Global Stability of Two-Station Queueing Networks

J. G. Dai, John H. Vande Vate

ABSTRACT This paper summarizes results of Dai and Vande Vate [15, 14] characterizing explicitly, in terms of the mean service times and average arrival rates, the global pathwise stability region of two-station open multiclass queueing networks with very general arrival and service processes. The conditions for pathwise global stability arise from two intuitively appealing phenomena: virtual stations and push starts. These phenomena shed light on the sources of bottlenecks in complicated queueing networks like those that arise in wafer fabrication facilities. We show that a two-station open multi-class queueing network is globally pathwise stable if and only if the corresponding fluid model is globally weakly stable. We further show that a two-station fluid model is globally (strongly) stable if and only if the average service times are in the interior of the global weak stability region. As a consequence, under stronger distributional assumptions on the arrival and service processes, the queueing network is globally stable in a stronger sense when the mean service times are in the interior of the global pathwise stability region. Namely, the underlying state process of the queueing network is positive Harris recurrent.

1 Introduction

Queueing networks offer an appealing method for modeling complex manufacturing processes. Unfortunately, they are themselves generally too complex for successful analysis. For example, the primary tool for evaluating the performance of a given dispatching rule is simulation. In fact, we generally resort to simulation even to determine whether a queueing network is stable under a given dispatching rule.

Even very simple queueing networks exhibit surprising and often counterintuitive behavior. In a surprising series of examples, Kumar and Seidman [28], Lu and Kumar [29] and Rybko and Stolyar [33] demonstrated queueing networks that cannot keep up with customer arrival rates under certain non-idling queueing disciplines even though the traffic intensity at each station is less than one. Bramson [4, 5] and Seidman [34] independently provided examples demonstrating the same remarkable behavior with the popular first-in–first-out (FIFO) queueing discipline.

These examples have inspired a number of investigations into the ca-

pacity of general queueing networks. For example, Kumar and Meyn [27], Dai [11], Chen [8], Down and Meyn [17], Chen and Zhang [10], Foss and Rybko [23], Bramson [6, 7] all proposed various sufficient conditions, which if satisfied, ensure that the network has adequate capacity.

In some specific instances, researchers have been able to characterize exact conditions. For example, Botvich and Zamyatin [3] determined exactly those average service rates able to keep up with a given average rate of customer arrivals in a specific two-station queueing network following a particular buffer priority discipline. Dumas [18] accomplished the same for a specific three-station queueing network. In both cases, the capacity of the network was less than that of the busiest server.

While these researchers have studied networks following specific queueing disciplines, we are interested in the capacity of queueing networks that may follow any non-idling queueing discipline, that is, any queueing discipline that requires servers to work whenever there is work for them to do. We summarize results in Dai and Vande Vate [15, 14] describing conditions under which the network is *globally stable*, i.e., has sufficient capacity regardless of the queueing discipline employed as long as it is non-idling. A globally stable queueing network will always have sufficient capacity to meet given customer arrival rates no matter what queueing disciplines the servers follow as long as they keep busy whenever there are customers available to serve. On the other hand, a queueing network that is not globally stable will not have adequate capacity to meet demand under some non-idling queueing disciplines.

These results show that among all non-idling queueing disciplines, the static buffer priority disciplines alone define the capacity of two-station queueing networks. In other words, one corollary of our results is the conclusion that static buffer priority queueing disciplines are extreme or "worst" in the class of all non-idling queueing disciplines for two-station queueing networks.

Two phenomena determine the capacity of two-station queueing systems: virtual stations and push starts. These two phenomena provide insight into the sources of bottlenecks in complicated networks like those in wafer fabrication facilities.

Virtual stations affect the global stability of queueing networks because, under some non-idling queueing disciplines, certain groups of buffers can never be served simultaneously even though they are served at different stations. Thus, just as at stations, the traffic intensities at these groups of buffers called *virtual stations* must be at most one.

Push starts affect the global stability of queueing networks because of their influence on virtual stations. In particular, if we give highest priority to the first few buffers customers visit, they pass through to the rest of the network at the same rate they arrive, but the servers have that much less capacity to dedicate to the rest of network. This does not influence the traffic intensities at the stations, but it dramatically affects the traffic

intensities at virtual stations.

Together, these two intuitively appealing phenomena characterize when a two-station queueing network is globally pathwise stable. We show that the corresponding two-station *fluid networks* are globally weakly stable under the same conditions. Thus, a two-station, open multi-class queueing network is globally pathwise stable if and only if the corresponding fluid model is globally weakly stable.

When the arrival and service processes satisfy stronger distributional assumptions, it is possible to establish stronger forms of stability like *Harris recurrence*. This paper also proves that under a non-idling queueing discipline if the mean service times are in the *interior* of the global pathwise stability region, the underlying state process is *positive* Harris recurrent and so the queueing network is stable in a stronger sense. Conversely, when the mean service times are outside the global pathwise stability region, there is a buffer priority discipline under which the total number of customers in the network diverges to infinity. Under this stronger notion of stability, we cannot conclude that the conditions are exact, however, because we do not know whether the network will be stable when the the mean service times are on the boundary of the global pathwise stability region. We conjecture that the network is not positive recurrent in this case.

Although these results provide compelling evidence in support of the belief that fluid models accurately describe the capacity of queueing networks, we do not prove the strong relationships between the two models directly. Rather, we observe the phenomena determining the capacity of the two-station fluid model and argue that they also determine the capacity of the two-station queueing network. This indirect argument highlights a more direct connection between the two models: In each case, the same queueing discipline gives rise to the same constraint on the global stability region.

Our proof of the necessity of our conditions for the global stability of the queueing network is direct. We show that if the system violates one of our conditions, the number of customers in the system goes to infinity.

Our proof of the sufficiency of our conditions to ensure the global stability of the queueing network relies on the fact that the queueing network is globally stable if the fluid model is. We determine when a two-station fluid network is stable by determining when there is a piecewise linear Lyapunov function for it. We formulate the problem of determining the coefficients of the Lyapunov function as a linear programming problem, which has unbounded objective values if and only if the coefficients and hence the Lyapunov function exist. Our linear program arises directly from the piecewise linear Lyapunov function introduced in Dai [12], which generalizes that of Botvich and Zamyatin [3] and is simpler than that independently formulated by Down and Meyn [17].

We transform our linear program into a parametric network flow problem in an acyclic network. The fluid network is globally stable if there is a

value of the parameter for which the minimum flow in this network is sufficiently small. Thus, sufficient conditions for global stability arise from the constraints imposed on the upper ideals of the partial order defined by the acyclic network in order to ensure there is a sufficiently small flow.

Recently, Bertsimas, Gamarnik and Tsitsiklis [2] showed that a two-station fluid network is globally stable if and only if a certain linear program has bounded objective value. We extend the results of Bertsimas et al. by stating explicitly in terms of the service times, necessary and sufficient conditions for a two-station fluid network to be stable under all non-idling dispatching policies. The explicit description of necessary and sufficient conditions for the stability of two-station fluid networks provides a number of corollaries not immediately available from the linear programming characterization of Bertsimas et al. [2]. Most important among these is a complete understanding of how virtual stations arise in two-station fluid networks. In addition, our conditions demonstrate that the global stable region of a two-station fluid network is *monotone*, i.e., reducing service times maintains global stability. This is not the case for stability with respect to a given dispatching policy. It is possible for a dispatching policy to be stable for a given fluid network, but instable when the service times are reduced. For fluid networks with more than two stations even the global stable region need not be monotone.

In Section 2 we introduce our two-station queueing model and define how to measure its capacity. In Section 3 we describe the corresponding fluid model. Section 4 states our main results, which we interpret and refine in terms of virtual stations and push starts in Section 5. This section proves the necessity of our conditions for global pathwise stability. In Section 7 we outline the proof that these conditions are also sufficient. Finally, in Section 6, we consider a stronger notion of stability in the queueing network, positive Harris recurrence, under additional assumptions on the arrival and service processes.

2 The Queueing Model

We consider a queueing network with two single-server stations, denoted A and B. The system serves a set $I = \{1, \ldots, n\}$ of n different types of customers. Type i customers arrive according to the exogenous arrival process $S_0^i = \{S_0^i(t), t \geq 0\}$, where $S_0^i(t)$ is the cumulative number of exogenous arrivals by time t.

Different types of customers may follow different routes, but each type i customer follows the same deterministic route visiting first one station and then the other a number of times before exiting the system. We number the visits by type i customers consecutively from 1 to c_i and let A_i denote those to station A and B_i, those to station B.

Following Kelly [26], we refer to type i customers during visit k (either waiting or being served) as *class* (i,k) customers. We assume the system can accommodate an unlimited number of class (i,k) customers and treat these customers as though they resided in a dedicated buffer with infinite capacity. Class (i,k) customers receive service at station $\sigma(i,k)$ according to the service process $S_k^i = \{S_k^i(t), t \geq 0\}$, where $S_k^i(t)$ is the cumulative number of service completions for class (i,k) customers if the server dedicates t units of *service* to the class.

The arrival processes and service processes can be random. We assume that they are defined on a probability space and satisfy a *strong law of large numbers*. That is, we assume that for almost every sample path, as $t \to \infty$,

$$S_0^i(t)/t \to \lambda_i = 1, \qquad \text{for each } i \in I, \text{ and} \tag{1.1}$$

$$S_k^i(t)/t \to \mu_k^i, \qquad \text{for each class } (i,k). \tag{1.2}$$

We interpret λ_i as the average arrival rate for type i customers and μ_k^i as the average service rate for class (i,k) customers. Since μ_k^i is the average service rate, we can interpret $m_k^i = 1/\mu_k^i$ as the average service time for class (i,k) customers. We assume that $\lambda_i = 1$ without loss of generality, since measuring the buffer levels for type i customers in units of λ_i and scaling the service processes accordingly (so that we scale the average service rate for class (i,k) customers to μ_k^i/λ_i and the average service time to $\lambda_i m_k^i$), normalizes the rates at which customers enter the system. This scaling amounts to nothing more than changing the units by which we measure the different types of customers.

We let $Q_k^i(t)$ denote the number of class (i,k) customers in the buffer (or being served) at time t and $T_k^i(t)$ the cumulative time in $[0,t]$ that server $\sigma(i,k)$ spends on class (i,k) customers. Note that $S_k^i(T_k^i(t))$ is the number of customers to complete class (i,k) service by time t and so:

$$Q_k^i(t) = Q_k^i(0) + S_{k-1}^i(T_{k-1}^i(t)) - S_k^i(T_k^i(t)) \tag{1.3}$$

for each class (i,k), where we model exogenous arrivals of type i customers by setting $T_0^i(t) = t$.

In general, a station serves many classes and so the server must decide which class and even which customer to serve next. A *queueing discipline* dictates which customer to work on each time the system changes state or experiences an *event*. Events occur when a customer arrives or a service is completed. Whitt [37] shows that simply changing how a system handles simultaneous events can dramatically affect its capacity. Thus, to reach any meaningful conclusions about the capacity of queueing networks, we must adopt some convention on simultaneous events. We assume that the queueing discipline responds to events one at a time, which is consistent with the way a single sequential processor would handle them.

We are primarily interested in preempt-resume, static buffer priority queueing disciplines as these alone determine the global stability of two-station queueing networks. A *preempt-resume* queueing discipline can interrupt service to one customer in order to serve another and later resume the interrupted service where it left off. *Static buffer priority queueing disciplines* simply stipulate that each station serve its different classes according to some fixed rank order and customers within a class are served on a first-come–first-served basis. Under a static buffer priority discipline, a server cannot work on a class unless there are no customers available in any higher priority class at the station.

After normalizing the average arrival rates, the traffic intensities at the stations are given by:

$$\rho_A = \sum_{i \in I} \sum_{k \in A_i} m_k^i \quad \text{and} \quad \rho_B = \sum_{i \in I} \sum_{k \in B_i} m_k^i. \tag{1.4}$$

They measure the nominal work imposed on the stations each unit of time. If the traffic intensity at some station exceeds 1, work for the station arrives faster than the server can complete it and so clearly the server and hence the system does not have sufficient capacity.

Even if each server individually has sufficient capacity, the system as a whole may not because the servers must interact: one station can only serve a customer after another station has finished. Thus, we extend the notion of the capacity of a server to define the capacity of the system. Just as we say a server has sufficient capacity if he can complete the work as quickly as it arrives, we say that the system has sufficient capacity if it can finish serving customers as quickly as they arrive. More formally, we let $D_i(t)$ denote the number of type i customers to complete their last service, namely class (i, c_i) service, by time t. Thus, $D_i(t) = S^i_{c_i}(T^i_{c_i}(t))$.

Definition 2.1 We say that the system has *sufficient capacity* or is *pathwise stable* if the long run average arrival and departure rates are equal, that is, if for almost every sample path,

$$\frac{D_i(t)}{t} \to \lambda_i \text{ as } t \to \infty \text{ for each type } i.$$

Thus, a queueing network is *globally pathwise stable* if, under each non-idling queueing discipline, $D_i(t)/t \to \lambda_i$ for each type i.

Pathwise stability, first introduced by El-Taha and Stidham [21] (see also El-Taha and Stidham [21] and Altman, Foss, Riehl and Stidham [1]), is very weak by conventional standards. The well-known $M/M/1$ queueing system, for example, is pathwise stable if and if only the average arrival rate λ does not exceed the average service rate μ. When $\lambda = \mu$, however, the system is *null* recurrent, not positive recurrent. In particular, it does not possess an equilibrium and is often considered "unstable". In Section 6,

we introduce a stronger notion of stability—*positive Harris recurrence*. An $M/M/1$ queueing system with $\lambda = \mu$ is not stable under this stronger notion.

Having sufficient capacity is necessary, but not sufficient to ensure that for almost every sample path the number of customers in the system is bounded over time: Even in an $M/M/1$ queueing system with $\rho := \lambda/\mu < 1$, the number of customers in the system is not bounded. Having sufficient capacity is enough to ensure the following properties of the long term behavior of the system. Note this result holds even when there are more than two stations.

Lemma 2.2 If an open multi-class queueing network has sufficient capacity, then for each class (i,k)

$$S_k^i(T_k^i(t))/t \;\to\; \lambda_i \quad \text{and} \tag{1.5}$$

$$T_k^i(t)/t \;\to\; \lambda_i m_k^i \tag{1.6}$$

as $t \to \infty$.

Proof. The definition of sufficient capacity ensures (1.5) for the last class for each type of customer. It is easy to show that if a class (i,k) satisfies (1.5) it also satisfies (1.6). Finally, a class feeding a class (i,k) satisfying (1.5) must itself satisfy this condition. For a more detailed proof, see Lemma 1.1 of Dai and Vande Vate [15]. □

For future reference, we define an *excursion* to be a block of consecutive visits to the same station. In the Lu-Kumar network of Figure 1 for example, there are two excursions at station A — each consisting of a single visit — and one excursion at station B consisting of visits 2 and 3.

We let E_i denote the set of excursions for type i customers and number these excursions consecutively from 1 to $|E_i|$. We partition E_i into E_A^i, the set of excursions at station A, and E_B^i, those at station B. Since an excursion at one station must be followed by an excursion at the other (unless it is the last excursion), one of these is the set of odd numbered excursions and the other is the set of even numbered excursions depending on where type i customers first enter the network.

We partition the visits of an excursion into the *last visit* and all the rest, which we call *first visits*. We let $\ell(i,e)$ denote the last visit and $f(i,e)$ the set of first visits in excursion e for type i customers. If an excursion consists of only one visit, that visit is the last visit and the excursion has no first visits. For example, in the Lu-Kumar network of Figure 1, $\ell(1,2) = 3$ and $f(1,2) = \{2\}$ while $\ell(1,1) = 1$ and $f(1,1) = \emptyset$.

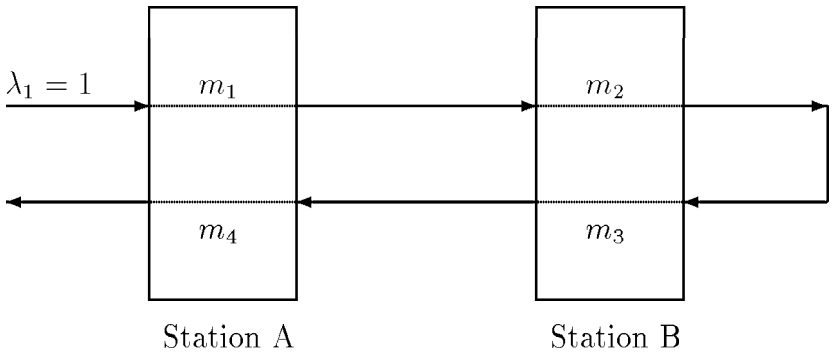

FIGURE 1. The Lu-Kumar Network.

3 The Fluid Model and Fluid Limits

Fluid models are continuous, deterministic approximations to discrete, stochastic queueing networks. The fluid model corresponding to our two-station queueing network replaces discrete customers arriving according to a random process with continuous fluids arriving at a constant rate. Type i fluid arrives at the constant rate λ_i and follows the same prescribed route as type i customers before exiting the system. As in the queueing model, we assume that $\lambda_i = 1$ without loss of generality.

We refer to type i fluids during visit k as *class* (i,k) *fluid* and we let $\overline{Q}_k^i(t)$ denote the volume of class (i,k) fluid in the buffer at time t. Server $\sigma(i,k)$ can process class (i,k) fluid at rate μ_k^i. This means that the server depletes class (i,k) fluid from the buffer at rate μ_k^i when he devotes all his efforts to serving that class. Equivalently, each unit of class (i,k) fluid requires service lasting $m_k^i = 1/\mu_k^i$ units of time.

We denote by $\overline{T}_k^i(t)$ the cumulative effort server $\sigma(i,k)$ has dedicated to class (i,k) up to time t. Since, $\overline{T}_k^i(t)$ is the cumulative time allocated to class (i,k) service, we refer to $\overline{T}(\cdot) = (\overline{T}_k^i(\cdot))$ as the *cumulative allocation*. Note that $\overline{T}_k^i(\cdot)$ is a non-decreasing function of time and $\mu_k^i \overline{T}_k^i(t)$ is the cumulative volume of fluid to complete class (i,k) service.

We let $I_A(t)$ denote the cumulative time station A is idle and $I_B(t)$, the cumulative time station B is idle up to time t. Note that $I_A(\cdot)$ and $I_B(\cdot)$ are non-decreasing functions of time as well.

Finally, we let $V_A(t) = \sum_{i\in I}\sum_{k\in A_i} \overline{Q}_k^i(t)$ denote the volume of fluid in the buffers at station A and $V_B(t) = \sum_{i\in I}\sum_{k\in B_i} \overline{Q}_k^i(t)$, the volume of fluid in the buffers at station B. The non-idling conditions can be expressed via the requirement that when $V_A(t) > 0$, i.e., when there is work at station A, the time derivative $\dot{I}_A(t) = 0$ and so station A is not accumulating

idle time. Note that since $I_A(\cdot)$ need not be everywhere differentiable, we only impose this condition on the *regular* points or points where $I_A(\cdot)$ is differentiable. Similarly, we express the non-idling condition at station B via the requirement that when $V_B(t) > 0$ and $I_B(\cdot)$ is differentiable at t, then $\dot{I}_B(t) = 0$.

Fluid models can also include other conditions representing the queueing discipline. As we are interested in the network under all non-idling queueing disciplines, however, the following equations define our fluid model:

$$\overline{Q}_k^i(t) = \overline{Q}_k^i(0) + \mu_{k-1}^i \overline{T}_{k-1}^i(t) - \mu_k^i \overline{T}_k^i(t) \text{ for each class } (i,k), \quad (1.7)$$

$$I_A(t) = t - \sum_{i \in I} \sum_{k \in A_i} \overline{T}_k^i(t), \quad (1.8)$$

$$I_B(t) = t - \sum_{i \in I} \sum_{k \in B_i} \overline{T}_k^i(t). \quad (1.9)$$

Note that we model exogenous arrivals of fluid type i by setting $\mu_0^i = \lambda_i = 1$ and $\overline{T}_0^i(t) = t$. In addition we require that:

$$\overline{Q}_k^i(t) \geq 0 \quad \text{for each class } (i,k) \quad (1.10)$$

$$\overline{T}_k^i(0) = 0 \quad \text{for each class } (i,k), \text{ and } \overline{T}_k^i(\cdot) \text{ is non-decreasing}, \quad (1.11)$$

$$I_A(0) = 0 \quad \text{and } I_A(\cdot) \text{ is non-decreasing}, \quad (1.12)$$

$$I_B(0) = 0 \quad \text{and } I_B(\cdot) \text{ is non-decreasing}, \quad (1.13)$$

$$\dot{I}_A(t) = 0 \quad \text{if } V_A(t) > 0 \text{ and } I_A(\cdot) \text{ is differentiable at } t, \text{ and} \quad (1.14)$$

$$\dot{I}_B(t) = 0 \quad \text{if } V_B(t) > 0 \text{ and } I_B(\cdot) \text{ is differentiable at } t. \quad (1.15)$$

A *fluid vector* is a vector $(\overline{Q}(\cdot), \overline{T}(\cdot))$, where $\overline{Q}(\cdot) = (\overline{Q}_k^i(\cdot))$ is a vector of buffer levels and $\overline{T}(\cdot) = (\overline{T}_k^i(\cdot))$ is a vector of allocations. A fluid vector $(\overline{Q}(\cdot), \overline{T}(\cdot))$ satisfying (1.7)–(1.15) is a *fluid solution*. The set of fluid solutions describes all possible trajectories of the fluid network under non-idling queueing disciplines.

The fluid network is said to be *weakly stable under non-idling queueing disciplines*, or simply *globally weakly stable*, if starting out empty, it remains empty, i.e., if every fluid solution to the system with

$$V_A(0) + V_B(0) = 0,$$

satisfies

$$V_A(t) + V_B(t) = 0,$$

for all $t \geq 0$.

The fluid network is said to be *(strongly) stable under non-idling queueing disciplines*, or simply *globally stable*, if there is some finite time $\tau > 0$

beyond which every fluid solution $(\overline{Q}(\cdot), \overline{T}(\cdot))$ that begins with one unit of fluid in the system, i.e., with

$$V_A(0) + V_B(0) = 1,$$

will be empty for all $t \geq \tau$, i.e., will satisfy

$$V_A(t) + V_B(t) = 0,$$

for all $t \geq \tau$.

In the remainder of this section, we discuss the relationships between fluid solutions and *fluid limits* or the limits of sample paths in a corresponding queueing network under a time and space scaling. Fluid limits provide a direct link between the discrete, stochastic queueing network and the continuous, deterministic fluid network; see, for example, Chen and Mandelbaum [9] and Dai [11]. To introduce fluid limits, we need to define a convergence notion in the path space. For each positive integer k, let $D^k[0,\infty)$ denote the set of functions from $[0,\infty)$ to $\mathbb{R}^k$ that are right continuous on $[0,\infty)$ and have left limits on $(0,\infty)$. Notice that for almost every sample path ω, the queue length process $\{Q(t,\omega), t \geq 0\}$ is an element in $D^k[0,\infty)$, where k is the total number of customer classes in the network. A sequence of functions $\{f_j\}$ in $D^k[0,\infty)$ is said to converge to $f \in D^k[0,\infty)$ uniformly on compact sets (u.o.c.) if for each $t > 0$,

$$\sup_{0 \leq s \leq t} |f_j(s) - f(s)| \to 0 \quad \text{as } j \to \infty,$$

where for a vector $x \in \mathbb{R}^k$, $|x|$ is the Euclidean norm of x.

For each sample path ω, each class (i,k) and each $t > s \geq 0$,

$$T_k^i(t,\omega) - T_k^i(s,\omega) \leq t - s.$$

Therefore the family

$$\{(T(r\,\cdot,\omega)/r, r \geq 1\}$$

is pre-compact under the u.o.c. topology. That is, for each sequence $\{r_j\}$ with $r_j \to \infty$ as $j \to \infty$, there is a subsequence $\{r_{j'}\} \subset \{r_j\}$ such that $T_k^i(r_{j'}t)/r_{j'}$ converges to a limit $\overline{T}_k^i(t)$ u.o.c. as $j' \to \infty$ for each class (i,k); see Royden [32]. It follows that $Q_k^i(r_{j'}t)/r_{j'}$ converges u.o.c. to a limit $\overline{Q}_k^i(t)$ for each class (i,k) as $j' \to \infty$. Any such limit $(\overline{Q}(\cdot), \overline{T}(\cdot))$ is said to be a *fluid limit*. Each fluid limit is a fluid solution satisfying (1.7)-(1.15). The following proposition follows immediately from the proof of Theorem 4.1 of Chen [8].

Proposition 3.1 For a given queueing discipline, the queueing network is pathwise stable if and only if the fluid limit $(\overline{Q}(\cdot), \overline{T}(\cdot))$ is unique and given by $\overline{Q}_k^i(t) = 0$ and $\overline{T}_k^i(t) = m_k^i t$ for each class (i,k).

One immediate corollary is the following result, which first appeared in Chen [8].

Corollary 3.2 If the fluid model is globally weakly stable, the corresponding queueing network is globally pathwise stable.

Lemma 3.3 Let C be a set of classes and $w = (w_k^i)_{(i,k)\in C}$ weights such that for each fluid limit $(\overline{Q}(\cdot), \overline{T}(\cdot))$:

$$\sum_{(i,k)\in C} w_k^i \overline{T}_k^i(t) \leq t$$

for all $t \geq 0$. If

$$\sum_{(i,k)\in C} w_k^i m_k^i > 1,$$

then the total volume of fluid in the fluid network diverges to infinity as $t \to \infty$. Therefore, the queueing network is not globally weakly stable, and for almost every sample path, the total number of customers in the network diverges to infinity as time $t \to \infty$.

Proof. Let $(\overline{Q}(\cdot), \overline{T}(\cdot))$ be a fluid limit. Because it is also a fluid solution with initial volume zero, it follows from (1.7) that for each class (i,k) we have

$$\sum_{\ell=1}^{k} \overline{Q}_\ell^i(t) = t - \mu_k^i \overline{T}_k^i(t).$$

Therefore,

$$\begin{aligned}\sum_{(i,k)\in C} w_k^i m_k^i \left(\sum_{\ell=1}^{k} \overline{Q}_\ell^i(t)\right) &= \left(\sum_{(i,k)\in C} w_k^i m_k^i\right) t - \sum_{(i,k)\in C} w_k^i \overline{T}_k^i(t) \\ &> \left(\sum_{(i,k)\in C} w_k^i m_k^i - 1\right) t.\end{aligned}$$

Thus, the total volume of fluid in the fluid network diverges to infinity as $t \to \infty$ and by Proposition 3.1 the queueing network is not globally weakly stable. Furthermore, It follows from Dai [13, Theorem 3.2] that the total number of customers in the system goes to infinity as $t \to \infty$. □

4 Main Results

We show that virtual stations and push starts define necessary and sufficient conditions for the global pathwise stability of any two-station, open multiclass queueing network.

Definition 4.1 *Virtual stations* are sets of classes satisfying:

1. No class of a first excursion is in a virtual station, i.e., a class (i,k) in a virtual station must be in one of the excursions numbered 2, 3, $\ldots$, $|E_i|$.

2. If the last class of an excursion is in a virtual station, then every class of that excursion is in the virtual station and if a first class of an excursion is in a virtual station, then every first class of that excursion is in the virtual station. Thus, a virtual station must have either none of the classes, all of the classes, or all but the last class of each excursion.

3. If any class of an excursion is in a virtual station, then the last class of the preceding excursion cannot be in the virtual station.

Definition 4.2 A *push start set* is a set F of classes satisfying:

1. If class (i,k) is in F, then each class (i,k'), where $1 \leq k' < k$, is also in F, and

2. If a first class of an excursion is in F, then every first class of the excursion is in F.

For a set C of classes, we let C_A denote the classes of C served at station A and C_B, those served at station B.

Theorem 4.3 A two-station open multi-class queueing network is globally pathwise stable if and only if for each push start set F and each virtual station C in the subnetwork consisting of classes not in F, we have

$$\frac{\sum_{(i,k)\in C_A} m_k^i}{1-\sum_{(i,k)\in F_A} m_k^i} + \frac{\sum_{(i,k)\in C_B} m_k^i}{1-\sum_{(i,k)\in F_B} m_k^i} \leq 1. \tag{1.16}$$

Furthermore, if (1.16) is violated for some sets F and C, then there is a static buffer priority discipline under which for almost every sample path, the total number of customers in the network diverges to infinity with time.

We also show that these same conditions are both necessary and sufficient to ensure the weak stability of the corresponding fluid model.

Theorem 4.4 A two-station fluid network is globally weakly stable if and only if the service times satisfy the conditions (1.16) of Theorem 4.3.

We further show that a two-station fluid network is globally (strongly) stable if and only if the service times satisfy the conditions of Theorem 4.3 with strict inequality.

queueing discipline that gives higher priority at station A to classes in C_A and at station B to classes in C_B a *C-priority discipline*. The following lemma was proved in Dai and Vande Vate [15, Proposition 3.1].

Lemma 5.1 Let C be a virtual station in an initially empty two-station queueing network. Under a C-priority discipline,

$$\left(\sum_{(i,k)\in C_A} Q_k^i(t) \right) \left(\sum_{(i,k)\in C_B} Q_k^i(t) \right) = 0$$