STOCHASTIC PROCESSING NETWORKS

Ruth J. Williams[1]

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

xxxxxx 0000. 00:1–24

Keywords

complex heterogeneous networks, stochastic variability, queueing, congestion control, resource sharing

Abstract

Stochastic processing networks arise as models in manufacturing, telecommunications, transportation, computer systems, the customer service industry and biochemical reaction networks. Common characteristics of these networks are that they have entities, such as jobs, packets, vehicles, customers or molecules, 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. Understanding, analyzing and controlling congestion in stochastic processing networks is the aim of the mathematical theory of queueing.

In this article, we begin by summarizing some of the highlights in the development of the theory of queueing prior to 1990; this includes some exact analysis and development of approximate models for certain queueing networks. We then describe some surprises of the early 1990s and ensuing developments of the last 25 years related to the use of approximate models for analyzing the stability and performance of multi-class queueing networks. We conclude with a description of recent developments for more general stochastic processing networks, and point to some open problems.

Contents

1. Introduction

Stochastic models of complex processing networks arise in a wide variety of applications in science and engineering, e.g., in manufacturing, transportation, telecommunications, computer systems, customer service facilities and biochemical reaction networks (Anderson & Kurtz 2011; Bertsekas & Gallager 1992; Goss & Peccoud 1998; Kelly & Yudovina 2014; Koole 2013; Kushner 2001; Meyn 2008; Srikant 2004; Yao 1994). These "stochastic processing networks" typically have entities, such as jobs, vehicles, packets, customers or molecules, that move along paths or routes, receive processing from various resources, such as servers, and are subject to the effects of stochastic variability through such variables as arrival times, processing times and routing protocols. Networks arising from modern applications are often highly complex and heterogeneous in that different classes of entities can share common network resources. Frequently the processing capacity of some resources is limited, i.e., there are bottlenecks, which results in congestion and delay due to entities waiting for processing. Understanding, analyzing and controlling congestion in stochastic processing networks is the aim of the mathematical theory of queueing.

Much of the research in queueing theory to date has focussed on models that fall into a broad subset of stochastic processing networks called multi-class queueing networks. Some exact analysis of these networks is available under certain structural and distributional restrictions. However, for networks with more general assumptions, one usually needs to resort to approximate models. There are various kinds of approximate models that are used, and these typically involve proving some kind of scaling limit to justify their use as approximations. For the approximations considered in this article, we keep the network structure fixed, while rescaling time and space. Here we shall consider two of the most common types of approximations, namely, (i) first order approximations, which are dynamical systems called fluid models, and (ii) second order approximations, which are diffusion models. The dynamics of the approximate models are constrained and hence non-linear, because we are modeling quantities that are non-negative, such as numbers of jobs of various classes that are waiting for processing. Fluid approximations involve a law of large numbers scaling and capture average dynamics, whereas diffusion approximations capture statistical fluctuations around average behavior and involve a central limit theorem type of rescaling. The approximations are justified by proving rigorous stochastic process limit theorems. The diffusion approximations we consider here will be for heavily loaded networks, where the rate at which work is input to the system is approximately balanced by the capacity of the system to process that work. Such networks are of particular interest because the effects of stochastic variability are often most apparent in this "heavy traffic regime".

At this point in time, there is a significant theory that justifies and uses fluid and diffusion approximations to characterize the stability and heavy traffic performance of multi-class queueing networks that use head-of-the-line (HL) processing. Under a HL protocol, entities of the same class that are waiting for processing by a server are served in the order in which they arrived to the server. Approximations for HL multi-class queueing networks have been developed over a number of years and there have been some surprises along the way. Furthermore, there are still some challenging problems to consider for these networks. However, in contrast, understanding the stability and performance of non-HL queueing networks, and more general stochastic processing networks, is only in its early stages of development, and is a very active area of current research. While the general approach of using fluid and diffusion approximations as part of an array of tools for studying these problems is still valid, and some use of this methodology has been made for certain applications, there is as yet no general theory for these more complex stochastic processing networks and there are many open questions.

In this article, we begin by summarizing some of the highlights in the development of the theory of queueing prior to 1990, this includes some exact analysis and development of approximate models for certain queueing networks. We then describe some surprises of the early 1990s and ensuing developments related to the use of approximate models for analyzing the stability and heavy traffic performance of multi-class queueing networks. We conclude with a description of recent developments for more general stochastic processing networks and describe various open problems.

1.1. Some Terminology

Frequently we shall refer to the entities processed in stochastic processing networks as jobs, but depending on the application context, we may alternatively use the terms customers, packets, vehicles or molecules. In our discussion in this article, we shall restrict to open networks in which entities arrive from outside the system and eventually depart the system.

Sections 2 and 3 focus on (multi-class) queueing networks. Such a network consists of a set of nodes (or stations). At each node there are jobs that need to be processed at the node and a server (or a pool of identical servers) capable of processing those jobs. We call the nodes "queues", following Kelly (1979). The term queue, as used here, includes the server(s), the jobs being processed, and the jobs waiting for processing at a node. Jobs at a node are classified into finitely many classes depending on their arrival characteristics, service requirements, and routing needs. Classes at different nodes have distinct labels. If there is more than one class of job at a node, the node will be called a multi-class queue, otherwise it will be called a single-class queue. When a job has finished service at a node, it either departs the system or changes class via a routing mechanism, which may be probabilistic. Acyclic queueing networks are those in which the routing is such that a job never returns to a node that it has previously visited. Feedforward queueing networks are those in which the queues are numbered and jobs are always routed from lower to higher numbered queues. We shall use the term "queueing networks with feedback" to mean queueing networks that allow arbitrary probabilistic routing.

In depicting a multi-class queue, we will find it helpful to relegate different classes of waiting jobs to different buffers so that there is one buffer for each class of job. Note that, in contrast to usage by some authors, the collection of buffers and server(s) at a node is what we are calling a queue. We think of the jobs in a buffer as being ordered according to their time of arrival to the buffer, with the job that arrived the longest time ago being at the head-of-the-line. Under a head-of-the-line (HL) service discipline, when a job is selected from a buffer for service, the job at the head-of-the-line is the one selected. Examples of HL disciplines are First-Come-First-Served (FCFS), where jobs are served in the order in which they arrive to a queue, and priority disciplines which give priority to some buffers over others, while serving jobs within a buffer in HL order. Examples of non-HL disciplines are Last-Come-First-Served (LCFS), in which the last job to arrive to a queue is served first, Random Selection for Service (RSS), in which waiting jobs are selected at random for service, and Processor Sharing (PS), which is an idealization of a round-robin or time-sharing type of discipline. (Here we consider LCFS with preemption, where a job in service will be suspended by a newly arriving job. Service of such a job is resumed when all jobs that arrived after it have been served.) Describing the dynamics of HL networks is frequently simpler than for non-HL networks. A multi-class queueing network is depicted in **Figure 2a**.

2. Some Exact Formulas in Queueing Theory

2.1. Early Developments for Single-Class Queues

The origins of queueing theory lie in the work of Danish mathematician, statistician and engineer, A. K. Erlang, who developed and analyzed models of telephone exchanges. Especially well known is the work (Erlang 1917), which contains his famous loss formula for the steady-state probability that a telephone call coming into a system with finitely many circuits will be blocked. This formula has been used in designing the capacity of telephone systems, and also in other applications, where there is a limited service capacity and no capacity for storing jobs awaiting service. Such systems are often referred to as loss systems (Hunt & Kurtz 1994; Kelly 1991; Zachary & Ziedins 2002).

For calls that arrive according to a Poisson arrival process of rate λ to an exchange with c circuits, where call lengths are independent and exponentially distributed with a mean of μ^{-1}, Erlang's loss formula gives the steady-state probability, B, that a typical arriving call will be blocked:

$$B = \frac{\frac{\lambda^c}{\mu^c c!}}{\sum_{j=0}^{c} \frac{\lambda^j}{\mu^j j!}}. \tag{1}$$

It was later realized that Erlang's loss formula still holds if the call lengths have an arbitrary distribution with finite mean (Sevastyanov 1957). Such insensitivity properties, which expand the applicability of models, have also been identified in other more complex stochastic processing networks. For a short survey and recent work on such insensitivity properties in communication network models, see the works of Bonald (2007) and Zachary (2007).

Work on queueing theory, especially related to telephony, continued after Erlang. Another important formula was developed in the early 1930s by Pollaczek (1930) and Khintchine (1932). For a system with a single server, this formula gives the following succinct expression for the steady-state mean number of jobs L in the system when arrivals follow a Poisson process of rate λ, there is unlimited capacity for storing waiting jobs, the jobs are served in the order in which they arrive to the system and the service times for individual jobs are independent of one another and have a common general distribution with finite mean $(1/\mu)$ and variance (σ_S^2):

$$L = \rho + \frac{\rho^2 + \lambda^2 \sigma_S^2}{2(1-\rho)}. \tag{2}$$

Here $\rho = \lambda/\mu$ is called the traffic intensity for the system; it measures the average load on the system and needs to be less than one for a steady-state distribution to exist. An important feature of this formula is that it shows that the mean number of jobs in the system depends not only on the mean arrival rate and service rate, but also on the variability of the service distribution.

With the successful use of analytical techniques in operations management and decision making during World War II, work in queueing theory, and its application in industry and business, expanded dramatically in the 1950s and 1960s. A summary notation A/S/c, introduced by Kendall (1953), very efficiently described key assumptions for many single-class queues. Here the letter A describes the interarrival time distribution, S the service time distribution and c the number of servers. For example, the letter M in the A or S position stood for Markovian, denoting either Poisson arrivals or independent, exponentially distributed service times, respectively, and the letter GI in the A or S position stood for independent, identically distributed random variables with a general distribution. (We note that some authors implicitly assume independence and then simply use G in place of GI.) Thus such a queue with a single server, Poisson arrivals and independent, exponentially distributed service times would be denoted M/M/1, and M/GI/1 would indicate a single-server queue with Poisson arrivals and independent, identically distributed service times having a general distribution. Later this notation was expanded to A/S/c/K/D to include the system capacity for holding jobs (K) and the service discipline (D) that specifies the order in which the processing of jobs should occur. If the capacity is omitted from the notation, it is assumed to be infinite. Examples of common service disciplines for a single-class queue are FCFS, LCFS, RSS and PS. Only FCFS is a HL discipline (see Section 1.1).

For single-class queues, closed form expressions for steady-state statistics and some time-dependent distributions were obtained, especially where either the arrivals were Poisson or the service times were independent and exponentially distributed, e.g., M/GI/1 or GI/M/1; in these situations, an embedded Markov chain could often be identified and analyzed. Many variations were considered and analyzed for single-class queues under certain assumptions, including situations where rates were state-dependent, arrivals or services occurred in batches, servers broke down, or customers grew impatient. A powerful formula, proved first by Little (1961), and now referred to as Little's law, relates the average number of jobs L in a stable queueing system to the average time W spent per job in the system:

$$L = \lambda W. \tag{3}$$

The conditions under which this formula holds have been generalized over the years and it is true for input-output systems in amazing generality (Stidham 1974).

For more on the exact analysis of single-class queues, see e.g., the book by Asmussen (2003). Henceforth, to simplify the exposition, we shall focus on queues that have an infinite capacity for holding waiting jobs.

2.2. Networks of Single-Class FCFS Queues

The first general results for networks of queues were obtained by Jackson (1957, 1963), motivated by manufacturing applications. The main result of Jackson's 1963 paper provides a beautiful decomposition result for the steady-state behavior of a network of finitely many single-class queues, where external arrivals are given by independent Poisson processes, service times are all independent of one another and exponentially distributed, the service discipline is FCFS, and jobs are routed probabilistically between queues. In particular, Jackson's result implies the following. Consider a network of N single-class queues where jobs arrive from outside the network and eventually leave the network, where the external arrivals to each queue are given by independent Poisson processes with arrival rate $\alpha_i \geq 0$ for queue i (and $\sum_{i=1}^N \alpha_i > 0$), where there is a single server for each queue that serves jobs in the order in which they arrive at the queue (i.e., FCFS order) such that, when there are $n_i > 0$ jobs at queue i, the service time for the job currently in service is exponentially distributed with parameter $1/\mu_i(n_i)$ (so that $\mu_i(n_i)$ is the effective rate at which the job is being served), and where a job upon completing service at queue i is routed next to queue j with probability P_{ij}. Assuming that the routing matrix P has spectral radius less than one and P' denotes the transpose of P, let $\lambda = (\lambda_1, \ldots, \lambda_N)'$ be the unique column vector solution of the traffic equation:

$$\lambda = \alpha + P'\lambda, \tag{4}$$

where $\alpha = (\alpha_1, \ldots, \alpha_N)'$ denotes the column vector of external arrival rates. Suppose that for each $i = 1, \ldots, N$, there is $\pi_i(0) > 0$ such that

$$\pi_i(n) = \pi_i(0) \frac{\lambda_i^n}{\prod_{k=1}^n \mu_i(k)}, \quad n = 1, 2, \ldots, \tag{5}$$

defines a probability distribution π_i on the non-negative integers. Then, there is a unique steady-state distribution for the number of jobs $Q = (Q_1, \ldots, Q_N)$ at each of the queues, which has the following product form:

$$\mathbb{P}(Q = (n_1, \ldots, n_N)) = \prod_{i=1}^N \pi_i(n_i). \tag{6}$$

Thus, in steady-state, the number of jobs in distinct queues are independent of one another and for $i = 1, \ldots, N$, the distribution of Q_i is the same as if the i^{th} queue were an isolated single-class queue fed by a Poisson arrival process of rate λ_i.

2.3. Networks of Multi-Class Queues and Product Form

Multi-class queues are simple examples of stochastic processing networks with resource sharing in that jobs of different classes share the processing capacity of a server (or pool of identical servers). In addition to the service disciplines mentioned in connection with single-class queues above, new possibilities arise for multi-class queues such as priority disciplines, which give priority to one job class over another.

While there was some interest in multi-class queues in the 1950s and 1960s, starting in the 1970s, the theory received a significant boost from the need to model computer systems. Indeed, this motivated the seminal work of Baskett et al. (1975) which provided an important generalization of Jackson's result by establishing sufficient conditions for a multi-class queueing network to have an explicit product form steady-state distribution, i.e., the queues in the network are mutually independent in steady-state and the steady-state distributions for the individual queues have exact expressions. Networks satisfying the conditions of Baskett et al. (1975) are commonly referred to as BCMP networks. In an open BCMP network, there are Poisson arrivals to each class, all of the multi-class queues operate under the same service discipline and the possible disciplines are restricted to four types: FCFS, LCFS, PS and infinitely many servers (IS). For the first three types, there is a single server at each queue. Under the FCFS discipline, the BCMP condition requires the service times to all be independent and exponentially distributed and all jobs at a given queue have the same mean service time, whereas for LCFS, PS and IS, the service times need not be exponentially distributed, but for a given class they are independent and identically distributed with Coxian distributions. Some state-dependence of arrival and service rates is allowed. An important aspect of the BCMP product form distributions is that they are insensitive in that the dependence on the service time distributions is only through the means of the distributions. The BCMP method of proof uses a notion of partial balance (a variant of the notion of detailed balance familiar for reversible Markov chains), which was introduced previously by Whittle (1968), in the context of migration models.

Strong probabilistic insight into the BCMP product form result was provided by Kelly (1979), who proved that a network of quasi-reversible multi-class queues has a product form steady-state distribution and that BCMP networks are examples of networks of quasi-reversible queues. The notion of a quasi-reversible queue was first identified by Muntz (1972) and further developed by Kelly (1979). The notion builds on a result of Burke (1956) for single-class queues: a stationary $M/M/c$ queue has the property that the departure process is a Poisson process with the same rate as the Poisson arrival process and the number of customers at the queue at time t is independent of the departure process up to that time. A quasi-reversible multi-class queue is a queue having a steady-state distribution such that the state of the stationary queue at time t is independent of the arrival times for each class of customer served at the queue subsequent to time t, and of the departure times for each class of customer served at the queue up to time t. The term quasi-reversible stems from the fact that a stationary quasi-reversible multi-class queue, when viewed in reverse time, behaves like a similar multi-class queue but where the parameters are usually different from those of the original queue. Kelly's insights also led to a clear understanding of certain insensitivity properties of networks of quasi-reversible queues.

Since the work of BCMP and Kelly, there have been various extensions and refinements giving sufficient conditions for multi-class queueing networks to have product form steady-state distributions, see e.g., (Chao, Miyazawa & Pinedo 1999). Indeed, new applications and extensions of the theory are still being made. An emerging area of application is to synthetic biology, where both single-class, see e.g., (Arazi, Ben-Jacob & Yechiali 2004; Elgart, Jia & Kulkarni 2010; Jia & Kulkarni 2011; Levine & Hwa 2007), and multi-class (Mather et al. 2010, 2011) queueing networks have been used as stochastic models. In particular, Mather et al. (2010, 2011) derived previously unreported formulas for the steady-state correlations between the numbers of jobs in distinct classes at a node in a multi-class queueing network and applied this to enzymatic networks.

3. Approximations for Heavily Loaded Queueing Networks

As we have indicated in the previous section, much of the early work in queueing theory focussed on exact analysis of individual queues and of networks of such queues that have product form steady-state distributions. This exact analysis required special assumptions on the distributions associated with the arrival and/or service of jobs, such as Poisson arrival processes and/or exponentially distributed service times. While insights from these investigations are still used and expanded upon, systems with general distributions for the arrivals and service typically cannot be analyzed exactly, and it is natural to seek more tractable approximate models for their analysis. Heavily loaded networks, where congestion is a compelling problem, are of particular interest. Over the years, a substantial theory has been built up of diffusion approximations to heavily loaded multi-class queueing networks with head-of-the-line service. However, there have been some surprises along the way, relating at least in part to stability questions. In this section, we provide a brief overview of this development. For more details up through the mid 1990s, we refer readers to (Williams 1996).

To simplify the description, we shall focus attention on situations where there is one server for each queue. For all but the last subsection, the queueing networks considered will be assumed to operate under HL service disciplines such as FCFS (across all buffers processed by a given server) and priority policies which give preference to some buffers over others. We shall not explicitly write HL before the term multi-class queueing networks in the first five subsections. In the last subsection, we shall discuss non-HL queueing networks with service disciplines such as LCFS and PS.

3.1. Approximations for Single-Class Queues and Networks

Early work of Kingman (1961, 1962, 1965), Prohorov (1963), Borovkov (1964, 1965), and others in the 1960s, emphasized approximation of steady-state distributions or of finite-dimensional distributions for waiting times and job counts for a single-class queue. A heavy traffic approximation for the entire job count process was first clearly provided in the work of Iglehart & Whitt (1970a), who considered a FCFS single-class queue (with multiple servers). Their work is generally regarded as the prototype for process-level heavy traffic approximations for single-class queueing networks. This work shows that a sequence of suitably rescaled job count processes for a single-class queue converges in distribution to a one-dimensional reflecting Brownian motion as the traffic intensity parameter tends to one. Here the rescaling is a central limit theorem type of scaling (for the n^{th} system in the sequence, time is accelerated by a large factor r_n^2 and the weight of an individual job is reduced by multiplying the job count by a factor $\frac{1}{r_n}$, which is typically of the order of $1-\rho_n$, where ρ_n is the traffic intensity parameter). Assumptions on the sequences of arrival and service processes require that they satisfy a functional central limit theorem and in particular, interarrival times and service times have finite first and second moments.

Subsequently, Iglehart & Whitt (1970b) generalized their results to acyclic networks of FCFS single-class queues, although the approximating diffusion had a complex description. Harrison (1978) simplified this description for two queues in tandem when he proved a heavy traffic limit theorem for this case and provided an elegant sample path representation for the limit process, which is a reflecting Brownian motion living in the positive quadrant of two-dimensional Euclidean space.

Reiman (1984a) proved the first general heavy traffic diffusion approximation result for networks of FCFS single-class queues with arbitrary probabilistic routing. These networks are sometimes referred to as generalized Jackson networks because they are generalizations of the (product form) networks considered by Jackson (1957, 1963), in the sense that the exponential distributions of the interarrival and service times used by Jackson are replaced by general distributions satisfying the finite first and second moment conditions needed for a functional central limit theorem to hold. In the work of Reiman (1984a), the limit process is a multi-dimensional reflecting Brownian motion that lives in the positive orthant of Euclidean space where the dimension of the space is the number of queues in the network. This diffusion process is constructed by applying a continuous mapping to a Brownian motion. The existence and uniqueness of this mapping, which solves the so-called Skorokhod reflection problem, was proved by Harrison & Reiman (1981). An attractive feature of the diffusions is that their characteristics are determined from just first and second moment information of the arrival and service processes for the network, together with a Markovian routing matrix.

3.2. Some Approximations for Multi-Class Queues and Networks

For multi-class queues, early on, Whitt (1971) proved a heavy traffic limit theorem for a multi-class queue with static priority service. He considered a queue with traffic intensity equal to one and two priority classes (high and low). He showed that the two-dimensional process that tracks the numbers of jobs of each class in the queue at any time, when renormalized with a central limit theorem type of scaling, converges in distribution to a process in which the component corresponding to the high priority class is identically zero and that corresponding to the low priority class is a one-dimensional driftless reflecting Brownian motion.

Subsequently, various authors (Chen & Mandelbaum 1991; Dai & Kurtz 1995; Johnson 1983; Peterson 1991; Reiman 1988) were able to prove heavy traffic limit theorems for more general situations, including for multi-class queues with feedback and for networks of multi-class queues. However, all of these works had restrictions and relied on the existence of a continuous mapping to define the limit reflecting Brownian motion process from a Brownian motion.

One of the most general of these results is due to Peterson (1991) who proved a heavy traffic limit theorem for networks of multi-class queues with feedforward deterministic routing and with either FCFS or static priority service. Peterson's work justifies a diffusion approximation for the job count process, which tracks the number of jobs of each class at each server. The dimension of this process is equal to the number of classes. Peterson's paper highlighted the importance for multi-class heavy traffic limit theorems of considering another process besides the job count process called the (immediate) workload process, which tracks the amount of work (measured in units of required service time) embodied in the jobs at each queue. The dimension of this process is equal to the number of queues (i.e., nodes) in the network, which in a multi-class queueing network is lower than the number of classes, as there are multiple classes associated with at least one queue. Peterson proved, under standard heavy traffic assumptions, that the renormalized workload process converges to a reflecting Brownian motion living in the positive orthant of a Euclidean space whose dimension is equal to the number of queues in the network. He showed also that the job count process converges weakly, with the same normalization as for the workload process, to a process that is a linear lifting of the reflecting Brownian motion, where the form of the linear lifting depends on the service discipline. This phenomenon, that the limit of the renormalized workload process determines the limit for the (typically higher dimensional) renormalized job count process, has been called state space collapse (SSC). This term was coined in earlier work on heavy traffic approximations for multi-class queues by Reiman (1984b, 1988).

After the work of Peterson (1991), a major challenge was to establish a general heavy traffic limit theorem for multi-class queueing networks with feedback (i.e., arbitrary probabilistic routing). While it was known that the diffusion approximations for such networks could not all be described as continuous mappings of Brownian motion, a theory of weak existence and uniqueness for the purported diffusion approximations was in place by the early 1990s (Taylor & Williams 1993) and so a conjecture for a limit theorem could be stated precisely (Harrison & Nguyen 1990, 1993). However, some surprising counterexamples produced in the early 1990s indicated that the behavior of multi-class queueing networks can be considerably more complex than that of their single-class counterparts.

3.3. Surprising Examples

Dai & Wang (1993) produced a surprising two-queue FCFS example which showed that not all multi-class queueing networks can have an approximation as conjectured by Harrison & Nguyen (1990, 1993). Around that time, there was also a growing interest in the stability of multi-class queueing networks. Simple two-queue deterministic examples with a priority service discipline due to Kumar & Seidman (1990) and Lu & Kumar (1991) showed that such networks need not always be stable under a natural load condition, which is normally phrased as saying the traffic intensity parameter is less than one at each queue. Rybko & Stolyar (1992) gave the first stochastic counterexample, for a two-queue network with a priority service discipline. It was a further surprise when two-queue counterexamples for the FCFS service discipline were given by Seidman (1994) (for deterministic interarrival and service times) and Bramson (1994) (for exponentially distributed interarrival and service times).

3.4. Stability of Multi-Class Queueing Networks

The aforementioned examples generated a great deal of interest in seeking sufficient conditions for stability of multi-class queueing networks. The first technique used was to find Lyapunov functions for proving positive recurrence of Markov processes describing the queueing networks themselves. There is much work in the early to mid-1990's in this vein, for a sample, see Kumar & Meyn (1995).

A significant advance was made when an alternative approach using approximate deterministic dynamical system models, called fluid models, was developed. Fluid models can be thought of as first-order approximate models for the queueing networks. Some solutions of these models can be obtained as limits of the original queueing networks under a functional law of large numbers type of rescaling, where the size of the initial condition is allowed to go to infinity. Fluid model solutions need not always be unique and a fluid model may have more solutions than can be obtained as fluid limits.

The fluid model approach was first introduced for a specific example of a two-queue network, with FCFS service and exponentially distributed interarrival and service times, by Rybko & 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 with their method when comparing solutions of the stochastic and deterministic systems. Subsequently, Stolyar (1995) and Dai (1995) independently developed criteria for the stability of multi-class queueing networks, in terms of the stability of fluid limits and fluid models, respectively. As mentioned above, 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. Dai's work was motivated in part by the knowledge that an analogous theorem holds for diffusions (Dupuis & Williams 1994). The results of Stolyar (1995) and Dai (1995) set off another flurry of work on establishing stability of queueing networks by analysing fluid models. In particular, Bramson (1996a,b) found entropy functions for fluid models associated with multi-class queueing networks operating under (i) the FCFS discipline, provided all jobs at a given queue have the same mean service time (Kelly-type), and (ii) a HL version of the PS discipline. This work was particularly notable as FCFS is a difficult discipline to analyze for multi-class queues because one needs to keep track of the order in which jobs arrive to each queue, hence the state descriptor is essentially infinite dimensional. For an excellent exposition of the use of fluid models for studying stability of multi-class queueing networks, see Bramson (2006).

3.5. Diffusion Approximations for Multi-Class Queueing Networks via State Space Collapse

While subcritical fluid models (with traffic intensity less than one at each queue) are used in establishing stability for multi-class queueing networks, critical fluid models (with traffic intensity equal to one at each queue) have been used as an important ingredient in establishing heavy traffic diffusion approximations for these networks. Indeed, in the late 1990s, Bramson (1998) and Williams (1998a) developed a modular approach to proving heavy traffic limit theorems for HL multi-class queueing networks. The idea of this approach is (i) to use asymptotic behavior of critical fluid models to prove a version of state space collapse called multiplicative state space collapse (Bramson 1998), and then (ii) to use the result of (i) to establish a diffusion approximation for the workload process, which can be lifted to a diffusion approximation for the job count process via multiplicative state space collapse (Williams 1998a).

The notion of multiplicative state space collapse (MSSC), introduced by Bramson (1998), is a generalization of the notion of state space collapse (SSC) used by Peterson (1991) in his work on feedforward networks. Loosely speaking, state space collapse holds if, in diffusion scale, the job count process can be approximately recovered from the (typically lower dimensional) workload process and that the precision of this approximation becomes exact in the heavy traffic limit. MSSC involves a normalization by the amount of work in the system. It is easier to verify MSSC than SSC and frequently one can prove SSC once MSSC is known.

The combination of the results of Bramson (1998) and Williams (1998a) yielded heavy traffic limit theorems for multi-class queueing networks that have FCFS service, provided they are of Kelly-type, and for a HL version of the PS discipline. These works also provided a general procedure that has since been utilized by various authors in proving heavy traffic diffusion approximations for multi-class queueing networks with other HL service disciplines, see e.g., (Ata & Lin 2008; Bramson & Dai 2001). The diffusion processes arising in these approximations are semimartingale reflecting Brownian motions (SRBMs) living in the positive quadrant of a suitable Euclidean space. These SRBMs are more general than the reflecting Brownian motions that arise as heavy traffic limits in the works of Peterson (1991); Reiman (1984a), which are defined via a continuous mapping from Brownian motion. For surveys describing various aspects of SRBMs, see Williams (1995) and Dieker (2010), and for an invariance principle that provided a key step in (Williams 1998a), see Williams (1998b).

3.6. Non-HL Queueing Networks

In contrast to the developed theory of stability and heavy traffic diffusion approximations for HL multi-class queueing networks, a similar theory for queueing networks operating under non-HL policies is much less developed. Part of the challenge in dealing with such non-HL policies is the need to develop a good mathematical framework for describing the stochastic dynamics of the network. In the last 15 years, measure-valued processes have been successfully used to describe and analyze some non-HL queueing systems. In particular, under general distributional assumptions, diffusion approximations have been developed for single-class queues operating under certain non-HL policies, including LCFS, PS, shortest remaining processing time (SRPT) and earliest-deadline-first (EDF) policies, see e.g., (Doytchinov, Lehoczky and Shreve 2001; Gromoll 2004; Gromoll & Kruk 2007; Gromoll, Kruk & Puha 2011; Kruk et al. 2011; Limic 2000; Puha 2015). The work of Kruk et al. (2004) treats acyclic multi-class networks under the EDF policy. However, it remains an outstanding problem to develop a general theory of stability and heavy traffic diffusion approximations for multi-class queueing networks with non-HL service disciplines.

4. More General Stochastic Processing Networks

This section begins with an explanation in broad terms of what a stochastic processing network is and gives examples to indicate their flexibility for modeling in applications. This is followed with a brief description of some questions and approaches associated with the analysis and control of stochastic processing networks, and a description of some recent developments, open problems and future directions for research. Henceforth we shall use SPN as an abbreviation for Stochastic Processing Network.

4.1. Network Structure

Three key components of the structure of an SPN are (i) buffers (or classes) for storing waiting jobs, (ii) resources for processing activities, and (iii) activities. Examples of resources are servers or even pools of identical servers. 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 resources in the process. [This notion is a stochastic model analogue of one used for dynamic deterministic production models by mathematicians and economists in the 1950s, see the works of Koopmans (1951) and Harrison (2002) for more on this.] An especially desirable feature of the notion of a processing activity is that it is very broad. In particular, it allows several familiar categories of SPNs to be included within one modeling framework. This includes multi-class 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 to produce a finished product (so-called fork-join networks). The notion of an SPN considered here is similar to that introduced by Harrison (2000, 2003), although we embrace a more general service discipline — Harrison restricts to service protocols that process jobs from the head-of-the-line (HL) of each buffer, whereas we allow for more general non-HL processing as well. A sample schematic for an SPN is shown in **Figure 1**.

4.2. Examples of Activities

To illustrate the flexibility of activities for modeling, in Figure 2 we show schematics for four examples of SPNs related to applications. The applications are to semiconductor chip fabrication, customer service centers, input-queued packet switches and bandwidth sharing in a data network. Below we describe each of these in a bit more detail. Here, unless indicated otherwise, the resources will be servers.

Figure 2a is a schematic for a multi-class queueing network. This is an SPN in which there is just one activity associated with each buffer and where that activity connects to just one server. On the other hand, a server may process multiple activities and hence multiple buffers. The servers thus partition the set of buffers in a multi-class queueing network. A semiconductor chip fabrication plant is an example of a system that can be modeled as a multi-class queueing network (Kumar & Kumar 2001). In this application, (very expensive) machines are the servers. To manufacture a semiconductor chip, a partially finished chip may need to make multiple passes through the same machine to receive similar processing at different stages in the production cycle. Chips in different stages of production are stored in distinct buffers while waiting for processing. Each machine may process from several different buffers, but each buffer is associated with a particular machine. An activity corresponds to a machine processing partially completed chips from a given buffer. The routing of the partially completed chips is typically deterministic but with a lot of feedback.

Figure 2b is a schematic for an SPN that one often sees in customer service situations such as telephone call centers, where service agents are cross-trained to have overlapping capabilities. In the situation depicted in the figure, flexible scheduling of customers is permitted in that customers from the middle buffer may be processed by either the top server or the middle server. Similarly, customers from the bottom buffer may be processed by the middle or bottom server. Which server a particular customer is sent to depends on the scheduling policy used for the network. In large telephone call centers, a pool of agents might be used in place of a single server, where all agents in a pool have the same processing capabilities (Gans, Koole & Mandelbaum 2003).

Figure 2c is a schematic for an SPN associated with a 2×2 input-queued packet switch. Such switches are ubiquitous in routers in the Internet. Their input consists of streams of packets coming from different sources and the switch routes these packets to their intended destinations. For an $N \times N$ switch, there are N sources and N destinations. Although in reality there is one buffer for each source, it is helpful for modeling to think of having N virtual buffers for each source, one for each of the N destinations. Then each arriving packet can be thought of as being stored in a virtual buffer associated with its source and destination while waiting to be routed to its destination by the switch. Thus there are N^2 virtual buffers where packets within a virtual buffer are queued in the order in which they arrived. The switch operates in discrete time and in each time slot it can route at most one packet from each source and at most one packet can be routed to each destination. To describe the scheduling, in each time slot, the switch chooses a bijection between the sources and destinations. When the virtual buffers associated with a bijection are all non-empty, choice of such a bijection results in the transfer of the maximum possible number of packets, N, in a time-slot. Such transfers correspond to the use of one of $N!$ "maximal" activities. Other activities (not shown), in which less than N packets are transferred in a time slot, result when a bijection is chosen for which one or more of the associated virtual buffers is empty. Figure 2c shows the two maximal activities corresponding to the two bijections: $1 \to 1$, $2 \to 2$ and $1 \to 2$, $2 \to 1$, for a 2×2 switch. Input-queued switches are examples of SPNs in which activities can simultaneously consume from more than one buffer.

Figure 2d is a schematic for an SPN associated with the transfer of data in a communication network such as the Internet (Srikant 2004). This schematic is for a so-called connection level model, introduced by Massoulié & Roberts (2000), which aims to capture the dynamics of file arrivals and departures for a data network. Files arrive to the network and each file has an associated deterministic route, consisting of a set of links that it needs to traverse before departing the network. Here the processing resources correspond to links in the network that have a finite bandwidth or rate at which data can be transmitted through them. Files awaiting transmission through the network are stored in buffers where all files in a given buffer are statistically indistinguishable in that they come from a common source, are destined for the same route through the network and their sizes are drawn from a common distribution. The model is at a time scale where propagation time through the network is infinitesimal, so that at any instant of time, a file being transmitted is utilizing the same bandwidth at all of the links on its route. In other words, processing files from a given buffer simultaneously uses (the same) resource capacity from all of the links on the route associated with that buffer. This "simultaneous resource possession" is depicted in Figure 2d. For example, the orange files, stored in the middle buffer, will be transmitted through the middle set of three links, simultaneously using the same bandwidth at these three resources, and the second and fourth links share their bandwidth capacity with more than one file type. Consequently, bandwidth allocated to the orange files will affect how much bandwidth is available at the second and fourth links to allocate to green and pink files, respectively. Of particular interest is how this simultaneous resource possession can cause entrainment of resources resulting in underutilization of network capacity.

4.3. Questions and Approaches

Study of SPNs with general distributional assumptions on interarrival and processing times 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 operating under a given policy, two major questions of interest are as follows.

(i) Is the system stable?
(ii) How does it perform when heavily loaded?

When considering problems of optimal control, one aims to

(iii) Find an operating policy that optimizes some performance criterion.

In answering (iii), one typically has to ultimately address questions (i) and (ii) for the system operating under a proposed optimal policy.

As we have described in Section 3, for HL multi-class queueing networks there is now a considerable theory and tools addressing the stability and heavy traffic performance of such networks via fluid and diffusion approximations. It is desirable to have an analogous theory for more general SPNs. Most of the developments to date in this direction have been for SPNs associated with optimal control or have been tied to specific applications such as packet switches and bandwidth sharing in data networks. In the next subsection, we describe some of these developments and some open problems. In the subsequent subsection, we outline some directions for future research.

As most of the developments to date have been for SPNs with HL service from each buffer (HL SPNs), in the following, the SPNs we consider will be assumed to process jobs from each buffer in an HL manner unless specifically indicated otherwise.

4.4. Recent Developments and Some Open Problems

4.4.1. Optimal Control. In the context of optimal control, mathematical models for SPNs with HL service and proposed approximating diffusion control problems have been introduced by Harrison (2000, 2003). While these proposed approximations are only formal (not rigorously justified), for a subclass of such SPNs called unitary networks, assuming infinite horizon discounted linear holding costs, Budhiraja & Ghosh (2012) have rigorously justified these approximations. (In unitary networks, each activity consumes from at most one buffer and uses at most one server for its processing (Bramson & Williams 2003). Multi-class queueing networks and parallel server systems are examples of unitary networks.) Also, in certain contexts related to applications, see e.g., (Ata & Kumar 2005; Ata & Lin 2008; Bell & Williams 2001, 2005; Dai & Lin 2008; Harrison 1998; Kushner & Chen 2000; Mandelbaum & Stolyar 2004; Stolyar 2004), diffusion approximations for SPNs have been justified.

One class of SPNs that have received considerable attention are parallel server systems. Control problems for these systems are sometimes called assignment problems. These systems are natural models for “one-pass” customer service facilities where there is no feedback and there is some choice in assigning jobs to servers. Figure 2b depicts a small example of such a system. A feature of these models is that flexibility in assigning jobs to servers can potentially permit workload to be dynamically distributed so as to optimize utilization of servers.

Complete Resource Pooling. In a seminal work on parallel server systems, building on notions of resource pooling introduced earlier by Kelly & Laws (1993) for multi-class queueing networks, Harrison & López (1999) identified a condition for optimal utilization of servers (in heavy traffic) called complete resource pooling. An important aspect of the work of Harrison & López (1999) is that the authors showed that complete resource pooling holds for a parallel server system only when a certain server-buffer graph associated with the system is connected (and in fact it is a tree, see (Williams 2000)). Under this condition, the diffusion control problem can be reduced to a control problem for a one-dimensional workload process and Harrison and López solved this problem for linear holding costs. In this solution, idleness of servers is only incurred in the diffusion control problem when there is no work for the servers to do. Subsequently, in successively more general works, Harrison (1998), Williams (2000) and Bell & Williams (2001, 2005), provided an interpretation for this solution in terms of a dynamic threshold policy for the original parallel server system and proved asymptotic optimality of this policy in the heavy traffic limit. For similar systems, with strictly convex holding costs, Stolyar (2004) and Mandelbaum & Stolyar (2004) proved asymptotic optimality of so-called max-weight and generalized $c\mu$-type policies, respectively.

Going beyond one-pass systems, Ata & Kumar (2005) and Dai & Lin (2008) considered optimal control of SPNs with feedback under a complete resource pooling condition. Ata and Kumar proved asymptotic optimality of a discrete review policy for unitary networks with linear holding costs. For more general SPNs with quadratic holding costs, Dai and Lin proved asymptotic optimality of maximum pressure policies, a type of back pressure policy. This followed earlier work of Dai & Lin (2005) showing that maximum pressure policies stabilize SPNs under nominal conditions in the sense that the total output rate is equal to the total input rate. This notion of stability is often called rate stability or it is said that the policy maximizes throughput. An attractive feature of the policies of Dai & Lin (2008), Mandelbaum & Stolyar (2004), and Stolyar (2004), is that they do not require knowledge of the arrival rates for execution of the policies.

Beyond Complete Resource Pooling. When complete resource pooling is not satisfied, there are a few examples of HL multi-class queueing networks (which are not parallel server systems) in which analytic solutions of approximating diffusion control problems have been used to suggest asymptotically optimal controls for the original networks, see e.g., (Budhiraja & Ghosh 2005; Harrison & Wein 1989; Kelly & Laws 1993; Laws 1992; Laws & Louth 1990; Martins, Shreve & Soner 1996; Wein 1990). These examples typically have a small number of queues and corresponding low workload dimension. Ata & Lin (2008) have studied the behavior of SPNs under a maximum pressure policy when a linear independence assumption is satisfied by workload vectors. Beyond analytic solutions, several methods for generically approximating solutions of the diffusion control problems have been proposed for various SPNs, see e.g., (Budhiraja, Chen & Rubenthaler 2014; Budhiraja & Ghosh 2012; Harrison 1996; Kushner & Dupuis 2001; Maglaras 2003). However, these generic methods typically do not take advantage of elegant structure revealed through solving the diffusion control problems. Consequently, it has remained an outstanding open problem to find and exploit structure of the diffusion control problems when complete resource pooling does not hold. As a step in this direction, in recent work, Pesic & Williams (2015) have explored partial pooling in parallel server systems based on a “forest of trees” structure that they derived for these systems, which generalizes the single tree structure of such systems under complete resource pooling. A new definition of workload introduced there potentially opens up avenues for further exploration.

4.4.2. Input-queued Switches. Input-queued switches, as described in Section 4.2, are examples of SPNs in which activities can simultaneously consume from more than one buffer. Various scheduling algorithms have been proposed for these switches that involve choosing matchings (bijections) between sources and destinations. See the work of Shah & Wischik (2006) for a summary. One algorithm that has received substantial study is the maximum weight matching (MWM) algorithm, initially shown to be stable for radio hop networks by Tassiulas & Ephremides (1992). Letting Q_{ij} denote the number of packets in the virtual buffer associated with source i and destination j, the MWM algorithm chooses a matching π (a bijection from $\{1,\dots,N\}$ to itself) that achieves the maximum value for the weight, $\sum_{i=1}^{N} Q_{i\pi(i)}$, associated with π. A variant of the MWM algorithm is the MWM-α algorithm for which $Q_{i\pi(i)}$ in the last expression is replaced by $Q^{\alpha}_{i\pi(i)}$ for $\alpha > 0$. An advantage of these algorithms is that they only depend on information about numbers of packets waiting to be processed and do not need information about arrival rates.

It has been shown under various assumptions on the inputs that an input-queued switch is rate stable under the MWM (Dai & Prabhakar 2000; McKeown, Anantharam & Walrand 1996) and MWM-α (Keslassy & McKeown 2001; Shah & Wischik 2006) algorithms. Shah & Wischik (2012) studied the behavior of fluid models and proved multiplicative state space collapse (MSSC) for input-queued switches operating under any one of a family of policies that includes the MWM-α, $\alpha > 0$, policies. Kang & Williams (2012) subsequently used this MSSC to prove a heavy traffic diffusion approximation under the MWM policy. It remains an open problem to prove such a diffusion approximation under the MWM-α policy for all $\alpha \neq 1$. In fact, Shah & Wischik (2012) proved their MSSC for acyclic networks of switches and it is an open problem to prove a diffusion approximation in that context as well for all $\alpha > 0$.

In other work on input-queued switches, some authors have sought bounds on the behavior of the expected value of the total number of packets under various matching algorithms, where the bounds are in terms of the switch size N and the load on the system ρ, especially as $N \to \infty$ and/or $\rho \to 1$. In particular, Shah, Tsitsiklis & Zhong (2011) conjectured that in certain limiting regimes, an optimal bound should be of the order of $N/(1-\rho)$. For the latest on this conjecture and other references, see the works of Shah, Walton & Zhong (2014), Shah, Tsitsiklis & Zhong (2014) and Maguluri & Srikant (2015).

4.4.3. Connection Level Models with Bandwidth Sharing. For the connection level models of Massoulié & Roberts (2000), described in Section 4.2, there has been considerable interest in understanding their stability and heavy traffic behavior when operated under "fair" bandwidth sharing policies. This is especially true for a family of such policies introduced by Mo & Walrand (2000), called α-fair policies, which are indexed by a parameter $\alpha \in (0, \infty)$. The case $\alpha = 1$ includes the notion of proportional fairness, which was introduced earlier by Kelly (1997). Assuming Poisson arrivals and exponentially distributed file sizes, the process that tracks the number of files on each route is a continuous time Markov chain and the model is equivalent in distribution to a HL SPN. Lyapunov functions constructed by de Veciana, Lee & Konstantopoulos (2001) for $\alpha = 1$ and for max-min fair (corresponding to $\alpha \to \infty$), and by Bonald & Massoulie (2001) for all $\alpha \in (0, \infty)$, imply positive recurrence of the Markov chain associated with this model when the average load on each resource is less than its capacity. Kang et al. (2009) proved that MSSC holds in heavy traffic for all $\alpha \in (0, \infty)$. Then assuming a local traffic condition, they used this to prove a diffusion approximation when $\alpha = 1$ and that for proportional fairness this diffusion has a product form stationary distribution. Ye & Yao (2012) subsequently showed how to remove the local traffic condition. It remains an open problem to prove a diffusion approximation for $\alpha \neq 1$; the main difficulty being the current lack of a general theory of well posedness for the approximating diffusion process. Several authors have considered variants of the connection level model and more general bandwidth sharing policies under the Poisson arrivals and exponential file size assumptions, see (Harrison et al. 2015) and the introduction to (Gromoll & Williams 2009).

When file sizes are generally distributed, the connection level model under a bandwidth sharing policy is a network generalization of a processor sharing queue where the bandwidth allocated to a route is shared equally amongst all of the files on a route. This is a non-HL

SPN, which can be described using a measure-valued process that keeps track of residual file sizes or ages of files (and also residual interarrival times if arrivals are not Poisson). Even the question of stability of such an SPN is challenging. Stability has been proved for certain values of α, e.g., for max-min fair by Bramson (2010), and for $\alpha = 1$ with Poisson arrivals and phase-type file size distributions by Lakshmikantha, Beck & Srikant (2004) for certain network topologies and generally by Massoulié (2007). A fluid limit theorem has been proved for this model by Gromoll & Williams (2009) and the invariant states of the fluid model for α-fair policies have been characterized there. Stability of the fluid model under α-fair policies has been established under a nominal load condition for linear networks and simple tree networks by (Gromoll & Williams 2008) and more generally by Paganini et al. (2012), under the assumption that fluid model solutions are sufficiently smooth that they have densities satisfying a partial differential equation. Lee (2008) established sufficient conditions for stability of the fluid model to imply positive recurrence of a Markovian measure-valued age process associated with the original connection level model. In summary, there is nearly a complete theory for stability of the connection level model under α-fair policies with general file size distributions. The matter of diffusion approximations is however open except for the case of $\alpha = 1$, which has recently been addressed by Vlasiou, Zhang & Zwart (2014) for file sizes having phase-type distributions.

4.5. Future Directions

As one can see from the previous section, there has been progress on studying stability, performance and control for some types of SPNs and some policies for operating them. It is likely that new developments will continue to be inspired by applications. However, at a general mathematical level, it remains an outstanding problem to develop a general theory of stability and heavy traffic diffusion approximations for SPNs. For HL policies, the model framework of Harrison (2000, 2003) can be used to describe the networks, and it is desirable to develop a theory parallel to what is now available for HL multi-class queueing networks. For non-HL policies, one has the same problem for SPNs as for multi-class queueing networks of developing a workable, general mathematical framework to describe the stochastic dynamics of these networks. Measure-valued processes appear to be a potentially useful modeling tool here. However, so far, this area is relatively underdeveloped. The connection level models with bandwidth sharing policies and general interarrival time and file size distributions are good test examples of such networks.

This article has only touched on some aspects of SPNs. Three other important directions in which research is being conducted for these models and which we have not treated here are “heavy tails”, “large deviations” and “many server limits”. The book by Whitt (2002) is a useful starting point for heavy tails, and “Big Queues” (Ganesh, O’Connell & Wischik 2004) is a good place to start for those interested in large deviations for SPNs. The study of many server limits for SPNs has been motivated principally by applications to large telephone call centers (Gans, Koole & Mandelbaum 2003). This has been a very active area of research in recent years. See the recent survey (Dai & He 2012) and references therein, for a start.

ACKNOWLEDGMENTS

Some parts of this article draw on unpublished notes prepared in conjunction with the Belz lectures delivered by the author at the University of Melbourne in September 2006 and on talks given by the author, slides for which are available at http://www.math.ucsd.edu/∼williams/research.html. Research of the author described here is supported in part by NSF grant DMS-1206772. The author thanks J. G. Dai, D. Lipshutz, J. M. Harrison, F. P. Kelly, A. L. Puha and K. Ramanan for helpful comments on a preliminary draft of this article.

LITERATURE CITED

Anderson DF, Kurtz TG. 2011. Continuous-time Markov chain models for chemical reaction networks. In *Design and Analysis of Biomolecular Circuits*, eds. H Koeppl, D Densmore, G Setti, M di Bernardo. New York: Springer, 3–42

Arazi A, Ben-Jacob E, Yechiali U. 2004. Bridging genetic networks and queueing theory. *Physica A-Statistical Mechanics and Its Applications* 332:585–616

Asmussen S. 2003. *Applied Probability and Queues*, 2nd ed. Springer-Verlag, New York

Ata B, Kumar S. 2005. Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. *Ann. Appl. Probab.* 15:331–391

Ata B, Lin W. 2008. Heavy traffic analysis of maximum pressure policies for stochastic processing networks with multiple bottlenecks. *Queueing Syst.* 59:191–235

Baskett F, Chandy KM, Muntz RR, Palacios FG. 1975. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* 22:248–260

Bell SL, Williams RJ. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Ann. Appl. Probab.* 11:608–649

Bell SL, Williams RJ. 2005. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: asymptotic optimality of a threshold policy. *Electronic J. of Probab.* 10:1044–1115

Bertsekas D, Gallager R. 1992. *Data Networks.* Englewood Cliffs, NJ: Prentice-Hall

Bertsimas D, Paschalidis IC, Tsitsiklis JN. 1994. Optimization of multiclass queueing networks: polyhedral and nonlinear characterizations of achievable performance. *Ann. Appl. Probab.* 4:43–75

Bonald T. 2007. Insensitive traffic models for communication networks. *Discrete Event Dynamic Systems* 17:405–421

Bonald T, Massoulie L. 2001. Impact of fairness on Internet performance. *SIGMETRICS/Performance*, 82–91

Borovkov A. 1964. Some limit theorems in the theory of mass service, I. *Theory Probab. Appl.* 9:550–565

Borovkov A. 1965. Some limit theorems in the theory of mass service, II. *Theory Probab. Appl.* 10:375–400

Bramson M. 1994. Instability of FIFO queueing networks. *Ann. Appl. Probab.* 4:414–431

Bramson M. 1996a. Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Syst.* 22:5–45

Bramson M. 1996b. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Syst.* 23:1–26

Bramson M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst.* 30:89–148

Bramson M. 2006. *Stability of Queueing Networks.* Lecture Notes in Mathematics, Vol. 1950. Springer-Verlag

Bramson M. 2010. Network stability under max-min fair bandwidth sharing. *Ann. Appl. Probab.* 20:1126–1176

Bramson M, Dai JG. 2001. Heavy traffic limits for some queueing networks. *Ann. Appl. Probab.* 11:49–90

Bramson M, Williams RJ. 2003. Two workload properties for Brownian networks. *Queueing Syst.* 45:191–221

Budhiraja A, Chen J, Rubenthaler S. 2014. A numerical scheme for invariant distributions of constrained diffusions. *Math. Oper. Res.* 39:262–289

Budhiraja A, Ghosh AP. 2005. A large deviations approach to asymptotically optimal control of crisscross network in heavy traffic. *Ann. Appl. Probab.* 15:1887–1935

Budhiraja A, Ghosh AP. 2012. Controlled stochastic networks in heavy traffic: convergence of value functions. *Ann. Appl. Probab.* 22:734–791

Burke PJ. 1956. The output of a queueing system. *Oper. Res.* 4:699–704

Chen H, Mandelbaum A. 1991. Stochastic discrete flow networks: diffusion approximations and bottlenecks. *Ann. Probab.* 19:1463–1519

Chao X, Miyazawa M & Pinedo M. 1999. *Queueing Networks: Customers, Signals and Product Form Solutions.* Wiley

Dai JG. 1995. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* 5:49–77

Dai JG, He S. 2012. Many server queues with customer abandonment: a survey of fluid and diffusion approximations. *J. Syst. Sci. Syst. Eng.* 21:1–36

Dai JG, Kurtz TG. 1995. A multiclass station with Markovian feedback in heavy traffic. *Math. Oper. Res.* 20:721–742

Dai JG, Lin W. 2005. Maximum pressure policies in stochastic processing networks. *Oper. Res.* 53:197–218

Dai JG, Lin W. 2008. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.* 18:2239–2299

Dai JG, Prabhakar B. 2000. The throughput of data switches with and without speedup. *Proceedings of IEEE INFOCOM* 2:556–564

Dai JG, Wang Y. 1993. Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Syst.* 13:41–46

de Veciana G, Lee TJ, Konstantopoulos T. 2001. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking* 9:2–14

Dieker AB. 2010. Reflected Brownian motion. In *Encyclopedia of Operations Research and Management Science*, Wiley

Doytchinov B, Lehoczky J, Shreve S. 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* 11:332-378.

Dupuis P, Williams RJ. 1994. Lyapunov functions for semimartingale reflecting Brownian motions. *Ann. Prob.* 22:680–702

Elgart V, Jia T, Kulkarni RV. 2010. Applications of Little's law to stochastic models of gene expression. *Physical Review E* 82:021901

Erlang AK. 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges (in Danish). *Elektrotkeknikeren* 13:5–13

Ganesh A, O'Connell N, Wischik D. 2004. Big Queues. *Lecture Notes in Mathematics*, vol. 1838. Springer-Verlag

Gans N, Koole G, Mandelbaum A. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* 5:79–141

Goss PJE, Peccoud J. 1998. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proceedings National Academy of Sciences* 95:6750–6755

Gromoll HC. 2004. Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.* 14:555–611

Gromoll HC, Kruk Ł. 2007. Heavy traffic limit for a processor sharing queue with soft deadlines. *Ann. Appl. Probab.* 17:1049–1101

Gromoll HC, Kruk L, Puha AL. 2011. Diffusion limits for shortest remaining processing time queues. *Stoch. Syst.* 1:1–16

Gromoll HC, Williams RJ. 2008. Fluid model for a data network with α-fair bandwidth sharing and general document size distributions: two examples of stability. In *Markov Processes and Related Topics: a Festschrift for Thomas G. Kurtz*, vol. 4 of *Inst. Math. Stat. Collect.* Inst. Math. Statist., Beachwood, OH, 253–265

Gromoll HC, Williams RJ. 2009. Fluid limits for networks with bandwidth sharing and general document size distributions. *Ann. Appl. Probab.* 19:243–280

Harrison JM. 1978. The diffusion approximation for tandem queues in heavy traffic. *Adv. Appl. Probab.* 10:886–905

Harrison JM. 1996. The BIGSTEP approach to flow management in stochastic processing networks. In *Stochastic Networks: Theory and Applications*, eds. FP Kelly, S Zachary, I Ziedins, vol. 4 of *Lecture Note Series*. Oxford University Press, 57–90

Harrison JM. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies. *Ann. Appl. Probab.* 8:822–848

Harrison JM. 2000. Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.* 10:75–103. Correction, 13:390–393, 2003

Harrison JM. 2002. Stochastic networks and activity analysis. In *Analytic Methods in Applied Probability*, ed. Y Suhov, In Memory of Fridrik Karpelevich. American Mathematical Society, Providence, RI

Harrison JM. 2003. A broader view of Brownian networks. *Ann. Appl. Probab.* 13:1119–1150

Harrison JM, López MJ. 1999. Heavy traffic resource pooling in parallel-server systems. *Queueing Syst.* 33:339–368

Harrison JM, Mandayam CV, Shah D, Yang Y. 2015. Resource sharing networks: Overview and an open problem. *Stochastic Systems* 5:1–32

Harrison JM, Nguyen V. 1990. The QNET method for two-moment analysis of open queueing networks. *Queueing Syst.* 6:1–32

Harrison JM, Nguyen V. 1993. Brownian models of multiclass queueing networks: current status and open problems. *Queueing Syst.* 13:5–40

Harrison JM, Reiman MI. 1981. Reflected Brownian motion on an orthant. *Ann. Prob.* 9:302–308

Harrison JM, Wein LM. 1989. Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Syst.* 5:265–280

Hunt P, Kurtz T. 1994. Large loss networks. *Stoch. Proc. Appl.* 53:363 – 378

Iglehart DL, Whitt W. 1970a. Multiple channel queues in heavy traffic I. *Adv. Appl. Probab.* 2:150–177

Iglehart DL, Whitt W. 1970b. Multiple channel queues in heavy traffic II: Sequences, networks, and batches. *Adv. Appl. Probab.* 2:355–369

Jackson JR. 1957. Networks of waiting lines. *Oper. Res.* 5:518–521

Jackson JR. 1963. Jobshop-like queueing systems. *Management Science* 10:131–142

Jia T, Kulkarni RV. 2011. Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Phys. Rev. Lett.* 106:058102

Johnson DP. 1983. *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks.* Ph.D. thesis, University of Wisconsin

Kang WN, Kelly FP, Lee NH, Williams RJ. 2009. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* 19:1719–1780

Kang WN, Williams RJ. 2012. Diffusion approximation for an input-queued switch operating under a maximum weight matching policy. *Stochastic Systems* 2:277–321

Kelly FP. 1979. *Reversibility and Stochastic Networks.* Wiley

Kelly FP. 1991. Loss networks. *Ann. Appl. Probab.* 1:319–378

Kelly FP. 1997. Charging and rate control for elastic traffic. *European Transactions on Telecommunications* 8:33–37

Kelly FP, Laws CN. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Syst.* 13:47–86

Kelly FP, Yudovina E. 2014. *Stochastic Networks.* Cambridge University Press

Kendall DG. 1953. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics* 24:338–354

Keslassy I, McKeown N. 2001. Analysis of scheduling algorithms that provide 100% throughput in input-queued switches. In *Proceedings of the 39th Annual Allerton Conference*

Khintchine AY. 1932. Mathematical theory of a stationary queue. *Matematicheskii Sbornik* 39:73–84

Kingman JFC. 1961. The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* 57:902–904

Kingman JFC. 1962. On queues in heavy traffic. *Journal of Royal Statist. Soc., Series B* 24:383–392

Kingman JFC. 1965. In *Proc. Symp. on Congestion Theory*, eds. WL Smith, *et al.* University of North Carolina Press, 137–159

Koole G. 2013. *Call Center Optimization.* MG books, Amsterdam

Koopmans TC, ed. 1951. *Activity Analysis of Production and Allocation.* New York: John Wiley and Sons

Kruk Ł, Lehoczky J, Ramanan K, Shreve S. 2011. Heavy traffic analysis for EDF queues with reneging. *Ann. Appl. Probab.* 21:484–545

Kruk Ł, Lehoczky J, Shreve S, Yeung SN. 2004. Earliest-deadline-first service in heavy-traffic acyclic networks. *Ann. Appl. Probab.* 14:1306–1352

Kumar PR, Meyn SP. 1995. Stability of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* 40:251–260

Kumar PR, Seidman TI. 1990. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control* 35:289–298

Kumar S, Kumar PR. 2001. Queueing network models in the design and analysis of semiconductor wafer fabs. *IEEE Transactions on Robotics and Automation* 17:548–561

Kushner HJ. 2001. *Heavy Traffic Analysis of Controlled Queueing and Communication Networks.* New York: Springer Verlag

Kushner HJ, Chen YN. 2000. Optimal control of assignment of jobs to processors under heavy traffic. *Stochastics and Stochastics Reports* 68:177–228

Kushner HJ, Dupuis P. 2001. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, New York, 2nd ed.

Lakshmikantha A, Beck CL, Srikant R. 2004. In *Conference on Information Sciences and Systems, Princeton*

Laws CN. 1992. Resource pooling in queueing networks with dynamic routing. *Adv. Appl. Probab.* 24:699–726

Laws CN, Louth GM. 1990. Dynamic scheduling of a four-station queueing networks. *Prob. Eng. Inf. Sci.* 4:131–156

Lee NH. 2008. *A Sufficient Condition for Stochastic Stability of an Internet Congestion Control Model in Terms of Fluid Model Stability.* Ph.D. thesis, University of California, San Diego

Levine E, Hwa T. 2007. Stochastic fluctuations in metabolic pathways. *Proceedings of the National Academy of Sciences* 104:9224–9229

Limic V. 2000. On the behavior of LIFO preemptive resume queues in heavy traffic. *Electron. Comm. Probab.* 5:13–27

Little JDC. 1961. A proof for the queuing formula: $L = \lambda W$. *Oper. Res.* 9:383–387

Lu SH, Kumar PR. 1991. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control* 36:1406–1416

Maglaras C. 2003. Continuous-review tracking policies for dynamic control of stochastic networks. *Queueing Syst.* 43:43–80

Maguluri S T, Srikant R. 2015. Queue length behavior in a switch under a maxweight algorithm.

Preprint

Mandelbaum A, Stolyar AL. 2004. Scheduling flexible servers with convex delay costs: heavy traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* 52:836–855

Martins LF, Shreve SE, Soner HM. 1996. Heavy traffic convergence of a controlled, multi-class queueing system. *SIAM Journal on Control and Optimization* 34:2133–2171

Massoulié L. 2007. Structural properties of proportional fairness: stability and insensitivity. *Ann. Appl. Probab.* 17:809–839

Massoulié L, Roberts JW. 2000. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems* 15:185–201

Mather WH, Cookson NA, Hasty J, Tsimring LS, Williams RJ. 2010. Correlation resonance generated by coupled enzymatic processing. *Biophysical Journal* 99:3172–3181

Mather WH, Hasty J, Tsimring LS, Williams RJ. 2011. Factorized time-dependent distributions for certain multiclass queueing networks and an application to enzymatic processing networks. *Queueing Syst.* 69:313–328

McKeown N, Anantharam V, Walrand J. 1996. Achieving 100% throughput in an input-queued switch. *Proceedings of IEEE INFOCOM*, 296–302

Meyn S. 2008. *Control Techniques for Complex Networks.* Cambridge: Cambridge University Press

Mo J, Walrand J. 2000. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking* 8:556–567

Muntz RR. 1972. Poisson departure process and queueing networks. IBM Research Report RC 4145, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.

Paganini F, Tang A, Ferragut A, Lachlan LH. 2012. Network stability under alpha fair bandwidth allocation with general file size distribution. *IEEE Trans. Automat. Control* 57:579–591

Pesic V, Williams RJ. 2015. Dynamic scheduling for parallel server systems in heavy traffic: Graphical structure, decoupled workload matrix and some sufficient conditions for solvability of the Brownian control problem. Preprint

Peterson WP. 1991. A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Oper. Res.* 16:90–118

Pollaczek F. 1930. Über eine aufgabe der wahrscheinlichkeitstheorie. *Mathematische Zeitschrift* 32:64–100

Prohorov Y. 1963. Transient phenomena in processes of mass service. *Litovsk. Mat. Sb.* 3:199–205. In Russian

Puha AL. 2015. Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *To appear in Ann. Appl. Probab.*

Reiman MI. 1984a. Open queueing networks in heavy traffic. *Math. Oper. Res.* 9:441–458

Reiman MI. 1984b. Some diffusion approximations with state space collapse. In *Modeling and Performance Evaluation Methodology*, eds. F Baccelli, G Fayolle. Berlin: Springer, 209–240

Reiman MI. 1988. A multiclass feedback queue in heavy traffic. *Adv. Appl. Probab.* 20:179–207

Rybko AN, Stolyar AL. 1992. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problems of Information Transmission* 28:199–220

Seidman TI. 1994. 'First come, first served' can be unstable! *IEEE Transactions on Automatic Control* 39:2166–2171

Sevastyanov BA. 1957. An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theory Probab. Appl.* 2:104–112

Shah D, Tsitsiklis JN, Zhong Y. 2011. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Syst.* 68:375–384

Shah D, Tsitsiklis JN, Zhong Y. 2014. On queue-size scaling for input-queued switches. Preprint

Shah D, Walton NS, Zhong Y. 2014. Optimal queue-size scaling in switched networks. *Ann. Appl. Probab.* 24:2207–2245

Shah D, Wischik D. 2006. Optimal scheduling algorithms for input-queued switches. *Proceedings of INFOCOM*

Shah D, Wischik D. 2012. Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse. *Ann. Appl. Probab.* 22:70–127

Srikant R. 2004. *The Mathematics of Internet Congestion Control.* Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA

Stidham S. 1974. A last word on $L = \lambda W$. *Oper. Res.* 22:417–421

Stolyar AL. 1995. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes and Related Fields* 1:491–512

Stolyar AL. 2004. Maxweight scheduling in a generalized switch: state space collapse and equivalent workload minimization in heavy traffic. *Ann. Appl. Probab.* 14:1–53

Tassiulas L, Ephremides A. 1992. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control* 37:1936–1948

Taylor LM, Williams RJ. 1993. Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probab. Theory Related Fields* 96:283–317

Vlasiou M, Zhang J, Zwart B. 2014. Insensitivity of proportional fairness in critically loaded bandwidth sharing networks. Preprint

Wein L. 1990. Scheduling networks of queues: heavy traffice analysis of a two-station network with controllable inputs. *Oper. Res.* 38:1065–1078

Whitt W. 1971. Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Probab.* 8:74–94

Whitt W. 2002. *Stochastic Process Limits.* New York: Springer

Whittle P. 1968. Equilibrium distributions for an open migration process. *J. Appl. Prob.* 5:567–571

Williams RJ. 1995. Semimartingale reflecting Brownian motions in the orthant. In *Stochastic Networks*, eds. FP Kelly, RJ Williams, vol. 71 of *The IMA volumes in mathematics and its applications.* New York: Springer

Williams RJ. 1996. On the approximation of queueing networks in heavy traffic. In *Stochastic Networks: Theory and Applications*, eds. FP Kelly, S Zachary, I Ziedins. Royal Statistical Society, Oxford University Press

Williams RJ. 1998a. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst.* 30:27–88

Williams RJ. 1998b. An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Syst.* 30:5–25

Williams RJ. 2000. On dynamic scheduling of a parallel server system with complete resource pooling. In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, eds. DR McDonald, SRE Turner, vol. 8 of *Fields Institute Communications*. American Mathematical Society

Yao DD. 1994. *Stochastic Modeling and Analysis of Manufacturing Systems.* Springer Series in Operations Research. New York: Springer

Ye H, Yao DD. 2012. A stochastic network under proportional fair resource control - diffusion limit with multiple bottlenecks. *Oper. Res.* 60:716–738

Zachary S. 2007. A note on insensitivity in stochastic networks. *J. Appl. Prob.* 44:238–248

Zachary S, Ziedins I. 2002. A refinement of the [H]unt-Kurtz theory of large loss networks, with an application to virtual partitioning. *Ann. Appl. Probab.* 12:1–22

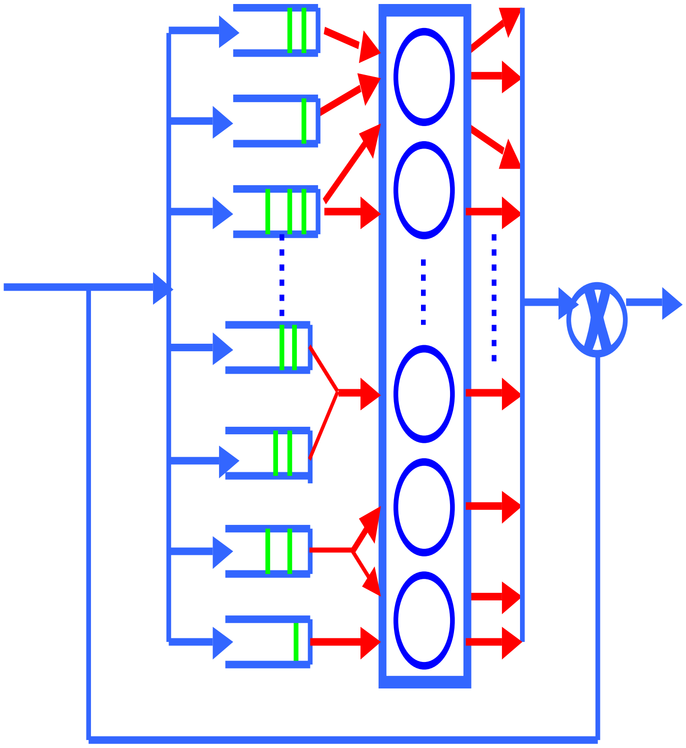

Figure 1

A sample schematic for a stochastic processing network. Here open-ended rectangles represent infinite capacity buffers for storing entities awaiting processing, circles represent resources (e.g., servers) for performing processing of activities, red arrows between buffers and resources represent activities, the circle with a cross indicates a routing mechanism whereby entities produced by a processing activity may be routed to various buffers in the network. Different types of activities are shown, illustrating their modeling flexibility. The green vertical bars indicate entities awaiting processing by an activity. The leftmost and rightmost arrows indicate inputs to and outputs from the network, respectively.

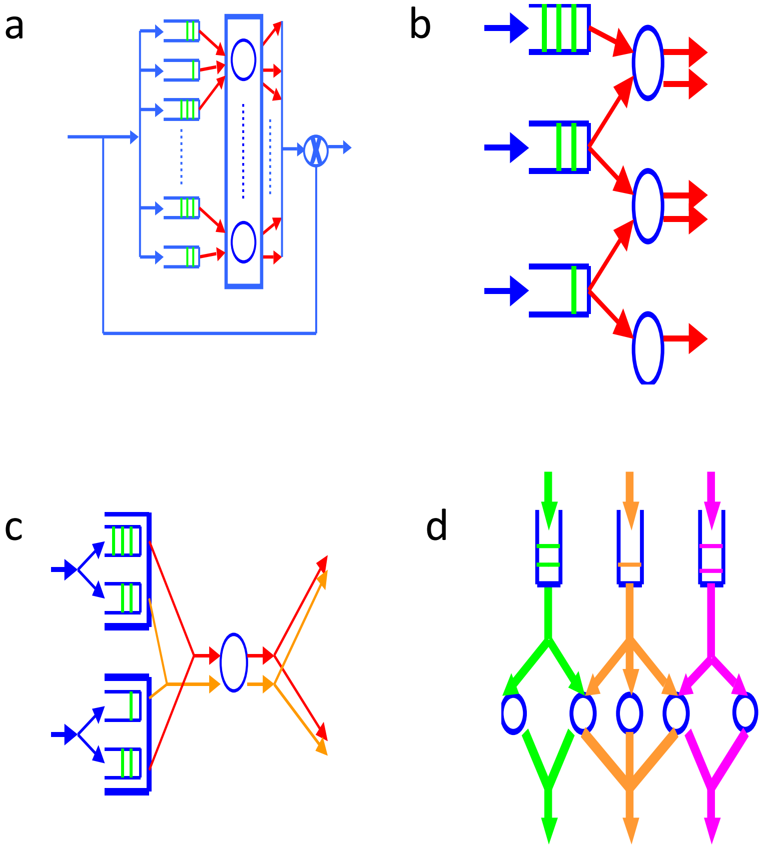

Figure 2

Schematics for four stochastic processing networks relating to applications. (a) Semiconductor chip fabrication. (b) Customer service center. (c) Input-queued packet switch. (d) Bandwidth sharing in a data network.