde{\mu}=-\tilde{R}e$, covariance matrix $a\tilde{\Omega}$ and reflection matrix $\tilde{R}$, where a is the unknown throughput rate of the queueing network. We have shown that such an RBM exists and is unique in the *weak* sense of Taylor and Williams [38] and Dai and Williams [13]. This section and the Appendix are focused on steady-state analysis of the RBM $\tilde{N}$. In QNET analysis of the network (see Section 8), the unknown throughput rate a will be estimated before steady-state analysis of $\tilde{N}$ is undertaken, and, therefore, the covariance matrix is simply denoted as $\tilde{\Omega}$ throughout this section and the Appendix. Recall that a probability measure $\tilde{\pi}$ on $\tilde{S}(n)$ is said to be a stationary distribution for the RBM $\tilde{N}$ if $\tilde{N}(t)$ has distribution $\tilde{\pi}$ for each $t > 0$ whenever $\tilde{N}(0)$ is randomized with distribution $\tilde{\pi}$.

PROPOSITION 7.1. *The RBM $\tilde{N}$ has a unique stationary distribution $\tilde{\pi}$, and $\tilde{\pi}$ has a density function with respect to the surface Lebesgue measure on $\tilde{S}(n)$.*

PROOF. Because the state space $\tilde{S}(n)$ is compact, the existence of a stationary distribution $\tilde{\pi}$ follows easily. See Theorem 5.3 of [24]. The uniqueness follows the arguments in Theorem 5.14 of [24]. Notice that in [24], the authors considered a special class of RBM's with reflection matrix $\tilde{R} = (I - \tilde{P}')$, where $\tilde{P}$ is an *irreducible stochastic* matrix. However, their arguments in Section 6 directly apply here. □

In the corollary to Proposition 7.3, we will show that for each $j \in \{1, \dots, J\}$, the limit

$$\lim_{t \to \infty} \frac{1}{t} E\left[\tilde{I}_j(t)\right] \equiv \tilde{\delta}_j \tag{7.1}$$

exists. The limit $\tilde{\delta}_j$ is interpreted as the long-run average idleness rate for server j in a queueing network. For the purpose of this paper, steady-state analysis of the RBM is confined to finding the stationary distribution $\tilde{\pi}$ and the vector $\tilde{\delta}$ of idleness rates.

Obviously, in order for our Brownian approximation to have any practical value, one must be able to find $\tilde{\pi}$. To that end, we reduce $\tilde{N}$ to an RBM N in the solid simplex $S = S(n)$ as described in Section 6. Because the state space S of N is compact, similar to the proof given in Theorems 5.3 and 5.14 of [24], we have the following proposition.

PROPOSITION 7.2. *The RBM N has a unique stationary distribution π, and π has a density function $p_0(x)$ with respect to Lebesgue measure dz on S.*

REMARK. It is easy to see that the stationary distribution π of N relates to the stationary distribution $\tilde{\pi}$ of $\tilde{N}$ by a simple projection: For any Borel set B in $\tilde{S}$,

$$\tilde{\pi}(B) = \int_S 1_B\left(z_1, \dots, z_{J-1}, n - \sum_{j=1}^{J-1} z_j\right) \pi(dz). \tag{7.2}$$

Before we state the basic adjoint relationship for p_0 and its associated boundary densities, let us introduce some additional notation. Define the second-order elliptic differential operator $\mathscr{G}$ via

$$\mathscr{G}f(x) = \frac{1}{2} \sum_{i=1}^{J-1} \sum_{j=1}^{J-1} \Omega_{ij} \frac{\partial^2}{\partial x_i \, \partial x_j} f(x) + \sum_{i=1}^{J-1} \mu_i \frac{\partial}{\partial x_i} f(x), \quad f \in C^2(S), \tag{7.3}$$

where $C^2(S)$ is the set of functions which, together with their first- and second-order derivatives, are continuous and bounded on S. Also, for each $j = 1, \dots, J$, define the directional derivative

$$\mathscr{D}_j f(x) = R^j \cdot \nabla f(x),$$

where R^j is the jth column of the reflection matrix R. The following proposition depends on Lemma 8.4 of [21], which can be proved by using Theorem 3.2 of Dai and Williams [13].

PROPOSITION 7.3. *For each $j \in \{1, \dots, J\}$, there exists a nonnegative integrable function p_j on F_j such that for each bounded Borel measurable function f on F_j,*

$$\lim_{t \to \infty} \frac{1}{t} E\left[\int_0^t f(N(s))\, dI_j(s)\right] = \int_{F_j} f p_j \, d\sigma_j, \tag{7.4}$$

where σ_j $(j = 1, \dots, J-1)$ is surface Lebesgue measure on F_j and σ_J is $(J-1)^{-1/2}$ times Lebesgue measure on F_J. Furthermore, $p \equiv (p_0, p_1, \dots, p_J)$ satisfies the following basic adjoint relationship:

$$\int_S \mathscr{G} f p_0 \, dz + \sum_{j=1}^{J} \int_{F_j} \mathscr{D}_j f p_j \, d\sigma_j = 0 \quad \textit{for all } f \in C^2(S). \tag{7.5}$$

PROOF. See Theorem 8.1 of [21]. □

COROLLARY. *For $j = 1, \dots, J$,*

$$\lim_{t \to \infty} \frac{1}{t} E[I_j(t)] = \int_{F_j} p_j \, d\sigma_j \equiv \delta_j. \tag{7.6}$$

Because $\tilde{I} = I$, (7.1) is true and

$$\delta = \tilde{\delta}. \tag{7.7}$$

PROOF. Specializing (7.4) by setting $f = 1$, we have (7.6) immediately. □

In light of (7.2) and (7.7), our focus is on computation of the stationary density $p_0(z)$ on S and its associated boundary densities p_j $(j = 1, \dots, J)$. The following proposition states that it is enough to work on the basic adjoint relationship (7.5).

PROPOSITION 7.4. *Let $p = (p_0, p_1, \dots, p_J)$ be any positive integrable functions, with p_0 being a probability density, satisfying the basic adjoint relationship* (7.5). *Then p_0 is the stationary density of the RBM and p_j $(j = 1, \dots, J)$ are associated boundary densities.*

PROOF. See Dai and Kurtz [11]. □

In the Appendix, an algorithm will be described to compute p from the basic adjoint relationship (7.5). This algorithm makes our entire QNET method practically feasible.

8. Naive QNET analysis and refined QNET analysis. To simplify notation, let us assume that the stations of our closed network are numbered so that

$$\rho_1 = \max\{\rho_1, \dots, \rho_J\}. \tag{8.1}$$

That is, no station of the network is more heavily loaded than station 1 (note that relative loadings of the various stations are not affected by the average throughput rate a). This state of affairs is often expressed by saying that station 1 is the "bottleneck station," or at least is tied for bottleneck status.

Rather than supposing that the throughput rate a is somehow known, let us simply assume that it has been estimated a priori as $a = z$, and that the covariance matrix in our Brownian system model (5.9)–(5.13) has consequently been set at $z\tilde{\Omega}$. From (7.1) and (7.7) we know that

$$\frac{1}{t}E\big[\tilde{I}(t)\big] \to \delta \quad \text{as } t \to \infty, \tag{8.2}$$

where δ is computed via (7.6) from the stationary distribution of an RBM whose state space is the solid simplex $S(n)$. Recall that the drift vector of the Brownian motion $\tilde{X}$ in (5.10) is $\tilde{\mu} = -\tilde{R}e$ and that (5.9) is the main equation $\tilde{N}(t) = \tilde{N}(0) + \tilde{X}(t) + \tilde{R}\tilde{I}(t)$. Thus, taking expectations in (5.9), dividing both sides by t, letting $t \to \infty$ and using (8.2), one obtains

$$-\tilde{R}e + \tilde{R}\delta = 0, \quad \text{or equivalently,} \quad \tilde{R}(e - \delta) = 0. \tag{8.3}$$

Now recall from (5.7) and (5.8) that $\tilde{R}$ is of rank $J - 1$ and its one-dimensional null space is spanned by the positive vector ρ. Combining this with (8.3), we have that

$$e - \delta = a\rho \quad \text{for some constant } a. \tag{8.4}$$

On the left side of (8.4), the jth component represents the average utilization for server j, so the constant a appearing on the right is obviously interpreted as the (initially unknown) average throughput rate for the closed network, as our choice of notation suggests. To repeat, (8.4) follows from the mathematical structure of our Brownian system model, regardless of how its elements may be interpreted, but our interpretation of the constant a derives from the original queueing context.

To derive a computational procedure from (8.4), let us write $\delta_1(z)$ to mean the value of δ_1 calculated via (7.6) when the covariance matrix of the RBM is set at $z\tilde{\Omega}$. The computed value of a will then be determined by the first component, say, of (8.4):

$$1 - \delta_1(z) = a\rho_1 \quad \text{or} \quad a = \frac{1 - \delta_1(z)}{\rho_1}. \tag{8.5}$$

If the job population is large, then one would expect idleness at the bottleneck (station 1) to be relatively rare, so one might plausibly set z initially at the theoretical maximum throughput rate:

$$\hat{a} = \max\{a \geq 0: a\rho \leq e\} = \frac{1}{\rho_1}. \tag{8.6}$$

Having set $z = \hat{a}$, one then obtains from (8.5) what we will call the *naive QNET estimate* of a:

$$\tilde{a} = \frac{1 - \delta_1(\hat{a})}{\rho_1}. \tag{8.7}$$

If one terminates with the naive estimate $\tilde{a}$, one has a computational procedure precisely analogous to that proposed by Harrison, Williams and Chen [24] for analysis of closed networks of the generalized Jackson type (see Section 10).

However, if the algorithm described in the Appendix can be run quickly enough (remember that the entire stationary density function must be reestimated for each new value of z), then one can obviously refine the foregoing procedure in iterative fashion. That is, one can set $a^0 = \hat{a}$ and then compute

$$a^{k+1} = \frac{1 - \delta_1(a^k)}{\rho_1} \quad \text{for } k = 0, 1, \ldots, \tag{8.8}$$

in the expectation that

$$a^k \to a^* \quad \text{as } k \to \infty. \tag{8.9}$$

The limiting value a^*, if it exists, will be a fixed point of the relationship (8.4), meaning that

$$e - \delta(a^*) = a^*\rho.$$

Assuming that the fixed point a^* exists and is unique, we will call it the *refined QNET estimate* of the unknown throughput rate a.

We conjecture that the function $\delta_1(\cdot)$ is continuous and strictly increasing on $(0, \infty)$, with $\delta_1(z) \downarrow 0$ as $z \downarrow 0$. If this is true, then there does indeed exist a unique fixed point a^*, as shown in Figure 2. Even more structure on $\delta_1(\cdot)$ will be required to prove convergence of the previously described iterative procedure, but practical experience to date suggests that effective convergence is obtained in just a few iterations, at least for population sizes large enough to give bottleneck utilization rates in the vicinity of 90%. Figure 2 also gives a graphical representation of the naive QNET estimate $\tilde{a}$. In our limited computational experience, the difference between the naive estimate $\tilde{a}$ and refined estimate a^* can be significant, especially for small and moderate population sizes. For example, with a moderate population size one might find that $\tilde{a} = 0.80$ whereas $a^* = 0.83$ (see Section 10), and such a difference is significant in most manufacturing applications. Hereafter, the term "refined QNET estimate" will be used to mean any estimate of system perfor-

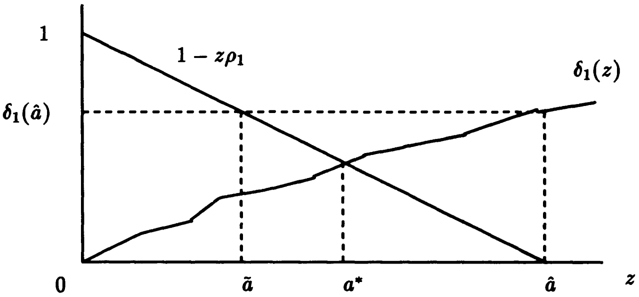

FIG. 2. *Schematic representation of the naive estimate $\tilde{a}$ and refined estimate a^*.*

mance derived from the Brownian model (5.9)–(5.13) with the system throughput rate a set equal to a^* in (5.10), not just to mean the throughput estimate a^* itself.

A closed network model is said to be *perfectly balanced* if $\rho = \kappa e$ for some $\kappa > 0$, or equivalently, if $\rho_1 = \cdots = \rho_J = \kappa$. Because $\tilde{R}\rho = 0$ in general by (5.8), one then has $\tilde{R}e = 0$ for a perfectly balanced system, and hence the Brownian motion $\tilde{X}$ in (5.10) has *zero* drift. For this special case there is an easy argument, involving a rescaling of time, to show that

$$\delta(z) = z\delta(1) \quad \text{for all } z > 0.$$

Thus the fixed point a^* in (8.8) can be computed without iteration via

$$a^* = \frac{1}{\rho_1 + \delta_1(1)}. \tag{8.10}$$

Note that $0 < a^* < \hat{a} = 1/\rho_1$.

9. Summary of performance analysis procedures. In principle, the data of our multiclass closed network are a population size n, a $K \times K$ switching matrix P, mean service times $\tau_1, \dots, \tau_J$ and the associated coefficients of variation $b_1, \dots, b_J$, plus a K-vector α and a $K \times K$ covariance matrix Δ that provide first- and second-moment information, respectively, on the job replacement process (see Section 2). Using the various formulas displayed in Sections 2 and 3, those data are eventually compressed into a $J \times J$ covariance matrix Γ, a $J \times J$ matrix M of routing information and a J-vector ρ of relative load factors, as in (3.5), (3.10) and (3.2). Given the interpretations for Γ, M and ρ provided in Section 3 and earlier in [18], one may be able to estimate these aggregate quantities directly from historical data, or through some theoretical shortcut, without a full elaboration of class-level information.

Now Γ, M and ρ are transformed to a $J \times J$ routing matrix $\tilde{R}$ and a $J \times J$ covariance matrix $\tilde{\Omega}$, each having rank $J - 1$, using the formulas in

(3.22), (5.5) and (5.10). It is $\tilde{R}$, $\tilde{\Omega}$ and n that serve as input data for the QNET method.

Given any initial estimate z for the unknown throughput rate a, one uses the algorithm described in Section 7 and the Appendix to compute (that is, approximate) the stationary distribution of an RBM with state space

$$\tilde{S}(n) = \{x \in \mathbb{R}_+^J \colon e'x = n\},$$

drift vector $\tilde{\mu} = -\tilde{R}e$, covariance matrix $z\tilde{\Omega}$ and reflection matrix $\tilde{R}$. (Recall that e is the J-vector of ones.) It is summary statistics derived from that stationary distribution that become our steady-state performance estimates for the original system. In particular, for a given value of z the algorithm returns a corresponding value, denoted in Section 8 by $\delta_1(z)$, for the idleness rate at station 1. From this, one can compute an improved estimate of the unknown throughput rate a via (see Section 8)

$$a = \frac{1 - \delta_1(z)}{\rho_1}. \tag{9.1}$$

As in Section 8, let the stations be numbered so that $\rho_1 = \max(\rho_1, \dots, \rho_J)$. In the QNET method we use an iterative procedure, beginning with the upper limit value $z = \hat{a} = 1/\rho_1$, to compute a fixed point a^* for (8.1), meaning that

$$a^* = \frac{1 - \delta_1(a^*)}{\rho_1}. \tag{9.2}$$

It is a^* that becomes our final estimate, or *refined QNET estimate*, of the throughput rate a. Finally, running the algorithm described in Section 7 and the Appendix with state space $\tilde{S}(n)$, drift vector $\tilde{\mu} = -\tilde{R}e$, covariance matrix $a^*\tilde{\Omega}$ and reflection matrix $\tilde{R}$, one obtains an approximation for the stationary distribution of the jobcount process $\tilde{N}$. This is our QNET estimate of the steady-state jobcount distribution, and summary statistics derived from it, such as average jobcounts at the various stations, are called *QNET performance estimates*. As we will show by example in Sections 11–13, one can use the queueing formula $L = \lambda W$ (Little's law) to derive estimates of average throughput time from average jobcount estimates.

10. Closed network models of the generalized Jackson type. To clarify the relationship between the QNET method and previous work in queueing theory, let us consider a simple model of the type introduced by Jackson [26] and by Gordon and Newell [14] in their original treatment of closed networks. In their model one has a single job class served at each station (thus $K = J$), and jobs change *station* in Markovian fashion. That is, we are given a $J \times J$ stochastic matrix $\tilde{P} = (\tilde{P}_{ij})$, and a job completing service at station i goes next to station j with probability $\tilde{P}_{ij}$, independent of all previous history. Because the switching matrix $\tilde{P}$ is stochastic, jobs never leave the system, and the population size remains constant at n. In Jackson's original paper on closed networks, service time distributions were assumed to be exponential, whereas we continue to allow general distributions. In other words, the model discussed in this section is the same as Jackson's with respect to job routing, but is more general with respect to service time distributions.

To map such a generalized Jackson network into the framework described in Section 2, we arbitrarily designate station J as the "exit station." A job leaving station J is considered to have completed a "route," and in accordance with that convention, we define a *reduced switching matrix* $P = (P_{ij})$ via

$$P_{ij} = \begin{cases} 0, & \text{if } i = J, \\ \tilde{P}_{ij}, & \text{otherwise.} \end{cases} \tag{10.1}$$

To avoid trivialities, we assume that $\tilde{P}$ is irreducible, and from this it follows that P is transient, as required in our general setup. Of course, the "exit" of a job after service completion at station J is followed immediately by insertion of a "replacement" job, and that replacement job enters the network at station j with probability

$$\alpha_j = \tilde{P}_{Jj}, \qquad j = 1, \dots, J, \tag{10.2}$$

independent of all previous history. Because the class designations of "replacement jobs" are independent and identically distributed, (2.4) gives us

$$\mathrm{Cov}[\zeta(r)] = \Delta r, \tag{10.3}$$

where

$$\Delta_{ij} = \begin{cases} \alpha_j(1 - \alpha_j), & \text{if } i = j, \\ -\alpha_i \alpha_j, & \text{if } i \neq j. \end{cases} \tag{10.4}$$

To review, (10.1), (10.2) and (10.4) show how the data P, α and Δ, required for our general formulation in Section 2, are determined from the original switching matrix $\tilde{P}$ of a generalized Jackson network. The other dáta required for our general formulation are the population size n, the mean service times $\tau_1, \dots, \tau_J$ and the associated coefficients of variation $b_1, \dots, b_J$. In the remainder of this section we show how various formulas presented in Sections 3–5 specialize in the case of a generalized Jackson network to give relatively simple expressions for the routing matrix $\tilde{R}$ and the covariance matrix $\tilde{\Omega}$ that ultimately serve as input data for the QNET method.

Set $Q = (I - P')^{-1}$ as in Section 2 and $\lambda = Q\alpha$ as in (3.2), so that λ_j represents the average number of stops at station j that a job makes on one "route." Readers may verify that, because station J is designated as the "exit station,"

$$\lambda_J = 1. \tag{10.5}$$

The defining characteristic of a generalized Jackson network is that $C = I$, and hence (3.10) specializes to give $M = TQ$, from which we obtain

$$R = M^{-1} = (I - P')T^{-1}. \tag{10.6}$$

Next, using (10.6) we reduce the definition (4.11) to

$$u = R\rho = (I - P')T^{-1}(T\lambda) = (I - P')Q\alpha = \alpha. \tag{10.7}$$

However, (5.2) then gives $c = e'\alpha = 1$, and the definition (5.5) simplifies to give

$$B = I - \alpha e' \tag{10.8}$$

and

$$\tilde{R} = BR = (I - \alpha e')(I - P')T^{-1}. \tag{10.9}$$

According to (10.1), P is identical to the stochastic matrix $\tilde{P}$ except that its Jth row is set to zero and, hence, $e'(I - P') = (0, \dots, 0, 1)$. Also, (10.2) says that α is the Jth column of $\tilde{P}'$, so (10.9) reduces to

$$\tilde{R} = (I - \tilde{P}')T^{-1}. \tag{10.10}$$

Recall that in Section 2 we defined covariance matrices $H^1, \dots, H^J$ in terms of P by means of formula (2.2). To derive a simplified expression for $\tilde{\Omega}$ in the case at hand, it will be convenient to define $J \times J$ covariance matrices $\tilde{H}^1, \dots, \tilde{H}^J$ via the same formula (2.2), but using $\tilde{P}$ in place of P. From (10.4) and the definitions (10.1) and (10.2), one then has

$$H^j = \tilde{H}^j \quad \text{for } j = 1, \dots, J - 1 \text{ and } \Delta = \tilde{H}^J. \tag{10.11}$$

Combining this with (10.5) and the definition (3.7) of H gives

$$\Delta + H = \tilde{H}, \tag{10.12}$$

where

$$\tilde{H} = \sum_{j=1}^{J} \lambda_j \tilde{H}^j. \tag{10.13}$$

Substituting (10.12) into the general formulas (3.5) and (3.6) gives

$$\begin{aligned} \Gamma &= (TQ)[\Delta + H + (I - P')D(I - P)](TQ)' \\ &= T(Q\tilde{H}Q' + D)T'. \end{aligned} \tag{10.14}$$

Of course, $\tilde{\Omega} = \tilde{R}\Gamma\tilde{R}'$ in general by (5.10), so (10.10) and (10.14) together imply that

$$\tilde{\Omega} = \tilde{R}\Gamma\tilde{R}' = \tilde{H} + (I - \tilde{P}')D(I - \tilde{P}). \tag{10.15}$$

Having computed $\tilde{R}$ and $\tilde{\Omega}$ from system data by means of (10.10) and (10.15), one can now apply the computational scheme outlined in Section 9 to estimate steady-state performance for a closed network of the generalized Jackson type. This method of performance analysis is the same as that proposed by Harrison, Williams and Chen [24] for generalized Jackson networks, except in the following regard. In [24], load factors or traffic intensity parameters $\rho_1, \dots, \rho_J$ were defined in such a way that max $\rho_j = 1$. In the language of this paper, that scaling convention is equivalent to fixing the a priori estimate of a in (5.10) at the upper limit value $\hat{a}$ (see Section 9);

hence, the scheme proposed in [24] is equivalent to what we have called naive QNET analysis. Unfortunately, readers who wish to verify the equivalence of our formulas and those in [24] will have to fight through formidable differences in notation that have arisen over a span of years in the development of Brownian models for ever more general classes of systems.

11. QNET analysis of a symmetric cyclic queue. Consider a J-station *cyclic* closed network. We assume that the network is populated by n *homogeneous* jobs circulating perpetually. This is a *generalized Jackson network* as discussed in Section 10. In such a network each server or station serves a single class of jobs, and thus the number of job classes K is equal to the number of stations J. We further assume that all servers are identical, having the same service time distribution with mean 1 and squared coefficient of variation b^2. This "simple" network is not amenable to exact analysis except when all the service time distributions are exponential.

The Brownian approximation for this network is particularly appealing because it can be analyzed exactly, without using numerical methods. We start by deriving the data for the Brownian model, a reflected Brownian motion in the simplex $\tilde{S}(n)$. First the routing matrix $\tilde{P}$ is given by $\tilde{P}_{i,i+1} = 1$ for $i = 1, \dots, J-1$ and $\tilde{P}_{J,1} = 1$. Because all the routing is deterministic and all servers are identical, from (10.10) and (10.15) one has the reflection matrix

$$\tilde{R} = I - \tilde{P}'$$

and the covariance matrix

$$a\tilde{\Omega} = ab^2(I - \tilde{P}')(I - \tilde{P}),$$

where a is the unknown throughput rate. Because, $\tilde{P}'\tilde{P} = I$, one has

$$a\tilde{\Omega} = ab^2(I - \tilde{P}) + (I - \tilde{P}')ab^2, \tag{11.1}$$

which is exactly the "product form" condition (4.10) of [24]. Thus the stationary distribution of the reflected Brownian motion is of exponential form and because the drift vector $\tilde{\mu} = -\tilde{R}e = 0$, the stationary distribution is uniform over the simplex $\tilde{S}(n)$. From (4.21) and (4.22) of [24],

$$E\left[\tilde{I}(t)\right] \sim \delta t \quad \text{as } t \to \infty,$$

where

$$\delta_i = ab^2 n^{-1}(J-1), \qquad i = 1, \dots, J.$$

Therefore, the refined throughput estimate a^* can be solved as a fixed point from (8.8) without iteration:

$$a^* = \left[1 + b^2 n^{-1}(J-1)\right]^{-1}. \tag{11.2}$$

Let us denote by T the long-run average time between consecutive entries into station 1 for any given job, calling this the average *cycle time* for the system. An application of Little's law ($L = \lambda W$) to the case at hand tells us that $n = aT$ and, hence, our QNET estimate of the average cycle time is

$$T = n/a^* = n + (J - 1)b^2. \tag{11.3}$$

To assess the numerical accuracy of (11.2) and (11.3), let us consider a system with $J = 4$ stations (the case $J = 3$ has also been investigated, and the results are very similar) and ask the following question for different values of the squared coefficient of variation b^2 of the service time: How large must one make the number of active jobs n in order to achieve a throughput rate $a = 0.9091$, or equivalently, a server utilization of 90.91% at each station? (We are interested in parameter combinations that give server utilization of about 90%. A nearby uneven value was chosen in order to get round values of n corresponding to round values of b^2.) Setting $J = 4$ and $a^* = 0.9091$ in (11.2), we estimate the required population size to be $n = 30b^2$, and (11.3) estimates the corresponding average cycle time to be $33b^2$. Those estimates are compared against simulation results for four different values of b^2, ranging from 0.5 to 10, in Table 1. The simulations were performed using SIMAN 3.5 [31]. In all cases, 10 replications were run and in each replication statistics were collected based on the first 30,000 cycles of jobs. In the simulation, we used an Erlang distribution for service times when $b^2 = 0.5$, and used a hyperexponential distribution with balanced mean when $b^2 > 1$. That is, when $b^2 > 1$ the service time distribution is exponential with mean $0.5/p$ with probability p and is exponential with mean $0.5/(1 - p)$ with probability $1 - p$, where

$$p = 0.5 + 0.5\left[(b^2 - 1)/(b^2 + 1)\right]^{1/2}.$$

In Table 1 and in all subsequent tables, as suggested by Reiman [33], the number in parentheses after each simulation result is the half-width of the 95% confidence interval, expressed as a percentage of the simulation average. The number in parentheses after each QNET estimate is the percentage error relative to the simulation average. This format makes it easy to determine the statistical significance of the errors. Table 1 indicates that QNET estimates of utilization are extremely accurate when compared with simulation estimates. Incidentally, we did not list the case where the common service

TABLE 1
Simulation estimates and QNET estimates of the utilization rate and mean cycle time for a cyclic network

		Utilization		Cycle time	
n	b^2	QNET	SIM	QNET	SIM
15	0.5	0.909 (1.01%)	0.900 (0.56%)	16.50 (−0.60%)	16.60 (0.60%)
60	2.0	0.909 (−0.54%)	0.914 (0.66%)	66.00 (0.61%)	65.60 (0.46%)
120	4.0	0.909 (−0.75%)	0.916 (0.76%)	132.00 (0.76%)	131.00 (0.76%)
300	10.0	0.909 (−0.75%)	0.916 (0.98%)	330.00 (1.23%)	326.00 (1.23%)

time distribution is exponential, in which case formula (11.2) is exact. Also listed in Table 1 is the QNET estimate and simulation estimate of mean cycle time. Again QNET estimates are quite accurate. Note that, to achieve the same utilization rate of 91%, the mean cycle time is 20 times longer for $b^2 = 10$ than for $b^2 = 0.5$.

For the cyclic queue studied in this section the upper limit throughput rate is $\hat{a} = 1/\rho_1 = 1$. Thus, if we use the naive QNET estimate (8.7) of a, we will have an estimated throughput rate

$$\tilde{a} = 1 - b^1(J - 1)n^{-1},$$

which is not accurate when the actual utilization rate a is below 85%, even for the case of exponential service time distribution. For example, when $n = 15$, $J = 4$ and the common service time distribution is exponential, our refined QNET estimate of server utilization (exact in this case) is 83%, whereas the naive QNET estimate is 80%.

12. A multiproduct two-station example.

Pictured in Figure 3 is a system with two machines producing three types of parts: part A, part B and part C. As indicated in Figure 3, the route of part A is: machine 1, machine 2, machine 1, machine 2. After the second service at machine 2, part A exits the system. Part B begins processing at machine 2 and then moves to machine 1, and then exits. Part C visits machine 1 and exits. The service discipline at each machine is assumed to be first-in–first-out (FIFO). To maintain a constant number n of active jobs, we will consider two different input control rules, referred to as *cyclic input* and *random input*, respectively. Under the former rule, as parts complete processing they are deterministically replaced by new parts of types A, B, C, A, B, C, A, B, C and so forth. Under the latter rule, each part that completes processing is replaced by a new one whose type is randomly determined, independent of all previous history, with the three different types being equally likely. The total number of parts n in the system is set at either 15 or 30 in our study.

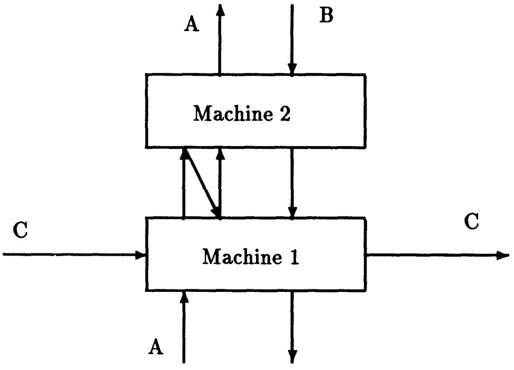

FIG. 3. *A two-machine system producing three types of products.*

To facilitate discussion, let us define different classes of jobs. Parts of type A at the first visit to machine 1 are called class 1 jobs, and at the first visit to machine 2 are called class 2 jobs. Parts of type A at the second visit to machines 1 and 2 are called class 3 and class 4 jobs, respectively. Parts of type B at machine 2 and machine 1 are called class 5 and class 6 jobs, respectively. Parts of type C belong to class 7. With this definition of job classes, the constituency matrix is given by

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

The transition matrix among classes is a 7×7 matrix P, with $P_{12} = P_{23} = P_{34} = P_{56} = 1$, and $P_{ij} = 0$ otherwise. The service times for jobs at station i are assumed to have a general distribution with mean τ_i and squared coefficient of variation b_i^2. We choose $\tau_1 = 0.75$ and $\tau_2 = 1$. With this choice of mean service times, the loads on the two machines are balanced, with $\rho_1 = \rho_2 = 1.0$, so the upper limit throughput rate is $\hat{a} = 1/\rho_1 = 1/\rho_2 = 1$. We consider different cases for this model. The service time at station 2 is assumed to be exponentially distributed (hence $b_2 = 1$). The service time at station 1 is either exponentially distributed ($b_1 = 1.0$) or constant ($b_1 = 0.0$). The total number of jobs allowed in the system is either $n = 15$ or $n = 30$. Rather than reporting the throughput rate a, one may equivalently speak in terms of the common utilization rate of the two machines, which is given by $u = a/\hat{a} = a$. The performance measures that we consider are the common utilization rate u and the mean throughput time for each type of part, denoted by T_A, T_B and T_C. (The throughput time for an individual part is defined as the elapsed time between its entry to and exit from the system.)

Both simulation results and QNET estimates for the various performance measures are shown in Table 2. The QNET estimates of machine utilization are extremely accurate for all cases, and the QNET estimates of mean throughput times are also very good in general. Both the QNET estimates

TABLE 2
Simulation estimates and QNET estimates of performance measures with random input and with cyclic input

n/b	Input		Utilization	T_A	T_B	T_C
30/1	Cyclic	QNET	0.972 (0.02%)	54.00 (0.00%)	27.00 (0.00%)	11.55 (1.31%)
		SIM	0.972 (0.51%)	54.00 (0.56%)	27.00 (0.37%)	11.40 (1.75%)
	Random	QNET	0.968 (0.07%)	54.25 (0.28%)	27.12 (0.09%)	11.62 (−0.65%)
		SIM	0.967 (0.10%)	54.10 (0.37%)	27.10 (0.37%)	11.70 (1.71%)
15/1	Cyclic	QNET	0.946 (−0.12%)	27.75 (−0.18%)	13.88 (−0.18%)	5.95 (3.30%)
		SIM	0.947 (0.21%)	27.80 (0.36%)	13.90 (0.72%)	5.75 (1.39%)
	Random	QNET	0.938 (0.05%)	28.00 (0.00%)	14.00 (0.00%)	6.00 (0.33%)
		SIM	0.937 (0.32%)	28.00 (0.36%)	14.00 (0.71%)	5.98 (1.17%)
15/0	Cyclic	QNET	0.968 (0.14%)	27.11 (−0.71%)	13.55 (1.13%)	5.81 (0.32%)
		SIM	0.967 (0.21%)	27.30 (0.37%)	13.40 (0.75%)	5.79 (1.90%)
	Random	QNET	0.960 (0.16%)	27.36 (−0.16%)	13.68 (0.57%)	5.86 (−2.52%)
		SIM	0.958 (0.21%)	27.40 (0.36%)	13.60 (0.29%)	6.01 (1.16%)

and simulation estimates predict that the mean throughput times are longer for the system with random input than that with cyclic input. However, the differences are small. Also notable is that the QNET estimates are generally better when $n = 30$ than when $n = 15$.

The QNET estimates displayed in Table 2, like those appearing in Table 1, were actually derived without recourse to numerical methods, because we are dealing with another special case where explicit formulas are available for steady-state analysis of the Brownian system model. (As the following discussion will reveal, the special feature of the current example is that it involves only two workstations and their workloads are balanced.) In the remainder of this section we will explain how the parameters of the Brownian system model were derived and how the performance estimates in Table 2 were computed from those parameters when the service time at station 1 is exponentially distributed.

The vector of average input rates in (2.3) is obviously $\alpha = (\frac{1}{3}, 0, 0, 0, \frac{1}{3}, 0, \frac{1}{3})'$, regardless of whether one has cyclic input or random input. The associated covariance matrix is simply $\Delta = 0$ if input is cyclic. If input is random, then (2.4) gives

$$\Delta_{\{1,5,7\}} = \frac{1}{9}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix},$$

where $\Delta_{\{1,5,7\}}$ is the 3×3 submatrix composed of rows 1, 5 and 7 and columns 1, 5 and 7; all the other elements of Δ are zero. The mean service times τ_i for two stations were specified earlier, and the associated coefficients of variation are $b_1 = b_2 = 1$, because each service time distribution is exponential. The 7×7 switching matrix P was specified earlier. Substituting these data into the general formulas in Sections 3 and 4, one eventually obtains

$$(12.1) \qquad M = \frac{1}{12}\begin{pmatrix} 11.25 & 6.00 \\ 9.00 & 16.00 \end{pmatrix} \text{ regardless of the input scheme,}$$

$$(12.2) \qquad \Gamma = \frac{1}{12}\begin{pmatrix} 9 & 0 \\ 0 & 0.12 \end{pmatrix} \text{ if input is cyclic}$$

and

$$(12.3) \qquad \Gamma = \frac{1}{12}\begin{pmatrix} 10.5 & 3.0 \\ 3.0 & 20.0 \end{pmatrix} \text{ if input is random.}$$

As noted earlier, the vector of relative load factors is $\rho = (1.0, 1.0)'$, regardless of the input scheme. Applying formulas (3.22), (5.5) and (5.10), one then obtains

$$(12.4) \qquad R = M^{-1} = \begin{pmatrix} 0.9375 & -0.5000 \\ -0.7500 & 1.3333 \end{pmatrix},$$

$$(12.5) \qquad \tilde{R} = \begin{pmatrix} 0.9796 & -0.9796 \\ -0.9796 & 0.9796 \end{pmatrix},$$

$$\tilde{\Omega} = \begin{pmatrix} 1.6793 & -1.6793 \\ -1.6793 & 1.6793 \end{pmatrix} \text{ if input is cyclic} \tag{12.6}$$

and

$$\tilde{\Omega} = \begin{pmatrix} 1.9592 & -1.9592 \\ -1.9592 & 1.9592 \end{pmatrix} \text{ if input is random.} \tag{12.7}$$

Finally, because $\rho = e$, we know from (5.8) that

$$\tilde{\mu} = -\tilde{R}e = 0. \tag{12.8}$$

As explained in Section 5, the two-dimensional jobcount process $\tilde{N}$ is to be approximated by an RBM with state space $\tilde{S}(n)$, drift vector $\tilde{\mu} = 0$, reflection matrix $\tilde{R}$ and covariance matrix $a\tilde{\Omega}$, where a is the unknown throughput rate and $\tilde{S}(n) = \{x \in \mathbb{R}_+^2\colon x_1 + x_2 = n\}$. But this simply means that either component of $\tilde{N}$ (that is, the jobcount at either station) is to be approximated by a one-dimensional RBM on the finite interval $[0, n]$ with drift parameter zero, reflection matrix $(r, -r)$, where

$$r = \tilde{R}_{11} = 0.9796, \tag{12.9}$$

and variance parameter $a\sigma^2$, where

$$\sigma^2 = \tilde{\Omega}_{11} = \begin{cases} 1.6793, & \text{if input is cyclic,} \\ 1.9592, & \text{if input is random.} \end{cases} \tag{12.10}$$

(Note that σ^2 is larger with random input than with cyclic input.) Let us suppose that the throughput rate is estimated as $a = z$. Then known results for one-dimensional RBM (see Chapter 5 of [16]) tell us that the stationary distribution is uniform, independent of z and the corresponding estimate of the steady-state idleness rate for machine 1 is $\delta_1(z) = z\sigma^2/(2rn)$. Because the RBM has zero drift, one can use formula (8.10) to compute the refined QNET estimate

$$a^* = \frac{1}{\rho_1 + \sigma^2/(2rn)} = \frac{1}{1 + \sigma^2/(1.9592n)}$$

for the system throughput rate, and then our estimate of machine utilization is $u^* = a^*/2$.

From the uniform stationary distribution of the foregoing RBM, the QNET method estimates the average total jobcount at each station to be $n/2$, so by Little's law the average number of jobs waiting at either station is estimated at $n/2 - u^*$. Then another application of Little's law gives an estimate of the mean waiting time per job visit at each station:

$$W_i = \frac{1}{a^*\gamma_i}\left(\frac{n}{2} - u^*\right), \qquad i = 1, 2, \tag{12.11}$$

where γ_i is the total arrival rate at station i when $a = 1$, calculated by means (3.2) and (3.3). (Readers may easily verify that $\gamma_1 = 4/3$ and $\gamma_2 = 1$.)

Finally, our QNET estimates of the mean sojourn times are given by

$$\begin{aligned} T_A &= 2(W_1 + W_2) + 1.0, \\ T_B &= (W_1 + W_2) + 1.0, \\ T_C &= W_1 + 1.0. \end{aligned}$$

13. A three-station example inspired by Solberg. Figure 4 depicts a three-station manufacturing system composed of one mill, one drill and one inspector. A unit of product A is milled first (mean service time 1.1) and then drilled (mean service time 1.5). After drilling there is a 50% chance that product A will require inspection (with a long mean service time of 5.0), and whether it is inspected or not, a unit of product A returns to milling (with mean service time of 1.1) before leaving the system. Type B jobs do not require any inspection. They are drilled first (mean service time 1.5) and then milled (mean service time 1.1). The number of active jobs is held constant at 20, and the replacement scheme is cyclic and deterministic $(A, B, A, B, \dots)$. A similar manufacturing model was considered by Solberg [37], who also included a very *lightly* loaded transport station. Because the lightly loaded station has little effect on the performance of the entire system, we consider a simplified version with the transport station deleted.

Enumerate milling, drilling and inspection as stations 1, 2 and 3, respectively. From the traffic equation (3.2), we see that $\rho = (1.65, 1.5, 1.25)'$, so milling is the bottleneck station. The upper limit throughput rate is $\hat{\alpha} = 1/1.65 = 0.6061$. We consider two cases for such a manufacturing system. In the first case, all the service times are exponentially distributed (case 1). In the second case, the service time at the bottleneck station (milling) has squared coefficient of variation 4 and the rest of the service time distributions are still exponential. In simulation, we used a hyperexponential distribution with balanced mean (case 2) and a gamma distribution (case 3) to fit the service time distribution at the milling station. Table 3 gives the QNET

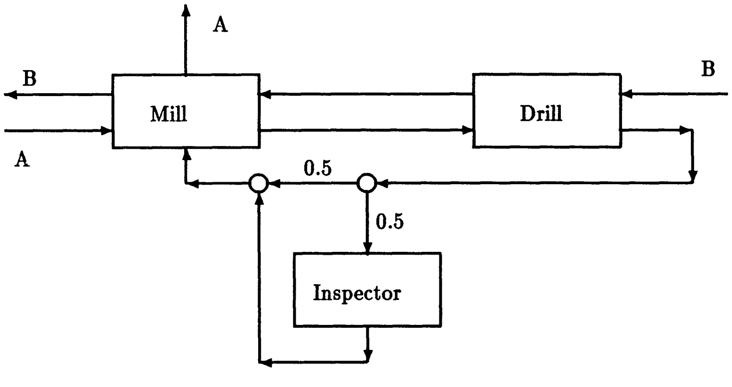

FIG. 4. *A manufacturing model inspired by Solberg.*

TABLE 3
Simulation estimates and QNET estimates of mean sojourn times and the utilization rate at the bottleneck station

Case	Mill utilization	Mean throughput time	
		Type *A*	Type *B*
1 QNET	0.980 (− 0.08%)	44.38 (− 0.26%)	22.95 (0.65%)
SIM	0.981 (0.10%)	44.50 (0.22%)	22.80 (0.44%)
2 QNET	0.928 (− 1.29%)	45.72 (0.92%)	25.40 (3.53%)
SIM	0.940 (0.21%)	45.30 (0.22%)	24.50 (0.82%)
3 QNET	0.928 (− 1.51%)	45.72 (0.70%)	25.40 (2.75%)
SIM	0.942 (0.21%)	45.40 (0.66%)	24.70 (0.81%)

estimates and simulation estimates of the mean throughput time for each product, as well as the utilization rate of the mill. Simulation results for cases 2 and 3 show that the *invariance principle* approximately holds, which says that steady system performance is insensitive to the actual distribution used to fit the service time distributions as long as the first two moments are fixed. However, the QNET and simulation estimates both show that high variability of the service time at the bottleneck station will substantially reduce the utilization rate of the system.

14. Extension to more general manufacturing systems. In Section 2, we described a closed network model that is quite general with respect to processing requirements (that is, the mix of jobs to be handled, their routes and their service times) but very restrictive with respect to processing capabilities. To be specific, it was assumed that each workstation consists of a single perfectly reliable server, but current work by Pich [32] shows how the model may be extended to allow what he calls "general workstation capabilities." To explain Pich's important generalization, let us suppose that each station consists of one or more identical machines (that is, machines play the role of servers) and that the service times of jobs are expressed in machine hours. It may be that machines break down from time to time and that machines are occasionally idled while they wait for technicians to become available for setup or repair activities, but three key assumptions allow one to cut through this complexity, at least for purposes of approximate analysis. The first is that workstations operate independently, which means, for example, that each station has its own dedicated operators and repair technicians.

To express the other two assumptions, let us suppose for the moment that station i is provided with an infinite queue of jobs to be processed, each belonging to some class k such that $s(k) = i$, and denote by $S_i(t)$ the total machine hours applied to the service requirements of those jobs up to time t, taking into account interruptions of all kinds. If the station consists of p machines and there are no service interruptions to be dealt with, then $S_i(t) = pt$ with probability 1, but in general $S_i(\cdot)$ is a stochastic process. The second key assumption is that the distribution of $S_i(\cdot)$ does not depend on the classes of jobs to be processed, which is reasonable in many circumstances. In a model with machine breakdown, for example, it is not required that jobs of different classes have the same service time distribution, but one must assume that breakdown frequency is independent of the job class being processed.

The final assumption in Pich's treatment of general workstation capabilities is that the cumulative service process $S_i(\cdot)$ satisfies a functional central limit theorem. That is, for each station i there exists a long-run average processing rate β_i and an associated variability parameter ω_i such that

$$E[S_i(t)] \sim \beta_i t \quad \text{and} \quad \mathrm{Var}[S_i(t)] \sim \omega_i^2 t$$

as $t \to \infty$ and, moreover, the centered and scaled processes $\{r^{-1/2}[S_i(rt) - r\beta_i t],\ t \geq 0\}$, indexed by $r = 1, 2, \dots,$ converge weakly as $r \to \infty$ to a Brownian motion with drift zero and variance ω_i^2. The pair (β_i, ω_i) provides a crude two-moment characterization of processing capabilities at station i. Pich [32] explains how these parameters are incorporated into a generalized Brownian approximation for a multiclass *open* network, and the story is virtually identical for our multiclass closed network. This generalization of the basic model developed in Section 2 is crucial for manufacturing applications, but to avoid excessive length, we will omit detailed further discussion. It should be mentioned in conclusion, however, that Pich's generalized Brownian model provides a rather crude approximation for stations with many servers. Such stations are less common in manufacturing than in communication systems and computer networks, but they do occur in some industries and, hence, there is a need for more research on refined approximations for many-server stations.

Another potential generalization of the model described in Section 2 concerns the incorporation of non-FIFO service disciplines. Harrison and Williams [23] explain in the Appendix of a recent paper how the Brownian model of a multiclass open network may be modified, at least in principle, to allow non-FIFO disciplines of a certain restrictive type, and that account carries over without change to multiclass closed networks. In broad terms, changes in service discipline manifest themselves as changes in the coefficient matrix M that relates the total workload process to the jobcount process in (3.16) or (4.8). The only non-FIFO disciplines that Harrison and Williams discuss in specific terms, however, are static priorities and processor sharing, and their proposed approximations are not supported by a rigorous heavy traffic limit theorem. Their attention is restricted entirely to *local* scheduling rules, which means that decisions about which job to take next at station i may depend only on the current composition of the job population *at that station*. This rules out the kind of global optimization that one might hope to accomplish with real-time computer control of a manufacturing system, where scheduling decisions (that is, sequencing decisions) at one station are based on congestion levels elsewhere in the network. On the other hand, decades of research on jobshop scheduling have yielded very little insight as to how such global information can be used effectively, so the restriction to simple local rules is usual in network performance analysis. Also, the analysis and numerical results in Harrison and Wein [20] and Chevalier and Wein [8] suggest that a static priority rule can be quite effective for maximizing the throughput of multiclass, single chain closed queueing networks.

15. An open problem. There is the open question of whether one can justify our approximating Brownian model of a multiclass closed network by a rigorous heavy traffic limit theorem like that proved by Chen and Mandelbaum [7] for closed networks of the generalized Jackson type. To do so, one must consider a sequence of multiclass closed networks with population size $n \to \infty$ and such that the loadings of the various stations are asymptotically balanced in an appropriate sense. We conjecture that the multidimensional jobcount processes for this sequence of networks converge after appropriate scaling to the jobcount process associated with a corresponding Brownian system model.

APPENDIX

Computing the stationary distribution. Continuing from Section 7, we are given a (μ, Ω, R) RBM N in the solid complex $S \equiv S(n) = \{z \in \mathbb{R}_+^{J-1}\colon e'z \le n\}$. This Appendix is concerned with an algorithm for computing the stationary distribution π and its associated vector δ of idleness rates. By a simple scaling argument, we can restrict the state space to be the *unit* solid simplex

$$S^* = S(1) = \{z \in \mathbb{R}_+^{J-1}\colon e'z \le 1\}.$$

In fact, let π^* and δ^* be the stationary distribution and associated vector of idleness rates for a $(n\mu, \Omega, R)$ RBM N^* in the unit solid simplex S^*. One can show that

$$\text{(A.1)} \qquad \delta = \frac{\delta^*}{n} \quad \text{and} \quad \pi(B) = \int_{S^*} 1_B(nz)\pi^*(dz) \quad \text{for } B \in \mathscr{B}_S,$$

where $\mathscr{B}_S$ is the Borel σ-field of S. In the remainder of this Appendix we will consider an RBM with data $(n\mu, \Omega, R)$ and state space $S = S^* = S(1)$, but the asterisks will be dropped in our notational system. Readers must bear in mind that the output π and δ eventually obtained are actually the quantities π^* and δ^* identified previously. First notice that Propositions 7.2 and 7.3 still hold and, therefore, our task is to find stationary density p_0 of π and its associated boundary densities p_i $(i = 1, \dots, J)$ of a $(n\mu, \Omega, R)$ RBM in S from the basic adjoint relationship (7.5).

The algorithm. Analogous to a general algorithm in [9] and [10], where the state space is a two-dimensional rectangle and a J-dimensional nonnegative orthant, respectively, we convert (7.5) into a compact form. Given an $f \in C_b^2(S)$, let

$$\mathscr{A}f \equiv (\mathscr{G}f, \mathscr{D}_1 f, \dots, \mathscr{D}_J f). \tag{A.2}$$

Also, let

$$dm \equiv (dz, d\sigma_1, \dots, d\sigma_J), \tag{A.3}$$

where dz is Lebesgue measure on S, σ_j $(j = 1, \dots, J-1)$ is surface Lebesgue measure on F_j and σ_J is $(J-1)^{1/2}$ times Lebesgue measure on F_J; see Proposition 7.3. For a subset E of $\mathbb{R}^{J-1}$, let $\mathscr{B}_E$ be the Borel σ-field of E and let $\mathscr{B}(E)$ denote the set of functions that are $\mathscr{B}_E$-measurable. Let

$$\begin{aligned} L^j(S, dm) \equiv \Bigg\{ g = (g_0, g_1, \dots, g_J) &\in \mathscr{B}(S) \\ &\times \mathscr{B}(F_1) \times \cdots \times \mathscr{B}(F_J): \\ &\int_S |g_0|^j \, dz + \sum_{i=1}^J \int_{F_i} |g_i|^j \, d\sigma_i < \infty \Bigg\}, \qquad j = 1, 2, \dots, \end{aligned} \tag{A.4}$$

and for $g \in L^1(S, dm)$, let

$$\int_S g \, dm \equiv \int_S g_0 \, dz + \sum_{i=1}^J \int_{F_i} g_i \, d\sigma_i.$$

For $g, h \in \mathscr{B}(S) \times \mathscr{B}(F_1) \times \cdots \times \mathscr{B}(F_J)$ we put

$$g \cdot h \equiv (g_0 h_0, g_1 h_1, \dots, g_J h_J),$$

and for $h > 0$ we put

$$g/h \equiv (g_0/h_0, g_1/h_1, \dots, g_J/h_J).$$

With this notation, the basic adjoint relationship (7.5) can be rewritten as

$$\int_S (\mathscr{A}f \cdot p) \, dm = 0 \quad \text{for all } f \in C_b^2(S). \tag{A.5}$$

Let $q = (q_0, q_1, \dots, q_J)$ be some given positive element in $L^1(S, dm)$. That is, $q_i > 0$ $(i = 0, \dots, J)$, $\int_S q_0(z) \, dz < \infty$ and $\int_{F_i} q_i \, d\sigma_i < \infty$ $(i = 1, \dots, J)$. The function q is called a *reference density*, whose role will be explained later. Given a reference density q, we define the *reference measure*

$$d\eta \equiv q \, dm = (q_0 \, dz, q_1 \, d\sigma_1, \dots, q_J \, d\sigma_J), \tag{A.6}$$

where the measure dm is defined in (A.3). Similar to the definition of $L^i(S, dm)$ and $\int_S g \, dm$ for $g \in L^1(S, dm)$, we can define $L^i(S, d\eta)$ and $\int_S g \, d\eta$ for $g \in L^1(S, d\eta)$. If we introduce a new unknown $r = p/q$, then the basic adjoint relationship (A.5) takes the form

$$\int_S (\mathscr{A}f \cdot r) \, d\eta = 0 \quad \text{for all } f \in C_b^2(S). \tag{A.7}$$

In the following, we actually develop an algorithm to solve for this new unknown r. Of course, once one has r, one can get the stationary density p via $p = r \cdot q$.

We denote by $L^2 \equiv L^2(S, d\eta)$ all the square integrable functions on S with respect to $d\eta$, taken with the usual inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Unless specified otherwise, all inner products and norms are taken in $L^2(S, d\eta)$. Because η is a finite measure and $\mathscr{A}f$ is bounded, we have $\mathscr{A}f \in L^2$ for each $f \in C_b^2(S)$. Assume for the moment that

$$r \in L^2(S, d\eta); \tag{A.8}$$

the basic adjoint relationship (A.7) dictates that we need only to find a $r \in L^2$ that is orthogonal to $\mathscr{A}f$ for each $f \in C_b^2(S)$, or, equivalently, we need only to find an r that is orthogonal to H, where H is the *closed subspace* of L^2 defined as

$$H \equiv \text{the closure of } \{\mathscr{A}f \colon f \in C_b^2(S)\}. \tag{A.9}$$

Let h^0 be the element of L^2 defined by

$$h^0 = (1, 1, \dots, 1). \tag{A.10}$$

An easy calculation gives

$$\begin{aligned} \langle r, h^0 \rangle &= \int_S p_0(z)\, dz + \sum_{i=1}^{J} \int_{F_i} p_i \, d\sigma_i \\ &\geq 1. \end{aligned}$$

Therefore, r is *not* orthogonal to h^0. Because r is orthogonal to H, therefore h^0 is not in H and thus $h^0 - \bar{h}^0$ is nontrivial, where $\bar{h}^0$ is the projection of h^0 onto H. Obviously, $h^0 - \bar{h}^0$ is orthogonal to H; therefore, there is a constant κ such that

$$r = \kappa(h^0 - \bar{h}^0). \tag{A.11}$$

In getting (A.11), we have used Proposition 7.4 and the conjecture that $h^0 - \bar{h}^0 \geq 0$. We do not have a general proof of this conjecture, and we assume this conjecture is true for the rest of this section.

Let

$$\{f_1, \dots, f_s\} \tag{A.12}$$

be a basis of an s-dimensional subspace of $C_b^2(S)$. Define $h^i = \mathscr{A}f_i$ $(i = 1, \dots, s)$. Then $\{h^1, \dots, h^s\}$ spans a finite-dimensional subspace H_s of H. Functions $h^1, \dots, h^s$ might be linearly *dependent*, depending on whether or not a constant is a linear combination of $f_1, \dots, f_s$. Assume that is the case. Then $\{h^1, \dots, h^{s-1}\}$ is a basis for H_s that has dimension $s - 1$. We still use $\bar{h}^0$ to denote the projection of h^0 onto H_s. By a standard approach to find a projection onto a finitely dimensional subspace, there exists $x = (x_1, \dots, x_{s-1})'$ such that

$$\bar{h}^0 = \sum_{i=1}^{s-1} x_i h^i$$

and x is the unique solution of the *normal* equation

$$Kx = b, \tag{A.13}$$

where $K_{ij} = \langle h^i, h^j \rangle$ for $1 \le i, j \le s - 1$ and $b_i = \langle h^i, h^0 \rangle$ $(i = 1, \dots, s - 1)$. The matrix K is symmetric and *positive definite*; therefore, (A.13) has a unique solution. Thus the sth approximation of r is

$$r = \kappa\left(h^0 - \sum_{i=1}^{s-1} h^i \right), \tag{A.14}$$

where κ is a normalizing constant such that $\int_S r_0(z) q_0(z)\, dz = 1$. Finally, the stationary density $p_0(z)$ is given by

$$p_0(z) = r_0(z) q_0(z)$$

and the idleness rate δ is given by

$$\delta_i = \langle r, \phi^i \rangle, \qquad i = 1, \dots, J,$$

where $\phi^i \in L^2$, whose only nonzero component is $\phi^i_i = 1.0$. Obviously, the choice of reference density q and the basis function f_i $(i = 1, \dots, s)$ will affect both the accuracy and efficiency of the algorithm. To present the one choice of q we used for this paper, we first give the condition under which the stationary density is of exponential form.

Exponential solutions. Define a $J \times (J - 1)$ matrix

$$L = \begin{pmatrix} I \\ -(J-1)^{-1/2} e' \end{pmatrix},$$

where e is the $(J - 1)$-dimensional (column) vector of ones and I is the $(J - 1) \times (J - 1)$ identity matrix. Let $D = \operatorname{diag}(LR)$ and $\Lambda = \operatorname{diag}(\Omega_{1,1}, \dots, \Omega_{J-1,J-1}, (J - 1)^{-1} e' \Omega e)$. By Lemma 6.1, the diagonal elements of $\tilde{R}$ are positive. Thus D is a positive diagonal matrix and hence is invertible. Performing the linear transformation exactly in the same way as in Harrison and Williams ([21], Section 9) or in Harrison, Williams and Chen ([24], Section 8) and then applying Theorem 6.1 of Harrison and Williams [22], one can show that p_0 is of exponential form if and only if

$$2L\Omega L' = LRD^{-1}\Lambda + \Lambda D^{-1} R' L'. \tag{A.15}$$

Using the fact that $e'\tilde{R} = 0$ and $e'\tilde{\Omega} = 0$, one can check that (A.15) is in turn equivalent to

$$2\tilde{\Omega} = \tilde{R}\tilde{D}^{-1}\tilde{\Lambda} + \tilde{\Lambda}\tilde{D}^{-1}\tilde{R}', \tag{A.16}$$

where $\tilde{D} = \text{diag}(\tilde{R})$ and $\tilde{\Lambda} = \text{diag}(\tilde{\Omega})$. Furthermore, under condition (A.16), the stationary density $p_0(z)$ is of exponential form $\exp(-\theta' z)$. Here,

$$\theta = -2n[\text{diag}(\Omega)]^{-1} D_1 R_1^{-1} \mu, \tag{A.17}$$

R_1 is the first $J - 1$ columns of R and $D_1 = \text{diag}(R_1)$. Notice that by Lemma 6.1, $R_1 = \tilde{R}_{|J}$ is admissible and hence is invertible.

Choice of reference density and basis functions. For the reference density, one may simply choose $q = 1$. However, that is not desirable most of the time. One reason for choosing a reference density q is to make condition (A.8) be satisfied. In theory, there is always such a choice. For example, if one chooses $q = p$, the resulting $r = 1 \in L^2$ can be recovered immediately. However, p is *unknown* a priori, so such a choice is unrealistic. Nevertheless, one may get some qualitative behaviors of p, such as the order of the singularities at the nonsmooth part of the boundaries, by some other means. Then one may build this information into the construction of q. In the two-dimensional case where $J = 3$, such information can indeed be obtained, at least for the driftless case; see [17]. Unfortunately, there are no analogous results for higher dimensional cases. Therefore, building singularities into the reference density q is not practical in general at this moment. However, there is a more important aspect that makes it necessary to introduce a reference density. The stationary density p_0 has large gradient when the drift vector $n\mu$ is large. ($n\mu$ is large whenever the network is not totally balanced and active job level n is large.) In fact, under a skew symmetry condition (A.15) on Ω and R, or equivalently (A.16) on $\tilde{\Omega}$ and $\tilde{R}$, the stationary density p_0 is of exponential form

$$p_0(z) = \kappa \exp(-\theta' z), \qquad z \in S, \tag{A.18}$$

where κ is a normalizing constant such that p_0 is a probability density and θ is given in (A.15). In general, θ is large in magnitude when $n\mu$ is large in magnitude, making the gradient of p_0 be large. Numerical experience shows that if we choose $q = 1$ for these cases, the convergence of the algorithm is very slow or it never converges at all. Formula (A.18) suggests that if we choose q to be the exponential in (A.18), then $r = 1$ or near 1 if the skew symmetry condition (A.15) or (A.16) is satisfied or nearly satisfied, and hence the convergence is very fast. Numerical experience thus far shows with this choice of q that the algorithm performs reasonably well *in general*, whether the condition (A.15) is satisfied or not.

Depending on the nature of the base functions f_i in (A.12), the algorithm can be further classified. If the f_i's are *piecewise* polynomials, then the algorithm is called a *finite element* algorithm. On the other hand, if the f_i's are *global* polynomials, then the algorithm is called a *spectral* algorithm. The finite element method has been used to solve a variety of PDE problems that have a variational formulation or, more generally, a Galerkin formulation; see [2]. The spectral method has been gaining increasing popularity in solving PDE's arising from meterology and fluid dynamics; see [15] and [5].

With the finite element approach, the normal matrix K in (A.13) is *sparse*, meaning that most of the entries in K are zeros, whereas with the spectral method, the matrix K is *full*. Further numerical investigations are needed in order to compare the efficiency and accuracy of these two approaches. For the purpose of this paper, we have implemented a spectral version of our algorithm by choosing

$$H_{s(k)} = \text{span of } \left\{\mathscr{A}f\colon f = z_1^{i_1} \cdots z_{J-1}^{i_{J-1}}, 0 < i_1 + i_2 + \cdots + i_{J-1} \le k\right\}.$$

Even for this particular subspace, how to choose a basis to get a well conditioned matrix K is quite an art. A basis naively chosen as $\{z_1, \dots, z_{J-1}, z_1^2, z_1 z_2, \dots, z_{J-2} z_{J-1}, z_{J-1}^2, \dots,\}$ is known to produce a K that has a big conditioning number, making the numerical solution of (A.13) unstable. However, it appears that for the test problems in Section 13, this implementation (with $k = 4$) performs reasonably well if we are interested in δ and the mean values of π, even with this poor choice of basis. However, it is known that a very high order of accuracy cannot be achieved with this basis.

No matter how one chooses a basis for the spectral method, one must be able to evaluate the following integrals efficiently:

$$\text{(A.19)} \qquad \begin{aligned} &\int_S z_1^{i_1} \cdots z_{J-1}^{i_{J-1}} \exp(-\theta' z)\, dz \quad \text{and} \\ &\int_{F_J} z_1^{i_1} \cdots z_{J-1}^{i_{J-1}} \exp(-\theta' z)\, d\sigma_J. \end{aligned}$$

When $\theta = 0$, which is the case when the network is fully balanced, it is easy to get

$$\int_S z_1^{i_1} \cdots z_{J-1}^{i_{J-1}}\, dz = \frac{i_1! \cdots i_{J-1}!}{(|i| + J - 1)!} \quad \text{and} \quad \int_{F_J} z_1^{i_1} \cdots z_{J-1}^{i_{J-1}}\, d\sigma_J = \frac{i_1! \cdots i_{J-1}!}{(|i| + J - 2)!},$$

where $|i| \equiv i_1 + \cdots + i_{J-1}$. In general, recursion formulas based on the divergence theorem can be found to compute these integrals.

Acknowledgments. We are indebted to Samuel Karlin for suggesting the direct proof of Lemma 3.1. We also thank Avi Mandelbaum for bringing reference 3 to our attention.

REFERENCES

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

[2] Becker, E. B., Carey, G. F. and Oden, J. T. (1981). *Finite Elements. The Texas Finite Element Series* **1**. Prentice-Hall, Englewood Cliffs, NJ.

[3] Berman, A. and Plemmons, R. J. (1979). *Nonnegative Matrices in the Mathematical Sciences*. Academic, New York.

[4] BERNARD, A. and EL KHARROUBI, A. (1991). Régulations déterministes et stochastiques dans le premier "orthant" de $\mathbb{R}^n$. *Stochastics Stochastics Rep.* **34** 149–167.

[5] CANUTO, C., HUSSAINI, M. Y., QUARTERONI, A. and ZANG, T. A. (1988). *Spectral Methods in Fluid Dynamics*. Springer, New York.

[6] CHEN, H., HARRISON, J. M., MANDELBAUM, A., VAN ACKERE, A. and WEIN, L. M. (1988). Empirical validation of a queueing network model for semiconductor wafer fabrication. *Oper. Res.* **36** 202–215.

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

[8] CHEVALIER, P. B. and WEIN, L. M. (1993). Scheduling networks of queues: Heavy traffic analysis of a multistation closed network. *Oper. Res.* **41**.

[9] DAI, J. G. and HARRISON, J. M. (1991). Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application. *Ann. Appl. Probab.* **1** 16–35.

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

[11] DAI, J. G. and KURTZ, T. G. (1993). Characterization of the stationary distribution for a semimartingale reflecting Brownian motion in a convex polyhedron. Unpublished.

[12] DAI, J. G. AND WANG, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems Theory Appl.* **13** 41–46.

[13] DAI, J. G. and WILLIAMS, R. J. (1993). Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. Unpublished manuscript.

[14] GORDON, W. J. and NEWELL, G. F. (1967). Closed queueing networks with exponential servers. *Oper. Res.* **5** 244–265.

[15] HALTINER, G. J. and WILLIAMS, R. T. (1980). *Numerical Prediction and Dynamical Meterology*. Wiley, New York.

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

[17] HARRISON, J. M., LANDAU, H. and SHEPP, L. A. (1985). The stationary distribution of reflected Brownian motion in a planar region. *Ann. Probab.* **13** 744–757.

[18] 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.

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

[20] HARRISON, J. M. and WEIN, L. M. (1990). Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* **38** 1052–1064.

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

[22] HARRISON, J. M. and WILLIAMS, R. J. (1987). Multidimensional reflected Brownian motions having exponential stationary distributions. *Ann. Probab.* **15** 115–137.

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

[24] HARRISON, J. M., WILLIAMS, R. J. and CHEN, H. (1990). Brownian models of closed queueing networks with homogeneous customer populations. *Stochastics* **29** 37–74.

[25] JACKSON, J. R. (1957). Networks of waiting lines. *Oper. Res.* **5** 518-521.

[26] JACKSON, J. R. (1963). Jobshop-like queueing systems. *Management Sci.* **10** 131–142.

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

[28] LAVENBERG, S. S. (1983). *Computer Performance Modeling Handbook*. Academic, New York.

[29] LAZOWSKA, E. D., ZAHORJAN, J., GRAHAM, G. S. and SEVCIK, K. C. (1984). *Quantitative System Performance*. Prentice-Hall, Englewood Cliffs, NJ.

[30] MANDELBAUM, A. (1993). The dynamic complementarity problem. *Math. Oper. Res.* To appear.

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

[32] PICH, M. (1992). Brownian models of open queueing networks with general station capabilities. Ph.D. thesis, Dept. Operations Research, Stanford Univ.

[33] REIMAN, M. I. (1990). Asymptotically exact decomposition approximations for queueing networks. *Oper. Res. Lett.* **9** 363–370.

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

[35] SAUER, C. H. and CHANDY, K. M. (1981). *Computer Systems Performance Modeling.* Prentice-Hall, Englewood Cliffs, NJ.

[36] SOLBERG, J. J. (1977). A mathematical model of computerized manufacturing systems. Paper presented at the 4th International Conference of Production Research, Tokyo.

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

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

[39] WHITT, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Sci.* **39**. To appear.

SCHOOL OF MATHEMATICS AND
INDUSTRIAL / SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305