Reflecting Diffusions and Queueing Networks

R. J. Williams[1]

1 Introduction

Queueing models are of interest for analyzing congestion and delay in complex processing networks such as those occurring in computer systems, telecommunications and manufacturing (see e.g., [BG92, Ya94]). Many of these networks can process more than one class of job at a given station (so-called *multiclass* networks) and/or have complex feedback structures. Generally such models cannot be analyzed exactly and it is natural to seek more tractable approximations. In connection with this, certain diffusion processes known as *semimartingale reflecting Brownian motions (SRBMs)* [RW88] have been proposed as approximations for heavily loaded queueing networks (see e.g., [Ha88, HN93]), and there is now a substantial theory for these diffusions (see the survey in [Wi95]). However, limit theorems justifying their role as approximations have only been proved for some networks (see the overview in [Wi96]). Indeed, since a surprising example of Dai and Wang [DWa93] it has been known that these approximations are not always valid for multiclass networks with feedback. A challenging open problem has been that of establishing general conditions under which SRBM approximations for open multiclass queueing networks are valid. Recent progress on this problem and related work is summarized here.

The paper is organized as follows. In §2, the existence and uniqueness theory for SRBMs is described, including an oscillation inequality [Wi97a] which is critical to establishing tightness of normalized queueing network processes. In §3, the model used here for an open multiclass queueing network is defined. In §4, the main theorem is stated which gives general sufficient conditions for a heavy traffic limit theorem, which justifies approximating an open multiclass queueing network by a SRBM [Wi97b]. One of the key conditions involves something called *"state space collapse"*. Bramson has recently given sufficient conditions for this to hold (see [Br97b] and his article [Br98] in this volume). New heavy traffic limit theorems for two interesting collections of networks are obtained by combining the above results. The paper concludes with some open problems in §5.

2 Semimartingale Reflecting Brownian Motions

Definition of a SRBM Let J be a positive integer, $\mathbb{R}_+^J \equiv \{x \in \mathbb{R}^J : x_j \geq 0 \text{ for } j = 1, \ldots, J\}$, $\mathcal{B}$ denote the σ-algebra of Borel subsets of $\mathbb{R}_+^J$, ν be a probability measure on $(\mathbb{R}_+^J, \mathcal{B})$, θ be a constant vector in $\mathbb{R}^J$, Γ be a $J \times J$ non-degenerate covariance matrix, and R be a $J \times J$ matrix.

[1] Research supported in part by the U.S. National Science Foundation.

Definition 2.1 *A semimartingale reflecting Brownian motion (SRBM) associated with the data (θ, Γ, R, ν) is a J-dimensional process W defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ such that*

$$W = X + RY \tag{1}$$

where W, X, Y are $\{\mathcal{F}_t\}$-adapted processes such that W has continuous paths in $\mathbb{R}_+^J$, X is a J-dimensional Brownian motion with drift vector θ, covariance matrix Γ, initial distribution ν, and $\{X(t) - X(0) - \theta t, \mathcal{F}_t, t \geq 0\}$ is a martingale, and Y is a J-dimensional process such that for each $j \in \{1, \ldots, J\}$, $Y_j(0) = 0$, Y_j is continuous and non-decreasing, and $\int_0^\infty 1_{(0,\infty)}(W_j(s))dY_j(s) = 0$, i.e., Y_j can increase only when W_j is zero.

Intuitively, such a SRBM behaves in the interior of the orthant $\mathbb{R}_+^J$ like a Brownian motion with initial distribution ν, constant drift θ and covariance matrix Γ, and it is confined to $\mathbb{R}_+^J$ by "pushing" at the boundary, where for $j = 1, \ldots, J$, the allowed direction of push on the relative interior of the boundary face $F_j = \{x \in \mathbb{R}_+^J : x_j = 0\}$ is given by the j^{th} column of the matrix R. At an intersection of faces, the allowed directions of "push" are given by the convex combinations of the push directions associated with the faces meeting there. For historical reasons, stemming from an alternative construction of the driftless process in one-dimension, the "pushing" at the boundary is called instantaneous reflection. However, it is more accurate to think of this action as deflection or regulation rather than some type of mirror reflection. The process Y is called the "pushing process" associated with W and it is related to the local time of W on the boundary. Since the state space for a SRBM is not smooth and the directions of reflection may be discontinuous at the non-smooth parts of the boundary, the general theory for diffusions with smooth boundary conditions [SV71] does not apply to SRBMs and one must develop a theory from first principles.

The above definition of a SRBM is in the spirit of weak solutions of stochastic equations. In particular, one is free to choose the filtered probability space and processes W, X, Y such that the above properties hold. Here the focus is on such weak solutions, since necessary and sufficient conditions for their existence and uniqueness are known, whereas only sufficient conditions are known for strong solutions. Furthermore, there are multiclass queueing networks (see the example due to Dai, Wang and Wang in Appendix A of [Wi97b]) whose SRBM approximants are not covered by the extant strong solution theory.

Existence and Uniqueness for SRBMs It is straightforward to see that a necessary condition for the existence of a SRBM associated with (θ, Γ, R, ν) for each probability measure ν on $(\mathbb{R}_+^J, \mathcal{B})$ is the following: at each point on the boundary of $\mathbb{R}_+^J$ there is a positive linear combination of the "push" directions that can be used there which points into the interior of $\mathbb{R}_+^J$. This geometric description can be expressed succinctly as the following algebraic condition: the matrix R is *completely-$\mathcal{S}$* if for each principal submatrix $\tilde{R}$ of R there is a vector $\tilde{x} \geq 0$ such that $\tilde{R}\tilde{x} > 0$. (Here inequalities are to be interpreted componentwise and a principal submatrix of R is obtained by deleting all rows and columns of R

with indices in some strict (possibly empty) subset of $\{1,\dots,J\}$.) In fact, R being completely-$\mathcal{S}$ is also sufficient for the existence and uniqueness in law of a SRBM. The following result is proved for $\nu = \delta_x$ (the unit mass at $x \in \mathbb{R}_+^J$) in [TW93] and is easily extended to all ν [Wi97a].

Theorem 2.1 *Suppose that R is completely-$\mathcal{S}$. There exists a SRBM associated with (θ,Γ,R,ν) and it is unique in law. Furthermore, the laws induced on the space of continuous paths in $\mathbb{R}_+^J$ by the SRBMs associated with $(\theta,\Gamma,R,\delta_x)$, $x \in S$, define a Feller continuous strong Markov process.*

Oscillation Inequality Solutions of a deterministic Skorokhod problem have been used to obtain strong constructions of SRBMs in some cases [DuI91, HR81]. While this Skorokhod problem will not have unique solutions for general completely-$\mathcal{S}$ matrices R [BEK91, Ma92], an oscillation inequality for a perturbed form of this problem can be used to establish tightness for suitable approximations to a SRBM. Indeed, this inequality can be used to show existence of a SRBM (using deflected random walk approximations having small inward jumps at the boundary) and the form obtained by restricting to continuous paths $x(\cdot)$ and setting $\epsilon = 0$ is used in the proof of uniqueness in law of a SRBM [TW93]. (This "continuous" case of the oscillation inequality first appeared in [BEK91].)

In the following statement of the oscillation inequality, for any $0 \le t_1 < t_2 < \infty$, $D([t_1,t_2],\mathbb{R}^J)$ denotes the set of functions $x : [t_1,t_2] \to \mathbb{R}^J$ that are right continuous on $[t_1,t_2)$ and have finite left limits on $(t_1,t_2]$ and $\mathrm{Osc}(x,[t_1,t_2]) = \sup\{|x(t)-x(s)| : t_1 \le s < t \le t_2\}$ for any $x \in D([t_1,t_2],\mathbb{R}^J)$, where $|a| = \max_{j=1}^J |a_j|$ for any $a \in \mathbb{R}^J$.

Theorem 2.2 [Wi97a] *Assume that R is completely-$\mathcal{S}$. Suppose that $\epsilon \ge 0$, $0 \le t_1 < t_2 < \infty$ and $w,x,y \in D([t_1,t_2],\mathbb{R}^J)$ are such that*

(i) *$w(t) = x(t) + Ry(t) \in \mathbb{R}_+^J$ for all $t \in [t_1,t_2]$,*

(ii) *for each $j \in \{1,\dots,J\}$, $y_j(t_1) \ge 0$, y_j is non-decreasing, and $\int_{[t_1,t_2]} 1_{(\epsilon,\infty)}(w_j(s))dy_j(s) = 0$.*

Then there is a constant $C > 0$, depending only on R, such that

$$\mathrm{Osc}(y,[t_1,t_2]) + \mathrm{Osc}(w,[t_1,t_2]) \le C(\mathrm{Osc}(x,[t_1,t_2]) + \epsilon). \tag{2}$$

This oscillation inequality plays a key role in establishing tightness of normalized queueing network processes approximating SRBMs (cf. §4).

Other Results and Extensions For further discussion of SRBMs, including weak versus strong solutions, conditions for recurrence, and characterization of stationary distributions, see the survey article [Wi95] and references therein. Semimartingale reflecting Brownian motions in convex polyhedrons (in contrast to the orthant) can arise as approximations to closed and capacitated queueing networks. The reader is referred to [DWi95] for sufficient conditions for the existence and uniqueness of such processes and to [DD97] for a related oscillation inequality and heavy traffic limit theorem. Semimartingale reflecting Brownian motions in polyhedrons also arise in other applications, e.g., in economic models of monetary exchange [FL98]. Reflecting Brownian motions (RBMs) that are not semimartingales have also been proposed as approximations to some particular queueing network models (see e.g., [DuR98b, KL93]). However, the theory of existence and uniqueness for these non-semimartingale RBMs is not as complete as for SRBMs, being restricted to the two-dimensional case [VW85] or to RBMs whose geometric data is a limit of that for SRBMs [DuR98a].

3 Open Multiclass Queueing Network Model

In an open queueing network, jobs arrive from outside the system, visit a finite number of stations where they receive service, and then exit the network. The model for an open multiclass queueing network used here is a generalization of one with a first-in-first-out (FIFO) service discipline considered in [HN93]. To simplify the exposition, attention is restricted to networks that are initially empty. For a more complete specification of the model, including a treatment of networks that are initially non-empty, see [Wi97b]. The model description is broken down into assumptions concerning the network structure, primitive stochastic processes (for exogenous arrivals, service times and routing), and the service discipline.

Network Structure The model has a fixed set $\{1,\ldots,J\}$ of stations with a single reliable server at each. At any given time, each job in the network belongs to one of a finite set $\mathcal{K}=\{1,\ldots,K\}$ of job classes. Each class is associated with exactly one station (where the class is to receive service). The deterministic many-to-one function mapping classes to stations is specified by a $J\times K$ constituency matrix C where $C_{jk}=1$ if class k is served at station j and $C_{jk}=0$ otherwise. At a given station, jobs of different classes may be distinguished by features such as the distributions of their service times, their routing characteristics, or their order of service. Upon completing service in a class, a job changes class in Markovian fashion. Each station serves at least one class and has an infinite buffer for storing jobs awaiting service there.

Stochastic Primitives The primitive stochastic processes for the model are (E,V,Φ) where E is a K-dimensional external arrival process, V is a K-dimensional cumulative service time process, $\Phi=(\Phi^1,\Phi^2,\ldots,\Phi^K)$ and Φ^k is a K-dimensional routing process for class $k\in\mathcal{K}$. More precisely, for each k and $t\geq 0$, $E_k(t)$ represents the number of exogenous arrivals to class k up to time t. It is assumed that $E_k\not\equiv 0$ for at least one k and for each such k, E_k is a renewal process derived from a sequence of positive i.i.d. interarrival times having finite mean and variance. For each class k and integer $n\geq 0$, $V_k(n)=\sum_{i=1}^n v_k(i)$ where $\{v_k(i)\}_{i=1}^\infty$ is a sequence of i.i.d. positive random variables with finite mean and variance, and $v_k(i)$ is interpreted as the service time for the i^{th} job that arrives to class k. To describe the Markovian routing, let $e_1,\ldots,e_K$ denote the non-negative unit basis vectors parallel to the K coordinate axes in $\mathbb{R}^K$ and let e_0 be the K-dimensional zero vector. For each class k and integer $n\geq 0$, $\Phi^k(n)=\sum_{i=1}^n\phi^k(i)$ where $\{\phi^k(i)\}_{i=1}^\infty$ is a sequence of i.i.d. random vectors taking values in $\{e_0,e_1,\ldots,e_K\}$ with $P(\phi^k(i)=e_l)=P_{kl}$, $k,l\in\mathcal{K}$, and P is a strictly substochastic $K \times K$ matrix. The interpretation of the routing vector $\phi^k(i)$ is that the i^{th} job to depart from class k is routed next to class l if $\phi^k(i) = e_l$ and it leaves the network if $\phi^k(i) = e_0$. The strict substochasticity of P ensures that jobs eventually leave the network. The processes $E_1, \dots, E_K, V_1, \dots, V_K, \Phi^1, \dots, \Phi^K$ are assumed to be mutually independent.

Service Discipline It remains to specify the order in which jobs are served at each station, i.e., the service discipline. Attention is confined to HL (head-of-the-line) service disciplines (cf. [Br97a, Wi97b]). (Other disciplines such as last-in-first-out or general processor sharing are also of interest, but the heavy traffic theory for networks with these disciplines is much less developed.) Firstly, an HL discipline is non-idling in the sense that a server is never idle when there are jobs waiting to be served at its station. In addition, jobs in each class are served on a first-in-first-out basis, i.e., service for each class is concentrated on the job at the head-of-the-line for that class. Each class receives a proportion (possibly zero) of the associated server's time, where this proportion may be random but is kept constant between changes in the arrival or departure processes, and these proportions depend in a measurable way on the "state" of the queueing network at the time of the last such change. (The "state" description includes such quantities as queue lengths, remaining service times of jobs at a station, amounts of time that jobs have been waiting in their current class, and the amount of time until the next exogenous arrival to each class cf. [Wi97b].) Common service disciplines included in the HL framework are FIFO (regardless of their class designation, jobs at a station are served in the order in which they arrived there), static priorities (classes at a station are ranked and jobs of a higher ranking class are always served before those of a lower ranking class), and HLPPS (head-of-the-line proportional processor sharing: each class at a station receives service in proportion to the number of jobs that are present in that class).

Descriptive Processes and Model Equations Let A, D be the K-dimensional processes such that $A_k(t)$ denotes the number of arrivals to, and $D_k(t)$ denotes the number of departures from, class k up to time t. The processes that are used to measure performance are a K-dimensional queue length process Z, a J-dimensional workload process W and a J-dimensional cumulative idletime process Y. For each class k, station j and time t, $Z_k(t)$ denotes the number of class k jobs that are in queue or being served at time t (the letter Z is mnemonic for the German *Zahl* or number), $W_j(t)$ denotes the amount of work for server j (measured in units of remaining service time) that is embodied in those jobs that are at station j at time t, $Y_j(t)$ denotes the total amount of time that server j has been idle up to time t.

The descriptive processes (A, D, W, Y, Z) satisfy the following equations:

$$A(t) = E(t) + \Phi(D(t)), \quad Z(t) = A(t) - D(t), \quad W(t) = CV(A(t)) - et + Y(t). \tag{3}$$

Here e is the J-dimensional vector of all ones and the k^{th} component of $\Phi(D(t))$ is to be read as $\sum_{l=1}^{K} \Phi_k^l(D_l(t))$ and the k^{th} component of $V(A(t))$ is to be read as $V_k(A_k(t))$. The equation for A indicates that the $A_k(t)$ arrivals to class k up to time t consist of $E_k(t)$ exogenous arrivals plus $\sum_{l=1}^{K} \Phi_k^l(D_l(t))$ arrivals obtained by feedback of some of the departures that have occurred up to time t. The equation for the workload process W expresses the fact that $\sum_{k\in\mathcal{K}} C_{jk}V_k(A_k(t))$ units of work have arrived for server j in $[0,t]$ and that this has been depleted by the amount of time $t - Y_j(t)$ that server j has been active in $[0,t]$. The fact that an HL discipline is non-idling implies that $\int_0^\infty 1_{(0,\infty)}(W_j(s))dY_j(s) = 0$ for all j.

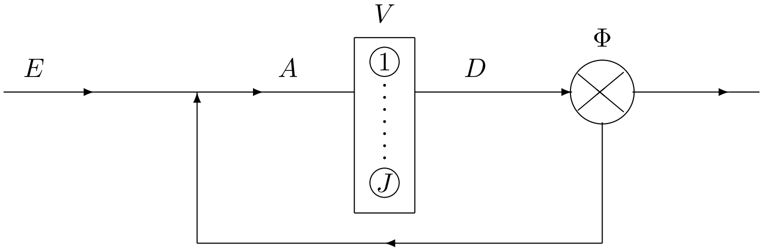

Figure 1: Schematic for an open multiclass queueing network

Note that the equations (3) do not give a complete description of the behavior of the queueing network. In particular one must add additional equations to provide information about the service discipline. For example, for the FIFO discipline one can add the relations: $D_k(t + W_j(t)) = A_k(t)$, for each class k and associated server j. Equations for other HL service disciplines will not be given here, since for the statement of the main theorem (Theorem 4.1), only a distillation of the service discipline is needed in the form of a $K \times J$ matrix Δ. Since this matrix is related to the heavy traffic behavior of networks, discussion of it is deferred to the next section.

Heavy Traffic The following notation is used in describing the notion of heavy traffic. Let α denote the K-dimensional long run average arrival rate vector for the exogenous arrival process E and let M denote the $K \times K$ diagonal matrix whose diagonal entries are the mean service times m_k for the classes $k \in \mathcal{K}$. Let λ be the unique solution of the "traffic flow" equation $\lambda = \alpha + P'\lambda$, i.e., $\lambda = (I - P')^{-1}\alpha$. Here $'$ denotes transpose. (To avoid degeneracies, it is assumed that $\lambda_k > 0$ for each k.) Define $\rho = CM\lambda$. The quantity λ_k is called the arrival rate for class k and ρ_j is called the traffic intensity parameter for station j. These are nominally the long run average rate at which jobs arrive to class k and the long run fraction of time that server j is busy, respectively. For single class networks, these nominal quantities represent actual long run quantities (provided $\rho_j \leq 1$ for all j). However, since the appearance of counterexamples in the early 1990s (see e.g., [LK91, RS91]), it has been known that this interpretation is not always valid for multiclass networks. Indeed, the question of whether these nominal quantities actually correspond to long run quantities is related to the stability properties of the queueing network. Rather than digressing to discuss this further here, the reader is referred to the articles on stability in [KW95], the references therein, and the article [Br98]. Here λ and ρ are simply regarded as useful parameters. Networks that are (nominally) heavily loaded or in heavy traffic are those in which ρ_j is close to one for each j. Such networks are the focus of attention in the next section.

4 Sufficient Conditions for a Heavy Traffic Limit Theorem

A Sequence of Networks Mathematically, to justify the approximation of a given heavily loaded open multiclass queueing network by a SRBM, we regard the network as being a member of a sequence of networks in which the traffic intensity vector ρ converges to e. Here, to simplify the exposition, the sequence is chosen so that only the distributions of the service times vary along the sequence where this variation is parametrized by the mean service times. (A more complex setup can be considered, allowing for more general variation of the distributions of all of the stochastic primitives along the sequence [Wi97b]. Although this implies a certain robustness of the approximation to small perturbations in the distributions of the stochastic primitives, for the purpose of stating a limit theorem that justifies the approximation of a fixed heavily loaded network, only the simpler setup described here is needed.)

Thus, we consider a sequence of networks indexed by r, which tends to infinity through a strictly increasing sequence of positive numbers. Each network in the sequence has the same basic structure as described in the previous section. Furthermore, J, K, C, E, Φ and the service discipline do not vary with r, and $v_k(i) = m_k^r u_k(i)$ where m_k^r is the mean service time for class k in the r^{th} network and $u_k(i)$ is a random variable independent of r that has mean one and finite variance. (To avoid degeneracies, it is assumed that $u_k(i)$ has positive variance for each class k. This assumption implies that the covariance matrix for the proposed SRBM approximant is non-degenerate. For other ways in which this can be achieved, see §5 of [Wi97b].) In the sequel, the superscript r is attached to all quantities that may depend on r.

Now assume the following heavy traffic conditions: as $r \to \infty$, $m_k^r \to m_k \in (0,\infty)$ for each $k \in \mathcal{K}$, such that $\gamma^r \equiv r(\rho^r - e) \to \gamma \in \mathbb{R}^J$. Define the diffusion scaled workload, cumulative idletime and queue length processes:

$$\hat{W}^r(t) = W^r(r^2t)/r, \qquad \hat{Y}^r(t) = Y^r(r^2t)/r, \qquad \hat{Z}^r(t) = Z^r(r^2t)/r. \tag{4}$$

The purpose of a heavy traffic limit theorem is to justify approximating $(\hat{W}^r, \hat{Y}^r, \hat{Z}^r)$ in distribution using a SRBM.

State Space Collapse A key feature of prior limit theorems in the multiclass setting [Wh71, Pe91, Re88] has been a phenomenon called state space collapse, which states that the diffusion scaled queue length process for each class k can be approximately recovered as a multiple of the associated station's diffusion scaled workload process. Here a slightly weaker notion called multiplicative state space collapse is used. This form suffices for our purposes and seems more amenable to verification (cf. [Br97b]). Here $\|f(\cdot)\|_T = \sup_{0\le t\le T} |f(t)|$ for any vector valued function f defined on $[0,T]$. (The notion of state space collapse is defined by omitting the denominator in (5) below.)

Definition 4.1 *Multiplicative state space collapse holds if there is a $K \times J$ matrix Δ such that for each $T \ge 0$,*

$$\frac{\|\hat{Z}^r(\cdot) - \Delta\hat{W}^r(\cdot)\|_T}{\|\hat{W}^r(\cdot)\|_T \vee 1} \to 0 \quad \textit{in probability as } r \to \infty, \tag{5}$$

where $a \vee b \equiv \max(a,b)$ for any two real numbers a, b.

Based on extant limit theorems, some conjectured forms of Δ for various service disciplines are described in [Wi97b]. In fact, one can show (see Appendix B in [Wi97b]) that a necessary condition for $\{(\hat{W}^r, \hat{Z}^r)\}$ to be C-tight under the FIFO service discipline is that (multiplicative) state space collapse holds with $\Delta = \Lambda C'$ where Λ is the $K \times K$ diagonal matrix with the entries of λ on its diagonal.

Sufficient Conditions for a Heavy Traffic Limit Theorem The main content of the following theorem is that for a sequence of open multiclass queueing networks as described above (with a general HL service discipline), multiplicative state space collapse plus the natural condition that the reflection matrix R for the purported SRBM approximant is well defined and completely-$\mathcal{S}$, is sufficient for a heavy traffic limit theorem to hold. Here $\Rightarrow$ denotes convergence in distribution of processes taking values in the space of paths that are right continuous with finite left limits, where this space is endowed with the usual Skorokhod topology.

Theorem 4.1 [Wi97b] *Suppose that multiplicative state space collapse holds and that the inverse matrix $R = (CM(I-P')^{-1}\Delta)^{-1}$ exists and is completely-$\mathcal{S}$. Then*

$$(\hat{W}^r, \hat{Y}^r, \hat{Z}^r) \Rightarrow (W^*, Y^*, Z^*) \quad \text{as } r \to \infty, \tag{6}$$

where W^* is a SRBM with data $(R\gamma, \Gamma, R, \delta_0)$ and associated pushing process Y^*, and $Z^* = \Delta W^*$. The covariance matrix Γ is a known quantity determined from C and the means and covariances of the stochastic primitives [Wi97b], and δ_0 denotes the unit mass at the origin in $\mathbb{R}^J_+$.

The proof of this theorem proceeds by showing tightness of the sequence $\{(\hat{W}^r, \hat{Y}^r, \hat{Z}^r)\}$ and uniqueness in law of any weak limit point. For the tightness, multiplicative state space collapse is combined with the oscillation inequality of Theorem 2.2. For the uniqueness of any weak limit point $(W^\dagger, Y^\dagger, Z^\dagger)$, one needs to show that $W^\dagger$ is a SRBM with associated pushing process $Y^\dagger$. In particular, the martingale property in the definition of a SRBM needs to be verified for $X^\dagger = W^\dagger - RY^\dagger$. This involves establishing a multiparameter stopping time property which is where the precise definition of a HL service discipline (including its measurable dependence on the "state") comes into play.

New Heavy Traffic Limit Theorems In a companion work to [Wi97b], Bramson [Br97b] (see also [Br98]) has given sufficient conditions for multiplicative state space collapse to hold. These conditions are in terms of the behavior of a balanced fluid model (a law of large numbers approximation for the sequence of heavily loaded queueing networks). In particular, using these conditions and his prior work on the fluid model behavior for FIFO Kelly type and HLPPS networks, Bramson [Br97b] has shown that multiplicative state space collapse holds for these two collections of networks. The qualifier "Kelly type" means that m_k depends only on the station j at which class k is served, i.e., the limiting mean service times are station-dependent, not class-dependent, quantities. In addition, it is known [DH93, Wi97b] that R is well defined and completely-$\mathcal{S}$ for these networks.

Combining the above results yields new heavy traffic limit theorems for these two collections of networks. In particular, the FIFO Kelly type network introduced by Dai, Wang and Wang (see Appendix A in [Wi97b]) can be approximated by a SRBM. This is particularly interesting since the continuous mapping (strong solution) approach used in most prior limit theorems cannot be applied to that example.

In independent work, Chen and Zhang [CZ97] have established a heavy traffic limit theorem for FIFO networks in which $G = CM(I - P')^{-1}P'\Lambda C'$ has spectral radius less than one. Although they do not use Theorem 4.1, they implicitly verify the conditions of that theorem for their case and avoid a continuous mapping argument in a similar manner to that in [Wi97b].

5 Open Problems

The results in [Br97b, Wi97b] reduce the problem of establishing heavy traffic limit theorems for open multiclass queueing networks with a HL service discipline to that of establishing multiplicative state space collapse through the study of balanced fluid models over long intervals of time and to verifying that the reflection matrix R is well defined and completely-$\mathcal{S}$. A compelling open problem is to identify new collections of networks that satisfy these conditions. In particular, it is natural to consider networks with static priority service disciplines (see the article [Br98] by Bramson for recent work in this direction). Another area for future investigation is heavy traffic behavior of networks with non-HL disciplines such as last-in-first-out and general processor sharing. Finally, the focus here has been on performance analysis for heavily loaded networks with a fixed structure. In some applications one may be able to vary such quantities as the service discipline or routing in a dynamic manner with the objective of optimizing some measure of performance. Again such problems frequently cannot be analyzed exactly and one may seek approximate models. An approach using approximate diffusion models has been advocated by some authors (see e.g., [HW89, KL93, Ku95]), but many open problems remain concerning justification and interpretation of such approximations in general.

References

[BEK91] Bernard, A., and El Kharroubi, A. (1991). Régulation de processus dans le premier orthant de $\mathbb{R}^n$. *Stochastics*, 34, 149–167.

[BG92] Bertsekas, D., and Gallagher, R. (1992). *Data Networks.* Prentice-Hall.

[Br97a] Bramson, M. (1997). Stability of two families of queueing networks and a discussion of fluid limits. To appear in *Queueing Syst.*

[Br97b] Bramson, M. (1997). State space collapse with application to heavy traffic limits for multiclass queueing networks. To appear in *Queueing Syst.*

[Br98] Bramson, M. (1998). State space collapse for queueing networks. *Proceedings of the International Congress of Mathematicians, Berlin, 1998,* this issue.

[CZ97] Chen, H., and Zhang, H. (1997). Diffusion approximations for some multiclass queueing networks with FIFO service disciplines. Preprint.

[DD97] Dai, J. G., and Dai, W. (1997). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. Preprint.

[DH93] Dai, J. G., and Harrison, J. M. (1993). The QNET method for two-moment analysis of closed manufacturing systems. *Ann. Appl. Prob.*, 3, 968–1012.

[DWa93] Dai, J. G., and Wang, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Syst.*, 13, 41–46.

[DWi95] Dai, J. G. and Williams, R. J. (1995). Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Theory Probab. Appl.*, 40, 1–40.

[DuI91] Dupuis, P., and Ishii, H. (1991). On the Lipschitz continuity of the solution mapping to the Skorokhod problem. *Stochastics,* 35, 31–62.

[DuR98a] Dupuis, P., and Ramanan, K. (1998). Convex duality and the Skorokhod problem, I & II. Submitted to *Prob. Theor. Rel. Fields.*

[DuR98b] Dupuis, P., and Ramanan, K. (1998). A Skorokhod problem formulation and large deviation analysis of a processor sharing model. To appear in *Queueing Syst.*

[FL98] Flandreau, M. (1998). The burden of intervention: externalities in multilateral exchange rate arrangements. To appear in *J. Intern. Econ.*

[Ha88] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, W. Fleming and P.-L. Lions (eds.), Springer-Verlag, 147–186.

[HN93] Harrison, J. M., and Nguyen, V. (1993). Brownian models of multiclass queueing networks: current status and open problems. *Queueing Syst.*, 13, 5–40.

[HR81] Harrison, J. M., and Reiman, M. I. (1981). Reflected Brownian motion on an orthant. *Ann. Probab.*, 9, 302–308.

[HW89] Harrison, J. M., and Wein, L. M. (1989). Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Syst.*, 5, 265–280.

[KL93] Kelly, F. P., and Laws, C. N. (1993). Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Syst.*, 13, 47–86.

[KW95] Kelly, F. P., and Williams, R. J. (eds.) (1995). *Stochastic Networks.* Vol. 71, Springer.

[Ku95] Kushner, H. J. (1995). A control problem for a new type of public transportation system, via heavy traffic analysis. In [KW95], 139–167.

[LK91] Lu, S. H., and Kumar, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Autom. Control*, 36, 1406–1416.

[Ma92] Mandelbaum, A. (1992). The dynamic complementarity problem. Preprint.

[Pe91] Peterson, W. P. (1991). Diffusion approximations for networks of queues with multiple customer types. *Math. Oper. Res.*, 9, 90–118.

[Re84] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* 9, 441–458.

[Re88] Reiman, M. I. (1988). A multiclass feedback queue in heavy traffic. *Adv. Appl. Prob.*, 20, 179–207.

[RW88] Reiman, M. I., and Williams, R. J. (1988–89). A boundary property of semimartingale reflecting Brownian motions. *Probab. Theory Relat. Fields,* 77, 87–97, and 80, 633.

[RS91] Rybko, A. N., and Stolyar, A. L. (1991). Ergodicity of stochastic processes describing the operation of an open queueing network. *Problemy Peredachi Informatsii*, 28, 2–26.

[SV71] Stroock, D. W., and Varadhan, S. R. S. (1971). Diffusion processes with boundary conditions. *Comm. Pure Appl. Math.*, 24, 147–225.

[TW93] Taylor, L. M., and Williams, R. J. (1993). Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probab. Theory Relat. Fields*, 96, 283–317.

[VW85] Varadhan, S. R. S., and Williams, R. J. (1985). Brownian motion in a wedge with oblique reflection. *Comm. Pure Appl. Math.*, 38, 405–443.

[Wh71] Whitt, W. (1971). Weak convergence theorems for priority queues: preemptive resume discipline. *J. Appl. Prob.*, 8, 74–94.

[Wi95] Williams, R. J. (1995). Semimartingale reflecting Brownian motions in the orthant. In [KW95], pp. 125–137.

[Wi96] Williams, R. J. (1996). On the approximation of queueing networks in heavy traffic. In *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins (eds.), Oxford University Press, Oxford, pp. 35–56.

[Wi97a] Williams, R. J. (1997). An invariance principle for semimartingale reflecting Brownian motions in an orthant. To appear in *Queueing Syst.*

[Wi97b] Williams, R. J. (1997). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. To appear in *Queueing Syst.*

[Ya94] Yao, D. D. (ed.) (1994). *Stochastic Modeling and Analysis of Manufacturing Systems.* Springer-Verlag.

Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla CA 92093-0112 USA