Stability and Convergence of Moments for Multiclass Queueing Networks via Fluid Limit Models

Jim Dai*
Georgia Institute of Technology

Sean Meyn[†]
University of Illinois

Drafted on April 13, 1994
Revised on March 8, 1995

Abstract

The subject of this paper is open multiclass queueing networks, which are common models of communication networks, and complex manufacturing systems such as wafer fabrication facilities. We provide sufficient conditions for the existence of bounds on long-run average moments of the queue lengths at the various stations, and we bound the rate of convergence of the mean queue length to its steady state value. Our work provides a solid foundation for performance analysis either by analytical methods or by simulation.

These results are applied to several examples including re-entrant lines, generalized Jackson networks, and a general polling model as found in computer networks applications.

Keywords: Multiclass queueing networks, ergodicity, general state space Markov processes, polling models, generalized Jackson networks, stability, performance analysis.

1 Introduction

The subject of this paper is open multiclass queueing networks, which are models of complex systems such as wafer fabrication facilities or communication networks. A simple example is illustrated in Figure 1. In this three station (i.e. machine) network, which might model a simple manufacturing system, one type of products are to be made. Jobs arrive at station 1 according to a general renewal process with arrival rate 1. Each job follows a deterministic route, and the station sequence that a job visits is 1, 2, 3, 2, 3, 2, 1, 3 and 1. Following Kelly [1], a job (or customer) class is defined to be the combination of a job type and a processing stage. Therefore, in this example, each machine processes three job classes. Unlike the networks described in

*School of Industrial and Systems Engineering, and School of Mathematics, Atlanta, GA 30332-0205. Research supported by NSF grants DMS-9203524 and DDM-9215233, and two grants from the Texas Instruments Corporation

[†]Coordinated Science Laboratory 1308 W. Main St., Urbana, IL 61801 Research supported in part by the NSF under Grant No. ECS 1-5-30679.

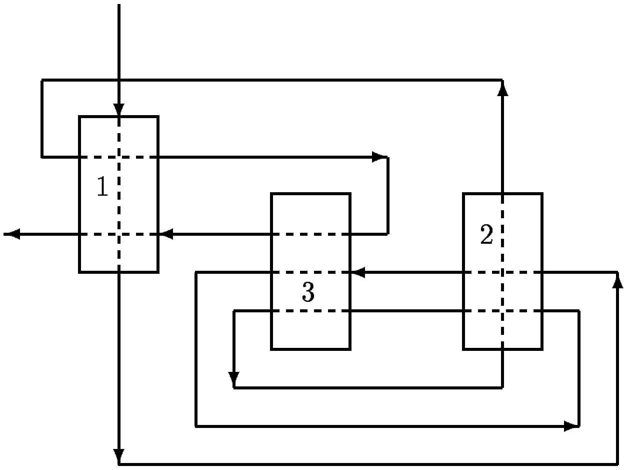

Figure 1. A three station network that may be unstable under certain priority service disciplines

[1], the external arrival processes are assumed to be general renewal processes, the processing requirements have class dependent distributions, and service discipline at each station can be general. Hence a job may require different processing requirements at subsequent visits to a machine.

We are interested in steady state performance criteria such as

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t |Q(s)|^p \, ds, \tag{1.1}$$

where $|Q(t)|$ is the total number of jobs in the system at time t. Specifically, we search for answers to the following questions:

(i) When is the long-run average (1.1) finite?

(ii) When does the pth moment $\mathsf{E}[|Q(s)|^p]$ converge to a steady state value as $s \to \infty$?

(iii) What is the rate of convergence of $\mathsf{E}[|Q(s)|]$ to stationarity?

Long run averages are frequently considered by system analysts in performance evaluation. In general, no analytical formula exists to describe the limit, and even approximations are difficult to obtain. Hence, computer simulations are still the primary tool available to estimate steady state performance measures such as (1.1).

Unfortunately, even stability of the network is frequently not known in advance, and recent research shows that in many multiclass networks, the limit supremum in (1.1) may be infinite for any $p > 0$, even though the usual capacity constraints for the network are satisfied (see e.g. Lu and Kumar [2], Bramson [3] and Seidman [4]). To illustrate the difficulties that can arise in a multiclass network, we present here the following simulation study of the network pictured in Figure 1. In these simulations, we make the simplifying assumption that each station has a common service time distribution with mean m_i, $i = 1, 2, 3$. We take $m_1 = m_2 = 0.3$ and $m_3 = 0.1$, and we assume that customers enter the network at rate 1. Therefore the nominal workload

case	running time	queue length at each station			utilization rate at each station			cycle time
		1	2	3	1	2	3	
(M)	1,000	41.07	91.74	0.10	0.73	0.82	0.25	137.66
	10,000	493.61	772.79	0.10	0.76	0.77	0.26	1289.58
	100,000	4993.16	7106.94	0.11	0.77	0.76	0.25	12446.21
(D1)	1,000	37.62	79.96	0.00	0.73	0.79	0.26	108.29
	10,000	483.49	718.17	0.00	0.77	0.77	0.26	1228.28
	100,000	4534.39	8301.29	0.00	0.74	0.79	0.26	13439.38
(D2)	1,000	0.40	0.30	0.04	0.90	0.90	0.30	2.85
	10,000	0.42	0.29	0.04	0.90	0.90	0.30	2.85
	100,000	0.42	0.29	0.04	0.90	0.90	0.30	2.85

Table 1. For (M) and (D1), average queue lengths at stations 1 and 2 grow without bound, while the queue length at station 3 nearly zero. The simulation (D2) is well behaved, even though the network differs from (D1) initially by only two jobs.

per unit of time for servers 1, 2 and 3 are 90%, 90% and 30%, respectively. Finally, we must specify the service discipline. At station 1, priority is given to customer classes in order $(9,7,1)$, where buffer 9 has highest priority. At station 2, priority is given to customer classes in order $(4,2,6)$, and at station 3 the service discipline is FIFO.

We have simulated this network under two distributional assumptions. In the first case (case (M)), all distributions are assumed to be exponential. In the second case (case (D)), all distributions are degenerate. That is, there is no randomness at all in the network. For (M) and (D2), the network is initially empty. For (D1), there are two jobs initially in front of buffer 1. The simulation is done using SIMAN 3.5 [5]. It is clear from Table 1 that the queue lengths in the simulations (M) and (D1) are unbounded, whereas in simulation (D2) the total customer population remains bounded. Figure 2 plots the queue length processes at stations 1 and 2 for system (M) in the first 10,000 units of simulation time. The plot again suggests that the *total* queue length cycles to infinity. Readers are referred to Dai and Weiss [6, Remark 2 in Section 6] and Gu [7] for more insight.

Previous research in this area has concentrated primarily on single station systems. See in particular the work of Keifer and Wolfowitz [8], Miyazawa [9], Daley and Rolski [10], and Sigman and Yao [11]. Polling models are treated in, for example, Altman et. al. [12, 13] or Georgiadis and Szpankowski [14]. Recently, more complex queueing networks have received greater attention. Generalized Jackson networks are treated in Borovkov [15], Sigman [16], Meyn and Down [17], and Baccelli and Foss [18]. Re-entrant lines are considered in Kumar et. al. [19, 20, 2, 21].

In recent years, significant progress has been made in two lines of research. The work of Meyn and Tweedie [22, 23, 24, 25] gives a framework for the analysis of continuous time, general state space Markov processes, which is in particular applicable to jump processes such as queueing and storage models. The work of Harrison [26], Chen and Mandelbaum [27, 28] and Harrison and Nguyen [29] has focused on the dynamics and sample path properties of multiclass networks. This work provides a valuable set of tools for network analysis which, in particular, has led to methods for

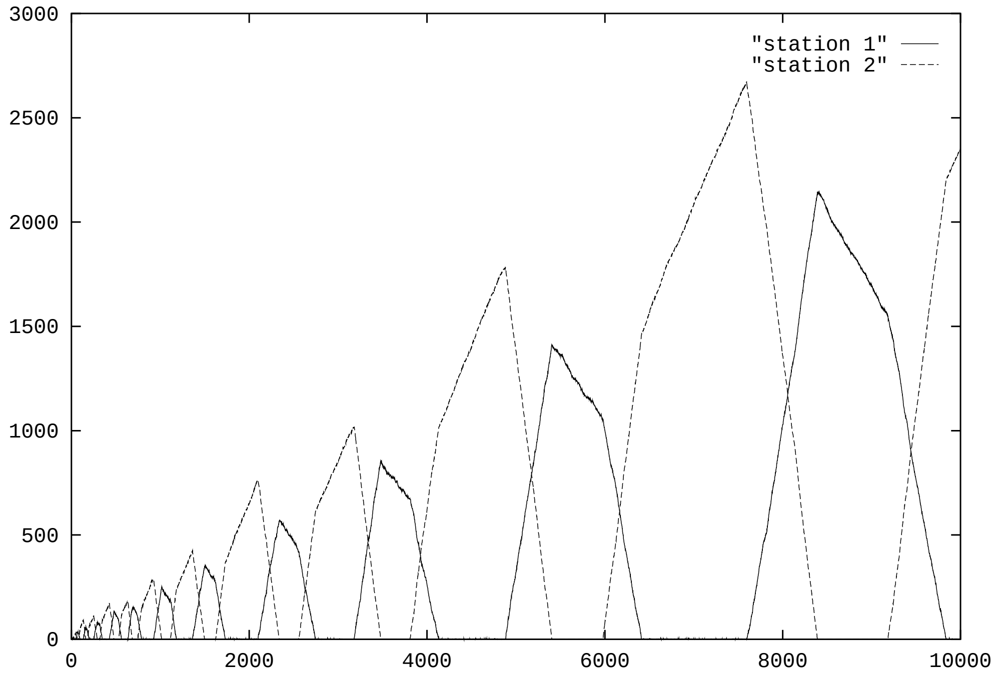

Figure 2. The queue lengths at each station oscillate with increasing magnitude: Mutual blocking between machines 1 and 2 results in instability.

approximation of networks by reflected Brownian motion models, and fluid approximations for multiclass networks. In this paper we merge these methods to obtain a general set of tools for answering such questions as (i)–(iii) above. This provides a solid foundation for performance analysis either by analytical methods or by simulation, and is sufficiently general to allow straightforward application to several diverse areas in which network models are used in practice.

In Section 2, we give precise definition of the network model considered, and Section 3 is devoted to fluid models and their stability. We summarize our main results and provide several examples in Section 4. In particular, in Section 4.3 we consider a token passing ring and provide a simple proof of stability and boundedness of moments for this important example in full generality. In Section 5, we obtain a bound for long-run average moments of the queue length process, and in Section 6 we strengthen these results to obtain convergence of steady state moments. The paper concludes with some discussion of future directions.

2 A Multiclass Network

2.1 Network Model

We consider a network composed of d single server stations, which we index by $i = 1, \ldots, d$. The network is populated by K classes of customers, where customers of class k arrive to the network via an exogenous arrival process with i.i.d. interarrival times $\{\xi_k(n), n \geq 1\}$. We allow $\xi_k(n) \equiv \infty$ for all n for some k, in which case we say that the external arrival process for customers of class k is *null.* We let $\mathcal{A}$ denote the set of classes with non-null exogenous arrivals. Hereafter, whenever external arrival processes are under discussion, only classes with non-null exogenous arrivals are considered. Class k customers require service at station $s(k)$. Their service times are also i.i.d., and are denoted $\{\eta_k(n), n \geq 1\}$. We assume that the buffers at each station have infinite capacity.

Routing is assumed to be *Bernoulli*, so that upon completion of service at station $s(k)$, a class k customer becomes a customer of class ℓ with probability $P_{k\ell}$, and exits the network with probability $1 - \sum_\ell P_{k\ell}$, independent of all previous history. To be more precise, let $\phi^k(n)$ be the routing vector for the nth class k customer who finishes service at station $s(k)$. The ℓth component of $\phi^k(n)$ is one if this customer becomes a class ℓ customer and zero otherwise. Therefore, $\phi^k(n)$ is a K-dimensional "Bernoulli random variable" with parameter P_k', where P_k denotes the kth row of $P = (P_{k\ell})$ (all vectors are envisioned as column vectors, and primes denote transpose). We assume that for each k the sequence $\phi^k = \{\phi^k(n), n \geq 1\}$ is i.i.d., and that $\phi^1, \ldots, \phi^K$ are mutually independent, as well as independent of the arrival and service processes. The transition matrix $P = (P_{k\ell})$ is taken to be transient. That is,

$$I + P + P^2 + \ldots \quad \text{is convergent.} \tag{2.1}$$

Condition (2.1) implies that all customers eventually leave the network. Hence the systems we consider are open queueing networks, although in our examples we show that some more general networks may also be included.

For future reference, let $\alpha_k = 1/\mathsf{E}[\xi_k(1)]$ and $\mu_k = 1/\mathsf{E}[\eta_k(1)]$ be the arrival rate and service rate for class k customers, respectively. The set $\mathcal{C}_i = \{k : s(k) = i\}$ is called the *constituency* for station i. We let C denote the $d \times K$ *incidence matrix*,

$$C_{ik} = \begin{cases} 1 & \text{if } s(k) = i \\ 0 & \text{otherwise.} \end{cases}$$

In light of assumption (2.1), $(I - P')^{-1}$ exists and is equal to

$$(I - P')^{-1} = (I + P + P^2 + \ldots)'.$$

Put $\lambda = (I - P')^{-1}\alpha$. One interprets λ_k as the *effective* arrival rate to class k. For each $i = 1, \ldots, d$ we define the *nominal load* for server i per unit of time as

$$\rho_i = \sum_{k \in \mathcal{C}_i} \lambda_k / \mu_k. \tag{2.2}$$

In vector form, we have $\rho = CM^{-1}\lambda$, where $M = \operatorname{diag}(\mu_1, \ldots, \mu_K)$.

This network description is quite standard, and may be found in numerous related papers (see e.g. [29]). The network pictured in Figure 1 is a particular kind of multiclass network, called *re-entrant line* by Kumar [19]. For this re-entrant line, we can designate customers in their kth stage of processing as class k customers, $k = 1, 2, \ldots, 9$. Using the notation introduced in the preceding paragraph, we have $d = 3$, $K = 9$, $\mathcal{A} = \{1\}$, $\alpha = (1, 0, 0, 0, 0, 0, 0, 0, 0)'$,

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$P = (P_{kl})$ with $P_{k,k+1} = 1$ for $k = 1, \ldots, 8$ and all other entries zero,

$$M^{-1} = \text{diag}(0.3, 0.3, 0.1, 0.3, 0.1, 0.3, 0.3, 0.1, 0.3),$$

$\lambda = (1,1,1,1,1,1,1,1,1)'$ and $\rho = (0.9, 0.9, 0.3)'$.

To fully describe a multiclass network, we must also specify how the server chooses among the various classes at a station. A *service discipline* at station i dictates which job will be served next when server i completes a service. We assume that service disciplines are non-idling (work-conserving), which means that a server works continuously whenever there is work to be done at the station. The server may split its capacity among classes at a station, but we assume that *at most one customer in each class can receive partial service time.* Examples include first-in–first-out (FIFO), buffer priority service disciplines (both preemptive and nonpreemptive), and head-of-line processor sharing (cf. [30, Section 2.1]).

2.2 A Markovian State

Now we define a state process for the network, which depends upon the particular service discipline which is employed. For example, under any preemptive resume buffer priority service discipline, the state $X(t)$ at time t may be defined as

$$X(t) = (Q_k(t), A_\ell(t), B_k(t)) : k = 1, \ldots, K, \ell \in \mathcal{A}) \in \mathbb{R}_+^{2K+|\mathcal{A}|}, \tag{2.3}$$

where $Q_k(t)$ is the queue length for class k customers, including the one being serviced, $B_k(t)$ is the residual service time for the class k customer that is in service, which is set to be a fresh class k service time if $Q_k(t) = 0$. The residual arrival time, which is equal to the remaining time until the next customer of class k arrives, is denoted $A_k(t)$. Both $B(t)$ and $A(t)$ are taken to be right continuous. State descriptions for other service disciplines can be defined similarly: Readers are referred to Dai [30, Section 2] and Section 4 below for other examples.

We let X denote the *state space* for the state process, which is by definition equal to the set of possible values for the state $X(t)$, and we let $x = (q, a, b)$ denote a generic state in X. The first component q captures the positions of customers in the network. It can be finite dimensional as in (2.3), or infinite dimensional as is the case for the FIFO service discipline (see [30]). We use $|q|$ to denote the total queue length in the network. The remaining components of x denote the residual interarrival times and the residual service times for each class. Because we assume that at most one customer in each class can receive partial service time, we have $a \in \mathbb{R}_+^K$ and $b \in \mathbb{R}_+^K$. For the sake of concreteness, readers can assume, for example, that the state process is of the form (2.3) described for a preemptive resume buffer priority service discipline. However, the reader should bear in mind that all discussion in the paper is far more general.

For a state $x = (q, a, b) \in \mathsf{X}$, we define the "norm" of x to be $|x| = |q| + |a| + |b|$. Whether or not this is an actual norm depends on the specific form of X, however, we assume throughout the paper that the sublevel set

$$C(n) = \{x \in \mathsf{X} : |x| \leq n\}$$

is a compact subset of X for any n. This condition is satisfied automatically in virtually all practical cases.

It was shown in Dai [30, Section 2.2] that for a wide class of service disciplines, $X = \{X(t), t \geq 0\}$ is a strong Markov process. This allows us to assume at our disposal the usual elements that constitute a Markovian environment for X. Formally, it is assumed hereafter that $((\Omega, \mathcal{F}), \mathcal{F}_t, X(t), \theta_t, \mathsf{P}_x)$ is a Borel right process on the measurable state space $(\mathsf{X}, \mathcal{B}_{\mathsf{X}})$. (For a definition of right process, see Sharpe [31].) In particular, $X = \{X(t), t \geq 0\}$ has right-continuous sample paths; it is defined on $(\Omega, \mathcal{F})$ and is adapted to $\{\mathcal{F}_t, t \geq 0\}$; $\{\mathsf{P}_x, x \in \mathsf{X}\}$ are probability measures on $(\Omega, \mathcal{F})$ such that for all $x \in \mathsf{X}$,

$$\mathsf{P}_x\{X(0) = x\} = 1,$$

and

$$\mathsf{E}_x \{f(X \circ \theta_\tau) \mid \mathcal{F}_\tau\} = \mathsf{E}_{X(\tau)} f(X) \quad \text{on} \quad \{\tau < \infty\}, \quad \mathsf{P}_x\text{-a.s.}, \tag{2.4}$$

where τ is any $\mathcal{F}_t$-stopping-time,

$$(X \circ \theta_\tau)(\omega) = \{X(\tau(\omega) + t, \omega), t \geq 0\},$$

and f is any real-valued bounded measurable function (the domain of f is the space of X-valued right-continuous functions on $[0, \infty)$, equipped with the Kolmogorov σ-field generated by cylinders).

We note that the assumption that a Markovian state exists implicitly imposes some constraints on the service discipline. In particular, the time homogeneity of the process X implies that the service discipline is also time homogeneous. Such constraints are not particularly restrictive if one is willing to take the state space X sufficiently large. For instance, the state space for the FIFO service discipline is substantially larger than the state space for a priority discipline, and more complex disciplines may result in still larger state space representations.

Let $P^t(x, D)$, $D \in \mathcal{B}_{\mathsf{X}}$, $t \geq 0$, be the transition probability of X, defined as

$$P^t(x, D) = \mathsf{P}_x(X(t) \in D).$$

A nonzero measure π on $(\mathsf{X}, \mathcal{B}_{\mathsf{X}})$ is *invariant* for X if π is σ-finite, and

$$\pi(D) = \int_{\mathsf{X}} P^t(x, D)\, \pi(dx), \quad \text{for all } D \in \mathcal{B}_{\mathsf{X}},\ t \geq 0.$$

An invariant measure π is said to be unique if the only invariant measures for X are positive scalar multiples of π.

The Markov process X is called *Harris recurrent* if there exists some probability measure ν on $(\mathsf{X}, \mathcal{B}_{\mathsf{X}})$, such that whenever $\nu(D) > 0$ and $D \in \mathcal{B}_{\mathsf{X}}$,

$$\mathsf{P}_x\{\tau_D < \infty\} \equiv 1,$$

where $\tau_D = \inf\{t \geq 0 : X_t \in D\}$ (see Kaspi and Mandelbaum [32] and Meyn and Tweedie [22]). If X is Harris recurrent then a unique invariant measure π exists, see for example Getoor [33]. If the invariant measure is finite, then it may be normalized to a probability measure; in this case X is called *positive Harris recurrent*. When X is positive Harris recurrent, we say the *service discipline is stable*. In this case, we use π

to denote the stationary distribution, we let $\mathsf{P}_\pi(\cdot) = \int_{\mathsf{X}} \mathsf{P}_x(\cdot)$, and we use E_π to denote the corresponding expectation operator, so that the process $X = \{X(t), t \geq 0\}$ is a strictly stationary process under P_π.

At first sight, it appears that Harris recurrence is a difficult property to verify, as it involves the hitting time τ_D for an uncountably infinite number of sets D. A set $D \in \mathcal{B}_{\mathsf{X}}$ is called *small* if there exists $t > 0$, a probability measure ν on $\mathcal{B}_{\mathsf{X}}$, and a $\delta > 0$ such that

$$P^t(x, A) \geq \delta\nu(A), \qquad x \in D, \; A \in \mathcal{B}_{\mathsf{X}}.$$

It is shown in [22] that if $\mathsf{P}_x\{\tau_D < \infty\} \equiv 1$ for just one closed small set, then the process is Harris recurrent. In queueing network models, typically, every compact subset of X is small, and in this case there is a strong relationship between topological formulations of stability, and the measure-theoretic recurrence properties defined here.

3 Fluid models and their stability

To give a formal definition of the fluid model, we first require a particularly transparent description of the network. For each k and n, define

$$\Phi^k(n) := \sum_{i=1}^{n} \phi^k(i).$$

Assume that the initial state of the network is $x = (q, a, b) \in \mathsf{X}$. Then for each k, define

$$\begin{aligned} E_k^x(t) &:= \max\{n \geq 0 : A_k^x(0) + \xi_k(1) + \ldots + \xi_k(n-1) \leq t\}, \\ S_k^x(t) &:= \max\{n \geq 0 : B_k^x(0) + \eta_k(1) + \ldots + \eta_k(n-1) \leq t\}, \end{aligned}$$

where the maximum of an empty set is defined to be zero. It is clear that all processes in discussion, except Φ, depend on x, and we use a superscript x to explicitly denote such dependency. Let $T_k^x(t)$ be the cumulative time that server $s(k)$ has spent on class k customers in $[0, t]$. We then have

$$Q_k^x(t) = Q_k^x(0) + E_k^x(t) + \sum_{\ell=1}^{K} \Phi_k^\ell(S_\ell^x(T_\ell^x)) - S_k^x(T_k^x(t)), \text{ for } k = 1, \ldots, K, \tag{3.1}$$

$$Q^x(t) = (Q_1^x(t), \ldots, Q_K^x(t))' \geq 0, \tag{3.2}$$

$$T^x(\cdot) = (T_1^x(\cdot), \ldots, T_K^x(\cdot))' \text{ is nondecreasing}, \tag{3.3}$$

$$I_i^x(t) = t - \sum_{k \in \mathcal{C}_i} T_k^x(t) \text{ is nondecreasing}, \quad i = 1, \ldots, d, \tag{3.4}$$

$$\int_0^\infty \sum_{k \in \mathcal{C}_i} Q_k^x(t)\, dI_i^x(t) = 0, \quad i = 1, \ldots, d, \tag{3.5}$$

Additional conditions on $(Q^x(\cdot), T^x(\cdot))$ that are specific to the queueing discipline. (3.6)

Equation (3.1) expresses the fact that the queue length for class k at time t is equal to initial queue length, plus cumulative external arrivals and cumulative internal arrivals to class k by time t, minus the cumulative departures from class k by time t. Condition (3.5) is the non-idling constraint that the cumulative idle time $I_i^x(\cdot)$ at station i

does not increase when the queue length at station i is positive. Examples are given below to illustrate the final condition (3.6). Readers are referred to Harrison [26] for more discussion on the system equations describing the dynamics of discrete queueing networks.

We now scale space and time to reduce the apparent fluctuation of the model. Consider the process

$$\bar{Q}^x(t) = \frac{1}{|x|}Q^x(|x|t), \tag{3.7}$$

where as usual $x \in \mathsf{X}$ is the initial condition. For large $|x|$, we will see that the normalized queue length process $\bar{Q}^x$ is approximated by a solution $\bar{Q}(t)$ of a set of integral equations. The scaling appeared in (3.7) is often called the *fluid scaling*, and any limit $\bar{Q}(t)$ is called a *fluid limit* of the queue length process.

Letting $|q| \to \infty$ while keeping the remaining components of the initial condition x fixed, any limit point of the normalized queue length process $\bar{Q}^x$ is a solution of the following fluid model (see Dai [30] for the precise procedure):

Definition 3.1 A (undelayed) fluid limit for a network under a specific service discipline is defined to be any solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ to the following equations, where $\bar{Q}(t) = (\bar{Q}_1(t), \dots, \bar{Q}_K(t))'$ and $\bar{T}(t) = (\bar{T}_1(t), \dots, \bar{T}_K(t))'$.

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \alpha_k t - \mu_k \bar{T}_k(t) + \sum_{\ell=1}^{K} P_{\ell k}\mu_\ell \bar{T}_\ell(t) \text{ for } k = 1, \dots, K, \tag{3.8}$$

$$\bar{Q}_k(t) \geq 0 \text{ for } k = 1, \dots, K, \tag{3.9}$$

$$\bar{T}_k(0) = 0 \text{ and } \bar{T}_k(\cdot) \text{ is nondecreasing for } k = 1, \dots, K, \tag{3.10}$$

$$\bar{I}_i(t) = t - \sum_{k \in C_i} \bar{T}_k(t) \text{ is nondecreasing for } i = 1, \dots, d, \tag{3.11}$$

$$\bar{I}_i(\cdot) \text{ increases at times } t \text{ when } \textstyle\sum_{k \in C_i} \bar{Q}_k(t) = 0 \text{ for } i = 1, \dots, d, \tag{3.12}$$

$$\text{Additional conditions on } (\bar{Q}(\cdot), \bar{T}(\cdot)) \text{ that are specific to the queueing discipline.} \tag{3.13}$$

The set of equations (3.8)–(3.13) is called the *fluid model*, and we let $\mathcal{Q}$ denote the collection of all solutions $(\bar{Q}(\cdot), \bar{T}(\cdot))$ of the fluid model.

In vector form, the fluid model takes the form

$$\begin{aligned}
&\bar{Q}(t) = \bar{Q}(0) + \alpha t - (I - P')M\bar{T}(t) \\
&\bar{Q}(t) \geq 0, \\
&\bar{T}(0) = 0 \text{ and } \bar{T}(\cdot) \text{ is nondecreasing}, \\
&\bar{I}(t) = te - C\bar{T}(t) \text{ is nondecreasing}, \\
&\int_0^\infty C\bar{Q}(t)\, d\bar{I}(t) = 0, \\
&\text{Additional conditions on } (\bar{Q}(\cdot), \bar{T}(\cdot)) \text{ that are specific to} \\
&\text{the queueing discipline.}
\end{aligned}$$

In general, (3.8)–(3.13) may not uniquely determine $(\bar{Q}(\cdot), \bar{T}(\cdot))$ because the queueing network may be sensitive to its initial configuration. That is, a slight change of the initial network configuration, negligible under fluid scaling, may completely change the subsequent behavior of the network (see for example cases (D1) and (D2) in Table 1).

If we let $|x| \to \infty$ without constraining any components of $x = (q, u, v)$, then we also obtain a fluid model, but in this case the residual arrival and service processes introduce a delay:

Definition 3.2 The *delayed* fluid model of a service discipline in a network with delay $(\bar{A}(0), \bar{B}(0)) \in \mathbb{R}_+^{K+|\mathcal{A}|}$ is defined to be the equations (3.9)–(3.13), together with

$$\bar{Q}(t) = \bar{Q}(0) + (\alpha t - \bar{A}(0))^+ - (I - P')M(\bar{T}(t) - \bar{B}(0))^+, \tag{3.14}$$

where for a $y \in \mathbb{R}$, $y^+ = (y + |y|)/2$.

The delay is important in subsequent analysis, but is largely irrelevant when addressing stability of the network.

For example, for a $G/G/1$ queue the (undelayed) fluid model is succinctly described by the differential equation

$$\frac{d}{dt}\bar{Q}(t) = (\lambda - \mu)\mathbb{1}(\bar{Q}(t) > 0), \qquad t \geq 0.$$

Condition (3.13) will be derived here for a buffer priority service discipline, and other examples will be considered in Section 4. Under a buffer priority service discipline, one can envision that customers in class k wait in their own buffer. Customers in distinct buffers have different service priorities, and we assume that there are no ties among classes. Within each buffer, customers are served in FIFO discipline. Let H_k denote the set of indices for all classes served at station $s(k)$ which have priority greater than or equal to that of class k, and let

$$\begin{aligned} \bar{T}_k^+(t) &= \sum_{\ell \in H_k} \bar{T}_\ell(t) \\ \bar{I}_k^+(t) &= t - \bar{T}_k^+(t), \\ \bar{Q}_k^+(t) &= \sum_{\ell \in H_k} \bar{Q}_\ell(t). \end{aligned}$$

Then $\bar{T}_k^+(t)$ is the cumulative amount of service in $[0, t]$ dedicated to customers whose classes are included in H_k, and $\bar{I}_k^+(t)$ is the total unused capacity that is available to serve customers whose class does not belong to H_k. Note that $\bar{I}_i(t)$ is a station level quantity representing the total unused capacity in $[0, t]$ by server i; whereas $\bar{I}_k^+(t)$ is a class level quantity. The priority service discipline requires that for every k, all the service capacity of station $s(k)$ is dedicated to classes in H_k, as long as the workload present in these buffers is positive. Thus we may express the additional condition (3.13) by the integral equation

$$\int_0^\infty \bar{Q}_k^+(t)\, d\bar{I}_k^+(t) = 0, \qquad 1 \leq k \leq K. \tag{3.15}$$

Although the constraint (3.15) appears to be almost obvious for a buffer priority discipline, in general one must check carefully that the condition (3.13) for the fluid limit model can indeed be obtained from (3.6) by the limiting procedure. For example, when the preemptive resume buffer priority service discipline is applied, condition (3.6) takes the form

$$\int_0^\infty Q_k^{x,+}(t)\, d(t - T_k^{x,+}(t)) = 0, \quad k = 1, \ldots, K, \tag{3.16}$$

where

$$Q_k^{x,+}(t) = \sum_{\ell \in H_k} Q_\ell^x(t) \quad \text{and} \quad T_k^{x,+}(t) = \sum_{\ell \in H_k} T_\ell^x(t).$$

By using Lemma 2.4 of Dai and Williams [34], one can show that upon taking limits, the identity (3.16) does indeed become (3.15).

Now we present a recent formulation of stability for the fluid model. In later sections, this stability property for the fluid model is shown to imply several probabilistic forms of stability for the network.

Definition 3.3 The fluid model is *stable* if there exists a fixed time t_0 such that $\bar{Q}(t) = 0$, $t \geq t_0$, for any $\bar{Q}(\cdot) \in \mathcal{Q}$ satisfying $|\bar{Q}(0)| = 1$.

We conclude this section with the following result of Chen [35, Theorem 5.3] that will be useful below.

Lemma 3.1 *If the fluid model defined by (3.8)–(3.13) is stable, then the delayed fluid model is also stable. That is, there exists $t_0 > 0$ such that $\bar{Q}(t) = 0$ for $t \geq t_0$, for any solution to the delayed fluid model whose initial condition $\bar{x}$ satisfies the bound $|\bar{x}| = |\bar{Q}(0)| + |\bar{A}(0)| + |\bar{B}(0)| \leq 1$.*

Now that stability is defined for the fluid model, we describe how these properties are reflected in the associated network.

4 Main Results and Examples

We describe here our main results, and give several examples. First we list three assumptions on the network:

(A1) $\xi_1, \ldots, \xi_K, \eta_1, \ldots, \eta_K$ are mutually independent, and i.i.d. sequences.

(A2) For some integer $p \geq 1$,

$$\mathsf{E}[\xi_\ell(1)^{p+1}] < \infty \text{ for } \ell \in \mathcal{A} \quad \text{and} \quad \mathsf{E}[\eta_k(1)^{p+1}] < \infty \text{ for } k = 1, \ldots, K,$$

(A3) The set $\{x \in \mathsf{X} : |x| = 0\}$ is a singleton, and for each $k \in \mathcal{A}$, there exists some positive function $q_k(x)$ on $\mathbb{R}_+$, and some integer j_k, such that

$$\mathsf{P}(\xi_k(1) \geq x) \quad > \quad 0 \quad \text{for all} \quad x > 0. \tag{4.1}$$

$$\mathsf{P}(\xi_k(1) + \cdots + \xi_k(j_k) \in dx) \quad \geq \quad q_k(x)\, dx \quad \text{and} \quad \int_0^\infty q_k(x)\, dx > 0. \tag{4.2}$$

Conditions (A1) and (A2) are quite standard, although the independence assumption (A1) can be relaxed: see the remark after Proposition 2.1 of Dai [30].

Condition (A3) is not needed for bounding moments, but is required to establish ergodicity of the network. Under this condition, the argument used in Lemma 3.4 of Meyn and Down [17] may be applied to deduce that all compact subsets of X are small. Frequently, milder conditions can be invoked to obtain this property for the network, and since this is the only reason that (A3) is introduced, we list here the following generalization:

(A3') For the Markov process X, every compact subset of X is small.

For example, for a G/G/1 queue with $\rho < 1$, Condition (A3') is satisfied if and only if the spread-out condition (4.2) holds (see Meyn and Tweedie [36]). Sigman [16] describes more general examples in which the unboundedness condition (4.1) is superfluous.

We may now highlight the main results of this paper.

Theorem 4.1 *Assume that the fluid model for a service discipline is stable, and that (A1) and (A2) hold. Then*

(i) *For some constant κ_p, and for each initial condition $x \in \mathsf{X}$,*

$$\lim_{t\to\infty} \sup \frac{1}{t} \int_0^t \mathsf{E}_x[|Q(t)|^p]\, ds \le \kappa_p,$$

where p is the integer used in (A2).

Assume further that (A3) or (A3') holds. Then, the service discipline is stable, and moreover, for each initial condition,

(ii) *The transient moments converge to their steady state values:*

$$\lim_{t\to\infty} \mathsf{E}_x[Q_k(t)^r] = \mathsf{E}_\pi[Q_k(0)^r] \le \kappa_r, \quad \textit{for } r = 1,\dots,p,\ k = 1,\dots,K.$$

(iii) *The first moment converges at rate t^{p-1}:*

$$\lim_{t\to\infty} t^{(p-1)} \left|\mathsf{E}_x[Q(t)] - \mathsf{E}_\pi[Q(0)]\right| = 0$$

(iv) *The strong law of large numbers holds:*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t Q_k^r(s)\, ds = \mathsf{E}_\pi[Q_k(0)^r], \quad \mathsf{P}_x\textit{-a.s.}, \quad \textit{for } r = 1,\dots,p,\ k = 1,\dots,K.$$

Proof The proof of these results, and several related results, follow from Theorems 5.5, 6.2, 6.3, and 6.4. □

To show how this result is applied in practice, and how it may be used to strengthen previous results in the networks area, we now consider several examples from the operations research and computer networks literature.

4.1 Generalized Jackson Networks

A generalized Jackson network is an example of an open queueing network in which only one customer class is serviced at a given station, so that $|C_k| = 1$ for all k. It is shown in Dai [30] that the associated model is stable, and hence all of the conclusions of Theorem 4.1 apply.

This generalizes Theorem 3.4 of Meyn and Down [17] where a similar conclusion is reached in the special case where $p = 2$, and stronger conditions are imposed upon the service processes. In the case where the interarrival times possess geometrically decaying tails, it is shown in [17] that the state process is geometrically ergodic. We do not know if the same conclusion can be reached by considering the fluid model.

The time averaged bound of Theorem 4.1 (i) is also obtained in Theorem 2.2 of [17] under very mild conditions on the arrival process. It is likely that this generality can also be inferred from the fluid model since the methods of Section 5 do not depend on the Markov property as much as as they do on properties of the conditional expectation.

4.2 Re-entrant Lines

Re-entrant lines are also a subclass of the models considered here, in which routing is deterministic: customers arrive to buffer one, where they wait in queue until service. After a service is completed, a customer moves on to buffer two, and so forth, until it finally reaches buffer K. After service is completed at this final queue, the customer leaves the network. Hence the routing probability is of the form $P_{k,k+1} = 1$ for all k, and it follows that this is also an open network.

Stability for such models depends crucially on the service discipline chosen. For the FBFS (first buffer-first served) service discipline it is shown in [27, 30] that the fluid model is stable. Also, the LBFS (last buffer-first served) discipline is treated in Kumar and Kumar [20] and Dai and Weiss [6], where again stability is demonstrated for general network topologies, whenever the nominal load is less than unity at each station. Theorem 4.1 thus shows that these models have bounded pth moments in steady state, under Assumptions (A1)–(A3).

4.3 Polling Models

So far, we have focused primarily on models from manufacturing applications. The methods developed here are of course also applicable to communication networks. We consider here one such model which has recently received a great deal of attention in the computer networks literature.

Consider an ℓ-limited token ring, as described in Altman et. al. [12, 13], Georgiadis and Szpankowski [14], or Kuehn [37]; see also Fricker and Jaïbi [38] and Borovkov and Shassburger [39]. The token passing ring is a single server station populated with K classes of customers. Each class has its own buffer at the station. Customers arrive to the kth buffer at rate α_k, and are serviced at rate μ_k when a service is in progress. After the service, a customer leaves the network. A token must be possessed by buffer k before the server can initiate a service session for class k customers. Once a service session starts for buffer k, the server continues to serve class k customers until ℓ_k class k customers depart or until the queue empties, whichever event occurs first. At the end of this service session, the token in the token passing ring travels from buffer k, to $k+1$ (mod K) with a switch-over time of $\eta^0_k(n)$ on its nth such transition. We refer the reader to Georgiadis and Szpankowski [14] or Takagi [40] for further details.

The polling model differs from the standard multiclass queueing network defined in Section 2 because extra switch over times for the token are required. During the time period that the token is in the process of switching to a new buffer, the server is idling, although there may be non-empty buffers in the network. This token interference certainly reduces the service capacity of the network.

We assume that the switch-over time sequence $\{(\eta^0_1(n), \ldots, \eta^0_K(n)), n \geq 1\}$ is i.i.d., and is independent of the arrival sequences and service sequences in (A1). Let the switch-over rate be $\mu_k^0 = \mathsf{E}[\eta_k^0(1)]^{-1} > 0$, $1 \leq k \leq K$. This network may be modeled as a continuous time-continuous state space Markov process

$$X(t)^\top = (Q_k(t), A_\ell(t), B_k(t), B_k^0(t), C(t)) : k = 1, \dots, K, \ell \in \mathcal{A}),$$

where $Q_k(t)$, $B_k(t)$, and $A_k(t)$ are defined as in (2.3), $B_k^0(t)$ is the remaining switch-over time between class k and class $k+1$ (mod K), set to be a fresh switch-over time if such a switch-over is not occurring at time t, and $C(t)$ indicates the number of services which have been started and/or completed during the current session of an active buffer. This variable is set to zero if a switch-over is in progress, when no buffer is active.

The crucial parameters of the network are defined as follows. The nominal load at queue k is $\beta_k = \alpha_k/\mu_k$, and the total load is $\rho_0 = \sum \beta_k$. The total mean switch-over time in a token cycle is defined to be

$$u^0 = \sum_{k=1}^{K} \mathsf{E}[\eta_k^0(1)] = \sum_{k=1}^{K} \frac{1}{\mu_k^0} \tag{4.3}$$

The queue length process $Q_k^x(t)$ and the cumulative service allocation process $T_k^x(t)$ for buffer k and for initial state x are defined as before. Let $T_k^{x,0}(t)$ be the cumulative time by time t that the token spends in switching from buffer k to $k+1$ (mod K). Suppose that the function $(\bar{Q}(\,\cdot\,), \bar{T}(\,\cdot\,), \bar{T}^0(\,\cdot\,))$ is a limit point of

$$\left(\frac{1}{|x|}Q^x(|x|t), \frac{1}{|x|}T^x(|x|t), \frac{1}{|x|}T^{x,0}(|x|t)\right) \tag{4.4}$$

when $|x| \to \infty$. Then $(\bar{Q}(t), \bar{T}(t), \bar{T}^0(t))$ is a (delayed) fluid limit of the token ring.

We show here that the delayed fluid model is stable under the load condition of [14]. Similar to [35], it is enough to prove that the undelayed fluid model is stable. This model is obtained by letting $q \to \infty$ in (4.4), while keeping the remaining components of the initial state x fixed. We summarize some important features of the undelayed fluid model in the following proposition:

Proposition 4.2 *Let $(\bar{Q}, \bar{T}, \bar{T}^0)$ be any fluid limit of (4.4), and assume that as $x \to \infty$ along some subsequence,*

$$\left(\frac{1}{|x|}Q_k^x(0), \frac{1}{|x|}A_k^x(0), \frac{1}{|x|}B_k^x(0), \frac{1}{|x|}B^{x,0}(0)\right) \to (\bar{Q}_k(0), 0, 0, 0),$$

$1 \leq k \leq K$. The fluid limit then has the following properties, where properties of a derivative hold whenever the derivative exists.

(i) *The busy time vectors $\bar{T}(t)$ and $\bar{T}^0(t)$ are increasing and continuous with $\bar{T}(0) = \bar{T}^0(0) = 0$;*

(ii) *For all $t \geq 0$,*

$$\sum_{k=1}^{K}[\bar{T}_k(t) + \bar{T}_k^0(t)] = t.$$

(iii) *For all $1 \leq k \leq K$,*

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \alpha_k t - \mu_k \bar{T}_k(t).$$

(iv) *For all* $1 \leq k \leq K$,

$$\dot{\bar{T}}_k(t) = \beta_k, \qquad \textit{whenever } \bar{Q}_k(t) = 0.$$

(v) *For all* k, j,

$$\mu_k^0 \bar{T}_k^0(t) = \mu_j^0 \bar{T}_j^0(t).$$

(vi) *For all* $1 \leq k \leq K$,

$$\mu_k \dot{\bar{T}}_k(t) = \ell_k \mu_k^0 \dot{\bar{T}}_k^0(t), \qquad \textit{whenever } \bar{Q}_k(t) > 0.$$

PROOF Let $(\bar{Q}, \bar{T}, \bar{T}^0)$ be a limit point of (4.4) as in Dai [30]. The results (i) and (ii) follow since the busy times for the network have these properties. Property (iii) is a special case of (3.14). When $\bar{Q}_k(t) = 0$ then $\dot{\bar{Q}}_k(t) = 0$ whenever the derivative exists, by positivity of $\bar{Q}_k(t)$. Hence in this case, from (iii),

$$0 = \dot{\bar{Q}}_k(t) = -\mu_k \dot{\bar{T}}_k(t) + \alpha_k,$$

which is (iv).

Now we check that (v) holds. Let $Q_k^{x,0}(t) = 1$ if the token is in transition from buffer k to $k+1$ (mod K) at time t and zero otherwise. Let

$$S_k^{x,0}(t) = \max\{n \geq 0 : B_k^0(0) + \eta_k^0(1) + \ldots + \eta_k^0(n-1) \leq t\}.$$

Then $S_k^{x,0}(T_k^{x,0}(t))$ is the number of transitions from k to $k+1$ (mod K) finished in $[0, t]$. It follows from the definitions that

$$Q_k^{x,0}(t) = Q_k^{x,0}(0) + S_{k-1}^{x,0}(T_{k-1}^{x,0}(t)) - S_k^{x,0}(T_k^{x,0}(t)).$$

Because $0 \leq Q_k^{x,0}(t) \leq 1$ for all t, and

$$\frac{1}{|x|} S_k^{x,0}(|x|t) \to \mu_k^0 t$$

uniformly on compact sets almost surely, we have that for all k (mod K),

$$\mu_k^0 \bar{T}_k^0(t) = \mu_{k+1}^0 \bar{T}_{k+1}^0(t).$$

It remains to prove that property (vi) holds. This is an application of the law of large numbers for i.i.d. sequences; the assumption that as long as there are at least ℓ_k customers at buffer k, exactly ℓ_k services occur; and the fact that one switch-over is completed in each token cycle. Now assume that $\bar{Q}_k(t) > 0$. By the Lipschitz continuity of $\bar{Q}(t)$, there is a small interval $[t, t+h]$ such that

$$\min_{t \leq s \leq t+h} \bar{Q}_k(s) > 0.$$

Then there is a subsequence $\{Q_k^{x_n}(|x_n|s)/|x_n|, n \geq 1\}$ such that

$$Q_k^{x_n}(|x_n|s)/|x_n| \to \bar{Q}_k(s)$$

uniformly on $[t, t+h]$ as $n \to \infty$. Therefore $Q_k^{x_n}(|x_n|s) \geq \ell_k$ for all large n and $s \in [t, t+h]$. Thus in each complete token cycle within $[t, t+h]$, exactly ℓ_k services occur at buffer k. Also, the number of token cycles within $[t, t+h]$ differs from $S_k^{x,0}(T_k^{x,0}(s)) - S_k^{x,0}(T_k^{x,0}(t))$ by at most one. It follows that $\ell_k[S_k^{x,0}(T_k^{x,0}(s)) - S_k^{x,0}(T_k^{x,0}(t))]$ differs from $S_k^x(T_k^x(s)) - S_k^x(T_k^x(t))$ by at most ℓ_k customer for all $s \in [t, t+h]$. Therefore, $\mu_k(\bar{T}_k(s) - \bar{T}_k(t)) = \ell_k \mu_k^0(\bar{T}_k^0(s) - \bar{T}_k^0(t))$, and hence $\mu_k \dot{\bar{T}}_k(t) = \ell_k \mu_k^0 \dot{\bar{T}}_k^0(t)$. □

To prove that the undelayed fluid model is stable, consider the *work in the system*, defined as

$$W(t) = \sum_{k=1}^{K} \frac{1}{\mu_k} \bar{Q}_k(t).$$

We have from (iii) that

$$W(t) = W(0) + \rho_0 t - \sum_{k=1}^{K} \bar{T}_k(t). \tag{4.5}$$

We now rewrite this by substituting an expression for the total switch-over time. Let $N(t)$ denote the common value of $\mu_j^0 \bar{T}_j^0(t)$, which is well defined by (vi). Then by (ii) and (4.3) we have that

$$\sum_{k=1}^{K} \bar{T}_k(t) = t - u^0 N(t), \tag{4.6}$$

and substituting this into (4.5) gives

$$W(t) = u^0 N(t) - (1 - \rho_0) t. \tag{4.7}$$

This shows that obtaining stability amounts to bounding the total switch over time $u^0 N(t)$.

For a fixed time t, let J denote those indices for which $\bar{Q}_i(t) > 0$. Then we have from (4.6), and Proposition 4.2 (iv) and (v), whenever the derivatives exist,

$$\begin{aligned} 1 - u^0 \dot{N}(t) &= \sum_{k \in J} \dot{\bar{T}}_k(t) + \sum_{k \in J^c} \dot{\bar{T}}_k(t) \\ &= \sum_{k \in J} \frac{\ell_k}{\mu_k} \mu_k^0 \dot{\bar{T}}_k^0(t) + \sum_{k \in J^c} \beta_k \end{aligned}$$

Substituting the definition of N then gives

$$[u^0 + \sum_{k \in J} \frac{\ell_k}{\mu_k}] \dot{N}(t) = [1 - \sum_{k \in J^c} \beta_k],$$

which together with (4.7) and the definition of ρ_0 shows that

$$\dot{W}(t) = \left\{ \frac{u^0[(1 - \rho_0) + \sum_{k \in J} \beta_k]}{u^0 + \sum_{k \in J} \frac{\ell_k}{\mu_k}} - (1 - \rho_0) \right\}.$$

After rearranging terms, this becomes

$$\dot{W}(t) = \left\{ \frac{\sum_{k \in J}[u^0 \beta_k - (1 - \rho_0) \frac{\ell_k}{\mu_k}]}{u^0 + \sum_{k \in J} \frac{\ell_k}{\mu_k}} \right\}.$$

and this is negative for any $\bar{Q}(t) \neq 0$ if and only if the following condition holds:

$$\beta_k < (1 - \rho_0)\frac{\ell_k}{\mu_k u^0}, \qquad 1 \le k \le K. \tag{4.8}$$

It follows from Lemma 2.2 of Dai and Weiss [6] that the fluid model is stable whenever (4.8) holds.

This together with Theorem 4.1 gives the following stability result for the network:

Theorem 4.3 *If the load condition (4.8) holds, then the pth moment of the queue lengths in the token ring are bounded as in Theorem 4.1 (i). If in addition (A3') holds, then the remaining conclusions of Theorem 4.1 follow.*

A converse to this result is supplied in [14], together with a related stability result in the special case where the arrival stream is Poisson. We note that a direct stability proof for this network is extremely difficult. See the aforementioned paper, or the two buffer analysis of Boxma and Groenendijk [41].

4.4 Modeling Machine Failures

Machine failures can be modeled by making a mild generalization of the general framework of Section 2. Consider for simplicity a single station system without reentry with a failure-prone server (machine). The failure is assumed to be *autonomous* as opposed to *operational*. See Harrison and Pich [42] and the references therein for further discussion. To model failures, we assume that there are two buffers at the station. Buffer one is an actual buffer in which customers await service. Buffer two models breakdowns: when buffer two is non-empty, a failure is in progress.

To give a complete picture it is necessary to introduce a second station with associated buffer three, so that the network takes on the form given in Figure 3. There is a single customer which travels back and forth between buffer 2 and buffer 3. When this fictitious customer is at station 1, a failure is in progress, and when this customer is at station 2 the server at station one is operating normally. We assume that the service discipline at station 1, the true machine, gives strict priority to buffer 2, and is preemptive.

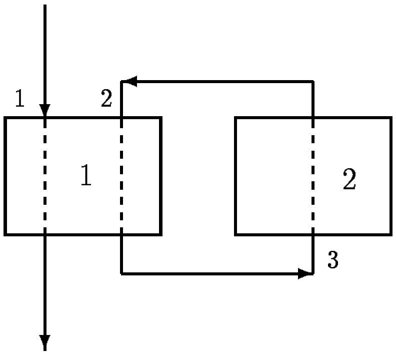

Figure 3. An open-closed mixed network modeling a machine with random failures. When the fictitious customer is at station 1, a failure is in progress, and when this customer is at station 2 the server at station one is operating normally.

Strictly speaking, this network falls outside of the framework described in Section 2 since it is a model which is neither open nor closed. However, the analysis in Section 4.3 can be carried over, the fluid model has a simple form, and the main results of this paper still apply. It is a mixed network as considered in Nguyen [43], but the fluid model considered here is different from that of [43] since in the example considered here, the total number of closed customers is fixed to be one. When all distributions are exponential, the stability of such systems was also studied by Ingenoso [44].

Because the analysis is analogous to that given in Section 4.3, we present here the (undelayed) fluid model for this example without proof. As always, the queue length process depends upon the busy time, service rates, and arrival rates. In this case it is of a particularly simple form due to the simplicity of the model:

$$\bar{Q}_1(t) = \bar{Q}_1(0) - \mu_1\bar{T}_1(t) + \lambda_1 t. \tag{4.9}$$

By the feedback nature of the failure process, the fictitious customer is either at buffer 2 or at buffer 3. Hence, we have $\bar{T}_2(t) + \bar{T}_3(t) = t$ for all t. Similar to the derivation in Section 4.3, we also have

$$\mu_2\bar{T}_2(t) = \mu_3\bar{T}_3(t),$$

and combining these two identities gives

$$\dot{\bar{T}}_2(t) = \frac{1/\mu_2}{1/\mu_2 + 1/\mu_3}. \tag{4.10}$$

Finally, we have the non-idling constraint:

$$\dot{\bar{T}}_1(t) + \dot{\bar{T}}_2(t) = 1 \qquad \text{when } \bar{Q}_1(t) > 0. \tag{4.11}$$

Combining (4.11) and (4.10) gives

$$\dot{\bar{T}}_1(t) = \frac{\mu_2}{\mu_2 + \mu_3} \qquad \text{when } \bar{Q}_1(t) > 0,$$

and substituting this into (4.9) gives, whenever the first buffer is non-empty,

$$\dot{\bar{Q}}_1(t) = \left[-\frac{\mu_1\mu_2}{\mu_2 + \mu_3} + \lambda \right]$$

Hence it follows from Lemma 2.2 of Dai and Weiss [6] that the fluid model is stable if the following load condition is satisfied:

$$\beta_1 = \frac{\lambda_1}{\mu_1} < \frac{\mu_2}{\mu_2 + \mu_3}.$$

Under this condition, it follows that the pth moment bounds obtained in Theorem 4.1 (i) hold for this model. Again, if in addition Condition (A3') may be verified, then each of the ergodic limits of Theorem 4.1 also hold.

5 Moments

The main result of this section shows how one obtains moments on the queue lengths under a stability condition on the fluid model. Because we only assume that ξ_k and

η_k are i.i.d., the model is not necessarily ψ-irreducible, and hence it is impossible to assert that the state process is Harris recurrent in this case, and this rules out many ergodic theorems [36]. However, we still find that the queue lengths are bounded in an L_p sense.

This section is divided into three parts. First we show that the network exhibits a contractive property if the fluid model is stable. Next, it is shown that such a contractive property implies strong bounds on the mean return time to a compact set. Finally, these results are used in Theorem 5.5 to show that the pth moment of the queue lengths is bounded on average under Assumptions (A1)–(A2).

5.1 A contraction property for the network

The main result given here is essentially the first step which allows us to connect stability of the fluid model with stability for $X = \{X(t), t \geq 0\}$. The following result asserts that for large x, after a time period which is proportional to the magnitude of the initial condition $X(0) = x$, the L_p norm of the queue lengths will be small when compared to their initial values.

Proposition 5.1 *Suppose that Assumptions (A1) and (A2) hold, and that the fluid model is stable. Then there exists $t_0 > 0$ such that*

$$\lim_{|x|\to\infty} \frac{1}{|x|^{p+1}} \mathsf{E}_x\Big[|\, X(t_0|x|)\, |^{p+1} \Big] = 0. \tag{5.1}$$

A large part of the proof of Proposition 5.1 is based upon the following.

Lemma 5.2 ([45], Theorem 5.1) *Let $\{\zeta(k) : k \in \mathbb{Z}_+\}$ be an i.i.d. sequence taking values in $(0,\infty)$, and let $E(t)$ denote the counting process $E(t) = \max(n \geq 1 : \zeta(1) + \cdots + \zeta(n-1) \leq t)$. If $\mathsf{E}[\zeta(1)] < \infty$, then for any integer $r \geq 1$,*

$$\lim_{t\to\infty} \mathsf{E}\Big[\Big(\frac{E(t)}{t}\Big)^r\Big] = \Big(\frac{1}{\mathsf{E}[\zeta_1]}\Big)^r.$$

Hence, under these conditions,

(a) *for any $\delta > 0$, $\sup_{t\geq\delta} \mathsf{E}\left[(E(t)/t)^r\right] < \infty$.*

(b) *The random variables*

$$\{(E(t)/t)^r : t \geq 1\}$$

are uniformly integrable.

□

Proof of Proposition 5.1. Assume that Assumptions (A1) and (A2) hold, and that the fluid model specified in (3.8)–(3.13) is stable. By Lemma 3.1, the corresponding delayed fluid model is stable. It then follows from Dai [30, Section 4] that there exists $t_0 \geq 1$ such that

$$\lim_{|x|\to\infty} \frac{1}{|x|} |Q^x(t_0|x|)| = 0 \quad \text{in probability.}$$

Because

$$\begin{aligned}\frac{1}{|x|}\left|Q^x(|x|t_0)\right| &\le 1+\sum_{k\in\mathcal{A}}\frac{1}{|x|}\left|E_k^x(|x|t_0)\right| \\ &\le 1+\sum_{k\in\mathcal{A}}\frac{1}{|x|}\left|E_k^0(|x|t_0)\right|,\end{aligned}$$

It follows from Lemma 5.2 that the collection of random variables

$$\left\{\frac{1}{|x|^{p+1}}\left|Q^x(t_0|x|)\right|^{p+1} : |x|\ge 1\right\}$$

is uniformly integrable, and hence

$$\lim_{|x|\to\infty}\frac{1}{|x|^{p+1}}\mathsf{E}_x\left|Q^x(t_0|x|)\right|^{p+1}=0.$$

It remains to show

$$\lim_{|x|\to\infty}\frac{1}{|x|^{p+1}}\mathsf{E}_x\left|A(t_0|x|)\right|^{p+1}=0, \tag{5.2}$$

$$\lim_{|x|\to\infty}\frac{1}{|x|^{p+1}}\mathsf{E}_x\left|B(t_0|x|)\right|^{p+1}=0. \tag{5.3}$$

For $x=(q,a,b)$, notice that because $t_0|x|\ge a_k$,

$$\begin{aligned}\frac{1}{|x|^{p+1}}\left|A_k^x(t_0|x|)\right|^{p+1} &\le \frac{1}{|x|^{p+1}}(\xi_k(E_k^x(|x|t_0)+1))^{p+1} \\ &\le \frac{1}{|x|^{p+1}}\sum_{i=1}^{E_k^x(|x|t_0)+1}(\xi_k(i))^{p+1} \\ &\le \frac{1}{|x|^{p+1}}\sum_{i=1}^{E_k^0(|x|t_0)+1}(\xi_k(i))^{p+1}.\end{aligned}$$

By Walds' identity,

$$\begin{aligned}&\frac{1}{|x|^{p+1}}\mathsf{E}\left[\sum_{i=1}^{E_k^0(|x|t_0)+1}(\xi_k(i))^{p+1}\right] \\ &= \frac{1}{|x|^{p+1}}\mathsf{E}\left[E_k^0(|x|t_0)+1\right]\mathsf{E}\left[(\xi_k(1))^{p+1}\right],\end{aligned}$$

which converges to zero as $|x|\to\infty$ by part (b) of Lemma 5.2 and the fact that $p>0$. Therefore

$$\lim_{|x|\to\infty}\frac{1}{|x|^{p+1}}\mathsf{E}_x\left|A_k^x(t_0|x|)\right|^{p+1}=0$$

and thus (5.2) holds. Similarly, we can show that (5.3) holds, and this proves the proposition. □

5.2 Bounds on mean return times

We now consider several consequences of Proposition 5.1, all of which concern the return time

$$\tau_C(\delta) = \min(t \geq \delta : X(t) \in C).$$

From now on the symbol C will be used to denote a subset of the state space X, instead of the incidence matrix defined in Section 2, and the symbol b will be used to denote a generic positive constant, instead of the residual service times.

Proposition 5.3 *Let X be the state process for the network, and suppose that Assumption (A1) and (A2) are satisfied. Then for some constant $c_{p+1} < \infty$, $\delta > 0$, and a compact set $C \subset \mathsf{X}$,*

$$\mathsf{E}_x\Big[\int_0^{\tau_C(\delta)} (1 + |X(t)|^p)\, dt\Big] \leq c_{p+1}(|x|^{p+1} + 1), \quad x \in \mathsf{X}. \tag{5.4}$$

PROOF Under the conditions of the proposition, it follows from Proposition 5.1 that there exists a compact set of the form $C = \{x : |x| \leq L\}$ such that for $x \in C^c$,

$$\mathsf{E}_x[|X(t_0|x|)|^{p+1}] \leq \frac{1}{2}|x|^{p+1}.$$

On letting $t(x) = t_0 \max(L, |x|)$, this bound may be written

$$\int P^{t(x)}(x, dy)|y|^{p+1} \leq \frac{1}{2}|x|^{p+1} + b\mathbb{1}_C(x), \quad x \in \mathsf{X}, \tag{5.5}$$

where b is a finite constant.

Define as in the proof of Theorem 2.1 (b) of [36], the sequence of stopping times $\sigma_0 = 0$, $\sigma_1 = t(x)$, and $\sigma_{k+1} = \sigma_k + \theta_{\sigma_k}\sigma_1$, $k \geq 1$, where θ is the shift operator on the sample space. The stochastic process $\hat{X}_k := X(\sigma_k)$, $k \geq 0$, is a Markov chain with transition kernel

$$\hat{P}(x, A) = \mathsf{P}_x\{X(t(x)) \in A\}, \qquad x \in \mathsf{X}, A \in \mathcal{B}_{\mathsf{X}},$$

and the bound (5.5) may be expressed

$$\int_{\mathsf{X}} \hat{P}(x, dy)U_{p+1}(y) \leq U_{p+1}(x) - \frac{1}{2}|x|^{p+1} + b\mathbb{1}_C(x)$$

with $U_{p+1}(x) = |x|^{p+1}$. From the Comparison Theorem [36, p. 337] we then have

$$\mathsf{E}_x\Big[\sum_{k=0}^{k_*-1} |X(\sigma_k)|^{p+1}\Big] = \mathsf{E}_x\Big[\sum_{k=0}^{k_*-1} |\hat{X}_k|^{p+1}\Big] \leq 2\Big\{|x|^{p+1} + b\mathbb{1}_C(x)\Big\}, \quad x \in \mathsf{X}, \tag{5.6}$$

where $k_* = \min(k \geq 1 : \hat{X} \in C)$.

To prove the proposition we first show that for some constant c_0,

$$\mathsf{E}_x\Big[\int_{\sigma_k}^{\sigma_{k+1}} (1 + |X(t)|^p)\, dt \mid \mathcal{F}_{\sigma_k}\Big] \leq c_0\left(|X(\sigma_k)|^{p+1} + 1\right), \qquad k \geq 0,\ x \in \mathsf{X}, \tag{5.7}$$

which by the strong Markov property amounts to

$$\mathsf{E}_x\Big[\int_0^{\sigma_1} (1 + |X(t)|^p)\, dt\Big] \leq c_0(|x|^{p+1} + 1), \qquad x \in \mathsf{X}. \tag{5.8}$$

Because $|X(t)| = |Q^x(t)| + |A^x(t)| + |B^x(t)|$, let us first look at

$$\mathsf{E}_x\left[\int_0^{\sigma_1}(A_k^x(s))^p\,ds\right].$$

For $x=(q,a,b)\in\mathsf{X}$, it is easy to check that

$$A_k^x(s)\le\begin{cases} a_k-s & \text{for } s\le a_k,\\ \xi_k(E_k^x(s)+1) & \text{for } s\ge a_k\ .\end{cases}$$

Therefore

$$(A_k^x(s))^p\le|x|^p+\sum_{i=1}^{E_k^0(s)+1}(\xi_k(i))^p.$$

Using Wald's identity and part (a) of Lemma 5.2, we have

$$\begin{aligned}\mathsf{E}_x\left[(A_k^x(s))^p\right] &\le |x|^p+\mathsf{E}\Big[\sum_{i=1}^{E_k^0(s)+1}(\xi_k(i))^p\Big]\\ &= |x|^p+\mathsf{E}\left[E_k^0(s)+1\right]\mathsf{E}\left[(\xi_k(1))^p\right]\\ &\le |x|^p+c_1(s+1)\mathsf{E}\left[(\xi_k(1))^p\right].\end{aligned}$$

Thus,

$$\mathsf{E}_x\left[\int_0^{\sigma_1}(A_k^x(s))^p\,ds\right]\le|x|^p\sigma_1+c_1(\sigma_1+(\sigma_1)^2/2)\mathsf{E}\left[(\xi_k(1))^p\right]\le c_2(|x|^{p+1}+1).\quad(5.9)$$

Similarly, we have

$$\mathsf{E}_x\left[\int_0^{\sigma_1}(B_k^x(s))^p\,ds\right]\le c_3(|x|^{p+1}+1).\qquad(5.10)$$

So, it remains to bound the integral of $|Q(t)|^p$. By ignoring customers that leave the network during the time interval $[0,t]$ we obtain the bound $|Q^x(t)|\le|Q^x(0)|+\sum_{k\in\mathcal{A}}E_k^x(t)$. By part (a) of Lemma 5.2, there exists some constant c_4 such that

$$\mathsf{E}_x[(E_k^x(t))^p]\le\mathsf{E}[(E_k^0(t))^p]\le c_4(t^p+1),\qquad k\in\mathcal{A},\ t\ge0,$$

and hence for constants c_5, $c_6<\infty$,

$$\begin{aligned}\mathsf{E}_x\Big[\int_0^{\sigma_1}|Q(t)|^p\,dt\Big] &\le c_5\sigma_1(|Q(0)|^p+\sigma_1^p)\\ &\le c_6(|x|^{p+1}+1),\qquad x\in\mathsf{X}.\end{aligned}$$

This together with (5.9) and (5.10), shows that (5.8) does hold.

Substituting the equivalent bound (5.7) into (5.6) we have for some $c_7<\infty$,

$$\mathsf{E}_x\Big[\sum_{k=0}^{\infty}\mathsf{E}_x\Big[\int_{\sigma_k}^{\sigma_{k+1}}(1+|X(t)|^p)\,dt\mid\mathcal{F}_{\sigma_k}\Big]\mathbb{1}\{k<k_*\}\Big]\le c_7\left[|x|^{p+1}+1\right].$$

By Fubini's theorem and the smoothing property of the conditional expectation, the LHS is precisely $\mathsf{E}_x[\int_0^{\sigma_{k_*}}(1+|X(t)|^p)\,dt]$. Since $\sigma_{k_*}\ge\tau_C(t_0L)$, this establishes the proposition. □

We now give a general statement concerning Markov processes which we require to apply Proposition 5.3.

Proposition 5.4 *Let X be a Borel right Markov process on X, let $f : \mathsf{X} \to \mathbb{R}_+$, and define for some $\delta > 0$, and a closed set $C \subseteq \mathsf{X}$*

$$V(x) := \mathsf{E}_x\Bigl[\int_0^{\tau_C(\delta)} f(X(t))\,dt\Bigr] \qquad x \in \mathsf{X},$$

If V is everywhere finite, and uniformly bounded on C, then there exists $\kappa < \infty$ such that

$$\frac{1}{t}\mathsf{E}_x[V(X(t))] + \frac{1}{t}\int_0^t \mathsf{E}_x[f(X(s))]\,ds \le \frac{1}{t}V(x) + \kappa, \quad t > 0,\ x \in \mathsf{X}. \tag{5.11}$$

PROOF We first demonstrate that for a constant $b < \infty$,

$$\mathsf{E}_x\Bigl[\int_0^{\tau_C(r)} f(X(t))\,dt\Bigr] \le V(x) + br, \qquad x \in \mathsf{X},\ r > 0. \tag{5.12}$$

Because the LHS is monotone in r, it is enough to establish the result for r of the form $n\delta$, where $n = 1, 2, \ldots$. The proof is by induction where we take $b = \sup_{x\in C} V(x)$. For $n = 1$, this is the definition of V. Supposing now that (5.12) holds for $r = n\delta$, we use the bound

$$\begin{aligned} \tau_C((n+1)\delta) &= \delta + \theta_\delta \tau_C(n\delta) \\ &\le \tau_C(\delta) + \theta_{\tau_C(\delta)}\tau_C(n\delta). \end{aligned}$$

So, for any x,

$$\begin{aligned} &\mathsf{E}_x\Bigl[\int_0^{\tau_C((n+1)\delta)} f(X(t))\,dt\Bigr] \\ &\quad\le\ \mathsf{E}_x\Bigl[\int_0^{\tau_C(\delta)} f(X(t))\,dt\Bigr] + \mathsf{E}_x\Bigl[\int_{\tau_C(\delta)}^{\theta_{\tau_C(\delta)}\tau_C(n\delta)} f(X(t))\,dt\Bigr]. \end{aligned}$$

The first term on the RHS is precisely $V(x)$. The second term on the RHS can be rewritten using the strong Markov property as follows:

$$\begin{aligned} &\mathsf{E}_x\Bigl[\int_0^{\tau_C((n+1)\delta)} f(X(t))\,dt\Bigr] \\ &\quad\le\ V(x) + \mathsf{E}_x\Bigl[\mathsf{E}_{X_{\tau_C(\delta)}}\Bigl[\int_0^{\tau_C(n\delta)} f(X(t))\,dt\Bigr]\Bigr] \\ &\quad\le\ V(x) + \sup_{x\in C}\mathsf{E}_x\Bigl[\int_0^{\tau_C(n\delta)} f(X(t))\,dt\Bigr] \\ &\quad\le\ V(x) + \sup_{x\in C} V(x) + nb \end{aligned}$$

where the last step uses the induction hypothesis. By induction we see that (5.12) does hold for any integer r, and by monotonicity it then follows that this inequality holds for all positive r, with a possibly larger constant b.

To prove the proposition we now write

$$\begin{aligned} \int P^t(x,dy)V(y) &= \mathsf{E}_x\Bigl[\int_t^{t+\theta_t\tau_C(\delta)} f(X(s))\,ds\Bigr] \\ &= V(x) - \mathsf{E}_x\Bigl[\int_0^t f(X(s))\,ds\Bigr] \\ &\quad + \mathsf{E}_x\Bigl[\int_{\tau_C(\delta)}^{t+\theta_t\tau_C(\delta)} f(X(s))\,ds\Bigr]. \end{aligned}$$

It may be shown by considering the two cases $\tau_C(\delta) \geq t$, and $\tau_C(\delta) < t$ separately that $t + \theta_t \tau_C(\delta) \leq \tau_C(\delta) + \theta_{\tau_C(\delta)}\tau_C(t)$. Thus by the strong Markov property as above,

$$\begin{aligned}
\int P^t(x,dy)V(y) &\leq V(x) - \int_0^t \mathsf{E}_x[f(X(s))]\,ds \\
&\quad + \mathsf{E}_x\Big[\mathsf{E}_{X_{\tau_C(\delta)}}\Big[\int_0^{\tau_C(t)} f(X(s))\Big]\,ds\Big] \\
&\leq V(x) - \int_0^t \mathsf{E}_x[f(X(s))]\,ds \\
&\quad + \sup_{x\in C} \mathsf{E}_x\Big[\int_0^{\tau_C(t)} f(X(s))\,ds\Big] \\
&\leq V(x) - \int_0^t \mathsf{E}_x[f(X(s))]\,ds \\
&\quad + \sup_{x\in C} V(x) + bt
\end{aligned}$$

where the last inequality uses the bound (5.12). Dividing both sides of this inequality by t, we obtain the desired result with $\kappa = \sup_{x\in C} V(x) + b$. □

5.3 L_p stability for the network

With these preliminaries complete, we may now present the main result of this section. Proposition 5.3 implies that the conditions of Proposition 5.4 are satisfied with $f(x) = 1 + |x|^p$, and this is sufficient to prove the theorem.

Theorem 5.5 *Suppose that Assumptions (A1) and (A2) hold, and that the fluid model is stable. Then there exists a constant $\kappa_p < \infty$ such that*

$$\frac{1}{t}\int_0^t \mathsf{E}_x[|Q(s)|^p]\,ds \leq \kappa_p\Big\{\frac{1}{t}|x|^{p+1} + 1\Big\}, \qquad t > 0,\ x \in \mathsf{X}. \tag{5.13}$$

In particular, for each initial condition,

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t \mathsf{E}_x[|Q(s)|^p]\,ds \leq \kappa_p.$$

□

6 Convergence

The inequality (5.4) is a form of *f-regularity* for the process, as defined in [36], which under general conditions is known to imply limit theorems such as strong laws of large numbers; mean ergodic theorems; functional central limit theorems; and laws of the iterated logarithm. In this section we develop limit theory which is most relevant to further systems analysis.

The first set of results involves convergence of polynomial moments of the queue length process to their steady state values. We then consider bounds on the rate of convergence to these steady state values, and give a proof of the Strong Law of Large Numbers. These results are crucial in simulation [46] and in evaluating performance of the network in steady state [20, 47, 48].

6.1 Preliminaries

The most convenient way of approaching convergence is to pose the problem in an operator theoretic framework, in which the transition kernel P^t is viewed as a linear operator from one function space to another. Recently it has been discovered that the following function space is particularly convenient when considering Harris recurrent processes. For a function $f\colon \mathsf{X} \to [1,\infty)$, define L_f^∞ to be the Banach space

$$L_f^\infty = \{g\colon \mathsf{X} \to \mathbb{R} : \text{for some } c < \infty,\ |g(x)| \leq cf(x),\ x \in \mathsf{X}\},$$

equipped with the norm $\|g\| = \sup_{x\in\mathsf{X}} |g(x)|/f(x)$.

We let R denote the resolvent kernel

$$R(x,A) = \int_0^\infty e^{-t} P^t(x,A)\,dt, \qquad x \in \mathsf{X},\ A \in \mathcal{B}_{\mathsf{X}},$$

and for a function g on X, we let $P^t g$ and Rg denote the functions $P^t g\,(x) = \int P^t(x,dy)g(y)$ and $Rg\,(x) = \int R(x,dy)g(y)$, respectively. In our definition of L_f^∞, we most typically take $f = f_q$ for some $q \geq 1$, where

$$f_q(x) := 1 + \int |y|^q R(x,dy). \tag{6.1}$$

In this case $P^t f \leq e^t f$, so that P^t is a bounded linear operator on L_f^∞ for any t.

If the stationary distribution π exists, we put $\pi(g) = \int_{\mathsf{X}} g(x)\,\pi(dx)$. Most of the results of this section involve bounding the convergence rate of $P^t g\,(x)$ to its steady state value $\pi(g)$ for $g \in L_f^\infty$. To obtain such results, it is most convenient to work with a version of the infinitesimal generator. Let $\mathcal{G}$ denote the extended generator for the process, as defined in Davis [49]. From Lemma 4.1 of Down, Meyn and Tweedie [50], the function Rh is in the domain of $\mathcal{G}$, and we have the formula

$$\mathcal{G}Rh = Rh - h \tag{6.2}$$

whenever $\int R(x,dy)|h(y)|$ is finite for each x. From this fact and the results of the previous section we obtain the following *Lyapunov drift* property for the network:

Proposition 6.1 *Suppose that Assumptions (A1) and (A2) hold, and that the fluid model is stable. Then there exists a function V_{p+1} in the domain of $\mathcal{G}$, a compact set $C \subset \mathsf{X}$, and a finite constant b such that*

$$\mathcal{G}V_{p+1} \leq -f_p(x) + b\mathbb{1}_C(x),$$

where f_p is defined in (6.1). The function V_{p+1} satisfies the bound $V_{p+1} \leq k_{p+1} f_{p+1}$ for a finite constant k_{p+1}.

PROOF From Proposition 5.3 and Proposition 5.4 it follows that (5.11) holds with $V(x) = \mathsf{E}_x[\int_0^{\tau_C(\delta)}(1 + |X(t)|^p)\,dt]$, and $f(x) = 1 + |x|^p$. Multiplying both sides of the inequality (5.11) by te^{-t} and integrating gives,

$$RV\,(x) \leq V(x) - f_p\,(x) + \kappa,$$

where f_p is given in (6.1). Since $f_p\,(x) \to \infty$ as $|x| \to \infty$, we may find a suitably large compact set $C \subset \mathsf{X}$, and some $b_0 < \infty$, such that

$$RV(x) \leq V(x) - \frac{1}{2} f_p(x) + b_0 \mathbb{1}_C(x). \tag{6.3}$$

This represents a "drift" for the Markov chain with transition kernel R, which easily translates to a drift property for the process X. On combining (6.3) and (6.2) we have with $V_{p+1} = 2RV$,

$$\mathcal{G}V_{p+1} = 2[RV - V] \leq -f_p(x) + 2b_0 \mathbb{1}_C(x),$$

which is the desired inequality.

The required bound on V_{p+1} follows directly from (5.4). □

We now apply this result to obtain convergence of moments.

6.2 Convergence of moments

If π is an invariant probability with $\pi(f) := \int f\, d\pi < \infty$, then we write $\Pi\colon L_f^\infty \to L_f^\infty$ as the simple projection operator $\Pi f = \pi(f)$. It is then of interest to know whether or not $\|\!|P^t - \Pi|\!\| \to 0$ as $t \to \infty$, where $\|\!| \cdot |\!\|$ is the induced operator norm. This turns out to be equivalent to exponential ergodicity [51, 52] which appears to be a difficult property to obtain, given only stability of the fluid model, although this behavior is typical of queueing models when f is taken to be an exponential (cf. [36, Section 16.4]). Instead we bound the pointwise norm defined as

$$||P^t(x, \cdot\,) - \pi(\,\cdot\,)||_f = \sup_{|g| \leq f} |\int \pi(dy) g(y) - \int P^t(x, dy) g(y)|, \qquad x \in \mathsf{X},$$

which is also called the *f-total variation norm* between the probability measures $P^t(x, \cdot\,)$ and $\pi(\,\cdot\,)$.

The following result shows that the f-total variation norm distance between the transient and steady state distributions converges to zero with only minor extra conditions beyond what was assumed in Theorem 5.5. The proof follows from the drift property Proposition 6.1 and Theorem 5.3 of [23].

Theorem 6.2 *Suppose that Assumptions (A1)–(A3) hold, and that the fluid model is stable. Then we have*

$$||P^t(x, \cdot\,) - \pi(\,\cdot\,)||_{f_p} \to 0, \qquad t \to \infty, \; x \in \mathsf{X}.$$

In particular, for each initial condition,

$$\lim_{t \to \infty} \mathsf{E}_x[|Q(t)|^p] = \mathsf{E}_\pi[|Q(0)|^p] < \infty$$

□

6.3 Rates of convergence

We now show how Theorem 6.2 may be strengthened to give rates of convergence of the first moment to its stationary value.

Theorem 6.3 *Suppose that Assumptions (A1)-(A3) hold, and that the fluid model is stable. Then with $f(x) = f_1(x)$ we have*

$$\lim_{t\to\infty} t^{(p-1)}\|P^t(x,\,\cdot\,) - \pi(\,\cdot\,)\|_f = 0, \qquad x \in \mathsf{X}.$$

In particular, for each initial condition,

$$\lim_{t\to\infty} t^{(p-1)}|\mathsf{E}_x[Q_t] - \mathsf{E}_\pi[Q_0]| = 0.$$

PROOF It is most convenient to first obtain the result for the skeleton chain $\hat{X} = \{\hat{X}_0, \hat{X}_1, \hat{X}_2, ...\}$, where $\hat{X}_i = X(i)$. As shown in [53], it is necessary to obtain a bound of the form

$$\mathsf{E}_x\Big[\sum_{n=0}^{\hat{\tau}_C-1} n^{p-1} f(\hat{X}_n)\Big] < \infty, \qquad x \in \mathsf{X}, \tag{6.4}$$

where $\hat{\tau}_C$ is the first entrance time, $\hat{\tau}_C = \min\{k \geq 1 : \hat{X}_k \in C\}$.

From Proposition 6.1 we have for any $q \leq p$,

$$\mathcal{G}V_{q+1} \leq -f_q(x) + b\mathbb{1}_C(x),$$

where f_q is defined in (6.1), C is a compact set, and the function V_q satisfies $V_q \leq k_q f_q$ for some constant k_q. From the definition of the generator, it then follows that

$$P^1V_{q+1}(x) - V_{q+1}(x) = \int_0^1 P^s\mathcal{G}V_{q+1}(x)\,ds \leq -\int_0^1 P^sf_q(x) + bP^s(x,C)\,ds$$

From the definition of f_q we have that $P^sf_q \leq e^sf_q$. Using this bound and the fact that $f_q(x) \to \infty$ as $|x| \to \infty$ we have for a possibly larger compact set C and constant b, and some constant $\delta > 0$,

$$P^1V_{q+1} \leq V_{q+1} - \delta f_q + b\mathbb{1}_C(x).$$

From the Comparison Theorem again (see [36, p. 337]), combined with the upper bound on V_{q+1}, it follows that for $q \leq p$,

$$\mathsf{E}_x\Big[\sum_{n=0}^{\hat{\tau}_C-1} f_q(\hat{X}_n)\Big] \leq d_qf_{q+1}(x), \qquad x \in \mathsf{X}, \tag{6.5}$$

where d_q is a finite constant.

Let $U_1 = f_1$, and for $q \geq 2$ define inductively

$$U_{q+1}(x) := \mathsf{E}_x\Big[\sum_{n=0}^{\hat{\tau}_C-1} U_q(\hat{X}_n)\Big], \qquad x \in \mathsf{X}.$$

Using (6.5), it is easy to verify by induction that

$$U_q(x) \leq D_qf_q(x),\ 2 \leq q \leq p+1, \qquad \text{where } D_q = \textstyle\prod_{1\leq i\leq q-1} d_q. \tag{6.6}$$

We also have just the lower bound that is needed to infer (6.4): For any $q \geq 2$,

$$U_q(x) \geq \frac{1}{(q-2)!}\mathsf{E}_x\Big[\sum_{n=0}^{\hat{\tau}_C-1} k^{q-2}f_1(\hat{X}_n)\Big] \tag{6.7}$$

We again prove this by induction. The case $q = 2$ is obvious. Assuming that (6.7) is valid for some arbitrary q, we may then write,

$$\begin{aligned} U_{q+1}(x) &:= \mathsf{E}_x\Big[\sum_{k=0}^{\hat\tau_C-1} U_q(\hat X_k)\Big] \\ &\geq \mathsf{E}_x\Big[\sum_{k=0}^{\hat\tau_C-1} \frac{1}{(q-2)!}\mathsf{E}_{X_k}\Big[\sum_{n=0}^{\hat\tau_C-1} n^{q-2} f_1(\hat X_n)\Big]\Big] \\ &= \mathsf{E}_x\Big[\sum_{k=0}^{\infty} \frac{1}{(q-2)!}\mathsf{E}\Big[\sum_{n=k}^{(k+\theta_k\hat\tau_C)-1} (n-k)^{q-2} f_1(\hat X_n) \mid \hat{\mathcal{F}}_k\Big]\mathbb{1}(k<\hat\tau_C)\Big] \end{aligned}$$

where $\hat{\mathcal{F}}_n = \sigma\{\hat X_0, \dots, \hat X_n\}$. The last equality follows from the Markov property, and the fact that $\theta_k\hat\tau_C = \min(j \geq 1 : \hat X_{k+j} \in C\}$.

On the event $\{k < \hat\tau_C\}$ we have that $k+\theta_k\hat\tau_C = \hat\tau_C$, and hence the above inequality may be written

$$\begin{aligned} U_{q+1}(x) &\geq \frac{1}{(q-2)!}\mathsf{E}_x\Big[\sum_{k=0}^{\infty} \mathsf{E}\Big[\sum_{n=k}^{\hat\tau_C-1} (n-k)^{q-2} f_1(\hat X_n)\mathbb{1}(k<\hat\tau_C) \mid \hat{\mathcal{F}}_k\Big]\Big] \\ &= \frac{1}{(q-2)!}\mathsf{E}_x\Big[\sum_{n=0}^{\hat\tau_C-1}\sum_{k=0}^{n} (n-k)^{q-2} f_1(\hat X_n)\Big] \end{aligned}$$

where the equality follows from Fubini's Theorem, and the smoothing property of the conditional expectation. Using the bound

$$\sum_{k=0}^{n}(n-k)^{q-2} = \sum_{k=0}^{n} k^{q-2} \geq \frac{1}{q-1}n^{q-1}$$

this then gives

$$U_{q+1}(x) \geq \frac{1}{(q-1)(q-2)!}\mathsf{E}_x\Big[\sum_{n=0}^{\hat\tau_C-1} n^{q-1} f_1(\hat X_n)\ \Big|\Big],$$

which is (6.7).

The bounds (6.7) and (6.6) taken together imply that for some constant c,

$$\mathsf{E}_x\Big[\sum_{n=0}^{\hat\tau_C-1} n^{p-1} f_1(\hat X_n)\Big] \leq c f_{p+1}(x), \qquad x \in \mathsf{X}.$$

This together with Theorem 2.1 of [53] implies that $n^{p-1}\|P^n(x,\,\cdot\,)-\pi\|_f \to 0$, $n \to \infty$, where $f = f_1$. Since we have from the bound $P^s f \leq e^s f$,

$$\|P^{n+s}(x,\,\cdot\,)-\pi\|_f \leq e^s\|P^n(x,\,\cdot\,)-\pi\|_f, \qquad 0 \leq s \leq 1,$$

this convergence rate for the skeleton carries over immediately to the process. □

6.4 Sample paths

We now consider limits involving the sample paths of the process.

Theorem 6.4 *Suppose that Assumptions (A1)–(A3) hold, and that the fluid model is stable. Let ν be any probability distribution on $(\mathsf{X}, \mathcal{B}_{\mathsf{X}})$, and π be the stationary distribution for X.*

(i) *For any $f : \mathsf{X} \to \mathrm{I\!R}_+$,*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X(s))\,ds = \pi(f) := \int f(x)\,\pi(dx) \quad \mathsf{P}_\nu\text{-}a.s. \tag{6.8}$$

(ii) *For any $f : \mathsf{X} \to \mathrm{I\!R}$ with $\pi(|f|) < \infty$, (6.8) holds.*

Proof The proof presented here is adapted from Meyn and Tweedie [36, Theorem 17.1.7] for a discrete time Markov chain. Under the stated assumptions of the theorem, $X = \{X(t), t \geq 0\}$ is positive Harris current. Let π be its stationary distribution.

By the Law of Large Numbers for stationary processes (see e.g. Doob [54, Theorem 2.1 in Chapter XI]), we have for any positive function $f : \mathsf{X} \to \mathrm{I\!R}_+$,

$$\mathsf{P}_\pi\left\{\frac{1}{t}\int_0^t f(X(s))\,ds \to \mathsf{E}_\pi\left[f(X(0)) \mid \Sigma_\pi\right]\right\} = 1,$$

where Σ_π is the σ-field of all P_π-invariant events. Following the proof of Proposition 17.1.4 and Theorem 17.1.5 of [36], we can show that Σ_π is P_π-trivial, and hence

$$\mathsf{E}_\pi\left[f(X(0)) \mid \Sigma_\pi\right] = \mathsf{E}_\pi\left[f(X(0))\right] = \pi(f) \quad \mathsf{P}_\pi\text{-a.s.}$$

By the definition of P_π it then follows that

$$\int_{\mathsf{X}} f_\infty(x)\,\pi(dx) = 1,$$

where $f_\infty(x) = \mathsf{P}_x\left\{\frac{1}{t}\int_0^t f(X(s))\,ds \to \pi(f)\right\}$.

A function $h : \mathsf{X} \to \mathrm{I\!R}$ is called *harmonic* if, for all $x \in \mathsf{X}$ and all $t \geq 0$,

$$\int P^t(x, dy)h(y) = h(x).$$

Using exact the same proof as in Theorem 17.1.5 of [36], one can show that the constants are the only bounded harmonic functions. Following the same proof as in Theorem 17.1.6 of [36], we can also show that f_∞ is harmonic, and hence $f_\infty(x) \equiv 1$.

We have thus established the desired limit

$$\mathsf{P}_x\left\{\frac{1}{t}\int_0^t f(X(s))\,ds \to \pi(f)\right\} = 1,$$

which implies that

$$\mathsf{P}_\nu\left\{\frac{1}{t}\int_0^t f(X(s))\,ds \to \pi(f)\right\} = 1$$

for any initial distribution ν. This shows that (i) holds, and (ii) follows from (i) by considering the positive part and the negative part of f respectively. □

Corollary 6.5 *Suppose that Assumptions (A1)–(A3) hold, and that the fluid model is stable. Let ν be any probability distribution on $(\mathsf{X}, \mathcal{B}_{\mathsf{X}})$, and π be the stationary distribution for X. For $k = 1, \ldots, K$, and $r = 1, \ldots, p$,*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t (Q_k(s))^r\,ds = \mathsf{E}_\pi[Q_k(0)^r] < \infty \quad \mathsf{P}_\nu\text{-}a.s.$$

7 Concluding Remarks

In this paper we have provided a set of sufficient conditions for stability which is applicable to virtually any network model found in practice. The question then naturally arises, how strong are these conditions? A partial converse is given in Meyn [55], but the complete characterization is still not fully understood. We have not seen an example of a stable network for which the fluid model is unstable, and it does seem likely that in most situations stability of the fluid model is also *necessary* in some sense for stability of the network. This is an important open problem which if answered positively would greatly simplify the counterexamples which have been recently devised. A direct proof of instability for a network can be extremely difficult (cf. Bramson [3, 56]), while the analysis of an associated fluid model is far simpler (cf. Dai and Weiss [6]).

Stability is largely a first step in a finer performance analysis of the network. For complex networks, there are presently two routes that one can follow. There is the LP approach of Kumar et. al. [20, 48] and Bertsimas et. al. [47] which is valid for exponential models, and there is also computer simulation. It is likely that both of these approaches may be improved using the methods developed here. In particular, it may be possible to extend the LP approach to include more general service and arrival processes by examining the generator for the network. Also, bounds on simulation error may be obtained given precise rates of convergence of the distributions of the network. We expect that such results may be obtained by using the results presented here combined with the recent methods of Lund and Tweedie [57].

References

[1] F. P. Kelly, "Networks of queues with customers of different types," *J. Appl. Probab.*, vol. 12, pp. 542–554, 1975.

[2] S. H. Lu and P. R. Kumar, "Distributed scheduling based on due dates and buffer priorities," *IEEE Transactions on Automatic Control*, vol. 36, pp. 1406–1416, 1991.

[3] M. Bramson, "Instability of FIFO queueing networks," *Annals of Applied Probability*, vol. 4, pp. 414–431, 1994.

[4] T. I. Seidman, "'First come, first served' can be unstable!," *IEEE Transactions on Automatic Control*, vol. 39, pp. 2166–2171, 1994.

[5] C. D. Pegden, *Introduction to SIMAN.* Sewickley, Pennsylvania: Systems Modeling Corporation, 1989.

[6] J. G. Dai and G. Weiss, "Stability and instability of fluid models for re-entrant lines," *Mathematics of Operations Research*, to appear.

[7] J. M. Gu, *Convergence and Performance for some Kelly-like Queueing Networks.* PhD thesis, University of Wisconsin, Madison, 1995.

[8] J. Keifer and J. Wolfwitz, "On the characteristics of the general queueing process, with applications to random walk," *Ann. Math. Statist.*, vol. 27, pp. 147–161, 1956.

[9] M. Miyazawa, "A formal approach to queueing processes in the steady state and their applications," *J. Appl. Probab.*, vol. 16, pp. 332–346, 1979.

[10] D. Daley and R. J. Rolski, "Finiteness of waiting-time moments in general stationary single-server queues," *Annals of Applied Probability*, vol. 2, pp. 987–1008, 1992.

[11] K. Sigman and D. D. Yao, "Finite moments for inventory processes," *Annals of Applied Probability*, vol. 4, pp. 765–778, 1994.

[12] E. Altman, P. Konstantopoulos, and Z. Liu, "Stability, monotonicity and invariant quantities in general polling systems," *Queueing Systems*, vol. 11, pp. 35–57, 1992.

[13] E. Altman and F. Spieksma, "Poling systems–moment stability of station times and central limit theorems." Submitted, 1993.

[14] L. Georgiadis and W. Szpankowski, "Stability of token passing rings," *Queueing Systems: Theory and Applications*, vol. 11, pp. 7–33, 1992.

[15] A. A. Borovkov, "Limit theorems for queueing networks," *Theory of Probability and its Applications*, vol. 31, pp. 413–427, 1986.

[16] K. Sigman, "The stability of open queueing networks," *Stoch. Proc. Applns.*, vol. 35, pp. 11–25, 1990.

[17] S. P. Meyn and D. Down, "Stability of generalized Jackson networks," *Annals of Applied Probability*, vol. 4, pp. 124–148, 1994.

[18] F. Baccelli and S. Foss, "Stability of Jackson-type queueing networks, I," *Queueing Systems: Theory and Applications*, vol. 17, pp. 5–72, 1994.

[19] P. R. Kumar, "Re-entrant lines," *Queueing Systems: Theory and Applications*, vol. 13, pp. 87–110, 1993.

[20] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. AC-39, pp. 1600–1611, August 1994.

[21] P. R. Kumar and S. P. Meyn, "Stability of queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. 40, pp. 251–260, February 1995.

[22] S. P. Meyn and R. L. Tweedie, "Stability of Markovian processes II: Continuous time processes and sample chains," *Adv. Appl. Probab.*, vol. 25, pp. 487–517, 1993.

[23] S. P. Meyn and R. L. Tweedie, "Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes," *Adv. Appl. Probab.*, vol. 25, pp. 518–548, 1993.

[24] S. P. Meyn and R. L. Tweedie, "Generalized resolvents and Harris recurrence of Markov processes," *Contemporary Mathematics*, vol. 149, pp. 227–250, 1993.

[25] S. P. Meyn and R. Tweedie, "A survey of Foster-Lyapunov techniques for general state space Markov processes," in *Proceedings of the Workshop on Stochastic Stability and Stochastic Stabilization, Metz, France*, June 1993.

[26] J. M. Harrison, "Brownian models of queueing networks with heterogeneous customer populations," *Proceedings of the IMA Workshop on Stochastic Differential Systems*, 1988. Springer-Verlag.

[27] H. Chen and A. Mandelbaum, "Discrete flow networks: Bottlenecks analysis and fluid approximations," *Mathematics of Operations Research*, vol. 16, pp. 408–446, 1991.

[28] H. Chen and A. Mandelbaum, "Hierarchical modeling of stochastic networks I: fluid models," in *Probability Models in Manufacturing Systems* (D. D. Yao, ed.), ch. 2, pp. 47–106, New York: Springer, 1994.

[29] J. M. Harrison and V. Nguyen, "Brownian models of multiclass queueing networks: Current status and open problems," *Queueing Systems: Theory and Applications*, vol. 13, pp. 5–40, 1993.

[30] J. G. Dai, "On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Annals of Applied Probability*, vol. 5, pp. 49–77, 1995.

[31] M. Sharpe, *General theory of Markov processes.* Boston: Academic Press, 1988.

[32] H. Kaspi and A. Mandelbaum, "Regenerative closed queueing networks," *Stochastics*, vol. 39, pp. 239–258, 1992.

[33] R. K. Getoor, "Transience and recurrence of Markov processes," in *Séminaire de Probabilités XIV* (J. Azéma and M. Yor, eds.), pp. 397–409, New York: Springer-Verlag, 1979.

[34] J. G. Dai and R. J. Williams, "Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons," *Theory of Probability and its Applications*, 1995.

[35] H. Chen, "Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines," *Annals of Applied Probability*, 1995. To appear.

[36] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability.* London: Springer-Verlag, 1993.

[37] P. Kuehn, "Multiqueue systems with nonexhaustive cycle service," *Bell Syst. Tech. J.*, vol. 58, pp. 671–698, 1979.

[38] C. Fricker and M. R. Jaïbi, "Monotonicity and stability of periodic polling models," *Queueing Systems: Theory and Applications*, vol. 15, pp. 211–238, 1994.

[39] Borovkov and Shassburger, "Ergodicity of a polling network," *Stoch. Proc. Appl.*, vol. 50, pp. 253–262, 1994.

[40] H. Takagi, *Analysis of Polling Systems.* Cambridge, MA: MIT Press, 1986.

[41] O. Boxma and W. Groenendijk, "Two queues with alternating service and switching times," vol. 7 of *CWI Monographs*, pp. 261–282, Amsterdam: North-Holland, 1988.

[42] J. M. Harrison and M. Pich, "Two-moment analysis of open queueing networks with general workstation capabilities," *Operations Research*, to appear.

[43] V. Nguyen, "Fluid and diffusion approximations of a two-station mixed queueing network," *Mathematics of Operations Research*, to appear.

[44] M. Ingenoso, *Stability analysis for certain queueing systems and multi-access communication channels.* PhD thesis, University of Wisconsin, 1993.

[45] A. Gut, *Stopped random walks: limit theorems and applications*, vol. 5 of *Applied probability.* New York: Springer-Verlag, 1988.

[46] D. L. Iglehart and G. S. Shedler, *Regenerative simulation of response times in networks of queues.* New York: Springer-Verlag, 1980.

[47] D. Bertsimas, I. C. Paschalidis, and Tsitsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance," *Annals of Applied Probability*, vol. 4, pp. 43–75, 1994.

[48] P. R. Kumar and S. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, 1995. To appear.

[49] M. Davis, *Markov Models and Optimization.* London: Chapman and Hall, 1993.

[50] D. Down, S. P. Meyn, and R. L. Tweedie, "Geometric and uniform ergodicity of Markov processes," *Annals of Probability*, to appear.

[51] S. P. Meyn and R. L. Tweedie, "Computable bounds for convergence rates of Markov chains," *Annals of Applied Probability*, vol. 4, pp. 981–1011, 1994.

[52] A. Hordijk and F. Spieksma, "On ergodicity and recurrence properties of a Markov chain with an application," *Adv. Appl. Probab.*, vol. 24, pp. 343–376, 1992.

[53] P. Tuominen and R. Tweedie, "Subgeometric rates of convergence of f-ergodic Markov chains," *Adv. Appl. Probab.*, vol. 26, pp. 775–798, 1994.

[54] J. L. Doob, *Stochastic Processes.* New York: John Wiley & Sons, 1953.

[55] S. P. Meyn, "Transience of multiclass queueing networks via fluid limit models," *Annals of Applied Probability*, submitted.

[56] M. Bramson, "Instability of FIFO queueing networks with quick service times," *Annals of Applied Probability*, vol. 4, pp. 693–718, 1994.

[57] R. Lund and R. Tweedie, "Geometric convergence rates for stochastically ordered Markov chains," *Mathematics of Operations Research*, to appear.