The Annals of Applied Probability
1996, Vol. 6, No. 3, 751–757

A FLUID LIMIT MODEL CRITERION FOR INSTABILITY OF MULTICLASS QUEUEING NETWORKS[1]

BY J. G. DAI

Georgia Institute of Technology

This paper studies the instability of multiclass queueing networks. We prove that if a fluid limit model of the queueing network is weakly unstable, then the queueing network is unstable in the sense that the total number of customers in the queueing network diverges to infinity with probability 1 as time $t \to \infty$. Our result provides a converse to a recent result of Dai which states that a queueing network is positive Harris recurrent if a corresponding fluid limit model is stable. Examples are provided to illustrate the usage of the result.

1. Introduction. This paper studies the transience of multiclass queueing networks. We prove that if a fluid limit model of the queueing network is weakly unstable, then the total number of customers in the queueing network diverges to infinity with probability 1 as time $t \to \infty$. The fluid limit model is weakly unstable if for almost every sample path there exists a time $\delta > 0$ such that the total volume of each fluid limit associated with the sample path is nonzero at δ. Our paper provides a converse to a recent result of Dai [2] which states that a queueing network is positive Harris recurrent if a corresponding fluid limit model eventually empties out regardless of the initial configuration.

The fluid limits in this paper have initial volume zero. They are obtained through a limiting procedure with the initial state of the queueing network *fixed* (to be empty, for example). Therefore, the fluid limits here are slightly different from the ones in Dai [2], where fluid limits were obtained through a limiting procedure when the initial state of the network goes to infinity. Meyn [6] obtained a similar converse result. The result in this paper is attractive because (a) our assumptions on the queueing network are much weaker than those imposed in [6], (b) we need only to check the instability of fluid limits for one initial condition, namely an empty system, instead of *every* initial condition and (c) there is no need to check the rate of divergence of a fluid limit. In Section 2, we introduce the multiclass queueing network with some minimal set of assumptions. The main theorem is proved in Section 3. To facilitate the checking of the weak instability of the fluid limit model, we show in Section 4 that any fluid limit is a fluid solution to a set of equations. Examples are provided in Section 5 to illustrate the usage of our theorem.

Received July 1995; revised February 1996.

[1]Research supported in part by NSF Grant DMI-94-57336 and matching funds from CACI Products Co. and AutoSimulations, Inc.

AMS 1991 *subject classifications.* Primary 60K25, 90B22; secondary 60K20, 90B35.

Key words and phrases. Multiclass queueing networks, instability, transience, Harris positive recurrent, fluid approximation, fluid model.

2. The multiclass queueing network. The queueing network under discussion has d single server stations and K customer classes. Each class k has its own exogenous arrival process $E_k = \{E_k(t),\ t \geq 0\}$, where $E_k(t)$ is the cumulative number of exogenous arrivals to class k by time t. We allow arrival processes for some classes to be null. Class k customers require service at station $\sigma(k)$ with service process $S_k = \{S_k(t),\ t \geq 0\}$, where $S_k(t)$ is the cumulative number of service completions for class k customers if class k receives t units of *service* from server $\sigma(k)$. Let $\Phi_l^k(n)$ be the number of transitions to class l among the first n class k service completions. We call $\Phi^k = \{\Phi^k(n),\ n \geq 1\}$ the routing process for class k.

When more than one class is served at a station, the server may have to choose the next customer to work on each time the network changes its state. A rule dictating such choices is called a queueing discipline or a dispatching rule. In this paper we allow queueing disciplines to be virtually arbitrary. [See condition (2.2) and the related discussions below.] Throughout this paper, we assume that there exist $\alpha_k \geq 0$, $\mu_k > 0$ and $P_{kl} \geq 0$ such that as $t \to \infty$ and $n \to \infty$, with probability 1,

$$\frac{E_k(t)}{t} \to \alpha_k \quad \text{for } k = 1, \ldots, K, \tag{2.1}$$

$$\frac{S_k(t)}{t} \to \mu_k \quad \text{for } k = 1, \ldots, K, \tag{2.2}$$

$$\frac{\Phi_l^k(n)}{n} \to P_{kl} \quad \text{for } k,\ l = 1, \ldots, K. \tag{2.3}$$

We further assume that $P = (P_{kl})$ is substochastic and has spectral radius less than 1, which is equivalent to the fact that $(I - P)$ is invertible. Assumption (2.1) is satisfied if interarrival times to class k are iid with mean $1/\alpha_k$. Assumption (2.2) is satisfied if class k service times are iid and at most κ class k customers can receive outstanding partial services, where κ is some constant. In general, the validity of assumption (2.2) depends on the queueing discipline used. Assumption (2.3) is satisfied if Markovian routing among classes is used. Note that it is satisfied even if routing is not Markovian. Consider, for example, the following routing procedure for class k: every odd number of class k service completion is sent to class l and every even number of class k service completion is sent to class l' for $l \neq l'$. In this case, assumption (2.3) is satisfied with $P_{kl} = P_{kl'} = 1/2$.

3. The main result. Assume that the network is initially empty. Let $Q(t) = (Q_1(t), \ldots, Q_K(t))'$, where $Q_k(t)$ is the number of class k jobs at time t and the prime denotes transpose. Let $T(t) = (T_1(t), \ldots, T_K(t))'$, where $T_k(t)$ is the cumulative amount of time that server $\sigma(k)$ has served class k customers by time t. Of course, $T(t)$ heavily depends on the service discipline used. Let ω be a fixed sample path such that (2.1)–(2.3) hold. In the following, we write down the explicit dependence on ω for some processes. For a sequence of functions $\{f_n(\cdot)\}$ on $[0, \infty)$, $f_n(t)$ is said to converge to $f(t)$ uniformly on

compact sets (u.o.c.) if for each $u > 0$,

$$\sup_{0 \le t \le u} |f_n(t) - f(t)| \to 0 \quad \text{as } n \to \infty.$$

For each $k = 1, \ldots, K$ and $t > s \ge 0$,

$$T_k(t, \omega) - T_k(s, \omega) \le t - s.$$

Therefore the family

$$\{T_k(r\cdot, \omega)/r,\ r \ge 1\}$$

is precompact under the u.o.c. topology. That is, for each sequence $\{r_n\}$ with $r_n \to \infty$ as $n \to \infty$, there is a subsequence $\{r_{n_j}\} \subset \{r_n\}$ such that $T_k(r_{n_j}t)/r_{n_j}$ converges to a limit $\bar{T}_k(t)$ u.o.c. as $j \to \infty$; see Royden [7].

PROPOSITION 3.1. *Suppose that $r_n \to \infty$ and $T(r_n t, \omega)/r_n$ converges to $\bar{T}(t)$ u.o.c. as $n \to \infty$. Then $Q(r_n t, \omega)/r_n$ converges to $\bar{Q}(t)$ u.o.c. as $n \to \infty$, where*

$$\bar{Q}(t) = \alpha t + (P' - I)\Delta \bar{T}(t), \tag{3.1}$$

and $\Delta = \operatorname{diag}(\mu_1, \ldots, \mu_K)$.

PROOF. The proposition readily follows from assumptions (2.1)–(2.3), Lemma 4.1 of Dai [2] and

$$Q(t) = E(t) + \sum_{k=1}^{K} \Phi^k(S_k(T_k(t))) - S(T(t)), \quad t \ge 0,$$

where $S(T(t)) = (S_1(T_1(t)), \ldots, S_K(T_K(t)))'$. □

The process $\bar{Q}(\cdot)$ in (3.1) is called a fluid limit starting from 0 for the sample path ω. Let $\mathscr{Q}_\omega$ be the set of fluid limits $\bar{Q}(\cdot)$. We call the collection of all fluid limits for all ω *the fluid limit model*. Notice that fluid limits defined here necessarily start from 0 because the initial state of the queueing network is fixed to be empty. However, in Dai [2] the fluid limits start from total volume 1 because the fluid limits there were obtained by letting the magnitude of the initial states go to infinity. Inspired by a definition of Chen [1] for weak stability of a fluid model, we give the following definition.

DEFINITION 3.1. The fluid limit model is *weakly unstable* if for each ω satisfying (2.1)–(2.3) there exists $\delta = \delta(\omega) > 0$ such that $\bar{Q}(\delta) \ne 0$ for each $\bar{Q}(\cdot) \in \mathscr{Q}_\omega$.

THEOREM 3.2. *If the fluid limit model is weakly unstable, the queueing network is unstable in the sense that, with probability* 1,

$$|Q(t)| \to \infty \quad \text{as } t \to \infty,$$

where, for a vector $x \in \mathbb{R}^K$, $|x| = \sum |x_k|$.

PROOF. Let ω be fixed so that (2.1)–(2.3) are satisfied. If the fluid limit is weakly unstable, there is a $\delta = \delta(\omega) > 0$ such that for each $\bar{Q} \in \mathcal{Q}_\omega$, $|\bar{Q}(\delta)| > 0$. We claim that

$$\liminf_{r\to\infty} |Q(r\delta, \omega)/r| > 0, \tag{3.2}$$

which, of course, implies that $\lim_{t\to\infty} |Q(t, \omega)| = \infty$. To prove (3.2), suppose that

$$\liminf_{r\to\infty} |Q(r\delta, \omega)/r| = 0.$$

There exists a sequence $\{r_n\}$ with $r_n \to \infty$ as $n \to \infty$ such that

$$|Q(r_n\delta, \omega)/r_n| \to 0. \tag{3.3}$$

Because $\{Q(r\cdot, \omega)/r, r \geq 1\}$ is precompact, there exists a subsequence $\{r_{n_m}\} \subset \{r_n\}$ such that $Q(r_{n_m}\cdot, \omega)/r_{n_m}$ converges u.o.c. to a limit $\bar{Q}(\cdot) \in \mathcal{Q}_\omega$. Hence,

$$|Q(r_{n_m}\delta, \omega)/r_{n_m}| \to |\bar{Q}(\delta)| > 0,$$

which contradicts (3.3). □

REMARK. Stolyar [8] proved that the set of fluid limits $\mathcal{Q}_\omega$ is precompact. Therefore, the weak instability implies the strong condition

$$\inf_{\bar{Q}\in\mathcal{Q}_\omega} |\bar{Q}(\delta)| > 0, \tag{3.4}$$

where δ is as in Definition 3.1.

COROLLARY 3.3. *Assume that the fluid limit model is weakly unstable. For each ω satisfying* (2.1)–(2.3) *and each $\varepsilon > 0$ with*

$$\varepsilon < \inf_{\bar{Q}\in\mathcal{Q}_\omega} |\bar{Q}(\delta)|,$$

there exists an $M(\omega) > 0$ such that

$$|Q(t, \omega)| \geq \frac{\varepsilon}{\delta(\omega)} t \quad \text{for } t \geq M(\omega).$$

4. Fluid solution. Often it is difficult to work with fluid limits directly. In this section, we introduce the notion of fluid solutions to a fluid model. Let us restrict the rest of the paper to non-idling queueing disciplines. That is, whenever there is a customer at a station, the server at the station should work. For a function f: $[0, \infty) \to \mathbb{R}^K$, $t > 0$ is a regular point of f if f is differentiable at t. We use $\dot{f}(t)$ to denote the derivative of f at time t.

PROPOSITION 4.1. *Any fluid limit* $\bar{Q}(\cdot)$ *together with the corresponding limit* $\bar{T}(\cdot)$ *must satisfy the following equations*:

$$\bar{Q}(t) = \alpha t + (P' - I)\Delta\bar{T}(t); \tag{4.1}$$

$$\bar{Q}(t) \geq 0; \tag{4.2}$$

$$\bar{T}(0) = 0 \text{ and each component of } \bar{T} \text{ is nondecreasing}; \tag{4.3}$$

$$\text{for each station } i,\ \bar{I}_i(\cdot) \text{ is nondecreasing where for each station } i,\ \bar{I}_i(t) \equiv t - \textstyle\sum_{k:\,\sigma(k)=i} \bar{T}_k(t); \tag{4.4}$$

$$\text{if } \sum_{k:\,\sigma(k)=i} \bar{Q}_k(t) > 0 \text{ at a regular point } t \text{ of } \bar{T}(\cdot), \text{ then } \dot{\bar{I}}_i(t) = 0; \tag{4.5}$$

$$\text{some additional equations for } (\bar{Q}(\cdot), \bar{T}(\cdot)) \text{ that are specific to the queueing discipline.} \tag{4.6}$$

PROOF. Equation (4.1) follows from (3.1). Equations (4.2)–(4.4) follow trivially from the queueing network counterparts. Equation (4.5) follows from (4.21) of Dai [2]. □

Equations (4.1)–(4.6) define a *fluid model*. Any solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ to equations (4.1)–(4.6) is a fluid solution to the fluid model. Proposition 4.1 shows that any fluid limit is a fluid solution.

DEFINITION 4.1. The fluid model is weakly unstable if there exists $\delta > 0$ such that for each fluid solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ starting from 0, $\bar{Q}(\delta) \neq 0$.

As a corollary to Proposition 4.1 and Theorem 3.2, we have

THEOREM 4.2. *If the fluid model is weakly unstable, the queueing network is unstable in the sense that, with probability* 1,

$$|Q(t)| \to \infty \quad \text{as } t \to \infty.$$

5. Examples.

5.1. *A general instability criterion.* Let $\lambda = (I - P')^{-1}\alpha$ and for each station i, define the traffic intensity at the station to be

$$\rho_i = \sum_{k:\,\sigma(k)=i} \lambda_k/\mu_k.$$

PROPOSITION 5.1. *If* $\rho_i > 1$ *for some station* i, *the fluid model is weakly unstable. Therefore the queueing network is unstable.*

PROOF. For any fluid solution $\bar{Q}(\cdot)$, let $Z(t) = (I - P')^{-1}\bar{Q}(t)$. We have

$$\begin{aligned}\sum_{k:\sigma(k)=i} \frac{1}{\mu_k} Z_k(t) &= \rho_i t - \sum_{k:\sigma(k)=i} \bar{T}_k(t) \\ &\geq \rho_i t - t = (\rho_i - 1)t \\ &> 0\end{aligned}$$

for $t > 0$. Thus, the fluid model is weakly unstable. By Theorem 4.2, the queueing network is unstable. □

5.2. *A seven-class reentrant line.* Consider the seven-class reentrant line pictured in Figure 1. Assume that the arrival rate to class 1 is $\alpha_1 = 1$. Let $m_k = 1/\mu_k$ be the average service time for a class k customer. For this network, we use a preempt–resume static buffer priority discipline. For a definition see, for example, Section 3.4 of Gross and Harris [5]. The discipline used gives priorities (1, 7, 5, 3) at station 1 with class 1 customers having the highest priority and class 3 customers having the lowest priority, and gives priorities (2, 4, 6) at station 2.

PROPOSITION 5.2. *If*

$$\frac{m_4}{1 - m_2} + \frac{m_7}{1 - m_1} > 1, \tag{5.1}$$

then the fluid limit model is weakly unstable, and hence the queueing network is unstable.

PROOF. Following the argument in Lemma 2.2 of Dumas [4], it can be verified that if the network starts empty, class 4 and class 7 can never be served simultaneously. Therefore,

$$T_4(t) + T_7(t) \leq t \quad \text{for all } t \geq 0.$$

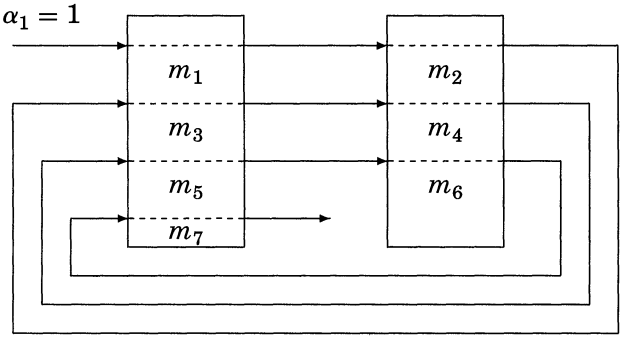

FIG. 1. *A seven-class reentrant line.*

Let $(\bar{Q}(\cdot), \bar{T}(\cdot))$ be a fluid limit. We have

$$\bar{T}_4(t) + \bar{T}_7(t) \le t \quad \text{for all } t \ge 0. \tag{5.2}$$

Hence the corresponding fluid model has an additional equation (5.2). Unfortunately, (5.2) is not sufficient for establishing the weak instability. Because class 1 has the highest priority at station 1 and class 2 has the highest priority at station 2, it was proved in Dai and VandeVate [3] that (5.2) could be further strengthened as follows:

$$\frac{\bar{T}_4(t)}{1-m_2} + \frac{\bar{T}_7(t)}{1-m_1} \le t \quad \text{for all } t \ge 0.$$

Let $Z(t) = (I - P')^{-1}\bar{Q}(t)$. It is easy to check that

$$\begin{aligned}
\frac{m_4 Z_4(t)}{1-m_2} + \frac{m_7 Z_7(t)}{1-m_1} &= \left(\frac{m_4}{1-m_2} + \frac{m_7}{1-m_1}\right)t - \left(\frac{\bar{T}_4(t)}{1-m_2} + \frac{\bar{T}_7(t)}{1-m_1}\right) \\
&\ge \left(\frac{m_4}{1-m_2} + \frac{m_7}{1-m_1}\right)t - t \\
&= \left(\frac{m_4}{1-m_2} + \frac{m_7}{1-m_1} - 1\right)t > 0
\end{aligned}$$

for $t \ge 0$. Thus the fluid limit model is weakly unstable and hence the queueing network is unstable. □

REFERENCES

[1] CHEN, H. (1995). Fluid approximations and stability of multiclass queueing networks. I. Work-conserving disciplines. *Ann. Appl. Probab.* **5** 637–665.

[2] 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.

[3] DAI, J. G. and VANDEVATE, J. (1996). Virtual stations and the capacity of two-station queueing networks. Preprint.

[4] DUMAS, V. (1995). A multiclass network with non-linear, non-convex, non-monotonic stability conditions. Preprint.

[5] GROSS, D. and HARRIS, C. M. (1985). *Fundamentals of Queueing Theory*. Wiley, New York.

[6] MEYN, S. P. (1995). Transience of multiclass queueing networks via fluid limit models. *Ann. Appl. Probab.* **5** 946–957.

[7] ROYDEN, H. L. (1988). *Real Analysis*, 3rd ed. Macmillan, New York.

[8] STOLYAR, A. (1994). On the stability of multiclass queueing networks. In *Proceedings of the Second International Conference on Telecommunication Systems—Modeling and Analysis* 23–35, Nashville, TN.

SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING
AND SCHOOL OF MATHEMATICS
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332-0205
E-MAIL: dai@isye.gatech.edu