STABILITY OF OPEN MULTICLASS QUEUEING NETWORKS VIA FLUID MODELS

J. G. Dai*

January 7, 2000

Abstract

This paper surveys recent work on the stability of open multiclass queueing networks via fluid models. We recapitulate the stability result of Dai [8]. To facilitate study of the converse of the stability result, we distinguish between the notion of *fluid limit* and that of *fluid solution.* We define the stability region of a service discipline and the global stability region of a network. Examples show that piecewise linear Lyapunov functions are powerful tools in determining stability regions.

Stability, queueing networks, fluid models, scheduling, performance analysis, Harris recurrence, heavy traffic, Brownian models.

1 Introduction

There has been a recent surge in studying stability/instability of multiclass queueing networks. See, for example, Lu and Kumar [21], Rybko and Stolyar [24], Whitt [27], Bramson [2, 3] and Seidman [25]. To show that the instability can occur even in a Kelly-type network, a network in which all customers visit a station have a common service time distribution, we consider the three station network pictured in Figure 1. Jobs (or customers) 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 [19], a job class is defined for each processing stage. Therefore, in this example, each station processes three job classes. Each class may have its own general service time distribution (thus a job may have

*School of Industrial and Systems Engineering, and School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0205. Research supported by NSF grants DMS-9203524 and DDM-9215233, and two grants from the Texas Instruments Corporation. Part of this work was done while the author was visiting the *Institute For Mathematics And Its Applications*, whose financial support is also acknowledged.

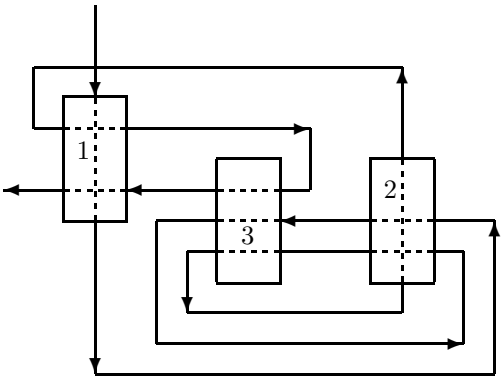

Figure 1: A three station network that may be unstable under certain priority service disciplines

different processing requirements on different visits to a station), and the service discipline at each station can be general.

Assume that at station 1, priority is given to customer classes in order $(9, 7, 1)$, where class 9 has the highest priority. At station 2, priority is given to customer classes in order $(4, 2, 6)$, and at station 3 the service discipline is first-in–first-out (FIFO). Assume further that the mean arrival rate to class 1 is 1, and the mean service time for each visit to stations 1 and 2 is 0.3 and for each visit to station 3 is 0.1. Therefore the nominal workloads per unit of time for servers 1, 2 and 3 are 90%, 90% and 30%, respectively. Dai and Meyn [9] 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 interarrival and service times are constants. Therefore 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. It appears from Table 1 that the average queue lengths in the simulations (M) and (D1) are growing without bound, whereas in simulation (D2) the total customer population seems 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 Section 7.1 or Dai and Weiss [12, Remark 3 in Section 6] for the insight into the instability revealed in these simulations (see also Chen [5, Theorem 4.5] and Gu [16].)

It is now known that the stability of a queueing network is closely related to that of the corresponding fluid model as in Rybko and Stolyar [24], Dai [8], Chen [5] and Stolyar [26]. (See Chen and Mandelbaum [6] for a survey on fluid models.) In this paper, we recapitulate the main result in Dai [8] which says that a service discipline in an open queueing network is stable if the corresponding fluid model eventually drains to zero starting from any initial condition. We

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.

carefully distinguish the notion of a *fluid limit* from that of a *fluid solution*. We believe that this distinction is helpful in studying the converse of the stability result. We also introduce definitions of the stability region of a service discipline and of the global stability region of a network. We show that the piecewise Lyapunov functions used in Botvitch and Zamyatin [1], Dai and Weiss [12] and Down and Meyn [14] provide a powerful tool in determining a stability region.

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 ($k = 1, \ldots, K$) 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 $\mathcal{E}$ 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* among classes, 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

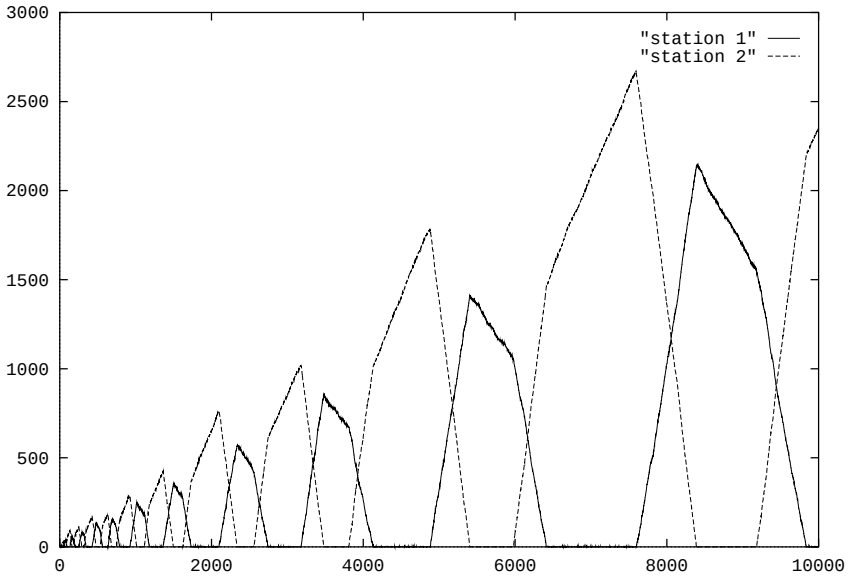

Figure 2: The queue lengths at each station oscillate with increasing magnitude: Mutual blocking between machines 1 and 2 results in instability.

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,

$$I + P + P^2 + \ldots \quad \text{is convergent.} \tag{1}$$

Condition (1) implies that all customers eventually leave the network. Hence the systems we consider are open queueing networks, although some more general networks may also be included (cf., Dai and Meyn [9]). This network description is quite standard, and may be found in numerous related papers (see, for example, Harrison and Nguyen [18]).

Throughout this paper, we assume that

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

(A2) $\mathsf{E}[\xi_\ell(1)] < \infty$ for $\ell \in \mathcal{E}$ and $\mathsf{E}[\eta_k(1)] < \infty$ for $k = 1, \ldots, K$.

(A3) For each $k \in \mathcal{E}$, there exists some nonnegative function $q_k(x)$ on $\mathbb{R}_+$ with $\int_0^\infty q_k(x)\,dx > 0$, and some integer j_k, such that

$$\begin{aligned} &\mathsf{P}(\xi_k(1) \geq x) > 0 \quad \text{for all} \quad x > 0, \\ &\mathsf{P}(\xi_k(1) + \cdots + \xi_k(j_k) \in dx) \geq q_k(x)\,dx. \end{aligned}$$

Conditions (A1) and (A2) are quite standard, although the independence assumption (A1) can be relaxed: see the remark after Proposition 2.1 of Dai [8]. Condition (A3) is required to establish ergodicity of the network. Under this condition, the argument used in Lemma 3.4 of Meyn and Down [22] may be applied to deduce that all compact subsets of a state space are *petite*. (For the definition of a petite set, see Section 4.1 of Meyn and Tweedie [23].) Frequently, milder conditions can be invoked to obtain this property for the network (see, for example, Assumption (A3') of Dai and Meyn [9].) Condition (A3) is *not* needed for bounding the moments of queue lengths (see Theorem 5.)

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 (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 workload* for server i per unit of time as

$$\rho_i = \sum_{k \in \mathcal{C}_i} \lambda_k/\mu_k. \tag{2}$$

In vector form, we have $\rho = CM\lambda$, where $M = \text{diag}(m_1, \ldots, m_K)$ and $m_k = 1/\mu_k$.

2.2 Service disciplines

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. For concreteness, at each station one of the following service disciplines is employed: first-in–first-out (FIFO), static buffer priority among classes (both preemptive and nonpreemptive), and head-of-line processor sharing among classes. Notice that under these service disciplines, the server may split its capacity among classes at a station, and at most one customer in each class can receive partial

service time. Readers will see that our approach actually can be applied to far more general service disciplines. In particular, processor sharing at station i among the first $r(k)$ customers in class k with $s(k) = i$ can be treated similarly. However, the true processor sharing discipline among all customers at a station is ruled out.

2.3 A Markovian state

Now we define a state process for the network, which depends upon the particular service discipline employed. Let $Q_k(t)$ be the queue length for class k customers, including the one being serviced. Let $Q(t) = (Q_1(t), \ldots, Q_K(t))'$. Thus, $Q(t)$ is the K-dimensional vector of *class-level* queue lengths at time t. For each station i, define

$$N_i(t) = \sum_{k \in \mathcal{C}_i} Q_k(t).$$

Then $N(t) = (N_1(t), \ldots, N_d(t))'$ is the d-dimensional vector of *station-level* queue lengths at time t.

Even when all distributions are exponential, under the FIFO service discipline, $\{Q_k(t) : k \in \mathcal{C}_i\}$ does not contain enough information to tell which customer will be served next at station i. In this case, we need to know how customers are lined up at station i. Thus, we define for station i,

$$\mathbb{Q}_i(t) = (k_{i,1}, k_{i,2}, \ldots, k_{i,N_i(t)}), \tag{3}$$

where $k_{i,j}$ is the class number for jth customer at station i. (If $N_i(t) = 0$, $\mathbb{Q}_i(t)$ is defined to be an empty list.) Put

$$\mathbb{Q}(t) = (\mathbb{Q}_1(t), \ldots, \mathbb{Q}_d(t)).$$

Then $\mathbb{Q}(t)$ tells exactly how customers are lined up at each station. It embodies more information than $Q(t)$ does. For static buffer priority and head-of-line processor sharing service disciplines, we simply let

$$\mathbb{Q}(t) = Q(t). \tag{4}$$

Therefore, in general, the state $X(t)$ at time t is

$$X(t) = (\mathbb{Q}(t), U(t), V(t)),$$

where $U(t) = (U_k(t) : k \in \mathcal{E})' \in \mathbb{R}_+^{|\mathcal{E}|}$ and $V(t) = (V_1(t), \ldots, V_K(t))' \in \mathbb{R}_+^K$. For $k \in \mathcal{E}$, $U_k(t)$ is the remaining time before the next class k customer will arrive from outside. For $k = 1, \ldots, K$, $V_k(t)$ is the remaining service time for the class k customer that is in service, which is set to be a fresh class k service time if $Q_k(t) = 0$. Both $U(t)$ and $V(t)$ are taken to be right continuous.

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 $x = (\mathbb{Q}, U, V)$ denote a generic state in X. Notice that the first component $\mathbb{Q}$ captures the

positions of customers in the network. It can be finite dimensional as in (4), or infinite dimensional as is the case for the FIFO service discipline. We use $|\mathbb{Q}|$ to denote the total number of jobs in the network, and for a $u \in \mathbb{R}^K$, $|u| = \sum_{k=1}^K |u_k|$. For a state $x = (\mathbb{Q}, U, V) \in \mathsf{X}$, we define the norm of x to be

$$|x| = |\mathbb{Q}| + |U| + |V|,$$

Let X be endowed with the natural induced topology. It is easy to check that the sublevel set

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

is a compact subset of X for any n.

It was shown in Dai [8, Section 2.2] that $X = \{X(t), t \geq 0\}$ 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 $((\Omega, \mathfrak{F}),\ \mathfrak{F}_t, X(t),\ \theta_t, \mathsf{P}_x)$ is a Borel right process on the measurable state space $(\mathsf{X}, \mathfrak{B}_{\mathsf{X}})$. In particular, $X = \{X(t), t \geq 0\}$ has right-continuous sample paths; it is defined on $(\Omega, \mathfrak{F})$ and is adapted to $\{\mathfrak{F}_t, t \geq 0\}$; $\{\mathsf{P}_x, x \in \mathsf{X}\}$ are probability measures on $(\Omega, \mathfrak{F})$ such that for all $x \in \mathsf{X}$,

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

and

$$\mathsf{E}_x\left\{f(X \circ \theta_\tau) \mid \mathfrak{F}_\tau\right\} = \mathsf{E}_{X(\tau)} f(X) \quad \text{on} \quad \{\tau < \infty\}, \quad \mathsf{P}_x\text{-a.s.}, \tag{5}$$

where τ is any $\mathfrak{F}_t$-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).

3 Discrete system dynamics

Let $x = (\mathbb{Q}(0), U(0), V(0))$ be the initial state of the network under a specified service discipline. In this section, we attach a superscript x to a symbol to explicitly denote the dependence on initial state x. In particular, $Q_k^x(t)$ is the queue length for class k customers at time t. For $\ell \in \mathcal{E}$ and $k = 1, \ldots, K$,

$$\begin{aligned} E_\ell^x(t) &= \max\{n \geq 1 : U_\ell(0) + \xi_\ell(1) + \ldots + \xi_\ell(n-1) \leq t\}, \quad t \geq 0, \\ S_k^x(t) &= \max\{n \geq 1 : V_k(0) + \eta_k(1) + \ldots + \eta_k(n-1) \leq t\}, \quad t \geq 0. \end{aligned}$$

It is easy to check that $E_\ell^x(t)$ is the number of exogenous arrivals to class ℓ by time t, and $S_k^x(r)$ is the number of service completions of class k customers if server $s(k)$ devotes r units of time to class k customers. Let $T_k^x(t)$ be the cumulative time that server $s(k)$ has devoted to class k customers by time t.

Then, $S_k^x(T_k^x(t))$ is the number of service completions for class k by time t. Recall the routing vectors $\phi^k(j)$ defined in Section 2. Let

$$\Phi^k(n) = \sum_{j=1}^{n} \phi^k(j). \tag{6}$$

Then $\Phi_\ell^k(n)$ is the number of class k customers routed to class ℓ among the first n class k service completions. It follows that for $k = 1, \ldots, K$,

$$Q_k^x(t) = Q_k^x(0) + E_k^x(t) + \sum_{\ell=1}^{K} \Phi_k^\ell(S_\ell^x(T_\ell^x(t))) - S_k^x(T_k^x(t)). \tag{7}$$

Let

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

Then $I_i^x(t)$ is the cumulative time that server i is idle in $[0, t]$. Besides (7), we have

$$Q^x(t) = (Q_1^x(t), \ldots, Q_K^x(t))' \geq 0, \quad t \geq 0, \tag{8}$$

$$T^x(t) = (T_1^x(t), \ldots, T_K^x(t))' \text{ is a non-decreasing and starts from 0,} \tag{9}$$

$$I^x(t) = (I_1^x(t), \ldots, I_d^x(t))' \text{ is non-decreasing,} \tag{10}$$

Because all of the allowable service disciplines are non-idling, the cumulative idle time $I_i^x(t)$ does not increase when $N_i^x(t) > 0$, where as before

$$N_i^x(t) = \sum_{k \in \mathcal{C}_i} Q_k^x(t).$$

That is,

$$\int_0^\infty N_i^x(t)\, dI_i^x(t) = 0. \tag{11}$$

Recall the constituency matrix C defined in Section 2. In vector form, (7)–(11) can be written as

$$Q^x(t) = Q^x(0) + E^x(t) + \sum_{\ell=1}^{K} \Phi^\ell(S_\ell^x(T_\ell^x(t))) - S^x(T^x(t)), \tag{12}$$

$$Q^x(t) \geq 0, \quad t \geq 0, \tag{13}$$

$$T^x(0) = 0, \text{ and } T^x(\cdot) \text{ is a non-decreasing,} \tag{14}$$

$$I^x(t) = et - CT^x(t) \text{ is non-decreasing,} \tag{15}$$

$$\int_0^\infty CQ^x(t)\, dI^x(t) = 0, \tag{16}$$

where, as usual, $S^x(T^x(t)) = (S_1^x(T_1^x(t)), \ldots, S_K^x(T_K^x(t)))'$. Notice that (12)–(16) hold for FIFO, buffer priority disciplines, and the head-of-line processor

sharing discipline that are considered in this paper. For each service discipline, there are additional equations that, together with (12)–(16), describe the network dynamics for the discipline. In the remainder of this section, we derive these extra equations for static buffer priority disciplines.

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. For concreteness, we assume the discipline is preemptive resume. 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

$$\begin{aligned} T_k^{x,+}(t) &= \sum_{\ell \in H_k} T_\ell^x(t) \\ I_k^{x,+}(t) &= t - T_k^{x,+}(t), \\ Q_k^{x,+}(t) &= \sum_{\ell \in H_k} Q_\ell^x(t). \end{aligned}$$

Then $T_k^{x,+}(t)$ is the cumulative amount of service in $[0,t]$ dedicated to customers whose classes are included in H_k, and $I_k^{x,+}(t)$ is the total unused capacity that is available to serve customers whose class does not belong to H_k. Note that $I_i^x(t)$ is a station level quantity representing the total unused capacity in $[0,t]$ by server i; whereas $I_k^{x,+}(t)$ 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 by the integral equation

$$\int_0^\infty Q_k^{x,+}(t)\, dI_k^{x,+}(t) = 0, \qquad 1 \le k \le K. \tag{17}$$

4 Fluid limit dynamics

Definition 1. A sequence of functions $f_n(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is said to be convergent to $f(\cdot)$ *uniformly on compact sets* (u.o.c.) if for every $t > 0$,

$$\sup_{0 \le s \le t} |f_n(s) - f(s)| \to 0$$

as $n \to \infty$.

In the following lemma, we present functional strong laws of large numbers for some processes defined. The proof can be found in Dai [8, Lemma 4.2]. Recall that a state $x = (\mathbb{Q}(0), U(0), V(0))$ has three components.

Lemma 1. *Assume that*

$$\lim_{|x| \to \infty} \frac{1}{|x|} U(0) = \bar{U} \quad \textit{and} \quad \lim_{|x| \to \infty} \frac{1}{|x|} V(0) = \bar{V}.$$

Then as $|x| \to \infty$, almost surely

$$\frac{1}{|x|}\Phi^k([|x|t]) \to P_k' t, \quad u.o.c. \tag{18}$$

$$\frac{1}{|x|}E_k^x(|x|t) \to \alpha_k(t-\bar{U}_k)^+, \quad u.o.c. \tag{19}$$

$$\frac{1}{|x|}S_k^x(|x|t) \to \mu_k(t-\bar{V}_k)^+, \quad u.o.c., \tag{20}$$

where $[t]$ is the integer part of t.

Lemma 2. *For a fixed sample path ω, for each sequence x_n of initial states with $|x_n| \to \infty$, there is a subsequence $\{x_{n_j}\}$ such that*

$$\begin{aligned}\frac{1}{|x_{n_j}|}(T_1^{x_{n_j}}(|x_{n_j}|\,t,\omega),\ldots T_K^{x_{n_j}}(|x_{n_j}|\,t,\omega)) \\ \to (\bar{T}_1(t,\omega),\ldots,\bar{T}_K(t,\omega))\end{aligned} \tag{21}$$

uniformly (with respect to t) on compact sets as $j \to \infty$.

Proof. Let ω be a fixed sample path. Let $0 \le s < t$. It is easy to check that

$$\frac{1}{|x|}T_k^x(|x|\,t,\omega) - \frac{1}{|x|}T_k^x(|x|s,\omega) \le t-s.$$

The lemma then follows easily. □

REMARK. In general the limit $\bar{T}(t,\omega) = (\bar{T}_1(t,\omega),\ldots,\bar{T}_K(t,\omega))'$ is random. That is $\bar{T}(t,\omega)$ may indeed depend on ω. However, from now on the dependence of ω is suppressed from the expression.

Theorem 1. *Consider a non-idling service discipline. For almost all sample paths ω and any sequence of initial states $\{x_n\}$ with $|x_n| \to \infty$, there is a subsequence $\{x_{n_j}\}$ such that*

$$\frac{1}{|x_{n_j}|}(Q^{x_{n_j}}(0), U^{x_{n_j}}(0), V^{x_{n_j}}(0)) \to (\bar{Q}(0), \bar{U}, \bar{V}), \tag{22}$$

$$\frac{1}{|x_{n_j}|}(Q^{x_{n_j}}(|x_{n_j}|\,t), T^{x_{n_j}}(|x_{n_j}|\,t)) \to (\bar{Q}(t), \bar{T}(t)) \quad u.o.c. \tag{23}$$

Furthermore, $(\bar{Q}, \bar{T})$ satisfies the following set of equations.

$$\bar{Q}(t) = \bar{Q}(0) + (\alpha t - \bar{U})^+ - (I-P)'M^{-1}(\bar{T}(t) - \bar{V})^+, \tag{24}$$

$$\bar{Q}(t) \ge 0, \tag{25}$$

$$\bar{T}(t) \text{ is non-decreasing and starts from zero,} \tag{26}$$

$$\bar{I}(t) = et - C\bar{T}(t) \text{ is non-decreasing,} \tag{27}$$

$$\int_0^\infty C\bar{Q}(t)\,d\bar{I}(t) = 0. \tag{28}$$

Proof. Recall first that $|x_n| = |Q^{x_n}(0)| + |U^{x_n}(0)| + |V^{x_n}(0)|$. Thus,

$$\frac{1}{|x_n|}|Q^{x_n}(0)| \le 1, \quad \frac{1}{|x_n|}|U^{x_n}(0)| \le 1, \quad \frac{1}{|x_n|}|V^{x_n}(0)| \le 1$$

for all n. Therefore by Lemma 2, there is a subsequence $\{n_j\}$ such that (22) and (21) hold. By Lemma 1 and (12), we have

$$\frac{1}{|x_{n_j}|}Q^{x_{n_j}}(|x_{n_j}|\, t) \to \bar{Q}(t),$$

where $\bar{Q}(t)$ satisfies (24). Conditions (25)–(27) follow from (13)–(15), whereas Condition (28) follows from (16) and Lemma 2.4 of Dai and Williams [13]. □

Theorem 1 holds for FIFO, buffer priority disciplines and the head-of-line processor sharing discipline, as well as many other non-idling disciplines. For a particular service discipline, a limit $(\bar{Q}(\cdot), \bar{T}(\cdot))$ usually satisfies more equations, in addition to (24)–(28). For a buffer priority discipline, let

$$\bar{Q}_k^+(t) = \sum_{\ell \in H_k} \bar{Q}_\ell(t),$$
$$\bar{T}_k^+(t) = \sum_{\ell \in H_k} \bar{T}_\ell(t).$$

Theorem 2. *Under the static preemptive resume buffer priority discipline, the limit* $(\bar{Q}(t), \bar{T}(t))$ *in Theorem 1 satisfies*

$$\int_0^\infty \bar{Q}_k^+(t)\, d(t - \bar{T}_k^+(t)) = 0, \quad k = 1, \ldots, K, \tag{29}$$

in addition to (24)–(28).

Proof. The theorem follows from (17) and Lemma 2.4 of [13]. □

Definition 2. A limit $(\bar{Q}(\cdot), \bar{T}(\cdot))$ in Theorem 1 is called a *fluid limit* under a service discipline with initial fluid level $\bar{Q}(0)$ and *delays* $\bar{U}$ and $\bar{V}$. We use $\mathfrak{L}$ to denote the set of such limits.

Definition 3. The *delayed fluid model* of a buffer priority service discipline in a network with delay $(\bar{U}, \bar{V}) \in \mathbb{R}_+^{|\mathcal{E}|+K}$ starting from $\bar{Q}(0)$ is defined to be the set of equations (24)–(29). Any solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ to equations (24)–(29) is called a *fluid solution* of the fluid model for the buffer priority discipline. We use $\mathfrak{M}$ to denote the collection of all solutions $(\bar{Q}(\cdot), \bar{T}(\cdot))$ of the fluid model.

Remark. For a given discipline, one can define the corresponding delayed fluid model similarly. The only change needed is to replace (29) in Definition 3 with a condition analogous to (29) that is specific to the discipline.

It is obvious that any fluid limit is a fluid solution to the fluid model. Therefore, we have

$$\mathfrak{L} \subset \mathfrak{M}. \tag{30}$$

When there is a single customer class served at each station, the fluid model has a unique solution. In this case, there is no need to differentiate between fluid limits and fluid solutions. However, in multiclass networks, it is quite typical that the fluid model has multiple solutions. The reason to distinguish a fluid limit from a fluid solution will be explained in the next section.

5 Stability of the fluid model and the queueing network

Definition 4. The delayed fluid model of a service discipline is *stable* if there exists a $\delta > 0$ such that for any fluid solution $(\bar{Q}(\cdot), \bar{T}(\cdot)) \in \mathfrak{M}$ with $|\bar{Q}(0)| + |\bar{U}| + |\bar{V}| = 1$, $\bar{Q}(t) \equiv 0$ for $t \geq \delta$.

When the delays $\bar{U}$ and $\bar{V}$ are zeros, the corresponding delayed fluid model is called the *undelayed* fluid model, or simply the fluid model. That is, the undelayed fluid model is defined by adding the following equation to (25)–(28)

$$\bar{Q}(t) = \bar{Q}(0) + \alpha t - (I - P)' M^{-1} \bar{T}(t). \tag{31}$$

If we let $|x| \to \infty$ while keeping $U(0)$ and $V(0)$ bounded, then the corresponding fluid limit is a solution to the undelayed fluid model. Assume that all distributions in the network are exponential. Because of the memoryless property of the distribution, for many service disciplines including FIFO and buffer priority disciplines, the residual interarrival times and service times are not needed in the state descriptions. Thus the corresponding fluid limit is always undelayed. However, under general distributional assumptions on the network, the corresponding fluid limit is delayed. Chen [5, Theorem 5.3] proved the following proposition.

Proposition 1. *If the undelayed fluid model is stable, then the delayed fluid model is stable.*

The emptying time δ in Definition 5.1 is independent of a particular fluid solution and initial state. Stolyar [26, Proposition 6] proved, however, that an apparent weaker condition implies the stability of a fluid model. The proof also follows from the remark following Proposition 3.3 of Dupuis and Williams [15].

Proposition 2. *The undelayed fluid model is stable if and only if there exists some norm $||\cdot||$ on $\mathbb{R}_+^K$ such that for any $\bar{Q} \in \mathfrak{M}$ with $||\bar{Q}(0)|| = 1$, there exists $t > 0$ such that $||\bar{Q}(t)|| < 1$.*

The following theorem was proved in Dai [8, Theorem 4.3].

Theorem 3. *A service discipline is positive Harris recurrent if the corresponding fluid limit model is stable.*

We conjecture that under certain service disciplines there are fluid solutions which cannot be achieved as fluid limits. It is therefore conceivable that there is a queueing network with certain service discipline, whose fluid limit model is stable but whose fluid model is not stable. This distinction is important in formulating a correct converse result of Theorem 3. The converse problem is still open although in many specific cases it has been shown that the instability of a fluid limit model implies the instability of the corresponding queueing network.

Definition 5. For a service discipline π in an open multiclass queueing network, its *stability region* is defined to be

$$\mathfrak{D}_\pi = \left\{(\alpha, m) \in \mathbb{R}_+^{|\mathcal{E}|+K} : \begin{array}{l}\text{such that the discipline } \pi \text{ is pos-}\\ \text{itive Harris recurrent}\end{array}\right\}. \tag{32}$$

REMARK. We believe the stability region for a non-idling discipline depends on α and m only, not on the second or higher moments.

Recall that, for concreteness, at each station one of the following service disciplines is employed: first-in–first-out (FIFO), static buffer priority among classes (both preemptive and nonpreemptive), and head-of-line processor sharing among classes. However, it is easy to see that more disciplines can be considered in the following definition.

Definition 6. The *global stability region* for an open multiclass queueing network is defined to be

$$\mathfrak{G} = \cap_\pi \mathfrak{D}_\pi, \tag{33}$$

where π ranges among all allowable service disciplines.

Similarly, using the stability notion defined in Definition 5.1, we can define the stability region $\bar{\mathfrak{D}}_\pi$ of a service discipline π for the fluid model and the global stability region $\bar{\mathfrak{G}}$ of the fluid model. Theorem 3 proves that

$$\bar{\mathfrak{D}}_\pi \subset \mathfrak{D}_\pi \quad \text{and} \quad \bar{\mathfrak{G}} \subset \mathfrak{G}.$$

Let

$$\mathfrak{D}_0 = \left\{(\alpha, m) \in \mathbb{R}_+^{|\mathcal{E}|+K} : \rho_i < 1, \quad i = 1, \ldots, d.\right\} \tag{34}$$

where $\rho = (\rho_1, \ldots, \rho_d)'$ is defined in (2). Obviously, we have

$$\mathfrak{D}_\pi \subset \mathfrak{D}_0 \quad \text{and} \quad \bar{\mathfrak{D}}_\pi \subset \mathfrak{D}_0.$$

Bramson [2, 3] proved the surprising result that

$$\mathfrak{D}_{\text{FIFO}} \neq \mathfrak{D}_0.$$

Similar results were proved by Lu and Kumar [21], Rybko and Stolyar [24] and Seidman [25].

6 Calculus of fluid models

A nice feature of the fluid model is that one can apply calculus to it. It follows from (26) and (27) that $\bar{T}(\cdot)$ is Lipschitz continuous, and hence by (31) $\bar{Q}(\cdot)$ is Lipschitz continuous. Therefore we have the following proposition.

Proposition 3. *The paths $\bar{Q}_k(\cdot)$ and $\bar{T}_k(\cdot)$ are absolutely continuous. Therefore they have derivatives almost everywhere with respect to the Lebesgue measure on $[0, \infty)$.*

A path $f(\cdot)$ is said to be regular at t if it is differentiable at t. We use $\dot{f}(t)$ to denote the derivative of $f(\cdot)$ at a regular point t.

The following simple lemma, which appeared in Dai and Weiss [12], turns out to be very useful for the analysis of stability of the fluid model.

Lemma 3. *Let $f(\cdot)$ be an absolutely continuous nonnegative function on $[0, \infty)$.*
(i) If $f(t) = 0$ and $\dot{f}(t)$ exists, then $\dot{f}(t) = 0$.
(ii) Assume that for some $\epsilon > 0$ and almost every regular point $t > 0$, whenever $f(t) > 0$ then $\dot{f}(t) \leq -\epsilon$. Then $f(t) = 0$ for all $t \geq \delta$, where $\delta = f(0)/\epsilon$. Furthermore, $f(\cdot)$ is nonincreasing, and hence once it reaches zero it stays there forever.

Throughout this paper, whenever the derivative of the fluid model is considered at time t, we assume t is a regular point of $\bar{Q}(\cdot)$ and $\bar{T}(\cdot)$. Let $a \in \mathbb{R}_+^K$ be fixed. Define

$$G(t) = C\mathrm{diag}(a)(I - P')^{-1}\bar{Q}(t),$$

and

$$f(t) = \max\{G_1(t), \ldots, G_d(t)\}.$$

Note that $f(t)$ is a piecewise linear function of $\bar{Q}(t)$. It is easy to check that $f(t)$ is nonnegative and Lipschitz continuous and hence absolutely continuous.

Theorem 4. *If there exists an $a = (a_1, \ldots, a_K)' > 0$ and $\epsilon > 0$ such that whenever $|\bar{Q}(t)| > 0$, $\dot{f}(t) \leq -\epsilon$, then the service discipline is stable.*

Proof. The proof follows from Lemma 3. □

Corollary 1. *Assume that in a two station network there exist $a = (a_1, \ldots, a_K)' > 0$ and $\epsilon_i > 0$ for $i = 1, 2$ such that $\dot{G}_i(t) \leq -\epsilon_i$ whenever $\bar{N}_i(t) > 0$. Furthermore, assume that $G_1(t) \leq G_2(t)$ whenever $\bar{N}_1(t) = 0$ and $G_2(t) \leq G_1(t)$ whenever $\bar{N}_2(t) = 0$, where $\bar{N}_i(t) = \sum_{k \in \mathcal{C}_i} \bar{Q}_k(t)$. Then, the service discipline is stable.*

7 Examples

7.1 The Lu-Kumar-Bramson-type network

A re-entrant line is a multiclass open queueing network, whose routing matrix is of the form $P_{k,k+1} = 1$ for $k = 1, \ldots, K-1$ and $P_{k,\ell} = 0$ otherwise and $\mathcal{E} = \{1\}$. Consider a two station re-entrant line, where all customers visit stations 1, 2, ..., 2, 1, ..., 1. Therefore, $s(1) = 1$, $s(k) = 2$ for $k = 2, \ldots, r$, and $s(k) = 1$ for $k = r+1, \ldots, K$. When $r = 3$ and $K = 4$, the resulting network is the Lu-Kumar network [21]. When $r = K-1$, the resulting network is the Bramson network [2]. We call this network a Lu-Kumar-Bramson-type network. For the Lu-Kumar network, assuming that $\alpha_1 = 1$, Dai and Weiss [12] showed that

$$\bar{\mathfrak{G}} = \{m \in \mathbb{R}^4_+ : m_1 + m_4 < 1, \quad m_2 + m_4 < 1, \quad m_2 + m_3 < 1\}.$$

Furthermore, the global stability region is realized by the Lu-Kumar buffer priority discipline that gives classes 2 and 4 higher priorities. That is,

$$\bar{\mathfrak{G}} = \bar{\mathfrak{D}}_{\mathrm{Lu-Kumar-buffer-priority}}.$$

Therefore, the Lu-Kumar priority discipline is the "worst" one in the sense that it gives the smallest stability region. The approach used in [12] was modified from Botvitch and Zamyatin [1]. The method has been further generalized by Dai and VandeVate [10]. For the Lu-Kumar-Bramson type networks, using Corollary 1, they proved that

$$\left\{ m \in \mathbb{R}^K_+ : \rho_1 < 1, \quad \rho_2 < 1, \quad \sum_{k=2}^{r-1} m_k + \sum_{k=r+1}^{K} m_k < 1 \right\} \subset \bar{\mathfrak{G}}. \tag{35}$$

Now consider the following buffer priority discipline, which generalizes the Lu-Kumar priority discipline. At station 1 priorities, in decreasing order, are given as K, $K-1$, ..., $r+1$ and 1. At station 2 priorities are given as $r-1, r-2, \ldots, 2$ and r. Under this buffer priority discipline, the network is effectively reduced to the original Lu-Kumar network with a new set of parameters $\tilde{m} = (\tilde{m}_1, \tilde{m}_2, \tilde{m}_3, \tilde{m}_4)' \in \mathbb{R}^4_+$, where

$$\tilde{m}_1 = m_1, \ \tilde{m}_2 = \sum_{k=2}^{r-1} m_k, \ \tilde{m}_3 = m_r, \ \tilde{m}_4 = \sum_{k=r+1}^{K} m_k.$$

Therefore, we can apply results on Lu-Kumar network in Dai and Weiss [12, Section 5] to ensure

$$\bar{\mathfrak{G}} = \left\{ m \in \mathbb{R}^K_+ : \rho_1 < 1, \rho_2 < 1, \sum_{k=2}^{r-1} m_k + \sum_{k=r+1}^{K} m_k < 1 \right\}.$$

The worst discipline is this generalized Lu-Kumar buffer priority discipline. For the queueing network under discussion, consider the generalized Lu-Kumar buffer priority preemptive resume discipline. Following Theorem 3 and (35), we have

$$\left\{ m \in \mathbb{R}^K_+ : \rho_1 < 1, \quad \rho_2 < 1, \quad \sum_{k=2}^{r-1} m_k + \sum_{k=r+1}^{K} m_k < 1 \right\} \subset \mathfrak{G}.$$

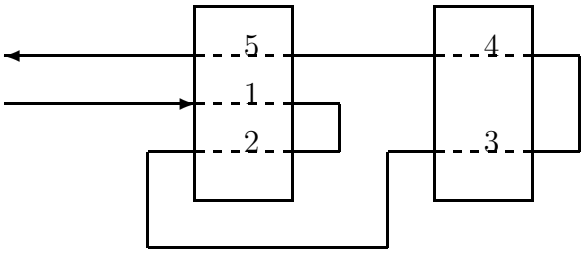

Figure 3: The Dai-Wang network

Harrison [17] made the following key observation. Suppose the network initially has customers only in buffers 1 and r. Then no two classes different from 1 and r can ever be worked on simultaneously. Consequently,

$$\sum_{k=2}^{r-1} T_k^x(t) + \sum_{k=r+1}^{K} T_k^x(t) \le t, \quad t \ge 0.$$

Using this observation, it is not difficult to argue that

$$\left\{ m \in \mathbb{R}_+^K : \rho_1 < 1, \quad \rho_2 < 1, \quad \sum_{k=2}^{r-1} m_k + \sum_{k=r+1}^{K} m_k < 1 \right\} = \mathfrak{G}.$$

When the service times at station 3 in the network pictured in Figure 1 are zero, the corresponding network is reduced to a Lu-Kumar-Bramson-type network with $r = 4$ and $K = 6$. Therefore $m_k = 0.25$, $k = 1, \ldots, 6$, is the critical value for stability when all service times have the same mean. This explains why the network in Figure 1 is unstable when service time at station 3 are small.

7.2 The Dai-Wang network

For the network pictured in Figure 3, again by using Corollary 1 and assuming $\alpha_1 = 1$, Dai and VandeVate [10] showed that

$$\left\{ m \in \mathbb{R}_+^5 : \rho_1 < 1, \quad \rho_2 < 1, \quad m_5 < (1 - m_1)(1 - m_3) \right\} \subset \bar{\mathfrak{G}}.$$

Now consider the priority discipline giving priorities, in decreasing order, at station one: 5, 1, 2 and at station two: 3, 4. We show, using the construction similar to the one in [12], that if $\rho_1 < 1$, $\rho_2 < 1$ and

$$(1 - m_1)(1 - m_3) \le m_5,$$

the discipline is not stable for the fluid model. Indeed, let the fluid model start from $\bar{Q}(0) = (1, 0, 0, 0, 0)'$. Then at $t_1 = 1/(\mu_1 - 1)$, $\bar{Q}(t_1) = (0, 1/(1 - m_1), 0, 0, 0)'$. Let t_2 be the first time that the fluid level at buffer 2 reaches zero.

Let ν be the departure rate from buffer 2 between t_1 and t_2. It follows from Dai and Weiss [12, Proposition 3.1] that

$$\nu = (1 - m_1)\mu_2.$$

We first check that $\nu > \mu_3$. In fact, if $\nu \leq \mu_3$, we have

$$\begin{aligned} m_1 + m_2 + m_5 &\geq m_1 + m_5 + \frac{1}{1-m_1}\frac{1}{m_3} \\ &\geq m_1 + \frac{1}{1-m_1} \geq 1, \end{aligned}$$

contradicting $\rho_1 < 1$. It is easy to check that $t_2 - t_1 = (1/(\nu-1))(1/(1-m_1))$ or equivalently,

$$t_2 = \frac{\nu}{\nu-1}\frac{1}{\mu_1 - 1} + \frac{1}{\nu - 1}.$$

and

$$\begin{aligned} \bar{Q}(t_2) &= (0, 0, 1 + t_2 - \mu_3(t_2 - t_1), \mu_3(t_2 - t_1), 0)' \\ &= (0, 0, (\nu - \mu_3)(1/(\nu-1))(1/(1-m_1)), \\ &\qquad \mu_3(1/(\nu-1))(1/(1-m_1)), 0)'. \end{aligned}$$

Let t_3 be the first time that buffer 3 reaches zero. At t_3,

$$\bar{Q}(t_3) = (0, 0, 0, 1 + t_3, 0)',$$

where

$$t_3 - t_2 = \frac{(\nu - \mu_3)(1/(\nu-1))(\mu_1/(\mu-1))}{\mu_3 - 1}.$$

We can check that

$$\begin{aligned} 1 + t_3 &= t_3 - t_2 + t_2 + 1 \\ &= \frac{1}{\mu_3 - 1}\frac{\mu_1}{\mu_1 - 1}\frac{\nu}{\nu - 1} - \frac{\mu_3}{\mu_3 - 1}\frac{\mu_1}{\mu_1 - 1}\frac{1}{\nu - 1} \\ &\quad + \frac{1}{\mu_1 - 1}\frac{\nu}{\nu - 1} + \frac{1}{\nu - 1} + 1 \\ &= \frac{\mu_3}{\mu_3 - 1}\frac{\mu_1}{\mu_1 - 1} = \frac{1}{1-m_1}\frac{1}{1-m_3}. \end{aligned}$$

Let $t_4 - t_3 = m_5/((1-m_1)(1-m_3))$. Then at t_4,

$$\bar{Q}(t_4) = (m_5/((1-m_1)(1-m_3)), 0, 0, 0, 0)'.$$

From t_4 the solution enters a new cycle with initial total fluid level

$$m_5/((1-m_1)(1-m_3)) \geq 1.$$

Hence the discipline is unstable for the fluid model. Therefore we have proved that

$$\bar{\mathfrak{G}} = \left\{ m \in \mathbb{R}^5_+ : \rho_1 < 1, \rho_2 < 1, m_5 < (1 - m_1)(1 - m_3) \right\},$$

and the worst discipline is the buffer priority discipline given earlier.

In [11], Dai and Wang showed that when $m_1 = 0.1\alpha_1$, $m_2 = 0.05\alpha_1$, $m_3 = 0.9\alpha_1$, $m_4 = 0.05\alpha_1$ and $m_5 = 0.8\alpha_1$, the corresponding Brownian model proposed by Harrison and Nguyen [18] does not exist under the FIFO discipline. Specializing the stability region to this case, we have

$$\bar{\mathfrak{G}} = \left\{ \alpha_1 \geq 0 : \alpha_1 < 10 \left(1 - 2\sqrt{2}/3 \right) \approx 0.57191 \right\}.$$

By solving an LP problem numerically and performing computer simulations, Down and Meyn [14] were able to predict the critical value to be 0.57191. By using a different LP, quadratic LP, Kumar and Meyn [20] had shown earlier that

$$\{ \alpha_1 \geq 0 : \alpha_1 < 0.55587 \} \subset \bar{\mathfrak{G}}.$$

The quadratic LP solution captures the greater part of the stability region. However, it was demonstrated that the quadratic LP method of Kumar and Meyn cannot give a sharp region. It is interesting to note that the critical utilization rates at both stations are

$$\rho_1 = \rho_2 < 9.5 \left(1 - 2\sqrt{2}/3 \right) \approx 0.543314,$$

which is far below one.

7.3 Re-entrant line without immediate feedback

When there is no immediate feedback in a two station re-entrant line, the station visitation sequence takes the simple form: 1, 2, 1, 2, ... It was shown in Dai and Weiss [12, Section 3] that for $K = 3$ when

$$\rho_1 = \alpha_1(m_1 + m_3) < 1 \quad \text{and} \quad \rho_2 = \alpha_1 m_2 < 1,$$

the conditions in Corollary 1 hold for any non-idling discipline. Therefore, we have for any non-idling discipline π,

$$\bar{\mathfrak{D}}_\pi = \bar{\mathfrak{G}} = \mathfrak{D}_0,$$

where $\mathfrak{D}_0$ is defined in (34). The argument was generalized in Dai and Vande-Vate [10] to networks with $K = 4$,

$$\bar{\mathfrak{G}} = \mathfrak{D}_0.$$

In both cases, we also have

$$\mathfrak{G} = \mathfrak{D}_0.$$

8 Moments

So far we have shown that the stability of a fluid model implies the positive Harris recurrence for the Markov process in the corresponding queueing network. In this section, we show that under some stronger moment conditions on interarrival and service times, we can obtain some stronger stability results for the queueing network. In particular, we can bound the moments of queue lengths, which are the primary performance measures of a network. Assume that

(A2') For some integer $p \geq 1$, $\mathsf{E}[\xi_\ell(1)^{p+1}] < \infty$ for $\ell \in \mathcal{E}$ and $\mathsf{E}[\eta_k(1)^{p+1}] < \infty$ for $k = 1, \ldots, K$.

Recently, Dai and Meyn [9] proved the following result.

Theorem 5. *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 $x \in \mathsf{X}$,*

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

where p is the integer used in (A2').

Assume further that (A3) holds. Then, the service discipline is stable with stationary distribution π, and moreover, for each initial condition,

(ii) *The transient moments converge to their steady state values: for $r = 1, \ldots, p$, $k = 1, \ldots, K$,*

$$\lim_{t\to\infty} \mathsf{E}_x[Q_k(t)^r] = \mathsf{E}_\pi[Q_k(0)^r] \leq \kappa_r.$$

(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: for $r = 1, \ldots, p$, $k = 1, \ldots, K$,*

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

9 Concluding remarks

For any given service discipline, it is a challenging problem to *characterize* its stability region. It appears that the piecewise linear Lyapunov function used in Botvitch and Zamyatin [1], Dai and Weiss [12] and Down and Meyn [14] is a powerful tool in determining a global stability region. For a buffer priority

discipline, Down and Meyn [14] were also able to apply the piecewise linear Lyapunov function technique. For the FIFO discipline, we refer readers to Chen and Zhang [7] and Bramson [4] for the latest developments.

Acknowledgments. I am grateful to Mike Harrison for his numerous comments on an earlier version of this paper. I thank Gideon Weiss and Sean Meyn for allowing me to cite recent joint work with them in this paper. I also thank an anonymous referee for suggesting many helpful improvements.

References

[1] D. D. Botvitch and A. A. Zamyatin, *Ergodicity of conservative communication networks*. Rapport de recherche 1772, INRIA, October 1992.

[2] M. Bramson, *Instability of FIFO queueing networks*, Annals of Applied Probability, (to appear).

[3] ——, *Instability of FIFO queueing networks with quick service times*, Annals of Applied Probability, (to appear).

[4] ——, *Private communications*. 1994.

[5] H. Chen, *Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines*, Annals of Applied Probability, (Submitted).

[6] H. Chen and A. Mandelbaum, *Hierarchical modeling of stochastic networks, Part I: fluid models*, In D.D. Yao (ed.), Applied Probability in Manufacturing Systems, (forthcoming).

[7] H. Chen and H. Zhang, *Fluid approximations and stability of multiclass queueing networks II: FIFO discipline*. in preparation.

[8] J. G. Dai, *On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models*, Annals of Applied Probability, (to appear).

[9] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Transactions on Automatic Control, (submitted).

[10] J. G. Dai and J. VandeVate, *Characterizing stability region for heterogeneous fluid models*. in preparation.

[11] J. G. Dai and Y. Wang, *Nonexistence of Brownian models of certain multiclass queueing networks*, Queueing Systems: Theory and Applications, 13 (1993), pp. 41–46.

[12] J. G. Dai and G. Weiss, *Stability and instability of fluid models for certain re-entrant lines*, Mathematics of Operations Research, (submitted).

[13] J. G. Dai and R. J. Williams, *Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons*, Theory of Probability and its Applications, (to appear).

[14] D. Down and S. Meyn, *Piecewise linear test functions for stability of queueing networks*, Proceedings of the 33rd Conference on Decision and Control, (1994). submitted.

[15] P. Dupuis and R. J. Williams, *Lyapunov functions for semimartingale reflecting Brownian motions*, Annals of Probability, (to appear).

[16] J. M. Gu, *Convergence and Performance for some Kelly-like Queueing Networks*, PhD thesis, University of Wisconsin, Madison, 1994.

[17] J. M. Harrison, *Private communications*. 1994.

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

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

[20] P. R. Kumar and S. Meyn, *Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies*. Preprint.

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

[22] S. P. Meyn and D. Down, *Stability of generalized jackson networks*, Annals of Applied Probability, 4 (1994), pp. 124–148.

[23] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes II: continuous time processes and sample chains*, Advances of Applied Probability, 25 (1993), pp. 487–517.

[24] A. N. Rybko and A. L. Stolyar, *Ergodicity of stochastic processes describing the operation of open queueing networks*, Problems of Information Transmission, 28 (1992), pp. 199–220.

[25] T. I. Seidman, *'First come, first served' can be unstable!*, IEEE Transactions on Automatic Control, (to appear).

[26] A. Stolyar, *On the stability of multiclass queueing networks*. Submitted to the Proceeding of Second Conference on Telecommunication Systems—Modeling and Analysis, Nashville, March 22–27, 1994.

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