Transience of Multiclass Queueing Networks via Fluid Limit Models

Sean P. Meyn*
University of Illinois

Abstract

This paper treats transience for queueing network models by considering an associated fluid model. If starting from any initial condition the fluid model explodes at a linear rate, then the associated queueing network with i.i.d. service times and a renewal arrival process explodes faster than any fractional power.

Keywords: Queueing networks, stability.

1 Introduction

There has been much recent interest in understanding the dynamics of queueing networks, and in particular their stability properties. Numerous techniques have been developed for verification of stability or ergodicity using a variety of methods. Of interest to us in the present paper is the recent approach based upon a fluid approximation.

Rybko and Stolyar [15] have recently examined the stability properties of a particular example by studying the properties of the associated fluid approximation. Dupuis and Williams obtained results of this kind for reflected Brownian motion [8], and these ideas were subsequently generalized in Dai [3] and Dai and Meyn [4]. These results show how to demonstrate the stability of the stochastic system by establishing the stability of a fluid approximation. In this paper we establish a converse result to obtain criteria for transience for stochastic queueing networks based upon a fluid model.

*Coordinated Science Laboratory 1308 W. Main St., Urbana, IL 61801.
Send electronic mail to: meyn@decision.csl.uiuc.edu.
Research supported in part by the NSF under Grant No. ECS 9403742, and by the University of Illinois Research Board under Grant No. 1-6-49749

Establishing transience of a queueing network appears to be at least as difficult as proving ergodicity. Several papers have appeared recently in which instability is established for a specific multiclass network (see for example Kumar et. al. [13, 10], Bramson [1] and Seidman [16]). Interestingly, in these examples transience occurs even though the usual load conditions are not violated.

The approach that we follow can be outlined as follows. For an associated fluid model, if the trajectories grow without bound, then a certain functional of the paths of the fluid model serves as a Lyapunov function. In Theorem 3.2 we show using weak convergence arguments that an analogous functional $W(n)$ for the stochastic queueing network satisfies the supermartingale property

$$\mathsf{E}_x[W(n+1) \mid \mathcal{F}_n] \leq W(n) - \frac{1}{|X(n)|^m}, \qquad n \geq 0,\ |X(n)| \geq c_0,$$

where the number $m \geq 1$ captures the rate of explosion for the fluid model, and c_0 is some postive constant. In Theorem 3.1 it is shown that this then gives an almost sure rate of explosion of the total customer population $|X(n)|$ of the network, and from this we conclude that the network is transient.

2 Network and Fluid Models

We consider a network composed of S single server stations and K buffers which are located among the server stations. We assume that there is a single exogenous arrival process with i.i.d. interarrival times $\{\xi(n), n \geq 1\}$. Customers at buffer k 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 customer at buffer k moves to buffer j with probability $p_{k,j}$, and exits the network with probability $1 - \sum_\ell p_{k\ell}$, independent of all previous history. Moreover, the nth arrival to the network enters buffer k with probability $p_{0,k}$, again independent of the history of the process. We assume that the network is *open*; that is, that all customers eventually leave the network.

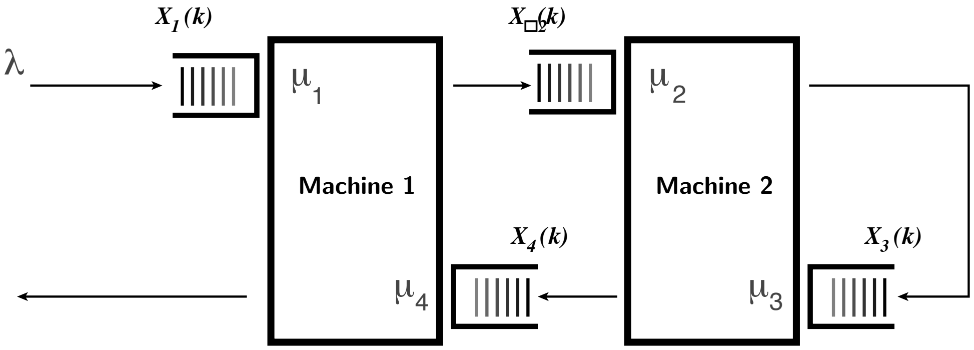

Figure 1. A Re-entrant Line

A simple example of such a network is the *re-entrant line*, in which routing is deterministic. An example is illustrated in Figure 1 consisting of two machines and four buffers, with a single server at each machine. In this example, $K = 4$ and $S = 2$.

Throughout the first part of this paper, we make the following restriction that

> $\xi, \eta_1, \ldots, \eta_K$ are mutually independent, *exponentially distributed*, i.i.d. sequences. We set $\mu_k = 1/\mathsf{E}[\eta_k(1)]$ and $\lambda = 1/\mathsf{E}[\xi(1)]$.

These assumptions have been imposed for ease of exposition – some discussion on how the exponential assumption can be relaxed is included in Section 5.

We will sample the process at virtual event times as in Lippman [12] to form a discrete time process $\mathbf{X} = \{X(n) : n \in \mathbb{Z}_+\}$ evolving on $\mathsf{X} = \mathbb{Z}_+^K$, where $\mathbf{X}$ is simply the vector of buffer lengths. The resulting process satisfies the *skip free* property

$$|X(n) - X(m)| \leq |n - m|, \qquad \text{for all } n,\, m. \tag{1}$$

We consider exclusively scheduling policies which are *state dependent*, so that the process $\mathbf{X}$ becomes a Markov chain. The norm $|\cdot|$ on the state space X will be taken to be the standard ℓ^1 norm, so that $|X(n)|$ denotes the total customer population at the time of the nth sampling.

For each $\sigma = 1, \ldots, S$ we let $C_\sigma = \{k : s(k) = \sigma\}$, and define the *nominal load* at station σ as

$$\rho_\sigma = \sum_{k \in \mathcal{C}_i} \lambda_k / \mu_k .$$

where $\{\lambda_k\}$ are found by solving the *traffic equations*

$$\lambda_k = \lambda p_{0,k} + \sum \lambda_j p_{j,k}, \qquad 1 \le k \le K.$$

It is clear that if the capacity constraint $\rho_\sigma < 1$ is violated for any $\sigma = 1, \ldots, S$, then the network will be transient, i.e. $\mathsf{P}_x\{|X(n)| \to \infty\} = 1$ for all x. We will adopt this as our definition of instability. Thus, in this paper we restrict our attention to the case when the capacity constraint is satisfied for each σ.

From the Markov chain $\mathbf{X}$ we can construct a continuous time process $\phi^x(t)$ as follows: If $|x|t$ is an integer, we set

$$\phi^x(t) = \frac{1}{|x|} X(|x|t). \tag{2}$$

For all other t, we define $\phi^x(t)$ so that it is continuous and piecewise linear in t. In view of the skip free property (1), we have for any x,

$$|\phi^x(t) - \phi^x(s)| \le |t - s|, \quad t, s \ge 0. \tag{3}$$

It follows trivially that for $p > 0$, $T < \infty$, the family of random variables

$$\{|\phi^x(t)|^p : x \ne 0, \quad 0 \le t \le T\} \qquad \text{is uniformly integrable.} \tag{4}$$

By (3) and (4) it follows that the processes $\{\phi^x : x \ne 0\}$ are tight in the function space $C[0, \infty]$.

The *fluid model* is defined to be the set of all weak limits

$$\begin{aligned} \mathit{\Phi} &= \{\phi : \phi^x \xrightarrow{w} \phi \quad \text{as } x \to \infty \text{ along some subsequence}\} \\ &= \bigcap_{n=1}^{\infty} \overline{\{\phi^x : |x| > n\}} \end{aligned}$$

where the bar denotes weak closure. Any particular $\phi \in \mathit{\Phi}$ is called a *fluid limit.*

It is shown in Dai, Chen and Mandelbaum [3, 2] that any $\phi \in \mathit{\Phi}$ satisfies a certain integral equation. For an M/M/1 queue where $S = K = 1$, the fluid model $\mathit{\Phi}$ consists of the single path

$$\phi(t) = \max[0, \phi(0) - (\mu - \lambda)t], \qquad t \ge 0,$$

where in this case $\phi(0) = |\phi(0)| = 1$. For the network described in Figure 1, the following differential equations are satisfied

$$\begin{array}{rcl} \dot{\phi}_1 & = & \lambda - \dot{T}_1\mu_1 \\ \dot{\phi}_2 & = & \dot{T}_1\mu_1 - \dot{T}_2\mu_2 \\ \dot{\phi}_3 & = & \dot{T}_2\mu_2 - \dot{T}_3\mu_3 \\ \dot{\phi}_4 & = & \dot{T}_3\mu_3 - \dot{T}_4\mu_4 \end{array} \tag{5}$$

where $\dot{T}_k$ is interpreted as the proportion of time that a server is busy on buffer k. Constraints on the variables T_k can be found through the particular scheduling policy employed.

We will see in an example below that the set of solutions to these integral or differential equations may be far larger than Φ, since in general it is difficult to completely characterize T_k. Consequently, our results will be strengthened if we focus on the true fluid limits.

For the exponential network described here, we have the following result from [4]:

Theorem 2.1 *Assume the fluid model Φ is stable, in the sense that for some fixed time t_0, and for any $\phi \in \Phi$, we have $\phi(t) = 0,\ t \geq t_0$. Then*

(i) *The transient moments converge to their steady state values: for any $r \geq 1$,*

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

(ii) *The first moment converges faster than any polynomial: for all n,*

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

(iii) *The strong law of large numbers holds: for all r,*

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

□

In the remainder of this paper we prove that a converse holds: If $\phi(t) \to \infty$ from any initial condition, then the process $X(n)$ also explodes.

3 Transience of Networks

The main result of this paper shows that the Markov chain $\mathbf{X}$ is transient whenever the fluid model Φ is unstable. The underlying idea is to construct a Lyapunov function for the fluid model, and from this construct a positive supermartingale $(W(n), \mathcal{F}_n : n \geq 0)$ for the network of the form

$$\mathsf{E}[W(X_{n+1}) \mid \mathcal{F}_n] \leq W(X_n) - g(X_n), \tag{6}$$

where $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$. The drift (6) will hold whenever X_n is "large" (when measured by W). This drift condition can also be expressed in the form

$$\mathsf{E}[\theta^1 W(X_n) - W(X_n) \mid \mathcal{F}_n] \leq -g(X_n),$$

where θ^1 is the usual backward shift operator on sample space (cf. Meyn and Tweedie [14]). It is this form that will be considered in the application below. The idea of transferring a Lyapunov function from a fluid model to the network is also used in [8, 4].

Note that (6) has the same form as Foster's criterion – it is only the nature of the set off which the drift occurs, and the use of the state dependent drift $-g(X_n)$ that is different. Note however that Foster's criterion gives positive recurrence, while the result below implies transience if g takes on strictly positive values.

We begin with the following general result, which is an extension of Theorem 8.0.2 (i) of [14].

Theorem 3.1 *Suppose that for the Markov chain* $\mathbf{X}$*, there exist positive functions* W *and* g *on* X*, and* $c_0 > 0$*, such that (6) holds whenever* $X_n \in A_{c_0} = \{x \in \mathsf{X} : W(x) \leq c_0\}$*.*

Then for all x*,*

$$\mathsf{P}_x \left\{ \sum_{n=0}^{\infty} g(X_n) < \infty \right\} \geq 1 - c_0^{-1} W(x). \tag{7}$$

Proof Let

$$\sigma = \min(n \geq 0 : X_n \in A_{c_0}^c) = \min(n \geq 0 : W(X_n) \geq c_0),$$

and define the adapted process $\{M_n : n \geq 0\}$ by $M(0) = W(X(0))$, and for $n \geq 1$,

$$M(n) = W(X(n \wedge \sigma)) + \sum_{k=0}^{(\sigma-1)\wedge(n-1)} g(X(k)),$$

where the sum $\sum_0^{-1}$ is interpreted as zero.

If $\sigma < n$ then $\sigma - 1 < n - 1$ also, so

$$M(n) = W(X(\sigma)) + \sum_{k=0}^{\sigma-1} g(X(k)) = M(\sigma) \qquad \text{on } \{\sigma < n\}.$$

Using this identity and the drift inequality (6) gives for $n \geq 1$,

$$\begin{aligned} \mathsf{E}_x[M(n) \mid \mathcal{F}_{n-1}] &= M(n-1)\mathbb{1}(\sigma < n) \\ &\quad + \mathsf{E}\left[W(X(n)) + \sum_{k=0}^{n-1} g(X(k)) \mid \mathcal{F}_{n-1}\right] \mathbb{1}(\sigma \geq n) \\ &\leq M(n-1)\mathbb{1}(\sigma < n) \\ &\quad + \left(W(X(n-1)) - g(X(n-1)) + \sum_{k=0}^{n-1} g(X(k))\right) \mathbb{1}(\sigma \geq n). \end{aligned}$$

After rearranging terms we see that $\mathsf{E}[M(n) \mid \mathcal{F}_{n-1}] \leq M(n-1)$, and hence $(M(n), \mathcal{F}_n)$ is a positive supermartingale.

From the Martingale Convergence Theorem, there exists a random variable $M(\infty)$ such that

(i) $M(n) \to M(\infty)$, $n \to \infty$, almost surely, and in the mean;

(ii) $(M(n), \mathcal{F}_n; 0 \leq n \leq \infty)$ is a positive supermartingale.

From the definition of $M(n)$, we evidentially have

$$\begin{aligned} M(\infty) &= \lim_{n\to\infty} W(X(n \wedge \sigma)) + \sum_{k=0}^{\sigma-1} g(X(k)) \\ &\geq c_0 \mathbb{1}(\sigma < \infty) + \sum_{k=0}^{\sigma-1} g(X(k)) \end{aligned} \tag{8}$$

where we have used the assumption that $W(X(n)) \geq c_0$ if $X(n) \notin A_{c_0}$.

From (8) and the supermartingale property (ii) we have

$$c_0 \mathsf{P}(\sigma < \infty) + \mathsf{E}_x\left[\sum_{k=0}^{\sigma-1} g(X(k))\right] \leq \mathsf{E}_x[M(\infty)] \leq \mathsf{E}_x[M(0))] = W(x).$$

From this bound we can infer

$$\mathsf{P}(\sigma < \infty) \leq c_0^{-1} W(x) \tag{9}$$

$$\mathsf{E}_x\left[\sum_{k=0}^{\infty} g(X(k)) \mathbb{1}(\sigma = \infty)\right] \leq W(x). \tag{10}$$

The first bound implies that $\mathsf{P}(\sigma = \infty) \geq 1 - c_0^{-1}\mathsf{E}[W(0)]$, and we have from the second bound that

$$\mathbb{1}(\sigma = \infty) \sum_{k=0}^{\infty} g(X(k)) < \infty \qquad a.s..$$

These conclusions prove the theorem. □

We may now state and prove our main result. In all of the applications we have considered, we have found that $m_0 = 1$ will suffice in the conditions of Theorem 3.2. In this case, the fluid model "explodes" at a linear rate, and the finiteness of the sum in (12) then implies that the process $X(n)$ explodes faster than any fractional power of n.

Theorem 3.2 *Suppose that for some $m_0 \geq 1$, $\delta_0 > 0$, $T_0 \geq 0$, for any $\phi \in \Phi$,*

$$|\phi(T)| \geq b(T), \quad T \geq T_0, \tag{11}$$

where $b(T) := \delta_0 \sqrt[m_0]{T}$. Then the Markov chain $\mathbf{X}$ is transient, and moreover for any $m > m_0$

$$\lim_{|x|\to\infty} \mathsf{P}_x \left\{ \sum_{k=1}^{\infty} \frac{1}{1 + |X_k|^m} < \infty \right\} = 1 \tag{12}$$

Proof To construct a function W satisfying the conditions of Theorem 3.1, set $W(x) = \mathsf{E}_x[\mathcal{W}]$, where the random variable $\mathcal{W}$ is defined as

$$\mathcal{W} = \sum_{n=|X(0)|T_0}^{|X(0)|T-1} [1 + |X(0)| + a|X(n)|]^{-m}$$

where a and $T > T_0$ are positive real numbers, and $m > m_0$. Here and in the remainder of this proof, we interpret the sum $\sum_{n=a}^{b}$ as $\sum_{a \leq n \leq b}$, even if a and b are not integers.

The random variable $\mathcal{W}$ has the appealing interpretation

$$|X(0)|^{m+1}\mathcal{W} \approx \int_{T_0}^{T} \left(\frac{1}{1 + a|\phi^x(s)|} \right)^m ds,$$

where the approximation becomes exact as $|X(0)| \to \infty$. The right hand side can be interpreted as an approximation to a Lyapunov function for the fluid limit model. Indeed, define for $\phi \in \Phi$,

$$\mathcal{V}(\phi) = \int_{T_0}^{T} \left(\frac{1}{1 + a|\phi(s)|} \right)^m ds.$$

Assuming the fluid model is unstable in the sense of Theorem 3.2, we can choose a and T so that $\mathcal{V}(\Theta^r\phi) \leq \mathcal{V}(\phi)$ for any $r > 0$. Analogous to Markov processes, the shift Θ is defined as $(\Theta^r\phi)(t) = \phi(t+r)$, $\phi \in \Phi$.

Using the representation

$$\mathcal{W} = \sum_{n=0}^{\infty}[1+|X(0)|+a|X(n)|]^{-m}\mathbb{1}(|X(0)|T > n)\mathbb{1}(|X(0)|T_0 \leq n),$$

we have

$$\begin{aligned}\theta^1\mathcal{W} &= \sum_{n=0}^{\infty}[1+|X(1)|+a|X(n+1)|]^{-m}\mathbb{1}(|X(1)|T > n)\mathbb{1}(|X(1)|T_0 \leq n)] \\ &= \sum_{n=1+|X(1)|T_0}^{|X(1)|T}[1+|X(1)|+a|X(n)|]^{-m}.\end{aligned}$$

We now break the difference $\theta^1\mathcal{W} - \mathcal{W}$ into three terms which are considered separately:

$$\theta^1\mathcal{W} - \mathcal{W} = A + B + C \tag{13}$$

where

$$\begin{aligned}A &= -[1+|X(0)|+a|X(T_0|X(0)|)]^{-m} &(14)\\ B &= \sum_{n=1+|X(1)|T_0}^{|X(0)|T-1}\Big\{[1+|X(1)|+a|X(n)|]^{-m} - [1+|X(0)|+a|X(n)|]^{-m}\Big\} &(15)\\ C &= \sum_{n=|X(0)|T}^{|X(1)|T}[1+|X(1)|+a|X(n)|]^{-m} &(16)\end{aligned}$$

The first term provides a negative contribution:

$$\liminf_{|X(0)|\to\infty}|X(0)|^m A \leq \inf_{\phi\in\Phi}\frac{-1}{(1+a|\phi(T_0)|)^m} \leq \frac{-1}{(1+a(1+\lambda T_0))^m}$$

We will show that this term dominates the remaining terms for suitable choices of a and T, whenever the initial condition is suitably large.

The second term is bounded using the Mean Value Theorem and the skip free property (1),

$$\begin{aligned}B &\leq \sum_{n=1+|X(1)|T_0}^{|X(0)|T-1}\Big\{[1+(|X(0)|-1)+a|X(n)|]^{-m} - [1+|X(0)|+a|X(n)|]^{-m}\Big\} \\ &\leq \sum_{n=1+|X(1)|T_0}^{|X(0)|T-1} m[|X(0)|+a|X(n)|]^{-m-1}.\end{aligned}$$

Multiplying both sides by $|X(0)|^m$, we see that

$$|X(0)|^m B \leq \frac{m}{|X(0)|} \sum_{n=1+|X(1)|T_0}^{|X(0)|T-1} \left[1 + a\frac{|X(n)|}{|X(0)|}\right]^{-m-1}. \tag{17}$$

From weak convergence, the family of random variables $\{|X(0)|^m B : |X(0)| \in \mathsf{X}\}$ is tight, and any weak limit may be bounded by a random variable of the form

$$m \int_{T_0}^{T} [1 + a|\phi(s)|]^{-(m+1)}\, ds,$$

where ϕ is possibly random, taking values in Φ. Since the right hand side of (17) is bounded, we have

$$\begin{aligned}\limsup_{|X(0)|\to\infty} \mathsf{E}_x[|X(0)|^m B] &\leq \sup_{\phi\in\Phi}\Bigl\{m \int_{T_0}^{T} [1 + a|\phi(s)|]^{-(m+1)}\, ds\Bigr\} \\ &\leq m \int_{T_0}^{T} [1 + ab(s)]^{-(m+1)}\, ds\end{aligned}$$

We now bound the third term. Since $C \leq 0$ when $|X(1)| \leq |X(0)|$, we only have to consider the case where $|X(1)| = |X(0)| + 1$. It then follows that

$$C \leq (T+1)[1 + |X(0)| + 1 + a(|X(n) - T|)]^{-m}.$$

Multiplying by $|X(0)|^m$ and taking limits gives a bound on the final term,

$$\limsup_{|X(0)|\to\infty} \mathsf{E}_x[|X(0)|^m C] \leq \sup_{\phi\in\Phi}(T+1)[1 + a|\phi(T)|]^{-m} \leq (T+1)[1 + ab(T)]^{-m}.$$

Putting these three bounds together gives

$$\begin{aligned}\limsup_{|X(0)|\to\infty} |X(0)|^m \mathsf{E}_x[\theta^1 \mathcal{W} - \mathcal{W}] &\leq -[1 + a(1+\lambda T_0)]^{-m} + m \int_0^{\infty} [1 + ab(s)]^{-(m+1)}\, ds \\ &\quad +(T+1)[1 + ab(T)]^{-m}. \end{aligned} \tag{18}$$

We may now specify a and T. First, choose a so large that

$$m \int_0^{\infty} [1 + ab(s)]^{-(m+1)}\, ds \leq \frac{1}{3}[1 + a(1+\lambda T_0)]^{-m}.$$

This is possible because of the larger exponent in the integrand on the left hand side. With a fixed, choose T so large that

$$\frac{(T+1)}{(1+ab(T))^m} \leq \frac{1}{3}\frac{1}{[1 + a(1+\lambda T_0)]^{-m}}$$

This is possible because of our assumption that $m > m_0 \geq 1$, and the definition of $b(T)$.

Hence from (18) we have

$$\limsup_{|X(0)|\to\infty} |X(0)|^m (PW(x) - W(x)) \leq -\frac{1}{3}\frac{1}{[1 + a(1 + \lambda T_0)]^{-m}} < 0, \tag{19}$$

which shows that the conditions of Theorem 3.1 hold with $g(x) = \text{const.}/(1 + |x|)^m$.

□

4 Examples

We now give to examples to illustrate the main results given above. It will be seen that some care must be taken when working with the fluid limit model.

Fluid integral equations do not characterize fluid limits Here we give an example of a network for which the differential equations used to describe the fluid limits admit solutions which tend to zero, even though the weak limits themselves always tend to infinity.

Consider the network described in Figure 1 under the following conditions

(i) The load conditions are satisfied:

$$\rho_1 = \frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_4} < 1, \qquad \rho_2 = \frac{\lambda}{\mu_2} + \frac{\lambda}{\mu_3} < 1.$$

(ii) Buffers 2 and 4 have priority at their respective machines.

(iii) The orderings $\mu_1 > \mu_2$; $\mu_3 > \mu_4$ are satisfied, and

$$\frac{\lambda}{\mu_2} + \frac{\lambda}{\mu_4} > 1$$

Under (ii) and (iii), priority is given to slow queues, and this causes alternate starvation of the machines, resulting in instability. This is shown in [13] for a deterministic model of the network. We give a proof of this result for the fluid model using a Lyapunov function approach.

Before we begin, we note that for any $\phi \in \Phi$, after either buffer 2 or 4 empties, these buffers can never again be active simultaneously until the fluid model empties.

For example, at the time that buffer 2 empties, buffer 4 becomes busy and is fed by buffer 3, which sends fluid to buffer 4 strictly faster than it can be processed. Since buffer 4 has priority over buffer 1, work to buffer 2 is completely cut off so that $\phi_3(t) = \dot{\phi}_3(t) = 0$ until buffer 4 finally empties. At this time, work moves from buffer 1 to buffer 2, and an analogous starvation of buffer 4 occurs. This phenomenon can be proved rigorously following Lemma 5.1 of [7]. It follows that after some finite transient, whenever the fluid model is nonempty,

$$\mathbb{1}(\phi_2(t) > 0) + \mathbb{1}(\phi_4(t) > 0) \leq 1, \qquad t \in \mathbb{Z}_+.$$

Define the *work* destined for buffers 2 and 4 as follows

$$W_2(t) = (\phi_1(t) + \phi_2(t))/\mu_2, \qquad W_4(t) = (\phi_1(t) + \phi_2(t) + \phi_3(t) + \phi_4(t))/\mu_4.$$

For example, the quantity $W_2(t)$ is the total amount of time that buffer 2 must spend to process the fluid which is in the system at time t. Letting $V(t) = W_2(t) + W_4(t)$, we have from the previous arguments that after buffer 2 or 4 first empties, for almost every t,

$$\frac{d}{dt}V(t) = \frac{\lambda}{\mu_2} - \mathbb{1}(\phi_2(t) > 0) + \frac{\lambda}{\mu_4} - \mathbb{1}(\phi_4(t) > 0) \geq \frac{\lambda}{\mu_2} + + \frac{\lambda}{\mu_4} - 1 > 0$$

On integrating both sides of this inequality, one sees that the conditions of Theorem 3.2 are satisfied with $m_0 = 1$.

The fluid model differential equations are given by (5). Because of the form of the buffer priority policy, $\dot{T}_2(t) = 1$ if $\phi_2(t) > 0$, and $\dot{T}_4(t) = 1$ if $\phi_4(t) > 0$. It is easy to constuct solutions to these equations which tend to zero if the constraint that $\phi_2(t)\phi_4(t)$ is eventually zero is removed. When attempting to establish transience through the fluid model, we see that one must be careful to eliminate any extraneous solutions to the fluid limit differential equations.

Drift vectors do not characterize fluid limits For a network of the form described here, it is popular to address stability through an analysis of the *drift vectors* defined by

$$\mathit{\Delta}(x) = \mathsf{E}[X(k+1) - X(k) \mid X(k) = x], \qquad x \in \mathsf{X}.$$

Two examples of skip-free random walks on $\mathbb{R}^3$ are analyzed in [9] for which stability can be addressed using a fluid model following the approach described in this paper. The vector field $\Delta(x)$ is identical in these two examples, yet one is transient with the process exploding along the x_2 axis, and the other is positive recurrent. The fluid models are of course very different. In the stable case, the fluid model approaches zero along the x_2 axis, while for the unstable model, the fluid limits explode at a linear rate along this axis.

Hence, the drift vectors *do not* describe the motion of the fluid model.

5 Generalizations and Conclusions

The assumptions imposed in Theorem 3.2 are just what one would expect from an unstable network, particularly with the choice of $m_0 = 1$. If the network is unstable, then a linear rate of explosion seems reasonable; and if the network explodes from one initial condition then, by irreducibility, one would expect the same behavior from all initial conditions. This is precisely what is observed for the M/M/1 queue when $\rho > 1$, and more elaborate examples are treated in [6, 7, 5, 11].

The main difficulty with this result is in verification. Ideally, the result would state that the queueing network is transient if the fluid model explodes from just one initial condition. Extensions in this direction are currently under investigation.

We have restricted to state dependent policies, but this can often be relaxed. More complex scheduling policies can be treated as long as a Markov state process and a corresponding fluid model may be constructed. We refer the reader to [4] for further discussion.

One strong assumption imposed in this paper is the distributional condition on the arrival and service processes. This is not essential, although it does greatly simplify the exposition. To generalize Theorem 3.2 to general i.i.d. services with a single renewal input, sample the process at the arrival epochs to form a general state space Markov chain $X(n) = \binom{Q(n)}{R(n)}$, where $Q(n)$ denotes the buffer of queue lengths, and $R(n)$ denotes the vector of residual service times, all at the time of the nth arrival. Sampling in this way preserves the *upper bound*,

$$Q(n+m) - Q(n) \leq m, \qquad n, m \in \mathbb{Z}_+,$$

and a stochastic lower bound may also be found. A fluid model may be constructed as before, although it is slightly more complex due to the fact that the residual service times introduce a delay [4]. In spite of this added complexity, the proof of Theorem 3.2 goes through in essentially the same way as presented here.

The assumptions of Theorem 3.2 can be verified under general conditions by solving an associated linear program [6, 7]. Although this method completely characterizes stability for all of the examples that we have investigated, it is not known if this approach characterizes stability in general. Such issues are also currently being explored.

Acknowledgement This paper was begun while visiting Ludwig Arnold at the *Institut Für Dynamische Systeme, the University of Bremen*, and Rolf Shaßberger, Serguei Foss, and Günter Last at the *Technical University of Braunschweig*. This research was largely completed at the *11th International Conference on Analysis and Optimization*, held in Sophia Antipolis, June 15–17, 1994, which was organized by Guy Cohen and Jean-Pierre Quadrat. The author is grateful to these individuals and their instituions for their support.

V. Dumas discovered an error in an earlier draft of the paper, and D. Down together with the annonymous reviewers supplied numerous comments. Their helpful advice is gratefully appreciated.

Research was supported in part by National Science Foundation grants ECS 94-03742 and ECS 92-16487, and by University of Illinois Research Board grant 1-6-49749.

REFERENCES

[1] M. Bramson. Instability of FIFO queueing networks. To appear in *Annals of Applied Probability*, 1995.

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

[3] J. G. Dai. On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.*, 5:49–77, 1995.

[4] J. G. Dai and S.P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Automat. Control*, 40:1–16, November 1995.

[5] J. G. Dai and G. Weiss. Stability and instability of fluid models for certain re-entrant lines. *Math. Operations Res.*, 1995. to appear.

[6] D. Down. *Stability of queueing networks*. PhD thesis, University of Illinois, Urbana, IL, 1994.

[7] D. Down and S. P. Meyn. The structure of fluid models and instability of queueing networks. to appear at the *European Control Conference 1995*, Rome, Italy, 1994.

[8] P. Dupuis and R. J. Williams. Lyapunov functions for semimartingale reflecting Brownian motions. *Ann. Appl. Probab.*, 22(2):680–702, 1994.

[9] M. Kotler. Drift vectors are not sufficient to determine recurrence of a markov chain on $\mathbb{Z}_+^3$. *J. Appl. Probab.*, 31:1098–1102, 1994.

[10] P. R. Kumar and T. I. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control*, AC-35(3):289–298, March 1990.

[11] S. Kumar and P. R. Kumar. Fluctuation smoothing policies are stable for stochastic re-entrant lines. To appear in *Proceedings of the 33rd IEEE Conference on Decision and Control*, December 1994.

[12] S. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.

[13] S. H. Lu and P. R. Kumar. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, 36(12):1406–1416, December 1991.

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

[15] A. N. Rybko and A. L. Stolyar. On the ergodicity of stohastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii*, 28:3–26, 1992.

[16] T. I. Seidman. First come first serve can be unstable. *IEEE Transactions on Automatic Control*, 39(10):2166–2170, October 1994.