Brownian Models of Feedforward Queueing Networks: Quasireversibility and Product Form Solutions

J. M. Harrison[1]
Stanford University

R. J. Williams[2]
University of California, San Diego

In Annals of Applied Probability, 2 (1992), 263–293.
This version is without figures. See the journal for a full version with figures.

Abstract

We consider a very general type of d-station open queueing network, with multiple customer classes and a more or less arbitrary service discipline at each station, but restricted by the requirement that customers always flow from lowered numbered stations to higher numbered ones. To approximate the behavior of such a queueing network under heavy traffic conditions, a corresponding *Brownian network model* is proposed, and it is shown that the approximating Brownian model reduces to a d-dimensional reflected Brownian motion W whose state space is the non-negative orthant. A necessary and sufficient condition for W to have a product form stationary distribution (that is, a stationary distribution with independent components), and a probabilistic interpretation for that condition, are given. Our interpretation involves a notion of *quasireversibility* analogous to that introduced by F. P. Kelly and elaborated by J. Walrand in their brilliant analysis of product form solutions for conventional queueing network models. Three illustrative queueing network models are discussed in detail, and the analysis of these examples shows how a Brownian network approximation may have a product form stationary distribution even when the original or exact model is intractable. Particularly intriguing in that regard are two examples involving non-Poisson inputs, deterministic routing, deterministic service times and processor sharing service disciplines.

AMS 1980 subject classifications. Primary 60J60, 60J70, 60K25.

[1] Research supported in part by Semiconductor Research Corporation Grant 84-01-046.
[2] Research supported in part by NSF Grants DMS 8657483 and 8722351, a grant from AT&T Bell Laboratories, and an Alfred P. Sloan Research Fellowship.

Key words and phrases. Multiclass queueing networks, feedforward, quasireversibility, product form, Brownian models.

1. Introduction

The object of study in this paper is a d-dimensional diffusion process $W = \{W(t), t \geq 0\}$ whose state space is the non-negative orthant. To be more specific, W is a d-dimensional *reflected Brownian motion,* also called regulated Brownian motion [5] or just RBM, the data for which are a d-dimensional drift vector μ, a $d \times d$ covariance matrix Σ, and a $d \times d$ *reflection matrix* R. A distinguishing feature of this paper is that we restrict attention to the case where R is lower triangular.

Apart from their intrinsic mathematical interest, processes like W are studied because they arise as diffusion approximations for the workload processes and queue length processes associated with open queueing networks [4, 17, 10, 16]. Thus the process W, or more often a family of processes that includes W, will be referred to in this paper as a *Brownian network model.* As we will explain later, an RBM with a lower triangular reflection matrix corresponds to what is called a *feedforward* queueing network, in which customers always flow from lower numbered stations to higher numbered ones.

It is the stationary or "steady-state" characteristics of a queueing model that are usually of greatest interest, and virtually all of the models that have been successfully analyzed in classical queueing network theory are models having a so-called "product form" stationary distribution. In the case of open networks, this means that the stationary distribution of the entire system is the product of independent marginal distributions associated with the individual stations. This state of affairs is often described by the statement that individual stations are "independent in equilibrium". Such queueing network models are said to have a "product form solution", and they are frequently referred to as "product form networks". For an approximating Brownian network model, the analogous property is that the stationary distribution of W be the product of independent marginal distributions for $W_1, \ldots, W_d$.

From extant results it is relatively easy to show that our Brownian network model has a product form stationary distribution if and only if the covariance matrix and reflection matrix of W satisfy a certain algebraic equation. Until now there has been no probabilistic interpretation of the algebraic condition, but in this paper we explain it in terms of *quasireversibility,* an analog of the probabilistic notion introduced by Kelly [13, 14] and elaborated by Walrand [22] in their brilliant analysis of product form solutions for conventional queueing network models.

For the simple case of a single Brownian service station, our analog of the Kelly-Walrand definition of quasireversibility was introduced and analyzed in the recent paper [11]. Here that notion is transported to the broader setting of *feedforward* Brownian network models, and it leads to a clear and simple interpretation of findings that previously seemed mysterious. Unfortunately, our analysis does not extend in any obvious way to networks with feedback, and treatment of the general case is left as a topic for future research.

The paper is organized as follows. In section 2 we define precisely the process W under study, and then in section 3 the algebraic condition required for a product form stationary distribution is derived. In section 4 we explain what is meant by a quasireversible station in the context of a feedforward Brownian network model. In section 5 we combine results from preceding sections to conclude that such a Brownian model has a product form stationary distribution if and only if each station is quasireversible, and we give a direct probabilistic proof of the "if" part of this result. Sections 6 through 8 are devoted to the analysis of three illuminating examples.

Section 8 is followed by a lengthy appendix in which we explain how parameters of a conventional queueing network model are used to determine the data of an approximating Brownian model. To be more precise, we consider a very general type of open queueing network and propose a scheme for associating with each such network a "natural" Brownian approximation. A limit theorem is described that would rigorously justify the proposed approximation scheme, but its proof is left as an open research problem. (Actually, there are several important open problems mentioned at different points in the appendix, and all of these must be resolved if a comprehensive limit theorem is to be proved.) Brownian approximations for multiclass open queueing networks have been proposed and discussed earlier [6, 7], but the treatment given in the appendix is more general, somewhat more complete and explicit, and slightly different in style. The most important added generality is the allowance of service disciplines other than first-in-first-out. In particular, the current treatment allows service stations with a processor sharing discipline, and processor sharing figures prominently in our examples (see sections 6 and 7). Most readers will want to at least scan the appendix before starting section 2, and to understand the examples discussed in sections 6 through 8 one must make frequent reference to the appendix.

For the most part, the mathematical development in sections 2 through 5 consists of recalling definitions, adapting old results to establish several preliminary propositions, and then assembling the pieces in a more or less obvious way. Strictly speaking, however, all of the results are new in at least some minor, technical sense, and there is one important new contribution that arises in section 4 and may not be evident to all readers. Our previous paper [11] dealt with a single Brownian service station that processes several *classes of customers,* and in a network context it is not at all obvious how to interpret the italicized phrase. In the definitional system advanced in section 4, what plays the role of a "customer class" is workload content for a particular downstream server, and this is essential for the sharpness of our final result (Theorem 5.1). If one defines quasireversibility of individual stations in terms of customer classes that were meaningful in the original queueing model, a much weaker theory is eventually obtained — a theory in which quasireversibility is sufficient but not necessary for a product form solution. In fact, a secret of success in formulating the Brownian model is to suppress all fine structure that may have been present in the original queueing model, taking as given just the drift vector, covariance matrix and reflection matrix of W.

It will become apparent that the conditions yielding a product form stationary distribution for a Brownian network model are very special, and readers might well ask why so much effort is being expended on this apparently narrow subject. One important reason is our general desire to establish solid, concrete connections between conventional queueing network models on the one hand, and Brownian network models on the other. Product form queueing networks are widely taught and widely accepted by even the most practically-oriented engineers as useful tools for system performance analysis [15, 20]. In contrast, "diffusion approximations" for complex queueing systems are often relegated to the category of inaccessible arcana, even by queueing theorists with a relatively high tolerance for mathematical theory. By elaborating on the one subject familiar to all students of queueing network theory — product form stationary distributions — we hope to hasten the acceptance of Brownian models as a "mainstream" topic in performance analysis.

We conclude this introduction with an account of some notational and terminological conventions used in this paper. Vectors, including the values of vector-valued processes, are regarded as column vectors. Vector (in)equalities are to be interpreted componentwise and a vector-valued function is non-decreasing (or non-increasing) if and only if each component has that property. For a vector v, $\mathrm{diag}(v)$ will denote the diagonal matrix whose diagonal entries are given by the components of v, and for a square matrix M, $\mathrm{diag}(M)$ will denote the diagonal matrix with the same diagonal entries as M. An n-dimensional process X will be called a (μ, Σ) Brownian motion if it is a Brownian motion with constant drift vector $\mu \in \mathbb{R}^n$ and $n \times n$ covariance matrix Σ.

2. The Brownian Network Model

In the appendix we describe a very general open queueing network model with multiple customer classes, arbitrary interarrival and service time distributions, and a more or less arbitrary queue discipline at each of the d nodes or stations that constitute the network ($d \geq 1$). As explained in the appendix, one may approximate such a queueing system by a corresponding *Brownian network model,* which is defined from a given Brownian motion ξ and a given random vector $W(0)$ by the following five relationships:

(2.1) $\quad W(t) = W(0) + \zeta(t) + Y(t),\ t \geq 0,$

(2.2) $\quad \zeta(t) = \xi(t) - G[W(t) - W(0)],\ t \geq 0,$

(2.3) $\quad W(t) \geq 0,\ t \geq 0,$

(2.4) $\quad Y(\cdot)$ is continuous and non-decreasing with $Y(0) = 0$,

and

(2.5) $\quad Y_i(\cdot)$ can increase only at times t for which $W_i(t) = 0, \quad i = 1, \ldots, d.$

The primitive elements of the Brownian network model are (i) a d-dimensional Brownian motion $\xi = \{\xi(t), t \geq 0\}$ called the *total workload netflow process,* (ii) a non-negative random d-vector $W(0)$ representing the *initial workload vector,* and (iii) a non-negative $d \times d$ matrix $G = (G_{ij})$ called the *workload contents matrix.* In contrast to the general system model described in the appendix, we assume in the body of the paper that G is lower triangular (that is, $G_{ij} = 0$ if $j \geq i$), corresponding to a so-called feedforward queueing network, where customers always flow from lower numbered stations to higher numbered ones, with no loops or cycles.

The drift vector of ξ is denoted by $-\theta$ (the reason for this sign convention will become apparent shortly), its covariance matrix is denoted by Γ, and the initial value is $\xi(0) = 0$. It is required that $W(0)$ and ξ be independent, and to avoid trivial complications we assume throughout that

(2.6) $\theta > 0$

and

(2.7) Γ is non-degenerate.

The relationships (2.1)–(2.5) serve to define (see below) two d-dimensional stochastic processes $W = \{W(t),\ t \geq 0\}$ and $Y = \{Y(t),\ t \geq 0\}$ in terms of primitive model elements. As explained in the appendix, $Y_i(t)$ represents cumulative server idleness at station i up to time t, and $W_i(t)$ represents the amount of work for servers at station i that is embodied in customers who occupy that station at time t. One interprets $\xi_i(t)$ as the amount of work for servers at station i that is embodied in customers who enter the network by time t (not all such customers will have reached station i by that time), minus the total amount of work that servers at station i are capable of completing by time t if they are never idle. One interprets θ_i as the service capacity at station i minus the average rate of workload input there, so (2.6) is a natural stability condition.

One interprets $G_{ij}W_j(t)$ as the amount of future work for servers at station i that is embodied in customers who occupy station j $(1 \leq j \leq i \leq d)$ at time t; thus the quantity $\zeta_i(t)$ defined by (2.2) represents the amount of work for servers at station i that is embodied in customers who have entered station i by time t, minus the total amount of work that servers at station i are capable of completing by time t if they are never idle. We call ζ the *immediate workload netflow process* for our Brownian network model.

To conclude this description of the Brownian network model, we now use the assumed triangularity of G to write out recursive formulas for ζ, Y and W in terms of our primitive model elements. From (2.1)–(2.5), we have for each $i = 1, \ldots, d$,

(2.8) $\zeta_i(t) = \xi_i(t) - \sum_{j<i} G_{ij}[W_j(t) - W_j(0)], \quad t \geq 0,$

(2.9) $W_i(t) = W_i(0) + \zeta_i(t) + Y_i(t), \quad t \geq 0,$

(2.10) $W_i(t) \geq 0, \quad t \geq 0,$

(2.11) $Y_i(\cdot)$ is continuous and non-decreasing with $Y_i(0) = 0$,

(2.12) $Y_i(\cdot)$ can increase only at times t for which $W_i(t) = 0$.

Putting $i = 1$ in (2.8) gives

$$(2.13) \qquad \zeta_1(t) = \xi_1(t), \quad t \geq 0,$$

and for each $i = 1, \ldots, d$, given ζ_i, it is known (see Chung-Williams [3], section 8.2) that the unique process Y_i satisfying (2.9)–(2.12) is

$$(2.14) \qquad Y_i(t) = \big(- \min_{0 \leq s \leq t} [W_i(0) + \zeta_i(s)]\big)^+, \quad t \geq 0.$$

Putting $i = 1$ in (2.14) and (2.9) gives us the constructive definition of Y_1 and W_1, respectively, and then for $i = 2, \ldots, d$, we use equations (2.8), (2.14) and (2.9) to define ζ_i, Y_i and W_i, respectively, by means of the obvious induction on i.

For future purposes it will be useful to note that, in our original description (2.1)–(2.5) of the Brownian system model, equations (2.1) and (2.2) can actually be compressed into the single relationship

$$(2.15) \qquad (I + G)W(t) = (I + G)W(0) + \xi(t) + Y(t), \quad t \geq 0.$$

What has been shown in the previous paragraphs is that (2.15) and (2.3)–(2.5) together uniquely define Y and W via a path-to-path mapping from $W(0)$ and ξ. We note that (Y, W) is adapted to the filtration generated by $W(0)$ and ξ.

3. RBM in an Orthant

Recall from the previous section that the immediate workload process W satisfies equation (2.15) where ξ is a $(-\theta, \Gamma)$ Brownian motion, $W(0) \geq 0$ and Y, W satisfy (2.3)–(2.5). Since the matrix G is lower triangular, $G^n = 0$ for all $n \geq d$, and so $I + G$ is invertible with

$$(3.1) \qquad R \equiv (I + G)^{-1} = I - G + G^2 - \ldots + (-G)^{d-1}.$$

Observe that $R_{ii} = 1$ for $i = 1, \ldots, d$. Let $X = R\xi$ and

$$(3.2) \qquad \mu = -R\theta \quad \text{and} \quad \Sigma = R\Gamma R'.$$

Then (2.15) is equivalent to

$$(3.3) \qquad W = W(0) + X + RY,$$

where X is a (μ, Σ) Brownian motion satisfying $X(0) = 0$. Conditions (3.3), (2.3)–(2.5) together are equivalent to (2.15), (2.3)–(2.5), and so given $W(0)$ and X, the pair of processes (Y, W) is the unique solution of (3.3), (2.3)–(2.5), and by construction and the invertibility of R, these processes are adapted to the filtration generated by $W(0)$, X. The process W is called a reflecting Brownian motion with data (S, μ, Σ, R), where $S = \mathbb{R}_+^d$ is the state space of the process. It behaves like a Brownian motion with drift μ and covariance matrix Σ in the interior of the positive orthant S. When W hits the boundary face $F_i = \{x \in \mathbb{R}_+^d : x_i = 0\}$ of S, the ith component Y_i of Y increases to give an instantaneous push to W in the direction of the ith column of R, so as to keep W in the orthant.

For the following, we need to be more precise about the probability space on which W is defined. We can realize W on the path space $\Omega \equiv C\big([0,\infty], \mathbb{R}^d\big)$ by letting $W(0) = \omega(0)$, $\xi(\cdot) = \omega(\cdot) - \omega(0)$ for all $\omega \in \Omega$, and $\{P_x,\ x \in S\}$ be the family of probability measures on $(\Omega, \mathcal{M})$, where $\mathcal{M} = \sigma\{\omega(s) : 0 \le s < \infty\}$, such that for each $x \in S$, under P_x, $\omega(\cdot)$ is a $(-\theta, \Gamma)$ Brownian motion starting from x. Then W, as defined in section 2 from $W(0)$ and ξ, together with $\{P_x,\ x \in S\}$, defines a strong Markov process on $(\Omega, \mathcal{M})$. The strong Markov property comes from the fact that by construction, for any stopping time $\tau < \infty$, $W(\tau + \cdot)$ is defined by the same path-to-path map applied to $(W(\tau), \xi(\cdot + \tau) - \xi(\tau))$ that defines $W(\cdot)$ from $(W(0), \xi(\cdot))$. Henceforth by W we shall mean the strong Markov process as defined above. We are interested in stationary distributions for W.

In a similar manner to that in sections 7 and 8 of [10], we can show the following proposition. Here two measures are equivalent if they are mutually absolutely continuous and the symbol $\approx$ will be used to denote such an equivalence. Let σ_i denote $(d-1)$-dimensional Lebesgue measure (i.e., surface measure) on the ith face $F_i \equiv \{x \in S :\ x_i = 0\}$ of S, and let $b\mathcal{B}(F_i)$ denote the set of real-valued bounded Borel measurable functions on F_i. Let $C_b^2(S)$ denote the set of real-valued functions that are twice continuously differentiable on some domain containing S and that together with their first and second partial derivatives are bounded on S.

Proposition 3.1. *Suppose π is a stationary distribution for W. Then π is unique and is equivalent to Lebesgue measure on S. Moreover, for each $i \in \{1, \ldots, d\}$, there is a finite Borel measure ν_i on F_i such that $\nu_i \approx \sigma_i$ and*

$$(3.4) \qquad E_\pi\left[\int_0^t f\big(W(s)\big) dY_i(s)\right] = t \int_{F_i} f d\nu_i \quad \text{for all} \quad f \in b\mathcal{B}(F_i),$$

where E_π *denotes expectation under* $P_\pi \equiv \int\limits_S \pi(dx)P_x$; *and for each* $f \in C_b^2(S)$,

(3.5) $\quad \int_S Lf \; d\pi \; + \sum_{i=1}^d \int_{F_i} D_i f \; d\nu_i = 0$

where

(3.6) $$Lf = \frac{1}{2}\sum_{i,j=1}^d \Sigma_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_{i=1}^d \mu_i \frac{\partial f}{\partial x_i},$$

(3.7) $\quad D_i f = v_i \cdot \nabla f, \quad i = 1, \ldots, d,$

for v_i *equal to the* i*th column of the "reflection" matrix* R.

Proof. The proof that π is unique and is equivalent to Lebesgue measure on S is the same as that in section 7 of [10]. The remainder of the proposition has the same statement as Theorem (8.1) in [10], with the exception that ν_i here is twice what it is in [10]. This difference in scale factor, which makes for a cleaner statement of the above proposition, does not affect the validity of the result. The proof given in [10] carries over for the R-matrices considered here once Lemmas (8.4) and (8.7) there are verified. For this, observe that W together with $\{P_x, x \in S\}$ defined on $(\Omega, \mathcal{M}, \{\mathcal{M}_t\})$, where $\mathcal{M}_t = \sigma\{\omega(s) : 0 \le s \le t\}$, is an SRBM with data (S, μ, Σ, R), as defined in [19]. It follows from [19] that Lemma (8.7) of [10] holds and that R is a completely-$\mathcal{S}$ matrix (see [19] for the definition). By the latter and Lemma 1 of Bernard and El Kharroubi [2], there is a constant $C_1 > 0$ such that for all i and t,

(3.8) $\quad Y_i(t) \le C_1 \max_{0 \le s \le t} |X(s) - X(0)|,$

and so there is a constant $C_2 > 0$ such that

$$E_x[Y_i(t)] \le C_2(t+1) \text{ for all } i, \ t \text{ and } x,$$

where E_x denotes expectation under P_x. We remark that in the case treated here, where G is lower triangular, one can verify (3.8) directly from the construction of Y_i and induction on i. Thus Lemma (8.4) of [10] holds for the RBM's treated here and so the measures ν_i are finite. The rest of the proof given in [10] goes through to yield (3.4) and (3.5). The latter comes from taking expectations under π in Itô's formula applied to f and W. □

Definition. We say that W has a *product form stationary distribution*, or that the Brownian network model has a *product form solution*, if W has a stationary distribution π whose density p relative to Lebesgue measure on S is of the form

$$(3.9) \qquad p(x) = p_1(x_1)\cdots p_d(x_d) \qquad \text{for } x = (x_1, \dots, x_d) \in S,$$

where $p_1, \dots, p_d$ are probability densities relative to Lebesgue measure on $\mathbb{R}_+$.

In [10], Brownian models of single-class open queueing networks were studied. The RBM's arising there are characterized as unique solutions of (3.3), (2.3)–(2.5) for R-matrices of the form $R = I - P'$, where P is a $d \times d$ matrix with non-negative entries, zeros on the diagonal and spectral radius strictly less than one. It was shown in [10] that a necessary and sufficient condition for these RBM's to have product form stationary distributions is that

$$(3.10) \qquad 2\Sigma_{ij} = -(P_{ji}\Sigma_{jj} + P_{ij}\Sigma_{ii}) \quad \text{for all } j \neq i,$$

or equivalently, that

$$(3.11) \qquad 2\Sigma = RD + DR',$$

where $D = \text{diag}\ (\Sigma)$. The following theorem is an analog of the result in [10] for the feedforward multiclass Brownian models considered in this paper. The class of RBM's arising here has non-empty intersection with that considered in [10], but it is by no means contained within it. The proof of Theorem 3.1 below parallels that in [10], so we shall not repeat the details here, but simply give an outline of the argument. The primary purpose of this paper is to give a probabilistic interpretation of the algebraic condition (3.12).

Theorem 3.1. *W has a product form stationary distribution if and only if*

$$(3.12) \qquad \Gamma_{ij} = \tfrac{1}{2} G_{ij}\Gamma_{jj} \quad \text{for} \quad 1 \leq j < i \leq d.$$

When (3.12) holds, the stationary distribution of W has the exponential density function (relative to Lebesgue measure):

$$(3.13) \qquad p(x) = \textstyle\prod_{i=1}^{d} \gamma_i \exp(-\gamma_i x_i) \quad \text{for} \quad x = (x_1, \dots, x_d) \in S,$$

where

$$(3.14) \qquad \gamma_i = 2\theta_i/\Gamma_{ii}, \quad i = 1, \dots, d.$$

Proof. For this proof only, let $\Lambda = \mathrm{diag}(\Gamma)$. We first verify that (3.11) is equivalent to

$$2\Gamma = \Lambda(I + G') + (I + G)\Lambda, \tag{3.15}$$

and if either (3.11) or (3.15) holds, $\Sigma_{ii} = \Gamma_{ii}$ for all i. First, suppose (3.11) holds. Premultiplying (3.11) by $R^{-1} = I + G$ and postmultiplying by $(R')^{-1}$ yields

$$2\Gamma = D(I + G') + (I + G)D. \tag{3.16}$$

But, since D is diagonal and G has zeros on its diagonal, this implies $\Gamma_{ii} = D_{ii} \equiv \Sigma_{ii}$ and substituting this in (3.16) yields (3.15). Thus (3.11) implies (3.15) and $\Sigma_{ii} = \Gamma_{ii}$ for all i. Similarly, by premultiplying (3.15) by $R = (I + G)^{-1}$ and postmultiplying by R', one can show that (3.15) implies (3.11) and $\Sigma_{ii} = \Gamma_{ii}$ for all i. We also note here that (3.12) is equivalent to (3.15), by the symmetry of Γ and triangularity of G.

The "if" part of Theorem 3.1 follows as in [10, Theorem (9.23)]. We briefly sketch the proof here. By performing a linear transformation of coordinates, we can transform W to an RBM in a polyhedral cone with covariance matrix equal to the identity. In [24], a sufficient condition for such an RBM to have an exponential form stationary distribution was given. When this condition is transformed back to the orthant (see [10, p. 110] with $D = I$ and $H = \Lambda^{-\frac{1}{2}}$ there), it becomes (3.11) and under this condition W has the stationary density given by (3.13)–(3.14), with Σ_{ii} in place of Γ_{ii} there. By the discussion in the first paragraph of this proof, (3.12) is equivalent to (3.11) and in the case that it holds, $\Sigma_{ii} = \Gamma_{ii}$. It follows that (3.12) is sufficient for W to have a product form stationary distribution, which is then given by (3.13)–(3.14).

For the "only if" part of the theorem, suppose π is a stationary distribution for W. Then (3.5) holds by Proposition 3.1. On substituting exponential functions into (3.5), as in [10, Theorem (9.3)], one derives a relationship between the Laplace transform of π and the Laplace transforms of the boundary measures ν_i. In precisely the same manner as in [10], one concludes from this that π is of product form only if its density p is of the exponential form (3.13)–(3.14) with Σ_{ii} in place of Γ_{ii} there, and that (3.11) holds. Again, from the first paragraph of this proof, it follows that (3.12) is necessary for W to have a product form stationary distribution and in this case (3.13)–(3.14) hold. □

4. Quasireversibility of a Brownian Service Station

To develop our interpretation of (3.12), it will be useful to consider the subnetwork composed of stations k through d only. Throughout this section k will be a fixed integer satisfying $1 \leq k < d$, and $\alpha = (\alpha_k, \ldots, \alpha_d)'$ will have the same distribution as $\xi^k \equiv (\xi_k, \ldots, \xi_d)'$, i.e., α will be a $(d-k+1)$-dimensional Brownian motion with drift vector $-\theta^k$ and covariance matrix Γ^k and $\alpha(0) = 0$, where $\theta^k = (\theta_k, \ldots, \theta_d)'$ and $\Gamma^k = (\Gamma_{ij})_{k \leq i,j \leq d}$. The process α represents the total workload netflow to the subnetwork under study, i.e., α plays the same role for this subnetwork as does ξ for the entire network.

Now let $Z(0)$ be a non-negative random variable that is independent of α, and define the one-dimensional RBM

(4.1) $\quad Z(t) = Z(0) + \alpha_k(t) + L(t), \quad t \geq 0,$

where

(4.2) $$L(t) = \left(- \min_{0 \leq s \leq t} \big(Z(0) + \alpha_k(s) \big) \right)^{+}.$$

Thus Z and L are defined in terms of $Z(0)$ and α_k in the same way that W_k and Y_k are defined in terms of $W_k(0)$ and ζ_k. (The relevant connection will be explained in the next section.) Since $\theta_k > 0$, Z has a (unique) stationary distribution with density

(4.3) $\quad p_k(x_k) = \gamma_k \exp(-\gamma_k x_k) \quad \text{for} \quad x_k \in \mathbb{R}_+,$

where

(4.4) $\quad \gamma_k = 2\theta_k / \Gamma_{kk}.$

Now suppose that $Z(0)$ is randomized to have the stationary distribution of Z. Let $\beta \equiv (\beta_{k+1}, \ldots, \beta_d)'$ be defined by

(4.5) $\quad \beta_i(t) = \alpha_i(t) - G_{ik}(Z(t) - Z(0)), \quad t \geq 0,\ i = k+1, \ldots, d.$

Theorem 4.1. *The following two statements are equivalent:*

(4.6) $\quad \{\beta(s) : 0 \leq s \leq t\}$ *is independent of* $Z(t)$ *for each fixed* $t > 0$.

(4.7) $\quad \Gamma_{ik} = \frac{1}{2} G_{ik} \Gamma_{kk} \quad$ *for* $\quad i = k+1, \ldots, d.$

Moreover, whenever (4.6)–(4.7) hold, we have

(4.8) $\quad \beta$ *is a Brownian motion with the same distribution as* ξ^{k+1}.

In fact, if $G_{ik} \neq 0$ *for some* $i \in \{k+1, \ldots, d\}$, *then (4.8) is equivalent to (4.6)–(4.7).*

Proof. Consider the model of a Brownian service station described in [11] with $A(t) = \alpha(t) + \tau t, \tau = (1, 0, \ldots, 0)'$, $\nu = 0$, $\delta = (1, G_{k+1,k}, \ldots, G_{dk})'$ and $N = [0 \mid I]$ there, where in the latter, 0 is the $(d-k)$-dimensional column vector of all zeros and I is the $(d-k) \times (d-k)$ identity matrix. We note that in [11], A was assumed to have a positive drift and here A will have some drift components that are negative. However, all that is important for the application of the results in [11] is that $X = \alpha_k$ there has a negative drift, which it will since the drift of α_k is $-\theta_k < 0$. With these choices, the processes, W, Y and ND in [11] correspond to Z, L and β here (note that the $+$ in equation (2.15) in [11] should be a $-$). Thus, by the proof of Theorem 3.1 in [11], (4.6) is equivalent to

(4.9) $\quad N\Gamma^k \tau = \frac{1}{2}(\tau' \Gamma^k \tau) N\delta,$

which in turn is equivalent to (4.7). Moreover, it is shown in the proof in [11] that (4.9) implies ND is a Brownian motion with the same distribution as NA, and that the converse holds if $N\delta$ has at least one non-zero component. It follows that (4.7) implies (4.8) and that the converse is true if $G_{ik} \neq 0$ for some $i \in \{k+1, \ldots, d\}$. $\square$

Definition. We say that station k is *quasireversible*, or is quasireversible when viewed in isolation, if (4.6)–(4.7) hold.

Remark. To paraphrase, quasireversibility of station k means that if we feed the station a vector Brownian "input" α distributed as ξ^k, then in equilibrium it produces a vector Brownian "output" β. Moreover β is distributed as ξ^{k+1} and $\{\beta(s) : 0 \leq s \leq t\}$, the output up to time t, is independent of $Z(t)$, the station's "state description" at time t.

5. Quasireversibility and Product Form Solutions

Comparing condition (3.12) with (4.7), we see that Theorem 3.1 can be restated in the following form.

Theorem 5.1. *The Brownian network model has a product form solution if and only if each station $k = 1, 2, \ldots, d-1$, is quasireversible. In this case, the stationary distribution of W has density function $p(x) = p_1(x_1) \ldots p_d(x_d)$ where $p_k(x_k)$ is the exponential density given by (4.3)–(4.4) for $k = 1, \ldots, d$.*

Remark. Since (4.7) is vacuous when $k = d$, let us say as a matter of definition that station d is automatically quasireversible. Then Theorem 5.1 can be stated more succinctly as follows. The Brownian network model has a product form solution if and only if each station is quasireversible.

Theorem 5.1 is essentially a restatement of Theorem 3.1, and it fulfills our promise of a "probabilistic interpretation" for the algebraic condition (3.12). To reinforce that interpretation, we now give a proof of the "if" part of Theorem 5.1.

Proof of the "if" part of Theorem 5.1. Suppose that for each $k \in \{1, \ldots, d-1\}$, station k is quasireversible, i.e., (4.7) holds. First consider station 1. We have $\zeta_1 = \xi_1$. By setting $k = 1$ and $\alpha = \xi$ in section 4, we obtain that Z, L there equal W_1, Y_1. Hence W_1 has stationary density p_1 given by (4.3) with $k = 1$, and by Theorem 4.1, since (4.7) holds, when W_1 is initialized with this stationary distribution we have for $\beta^1 \equiv (\beta^1_2, \ldots, \beta^1_d)'$ where $\beta^1_i \equiv \xi_i - G_{i1}(W_1 - W_1(0))$, $i = 2, \ldots, d$, β^1 is a $(-\theta^2, \Gamma^2)$ Brownian motion and for each $t > 0$, $\{\beta^1(s) : 0 \le s \le t\}$ is independent of $W_1(t)$.

For a proof by induction on the station index, we make the induction hypothesis that for j fixed such that $1 \le j < d$, when $W_1(0), \ldots, W_j(0)$ are independent random variables (also independent of ξ) with the density of $W_k(0)$ being given by (4.3) for $k = 1, \ldots, j$, then

(5.1) $\quad \big(W_1(t), \ldots, W_j(t)\big)$ has the same distribution as $\big(W_1(0), \ldots, W_j(0)\big)$ for each $t > 0$,

and for $\beta^j \equiv (\beta^j_{j+1}, \ldots, \beta^j_d)$ defined by

(5.2) $\quad \beta^j_i = \xi_i - \sum_{l \le j} G_{il}(W_l - W_l(0)), \quad i = j+1, \ldots, d,$

(5.3) $\quad \beta^j$ is a $(-\theta^{j+1},\ \Gamma^{j+1})$ Brownian motion and for each $t > 0$, $\{\beta^j(s) : 0 \le s \le t\}$ is independent of $\big(W_1(t), \ldots, W_j(t)\big)$.

Let $W_1(0), \ldots, W_j(0)$ have the distributions described above. Setting $k = j+1$ and $\alpha = \beta^j$

in section 4, we see that $\alpha_k = \beta^j_{j+1} = \xi_{j+1} - \sum\limits_{l<j+1} G_{j+1,l}(W_l - W_l(0)) = \zeta_{j+1}$ and hence Z, L in section 4 equal W_{j+1}, Y_{j+1}. Thus, W_{j+1} has stationary density p_{j+1} given by (4.3) with $k = j+1$. Suppose $W_{j+1}(0)$ is chosen independent of $\big(W_1(0), \ldots, W_j(0), \xi\big)$ with density p_{j+1}. Then $W_{j+1}(t)$ has the same distribution as $W_{j+1}(0)$ and by Theorem 4.1, for $\beta^{j+1} \equiv (\beta^{j+1}_{j+2}, \ldots, \beta^{j+1}_d)'$, defined by

$$\begin{aligned}\beta^{j+1}_i &= \alpha_i - G_{i,j+1}\big(W_{j+1} - W_{j+1}(0)\big) \\ &= \xi_i - \sum_{l \le j+1} G_{il}\big(W_l - W_l(0)\big), \;\; i = j+2, \ldots, d,\end{aligned}$$

β^{j+1} is a $(-\theta^{j+2}, \Gamma^{j+2})$ Brownian motion, and for each $t > 0$, $\beta^{j+1}\big|_{[0,t]}$ is independent of $W_{j+1}(t)$. Now for $t > 0$ fixed, $\Big(W_{j+1}(0), \alpha\big|_{[0,t]}\Big)$ is independent of $\big(W_1(t), \ldots, W_j(t)\big)$. Since $W_{j+1}(t)$ is determined by $\Big(W_{j+1}(0),\ \alpha_k\big|_{[0,t]}\Big)$, it follows that $W_{j+1}(t)$ is independent of $\big(W_1(t), \ldots, W_j(t)\big)$ and hence $\big(W_1(t), \ldots, W_{j+1}(t)\big)$ has the same product form distribution as $\big(W_1(0), \ldots, W_{j+1}(0)\big)$. Also $(\beta^{j+1},\ W_{j+1})\big|_{[0,t]}$ is determined by $\Big(W_{j+1}(0),\ \alpha\big|_{[0,t]}\Big)$ and so is independent of $\big(W_1(t), \ldots, W_j(t)\big)$. In turn, $\beta^{j+1}\big|_{[0,t]}$ is independent of $W_{j+1}(t)$, and so $\beta^{j+1}\big|_{[0,t]}$ is independent of $\big(W_1(t), \ldots, W_{j+1}(t)\big)$.

Summarizing the above, we have shown that when $W_1(0), \ldots, W_{j+1}(0)$ are independent (and independent of ξ) with the density of $W_k(0)$ being given by (4.3) for $k = 1, \ldots\, j+1$, then (5.1) and (5.3) hold with $j+1$ in place of j. This completes the induction step. It follows by induction that when $W_1(0), \ldots, W_d(0)$ are chosen independent, with the density of $W_k(0)$ given by (4.3) for $k = 1, \ldots, d$, then $\big(W_1(t), \ldots, W_d(t)\big)$ has the same distribution as $\big(W_1(0), \ldots, W_d(0)\big)$, i.e., the product form density $p(x) = p_1(x_1) \ldots p_d(x_d)$, where $p_k(x_k)$ is given by (4.3) for $k = 1, \ldots, d$, is a stationary density for W, and hence W has a product form stationary distribution. $\square$

6. A Simple Example

The remainder of this paper is devoted to analysis of three examples, all of which are queueing networks of the type described in the appendix. In the context of each example we will discuss parameter combinations that yield a product form stationary distribution for the approximating Brownian network model. That is, we discuss parameter combinations such that the covariance matrix Γ and the workload contents matrix G of the approximating

Brownian network model jointly satisfy the product form condition (3.12). The matrices Γ and G will be calculated from elemental model parameters by means of formulas (A.51) and (A.49), respectively. These formulas do not involve the numbers of servers at the various stations, so we will not specify values for the parameters $c_1, \ldots, c_d$ except to say that each station is assumed to have enough servers to satisfy the stability condition $\theta_i > 0$ where θ_i is defined by (A.14).

As a first example, consider the network pictured in Figure 1. This is a *generalized Jackson network*, where the number of customer classes n equals the number of service stations d. That is, in a generalized Jackson network there is a single customer class associated with each service station, and customers change *station* in Markovian fashion. For the series network pictured in Figure 1, the switching probabilities are $P_{12} = P_{23} = 1$ and $P_{ij} = 0$ otherwise. Following the notational convention used in the appendix, we denote by τ_i and b_i the mean and the coefficient of variation, respectively, for the service time distribution at station i. Also, let α be the exogenous input rate to station 1, and let a be the coefficient of variation for the interarrival time distribution. It follows that the average arrival rate of customers to each station i is $\lambda_i = \alpha$. (In general, λ_i denotes the average arrival rate to customer class i, but in this case there is a one-to-one correspondence between customer classes and service stations.)

Figure 1.

Three Queues in Series

Let us now consider formula (A.51) for the asymptotic covariance matrix Γ of the total workload netflow process ξ. For a generalized Jackson network the constituency matrix C is simply $C = I$ (the $d \times d$ identity matrix). Also, recalling that H is the asymptotic covariance matrix of the "switching noise" process V defined by (A.26), readers may verify that $H = 0$ for any multiclass network with deterministic routing. That is, $H = 0$ whenever $P_{ij} = 0$ or 1 for all (i, j) pairs. Thus, for the network under discussion, (A.51) reduces to

$$\Gamma = \Lambda + (TB)K(TB)' \tag{6.1}$$

where

$$(6.2)\qquad \Lambda = \text{diag }(\alpha\tau_1^2 b_1^2,\ \alpha\tau_2^2 b_2^2,\ \alpha\tau_3^2 b_3^2),$$

$$(6.3)\qquad T = \text{diag }(\tau_1, \tau_2, \tau_3),$$

$$(6.4)\qquad K = \text{diag }(\alpha a^2, 0, 0),$$

and

$$(6.5)\qquad B = (I + P + P^2)' = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Combining (6.1) through (6.5), one arrives at

$$(6.6)\qquad \Gamma = \alpha \begin{bmatrix} (a^2 + b_1^2)\tau_1^2 & a^2\tau_1\tau_2 & a^2\tau_1\tau_3 \\ a^2\tau_1\tau_2 & (a^2 + b_2^2)\tau_2^2 & a^2\tau_2\tau_3 \\ a^2\tau_1\tau_3 & a^2\tau_2\tau_3 & (a^2 + b_3^2)\tau_3^2 \end{bmatrix}.$$

To determine the workload contents matrix G, we must first specify parameters δ_i ($i = 1, 2, 3$) that reflect the service disciplines at the three stations. Because there is just one customer class served at each station, formula (A.34) specializes in the case at hand to give

$$(6.7)\qquad \delta_i = 1/\tau_i \quad \text{with FIFO at station } i.$$

Similarly, for the processor sharing (PS) discipline, formulae (A.60)–(A.61) specialize to give

$$(6.8)\qquad \delta_i = 2/\tau_i(1 + b_i^2) \quad \text{with PS at station } i.$$

Recalling that $\Delta \equiv \text{diag }(\delta_1, \delta_2, \delta_3)$, readers may verify that (A.49) reduces in the case at hand to

$$(6.9)\qquad G = \begin{pmatrix} 0 & 0 & 0 \\ \tau_2\delta_1 & 0 & 0 \\ \tau_3\delta_1 & \tau_3\delta_2 & 0 \end{pmatrix}.$$

For an intuitive understanding of (6.9), note that formula (A.58), specialized to generalized Jackson networks, identifies $1/\delta_j$ as the average amount of remaining work for servers at station j embodied in a customer occupying station j. Thus, for $1 \leq j < i \leq 3$, $\tau_i \delta_j$ represents the average amount of future work for station i embodied in a unit of immediate work at station j, which is the general interpretation of G_{ij} given in the appendix.

Our general condition (3.12) for a product form stationary distribution is that $\Gamma_{ij} = \frac{1}{2} G_{ij} \Gamma_{jj}$ for $1 \leq j < i \leq 3$. Using (6.6) and (6.9), we see that this reduces to

(6.10) $\quad a^2 \tau_i \tau_j = \frac{1}{2} \tau_i \delta_j (a^2 + b_j^2) \tau_j^2 \quad \text{for} \quad 1 \leq j < i \leq 3,$

or equivalently,

(6.11) $\quad \delta_j = 2a^2 / \tau_j (a^2 + b_j^2) \quad \text{for} \quad j = 1, 2.$

Let us assume that the service discipline at both station 1 and station 2 is either FIFO or PS. Comparing (6.11) with (6.7) and (6.8), one arrives at the following requirements for (6.11) to hold:

(6.12) if station j has a FIFO discipline, then $b_j = a$;

or

(6.13) if station j has a PS discipline, then either $a = 1$ or else $b_j = 0$.

In other words, the parameter combinations yielding a product form solution for the approximating Brownian network model are precisely those given in Table 1 below.

Table 1.

Conditions Yielding a Product Form Solution for the Approximating Brownian Network Model

These findings are in some respects predictable. For example, there are no restrictions on either the service discipline or the service time distribution at station 3, which one would expect because that is an "exit node". Also, consider the case where input to the network is Poisson (implying $a = 1$) and each non-exit node $j = 1, 2$ satisfies one of the following two descriptions: either the service discipline is PS, or else the service time distribution is exponential (implying $b_j = 1$) and the discipline is FIFO. It is known [1, 13] that such a network has a product from stationary distribution, and Table 1 confirms that the corresponding Brownian network model also has a product form solution, as one would expect.

What is striking about (6.12) and (6.13) is that one may obtain a product form solution for the Brownian network model under weaker conditions: for a non-exit node j with FIFO discipline it is only required that $b_j = a$; and with non-Poisson input and a PS discipline at a non-exit node, one may still obtain a product form solution if the service times at that station are deterministic. As explained in the appendix, the Brownian network models described in this paper can be rigorously justified as heavy traffic limits when all nodes have FIFO discipline, and product form conditions that generalize (6.12) have appeared in earlier papers [8,10]. In contrast, our proposed method for representing the PS discipline in a Brownian network model is based on conjecture; there is as yet no rigorous heavy traffic limit theory for queueing systems with PS discipline, and such a theory is needed to fully justify the analysis presented here.

7. A Multiclass Example

Consider now the three-station network pictured in Figure 2. There are a total of four customer classes, and the exogenous inputs for class 1 and class 2 are assumed to be independent renewal processes. We denote by α_k the average input rate for class k customers, and by a_k the coefficient of variation for the class k interarrival time distribution ($k = 1, 2$). Thus the asymptotic covariance matrix for the four-dimensional input process $I = \{I(t),\ t \geq 0\}$ is

$$(7.1) \qquad K = \text{diag}\ (\alpha_1 a_1^2,\ \alpha_2 a_2^2,\ 0,\ 0).$$

The constituency of station 1 consists of customer classes 1 and 2, whereas only class 3 is served at station 2, and only class 4 is served at station 3. Thus we have the constituency matrix

$$(7.2) \qquad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The 4×4 switching matrix P has $P_{13} = P_{24} = 1$ and $P_{ij} = 0$ otherwise, and as in our previous example, such deterministic switching implies that

$$(7.3) \qquad H = 0.$$

Figure 2.

A Multiclass Network with Deterministic Switching

The overall arrival rates λ_k for the customer classes are simply $\lambda_1 = \lambda_3 = \alpha_1$ and $\lambda_2 = \lambda_4 = \alpha_2$, so formula (A.17) for the covariance matrix Λ reduces to

$$(7.4) \qquad \Lambda = \text{diag }(\alpha_1 \tau_1^2 b_1^2,\ \alpha_2 \tau_2^2 b_2^2,\ \alpha_1 \tau_3^2 b_3^2,\ \alpha_2 \tau_4^2 b_4^2).$$

From the switching probabilities P_{ij} specified above, it follows that

$$(7.5) \qquad B = I + P' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

and as in the appendix we set

$$(7.6) \qquad T = \text{diag }(\tau_1, \tau_2, \tau_3, \tau_4).$$

Substituting (7.1)–(7.6) into (A.51) and simplifying, one finds that the 3×3 covariance matrix Γ has

(7.7) $\Gamma_{11} = \alpha_1(a_1^2 + b_1^2)\tau_1^2 + \alpha_2(a_2^2 + b_2^2)\tau_2^2,$

(7.8) $\Gamma_{21} = \alpha_1 a_1^2 \tau_1 \tau_3,$

and

(7.9) $\Gamma_{31} = \alpha_2 a_2^2 \tau_2 \tau_4.$

On the other hand,

$$\Delta = \begin{bmatrix} \delta_1 & 0 & 0 \\ \delta_2 & 0 & 0 \\ 0 & \delta_3 & 0 \\ 0 & 0 & \delta_4 \end{bmatrix}$$

by definition, and one then deduces from (A.49) that the workload contents matrix G has the simple form

(7.10) $$G = \begin{bmatrix} 0 & 0 & 0 \\ \tau_3\delta_1 & 0 & 0 \\ \tau_4\delta_2 & 0 & 0 \end{bmatrix}.$$

Our general criterion (3.12) for a product form stationary distribution requires in this case that

(7.11) $\Gamma_{i1} = \frac{1}{2} G_{i1} \Gamma_{11}$ for $i = 2, 3,$

and by (7.8)–(7.10) this is equivalent to

(7.12) $\alpha_k a_k^2 \tau_k = \frac{1}{2} \delta_k \Gamma_{11}$ for $k = 1, 2.$

The product form criterion (7.12) involves not only the first and second moments of the interarrival and service time distributions, but also the service discipline at station 1, as manifested in the constants δ_1 and δ_2. To simplify subsequent discussion, let us *assume until further notice that both input processes are Poisson*, implying that $a_1 = a_2 = 1$. Then (7.12) reduces to the requirement that

(7.13) $\delta_k = (\lambda_k \tau_k) d_1$ for $k = 1, 2,$

where

(7.14) $d_1 = 2/\Gamma_{11} = 2/\{\lambda_1(1 + b_1^2)\tau_1^2 + \lambda_2(1 + b_2^2)\tau_2^2\}.$

Comparing this with (A.60)–(A.61), we see that (7.13) holds if one assumes a processor sharing (PS) discipline at station 1, and this is as one would expect, because in that case the original queueing network model is known to have a product form stationary distribution. A further implication of (7.13)–(7.14) is that, if one assumes some other service discipline at station 1, *the Brownian network model can only have a product form solution if that discipline gives the same values for δ_1 and δ_2 as does the PS discipline.*

For example, if the service discipline at station 1 is FIFO, we know from (A.34) that

$$(7.15) \qquad \delta_k = \lambda_k/\rho_1 \quad \text{for } k = 1, 2,$$

where $\rho_1 = \lambda_1\tau_1 + \lambda_2\tau_2$. Substituting (7.15) into (7.13) gives the product form criterion

$$(7.16) \qquad \tau_k = 1/d_1\rho_1 \quad \text{for } k = 1, 2.$$

This obviously requires that $\tau_1 = \tau_2 = \tau$, in which case $\rho_1 = (\lambda_1 + \lambda_2)\tau$ and (7.14) reduces to

$$(7.17) \qquad d_1 = 2/\{\rho_1 + (\lambda_1 b_1^2 + \lambda_2 b_2^2)\tau\}\tau.$$

Then the product form condition (7.16) becomes

$$(7.18) \qquad b^2 \equiv \left(\tfrac{\lambda_1}{\lambda_1+\lambda_2}\right) b_1^2 + \left(\tfrac{\lambda_2}{\lambda_1+\lambda_2}\right) b_2^2 = 1.$$

To summarize, if one assumes that both inputs are Poisson and the service discipline at station 1 is FIFO, then the approximating Brownian network model has a product form solution *if and only if* $\tau_1 = \tau_2$ and $b = 1$. To have $b = 1$ it is obviously sufficient that $b_1 = b_2 = 1$. That would be true if classes 1 and 2 had a common exponential service time distribution, in which case the original queueing network model is known to have a product form solution. However, (7.18) may hold even when classes 1 and 2 have distinct, non-exponential service time distributions, and in such cases the original model does *not* generally have a product form solution.

As a final note, let us consider again the case where station 1 has a PS discipline. We concluded that in this case the approximating Brownian network model has a product form solution, as does the original queueing network model, regardless of the service time distributions for classes 1 and 2. However, that conclusion is very much dependent on the assumption of Poisson inputs. If one takes either a_1 or a_2 to be different from 1 in the Brownian network model, a product form solution is no longer guaranteed, but neither is it impossible. To illustrate the latter point, consider the symmetric case with $\alpha_1 = \alpha_2 = \alpha > 0$, $a_1 = a_2 = a > 0$, $\tau_1 = \tau_2 = \tau$ and $b_1 = b_2 = 0$. Formula (7.7) then simplifies to give $\Gamma_{11} = 2\alpha a^2\tau^2$, and the product form condition (7.12) reduces to

(7.19) $\quad \delta_k = 1/\tau \quad$ for $k = 1, 2$.

If one assumes a PS discipline at station 1, it can be verified from (A.60) and (A.61) that (7.19) holds, so the Brownian network model has a product form solution. We presume that in cases like this, with non-Poisson input and deterministic services, the original queueing network model does not have a product form solution, but that issue has not been investigated.

8. An Example with Correlated Inputs

As a final example, consider the two-station network pictured in Figure 3. Here one has exogenous input processes $I_1 = I_2 = N$, and we take N to be a Poisson process with intensity parameter α. In other words, pairs of customers arrive in Poisson fashion at average rate α, and one member of each pair goes directly to station 2, whereas the other member requires a service at station 1 before proceeding to station 2. It follows that $\alpha_1 = \alpha_2 = \alpha$ and that the asymptotic covariance matrix of the two-dimensional input process I is

(8.1) $\quad K = \begin{bmatrix} \alpha & \alpha \\ \alpha & \alpha \end{bmatrix}.$

Class k customers are defined to be those visiting station k ($k = 1, 2$) and the 2×2 switching matrix P is given by

(8.2) $\quad P_{12} = 1$ and $P_{ij} = 0$ otherwise.

Figure 3.

A Two-Station Network with Perfectly Correlated Inputs

As in our previous examples, this deterministic switching implies that the switching noise process V has asymptotic covariance matrix $H = 0$. Obviously, $\lambda_1 = \alpha$ and $\lambda_2 = 2\alpha$, so (A.17) reduces to

$$\Lambda = \text{diag}\ (\alpha\tau_1^2 b_1^2,\ 2\alpha\tau_2^2 b_2^2). \tag{8.3}$$

The constituency matrix C is simply the 2×2 identity matrix, and

$$B = I + P' = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{8.4}$$

From these data and the general formula (A.51), readers may verify that

$$\Gamma = \begin{bmatrix} \alpha(1+b_1^2)\tau_1^2 & 2\alpha\tau_1\tau_2 \\ & \\ 2\alpha\tau_1\tau_2 & 2\alpha(2+b_2^2)\tau_2^2 \end{bmatrix}. \tag{8.5}$$

Assuming a FIFO service discipline at station 1, formula (A.34) gives $\delta_1 = 1/\tau_1$, and then (A.49) gives the workload contents matrix

$$G = \begin{bmatrix} 0 & 0 \\ \tau_2\delta_1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \tau_2/\tau_1 & 0 \end{bmatrix}. \tag{8.6}$$

Our general product form condition (3.12) is $\Gamma_{21} = \frac{1}{2}G_{21}\Gamma_{11}$, and by (8.5) and (8.6) that reduces to

$$b_1^2 = 3. \tag{8.7}$$

That is, with a FIFO service discipline at station 1, the Brownian network model has a product form stationary distribution if and only if $b_1^2 = 3$, although the original queueing network model appears to be intractable (we have not actually investigated this matter carefully), regardless of the service time distribution at station 1. Using formulas (A.60) and (A.61), interested readers may verify that if one assumes a PS discipline at station 1, it is impossible to satisfy the product form condition (3.12), regardless of the service time distribution at station 1.

Appendix: The Brownian Model of a Multiclass Open Network

In this appendix we describe a very general class of conventional queueing network models, and we explain how one approximates such a system by a Brownian network model of the type defined in section 2. Only *open* queueing networks are considered, in which customers arrive from outside the system and return to the outside world after a finite number of required "services" have been completed. In this appendix, customer routing is allowed to be arbitrary, but in the body of the paper only feedforward networks are considered.

Following the pattern established in [6] and section 5 of [7], consider a structured network model with *service stations* indexed by $i, j = 1, \ldots, d$ and *customer classes* indexed by $k, l = 1, \ldots, n$. Each class k has its own exogenous input process $I_k = \{I_k(t),\ t \geq 0\}$ (possibly null), and in the obvious way we denote by I the n-dimensional process with components $I_1, \ldots, I_n$. One interprets $I_k(t)$ as the number of class k customers who arrive from the outside by time t, and it is assumed that $I_k(0) = 0$. We also assume that there exists an n-vector α and an $n \times n$ covariance matrix K such that

$$\text{(A.1)} \qquad E[I(t)] \sim \alpha t \quad \text{and} \quad \operatorname{Cov}[I(t)] \sim Kt \quad \text{as} \quad t \to \infty,$$

where $\sim$ means "is asymptotic to", in the sense that the term on the left of this sign divided by the term on the right tends to 1 as $t \to \infty$. (In the second expression, this must be done component by component for the matrices involved.) In addition, a rigorous justification of the Brownian network model (see below) requires that I satisfy a functional central limit theorem, but we postpone discussion of that requirement until later.

It is assumed that $\alpha_k > 0$ for at least one class k. Customers of class k require service at a specific station $s(k)$, and their service times there are independent and identically distributed (IID) with mean $\tau_k > 0$ and coefficient of variation (that is, standard deviation divided by mean) b_k. The service time sequences for the various classes are assumed to be independent of one another and also of the arrival process I. The probability that a class k customer, upon completion of service at station $s(k)$, will turn next into a customer of class l is P_{kl}, and the probability that a class k customer will exit the system after completing service is $1 - \Sigma_l P_{kl}$, independent of all previous history. The $n \times n$ Markov switching matrix $P = (P_{kl})$ is assumed to be transient, which simply means that all arriving customers eventually leave the system. Let $\mathcal{C}(i)$ be the set of all customer classes k such that $s(k) = i$. We call $\mathcal{C}(i)$ the *constituency* of station i, and it is assumed that $\mathcal{C}(i)$ is non-empty for $i = 1, \ldots, d$.

Our assumptions with regard to customer routing are extremely weak. In particular, there is little or no loss of generality in the assumption that customers switch *classes* in Markovian fashion, or that the different classes have independent IID service time sequences, because the number of classes n can be made arbitrarily large; see section 2 of [6] and section 5 of [7] for discussion of this point. Completing the description of our queueing network model, it is assumed that station i consists of c_i identical servers working in parallel ($c_i \geq 1$) and each station i employs a work-conserving service discipline that is static and only uses information about customers present at station i when the scheduling decision (or priority decision) is made. This characterization of admissible service disciplines is admittedly vague, and nothing more will be said on the matter at present, except for the following: three illustrative disciplines of the type we intend to include in this discussion are first-in-first-out scheduling of the station, a static priority ranking (either preemptive-resume or non-preemptive) of the customer classes served at the station, and the "processor sharing" discipline to be discussed later.

In addition to the stochastic processes described above, let $Y_i(t)$ denote the cumulative server idleness at station i up to time t (a sum over the c_i servers who work in parallel there), and let $Q_k(t)$ denote the number of class k customers who are present at station $s(k)$ at time t, either waiting or being served. Also, let $A_k(t)$ be the number of customers who "enter" class k (external arrivals plus internal transitions) up to time t, and let $S_k(m)$ be the sum of the service times for the first m of those arrivals. It will be useful to define the *immediate workload input process* for class k,

$$L_k(t) = S_k\big(A_k(t)\big), \tag{A.2}$$

and the *immediate workload netflow* process for station i,

$$\zeta_i(t) = \sum_{k \in \mathcal{C}(i)} L_k(t) - c_i t. \tag{A.3}$$

Let $W_i(t)$ be the *immediate workload* at time t for servers at station i, equal to the sum of the impending service times of customers who are queued at the station at time t, plus the remaining service times of those customers (if any) who are being serviced there at time t. For any work-conserving service discipline, one then has that

$$W_i(t) = W_i(0) + \zeta_i(t) + Y_i(t). \tag{A.4}$$

To express the system equations (A.2)–(A.4) in more compact form, it will be convenient to define a $d \times n$ *constituency matrix* C via

$$\text{(A.5)} \qquad C_{ik} = \begin{cases} 1 & \text{if } s(k) = i \\ 0 & \text{otherwise,} \end{cases}$$

and an $n \times n$ diagonal matrix

$$\text{(A.6)} \qquad T = \text{diag }(\tau_1, \ldots, \tau_n).$$

Also, let us define

$$\text{(A.7)} \qquad \hat{S}_k(m) = S_k(m) - m\tau_k \quad \text{ and } \quad U_k(t) = \hat{S}_k\big(A_k(t)\big),$$

so that (A.2) can be rewritten in the form

$$\text{(A.8)} \qquad L_k(t) = U_k(t) + \tau_k A_k(t).$$

Defining vector stochastic processes (some d-dimensional and some n-dimensional) I, A, L, Y, X, W and U in the obvious way, we can restate (A.8), (A.3) and (A.4) in vector form as

$$\text{(A.9)} \qquad L(t) = U(t) + TA(t),$$

$$\text{(A.10)} \qquad \zeta(t) = CL(t) - ct,$$

and

$$\text{(A.11)} \qquad W(t) = W(0) + \zeta(t) + Y(t).$$

Recall that P is a transient Markov matrix by assumption. Thus we can define the fundamental matrix

$$\text{(A.12)} \qquad B = (I - P')^{-1} = (I + P + P^2 + \ldots)',$$

where I here is the $n \times n$ identity matrix, not to be confused with the process $I(t)$. One interprets B_{kl} as the average number of visits to class k made by a customer who starts in class l. Thus, defining an n-vector $\lambda = (\lambda_k)$ via

$$\text{(A.13)} \qquad \lambda = B\alpha,$$

we see that λ_k represents the long-run average number of customer visits to class k per unit time, assuming that every station has enough capacity to handle the workload imposed on it. With that proviso, the long-run average rate of workflow into station i will be $\Sigma_{k\in\mathcal{C}(i)}\ \lambda_k\tau_k$ and hence the "excess capacity" at station i is

$$\text{(A.14)} \qquad \theta_i = c_i - \sum_{k\in\mathcal{C}(i)} \lambda_k\tau_k.$$

Hereafter it is assumed that $\theta_i > 0$ for all i, in which case one expects that

$$\text{(A.15)} \qquad E[A(t)] \cong \lambda t \quad \text{for large } t,$$

where $\cong$ means "is approximately equal to". From this and (A.7) it follows that

$$\text{(A.16)} \qquad \text{Cov }[U(t)] \cong \Lambda t \quad \text{for large } t,$$

where

$$\text{(A.17)} \qquad \Lambda = \text{diag }(\lambda_1\tau_1^2 b_1^2, \ldots, \lambda_n\tau_n^2 b_n^2).$$

Let us denote by $D_k(t)$ the number of class k customers who complete service by time t (the class k *departure* process) and by $F_k(t)$ the number of customers who enter class k by means of internal transition, as opposed to external arrivals, during the interval $[0,t]$. (The letter F is mnemonic for *feedback*.) Defining n-dimensional processes D and F in the obvious way, one then has as a matter of definition that

$$\text{(A.18)} \qquad Q(t) = Q(0) + A(t) - D(t)$$

and

$$\text{(A.19)} \qquad A(t) = I(t) + F(t).$$

Our next task is to connect the feedback process F with the departure process D, and for that purpose let $\{\phi^m(1), \phi^m(2), \ldots\}$ be a sequence of IID *routing vectors* for customers completing visits to class m; the kth component of the vector equals 1 if the customer goes next to class k and equals zero otherwise. Denoting by ϕ^m a generic element of the sequence, it follows that

$$\text{(A.20)} \qquad E(\phi^m) = P_m' \quad \text{and} \quad \text{Cov }(\phi^m) = H^m,$$

where P_m is the mth row of P (thus P_m' is a column vector) and H^m is the $n \times n$ matrix defined by

$$\text{(A.21)} \qquad H^m_{kl} = \begin{cases} P_{mk}\,(1 - P_{mk}) & \text{if } k = l \\ \\ -P_{mk}P_{ml} & \text{if } k \neq l. \end{cases}$$

Also let $\hat{\phi}^m(r)$ be the centered random vector

$$\text{(A.22)} \qquad \hat{\phi}^m(r) = \phi^m(r) - P'_m,$$

and define the n-dimensional cumulative sums

$$\text{(A.23)} \qquad \Phi^m(r) = \sum_{i=1}^{r} \phi^m(i) \quad \text{and} \quad \hat{\Phi}^m(r) = \sum_{i=1}^{r} \hat{\phi}^m(i).$$

Then one has the key representation

$$\text{(A.24)} \qquad F(t) = \sum_{m=1}^{n} \Phi^m\big(D_m(t)\big) = \sum_{m=1}^{n} [\hat{\Phi}^m\big(D_m(t)\big) + P'_m D_m(t)],$$

and (A.24) can be rewritten in the compact form

$$\text{(A.25)} \qquad F(t) = V(t) + P'D(t)$$

where

$$\text{(A.26)} \qquad V(t) = \sum_{m=1}^{n} \hat{\Phi}^m\big(D_m(t)\big).$$

Given (A.15) and the stability condition $\theta > 0$, one naturally expects that

$$\text{(A.27)} \qquad E[D(t)] \cong E[A(t)] \cong \lambda t \quad \text{for large } t,$$

and from this and (A.26) it follows that

$$\text{(A.28)} \qquad \text{Cov}\,[V(t)] \cong Ht \quad \text{for large } t,$$

where

$$\text{(A.29)} \qquad H = \sum_{m=1}^{n} \lambda_m H^m.$$

Before an approximating Brownian network model can be proposed, it remains to connect the queue length process Q with the immediate workload process W, and that relationship depends critically on the service disciplines employed at the various stations.

The final data of our Brownian network model will be non-negative constants $\delta_1, \ldots, \delta_n$ such that

$$\text{(A.30)} \qquad \sum_{k \in \mathcal{C}(i)} \delta_k > 0 \quad \text{for } i = 1, \ldots, d.$$

A key hypothesis underlying the proposed approximation is that the service disciplines manifest themselves in the relationship

$$\text{(A.31)} \qquad Q_k(t) \cong \delta_k W_i(t) \quad \text{for each } k \in \mathcal{C}(i),$$

which can be stated more compactly as

$$\text{(A.32)} \qquad Q(t) \cong \Delta W(t),$$

where Δ is the $n \times d$ matrix defined by

$$\text{(A.33)} \qquad \Delta_{ki} = \begin{cases} \delta_k & \text{if } k \in \mathcal{C}(i) \\ 0 & \text{otherwise.} \end{cases}$$

The approximation (A.32) is a key to the tractability of the Brownian network model. It is precisely analogous to the relationships hypothesized in [11] to connect the queue length process and server workload process in a Brownian model of a single service station, and as we will discuss later, existing limit theorems suggest that it can be rigorously justified under heavy traffic conditions, at least for certain familiar service disciplines. If station i employs a FIFO service discipline, it is clear from existing theory that one should choose

$$\text{(A.34)} \qquad \delta_k = \lambda_k / \sum_{l \in \mathcal{C}(i)} \lambda_l \tau_l \quad \text{for all } \; k \in \mathcal{C}(i) \text{ and } i = 1, \ldots, d$$

when forming the Brownian network model, and we will discuss later the appropriate choices to represent other disciplines.

In describing the Brownian network model we will use the same symbols employed in the description above, with the understanding that each process is interpreted just as before. The primitive elements of the Brownian model are a non-negative random d-vector $W(0)$ and three *independent* n-dimensional Brownian motions I, U and V, which are also independent of $W(0)$, with $I(0) = U(0) = V(0) = 0$ and the following parameters:

$$
\text{(A.35)}\qquad
\begin{array}{l}
I \text{ has drift } \alpha \text{ and covariance matrix } K; \\
U \text{ has drift } 0 \text{ and covariance matrix } \Lambda; \text{ and} \\
V \text{ has drift } 0 \text{ and covariance matrix } H.
\end{array}
$$

We call I the exogenous *input process* as before, and given the earlier definitions of U and V in terms of centered random variables, one might reasonably describe them as a *service noise process* and a *switching noise process*, respectively. The system equations for the Brownian network model are the following:

$$\text{(A.36)}\qquad A(t) = I(t) + F(t),$$

$$\text{(A.37)}\qquad L(t) = U(t) + TA(t),$$

$$\text{(A.38)}\qquad \zeta(t) = CL(t) - ct,$$

$$\text{(A.39)}\qquad W(t) = W(0) + \zeta(t) + Y(t),$$

$$\text{(A.40)}\qquad Q(t) = \Delta W(t),$$

$$\text{(A.41)}\qquad D(t) = A(t) - [Q(t) - Q(0)],$$

and

$$\text{(A.42)}\qquad F(t) = V(t) + P'D(t).$$

Each of these relationships, except (A.40), is a repetition of an equation appearing in the earlier discussion, and (A.40) is obtained from (A.32) upon replacing the approximate equality by equality. Completing the specification of the Brownian network model, the following two relationships characterize the cumulative idleness process Y:

$$\text{(A.43)}\qquad Y \text{ is continuous and non-decreasing with } Y(0) = 0,$$

and

$$\text{(A.44)}\qquad Y_i \text{ can increase only at times } t \text{ for which } W_i(t) = 0, \quad i = 1, \ldots, d.$$

Of course, (A.43) is a natural physical restriction. Condition (A.44) says that cumulative server idleness at station i only increases when the station is devoid of customers. This is exactly true for single-server stations, and we take the point of view that it is an acceptable idealization in the case of multi-server stations. One can rigorously defend that point of view under "heavy traffic" assumptions, but (A.44) may represent a substantial compromise with reality for stations that are lightly loaded and/or have many servers.

Once the Brownian network model has been described by (A.35)–(A.44), two questions naturally arise. First, does there exist a family of processes that satisfies these relationships, and is that family in any sense unique? Second, can the Brownian network model be rigorously justified as a "heavy traffic limit" of the conventional network model described earlier? Turning first to the former issue, one can substitute (A.40)–(A.42) into (A.36) to obtain

$$A(t) = I(t) + P'A(t) - P'\Delta[W(t) - W(0)] + V(t). \tag{A.45}$$

Recalling that $B \equiv (I - P')^{-1}$, we now solve for $A(t)$ in terms of the other quantities:

$$A(t) = B[I(t) + V(t)] - BP'\Delta[W(t) - W(0)]. \tag{A.46}$$

Next, defining

$$\xi(t) = CU(t) + (CTB)I(t) + (CTB)V(t) - ct, \tag{A.47}$$

one may substitute (A.37) and then (A.46) into (A.38) to arrive at

$$\zeta(t) = \xi(t) - G[W(t) - W(0)], \tag{A.48}$$

where G is the $d \times d$ matrix defined by

$$G = CTBP'\Delta. \tag{A.49}$$

Finally, adding $W(0) + Y(t)$ to both sides of (A.48) and substituting (A.39) on the left, we conclude that

$$W(t) = W(0) + \xi(t) - G[W(t) - W(0)] + Y(t). \tag{A.50}$$

To understand the significance of (A.50), note first that (A.47) defines ξ entirely in terms of primitive model elements. More specifically, it follows from (A.35) that ξ is a d-dimensional Brownian motion with $\xi(0) = 0$, covariance matrix

$$\Gamma \equiv C\Lambda C' + (CTB)(K + H)(CTB)', \tag{A.51}$$

and drift vector $CTB\alpha - c$. Using (A.13), one may re-express the drift vector as $CT\lambda - c$ but this is just $-\theta$, where θ is the d-vector of "excess capacities" defined by (A.14). As we will see shortly, both ξ and the matrix G appearing in (A.50) have a ready interpretation. For the moment, however, the important point is that (A.50) and the two restrictions (A.43) and (A.44) that characterize Y contain all of the information about W and Y that is present in our original model description (A.35)–(A.44). One may naturally ask, given $W(0)$ and ξ, is there a pair (W,Y) satisfying (A.50), (A.43) and (A.44), and is that pair unique? In section 3 we show that the answer is affirmative for the relatively easy special case where G is lower triangular, corresponding to a feedforward queueing network. For general G, the question is much more complicated, and we do not actually know the answer, but it is worthwhile to say just a bit more about this fundamental issue. To avoid trivial complications, assume that Γ is non-degenerate and that $R \equiv (I+G)^{-1}$ exists. Then (A.50) can be re-expressed as

$$\text{(A.52)} \qquad W(t) = W(0) + R\xi(t) + RY(t)$$

or

$$\text{(A.53)} \qquad W(t) = W(0) + X(t) + RY(t),$$

where $X = R\xi$ is a d-dimensional Brownian motion with drift vector $\mu = -R\theta$ and non-degenerate covariance matrix $\Sigma = R\Gamma R'$. For the purpose of analyzing such a process, a minimal assumption is that $\{X(t) - \mu t,\ t \geq 0\}$ be a martingale with respect to a filtration to which W and Y are adapted. These conditions on X, together with (A.53), (A.43) and (A.44) identify W as a so-called *semimartingale reflected Brownian motion* (SRBM) with state space $\mathbb{R}^d_+$. It follows from the works of Reiman and Williams [19] and Taylor and Williams [21] that there exists a triple (W,X,Y) defined on some probability space and satisfying these properties *if and only if* R is what is called a *completely-$\mathcal{S}$* matrix, in which case W and Y are *unique in distribution* given $W(0)$. An important open research question is whether the R-matrices derived from queueing networks by means of the process described here are automatically of this class. That question, in turn, involves the issue of which Δ matrices can legitimately arise from queueing network models, and we will return to that subject shortly.

For an interpretation of the key relationship (A.50), let us return briefly to the conventional queueing network model described earlier. For the purposes of this paragraph only, let $\xi_i(t)$ be the *total* amount of work that servers at station i must do to complete the processing of all customers who enter the network by time t, minus $c_i t$ (the total amount of work that servers at station i can complete by time t if they are never idle). Defining a d-dimensional process ξ in the obvious way, we call ξ the *total workload netflow process* for the queueing network. Arguing exactly as in section 5 of [7], one can show that $E[\xi(t)] \sim -\theta t$ and Cov $[\xi(t)] \sim \Gamma t$ as $t \to \infty$. In the Brownian network model, ξ is represented by the Brownian motion on the right side of (A.47), whose drift vector and covariance matrix we have shown to be $-\theta$ and Γ, respectively. Next, from (A.12) and (A.49), it follows that $G = M\Delta$, where M is the $d \times n$ matrix defined by

$$M = CT(P + P^2 + \ldots)'.$$

One interprets M_{ik} as the average amount of *future work* required from servers at station i to complete processing of a customer currently in class k, where "future work" means work remaining after the customer's next class transition, or equivalently, after completion of the customer's impending class k service. When one equates $Q(t)$ with $\Delta W(t)$ in accordance with (A.40), one obtains

$$\text{(A.54)} \qquad GW(t) = M\Delta W(t) = MQ(t),$$

and the ith component of $MQ(t)$ represents the expected future work for servers at station i embodied in customers now present anywhere in the network. Thus G_{ij} *represents the average amount of future work for station i embodied in a unit of immediate work at station j*. In the Brownian network model, *actual* future work per unit of immediate work is simply equated with *average* future work per unit of immediate work, which results in equation (A.48) for the immediate workload netflow process ζ. Given the interpretations of ξ and G developed in this paragraph, of course, one can simply take these to be primitive model elements, and use (A.48) to directly define the immediate workload netflow ζ, arriving at the *reduced form* Brownian network model laid out in section 2, as opposed to the *extensive form* Brownian model (A.35)–(A.44). Such a direct formulation of the reduced form Brownian model was advocated in [7].

There have been repeated references in this paper to "heavy traffic limit theorems" that rigorously justify Brownian network models as weak limits of conventional queueing networks. What we have described in this appendix is an approximation scheme that goes far beyond anything one can justify on the basis of existing heavy traffic results, but we conjecture that a limit theory can in fact be developed to provide formal justification of the proposed approximation. The following paragraphs elaborate on this conjecture.

Consider a multiclass queueing network model of the type described earlier in this appendix, and assume that θ_i is small but positive for each station i. (That is, consider a stable system *in heavy traffic*). One can then choose a large integer N such that

(A.55) $\quad \beta \equiv N^{\frac{1}{2}}\theta > 0$ is of moderate size.

What one would like to show is that the d-dimensional scaled workload process W^N defined by

$$\text{(A.56)} \qquad W^N(t) = N^{-\frac{1}{2}}W(Nt)\ ,\ t \geq 0,$$

is well approximated by a d-dimensional RBM with appropriately chosen parameters. To obtain such a conclusion one needs to assume something more about the vector input process I, in addition to existence of an asymptotic mean vector α and an asymptotic covariance matrix K. One natural assumption, but certainly not the weakest possible, is that the centered and scaled input process I^N defined by

$$\text{(A.57)} \qquad I^N(t) = N^{-\frac{1}{2}}[I(Nt) - N\alpha t],\ t \geq 0,$$

behaves approximately as a $(0, K)$ Brownian motion. If I has independent renewal inputs (in that case K is diagonal), such a statement is justified by the familiar functional central limit theorem (FCLT) for renewal processes, and a similar FCLT can be proved for many other structured models of input flows.

To be more precise, one wants to consider a sequence of queueing networks indexed by $N = 1, 2, \ldots$ whose excess capacity vectors θ^N satisfy $N^{\frac{1}{2}}\theta^N \to \beta > 0$ as $N \to \infty$, and then to show that the scaled workload processes W^N associated with the successive systems converge weakly to a specified RBM as $N \to \infty$. (If each system in the sequence has the same master covariance matrix Γ and the same workload contents matrix G, then the limiting RBM will be one with reflection matrix $R = (I + G)^{-1}$, assuming this exists, covariance matrix $\Sigma = R\Gamma R'$, drift vector $\mu = -R\beta$, and state space $S = \mathbb{R}^d_+$.) More generally, given appropriate restrictions on the service disciplines (see below) we conjecture that the entire vector of processes $I, A, F, L, U, \zeta, W, Y, Q, D$ and V, after proper centering and scaling, converges weakly to the analogous vector of processes associated with the approximating Brownian network model. That is precisely the sort of result obtained by Peterson [16] for feedforward networks having multiple customer types and deterministic routing. The result that we are conjecturing here would generalize Peterson's limit theorem by allowing probabilistic switching among customer classes, including the possibility of feedback, and multi-server stations. Previous work on heavy traffic theory suggests that the extension to multi-server stations is relatively easy, whereas the extension to networks *with feedback* involves profound difficulties; by using induction, one can reduce the analysis of a feedforward network to analysis of single stations, but a new approach must be found to treat the general case with feedback. Also, as the next paragraph suggests, service disciplines have a definite influence on the RBM that one obtains as a heavy traffic limit, and it is not clear thus far how to even *state* a limit theorem for a network with "general" service disciplines, let alone prove it.

In describing the Brownian network model that approximates a given conventional queueing network, we have specified that the queue length process Q be related to the workload process W via

$$Q_k(t) = \delta_k W_i(t) \quad \text{for all} \quad k \in \mathcal{C}(i) \quad \text{and} \quad i = 1, \ldots, d, \tag{A.58}$$

where $\{\delta_k,\ k \in \mathcal{C}(i)\}$ are constants (not all zero) reflecting the service discipline at station i. The simple relationship (A.58) does in fact characterize the Brownian network model obtained by Peterson [16] as a heavy traffic limit, at least for the case of a static priority ranking at each station. To be more precise, for each station i, let $\mathcal{L}(i)$ be a non-empty subset of the constituency $\mathcal{C}(i)$, and suppose that customers at station i are granted admission to service in accordance with a static priority ranking, classes in $\mathcal{L}(i)$ being tied for lowest priority. This means that classes $k \in \mathcal{L}(i)$ are served on a first-in-first-out basis at station i, and customers of all *other* classes in $\mathcal{C}(i)$ are given priority over classes in $\mathcal{L}(i)$. Peterson [16] showed that (A.58) holds in the Brownian network model that he obtains as a heavy traffic limit, where

$$\delta_k = \begin{cases} \lambda_k / \sum\limits_{l \in \mathcal{L}(i)} \lambda_l \tau_l & \text{if } k \in \mathcal{L}(i), \\ 0 & \text{otherwise.} \end{cases} \tag{A.59}$$

Of course, ordinary FIFO scheduling corresponds to the special case where $\mathcal{L}(i) = \mathcal{C}(i)$, and we conjecture that (A.58) holds for many other types of "local" scheduling rules (that is, scheduling rules that depend only on the current mix of customers at the station being scheduled) if the constants δ_k are chosen correctly. As an example, consider a single-server station i that uses the so-called *processor sharing* rule, which means that when a total of m customers are present at station i, work is done constantly on each of those customers at rate $1/m$ (that is, each customer receives one mth of the server's total attention). This rule may be viewed as a limit of the so-called *round-robin discipline* with service increment ϵ. In that discipline the customer at the head of the queue at station i receives ϵ time units of service, and if that does not suffice to complete the customer's service requirement, he is sent to the end of the queue and must work his way up to the head again to receive another service increment. Assuming that new arrivals to station i join the end of the queue, one may think of processor sharing as the limit of this discipline as $\epsilon \downarrow 0$. Reiman [18] has proved a heavy traffic limit theorem for a single-station queueing model with multiple customer classes and probabilistic feedback, which includes the round-robin discipline as a special case. A formal analysis of his limiting Brownian station model leads one to conjecture that, for a queueing network in which station i has processor sharing, the limiting Brownian network model satisfies (A.58) with

(A.60) $\quad \delta_k = d_i(\lambda_k \tau_k) \quad \text{for all } k \in \mathcal{C}(i),$

where

(A.61) $\quad d_i = 2[\sum_{k\in\mathcal{C}(i)} \tau_k^2(1+b_k^2)\lambda_k]^{-1}.$

To repeat, we conjecture that (A.58), (A.60) and (A.61) hold for a Brownian network model obtained as a heavy traffic limit of a conventional model with processor sharing at station i; there is no rigorous limit theorem to justify this assertion even in a single-station setting, to say nothing of a full-blown network setting. The intuitive content of (A.60) is that the *queue lengths* for various classes served at station i remain at all times proportional to the average contributions those classes make to the overall workload at the station. For a derivation of the proportionality constant d_i, let F_k be the distribution function for class k service times and consider the corresponding residual lifetime distribution

$$G_k(x) = \tau_k^{-1} \int_0^x [1 - F_k(y)]dy, \quad x \geq 0.$$

The mean of the residual lifetime distribution is

(A.62) $\quad r_k \equiv \int_0^\infty x dG_k(x) = \frac{1}{2}(1+b_k^2)\tau_k \ , \ k \in \mathcal{C}(i).$

In the Brownian model of a station with processor sharing, the distribution of remaining service time among class k customers who occupy the station is at all times equal to the residual lifetime distribution G_k, and thus the average amount of remaining work to be done per class k customer is at all times equal to r_k. Combining (A.62) with (A.59) and (A.60), we conclude that the immediate workload for station i at time t is

$$\sum_{k\in\mathcal{C}(i)} r_k Q_k = \sum_{k\in\mathcal{C}(i)} r_k d_i \lambda_k \tau_k W_i(t),$$

and upon equating that expression to $W_i(t)$ and solving for d_i, one obtains (A.61).

The heavy traffic limit theorem that was conjectured earlier in this appendix involved a structured multiclass model with Markovian switching among customer classes and independent IID service time sequences. In the end, however, we obtain an approximating Brownian network model that is built from just two primitive elements: a Brownian motion ξ that represents the total workload netflow process; and a matrix G whose (i,j)th element represents the average amount of future work for station i embodied in a unit of immediate work for station j, taking into account the service discipline employed at station j. This suggests that a heavy traffic limit theorem might be obtainable with much weaker assumptions than described earlier. Without even introducing the notion of customer classes, one might directly hypothesize (a) a functional central limit theorem for the total workload netflow process, and (b) some sort of functional strong law of large numbers that involves a long-run average workload contents matrix G. There is reason to believe that under such weak assumptions one could still prove convergence to the approximating Brownian network model proposed in this paper. The exposition of Brownian network models in sections 1–3 of [7] was based on just such a minimalist approach.

References

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

[2] Bernard, A., and El Kharroubi, A. Régulation de processus dans le premier orthant de $\mathbb{R}^n$, to appear in *Stochastics.*

[3] Chung, K.L., and Williams, R.J. (1990). *Introduction to Stochastic Integration*, Birkhäuser, Boston.

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

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

[6] Harrison, J.M. (1988). Brownian models of queueing networks with heterogeneous customer populations, in: *Stochastic Differential Systems, Stochastic Control Theory and their Applications*, W. Fleming and P.-L. Lions (Eds.), vol. 10 in the IMA Series on Mathematics and its Applications (Springer, New York) pp. 147–186.

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

[8] Harrison, J. M. and Reiman, M. I. (1981). On the distribution of multidimensional reflected Brownian motion, SIAM J. *Appl. Math.* **41** 345–361.

[9] Harrison, J.M., and Reiman, M.I. (1981). Reflected Brownian motion on an orthant, *Annals of Probability* **9** 302–308.

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

[11] Harrison, J.M., and Williams, R.J. (1990). On the quasireversibility of a multiclass Brownian service station, *Annals of Probability* **18** 1249-1268.

[12] Jackson, J. R. (1963). Jobshop-like queueing systems. *Management Science* **10** 131–142.

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

[14] Kelly, F.P. (1982). Networks of quasireversible nodes. In *Applied Probability and Computer Science: The Interface* (R. L. Disney and T. J. Ott, eds.) **1** 3-29. Birkhauser, Boston.

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

[16] Peterson, W.P. (1990). A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Ops. Rsch.* **16** (1991) 1-28.

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

[18] Reiman, M. I. (1988). A multiclass feedback queue in heavy traffic. *Adv. Appl. Prob.* **20** 179–207.

[19] Reiman, M.I., and Williams, R.J. A boundary property of semimartingale reflecting Brownian motions, *Prob. Theor. Rel. Fields* **77** (1988), 87–97; **80** (1989), 633.

[20] Sauer, C. H., and Chandy, K. M. (1981). *Computer System Performance Modeling*, Prentice-Hall, Englewood Cliffs, N.J.

[21] Taylor, L. M., and Williams, R. J. (1990). Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant, in preparation.

[22] Walrand, J. (1983). A probabilistic look at networks of quasi-reversible queues. *IEEE Trans. on Information Theory* **IT-29** 825–831.

[23] Walrand, J. (1988). *An Introduction to Queueing Networks.* Prentice-Hall, Englewood Cliffs, N.J.

[24] Williams, R.J. (1987). Reflected Brownian motion with skew symmetric data in a polyhedral domain, *Prob. Theor. Rel. Fields* **75** 459–485.