State Space Collapse for Queueing Networks

Maury Bramson*

Abstract. The diffusive limits of queueing networks, known as heavy traffic limits, are a topic of continuing interest. An important ingredient in such work is the demonstration of state space collapse, which says that, in the limit, the process must live on an appropriate subspace. In [Wi98b], conditions are given under which state space collapse suffices for heavy traffic limits. Here, we discuss how state space collapse can be reduced to the problem of showing stability for the fluid model which is the deterministic analog of the queueing networks under consideration. We discuss specific cases, such as first-in first-out (FIFO) networks of Kelly type and certain static priority networks.

1991 Mathematics Subject Classification: Primary 60K25.

1 Introduction

Queueing networks constitute a general family of stochastic processes. In such models, one envisions "customers", such as people, products or some task to be performed, as being lined up at the different queues, or *stations*, of a network. When service of a customer at a station is completed, the customer moves to another station or leaves the network. Customers are also assumed to enter the network at various stations. This behavior will, in general, be random, with random variables corresponding to the choice of the next station when service at a station is completed, to the service times at stations, and to the interarrival times for customers entering the network. The evolution of such a network can be formulated as a continuous time Markov process. Two basic topics for queueing networks concern (1) obtaining criteria for when this Markov process is positive recurrent and (2) deciding when a sequence of networks, under diffusive scaling, converges to a reflecting Brownian motion. The criteria, in the two cases, are related. In this survey, we discuss both topics, with emphasis on the latter.

In many situations, it is important to permit more than one type of behavior for customers at a given station. (For example, patients at the receptionist's desk of a doctor's office will follow different rules, depending on whether they are checking in or out.) To allow for this, one distinguishes between different *classes* or *buffers* at a station; customers in the same class are subject to the same random rules for service and routing to the next class. A queueing network is *single class* if only one class is assigned to each station; otherwise, it is *multiclass*. One can

*The author was supported in part by the National Science Foundation.

also classify a queueing network based on whether or not it allows *feedback*, that is, output from a station can eventually become part of its input. This will occur, for example, when customers repeatedly visit a station along some preassigned route. Not surprisingly, answers for (1) and (2) above will be easiest to obtain for single class networks without feedback, and most difficult for multiclass networks with feedback.

The limits in (2), which are referred to as heavy traffic limits (HTL), have been investigated over the past three decades. Presently, HTL theory remains incomplete for multiclass networks. An important concept in this context is state space collapse (SSC). When SSC holds, customers in the different classes at a station occur (asymptotically) in fixed proportions. Such behavior enables one to generalize HTL results from single class networks to multiclass networks. This is done in [Wi98a]. It is also shown there that SSC follows from a somewhat weaker concept, multiplicative state space collapse (MSSC). This work is summarized in the article [Wi98b] in this volume.

Here, we discuss certain settings where one can demonstrate MSSC. These include well-known families of networks, such as first-in first-out networks of Kelly type. More generally, sufficient conditions for MSSC are given by the convergence of the solutions of fluid model equations which are associated with the networks in question. Such criteria hold, for example, for static priority networks.

The remainder of this article is organized as follows. In Section 2, we summarize the basic notation and definitions for queueing networks. Section 3 discusses the stability of queueing networks. Fluid models, the main tool for demonstrating stability, are introduced here. Section 4 discusses heavy traffic limits. Emphasis there is placed on recent work, in [Br98, Wi98a], which employs state space collapse.

2 NOTATION AND DEFINITIONS

We make use here of concepts and notation employed in the article [Wi98b] in this volume, which the reader should consult for more detail. The variable j, $j = 1, \dots, J$, will denote the stations of the network under consideration, and k, $k = 1, \dots, K$, will denote the classes of the network. We use $\mathcal{C}(j)$ for the set of classes belonging to a station j, and $s(k)$ for the station to which class k belongs. At each station there is a single server. This server will always by *non-idling*, that is, the server will remain busy as long as there are customers present at its station.

The triple $(E(\cdot), V(\cdot), \Phi(\cdot))$ contains the random input of the network. The random vector $E(t) = \{E_k(t),\ k = 1, \dots, K\}$ denotes the number of external arrivals by time t, $t \geq 0$, and $V(\mathbf{n}) = \{V_k(n_k),\ k = 1, \dots, K\}$, $\mathbf{n} = (n_1, \dots, n_K)$, denotes the cumulative service times for the first n_k customers in each class. The random matrix $\Phi(\mathbf{n})$, with rows $\Phi^k(n_k)$, $k = 1, \dots, K$, denotes the cumulative routing process after n_k departures from each class k. As in [Wi98b], summands of these quantities are assumed, in each case, to be independent and identically distributed, with the different sequences also being independent of one another. The triple (α, M, P) is the deterministic analog of $(E(\cdot), V(\cdot), \Phi(\cdot))$. The mean vector α gives the external arrival rates at the different classes; the $K \times K$ diagonal matrix M has the mean service times m_k as its diagonal entries. The matrix $P = \{P_{k\ell},\ k,\ell = 1,\dots,K\}$ gives the probability of a customer being routed from one class to another. In many interesting cases, the routing of the queueing network will be deterministic, with all customers entering the system at the same class, and moving along a given route, until they exit from the system. Such networks are referred to as *re-entrant lines.*

We will consider here only open networks, that is, networks for which the matrix

$$Q \stackrel{\text{def.}}{=} (I - P')^{-1} = I + P' + (P')^2 + \dots \tag{2.1}$$

is finite. (" $'$ " denotes the transpose.) This means that customers at any class are capable of ultimately leaving the network. To investigate these networks, one employs the solutions $\lambda_\ell,\ \ell = 1,\dots,K$, of the *traffic equations*

$$\lambda_\ell = \alpha_\ell + \sum_{k=1}^{K} \lambda_k P_{k\ell}, \tag{2.2}$$

or equivalently, in vector form, of $\lambda = \alpha + P'\lambda$. (All vectors in this article are to be interpreted as column vectors.) Solving (2.2), one obtains $\lambda = Q\alpha$. The term λ_k is the *nominal arrival rate* for class k; to avoid degeneracies, we assume that $\lambda_k > 0$ for all k. Employing m and λ, one defines the *traffic intensity* ρ_j for the jth server as

$$\rho_j = \sum_{k\in\mathcal{C}(j)} m_k\lambda_k, \tag{2.3}$$

with ρ being the corresponding vector. The condition $\rho_j < 1,\ j = 1,\dots,J$, is required for each station, when nonempty, to serve customers, in the long run, more rapidly than they enter the station. When this holds, the network is *strictly subcritical.* When $\rho_j = 1$ for each j, the network is referred to as being *critical* or *balanced.*

Associated with each queueing network is a *discipline*, which specifies the order in which customers receive service. We consider here only *head-of-the-line* (HL) disciplines, where only the first customer in each class may receive service at a given time. For multiclass networks, the proportion of service to be devoted to each class needs to be specified. Examples of disciplines which we will discuss are first-in first-out (FIFO), where the first customer at a station receives all of the service irrespective of its class; head-of-the-line proportional processor sharing (HLPPS), where the amount of service allocated to the first customer in each class is proportional to the number of customers in that class, and static priority disciplines, where classes are assigned a strict ranking, and customers of higher ranked classes are always served first. In the setting of re-entrant lines, examples of static priority disciplines are first-buffer-first-served (FBFS) and last-buffer-first-served (LBFS), where customers at the earlier, respectively latter, classes have priority. When the queueing network is single class, and the service and interarrival times are exponentially distributed, it is referred to as a Jackson network. When the restriction on the service and interarrival times is removed, it is called a generalized Jackson network.

Once a discipline has been given, the triple $(E(\cdot), V(\cdot), \Phi(\cdot))$ and the initial data uniquely specify the evolution of a queueing network along each realization. This defines an underlying Markov process. When this process is positive recurrent, the queueing network is said to be *stable*. Depending on the discipline, the description of the state space can be a bit of a notational burden. We avoid such details here.

3 STABILITY AND FLUID MODELS

A necessary condition for a queueing network to be stable is that it be strictly subcritical. For a long while, it was generally believed that the condition is also sufficient. This is now known to be false ([Br94], [LuKu91], [RySt92] and [Se94]). It is possible for the flow of customers through a network to synchronize so that, at a given time, customers are clustered at specific parts of the network. This permits individual stations to be periodically "starved" for work, which reduces their long-term efficiency. At the end of each additional cycle, the number of customers in the network will then be, on the average, a multiple of the number for the previous cycle, which produces geometric growth (as measured in cycles).

For many disciplines, however, a queueing network is stable whenever it is strictly subcritical. Fluid models are the main tool for showing this. They allow one, in essence, to replace a queueing network with its continuous deterministic analog of mass flowing through the system. It is typically a considerably easier problem to show stability in this deterministic setting. Under mild conditions on the service and interarrival distributions, the stability of the original queueing network will then follow.

The basic idea is to describe the evolution of a queueing network by a set of equations. One then analyzes the solutions of the corresponding set of deterministic equations, where random quantities have been replaced by their means. One needs to show that the "queue length" vector for such solutions is 0 after a fixed time. It then follows that the queueing network is stable.

In order to describe the evolution of a queueing network, one employs random vectors such as $A(t)$, $D(t)$, $W(t)$, $Y(t)$ and $Z(t)$. The vector $A(t)$ denotes the number of arrivals by time t, $D(t)$ denotes the number of departures, and $Z(t)$ is the number of customers at time t. These three quantities are all class vectors, with components corresponding to the individual classes. The vectors $W(t)$ and $Y(t)$ are both station vectors, with $W(t)$ being the immediate workload (the future time required to serve customers currently at each station), and $Y(t)$ is the cumulative idletime. Typically, the choice of exactly which quantities one employs depends on the particular setting. We will denote the corresponding n-tuple by $\mathfrak{X}(t)$; in the above setting,

$$\mathfrak{X}(t) = (A(t), D(t), W(t), Y(t), Z(t)). \tag{3.1}$$

One connects these quantities together by *queueing network equations*, which include

$$A(t) = E(t) + \sum_k \Phi^k(D_k(t)), \tag{3.2}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{3.3}$$
$$W(t) = CV(A(t) + Z(0)) - et + Y(t) \tag{3.4}$$
$$\int_0^\infty 1_{(0,\infty)}(W_j(s))dY_j(s) = 0, \quad j = 1, \dots, J, \tag{3.5}$$

for $t \geq 0$. Here, e is the J-vector of all 1's, and C is the $J \times K$ matrix with $C_{jk} = 1$ for $k \in \mathcal{C}(j)$, and $C_{jk} = 0$ otherwise. An additional equation or two is required for the discipline of the network. For instance, for the FIFO discipline, one employs

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t), \quad k = 1, \dots, K, \tag{3.6}$$

for $t \geq 0$.

For our purposes, the exact nature of the equations (3.2)–(3.6) is not too important. One should think of there as being enough equations to determine the evolution of the queueing network. These equations are used in conjunction with their deterministic analogs, known as *fluid model equations*, which are obtained by replacing $(E(\cdot), V(\cdot), \Phi(\cdot))$ by (α, M, P). The analogs of (3.2)–(3.6) are then given by

$$\bar{A}(t) = \alpha t + P'\bar{D}(t), \tag{3.7}$$
$$\bar{Z}(t) = \bar{Z}(0) + \bar{A}(t) - \bar{D}(t), \tag{3.8}$$
$$\bar{W}(t) = CM(\bar{A}(t) + \bar{Z}(0)) - et + \bar{Y}(t), \tag{3.9}$$
$$\int_0^\infty 1_{(0,\infty)}(\bar{W}_j(s))d\bar{Y}_j(s) = 0, \quad j = 1, \dots, J, \tag{3.10}$$
$$\bar{D}_k(t + \bar{W}_j(t)) = \bar{Z}_k(0) + \bar{A}_k(t), \quad k = 1, \dots, K, \tag{3.11}$$

for $t \geq 0$. (To distinguish the solutions of the fluid model equations, we employ overbar notation for the variables in this context.) We also write $\bar{\mathfrak{X}}(t)$ for the analog of (3.1). Such solutions are referred to as fluid model solutions. We restrict our attention to solutions with continuous and nonnegative components, where $\bar{A}(t)$, $\bar{D}(t)$ and $\bar{Y}(t)$ are nondecreasing.

The solutions of the equations (3.2)–(3.6) and (3.7)–(3.11) are connected via the *fluid limits* of $\mathfrak{X}(t)$. These are the limits obtained by applying hydrodynamic scaling to $\mathfrak{X}(t)$, i.e., by scaling the weight of individual customers and time proportionately. (We avoid the technical details here.) Fluid limits are solutions of the fluid model equations; solution of the latter will give information about the original queueing network. The fluid model is said to be *stable* if, for a given $\delta > 0$ and all solutions of the fluid model equations, $\bar{Z}(t) = 0$ for $t \geq \delta|\bar{Z}(0)|$. ($|\cdot|$ denotes the sum of the coordinates.) Since the solutions of a fluid model correspond to a queueing network with the randomness removed, stability of the fluid model says that, in essence, the total number of customers in the queueing network has a net negative drift.

Using elementary properties of Markov processes on general state spaces, it is shown in [Da95] that, under mild assumptions on the service and interarrival times, a queueing network is stable whenever the corresponding fluid model is stable. (Versions of these ideas were first employed in [RySt92].) This enables one to indirectly study a queueing network by means of the corresponding fluid model equations. In particular, the distributions of the service and interarrival times do not occur in this setting. This enables one, for example, to simply demonstrate the stability of strictly subcritical generalized Jackson networks, whereas a direct argument is quite tedious. The stability of strictly subcritical FIFO networks of Kelly type is another application. (The latter condition means that $m_k = m_\ell$ whenever $s(k) = s(\ell)$.) In general, strictly subcritical FIFO networks which are not of Kelly type need not be stable.

4 HEAVY TRAFFIC LIMITS

Some background

In the introduction, we briefly discussed heavy traffic limits. Here, we go into more detail. The basic setup for HTLs consists of a sequence of queueing networks, with the accompanying n-tuples $\mathfrak{X}^r(t)$ and queueing network equations. One scales the quantities $W^r(t)$ and $Z^r(t)$, setting $\hat{W}^r(t) = W^r(r^2 t)/r$ and $\hat{Z}^r(t) = Z^r(r^2 t)/r$. The goal is to show that

$$\hat{W}^r(\cdot) \Rightarrow W^*(\cdot) \quad \text{as } r \to \infty, \tag{4.1}$$

where $W^*(\cdot)$ is a semimartingale reflecting Brownian motion (SRBM). The functions $\hat{W}^r(\cdot)$ take values in the space of J-dimensional right continuous functions with left limits, which is equipped with the usual Skorokhod topology, and "$\Rightarrow$" denotes weak convergence.

SRBMs and related concepts are defined in [Wi98b]. Intuitively, the SRBM $W^*(\cdot)$ behaves like a Brownian motion in the interior of the orthant $\mathbb{R}^J_+$; its drift and its covariance matrix are given by appropriate limits of the first two moments of the summands of the triples $(E^r(\cdot), V^r(\cdot), \Phi^r(\cdot))$, and by the discipline of the networks. It is confined to $\mathbb{R}^J_+$ by pushing on the boundary in the directions given by a reflection matrix R (also determined by the above quantities), according to the local time spent there. In order for such a process $W^*(\cdot)$ to exist, R needs to be completely-$\mathcal{S}$.

HTLs have been investigated over the past three decades; a summary of the subject is given in [Wi96, Wi98b]. Implicit in the formulation of (4.1) is the assumption that the states of the corresponding networks are, for large r, essentially given by $\hat{W}^r(t)$ at time t. More detailed information about the system, such as $\hat{Z}^r(t)$, should not be necessary to study the evolution of the limit $W^*(t)$. This type of behavior is known as state space collapse. (The term was used in [Re84a]; related ideas go back to [Wh71].) For our purposes, the relevant variant is *multiplicative state space collapse*, that is

$$\frac{\|\hat{Z}^r(\cdot) - \Delta \hat{W}^r(\cdot)\|_T}{\max(\|\hat{W}^r(\cdot)\|_T, 1)} \to 0 \quad \text{in probability} \tag{4.2}$$

as $r \to \infty$. Here, Δ is an appropriate linear map from $\mathbb{R}^J$ to $\mathbb{R}^K$, which depends on the service discipline; $\|\cdot\|_T$ is the uniform norm over $[0, T]$.

HTLs as in (4.1) need not exist, even for standard disciplines such as FIFO. It was shown in [DaNg94, DaWa93, Wh93] that this is the case for certain sequences of FIFO networks; the problem is related to the limiting reflection matrix R not being completely-$\mathcal{S}$. Another potential problem is the lack of MSSC. These problems need to be faced when dealing with multiclass networks with feedback. (When a network is single class, these problems do not arise, and HTLs exist ([Re84b])). This is also the case when the network is *feedforward*, that is, an ordering among the stations is possible so that customers at lower ranked stations always go to higher numbered stations.) The general theory for multiclass networks is presently incomplete. Below, we summarize some recent work on the subject which uses MSSC and the fluid model equations introduced in Section 3.

Reduction to fluid model equations

In [Wi98a, Br98], HTLs are demonstrated for certain families of multiclass networks. The reasoning employed there can be broken into three "modules", which are essentially independent. The first module, which is worked out in [Wi98a], uses MSSC and the completely-$\mathcal{S}$ condition to derive HTLs. Solutions of the balanced fluid model equations corresponding to the limiting triple (α, M, P), obtained from (α^r, M^r, P^r), are employed in the second module. It is shown in [Br98], that MSSC holds whenever such solutions have "nice" asymptotic behavior. The third module consists of deriving the desired asymptotics for these solutions, and verifying that R is completely-$\mathcal{S}$. Both of these conditions, in the last step, are not trivial in general. They are, though, substantial reductions from MSSC. In this subsection, we discuss the appropriate framework for the second module. We also mention some specific disciplines where the conditions in the third module can be verified.

In order to state our results for MSSC, we need to overcome some technical difficulties. The specific discipline must be known in order to be able to write down all of the relevant queueing network or fluid model equations, such as (3.6). If one wishes to state results on MSSC at the general level of HL processes, it is more convenient to instead work with *cluster points*. These are, in the setting of MSSC, the analog of the fluid limits, which were mentioned briefly in Section 3. Rather than complicate matters, we restrict ourselves here to several more concrete families where we can work directly with the corresponding fluid model equations. Also, as in [Wi98b], we assume that $Z^r(0) = 0$ for the sequences of queueing networks under consideration, in order to simplify formulation of the results.

Associated with a sequence of queueing networks are the triples $(E^r(\cdot), V^r(\cdot), \Phi^r(\cdot))$. We assume here that the corresponding means (α^r, M^r, P^r) satisfy

$$\alpha^r \to \alpha, \quad M^r \to M, \quad P^r \to P \quad \text{as } r \to \infty, \tag{4.3}$$

and that the limit (α, M, P) is balanced. One also needs a uniformity condition on the second moments of the service and interarrival distributions for the sequence. The latter conditions can be ensured, for example, by not allowing $E^r(\cdot)$ or $\Phi^r(\cdot)$ to vary with r, and only allowing the components of $V^r(\cdot)$ to vary by scalar multiples, as is done in [Wi98b]. In order to obtain HTLs from MSSC, as in [Wi98a, Wi98b], one will need to strengthen (4.3) so that $r(\rho^r - e) \to \gamma$ as $r \to \infty$, for some γ, also holds, although this is not needed for MSSC itself. ($R\gamma$ will be the drift of the HTL.)

We first consider a sequence of queueing networks, with a fixed static priority discipline. As mentioned above, we assume that $Z^r(0) = 0$ for all r. We also assume that (4.3) holds, that (α, M, P) is balanced, and that the second moment conditions referred to above hold. Let $\bar{Z}(t)$ denote the queue length for solutions of the corresponding fluid model equations for the specific discipline. We further assume that for all solutions with $|\bar{Z}(0)| \leq 1$,

$$|\bar{Z}(t) - \bar{Z}(\infty)| \leq H(t) \tag{4.4}$$

holds for a fixed function $H(t)$, with $H(t) \to 0$ as $t \to \infty$, and for appropriate $\bar{Z}(\infty)$ (depending on $\bar{Z}(0)$) of the form

$$\bar{Z}(\infty) = \Delta \bar{W} \quad \text{for some } \bar{W} \in \mathbb{R}^J. \tag{4.5}$$

It is shown in [Br98], that MSSC follows under these conditions. In [BrDa98], (4.4)–(4.5) are verified for several disciplines, such as FBFS and LBFS. Since one can also show that the R matrix is completely-$\mathcal{S}$ in both cases, the corresponding HTLs follow. (HTLs for FBFS networks are also shown in [ChZh96].)

One can also obtain HTLs for sequences of FIFO networks of Kelly type and HLPPS networks by investigating the corresponding fluid models. The basic procedure is the same as above. In each case, one can, in fact, demonstrate (4.4) with $H(t) = B_1 e^{-B_2 t}$, for appropriate B_1 and $B_2 > 0$. MSSC therefore follows. Since the R matrix will always be completely-$\mathcal{S}$ in both cases, (4.1) holds for appropriate $W^*(t)$. The arguments for showing (4.4) for the two models are related. One obtains an entropy function $\mathcal{H}(t)$ which converges exponentially fast to 0; the states with entropy 0 will satisfy (4.5). The function for FIFO fluid models of Kelly type is

$$\mathcal{H}(t) = \sum_k \int_t^{t+\bar{W}_j(t)} h_k(\bar{D}'_k(r))dr. \tag{4.6}$$

Its asymptotic behavior is analyzed in [Br96] by employing the equations (3.7)–(3.11).

So far, we have not identified the linear map Δ, which "lifts" $\mathbb{R}^J$ to $\mathbb{R}^K$. For the above disciplines, this is easy to do, since ΔW, for $W \in \mathbb{R}^J_+$, will be among the states that remain invariant under the evolution of the corresponding fluid model. Clearly, for static priority disciplines, $(\Delta W)_k = 0$ at all coordinates except where k is the lowest ranked class at its station $j = s(k)$, in which case $(\Delta W)_k = W_j/m_k$. For FIFO networks, $(\Delta W)_k = \lambda_k W_j$, where λ is as in (2.2), and for HLPPS networks,

$$(\Delta W)_k = \frac{\lambda_k m_k W_j}{\sum_{\ell \in \mathcal{C}(j)} \lambda_\ell m_\ell^2}. \tag{4.7}$$

One can see why, in principle, MSSC should follow from the limiting behavior of the fluid model solutions, as in (4.4)–(4.5), by comparing the evolution of the queue length vector $Z(t)$ under hydrodynamic scaling with its behavior under diffusive scaling. (Some poetic license is taken in phrasing the following steps.) Fluid limits, which are solutions of the fluid model equations, arise from hydrodynamic scaling. So, for large t, the components $\tilde{Z}_k^r(t)$ of $\tilde{Z}^r(t) \stackrel{\text{def.}}{=} Z^r(rt)/r$, as $r \to \infty$, will be in the proportions prescribed by Δ. Recalling that $\hat{Z}^r(t) = Z^r(r^2 t)/r$, this implies that $\hat{Z}^r(T_r) = \tilde{Z}^r(rT_r)$, as $r \to \infty$, collapses to the subspace given by Δ, if T_r is chosen so that $rT_r \to \infty$ sufficiently slowly as $r \to \infty$. (One needs the growth of rT_r to be slow enough to avoid the contribution of noise from random fluctuations of $Z^r(r^2T_r)$.) One is, moreover, entitled to restart the processes $\tilde{Z}^r(t)$ at times $i = 1, 2, \ldots$, with

$$\tilde{Z}^{r,i}(t) \stackrel{\text{def.}}{=} Z^r(r(t+i))/r. \tag{4.8}$$

Chopping up the interval $[0, r^2T]$, $T > 0$, from the original time scale into rT pieces, it suffices to analyze the fluid limits corresponding to each of these processes in order to demonstrate MSSC. Under the second moment conditions on the service and interarrival distributions that have already been made, the exceptional events where any of these processes is ill behaved, and the desired collapse does not occur, will have small probability for large r. Also, the assumption $\hat{Z}^r(0) = 0$ ensures that $\hat{Z}^r(t)$ remains close to 0 at small times. Therefore, for a typical realization, $\hat{Z}^r(t)$ collapses to the desired subspace for all $t \in [0, T]$. This reasoning (when carefully carried out) will demonstrate MSSC.

References

[Br94] Bramson, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.*, 4, 414–431.

[Br96] Bramson, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems*, 22, 5–45.

[Br98] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, to appear.

[BrDa98] Bramson, M. and Dai, J. (1998). Heavy traffic limits for some queueing networks. In preparation.

[ChZh96] Chen, H. and Zhang, H. (1996). Diffusion approximations for re-entrant lines with a first-buffer-first-served priority discipline. *Queueing Systems*, 23, 177–195.

[Da95] Dai, J. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid models. *Ann. Appl. Probab.*, 5, 49–77.

[DaNg94] Dai, J. and Nguyen, V. (1994). On the convergence of multiclass queueing networks in heavy traffic. *Ann. Appl. Probab.*, 4, 26–42.

[DaWa93] Dai, J. and Wang, Y. (1993). Nonexistence of Brownian models for certain multiclass queueing networks. *Queueing Systems*, 13, 41–46.

[LuKu91] Lu, S.H. and Kumar, P.R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Autom. Control*, 36, 1406–1416.

[Re84a] Reiman, M.I. (1984). Some diffusion approximations with state space collapse. *Proceedings International Seminar on Modeling and Performance Evaluation Methodology*, Lecture Notes in Control and Informational Sciences, F. Baccelli and G. Fayolle (eds.), Springer, New York, 209–240.

[Re84b] Reiman, M.I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9, 441–458.

[RySt92] Rybko, S. and Stolyar, A. (1992). Ergodicity of stochastic processes that describe the functioning of open queueing networks. *Problems Inform. Trans.*, 28, 3–26 (in Russian).

[Se94] Seidman, T.I. (1994). "First come, first served" can be unstable! *IEEE Trans. Automat. Control*, 39, 2166–2171.

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

[Wh93] Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Science*, 39, 1020–1028.

[Wi96] Williams, R.J. (1996). On the approximation of queueing networks in heavy traffic. *Stochastic Networks, Theory and Applications*, Royal Statistical Society Lecture Note Series, F.P. Kelly, S. Zachary, I. Ziedlins (eds.), Clarendon Press, Oxford, 35–56.

[Wi98a] Williams, R.J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, to appear.

[Wi98b] Williams, R.J. (1998). Reflecting diffusions and queueing networks. *Proceedings of the International Congress of Mathematicians*, this issue.

Maury Bramson
School of Mathematics
University of Minnesota
Minneapolis, MN 55455