

ON STOCHASTIC PROCESSING NETWORKS*

RUTH J. WILLIAMS †

November 7, 2010

*Based on a short course given at the University of Melbourne, September 4-6, 2006, 4.15pm-6.15pm daily. Jointly sponsored by the Belz fund and MASCOS. Copyright R. J. Williams, 2006.

†Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92093-0112, USA, email: williams@math.ucsd.edu

Abstract

Stochastic processing networks arise as models in manufacturing, telecommunications, computer systems and the service industry. Common characteristics of these networks are that they have entities, such as jobs, customers or packets, that move along routes, wait in buffers, receive processing from various resources, and are subject to the effects of stochastic variability through such quantities as arrival times, processing times, and routing protocols. Networks arising in modern applications are often highly complex and heterogeneous. Typically, their analysis and control present challenging mathematical problems. One approach to these challenges is to consider approximate models.

In the last 15 years, significant progress has been made on using approximate models to understand the stability and performance of a class of stochastic processing networks called open multiclass HL queueing networks. Here HL stands for a non-idling service discipline that is head-of-the-line, i.e., jobs are drawn from a buffer in the order in which they arrived. Examples of such disciplines are FIFO (first-in-first-out) and static priorities. First order (functional laws of large numbers) approximations called fluid models have been used to study the stability of these networks, and second order (functional central limit theorem) approximations which are diffusion models, have been used to analyze the performance of heavily congested networks. The interplay between these two levels of approximation has been an important theme in this work.

In contrast to this progress, optimal control of multiclass HL queueing networks is an active area of research, and performance analysis and optimal control of more general stochastic processing networks are still in their early stages of development.

These lectures will progress from motivating the study of stochastic processing networks, through describing some significant developments of the last 15 years, and will end with some current research topics.

Contents

1	Stochastic Processing Networks: What, Why and How?	5
1.1	Stochastic Processing Network Structure	5
1.2	Questions and Approaches	6
1.3	Two Illustrative Examples	7
1.4	Exercises Related to Introductory Lecture	8
2	Open Multiclass HL Queueing Networks	11
2.1	Open Multiclass HL Queueing Network Structure	11
2.2	Stability	13
2.2.1	Markov Model	14
2.2.2	Fluid Model	15
2.2.3	Sufficient Conditions for Stability via Fluid Models	16
2.2.4	Stability and Equilibria for Some Fluid Models	16
2.2.5	Exercises Related to Fluid Models	16
2.3	Performance in Heavy Traffic	20
2.3.1	Semimartingale Reflecting Brownian Motions	20
2.3.2	Diffusion Approximations via State Space Collapse	22
2.3.3	Sufficient Conditions for State Space Collapse	28
3	Control of Stochastic Processing Networks	30
A	Background	31
A.1	Notation	31
A.2	Stochastic Processes	31
A.3	Path Spaces	32
A.3.1	Definitions and Topologies for \mathbb{C}^d and \mathbb{D}^d	32
A.3.2	Compact Sets in \mathbb{D}^d	33
A.4	Weak Convergence of Probability Measures	35
A.4.1	General Definitions and Results	35
A.4.2	Tightness of Probability Measures on \mathbb{D}^d	36
A.5	Convergence in Distribution for Stochastic Processes	37

A.5.1	Types of Convergence	37
A.5.2	Tightness and Continuity of Limit Processes	37
A.5.3	Continuous Mapping Theorem	39
A.6	Functional Central Limit Theorems	39
A.7	Real Analysis Lemma	40

1 Stochastic Processing Networks: What, Why and How?

This introductory lecture will explain what a stochastic processing network is, why such networks models are of interest in applications, and outline some approaches and challenges associated with the analysis and control of these networks. A pdf copy of the lecture is available by clicking [here](#).

The following subsections contain some further explanations and exercises. These are not intended to replace or replicate the lecture, but rather provide additional commentary.

In addition to studying the material for this lecture, students are encouraged to read the background material in the appendix and to do the exercises at the end of this section.

1.1 Stochastic Processing Network Structure

Stochastic processing networks are used as models for complex processing systems involving dynamic interactions subject to uncertainty. Examples of application domains include manufacturing, the service industry, telecommunications, and computer systems. Typically these networks have entities, such as jobs, customers, or packets, that move along paths or routes, wait in buffers, receive processing from various resources, and that are subject to the effects of stochastic variability through such variables as arrival times, processing times, and routing protocols.

Three key components of the structure of a stochastic processing network are the buffers (or classes) for storing waiting jobs, the resources (or servers) for processing jobs and the processing activities. The *activities* embody the processing capabilities of the network. In abstract terms, an activity consumes from certain classes, produces for certain (possibly different) classes, and uses certain servers in the process. [This notion is an analogue of one used for dynamic *deterministic* production models by Dantzig and others in the 1950s (see [Harrison \(2002\)](#) for more on this connection).] An especially

desirable feature of the notion of a processing activity is that it is broad enough to apply to several familiar categories of processing networks, including open multiclass queueing networks, processing facilities with alternate routing capabilities, and manufacturing plants in which multiple components may be combined to produce new components or in which components may be split up so that different parts undergo different types of processing and processed parts may be combined later on to produce a finished product (so-called fork-join networks). The notion of a stochastic processing network considered here is similar to that used by [Harrison \(2000, 2003\)](#), although we have a somewhat more general notion — Harrison restricts to head-of-the-line service whereas we allow for non-head-of-the-line service as well.

Examples of applications in which different types of activity structures arise are given in the introductory lecture. These include complex semiconductor wafer fabrication, call centers or differentiated service centers, an input queued packet switch and an Internet congestion control model for bandwidth sharing.

1.2 Questions and Approaches

Study of stochastic processing networks typically focusses on the analysis of a system under a fixed operating policy or on the design of an optimal (or near optimal) control policy. Given a fixed system, two main questions of interest are

- (i) when is the system stable?
- (ii) how does it perform when heavily loaded?

Even when considering problems of optimal control, one typically has to ultimately address these two questions for the system operating under a proposed good policy. To illustrate what kinds of insight one might hope to obtain, we indicate answers to these two questions for the classic example of a single server, single class queue.

For general stochastic processing networks (SPNs), exact analysis is usually not possible and approximate models are frequently used in addressing the challenges associated with the performance analysis and control of SPNs. The last 15 years have seen some surprises and major advances associated with the stability and heavy traffic performance of open multiclass HL queueing networks. In particular, there is now a fairly well developed theory of fluid (first order) and diffusion (second order) approximations for these networks. We will describe the mathematics associated with these approximations. The interplay between these two levels of approximation will be an important theme in this development.

In contrast to the situation described above, performance and control for more general stochastic processing networks is an active area of current research. While the general philosophy of using fluid and diffusion approximations to study these problems is a reasonable one, there are many open questions to be resolved. We will describe some of the developments for control of stochastic processing networks and give some applications.

1.3 Two Illustrative Examples

This introductory lecture concludes with two examples to illustrate the kinds of difficulties that can arise in studying stability (Rybko-Stolyar example) and the kind of solutions that can arise in studying control problems for stochastic processing networks.

1.4 Exercises Related to Introductory Lecture

1. Let $\mathbb{D}_+ = \{x : [0, \infty) \rightarrow \mathbb{R}, x \text{ has r.c.l.l. paths, } x(0) \geq 0\}$. Endow \mathbb{D}_+ with the Skorokhod J_1 -topology (see background notes). Prove that the function $\psi : \mathbb{D}_+ \rightarrow \mathbb{D}_+$ defined by

$$\psi(x)(t) = \sup\{x^-(s) : 0 \leq s \leq t\}, \quad t \geq 0, x \in \mathbb{D}_+,$$

is continuous. Here $y^- = \max(-y, 0)$, the negative part of y for $y \in \mathbb{R}$.

2. Consider a single server $GI/GI/1$ queue operating under the FIFO service discipline. For simplicity, assume that the system starts empty. The *workload process* W (also equal to the virtual waiting time process) for this model can be expressed as

$$W(t) = V(E(t)) - t + Y(t), \quad t \geq 0,$$

where

$$V(n) = \sum_{j=1}^n v_j, \quad n \geq 0,$$

is the cumulative sum process for the i.i.d. service times $\{v_j\}_{j=1}^\infty$, E is the exogenous renewal arrival process such that $E(t)$ is the number of new arrivals up to time t , and $Y(t)$ is the cumulative amount of time that the server has been idle up to time t . It can be shown that $Y = \psi(X)$ where

$$X(t) = V(E(t)) - t, \quad t \geq 0,$$

and ψ is the mapping described in Exercise 1. Assume that the i.i.d. sequences of interarrival times and service times are mutually independent and their members have finite first and second moments. Denote the mean arrival rate by λ , the mean service time by m , the variance of the interarrival times by σ_a^2 and the variance of the service times by σ_s^2 .

(i) Use the functional central limit theorems for V and E and the Skorokhod representation theorem to conclude that when $\rho = \lambda m = 1$, for

$$\widehat{X}^r(t) = \frac{X(r^2t)}{r}, \quad t \geq 0, r > 0,$$

we have that \widehat{X}^r converges in distribution to X^* , a one-dimensional Brownian motion, as r tends to infinity through a sequence. The Brownian motion has zero drift. What is the variance parameter of the Brownian motion?

- (ii) Now use the result of Exercise 1, together with Exercise 2(i), and possibly the Skorokhod representation theorem again, to conclude that, as r tends to infinity through a sequence, \widehat{W}^r converges in distribution to W^* , where

$$W^* = X^* + Y^*$$

and $Y^* = \psi(X^*)$. This process W^* is called a one-dimensional *reflected Brownian motion* and Y^* is called its *local time* process at the origin.

3. Consider the Rybko-Stolyar example described in the introductory lecture. This model has Poisson arrivals at rate 1 for each of classes 1 and 3, and i.i.d. exponential service times for each class (where the mean service time can be different for different classes). Classes 2 and 4 have highest priority at their respective stations. Suppose that the system starts empty.

- (i) Convince yourself that $Q_2(t)Q_4(t) = 0$ for all $t \geq 0$ where $Q_i(t)$ denotes the number of jobs in queue i (including those in service) at time t .
- (ii) In view of (i), classes 2 and 4 act as if they are two classes being served by a single server (since at most one of the classes can be served at any one time). This is often referred to as a *virtual station*. For $i = 1, 2, 3, 4$, let $L_i(t)$ denote the cumulative amount of class i work (measured in units of required service time) associated with the exogenous arrivals to the system up to time t . Let $E_i(t)$ denote the cumulative number of exogenous arrivals to class i up to time t and let v_j^i denote the service time associated with the j^{th} job to be processed from class i by the relevant server. Then, for $i = 2, 4$, and $t \geq 0$,

$$L_2(t) = \sum_{j=1}^{E_1(t)} v_j^2$$

$$L_4(t) = \sum_{j=1}^{E_3(t)} v_j^4.$$

The amount of class 2 plus class 4 work associated with the exogenous arrivals to the system by time t is $L_2(t) + L_4(t)$. The amount of this work that is still in the system at time t is at least

$$M(t) = L_2(t) + L_4(t) - t,$$

since this work can be dispensed at a maximum rate of one as only one of the classes 2 or 4 can be served at any one time. Let m_i denote the mean service time for class i , and assume that $m_2 + m_4 > 1$ as in the Rybko-Stolyar example. Use laws of large numbers to show that

$$\lim_{t \rightarrow \infty} M(t)/t = m_2 + m_4 - 1 > 0$$

almost surely. Hence, this queueing system is not stable.

2 Open Multiclass HL Queueing Networks

This section will review developments of the last 15 years which have resulted in a useful approach to analyzing the stability and heavy traffic performance of open multiclass HL queueing networks. In particular, (subcritical) fluid models can be used to determine sufficient conditions for the stability of open multiclass HL queueing networks, and suitable asymptotic behavior of critical (or balanced) fluid models, coupled with an algebraic condition on the first order queueing model data, has been shown to be sufficient to imply the validity of a heavy traffic diffusion approximation for an open multiclass HL queueing network.

A pdf copy of the lectures related to this section is available by [clicking here](#). In the following subsections, we provide some additional notes and exercises related to these lectures.

2.1 Open Multiclass HL Queueing Network Structure

In the terminology used here for multiclass queueing networks, each resource (or station) will consist of a single server and each class (or buffer) can be served by just one server. However, a server may process jobs from more than one class; this is the source of the term “multiclass”. Due to this structure, the mapping from classes to servers is many-to-one. In an open network, entities arrive from external sources and eventually leave the network. An HL scheduling policy is one that is non-idling (or work conserving) and head-of-the-line (some authors just use HL for head-of-the-line, but here we shall also implicitly include the non-idling assumption). Non-idling means that a server will never idle unless there is no work waiting for it to process. Head-of-the-line assumes that entities *within* a buffer are ordered and selected for processing by a server on a first-in-first-out (FIFO) basis. Examples of HL policies are FIFO (across all buffers processed by a given server), and static priority policies which give preference to some buffers over others. Examples of non-HL policies are processor sharing (all entities waiting for processing

by a given server receive an equal share of the processing capacity of the server), and last-in-first-out (LIFO) with preemption.

An open multiclass HL queueing network is specified by describing the network structure (classes and servers and activities connecting them), the stochastic primitives (exogenous arrival process E , cumulative service time process V and routing process Φ), and the HL service discipline for the network. Performance processes of interest include the queue length process Q (one component for each class), the workload process W (one component for each server) and the cumulative idletime process Y (one component for each server). It is also convenient to keep track of the cumulative arrivals A and departures D for each class and the cumulative amount of service time given to each class T . The renewal process associated with the cumulative service process V is denoted by S . The stochastic primitives are assumed to be generated by i.i.d. sequences of random variables. For fluid limit results, they need to satisfy functional laws of large numbers and for the heavy traffic diffusion approximation results they need to satisfy functional central limit theorems. Equations satisfied by the performance processes are used to describe the behavior of the network. There is a common set of equations satisfied by all open multiclass HL networks and then typically one adds some additional equations that are specific to the HL service discipline of interest. Equations for the FIFO (across all classes) and static priority service disciplines are well known. These equations need not be a complete discrete-event type description of the system dynamics; rather they suffice for the stability and performance analysis considered here. The setup for an open multiclass HL queueing network is described in more detail in [Williams \(1998b\)](#).

Using the first order parameters (α, m, P) for an open multiclass HL queueing network, one can specify the traffic equation

$$\lambda = \alpha + P'\lambda$$

for the nominal long run average arrival rate vector λ . The traffic intensity

vector ρ is then given by

$$\rho = CM\lambda,$$

where C is the matrix specifying the many-to-one mapping from classes to servers and $M = \text{diag}(m)$. The queueing network is said to be strictly subcritical if $\rho < 1$, subcritical if $\rho \leq 1$, and critical if $\rho \leq 1$ with $\rho_k = 1$ for at least one k .

2.2 Stability

In the early 1990's several counterexamples appeared showing that open multiclass HL queueing networks need not always be stable even though the traffic intensity parameter is less than one at each station. A two station counterexample given by [Kumar and Seidman \(1990\)](#) used a complicated dynamic (non-idling) policy. Subsequent two-station counterexamples involving static priority policies were given by [Lu and Kumar \(1991\)](#) (deterministic interarrival and service times, reentrant line), and [Rybko and Stolyar \(1992\)](#) (exponential interarrival and service times, no feedback, but traffic flow in opposite directions). The first counterexamples with the FIFO service discipline were also two station examples and were given by [Seidman \(1994\)](#) (deterministic interarrival and service times) and [Bramson \(1994\)](#) (exponential interarrival and service times). FIFO is a particularly difficult discipline to analyze because one needs to keep track of the order in which jobs arrive to each station, hence the state descriptor is necessarily large.

Once counterexamples appeared, there was a rush to find sufficient conditions for stability (i.e., positive recurrence) of open multiclass HL queueing networks. The first technique used was to find Lyapunov functions for Markov processes describing the queueing networks themselves. There is much work in the early to mid 1990's in this vein, including work by Kumar, Kumar and Meyn, Bertsimas et al.,

An alternative approach using fluid models was soon developed. This had the advantage of reducing the problem of obtaining sufficient conditions for stability to the study of deterministic dynamical systems called fluid models.

This approach was first introduced for a specific example of a two station FIFO queueing network with exponential interarrival and service times by [Rybko and Stolyar \(1992\)](#). As these authors pointed out, their procedure was, in principle, quite general in nature. However, for all but the simplest systems, technical problems arise when comparing solutions of the stochastic and deterministic systems. In 1995, independently, [Stolyar \(1995\)](#) and [Dai \(1995\)](#) developed criteria for the stability of open multiclass HL queueing networks, in terms of the stability of fluid limits and fluid models. [Stolyar \(1995\)](#) considered the former, whereas [Dai \(1995\)](#) focussed on the latter. (Fluid limits are fluid model solutions, but the converse need not be true. However, fluid limits can be awkward to work with as their characterization can be a difficult issue.) Also, [Stolyar \(1995\)](#) assumed exponential distributions for the interarrival and service times, whereas [Dai \(1995\)](#) considered more general distributions. The material presented in this section is based on the approach taken in [Dai \(1995\)](#) (see also the exposition in [Bramson \(1998a, 2006\)](#)). (Dai's work was motivated in part by knowledge that an analogous theorem holds for diffusions (cf. [Dupuis and Williams \(1994\)](#)). The results in [Stolyar \(1995\)](#) and [Dai \(1995\)](#) set off another avalanche of work on establishing stability of queueing models by analysing fluid models. Some authors involved in this were Meyn; Bertsimas; Hong Chen; Dai, Hasenbein and Vande Vate; Foss; Stolyar; Bramson;

2.2.1 Markov Model

Various authors use different notions to describe stability. Some use a rather weak notion related to throughput which roughly requires that (assuming existence of long run average rates), the long run departure rate is equal to the long run arrival rate. A more rigorous and stronger form of stability requires that a suitable *Markov* model for the system is positive Harris recurrent. We adopt this latter notion here. (We note that the Markovian state descriptor is really only needed here to make this notion of stability precise. Later on, when we discuss heavy traffic diffusion approximations, we will not need the

full details of this state descriptor.)

A Markovian model for an open multiclass queueing network operating under a fairly broad range of HL policies is described in [Bramson \(1998a\)](#) and also in the recent Saint-Flour notes by [Bramson \(2006\)](#). The state descriptor keeps track of the number of jobs in each class and the ages of those jobs, the residual interarrival times and residual service times of jobs currently being served, as well as the current rate at which each class is being served. Rates of service are allowed to change only when there is a new arrival to or new departure from some class, and rates are assumed to be measurable functions of the state. All of the server's effort for a class is given to the job at the head-of-the-line. The state space is chosen so as to be a locally compact, separable metric space (these assumptions are needed to apply recurrence theory from work of [Meyn and Tweedie \(1993\)](#)).

Remark. We have described here the state descriptor used by Bramson. This suffices for many HL policies. However, we note that for some HL policies, one might want to keep track of even more information and then one may pursue a similar development to that of Bramson. For instance, one may wish to keep track of the age of a job in the network, in addition to the age in a class, e.g., Global-FIFO or first-in-system-first served needs the arrival time to the *network* for each job. Also, some due date based policies need one to keep track of due date information, which may be dynamic, as the manager may update the due date. One may also want to allow randomized policies, so that the processing rate may also depend on a random variable.

2.2.2 Fluid Model

A fluid model is a formal functional law of large numbers approximation for the queueing model. In fact, with suitable initial conditions, one can obtain fluid model solutions as limit points of queueing network quantities under law of large numbers scaling. Depending on the model, fluid model solutions need not be unique and a fluid model may have more solutions than can be obtained as fluid limit points.

2.2.3 Sufficient Conditions for Stability via Fluid Models

Definition 2.1. A fluid model is stable if there exists $t_0 > 0$ such that for any fluid model solution, $\bar{Q}(t) = 0$ for all $t \geq t_0|\bar{Q}(0)|$.

Theorem 2.2. (*Dai (1995)*) *Fix an open multiclass HL queueing network and consider “the” associated fluid model. Suppose that the interarrival times have unbounded support and satisfy a “spread-out” assumption. If the fluid model is stable, then a Markov process describing the queueing network is positive Harris recurrent.*

This result is originally due to [Dai \(1995\)](#). A nice exposition of the proof is given in a paper by [Bramson \(1998a\)](#) (also see the recent Saint-Flour lecture notes by [Bramson \(2006\)](#)). This uses results on positive Harris recurrence from [Meyn and Tweedie \(1993\)](#), as well as verification of a petite set property due to [Meyn and Down \(1994\)](#). [Bramson \(1998a\)](#) has a nice piece of his proof which allows one to restrict to fluid models without delay.

2.2.4 Stability and Equilibria for Some Fluid Models

[Bramson \(1996a,b\)](#) found an entropy function for the fluid models associated with FIFO Kelly type and for HLPPS (head-of-the-line proportional processor sharing) open multiclass queueing networks. He used these functions to show convergence of fluid model solutions to equilibrium states for subcritical fluid models. For the strictly subcritical case, the only equilibrium state is the zero state, and so when coupled with Dai’s theorem, this result of Bramson implies stability of open multiclass queueing networks under these two policies. Note that the state descriptor used for FIFO keeps track of more than just the queue length and workload.

2.2.5 Exercises Related to Fluid Models

1. Under the mild conditions of the theorem of [Dai \(1995\)](#), this exercise enables a proof of stability for open single class HL queueing networks when

the traffic intensity is less than one at each station. The Lyapunov function used for the fluid model here is based on one discovered for reflected Brownian motions by [Chen \(1996\)](#). It is quite surprising that a linear Lyapunov function suffices here.

Consider an open single class HL queueing network (sometimes also called a generalized Jackson network). Here C is the $\mathbb{K} \times \mathbb{K}$ identity matrix, $M = \text{diag}(m)$, P is a (transient) transition matrix with spectral radius strictly less than one, and the network operates under the FIFO service discipline (the only HL discipline for a single class network). Suppose that $\rho < \mathbf{1}$, i.e., $\lambda < \mu$.

- (i) Write down the equations for the fluid model.
- (ii) Show that the fluid queue length satisfies

$$\bar{Q}(t) = \bar{Q}(0) + R(\lambda - \mu)t + R\tilde{Y}(t), \quad t \geq 0,$$

where $R = I - P'$, $\mu_i = m_i^{-1}$ for each $i = 1, \dots, \mathbb{K}$, and $\tilde{Y} = M^{-1}\bar{Y}$. Here $'$ denotes transpose. Let $h > 0$ be a \mathbb{K} -dimensional vector with all entries strictly positive. Define

$$u' = h'R^{-1} = h'(I - P')^{-1}.$$

Then u is well defined and $u > 0$, since the transition matrix P has spectral radius strictly less than one. Also, $u'R = h' > 0$.

Consider a , a strict subset of $\{1, \dots, \mathbb{K}\}$ (not empty and not equal to the whole set). Let b denote the complement of a in $\{1, \dots, \mathbb{K}\}$. Partition R in the obvious way:

$$\begin{pmatrix} R_a & R_{ab} \\ R_{ba} & R_b \end{pmatrix}. \tag{1}$$

Then R_a is of a similar form to R and so R_a^{-1} exists and has all non-negative entries; similarly for R_b^{-1} . The matrices R_{ab} and R_{ba} have

all non-positive entries. Since $u'R > 0$, using obvious notation for subvectors of u , we have

$$u'_a R_a + u'_b R_{ba} > 0, \quad u'_a R_{ab} + u'_b R_b > 0. \quad (2)$$

Multiply the first equation on the right by $R_a^{-1} R_{ab}$ (which has all non-positive entries) to obtain

$$u'_a R_{ab} \leq -u'_b R_{ba} R_a^{-1} R_{ab}.$$

Then substituting this in the other equation, we obtain

$$u'_b R_b - u'_b R_{ba} R_a^{-1} R_{ab} > 0. \quad (3)$$

- (iii) Consider a regular point $t > 0$ for the fluid model: \bar{Q}, \bar{Y} are both differentiable at t . Let $a = \{i : \bar{Q}_i(t) = 0\}$. Then, if a is a strict non-empty subset of $\{1, \dots, \mathbb{K}\}$, the derivative of each component of \bar{Q}_a at t must be zero (each component is non-negative and zero at t and so must have a derivative of zero there - consider the one-sided derivatives from the left and the right). Furthermore, only components of \bar{Y} indexed by a can increase at t (due to the non-idling nature of the policy). Thus, we have

$$\dot{\bar{Q}}_a(t) = R_a(\lambda_a - \mu_a) + R_{ab}(\lambda_b - \mu_b) + R_a \dot{\bar{Y}}_a(t) = 0 \quad (4)$$

$$\dot{\bar{Q}}_b(t) = R_{ba}(\lambda_a - \mu_a) + R_b(\lambda_b - \mu_b) + R_{ba} \dot{\bar{Y}}_a(t) \quad (5)$$

$$= R_b(\lambda_b - \mu_b) - R_{ba} R_a^{-1} R_{ab}(\lambda_b - \mu_b), \quad (6)$$

where $\dot{}$ denotes time derivative and we have substituted the expression for $\dot{\bar{Y}}_a(t)$ from the first equation into the second equation. Thus, when a is a strict non-empty subset of $\{1, \dots, \mathbb{K}\}$, using (3) and the fact that $\lambda < \mu$, we conclude that

$$\frac{d}{dt} u' \bar{Q}(t) = u'_b (R_b - R_{ba} R_a^{-1} R_{ab})(\lambda_b - \mu_b) < 0.$$

If a is empty, then \tilde{Y} cannot increase at t and so

$$\frac{d}{dt}u'\bar{Q}(t) = h'R^{-1}\dot{\bar{Q}}(t) = h'(\lambda - \mu) < 0,$$

since $h > 0$ and $\lambda < \mu$. If a is the whole of $\{1, \dots, \mathbb{K}\}$, $\dot{\bar{Q}}(t) = 0$ and so $\frac{d}{dt}u'\bar{Q}(t) = 0$ then. Conclude that there is $t_0 > 0$ such that $\bar{Q}(t) = 0$ for all $t \geq t_0|\bar{Q}(0)|$.

(iv) State the conclusion that follows from Dai's theorem.

2. Consider the fluid model equations for a FIFO multiclass queueing network. Suppose that $\rho = \mathbf{1}$. Find the invariant states for this fluid model. That is, find $(\bar{Q}, \bar{W}, \bar{A}, \bar{D}, \bar{T}, \bar{Y})$ such that $\bar{Q}(t) = \bar{Q}(0)$ for all $t \geq 0$.

3. Consider a fluid model corresponding to the Rybko-Stolyar example with a static buffer priority policy as described in the first lecture. Initialize this model with $\bar{Q}_1(0) = \bar{Q}_3(0) = 1, \bar{Q}_2(0) = \bar{Q}_4(0) = 0$. Show that there can be more than one fluid model solution starting from this state. (The fluid model equations relating to the static priority discipline have the form

$$t - \bar{T}_i^+(t) \text{ can only increase when } \bar{Q}_i(t) = 0,$$

where $\bar{T}_i^+(t) = \sum_{j \geq i} \bar{T}_j(t)$ is the cumulative amount of time allocated to buffers that have equal or higher priority than i and are served by the same server as buffer i .)

2.3 Performance in Heavy Traffic

In this section, we describe sufficient conditions under which performance measures for open multiclass HL queueing networks can be approximated by certain diffusion processes called semimartingale reflecting Brownian motions.

2.3.1 Semimartingale Reflecting Brownian Motions

A survey describing various aspects of SRBMs is in [Williams \(1995\)](#).

Existence and Uniqueness.

Semimartingale reflecting Brownian motions in the orthant arise as diffusion approximations for the workload processes in heavily loaded open multiclass HL queueing networks. For some networks (principally those that are single class or are multiclass with a feedforward structure), these diffusions can be constructed by a continuous path-to-path mapping from a driving Brownian motion. This is often referred to as the solution of the Skorokhod problem, after Skorokhod who identified the explicit form of the mapping in one-dimension. However, in general, there is no known continuous mapping construction for SRBMs. Indeed, when mapping of continuous paths (not just Brownian paths) is considered, there are examples to show that there can be non-uniqueness for solutions of the Skorokhod problem, see [Mandelbaum \(1992\)](#); [Bernard and El Kharroubi \(1991\)](#).

The Skorokhod problem approach is a “strong” approach to proving existence and uniqueness of SRBMs via a continuous path-to-path mapping and it is limited in scope. In general, reflection matrices coming from open multiclass HL queueing networks that are not feedforward do not satisfy these conditions, and so an alternative approach to constructing the reflecting processes for these networks is needed. Such an approach, using “weak” solutions, is described below.

A necessary condition for the existence of an SRBM starting from each point in the orthant is that the reflection matrix be completely- S , see [Reiman and Williams](#)

(1988). It turns out that this is also sufficient for “weak” existence and uniqueness of the process, see [Taylor and Williams \(1993\)](#).

Theorem 2.3. *There is an SRBM W starting from each point in \mathbb{R}_+^K if and only if the reflection matrix R is completely- S . In this case, each such SRBM is unique in law and these laws define a continuous strong Markov process.*

There is a generalization of the existence and uniqueness results to polyhedral domains by [Dai and Williams \(1995\)](#). More generally, some applications involving stochastic processing networks have conjectured diffusion approximations living in piecewise smooth domains, cf. [Kelly and Williams \(2004\)](#); [Shah and Wischik \(2006\)](#). There is some theory for strong constructions of such diffusions, see [Dupuis and Ishii \(1993\)](#).

Properties of SRBMs.

A sufficient condition for positive recurrence of SRBMs was given in [Dupuis and Williams \(1994\)](#). It requires that all solutions of the linear Skorokhod problem (with driving drift path) are attracted to the origin. The proof uses a Lyapunov function. An alternative proof has been given by Yager and Whitley, using fluid models. If θ is the drift and $R = I - P'$, P a transition matrix for a transient Markov chain, $\text{diag}(P) = 0$, then the SRBM is positive recurrent if and only if $R^{-1}\theta < 0$.

Stationary distributions for SRBMs can be characterized through a Basic Adjoint Relation (BAR) (an integrated form of a partial differential equation). In two dimensions, some techniques from complex analysis can be employed. Some works along these lines are by [Foddy \(1983\)](#); [Harrison et al. \(1985\)](#); [Trefethen and Williams \(1986\)](#) and most recently by [Harrison \(2006\)](#). Conditions for product form stationary distributions have been given by [Harrison and Williams \(1987, 1992\)](#). Some numerical methods are available for solving (BAR): see work of [Dai and Harrison \(1991, 1993\)](#); [Shen et al. \(2002\)](#). However, it is desirable to have improved methods here.

Oscillation Inequality and Invariance Principle.

An oscillation inequality for a perturbed Skorokhod problem is useful for proving weak existence of reflecting Brownian motions when R is completely- S and for proving convergence of queueing network processes to semimartingale reflecting Brownian motions. The incorporation of the δ threshold regions near the boundaries where idletime may increase also makes this inequality useful for establishing results for queueing systems with threshold type controls. The paper developing the oscillation inequality for SRBMs in the orthant and an associated invariance principle is [Williams \(1998a\)](#). A generalization of this to piecewise smooth domains is in [Kang and Williams \(2006\)](#).

2.3.2 Diffusion Approximations via State Space Collapse

For a full development of this topic, see [Williams \(1998b\)](#).

Queueing Model Equations.

For simplicity, we start the system empty. For $t \geq 0$,

$$\begin{aligned} A(t) &= E(t) + \sum_{j=1}^{\mathbb{I}} \Phi^j(D_j(t)) \\ Q(t) &= A(t) - D(t) \\ W(t) &= CV(A(t)) - \mathbf{1}t + Y(t). \end{aligned}$$

Additional equations for FIFO.

For $1 \leq k \leq \mathbb{K}$, $i \in k$, and $t \geq 0$,

$$D_i(t + W_k(t)) - D_i(t) = Q_i(t).$$

Consider a sequence of systems indexed by r tending to infinity through a sequence. Put a superscript of r on each parameter and process that depends on r . For each $r > 0$, let λ^r be the solution of the traffic equation

$$\lambda^r = \alpha^r + (P^r)' \lambda^r,$$

and let $P^{j,r}$ denote the j^{th} row of P^r .

Diffusion Scaling.

For each $r > 0$ and $t \geq 0$, $j \in \{1, \dots, \mathbb{I}\}$, let

$$\begin{aligned}\widehat{E}^r(t) &= \frac{E^r(r^2t) - \alpha^r r^2t}{r}, \\ \widehat{A}^r(t) &= \frac{A^r(r^2t) - \lambda^r r^2t}{r}, \\ \widehat{D}^r(t) &= \frac{D^r(r^2t) - \lambda^r r^2t}{r}, \\ \widehat{V}^r(t) &= \frac{V^r(\lfloor r^2t \rfloor) - m^r r^2t}{r}, \\ \widehat{\Phi}^{j,r}(t) &= \frac{\Phi^{j,r}(\lfloor r^2t \rfloor) - (P^{j,r})' r^2t}{r}, \\ \widehat{Q}^r(t) &= \frac{Q^r(r^2t)}{r}, \\ \widehat{W}^r(t) &= \frac{W^r(r^2t)}{r}, \\ \widehat{Y}^r(t) &= \frac{Y^r(r^2t)}{r}.\end{aligned}$$

Double fluid scaled processes.

For $r > 0$ and $t \geq 0$, let

$$\begin{aligned}\bar{\bar{E}}^r(t) &= \frac{E^r(r^2t)}{r^2}, \\ \bar{\bar{A}}^r(t) &= \frac{A^r(r^2t)}{r^2}, \\ \bar{\bar{D}}^r(t) &= \frac{D^r(r^2t)}{r^2}, \\ \bar{\bar{Q}}^r(t) &= \frac{Q^r(r^2t)}{r^2}.\end{aligned}$$

Diffusion Scaled Queueing Model Equations for r^{th} system.

After some manipulations, for $r > 0$, $t \geq 0$,

$$\begin{aligned}\widehat{Q}^r(t) &= \widehat{A}^r(t) - \widehat{D}^r(t) \\ &= \widehat{E}^r(t) + \sum_{j=1}^{\mathbb{I}} \widehat{\Phi}^{j,r} \left(\bar{\bar{D}}_j^r(t) \right) - (I - (P^r)') \widehat{D}^r(t),\end{aligned}$$

and

$$\begin{aligned}\widehat{W}^r(t) &= C\widehat{V}^r\left(\bar{A}^r(t)\right) + CM^r\left(\widehat{Q}^r(t) + \widehat{D}^r(t)\right) \\ &\quad + (CM^r\lambda^r - \mathbf{1})rt + \widehat{Y}^r(t).\end{aligned}$$

We eliminate the diffusion scaled departure process from the diffusion scaled workload equation, by solving the diffusion scaled queue length equation for the diffusion scaled departure process. Specifically, for $r > 0$ and $t \geq 0$,

$$\widehat{D}^r(t) = (I - (P^r)')^{-1} \left(\widehat{E}^r(t) + \sum_{j=1}^{\mathbb{I}} \widehat{\Phi}^{j,r} \left(\bar{D}_j^r(t) \right) - \widehat{Q}^r(t) \right),$$

and on substituting this into the diffusion scaled workload equation we obtain after some simplification that for $r > 0$ and $t \geq 0$,

$$\widehat{W}^r(t) = \widehat{\chi}^r(t) + \widehat{\gamma}^r t - CM^r((I - (P^r)')^{-1} (P^r)' \widehat{Q}^r(t) + \widehat{Y}^r(t)),$$

where

$$\begin{aligned}\widehat{\chi}^r(t) &= C\widehat{V}^r\left(\bar{A}^r(t)\right) \\ &\quad + CM^r(I - (P^r)')^{-1} \left(\widehat{E}^r(t) + \sum_{j=1}^{\mathbb{I}} \widehat{\Phi}^{j,r} \left(\bar{D}_j^r(t) \right) \right),\end{aligned}$$

and

$$\widehat{\gamma}^r = (CM^r\lambda^r - \mathbf{1})r.$$

Heavy Traffic Assumptions.

Assume that as $r \rightarrow \infty$,

1. $\alpha^r \rightarrow \alpha$,
2. $M^r \rightarrow M$,

3. $P^r \rightarrow P$,

4. $\gamma^r \rightarrow \gamma$,

where P is a substochastic matrix with spectral radius strictly less than one.

We also impose the following functional central limit theorem assumption on the rescaled stochastic primitives. As $r \rightarrow \infty$,

$$(\widehat{E}^r(\cdot), \widehat{V}^r(\cdot), \widehat{\Phi}^r(\cdot)) \Rightarrow (\widehat{E}(\cdot), \widehat{V}(\cdot), \widehat{\Phi}(\cdot)),$$

where $(\widehat{E}(\cdot), \widehat{V}(\cdot), \widehat{\Phi}(\cdot))$ is a multidimensional Brownian motion with certain statistics and independence assumptions.

Diffusion Scaled FIFO Equations.

For $1 \leq k \leq \mathbb{K}$, $i \in k$, $r > 0$, and $t \geq 0$,

$$\widehat{D}_i^r(t + \bar{W}_k^r(t)) - \widehat{D}_i^r(t) = \widehat{Q}_i^r(t) - \lambda_i^r \widehat{W}_k^r(t).$$

Exercise. Consider the sequence of open multiclass queueing networks operating under the FIFO service discipline. Suppose that $(\widehat{Q}^r, \widehat{W}^r) \Rightarrow (\widehat{Q}, \widehat{W})$ as $r \rightarrow \infty$ where \widehat{Q} and \widehat{W} are continuous processes. Conclude from the expression for \widehat{D}^r that $\widehat{D}^r/r \Rightarrow 0$ as $r \rightarrow \infty$ and then use that to conclude that \widehat{D}^r converges in distribution to a continuous process as $r \rightarrow \infty$. Then use this and the diffusion scaled FIFO equation to conclude that for all $1 \leq k \leq \mathbb{K}$ and $i \in k$,

$$\widehat{Q}_i = \lambda_i \widehat{W}_k,$$

where λ is the unique solution of $\lambda = \alpha + P'\lambda$. The above relationship implies that in the limit the queuelength state space can be collapsed from \mathbb{L} dimensions to \mathbb{K} dimensions. This is an example of what has come to be known as *state space collapse*.

Definition 2.4. (State Space Collapse). There exists $\{\Delta^r\}_{r>0} \subseteq \mathbb{R}_+^{\mathbb{L} \times \mathbb{K}}$ with

$$\Delta_{ik}^r = \begin{cases} \delta_i^r, & \text{if } i \in k, \\ 0, & \text{otherwise,} \end{cases}$$

such that, for each $T > 0$,

$$\|\widehat{Q}^r(\cdot) - \Delta^r \widehat{W}^r(\cdot)\|_T \rightarrow 0$$

in probability as $r \rightarrow \infty$, where $\|x\|_T = \sup_{0 \leq t \leq T} |x(t)|$ for $x \in \mathbb{D}^{\mathbb{I}}$.

Example. In the case of FIFO, the natural candidate for Δ^r is given by

$$\delta_i^r = \frac{\lambda_i^r}{\sum_{\ell \in k} \lambda_\ell^r m_\ell^r},$$

$i \in k, 1 \leq k \leq \mathbb{K}$. Then, by the heavy traffic assumptions, for each $1 \leq i \leq \mathbb{I}$, $\delta_i^r \rightarrow \lambda_i$ as $r \rightarrow \infty$. A natural candidate for $\{\Delta^r\}_{r>0}$ also exists for static priorities and HLPPS. In all of these cases, for each $r > 0$, $CM^r \Delta^r = I$.

Suppose that one can prove that state space collapse holds and that $CM^r \Delta^r = I$. Then, in the diffusion scaled workload equation, for $r > 0$ and $t \geq 0$, $\widehat{Q}^r(t)$ can be replaced with $\Delta^r \widehat{W}^r(t)$ plus a small error $\widehat{\epsilon}^r(t)$ whose supremum on each compact time interval converges to zero in probability as $r \rightarrow \infty$. Thus, for $r > 0$ and $t \geq 0$,

$$\begin{aligned} \widehat{W}^r(t) &= \widehat{\chi}^r(t) + \widehat{\gamma}^r t \\ &\quad - CM^r((I - (P^r)')^{-1} (P^r)') \left(\Delta^r \widehat{W}^r(t) + \widehat{\epsilon}^r(t) \right) + \widehat{Y}^r(t). \end{aligned}$$

For $r > 0$, let $G^r = CM^r((I - (P^r)')^{-1} (P^r)') \Delta^r$. In order to solve for $\widehat{W}^r(t)$, we require that $I + G^r$ is invertible for all r sufficiently large. In fact we shall need the following to ensure that the limiting diffusion is well defined.

The Completely-S Assumption.

As $r \rightarrow \infty$, $\Delta^r \rightarrow \Delta$ and for $G = CM(I - P')^{-1} P' \Delta$, $I + G$ is invertible and $R = (I + G)^{-1}$ is completely-S.

It can be shown that $\{(\widehat{W}^r, \widehat{Y}^r)\}_{r>0}$ is C -tight using the oscillation inequality. Then one needs to characterize the limits points. The challenge here is to verify the martingale property that is needed for an SRBM. This boils down to a stopping time argument, which is valid for HL disciplines and the i.i.d. assumptions on the stochastic primitive processes.

In fact it turns out that the following slightly weaker form of state space collapse is easier to verify and is sufficient to work with to prove the diffusion approximation result.

Definition 2.5. (Multiplicative State Space Collapse) There exists $\{\Delta^r\}_{r>0} \subseteq \mathbb{R}_+^{\mathbb{I} \times \mathbb{K}}$ with

$$\Delta_{i,k}^r = \begin{cases} \delta_i^r, & \text{if } i \in k, \\ 0, & \text{otherwise,} \end{cases}$$

such that, for each $T > 0$,

$$\frac{\|\widehat{Q}^r(\cdot) - \Delta^r \widehat{W}^r(\cdot)\|_T}{\|\widehat{W}^r(\cdot)\|_T \vee 1} \rightarrow 0$$

in probability as $r \rightarrow \infty$.

The following is the main diffusion approximation result.

Theorem 2.6. *Consider a sequence of HL queueing networks that satisfies*

1. *the Heavy Traffic Assumptions,*
2. *Multiplicative State Space Collapse, and*
3. *the Completely-S Assumption.*

Then, as $r \rightarrow \infty$,

$$(\widehat{W}^r, \widehat{Q}^r) \Rightarrow (\widehat{W}, \widehat{Q}),$$

where \widehat{W} is a SRBM and $\widehat{Q} = \Delta \widehat{W}$.

Examples.

The conditions in the theorem hold for FIFO Kelly type and HLPPS (cf. [Bramson \(1998a\)](#); [Williams \(1998b\)](#)) with suitable initial conditions. They also hold for some static priorities, e.g., reentrant lines with LBFS (cf. [Bramson and Dai \(2001\)](#)). A specific example of [Dai et al. \(1992\)](#) is a FIFO

Kelly type reentrant line described in the lectures. In this example, $\mathbb{K} = 3$ and $\mathbb{I} = 6$. The routing is such that $1, 6 \in 1$, $2, 3 \in 2$ and $4, 5 \in 3$. Arrivals in this network are Poisson rate 1 and service times are all exponential with rate $1/2$. In this case,

$$R = \begin{pmatrix} 1 & 2/5 & -6/5 \\ -2/3 & 1 & 4/5 \\ 2/9 & -8/5 & 1 \end{pmatrix}.$$

A submatrix obtained by deleting row one and column one is of a type considered by Mandelbaum which exhibits nonuniqueness of pathwise solutions for the Skorokhod problem. Nevertheless, the original network satisfies the conditions of the heavy traffic limit theorem, and the R matrix has uniqueness in distribution for the associated SRBM.

2.3.3 Sufficient Conditions for State Space Collapse

For more details on this aspect, see [Bramson \(1998b\)](#).

Bramson has shown that for a heavily loaded open multiclass HL queueing network, if fluid model solutions starting in a compact set converge uniformly to an invariant manifold (defined by equilibrium fluid model solutions), then under suitable initial conditions, multiplicative state space collapse holds. (Using his entropy functions, he has verified the uniform convergence to an invariant manifold for fluid models of FIFO Kelly type and for HLPPS networks.)

A critical idea (which originated with Mike Harrison) in this development is to view a diffusion scaled process $\{\widehat{U}^r(t) = U^r(r^2t)/r, t \in [0, T]\}$ (for fixed $T > 0$) as derived from $[rT] + 1$ fluid scaled processes $\{\bar{U}^{r,m}(s) = U^r(r(m+s))/r, s \geq 0\}$, $m = 0, 1, \dots, [rT]$. The paths of these fluid scaled processes (over fixed compact time intervals) have cluster points (as $r \rightarrow \infty$) which are fluid model solutions. The assumed uniform asymptotic behavior (convergence to equilibrium) of these fluid models solutions, together with “relative compactness” of the paths of the fluid scaled processes, yields that

there is an integer L' such that, provided $\bar{U}^{r,m}(0)$ lies in a fixed compact set, for all r sufficiently large, for $m = 0, 1, \dots, [rT]$ and $s \in [L', L' + 1]$, $\bar{U}^{r,m}(s)$ is close to an equilibrium state for the fluid model. This equilibrium state satisfies “complete” state space collapse. For $t \in [L'/r, T]$, let $m_r = [rt - L']$, then $\hat{U}^r(t) = \bar{U}^{r,m_r}(rt - m_r)$ where $rt - m_r \in [L', L' + 1)$, and for r sufficiently large (not depending on t), $\bar{U}^{r,m_r}(rt - m_r)$ is close to satisfying complete state space collapse. For $t \in [0, L'/r]$, the initial conditions are used to obtain the state space collapse estimate.

A wrinkle in the above sketch of a proof of the state space collapse is the phrase “provided $\bar{U}^{r,m}(0)$ lies in a fixed compact set”. This may not be ensured with the simple fluid scaling described above, because we are looking at so many initial points as m varies ($[rT] + 1$ of them). (Indeed, we expect $\hat{U}^r(\cdot)$ to move around like a diffusion.) To take care of this, one uses a slightly more elaborate scaling: $\bar{U}^{r,m}(t) = U^r(rm + u_{r,m}t)/u_{r,m}$ where $u_{r,m} = |U^r(rm)| \vee r$ for a suitable “norm” $|\cdot|$ on $U^r(rm)$. This then ensures that the initial values of the $\bar{U}^{r,m}$ lie in a compact set (the “norm unit ball”) and with some more technical work this yields multiplicative state space collapse (rather than state space collapse).

For the proof one needs that with high probability, increments of the fluid scaled primitive processes (restarted at m , for each $m = 0, 1, 2, \dots, [rT]$) stay close to their averages over intervals of length $O(1)$ in fluid scale; this follows from the (second) moment assumptions needed for the functional central limit theorems.

3 Control of Stochastic Processing Networks

For the powerpoint file related to these lectures, [click here](#). As introductory references on this topic, see [Harrison and Van Mieghem \(1997\)](#); [Harrison \(2000\)](#); [Bramson and Williams \(2003\)](#); [Bell and Williams \(2005\)](#).

A Background

A.1 Notation

For each integer $d \geq 1$, let \mathbb{R}^d denote d -dimensional Euclidean space and let \mathbb{R}_+^d denote the d -dimensional non-negative orthant

$$\{x \in \mathbb{R}^d : x = (x_1, \dots, x_d), x_i \geq 0 \text{ for } i = 1, \dots, d\}.$$

We will use the following norm on \mathbb{R}^d :

$$|x| \equiv \max_{1 \leq i \leq d} |x_i|, \quad x \in \mathbb{R}^d. \quad (7)$$

We endow \mathbb{R}^d with the σ -algebra of Borel sets. When $d = 1$, we will omit the qualifying superscript d .

A.2 Stochastic Processes

A d -dimensional stochastic process is a collection of random variables $X = \{X(t), t \geq 0\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in \mathbb{R}^d . In particular, for each $t \geq 0$, the function $X(t) : \Omega \rightarrow \mathbb{R}^d$ is measurable, where Ω is endowed with the σ -algebra \mathcal{F} and \mathbb{R}^d is endowed with the Borel σ -algebra. We shall often write $X(t, \omega)$ or $X_t(\omega)$ in place of $X(t)(\omega)$ for $t \geq 0, \omega \in \Omega$. Such a stochastic process X is said to have r.c.l.l. (sample) paths if for each $\omega \in \Omega$, the function $t \rightarrow X(t, \omega)$ from $[0, \infty)$ into \mathbb{R}^d is right continuous on $[0, \infty)$ and has finite limits from the left on $(0, \infty)$. Here r.c.l.l. stands for “right continuous with finite left limits”. The acronym càdlàg for the equivalent French phrase is used by some authors instead of r.c.l.l. The stochastic process X is said to have continuous (sample) paths if for each $\omega \in \Omega$, the function $t \rightarrow X(t, \omega)$ is continuous from $[0, \infty)$ into \mathbb{R}^d .

We will be concerned with the construction of, and convergence in distribution of, d -dimensional stochastic processes having continuous paths or r.c.l.l. paths. The next few sections summarize some basic definitions and properties needed for this. For more details, the reader is referred to [Billingsley \(1999\)](#) or [Ethier and Kurtz \(1986\)](#) or [Jacod and Shiryaev \(1987\)](#).

A.3 Path Spaces

A.3.1 Definitions and Topologies for \mathbb{C}^d and \mathbb{D}^d

For $d \geq 1$, let $\mathbb{C}^d \equiv C([0, \infty), \mathbb{R}^d)$ denote the space of continuous functions from $[0, \infty)$ into \mathbb{R}^d . When $d = 1$, we shall suppress the superscript d . We endow \mathbb{C}^d with the topology of uniform convergence on compact time intervals. Let \mathbb{D}^d denote the space of functions from $[0, \infty)$ into \mathbb{R}^d that are right continuous on $[0, \infty)$ and have finite left limits on $(0, \infty)$. When $d = 1$, we shall suppress the superscript d . An element $x \in \mathbb{D}^d$ only has discontinuities of jump type and there are only countably many points in $(0, \infty)$ where x has a jump discontinuity. We endow \mathbb{D}^d with the (Skorokhod) J_1 -topology. There is a metric m_{J_1} on \mathbb{D}^d which induces this topology and under which the space is a complete, separable metric space (i.e., a Polish space). For our purposes, we do not need to know the precise form of this metric, rather it will suffice to characterize convergence of sequences in the J_1 -topology. For this, let Γ denote the set of functions $\gamma : [0, \infty) \rightarrow [0, \infty)$ that are strictly increasing and continuous with $\gamma(0) = 0$ and $\lim_{t \rightarrow \infty} \gamma(t) = \infty$. (In particular, γ maps $[0, \infty)$ onto $[0, \infty)$.) A sequence $\{x_n\}_{n=1}^\infty$ in \mathbb{D}^d converges to $x \in \mathbb{D}^d$ in the J_1 -topology if and only if for each $T > 0$ there is a sequence $\{\gamma_n\}_{n=1}^\infty$ (possibly depending on T) in Γ such that

$$\sup_{0 \leq t \leq T} |\gamma_n(t) - t| \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad \text{and} \quad (8)$$

$$\sup_{0 \leq t \leq T} |x_n(\gamma_n(t)) - x(t)| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (9)$$

If x is in \mathbb{C}^d , $\{x_n\}_{n=1}^\infty$ converges to x in the J_1 -topology if and only if $\{x_n\}_{n=1}^\infty$ converges to x uniformly on compact time intervals. In particular, with the topologies described above, \mathbb{C}^d is a topological subspace of \mathbb{D}^d . For later use, for each $T \geq 0$, we define

$$\|x\|_T = \sup_{t \in [0, T]} |x(t)|, \quad \text{for } x \in \mathbb{D}^d. \quad (10)$$

We note in passing that one may endow \mathbb{D}^d with the topology of uniform convergence on compact time intervals. However, this topology is finer than

the J_1 -topology and \mathbb{D}^d with this topology is not separable.

Remark. The reader is cautioned here that for positive integers d and k , the product space $\mathbb{C}^d \times \mathbb{C}^k$ is the same topologically as the space \mathbb{C}^{d+k} , because a sequence of functions $\{x^n\}$ in \mathbb{C}^d say, converges uniformly on each compact time interval if and only if each component function $x_i^n, i = 1, \dots, d$, converges uniformly on each compact time interval. On the other hand, the product space $\mathbb{D}^d \times \mathbb{D}^k$ is not the same topologically as the space \mathbb{D}^{d+k} , because in defining the J_1 -topology a change of time scale (given by functions γ_n , cf. (8)), is used and this may be different for two sequences taking values in the two spaces \mathbb{D}^d and \mathbb{D}^k , respectively.

For the purpose of defining measurability, we endow \mathbb{C}^d with the Borel σ -algebra associated with the topology of uniform convergence on compact time intervals; this agrees with the σ -algebra $\mathcal{C}^d = \sigma\{x(t) : x \in \mathbb{C}^d, 0 \leq t < \infty\}$ which is the smallest σ -algebra on \mathbb{C}^d such that for each $t \geq 0$, the projection mapping $x \rightarrow x(t)$, from \mathbb{C}^d into \mathbb{R}^d , is measurable. We endow the space \mathbb{D}^d with the Borel σ -algebra associated with the J_1 -topology; this agrees with the σ -algebra $\mathcal{D}^d = \sigma\{x(t) : x \in \mathbb{D}^d, 0 \leq t < \infty\}$. As usual, when $d = 1$, we shall omit the superscript d .

Consider a d -dimensional stochastic process X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If X has r.c.l.l. (resp. continuous) paths, then $\omega \rightarrow X(\cdot, \omega)$ defines a measurable mapping from (Ω, \mathcal{F}) into $(\mathbb{D}^d, \mathcal{D}^d)$ (resp. $(\mathbb{C}^d, \mathcal{C}^d)$) and it induces a probability measure π on $(\mathbb{D}^d, \mathcal{D}^d)$ (resp. $(\mathbb{C}^d, \mathcal{C}^d)$) via $\pi(A) = P(\{\omega : X(\cdot, \omega) \in A\})$ for all $A \in \mathcal{D}^d$ (resp. \mathcal{C}^d). This probability measure π is called the *law* of X .

A.3.2 Compact Sets in \mathbb{D}^d

To develop a criterion for relative compactness of probability measures associated with d -dimensional stochastic processes having r.c.l.l. paths, we need a characterization of the (relatively) compact sets in \mathbb{D}^d with the J_1 -topology.

Definition A.1. For each $x \in \mathbb{D}^d$, $\delta > 0$ and $T > 0$, define

$$w'(x, \delta, T) = \inf_{\{t_i\}} \max_i \sup_{s, t \in [t_{i-1}, t_i)} |x(s) - x(t)|$$

where $\{t_i\}$ ranges over all partitions of the form $0 = t_0 < \dots < t_{n-1} < T \leq t_n$ with $\min(t_i - t_{i-1}) > \delta$ and $n \geq 1$.

Remark. This definition is slightly different from that in Billingsley (1999) and is taken from Ethier and Kurtz (1986), p. 122. The notation $w'(x, \delta, T)$ is used rather than simply $w(x, \delta, T)$ because $w(x, \delta, T)$ is commonly used for the usual modulus of continuity. The quantity $w'(x, \delta, T)$ is often called a modified modulus of continuity.

Proposition A.2. For each $x \in \mathbb{D}^d$ and $T > 0$,

$$\lim_{\delta \rightarrow 0} w'(x, \delta, T) = 0.$$

For each $\delta > 0$ and $T > 0$, $w'(\cdot, \delta, T)$ is a Borel measurable function from \mathbb{D}^d into \mathbb{R} .

Proof. See Ethier and Kurtz (1986), Lemma 3.6.2. □

Proposition A.3. A set $A \subset \mathbb{D}^d$ is relatively compact if and only if the following two conditions hold for each $T > 0$:

- (i) $\sup_{x \in A} \|x\|_T < \infty$,
- (ii) $\lim_{\delta \rightarrow 0} \sup_{x \in A} w'(x, \delta, T) = 0$.

Proof. See Ethier and Kurtz (1986), Theorem 3.6.3. □

Remark. One can replace condition (i) by the following and then the same result holds:

- (i)' for each rational $t \in [0, T]$, $\sup_{x \in A} |x(t)| < \infty$.

A.4 Weak Convergence of Probability Measures

A.4.1 General Definitions and Results

Let (\mathbb{M}, m) be a complete separable metric space where \mathbb{M} denotes the set and m is the metric. Let \mathcal{M} denote the σ -algebra of Borel sets associated with the topology induced on \mathbb{M} by m .

Definition A.4. A sequence of probability measures $\{\pi_n\}_{n=1}^\infty$ on $(\mathbb{M}, \mathcal{M})$ converges weakly to a probability measure π on $(\mathbb{M}, \mathcal{M})$ if and only if

$$\int_{\mathbb{M}} f d\pi_n \longrightarrow \int_{\mathbb{M}} f d\pi \quad \text{as } n \rightarrow \infty$$

for each bounded continuous function $f : \mathbb{M} \rightarrow \mathbb{R}$.

Definition A.5. A family of probability measures Π on $(\mathbb{M}, \mathcal{M})$ is (weakly) relatively compact if each sequence $\{\pi_n\}_{n=1}^\infty$ in Π has a subsequence that converges weakly to a probability measure π on $(\mathbb{M}, \mathcal{M})$.

Definition A.6. A family of probability measures Π on $(\mathbb{M}, \mathcal{M})$ is tight if for each $\varepsilon > 0$ there is a compact set A in \mathbb{M} such that

$$\pi(A) > 1 - \varepsilon \quad \text{for all } \pi \in \Pi.$$

Theorem A.7 (Prohorov's Theorem). *A family of probability measures on $(\mathbb{M}, \mathcal{M})$ is tight if and only if it is (weakly) relatively compact.*

Proof. See [Billingsley \(1999\)](#), Theorems 5.1 and 5.2, or [Ethier and Kurtz \(1986\)](#), Theorem 3.2.2. □

Remark. The “if part” of this proposition uses the fact that the metric space (\mathbb{M}, m) is separable and complete.

Corollary A.8. *Suppose that $\{\pi_n\}_{n=1}^\infty$ is a tight family of probability measures on $(\mathbb{M}, \mathcal{M})$ and that there is a probability measure π on $(\mathbb{M}, \mathcal{M})$ such that each weakly convergent subsequence of $\{\pi_n\}_{n=1}^\infty$ has limit π . Then $\{\pi_n\}_{n=1}^\infty$ converges to π .*

Proof. See Billingsley (1999), p. 59. □

The following theorem is often useful for reducing arguments about convergence of probability laws associated with stochastic processes to real analysis arguments based on almost sure convergence of equivalent distributional representatives for those processes.

Theorem A.9 (Skorokhod Representation Theorem). *Suppose that π and $\{\pi_n\}_{n=1}^\infty$ are all probability measures on $(\mathbb{M}, \mathcal{M})$ and that $\{\pi_n\}_{n=1}^\infty$ converges weakly to π . Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which are defined \mathbb{M} -valued random variables X and $\{X^n\}_{n=1}^\infty$ such that X has distribution π , X^n has distribution π_n for $n = 1, 2, \dots$, and $X^n \rightarrow X$ \mathbb{P} -a.s. as $n \rightarrow \infty$.*

Proof. See Ethier and Kurtz (1986), Theorem 3.1.8. □

A.4.2 Tightness of Probability Measures on \mathbb{D}^d

Combining the characterization of relative compactness in \mathbb{D}^d with Prohorov's Theorem yields the following.

Theorem A.10. *A sequence of probability measures $\{\pi_n\}_{n=1}^\infty$ on $(\mathbb{D}^d, \mathcal{D}^d)$ is tight if and only if for each $T > 0$ and $\varepsilon > 0$,*

$$(i) \quad \lim_{K \rightarrow \infty} \limsup_{n \rightarrow \infty} \pi_n(\{x \in \mathbb{D}^d : \|x\|_T \geq K\}) = 0,$$

$$(ii) \quad \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \pi_n(\{x \in \mathbb{D}^d : w'(x, \delta, T) \geq \varepsilon\}) = 0.$$

Proof. See Ethier and Kurtz (1986), Theorem, 3.7.4, or Billingsley (1999), Theorem 16.8. □

Remark. Condition (i) can be replaced by the following and then the above theorem still holds:

$$(i)' \quad \text{for each rational } t \in [0, T], \quad \lim_{K \rightarrow \infty} \limsup_n \pi_n(\{x \in \mathbb{D}^d : |x(t)| \geq K\}) = 0.$$

A.5 Convergence in Distribution for Stochastic Processes

A.5.1 Types of Convergence

Suppose that $X, X^n, n = 1, 2, \dots$, are d -dimensional stochastic processes with r.c.l.l. paths (these processes may be defined on different probability spaces). Let π denote the law of X and for each $n = 1, 2, \dots$, let π_n denote the law of X^n . The sequence of processes $\{X^n\}_{n=1}^\infty$ converges in distribution to X if and only if $\{\pi_n\}_{n=1}^\infty$ converges weakly to π . Some authors abuse terminology and say that $\{X^n\}_{n=1}^\infty$ converges weakly to X . We denote such convergence in distribution by $X^n \Rightarrow X$ as $n \rightarrow \infty$.

Suppose that X and $\{X^n\}_{n=1}^\infty$ are all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\{X^n\}_{n=1}^\infty$ converges to X almost surely (resp. in probability) if and only if $\lim_{n \rightarrow \infty} m_{J_1}(X^n(\omega), X(\omega)) = 0$ for \mathbb{P} -almost every $\omega \in \Omega$ (resp. for each $\varepsilon > 0$, $\lim_{n \rightarrow \infty} \mathbb{P}(\omega \in \Omega : m_{J_1}(X^n(\omega), X(\omega)) \geq \varepsilon) = 0$). (Here m_{J_1} is the metric introduced previously which induces the J_1 -topology on \mathbb{D}^d .) We denote the almost sure convergence by $X^n \rightarrow X$ a.s. as $n \rightarrow \infty$, and we denote the convergence in probability by $X^n \rightarrow X$ in prob. as $n \rightarrow \infty$. If $\{X^n\}_{n=1}^\infty$ converges to X almost surely or in probability, then $\{X^n\}_{n=1}^\infty$ converges in distribution to X . Conversely, if $\{X^n\}_{n=1}^\infty$ converges in distribution to X and X is a.s. constant, then $\{X^n\}_{n=1}^\infty$ converges in probability to X as $n \rightarrow \infty$.

A.5.2 Tightness and Continuity of Limit Processes

The following criterion due to Aldous often provides a convenient mechanism for verifying tightness of the laws associated with d -dimensional stochastic processes having r.c.l.l. paths.

Theorem A.11. *Let $\{X^n\}_{n=1}^\infty$ be a sequence of d -dimensional stochastic processes with r.c.l.l. paths. Then the probability measures induced on \mathbb{D}^d by $\{X^n\}_{n=1}^\infty$ are tight if the following two conditions hold for each $T > 0$:*

$$(i) \lim_{K \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P}(\|X^n\|_T \geq K) = 0,$$

(ii) for each $\varepsilon > 0, \eta > 0$, there are positive constants $\delta_{\varepsilon, \eta}$ and $n_{\varepsilon, \eta}$ such that for all $0 < \delta \leq \delta_{\varepsilon, \eta}$ and $n \geq n_{\varepsilon, \eta}$,

$$\sup_{\tau \in \mathcal{T}_{[0, T]}^f} \mathbb{P}(|X^n(\tau + \delta) - X^n(\tau)| \geq \varepsilon) \leq \eta$$

where $\mathcal{T}_{[0, T]}^f$ denotes the set of all stopping times relative to the filtration generated by X^n that take values in a finite subset of $[0, T]$.

Proof. See Billingsley (1999), Theorem 16.10. \square

The following proposition provides a useful criterion for checking when a sequence of d -dimensional stochastic processes with r.c.l.l. paths has associated laws that are tight and whose limit points are concentrated on the set of continuous paths \mathbb{C}^d . (We say such a sequence of laws is C -tight and (with an abuse of terminology) we sometimes say the sequence of processes is C -tight.)

Proposition A.12. *Let $\{X^n\}_{n=1}^\infty$ be a sequence of d -dimensional processes with r.c.l.l. paths. Then the sequence of probability measures $\{\pi_n\}_{n=1}^\infty$ induced on \mathbb{D}^d by $\{X^n\}_{n=1}^\infty$ is tight and any weak limit point of this sequence is concentrated on \mathbb{C}^d if and only if the following two conditions hold for each $T > 0$ and $\varepsilon > 0$:*

$$(i) \lim_{K \rightarrow \infty} \limsup_{n \rightarrow \infty} P(\|X^n\|_T \geq K) = 0,$$

$$(ii) \lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} P(w(X^n, \delta, T) \geq \varepsilon) = 0,$$

where for $x \in \mathbb{D}^d$,

$$w(x, \delta, T) = \sup \left\{ \sup_{u, v \in [t, t+\delta]} |x(u) - x(v)| : 0 \leq t < t + \delta \leq T \right\}. \quad (11)$$

Proof. See Proposition VI.3.26 in Jacod and Shiryaev (1987). \square

A.5.3 Continuous Mapping Theorem

Theorem A.13 (Continuous Mapping Theorem). *For fixed positive integers d and k , let h_n , $n = 1, 2, \dots, h$ be measurable functions from $(\mathbb{D}^d, \mathcal{D}^d)$ into $(\mathbb{D}^k, \mathcal{D}^k)$. Define*

$$C_h = \{x \in \mathbb{D}^d : h_n(x_n) \rightarrow h(x) \text{ in } \mathbb{D}^k \text{ whenever } x_n \rightarrow x \text{ in } \mathbb{D}^d\}.$$

Let X , X^n , $n = 1, 2, \dots$, be d -dimensional stochastic processes with r.c.l.l. paths such that $X^n \Rightarrow X$ as $n \rightarrow \infty$ and $\mathbb{P}(X \in C_h) = 1$. Then $h_n(X^n) \Rightarrow h(X)$ as $n \rightarrow \infty$.

Exercise A.14. Prove the continuous mapping theorem. Hint: use the Skorokhod Representation Theorem.

A.6 Functional Central Limit Theorems

Theorem A.15 (Donsker's Theorem). *Let v_1, v_2, \dots be independent, identically distributed (i.i.d.) real-valued random variables with mean $\mu \in (-\infty, \infty)$ and variance $\sigma^2 \in (0, \infty)$. Let*

$$\widehat{V}^n(t) = \frac{1}{\sigma\sqrt{n}} \left(\sum_{i=1}^{\lfloor nt \rfloor} v_i - \mu nt \right), \quad t \geq 0. \quad (12)$$

Then $\widehat{V}^n \Rightarrow W$, where W is a standard one-dimensional Brownian motion.

Proof. This is usually shown in two steps: (i) convergence of finite dimensional distributions, and (ii) tightness. See Billingsley (1999), Theorem 14.1, or Ethier and Kurtz (1986), Theorem 5.1.2. \square

Remark. This result is also called a “Functional Central Limit Theorem (FCLT)” or “Invariance Principle”.

Remark. There are extensions of Donsker's theorem that relax the i.i.d. requirements.

Theorem A.16. (Functional Central Limit Theorem for Renewal Processes) Let u_1, u_2, \dots be i.i.d. positive random variables with mean $\lambda^{-1} \in (0, \infty)$ and variance $\sigma^2 \in (0, \infty)$. For each $t \geq 0$, define

$$N(t) = \sup\{k \geq 0 : u_1 + \dots + u_k \leq t\}$$

and set

$$\widehat{N}^n(t) = \frac{1}{\lambda^{3/2}\sigma\sqrt{n}} (N(nt) - \lambda nt).$$

Then $\widehat{N}^n \Rightarrow W$ as $n \rightarrow \infty$, where W is a standard one-dimensional Brownian motion.

Proof. This follows from Donsker's theorem by a clever inversion argument, see [Billingsley \(1999\)](#), Theorem 14.6. \square

A.7 Real Analysis Lemma

Lemma A.17. Let $f : [0, \infty) \rightarrow \mathbb{R}$ be a right continuous function that is of bounded variation on each finite time interval. Then, for each $t \geq 0$,

$$(f(t))^2 - (f(0))^2 = 2 \int_{(0,t]} f(s) df(s) - \sum_{0 < s \leq t} (\Delta f(s))^2. \quad (13)$$

Proof. By [Folland \(1984\)](#), p. 103, we have

$$(f(t))^2 - (f(0))^2 = \int_{(0,t]} (f(s) + f(s-)) df(s) \quad (14)$$

$$= 2 \int_{(0,t]} f(s) df(s) + \int_{(0,t]} (f(s-) - f(s)) df(s) \quad (15)$$

$$= 2 \int_{(0,t]} f(s) df(s) + \sum_{0 < s \leq t} (f(s-) - f(s)) \Delta f(s) \quad (16)$$

$$= 2 \int_{(0,t]} f(s) df(s) - \sum_{0 < s \leq t} (\Delta f(s))^2, \quad (17)$$

where the third equality follows from the fact that the continuous part of f does not charge the countable set of times s at which $f(s) \neq f(s-)$. \square

References

- S. L. Bell and R. J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: asymptotic optimality of a threshold policy. *Electronic J. of Probability*, 10:1044–1115, 2005. URL <http://www.math.ucsd.edu/~williams/dyn/multiserv.html>.
- A. Bernard and A. El Kharroubi. Régulations déterministes et stochastiques dans le premier “orthant” de \mathbb{R}^n . *Stochastics and Stochastics Reports*, 34: 149–167, 1991.
- P. Billingsley. *Convergence of probability measures*. John Wiley & Sons Inc., New York, second edition, 1999.
- M. Bramson. Stability and heavy traffic limits for queueing networks. In *Proceedings of the XXXVIth International Probability Summer School, Saint-Flour, France*. Springer-Verlag, 2006. URL <http://www.springer.com/mathematics/algebra/book/978-3-540-68895-2>.
- M. Bramson. Instability of FIFO queueing networks. *Annals of Applied Probability*, 4:414–431, 1994.
- M. Bramson. Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems: Theory and Applications*, 22:5–45, 1996a.
- M. Bramson. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems: Theory and Applications*, 23:1–26, 1996b.
- M. Bramson. Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems: Theory and Applications*, 28:7–31, 1998a.
- M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Sys-*

- tems: *Theory and Applications*, 30:89–148, 1998b. URL <http://www.math.ucsd.edu/~williams/diff/diff.html>.
- M. Bramson and J. G. Dai. Heavy traffic limits for some queueing networks. *Annals of Applied Probability*, 11:49–90, 2001.
- M. Bramson and R. J. Williams. Two workload properties for Brownian networks. Technical report, Department of Mathematics, University of California, San Diego, 2003.
- H. Chen. A sufficient condition for the positive recurrence of a semimartingale reflecting Brownian motion in an orthant. *Annals of Applied Probability*, 6:758–765, 1996. URL <http://projecteuclid.org/Dienst/UI/1.0/Summarize/euclid.aoap/1034968226?abstract>
- J. G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, 5:49–77, 1995.
- J. G. Dai and J. M. Harrison. Steady-state analysis of RBM in a rectangle: numerical methods and a queueing application. *Annals of Applied Probability*, 1:16–35, 1991.
- J. G. Dai and J. M. Harrison. The QNET method for two-moment analysis of closed manufacturing systems. *Annals of Applied Probability*, 3:968–1012, 1993.
- J. G. Dai and R. J. Williams. Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Theory of Probability and its Applications*, 40:3–53, 1995. Correction, 50:346–347, 2006.
- J. G. Dai, G. Wang, and Y. Wang. Nonuniqueness of the Skorohod problem arising from FIFO Kelly type network. *Private Communication*, 1992.
- P. Dupuis and H. Ishii. Sdes with oblique reflection on non-smooth domains. *Annals of Probability*, 21:554–580, 1993.

- P. Dupuis and R. J. Williams. Lyapunov functions for semimartingale reflecting Brownian motions. *Annals of Probability*, 22:680–702, 1994.
- S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.
- M. Foddy. *Analysis of Brownian motion with drift, confined to a quadrant by oblique reflection*. PhD thesis, Department of Mathematics, Stanford University, 1983.
- G. B. Folland. *Real Analysis*. Wiley, New York, 1984.
- J. M. Harrison. Brownian models of open processing networks: canonical representation of workload. *Annals of Applied Probability*, 10(1):75–103, 2000. ISSN 1050-5164. Correction, 13:390–393, 2003.
- J. M. Harrison. Stochastic networks and activity analysis. In Y. Suhov, editor, *Analytic Methods in Applied Probability*, In Memory of Fridrik Karpelevich, Providence, RI, 2002. American Mathematical Society.
- J. M. Harrison. A broader view of Brownian networks. *Annals of Applied Probability*, 13(3):1119–1150, 2003.
- J. M. Harrison and J. A. Van Mieghem. Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Annals of Applied Probability*, 7:747–771, 1997.
- J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, 22:77–115, 1987.
- J. M. Harrison and R. J. Williams. Brownian models of feedforward queueing networks: Quasireversibility and product form solutions. *Annals of Applied Probability*, 2:263–293, 1992.

- J. M. Harrison, H. Landau, and L. A. Shepp. The stationary distribution of reflected Brownian motion in a planar region. *Annals of Probability*, 13: 744–757, 1985.
- J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer-Verlag, New York, 1987.
- W. Kang and R. J. Williams. An invariance principle for semimartingale reflecting Brownian motions in domains with piecewise smooth boundaries. 2006. URL <http://www.math.ucsd.edu/~williams/invariance/kangpiecewise.html>. Preprint.
- F. P. Kelly and R. J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Probab.*, 14(3):1055–1083, 2004. ISSN 1050-5164. URL <http://www.math.ucsd.edu/~williams/bandwidth/kwfluid.html>.
- P. R. Kumar and T. I. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control*, AC-35:289–298, 1990.
- S. H. Lu and P. R. Kumar. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, 36:1406–1416, 1991.
- A. Mandelbaum. The dynamic complementary problem. *Preprint*, 1992.
- S. P. Meyn and D. Down. Stability of generalized Jackson networks. *Annals of Applied Probability*, 4:124–148, 1994.
- S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993.

- M. I. Reiman and R. J. Williams. A boundary property of semimartingale reflecting Brownian motions. *Probability Theory and Related Fields*, 77: 87–97, 1988. Correction, 80:633, 1989.
- A. N. Rybko and A. L. Stolyar. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problems of Information Transmission*, 28:199–220, 1992.
- T. I. Seidman. ‘First come, first served’ can be unstable! *IEEE Transactions on Automatic Control*, 39:2166–2171, 1994.
- D. Shah and D. Wischik. Optimal scheduling algorithms for input-queued switches. 2006. Preprint.
- X. Shen, H. Chen, J. G. Dai, and W. Dai. The finite element method for computing the stationary distribution of an srbm in a hypercube with applications to finite buffer queueing networks. *Queueing Systems: Theory and Applications*, 42:33–62, 2002.
- A. L. Stolyar. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes and Related Fields*, 1:491–512, 1995.
- L. M. Taylor and R. J. Williams. Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probability Theory and Related Fields*, 96:283–317, 1993.
- L. N. Trefethen and R. J. Williams. Conformal mapping solution of Laplace’s equation on a polygon with oblique derivative boundary conditions. *Journal of Computational and Applied Mathematics*, 14:227–249, 1986.
- R. J. Williams. Semimartingale reflecting Brownian motions in the orthant. In F. P. Kelly and R. J. Williams, editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 125–137, New York, 1995. Springer. URL <http://www.math.ucsd.edu/~williams/talks/belz/srbmsurvey.pdf>.

- R. J. Williams. An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems: Theory and Applications*, 30:5–25, 1998a. URL <http://www.math.ucsd.edu/~williams/diff/diff.html>.
- R. J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems: Theory and Applications*, 30:27–88, 1998b. URL <http://www.math.ucsd.edu/~williams/diff/diff.html>.