The Annals of Applied Probability
1995, Vol. 5, No. 3, 637–665

FLUID APPROXIMATIONS AND STABILITY OF MULTICLASS QUEUEING NETWORKS: WORK-CONSERVING DISCIPLINES[1]

BY HONG CHEN

University of British Columbia

This paper studies the fluid approximation (also known as the functional strong law of large numbers) and the stability (positive Harris recurrence) for a multiclass queueing network. Both of these are related to the stabilities of a linear fluid model, constructed from the first-order parameters (i.e., long-run average arrivals, services and routings) of the queueing network. It is proved that the fluid approximation for the queueing network exists if the corresponding linear fluid model is weakly stable, and that the queueing network is stable if the corresponding linear fluid model is (strongly) stable. Sufficient conditions are found for the stabilities of a linear fluid model.

1. Introduction. The subject of this paper is multiclass queueing networks [Baskett, Chandy, Muntz and Palacios (1975), Kelly (1982) or Harrison (1988)]. The functional strong law of large numbers theorem, also known as the fluid approximation theorem, is proved for a class of multiclass queueing networks under work-conserving service disciplines. Furthermore, the fluid approximation result is used to prove the stability of a class of multiclass queueing networks under work-conserving service disciplines.

The fluid approximation is proved for open generalized Jackson networks by Johnson (1983), and for both open and closed generalized Jackson networks by Chen and Mandelbaum (1991a). Essentially, all work on a diffusion approximation implicitly implies a fluid approximation result, although more assumptions may be assumed. Readers are referred to Glynn (1990) for references on the diffusion approximation. The limit of the fluid approximation is a class of fluid models, which has been used for a wide variety of dynamic systems [Newell (1982), Kleinrock (1976), Vandergraft (1983), Mitra (1986), Kosten (1986) and Chen and Yao (1993)]. See Chen and Mandelbaum (1994) for a survey of fluid models.

There is a lot of literature on the stability of queueing networks. The results are mostly on Jackson networks, generalized Jackson networks and Kelly networks, which include primarily single class networks with some exceptions for restrictive cases. For Jackson networks and Kelly networks, the stability

Received January 1994; revised July 1994.

[1]Research supported in part by NSERC Grant OGP0089698 and by the UBC Killam Research Fellowship. Part of the research was done while the author was visiting the Hong Kong University of Science and Technology.

AMS 1991 *subject classifications*. Primary 60F17, 60K25, 60G17; secondary 60F15, 60K30, 90B22.

Key words and phrases. Multiclass queueing networks, fluid models, fluid approximations, stability, positive Harris recurrent and work-conserving service disciplines.

is established by explicitly determining the invariant distribution [Jackson (1963), Baskett, Chandy, Muntz and Palacios (1975), Kelly (1982), among others], and for generalized Jackson networks, the stability is established by proving directly the positive Harris recurrence of the underlying Markov process [Borovkov (1986), Sigman (1990), Chang, Thomas and Kiang (1993), Meyn and Down (1994), Baccelli and Foss (1994) and references in their papers]. Recently, Kumar and Meyn (1993) established the stability of a class of multiclass queueing networks with Poisson arrivals and exponential service times by constructing a Lyapunov function.

It was widely believed that under the usual traffic intensity condition (i.e., the nominal load is less than one at all stations), the queueing network would be stable under all work-conserving service disciplines. Recently, several examples have been found that the usual condition is *not* sufficient for the stability of multiclass queueing networks. Kumar and Seidman (1990) found a deterministic model which is unstable under exhaustive service disciplines. Rybko and Stolyar (1992) proved a stochastic (queueing) system that corresponds to one of the examples in Kumar and Seidman (1990), which is unstable under a certain priority service discipline, but is stable under FIFO (first-in first-out) discipline. Recently, Seidman (1993) and Bramson (1994) demonstrated the instability of FIFO discipline for a deterministic system and a stochastic system, respectively. Closely related to the stability is the limit theorem for multiclass queueing networks. It may be expected that the heavy traffic limit exists when the network parameters are at the boundary between the stable and the unstable regions (i.e., under heavy traffic condition). Dai and Wang (1993) constructed a counterexample where the proposed Brownian limit does not exist, and Whitt (1993) provided a similar example.

In analyzing a two-station multiclass Markovian network, Rybko and Stolyar (1992) link the stability of the queueing network to the stability of its fluid approximation limit. The latter linkage was generalized by Dai (1995), to cover more general networks, service disciplines and distributions for arrival and service times. Roughly speaking, Dai (1995) established that stability of the fluid approximation limit is sufficient for the stability of a multiclass queueing network, thus providing a general approach to the stability of a multiclass queueing network. In particular, his result, combined with some earlier results on fluid approximations, implies the stability of generalized open Jackson networks and feedforward multiclass queueing networks.

Our analysis starts with a heterogeneous *linear fluid model* [introduced in Chen (1988) and Chen and Mandelbaum (1991b)]. The linear fluid model is called *weakly stable* if the fluid (inventory) level remains zero when its initial level is zero, and is called (*strongly*) *stable* if the fluid level reaches zero in a finite amount of time for any given initial inventory levels. Sufficient conditions are identified for the weak stability and the stability of the fluid network under a work-conserving condition. The approach taken is to use a Lyapunov function analogous to Kumar and Meyn (1993). [The use of a Lyapunov function for a fluid model can also be found in Dupuis and Williams (1993).] It is established that the weak stability and the stability of a linear

fluid model are sufficient for the existence of the fluid approximation and for the stability of the corresponding queueing network, respectively.

The notion of the stability mentioned above is the positive Harris recurrence for the underlying Markov process that describes the dynamics of the queueing network [Mèyn and Tweedie (1993) and Dai (1995)]. Some weaker notion of stabilities may also be of interest, such as Harris recurrence (which includes both null and positive Harris recurrences) and a pathwise stability [a term communicated to me by Shaler Stidham; see Altman, Foss, Riehl and Stidham (1994)]. Pathwise stability means that the long-run average departures must equal the long-run average arrivals at each station with probability 1. In a Jackson network, the queue length process is positive Harris recurrent if and only if the corresponding fluid network is stable, and, in addition, the following are equivalent: (a) the queue length process is Harris recurrent, (b) the queueing network is pathwise stable, (c) the fluid limit of the queue length process is zero and (d) the corresponding fluid network is weakly stable. In this paper, we establish that (d) implies (b) and (c) for a multiclass queueing network under work-conserving service disciplines. [It is obvious that (c) always implies (b).]

The primary contribution of this paper is to prove the fluid approximation for a class of multiclass queueing networks (Theorems 4.1 and 3.5) and to prove stability for a class of multiclass queueing networks (Corollary 5.3 and Theorem 3.5) [which extends Kumar and Meyn (1993) to non-Poisson arrival and nonexponential service distributions]. In addition, the sufficient condition for the stability of multiclass queueing networks in Dai (1995) is simplified from the stability of a *piecewise-linear* fluid limit model to the stability of a *linear* fluid model (Theorem 5.2).

The paper is structured as follows. In Section 2 we introduce the model of a heterogeneous linear fluid network, operating under a work-conserving policy. Then we establish sufficient conditions for the weak stability and the stability of the fluid network in Section 3. The multiclass queueing network is introduced and its fluid approximation is established in Section 4. The stability of a class of multiclass queueing networks is the topic of Section 5. Section 6 concludes with some remarks. Readers may first read through the example in Section 4.3 that was studied by Kumar and Seidman (1990) and Rybko and Stolyar (1992) and that motivated this research.

Some notation and conventions used throughout the paper are as follows. Vectors are understood to be column vectors; the transpose of a vector or a matrix is denoted by a prime. For a real number x, $x^- = \min\{x, 0\}$ and $x^+ = \max\{x, 0\}$. The I-dimensional Euclidean space is denoted by $\Re^I$. When the letter e denotes a vector, it stands for the vector of ones. Vector (in)equalities are to be interpreted componentwise. All functions are assumed right-continuous with left limits (RCLL functions). The abbreviation "u.o.c." stands for "uniformly on compact sets." For a vector $\mu' = (\mu_1, \ldots, \mu_K)$, the matrix $\text{diag}(\mu)$ denotes the $K \times K$ diagonal matrix with diagonal elements $\mu_1, \ldots, \mu_K$. The indicator function of a set S is the function $1_S(z)$ which equals 1 for z in S and 0 otherwise.

2. A linear heterogeneous fluid model. The fluid model consists of a set of J buffers, indexed by $j = 1, \ldots, J$. Each buffer has an infinite storage capacity, and buffers are interconnected by pipes to form a network within which several classes of fluids are simultaneously circulating. (Though "types" is a better word for fluids, we use "classes" for consistency with the terminology in queueing networks.) There are K classes of fluids, indexed by $k = 1, \ldots, K$. Fluid of class k, hereafter referred to as fluid k, resides exclusively in buffer $j = \sigma(k)$, where $\sigma(\cdot)$ is a many-to-one mapping from classes to buffers. We denote by $C(j) = \{k\colon\ \sigma(k) = j\}$ the set of classes that reside in buffer j, and by $C = (c_{jk})$ a $J \times K$ matrix with $c_{jk} = 1$ when $\sigma(k) = j$ and $c_{jk} = 0$ otherwise; the latter is referred to as a constituent matrix. Without loss of generality, it is assumed that $C(j)$ is nonempty for all $j = 1, \ldots, J$.

The fluid network is described by two K-dimensional nonnegative vectors $Q(0) = (Q_k(0))$ and $\alpha = (\alpha_k)$, one K-dimensional positive vector $\mu = (\mu_k)$, one $K \times K$ substochastic matrix $P = (p_{ik})$ with a spectral radius less than unity and the $J \times K$ constituent matrix C. Such a fluid network is referred to as fluid network (α, μ, P, C) with initial fluid level $Q(0)$. Vectors $Q(0)$, α and μ are referred to as the *initial fluid level*, the *exogenous inflow rate* and the *potential outflow rate*, respectively. Matrix P is referred to as the *flow-transfer matrix*. Thus, $\alpha_k t$ indicates the cumulative exogenous inflow of fluid k during the time interval $[0, t]$. The component p_{ik} of matrix P indicates the fraction of the outflow of fluid i that turns into fluid k, and $1 - \sum_{k=1}^{K} p_{ik}$ indicates the fraction that leaves the network. The interpretation of the potential outflow rate is given shortly.

Let $T_k(t)$ be the cumulative amount of time allocated to processing fluid k by buffer $\sigma(k)$, during $[0, t]$. Then $\mu_k T_k(t)$ indicates the cumulative amount of outflow of fluid k during $[0, t]$. It is assumed that the maximum time that can be allocated to processing fluid k during any time interval $[s, t]$ $(t \geq s \geq 0)$ is $(t - s)$. Thus, $\mu_k(t - s)$ is potentially the maximum possible amount of outflow of fluid k in that duration. The K-dimensional process $T = (T_k)$ with $T_k = \{T_k(t),\ t \geq 0\}$ is referred to as *allocation process*.

The process of primary interest is the *fluid level process* $Q = (Q_k)$ with $Q_k = \{Q_k(t),\ t \geq 0\}$, where $Q_k(t)$ is the fluid level of fluid k at time t. Given (α, μ, P, C), $Q(0)$ and the allocation process T, $Q_k(t)$ is given by the flow-balance relations

$$Q_k(t) = Q_k(0) + \alpha_k t + \sum_{i=1}^{K} \mu_i T_i(t) p_{ik} - \mu_k T_k(t) \geq 0. \tag{2.1}$$

The total busy time of buffer j during $[0, t]$ equals $\sum_{k \in C(i)} T_k(t)$. Thus, its cumulative *idle time* (unused capacity) is

$$U_j(t) = t - \sum_{k \in C(j)} T_k(t).$$

Clearly, it must be true that

$$T_k(\cdot) \text{ is nondecreasing with } T_k(0) = 0, \qquad k = 1, \ldots, K, \tag{2.2}$$

(2.3) $U_j(\cdot)$ is nondecreasing, $j = 1, \ldots, J$.

An allocation T is called *feasible* if it satisfies (2.1)–(2.3), and a feasible allocation is *work-conserving* if it also satisfies

(2.4) $U_j(\cdot)$ is increasing at time t only when $Z_j(t) = 0$, $j = 1, \ldots, J$,

where

$$Z_j(t) = \sum_{k \in C(j)} Q_k(t)$$

is the total fluid level in buffer j at time t. We note that any feasible allocation process T must be Lipschitz continuous, as must be its associated fluid level process.

To summarize in a vector form, an allocation T is work-conserving if and only if, together with the associated fluid level process Q, it satisfies for $t \geq 0$,

(2.5) $$Q(t) = Q(0) + \alpha t - [I - P']MT(t) \geq 0,$$

(2.6) $$dT(t) \geq 0 \quad \text{with } T(0) = 0,$$

(2.7) $$d[et - CT(t)] \geq 0,$$

(2.8) $$[CQ(t)]'d[et - CT(t)] = 0,$$

where $M = \text{diag}(\mu)$ is a $K \times K$ diagonal matrix. The associated fluid level process is referred to as a work-conserving fluid level process. The pair (T, Q) is simply referred to as a work-conserving pair.

It is easy to see that the set of feasible allocation processes is not unique. For example, both $T(t) \equiv 0$ and $T(t) = M^{-1}[I - P']^{-1}\alpha t$ are feasible. On the other hand, it is known that the set of work-conserving allocations is always unique when $J = K$ [Chen and Mandelbaum (1991a)]. However, this may not be the case in general. In the next section, a sufficient condition is found under which the set of work-conserving allocation processes is unique for the case $Q(0) = 0$. In general, additional constraints on the allocation process are necessary for uniqueness. Such constraints may correspond to various service disciplines in queueing networks. (See the discussion in Section 6.)

The existence of a work-conserving allocation process is stated as follows (the proof is given in the Appendix).

THEOREM 2.1. *For any linear fluid network (α, P, μ, C) with $Q(0)$, there exists at least one work-conserving allocation.*

REMARKS.

1. The proof of the theorem actually implies a stronger result. Let I^0 be a J-dimensional process; each of its component $I_j^0 = \{I_j^0(t),\ t \geq 0\}$ is nonde-

creasing with $I_j^0(0) = 0$ and is Lipschitz continuous with a Lipschitz constant 1. Then given I^0, there always exists an allocation process satisfying (2.5) and (2.6), and

$$d[I^0(t) - CT(t)] \geq 0, \tag{2.7'}$$

$$[CQ(t)]'d[I^0(t) - CT(t)] = 0. \tag{2.8'}$$

The theorem is a special case with $I^0(t) = et$. This generalization is very useful for the construction of an allocation that utilizes the remaining capacity described by I^0. In particular, given any feasible allocation T^0, there exists a work-conserving allocation $T \geq T^0$. [In fact, $T - T^0$ is an allocation satisfying (2.5) and (2.6) with $Q(0) = 0$ and (2.7') and (2.8') with $I^0(t) = et - CT^0(t)$.]

2. The existence theorem can also be extended to a nonlinear fluid model, where the linear cumulative exogenous arrival vector αt in (2.5) is replaced by a nonlinear nondecreasing vector $\alpha(t)$.

This section is concluded with two important properties of a linear fluid network; their proofs are straightforward.

PROPERTY 1 (Shift property). Suppose that (T, Q) is a work-conserving pair for fluid network (α, μ, P, C) with an initial fluid level $Q(0)$. For any fixed $s \geq 0$, let $\tilde{T}(t) = T(t+s) - T(s)$ and $\tilde{Q}(t) = Q(t+s)$. Then $(\tilde{T}, \tilde{Q})$ is a work-conserving pair for the same fluid network with initial fluid level $Q(s)$.

PROPERTY 2 (Scale property). Suppose that (T, Q) is a work-conserving pair for fluid network (α, μ, P, C) with an initial fluid level $Q(0)$. Then $\tilde{T}(t) = \theta T(t/\theta)$ and $\tilde{Q}(t) = \theta Q(t/\theta)$ give a work-conserving pair for the same fluid network with the initial fluid level $\theta Q(0)$.

3. Stability of the linear fluid network. A fluid network (α, μ, P, C) is said to be *weakly stable* if the set of work-conserving allocations is unique with $Q(t) = 0$ for all $t \geq 0$ when $Q(0) = 0$, and is said to be (*strongly*) *stable* if there exists a finite time t_0 such that $Q(t_0) = 0$ for *all* work-conserving fluid level processes Q with $e'Q(0) = 1$. The notion of the stabilities here is about the stabilities of a fluid network under *all* work-conserving conditions, and a brief discussion of the stabilities under more restrictive conditions is in Section 6.

The presentation of this section is arranged into four subsections: the first subsection proves that stability implies weak stability and gives some necessary conditions for stabilities; the second and the third present, respectively, sufficient conditions for weak and strong stability; the last proves some additional properties of stability, including a monotonicity property.

3.1. *Properties and necessary conditions for stabilities.*

THEOREM 3.1. *A necessary condition for a fluid network* (α, μ, P, C) *to be weakly stable is that*

$$\rho := C\beta \leq e, \tag{3.1}$$

where

$$\beta := M^{-1}[I - P']^{-1}\alpha. \tag{3.2}$$

If the network is weakly stable, then when $Q(0) = 0$, *the unique work-conserving allocation is given by* $T(t) = \beta t$, *and its associated fluid level process is given by* $Q(t) = 0$.

REMARKS.

1. The J-dimensional vector ρ is known as the *traffic intensity* for the network. Throughout the paper, assume that $\beta > 0$. This is without loss of generality, since otherwise, those classes that correspond to a zero coordinate of β can be removed from the network. (See Section 4.1 for more comments on the traffic intensity.)
2. Condition (3.1) is sufficient for the case when $J = K$ [in this case the fluid network is known as a *homogeneous* fluid network; see Chen and Mandelbaum (1991a)], but it is *not* sufficient in more general cases, which is demonstrated by the example in Section 4.3.
3. Condition (3.1) is sufficient for the case when $J = 1$. The proof is by observing that

$$Z(t) := M^{-1}[I - P']^{-1}Q(t) = (\rho - e)t + U(t) \geq 0$$

and $Z(t)\,dU(t) = 0$ and by the uniqueness of one-dimensional reflection mapping. The argument can be extended to a feedforward network [Peterson (1991)], which implies that condition (3.1) is also sufficient in this case.

PROOF OF THEOREM 3.1. Multiplying both sides of (2.5) by $CM^{-1}[I - P']^{-1}$ yields

$$CM^{-1}[I - P']^{-1}Q(t) = (\rho - e)t + U(t) \geq (\rho - e)t,$$

which clearly implies the theorem. □

THEOREM 3.2. *A necessary condition for a fluid network* (α, μ, P, C) *to be stable is that*

$$\rho = C\beta = CM^{-1}[I - P']^{-1}\alpha < e. \tag{3.3}$$

REMARKS.

1. Condition (3.3) is also sufficient for the homogeneous fluid network to be stable [see Chen and Mandelbaum (1991a)], but the example in Section 4.3 indicates that condition (3.3) is *not* sufficient in the more general case.
2. Condition (3.3) is sufficient for the single station case ($J = 1$) and the feedforward network. The justification is similar to the one for Remark 3 after Theorem 3.1.

PROOF OF THEOREM 3.2. Multiplying both sides of (2.5) by $CM^{-1}[I - P']^{-1}$ yields

$$\begin{aligned} CM^{-1}[I - P']^{-1}Q(t) &= CM^{-1}[I - P']^{-1}Q(0) + (\rho - e)t + U(t) \\ &\geq CM^{-1}[I - P']^{-1}Q(0) + (\rho - e)t. \end{aligned}$$

In the above, taking a $Q(0)$ with $Q_j(0) > 0$ for $\rho_j \geq 1$ proves the theorem. □

THEOREM 3.3. *If a fluid network is stable, then it must be weakly stable.*

PROOF. When the fluid network is stable, by Properties 1 and 2 (the shift property and the scale property), there exists a t_0 such that if $e'Q(s) = \varepsilon$,

$$Q(s + t) = 0 \quad \text{for } t \geq \varepsilon t_0. \tag{3.4}$$

Suppose that the fluid network is not weakly stable. Then there exists a fluid level process Q with $Q(0) = 0$ and a $\tau > 0$ such that $e'Q(\tau) = \delta > 0$. Pick $\varepsilon << \delta$, and let $\tau_0 = \max\{t \leq \tau\colon e'Q(t) = \varepsilon\}$ (which is well defined since Q is continuous). Since Q is Lipschitz continuous,

$$|e'Q(t) - e'Q(\tau_0)| \leq \theta(t - \tau_0),$$

for all $t \geq \tau_0$, where θ is a finite constant. The above inequality implies that

$$\begin{aligned} e'Q(t) &< e'Q(\tau_0) + \theta\varepsilon t_0 = (1 + \theta t_0)\varepsilon \\ &< \delta = e'Q(\tau) \quad \text{for all } t \in [\tau_0, \tau_0 + \varepsilon t_0], \end{aligned} \tag{3.5}$$

where in the last inequality we assume that we have picked ε small enough. Inequality (3.5) implies that $\tau_0 + \varepsilon t_0 < \tau$. On the other hand, by (3.4), we have $e'Q(\tau_0 + \varepsilon t_0) = 0$. Therefore, there must be $e'Q(s) = \varepsilon$ for some $\tau_0 + \varepsilon t_0 < s < \tau$, contradicting the definition of τ_0. □

3.2. *A sufficient condition for weak stability.* A $K \times K$ symmetric matrix A is called a *strictly copositive* matrix if, for all $x \in \Re^K$ and $x \geq 0$, $x'Ax \geq 0$ and $x'Ax = 0$ only when $x = 0$ [see Cottle, Habetler and Lemke (1970)].

THEOREM 3.4. *The fluid network (α, μ, P, C) is weakly stable if there exists a $K \times K$ symmetric strictly copositive matrix $A = (a_{ik})$ such that, for $k =$*

$1, \ldots, K$,

$$\sum_{i=1}^{K} \alpha_i a_{ik} - \min_{i \in C(\sigma(k))} h_{ik} - \sum_{\substack{j=1 \\ j \neq \sigma(k)}}^{J} \Big(\min_{i \in C(j)} h_{ik} \Big)^- \leq 0, \tag{3.6}$$

where $H = (h_{ik}) = M[I - P]A$, *or equivalently,*

$$h_{ik} = \mu_i \Big[a_{ik} - \sum_{l=1}^{K} p_{il} a_{lk} \Big].$$

REMARKS.

1. The existence of a strictly copositive matrix A such that $d[Q'(t)AQ(t)] = 0$ is in fact a necessary and sufficient condition for weak stability. Thus, the necessary condition may be improved if inequalities (3.7) and (3.9) can be tightened.
2. The verification of condition (3.6) can be formulated into a linear programming problem (see Remark 2 after Theorem 3.5).

PROOF OF THEOREM 3.4. Suppose that Q is a work-conserving fluid level process of the fluid network (α, μ, P, C) with $Q(0) = 0$. Let $f(t) = Q'(t)AQ(t)$. Clearly, $f(0) = 0$. If we can show that $df(t) = 0$ for all $t \geq 0$, then $f(t) = 0$ for all $t \geq 0$. This implies that $Q(t) = 0$ for all $t \geq 0$, in view of matrix A being strictly copositive. Thus, it suffices to show that $df(t) = 0$ for all $t \geq 0$.

Using (2.5) yields

$$\begin{aligned}
\tfrac{1}{2} df(t) &= (\alpha'\, dt - dT'(t) M[I - P]) AQ(t) \\
&= \alpha' AQ(t)\, dt - \sum_{k=1}^{K} \sum_{i=1}^{K} dT_i(t) h_{ik} Q_k(t) \\
&= \alpha' AQ(t)\, dt - \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i \in C(j)} h_{ik} Q_k(t)\, dT_i(t) \\
&\leq \alpha' AQ(t)\, dt - \sum_{k=1}^{K} \sum_{j=1}^{J} \Big(\min_{i \in C(j)} h_{ik} \Big) Q_k(t) d(t - U_j(t)) \qquad (3.7) \\
&\leq \alpha' AQ(t)\, dt - \sum_{k=1}^{K} \Big[\min_{i \in C(\sigma(k))} h_{ik} + \sum_{\substack{j=1 \\ j \neq \sigma(k)}}^{J} \Big(\min_{i \in C(j)} h_{ik} \Big)^- \Big] Q_k(t)\, dt \qquad (3.8) \\
&= \sum_{k=1}^{K} \Big[\sum_{i=1}^{K} \alpha_i a_{ik} - \min_{i \in C(\sigma(k))} h_{ik} - \sum_{\substack{j=1 \\ j \neq \sigma(k)}}^{J} \Big(\min_{i \in C(j)} h_{ik} \Big)^- \Big] Q_k(t)\, dt \leq 0,
\end{aligned}$$

where $U_j(t) = t - \sum_{i \in C(j)} T_i(t)$ is used in (3.7), inequality (3.8) follows from

$$Q_k(t)\, dU_j(t) = 0 \quad \text{for } j = \sigma(k)$$

[the work-conserving condition (2.4)], and

$$(3.9)\qquad \Big(\min_{i\in C(j)} h_{ik}\Big) Q_k(t)\, dU_j(t) \le \Big(\min_{i\in C(j)} h_{ik}\Big)^+ Q_k(t)\, dt \quad \text{for } j \ne \sigma(k). \qquad \square$$

3.3. *A sufficient condition for stability.*

THEOREM 3.5. *The fluid network* (α, μ, P, C) *is stable if there exists a* $K \times K$ *symmetric strictly copositive matrix* $A = (a_{ik})$ *such that, for* $k = 1, \dots, K$,

$$(3.10)\qquad \sum_{i=1}^{K} \alpha_i a_{ik} - \min_{i\in C(\sigma(k))} h_{ik} - \sum_{\substack{j=1 \\ j\ne\sigma(k)}}^{J} \Big(\min_{i\in C(j)} h_{ik}\Big)^- < 0,$$

where H *is as defined in Theorem* 3.4.

REMARKS.

1. Condition (3.10) is the same as the stability condition in Theorem 7 of Kumar and Meyn (1993) for queueing networks. Therefore, the fluid networks that correspond to the stable queueing networks in Kumar and Meyn (1993) are stable.
2. Kumar and Meyn (1993) formulated the problem of verifying condition (3.10) as a linear programming problem. There, several examples are given to show the stability of some queueing networks.

PROOF OF THEOREM 3.5. Suppose that Q is a work-conserving fluid level process of the fluid network (α, μ, P, C) with $e'Q(0) = 1$. Let $f(t) = Q'(t)AQ(t)$ and

$$-\theta_k = \sum_{i=1}^{K} \alpha_i a_{ik} - \min_{i\in C(\sigma(k))} h_{ik} - \sum_{\substack{j=1 \\ j\ne\sigma(k)}}^{J} \Big(\min_{i\in C(j)} h_{ik}\Big)^-.$$

Then $\theta = (\theta_k) > 0$. It follows from the proof of Theorem 3.4 that

$$(3.11)\qquad df(t) \le -\theta' Q(t)\, dt,$$

for all work-conserving fluid level processes Q with $e'Q(0) = 1$. Let $u(t) = \sqrt{f(t)}$. Then it follows from (3.11) that

$$(3.12)\qquad du(t) \le -\gamma\, dt \quad \text{for all } t \ge 0,$$

whenever $e'Q(t) > 0$, with

$$\gamma = \frac{1}{2} \inf_{\substack{x\ge 0 \\ x\ne 0}} \frac{\theta' x}{\sqrt{x'Ax}} = \frac{1}{2} \inf_{\substack{x\ge 0 \\ e'x=1}} \frac{\theta' x}{\sqrt{x'Ax}}.$$

Clearly, $\gamma > 0$. Thus, (3.12) implies that $Q(t) = 0$ for $t \ge u(0)/\gamma$. $\square$

3.4. *Some additional properties for stability.* The first theorem is a monotone property for the stability, and the second is on the stability of a subnetwork of a stable fluid network.

THEOREM 3.6. *Suppose that a fluid network* (α^1, P, μ, C) *is weakly (respectively, strongly) stable. Then the fluid network* (α^2, P, μ, C) *with* $\alpha^2 \leq \alpha^1$ *is also weakly (respectively, strongly) stable.*

REMARK. It is plausible that a similar property also holds for varying the service rate μ and the flow-transfer matrix P: increasing the service rate and decreasing the flow-transfer matrix preserve the (weak and strong) stabilities. However, this does not hold in general. Counterexamples have been found by Maury Bramson and were brought to my attention by Jim Dai.

PROOF OF THEOREM 3.6. We prove the weak stability case only; the proof for the stability case is similar. Suppose that the fluid network (α^2, P, μ, C) is not stable. Then there exists a work-conserving allocation T^2 such that its associated fluid level process $Q^2 \not\equiv 0$, but $Q^2(0) = 0$.

From Remark 1 after Theorem 2.1, there exist an allocation process T^0 and its associated fluid level process Q^0 for the fluid network $(\alpha^1 - \alpha^2, P, \mu, C)$ satisfying (2.5) and (2.6) and (2.7′) and (2.8′) with $I^0(t) = U^2(t) \equiv et - CT^2(t)$.

Let $T^1(t) = T^2(t) + T^0(t)$ and $Q^1(t) = Q^2(t) + Q^0(t)$. If we can prove that T^1 and Q^1 are a work-conserving pair for the fluid network (α^1, P, μ, C), then we reach a contradiction that the fluid network is weakly stable, since $Q^1(0) = 0$, but $Q^1 \geq Q^2 \not\equiv 0$. For the pair, a direct verification shows that it satisfies (2.5)–(2.7). For (2.8), note that

$$\begin{aligned}(CQ^1(t))'d[et - CT^1(t)] &= (CQ^2(t) + CQ^0(t))'d[U^2(t) - CT^0(t)] \\ &= (CQ^2(t))'d[U^2(t) - CT^0(t)] = 0,\end{aligned}$$

where the second equality is due to (2.8′). The last equality of the above equation follows from the statement that if $(CQ^2(t))_j > 0$, then $(U^2)_j$ must not increase at time t (since T^2 and Q^2 are a work-conserving pair). In this case, $(CT^0)_j$ must not increase at time t as well (since $U^2 - CT^0$ is nondecreasing). Therefore, it must be that $(d[U^2(t) - CT^0(t)])_j = 0$. □

Consider a fluid network (α, P, μ, C). Let a be a subset of its classes and let b be the complement of a. A subnetwork of this fluid network consisting of classes a is constructed as follows. The set of stations for the subnetwork is given by $s(a) = \{\sigma(k)\colon\ k \in a\}$. Its constituent matrix $\tilde{C}_a$ is a submatrix of C with row indices $s(a)$ and column indices a. The exogenous inflow rate and the flow-transfer matrix are, respectively, given by

$$\begin{aligned}\tilde{\alpha}_a &= \alpha_a + P'_{ba}[I - P'_b]^{-1}\alpha_b, \\ \tilde{P}_a &= P_a + P_{ab}[I - P_b]^{-1}P_{ba}.\end{aligned}$$

The construction of the fluid network $(\tilde{\alpha}_a, \tilde{P}_a, \mu_a, \tilde{C}_a)$ here is a generalization of the single class $(J = K)$ case in Chen and Mandelbaum (1991a), to which readers are referred for the interpretation of $\tilde{\alpha}$ and $\tilde{P}$. There it is proved that the original network and the constructed network have the same traffic intensities and hence are either both stable or both weakly stable or both unstable. The generalization here is as follows. Its proof resembles that in Chen and Mandelbaum (1991a) and thus is omitted.

THEOREM 3.7. *If the fluid network (a, P, μ, C) is weakly stable (respectively, stable), then the fluid network $(\tilde{\alpha}_a, \tilde{P}_a, \mu_a, \tilde{C}_a)$ is weakly stable (respectively, stable).*

4. Multiclass queueing networks and their fluid approximations. In this section, it is established that the fluid approximation for a multiclass queueing network exists if its corresponding multiclass fluid network is weakly stable. Conversely, it is proved that if a multiclass queueing network has a traffic intensity less than unity, and its fluid approximation exists, the fluid limit for the queue length process must be zero. Finally, a two-station network example is given whose parameters satisfy traffic condition (3.3). However, the corresponding queueing network does not have a fluid limit, and the corresponding fluid network is not weakly stable and hence is not stable.

4.1. *Multiclass queueing networks.* The queueing network consists of J single-server stations indexed by $j = 1, \ldots, J$ and K job classes indexed by $k = 1, \ldots, K$. Jobs of class k are exclusively served by station $\sigma(k)$ with $\sigma(\cdot)$ as in Section 2.

The evolution of the network is modelled in terms of a random vector and three stochastic processes, all defined on a common probability space. These are a K-dimensional vector $Q(0) = (Q_k(0))$, two K-dimensional counting processes $A = \{A(t), t \geq 0\}$ and $S = \{S(t), t \geq 0\}$ and a $K \times K$ matrix process $R = \{R(n), n = 1, 2, \ldots\}$. The components of all these vectors and vector and matrix processes take nonnegative integer values. Vector $Q(0)$ is referred to as an *initial queue length* vector; its kth component indicates the number of class k jobs initially present in the network. Process A is referred to as an *arrival* process with $A_k(t)$ indicating the number of class k jobs that arrive exogenously during the time interval $(0, t]$, and process S is referred to as a *service* process with $S_k(t)$ indicating the number of class k jobs that can be served by station $\sigma(k)$ during the first t units of time allocated to the service of class k jobs. Process R is referred to as a *routing sequence*; its (i, k)th component evaluated at n, $R_k^i(n)$, indicates the number of class i jobs, among the first n class i jobs served at station $\sigma(i)$, that switch into class k jobs. Note that $n - \sum_{k=1}^{K} R_k^i(n)$ must be nonnegative, indicating the number of class i jobs, among the first n class i jobs served, that leave the network.

Let $Q_k(t)$ denote the number of class k jobs in the network at time $t \geq 0$ (either served or waiting), and let $T_k(t)$ denote the cumulative amount of time that has been allocated to their service during $[0, t]$. Then the *queue*

length process of class k, $Q_k = \{Q_k(t),\ t \geq 0\}$, and the *allocation processes* $T_k = \{T_k(t),\ t \geq 0\}$ are related via the *flow-balance* relations

$$Q_k(t) = Q_k(0) + A_k(t) + \sum_{i=1}^{K} R_k^i(S_i(T_i(t))) - S_k(T_k(t)) \geq 0. \tag{4.1}$$

The total queue length at station j, $Z_j = \{Z_j(t),\ t \geq 0\}$, and the cumulative busy time of station j, $B_j = \{B_j(t),\ t \geq 0\}$, can be represented by

$$Z_j(t) = \sum_{k \in C(j)} Q_k(t) \quad \text{and} \quad B_j(t) = \sum_{k \in C(j)} T_k(t),$$

where $C(j) = \{k\colon\ \sigma(k) = j\}$ is the constituent set as defined in Section 2. The *idleness* process (unused capacity) of station j, $U_j = \{U_j(t),\ t \geq 0\}$, is given by $U_j(t) = t - B_j(t)$, $j = 1, \ldots, J$. Clearly, it must be true that

$$T_k(\cdot) \text{ is nondecreasing with } T_k(0) = 0, \qquad k = 1, \ldots, K, \tag{4.2}$$

$$U_j(\cdot) \text{ is nondecreasing}, \qquad j = 1, \ldots, J. \tag{4.3}$$

An allocation $T = (T_k)$ is called *feasible* if it satisfies (4.1)–(4.3), and a feasible allocation is *work-conserving* if it also satisfies

$$U_j(t) = \int_0^t 1[Z_j(u) = 0]\, du, \qquad t \geq 0,$$

for $j = 1, \ldots, J$, which is equivalent to

$$Z_j(t)\, dU_j(t) = 0, \qquad t \geq 0. \tag{4.4}$$

In words, each station j idles only when there are no jobs at station j, or as stated in (2.4).

The fact that both U_j and B_j are monotone implies that both are Lipschitz continuous, and consequently all the T_k's are as well (the Lipschitz constant equals unity). The existence of a feasible allocation is clear: simply take $T \equiv 0$. The construction of work-conserving allocations is postponed to the end of this subsection.

Hereafter we assume that our queueing network has an *exogenous arrival-rate* vector α, a potential *service-rate* vector μ and a *routing-rate* matrix $P = (p_{jk})$. These are, respectively, given by

$$\lim_{t \to \infty} \frac{A(t)}{t} = \alpha, \tag{4.5}$$

$$\lim_{t \to \infty} \frac{S(t)}{t} = \mu, \tag{4.6}$$

$$\lim_{t \to \infty} \frac{1}{n} R(n) = P. \tag{4.7}$$

We also assume that the routing matrix P is substochastic with a spectral radius less than unity, thus restricting attention to *open* networks.

Let $\lambda = [I - P']^{-1}\alpha$, $\beta = M^{-1}\lambda = M^{-1}[I - P']^{-1}\alpha$ and $\rho = C\beta = CM^{-1}[I - P']^{-1}\alpha$, where C is the constituent matrix and $M = \text{diag}(\mu)$, both the same as defined in Section 2. Vector ρ is known as the *traffic intensity*. Note that λ is the solution to

$$\lambda = \alpha + P'\lambda.$$

This equation is known as a *traffic equation*. For single class queueing networks ($J = K$ case), it is proved that when $\rho \leq e$, λ_k, the kth component of λ, gives the long-run average arrival rate of class k jobs, derived from both exogenous and endogenous arrivals; thus, λ is known as an *effective arrival rate*. The counterexamples by Rybko and Stolyar (1992), Seidman (1993) and Bramson (1994) imply that, for multiclass queueing networks, λ may not be the effective arrival rate (meaning the long-run average arrival rate), even when $\rho < e$. Identifying the effective arrival rate in a multiclass queueing network is further complicated by its dependence on service disciplines. Chen and Mandelbaum (1991b) contains a preliminary discussion on the traffic equation and the effective arrival rate for multiclass queueing networks, in particular, under the first-in first-out service disciplines.

Now we consider the construction of work-conserving allocation processes for queueing networks under several service disciplines. The FIFO discipline is discussed in detail in Chen and Zhang (1994), and will not be discussed here.

The head-of-line processor sharing discipline can be constructed as follows [see Johnson (1983)]:

$$T_k(t) = \int_0^t \frac{1[Q_k(u) > 0]}{\sum_{i \in C(\sigma(k))} 1[Q_i(u) > 0]}\, du, \tag{4.8}$$

where the integrand is taken to be zero if its denominator equals zero. By induction on all jump points (starting from time 0), it can be shown that there exists a unique allocation process satisfying relations (4.1)–(4.4) and (4.8), that is, relations (4.1)–(4.4) and (4.8) well define the head-of-line processor sharing allocation process.

Next, we construct an allocation process for a priority discipline with preemption. Without loss of generality, we assume that $C(j) = \{k\colon\ \sigma(k) = j\} = \{k_{j1}, \ldots, k_{jn_j}\}$ and class k_{jl} has higher priority than class $k_{jl'}$ if $1 \leq l \leq l' \leq n_j$, $j = 1, \ldots, J$. Then for $j = 1, \ldots, J$,

$$T_{k_{j1}}(t) = \int_0^t 1[Q_{k_{j1}}(u) > 0]\, du, \tag{4.9}$$

$$T_{k_{jl}}(t) = \int_0^t 1[Q_{jl'}(u) = 0,\ l' = 1, \ldots, l-1, Q_{jl}(u) > 0]\, du, \tag{4.10}$$

$$l = 2, \ldots, n_j.$$

Similarly it can be shown that there exists a unique allocation process satisfying (4.1)–(4.4) and (4.9)–(4.10), which describes the dynamics of the priority discipline with preemption.

The construction of an allocation process under a priority discipline without preemption is more involved, and we leave it to the Appendix. Again, the allocation process satisfies relations (4.1)–(4.4), with some additional relations.

Finally, we point out that it may *not* be true that all work-conserving service disciplines are described by relations (4.1)–(4.4). The dynamics of some processor sharing disciplines may not take the form given by relations (4.1)–(4.4). In particular, consider the processor sharing based on jobs (i.e., at any time t, every job residing at a station shares the same proportion of the processing capacity of that station). If the time interval between two consecutive jumps of the service (counting) process S_k indicates the service time for a class k job, then $S_k(T_k(t))$ does *not* indicate the number of departed class k jobs. Therefore, relations (4.1)–(4.4) may not describe the dynamics for all work-conserving service disciplines. In this paper, we restrict our attention to the work-conserving service disciplines that can be described by relations (4.1)–(4.4).

4.2. *Fluid approximations for multiclass queueing networks.* Consider a multiclass queueing network as described in Section 4.1, which operates under work-conserving disciplines and satisfies (4.5)–(4.7). As in Section 4.1, the queueing length process, the allocation process, the busy time process and the idle time process are denoted by Q, T, B and U, respectively. Corresponding to the queueing network, a fluid network (α, μ, P, C) can be constructed, where parameters α, μ and P are from (4.5)–(4.7), and the matrix C is from the constituent matrix of the queueing network.

THEOREM 4.1. *If the corresponding fluid network (α, μ, P, C) is weakly stable, then the multiclass queueing network under all work-conserving service disciplines has the same fluid limits:*

$$\frac{1}{n}(Q(nt), T(nt), B(nt), U(nt)) \to (\overline{Q}(t), \overline{T}(t), \overline{B}(t), \overline{U}(t)), \quad \text{u.o.c.}, \tag{4.11}$$

as $n \to \infty$, where $\overline{Q}(t) = 0$, $\overline{T}(t) = \beta t$, $\overline{B}(t) = \rho t$ and $\overline{U}(t) = (e - \rho)t$.

REMARKS.

1. The convergence (4.11) is also known as the functional strong law of large numbers theorem for multiclass queueing networks. This generalizes the result of Chen and Mandelbaum (1991a) for a single class queueing network, where condition (3.3) is proved to be sufficient for the convergence (4.11).
2. The existence of a fluid limit in the theorem is for *all* work-conserving service disciplines, thus, more restrictive than a specific service discipline. For example, it may well be the case where for given parameters (α, μ, P, C), the queue length process has a fluid limit under one work-conserving service discipline, but does not have a fluid limit under another work-conserving service discipline. (See Section 6 for more discussion.)

3. The traffic intensity condition (3.1), even the stronger condition (3.3), is not sufficient for the existence of the fluid limit (4.11), which is demonstrated by the example in Section 4.3.

PROOF OF THEOREM 4.1. First, as in Section 2.3 of Chen and Mandelbaum (1991a), existence of the long-run averages (4.5)–(4.7) is equivalent to

$$\frac{1}{n}A(nt) \to \alpha t, \quad \text{u.o.c.}, \tag{4.12}$$

$$\frac{1}{n}S(nt) \to \mu t, \quad \text{u.o.c.}, \tag{4.13}$$

$$\frac{1}{n}R(\lfloor nt \rfloor) \to Pt, \quad \text{u.o.c.}, \tag{4.14}$$

as $n \to \infty$. Clearly, the convergence,

$$\frac{1}{n}Q(0) \to 0, \quad \text{a.s.}, \tag{4.15}$$

holds.

Let $\overline{T}^n(t) = T(nt)/n$. Then clearly, for any $t \geq s \geq 0$,

$$0 \leq \overline{T}_k^n(t) - \overline{T}_k^n(s) \leq t - s.$$

Hence $\{\overline{T}^n(t),\ n \geq 1\}$ are uniformly Lipschitz. By the Arzela–Ascoli theorem, any subsequence of $\overline{T}^n(t)$ has a u.o.c. convergent subsequence. Next, we can show that any limit $\overline{T}$ of any subsequence of $\overline{T}^n$ must be a work-conserving allocation for the fluid network (α, μ, P, C) with zero initial fluid level. Then, by Theorem 3.1 and the weak stability of the fluid network, the limit must be $\overline{T}(t) = \beta t$, thus proving the convergence of $\overline{T}^n$. The rest follow immediately. □

THEOREM 4.2. *Suppose that a multiclass queueing network under a (specific) work-conserving service discipline has the fluid limit*

$$\frac{1}{n}(Q(nt), T(nt)) \to (\overline{Q}(t), \overline{T}(t)), \quad \textit{u.o.c., as } n \to \infty, \tag{4.16}$$

and that

$$\rho = CM^{-1}[I - P']^{-1}\alpha \leq e. \tag{4.17}$$

Then $\overline{Q}(t) = 0$ *and* $\overline{T}(t) = \beta t$.

REMARK. It is easy to show that the fluid limit of a queueing network, if it exists, must be a fluid network, but it remains to show that a fluid network must be the limit of a queueing network. If the latter is the case, the theorem implies that the converse of Theorem 4.1 holds; that is, if the fluid limit exists for a queueing network under all work-conserving service disciplines, then the corresponding fluid network must be weakly stable.

PROOF OF THEOREM 4.2. Since T is nondecreasing and Lipschitz continuous, it follows from convergence (4.16) that $T(t)/t$ converges to $b := \overline{T}(1)$ as $t \to \infty$. Note that $T(nt)/n = (T(nt)/(nt))t$; thus, $\overline{T}(t) = bt$. Next, scaling the time in (4.1)–(4.4) by a factor of n and scaling all the processes by a factor of $1/n$, and then taking limits as $n \to \infty$, yields (in vector form)

$$q := \alpha - [I - P']Mb \geq 0, \tag{4.18}$$

$$b \geq 0, \tag{4.19}$$

$$e - Cb \geq 0, \tag{4.20}$$

$$[Cq]'[e - Cb] = 0, \tag{4.21}$$

where we used the fact that the limiting processes $(\overline{T}, \overline{Q}, \overline{U})$ must jointly satisfy (2.1)–(2.4), and that $\overline{T}(t) = bt$, $\overline{U}(t) = (e - Cb)t$ and $\overline{Q}(t) = qt$.

Now it suffices to show that (4.18)–(4.21) have a unique solution $b = \beta$ and $q = 0$. If $q_k > 0$ for some index k, say k_0, then clearly $[Cq]_{\sigma(k_0)} > 0$; this, together with (4.21), implies

$$\sum_{k \in C(\sigma(k_0))} b_k = 1. \tag{4.22}$$

On the other hand, using (4.18) and $q_{k_0} > 0$ yields

$$\beta - b = M^{-1}[I - P']^{-1}q \geq 0,$$

$$(\beta - b)_{k_0} = (M^{-1}[I - P']^{-1}q)_{k_0} > 0,$$

implying

$$\rho_{\sigma(k_0)} = \sum_{k \in C(\sigma(k_0))} \beta_k > \sum_{k \in C(\sigma(k_0))} b_k = 1,$$

where (4.22) is used in the last equality. The above inequality contradicts the assumption (4.17). Now it is proved that $q = 0$ must hold; this clearly implies $b = \beta$, which satisfies (4.18)–(4.21). □

4.3. *An example.* The example, adapted from Kumar and Seidman (1990), is a network with $J = 2$, $K = 4$, $C(1) = \{1, 4\}$, $C(2) = \{2, 3\}$ and

$$\alpha = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mu = \begin{pmatrix} 4 \\ 3/2 \\ 4 \\ 3/2 \end{pmatrix}, \qquad P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.23}$$

In this case,

$$\beta = M^{-1}[I - P']^{-1}\alpha = \begin{pmatrix} 1/4 \\ 2/3 \\ 1/4 \\ 2/3 \end{pmatrix} \quad \text{and} \quad \rho = C\beta = \begin{pmatrix} 11/12 \\ 11/12 \end{pmatrix}.$$

Therefore, condition (3.3) prevails.

THEOREM 4.3. *The fluid network* (α, P, μ, C) *defined above is not weakly stable and hence is not stable.*

THEOREM 4.4. *Let a queueing network have deterministic interarrival and service times with parameters as defined above. Suppose the initial queue length* $Q(0) = (\gamma, 0, 0, 0)$. *Then for any given* γ *large enough, there exists a work-conserving service discipline such that the fluid limit* (4.11) *does not exist. In particular, we have*

$$\limsup_{t\to\infty} \frac{Q(t)}{t} = \left(\frac{1}{2}, \frac{3}{7}, \frac{1}{2}, \frac{3}{7}\right) \quad \textit{and} \quad \liminf_{t\to\infty} \frac{Q(t)}{t} = 0, \tag{4.24}$$

$$\limsup_{t\to\infty} \frac{T(t)}{t} = \beta \quad \textit{and} \quad \liminf_{t\to\infty} \frac{T(t)}{t} = \left(\frac{1}{8}, \frac{1}{3}, \frac{1}{8}, \frac{1}{3}\right), \tag{4.25}$$

where both lim sup *and* lim inf *are taken coordinatewise.*

REMARKS.

1. Although our proof calls for $\gamma > 5$, a more tedious argument proves that $\gamma \geq 2$ is sufficient for the theorem to hold.
2. The fluid limit (4.11) does exist if the queueing network specified in the theorem has an initial queue length $Q(0) = (0, 0, 0, 0)$ (i.e., the system starts empty).

The proof of Theorem 4.3 follows from Theorems 4.1 and 4.4, while the proof of Theorem 4.4 is quite tedious and therefore is included in the Appendix.

Through analyzing the fluid network, we give a heuristic proof for Theorems 4.3 and 4.4, which may provide some insight into the stability of this network. (The heuristic is actually rigorous for the stability part of Theorem 4.3.) Let the initial fluid level $Q(0) = (\gamma, 0, 0, 0)$ with $\gamma > 0$. The dynamics of the fluid network is as follows: at time $t = 0$, buffer 1 processes fluid 1 and buffer 2 processes fluid 2, both at full capacity, until time $t_1 = \gamma/3$. Then $Q(t_1) = (0, 5\gamma/6, \gamma/3, 0)$, buffer 1 is forced to process at one quarter of its capacity, since it is empty, and buffer 2 processes at its full capacity. This continues until time $t_2 = 2\gamma$, at which $Q(t_2) = (0, 0, 2\gamma, 0)$, that is, the inventory levels for both fluids 1 and 2 are zero. Next, we switch the roles of buffers 1 and 2, fluids 1 and 3 and fluids 2 and 4. Similarly we continue until the inventory levels for both fluids 3 and 4 are zero, to obtain that $Q(6\gamma) = (4\gamma, 0, 0, 0)$. The

fluid level is four times the initial fluid level. Continuing with this process, we find the following work-conserving pair:

$$Q(t) = \begin{cases} \left(\gamma_n - 3(t-t_n), \frac{5}{2}(t-t_n), t-t_n, 0\right), & t_n \le t < t_n + \frac{1}{3}\gamma_n, \\ \left(0, \gamma_n - \frac{1}{2}(t-t_n), t-t_n, 0\right), & t_n + \frac{1}{3}\gamma_n \le t < t_n + 2\gamma_n, \\ \left(t-t_n-2\gamma_n, 0, 2\gamma_n - 3(t-t_n-2\gamma_n), \right. & \\ \qquad \left. \frac{5}{2}(t-t_n-2\gamma_n)\right), & t_n + 2\gamma_n \le t < t_n + \frac{8}{3}\gamma_n, \\ \left(t-t_n-2\gamma_n, 0, 0, \frac{5}{3}\gamma_n - \frac{1}{2}\left(t-t_n-\frac{8}{3}\gamma_n\right)\right), & t_n + \frac{8}{3}\gamma_n \le t < t_{n+1}, \end{cases}$$

and

$$\frac{dT(t)}{dt} = \begin{cases} (1,1,0,0), & t_n \le t < t_n + \frac{1}{3}\gamma_n; \\ \left(\frac{1}{4},1,0,0\right), & t_n + \frac{1}{3}\gamma_n \le t < t_n + 2\gamma_n; \\ (0,0,1,1), & t_n + 2\gamma_n \le t < t_n + \frac{8}{3}\gamma_n; \\ \left(0,0,\frac{1}{4},1\right), & t_n + \frac{8}{3}\gamma_n \le t < t_{n+1}, \end{cases}$$

where $t_n = 2(4^n - 1)\gamma$, $\gamma_n = 4^n\gamma$ and $t_{n+1} = t_n + 6\gamma_n$. This clearly implies that the network is not stable. In addition, it implies

$$Q(t_n) = (\gamma_n, 0, 0, 0) \quad \text{and} \quad Q(t_n + 2\gamma_n) = (0, 0, 2\gamma_n, 0),$$

which implies that $Q(n\gamma)/n$ does *not* converge, since one of its subsequences, with $n = 2(4^k - 1)$, converges to $(1/2, 0, 0, 0)$ as $k \to \infty$, and one of the other subsequences, with $n = 4 \times 4^k - 2$, converges to $(0, 0, 1/2, 0)$ as $k \to \infty$. Similarly, the process $T(nt)/n$ does not converge. Thus, we prove that the fluid limit (4.11) does not exist for the fluid level and allocation processes with initial inventory level $Q(0) = (\gamma, 0, 0, 0)$.

5. Stability of multiclass queueing networks. Based on Theorem 4.3 in Dai (1994), we establish that the stability of the corresponding fluid network suffices for the stability of a multiclass queueing network.

Throughout this section, the queueing network is as described in Section 4. Assume that both the arrival and the service processes are renewal processes, and the routing matrix process R is constructed from an iid sequence [i.e., $\{R(n) - R(n-1),\ n \ge 1\}$ is an iid sequence, where $R(0) = 0$]. Also assume that the arrival process A, the service process S and the routing process R are mutually independent. Furthermore, assume that the interarrival times of the arrival process are unbounded and spread out (thus, excluding deterministic interarrival times).

A queueing network under a specific service discipline is *stable* if the underlying Markov process that describes the dynamics of the queueing network is positive Harris recurrent. That is, the Markov process has a unique invariant probability measure. Readers are referred to Dai (1995) for a precise definition.

The specific form of the underlying Markov process may vary with service disciplines. Readers are referred to Dai (1995) for a list that includes FIFO, priorities and head-of-line processor-sharing disciplines.

The following theorem is a variation of Theorem 4.3 in Dai (1995), where the workload formulation is replaced by the fluid level formulation. [It is consistent with the latest version of Dai (1995).]

THEOREM 5.1. *The queueing network is stable if there exists a constant* t_0 *that depends on* (α, μ, P, C) *only, such that, for any* (Q, Z, T, U) *and* v *satisfying* (5.1)–(5.5) *below,* $Z(t) = 0$ *for all* $t \geq t_0$:

$$Q(t) = Q(0) + Mv + \operatorname{diag}(\alpha)(et - u)^+ + P'M[T(t) - v]^+ - MT(t) \geq 0, \tag{5.1}$$

$$dT(t) \geq 0 \quad \text{with } T(0) = 0, \tag{5.2}$$

$$U(t) := et - CT(t) \quad \text{and} \quad dU(t) \geq 0, \tag{5.3}$$

$$Z(t) := CQ(t) \quad \text{and} \quad Z'(t)\,dU(t) = 0, \tag{5.4}$$

$$e'[M^{-1}Q(0) + v] \leq 1, \tag{5.5}$$

where $u \geq 0$ *and* $v \geq 0$.

REMARKS.

1. In Dai (1995), any processes (Q, Z, T, U) satisfying (5.1)–(5.5) are referred to as a fluid limit model for the multiclass queueing network. The fluid limit model is said to be *stable* if there exists a finite t_0 such that for any fluid limit, $Q(t) = 0$ or equivalently $Z(t) = 0$ for all $t \geq t_0$.
2. The theorem is proved for service disciplines such as FIFO, priority with and without preemption and head-of-line processor sharing. It has not been proved for more general processor sharing.

Theorem 5.1 [which is from Theorem 4.3 in Dai (1995)] significantly simplifies the process of verifying the stability of multiclass queueing networks to the verification of the stability of a piecewise linear fluid network. In fact, the verification of the latter can be further simplified to the verification of a *linear* fluid model, which is summarized as follows. (The proof is given at the end of this section.)

THEOREM 5.2. *If the linear fluid model corresponding to the queueing network is stable, then the queueing network is stable.*

REMARKS. Combining this theorem and Remark 2 after Theorem 3.2 provides a simple proof for the stability of the generalized Jackson network and the feedforward queueing network, under the traffic intensity condition (3.3). This theorem is also used in Dai (1995) to reach the same conclusion.

COROLLARY 5.3. *If condition* (3.10) *holds, then the queueing network under all work-conserving service disciplines that are included in Theorem* 5.1 *is stable.*

REMARK. This generalizes the stability result of Kumar and Meyn (1993) to nonexponential interarrival and service times. On the verification of condition (3.10) and the examples of stable multiclass queueing networks, see Remark 2 that follows Theorem 3.5.

As a preparation for the proof of Theorem 5.2, we prove the following lemma.

LEMMA 5.4. *Consider a linear fluid network* (α, μ, P, C) *and suppose that its traffic intensity* $\rho < e$. *Then for any work-conserving total fluid level process* $Z(t) = CQ(t)$ *and* $j = 1, \ldots, J$, *the set* $\{t\colon Z_j(t) = 0\}$ *is unbounded. In other words, there does not exist a finite* t_0 *such that* $Z_j(t) > 0$ *for all* $t \geq t_0$.

PROOF. If such a t_0 exists, from (2.7), it must be true that

$$d\left[t - \sum_{k \in C(j)} T_k(t)\right] = 0 \quad \text{for all } t \geq t_0. \tag{5.6}$$

On the other hand, it follows from (2.5) that

$$CT(t) \leq CM^{-1}[I - P']^{-1}Q(0) + \rho t.$$

This, together with $\rho < e$, contradicts (5.6). □

PROOF OF THEOREM 5.2. Consider T and Q satisfying (5.1)–(5.5). Suppose that there exists a finite t' such that $et \geq u$ and

$$T(t) \geq v \quad \text{for } t \geq t'. \tag{5.7}$$

Then (5.1) takes the form

$$Q(t) = \tilde{Q}(0) + \alpha t - [I - P']MT(t) \quad \text{for } t \geq t', \tag{5.8}$$

where

$$\tilde{Q}(0) = Q(0) - \operatorname{diag}(\alpha)u + [I - P']Mv.$$

Now using the shift property of the linear fluid model (in Section 2) yields that there exists a finite $t_0 > t'$, such that $Q(t) = 0$ for $t \geq t_0$. This, combined with Theorem 5.1, proves the theorem.

Returning to (5.7), we actually prove a stronger result that $T_k(\infty) = \infty$ for all k (noting that T_k is nondecreasing). Otherwise, there exists an allocation T satisfying (5.1)–(5.5) such that

$$a = \{k\colon T_k(\infty) < \infty\} \neq \varnothing.$$

Let b be the complement of a.

Consider $t \geq \max_{1 \leq k \leq K} u_k$ and write the subblock of (5.1) corresponding to a:

$$\begin{aligned} Q_a(t) = \tilde{Q}_a(0) + \alpha_a t + P'_{ba}M_b[T_b(t) - v_b]^+ \\ + P'_aM_a[T_a(t) - v_a]^+ - M_aT_a(t) \geq 0, \end{aligned} \tag{5.9}$$

where $\overline{Q}_a(0)$ is the corresponding subblock of

$$\tilde{Q}(0) = Q(0) + Mv - \text{diag}(\alpha)u.$$

Since $T_k(\infty) = \infty$, for $k \in b$, and $T_k(\infty) < \infty$, for $k \in a$, it follows from (5.9) we must have $\alpha_a = 0$ and $P_{ba} = 0$; otherwise, at least one of the coordinates of $Q_a(t)$ must approach infinity as $t \to \infty$, contradicting Lemma 5.4. Writing the traffic equation $\lambda = \alpha + P'\lambda$ with coordinates in a yields

$$\lambda_a = \alpha_a + P'_{ba}\lambda_b + P'_a\lambda_a.$$

When $\alpha_a = 0$ and $P_{ba} = 0$, the above implies that $\lambda_a = 0$, contradicting our assumption that $\beta = M^{-1}\lambda > 0$ (see Remark 1 after Theorem 3.1). □

6. Concluding remarks. Based on Dai (1995), we relate the stability of a multiclass queueing network to the stability of a linear fluid network. In addition, we relate the existence of the strong law-of-large-numbers theorem (fluid approximations) for a multiclass queueing network to the weak stability of a linear fluid network. Using an approach similar to Kumar and Meyn (1993), sufficient conditions are identified for both stability and weak stability of a fluid network.

The notions of stability considered here are concerned with general work-conserving service disciplines. However, a queueing network (as well as a fluid network) may be stable under one service discipline, but not under another. A two-station queueing network in Rybko and Stolyar (1992) provides a good example where the network is stable under FIFO but is unstable under a priority service discipline. Thus, the notions of stability considered are more restrictive when applied to a queueing network (respectively, a fluid network) under a specific service discipline (respectively, a specific class of allocation processes). However, the idea of the current approach can be extended to those cases as well. Chen and Zhang (1994) investigate a multiclass queueing (and the corresponding fluid) network under FIFO service discipline. As a simple example, we illustrate the idea by considering a multiclass queueing network under a priority service discipline with preemption. In this case, an allocation is feasible if in addition to (4.1)–(4.4), it also satisfies (4.9) and (4.10). The dynamics of the corresponding linear fluid network is described by (A.1)–(A.3) (in the Appendix), in addition to (2.5)–(2.8); these relations also define the feasibility of the allocation process for the fluid network. Accordingly, we can define the weak stability and the stability of the linear fluid network, and then establish that weak stability and strong stability are sufficient for the fluid approximation and the stability of queueing network, respectively. As the set of feasible allocation processes is smaller in this case, the conditions for both weak and strong stabilities are in general weaker than conditions (3.6) and (3.10), respectively.

Kumar and Meyn (1993) developed a linear programming approach to identify the given parameters of a queueing network that satisfy the sufficient condition (3.10) for the stability of the network. However, so far there is no explicit condition on the given parameters such that a fluid network is stable,

which is sufficient for the stability of the corresponding queueing network. Note that stability is weaker but less explicit than condition (3.10). It is our belief that the stability of a fluid network is also necessary for the stability of the corresponding queueing network (under all work-conserving service disciplines), at least for the Markovian network.

APPENDIX

A.1. Proof of Theorem 2.1. First we state a lemma, whose proof is elementary and hence omitted.

LEMMA A.1.

1. *If f is a Lipschitz continuous function, then*

$$h(t) = \sup_{0\le u\le t} f(u)$$

is also Lipschitz continuous and both have the same Lipschitz constants.
2. *If f and g are Lipschitz continuous, then $f - g$ is Lipschitz continuous.*

PROOF OF THEOREM 2.1. Let $C(j) = \{k\colon \sigma(k) = j\} = \{k_{j1}, \dots, k_{jn_j}\}$, $j = 1, \dots, J$. Note that $n_j \ge 1$ for all j and $\sum_{j=1}^{J} n_j = K$. Hence,

$$d := \max\{n_j\colon\ j = 1, \dots, J\} \le K - J + 1.$$

Let $h_l = \{k_{jl}\colon\ j = 1, \dots, J\}$ be a subset of class indices, and let $s(h_l) = \{s(k_{jl}) : j = 1, \dots, J\}$ be a subset of station indices, where both k_{jl} and $s(k_{jl})$ are understood to be null elements if $l > n_j$. Hence, both h_l and $s(h_l)$ may have less than J elements but both h_1 and $s(h_1)$ have exactly J elements and $s(h_1) = \{1, \dots, J\}$ is the set of all station indices.

First, we prove that there exists a process T satisfying

$$I^l_{s(h_l)}(t) = \sup_{0\le u\le t} \left\{ I^{l-1}_{s(h_l)}(u) - (M^{-1}[Q(0) + \alpha u + P'MT(u)])_{h_l} \right\}^+, \qquad l = 1, \dots, d, \tag{A.1}$$

$$I^0(t) = et, \tag{A.2}$$

$$T_{h_l}(t) = I^{l-1}_{s(h_l)}(t) - I^l_{s(h_l)}(t), \qquad l = 1, \dots, d. \tag{A.3}$$

Consider a sequence of K-dimensional processes T^n (with $T^0 \equiv 0$) generated by the following iteration:

$$I^{l,n+1}_{s(h_l)}(t) = \sup_{0\le u\le t} \left\{ I^{l-1,n+1}_{s(h_l)}(u) - (M^{-1}[Q(0) + \alpha u + P'MT^n(u)])_{h_l} \right\}^+, \qquad l = 1, \dots, d, \tag{A.4}$$

$$I^{0,n+1}(t) = et, \tag{A.5}$$

$$T_{h_l}^{n+1}(t) = I_{s(h_l)}^{l-1,n+1}(t) - I_{s(h_l)}^{l,n+1}(t), \qquad l = 1, \ldots, d. \tag{A.6}$$

By Lemma A.1, it can be shown inductively that $T_k^n(\cdot)$ and $I_j^{l,n}(\cdot)$ are nondecreasing with $T_k^n(0) = 0$ and are Lipschitz continuous with unity Lipschitz constants for all $k = 1, \ldots, K$, $l = 1, \ldots, d$ and $j = 1, \ldots, J$. Thus, the sequence $\{T^n,\ n \geq 0\}$ is equicontinuous and there exists a u.o.c. convergent subsequence. The limit denoted by T must satisfy (A.1)–(A.3) and its coordinate $T_k(\cdot)$ must be nondecreasing and Lipschitz continuous with Lipschitz constant 1. [We could not prove that (A.4)–(A.6) defines a contraction mapping; if so, then there exists a unique solution to (A.4)–(A.6). Note that the allocation process satisfying (A.4)–(A.6) appears to correspond to the priority service discipline. Given the counterexample of a two-station network in Kumar and Seidman (1990), it might well be that the mapping defined by (A.4)–(A.6) is not a contraction mapping.]

Now we show that the process T that satisfies (A.1)–(A.3) is a work-conserving allocation process. First, by our induction, T satisfies (2.6). Next, verifying that T satisfies the nonnegativity (2.5), we use (A.3) to yield

$$\begin{aligned} (M^{-1}Q(t))_{h_l} &= (M^{-1}[Q(0) + \alpha t + P'MT(t)])_{h_l} - T_{h_l}(t) \\ &= (M^{-1}[Q(0) + \alpha t + P'MT(t)])_{h_l} \\ &\quad - I_{s(h_l)}^{l-1}(t) + I_{s(h_l)}^{l}(t) \geq 0, \end{aligned} \tag{A.7}$$

where the last inequality follows from (A.1), thus proving the allocation satisfies (2.5).

By (A.2) and (A.3), it can be checked directly that

$$\begin{aligned} U_{s(h_d)}(t) &= I_{s(h_d)}^{d}(t), \\ U_{s(h_{l-1})\setminus s(h_l)}(t) &= I_{s(h_{l-1})\setminus s(h_l)}^{l-1}(t), \qquad l = d-1, \ldots, 2. \end{aligned}$$

By (A.1) and (A.3),

$$U_{s(h_d)(t)} = I_{s(h_d)}^{d}(t) \tag{A.8}$$

$$= \sup_{0 \leq u \leq t} \{I_{s(h_d)}^{d-1}(u) - (M^{-1}[Q(0) + \alpha u + P'MT(u)])_{h_d}\}^+ \tag{A.9}$$

$$= I_{s(h_d)}^{d-1}(t) - T_{h_d}(t). \tag{A.10}$$

It follows from (A.9) that $dU_{s(h_d)}(t) \geq 0$. By (A.1) and (A.7)–(A.9), coordinatewise the increase of $U_{s(h_d)}$ at time t implies $Q_{h_k}(t) = 0$ and implies the increase of $I_{s(h_d)}^{d-1}$ at time t. Next, note that, for $l = d-1, \ldots, 1$,

$$I_{s(h_l)}^{l}(t) = \sup_{0 \leq u \leq t} \{I_{s(h_l)}^{l-1}(u) - (M^{-1}[Q(0) + \alpha u + P'MT(u)])_{h_l}\}^+.$$

This, together with (A.1) and (A.7), implies that coordinatewise $I^l_{s(h_l)}$ increases at time t only when $Q_{h_l}(t) = 0$ and $I^l_{s(h_{l-1})}$ increases at time t. Thus, the allocation process satisfies (2.8). Therefore, it is a work-conserving allocation. □

A.2. Construction of the allocation for priority without preemption. For a pure jump function f, define its last jump time

$$l(t; f) = \sup\{s \le t:\ f(s-) - f(s) \ge 1\}.$$

We may simply write it as $l(t)$ when no confusion arises. Similar to the priority with preemption case, we assume that $C(j) = \{k_{j1}, \ldots, k_{jn_j}\}$ and the priorities for each class are in a decreasing order: $k_{j1}, \ldots, k_{jn_j}$. Define first the indicator sets: for $j = 1, \ldots, J$ and $l = 1, \ldots, n_j$,

$$\tilde{\mathcal{I}}_{k_{jl}}(t) = \{\exists\ u:\ l(t; Q_{k_{jl}}) \le u \le t \text{ such that } Q_{k_{jl'}}(u) = 0,\ l' = 1, \ldots, l-1,\ Q_{k_{jl}}(u) > 0\},$$

and then

$$\mathcal{I}_{k_{jn_j}}(t) = \tilde{\mathcal{I}}_{k_{jn_j}}(t),$$
$$\mathcal{I}_{k_{jl}}(t) = \tilde{\mathcal{I}}_{k_{jl}}(t) \cap (\mathcal{I}_{k_{j,l+1}}(t))^c, \qquad l = n_j - 1, \ldots, 1.$$

The allocation process is given by

$$T_{kl}(t) = \int_0^t 1[\mathcal{I}_{k_{jl}}(u)]\, du, \qquad l = 1, \ldots, n_j,\ j = 1, \ldots, J. \tag{A.11}$$

It can be shown that there exists a unique allocation process satisfying (4.1)–(4.4) and (A.11), which describes the dynamics of the priority discipline without preemption.

A.3. Proof of Theorem 4.4. The state of the queueing network is described jointly by the queue length Q and the residual times $((r_1, r_3), (s_1, s_2, s_3, s_4))$, where r_k and s_k are residual interarrival and residual service times, respectively, It is assumed that $0 < r_k \le 1$ and $0 \le s_k \le 1/4$, where $s_k = 0$ is interpreted as there being no class k jobs in the network.

We first state a lemma, whose proof is postponed to the end of the Appendix.

LEMMA A.2. *Suppose that* $Q(0) = (\gamma_0, 0, \delta_0, 0)$ *with* $\gamma_0 > 5$ *and* δ_0 *being either* 0 *or* 1. *Then there exists a time* t_1 *such that* $Q(t_1) = (\delta_1, 0, \gamma_1, 0)$, *where* δ_1 *is either* 0 *or* 1,

$$t_1 \le 2\gamma_0 + 21 \quad \textit{and} \quad \gamma_1 \ge 2\gamma_0 - 5. \tag{A.12}$$

PROOF OF THEOREM 4.4. Note that the queueing network is symmetric when we switch the roles of classes 1 and 3, classes 2 and 4 and stations 1 and 2. Now applying Lemma A.2 repeatedly, we would have a

sequence of $(t_n, \gamma_n, \delta_n)$ such that $Q(t_{2m}) = (\gamma_{2m}, 0, \delta_{2m}, 0)$ and $Q(t_{2m+1}) = (\delta_{2m+1}, 0, \gamma_{2m+1}, 0)$, where δ_n is either 0 or 1,

$$t_{n+1} - t_n \le 2\gamma_n + 21, \tag{A.13}$$

$$\gamma_{n+1} \ge 2\gamma_n - 5, \tag{A.14}$$

$t_0 = 0$ and $\gamma_0 = \gamma$. It follows from (A.13) and (A.14) that

$$\gamma_n \ge (\gamma - 5)2^n + 5,$$
$$t_n \le 2(\gamma - 5)(2^n - 1) + 31n.$$

This clearly implies the lim sup in (4.24), whereas the lim inf in (4.24) is implied by Lemma 5.4 (which can also be observed from the proof for Lemma A.2). The convergence (4.25) follows from (4.1) (also refer to the proofs of Theorem 4.1 and Lemma A.2). The convergences (4.24) and (4.25) clearly indicate that the fluid limit (4.11) does not exist. □

PROOF OF LEMMA A.2. The initial queue length is $Q(0) = (\gamma_0, 0, \delta_0, 0)$, and $\delta_0 = 1[s_3 > 0]$. Based on the residual arrival and service times, there are three possible cases.

Case (i): $s_1 \le s_3$ *or* $s_3 = 0$. At time s_1, we have

$$Q(s_1) = (\gamma_0 - 1 + 1[r_1 < s_1], 1, 1[s_3 > 0] + 1[r_3 < s_1], 0),$$

where r's and s's are the residual interarrival and service times at time 0. From time s_1, station 2 switches to serving class 2 jobs, and station 1 serves class 1 jobs (whenever they are available). This continues until $\gamma_0 + n_1 + 1$ jobs complete service at station 2, where n_1 is determined from

$$3s_1 - 3r_1 + 2\gamma_0 - \tfrac{7}{4} < n_1 \le 3s_1 - 3r_1 + 2\gamma_0 - \tfrac{3}{4}.$$

The above service policy is work-conserving. Note that at time $n_1 + r_1$, there are $n_1 + 1$ exogenous arrivals of class 1 jobs. The time at which all $\gamma_0 + n_1 + 1$ class 1 jobs complete service at station 1 is $n_1 + r_1 + 1/4$, and and at station 2 is

$$\begin{aligned} t_1 &:= \max\{s_1 + \tfrac{2}{3}(\gamma_0 + n_1 + 1), n_1 + r_1 + \tfrac{11}{12}\} \\ &= s_1 + \tfrac{2}{3}(\gamma_0 + n_1 + 1), \end{aligned}$$

where the last equality follows from our choice of n_1.

At time t_1,

$$Q(t_1) = (1[t_1 \ge n_1 + r_1 + 1], 0, \gamma_1, 0),$$

where $\gamma_1 = \lfloor t_1 - r_3 \rfloor + 1$. By our choice of n_1, we have

$$t_1 < n_1 + r_1 + 1 + \tfrac{1}{4} \le 2\gamma_0 + 4, \tag{A.15}$$

$$\gamma_1 \ge t_1 \ge n_1 > 2\gamma_0 - 5. \tag{A.16}$$

Case (ii): $s_1 > s_3 > 0$ *and* $r_3 + 1/4 > s_3 + 2/3$ (*implying* $r_3 > 1/4 \geq s_3$). In this case, station 2 first completes serving a class 3 job at s_3, and this job proceeds to station 1. At this moment, station 1 switches to serving this class 4 job, which completes service at station 1 before another class 3 job completes service at station 2. Thus, at time $s_3 + 2/3$,

$$Q(s_3 + \tfrac{2}{3}) = (\gamma_0 + 1[r_1 \leq s_3 + \tfrac{2}{3}] + 1[r_1 + 1 \leq s_3 + \tfrac{2}{3}], 0, 1[r_3 \leq s_3], 0);$$

this reduces to Case (i). Applying the result of Case (i), we have $t_1 \leq 2\gamma_0 + 10$ and $\gamma_1 \geq 2\gamma_0 - 5$, with t_1 and γ_1 similarly defined.

Case (iii): $s_1 > s_3 > 0$ *and* $\max\{r_3, s_3\} + 1/4 \leq s_3 + 2/3$. Pick n_0 such that

$$\tfrac{1}{4} + 3s_3 - 3r_3 < n_0 \leq 3s_3 - 3r_3 + \tfrac{5}{4}.$$

At stations 1 and 2, class 4 and class 3 jobs have preemptive priorities over class 1 and class 2 jobs, respectively. This continues until $n_0 + 2$ class 4 jobs complete service at station 1. First note that at time $n_0 + r_3$, there are $n_0 + 1$ arrivals of class 3 jobs. Also note that there is a class 3 job initially in the network; hence, the time at which all $n_0 + 2$ jobs complete service at station 2 is

$$\begin{aligned} t_0' &:= \max\{s_3 + \tfrac{2}{3}(n_0 + 2), n_0 + r_3 + \tfrac{11}{12}\} \\ &= s_3 + \tfrac{2}{3}(n_0 + 2) \leq 4.25, \end{aligned}$$

where we used the definition of n_0 in the last equality and inequality. At time t_0',

$$Q(t_0') = \left(\gamma_0 + 1 + \sum_n 1[n + r_1 \leq t_0' < n + r_1 + 1], 0, 1[t_0 \geq n_0 + r_3 + 1], 0\right).$$

Now as in Case (ii), we can use the result of Case (i) to obtain that

$$t_1 \leq 2\gamma_0 + 21 \quad \text{and} \quad \gamma_1 \geq 2\gamma_0 - 3. \qquad \square$$

Acknowledgments. This research started in early 1988 when I visited Avi Mandelbaum at Technion. Part of the results here are from Chen and Mandelbaum (1991b). I would like to thank Avi for his help and many useful discussions. Jim Dai and Ward Whitt sent me the paper by Rybko and Stolyar (1992), and Jim shared with me a preliminary draft of Dai (1995); their help is greatly appreciated. Jim and a referee pointed out an error in the proof of Theorem 5.2, and a referee pointed out the current simpler proof for Theorem 3.3. I thank the referees for many helpful suggestions.

REFERENCES

ALTMAN, E., FOSS, S. G., RIEHL, E. and STIDHAM, S. (1994). Performance bounds and pathwise stability for generalized vacation and polling systems. Technical Report UNC/OR/TR-4-3, Dept. Operations Research, Univ. North Carolina.

BACCELLI, F. and FOSS, S. G. (1994). Stability of Jackson-type queueing networks. *Queueing Syst.* **17**.

BASKETT, F., CHANDY, K. M., MUNTZ, R. R. and PALACIOS, F. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22** 248–260.

BOROVKOV, A. A. (1986). Limit theorems for queueing networks. *Theory Probab. Appl.* **31** 413–427.

BRAMSON, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4** 414–431.

CHANG, C. S., THOMAS, J. A. and KIANG, S. H. (1993). On the stability of open networks: an unified approach by stochastic dominance. Preprint.

CHEN, H. (1988). A heterogeneous fluid model and its applications in multiclass queueing networks. Award-winning paper (Honorable mention), George Nicholson Student Paper Competition, Operations Research Society of America.

CHEN, H. and MANDELBAUM, A. (1991a). Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16** 408–446.

CHEN, H. and MANDELBAUM, A. (1991b). Open heterogeneous fluid networks, with applications to multiclass queues. Preprint.

CHEN, H. and MANDELBAUM, A. (1994). Hierarchical modelling of stochastic networks, part I: fluid models. In *Stochastic Modeling and Analysis of Manufacturing Systems* (D. D. Yao, ed.) 47–105.

CHEN, H. and YAO, D. (1993). Dynamic scheduling of a multi-class fluid network. *Oper. Res.* **41** 1104–1115.

CHEN, H. and ZHANG, H. (1994). Stability of multiclass queueing networks under FIFO discipline. Unpublished manuscript.

COTTLE, R. W., HABETLER, G. J. and LEMKE, C. E. (1970). On classes of copositive matrices. *Linear Algebra Appl.* **3** 295–310.

DAI, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

DAI, J. G. and WANG, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems Theory Appl.* **13** 41–46.

DUPUIS, P. and WILLIAMS, R. J. (1994). Lyapunov functions for semimartingale reflected Brownian motions. *Ann. Probab.* **22** 680–702.

GLYNN, P. (1990). Diffusion approximations. In *Handbooks in Operations Research and Management Science, 2: Stochastic Models* (D. P. Heyman and M. J. Sobel, eds.) 145–198. North-Holland, Amsterdam.

HARRISON, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Their Applications* (W. Fleming and P. L. Lions, eds.) 147–186. Springer, Berlin.

JACKSON, J. R. (1963). Jobshop-like queueing systems. *Management Sci.* **10** 131–142.

JOHNSON, D. P. (1983). Diffusion approximations for optimal filtering of jump processes and for queueing networks. Ph.D dissertation, Univ. Wisconsin.

KELLY, F. P. (1982). Networks of quasireversible nodes. In *Applied Probability and Computer Science: The Interface* (R. L. Disney and T. J. Ott, eds.) **1** 3–29. Birkhäuser, Boston.

KLEINROCK, L. (1976). *Queueing Systems II: Computer Applications*. Wiley, New York.

KOSTEN, L. (1986). Liquid models for a type of information buffer problems. *Delft Progress Report* **11** 71–86.

KUMAR, P. R. and MEYN, S. P. (1993). Stability of queueing networks and scheduling policies. Preprint.

KUMAR, P. R. and SEIDMAN, T. I. (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control* **35** 289–298.

MEYN, S. P. and DOWN, D. (1994). Stability of generalized Jackson networks. *Ann. Appl. Probab.* **4** 124–148.

MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes II: continuous time processes and sampled chains. *Adv. in Appl. Probab.* **25** 518–548.

MITRA, D. (1986). Stochastic theory of a fluid model of multiple failure-susceptible producers and consumers coupled by a buffer. *Adv. in Appl. Probab.* **20** 646–676.

NEWELL, G. F. (1982). *Applications of Queueing Theory*. Chapman and Hall, London.

PETERSON, W. P. (1991). Diffusion approximations for networks of queues with multiple customer types. *Math. Oper. Res.* **16** 90–118.

RYBKO, A. N. and STOLYAR, A. L. (1992). Ergodicity of stochastic processes describing the operations of open queueing networks. *Problemy Peredachi Informatsii* **28** 2–26.

SEIDMAN, T. I. (1993). "First come first serve" is unstable. Preprint.

SIGMAN, K. (1990). The stability of open queueing networks. *Stochastic Process Appl.* **35** 11–25.

VANDERGRAFT, J. M. (1983). A fluid flow model of networks of queues. *Management Sci.* **29** 1198–1208.

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

2053 MAIN MALL
COMMERCE AND BUSINESS ADMINISTRATION
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, BC V6T 1Z2
CANADA