The Annals of Probability
1990, Vol. 18, No. 3, 1249–1268

ON THE QUASIREVERSIBILITY OF A MULTICLASS BROWNIAN SERVICE STATION[1]

BY J. M. HARRISON AND R. J. WILLIAMS

Stanford University and University of California at San Diego

The object of study in this paper is a Brownian model of a multiclass service station. Such Brownian models arise as heavy traffic limits of conventional queueing models in which several different types or classes of customers are processed through a common service facility. Assuming that the Brownian service station is initialized with its stationary distribution, four different model characteristics are shown to be equivalent, and the station is said to be quasireversible if these equivalent conditions pertain. Three of the four conditions characterize the vector departure process from the Brownian service station, and our definition of quasireversibility parallels that proposed by F. P. Kelly for conventional queueing models. The last of our four conditions is expressed directly in terms of primitive model parameters, so one may easily determine from basic data whether or not a Brownian station model is quasireversible. Rather than characterizing the complete vector of departure processes from a Brownian service station, we prove a more general theorem expressed in terms of arbitrary linear combinations of the departure processes; this yields a generalized notion of quasireversibility that will play an important role in future work. To be more specific, in a future paper on multiclass Brownian *network* models, it will be shown that there is an intimate relationship between product form stationary distributions and the generalized notion of quasireversibility developed here.

1. Introduction. In the now classical theory of product form queueing networks, an important role is played by the station-level property that F. P. Kelly calls quasireversibility. This concept was first identified by Muntz [19], independently discovered and used by Kelly [14, 15] and later elaborated in important ways by Kelly [18] and Walrand [25]. For a systematic treatment of the basic concept and its role in queueing network theory, readers may consult the books by Kelly [17], Whittle [28] and Walrand [26], the last of which includes an extensive and up-to-date list of references.

Several different but ultimately equivalent definitions of quasireversibility can be found in the literature of applied probability; the one that serves best for our purposes is the following. (Because this is just background discussion, we shall be relatively informal, seeking to communicate the spirit of the mathematical theory rather than its precise content.) Consider a single-station queueing system, perhaps with multiple servers, in which customers of various classes arrive according to independent Poisson processes. Let us assume that the state of the system can be described as a Markov chain $X = \{X(t),\ t \geq 0\}$

Received October 1988; revised April 1989.

[1] Research supported in part by NSF Grants DMS-86-57483 and 87-22351, and by the Emerson Electric Faculty Fellowship, Graduate School of Business, Stanford University.

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

Key words and phrases. Multiclass queues, quasireversibility, Brownian models.

with finite or countably infinite state space, and that arrivals after time t are independent of the state $X(t)$. Finally, assuming that the Markov chain X admits a stationary distribution, let us suppose that $X(0)$ is randomized with that stationary distribution. The queueing system is said to be quasireversible if

(i) departures up to time t are independent of the state $X(t)$,

which is known ([18], pages 7–8; [26], page 90) to imply that

(ii) the departure streams for the various customer classes are independent Poisson processes.

It is not generally true that (ii) implies (i); see, for example, [17], Exercise 3.2.3, page 71.

The celebrated output theorem of Burke [2, 3] shows that an $M/M/s$ queueing system is quasireversible in this sense, and Burke's output theorem can be applied inductively to show that a network of exponential service stations has a product form stationary distribution, provided that all external arrival processes are Poisson and there is no feedback in the routing of customers. That is, for such a network the population sizes at the various stations are independent in equilibrium, and the stationary distribution for any one station in isolation is that of an $M/M/s$ queue. This result was extended by Jackson [11] to allow Markovian routing with feedback, but Jackson's method of proof made no mention of any output characteristics of individual stations.

Jackson's product form result was greatly extended in a famous paper by Baskett, Chandy, Muntz and Palacios [1], referred to hereafter as BCMP. They generalized Jackson's network model to allow virtually arbitrary customer routing, and they also showed that a nonexponential service time distribution can be allowed at any given station if it occurs in conjunction with a queue discipline or service discipline of a very particular type (for example, processor sharing). Even with the model generalized in this way, BCMP showed that the network has a product form stationary distribution, and they wrote out the stationary distribution in explicit formulas. Some readers have been seduced by this remarkable result into thinking that "essentially all" queueing networks have product form stationary distributions. That is certainly not true, but, just as certainly, the BCMP paper is one of the seminal achievements in queueing theory of the last 20 years.

As in Jackson's work, there was no mention of station-level output characteristics in the BCMP paper, but the earlier unpublished work of Muntz [19] shows that the authors were aware of such characteristics and of their connection with product form stationary distributions. Working independently of BCMP, Kelly [14, 15] discovered many of the same results, extending Jackson's theory of product form networks to allow multiple customer types and general routing. Moreover, Kelly did explicitly connect the product form property of a network, even one with feedback, to station-level output characteristics. It was shown in Kelly [17] that each station of a BMCP network is quasireversible when viewed in isolation. That is, if the station is removed from the network and driven with independent Poisson inputs, at average arrival rates consistent with the network's data, then its output streams have property (i) above, and hence property (ii). Kelly further showed that a network of quasireversible stations always has a product form stationary distribution. Thus Kelly's work did much to explain the essential nature of product form results for queueing networks, and the theory of quasireversibility has developed further since then and is still developing.

A separate stream of mathematical research over the last 20 years has resulted in the development of queueing network models built from Brownian motion, which were originally motivated or justified as heavy traffic limits of conventional network models. Developments in this area were summarized in two recent papers by Harrison and Williams [8, 9]. In both of these papers a major emphasis was placed on Brownian network models that have product form stationary distributions, because those are the only Brownian network models for which there currently exist results that are useful in practical performance analysis. However, no connection was made between product form Brownian networks and Kelly's notion of quasireversibility. In this paper we begin the process of making that connection, by developing a theory of quasireversible Brownian service stations. That is, we define a multiclass Brownian service station and propose a definition of quasireversibility in that context, showing that several different forms of the definition are equivalent, including one expressed in terms of primitive model parameters. Rather than characterizing the complete vector of departure processes from a Brownian service station, we prove a more general theorem expressed in terms of arbitrary linear combinations of the departure processes; this yields a generalized notion of quasireversibility that will play an important role in future work. To be more specific, in a future paper on multiclass Brownian *network* models, it will be shown that there is an intimate relationship between product form stationary distributions and the generalized notion of quasireversibility developed here.

The Brownian model of a single station that we describe in this paper allows multiple customer classes, whereas most previous work on Brownian network models has been restricted to the case of a single customer class. Thus we are both anticipating and laying groundwork for the development of a multiclass Brownian network theory. Peterson [21] and Reiman [23] have begun the development of that theory, and completion of the task is a top priority if Brownian network models are ever to have a really substantial impact on the world of practical performance analysis.

The remainder of this paper is organized as follows. Section 2 gives the precise mathematical definition of a multiclass Brownian service station, with relatively little in the way of interpretation or justification. Then in an appendix we recapitulate the heavy traffic limit theory that motivates this Brownian model and provides a clear interpretation for each of its constituents. Our main theorem is stated and proved in Section 3. As a corollary of this result, we deduce that two output characteristics, analogous to (i) and

(ii) above, are equivalent to one another, and are in turn equivalent to certain relationships among the parameters of the model.

In the sequel, an m-dimensional process U will be called a Brownian motion if there is a driftless Brownian motion V, with independent components starting from zero, such that V is independent of $U(0)$ and for $i = 1, \dots, m$, after subtraction of a linear drift and initial value, the ith component U_i of U can be written as a linear combination of the components of V. This allows the components of U to have drift and to be degenerate and/or dependent. We shall also use the following notation in the sequel. Vectors, including the values of vector-valued processes, will be regarded as column vectors, unless specifically indicated otherwise. For an m-dimensional vector v, the components of v will be denoted $v_1, \dots, v_m$, and diag(v) will denote the $m \times m$ diagonal matrix whose diagonal entries are given by these components.

2. The Brownian model of a multiclass station. Customer classes are indexed by $i = 1, \dots, n$ and for each class i we take as given an average arrival rate $\lambda > 0$, a mean service time $\tau_i > 0$ and a coefficient of variation $\beta_i \geq 0$ for class i service times. (That is, β_i represents the standard deviation of the class i service time distribution divided by the mean service time τ_i.) Define the n-vectors $\lambda = (\lambda_1, \dots, \lambda_n)'$ and $\tau = (\tau_1, \dots, \tau_n)'$, where prime denotes transpose. Also given are an $n \times n$ covariance matrix G and a nonnegative n-vector δ whose ith component δ_i is strictly positive for at least one class $i \in \{1, \dots, n\}$. As readers will see shortly, G is the covariance matrix of our vector arrival process and δ reflects the service discipline employed at the Brownian service station.

The probabilistic primitives in our model are a nonnegative random variable $W(0)$, an n-dimensional Brownian motion A with

$$(2.1) \qquad E[A(t)] = \lambda t \quad \text{and} \quad \mathrm{Cov}[A(t)] = Gt$$

and an n-dimensional Brownian motion S with *independent* components $S_1, \dots, S_n$ such that

$$(2.2) \quad E[S_i(t)] = \tau_i t \quad \text{and} \quad \mathrm{Var}[S_i(t)] = (\tau_i \beta_i)^2 t, \quad \text{for } i = 1, \dots, n.$$

Both A and S start from the origin [that is, $A(0) = S(0) = 0$] and $W(0)$, A, S are mutually independent. Readers should interpret $W(0)$ as the initial server workload (see below). As explained in the Appendix, when one formulates the Brownian analog of a conventional queueing model, A_i replaces the cumulative arrival process for class i customers and S_i replaces the partial sums process for class i service times. Most previous work on Brownian system models has focused on the case where A, like S, has independent components. However, from a mathematical standpoint, it is both natural and pleasing to let A have an arbitrary convariance structure, and that added generality will actually prove to be important in our later study of multiclass Brownian network models.

For future purposes, it will be convenient to adopt the representations

$$A(t) = \lambda t + \xi(t) \quad \text{and} \quad S(t) = \tau t + \nu(t), \tag{2.3}$$

where ξ and ν are independent n-dependent driftless Brownian motions starting from the origin, with covariance matrices G and $\operatorname{diag}((\tau_1\beta_1)^2, \dots, (\tau_n\beta_n)^2)$, respectively. Also, let

$$\mu_i = \tau_i^{-1} \quad \text{and} \quad \rho_i = \lambda_i\tau_i \quad \text{for } i = 1, \dots, n. \tag{2.4}$$

Obviously, μ_i represents the average service rate that the server can achieve when devoting all of his or her time to class i, and ρ_i is the contribution that class i makes to the traffic intensity $\rho = \rho + \cdots + \rho_n$ at the station. As explained in the Appendix, the random variable

$$L_i(t) = \rho_i t + \tau_i\xi_i(t) + \nu_i(\lambda_i t) \quad \text{for } i \in \{1, \dots, n\} \tag{2.5}$$

represents the total amount of server work (expressed in units of time) required to complete the processing of class i customers who arrive up to time t. (The letter L is mnemonic for *load*.) Now let

$$X(t) = \sum_{i=1}^{n} L_i(t) - t,$$

calling X the *workload netflow process* for the multiclass station. Substitution of (2.5) in the above yields

$$X(t) = -(1-\rho)t + \sum_{i=1}^{n} \left[\tau_i\xi_i(t) + \nu_i(\lambda_i t)\right]. \tag{2.6}$$

By (2.3), (2.4) and (2.6), we have the alternative expression for X which will also be used in the sequel:

$$X(t) = \tau' A(t) + \sum_{i=1}^{n} \nu_i(\lambda_i t) - t. \tag{2.7}$$

It follows from (2.3), (2.6) and (2.7) that X is a one-dimensional Brownian motion starting from zero with

$$E[X(t)] = -(1-\rho)t \tag{2.8}$$

and

$$\operatorname{Var}[X(t)] = \sigma^2 t, \quad \text{where } \sigma^2 = \tau' G \tau + \sum_{i=1}^{n} \lambda_i(\tau_i\beta_i)^2. \tag{2.9}$$

An important special case, and the only one dealt with in the Appendix to this paper, is that where

$$G = \operatorname{diag}\left(\lambda_1\alpha_1^2, \dots, \lambda_n\alpha_n^2\right), \tag{2.10}$$

corresponding to a conventional queueing model where classes $1, \dots, n$ have independent renewal input processes and the class i interarrival time distribution has coefficient of variation α_i. In this case, readers may verify that

$$\sigma^2 = \sum_{i=1}^{n} \lambda_i \tau_i^2 (\alpha_i^2 + \beta_i^2). \tag{2.11}$$

We now define the *unfinished workload process* W. Recall that $W(0)$ is a nonnegative random variable (the initial server workload) that is independent of the pair (A, S) and hence independent of X. By applying to $W(0) + X$ the path-to-path mapping that Harrison [7] calls the *one-sided regulator*, more often referred to as the *one-sided reflection mapping*, we obtain

$$W(t) = W(0) + X(t) + Y(t), \qquad t \geq 0, \tag{2.12}$$

where

$$Y(t) = \sup_{0 \leq s \leq t} (W(0) + X(s))^{-}. \tag{2.13}$$

It follows from these definitions that $W(t) \geq 0$ for all $t \geq 0$, that Y is continuous and increasing with $Y(0) = 0$ and that Y increases only at times t where $W(t) = 0$. One interprets $Y(t)$ as the cumulative server idleness up to time t and $W(t)$ as the total time required to complete processing of customers who are present at the station at time t. We shall define the n-dimensional *queue length process*

$$Q(t) = \delta W(t), \tag{2.14}$$

interpreting the ith component $Q_i(t)$ as the total number of class i customers present at the station at time t. The critical definition (2.14) says that, in the idealized Brownian model, for each i, the queue length for class i customers is proportional to the total server workload. This feature is a key to the model's tractability and, as explained in the Appendix, it can be rigorously justified by heavy traffic limit theorems for various queue disciplines. In particular, the familiar first-in-first-out (FIFO) discipline gives $\delta_i = \lambda_i/\rho$ for each class i, and a complete priority ranking of the classes yields $\delta_k = \mu_k$ for whatever class k is given lowest priority, with $\delta_i = 0$ for all other classes i; other disciplines give other constants of proportionality. The assumption that $\delta_i > 0$ for at least one i corresponds to the fact that nonzero workload must be attributable to at least one of the customer classes. To complete the specification of the multiclass Brownian station model, we define the n-dimensional *departure process*

$$D(t) = A(t) + (Q(t) - Q(0)), \tag{2.15}$$

interpreting the ith component D_i as the cumulative departure process for class i.

3. The main result. Hereafter it is assumed that $\rho < 1$ and $\sigma^2 > 0$. The process W defined in the previous section is a reflected Brownian motion on the positive half-line, also called regulated Brownian motion by Harrison [7], with drift parameter $-(1 - \rho)$, variance parameter σ^2 and a lower reflecting barrier at zero. As shown in Chapter 5 of Harrison [7], the unique stationary distribution of the Markov process W has density function

$$(3.1) \qquad p(w) = \eta e^{-\eta w}, \qquad w \geq 0, \quad \text{where } \eta = 2(1-\rho)/\sigma^2.$$

That is, the stationary distribution is exponential with mean $\sigma^2/2(1-\rho)$. Hereafter we assume that the initial server workload $W(0)$ is randomized with this stationary distribution.

Our main result is Theorem 3.1. Before stating this, we need to introduce the following notation.

NOTATION. Suppose U and V are a k-dimensional continuous semimartingale and an m-dimensional continuous semimartingale, respectively, with respect to a given filteration. We denote by $\langle U, V \rangle$ the $(k \times m)$-matrix-valued process whose (i,j)th component is the mutual variation process of the pair (U_i, V_j). That (i,j)th component will be denoted simply by $\langle U_i, V_j \rangle$.

Readers should note that the mutual variation processes referred to above are independent of the particular semimartingale representations one chooses for U and V [12, Theorem I.4.47]. Also, if (U, V) is a $(k+m)$-dimensional Brownian motion starting from the origin, then $\langle U_i, V_j \rangle_t$ is simply the convariance of the bivariate normal pair $(U_i(t), V_j(t))$. Finally, note that the $(2n+1)$-dimensional process $(A_1, \dots, A_n, D_1, \dots, D_n, X)$, formed from the components of A, D, and X, is a $(2n+1)$-dimensional continuous semimartingale with respect to the filtration generated by A, X and $W(0)$, and it starts from the origin.

THEOREM 3.1. *Let N be a $k \times n$ matrix $(k \geq 1)$ and suppose $N\delta$ has at least one nonzero component. The following four statements are equivalent.*

(i) *For each fixed t, the k-dimensional process $\{ND(s)\colon 0 \leq s \leq t\}$ is independent of $W(t)$.*

(ii) *ND is a Brownian motion.*

(iii) $\langle ND, ND \rangle = \langle NA, NA \rangle$.

(iv) $NG\tau = \frac{1}{2}\sigma^2 N\delta$.

DEFINITION. The multiclass Brownian station model is said to be *quasireversible with respect to N*, or simply *N-quasireversible*, if conditions (i)–(iv) hold. It is said to be *quasireversible* if it is quasireversible with respect to the $n \times n$ identity matrix I.

REMARK 1. If $N\delta = 0$, then (ii) and (iii) automatically hold and (i) and (iv) are equivalent, by the proof below. In this case, (i) and (iv) hold if and only if $NG\tau = 0$.

REMARK 2. In Lemma 3.2 it will be shown that $E[D(t)] = \lambda t = E[A(t)]$, without any special assumptions on the model data. Thus, $E[ND(t)] = N\lambda t = E[NA(t)]$ in all cases. Combining this with (ii) and (iii) of Theorem 3.1, we have the following: The multiclass Brownian service station is N-quasireversible if and only if ND has the same distribution as NA.

REMARK 3. Specializing to the case where G is diagonal (i.e., the components of A are independent) and further taking $N = I$, let us consider the output property (i) of Theorem 3.1. This is obviously analogous to condition (i) of Section 1, which serves to define quasireversibility for conventional queueing models and condition (ii) of Theorem 3.1 is similarly anologous to condition (ii) of Section 1. Readers should note that (i) and (ii) are equivalent for the Brownian model, whereas for conventional models (i) implies (ii), but the converse is not true in general. Also, it is a notable feature of Theorem 3.1 that condition (iv), involving model parameters rather than stochastic processes, is shown to be *necessary and sufficient* for the output properties to hold. No analogous result is known for conventional queueing models, but in a sense this discrepancy is a matter of definition: The class of Brownian station models, as we define that term, is relatively narrow, whereas the class of conventional queueing models is so large and amorphous that one can hardly imagine a result analogous to Theorem 3.1 in that setting. Nevertheless, we believe that our definition of a Brownian station model is broad enough to capture the heavy traffic limits of most single-station queueing models that have been identified in the literature of applied probability; the sharp characterization achieved in Theorem 3.1 testifies to the parsimony of Brownian models, where differences between systems are reflected in parameter values rather than model structure.

EXAMPLE. Consider the special case when (2.10) holds, corresponding to a conventional queueing model with independent renewal inputs for the various customer classes, and further suppose that $\delta_i = \lambda_i/\rho$ for each class i, corresponding to a FIFO service discipline. The variance parameter σ^2 is then given by, formula (2.11) and, specializing to the case $N = I$, readers may verify that condition (iv) of Theorem 3.1 reduces to

$$\tau_i\alpha_i^2 = \frac{\sigma^2}{2\rho} = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{\rho_j}{\rho}\right)\left(\tau_j\alpha_j^2 + \tau_j\beta_j^2\right) \quad \text{for } i = 1,\dots,n. \tag{3.2}$$

Recalling that $\rho = \rho_1 + 2 \cdots + \rho_n$, one finally deduces that, for the special case in question, the Brownian station model is quasireversible if and only if

$$\tau_i\alpha_i^2 = \sum_{j=1}^{n}\left(\frac{\rho_j}{\rho}\right)\left(\tau_j\beta_j^2\right) \quad \text{for } i = 1,\dots,n. \tag{3.3}$$

If we further assume that $\tau_i = \tau_1$ for each class i, then (3.3) reduces to

$$\alpha_i^2 = \sum_{i=1}^{n}\left(\frac{\rho_j}{\rho}\right)\beta_j^2 \quad \text{for } i = 1,\dots,n. \tag{3.4}$$

That is, the model is quasireversible if and only if the squared coefficient of variation (SCV) for *each* interarrival time distribution equals the indicated *weighted average* of the SCV's for the various service time distributions.

As preparation for the proof of Theorem 3.1, we first develop some alternative forms of of properties (i) and (ii).

Consider the $(n+1)$-dimensional Brownian motion $(A_1, \dots, A_n, X)$. From the expression (2.7) for X and the independence of A from ν, we have

$$\langle A, X\rangle_t = \langle A, \tau' A\rangle_t = G\tau t. \tag{3.5}$$

Recall the $\langle X, X\rangle_t = \sigma^2 t$, where $\sigma^2 > 0$ by assumption. Thus

$$\eta \equiv G\tau/\sigma^2 = \langle A, X\rangle_t / \langle X, X\rangle_t \quad \text{for all } t > 0. \tag{3.6}$$

Now let

$$B = A - \eta X. \tag{3.7}$$

Then B is a Brownian motion, being constituted from the $(2n)$-dimensional Brownian motion $(A_1, \dots, A_n, \nu_1, \dots, \nu_n)$ plus linear drifts. By the choice of η, the correlation of B_i with X is zero for all i and it follows that B is independent of X. Moreover, since $W(0)$ is independent of (A, S), $W(0)$ is independent of the pair (B, X). It follows that B is independent of $(X, W(0))$. Now, from the above, $A = B + \eta X$, and by (2.12) and (2.14), $Q(t) - Q(0) = \delta(X(t) + Y(t))$. Hence, D given by (2.15) may be rewritten as

$$D = B - C, \tag{3.8}$$

where

$$C = \gamma X + \delta Y \tag{3.9}$$

for $\gamma \equiv \delta - \eta$. Since B is independent of $(X, W(0))$ and Y and W are determined by the latter [see (2.12)–(2.13)], it follows that B is independent of (C, W).

LEMMA 3.2. *For each $t \geq 0$,*

$$E[D(t)] = \lambda t \quad \textit{and} \quad \langle D, D\rangle_t = \langle B, B\rangle_t + \Gamma\sigma^2 t, \tag{3.10}$$

where Γ is a $n \times n$ matrix such that $\Gamma_{ij} = \gamma_i\gamma_j$ for all $i, j \in \{1, \dots, n\}$.

PROOF. By taking expectations in the representation

$$D(t) = A(t) - \delta(W(t) - W(0)), \tag{3.11}$$

and recalling that W is initialized with its stationary distribution, we see that $E[D(t)] = E[A(t)] = \lambda t$. The mutual variation of D follows from the decomposition

$$D + B - \gamma X - \delta Y, \tag{3.12}$$

where B and X are independent Brownian motions and Y is locally of bounded variation. □

Lemma 3.3. *Let N be a $k \times n$ matrix for some $k \geq 1$. Then we have the following.*

(i) *For each t, $\{ND(s)\colon 0 \leq s \leq t\}$ is independent of $W(t)$ is and only if $\{NC(s)\colon 0 \leq s \leq t\}$ is independent of $W(t)$.*

(ii) *ND is a Brownian motion if and only if NC is a Brownian motion.*

Proof. For (i), note that NB is independent of (NC, W). One can use this, together with characteristic functions, to verify that for any times $s_1, \dots, s_n, t$: $(ND(s_1), \dots, ND(s_n))$ is independent of $W(t)$ if and only if $(NC(s_1), \dots, NC(s_n))$ is independent of $W(t)$. Then (i) follows.

For (ii), note that NB is independent of NC. The "if" part follows immediately from this. The "only if" part can be proved using characteristic functions to verify that the increments of NC are independent and multivariate normally distributed with the appropriate parameters. For the latter one makes use of Lemma 3.2. □

A key step toward proving Theorem 3.1 is to consider (i) and (ii) restricted to one component at a time. From (3.9) we have

$$NC = (N\gamma)X + (N\delta)Y.$$

Thus, by Lemma 3.3 (with $e_i' \cdot N$ in place if N, were e_i is the unit vector in the ith coordinate direction) the question of whether $\{(ND)_i(s),\ 0 \leq s \leq t\}$ is independent of $W(t)$ or whether $(ND)_i$ is a Brownian motion always reduces to the same question for a one-dimensional process of the form

$$Z = aX + bY, \tag{3.13}$$

where a and b are constants.

At first glance, since X is a Brownian motion and Y is an increasing nonlinear process, the question of whether Z is a Brownian motion may seem trivial. Indeed, with respect to the filteration generated by $(X, W(0))$, Z is the sum of the continuous martingale $a(X(t) + (1 - \rho)t)$ and the continuous locally bounded variation process $bY(t) - a(1 - \rho)t$. In fact, the latter is nonlinear. So how can Z be a Brownian motion? If $Z(t) + (a - b)(1 - \rho)t$ is a continuous local martingale with respect to its *own* filtration, then it will be a Brownian motion because its quadratic variation is the same as that of aX [5, Theorem 6.1]. For this to occur, the filteration of Z must be strictly smaller than that generated by $(X, W(0))$. Thus, from an abstract perspective, the problem of whether Z is a Brownian motion might be viewed as a question of measurability or of lifting of diffusions [4]. However, we do not take such an abstract view here; rather, we use direct methods that center around a time reversal argument to obtain the following.

Theorem 3.4. (i) *$\{Z(s),\ 0 \leq s \leq t\}$ is independent of $W(t)$ for each t if and only if $b = 2a$.*

(ii) *Z is a Brownian motion if and only if either $b = 0$ or $b = 2a$.*

Before proving this, we develop some preliminary results. The first of these will be used in proving the "only if" part of (ii).

LEMMA 3.5. *Suppose $a = 1$ and $b \geq 1$. Then Z is a Brownian motion only if $b = 2$.*

PROOF. Suppose $Z = X + bY$ is a Brownian motion. Substituting for Y from (2.12), we obtain the following alternative representation for Z:

$$Z(t) = (1-b)X(t) + b(W(t) - W(0)). \tag{3.14}$$

In a similar manner to that in Lemma 3.2, it follows from the above representations of Z, that Z has drift $(b-1)(1-\rho)$ and the same variance parameter σ^2 as X. It is known [13, page 197] that for $b > 1$, the reflected minimum, $-\min_{0 \leq s < \infty} Z(s)$ of such a Brownian motion with positive drift is exponentially distributed with parameter $2(b-1)(1-\rho)/\sigma^2$. If $b = 1$, Z is driftless and so has no minimum. On the other hand, by substituting for X from (2.12) in $Z = X + bY$, we obtain

$$Z(t) = W(t) - W(0) + (b-1)Y(t) \geq -W(0), \tag{3.15}$$

where the last inequality follows because $b \geq 1$ and W and Y are nonnegative. Indeed, since Y does not increase until W first reaches zero, $-W(0)$ is the minimum of Z. If $b = 1$, this contradicts the fact that Z has no minimum. For $b > 1$, by comparing the exponential distribution of the reflected minimum of Z with that of $W(0)$, given by (3.1), we see that $b - 1 = 1$. Thus, $b = 2$ is necessary for Z to be a Brownian motion when $a = 1$ and $b \geq 1$. □

We now define some processes obtained by time reversal on a fixed time interval $[0, t]$. These will play a key role in the proof of Theorem 3.4. For $t \geq 0$ fixed and $0 \leq s \leq t$, define

$$Z^*(s) = Z(t) - Z(t-s), \tag{3.16}$$

$$W^*(s) = W(t-s), \tag{3.17}$$

$$Y^*(s) = Y(t) - Y(t-s). \tag{3.18}$$

We shall not use an extra notation to indicate the dependence of these processes on t, for in all applications we shall fix t first and then consider these processes on the time interval $[0, t]$.

LEMMA 3.6. *Fix $t \geq 0$. Let $W^* = \{W^*(s): 0 \leq s \leq t\}$ and $Y^* = \{Y^*(s): 0 \leq s \leq t\}$ be defined as in* (3.17)–(3.18). *Then W^* is a stationary reflecting Brownian motion on the positive half-line with drift $-(1-\rho)$ and variance parameter σ^2. Moreover, Y^* is the local time of W^* at the origin.*

PROOF. Since the one-dimensional reflected Brownian motion W is reversible as a Markov process when initialized with its stationary distribution [20], it follows that the time reversal W^* of W on the time interval $[0, t]$ is equivalent in law to W on $[0, t]$. The local time property of Y^* comes from the fact that Y is the local time of W at the origin and can be characterized as a limit of occupation time [5, Chapter 7]. Specifically for $0 \le s \le t$, we have almost surely

$$\begin{aligned} Y^*(s) = Y(t) - Y(t-s) &= \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{t-s}^{t} 1_{[0,\varepsilon]}(W(u))\,du \\ &= \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{0}^{s} 1_{[0,\varepsilon]}(W^*(u))\,du. \end{aligned} \tag{3.19}$$

Since Y^* is continuous on $[0, t]$ and the last limit above has a continuous version which is the local time of W^* on $[0, t]$, Y^* is indistinguishable from this local time. □

COROLLARY 3.7. *For $t \ge 0$ fixed, (W^*, Y^*) is equivalent in law to (W, Y) on the time interval $[0, t]$.*

PROOF. The local time $\{Y(s),\ 0 \le s \le t\}$ is a functional of $\{W(s),\ 0 \le s \le t\}$ and so the joint law of (W, Y) on the time interval $[0, t]$ is determined by that of W on this interval. Since the law of (W^*, Y^*) is determined in the same way from the law of W^*, the desired result follows from Lemma 3.6. □

LEMMA 3.8. *Fix $t \ge 0$ and let $Z^* = \{Z^*(s)\colon 0 \le s \le t\}$ and $W^* = \{W^*(s)\colon 0 \le s \le t\}$ be defined by* (3.16)–(3.17). *Then*

(i) *$\{Z(s),\ 0 \le s \le t\}$ is independent of $W(t)$ if and only if Z^* is independent of $W^*(0)$.*

(ii) *$\{Z(s),\ 0 \le s \le t\}$ is a Brownian motion if and only if Z^* is a Brownian motion.*

PROOF. The equivalence in part (i) is easy, since $\{Z(s),\ 0 \le s \le t\}$ is recoverable from Z^* as $(Z^*)^*$. One uses the fact that Z starts from the origin for this.

The equivalence in (ii) comes from the characterization of Brownian motion on a finite time interval as a continuous process with stationary independent increments. Note here that both Z and Z^* start from the origin. □

PROOF OF THEOREM 3.4. We first observe that for $t \ge 0$ fixed,

$$\begin{aligned} Z^*(s) = &-a(W^*(s) - W^*(0) - Y^*(s)) \\ &+ (b - 2a)Y^*(s) \quad \text{for } s \in [0, t]. \end{aligned} \tag{3.20}$$

By Corollary 3.7, (3.20) and (2.12), $(Z^*, W^*(0))$ is equivalent in law to $(-aX + (b - 2a)Y, W(0))$ on $[0, t]$. Note, unlike the definition of the former pair, the definition of the latter pair does not vary with t.

For the proof of (i), observe that by Lemma 3.8(i) and paragraph above, $\{Z(s),\ 0 \leq s \leq t\}$ is independent of $W(t)$ for each t if and only if $-aX + (b - 2a)Y$ is independent of $W(0)$. Recall X is independent of $W(0)$. Thus, $-aX + (b - 2a)Y$ is independent of $W(0)$ if $b = 2a$. On the other hand, since Y, given by (2.13), is nonnegative and does not increase until $W(0) + X$ first reaches zero, for $t > 0$ fixed and $b > 2a$, the conditional probability $P(-aX_t + (b - 2a)Y_t > 1 | W(0) = w)$ tends to $P(-aX_t > 1)$ as $w \to \infty$, whereas it is strictly greater than $P(-aX_t > 1)$ for $w = 0$. For $b < 2a$, the conditional probability with < -1 in place of > 1 shows a similar disparity of values depending on the value of w. It follows that for $b \neq 2a$, $-aX + (b - 2a)Y$ is not independent of $W(0)$. Hence (i) follows.

For the proof of (ii), observe that if $b = 0$, then $Z = aX$ is a Brownian motion. On the other hand, if $a = 0$, then $Z = bY$ is an increasing process and so can only be a Brownian motion if it degenerates to a linear drift. But, Y is nonlinear, since Y is zero until W first reaches zero and then it starts to increase. Thus, bY is a Brownian motion if and only if $b = 0$. We have thus verified (ii) when $b = 0$ or $a = 0$, and so for the remainder of the proof we shall assume $a \neq 0$ and $b \neq 0$.

For the "only if" statement in (ii), suppose Z is a Brownian motion. Then so is $X + a^{-1}bY$, and it follows from Lemma 3.5 that we must have $b = 2a$ if $a^{-1}b \geq 1$. On the other hand, if $a^{-1}b < 1$, then by Lemma 3.8(ii), for each $t \geq 0$, $Z^* = \{Z^*(s)\colon 0 \leq s \leq t\}$ is a Brownian motion on $[0, t]$. Hence, by the first paragraph of this proof, $-aX + (b - 2a)Y$ is a Brownian motion on $[0, \infty)$. Since $a^{-1}(2a - b) > 1$ when $a^{-1}b < 1$, it then follows from Lemma 3.5 that $a^{-1}(2a - b) = 2$, i.e., $b = 0$, a case we have already excluded. Thus, we have proved the "only if" statement in (ii). For the "if" statement, suppose $b = 2a$. Then by the first paragraph of this proof, for each $t \geq 0$, Z^*, being equivalent in law to $\{-aX(s),\ 0 \leq s \leq t\}$, is a Brownian motion, and hence by Lemma 3.8(ii), Z is a Brownian motion on $[0, \infty)$. This completes the proof of (ii) in Theorem 3.4. □

Before proceeding with the proof of Theorem 3.1, we pause to note the connection of the "if" part of Theorem 3.4(ii) with Williams' [29] path decomposition of a one-dimensional Brownian motion with positive drift. For this, suppose $a = 1$ and $b = 2$. Then $Z = X + 2Y = W - W(0) + Y$ is a Brownian motion starting from the origin in $\mathbb{R}$ with drift $1 - \rho$ and variance parameter σ^2 (cf. Lemma 3.5). Define the stopping time

$$T = \inf\{t \geq 0\colon X(t) = -W(0)\}.$$

Then, since Y does not increase until the time T when W first reaches the origin, Z may be decomposed:

$$Z(t) = \begin{cases} X(t) & \text{for } 0 \leq t \leq T, \\ -W(0) + \hat{Z}(t - T) & \text{for } T \leq t < \infty, \end{cases} \tag{3.21}$$

where

$$\hat{Z}(t) = \hat{X}(t) + 2\hat{Y}(t)$$

for

$$\hat{X}(t) = X(t+T) - X(T),$$
$$\hat{Y}(t) = \sup_{0 \le s \le t} (-\hat{X}(s)).$$

Here $\hat{X}$ is a Brownian motion starting from the origin with drift $-(1-\rho)$ and variance parameter σ^2, and $\hat{Y}$ is the maximum process of $-\hat{X}$. If follows from the results of Rogers and Pitman [24] that $\hat{Z}$ is a diffusion process on $\mathbb{R}_+$, with infinitesimal generator

$$\frac{\sigma^2}{2}\frac{d^2}{dx^2} + (1-\rho)\coth\left(\frac{1-\rho}{\sigma^2}x\right)\frac{d}{dx}. \tag{3.22}$$

In fact, $\hat{Z}$ is equivalent in law to the radial part of a three-dimensional Brownian motion with drift of magnitude $(1-\rho)$. Moreover, $\hat{Z}$ is independent of $W(0)$ and $Z(\cdot \wedge T)$. Thus, the above gives an alternative verification of Williams' [29, Theorem 2.1] path decomposition of a Brownian motion with positive drift.

Remark. As pointed out by the referee, if one assumes the results of Rogers and Pitman [24], Theorem 3.4 can be proved without using time reversal. We chose to give the time reversal argument here because it is independent of the results of [24] and indicates that there is a connection between the notions of time reversal and quasireversibility.

Proof of Theorem 3.1. Recall that $\gamma = \delta - \eta$ and η is given by (3.6). Thus, the condition $N\delta = 2N\gamma$ is equivalent to $N\delta = 2N\eta$, which is equivalent to condition (iv) of Theorem 3.1.

We first prove the equivalence of (i) and (iv). This in fact does not require the assumption that $(N\delta)_i \neq 0$ for some i. If (i) holds, then by Lemma 3.3(i) and Theorem 3.4(i), $N\delta = 2N\gamma$, which is equivalent to (iv). On the other hand, if (iv) holds, then

$$NC = N\gamma(X + 2Y), \tag{3.23}$$

where $\{(X+2Y)(s),\ 0 \le s \le t\}$ is independent of $W(t)$ for each $t \ge 0$, by Theorem 3.4(i). Then (i) follows from this and Lemma 3.3(i).

We next prove that (ii) is equivalent to (iv). First suppose (iv) holds. Then NC is given by (3.23), where by Theorem 3.4(ii), $X + 2Y$ is a Brownian motion and so NC is a (degenerate) Brownian motion. Then, ND is a Brownian motion by Lemma 3.3(ii). Thus, (iv) implies (ii). Conversely, suppose (ii) holds. Then, by Lemma 3.3(ii), NC is a Brownian motion and by Theorem 3.4(ii), for each $i \in \{1,\dots,k\}$, either $(N\delta)_i = 0$ or $(N\delta)_i = 2(N\gamma)_i$. For a proof by contradiction, suppose (iv) does not hold, i.e., $(N\delta)_j \neq 2(N\gamma)_j$ for some $j \in \{1,\dots,k\}$. Then, $(N\delta)_j = 0$, but $(N\gamma)_j \neq 0$. Let $i \in \{1,\dots,k\}$ such that

$(N\delta)_i \neq 0$. We must have $(N\delta)_i = 2(N\gamma)_i$ and $(N\gamma)_i \neq 0$. Then, $((NC)_i, (NC)_j) = ((N\gamma)_i(X + 2Y), (N\gamma)_j X)$ is a Brownian motion and since $(N\gamma)_i(N\gamma)_j \neq 0$, it follows by taking appropriate linear combinations that Y is a Brownian motion. But, as shown in the third paragraph of the proof of Theorem 3.4, this nonlinear increasing process cannot be a Brownian motion. Thus we have obtained the desired contradiction and it follows that (ii) implies (iv).

Finally, we prove the equivalence of (iii) and (iv). By Lemma 3.2 and the bilinearity of $\langle \cdot , \cdot \rangle$, we have

$$\langle (ND)_i, (ND)_j \rangle_t = \langle (NB)_i, (NB)_j \rangle_t + (N\gamma)_i (N\gamma)_j \sigma^2 t. \tag{3.24}$$

Here $(NB)_i = (NA)_i - (N\eta)_i X$ where by (3.6), $(N\eta)_i = \langle (NA)_i, X \rangle_t / \langle X, X \rangle_t$ for $t > 0$, and $\langle X, X \rangle_t = \sigma^2 t$. Hence

$$\begin{aligned} \langle (NB)_i, (NB)_j \rangle_t &= \langle (NA)_i, (NA)_j \rangle_t - (N\eta)_i \langle (NA)_j, X \rangle_t \\ &\quad - (N\eta)_j \langle (NA)_i, X \rangle_t + (N\eta)_i (N\eta)_j \langle X, X \rangle_t \\ &= \langle (NA)_i, (NA)_j \rangle_t - (N\eta)_i (N\eta)_j \sigma^2 t, \end{aligned} \tag{3.25}$$

and so

$$\begin{aligned} \langle (ND)_i, (ND)_j \rangle_t &= \langle (NA)_i, (NA)_j \rangle_t \\ &\quad + ((N\gamma)_i (N\gamma)_j - (N\eta)_i (N\eta)_j) \sigma^2 t. \end{aligned} \tag{3.26}$$

Note that (iv) is equivalent to $N\gamma = N\eta$ and so (iv) clearly implies (iii). Conversely, suppose (iii) holds. Then

$$(N\gamma)_i (N\gamma)_j = (N\eta)_i (N\eta)_j \quad \text{for all } i, j \in \{1, \dots, k\}. \tag{3.27}$$

Now let $i \in \{1, \dots, k\}$ such that $(N\delta)_i \neq 0$. By setting $i = j$ in (3.27), we see that $(N\gamma)_i = (N\eta)_i$ or $(N\gamma)_i = -(N\eta)_i$. However, the last equality is equivalent to $(N\delta)_i = 0$, which we have excluded by the choice of i. Thus, $(N\gamma)_i = (N\eta)_i$, or equivalently, $(N\delta)_i = 2(N\gamma)_i$, where $(N\gamma)_i \neq 0$, since $(N\delta)_i \neq 0$. Then, considering (3.27) for all $j \neq i$, we obtain $(N\gamma)_j = (N\eta)_j$. It follows that (iii) implies (iv). □

APPENDIX

Brownian models as heavy traffic approximations. In this appendix we describe a conventional queueing model of a multiclass service station with independent renewal inputs and we explain how one approximates such a system by a multiclass Brownian model of the type defined in Section 2. Also, the heavy traffic limit theory that justifies such an approximation is briefly and informally reviewed. Our account is based loosely on results reported in [6, 10, 21, 23, 27], but there are many other previous papers that might equally well be cited.

Consider a single-server station at which customers of classes $1, \dots, n$ arrive according to independent renewal processes. Let $\{A_i(t),\ t \geq 0\}$ be the arrival process for class i, with $A_i(0) = 0$, let $\{s_i(k),\ k = 0, 1, \dots\}$ be the corresponding i.i.d. sequence of nonnegative service time random variables $(i = 1, \dots, n)$. The service time sequences for the different customer classes are assumed to be mutually independent of one another and of the arrival processes. We denote by λ_i the mean arrival rate for class i customers and by α_i the coefficient of variation for the class i interarrival times, assuming that the interarrival time distribution has finite second moment. It is well known that

$$\text{(A.1)} \qquad E[A_i(t)] \sim \lambda_i t \quad \text{and} \quad \mathrm{Var}[A_i(t)] \sim (\lambda_i \alpha_i^2) t$$

as $t \to \infty$ $(i = 1, \dots, n)$, where $\sim$ denotes "is asymptotic to" in the usual sense that the term on the left of this sign, when divided by the term on the right, tends to 1 as $t \to \infty$. Next, let τ_i and β_i denote the mean and the coefficient of variation, respectively, of the class i service time distribution, assuming that distribution has finite second moment $(i = 1, \dots, n)$. Define the service time partial sums $S_i(k) = s_i(1) + \cdots + s_i(k)$ for $k = 1, 2, \dots,$ with $S_i(0) = 0$ by convention $(i = 1, \dots, n)$. Thus

$$\text{(A.2)} \qquad E[S_i(k)] = \tau_i k \quad \text{and} \quad \mathrm{Var}[S_i(k)] = (\tau_i \beta_i)^2 k$$

for $i = 1, \dots, n$ and $k = 0, 1, \dots$. The total server work embodied in class i customers who arrive up to time t is given by

$$\text{(A.3)} \qquad L_i(t) = S_i(A_i(t)), \qquad t \geq 0,$$

and we then define the workload netflow process

$$\text{(A.4)} \qquad X(t) = \sum_{i=1}^{n} L_i(t) - t, \qquad t \geq 0.$$

Let us denote by $W(t)$ the amount of server work (expressed in units of time) required to complete processing of all customers who remain in the system at time t, taking the initial workload $W(0)$ to be an arbitrary nonnegative random variable. Under very general conditions (see below), the workload process W, which is also called the *virtual waiting time process* in queueing theory, is given by

$$\text{(A.5)} \qquad W(t) = W(0) + X(t) + Y(t), \qquad t \geq 0,$$

where

$$\text{(A.6)} \qquad Y(t) = \sup_{0 \leq s \leq t} (W(0) + X(s))^-, \qquad t \geq 0.$$

For (A.5)–(A.6) to hold, one need only assume a work conserving queue discipline, which means that (a) the server continues to work at full capacity as long as any customer remains in the system and (b) the service times of arriving customers are given by the sequences $\{s_i(k)\}$, regardless of the order or manner in which those customers may be served. Nothing more will be said about the queue discipline at this point, but readers may think in terms of a standard first-in-first-out (FIFO) discipline, in which customers are served in the order of their arrival, without regard to class.

To establish the connection between the conventional queueing model described above and our multiclass Brownian station model, we first define the centered processes

$$\hat{A}_i(t) = A_i(t) - \lambda_i t, \qquad t \geq 0 \tag{A.7}$$

and

$$\hat{S}_i(k) = S_i(k) - k\tau_i, \qquad k = 0, 1, \ldots \tag{A.8}$$

for $i = 1, \ldots, n$, and we write

$$W = \phi(W(0) + X), \tag{A.9}$$

where ϕ is simply the path-to-path mapping (the *one-sided reflection mapping* or *one-sided regulator*) defined by (A.5)–(A.6). Also, for future purposes, we extend $\hat{S}_i(\cdot)$ to a process with time domain $[0, \infty)$ via

$$\hat{S}_i(t) = \hat{S}_i(k) \quad \text{for } k = 0, 1 \ldots \text{ and } t \in [k, k+1). \tag{A.10}$$

Thus, $\hat{S}_i$ is piecewise constant and right continuous, and from (A.2) it is obvious that

$$E[\hat{S}_i(t)] = 0 \quad \text{and} \quad \operatorname{Var}[\hat{S}_i(t)] \sim (\tau_i \beta_i)^2 t \tag{A.11}$$

as $t \to \infty$ $(i = 1, \ldots, n)$.

Now observe that the basic relationship (A.3) can be expressed in terms of centered processes as

$$\begin{aligned} L_i(t) &= \tau_i A_i(t) + \hat{S}_i(A_i(t)) \\ &= \rho_i t + \tau_i \hat{A}_i(t) + \hat{S}_i(A_i(t)) \end{aligned} \tag{A.12}$$

and consequently

$$X(t) = -(1-\rho)t + \sum_{i=1}^{n} \left[\tau_i \hat{A}_i(t) + \hat{S}_i(A_i(t))\right], \tag{A.13}$$

where $\rho_i = \lambda_i \tau_i$ and $\rho = \rho_1 + \cdots + \rho_n$, as in Section 2. Because we have assumed independent renewal input processes in specifying our conventional queueing model, the corresponding Brownian system model has input covariance matrix $G = \operatorname{diag}(\lambda_1 \alpha_1^2, \ldots, \lambda_n \alpha_n^2)$. By comparing the development in Section 2 against the definitions laid out in this Appendix, readers will see that the Brownian model differs from the conventional model, at least thus far, only in the following regards. First, in forming the Brownian model, the centered arrival process $\hat{A}_i$ is replaced by a driftless Brownian motion ξ_i whose variance parameter matches the asymptotic variance of $\hat{A}_i$. Second, the centered process $\hat{S}_i \circ A_i$ is replaced by $\nu_i \circ \Lambda_i$, where $\Lambda_i(t) = \lambda_i t$ and ν_i is a driftless Brownian motion whose variance parameter matches the asymptotic variance of $\hat{S}_i$.

To rigorously justify the substitutions described above, one must consider a rescaling of time and state space. Specifically, for a large integer N let us define the scaled processes

$$\xi_i^N(t) = N^{-1/2}\hat{A}_i(Nt), \qquad \nu_i^N(t) = N^{-1/2}\hat{S}_i(Nt)$$

and the scaled random time-change

$$\Lambda_i^N(t) = N^{-1}A_i(Nt).$$

Using standard results in weak convergence theory, it is easy to show that

$$\left(\xi_i, \nu_i^N \circ \Lambda_i^N; i = 1, \dots, n\right) \Rightarrow \left(\xi_i, \nu_i \circ \Lambda_i; i = 1, \dots, n\right)$$

as $N \to \infty$, where $\Rightarrow$ signifies weak convergence in an appropriate function space. From this it follows that

$$\text{(A.14)} \qquad \hat{X}^N \equiv \sum_{i=1}^{n} \left(\tau_i \xi_i^N + \nu_i^N \circ \Lambda_i^N\right) \Rightarrow \hat{X} \quad \text{as } N \to \infty,$$

where $\hat{X}$ is a driftless Brownian motion whose variance parameter σ^2 is given by formula (2.11), corresponding to a Brownian station model in which $\xi_1, \dots, \xi_n$, $\nu_1, \dots, \nu_n$ are independent. Now suppose that the traffic intensity ρ of the queueing model is less than but close to 1. Let N be a large integer such that $\vartheta \equiv N^{1/2}(1 - \rho) > 0$ is of moderate size and define the scaled process

$$\text{(A.15)} \qquad X^N(t) = N^{-1/2}X(Nt) \quad \text{and} \quad W^N(t) = N^{-1/2}W(Nt).$$

Recall from (A.9) that $W = \phi(W(0) + X)$, where ϕ is the one-sided regulator. Assuming for simplicity that $W(0) = 0$ (the case of a positive initial workload involves a bit more care), it is easy to show that

$$\text{(A.16)} \qquad W^N = \phi(X^N).$$

That is, the one-sided regulator ϕ commutes with the scaling of time and state embodied in (A.15). Because $X^N(t) = \hat{X}^N(t) - \vartheta t$, the limit (A.14) justifies approximation of X^N by a Brownian motion with drift parameter $-\vartheta$ and variance parameter σ^2, and then because ϕ is appropriately continuous, (A.16) justifies approximation of W^N by reflected or regulated Brownian motion with the same parameters. One may compactify the latter statement by saying that (A.14) and (A.16) justify the approximation of W^N by a $(-\vartheta, \sigma^2)$ reflected Brownian motion (or RBM) with state space $[0, \infty)$.

Now let $Q_i(t)$ denote the number of class i customers present at time t and let $D_i(t)$ denote the total number of class i departures up to time t. To relate Q (and hence D) to W, one must specify a queue discipline. Assuming a standard FIFO discipline to begin with, let us return to the heavy traffic scenario described in the previous paragraph and define the scaled process

$$Q_i^N(t) = N^{-1/2}Q_i(Nt).$$

By specializing the limit theorems of Peterson [21] or Reiman [23] , one can justify the approximation

$$\text{(A.17)} \qquad Q_i^N \simeq \delta_i W^N,$$

for large N, where $\delta_i = \lambda_i/\rho$ in the case of FIFO. More generally, suppose that the classes are served according to a static priority ranking and that classes $1, \dots, m$ are tied for bottom priority ($1 \leq m \leq n$). This means that customers of classes $1, \dots, m$ are served on a first-in-first-out basis and customers of classes $m+1, \dots, n$ are given priority over classes $1, \dots, m$. The aforementioned limit theorems can be invoked to justify (A.17) with

$$\delta_i = \begin{cases} \lambda_i/(\rho_1 + \cdots + \rho_m) & \text{for } i = 1, \dots, m, \\ 0 & \text{for } i = m+1, \dots, n. \end{cases}$$

To be precise, one considers a sequence of systems with $\rho \uparrow 1$ and $N \to \infty$ in such a way that $N^{1/2}(1-\rho) \to \vartheta > 0$. The limit theory shows that

$$\text{(A.18)} \qquad \left(W^N; Q_i^N \colon i = 1, \dots, n\right) \Rightarrow \left(W^\dagger; \delta_i W^\dagger \colon i = 1, \dots, n\right),$$

where $W^\dagger$ is the $(-\vartheta, \sigma^2)$ RBM referred to in the previous paragraph. It appears that other types of queue disciplines will give similar results with different weights δ_i, but relatively little work has been done on this interesting topic thus far. Also, although we have spoken in terms of a single-server station, the results of Iglehart and Whitt [10] and of others strongly suggest that multiserver stations give rise to exactly the same class of Brownian station models as their heavy traffic limits.

The preceeding discussion can be summarized and extended in minor and obvious ways as follows. Under heavy traffic conditions, if one first scales and centers appropriately, the joint distribution of the processes

$$(W; A_i, Q_i, D_i \colon i = 1, \dots, n)$$

is well approximated by the joint distribution of the processes that were denoted by the same letters in Section 2. The proviso about centering and scaling, although important for purposes of rigorous proofs, actually plays no role when one seeks to interpret or apply the approximation in a concrete setting. That is, for all practical purposes the apparently naive exposition given in Section 2 provides a general, systematic and correct procedure for development of an approximate Brownian system model.

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. Assoc. Comput. Mach.* **22** 248–260.

[2] Burke, P. J. (1956). The output of a queueing system. *Oper. Res.* **4** 699–704.

[3] Burke, P. J. (1968). The output of a stationary *M/M/s* queueing system. *Ann. Math. Statist.* **39** 1144–1152.

[4] Carverhill, A. (1988). Conditioning a lifted stochastic system in a product space. *Ann. Probab.* **16** 1840–1853.

[5] Chung, K. L. and Williams, R. J. (1983). *Introduction to Stochastic Integration*. Birkhäuser, Boston.

[6] Harrison, J. M. (1973). A limit for priority queues in heavy traffic. *J. Appl. Probab.* **10** 907–912.

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

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

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

[10] IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic. I, II. *Adv. in Appl. Probab.* **2** 150–177, 355–364.

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

[12] JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. Springer, New York.

[13] KARATZAS, I. and SHREVE, S. E. (1988). *Brownian Motion and Stochastic Calculus*. Springer, New York.

[14] KELLY, F. P. (1975). Networks of queues with customers of different types. *J. Appl. Probab.* **12** 542–554.

[15] KELLY, F. P. (1976). Networks of queues. *Adv. in Appl. Probab.* **8** 416–432.

[16] KELLY, F. P. (1976). The departure process from a queueing system. *Math. Proc. Cambridge Philos. Soc.* **80** 283–286.

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

[18] 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. Birkhäuser, Boston.

[19] MUNTZ, R. R. (1972). Poisson departure processes and queueing networks. IBM Research Report RC 4145, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.

[20] NAGASAWA, M. (1961). The adjoint process of a diffusion with reflecting barrier. *Kodai Math. Sem. Rep.* **13** 235–248.

[21] PETERSON, W. P. (1990). A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Oper. Res.* To appear.

[22] REIMAN, M. I. (1987). A network of priority queues in heavy traffic: One bottleneck station. Technical Memorandum, AT & T Bell Laboratories, Murray Hill, N.J.

[23] REIMAN, M. I. (1988). A multiclass feedback queue in heavy traffic. *Adv. in Appl. Probab.* **20** 179–207.

[24] ROGERS, L. C. G. and PITMAN, J. W. (1981). Markov functions. *Ann. Probab.* **9** 573–582.

[25] WALRAND, J. (1983). A probabilistic look at networks of quasi-reversible queues. *IEEE Trans. Inform. Theory* **IT-29** 825–831.

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

[27] WHITT, W. (1971). Weak convergence theorems for priority queues: Preemptive resume discipline. *J. Appl. Probab.* **8** 74–94.

[28] WHITTLE, P. (1986). *Systems in Stochastic Equilibrium*. Wiley, New York.

[29] WILLIAMS, D. (1974). Path decomposition and continuity of local time for one-dimensional diffusions. I. *Proc. London Math. Soc.* (3) **28** 738–768.

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA AT SAN DIEGO
LA JOLLA, CALIFORNIA 92093