MATHEMATICS OF OPERATIONS RESEARCH
Vol. 21, No. 1, February 1996
Printed in U.S.A.

STABILITY AND INSTABILITY OF FLUID MODELS FOR REENTRANT LINES

J. G DAI AND G. WEISS

Reentrant lines can be used to model complex manufacturing systems such as wafer fabrication facilities. As the first step to the optimal or near-optimal scheduling of such lines, we investigate their stability. In light of a recent theorem of Dai (1995) which states that a scheduling policy is stable if the corresponding fluid model is stable, we study the stability and instability of fluid models. To do this we utilize piecewise linear Lyapunov functions. We establish stability of First-Buffer-First-Served (FBFS) and Last-Buffer-First-Served (LBFS) disciplines in all reentrant lines, and of all work-conserving disciplines in any three buffer reentrant lines. For the four buffer network of Lu and Kumar we characterize the stability region of the Lu and Kumar policy, and show that it is also the *global* stability region for this network. We also study stability and instability of Kelly-type networks. In particular, we show that *not* all work-conserving policies are stable for such networks; however, all work-conserving policies are stable in a ring network.

1. Introduction. Consider a multiclass queueing network with a single route for all customers. There are I stations (nodes) in the network. All the customers follow a fixed deterministic K stage route through the network. We shall number the stages so that each customer will enter the system in stage 1, and on completion of stage k will move to stage $k+1, k=1,\ldots,K-1$, and leave the system on completion of stage K. We designate those customers on the kth stage of the route (the kth visit along the route) as class k customers. We envision class k customers waiting in buffer k, which is assumed to have infinite capacity (when customers are served in FIFO rule, it is enough to have one buffer for each station). For each k let $\sigma(k)$ be the station number that class k customers visit (the station serving stage k). This model is called a reentrant line, see Kumar (1993). A distinctive feature of a reentrant line is that customers may visit a particular station more than once, so that each node serves several classes.

Let $\{\xi(n),\ n \geq 1\}$ be a sequence of positive random variables. The nth random variable $\xi(n)$ is interpreted as the interarrival time between the $(n-1)$th customer arrival and the nth customer arrival from outside; the first customer arrives at time $\xi(1)$. The service times at different stages (visits) for the nth customer are $\eta_1(n),\ldots,\eta_K(n)$. We make the following three assumptions. First we assume that

$$\{(\xi(n),\eta_1(n),\ldots,\eta_K(n)),\ n \geq 1\} \quad \text{is an iid sequence.} \tag{1.1}$$

Then we assume that all the random variables have finite first moments. That is,

$$E[\xi(1)] < \infty \quad \text{and} \quad E[\eta_k(1)] < \infty \quad \text{for } k = 1,\ldots,K. \tag{1.2}$$

Received February 25, 1994; revised November 20, 1994.
AMS 1991 subject classification. Primary: 60K25; Secondary: 60K20, 90B22, 90B35.
OR/MS Index 1978 subject classification. Primary: 697 Queues/Networks.
Key words. Stability, unstable networks, fluid models, piecewise linear Lyapunoc functions, reentrant lines, multiclass queueing networks, scheduling policies, Harris recurrence.

0364-765X/96/2101/0115/$01.25

Finally, we assume that the interarrival times are unbounded and their distribution is spread out. That is, for any $x > 0$,

$$\text{Proba}\{\xi(1) \geq x\} > 0, \tag{1.3}$$

and for some integer $n > 0$ and some function $p(x) \geq 0$ on $\mathbb{R}_+$ with $\int_0^\infty p(x)\,dx > 0$,

$$\text{Proba}\left\{a \leq \sum_{i=1}^{n} \xi(i) \leq b\right\} \geq \int_a^b p(x)\,dx, \quad \text{for any } 0 \leq a < b. \tag{1.4}$$

Note that in (1.1), we allow the service times at different stages of visits to have arbitrary dependency. This feature is useful for certain applications, notably in computer communications and manufacturing systems. There the length of a computer message or the size of a manufacturing lot may be random. However, the service times in general are proportional to the message length or lot size, and therefore are positively correlated. We also allow dependence of the service times on the previous interarrival time.

Let $C_i = \{k\colon \sigma(k) = i\}$. The set C_i is called the *constituency* of station i. Let C be the $I \times K$ incidence matrix,

$$C_{ik} = \begin{cases} 1 & \text{if } \sigma(k) = i, \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

Without loss of generality, we assume that

$$E[\xi(1)] = 1. \tag{1.6}$$

Let $m_k = E[\eta_k(1)]$ be the mean service time for class k customers. We assume that there is a single reliable server at each station. For each $i = 1, \ldots, I$, let

$$\rho_i = \sum_{k \in C_i} m_k.$$

We call ρ_i the *nominal workload* for server i per unit of time (recall that the arrival rate is normalized to be one in (1.6)). Throughout this paper, we assume

$$\rho_i < 1 \quad \text{for } i = 1, \ldots, I. \tag{1.7}$$

Nothing has been said yet about a queueing discipline, which dictates the order in which customers are served at each station (we will use queueing discipline and scheduling policy interchangeably). Specific queueing disciplines will be discussed in later sections. We assume that all disciplines are *work-conserving*. That is, server i works at full speed whenever there is work to do at station i.

In Dai (1995), the author introduced a stochastic process $\{X(t),\ t \geq 0\}$ that describes the dynamics of the queueing network under a specific queueing discipline. For each $t \geq 0$, $X(t) = (X_1(t), \ldots, X_I(t))$, where $X_i(t)$ is the state at station i at time t. The exact definition of state depends on the particular queueing discipline. For example, if first-in-first-out (FIFO) discipline is used at each station, then one needs to take

$$X_i(t) = \left(\left(x_{i1}(t), \ldots, x_{in_i(t)}(t)\right), u(t), v_{x_{i1}(t)}(t)\right),$$

where $n_i(t)$ is the queue length (including the possible one being served) at station i, $x_{ij}(t)$ is the class number of the jth customer at the station, $u(t)$ is the residual time for the next external customer to arrive, $v_{x_{i1}(t)}(t)$ is the residual service time for the customer being serviced; if $n_i(t) = 0$, $v_{x_{i1}(t)}(t)$ is taken to be zero. It was shown in Dai (1995) that under conditions (1.1) and (1.2), $X = \{X(t), t \geq 0\}$ is a strong Markov process. Readers are referred to §2.2 of Meyn and Tweedie (1993) or §3 of Dai (1995) for the definition of positive Harris recurrence for a strong Markov process.

DEFINITION 1.1. A queueing discipline is *stable* if the underlying Markov process $X = \{X(t), t \geq 0\}$ describing the dynamics of the network is positive Harris recurrent.

Positive recurrence for X implies the existence of a unique stationary distribution ϕ for X. It also implies the strong law of large numbers for the sample paths of X. In particular,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t Q_k(s)\,ds = \int_S f_k(x)\,\phi(dx) \quad \text{almost surely}$$

for any initial configuration, where $Q_k(t)$ is the queue length of class k customers at time t, and $f_k(x)$ is the queue length of class k customers for a state x in the state space S.

For a long time, researchers have believed that condition (1.7) is necessary and sufficient for a queueing discipline to be stable. This premise was based on the study of generalized Jackson networks and some special multiclass networks. In a generalized Jackson network, also called a single class network, all the customers that visit a particular node or station of the network are essentially indistinguisable. When all interarrival time and service time distributions are exponential, Jackson (1957) found the stationary distribution for the network and consequently established the stability. When these distributions are general, various stability results were obtained; see Borovkov (1986), Sigman (1990), Meyn and Down (1994), Baccelli and Foss (1994) and Chang, Thomas and Kiang (1994). Reentrant lines belong to the wider class of so called multiclass or non-homogeneous customer networks, because although all customers follow a single route, stations may contain customers at different stages along their route. Under some special distributional assumptions on interarrival times and service times, some scheduling disciplines in multiclass networks like BCMP (1975) and Kelly (1979) have been shown to be stable again by explicitly finding the stationary distributions.

The belief that (1.7) is sufficient for stability has been shattered by a series of brilliant examples considering multiclass networks. Notably, Bramson (1994) has recently presented an example of a two station reentrant line, with exponential interarrival and service distributions, for which (1.7) holds and yet FIFO is unstable. In an earlier paper, Lu and Kumar (1991) have shown that a particular buffer priority discipline is unstable for a two station reentrant line with *deterministic* arrivals and services. (We analyze this network in §5 of this paper.) Analogous instability results were obtained in Rybko and Stolyar (1992) and Seidman (1994), for more general (not reentrant) networks with multi-type customer arrivals. Readers are referred to §1 of Dai (1995) for a more detailed account of recent developments of the subject. In the following definition, $\mu_0 = 1$, $\mu_k = 1/m_k$ for $k = 1, \ldots, K$ and $T_0(t) = t$ for $t \geq 0$.

DEFINITION 1.2. By a fluid model of a queueing discipline in a reentrant line we mean any solution $(Q(\cdot), T(\cdot))$ to the following equations, where $Q(t) = (Q_1(t), \ldots, Q_K(t))'$ and $T(t) = (T_1(t), \ldots, T_K(t))'$.

$$Q_k(t) = Q_k(0) + \mu_{k-1}T_{k-1}(t) - \mu_k T_k(t) \quad \text{for } k = 1, \ldots, K, \tag{1.8}$$

$$Q_k(t) \geq 0 \quad \text{for } k = 1, \ldots, K, \tag{1.9}$$

$$T_k(0) = 0 \quad \text{and} \quad T_k(\cdot) \text{ is nondecreasing for } k = 1, \ldots, K, \tag{1.10}$$

$$B_i(t) = \sum_{k \in C_i} T_k(t) \quad \text{for } i = 1, \ldots, I, \tag{1.11}$$

$$U_i(t) = t - B_i(t) \quad \text{is nondecreasing for } i = 1, \ldots, I, \tag{1.12}$$

$$xU_i(\cdot) \text{ increases only at times } t \text{ when } \sum_{k \in C_i} Q_k(t) = 0 \quad \text{for } i = 1, \ldots, I, \tag{1.13}$$

(1.14) some additional conditions on $(Q(\cdot), T(\cdot))$ that are specific to the queueing discipline.

In the queueing network context, $Q_k(t)$ is the queue length for class k at time t with initial class k queue length $Q_k(0)$, $T_k(t)$ is the cumulative amount of time the server $\sigma(k)$ spends on class k customers in $[0, t]$, $B_i(t)$ is the cumulative amount of time that server i is busy in $[0, t]$, $U_i(t)$ is the cumulative amount of time that server i is idle in $[0, t]$. Condition (1.13) is the work-conserving assumption we made earlier. It is equivalent to

$$\int_0^\infty \left(\sum_{k \in C_i} Q_k(t) \right) dU_i(t) = 0 \quad \text{for } i = 1, \ldots, I.$$

Note that in (1.8) the arrival process is replaced by the deterministic continuous fluid flow $\mu_0 T_0(t) = t$, and the cumulative service completions at buffer k are replaced by the deterministic continuous $\mu_k T_k(t)$.

For future reference, by a *work-conserving fluid model* we mean any solution satisfying (1.8)–(1.13). It was shown in §4 of Dai (1995) that any fluid limit of a queueing network under a work-conserving queueing discipline is a fluid model satisfying (1.8)–(1.13) in Definition 1.2.

For a particular queueing discipline, the corresponding fluid model may have additional complementary conditions (1.14). *These additional conditions must be justified by taking "fluid limit" from corresponding conditions of the queueing network.* This will be done in this paper, for some buffer priority policies, see (4.4) in §4; more details and further examples are contained in Dai (1995). Even under these additional constraints, in general (1.8)–(1.14) may not uniquely determine $(Q(\cdot), T(\cdot))$. In the nonunique case, the queueing network is sensitive to the initial configuration. That is, a slight change of initial network configuration (negligible under fluid scaling) will completely change the dynamics of the network. Readers are referred to §5 below, and to the examples of Whitt (1993) for more insight. For $q = (q_1, \ldots, q_K)' \in \mathbb{R}_+^K$, let $|q| \equiv \sum_{k=1}^K q_k$.

DEFINITION 1.3. The fluid model corresponding to a queueing discipline is stable if there exists a time $\delta > 0$ such that for any solution $Q(\cdot)$ satisfying (1.8)–(1.14) and $|Q(0)| = 1$,

$$Q_k(t) \equiv 0, \quad t \geq \delta, \quad k = 1, \ldots, K.$$

REMARK. Recently, Stolyar (1994) proved that this definition of stability is equivalent to an apparently weaker condition that for each initial condition, there exists a t such that $\|Q(t)\| < \|Q(0)\|$, where $\|Q(t)\|$ is some norm, e.g. total queue length, of $Q(t)$.

The following theorem was proved in Theorem 4.3 of Dai (1995). The refinement to the current form is due to Theorem 5.2 of Chen (1995).

THEOREM 1.1. *A queueing discipline is stable if the corresponding fluid model is stable.*

REMARK. The theorem was proved by Rybko and Stolyar (1992) for a two station network with exponential interarrival and service distributions. Related work can also be found in Stolyar (1994).

The primary purpose of this paper is to use Theorem 1.1 to investigate the stability and instability of various queueing disciplines under (1.7). We show that for three buffer reentrant lines any work-conserving policy is stable. For buffer priority disciplines, we study the flow in different segments of a network. As a consequence, we prove that in any reentrant line, First-Buffer-First-Served (FBFS) and Last-Buffer-First-Served (LBFS) disciplines are stable. This result is analogous to the one by Kumar (1993), who considered discrete deterministic systems. The most interesting result of this paper is perhaps the *characterization* of the stability region of the Lu-Kumar network (1991). We also study the stability of Kelly-type networks. In particular, we show that instability can occur in a fluid model even if mean service times at different visits to a station are the same. Both the fluid version and queueing network version of this result were recently proved by Gu (1995). However, when a Kelly-type network has ring topology, we show that all work-conserving policies are stable.

Obviously, stability is a first issue one needs to address if one wishes to study optimal or near-optimal scheduling of a reentrant line. However, to determine the stability region for any given discipline in a multiclass network seems very difficult at the moment. It is still an open question whether the stability regions of a queueing network and the corresponding fluid network are the same. We hope that examples treated in this paper demonstrate that working with fluid models is a viable method to solve stability questions of queueing networks.

2. Preliminaries. In this section, we collect some intermediate results and introduce some more notation. It is clear from (1.10)–(1.12) that $T_k(\cdot)$, $B_i(\cdot)$ and $U_i(\cdot)$ are Lipschitz continuous. Hence we have the following proposition.

PROPOSITION 2.1. *The paths $Q_k(\cdot)$, $T_k(\cdot)$, $B_i(\cdot)$ and $U_i(\cdot)$ are absolutely continuous. Therefore they have derivatives almost everywhere with respect to the Lebesgue measure on* $[0, \infty)$.

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

REMARK. In later sections, whenever a derivative of a path at time t is considered, it is always assumed that t is a regular point for Q, T and B.

The following elementary lemma is useful.

LEMMA 2.2. *Let g be an absolutely continuous nonnegative function, and let $\dot{g}$ denote its derivative, whenever it exists.*

(i) If $g(t) = 0$ and $\dot{g}(t)$ exists, then $\dot{g}(t) = 0$.

(ii) Assume the condition that for some $\epsilon > 0$ and almost everywhere at regular points $t > 0$, whenever $g(t) > 0$ then $\dot{g}(t) < -\epsilon$. Then $g(t) = 0$ for all $t \geq \delta$, where $\delta = g(0)/\epsilon$. Furthermore, $g(\cdot)$ is nonincreasing, and hence once it reaches zero it stays there forever.

For future references, we define some processes related to the queue length process. First, let

$$W(t) = CMQ(t), \tag{2.1}$$

where $M = \text{diag}(m_1, \ldots, m_K)$. The ith component of $W(t)$ is

$$W_i(t) = \sum_{k \in C_i} m_k Q_k(t).$$

In the queueing network context, $W_i(t)$ is interpreted as the expected immediate workload at station i given that the queue length at time t is $Q(t)$. In our model, we shall call $W_i(t)$ *immediate volume* or simply *volume* at station i at time t. If no more fluids arrive at any of the buffers k, $k \in C_i$, after time t, server i will be busy $W_i(t)$ units of time to clear out all the fluids currently at station i. For future reference, we define

$$Q_k^+(t) = \sum_{l=1}^{k} Q_l(t), \quad \text{for } k = 1, \ldots, K. \tag{2.2}$$

3. A three buffer reentrant line. Consider a three buffer two station reentrant line pictured in Figure 1. Using the notation set up in §1, we have $I = 2$ and $K = 3$. The condition (1.7) now reads

$$\rho_1 = m_1 + m_3 < 1 \quad \text{and} \quad \rho_2 = m_2 < 1. \tag{3.1}$$

For this network, Kumar (1993) conjectured that FIFO queueing discipline is stable under (3.1) when all distributions are exponential. Wang (1993) proved that the corresponding fluid network is stable under FIFO discipline, which, by Theorem 1.1, confirms the conjecture. In this section we prove stability under any work-conserving queueing discipline.

THEOREM 3.1. *For the three buffer two station reentrant line, if* (3.1) *holds then the fluid model is stable for every work conserving policy.*

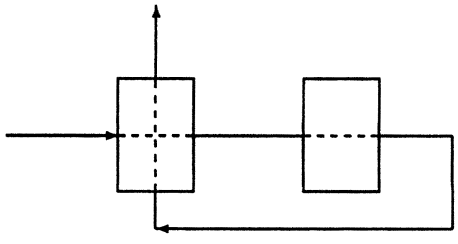

FIGURE 1. A three buffer two station reentrant line.

PROOF. This theorem can be proved directly by considering the two cases (a) $m_1 + m_2^{-1}m_3 < 1$ and (b) $m_1 + m_2^{-1}m_3 \geq 1$ separately; one can easily find a sharp upper bound on the emptying time for each case. The proof that we present here is shorter, and has wider applications. The key idea is to construct an appropriate Lyapunov function that is piecewise linear in terms of the K-dimensional queue length processes. This form of Lyapunov function was first advanced by Botvitch and Zamyatin (1992) and has the potential to be further generalized. In later sections, we make further use of this type of Lyapunov function.

Recall that $Q_k^+(t) = \Sigma_{l=1}^k Q_l(t)$. Let $\theta = m_1/(m_1 + m_3)$. Because $\rho_1 < 1$, we have $m_1 < \theta < 1$. Define

$$G_1(t) = \theta Q_1^+(t) + (1 - \theta)Q_3^+(t),$$

$$G_2(t) = Q_2^+(t).$$

On $W_1(t) = 0, G_1(t) = (1 - \theta)Q_2^+(t) \leq G_2(t)$. Similarly, on $W_2(t) = 0, G_2(t) = Q_1^+(t) \leq Q_1^+(t) + (1 + \theta)Q_3(t) = G_1(t)$. Note that, from (1.8),

$$G_1(t) = G_1(0) + t - \theta\mu_1 T_1(t) - (1 - \theta)\mu_3 T_3(t)$$

$$= G_1(0) + t - B_1(t)/\rho_1.$$

Because $\dot{B}_1(t) = 1$ when $W_1(t) > 0$, we have $\dot{G}_1(t) = -(1/\rho_1 - 1) < 0$ whenever $W_1(t) > 0$. Similarly, $\dot{G}_2(t) = -(1/\rho_2 - 1) < 0$ whenever $W_2(t) > 0$. Let

$$G(t) = \max\{G_1(t), G_2(t)\}$$

and t be a regular point of $G(\cdot)$, $G_1(\cdot)$ and $G_2(\cdot)$. Assume that $G(t) > 0$. It follows from Lemma 3.2 below that $\dot{G}(t) \leq -\epsilon$, where

$$\epsilon = \min\{1/\rho_1 - 1, 1/\rho_2 - 1\} > 0.$$

Therefore, by Lemma 2.2, $G(t) \equiv 0$ for $t \geq G(0)/\epsilon$. We assume initial total queue length $|Q(0)| = 1$. This implies that $G(0) \leq 1$, and hence $G(t) \equiv 0$ for $t \geq \max\{\rho_1/(1 - \rho_1), \rho_2/(1 - \rho_2)\}$. But $G(t) \equiv 0$ implies $Q(t) \equiv 0$. □

REMARK. The line $1 \to 2 \to 1$ is the most complex three buffer reentrant line. Because any other three buffer reentrant line is a feedforward network in a sense defined in §6 of Dai (1995), it follows from §§5 and 6 of Dai (1995) that any other three buffer reentrant line will also be stable under every work conserving policy.

LEMMA 3.2. *Let $G_i(t)$ be a nonnegative linear function of $(Q_1(t), \ldots, Q_K(t))$, $i = 1, \ldots, I$. Assume the following two properties*:

(a) *For each $i = 1, \ldots, I$, there exists $\epsilon_i > 0$ such that $W_i(t) > 0$ implies $\dot{G}_i(t) \leq -\epsilon_i$.*

(b) *For each $i = 1, \ldots, I$, if $W_i(t) = 0$ then $G_i(t) \leq \min_{j \neq i} G_j(t)$.*

Let $G(t) = \max\{G_1(t), \ldots, G_I(t)\}$. Then $G(\cdot)$ is an absolutely continuous nonnegative function. Furthermore, if t is a regular point of $G(\cdot), G_1(\cdot), \ldots, G_I(\cdot)$ such that $G(t) > 0$, then $\dot{G}(t) \leq -\epsilon$, where $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_I\}$.

PROOF. Because each $G_i(\cdot)$ is Lipschitz continuous, so is $G(\cdot)$. Hence $G(\cdot)$ is an absolutely continuous nonnegative function. Let t be a regular point of $G(\cdot), G_1(\cdot), \ldots, G_I(\cdot)$ such that $G(t) > 0$. Assume that

$$G_{i_1}(t) = \cdots = G_{i_k}(t) = G(t) \quad \text{and} \quad G_i(t) < G(t) \text{ for } i \notin \{i_1, \ldots, i_k\}.$$

By the regularity of $G(\cdot), G_{i_1}(\cdot), \ldots, G_{i_k}(\cdot)$ at t, we have

$$\dot{G}(t) = \dot{G}_{i_1}(t) = \cdots = \dot{G}_{i_k}(t). \tag{3.2}$$

To see (3.2), (which is a well-known property of the maximum function) consider i_j, $1 \le j \le k$ and choose two sequences $\{t_n\}$ and $\{s_n\}$ such that $s_n < t, t_n > t$, and $s_n \to t, t_n \to t$. Then

$$\dot{G}(t) = \lim_{n\to\infty} \frac{G(s_n) - G(t)}{s_n - t} \le \lim_{n\to\infty} \frac{G_{i_j}(s_n) - G_{i_j}(t)}{s_n - t} = \dot{G}_{i_j}(t),$$

$$\dot{G}(t) = \lim_{n\to\infty} \frac{G(t_n) - G(t)}{t_n - t} \ge \lim_{n\to\infty} \frac{G_{i_j}(t_n) - G_{i_j}(t)}{t_n - t} = \dot{G}_{i_j}(t).$$

Because $G(t) > 0$, there must be some j for which $W_j \ne 0$, and by property (b), if $k < I$, then $W_i = 0$ implies $i \notin \{i_1, \ldots, i_k\}$. Hence there must be $1 \le j \le k$ such that $W_{i_j}(t) > 0$. But this implies that one of the inequalities $\dot{G}_{i_j}(t) \le -\epsilon_{i_j}$, $j = 1, \ldots, k$, is true. Thus $\dot{G}(t) \le -\epsilon$. □

4. Buffer priority disciplines. In this section we introduce buffer priority policies. Class k has higher priority than class l if $\pi(k) < \pi(l)$, where the priorities π are a permutation of $1, \ldots, K$. To be more precise, under the π priority policy, customers within a class are served following FIFO discipline. If a higher priority class customer arrives at a node while a lower priority class customer is in service, the lower priority class customer is instantly preempted. When service on all higher priority class customers at the node is completed, the server resumes service on the lower priority class customer. Such an implementation of the priority policy is called a preemptive resume discipline.

We describe the dynamics of the reentrant line for some fixed preemptive resume priorities π. Let:

$$H_k = \{l: l \in C_{\sigma(k)}, \pi(l) \le \pi(k)\}.$$

Let

$$T_k^+(t) = \sum_{l \in H_k} T_l(t), \tag{4.1}$$

$$U_k^+(t) = t - T_k^+(t). \tag{4.2}$$

$T_k^+(t)$ is the cumulative amount of service in $[0, t]$ dedicated to customers whose classes are in H_k, and $U_k^+(t)$ is the total unused capacity of server $\sigma(k)$ in $[0, t]$ which is available to serve lower priority customers whose classes are not in H_k. Note that $U_i(t)$ is a station level quantity representing the total capacity unused by server i in $[0,t]$, and this quantity has a real meaning in that the server is actually idle for that amount of time; in contrast, $U_k^+(t)$ is a class level quantity.

The priority policy requires that for every k all the service capacity of station $\sigma(k)$ is dedicated to classes in H_k, as long as the workload present in those buffers is positive. Let

$$W_k^+(t) = \sum_{l \in H_k} m_l Q_l(t) \tag{4.3}$$

denote the total *local immediate workload* at station $\sigma(k)$ in buffers $k \in H_k$. Again $W_k^+(t)$ is a class level quantity as opposed to station i immediate volume $W_i(t)$. In the queueing network context, $W_k^+(t)$ corresponds to the local workload at station $\sigma(k)$ at time t embodied by customers whose classes are in H_k. We assume preemptive resume policy; therefore $U_k^+(\cdot)$ increases only at times t such that $W_k^+(t) = 0$, or equivalently,

$$\int_0^\infty W_k^+(t)\, dU_k^+(t) = 0. \tag{4.4}$$

Proceeding in exactly the same way as in Dai (1995), the complementary condition carries over to the fluid limit model. Therefore, we have

PROPOSITION 4.1. *A fluid model corresponding to a preemptive resume priority policy π must satisfy* (4.4) *in addition to* (1.8)–(1.12).

Notice that for a preemptive resume priority discipline, the extra condition (1.14) takes the form in (4.4). It is easy to see that condition (4.4) supersedes condition (1.13). For $k = 1, \ldots, K$, at all regular points t put

$$a_k(t) = \mu_{k-1}\dot{T}_{k-1}(t), \quad d_k(t) = \mu_k \dot{T}_k(t). \tag{4.5}$$

We call $a_k(t)$ the in-flow rate into buffer k at time t and $d_k(t)$ the out-flow rate from buffer k at time t; recall the convention that $\mu_0 = 1$ and $T_0(t) = t$.

PROPOSITION 4.2. *For the fluid model* (1.8)–(1.12), (4.4), *the following properties hold*: (a) $a_k(t) = d_{k-1}(t)$ *for* $k = 1, \ldots, K$, *where* $d_0(t) = 1$. (b) *If* $Q_k(t) = 0$, *then the in-flow rate and out-flow rate for buffer k are equal, i.e.,* $a_k(t) = d_k(t)$. (c) *At each node at most one non-empty buffer, the highest priority non-empty buffer, can have positive out-flow rate. If* k_0 *is the highest non-empty class at station* $\sigma(k_0)$, *then*

$$\sum_{k \in H_{k_0}} m_k d_k = 1. \tag{4.6}$$

The out-flow rate from every buffer at the node with lower priority than the highest non-empty buffer is 0. *That is, if* $\pi(l) > \pi(k_0)$ *and classes* l *and* k_0 *are served at the same station, then* $d_l(t) = 0$.

PROOF. Assertion (a) follows directly from (4.5). If $Q_k(t) = 0$, it follows from Lemma 2.2 that $Q_k(t) = 0$, and hence, by (1.8), (b) is true. To prove (c), let k_0 be the highest priority nonempty class at station i. Since, $Q_{k_0}(t) > 0$, it follows from (4.4) that $\dot{T}_{k_0}^+(t) = 1$. It then follows from $T_{k_0}^+(t) = \Sigma_{l \in H_{k_0}} m_l(\mu_l T_l(t))$ that (4.6) holds. If $\pi(l) > \pi(k_0)$ and classes l and k_0 are both served in the same station i, it again follows from (4.4) that $\dot{T}_l^+(t) = 1$. Hence

$$\dot{T}_l(t) \le \dot{T}_l^+(t) - \dot{T}_{k_0}^+(t) = 0,$$

and therefore $d_l(t) = 0$. □

DEFINITION 4.1. *If* $\pi(k) = k$, *the corresponding buffer priority discipline is called First-Buffer-First-Served (FBFS). If* $\pi(k) = K + 1 - k$, *then the corresponding buffer priority discipline is called Last-Buffer-First-Served (LBFS).*

In the following two theorems, we prove that the fluid models corresponding to FBFS and LBFS disciplines are stable. Kumar (1993) proved analogous theorems for discrete deterministic systems.

THEOREM 4.3. *The fluid model corresponding to the FBFS discipline is stable.*

PROOF. In the proof we make the following inductive hypothesis on k: Assume that at time t_{k-1} all the buffers $1, \ldots, k-1$ are empty and that they shall stay empty for $t > t_{k-1}$. With the convention that $t_0 = 0$, and $Q_0(t) = 0$, the assumption for $k = 1$ is trivially true. Let the content of buffer k at time t_{k-1} be $Q_k(t_{k-1}) > 0$. This is bounded above by $|Q(0)| + t_{k-1}$. Then, since for $t > t_{k-1}$ we assume buffers $1, \ldots, k-1$ remain empty, it follows that if $Q_k(t) > 0$ for $t \geq t_{k-1}$ then it is the first nonempty buffer, and so, by Proposition 4.2, $a_k(t) = 1$ and

$$d_k(t) = \frac{1 - \sum_{l \in H_k \setminus \{k\}} m_l}{m_k} > 1.$$

Therefore, $\dot{Q}_k(t) = -\mu_k(1 - \sum_{l \in H_k} m_l) < 0$, and hence, by Lemma 2.2, buffer k will be empty at time t_k, where $t_k - t_{k-1} = Q_k(t_{k-1})m_k/(1 - \sum_{l \in H_k} m_l)$ and will stay empty at all times after t_k; this completes the inductive step. To clinch the proof we perform the explicit calculations. We start with $|Q(0)| = 1$. We have $t_1 = Q_1(0)/(\mu_1 - 1) \leq 1/(\mu_1 - 1)$. Assume that

$$|Q(t_{k-1})| \leq \theta_{k-1} \equiv \prod_{l=1}^{k-1} \frac{1 - \sum_{j \in H_l \setminus \{l\}} m_j}{1 - \sum_{j \in H_l} m_j} \quad \text{and} \quad t_k - t_{k-1} \leq \theta_{k-1} \frac{m_k}{1 - \sum_{j \in H_k} m_j}.$$

Then

$$|Q(t_k)| \leq |Q(t_{k-1})| + (t_k - t_{k-1}) \leq \theta_{k-1}\left(1 + \frac{m_k}{1 - \sum_{j \in H_k} m_j}\right) \equiv \theta_k,$$

and

$$t_{k+1} - t_k \leq Q_{k+1}(t_k)m_{k+1}/\left(1 - \sum_{l \in H_{k+1}} m_l\right) \leq \theta_k \frac{m_{k+1}}{1 - \sum_{j \in H_{k+1}} m_j}.$$

Therefore the fluid model will reach $Q(t) = 0$ and remain zero thereafter no later than at

$$\delta = \sum_{k=1}^{K} \left(m_k \frac{\prod_{l=1}^{k-1}(1 - \sum_{j \in H_l \setminus \{l\}} m_j)}{\prod_{l=1}^{k}(1 - \sum_{j \in H_l} m_j)} \right). \qquad \square$$

THEOREM 4.4. *The fluid model corresponding to the LBFS discipline is stable.*

PROOF. Let $G(t) = |Q(t)|$. We will analyze $\dot{G}(t)$ at all regular points. First, because $|Q(t)| = |Q(0)| + t - \mu_K T_K(t)$, we have $\dot{G}(t) = 1 - d_K(t)$. Assume that $G(t) = |Q(t)| > 0$. Let k_0 be the last (highest index) nonempty buffer in the system. By Proposition 4.2, $d_{k_0}(t) = d_{k_0+1}(t) = \cdots = d_K(t)$. Since buffer k_0 is nonempty at time t, we have $\dot{T}^+_{k_0}(t) = 1$. On the other hand, we have

$$\dot{T}^+_{k_0}(t) = \sum_{k \in H_{k_0}} m_k d_k(t) = d_K(t) \sum_{k \in H_{k_0}} m_k.$$

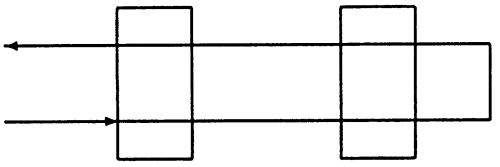

FIGURE 2. A reentrant line considered by Lu and Kumar.

Therefore by Proposition 4.2 we have $d_{k_0}(t) = \cdots = d_K(t) = \lambda$ where

$$\lambda = \frac{1}{\sum_{k \in H_{k_0}} m_k}.$$

Let

$$\hat{\lambda} = \frac{1}{\max_{i=1,\ldots,I}\{\sum_{l \in C(i)} m_l\}}.$$

Then by (1.7), $\lambda \geq \hat{\lambda} > 1$, and so we have shown that for all regular time points at which $G(t) = |Q(t)| > 0, \dot{G}(t) < 1 - \hat{\lambda} < 0$. By Lemma 2.2, the system will empty at the latest by time $\delta = |Q(0)|/(\hat{\lambda} - 1)$ and stay empty thereafter. □

REMARK. This theorem was also proved recently by Kumar and Kumar (1994).

5. Lu-Kumar example. Consider the reentrant line pictured in Figure 2. The condition (1.7) now reads

$$\rho_1 = m_1 + m_4 < 1 \quad \text{and} \quad \rho_2 = m_2 + m_3 < 1. \tag{5.1}$$

Lu and Kumar (1991) studied this network, and showed that for deterministic arrivals and services and a particular choice of the parameters, the priority policy that gives higher priority to classes 2 and 4 is unstable even though (5.1) holds. The following theorem characterizes the stability region for the priority policy used by Lu and Kumar.

THEOREM 5.1. *Assume that* (5.1) *is satisfied. There exists an unstable work-conserving policy for the fluid network if and only if*

$$m_2 + m_4 \geq 1. \tag{5.2}$$

REMARK 1. As we shall see in the proof, this theorem exactly characterizes the region of the parameters for which the Lu and Kumar priority policy will possess a stable fluid model. Furthermore, the theorem also states that in the region in which the Lu and Kumar priority policy has a stable fluid model, every work conserving policy will have a stable fluid model. We call such a region a *global stability region*. In particular, we see from this theorem that the stability region of the Lu and Kumar policy is the smallest among all work conserving policies.

REMARK 2. It is an open problem to determine the stability region for FIFO discipline.

PROOF OF THEOREM 5.1. The proof consists of two unrelated parts. In the first part we demonstrate that the fluid model of the Lu and Kumar policy is unstable if (5.2) holds. In the second part we show that all work conserving fluid models are stable, if $m_2 + m_4 < 1$ holds.

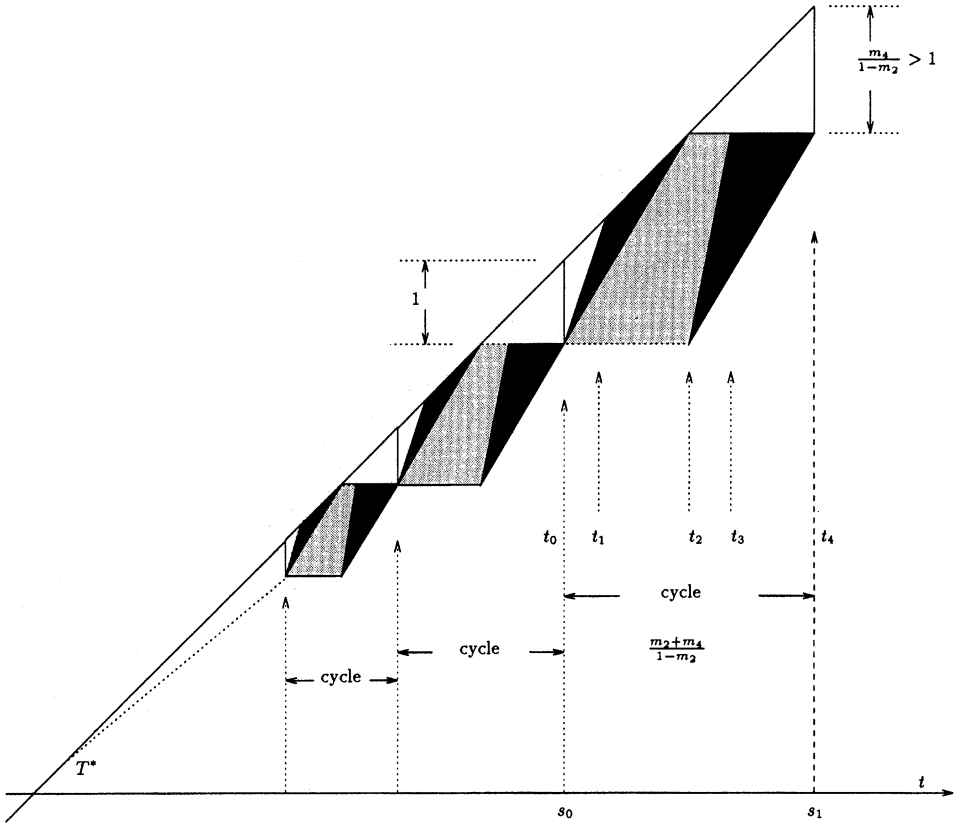

FIGURE 3. The dynamics of a divergent Lu-Kumar fluid network. The four different shadings in a cycle represent fluids in four buffers. The fluid level in a buffer is the distance between the upper boundary and the lower boundary of the shaded region.

PART I. Assume that (5.2) holds. We consider the Lu and Kumar priority policy that gives classes 2 and 4 higher priorities. Assume that the fluid model starts with $Q(0) = (1, 0, 0, 0)$. One can construct a solution $Q = \{Q(t), t \geq 0\}$ to the fluid model satisfying (1.8)–(1.12) and (4.4), and write $(Q(\cdot), T(\cdot))$ explicitly. Figure 3 is a graphical representation of the dynamics of such a solution. It shows the diverging cycles described in this proof. Here we list the changes of state at certain times. At $t_1 = 1/(\mu_1 - 1)$,

$$Q(t_1) = (0, (\mu_1 - \mu_2)t_1, \mu_2 t_1, 0).$$

At $t_2 = 1/(\mu_2 - 1)$,

$$Q(t_2) = (0, 0, 1/(1 - m_2), 0).$$

At t_3, where $t_3 - t_2 = m_3/(1 - m_2)$,

$$Q(t_3) = (t_3 - t_2, 0, 0, (\mu_3 - \mu_4)(t_3 - t_2)).$$

At t_4, where $t_4 - t_2 = m_4/(1 - m_2)$,

$$Q(t_4) = (m_4/(1 - m_2), 0, 0, 0).$$

Note that $t_4 = (m_2 + m_4)/(1 - m_2)$. By a scaling argument, we have

$$Q(s_n) = \left((m_4/(1-m_2))^n, 0, 0, 0\right), \quad \text{where}$$

$$s_n = s_{n-1} + (m_4/(1-m_2))^{n-1}(m_2 + m_4)/(1 - m_2).$$

It is obvious that if $m_2 + m_4 > 1$, as $n \to \infty$,

$$s_n \to \infty \qquad |Q(s_n)| \to \infty.$$

If $m_2 + m_4 = 1$, one obtains a periodic behavior,

$$Q(n/(1 - m_2)) = (1, 0, 0, 0).$$

In either case, this solution is unstable and so the fluid model for this priority policy is unstable. At the time points s_n, $|Q(s_n)|$ exhibits linear growth:

$$\frac{|Q(s_n)| - |Q(s_{n-1})|}{s_n - s_{n-1}} = \frac{m_2 + m_4 - 1}{m_2 + m_4}.$$

As seen in Figure 3, the time points s_n are local minima of $|Q(t)|$; at the local maxima (such as the point t_2), the linear growth rate is $(m_2 + m_4 - 1)/m_4$.

REMARK 3. Consider again Figure 3. Note that by backwards extrapolation we can find a point T^* which is singular: While the fluid network is empty at T^*, we found a solution $Q(\cdot)$ that is nonempty and divergent for $t > T^*$. Notice that $Q(t) \equiv 0$, $t \geq T^*$ is obviously also a solution. The non-uniqueness of solutions to the fluid equations suggests that a small change of initial configuration may cause the original network to follow drastically different sample paths. Such phenomena have been observed by Whitt (1993) for the FIFO Lu-Kumar network.

PART II. Now we assume that (5.2) does not hold, that is $m_2 + m_4 < 1$. Let

$$G_1(t) = \theta_1 Q_1^+(t) + (1 - \theta_1) Q_4^+(t), \tag{5.3}$$

$$G_2(t) = \theta_2 Q_2^+(t) + (1 - \theta_2) Q_3^+(t). \tag{5.4}$$

We are going to show that for an appropriate choice of $0 < \theta_i < 1$, $i = 1, 2$, $G_1(t)$ and $G_2(t)$ satisfy the conditions (a) and (b) in Lemma 3.2 and therefore

$$G(t) = \max\{G_1(t), G_2(t)\}$$

satisfies the condition in Lemma 2.2.

We start with conditions (b). If $W_2(t) = 0$ then,

$$G_2(t) = Q_1(t) \leq Q_1(t) + (1 - \theta_2) Q_4(t) = G_1(t).$$

If $W_1(t) = 0$ then,

$$G_1(t) = (1 - \theta_1)(Q_2(t) + Q_3(t)),$$

$$G_2(t) = Q_2(t) + (1 - \theta_2) Q_3(t),$$

and we can assure $G_1(t) \leq G_2(t)$ by taking $1 - \theta_1 \leq 1 - \theta_2$. Hence condition (b) holds if $\theta_2 \leq \theta_1$.

Consider now condition (a). It follows from (1.8) that

$$G_1(t) = G_1(0) + t - \theta_1 \mu_1 T_1(t) - (1 - \theta_1)\mu_4 T_4(t),$$

$$G_2(t) = G_2(0) + t - \theta_2 \mu_2 T_2(t) - (1 - \theta_2)\mu_3 T_3(t),$$

and taking derivatives, at all regular points (that is almost surely)

$$\dot{G}_1(t) = 1 - \theta_1 \mu_1 \dot{T}_1(t) - (1 - \theta_1)\mu_4 \dot{T}_4(t),$$

$$\dot{G}_2(t) = 1 - \theta_2 \mu_2 \dot{T}_2(t) - (1 - \theta_2)\mu_3 \dot{T}_3(t).$$

If $W_1(t) > 0$, then by work conservation, $\dot{T}_1(t) + \dot{T}_4(t) = 1$. Hence if we choose $m_1 < \theta_1 < 1 - m_4$ then $\theta_1 \mu_1 > 1$, and $(1 - \theta_1)\mu_4 > 1$, and we get

$$\dot{G}_1(t) \leq -\min\{\theta_1 \mu_1 - 1, (1 - \theta_1)\mu_4 - 1\} < 0.$$

Similarly, if $W_2(t) > 0$, a choice $m_2 < \theta_2 < 1 - m_3$ implies

$$\dot{G}_2(t) \leq -\min\{\theta_2 \mu_2 - 1, (1 - \theta_2)\mu_3 - 1\} < 0.$$

Let

$$\delta_1 = (1 - m_1 - m_4)/2,$$

$$\delta_2 = (1 - m_2 - m_3)/2,$$

$$\delta_3 = (1 - m_2 - m_4)/3.$$

By our assumptions, $\delta_1, \delta_2, \delta_3$ are all positive, and the choice

$$\theta_1 = 1 - m_4 - \min\{\delta_1, \delta_3\},$$

$$\theta_2 = m_2 + \min\{\delta_2, \delta_3\}$$

will satisfy all the conditions. □

REMARK 4. Given the set of parameters m_k, $k = 1, \ldots, 4$, we can set up a linear program for the values θ_1, θ_2:

$$\begin{aligned} &\max && \epsilon \\ &\text{subject to} && \epsilon \leq \theta_1 \mu_1 - 1, \\ &&& \epsilon \leq (1 - \theta_1)\mu_4 - 1, \\ &&& \epsilon \leq \theta_2 \mu_2 - 1, \\ &&& \epsilon \leq (1 - \theta_2)\mu_3 - 1, \\ &&& \theta_2 \leq \theta_1. \end{aligned} \tag{5.5}$$

Under the stability conditions, the solution of this linear program will give us appropriate θ_1, θ_2 and an $\epsilon > 0$, so that the system starting from $|Q(0)| = 1$ will be guaranteed to empty by time $t = 1/\epsilon$, for all initial configurations, and all work conserving policies. The value $1/\epsilon$ is a sharp bound.

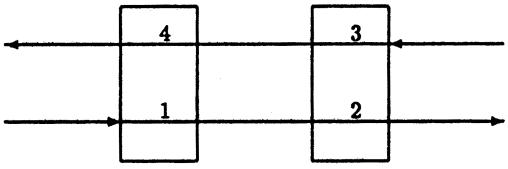

FIGURE 4. A network with two types of jobs.

REMARK 5. Traditionally, test functions like $G(t)$ constructed in the proof are called Lyapunov functions. The function $G(t)$ is of the form $\tilde{G}(Q_1(t), Q_2(t), Q_3(t), Q_4(t))$, where $\tilde{G}(q)$ is some piecewise linear function of $q \in \mathbb{R}_+^4$. Therefore $G(t)$ is also called a piecewise linear Lyapunov function in Down and Meyn (1994b). The Lyapunov function constructed here is similar to the one used in Botvitch and Zamyatin (1992) for a queueing network with exponential distributions.

REMARK 6. For the network pictured in Figure 4, we can also obtain the global stability region. Assume that

$$\rho_1 = m_1 + m_4 < 1,$$

$$\rho_2 = m_2 + m_3 < 1.$$

There exists an unstable policy for the fluid model if and only if

$$m_2 + m_4 \geq 1. \tag{5.6}$$

In fact, when (5.6) does not hold, one can find $0 < \theta_i < 1$, $i = 1, 2$, such that

$$G_1(t) = G_1(0) + \theta_1 Q_1(t) + (1 - \theta_1)(Q_3(t) + Q_4(t)),$$

$$G_2(t) = G_2(0) + \theta_2 Q_3(t) + (1 - \theta_2)(Q_1(t) + Q_2(t))$$

satisfies all the conditions in Lemma 3.2. Hence all work-conserving policies are stable when $m_2 + m_4 < 1$ (while this paper was being written up, Dumas (1993) proved the same result by using a more complicated method). On the other hand, if $m_2 + m_4 \geq 1$ and $Q(0) = (1, 0, 0, 0)$, one can show that, using a priority policy giving classes 2 and 4 higher priorities,

$$Q(m_2/((1 - m_4)(1 - m_2))) = \left(\frac{m_2}{1 - m_4}\frac{m_4}{1 - m_2}, 0, 0, 0\right).$$

Therefore the corresponding fluid model is unstable.

REMARK 7. For the network pictured in Figure 4, Rybko and Stolyar (1992) established the stability region for FIFO discipline to be $\rho_i < 1$, $i = 1, 2$. Kumar and Seidman (1990) considered a discrete deterministic version of this system, where they first introduced the priority discipline used above.

6. Kelly-Type networks. In this section, we consider a special class of reentrant lines, where *mean service times at different visits to a station are the same*. Let β_i denote the mean service time at station i, $i = 1, \ldots, I$. If all distributions are exponential and FIFO discipline is used at each station, Kelly (1979) proved that such networks are stable under the usual traffic condition (1.7). In fact, he was able to explicitly find the product form stationary distribution in this case. We use *Kelly-type network* here to

denote a network in which each visit of a customer to station i has mean service time β_i, regardless of the class designations.

THEOREM 6.1. *Consider a two station Kelly-type reentrant network. Assume that routing does not have immediate feedback. Then any work-conserving policy is stable.*

PROOF. First let us consider the case when $K = 2n$. Let β_i, $i = 1, 2$, be the mean service time at station i. Then the traffic condition (1.7) is reduced to

$$n\beta_i < 1, \qquad i = 1, 2. \tag{6.1}$$

Let

$$G_1(t) = \sum_{l=1}^{n} Q^+_{2l-1}(t) = G_1(0) + nt - (1/\beta_1)B_1(t),$$

$$G_2(t) = \sum_{l=1}^{n} Q^+_{2l}(t) = G_2(t) + nt - (1/\beta_2)B_2(t).$$

When $W_1(t) = 0$, we have

$$G_1(t) = \sum_{l=2}^{n} Q^+_{2l-2}(t) \le \sum_{l=1}^{n} Q^+_{2l}(t) = G_2(t).$$

On the other hand, when $W_2(t) = 0$,

$$G_2(t) = \sum_{l=1}^{n} Q^+_{2l-1}(t) = G_1(t).$$

One can check that all conditions in Lemma 3.2 on $G_1(t)$ and $G_2(t)$ are satisfied. Hence $G(t) = \max\{G_1(t), G_2(t)\}$ is a Lyapunov function satisfying conditions in Lemma 2.2. Therefore $G(t) \equiv 0$ and thus $Q(t) \equiv 0$ for $t \ge \delta$ for some $\delta > 0$. The case when $K = 2n + 1$ can be proved similarly. □

REMARK 1. Using the exact same proof, all work-conserving policies in the Kelly-type reentrant two station example of Dai and Wang (1993) are stable.

REMARK 2. For a more general Kelly-type fluid network, not all work-conserving policies are stable. In fact, consider a two station reentrant line whose visitation sequence is 1, 2, 2, 2, 1, 1 and exit. Let all the mean service times be 0.3, and therefore the nominal workload at each station is 0.90. We consider a static priority discipline as defined in the following table.

	Station 1 Priorities	Station 2 Priorities
Highest	class 6	class 3
2nd	class 5	class 2
3rd	class 1	class 4

It is clear that by the priorities assigned, the network is equivalent to a Lu-Kumar network with four customer classes, where classes 2 and 4 have higher priorities, and the mean service times in the new network are $m_1 = 0.3$, $m_2 = 0.6$, $m_3 = 0.3$ and $m_4 = 0.6$. It follows from Theorem 5.1 that such fluid network is unstable.

This shows that the assumption of no direct feedback in Theorem 6.1 cannot be relaxed. Both the fluid model version and the queueing network version of this remark and the next one were recently proved by Gu (1995).

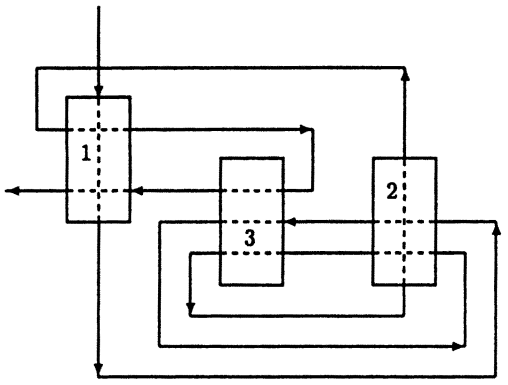

FIGURE 5. An unstable Kelly-type network.

REMARK 3. Theorem 6.1 cannot be generalized to Kelly-type networks with more than two stations. In fact, if we add an additional station in the preceding network, we can construct a Kelly-type network without immediate feedback that is unstable. More specifically, consider a three station network whose visitation sequence is $1, 2, 3, 2, 3, 2, 1, 3, 1$ and exit, see Figure 5. Let $\beta_i = 0.3$, $i = 1, 2$. Consider again a priority policy, where the priorities at stations 1 and 2 are kept the same as before and the priorities at station 3 are arbitrarily assigned. Then for sufficiently small β_3, the scheduling policy is unstable. See Gu (1995) for a proof.

Now we present a generalization of Theorem 6.1 for networks with a special topology. Consider a unidirectional ring network with J types of customers. Type j customers enter the network at some station i, and then follow a deterministic route with visitation sequence: $i, i+1, \ldots, i+n_j$ for some $n_j \geq 1$. We use the convention that whenever $i + k > I$, station $i + k$ is understood as station $i + k - I$. An example of a symmetric four station network is pictured in Figure 6. When $J > 1$, strictly speaking, this ring network is not a reentrant line as introduced in §1. We assume that the mean service times at station i are all the same, equal to β_i, $i = 1, \ldots, I$.

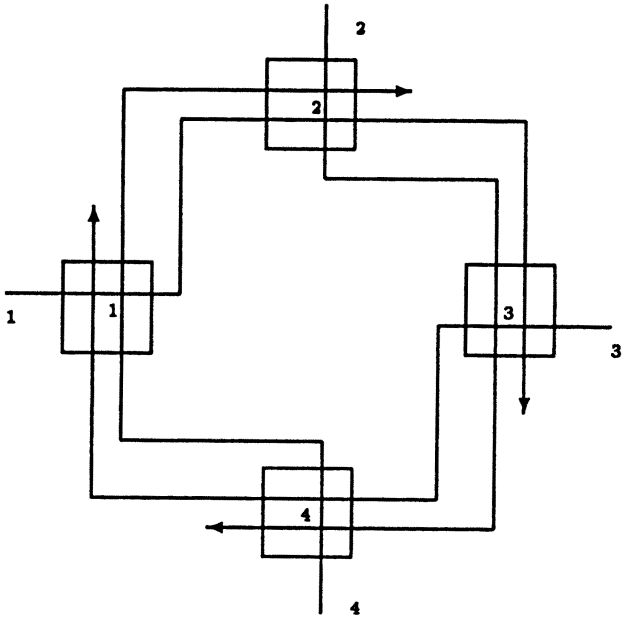

FIGURE 6. A four station symmetric network.

Hence the network is of Kelly-type. We use $\sigma(j,k)$ to denote the station number that type j customers visit at stage k of their route. We assume that the sequence of interarrival times and service times for type j customers satisfies conditions (1.2)–(1.4), $j = 1,\ldots,J$ and furthermore these J sequences are independent. Let α_j be the arrival rate of type j customers and let λ_i be the nominal total arrival rate to station i, which is defined to be the summation of $l_{ij}\alpha_j$ over all types j that visit station i for l_{ij} times, $i = 1,\ldots,I$. Assume that

$$\rho_i \equiv \lambda_i \beta_i < 1, \qquad i = 1,\ldots,I. \tag{6.2}$$

THEOREM 6.2. *Fluid models of the unidirectional ring network, under any work-conserving policy, are stable, if* (6.2) *holds.*

REMARK. Tassiulas and Georgiadis (1993) proved an analogous theorem when interarrival times and service times are deterministically constrained as in Cruz (1991). Using Theorem 4.3 of Dai (1995), Theorem 6.2 implies that under (6.2) any work-conserving policy is positive Harris recurrent.

PROOF OF THEOREM 6.2. Let $Q_{jk}(t)$ be the queue length of type j customers at stage k,

$$Q_{jk}^{+}(t) = \sum_{l=1}^{k} Q_{jl}(t), \qquad k = 1,\ldots,n_j \quad \text{and} \quad j = 1,\ldots,J.$$

For station i, $i = 1,\ldots,I$, define

$$G_i(t) = \sum_{(j,k);\ \sigma(j,k)=i} Q_{jk}^{+}(t) = G_i(0) + \lambda_i t - 1/\beta_i B_i(t),$$

$$W_i(t) = \beta_i \sum_{(j,k):\ \sigma(j,k)=i} Q_{jk}(t).$$

Then $W_i(t) > 0$ implies $\dot{G}_i(t) = -(1/\beta_i - \lambda_i) < 0$. Let

$$G(t) = \max\{G_1(t),\ldots,G_I(t)\}.$$

Assume that $G(t) > 0$ and t is a regular point for $G(\cdot)$ and $G_i(\cdot)$'s. Let $i_1,\ldots,i_l$ be stations such that

$$G_{i_1}(t) = G_{i_2}(t) = \cdots = G_{i_l}(t) = G(t)$$

and $G_i(t) < G(t)$ for $i \notin \{i_1,\ldots,i_l\}$. Then

$$\dot{G}(t) = \dot{G}_{i_1}(t) = \cdots = \dot{G}_{i_l}(t).$$

If $\{i_1,\ldots,i_l\} = \{1,\ldots,I\}$, then there exists at least one station i such that $W_i(t) > 0$, and hence $\dot{G}(t) \le -\epsilon$, where

$$\epsilon = \min_{1\le i\le I} (1/\beta_i - \lambda_i).$$

Otherwise, there exists an $i \notin \{i_1,\ldots,i_l\}$. Choose i such that $i \notin \{i_1,\ldots,i_l\}$, but

$i + 1 \in \{i_1, \ldots, i_l\}$. We claim that $W_{i+1}(t) > 0$. In fact, if $W_{i+1}(t) = 0$, we have

$$G_{i+1}(t) = \sum_{(j,k):\ \sigma(j,k)=i+1} Q_{jk}^{+}(t) = \sum_{(j,k):\ \sigma(j,k)=i+1} Q_{j(k-1)}^{+}(t)$$

$$\leq \sum_{(j,k):\ \sigma(j,k)=i} Q_{jk}^{+}(t) < G(t)$$

which is a contradiction to $\{i_1, \ldots, i_l\}$. But $W_{i+1}(t) = 0$ implies

$$\dot{G}(t) = \dot{G}_{i+1}(t) \leq -\epsilon.$$

It follows from Lemma 2.2 that the fluid model is stable. □

7. Concluding remarks. It is yet to be established that a queueing discipline is positive Harris recurrent if and only if the corresponding fluid model is stable. It is evident that fluid model is more tractable than the stochastic system itself in studying stability and instability of a queueing network. Admittedly to determine the stability region for any given discipline in a fluid model seems a formidable task, but it appears that the piecewise linear Lyapunov functions advanced by Botvitch and Zamyatin (1992) offer a promising approach. In constructing a quadratic Lyapunov function for queueing networks, Kumar and Meyn (1995a, 1995b) used a linear program approach to find the coefficients of the quadratic function. Their work was adapted to fluid networks by Chen (1995). For the three buffer reentrant lines, their quadratic Lyapunov function cannot determine the exact stability region, where our piecewise linear Lyapunov functions yield a sharp stability region. It would be interesting to compare in general the quadratic function approach to Kumar and Meyn (1995a, 1995b) and Chen (1995) with some generalization of the Botvitch-Zamyatin's approach. For preliminary work in this direction, readers are referred to recent work by Down and Meyn (1994a, 1994b) and Dai and Vande Vate (1995a, 1995b).

Acknowledgement. The research of the first author supported by NSF grants DMS-9203524 and DDM-9215233, and two grants from the Texas Instruments Corporation. The research of the second author supported by NSF grants DDM-8914863 and DDM-9215233, and the fund for the promotion of research at the Technion. We are grateful to Bruce Hajek and Leandros Tassiulas for helpful discussions on Theorem 6.2. We thank Vincent Dumas for sending us a paper by Botvich and Zamyatin. We also thank an anonymous referee for many helpful remarks.

References

Baccelli, F. and S. Foss (1994). Stability of Jackson-type queueing networks, I. *Queueing Systems: Theory and Applications* **17** 5–72.

Baskett, F., K. M. Chandy, R. R. Muntz and F. G. Palacios (1975). Open, closed and mixed networks of queues with different classes of customers. *J. ACM* **22** 248–260.

Borovkov, A. A. (1986). Limit theorems for queueing networks. *Theory of Probability and its Applications* **31** 413–427.

Botvich, D. D. and A. A. Zamyatin (1992). Ergodicity of conservative communication networks. Rapport de recherche 1772, INRIA.

Bramson, M. (1994). Instability of FIFO queueing networks. *Annals of Applied Probability* **4** 414–431.

Chang, C. S., J. A. Thomas and S.-H. Kiang (1994). On the stability of open networks: a unified approach by stochastic dominance. *Queueing Systems: Theory and Applications* **15** 239–260.

Chen, H. (1995). Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines. *Annals of Applied Probability* (to appear).

Cruz, R. L. (1991). A calculus for network delay, part II: network analysis. *IEEE Transactions of Information Theory* **37** 132–141.

Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid models. *Annals of Applied Probability* **5** 49–77.

___ and J. Vande Vate (1995a). The stability of two-station queueing networks. Preprint.

___ and ___ (1995b). The stability of two-station fluid networks. Preprint.

___ and Y. Wang (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems: Theory and Applications* **13** 41–46.

Down, D. and S. Meyn (1994a). Piecewise linear test functions for stability of queueing networks. *Proceedings of the 33rd Conference on Decision and Control* 2069–2074.

___ and ___ (1994b). A survey of Markovian methods for stability of networks (preprint).

Dumas, V. (1993). Harris ergodicity of a multiclass queueing network via its associated fluid model (preprint).

Gu, J. M. (1995). *Convergence and Performance for Some Kelly-like Queueing Networks*, Ph.D. thesis, University of Wisconsin, Madison.

Jackson, J. R. (1957). Networks of waiting lines. *Oper. Res.* **5** 518–521.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*, Wiley, New York.

Kumar, P. R. (1993). Reentrant lines. *Queueing Systems: Theory and Applications* **13** 87–110.

___ and S. Meyn (1995a). Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* (to appear).

___ and ___ (1995b). Stability of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* **40** 251–260.

___ and T. I. Seidman (1990). Dynamic instabilities and stabilization methods in distributed read-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control* **AC-35** 289–298.

Kumar, S. and P. R. Kumar (1994). Fluctuation smoothing policies are stable for stochastic reentrant lines. *J. Discrete Event Dynamic Systems: Theory and Appl.* (to appear).

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

Meyn, S. P. and D. Down (1994). Stability of generalized Jackson networks. *Annals of Applied Probability* **4** 124–148.

___ and R. L. Tweedie (1993). Stability of Markovian processes II: Continuous time processes and sample chains. *Adv. Appl. Probab.* **25** 487–517.

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

Seidman, T. I. (1994). 'First come, first served' can be unstable! *IEEE Transactions on Automatic Control* **39** 2166–2171.

Sigman, K. (1990). The stability of open queueing networks. *Stoch. Proc. Applns.* **35** 11–25.

Stolyar, A. (1994). On the stability of multiclass queueing networks. Proceeding of the Second Conference on Telecommunication Systems—Modeling and Analysis, Nashville, March 22–27.

Tassiulas, L. and L. Georgiadis (1993). Any work-conserving policy stabilizes the ring with spatial reuse (preprint).

Wang, Y. (1993). Private communications.

Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Sci.* **39** 1020–1028.

J. G. Dai: School of Industrial and Systems Engineering and School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332-0205; e-mail: dai@isye.gatech.edu

G. Weiss: Department of Statistics, The University of Haifa, Haifa, Israel; e-mail: gweiss@stat.haifa.ac.il