Stability, Capacity, and Scheduling of Multiclass Queueing Networks

A THESIS
Presented to
The Academic Faculty

by

John Jay Hasenbein

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Industrial and Systems Engineering

Georgia Institute of Technology
September 1998

Stability, Capacity, and Scheduling

of Multiclass Queueing Networks

Approved:

Prof. Jim Dai, Co-chairman

Prof. John Vande Vate, Co-chairman

Prof. Richard Serfozo

Prof. Mark Spearman

Prof. Yorai Wardi

Date Approved by Chairman _____

Dedicated to

my parents

Acknowledgments

I would like to acknowledge the support of the National Science Foundation during the course of my research. In particular, the research herein was supported by NSF Grant DMI-9457336.

My first thanks must go to my advisors, Jim Dai and John VandeVate. They were always willing to provide advice, academic and otherwise to help me along my way. I have to attribute most of my success to Jim's guidance through most of my years here, he made following the right path easy.

Of course, I also need to thank an excellent batch of fellow graduate students, who I will not mention by name for fear of leaving someone out. They made the social, academic and all other aspects of my years at Georgia Tech the most enjoyable of my life. This thesis is dedicated to my parents, I aspire to guide my children as well as they guided me.

Finally, to Carolina Barcenas for all of her love and support during this time ... gracias de mi corazón.

Contents

List of Figures

Summary

Manufacturing processes with complex routing, feedback, and varied processing times can be modeled as multiclass queueing networks. Since a refined analysis of these networks is generally difficult, we conduct an investigation of macroscopic properties such as stability and capacity. In conjunction with this, we gain some insight into which scheduling rules should be avoided in such networks. Our stability analysis involves both discrete stochastic queueing networks and their continuous deterministic counterparts, multiclass fluid models.

The contribution of this thesis consists primarily of three parts. First we derive necessary and sufficient conditions for global stability of a class of two-station fluid networks with proportional routing. Next, we obtain necessary conditions for the global stability of multiclass queueing networks with deterministic routing and an arbitrary number of stations. Finally, we undertake an in-depth investigation of the stability properties of a particular three-station fluid network. We are able to obtain the the monotone global stability region for this network and we demonstrate a number of properties which show a contrast with the two-station case. We also discuss how these results relate to the ideas of capacity and scheduling in such networks.

Chapter 1

Introduction

With the advent in recent years of highly complex manufacturing, communication, and computer systems has come a desire for in-depth analysis and control of such systems. In many of these systems random influences have an important effect on system behavior and thus engineers must rely on stochastic models to gain insight. Probably the most useful class of models have been termed *queueing networks* or perhaps more appropriately *stochastic processing networks*. While these stochastic models provide a better reflection of reality than their deterministic counterparts, they suffer from the disadvantage of being very difficult to analyze. Nonetheless, it is still possible to gain important insight by analyzing certain macroscopic properties of these networks. Such insight can then be applied to the more efficient design, operation and control of the real-life systems.

In this study, we are primarily motivated by semiconductor wafer fabrication processes, which have unique characteristics and hence present unique difficulties in analysis and control. The main distinguishing features of such systems are two: the highly *reentrant* nature of the processing routes through which jobs traverse the workstations and the rather varied processing, batching, and setup requirements needed at different stages. The mathematical models we consider, the multiclass queueing network and its corresponding fluid model, are simplified models that take into account the reentrant nature and varied processing requirements of these systems. Despite this simplification, we discover a number of surprising

and counterintuitive properties in our analysis, which reflect on the performance of the real systems.

Unfortunately, these multiclass queueing networks present an example of networks for which a refined probabilistic analysis is generally difficult. With this in mind, we conduct an investigation of what one might call the "first–order" properties of stability and capacity. Our main focus is on the *global stability* of queueing and fluid networks. Roughly speaking, a network is globally stable if it has enough resources to handle incoming work when operating under any reasonable (non-idling) scheduling rule. A related concept is the *capacity* of these systems, which is the maximum sustainable throughput under any non-idling policy. Stability and capacity are intimately related to scheduling policies and as a result we gain some intuition into which scheduling rules should be avoided in these networks.

Stability and capacity analysis of queueing networks was perhaps thought to be a moot subject in the area after the pioneering work of Jackson [24] and Kelly [26] indicated that such analysis was a relatively trivial matter which simply depended on the traffic intensity at each station. Essentially, this simplistic analysis implies that the stability and capacity of the network depends only on looking at the capacity of each station in the network individually. Renewed interest in this area was sparked by primarily two factors: a series of counterexamples demonstrating that the station traffic intensities may not be sufficient to determine the stability region, and insight into the close relationship between discrete queueing networks and their associated fluid models.

In the first area, Kumar and Seidman [28], Lu and Kumar [30], and Rybko and Stolyar [32] gave examples of queueing networks that are unstable under certain non-idling disciplines, even if the traffic intensity is less than one at all stations. Later, Bramson [4] and Seidman [33] demonstrated that the same phenomenon could occur in both stochastic and deterministic networks under the popular first-come-first-served (FCFS) queueing

discipline.

This work inspired further investigations into the stability regions of queueing models under various scheduling policies and also spurred work on the relationship of such networks with their fluid counterparts. In a remarkable series of papers, Rybko and Stolyar [32], Dai [12, 13], and Meyn [31] demonstrated that fluid models could be an extremely powerful tool for determining the stability region of a wide range of queueing networks. Refinements and some extensions to this new tool were later provided in Chen [7]. Roughly speaking, the results of Dai [12] indicate that a queueing network will be stable if its fluid model counterpart is stable. Dai [13] and Meyn [31] provide partial converses, but recent work by Bramson [**?**] indicates that a full converse may not hold.

As a consequence of this newly developed theory, a handful of results, for both the fluid and discrete models, have been obtained concerning the stability of either specific networks (see Botvich and Zamyatin [3] or Dumas [21]) or of certain policies (Kumar and Kumar [29], Kumar [27], Bramson [6]). Of particular interest among researchers have been the stability properties of networks under priority disciplines (Dai and Weiss [17] and Chen and Zhang [11]) and the FCFS discipline (Chen and Zhang [10] and Bramson [5]). In this study, we will be primarily interested in the *global stability region*, i.e. the region in which a network with a specific (but arbitrary) topology will be stable under any non-idling policy. In this direction, researchers have been able to develop fairly sophisticated tools, such as Lyapunov functions, to analyze the stability region of the fluid models, and thereby resolve the stability issue for some discrete models.

The contribution of this thesis consists primarily of three parts and we now provide the background for each part. Perhaps the most sweeping advance in understanding the global stability properties of queueing networks was provided in a series of papers by Dai and VandeVate [15, 14, 16]. These papers provided **exact** (i.e. necessary and sufficient)

stability conditions for all two-station fluid networks with non-branching (deterministic) routing. In particular, the stability conditions are given explicitly in terms of the service and arrival rates of the network. More importantly, they were able to provide an intuitive explanation of these conditions via the phenomena of push-starts and virtual stations. A major consequence of their results is that for the class of networks they study it turns out that the global stability region is determined by the stability properties under a comparatively simple, finite set of policies known as static buffer priority disciplines. In establishing these conditions for two-station fluid networks, they also were able to provide, in some cases, exact stability conditions for the corresponding queueing networks. At about the same time, Bertsimas, Gamarnik, and Tsitsiklis [2] independently developed an LP which could sharply determine the stability region of **all** two-station multiclass fluid networks. Unfortunately, this result did not provide as much insight, since the stability region could not be expressed as an explicit function of the rates in the network.

The results of Chapter 4 attempt to narrow the gap between the results of Bertsimas et. al. and Dai and VandeVate's work on two-station fluid networks. Specifically, we derive necessary and sufficient conditions for global stability of a large class of networks with **proportional** (probabilistic) routing. Once again, these conditions can be explained intuitively in terms of push-starts and virtual stations. As in Dai and VandeVate's study, this also establishes the importance of the static buffer priority policies for the larger class of networks we investigate. The fluid networks we study arise from fluid approximations of multiclass queueing networks with probabilistic routing. Thus, the conditions we derive, in some cases, yield necessary and sufficient conditions for the global stability of the associated stochastic networks.

Of course, the next natural question to investigate is the global stability properties of queueing and fluid networks with more than two stations, a topic which is addressed in

Chapters 5 and 6. In Chapter 5, we see that necessary conditions for global stability of both queueing and fluid networks can be obtained by extending Dai and VandeVate's idea of virtual stations to multi-station networks. We show that these necessary conditions, which again can be expressed explicitly in terms of the rate of the network, arise naturally from a new phenomenon, termed *pseudostations*. Although the conditions we derive are not sufficient for stability in general, they give important insight into the third major part of our study. It should be noted that Dumas [19, 20, 21] made independent observations along these lines and our work represents a strengthening of the conditions derived by him via his concept of 'unessential' states.

In Chapter 6 we undertake an in-depth investigation of the stability properties of a particular three-station network. Although we focus on only one example, we gain considerable information through this network. We investigate the stability regions of the possible static buffer priority rules and their relation to the global stability region. We find that, unlike the two-station case, the static buffer priority policies are no longer the extremal policies in this network. That is, the worst behavior of the network does not necessarily occur under such policies. In addition, we discover an important and surprising property of our network, which again contrasts with the two station case. We show that the global stability region is not *monotone* in terms of the mean service times, meaning that increasing the efficiency of a station may destroy the global stability of the system. Furthermore, we go on to obtain the exact monotone global stability region of the network. Our results also provide counterexamples to two recent conjectures in the literature.

As noted above, an LP developed by Bertsimas, Gamarnik, and Tsitsiklis [2] can sharply determine the stability region for two-station fluid networks. It was conjectured that this LP would also work for multi-station networks. We are able to show that their LP does not characterize either the global stability region or even the monotone global stability region

of our three-station network.

The other conjecture was put forth in a paper by Chen and Zhang [11]. In that paper, they developed an LP which yielded sufficient stability conditions for networks operating under a static buffer priority policy and also gave the exact stability region for several such networks. Unfortunately, our results also show that their LP method does not characterize the stability region of our three-station network under a static buffer priority discipline.

Finally, we provide a brief outline of this dissertation. In Chapter 2 we describe the queueing and fluid networks models which are the focus of our study. The theoretical framework for stability analysis is provided in Chapter 3. New theoretical results are provided in the next three chapters, as outlined above. Finally, we provide some further directions for research in Chapter 7.

Chapter 2

The Queueing and Fluid Network Models

2.1 The Queueing Model

We now introduce the queueing model that will be the focus of our investigations. We start with a general model, which is often referred to in the literature as an *open multiclass queueing network* (OMQN). In some of our work, we only deal with special cases of an OMQN and thus we define several submodels in Section 2.4. Unless otherwise stated, all vectors should be envisioned as column vectors and any inequalities between vectors should be interpreted componentwise.

Our queueing network consists of d single-server stations, denoted $1, 2, \ldots, d$ and K customer classes, labeled similarly. Each customer class k may incur exogenous arrivals according to the process $E_k = \{E_k(t), t \geq 0\}$ where $E_k(t)$ is a counting process indicating the number of arrivals to class k in the interval $[0, t]$. We allow some of the E_k processes to be null, in which case the corresponding classes do not incur exogenous arrivals.

Customers in class k of service require service at station $\sigma(k)$ and are served according to the service process $S_k = \{S_k(t), t \geq 0\}$, with $S_k(t)$ indicating the cumulative number of services for class k customers if server $\sigma(k)$ devotes t units of time to serving this class. Note that several customer classes may be served at the same station, thus the term "multiclass" network. When a customer completes service at a station it either leaves the network or is routed to another station. Let $\Phi_\ell^k(n)$ denote the the number of class k customers routed to

class ℓ, from the first n class k service completions. The routing process for class k is then given by $\Phi^k = \{\Phi^k(n), n \geq 1\}$.

We will assume the arrival, service, and routing processes are defined on a probability space and that the arrival and service processes are right continuous. We further suppose that all three processes satisfy the following *strong law of large numbers*. As $t \to \infty$ and $n \to \infty$, with probability one:

$$\frac{E_k(t)}{t} \to \alpha_k \quad \text{for } k = 1, \ldots, K \tag{2.1.1}$$

$$\frac{S_k(t)}{t} \to \mu_k \quad \text{for } k = 1, \ldots, K \tag{2.1.2}$$

$$\frac{\Phi_\ell^k(n)}{n} \to p_{k\ell} \quad \text{for } k, \ell = 1, \ldots, K \tag{2.1.3}$$

We assume that $0 \leq \alpha < \infty$ and $0 < \mu < \infty$. The matrix $P = (p_{k\ell})$, which is substochastic by definition, is usually referred to as the *routing matrix*. In this study, we only consider open queueing networks, and thus we impose the condition that P has spectral radius less than one. One consequence of this assumption is that the matrix $(I - P')$ is invertible with

$$(I - P')^{-1} = I + P + P^2 + \ldots$$

This restriction ensures that all customers will eventually leave the network.

Assumptions (2.1.1) and (2.1.2) are rather mild and roughly speaking, hold when E_k and S_k are renewal processes. We now make this statement more precise. Let $\mathcal{E}$ denote the set of classes with non-null exogenous arrivals. Suppose each arrival process $E_k, k \in \mathcal{E}$ is characterized by interarrival times $\xi_k = \{\xi_k(n), n \geq 1\}$ and the service processes S_k are

characterized by service times $\nu_k = \{\nu_k(n), n \geq 1\}$. Further, assume $\xi_1, \ldots, \xi_K, \nu_1, \ldots, \nu_K$ are mutually independent iid sequences with finite first moments. In this case E_k will satisfy (2.1.1). Also, if there is an upper bound on the number of class k customers which can have outstanding partial services, then S_k satisfies (2.1.2). We further note that assumption (2.1.3) holds if customers are routed according to the usual Markovian routing scheme, but in fact will hold under more general schemes. For example, consider customers who complete service at class 1. Suppose we route the first 5 customers to class 2, the next 5 to class 3, the next 5 to class 2, and continue in this manner. This type of routing scheme will satisfy assumption (2.1.3) with $P_{12} = P_{13} = 1/2$.

In later sections, we may need to consider stronger conditions on the the arrival and service processes. Again consider the characterizations of E_k and S_k given above. The following conditions were introduced in Dai [12].

Assumption 2.1.1. *The interarrival times are* unbounded *and* spread out *if for each $k \in \mathcal{E}$, there exists some integer $j_k > 0$ and some function $p_k(x) \geq 0$ on $\mathbb{R}_+$ with $\int_0^\infty p_k(x)\,dx > 0$, such that*

$$P\{\xi_k(1) \geq x\} > 0 \quad \textit{for any } x > 0$$

and

$$P\left\{a \leq \sum_{i=1}^{j_k} \xi_k(i) \leq b\right\} \geq \int_a^b p_k(x)\,dx \quad \textit{for any } 0 \leq a < b$$

2.2 Queueing Disciplines and Dynamics

We next consider the dynamic equations that govern the evolution of the queueing network. We let $Q_k(t)$ be the number of customers in class k (waiting or in service) at time t and set $Q(t) = (Q_1(t), \ldots, Q_K(t))$. The allocation process is $T(t) = (T_1(t), \ldots T_K(t))$, where $T_k(t)$ is the cumulative amount of time during the interval $[0, t]$ that server $\sigma(k)$ spends serving

class k customers. This allocation process will depend on the type of queueing discipline employed in the network. For notational convenience we also define the *idle time process* for each station i as

$$U_i(t) \equiv t - \sum_{k:\sigma(k)=i} T_k(t)$$

.

With these definitions we are prepared to write down the following dynamical equations for our network:

$$Q(t) = Q(0) + E(t) + \sum_{k=1}^{K} \Phi^k(S_k(T_k(t))) - S(T(t)) \quad \text{for } t \geq 0 \tag{2.2.1}$$

$$Q(t) \geq 0 \quad \text{for } t \geq 0 \tag{2.2.2}$$

$$T(0) = 0 \quad \text{and } T_k(\cdot) \text{ is nondecreasing for } 1 \leq k \leq K \tag{2.2.3}$$

$$U_i(\cdot) \quad \text{is nondecreasing for each station } i \tag{2.2.4}$$

where $S(T(t)) \equiv (S_1(T_1(t)), \ldots, S_K(T_K(t)))$. Equation (2.2.1) implies that the queue length process $Q(t)$ will also be right continuous. The last three equations are natural assumptions given our definitions of $Q(t)$ and $T(t)$.

The concept of a queueing discipline, which we turn to next, is of paramount importance in our study. When the queueing network undergoes a change of state, i.e. a customer arrives or completes service at a station, the server must determine which customer to work on next. The rule that the server uses to make this decision is known as a *queueing discipline* or *dispatch policy*. If we specify that our servers must act according to a particular policy or class of policies, this will impose additional restrictions on the allowable allocation processes $T(t)$ that determine the behavior of the queue length process $Q(t)$. When such restrictions

apply, we will need to augment (2.2.1)–(2.2.4) with additional equations. If simultaneous events are allowed in our model, then certain pathological behavior can occur depending on how deadlocks are broken by the system (see Whitt [34] for example). Hence, to avoid these pathologies, we assume that such events are treated as if they occurred sequentially, rather than simultaneously. For example, if two customers finish service at the same time, we assume one service is completed before the other, and at this time implement all system logic that follows due to this completion.

We now define the set of *non-idling* dispatch policies to be those under which a server must do work whenever there are customers waiting to be served. This class of policies imposes the additional conditions:

$$\int_0^\infty Z_i(t)\, dU_i(t) = 0 \quad i = 1, \ldots, d \tag{2.2.5}$$

where

$$Z_i(t) := \sum_{k:\sigma(k)=i} Q_k(t) \quad i = 1, \ldots, d$$

on our dynamical equations. From now on, we will assume that our networks are operating within the class of non-idling disciplines.

Another class of dispatch policies that we will be dealing with frequently is the class of *static buffer priority* policies. At a server employing such a discipline, customers are served according to some fixed ranking of the classes and on a first-come-first-served basis (FCFS) within classes. This class of policies will play a pivotal role in later sections. We describe a buffer priority policy π as a permutation of the classes $1, \ldots, K$ in the network. Classes listed first in the permutation have higher priority than those listed later. For example, in the network pictured in Figure 1, the priority policy that gives highest priority to classes 4, 2, and 6 is written as $\pi_{\{4,2,6,1,5,3\}}$. Since the lowest priority classes can be dropped from

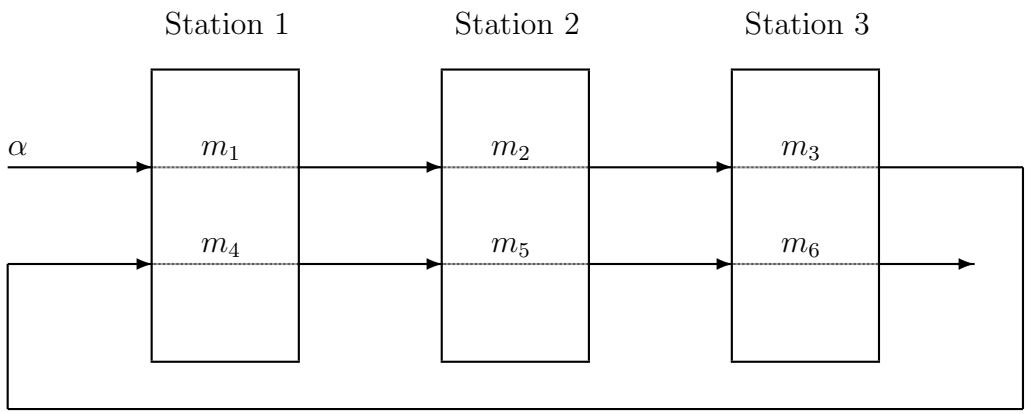

Figure 1: A simple priority network

the list without ambiguity we will write this same policy as $\pi_{\{4,2,6\}}$. Furthermore, if class k has higher priority than class ℓ, then we write $\pi(k) < \pi(\ell)$. For simplicity of exposition, we assume that the priority disciplines described above are *preempt-resume*, this means that a server can interrupt service to one customer to serve a higher priority customer and then resume service at a later time.

In a reentrant line (see Section 2.4), the static buffer priority policy for which $\pi(k) < \pi(\ell)$ if and only if $k > \ell$ is called last-buffer-first-served (LBFS). For reentrant lines, this is equivalent to a commonly seen policy in industry often called shortest-(expected)remaining-processing-time-first or first-in-system-first-out (FISTFO). The static buffer priority policy for which $\pi(k) < \pi(\ell)$ if and only if $k < \ell$ is called first-buffer-first-served (FBFS).

We need to describe additional equations that the queueing network must obey if it is

operating under a fixed static buffer priority policy π. To this end we let:

$$H_k = \{\ell : \sigma(\ell) = \sigma(k), \pi(\ell) \leq \pi(k)\} \tag{2.2.6}$$

$$T_k^+(t) = \sum_{\ell \in H_k} T_\ell(t) \tag{2.2.7}$$

$$U_k^+(t) = t - T_k^+(t) \tag{2.2.8}$$

$$Z_k^+(t) = \sum_{\ell \in H_k} Q_\ell(t) \tag{2.2.9}$$

Now a network operating under a preempt-resume static buffer priority policy π must obey:

$$\int_0^\infty Z_k^+(t) dU_k^+(t) = 0 \quad \text{for all } k = 1, \ldots, K \tag{2.2.10}$$

This expression ensures that a lower priority customer cannot receive service if there is positive workload due to higher priority customers.

2.3 Fluid Models

In this section we introduce the notion of a fluid network. A fluid network is a **continuous deterministic** dynamical model that is an analog to the **discrete stochastic** queueing network. In the fluid network the notion of discrete customers is replaced by the notion of fluids or customer mass. The connection between the two models will become more apparent in Chapter 3.

We start by briefly reviewing the notion of a fluid limit, for more details, we refer the reader to Chen and Mandelbaum [8] and Dai [12, 13]. Consider the joint process $(Q(t), T(t))$,

where $Q(t) = (Q_k^i(t))$ and $T(t) = (T_k^i(t))$ are the vector-valued queue length and allocations processes as previously defined. We say that this process has a *fluid limit* $(\bar{Q}(t), \bar{T}(t))$ if for some sample path ω and a sequence $r_n \to \infty$:

$$\left(\frac{Q(r_n t, \omega)}{r_n}, \frac{T(r_n t, \omega)}{r_n}\right) \longrightarrow \left(\bar{Q}(t), \bar{T}(t)\right) \quad \text{uniformly on compact sets} \tag{2.3.1}$$

The type of scaling in time and space used in (2.3.1) will be referred to as *fluid scaling*. Under our assumptions (2.1.1)–(2.1.3) on the queueing network, it can be shown that the fluid limits $\bar{Q}(t)$ and $\bar{T}(t)$, if they exist, must satisfy a set of fluid equations. This set of dynamical equations that corresponds with a queueing model will be collectively referred to as the *fluid model*, which we describe below.

In the fluid model that corresponds to the queueing network described in Section 2.1, we once again have d single-server stations and K classes of fluids that are processed at the various stations. Fluid of class k may arrive from the outside at rate $\alpha_k > 0$. Also, class k fluid requires processing at station $\sigma(k)$ and can be processed at the maximum rate of $0 < \mu_k < \infty$ if station $\sigma(k)$ devotes all of its effort to processing class k fluid. The service time for a class k fluid is $m_k = 1/\mu_k$, i.e. the time it takes to process one unit of fluid. As before, multiple classes may be served at a single station. After class k fluid is processed at a station it is routed to another station or stations according to the routing matrix P. If $p_{k\ell} = 1$, then all of the class k fluid is routed to class ℓ. In the case where $0 < p_{k\ell} < 1$, we have *proportional routing* in our network, i.e. a proportion $p_{k\ell}$ of class k fluid is routed to class ℓ. This proportional routing is analogous to probabilistic routing in the stochastic network. Any fluid that is not routed to another class leaves the network.

Let us denote the amount of class k fluid in the network by $\bar{Q}_k(t)$ and let $\bar{T}_k(t)$ denote the amount of time server $\sigma(k)$ devotes to class k fluid in the interval $[0, t]$. Again, the

allocation process $\bar{T}(\cdot)$ will depend heavily on the dispatch policy employed in the network. In general, we will use the bar notation to denote fluid quantities. In Chapters 4 and 6 we will be concerned only with fluid networks and thus drop this notation to avoid cluttered formulas.

With these definitions we are ready to write down the dynamical equations for our fluid network that are analogous to (2.2.1)–(2.2.4):

$$\bar{Q}(t) = \bar{Q}(0) + \alpha t + (I - P')\Delta\bar{T}(t) \quad \text{for } t \geq 0 \tag{2.3.2}$$

$$\bar{Q}(t) \geq 0 \quad \text{for } t \geq 0 \tag{2.3.3}$$

$$\bar{T}(0) = 0 \quad \text{and } \bar{T}_k(\cdot) \text{ is nondecreasing for } 1 \leq k \leq K \tag{2.3.4}$$

$$\bar{U}_i(\cdot) \quad \text{is nondecreasing for each station } i \tag{2.3.5}$$

where $\Delta = \text{diag}(\mu)$ is the diagonal matrix of the service rates and

$$\bar{U}_i(t) \equiv t - \sum_{k:\sigma(k)=i} \bar{T}_k(t)$$

We can consider dispatch policies for the fluid network that are analogous to those in the queueing network. Once again, a non-idling policy is one in which station i must work at full speed whenever there is a positive fluid level at station i. We can express this constraint

on the allowable allocation processes by

$$\int_0^\infty \bar{Z}_i(t)\, d\bar{U}_i(t) = 0 \quad i = 1, \ldots, d \tag{2.3.6}$$

where

$$\bar{Z}_i(t) := \sum_{k:\sigma(k)=i} \bar{Q}_k(t) \quad i = 1, \ldots, d$$

The class of static buffer priority disciplines for fluid networks is analogous to those in the queueing model. Once again, the classes at a station are assigned a fixed ranking, and the server cannot devote any effort to processing class k fluid unless the fluid level is zero for all classes with a higher ranking. We define fluid quantities that are analogous to those in the discrete model:

$$\bar{T}_k^+(t) = \sum_{\ell \in H_k} \bar{T}_\ell(t) \tag{2.3.7}$$

$$\bar{U}_k^+(t) = t - \bar{T}_k^+(t) \tag{2.3.8}$$

$$\bar{Z}_k^+(t) = \sum_{\ell \in H_k} \bar{Q}_\ell(t) \tag{2.3.9}$$

Again, a fluid network operating under a preempt-resume static buffer priority policy π must obey:

$$\int_0^\infty \bar{Z}_k^+(t) d\bar{U}_k^+(t) = 0 \quad \text{for all } k = 1, \ldots, K \tag{2.3.10}$$

2.4 Submodels

We will find it useful in subsequent sections to deal with a number of special cases or submodels, of an OMQN. These submodels are primarily differentiated by the type of routing scheme or routing matrix allowed. The definitions outlined in this section apply to both fluid and discrete networks.

The first definition is adapted from Chen and Yao [9], which studies networks with an acyclic transfer mechanism.

Definition 2.4.1. Let

$$\mathcal{I}(k) \equiv \{i = 1, \ldots, K : p_{i\ell_1} p_{\ell_1\ell_2} \cdots p_{\ell_n k} > 0 \quad \text{for some } \ell_1, \ldots, \ell_n\}$$

i.e. $\mathcal{I}(k)$ is the set of classes from which k is reachable. A multiclass network is said to be an *acyclic transfer mechanism network* (ACTN) iff $k \notin \mathcal{I}(k)$ for all classes $k = 1, \ldots, K$.

Essentially, customers or fluid in an ACTN cannot pass through any given buffer more than once. A network that is an ACTN is depicted in Figure 2.

Definition 2.4.2. A multiclass network is a *strictly branching network (SBN)* if

- It is an ACTN.
- For every class k there exists at most one class ℓ, such that $p_{\ell k} > 0$.
- For every class $k, \alpha_k > 0$ iff there are no classes ℓ, such that $p_{\ell k} > 0$.

A simple example of a strictly branching network is pictured in Figure 3. The class of SBN's will be the main focus of study in Chapter 4. In that chapter, we will also draw a

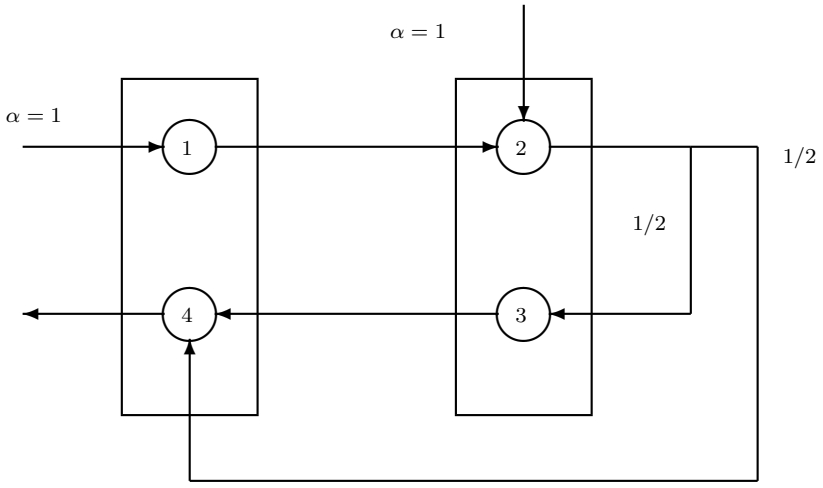

Figure 2: An acyclic transfer mechanism network

connection between the class of ACTN's and SBN's.

Definition 2.4.3. A multiclass network is a *multi-type network (MTN)* if

- It is a strictly branching network.
- $p_{\ell k}$ is 0 or 1 for every $1 \leq \ell, k \leq K$.

Essentially, a discrete MTN employs deterministic routing, i.e. there are a number of products or *types* of customers, each of which follows a deterministic route through the network. A fluid MTN does not employ proportional routing. We will study the behavior of both fluid and discrete MTN's in Chapter 5. A simple MTN is portrayed in Figure 4.

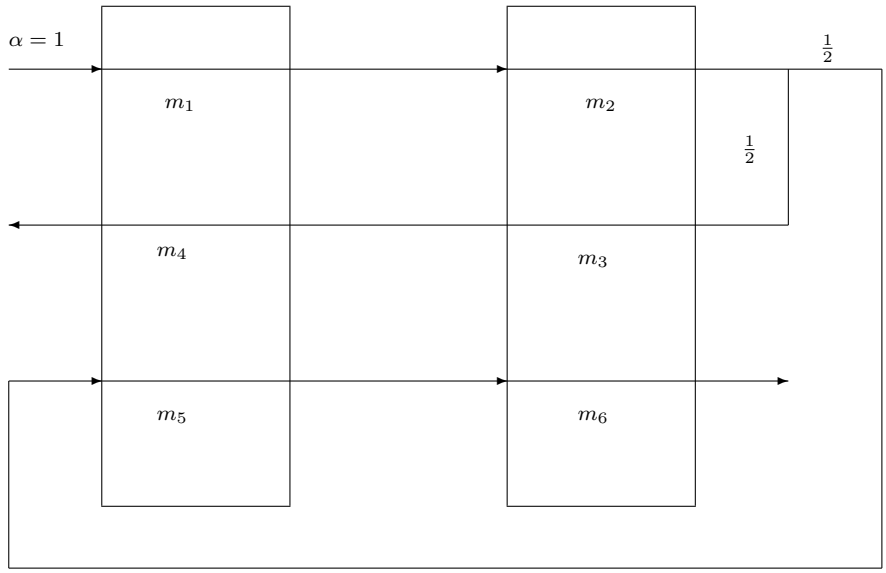

Figure 3: A strictly branching network

Definition 2.4.4. A multiclass network is a *reentrant line* if

- It is a multitype network.
- The classes $1, \ldots, K$ can be labeled such that $p_{k,k+1} = 1$ for all $1 \leq k < K$.

In other words, a reentrant line is a MTN with only one type of customer. Chapter 6 investigates the properties of the fluid reentrant line pictured in Figure 1.

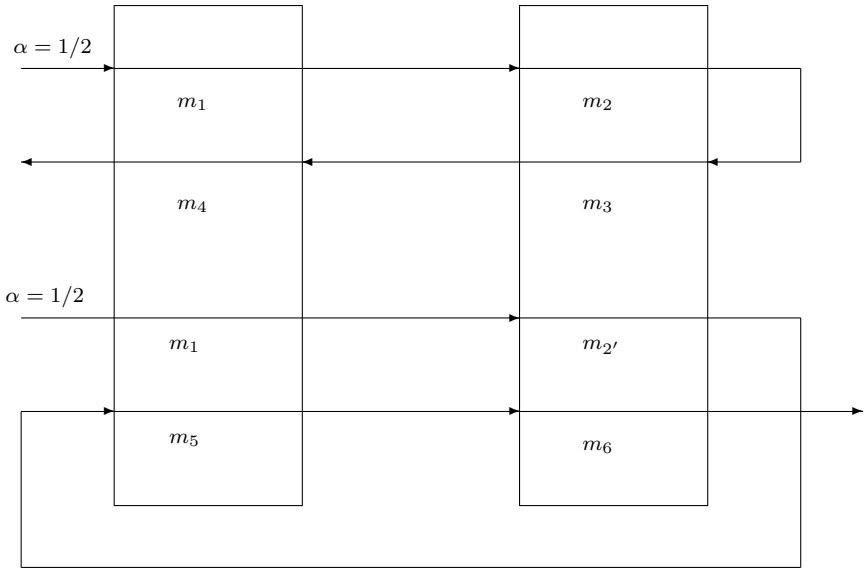

Figure 4: A multi-type network

Chapter 3

Stability and Capacity

3.1 Stability and Capacity

As mentioned in Chapter 1, the objective of our study is to investigate the macroscopic property of queueing and fluid networks that has been termed stability. In this chapter, we define notions of stability for queueing and fluid networks and explain how fluid networks can be exploited to more easily analyze the stability of their stochastic counterparts. In Section 3.3 we discuss an approach to the stability analysis of fluid networks via Lyapunov functions, specifically piecewise linear Lyapunov functions. This will then set the stage for more in-depth investigations of stability in subsequent chapters.

One notion of stability that we will explore is most easily connected to the notion of positive recurrence for Markov chains. First, consider a multiclass queueing network in which the arrival processes E_k and the service processes S_k are Poisson. In this case, with an appropriate state space $\{X(t), t \geq 0\}$ (which may depend on the service discipline), the state process of the queueing network can be modeled as a Markov process with a countable state space and stability of this process is equated with positive recurrence. For the case of non-Poisson arrival and service processes, we may augment the state space (i.e. add residual service and arrival times) to retain the Markov property. In fact, it can be shown that the state process $\{X(t), t \geq 0\}$ is a strong Markov process given appropriate assumptions on the input and service processes and a proper augmentation of the state space (see Dai [12]).

For such a Markov process, we may then consider the notion of positive Harris recurrence as a measure of stability. Thus, one sensible definition of stability is as follows:

Definition 3.1.1. A queueing network is *stable* if the associated state process $\{X(t), t \geq 0\}$ is positive Harris recurrent.

More information on positive Harris recurrence can also be found in Dai [12]. We note that in the case of a Markov chain, the notion of positive Harris recurrence coincides with the more familiar notion of positive recurrence.

To introduce a different notion of stability for the queueing network we need to define the concepts of effective arrival rates and traffic intensity. For each class k, the *effective arrival rate* λ_k indicates the long-term arrival rate to class k, due to both internal and external arrivals, that we would expect if the system is in a steady state. We present the vector traffic equation:

$$\lambda = \alpha + \lambda P' \tag{3.1.1}$$

Due to our assumption on P, equation (3.1.1) always has a unique solution. We thus define the vector of effective arrival rates as:

$$\lambda = (I - P')^{-1}\alpha$$

Next, for each station $i = 1, \ldots, d$ let

$$\rho_i = \sum_{k:\sigma(k)=i} \lambda_k m_k$$

We will refer to ρ_i as the nominal workload or traffic intensity for server i. Note that for

later purposes, the definitions for ρ and λ apply to both queueing and fluid networks.

Next we let $D_k(t) = S_k(T_k(t))$ be the departure process from each class k in the queueing network, i.e. $D_k(t)$ is the number of class k customers which have completed service in $[0, t]$. It is reasonable to expect that in a stable system, the long-run departure and arrival rates are equal. We thus consider the idea of pathwise stability, a concept first introduced by El-Taha and Stidham [22].

Definition 3.1.2. A queueing network is said to be *pathwise stable* if for every class k:

$$\frac{D_k(t)}{t} \longrightarrow \lambda_k \quad \text{as } t \to \infty \tag{3.1.2}$$

with probability one.

We note that the concept of pathwise stability for a queueing network is generally speaking weaker than the notion of stability. For example, consider an $M/M/1$ queueing network with arrival rate α and service rate μ. This system is both stable and pathwise stable when $\alpha < \mu$ and is neither when $\alpha > \mu$. When $\alpha = \mu$ it is well known that the underlying Markov process is null recurrent and thus by our definition not stable. In this case, though, the system is pathwise stable.

For the fluid model, which is a deterministic system, we characterize stability in terms of fluid solutions. A pair $(\bar{Q}(\cdot), \bar{T}((\cdot))$ is a fluid solution if it satisfies (2.3.2)–(2.3.5) and any additional equations imposed by the class of disciplines under study. With this in hand, we introduce below several stability notions related to the fluid network.

Definition 3.1.3. A fluid network is *globally stable* if there exists a time $\delta > 0$ such that for each non-idling fluid solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ satisfying (2.3.2)–(2.3.6) with $|\bar{Q}(0)| = 1$, $\bar{Q}(t) = 0$

for all $t \geq \delta$.

Definition 3.1.4. A fluid network under a static buffer priority discipline π is *stable* if there exists a time $\delta > 0$ such that for each fluid solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ satisfying (2.3.2)–(2.3.6) and (2.3.10) with $|\bar{Q}(0)| = 1$, $\bar{Q}(t) = 0$ for $t \geq \delta$.

Definition 3.1.5. A fluid solution $(\bar{Q}(\cdot), \bar{T}(\cdot))$ is *unstable* if there is no $\delta > 0$ such that $\bar{Q}(t) = 0$ for all $t \geq \delta$.

Definition 3.1.6. The fluid model under a given policy or class of policies is *weakly unstable* if there exists a time $\delta > 0$ such that $\bar{Q}(\delta) \neq 0$ for each fluid solution $\bar{Q}(\cdot)$ with $|\bar{Q}(0)| = 0$.

Definition 3.1.7. For given $\alpha = (\alpha_i) > 0$, the *global stability region* $\mathcal{D}_\infty$ of a fluid network is the set of positive service times $m = (m_k)$ for which a fluid network is globally stable. For a given $\alpha = (\alpha_i) > 0$ and a static buffer priority discipline π, the *stability region* $\mathcal{D}_\pi$ of a fluid network under the discipline is the set of positive service times $m = (m_k)$ for which the fluid network under the discipline is stable.

Definition 3.1.8. For a given arrival vector $\alpha = (\alpha_i) > 0$, the *monotone global stability region* $\mathcal{M}_\infty$ of the fluid network is the set of positive service time vectors m such that the fluid network is globally stable for all positive service time vectors $\tilde{m} \leq m$.

Note that our stability definitions require that **all** fluid solutions converge to zero after some finite time, while the notion of weakly unstable requires that all fluid solutions "pop up" uniformly at some time. Thus, the two definitions do not cover all possible behavior of the fluid solutions.

We next focus on the concept of the capacity of a queueing or fluid network. If a given system is stable, then we know that in the long-run the amount of customers or fluid in

the system remains "reasonable" or at least does not diverge to infinity. This implies that over time the rate at which customers are processed is roughly equal to the arrival rate of these customers. In industry, this processing rate is often called throughput. One notion of interest to a system designer may be the maximum sustainable throughput for a particular system, which we term capacity.

For a reentrant line, there is only one arrival rate α and thus only one throughput rate for such a system. For the case of a reentrant line we can define the capacity of the system as follows:

Definition 3.1.9. The capacity Λ_π of a reentrant line operating under the dispatch policy π, with fixed routing and processing rate vector μ is given by

$$\Lambda_\pi = \sup\{\alpha > 0 : \text{the network with arrival rate } \alpha \text{ is stable under } \pi\} \tag{3.1.3}$$

Definition 3.1.10. The capacity Λ_∞ of a reentrant line operating under the class of non-idling dispatch policies, with fixed routing and processing rate vector μ is given by

$$\Lambda_\infty = \sup\{\alpha > 0 : \text{the network with arrival rate } \alpha \text{ is globally stable}\} \tag{3.1.4}$$

Since determining the global stability region immediately gives the capacity Λ_∞ for a fixed network, it is the global stability region of various networks, both fluid and discrete that will be the primary object of study in this dissertation.

3.2 Stability Analysis

It is often difficult to verify directly the stability of a discrete stochastic network. In this section, we describe the connection between the stability of a queueing model and its fluid analog, which will provide a powerful tool for the stability analysis of such stochastic networks. The first result we quote is from Dai [12]:

Theorem 3.2.1. *Suppose Assumption 2.1.1 holds for the service and arrival processes in the queueing model, then the queueing model under a specified discipline or class of disciplines is stable if the corresponding fluid model is stable.*

A partial converse to the above theorem, which we will use in later sections, was provided in Dai [13]:

Theorem 3.2.2. *If the corresponding fluid model is weakly unstable, then the queueing network is unstable in the sense that with probability one:*

$$|Q(t)| \longrightarrow \infty \quad as\ t \to \infty$$

Remarks: The corresponding fluid model is described by (2.3.2)–(2.3.6) and any additional equations that are necessary to specify the class of queueing disciplines. We refer to Theorem 3.2.2 as a partial converse because of the gap between the notions of stability and weak instability discussed in Section 3.1.

It is well-known (see, for example, Chen [7]) that no fluid solutions is stable unless the

traffic intensity or work arriving per unit time for each station is less than 1, i.e.

$$\rho_i < 1 \quad \text{for} \quad i = 1, \ldots, d \tag{3.2.1}$$

Hence these conditions are necessary for stability of the fluid model. If the conditions (3.2.1) hold, we say that the *usual traffic conditions* are satisfied.

Now, if we have $\rho_i > 1$ for any station i, then in fact all fluid solutions will diverge to infinity. By virtue of Theorem 3.2.2, any associated queueing model will be neither stable or pathwise stable. Thus the weaker traffic conditions

$$\rho_i \leq 1 \quad \text{for} \quad i = 1, \ldots, d \tag{3.2.2}$$

are necessary for stability of both the fluid and discrete networks. Our objective is to investigate what other conditions may be necessary and/or sufficient for global stability of a given class of networks.

In Chapter 4, we will derive necessary and sufficient conditions for stability of a class of two-station fluid models. Hence, these results provide sufficient conditions for the stability of the associated queueing models, via Theorem 3.2.1. In some cases these conditions can also be shown to be necessary for the queueing model.

Necessary conditions for stability of multi-station queueing models are derived in Chapter 5. Essentially, we show that the associated fluid model is unstable if our conditions do not hold and thus Theorem 3.2.2 implies instability of the queueing model as well.

Finally, in Chapter 6, we undertake an in-depth investigation of the stability region of a certain three-station fluid network. In particular, we obtain sufficient conditions for stability

of this network and also derive the exact monotone global stability region. Unfortunately, these results tell us little about the associated queueing model, except to also yield some sufficient conditions, by again invoking Theorem 3.2.1.

3.3 Lyapunov Functions

More often than not, it is impossible to investigate directly the behavior of all admissible fluid solutions in order to determine the stability of a class of fluid networks. For this reason, we depend heavily on the concept of a potential function or Lyapunov function to analyze the behavior of fluid solutions. Suppose we can find a function $f(\cdot)$ of the fluid level vector $\bar{Q}(t)$ such that

- $f(\bar{Q}(t)) \geq 0 \quad$ for all $t \geq 0$
- $f(\bar{Q}(t)) = 0$ implies $|\bar{Q}(t)| = 0$
- $f(\bar{Q}(t)) = 0$ for all fluid solutions $\bar{Q}(t)$ and all $t > \delta$, where δ is a fixed time greater than zero.

Such a function $f(\cdot)$ is called a *Lyapunov function* for the fluid model. Often we will denote $f(\bar{Q}(t))$ as simply $f(t)$. It is clear from our definition of stability that the fluid model will be stable if a Lyapunov function exists for that network. However, this fact is not particularly useful to establish necessary and sufficient stability conditions, since obtaining sufficient conditions which are also necessary may require us to search over all possible functions $f(\cdot)$. Instead, we will restrict our search to a particular class of Lyapunov functions, specifically piecewise linear functions. Existence of such of a function can then establish sufficient conditions, which in some cases, can be shown to be necessary by other means.

In order to proceed, we need to review some analytical results. We call a function $g(\cdot)$ *regular* at t if it is differentiable at t and we will use $\dot{g}(t)$ to denote the derivative of $g(\cdot)$

at such a regular point. Any expressions involving $\dot{g}(t)$ assume that t is a regular point of g. Also, we recall that any absolutely continuous function on $[0, \infty)$ is regular almost everywhere with respect to the Lebesgue measure.

Now, we quote the following basic lemma from Dai and Weiss [17]:

Lemma 3.3.1. *Let g be an absolutely continuous non-negative function.*

1. *If $g(t) = 0$ and t is regular, then $\dot{g}(t) = 0$.*
2. *Suppose there exists an $\epsilon > 0$ such that for every t regular $g(t) > 0$ implies $\dot{g}(t) \leq -\epsilon$. Then $g(t) = 0$ for all $t \geq \delta$, where $\delta = g(0)/\epsilon$. Furthermore, $g(\cdot)$ is nonincreasing and hence once it reaches zero it stays there forever.*

Next we note that the fluid dynamical equation (2.3.5) implies that $\bar{T}_k(\cdot)$ and $\bar{U}_i(\cdot)$ are Lipschitz continuous and thus absolutely continuous. Then (2.3.2) gives us that $\bar{Q}_k(\cdot)$ is also absolutely continuous.

In Chapters 4 and 6 we will introduce different but related Lyapunov functions to aid in our stability analysis of certain classes of fluid networks. All of the Lyapunov functions we will be dealing with are *max-linear* functions of the fluid buffer levels $\bar{Q}_k(\cdot)$, i.e. the proposed functions are of the form

$$f(\bar{Q}(t)) = \max\{f_1(\bar{Q}(t)), f_2(\bar{Q}(t)), \ldots, f_N(\bar{Q}(t))\}$$

where $f_1(\bar{Q}(t)), f_2(\bar{Q}(t)), \ldots, f_N(\bar{Q}(t))$ are linear functions of the buffer levels and N is a finite index. It is clear that the Lipschitz continuity of $f(\cdot)$ then follows from the Lipschitz continuity of $Q(\cdot)$. The following proposition, which combines the above facts and lemma, will be extremely useful throughout our analysis.

Proposition 3.3.2. *Let $f(\cdot)$ be a non-negative max-linear function of $\bar{Q}(\cdot)$ and suppose*

1. *$f(\bar{Q}(t)) = 0$ implies $|\bar{Q}(t)| = 0$*

2. *There exists an $\epsilon > 0$*

$$\frac{df(\bar{Q}(t))}{dt} \leq -\epsilon \tag{3.3.1}$$

for each time t that is regular for $\bar{T}(\cdot)$ and $f(\bar{Q}(\cdot))$ with $|\bar{Q}(t)| > 0$

then $f(\cdot)$ is absolutely continuous and is a Lyapunov function for the fluid model.

Chapter 4

Stability of Fluid Networks with Proportional Routing

4.1 Introduction

In this chapter we derive necessary and sufficient conditions for global stability of two-station fluid networks with proportional routing. Our analysis generalizes the results obtained by Dai and VandeVate [16] and provides more insight into the results of Bertsimas, Gamarnik, and Tsitsklis [2]. Unfortunately, the results do not extend to the full class of two-station OMQN's, but rather to the class of ACTN's defined in Chapter 2. We note that the class of SBN's is a subset of the class of ACTN's. In particular, an ACTN only requires the first condition in the definition of SBN's. However, as we will discuss further in Section 4.7, any ACTN can be equivalently relabeled as a SBN and thus if we obtain the global stability region for the class of SBN's we have obtained the stability region for the class of ACTN's.

Our procedure for obtaining the exact global stability conditions is as follows. To prove sufficiency we examine a certain class of piecewise linear Lyapunov functions and show that such a Lyapunov function exists if the conditions are satisfied; then, necessity of the conditions is shown via the intuitively appealing phenomena of "virtual stations" and "pushstarts", originally introduced by Dai and VandeVate [16].

In Section 4.2 we give the framework for our model. Sections 4.3 through 4.6 provide the sufficiency arguments for our main theorem. The necessity is proven in Section 4.7, completing the proof of our main result. Since we only deal with fluid quantities in this

chapter, we will drop the "bar" notation of the previous two chapters.

4.2 Definitions and Notation

We reproduce a review of the Minimum Flow problem from Dai and VandeVate [16], since it will play an important role in proving our sufficiency result. See Ahuja *et al.* [1] for further background on network flow problems.

Consider a directed network (N, E) with node set N and edge set E. We distinguish two vertices s, the source, and t, the sink. Given (possibly infinite) lower bounds $\ell = (\ell_{ij})$ and upper bounds $u = (u_{ij})$, we wish to find a minimum flow from the source s to the sink t subject to flow conservation constraints and edge capacity constraints. Thus, the minimum flow problem is:

$$\text{minimize } v$$

$$\text{subject to}$$

$$\sum_{j \in N} x_{sj} - \sum_{j \in N} x_{js} = v \tag{4.2.1}$$

$$\sum_{j \in N} x_{ij} - \sum_{j \in N} x_{ji} = 0 \quad \text{for each node } i \in N \setminus \{s, t\} \tag{4.2.2}$$

$$\sum_{j \in N} x_{tj} - \sum_{j \in N} x_{jt} = -v \tag{4.2.3}$$

$$\ell_{ij} \leq x_{ij} \leq u_{ij} \qquad \text{for each edge } (i, j) \in E. \tag{4.2.4}$$

Suppose (x, v) satisfies (4.2.1)–(4.2.4). We refer to the vector x as a *feasible flow* and the value v as the *value of the flow* x. A *minimum flow* is a feasible flow with smallest value

among all feasible flows.

An s,t-*cut* in the network (N,E) is a partition of N into two sets S and T with $s \in S$ and $t \in T$. The *capacity* of the cut (S,T), denoted $c(S,T)$, is:

$$c(S,T) = \sum_{(i,j)\in E: i\in S, j\in T} \ell_{ij} - \sum_{(i,j)\in E: i\in T, j\in S} u_{ij}.$$

Note that our definition of capacity interchanges the roles of upper and lower bounds in the usual definition as applied to the maximum flow problem. This definition is appropriate for the minimum flow problem and is sometimes referred to as the *floor* of a cut. A *maximum* s,t-*cut* is one with largest capacity among all s,t-cuts. Theorem 4.2.1 is a classic result of network flows and can be found in Ahuja *et. al.* [1, Exercise 6.18, pp. 202].

Theorem 4.2.1. *The value of a minimum flow equals the capacity of a maximum* s,t*-cut.*

Next we introduce some definitions and notation required to present our analysis. In this chapter we will deal only with the class of two-station fluid SBN's. We will arbitrarily label one station A and the other station B.

At times, we will need to consider the *split network equivalent* or SNE of a strictly branching network. The SNE is a multitype network in which there is a type for every possible route in the original network. The service rate for each class in the SNE is the same as in the original network, however, the arrival rates are set to the effective arrival rate of the exit class of each possible route in the original network. The SNE of the network in Figure 5 is pictured in Figure 6.

The term "equivalent" may be misleading. It is tempting to think that the global stability conditions for the SNE are the same as the original network, but this is not the case. The relation between the SNE and the original network is discussed further in Section

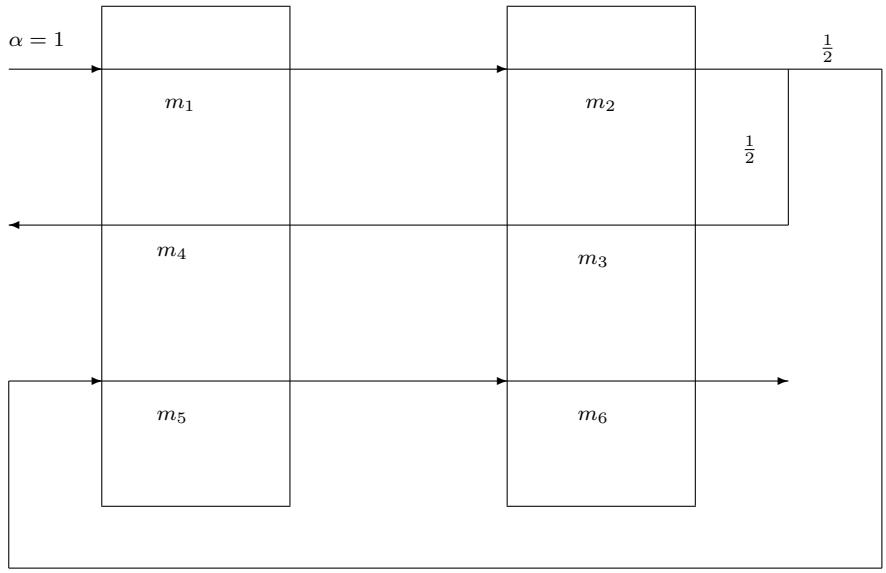

Figure 5: A strictly branching network

4.7.

With the routing restriction for a SBN, we can think of the network serving a set of I different fluid types. After a fluid type is processed at a station, it is then proportionally routed to any number of other stations for service. Since the routing structure does not allow a fluid to revisit a class, we can label each type with a finite set of class labels. For notational ease, we append a type label and speak of class (i, k) fluids. Accordingly, we also add a type label to the service time m_k^i and μ_k^i. Finally, define

$$A^i = \{(i,k) : \sigma[(i,k)] = A\}$$

$$B^i = \{(i,k) : \sigma[(i,k)] = B\}$$

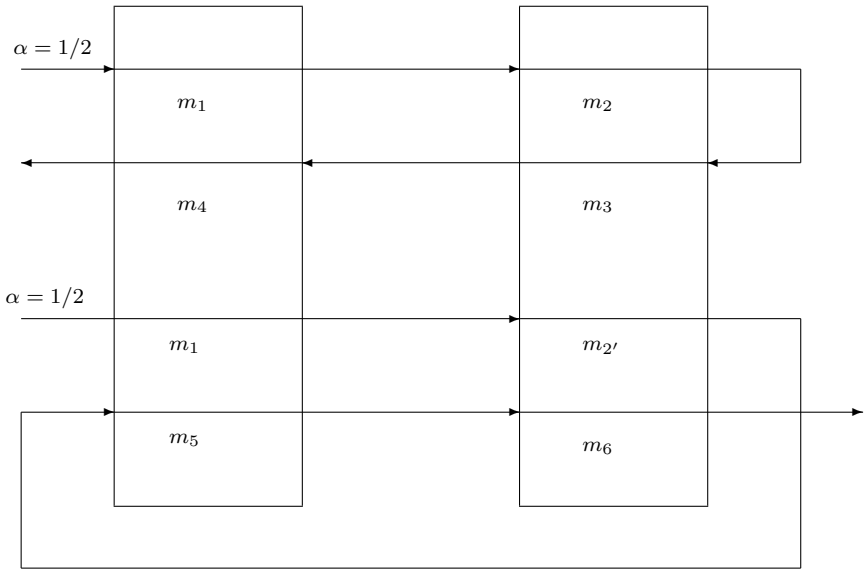

Figure 6: The SNE for the network of Figure 5

i.e. A^i is the set of classes of type i, served at station A.

The routing in a SBN induces a special structure on the classes in the network. Within a type i if class (i,k) must be visited before class (i,j), then we will write $(i,k) \prec (i,j)$ or in some contexts, simply $k \prec j$. So, we see that the routing induces a partial order on the classes in the network and that $\prec$ satisfies the usual partial order relations. Specifically, note that within a fluid type, two arbitrary classes need not possess the trichotomy property ($k \succ j$ or $k \prec j$ or $k = j$), unlike the case in a multitype network. Also, by our definition, the minimal element within a type is the only one that has a nonzero exogenous arrival rate.

It will be useful to group sets of classes at a station into *excursions*, which are blocks of consecutive visits to a station. The e^{th} excursion of type i classes will be denoted by $[i,e]$

and the classes in excursion $[i, e]$ will be denoted by $E[i, e]$. Each set $E[i, e]$ can be further divided into the last class in the excursion $\ell[i, e]$ and the rest of the classes $f[i, e]$, called *first classes*. When we deal with a set of excursions, we will use curly braces to indicate this. For example, the set of possible excursions that directly follow excursion $[i, e]$ will be denoted by $\{i, e^+\}$. Also, we let $f\{i, e^+\}$ denote the set of classes which are in $E\{i, e^+\}$ and which are not last classes in any possible excursion.

In Figure 7, there are two possible excursions at station A, one consisting of class 1 only and one consisting of classes 7, 8, and 9. At station B there are also two possible excursions, one consists of classes 2 and 3, which occurs if the customer leaves station B after its service at class 3. The other possibility is an excursion which includes classes 2 through 6, which occurs if the customer remains at station B after its service at class 3. In particular, we note that the classes 4, 5, and 6 alone do not comprise an excursion.

We need to introduce some definitions related to excursions in order to state our results succinctly:

Definition 4.2.1. A set X of excursions such that for each type $i \in I$ if $[i, e] \in X$ then $\{i, e^+\} \cap X = \emptyset$ is said to be *separating*. A separating set X is called A-strictly separating if it contains no first excursions at station A. We define *B-strictly separating* sets similarly.

Note that the set of excursions at station A is B-strictly separating and the set of excursions at station B is A-strictly separating. We call these two sets *trivial* separating sets.

Definition 4.2.2. Each separating set S of excursions induces a collection $V(S)$ which is

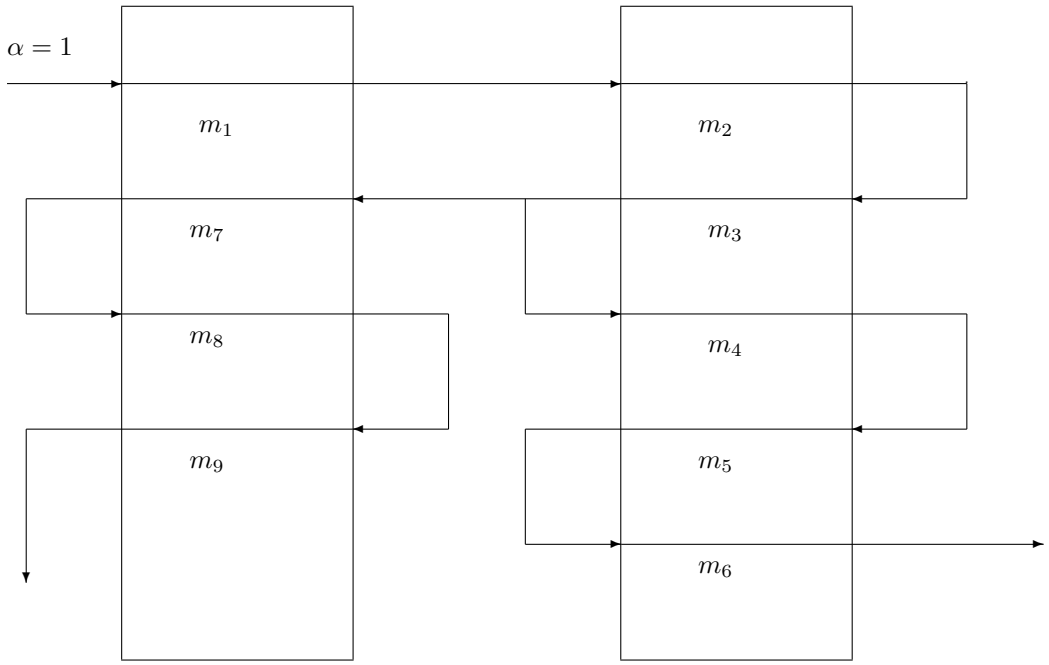

Figure 7: A strictly branching example

described by the following:

$$V(S) = \left(\cup_{[i,e]\in S} E[i,e]\right) \bigcup \left(\cup_{[i,e]\in E^i \setminus S} \hat{f}\{i,e^+\}\right) \bigcup \left(\cup_{[i,e]\in S} \overline{f}\{i,e^+\}\right)$$

where $\hat{f}\{i,e^+\}$ is the set of first classes which follow excursion $[i,e]$ and occur at a different station than $[i,e]$, $\overline{f}\{i,e^+\}$ is the set of first classes which follow excursion $[i,e]$ and occur at the same station as $[i,e]$. When S is strictly separating we refer to $V(S)$ as a *virtual station*.

Definition 4.2.3. We use the notation $\mathcal{E}$ to denote a collection of excursions which have

the property that for any (i,k) and (i,j) in $\mathcal{E}$ and which are not contained in the same excursion, $(i,k) \not\prec (i,j)$. We refer to such a set of excursions $\mathcal{E}$ as an *antichain*. Next, we let $F^{\preceq}(\mathcal{E})$ denote the collection:

$$F^{\preceq}(\mathcal{E}) = \bigcup_{[i,e^i]\in\mathcal{E}} \{(i,k) : i \in I, k \preceq j\}.$$

We let $F^{\prec}(\mathcal{E})$ denote the collection:

$$F^{\prec}(\mathcal{E}) = \bigcup_{[i,e^i]\in\mathcal{E}} \{(i,k) : i \in I, k \prec j\}.$$

We also adopt the notation, that for each subset of classes X:

$$\lambda m(X) = \sum_{(i,k)\in X} \lambda_k^i m_k^i.$$

The primary result of this chapter is the following:

Theorem 4.2.2. *A two-station fluid SBN is globally stable iff $\rho_A < 1$, $\rho_B < 1$ and for each antichain $\mathcal{E}$ and each separating set S, we have*

$$\frac{\lambda m(V_A(S) \setminus F_A^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_A^{\prec}(\mathcal{E}))} + \frac{\lambda m(V_B(S) \setminus F_B^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E}))} < 1, \tag{4.2.5}$$

We will refer to these conditions as virtual workload conditions. If such a virtual condition is violated for set of classes, we call this set of classes a *virtual bottleneck*. We will comment on the applicability of Theorem 4.2.2 at the end of Chapter 5

4.3 The Lyapunov function

In this section, we introduce the class of piecewise linear Lyapunov functions that we will use to prove the sufficiency of the conditions in Theorem 4.2.2. This Lyapunov function is a natural extension of the function used by Dai and VandeVate [16] and we will follow their methodology closely, making appropriate alterations where needed.

First, we let $Z_k^i(t)$ denote the amount of fluid that has entered the network by time t and will receive class (i,k) service eventually:

$$Z_k^i(t) = Z_k^i(0) + \lambda_k^i t - D_k^i(t) = \sum_{j \preceq k} \left(\frac{\lambda_k^i}{\lambda_j^i} \right) Q_j^i(t)$$

where $D_k^i(t)$ is the amount of class (i,k) fluid that has been serviced in the interval $[0,t]$. *Remark*: In an OMQN we would define $Z(t) = (I - P')^{-1}Q(t)$. For an SBN it is possible to label the classes such that $(I - P')^{-1}$ is lower triangular, allowing us to write $Z(t)$ in the more explicit form above. In fact, most of our analysis relies on this special structure and it is for this reason we restrict our study in this chapter to SBN's.

Now, for a given x we define:

$$G(x,t) = \max\{G_A(x,t), G_B(x,t)\}$$

where

$$G_A(x,t) = \sum_{i\in I}\sum_{k\in A^i} x_k^i Z_k^i(t) \tag{4.3.1}$$

$$G_B(x,t) = \sum_{i\in I}\sum_{k\in B^i} x_k^i Z_k^i(t) \tag{4.3.2}$$

If we set $x_k^i := m_k^i$ for each class (i,k), then we can interpret $G_A(m,t)$ as the total workload for station A in the system at time t. Thus, in general $G_A(x,t)$ is the total weighted workload in the system at time t for station A.

We would like to check for which values of x_k^i the $G(x,t)$ as defined above will be a Lyapunov function. Theorem 4.3.1, originally proven in Dai and Weiss [17], simplifies this analysis.

Theorem 4.3.1. *Suppose $G(\cdot)$ satisfies the following:*

$$G_A(x,t) \le G_B(x,t) \quad \textit{whenever} \quad \sum_{i\in I}\sum_{k\in A^i} Q_k^i(t) = 0 \tag{4.3.3}$$

$$G_B(x,t) \le G_A(x,t) \quad \textit{whenever} \quad \sum_{i\in I}\sum_{k\in B^i} Q_k^i(t) = 0 \tag{4.3.4}$$

$$\frac{\partial G_A(x,t)}{\partial t} \le -\epsilon \quad \textit{whenever} \quad \sum_{i\in I}\sum_{k\in A^i} Q_k^i(t) > 0 \tag{4.3.5}$$

$$\frac{\partial G_B(x,t)}{\partial t} \le -\epsilon \quad \textit{whenever} \quad \sum_{i\in I}\sum_{k\in B^i} Q_k^i(t) > 0 \tag{4.3.6}$$

where the derivative conditions hold only at regular points t, then $G(x,t)$ is a Lyapunov function.

We will refer to (4.3.3)–(4.3.6) as the Dai-Weiss conditions.

4.4 A Linear Programming Formulation

Now, modifying the development in Dai and VandeVate [16] we can transform the problem of finding weights x for which $G(\cdot)$ will satisfy the Dai-Weiss conditions into checking the feasibility of a linear programming problem.

First, we note that when

$$\sum_{i\in I}\sum_{k\in A^i} Q_k^i(t) = 0 \tag{4.4.1}$$

G_A reduces to:

$$\sum_{i\in I}\sum_{k\in A^i}\left[x_k^i \sum_{j\in B^i, j\prec k}\left(\frac{\lambda_k^i}{\lambda_j^i}\right) Q_j^i(t)\right] = \sum_{i\in I}\sum_{j\in B^i}\left[Q_j^i(t)\sum_{k\in A^i, k\succ j}\left(\frac{\lambda_k^i}{\lambda_j^i}\right) x_k^i\right]$$

and G_B becomes:

$$\sum_{i\in I}\sum_{k\in B^i}\left[x_k^i \sum_{j\in B^i, j\preceq k}\left(\frac{\lambda_k^i}{\lambda_j^i}\right) Q_j^i(t)\right] = \sum_{i\in I}\sum_{j\in B^i}\left[Q_j^i(t)\sum_{k\in B^i, k\succeq j}\left(\frac{\lambda_k^i}{\lambda_j^i}\right) x_k^i\right]$$

So, $G_A(x,t) \le G_B(x,t)$ for all $Q(\cdot) \ge 0$ satisfying (4.4.1) iff:

$$\sum_{k\in A^i, k\succ j}\left(\frac{\lambda_k^i}{\lambda_j^i}\right) x_k^i \le \sum_{k\in B^i, k\succeq j}\left(\frac{\lambda_k^i}{\lambda_j^i}\right) x_k^i$$

for each $i \in I$ and $\ell \in B^i$. Multiplying through by λ^i_j we obtain:

$$\sum_{k \in A^i, k \succ j} \lambda^i_k x^i_k \leq \sum_{k \in B^i, k \succeq j} \lambda^i_k x^i_k \tag{4.4.2}$$

again for each $i \in I$ and $j \in B^i$. Since we only consider non-negative x, it is sufficient to require (4.4.2) hold only for $j = \ell[i, e]$ for each $i \in I$ and each possible excursion $[i, e]$ at station B.

Similarly, enforcing (4.3.4) yields:

$$\sum_{k \in B^i, k \succ j} \lambda^i_k x^i_k \leq \sum_{k \in A^i, k \succeq j} \lambda^i_k x^i_k$$

for $j = \ell[i, e]$ for each $i \in I$ and for each possible excursion $[i, e]$ at station A. Note that above we used the fact that each possible excursion can be identified by a unique last class.

We next transform the Dai–Weiss derivative conditions (4.3.5) and (4.3.6) into inequalities involving x. When

$$\sum_{i \in I} \sum_{k \in A^i} Q^i_k(t) > 0$$

the non-idling conditions requires:

$$\sum_{i \in I} \sum_{k \in A^i} m^i_k \dot{D}^i_k(t) = 1 \tag{4.4.3}$$

We also have,

$$\begin{aligned}
\dot{G}_A(t) &= \sum_{i\in I}\sum_{k\in A^i} x_k^i \dot{Z}_k^i(t) \\
&= \sum_{i\in I}\sum_{k\in A^i} x_k^i \left(\lambda_k^i - \dot{D}_k^i(t)\right) \\
&= \sum_{i\in I}\sum_{k\in A^i} \lambda_k^i x_k^i - \sum_{i\in I}\sum_{k\in A^i} x_k^i \dot{D}_k^i(t)
\end{aligned}$$

So, $\dot{G}_A(x,t) \leq -\epsilon$ for each $D(t)$ satisfying (4.4.3) iff:

$$\sum_{i\in I}\sum_{k\in A^i} \lambda_k^i x_k^i + \epsilon \leq x_j^i/m_j^i$$

for each $i \in I$ and $\ell \in A^i$.

Similarly, (4.3.6) will be satisfied iff

$$\sum_{i\in I}\sum_{k\in B^i} \lambda_k^i x_k^i + \epsilon \leq x_j^i/m_j^i$$

for each $i \in I$ and $j \in B^i$.

We now combine the four sets of inequalities derived above to yield the following LP:

$$\text{maximize } \epsilon \tag{4.4.4}$$

subject to:

$$\sum_{k\in A^i, k\succ \ell[i,e]} \lambda_k^i x_k^i - \sum_{k\in B^i, k\succeq \ell[i,e]} \lambda_k^i x_k^i \leq 0 \text{ each } i\in I \text{ and } [i,e]\in E_B^i \tag{4.4.5}$$

$$\sum_{k\in B^i, k\succ \ell[i,e]} \lambda_k^i x_k^i - \sum_{k\in A^i, k\succeq \ell[i,e]} \lambda_k^i x_k^i \leq 0 \text{ each } i\in I \text{ and } [i,e]\in E_A^i \tag{4.4.6}$$

$$\sum_{i\in I}\sum_{k\in A^i} \lambda_k^i x_k^i - x_j^i/m_j^i + \epsilon \leq 0 \text{ for } i\in I, j\in A^i \tag{4.4.7}$$

$$\sum_{i\in I}\sum_{k\in B^i} \lambda_k^i x_k^i - x_j^i/m_j^i + \epsilon \leq 0 \text{ for } i\in I, j\in B^i \tag{4.4.8}$$

$$x, \epsilon \geq 0 \tag{4.4.9}$$

The LP above is easier to work with if set $\lambda_k^i x_k^i = y_k^i$ to obtain the transformed linear program (TLP):

$$\text{maximize } \epsilon \tag{4.4.10}$$

subject to:

$$\sum_{k\in A^i, k\succ \ell[i,e]} y_k^i - \sum_{k\in B^i, k\succeq \ell[i,e]} y_k^i \leq 0 \text{ each } i\in I \text{ and excursion } [i,e]\in E_B^i \tag{4.4.11}$$

$$\sum_{k\in B^i, k\succ \ell[i,e]} y_k^i - \sum_{k\in A^i, k\succeq \ell[i,e]} y_k^i \leq 0 \text{ each } i\in I \text{ and excursion } [i,e]\in E_A^i \tag{4.4.12}$$

$$\sum_{i\in I}\sum_{k\in A^i} y_k^i - y_j^i/(\lambda_j^i m_j^i) + \epsilon \leq 0 \text{ for } i\in I, j\in A^i \tag{4.4.13}$$

$$\sum_{i\in I}\sum_{k\in B^i} y_k^i - y_j^i/(\lambda_j^i m_j^i) + \epsilon \leq 0 \text{ for } i\in I, j\in B^i \tag{4.4.14}$$

$$y, \epsilon \geq 0 \tag{4.4.15}$$

Remark: If a strictly branching network has yield loss only (i.e. for all (i,k), $p_{(i,k),(i,j)} > 0$ for only one class (i,j)) then the above TLP has the same form as the Dai-Vande Vate LP and thus Theorem 4.2.2 follows immediately from their results.

The next proposition follows as an immediate generalization of Proposition 4.1 in Dai and Vande Vate [16]:

Proposition 4.4.1. *If the TLP given by (4.4.10)–(4.4.14) has unbounded objective values, then $G(x,t)$ is a Lyapunov function for each solution $(x,\epsilon) > 0$.*

4.5 A Network Flows Formulation

The feasibility analysis of the TLP (4.4.10)–(4.4.14) can be achieved more easily by transforming it into a parametric network flow problem. As per the Dai-Vande Vate framework, we can assume

$$\sum_{i\in I}\sum_{k\in A^i} y_k^i + \epsilon = 1 \tag{4.5.1}$$

$$\sum_{i\in I}\sum_{k\in B^i} y_k^i + \epsilon = \beta \tag{4.5.2}$$

Then, (4.4.13) and (4.4.14) become:

$$y_k^i \geq \lambda_k^i m_k^i \text{ for } i \in I \text{ and } k \in A^i \tag{4.5.3}$$

$$y_k^i \geq \beta\lambda_k^i m_k^i \text{ for } i \in I \text{ and } k \in B^i \tag{4.5.4}$$

Next, we add slacks $s = (s_e^i)$ and write (4.4.11)–(4.4.12) as:

$$\sum_{k\in A^i, k\succ\ell[i,e]} y_k^i - \sum_{k\in B^i, k\succeq\ell[i,e]} y_k^i + s_e^i = 0 \text{ each } i \in I \text{ and excursion } [i,e] \in E_B^i \tag{4.5.5}$$

$$\sum_{k\in B^i, k\succ\ell[i,e]} y_k^i - \sum_{k\in A^i, k\succeq\ell[i,e]} y_k^i + s_e^i = 0 \text{ each } i \in I \text{ and excursion } [i,e] \in E_A^i \tag{4.5.6}$$

For each possible excursion $[i, e]$ at Station B, we take its corresponding equation of the form of (4.5.5) and add to it all equations of the type (4.5.6) that correspond to successor excursions in $\{i, e_A^+\}$ and subtract all equations of the type (4.5.5) that correspond to

successor excursions in $\{i, e_B^+\}$. We then multiply by -1 to obtain:

$$-\sum_{k\in f\{i,e_A^+\}} y_k^i + \sum_{k\in f\{i,e_B^+\}} y_k^i + y_{\ell[i,e]}^i - s_e^i + \sum_{k\in\{i,e_B^+\}} s_k^i - \sum_{k\in\{i,e_A^+\}} s_k^i = 0$$

Similarly, we obtain

$$\sum_{k\in f\{i,e_B^+\}} y_k^i - \sum_{k\in f\{i,e_A^+\}} y_k^i - y_{\ell[i,e]}^i + s_e^i - \sum_{k\in\{i,e_A^+\}} s_k^i + \sum_{k\in\{i,e_B^+\}} s_k^i = 0$$

for excursions $[i, e]$ at station A.

These transformations give us the following network flow problem:

$$\text{maximize } \epsilon \tag{4.5.7}$$

subject to:

$$\sum_{k\in f\{i,e_B^+\}} y_k^i - \sum_{k\in f\{i,e_A^+\}} y_k^i - y_{\ell[i,e]}^i + s_e^i - \sum_{k\in\{i,e_A^+\}} s_k^i + \sum_{k\in\{i,e_B^+\}} s_k^i = 0 \text{ for } [i,e] \in E_A^i \tag{4.5.8}$$

$$-\sum_{k\in f\{i,e_A^+\}} y_k^i + \sum_{k\in f\{i,e_B^+\}} y_k^i + y_{\ell[i,e]}^i - s_e^i + \sum_{k\in\{i,e_B^+\}} s_k^i - \sum_{k\in\{i,e_A^+\}} s_k^i = 0 \text{ for } [i,e] \in E_B^i \tag{4.5.9}$$

$$\sum_{i\in I}\sum_{k\in A^i} y_k^i + \epsilon = 1 \tag{4.5.10}$$

$$-\sum_{i\in I}\sum_{k\in B^i} y_k^i - \epsilon = -\beta \tag{4.5.11}$$

$$y_k^i \geq \lambda_k^i m_k^i \text{ for } i \in I \text{ and } k \in A^i \tag{4.5.12}$$

$$y_k^i \geq \beta\lambda_k^i m_k^i \text{ for } i \in I \text{ and } k \in B^i \tag{4.5.13}$$

$$y, s, \epsilon \geq 0 \tag{4.5.14}$$

The network flow problem above can be characterized as follows. It has:

- A node for each possible excursion $[i, e]$ corresponding to (4.5.8) and (4.5.9).
- A node for station A and station B corresponding to (4.5.10) and (4.5.11).
- A node called *the root* corresponding to the redundant constraint

$$\sum_{i\in I}\left(\sum_{k\in B^i\cap f[1,i]} y_k^i - \sum_{k\in A^i\cap f[1,i]} y_k^i\right) + \sum_{i\in I:\ell[1,i]\in B^i} s_1^i - \sum_{i\in I:\ell[1,i]\in A^i} s_1^i = \beta - 1$$

  obtained by adding (4.5.8)–(4.5.11) and multiplying by -1.

The edges in the network are the following:

1. An edge from the node for station A to the node for possible excursion $[i, e]$ at station A. This corresponds to the variable $y^i_{\ell[i,e]}$ and has lower bound $\lambda^i_{\ell[i,e]} m^i_{\ell[i,e]}$.
2. An edge from the node for possible excursion $[i, e]$ at station B to the node for station B. This corresponds to the variable $y^i_{\ell[i,e]}$ and has lower bound $\beta\lambda^i_{\ell[i,e]} m^i_{\ell[i,e]}$.

3. An edge from the node for station A to the node for possible excursion $[i, e]$ at station B for each class (i, k) in $f\{i, e_A^+\}$. These edges correspond to the variables y_k^i for the classes in $f\{i, e_A^+\}$. The edge for class (i, k) has lower bound $\lambda_k^i m_k^i$.

4. An edge into the node for station B from the node for possible excursion $[i, e]$ at station B for each class (i, k) in $f\{i, e_B^+\}$. These edges correspond to the variables y_k^i for the classes in $f\{i, e_B^+\}$. The edge for class (i, k) has lower bound $\beta\lambda_k^i m_k^i$.

5. An edge from the node for possible excursion $[i, e]$ at station A to the node for station B for each class (i, k) in $f\{i, e_B^+\}$. These edges correspond to the variables y_k^i for the classes in $f\{i, e_B^+\}$. The edge for class (i, k) has lower bound $\beta\lambda_k^i m_k^i$.

6. An edge into the node for possible excursion $[i, e]$ at station A from the node for station A for each class (i, k) in $f\{i, e_A^+\}$. These edges correspond to the variables y_k^i for the classes in $f\{i, e_A^+\}$. The edge for class (i, k) has lower bound $\lambda_k^i m_k^i$.

7. An edge from the node for station A to the root for each class (i, k) in $f[1, i]$ served at station A. These edges correspond to the variables y_k^i for the classes in $f[1, i]$ served at station A. The edge for class (i, k) has lower bound $\lambda_k^i m_k^i$.

8. An edge from the root to the node for station B for each class (i, k) in $f[1, i]$ served at station B. These edges correspond to the variables y_k^i for the classes in $f[1, i]$ served at station B. The edge for class (i, k) has lower bound $\beta\lambda_k^i m_k^i$.

9. An edge from the node for excursion $[1, i]$ at station A to the root. This edge corresponds to the variable s_1^i and has lower bound 0.

10. An edge from the root to the node for excursion $[1, i]$ at station B. This edge corresponds to the variable s_1^i and has lower bound 0.

11. An edge from the node for each possible excursion $[i, e]$ at station A to the node for the preceding excursion (which is unique) at station B (if $[i, e-1] \in B_i$). These edges correspond to the variables s_e^i and have lower bounds 0.

12. An edge from the node for each possible excursion $[i, e]$ at station A to the node for the preceding excursion (which is unique) at station A (if $[i, e-1] \in A_i$). These edges correspond to the variables s_e^i and have lower bounds 0.

13. An edge to the node for each possible excursion $[i, e]$ at station B from the node for the preceding excursion (which is unique) at station A (if $[i, e-1] \in A_i$). These edges correspond to the variables s_e^i and have lower bounds 0.

14. An edge to the node for each possible excursion $[i, e]$ at station B from the node for the preceding excursion (which is unique) at station B (if $[i, e-1] \in B_i$). These edges correspond to the variables s_e^i and have lower bounds 0.

15. An edge from the node for station A to the node for station B. This edge corresponds to the variable ϵ and has lower bound 0.

As in Dai–Vande Vate, we convert our problem into a Minimum Flow Problem. We retain some other Dai–Vande Vate conventions. Given an A, B-cut (L, R), we let L_A denote the excursions in L that are served at Station A and L_B denote those served at Station B. Similarly, we let R_A denote the excursions in R served at Station A and R_B denote those at Station B.

We refer to an A, B-cut with the root in L as an L-cut. An A, B-cut with the root in R is a R-cut. Note that since the upper bound on each edge is infinite, an A, B-cut (L, R) in this network has finite capacity if and only if no edge extends from a node in R to a node in L, i.e., if and only if (L, R) satisfies:

1. If $[i, e] \in L_A$, then possible excursions $\{i, e_A^+\}$ are in L_A,

2. If $[i, e] \in R_B$, then possible excursions $\{i, e_B^+\}$ are in R_B,

3. If $[i, e] \in L_B$, then possible excursions $\{i, e_A^+\}$ are in L_A,

4. If $[i, e] \in R_A$, then possible excursions $\{i, e_B^+\}$ are in R_B,

5. If (L, R) is an R-cut, then $[i, 1] \cap L_B = \emptyset$ for each type i, and

6. If (L, R) is an L-cut, then $[i, 1] \cap R_A = \emptyset$ for each type i.

Otherwise, the capacity of the cut is $-\infty$. Thus, we have the following lemma, which allows us to speak in terms of separating sets rather than cuts.

Lemma 4.5.1. *An L-cut (L, R) has finite capacity only if $L_B \cup R_A$ is an A-strictly separating set. Similarly, an R-cut (L, R) has finite capacity only if $L_B \cup R_A$ is a B-strictly separating set.*

We can proceed as in Dai and Vande Vate to obtain the theorem below, which expresses the stability of the network in terms of the "cut conditions."

Theorem 4.5.2. *A two-station fluid network with service times m and effective arrival rates $\lambda = (\lambda_k)_{k \in K}$ satisfying the nominal workload conditions is globally stable if for each non-trivial A-strictly separating set S' and non-trivial B-strictly separating set S,*

$$\frac{\lambda m(V_A(S'))}{1 - \lambda m(V_B(S'))} < \frac{1 - \lambda m(V_A(S))}{\lambda m(V_B(S))}. \tag{4.5.15}$$

We call the conditions in (4.5.15) the *cut conditions*.

4.6 Equivalence

In this section we endeavor to prove the following lemma that, in conjunction with Theorem 4.5.2, proves the sufficiency part of Theorem 4.2.2.

Lemma 4.6.1. *If the arrival rates and service times satisfy the virtual workload conditions, then they also satisfy the cut conditions.*

Proof. Each cut condition is defined by a pair of non-trivial separating sets, S and S'. We show that the cut conditions induced by the pair (S, S') is implied by a pair of virtual workload conditions.

We first need to chose an antichain $\mathcal{E}$, which is induced by the pair (S, S'). For each type i, we choose a set $\mathcal{E}^i$ that will usually contain several elements, due to the branching nature of our networks. In the multitype case, $\mathcal{E}^i$ contains only one element for each $i \in I$.

We now specify how the set $\mathcal{E}^i$ is chosen. For each type, consider the SNE. Any class that is in S or S' in the original network remains in these sets in the SNE. For each type in the SNE, we now have a number of sub-types, one for each possible route in the original network. For each sub-type j we chose an index $\hat{e}^i_j$ according to the Dai-Vande Vate rules, i.e. we let $\hat{e}^i_j \geq 1$ be the largest index r such that

1. $\{[i, e] \in E^i_A : e \preceq r\} \cap S' = \emptyset$,
2. $\{[i, e] \in E^i_B : e \prec r\} \subseteq S'$ and
3. $[i, r] \notin S'$.

We chose e^i_j similarly, letting e^i_j be the largest index r such that

4. $\{[i, e] \in E^i_A : e \prec r\} \subseteq S$,
5. $\{[i, e] \in E^i_B : e \preceq r\} \cap S = \emptyset$ and

6. $[i, r] \notin S$.

Now, for each (i, j), we set $r_j^i = \min\{\hat{e}_j^i, e_j^i\}$. We then let $\mathcal{E}^i = \cup_j r_j^i$. Note that the r_j^i's need not be distinct. We set $\mathcal{E} = \cup_{i \in I} \mathcal{E}^i$.

We note that the set $\mathcal{E}$ is an antichain and has the following properties:

a) $\mathcal{E} \cap (S \cup S') = \emptyset$

b) $\left(\bigcup_{r \in \mathcal{E}} \{[i, e] \in E_A^i : i \in I, e \preceq r\} \right) \cap S' = \emptyset$

c) $\bigcup_{r \in \mathcal{E}} \{[i, e] \in E_B^i : i \in I, e \prec r\} \subseteq S'$

d) $\left(\bigcup_{r \in \mathcal{E}} \{[i, e] \in E_B^i : i \in I, e \preceq r\} \right) \cap S = \emptyset$

e) $\bigcup_{r \in \mathcal{E}} \{[i, e] \in E_A^i : i \in I, e \prec r\} \subseteq S$

These properties allow us to go through analogous algebra to Dai and Vande Vate to obtain the cut condition from the virtual workload conditions. The antichain $\mathcal{E}$ and the separating set S' induce the virtual workload condition:

$$\frac{\lambda m(V_A(S') \setminus F_A^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_A^{\prec}(\mathcal{E}))} + \frac{\lambda m(V_B(S') \setminus F_B^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E}))} < 1. \tag{4.6.1}$$

Similarly, $\mathcal{E}$ and the separating set S induce the virtual workload condition:

$$\frac{\lambda m(V_A(S) \setminus F_A^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_A^{\prec}(\mathcal{E}))} + \frac{\lambda m(V_B(S) \setminus F_B^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E}))} < 1. \tag{4.6.2}$$

We show that (4.6.1) and (4.6.2) imply the cut condition for the pair (S', S).

From (4.6.1) we have that

$$\frac{\lambda m(V_A(S') \setminus F_A^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E})) - \lambda m(V_B(S') \setminus F_B^{\preceq}(\mathcal{E}))} < \frac{1 - \lambda m(F_A^{\prec}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E}))}$$

and from (4.6.2) we have that

$$\frac{1 - \lambda m(F_A^{\prec}(\mathcal{E})) - \lambda m(V_A(S) \setminus F_A^{\preceq}(\mathcal{E}))}{\lambda m(V_B(S) \setminus F_B^{\preceq}(\mathcal{E}))} > \frac{1 - \lambda m(F_A^{\prec}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E}))}.$$

Thus,

$$\frac{1 - \lambda m(F_A^{\prec}(\mathcal{E})) - \lambda m(V_A(S) \setminus F_A^{\preceq}(\mathcal{E}))}{\lambda m(V_B(S) \setminus F_B^{\preceq}(\mathcal{E}))} > \frac{\lambda m(V_A(S') \setminus F_A^{\preceq}(\mathcal{E}))}{1 - \lambda m(F_B^{\prec}(\mathcal{E})) - \lambda m(V_B(S') \setminus F_B^{\preceq}(\mathcal{E}))}. \quad (4.6.3)$$

Now, (a) implies

$$\lambda m(V_A(S) \setminus F_A^{\preceq}(\mathcal{E})) + \lambda m(F_A^{\prec}(\mathcal{E})) \geq \lambda m(V_A(S)).$$

Next we note that we have the following

1. $\left(\cup_{[i,e] \in S} E[i,e]\right) \cap F_B^{\preceq}(\mathcal{E}) = \emptyset$, directly from (d)
2. $\left(\cup_{[i,e] \in E^i \setminus S} \hat{f}\{i,e\}\right) \cap F_B^{\preceq}(\mathcal{E}) = \emptyset$, directly from (e)
3. $\left(\cup_{[i,e] \in S} \overline{f}\{i,e\}\right) \cap F_B^{\preceq}(\mathcal{E}) = \emptyset$, from 1.

Recalling the definition of $V(S)$, the above imply that

$$V_B(S) \cap F_B^{\preceq}(\mathcal{E}) = \emptyset$$

Thus, we have

$$\lambda m(V_B(S) \setminus F_B^{\preceq}(\mathcal{E})) = \lambda m(V_B(S)).$$

Once again from (a) we have

$$\lambda m(V_B(S') \setminus F_B^{\preceq}(\mathcal{E})) + \lambda m(F_B^{\preceq}(\mathcal{E})) \geq \lambda m(V_B(S')).$$

Using a similar argument as before, we have that (b) and (c) imply,

$$\lambda m(V_A(S') \setminus F_A^{\preceq}(\mathcal{E})) = \lambda m(V_A(S')).$$

Thus, (4.6.3) implies that

$$\frac{1 - \lambda m(V_A(S))}{\lambda m(V_B(S))} > \frac{\lambda m(V_A(S'))}{1 - \lambda m(V_B(S'))},$$

which is exactly the cut condition for the pair S' and S. □

4.7 Necessity

We first claim that any ACTN can be equivalently relabeled as an SBN by adding a finite number of class labels. For example, if more than one class feeds a particular buffer, we

simply divide the incoming fluid into different buffers, depending on the buffer in which the fluid was last processed. Once the network has been relabeled, we note that any allocation process that was feasible in the original network is feasible in the relabeled network and vice versa. The main point to note here is that in the ACTN, the set of non-idling policies include those which determine processing priority at a buffer based on the processing history of the fluids in that buffer. Thus any allocation policy implemented in the SBN can also be implemented in the ACTN. In terms of global stability, the ACTN and corresponding SBN are equivalent. Hence, Theorem 4.2.2 provides necessary and sufficient conditions for global stability of any ACTN, after performing the appropriate transformation to an SBN.

The sufficiency of the conditions in Theorem 4.2.2 was proved in the last several sections. We now turn our attention to proving the necessity via the next few lemmas. The necessity essentially follows from Dai and VandeVate's arguments with minor adjustments.

The first lemma we need is Lemma 7.1 from Dai and Vande Vate [16]:

Lemma 4.7.1. *Let C be a set of classes such that*

$$\lambda m(C) \geq 1.$$

Each non-idling fluid solution $(Q(\cdot), T(\cdot))$ satisfying

$$\sum_{(i,k)\in C} \dot{T}(t) \leq 1 \tag{4.7.1}$$

for each regular point t is unstable.

Proof. The proof for the class of networks we consider is analogous to that given in Dai and

Vande Vate. □

The set C in Lemma 4.7.1 represents classes that form a virtual station. Under an appropriate static buffer priority policy, only one of the classes in C may be served at any time.

Next, we adapt another lemma from Dai and Vande Vate [16] to networks with proportional routing. This lemma enables us to consider the effect of "push-starting" some set classes that occur early in the route of a fluid type.

We consider a collection of excursions $\mathcal{E}$ that partitions the classes into those of $F^{\prec}(\mathcal{E})$ and the remainder that we denote as $R(\mathcal{E})$.

Let

$$\tilde{m}_k^i = m_k^i / \big(1 - \lambda m(F_A^{\prec}(\mathcal{E}))\big) \text{ for } (i,k) \in R_A(\mathcal{E}), \tag{4.7.2}$$

$$\tilde{m}_k^i = m_k^i / \big(1 - \lambda m(F_B^{\prec}(\mathcal{E}))\big) \text{ for } (i,k) \in R_B(\mathcal{E}). \tag{4.7.3}$$

Consider the induced fluid model on the classes of $R(\mathcal{E})$:

$$Q_k^i(t) = Q_k^i(0) + \tilde{\mu}_{k-1}^i T_{k-1}^i(t) - \tilde{\mu}_k^i T_k^i(t) \geq 0, \quad t \geq 0, \ (i,k) \in R(\mathcal{E}), \tag{4.7.4}$$

$$T_k^i(0) = 0 \text{ and } T_k^i(\cdot) \text{ is nondecreasing}, \quad (i,k) \in R(\mathcal{E}), \tag{4.7.5}$$

$$t - \sum_{(i,k)\in R_A(\mathcal{E})} T_k^i(t) \text{ is nondecreasing}, \tag{4.7.6}$$

$$t - \sum_{(i,k)\in R_B(\mathcal{E})} T_k^i(t) \text{ is nondecreasing}, \tag{4.7.7}$$

$$\sum_{(i,k)\in R_A(\mathcal{E})} \dot{T}_k^i(t) = 1 \text{ whenever } \sum_{(i,k)\in R_A(\mathcal{E})} Q_k^i(t) > 0 \text{ and } t \text{ is regular}, \tag{4.7.8}$$

$$\sum_{(i,k)\in R_B(\mathcal{E})} \dot{T}_k^i(t) = 1 \text{ whenever } \sum_{(i,k)\in R_B(\mathcal{E})} Q_k^i(t) > 0 \text{ and } t \text{ is regular}, \tag{4.7.9}$$

where, $\tilde{\mu}_k^i = 1/\tilde{m}_k^i$ for $(i,k) \in R(\mathcal{E})$. For each type $i \in I$, we let $\tilde{\mu}_{\ell[i,e_i-1]}^i = \lambda_i$ and $T_{\ell[i,e_i-1]}^i(t) = t$ to model the arrivals to the induced fluid network. Note that in the induced network the effective arrival rates to the remaining classes are the same as in the original network.

Lemma 4.7.2. *If the fluid model (4.7.4)–(4.7.9) is unstable, then the fluid model for the full fluid network is unstable.*

Proof. The proof of this lemma is again analogous to the proof of Lemma 7.2 in Dai and Vande Vate [16]. □

Necessity Proof of Theorem 4.2.2. With Lemma 4.7.2 in hand, it is sufficient to show that

if the virtual station $V(S)$ corresponding to some strictly separating set S satisfies

$$\lambda m(V(S)) \geq 1, \tag{4.7.10}$$

then there is an unstable fluid solution. The set of classes $V(S)$ is a pseudostation in any corresponding queueing network and so there exists a static buffer priority policy under which no two classes in $V(S)$ can be served simultaneously. A description of pseudostations and details on the construction of such a priority policy are described in the next chapter. Thus, if we examine fluid limits $(Q(\cdot), T(\cdot))$ of the type defined in Dai [13], we see that they are fluid solutions which satisfy (4.7.1). By Lemma 4.7.1, the fluid network is not globally stable. This completes our necessity proof and thus the proof of Theorem 4.2.2. □

Of course, the next natural step is to examine networks with three or more stations. In the next chapter we will see that the virtual station conditions can be shown to be necessary for multi-station networks. In Chapter 6 we will examine the global stability region of a three-station network in great depth.

Chapter 5

Necessary Conditions for Global Stability

5.1 Introduction

In this chapter we derive necessary conditions for global stability of multi-station stochastic MTN's. The conditions are given explicitly in terms of the average service and arrival rates of the network. We show that if the conditions are not satisfied, then all fluid solutions of the corresponding fluid network are positive for all $t > 0$. This implies that the fluid model and the queueing model are unstable and yields a partial extension of the results obtained in Chapter 4.

Since we are dealing only with MTN's we will make slight alterations in our notation to simplify expressions. We recall that in an MTN, we can think of the network as serving a set $I = \{1, \ldots, N\}$ of N different types of customers. In an MTN, the route for a given customer is deterministic but arbitrary, that is each customer type may follow a different route. The visits a customer of type i makes are numbered from 1 to c_i. As in the previous chapter, we refer to type i customers during visit k (either waiting or being served) as *class* (i, k) customers and we let A denote the set of all classes in the queueing network. For a subset C of A, the class of C-priority disciplines consists of the preempt-resume static buffer priority policies that give highest priority to classes in C and lower priority to the remainder of the classes. In examples in which there is only one type of customer in the network, we will often refer to a class only by its second index k.

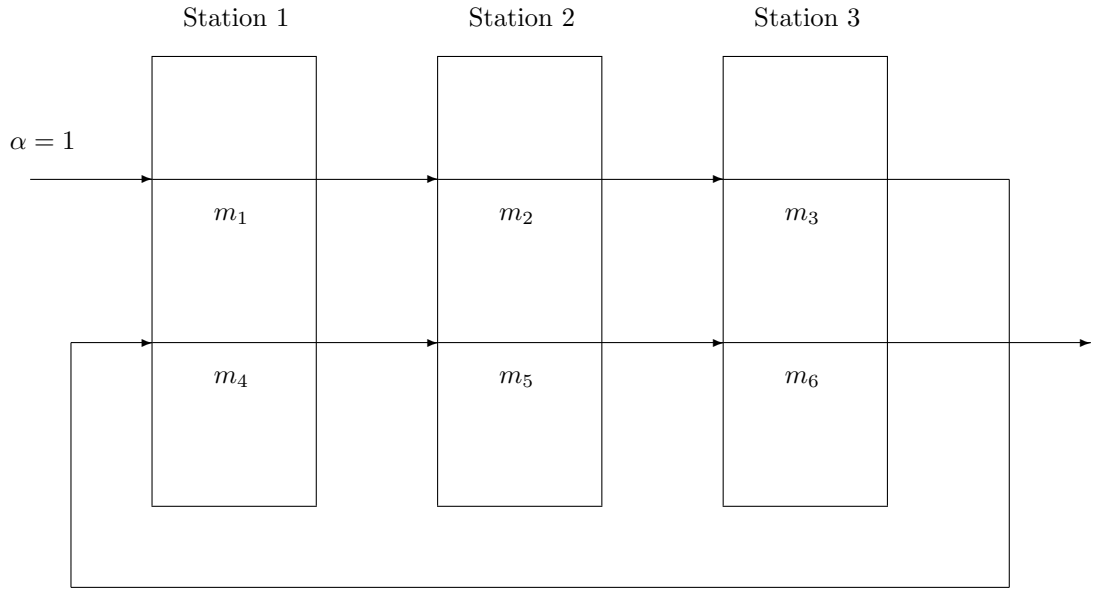

Figure 8: A six-class network

We let $A_j^i = \{k : \sigma(i,k) = j\}$, i.e. A_j^i is the set of visits a type i customer makes to station j. For example, in the network pictured in Figure 8, which has one type of customer and six classes of customers, we have $\sigma(1,3) = \sigma(1,6) = 3$, so $A_3^1 = \{(1,3),(1,6)\}$, which for our reentrant line we write as $A_3^1 = \{3,6\}$.

For clarity, we will append a type designation i to the service processes $S(\cdot)$, i.e. $S_k^i = \{S_k^i(t), t \geq 0\}$, with $S_k^i(t)$ being the cumulative number of service completions for class (i,k) if t units of time are dedicated to serving this class. We now label the exogenous arrival process for class k $S_0^i = \{S_0^i(t), t \geq 0\}$, where $S_0^i(t)$ is the cumulative number of exogenous arrivals by time t.

5.2 Pseudostations

Harrison and Nguyen [23] and Dumas [21] first observed that under some static priority queueing disciplines, certain classes in two-station queueing networks may not receive service at the same time, even though they may be served at different stations. Dai and VandeVate [15] termed such groups of classes virtual stations. They demonstrate how this phenomenon can be used to obtain necessary and sufficient conditions for the global stability of many two-station networks. In this chapter, we show how their idea of virtual stations can be extended to obtain necessary conditions for d-stations networks.

In a d-station network, we shall see that under some static buffer priority disciplines, certain sets of classes form *pseudostations*. We will show that the classes in such a pseudostation cannot all receive service at the same time, even though they may be served at different stations.

The following proposition characterizes which sets of classes form pseudostations, although we will give a more restrictive definition of pseudostations later on.

Proposition 5.2.1. *Let C be a set of classes in an initially empty d-station open multitype queueing network and suppose C satisfies:*

1. *For each type i, class $(i,1) \notin C$,*
2. *If class $(i,k) \in C$ and $\sigma(i,k-1) = \sigma(i,k)$, then class $(i,k-1) \in C$*
3. *If class $(i,k) \in C$ and $\sigma(i,k-1) \neq \sigma(i,k)$, then class $(i,k-1) \notin C$ and $\sigma(i,k-1) \in \sigma(C)$*

Let C_j be the classes of C served at station j. Under any preempt-resume static buffer

priority queueing discipline that gives higher priority at station j to the classes of C_j,

$$\prod_{j\in\sigma(C)}\left(\sum_{(i,k)\in C_j} Q_k^i(t)\right) = 0 \tag{5.2.1}$$

for all $t \geq 0$. Where we define,

$$\sigma(C) = \{j : \sigma(i,k) = j \text{ for some } (i,k) \in C\}$$

that is $\sigma(C)$ is the set of stations that serve at least one class in C.

Proof. Proceeding by contradiction, we suppose that there is a time $t > 0$ such that

$$\prod_{j\in\sigma(C)}\left(\sum_{(i,k)\in C_j} Q_k^i(t)\right) > 0$$

and let τ be the first such time t. Then τ must coincide with the time of an event. Note that for each $j \in \sigma(C)$ there must then exist a class $(i_j^*, k_j^*) \in C_j$, such that $Q_{k_j^*}^{i_j^*}(\tau) > 0$. Recall from Section 2.1 that the queue length process is right continuous, so that $\tau = \tau_n$ coincides with the time of the nth event. This implies that the system does not undergo a state change during the time interval (τ_{n-1}, τ_n) and so

$$\prod_{j\in\sigma(C)}\left(\sum_{(i,k)\in C_j} Q_k^i(\tau_{n-1})\right) > 0.$$

By definition of τ_n, since there are no simultaneous events at τ_n, it must then be the case

that

$$\sum_{(i,k)\in C_m} Q_k^i(\tau_{n-1}) = 0 \qquad \text{for some } m \in \sigma(C) \text{ and} \tag{5.2.2}$$

$$\sum_{(i,k)\in C_j} Q_k^i(\tau_{n-1}) > 0 \qquad \forall j \in \sigma(C), j \neq m \tag{5.2.3}$$

We see that (5.2.2) implies that $Q_{k_m^*}^{i_m^*}(\tau_{n-1}) = 0$, thus the event that occurs at time τ_n must be an arrival to class (i_m^*, k_m^*) from class $(i_m^*, k_m^* - 1)$. Therefore, $Q_{k_m^*-1}^{i_m^*}(\tau_{n-1}) > 0$, which means that class $(i_m^*, k_m^* - 1) \notin C_m$ and hence $\sigma(i_m^*, k_m^* - 1) \neq m$.

Now note that we have class $(i_m^*, k_m^*) \in C$ with $\sigma(i_m^*, k_m^*) \neq \sigma(i_m^*, k_m^* - 1)$ and thus condition 3 in Proposition 5.2.1, indicates that class $(i_m^*, k_m^* - 1) \notin C$ and there exists $l \in \sigma(C), l \neq m$ such that $\sigma(i_m^*, k_m^* - 1) \in C_l$. However, from (5.2.3) we have that

$$\sum_{(i,k)\in C_l} Q_k^i(\tau_{n-1}) > 0$$

so there is a class (i_l^*, k_l^*) at station l with $Q_{k_l^*}^{i_l^*}(\tau_{n-1}) > 0$ which has higher priority than class $(i_m^*, k_m^* - 1)$. Hence, server l cannot service customers in class $(i_m^*, k_m^* - 1)$ during (τ_{n-1}, τ_n). This contradicts our earlier conclusion that the event that occurs at time τ_n is an arrival to class (i_m^*, k_m^*). □

Classes in a pseudostation do not behave exactly like Dai and VandeVate's virtual stations, since several of the classes may be serviced at the same time (if $|\sigma(C)| = 2$ then C acts like the virtual stations described in Dai and VandeVate [15]). However, not *all* the classes in a pseudostation may be served at the same time, so these sets of classes have a hidden effect on the capacity of the queueing system.

Figure 8 shows perhaps the simplest example of a pseudostation in a 3-station network. In this case, classes 2, 4 and 6 form a pseudostation. Under the static buffer priority policy that gives classes 2,4, and 6 the highest priority, a maximum of two of these classes may be serviced simultaneously if the network starts with zero customers at time zero.

It will be seen later that many of the choices for C in Proposition 5.2.1 give redundant necessary conditions. Unfortunately, not all of these redundancies can be eliminated by adding simple restrictions on the admissible sets C. We can, however, eliminate some by adapting the Dai-VandeVate virtual station definition to suit our purposes.

We now give a more precise description of classes that form pseudostations.

Definition 5.2.1. A set of classes C forms a *pseudostation* if it satisfies the following:

- No class contained in a first excursion is in C.
- If the last visit of an excursion is in C, then all the visits of that excursion are in C.
- If any first visit of an excursion is in C, then all of the first visits of that excursion are also in C.
- If any class of an excursion is in C, then the last visit of the previous excursion is not in C and the previous excursion's server is in $\sigma(C)$.

To see that it is sufficient to consider only pseudostations in obtaining necessary conditions, we introduce the concept of a vertical subset. A set C' is a *vertical subset* of another set C if, $C' \subseteq C$ and $\sigma(C) = \sigma(C')$.

Now we note that any set C satisfying the conditions in Proposition 5.2.1 is a vertical subset of a pseudostation. An examination of Theorem 5.3.1 shows that the necessary conditions generated by any pseudostation will imply those generated by a vertical subset, thus we need focus only on pseudostations.

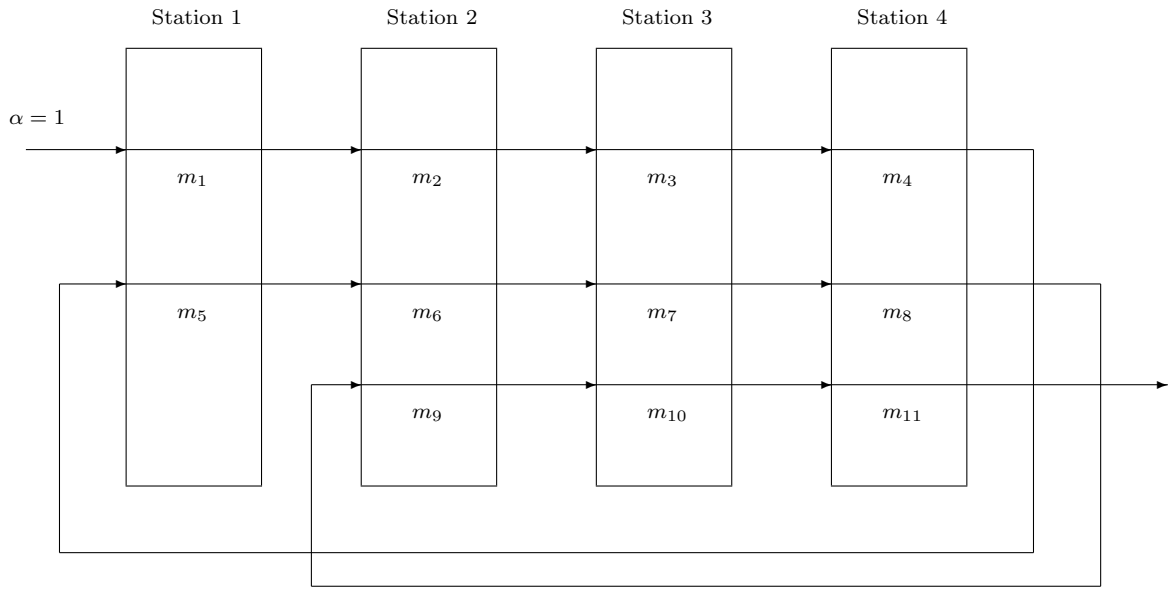

Figure 9: An eleven-class network

However, even if we restrict the sets C in Proposition 5.2.1 to those sets which form pseudostations, we may still have redundant conditions. In Figure 9, classes 7, 9, and 11 form a pseudostation, as do classes 5, 7, 9 and 11, but neither is a vertical subset of the other. From Theorem 5.3.1, classes 7, 9, and 11 give the condition

$$m_7 + m_9 + m_{11} \leq 2 \tag{5.2.4}$$

and classes 5, 7, 9, and 11 give

$$m_5 + m_7 + m_9 + m_{11} \leq 3 \tag{5.2.5}$$

Now, the usual traffic condition $\rho_1 < 1$ implies that $m_5 < 1$ so (5.2.4) is stronger than (5.2.5). Thus, in this case it is the smaller set that generates the stronger condition. It is difficult to concisely expand the definition of pseudostations to further restrict the classes C under consideration to avoid such redundant conditions.

5.3 Necessary Conditions for Stability

We now present our main result. As before, we use C to denote a subset of classes, and A to denote the set of all classes. We define $|x| = \sum |x_k|$ for a vector $x \in \mathbb{R}^K$ and for a set C, $|C|$ indicates the cardinality of C. Also, for a given set X we let

$$1_X(s) = \begin{cases} 1 & \text{if } s \in X \\ 0 & \text{otherwise} \end{cases}$$

Theorem 5.3.1. *In a d-station open multitype queueing network, if for any set of classes C satisfying:*

1. *For each type i, class $(i,1) \notin C$,*
2. *If class $(i,k) \in C$ and $\sigma(i,k-1) = \sigma(i,k)$, then class $(i,k-1) \in C$*
3. *If class $(i,k) \in C$ and $\sigma(i,k-1) \neq \sigma(i,k)$, then class $(i,k-1) \notin C$ and $\sigma(i,k-1) \in \sigma(C)$*

$$\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j}\lambda_k^i m_k^i > |\sigma(C)| - 1, \tag{5.3.1}$$

where C_j the set of classes of C served at station j, then with probability one,

$$|Q(t)| \to \infty \quad as\ t \to \infty, \tag{5.3.2}$$

if the network operates under a C-priority discipline.

Proof. Consider then a set C which satisfies 1–3 in Theorem 5.3.1 and for which (5.3.1) holds. We will show that the number of customers diverges to infinity almost surely under a C-priority discipline.

First, from the definition of $T_k^i(t)$, we have for each $j \in \sigma(C)$ that

$$\sum_{(i,k)\in C_j} T_k^i(t) \le \int_0^t 1_{\{\tau:\sum_{(i,k)\in C_j} Q_k^i(\tau)>0\}}(s)\, ds \tag{5.3.3}$$

Since (5.3.3) consists of $|\sigma(C)|$ equations (one for each station j that contains classes in C), upon summing we obtain,

$$\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} T_k^i(t) \le \int_0^t \sum_{j\in\sigma(C)} 1_{\{\tau:\sum_{(i,k)\in C_j} Q_k^i(\tau)>0\}}(s)\, ds \tag{5.3.4}$$

Now, under the queueing discipline we have imposed on the network, Proposition 5.2.1 implies

$$\sum_{j\in\sigma(C)} 1_{\{\tau:\sum_{(i,k)\in C_j} Q_k^i(\tau)>0\}}(s) \le |\sigma(C)| - 1 \quad \text{for all } s \ge 0$$

This and (5.3.4) yield

$$\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} T_k^i(t) \le (|\sigma(C)|-1)\cdot t \tag{5.3.5}$$

Taking the fluid limit (5.3.5) becomes:

$$\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} \bar{T}_k^i(t) \le (|\sigma(C)|-1)\cdot t \tag{5.3.6}$$

where $(\bar{Q}(t),\bar{T}(t))$ is a fluid limit as defined in Section 2.3. If we let $\bar{Z}(t) = (I-P')^{-1}\bar{Q}(t)$, then the fluid dynamical equations give

$$\Delta^{-1}\bar{Z}(t) = \Delta^{-1}(I-P')^{-1}\alpha t - \bar{T}(t) = \Delta^{-1}\lambda t - \bar{T}(t)$$

. This yields,

$$\begin{aligned}
\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} m_k^i \bar{Z}_k^i(t) &= \left(\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} \lambda_k^i m_k^i\right)\cdot t - \left(\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} \bar{T}_k^i(t)\right) \\
&\ge \left(\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} \lambda_k^i m_k^i\right)\cdot t - (|\sigma(C)|-1)\cdot t \\
&= \left[\sum_{j\in\sigma(C)}\sum_{(i,k)\in C_j} \lambda_k^i m_k^i - (|\sigma(C)|-1)\right]\cdot t
\end{aligned}$$

Note that we used (5.3.6) to obtain the inequality. For $t>0$, the last expression is strictly

greater than zero by assumption, so

$$\sum_{j \in \sigma(C)} \sum_{(i,k) \in C_j} m_k^i \bar{Z}(t) > 0 \text{ for all } t > 0$$

Thus $|\bar{Q}(t)| > 0$ for all $t > 0$ and the fluid model is weakly unstable. Invoking Theorem 3.2.2 gives us the desired result. □

The classes C that satisfy Theorem 5.3.1 may generate redundant conditions. As seen in Section 5.2, we can eliminate some redundancies by restricting the admissible sets C to pseudostations. However, even with redundancies eliminated, it seems that the number of necessary conditions in a d-station network will grow quickly, perhaps exponentially, with the number of classes.

With this in mind, make some brief comments about the applicability of Theorem 5.3.1, which also apply to Theorem 4.2.2. First, we note that although a full model of a typical wafer fabrication facility or other complex system may contain scores of stations and hundereds of classes, the full model is generally not needed to accurately represent the real life system, at least with regard to stability and capacity issues. Often, only five or six machine groups are highly to moderately utilized and it may be that only these groups are necessary for an accurate stability analysis.

Now, even in a reduced system, there may be a large number of groups of classes which form virtual stations. Since each such virtual station (possibly paired with the antichain $\mathcal{E}$) has a corresponding virtual workload condition, the time required to check all such conditions could grow rapidly with the size of the network, even with an automated procedure to check such conditions. However, in systems with more than a handful of stations and classes, classes with short processing times can often be safely ignored when

checking for virtual bottelnecks. Moreover, condition (4.2.5) is easily checked for any set of classes which come under suspicion as a source of trouble in the system. With these considerations, we believe that it is quite reasonable to implement efficient computational procedures to aid in the evaluation of stability and capacity issues for these systems.

5.4 A Capacity Example

In this section, we present an example of how the phenomenon of pseudostations can affect capacity calculations in a multiclass network. We again consider the reentrant line pictured in Figure 8. Now, suppose the service time vector is fixed at:

$$m = (0.1, 0.7, 0.1, 0.7, 0.1, 0.7).$$

The usual method to evaluate the capacity of this network is to check the usual traffic conditions and find the maximum sustainable input rate α. In industry, this is commonly referred to as *bottleneck analysis*. The usual traffic conditions for this network are:

$$\begin{aligned} \alpha\,(m_1 + m_4) &< 1 \\ \alpha\,(m_2 + m_5) &< 1 \\ \alpha\,(m_3 + m_6) &< 1 \end{aligned}$$

So, an upper bound on the global capacity Λ_∞, as defined in (3.1.4) would then be given by:

$$\Lambda_\infty \leq \min\left(\frac{1}{m_1 + m_4}, \frac{1}{m_2 + m_5}, \frac{1}{m_3 + m_6}\right) = 1.25$$

Essentially, this calculation considers each station individually and the capacity of the system depends upon the capacity of the slowest machine. Unfortunately, the potential effect of pseudostations may constrain the system more than the slowest machine. Theorem 5.3.1 indicates that necessary conditions for global stability are in fact:

$$\begin{aligned} \alpha\left(m_1 + m_4\right) &< 1 \\ \alpha\left(m_2 + m_5\right) &< 1 \\ \alpha\left(m_3 + m_6\right) &< 1 \\ \alpha\left(m_2 + m_4 + m_6\right) &< 2 \end{aligned}$$

Hence, a revised capacity calculation incorporating the pseudostation condition yields

$$\begin{aligned} \Lambda_\infty &\leq \min\left(\frac{1}{m_1 + m_4}, \frac{1}{m_2 + m_5}, \frac{1}{m_3 + m_6}, \frac{2}{m_2 + m_4 + m_6}\right) \\ &\approx 0.9524 \end{aligned}$$

We see that the original capacity calculation is off by at least 30 percent. In fact, the global stability region of this network is unknown and so the actual capacity may be even smaller than 0.9524. In the next chapter we further investigate the stability of the network in this example.

Chapter 6

Stability of a Three-station Fluid Network

6.1 The Fluid Network and Its Stability

In this chapter, we investigate the stability of a particular three-station fluid network. We recall that in Chapter 4 we were able to derive necessary and sufficient conditions for stability of the class of two-station fluid ACTN's. In Chapter 5, we were able to extend the necessity arguments to fluid (and discrete) networks with an arbitrary number of stations. In our analysis of the three-station network we extend the sufficiency arguments of Chapter 4 in an attempt to find necessary and sufficient stability conditions.

The three-station fluid network under consideration is depicted in Figure 10. Unless otherwise noted, all comments about fluid networks are specific to this three-station network. Also, as in Chapter 4, we again drop the "bar" notation, since all quantities in this chapter will be fluid quantities. Fluid comes to this network at the rate of α units per unit of time and is served at each station in turn starting with station 1. After processing at station 3, fluid returns to station 1 and is again served by each station in turn before it leaves the system. Thus, each unit of fluid is processed six times, twice at each station, before it leaves the system. Note that in this example $\lambda_k = \alpha$ for all six classes and thus from now on we drop the subscript and deal with the single effective arrival rate $\lambda = \alpha$. If we set $T_0(t) = t$

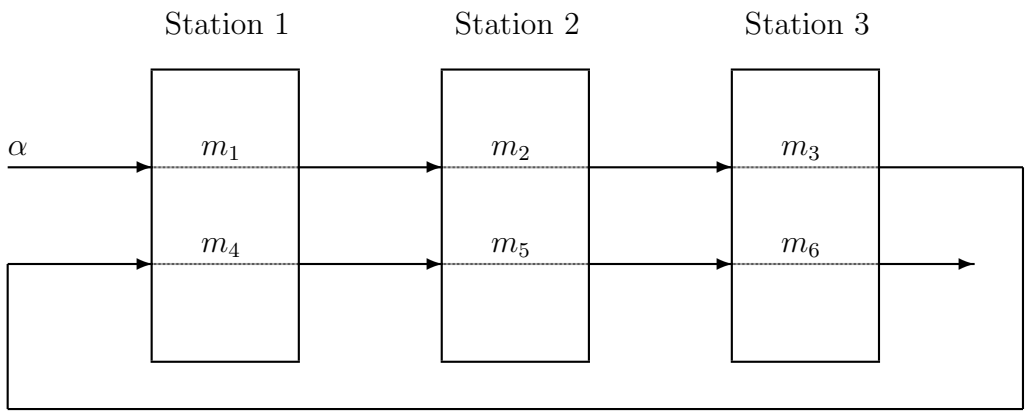

Figure 10: A three-station fluid network

and $\mu_0 = \alpha$ then the fluid dynamical equations (2.3.2)–(2.3.5) for this network simplify to:

$$Q_k(t) = Q_k(0) + \mu_{k-1}T_{k-1}(t) - \mu_k T_k(t), \quad t \geq 0, \quad k = 1, 2, \ldots, 6, \tag{6.1.1}$$

$$Q_k(t) \geq 0, \quad t \geq 0, \quad k = 1, 2, \ldots, 6, \tag{6.1.2}$$

$$T_k(\cdot) \text{ is nondecreasing}, \quad k = 1, 2, \ldots, 6, \tag{6.1.3}$$

$$U_i(\cdot) \text{ is nondecreasing}, \quad i = 1, 2, 3, \tag{6.1.4}$$

Once again, the static buffer priority disciplines will play an important role. We note that there are eight static buffer priority disciplines associated with our three-station fluid network. They are: $\pi_{\{1,2,3\}}$, $\pi_{\{1,2,6\}}$, $\pi_{\{1,5,3\}}$, $\pi_{\{1,5,6\}}$, $\pi_{\{4,2,3\}}$, $\pi_{\{4,2,6\}}$, $\pi_{\{4,5,3\}}$, and $\pi_{\{4,5,6\}}$.

For this network, we can also express the dynamical equations (2.3.10) that describe these policies in a simpler form. We let $\pi(i)$ denote the high priority class at station i under

the static buffer priority discipline π. With this notation, our three-station fluid network under the static buffer priority discipline π requires the additional equations:

$$\dot{T}_{\pi(i)}(t) = 1 \quad \text{if } Q_{\pi(i)}(t) > 0, \quad i = 1, 2, 3 \tag{6.1.5}$$

for each regular point t of $T(\cdot)$. These conditions simply stipulate that if fluid has accumulated in a station's higher priority buffer, the station must allocate all its effort to that buffer. Any solution $(Q(\cdot), T(\cdot))$ to (6.1.1)–(6.1.5) is a fluid solution under the discipline π.

In this chapter we will investigate the stability properties of the three-station network in Figure 10. In Section 3.2 we introduced the global stability region $\mathcal{D}_\infty$, the stability region $\mathcal{D}_\pi$ and the monotone global stability region $\mathcal{M}_\infty$. Before we present the theorems that will be proven in this chapter, we remark on the relationship between some of these regions.

First, for our three-station network, the region in which the usual traffic conditions hold is given by

$$\mathcal{D}_0 \equiv \{m \in \mathbb{R}^6_+ : m > 0, \lambda(m_1 + m_4) < 1, \lambda(m_2 + m_5) < 1, \lambda(m_3 + m_6) < 1\}$$

Since these conditions are necessary for stability, we have the following relationship

$$\mathcal{D}_\infty \subseteq \mathcal{D}_\pi \subseteq \mathcal{D}_0$$

Clearly, the monotone global stability region is contained in the global stability region. Thus,

$$\mathcal{M}_\infty \subseteq \mathcal{D}_\infty \subseteq \cap_\pi \mathcal{D}_\pi \subseteq \mathcal{D}_0, \tag{6.1.6}$$

where, hereafter, the intersection above is over all eight static buffer priority disciplines.

We will show that the global stability region of the network depicted in Figure 10 is not monotone. Thus, the network can be globally stable under one vector m of service times, but not be globally stable when some of the service times are reduced, i.e., not be globally stable under a service time vector $\tilde{m} \leq m$.

6.2 Results for the Three-station Network

To state our first theorem, we define the following system of linear constraints, which as we show in Section 6.4 is closely related to a piecewise linear Lyapunov function for our

three-station fluid network:

$$\lambda(x_1 + x_4) < x_1\mu_1, \tag{6.2.1}$$

$$\lambda(x_1 + x_4) < x_4\mu_4, \tag{6.2.2}$$

$$\lambda(x_2 + x_5) < x_2\mu_2, \tag{6.2.3}$$

$$\lambda(x_2 + x_5) < x_5\mu_5, \tag{6.2.4}$$

$$\lambda(x_3 + x_6) < x_3\mu_3, \tag{6.2.5}$$

$$\lambda(x_3 + x_6) < x_6\mu_6, \tag{6.2.6}$$

$$x_4 \leq x_3 + x_6, \tag{6.2.7}$$

$$x_5 \leq x_4, \tag{6.2.8}$$

$$x_2 + x_5 \leq x_1 + x_4, \tag{6.2.9}$$

$$x_3 + x_6 \leq x_2 + x_5, \tag{6.2.10}$$

$$x_6 \leq x_5. \tag{6.2.11}$$

Theorem 6.2.1. *The global stability region of the fluid network in Figure 10 is not monotone, i.e., $\mathcal{M}_\infty \neq \mathcal{D}_\infty$. Furthermore, for a positive service time vector m, the following are equivalent.*

1. *The vector m is in the monotone global stability region $\mathcal{M}_\infty$.*

2. *There exists* $x = (x_1, \ldots, x_6) > 0$ *satisfying (6.2.1)–(6.2.11).*

3. *The vector* m *belongs to*

$$\mathcal{D}_0 \cap \left\{ m \in \mathbb{R}_+^6 : \lambda m_2 + \lambda^2 m_4 m_6 < 1 \right\}.$$

We leave the proof of Theorem 6.2.1 to Section 6.4.

The system of linear constraints (6.2.1)–(6.2.11) derived from our piecewise linear Lyapunov function provides conditions sufficient to ensure that a service time vector m is in the global stability region. In fact we show that, together with the usual traffic conditions, the single additional condition:

$$\lambda m_2 + \lambda^2 m_4 m_6 < 1 \tag{6.2.12}$$

is sufficient to ensure global stability.

To obtain conditions necessary for global stability, we construct unstable fluid solutions. The next theorem shows that when $m_4 > m_3$, the additional condition (6.2.12) is also necessary to ensure global stability. When $m_4 \leq m_3$, however, new conditions arise. First, condition (6.2.17) ensures that work will arrive at station 1 at least as quickly as the station processes it. Otherwise, station 1 will eventually empty and thereafter remain empty, essentially reducing the system to a two-station network. The proof of Theorem 6.2.2, given in Section 6.3, involves the construction of unstable fluid solutions under dynamic disciplines that give different sets of buffers higher priority at different times. When $m_4 \leq m_3$, condition (6.2.16), the strongest necessary condition we could obtain from these disciplines, is weaker than our sufficient condition (6.2.12). It is unclear whether or not the fluid network

is globally stable when the mean service time vector m satisfies, $m \in \mathcal{D}_0$, $m_4 \leq m_3$ and

$$\left(\frac{\lambda m_1 m_3 + m_4 - m_3}{m_1 + m_4 - m_3}\right) \frac{\lambda m_6}{1 - \lambda m_2} < 1 \leq \lambda m_4 \frac{\lambda m_6}{1 - \lambda m_2}.$$

Theorem 6.2.2. *If the service time vector of the fluid network in Figure 10 satisfies*

$$m_4 > m_3, \text{ and} \tag{6.2.13}$$

$$\lambda m_2 + \lambda^2 m_4 m_6 \geq 1, \tag{6.2.14}$$

or if it satisfies

$$m_4 \leq m_3, \tag{6.2.15}$$

$$\left(\frac{\lambda m_1 m_3 + m_4 - m_3}{m_1 + m_4 - m_3}\right) \frac{\lambda m_6}{1 - \lambda m_2} \geq 1, \text{ and} \tag{6.2.16}$$

$$\lambda m_1 + \frac{m_4}{m_3} \geq 1 \tag{6.2.17}$$

there is an unstable (non-idling) fluid solution.

Bertsimas, Gamarnik and Tsitisklis [2] developed an LP for testing the global stability of a fluid network. For two-station fluid networks, their LP has optimal objective value 0 if and only if the network is globally stable with the given arrival and service rates. For the three-station network in Figure 10, their LP is:

$$\max \quad \tau_1 + \tau_2 + \tau_3 \tag{6.2.18}$$

subject to

$$\lambda\tau_1 - \mu_1\tau_{11} \leq 0, \tag{6.2.19}$$

$$\mu_{k-1}\tau_{k-1,\sigma(k)} - \mu_k\tau_{k,\sigma(k)} \leq 0 \quad k = 2, 3, \ldots, 6, \tag{6.2.20}$$

$$\sum_{k:\sigma(k)=i} \tau_{ki} = \tau_i \quad i = 1, 2, 3, \tag{6.2.21}$$

$$\sum_{k:\sigma(k)=j} \tau_{ki} \leq \tau_i \quad j, i \in \{1, 2, 3\} \tag{6.2.22}$$

$$j \neq i,$$

$$\lambda(\tau_1 + \tau_2 + \tau_3) - \mu_1(\tau_{11} + \tau_{12} + \tau_{13}) = 0, \tag{6.2.23}$$

$$\mu_{k-1}(\tau_{k-1,1} + \tau_{k-1,2} + \tau_{k-1,3}) - \mu_k(\tau_{k1} + \tau_{k2} + \tau_{k3}) = 0 \quad k = 2, 3, \ldots, 6, \tag{6.2.24}$$

$$\tau_i, \tau_{ji} \geq 0 \quad i = 1, 2, 3 \tag{6.2.25}$$

$$j = 1, \ldots, 6$$

Theorem 6.2.3. *The LP of Bertsimas, Gamarnik and Tsitisklis [2] does not provide a sharp characterization of (monotone) global stability for networks with more than two stations.*

We prove this theorem in Section 6.5 by demonstrating a service time vector m in the monotone global stability region $\mathcal{M}_\infty$ (with arrival rate $\alpha = 1$) for which the LP (6.2.18)–(6.2.26) of Bertsimas, Gamarnik and Tsitisklis [2] has unbounded objective value.

Theorem 6.2.4, which is proved in Section 6.6, shows that the stability regions of all but one of the static buffer priority disciplines are defined by the usual traffic conditions. The stability region of one static buffer priority discipline, $\pi_{\{4,2,6\}}$ involves conditions more restrictive than the usual traffic conditions, but strictly contains the global stability region.

Theorem 6.2.4. *(a) For any static buffer priority discipline* $\pi \neq \pi_{\{4,2,6\}}$, $\mathcal{D}_\pi = \mathcal{D}_0$.
(b) $\mathcal{D}_{\pi_{\{4,2,6\}}} \neq \mathcal{D}_0$.
(c) $\mathcal{D}_{\pi_{\{4,2,6\}}} \neq \mathcal{D}_\infty$.

An immediate consequence of Theorem 6.2.4 is the following corollary. Unlike their two-station counterparts the global stability regions of fluid networks with more than two stations need not be defined by the static buffer priority disciplines.

Corollary 6.2.5. $\mathcal{D}_\infty \neq \cap_\pi \mathcal{D}_\pi$.

Chen and Zhang [11] employed linear Lyapunov functions to study the stability of a fluid network under static buffer priority disciplines. They introduced a linear program, described in Lemma 6.6.1, that is related to the linear Lyapunov functions and showed that if this LP has strictly positive objective value, the fluid network is stable under the given discipline. Theorem 6.2.6 shows that the converse is not true.

Theorem 6.2.6. *The LP of Chen and Zhang [11] need not provide a sharp characterization of stability for fluid networks under static buffer priority disciplines.*

We prove this theorem in Section 6.6 by demonstrating a service time vector m in the

global stability region $\mathcal{M}_\infty$ (with arrival rate $\alpha = 1$), for which the LP of Chen and Zhang has optimal objective value 0.

6.3 Instability of the Fluid Network

To obtain conditions necessary to ensure global stability, we describe disciplines and construct unstable fluid solutions for a broad range of service times. These unstable fluid solutions explicitly demonstrate that the system is unstable over the range of service times. We offer two closely related disciplines. The first, given in Part (a) of the proof, demonstrates conditions under which the fluid network is not globally stable when $m_4 > m_3$. The second, given in Part (b) of the proof, provides similar conditions for the case when $m_4 \leq m_3$.

Proof of Theorem 6.2.2. **Part (a):** We assume that the mean service vector $m > 0$ satisfies (6.2.13)–(6.2.14). We further assume that the usual traffic conditions (3.2.1) hold. Otherwise, any non-idling solution is unstable.

For each subset $S \subseteq \{1, 2, \ldots, 6\}$ we define:

$$Q_S(t) = \sum_{i \in S} Q_i(t).$$

We construct an unstable fluid solution using a discipline under which the priorities among the classes at each station may change depending on the levels of fluid in the buffers. We set $s_0 = 0$ and let $[s_{i-1}, s_i]$, $i = 1, 2, \ldots$ be intervals in which the buffer priorities are constant. We use t_i to denote the length of the ith interval, so $t_i = s_i - s_{i-1}$. We also let d_k denote the departure rate from buffer k during a given interval.

We first note that the usual traffic conditions, along with (6.2.14) imply that

$$\mu_5 > \max\{\mu_4, \mu_6\} \text{ and} \tag{6.3.1}$$

$$\mu_2 < \min\{\mu_1, \mu_3\}. \tag{6.3.2}$$

We start at initial time s_0 and assume $Q_{\{1,2,3\}}(s_0) = 0$, $Q_{\{4,5\}}(s_0) > 0$ and $Q_6(s_0) \geq 0$.

Step 1. We begin by giving classes 1, 5, and 6 higher priority. We set $s_1 = \min\{t \geq s_0 : Q_5(t) = 0, Q_6(t) = Q_6(s_0)\}$. If $Q_5(s_0) = 0$ then $s_1 = s_0$ and we go directly to Step 2. Otherwise, since $\mu_6 < \mu_5$, buffer 6 begins to accumulate fluid and thus $d_6 = \mu_6$ in $[s_0, s_1]$. This implies that $d_3 = 0$ during this interval. We note further that $Q_1(s_1) = 0$ because buffer 1 has priority. So, we have that

$$\dot{Q}_{\{1,2,3\}}(t) = \lambda \quad \text{and} \quad \dot{Q}_{\{4,5,6\}}(t) = -\mu_6 \qquad \text{for } s_0 \leq t \leq s_1.$$

The above imply

$$\dot{Q}_{\{1,2,3\}}(t) + \lambda m_6 \dot{Q}_{\{4,5,6\}}(t) = 0 \quad \text{for } s_0 \leq t \leq s_1,$$

hence

$$Q_{\{2,3\}}(s_1) + \lambda m_6 Q_4(s_1) = \lambda m_6 Q_{\{4,5\}}(s_0). \tag{6.3.3}$$

Step 2. In the next period we give buffers 3, 4, and 5 higher priority. We set $s_2 = \min\{t \geq s_1 : Q_3(t) + Q_4(t) = 0\}$. If $Q_3(s_1) + Q_4(s_1) = 0$ then $s_2 = s_1$ and we go directly to Step 3. Otherwise, since $\mu_4 < \mu_3$, buffer 3 will empty before buffer 4. So, by our priority

scheme in $[s_1, s_2]$, we must have $d_4 = \mu_4$ and $d_1 = 0$ in $[s_1, s_2]$. Also, $Q_5(s_2) = 0$ since buffer 5 has priority and $\mu_4 < \mu_5$. Thus,

$$\dot{Q}_1(t) = \lambda \quad \text{and} \quad \dot{Q}_{\{2,3,4\}}(t) = -\mu_4 \qquad \text{for } s_1 \le t \le s_2.$$

The above imply

$$\dot{Q}_1(t) + \lambda m_4 \dot{Q}_{\{2,3,4\}}(t) = 0 \quad \text{for } s_1 \le t \le s_2,$$

hence

$$Q_1(s_2) + \lambda m_4 Q_2(s_2) = \lambda m_4 Q_{\{2,3,4\}}(s_1). \tag{6.3.4}$$

Step 3. In the final period, we let buffers 1, 2, and 3 have higher priority. We set $s_3 = \min\{t \ge s_2 : Q_2(t) = 0\}$. Notice that buffer 1 will empty before buffer 2 since $\mu_2 < \mu_1$. So we will have $d_2 = \mu_2$ and $d_5 = 0$ in $[s_2, s_3]$. Further, $Q_3(s_3) = 0$ since buffer 3 has high priority and $\mu_2 < \mu_3$. Thus,

$$\dot{Q}_{\{1,2\}}(t) = \lambda - \mu_2 \dot{Q}_{\{3,4,5\}}(t) = \mu_2 \qquad \text{for } s_2 \le t \le s_3.$$

The above imply

$$\dot{Q}_{\{3,4,5\}}(t) + \dot{Q}_{\{1,2\}}(t)/(1 - \lambda m_2) = 0 \quad \text{for } s_2 \le t \le s_3,$$

hence

$$Q_{\{4,5\}}(s_3) = \frac{Q_{\{1,2\}}(s_2)}{1 - \lambda m_2}. \tag{6.3.5}$$

Step 4. Now from equations (6.3.3)–(6.3.5) and the fact that $\lambda m_i < 1$ from the usual

traffic conditions we have

$$\begin{aligned}
Q_{\{4,5\}}(s_3) &= \frac{Q_{\{1,2\}}(s_2)}{1-\lambda m_2} \\
&= \frac{Q_1(s_2)+Q_2(s_2)}{1-\lambda m_2} \\
&\geq \frac{Q_1(s_2)+\lambda m_4 Q_2(s_2)}{1-\lambda m_2} \\
&= \frac{\lambda m_4 Q_{\{2,3,4\}}(s_1)}{1-\lambda m_2} \\
&= \frac{\lambda m_4 (Q_{\{2,3\}}(s_1)+Q_4(s_1))}{1-\lambda m_2} \\
&\geq \frac{\lambda m_4 (Q_{\{2,3\}}(s_1)+\lambda m_6 Q_4(s_1))}{1-\lambda m_2} \\
&= \frac{\lambda^2 m_4 m_6}{1-\lambda m_2} Q_{\{4,5\}}(s_0).
\end{aligned}$$

We remark that if either interval 1 or 2 is "null", the result still holds, by a similar (simpler) chain of inequalities.

Now, by condition (6.2.14) we conclude

$$Q_{\{4,5\}}(s_3) \geq Q_{\{4,5\}}(s_0).$$

Recalling that $Q_{\{1,2,3\}}(s_3) = 0$ under our policy, the above implies that the fluid solutions constructed under our discipline are unstable, proving that the network is not globally stable.

Part (b): Next we assume that the mean service time vector $m > 0$ satisfies (6.2.15)–(6.2.17). We begin by noting that (6.3.1) and (6.3.2) still hold under (6.2.15)–(6.2.17). We only need alter Steps 2 and 4 in the proof of Part (a). In particular, equations (6.3.3) and (6.3.5) continue to hold. We present the revised Steps 2' and 4' below.

Step 2′. In this period we give buffers 3, 4 and 5 higher priority. We again set $s_2 = \min\{t \geq s_1 : Q_3(t) + Q_4(t) = 0\}$. Without loss of generality, we suppose that buffer 4 drains before buffer 3, otherwise we may employ the proof used in Part (a). Also, as before, if $Q_3(s_1) + Q_4(s_1) = 0$, then $s_2 = s_1$ and we go directly to Step 3.

Let us denote the time at which buffer 4 empties as r (with $s_1 \leq r \leq s_2$). As before, we must have $d_4 = \mu_4$ and $d_1 = 0$ in $[s_1, r]$. Thus

$$\dot{Q}_1(t) = \lambda \quad \text{and} \quad \dot{Q}_{\{2,3,4\}}(t) = -\mu_4 \qquad \text{for } s_1 \leq t \leq r. \tag{6.3.6}$$

The above imply

$$\dot{Q}_1(t) = -\lambda m_4 \dot{Q}_{\{2,3,4\}}(t) \quad \text{for } s_1 \leq t \leq r$$

and this yields

$$Q_1(r) + \lambda m_4 Q_{\{2,3,4\}}(r) - \lambda m_4 Q_{\{2,3,4\}}(s_1) = 0. \tag{6.3.7}$$

Now during $[r, s_2]$, we have that $d_4 = d_3 = \mu_3$ and by work conservation $d_1 = \hat{d}_1 := \frac{1}{m_1}(1 - \mu_3 m_4)$. Note that $\hat{d}_1 \leq \lambda$ by (6.2.17). Thus, for this part of the interval, we have

$$\dot{Q}_1(t) = \lambda - \hat{d}_1 \quad \text{and} \quad \dot{Q}_{\{2,3,4\}}(t) = \hat{d}_1 - \mu_3 \qquad \text{for } r \leq t \leq s_2.$$

The above imply

$$\dot{Q}_1(t) + \frac{\lambda - \hat{d}_1}{\mu_3 - \hat{d}_1} \dot{Q}_{\{2,3,4\}}(t) = 0 \quad \text{for } r \leq t \leq s_2$$

and this gives

$$Q_1(s_2) - Q_1(r) + \kappa Q_2(s_2) - \kappa Q_{\{2,3,4\}}(r) = 0 \tag{6.3.8}$$

where we have set

$$\kappa = \frac{\lambda - \hat{d}_1}{\mu_3 - \hat{d}_1} = \frac{\lambda m_1 m_3 + m_4 - m_3}{m_1 + m_4 - m_3}.$$

Now, adding (6.3.7) and (6.3.8) and rearranging:

$$Q_1(s_2) + \kappa Q_2(s_2) = \kappa Q_{\{2,3,4\}}(s_1) + (\lambda m_4 - \kappa)[Q_{\{2,3,4\}}(s_1) - Q_{\{2,3,4\}}(r)].$$

A little algebra shows that $\kappa \leq \lambda m_4$ and $Q_{\{2,3,4\}}(s_1) \geq Q_{\{2,3,4\}}(r)$ by virtue of (6.3.6). Thus, we have

$$Q_1(s_2) + \kappa Q_2(s_2) \geq \kappa Q_{\{2,3,4\}}(s_1).$$

Step 4′.

$$\begin{aligned}
Q_{\{4,5\}}(s_3) &= \frac{Q_{\{1,2\}}(s_2)}{1 - \lambda m_2} \\
&= \frac{Q_1(s_2) + Q_2(s_2)}{1 - \lambda m_2} \\
&\geq \frac{Q_1(s_2) + \kappa Q_2(s_2)}{1 - \lambda m_2} \\
&\geq \frac{\kappa Q_{\{2,3,4\}}(s_1)}{1 - \lambda m_2} \\
&= \frac{\kappa(Q_{\{2,3\}}(s_1) + Q_4(s_1))}{1 - \lambda m_2} \\
&\geq \frac{\kappa(Q_{\{2,3\}}(s_1) + \lambda m_6 Q_4(s_1))}{1 - \lambda m_2} \\
&= \frac{\kappa \lambda m_6}{1 - \lambda m_2} Q_{\{4,5\}}(s_0).
\end{aligned}$$

By our assumptions we can conclude

$$Q_{\{4,5\}}(s_3) \geq Q_{\{4,5\}}(s_0),$$

which again implies the instability of our fluid solution.

□

6.4 Piecewise Linear Lyapunov Functions

In this section we prove Theorem 6.2.1 showing that the global stability region of our three-station network is not monotone and characterizing its monotone global stability region. We first introduce the piecewise linear Lyapunov functions we use to establish conditions sufficient to ensure global stability. Given $x = (x_k) > 0$ and a fluid solution $Q(\cdot)$, let

$$f_i(x, Q(t)) = \sum_{\sigma(k)=i} x_k Q_k^+(t), \quad i = 1, 2, 3,$$

where $Q_k^+(t) = \sum_{\ell=1}^{k} Q_\ell(t)$. Further, let

$$f(x, Q(t)) = \max\{f_1(x, Q(t)), f_2(x, Q(t)), f_3(x, Q(t))\}.$$

We often write $f(Q(t))$ in place of the more cumbersome $f(x, Q(t))$. Clearly, $f(Q(t))$ is a piecewise linear function of $Q(t) = (Q_k(t))$.

The next lemma suggests a way in which to construct piecewise linear Lyapunov functions. This type of construction was introduced by Botvich and Zamyatim [3] for a two-station network. It was independently generalized by Dai and Weiss [17] and Down and Meyn [18].

Lemma 6.4.1. *Suppose there exists $x = (x_k) > 0$, $t_0 \geq 0$ and $\epsilon > 0$ such that for each non-idling fluid solution $(Q(\cdot), T(\cdot))$ and each regular point $t > t_0$ of $T(\cdot)$, the following hold for each $i = 1, 2, 3$:*

$$\frac{df_i(x, Q(t))}{dt} \leq -\epsilon \text{ whenever } Z_i(t) > 0, \tag{6.4.1}$$

$$f_i(x, Q(t)) \leq \max\{f_j(x, Q(t)) : j \in \{1, 2, 3\}, j \neq i\} \text{ whenever } Z_i(t) = 0, \tag{6.4.2}$$

$$\max\{f_j(Q(t)) : j \in \{1, 2, 3\}, j \neq i\} \leq f_i(Q(t)) \text{ whenever } \sum_{j \neq i} Z_j(t) = 0. \tag{6.4.3}$$

Then f is a piecewise linear Lyapunov function.

Proof. Let t be a regular point of f and T with $Q(t) \neq 0$. We show that (3.3.1) holds. Because $Q(t) \neq 0$ and (6.4.2)–(6.4.3) hold, there exists an index $i \in \{1, 2, 3\}$ such that $f_i(Q(t)) = f(Q(t))$ and $Z_i(t) > 0$. From the proof of Lemma 3.2 of Dai and Weiss [17], we have

$$\frac{df(Q(t))}{dt} = \frac{df_i(Q(t))}{dt}.$$

Then the conditions in Proposition 3.3.2 follows from (6.4.1) and the definition of f and the conclusion follows. □

Lemma 6.4.2. *If there is $x = (x_k) > 0$ satisfying the linear constraints (6.2.1)–(6.2.11), then there exists $\epsilon > 0$ such that (6.4.1)–(6.4.3) hold and hence, f is a piecewise linear*

Lyapunov function.

Proof. Let $t_0 = 0$ and let $x = (x_k) > 0$ satisfy (6.2.1)–(6.2.11). Define ϵ to be the minimum of the following 6 terms:

$$x_1\mu_1 - \lambda(x_1 + x_4), \quad x_4\mu_4 - \lambda(x_1 + x_4),$$

$$x_2\mu_2 - \lambda(x_2 + x_5), \quad x_5\mu_5 - \lambda(x_2 + x_5),$$

$$x_3\mu_3 - \lambda(x_3 + x_6), \quad x_6\mu_6 - \lambda(x_3 + x_6).$$

Clearly, $\epsilon > 0$. Let $Q_k^+(t) = \sum_{\ell=1}^{k} Q_\ell(t)$ and consider a non-idling fluid solution $(Q(\cdot), T(\cdot))$ and a time $t > 0$ that is regular for $T(\cdot)$. Observe that the amount of fluid in buffers 1 through k is

$$Q_k^+(t) = Q_k^+(0) + \lambda t - \mu_k T_k(t).$$

Hence

$$f_1(Q(t)) = f_1(0) + (x_1 + x_4)\lambda t - x_1\mu_1 T_1(t) - x_4\mu_4 T_4(t)$$

and

$$\frac{df_1(Q(t))}{dt} = \lambda(x_1 + x_4) - x_1\mu_1\dot{T}_1(t) - x_4\mu_4\dot{T}_4.$$

If $Z_1(t) > 0$, it follows from (6.4.2) that since $(Q(\cdot), T(\cdot))$ is non-idling, $\dot{U}_1(t) = 0$ or $\dot{T}_1(t) + \dot{T}_4(t) = 1$. Thus, by the definition of ϵ,

$$\dot{f}_1(t) \leq -\epsilon \quad \text{when } Z_1(t) > 0.$$

Similar analysis for $i = 2$ and $i = 3$ shows that (6.4.1) holds.

We next establish (6.4.2). When $Z_1(t) = 0$,

$$f_1(Q(t)) = x_4(Q_2(t) + Q_3(t)) \text{ and}$$

$$f_3(Q(t)) = x_3(Q_2(t) + Q_3(t)) + x_6(Q_2(t) + Q_3(t) + Q_5(t) + Q_6(t)),$$

and Equation (6.2.7) ensures that $f_1(Q(t)) \leq f_3(Q(t))$. When $Z_2(t) = 0$,

$$f_2(Q(t)) = x_2Q_1(t) + x_5(Q_1(t) + Q_3(t) + Q_4(t)),$$

$$f_1(Q(t)) = x_1Q_1(t) + x_4(Q_1(t) + Q_3(t) + Q_4(t))$$

and Equations (6.2.8)–(6.2.9) ensure that $f_2(Q_2(t)) \leq f_1(Q_(t))$. When $Z_3(t) = 0$,

$$f_3(Q(t)) = x_3(Q_1(t) + Q_2(t)) + x_6(Q_1(t) + Q_2(t) + Q_4(t) + Q_5(t)),$$

$$f_2(Q(t)) = x_2(Q_1(t) + Q_2(t)) + x_5(Q_1(t) + Q_2(t) + Q_4(t) + Q_5(t))$$

and Equations (6.2.10)–(6.2.11) ensure that $f_3(Q(t)) \leq f_2(Q(t))$.

Finally, we establish (6.4.3). When $Z_1(t) = 0$ and $Z_2(t) = 0$,

$$f_1(Q(t)) = x_4Q_3(t),$$

$$f_2(Q(t)) = x_5Q_3(t),$$

$$f_3(Q(t)) = x_3Q_3(t) + x_6(Q_3(t) + Q_6(t)).$$

Equation (6.2.7) ensures that $f_1(Q(t)) \leq f_3(Q(t))$ and Equations (6.2.7) and (6.2.8) ensure that $f_2(Q(t)) \leq f_3(Q(t))$. The remaining cases of (6.4.3) can be verified similarly. □

Remark 6.4.1. (a) In general, condition (6.4.2) generates non-linear constraints on $x = (x_k)$. However, for our network, the linear constraints arising from (6.4.3) imply condition (6.4.2) and so we have the set of linear constraints (6.2.1)–(6.2.11) associated with our piecewise linear Lyapunov function.

(b) For a d-station generalization of our fluid network in which fluid repeatedly visits all of the stations in a fixed order, there is an analogous natural set of linear constraints associated with a piecewise linear Lyapunov function. Further, it is not difficult to obtain explicit conditions in terms of the service times and arrival rate characterizing exactly when the linear constraints admit a solution x.

(c) The existence of a solution x to the system of linear constraints (6.2.1)–(6.2.11) ensures the existence of a piecewise linear Lyapunov function satisfying conditions (6.4.1)–(6.4.3). The converse, however, does not hold; see Lemma 6.4.4.

Lemma 6.4.3. *The linear constraints (6.2.1)–(6.2.11) admit a feasible solution* $x = (x_k) > 0$ *if and only if*

$$\lambda(m_1 + m_4) < 1, \tag{6.4.4}$$

$$\lambda(m_2 + m_5) < 1, \tag{6.4.5}$$

$$\lambda(m_3 + m_6) < 1, \tag{6.4.6}$$

$$\lambda m_2 + \lambda^2 m_4 m_6 < 1. \tag{6.4.7}$$

Proof. Given $(x_1, \ldots, x_6) > 0$, let

$$y_1 = \frac{x_4}{x_1 + x_4}, \quad y_2 = \frac{x_5}{x_2 + x_5}, \quad y_3 = \frac{x_6}{x_3 + x_6}.$$

Then $(x_1, \ldots, x_6) > 0$ satisfies (6.2.1)–(6.2.11) if and only if $(y_1, y_2, y_3, x_4, x_5, x_6) > 0$ satisfies

$$\lambda m_1 < 1 - y_1, \tag{6.4.8}$$

$$\lambda m_4 < y_1, \tag{6.4.9}$$

$$\lambda m_2 < 1 - y_2, \tag{6.4.10}$$

$$\lambda m_5 < y_2, \tag{6.4.11}$$

$$\lambda m_3 < 1 - y_3, \tag{6.4.12}$$

$$\lambda m_6 < y_3, \tag{6.4.13}$$

$$y_3 x_4 \leq x_6, \tag{6.4.14}$$

$$x_5 \leq x_4, \tag{6.4.15}$$

$$x_5 y_1 \leq x_4 y_2, \tag{6.4.16}$$

$$x_6 y_2 \leq x_5 y_3, \tag{6.4.17}$$

$$x_6 \leq x_5. \tag{6.4.18}$$

The vector $(y_1, y_2, y_3, x_4, x_5, x_6) > 0$ satisfies (6.4.8)–(6.4.18) if and only if $\alpha(y_1, y_2, y_3, x_4, x_5, x_6)$ satisfies (6.4.8)–(6.4.18) for each positive scalar α. Choosing $\alpha = 1/x_4$, we see that there is a strictly positive solution to (6.4.8)–(6.4.18) if and only if there is $(y_1, y_2, y_3, x_5, x_6) > 0$ satisfying

$$\lambda m_4 < y_1 < 1 - \lambda m_1, \tag{6.4.19}$$

$$\lambda m_5 < y_2 < 1 - \lambda m_2, \tag{6.4.20}$$

$$\lambda m_6 < y_3 < 1 - \lambda m_3, \tag{6.4.21}$$

$$y_3 \leq x_6, \tag{6.4.22}$$

$$x_5 \leq 1 \tag{6.4.23}$$

$$x_5 \leq \frac{y_2}{y_1}, \tag{6.4.24}$$

$$x_6 \leq x_5 \frac{y_3}{y_2}, \tag{6.4.25}$$

$$x_6 \leq x_5. \tag{6.4.26}$$

The existence of $(y_1, y_2, y_3, x_5, x_6) > 0$ satisfying (6.4.19)–(6.4.26) is equivalent to the existence of $(y_1, y_2, y_3, x_5) > 0$ satisfying

$$\lambda m_4 < y_1 < 1 - \lambda m_1, \tag{6.4.27}$$

$$\lambda m_5 < y_2 < 1 - \lambda m_2, \tag{6.4.28}$$

$$\lambda m_6 < y_3 < 1 - \lambda m_3, \tag{6.4.29}$$

$$x_5 \leq 1 \tag{6.4.30}$$

$$x_5 \leq \frac{y_2}{y_1}, \tag{6.4.31}$$

$$y_2 \leq x_5, \tag{6.4.32}$$

$$y_3 \leq x_5, \tag{6.4.33}$$

which is equivalent to the existence of (y_1, y_2, y_3) satisfying

$$\lambda m_4 < y_1 < 1 - \lambda m_1, \tag{6.4.34}$$

$$\lambda m_5 < y_2 < 1 - \lambda m_2, \tag{6.4.35}$$

$$\lambda m_6 < y_3 < 1 - \lambda m_3, \tag{6.4.36}$$

$$y_1 y_3 \leq y_2. \tag{6.4.37}$$

Finally, the existence of (y_1, y_2, y_3) satisfying (6.4.34)–(6.4.37) is equivalent to (6.4.4)–(6.4.7). □

The following lemma establishes an alternate set of conditions sufficient to ensure global stability in our three-station fluid network.

Lemma 6.4.4. *If*

$$\lambda m_1 + m_4/m_3 < 1, \tag{6.4.38}$$

$$\lambda(m_2 + m_5) < 1, \tag{6.4.39}$$

$$\lambda(m_3 + m_6) < 1, \tag{6.4.40}$$

the fluid network of Figure 10 is globally stable.

Proof. Assume that (6.4.38) holds. We first show that there is $t_0 > 0$ such that for all non-idling fluid solutions $(Q(\cdot), T(\cdot))$ with $|Q(0)| = 1$, $Z_1(t) = 0$ for each time $t \geq t_0$. We then separately show that there is $t_1 > t_0$ such that for all non-idling fluid solutions $(Q(\cdot), T(\cdot))$ with $|Q(0)| = 1$, $Z_2(t) + Z_3(t) = 0$ for each time $t \geq t_1$ and hence that the network is globally stable.

Let $(Q(\cdot), T(\cdot))$ be any non-idling fluid solution with $|Q(0)| = 1$. Let

$$g(t) = m_1 Q_1(t) + m_4 Q_4(t).$$

From (6.1.1)–(6.1.4),

$$g(t) = g(0) + \lambda m_1 t - T_1(t) + m_4\mu_3 T_3(t) - T_4(t).$$

Therefore, for any regular t with $g(t) > 0$,

$$\dot{g}(t) = \lambda m_1 + m_4\mu_3\dot{T}_3(t) - (\dot{T}_1(t) + \dot{T}_4(t)) \leq \lambda m_1 + m_4\mu_3 - 1 < 0.$$

Therefore $g(t) = 0$ for all $t \geq g(0)/(1 - \lambda m_1 - m_4\mu_3)$. Since $g(0) \leq \max\{m_1, m_4\}$, we have $Z_1(t) = 0$ for $t \geq t_0$, where

$$t_0 = \frac{\max\{m_1, m_4\}}{1 - \lambda m_1 - m_4\mu_3}.$$

To show that buffers at stations 2 and 3 eventually empty, we consider times $t \geq t_0$ and specialize the proof of Lemma 6.4.1 to the case where $Z_1(t) = 0$ and $\dot{Q}_1(t) = \dot{Q}_4(t) = 0$. First, observe that since $Z_1(t) = 0$ for $t \geq t_0$, (6.4.1) is vacuously satisfied for $i = 1$. Similarly, (6.4.3) is trivially satisfied for $i = 1$. Then arguments analogous to those used in the proof of Lemma 6.4.2 show that (6.4.1)–(6.4.3) hold if there exists $(x_2, x_3, x_5, x_6) > 0$

satisfying

$$\lambda(x_2 + x_5) < \mu_2 x_2, \tag{6.4.41}$$

$$\lambda(x_2 + x_5) < \mu_5 x_5, \tag{6.4.42}$$

$$\lambda(x_3 + x_6) < \mu_3 x_3, \tag{6.4.43}$$

$$\lambda(x_3 + x_6) < \mu_6 x_6, \tag{6.4.44}$$

$$x_5 \leq x_3 + x_6, \tag{6.4.45}$$

$$x_3 + x_6 \leq x_2 + x_5, \tag{6.4.46}$$

$$x_6 \leq x_5. \tag{6.4.47}$$

Finally, arguments similar to those used in the proof of Lemma 6.4.3 show that there exists $x > 0$ satisfying (6.4.41)–(6.4.47) if and only if the usual traffic conditions (6.4.39)–(6.4.40) at stations 2 and 3 hold. Therefore, the lemma follows from Lemma 6.4.1. □

Remark 6.4.2. For two-station networks, there is $x > 0$ satisfying the linear constraints arising from our piecewise linear Lyapunov functions if and only if the fluid network is globally stable. This is not the case for networks with more than two stations and Lemma 6.4.4 illustrates one way in which the network can be globally stable even when the linear system (6.2.1)–(6.2.11) admits no positive solution.

We are now prepared to prove our main result, Theorem 6.2.1, showing that the global stability region of our three-station network is not monotone in the service times and characterizing its monotone global stability region both in terms of the solvability of the linear

system (6.2.1)–(6.2.11) and in terms of explicit constraints on the service times and arrival rate.

Proof of Theorem 6.2.1. We first show that (b), the existence of a solution $x > 0$ to the linear system (6.2.1)–(6.2.11), implies (a), that $m \in \mathcal{M}_\infty$. We proved the equivalence of (b) and (c) in Lemma 6.4.3. Then we show that (a) implies (c), thus proving the equivalence of (a), (b) and (c).

Suppose that $m > 0$ is a service time vector for which there exists an $x = (x_k) > 0$ satisfying (6.2.1)–(6.2.11). By Lemma 6.4.2, f is a piecewise linear Lyapunov function proving that m is in the global stability region. To see that m is in the monotone global stability region, observe that for each $0 < \tilde{m} \leq m$, $\tilde{\mu} = (1/\tilde{m}_k) \geq \mu$ and x satisfies (6.2.1)–(6.2.11) with μ replaced by $\tilde{\mu}$. Thus, $f(x, Q(\cdot))$ is also a piecewise linear Lyapunov function proving that $\tilde{m}$ is in the global stability region as well.

Consider a service time vector $m > 0$ such that

$$m \notin \mathcal{D}_0 \cap \{m \in \mathbb{R}_+^d : \lambda m_2 + \lambda^2 m_4 m_6 < 1\}.$$

To show that (a) implies (c), it is enough to show that $m \notin \mathcal{M}_\infty$. If $m \notin \mathcal{D}_0$, then m is clearly not in the global stability region and hence not in $\mathcal{M}_\infty$. So, suppose that m is in $\mathcal{D}_0$ and $\lambda m_2 + \lambda^2 m_4 m_6 \geq 1$. If $m_4 > m_3$, then it follows from Theorem 6.2.2 that m is not in the global stability region and hence not in the monotone global stability region. If $m_4 \leq m_3$, let

$$\tilde{m} = (m_1, m_2, \tilde{m}_3, m_4, m_5, m_6)$$

where $0 < \tilde{m}_3 < m_4 \leq m_3$. Clearly, $\tilde{m} \leq m$ and, by Theorem 6.2.2, $\tilde{m}$ is not in the global stability region. Therefore, m is not in the monotone global stability region of the fluid

network.

Finally, we show that the global stability region $\mathcal{D}_\infty$ is not monotone. Let $\alpha = 1$ and consider the service times

$$m = (0.1, 0.85, 0.5, 0.4, 0.1, 0.4).$$

Since $\lambda m_1 + m_4/m_3 = 0.9 < 1$, it follows from Lemma 6.4.4 that the fluid network is globally stable. Now, suppose that server 3 works faster on class 3 fluids and so the service time m_3 is reduced to $\tilde{m}_3 = 0.1$, for example. The other service times remain unchanged. That is,

$$\tilde{m} = (0.1, 0.85, 0.1, 0.4, 0.1, 0.4).$$

Since $\tilde{m}_4 > \tilde{m}_3$ and $\lambda\tilde{m}_2 + \lambda^2\tilde{m}_4\tilde{m}_6 = 1.01 > 1$, it follows from Theorem 6.2.2 that the network is not globally stable when the service time vector is $\tilde{m}$. □

6.5 The Power of the LP by Bertsimas, Gamarnik and Tsitsiklis

Based on a path decomposition approach, Bertsimas, Gamarnik and Tsitsiklis [2] proposed a linear program (LP) to determine whether a particular service time vector m is in the global stability region. They proved that for two-station networks, the LP has bounded objective value if and only if the network is globally stable. They further conjectured that the same would be true for general networks.

In this section we prove that their LP does not provide a sharp characterization of the global stability region or the monotone global stability region of the fluid network in

Figure 10.

Proof of Theorem 6.2.3. When $\alpha = 1$ the service time vector

$$m = (0.5, 0.5, 0.5, 0.4, 0.01, 0.4)$$

is in $\mathcal{M}_\infty$. Therefore, the fluid network with these service times and arrival rate $\alpha = 1$ is globally stable. However, for the service time vector m, a non-zero feasible solution to the LP (6.2.18)–(6.2.26) is given by

$$\tau_1 = \tau_2 = \tau_3 = 10$$

$$\begin{array}{lll} \tau_{11} = 5 & \tau_{21} = 5 & \tau_{31} = 6.25 \\ \tau_{12} = 7 & \tau_{22} = 10 & \tau_{32} = 7 \\ \tau_{13} = 3 & \tau_{23} = 0 & \tau_{33} = 1.75 \\ \tau_{41} = 5 & \tau_{51} = 0.3 & \tau_{61} = 3.75 \\ \tau_{42} = 0 & \tau_{52} = 0 & \tau_{62} = 0 \\ \tau_{43} = 7 & \tau_{53} = 0 & \tau_{63} = 8.25. \end{array}$$

Thus, the LP of Bertsimas, Gamarnik and Tsitsiklis [2] has unbounded objective value. □

6.6 Static Buffer Priority Disciplines

Chen and Zhang [11] employed linear Lyapunov functions to study the stability of fluid networks under static buffer priority disciplines. They showed that if an LP related to their linear Lyapunov function has positive objective value, the fluid network is stable under the discipline. In this section, we show that the converse is not true. Namely, we demonstrate service times m in $\mathcal{D}_{\pi_{\{4,2,6\}}}$, the stability region of our three-station network under the discipline that gives higher priorities to classes 2, 4 and 6, for which the LP of

Chen and Zhang has maximum objective value 0. Thus, their LP does not provide a sharp characterization of the stability of a priority fluid network.

For each $x = (x_k) > 0$ and fluid solution $(Q(\cdot), T(\cdot))$ under the priority discipline $\pi_{\{4,2,6\}}$ define

$$f(x, Q(t)) = \sum_{k=1}^{6} x_k Q_k(t).$$

Clearly, for fixed x, f is a linear function of $Q(t)$. We often write $f(Q(t))$ in place of the more cumbersome $f(x, Q(t))$.

If, for each fluid solution $(Q(\cdot), T(\cdot))$ under the discipline $\pi_{\{4,2,6\}}$ and regular point t such that $Q(t) \neq 0$,

$$\frac{df(Q(t))}{dt} \leq -\epsilon < 0, \tag{6.6.1}$$

then $f(Q(t)) = 0$, and hence $Q(t) = 0$, for all $t \geq f(Q(0))/\epsilon$. In this case, f is a linear Lyapunov function proving that the network is stable under the discipline $\pi_{\{4,2,6\}}$.

For each regular point t of the fluid solution $(Q(\cdot), T(\cdot))$

$$\frac{df(Q(t))}{dt} = \sum_{k=1}^{6} x_k \dot{Q}_k(t) = \sum_{k=1}^{6} x_k (d_{k-1} - d_k),$$

where $d_k = \mu_k \dot{T}_k(t)$ for $k = 1, 2, \ldots, 6$ and $d_0 = \lambda$. To ensure (6.6.1), we impose the linear constraint

$$\sum_{k=1}^{6} x_k (d_{k-1} - d_k) + \epsilon \leq 0 \tag{6.6.2}$$

on x for each feasible choice of $(d_1, d_2, \ldots, d_6)$.

The feasible values of $(d_1, d_2, \ldots, d_6) \geq 0$ depend on the fluid state $Q(t)$ in the following

ways:

$$d_k = d_{k-1} \text{ if } Q_k(t) = 0, \quad k = 1, 2, \ldots, 6, \tag{6.6.3}$$

$$d_1 = 0 \text{ if } Q_4(t) > 0, \tag{6.6.4}$$

$$d_5 = 0 \text{ if } Q_2(t) > 0, \tag{6.6.5}$$

$$d_3 = 0 \text{ if } Q_6(t) > 0, \tag{6.6.6}$$

$$d_1 m_1 + d_4 m_4 = 1 \text{ if } Z_1(t) > 0, \tag{6.6.7}$$

$$d_2 m_2 + d_5 m_5 = 1 \text{ if } Z_2(t) > 0, \tag{6.6.8}$$

$$d_3 m_3 + d_6 m_6 = 1 \text{ if } Z_3(t) > 0. \tag{6.6.9}$$

Equation (6.6.3) follows from Proposition 4.2 of Dai and Weiss [17]. Equations (6.6.4)-(6.6.6) follow from (6.1.5). Finally, equations (6.6.7)-(6.6.9) follow from (2.3.6). We refer to the set of all non-negative vectors $d = (d_1, d_2, \ldots, d_6)$ that satisfy (6.6.3)–(6.6.9) for some $Q(t) \geq 0$ as $\mathcal{T}_{\pi_{\{4,2,6\}}}$.

Lemma 6.6.1 is an immediate consequence of (6.6.2), it specializes the LP criterion of Chen and Zhang [11] to our three-station network.

Lemma 6.6.1. *If the following LP has positive objective value:*

$$\max \epsilon \tag{6.6.10}$$

subject to:

$$\sum_{k=1}^{6} x_k \leq 1, \tag{6.6.11}$$

$$\sum_{k=1}^{6} x_k (d^s_{k-1} - d^s_k) + \epsilon \leq 0 \text{ for each } d^s \in \mathcal{T}_{\pi_{\{4,2,6\}}}, \tag{6.6.12}$$

$$x = (x_k) \geq 0, \tag{6.6.13}$$

then the fluid network is stable under the static buffer priority discipline $\pi_{\{4,2,6\}}$ *and so* $m \in \mathcal{D}_{\pi_{\{4,2,6\}}}$.

We next show that the converse of Lemma 6.6.1 is not true and hence that the LP of Chen and Zhang does not provide a sharp characterization of stability under static priority disciplines.

Proof of Theorem 6.2.6. Let $\lambda = 1$ and let

$$m = (0.001, 0.18, 0.001, 0.9, 0.001, 0.9)$$

be the service time vector. Clearly, m satisfies the usual traffic conditions (3.2.1). Since

$$m_2 + m_4 m_6 = 0.99 < 1,$$

by Theorem 6.2.1, m is in the monotone global stability region, and hence in $\mathcal{D}_{\pi_{\{4,2,6\}}}$.

To show that there is no solution to the LP (6.6.10)–(6.6.13) with positive objective value, we demonstrate a feasible solution to the dual problem with objective value 0. The dual of (6.6.10)–(6.6.13) is:

$$\min \beta \tag{6.6.14}$$

subject to:

$$\sum_{s \in \mathcal{T}_{\pi_{\{4,2,6\}}}} y_s = 1, \tag{6.6.15}$$

$$\sum_{s \in \mathcal{T}_{\pi_{\{4,2,6\}}}} y_s(d^s_{k-1} - d^s_k) + \beta \geq 0 \text{ for each } k = 1, 2, \ldots, 6, \tag{6.6.16}$$

$$y = (y_s) \geq 0 \tag{6.6.17}$$

Our solution involves the seven states described in Table 1.

case	state	departure rates
1	$Q_2(t) > 0, Q_4(t) > 0, Q_6(t) > 0$	$d_1 = d_3 = d_5 = 0, d_2 = \mu_2, d_4 = \mu_4, d_6 = \mu_6$
2	$Q_2(t) > 0, Q_3(t) > 0, Q_4(t) > 0$	$d_1 = d_5 = d_6 = 0, d_2 = \mu_2, d_3 = \mu_3, d_4 = \mu_4$
3	$Q_2(t) > 0, Q_4(t) > 0$	$d_1 = d_5 = d_6 = 0, d_2 = d_3 = \mu_2, d_4 = \mu_4$
4	$Q_4(t) > 0, Q_5(t) > 0, Q_6(t) > 0$	$d_1 = d_2 = d_3 = 0, d_4 = \mu_4, d_5 = \mu_5, d_6 = \mu_6$
5	$Q_4(t) > 0$	$d_1 = d_2 = d_3 = 0, d_4 = d_5 = d_6 = \mu_4$
6	$Q_1(t) > 0, Q_2(t) > 0, Q_6(t) > 0$	$d_3 = d_4 = d_5 = 0, d_1 = \mu_1, d_2 = \mu_2, d_6 = \mu_6$
7	$Q_6(t) > 0$	$d_3 = d_4 = d_5 = 0, d_1 = d_2 = 1, d_6 = \mu_6$

Table 1: Departure rates for the seven states used in our dual solution. Note that the state only lists the highest priority class at each station with positive buffer level.

Tedious algebra establishes that

$$y_6 = \frac{m_1 m_4}{1 - m_1} \approx 0.00090 \tag{6.6.18}$$

$$y_7 = \frac{1 - m_1 - m_4}{1 - m_1} \approx 0.09910 \tag{6.6.19}$$

$$y_2 = \frac{\mu_6(1 - m_2) + m_4\mu_5(m_1 - m_2)/(m_1 - 1) - 1}{\mu_6(1 - m_2\mu_3) + m_4\mu_5(\mu_5 - \mu_4)} \approx 0.00018 \tag{6.6.20}$$

$$y_3 = m_2 - m_2\mu_3 y_2 \approx 0.14772 \tag{6.6.21}$$

$$y_4 = \frac{m_1 - m_2}{1 - m_1} m_4 + m_4\mu_5 y_2 \approx 0.00013 \tag{6.6.22}$$

$$y_5 = m_4 - m_4\mu_5 y_2 \approx 0.73861 \tag{6.6.23}$$

$$y_1 = \frac{m_2 - m_1}{1 - m_1} m_4 - m_2 - (1 - m_2\mu_3)\pi_2 \approx 0.01336 \tag{6.6.24}$$

and $y_s = 0$ otherwise describes a feasible solution to the dual problem (6.6.14)–(6.6.17) with $\beta = 0$ proving that there is no solution to the LP (6.6.10)–(6.6.13) with positive objective value. □

Nevertheless, linear Lyapunov functions remain a powerful tool for establishing the global stability of priority networks. In fact, we rely on this tool to prove that the stability regions of the static buffer priority disciplines to not characterize the global stability region of a network with more than two stations.

Dai and Vande Vate [16] showed that the global stability region of a two-station fluid network is determined by static buffer priority disciplines. We show that this is not the case for fluid networks with more than two stations. This helps explain why we required the dynamic disciplines used in the proof of Theorem 6.2.1 to characterize the monotone

global stability region of our three-station network.

We first show that the stability region of the network under all but one of the static buffer priority disciplines is determined by the usual traffic conditions at each station. Thus, stability under the remaining static buffer priority discipline $\pi_{\{4,2,6\}}$ implies stability under all static buffer priority disciplines. We then demonstrate a service time vector m that is not in the global stability region, but is in the stability region under the discipline $\pi_{\{4,2,6\}}$. This shows that the global stability region of a fluid network with more than two stations is determined by a richer family of disciplines than simply the static buffer priority disciplines.

We show that every fluid solution under a discipline that gives priority to class 1 over class 4 reduces to a fluid solution in the five-class network in Figure 11 obtained by deleting class 1. Similarly, every fluid solution under a discipline that gives priority to class 3 over class 6 eventually reduces to a fluid solution in the five-class network in Figure 12 obtained by deleting class 6.

We start by showing that the global stability regions of these two five-class subnetworks are defined by the usual traffic conditions at each station.

Lemma 6.6.2. *The five-class three-station fluid network in Figure 11 is globally stable so long as the traffic intensity at each station is less than one.*

Proof. Consider the fluid network in Figure 11. For a given $x = (x_1, \ldots, x_5)' > 0$, let

$$f_1(x, Q(t)) = x_3 Q_3^+(t),$$

$$f_2(x, Q(t)) = x_1 Q_1^+(t) + x_4 Q_4^+(t),$$

$$f_3(x, Q(t)) = x_2 Q_2^+(t) + x_5 Q_5^+(t),$$

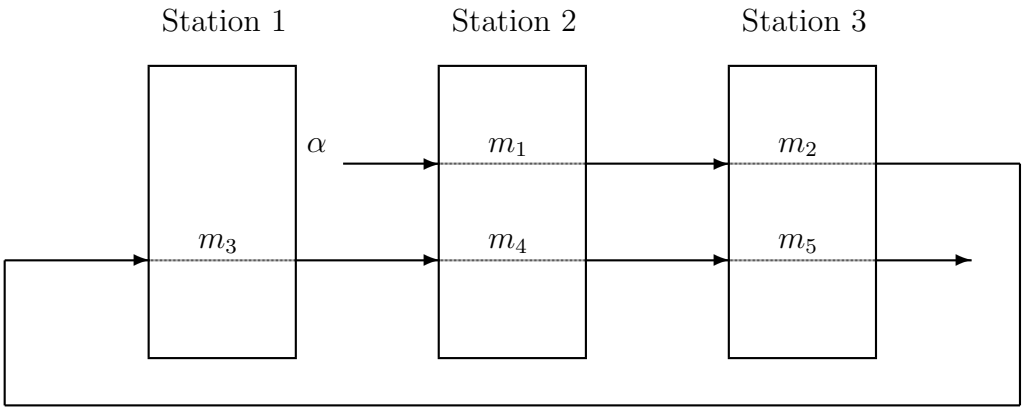

Figure 11: The five-class network obtained by deleting class 1 from the six-class fluid network

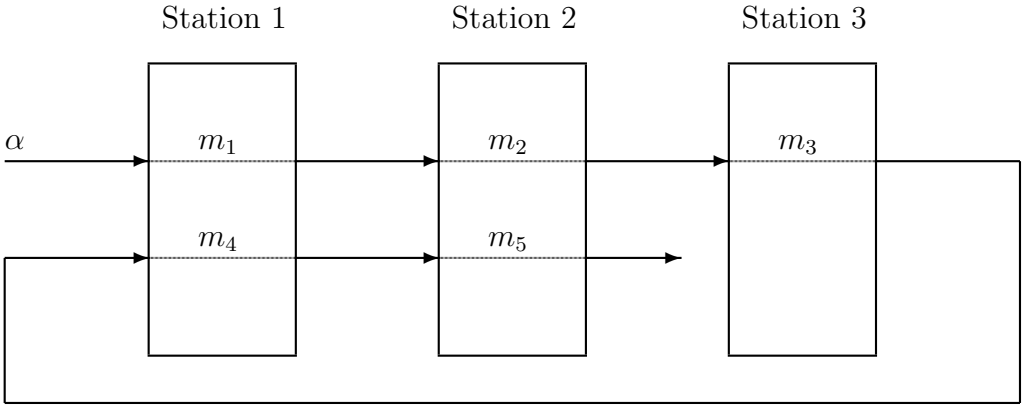

Figure 12: The five-class network obtained by deleting class 6 from the six-class fluid network

where, as before, $Q_k^+(t) = \sum_{\ell=1}^{k} Q_\ell(t)$. If, for each non-idling fluid solution $(Q(\cdot), T(\cdot))$ of the network, f_1, f_2 and f_3 satisfy conditions (6.4.1)–(6.4.3), it follows from the proof of Lemma 6.4.1 that the fluid network in Figure 11 is globally stable.

Mimicking the proof of Lemma 6.4.2, (6.4.1)–(6.4.3) hold if there is $x = (x_1, \ldots, x_5) > 0$ satisfying

$$\lambda(x_1 + x_4) < x_1\mu_1, \tag{6.6.25}$$

$$\lambda(x_1 + x_4) < x_4\mu_4, \tag{6.6.26}$$

$$\lambda(x_2 + x_5) < x_2\mu_2, \tag{6.6.27}$$

$$\lambda(x_2 + x_5) < x_5\mu_5, \tag{6.6.28}$$

$$\lambda x_3 < x_3\mu_3, \tag{6.6.29}$$

$$x_4 \leq x_3, \tag{6.6.30}$$

$$x_5 \leq x_4, \tag{6.6.31}$$

$$x_2 + x_5 \leq x_1 + x_4, \tag{6.6.32}$$

$$x_3 \leq x_2 + x_5. \tag{6.6.33}$$

Employing the techniques used in the proof of Lemma 6.4.3, we conclude that there is $x > 0$

satisfying (6.6.25)–(6.6.33) if and only if

$$\lambda(m_1 + m_4) < 1,$$

$$\lambda(m_2 + m_5) < 1,$$

$$\lambda m_3 < 1.$$

This proves the lemma for the network in Figure 11. □

The corresponding result for the network in Figure 12 follows immediately from re-numbering the stations.

Corollary 6.6.3. *The five-class three-station fluid network in Figures 12 is globally stable so long as the traffic intensity at each station is less than one.*

Lemma 6.6.4. *The stability region for any non-idling discipline that gives priority to class 3 over class 6 is* $\mathcal{D}_0$.

Proof. Consider $m \in \mathcal{D}_0$. Any fluid solution $(Q(\cdot), T(\cdot))$ under the priority discipline satisfies (6.1.1)–(2.3.6). In addition, $(Q(\cdot), T(\cdot))$ satisfies $\dot{T}_3(t) = 1$ for each regular point t such that $Q_3(t) > 0$. Therefore, $(Q_1(t), \ldots, Q_5(t))$ together with $(T_1(t), \ldots, T_5(t))$ is a fluid solution to the five-class fluid network in Figure 12 and, by Corollary 6.6.3, there exists $\delta > 0$ such that $(Q_1(t), \ldots, Q_5(t)) = 0$ for $t \geq \delta$. After δ, the input rate to buffer 6 is λ. If $Q_6(t) > 0$ for a regular point $t > \delta$, the departure rate d_6 from buffer 6 satisfies $\lambda m_3 + d_6 m_6 = 1$. Thus, $d_6 = \mu_6(1 - \lambda m_3)$, which is faster than the input rate λ. Hence buffer 6 will be empty

by

$$\frac{Q_6(0)+\delta}{\mu_6(1-\lambda m_3)-\lambda}.$$

Therefore, m is in the stability region. □

Lemma 6.6.5. *The stability region for any non-idling discipline that gives priority to class 1 over class 4 is* $\mathcal{D}_0$.

Proof. Consider $m \in \mathcal{D}_0$. Any fluid solution $(Q(\cdot), T(\cdot))$ under the priority discipline satisfies (6.1.1)–(2.3.6). In addition, $(Q(\cdot), T(\cdot))$ satisfies $\dot{T}_1(t) = 1$ for each regular point t such that $Q_1(t) > 0$. Because $\lambda m_1 < 1$, $Q_1(t) = 0$ for $t \geq \delta_0 = Q_1(0)/(\mu_1 - \lambda)$. For notational convenience, we assume $Q_1(0) = 0$ and hence $\delta_0 = 0$. From (6.1.1)–(6.1.4), we have $\mu_1 T_1(t) = \lambda t$ and hence

$$Q_2(t) = Q_2(0) + \lambda t - \mu_2 T_2(t),$$

$$Q_3(t) = Q_3(0) + \mu_2 T_2(t) - \mu_3 T_3(t),$$

$$Q_4(t) = Q_4(0) + \mu_3 T_3(t) - \mu_4 T_4(t),$$

$$Q_5(t) = Q_5(0) + \mu_4 T_4(t) - \mu_5 T_5(t),$$

$$Q_6(t) = Q_6(0) + \mu_5 T_5(t) - \mu_6 T_6(t),$$

and

$$\dot{T}_2(t) + \dot{T}_5(t) = 1 \qquad \text{if } Q_2(t) + Q_5(t) > 0,$$

$$\dot{T}_3(t) + \dot{T}_6(t) = 1 \qquad \text{if } Q_3(t) + Q_6(t) > 0,$$

$$\lambda m_1 + \dot{T}_4(t) = 1 \qquad \text{if } Q_4(t) > 0.$$

Let $\tilde{T}_4 = T_4(t)/(1 - \lambda m_1)$, $\tilde{m}_4 = m_4/(1 - \lambda m_1)$ and $\tilde{\mu}_4 = 1/\tilde{m}_4$. Then, we have

$$Q_2(t) = Q_2(0) + \lambda t - \mu_2 T_2(t),$$

$$Q_3(t) = Q_3(0) + \mu_2 T_2(t) - \mu_3 T_3(t),$$

$$Q_4(t) = Q_4(0) + \mu_3 T_3(t) - \tilde{\mu}_4 \tilde{T}_4(t),$$

$$Q_5(t) = Q_5(0) + \tilde{\mu}_4 \tilde{T}_4(t) - \mu_5 T_5(t),$$

$$Q_6(t) = Q_6(0) + \mu_5 T_5(t) - \mu_6 T_6(t),$$

and

$$\dot{T}_2(t) + \dot{T}_5(t) = 1 \qquad \text{if } Q_2(t) + Q_5(t) > 0,$$

$$\dot{T}_3(t) + \dot{T}_6(t) = 1 \qquad \text{if } Q_3(t) + Q_6(t) > 0,$$

$$\dot{\tilde{T}}_4(t) = 1 \qquad \text{if } Q_4(t) > 0.$$

Therefore $(Q_2(t), \ldots Q_6(t))$ together with $(T_2(t), T_3(t), \tilde{T}_4(t), T_5(t), T_6(t))$ is a fluid solution to the five-class fluid network in Figure 11 with service times $(m_2, m_3, \tilde{m}_4, m_5, m_6)$. Since $m \in \mathcal{D}_0$, we have

$$\lambda \tilde{m}_4 < 1,$$

$$\lambda(m_2 + m_5) < 1,$$

$$\lambda(m_3 + m_6) < 1.$$

It follows from Lemma 6.6.2 that $(Q_2(t), \ldots, Q_6(t)) = 0$ for $t > \delta$ for some $\delta > 0$. □

Proof of Theorem 6.2.4. **Part (a):** By Lemma 6.6.4, $\mathcal{D}_\pi = \mathcal{D}_0$ for $\pi = \pi_{\{1,2,3\}}, \pi_{\{1,5,3\}}, \pi_{\{4,2,3\}}, \pi_{\{4,5,3\}}$. By Lemma 6.6.5, $\mathcal{D}_\pi = \mathcal{D}_0$ for $\pi = \pi_{\{1,2,6\}}, \pi_{\{1,5,6\}}$. The static buffer priority discipline $\pi_{\{4,5,6\}}$ corresponds to the last-buffer-first-served priority discipline, whose stability region Dai and Weiss [17] showed to be $\mathcal{D}_0$.

Part (b): Let $\lambda = 1$. In Chapter 5 we proved that under the preemptive-resume priority discipline $\pi_{\{4,2,6\}}$ in the corresponding queueing network, classes 2, 4 and 6 constitute a pseudostation, in which at most two classes of jobs can be processed simultaneously. As a consequence, if $m_2 + m_4 + m_6 > 2$, then the total number of jobs in the system grows linearly with time. Furthermore, any fluid limit as taken in Dai [13] grows linearly with time. Because such a fluid limit is a fluid solution to equations (6.1.1)–(2.3.6), the fluid model is unstable under the discipline. The service time vector,

$$m = (0.1, 0.8, 0.1, 0.8, 0.1, 0.8),$$

for example, is in $\mathcal{D}_0$, but since $m_2 + m_4 + m_6 = 2.4 > 2$, $m \notin \mathcal{D}_\pi$.

Part (c): Let $\lambda = 1$ and consider the service time vector

$$m = (0.1, 0.8, 0.1, 0.45, 0.1, 0.45).$$

It is easy to check that

$$m_4 > m_3,$$

$$\lambda m_2 + \lambda^2 m_4 m_6 = 1.0025 > 1.$$

and so, by Theorem 6.2.2, m is not in the global stability region.

We now proceed to show that

$$x = (139, 139, 59, 63, 27, 27)$$

satisfies the linear constraints in (6.6.10)–(6.6.13) with $\epsilon = 1$. Hence, $m \in \mathcal{D}_{\pi_{\{4,2,6\}}}$, thus completing the proof.

Recall that to generate the vectors $d^s \in \mathcal{T}_{\pi_{\{4,2,6\}}}$, we solve (6.6.3)–(6.6.9) for each of the possible cases. These cases reduce to the following three at each station:

1. The higher priority buffer has positive fluid level,
2. Only the lower priority buffer has positive fluid level,
3. Both buffers are empty.

These three cases at each of the three stations lead to the 26 cases listed in Table 2 (there is no need to consider the case in which all the buffers are empty).

Case	Station A	Station B	Station C
1	None	None	3
2	None	None	6
3	None	5	None
4	None	5	3
5	None	5	6
6	None	2	None
7	None	2	3
8	None	2	6
9	1	None	None
10	1	None	3
11	1	None	6
12	1	5	None
13	1	5	3
14	1	5	6
15	1	2	None
16	1	2	3
17	1	2	6
18	4	None	None
19	4	None	3
20	4	None	6
21	4	5	None
22	4	5	3
23	4	5	6
24	4	2	None
25	4	2	3
26	4	2	6

Table 2: Enumeration of 26 states: each state corresponds to a different set of highest priority non-empty buffers.

If the solution d for a case (and the solution is unique for each case) does not satisfy

$$d_1 m_1 + d_4 m_4 \leq 1, \tag{.34}$$

$$d_2 m_2 + d_5 m_5 \leq 1, \tag{.35}$$

$$d_3 m_3 + d_6 m_6 \leq 1, \text{ and} \tag{.36}$$

$$d_i \geq 0 \text{ for } i = 1, 2, \ldots, 6 \tag{.37}$$

then the corresponding state is not feasible and hence not in $\mathcal{T}_{\pi_{\{4,2,6\}}}$. Otherwise, we include d in $\mathcal{T}_{\pi_{\{4,2,6\}}}$. Table 3 shows the departure rates d in each case and Table 4 shows the departure rates for the service times $m = (0.1, 0.8, 0.1, 0.45, 0.1, 0.45)$ used in Part (c) of the proof of Theorem 6.2.4. Table 5 shows both the rates of change in the buffer levels $\dot{Q}$ and the value of

$$df(x, Q(t))/dt = \sum_{k=1}^{6} \dot{Q}_k x_k,$$

where $x = (139, 139, 59, 63, 27, 27)$, for each regular state. This demonstrates that $f(x, Q(t))$ is a linear Lyapunov function proving the network is stable under the static buffer priority discipline $\pi_{\{4,2,6\}}$. □

Case	Departure Rate
1	$d_2 = d_1, d_4 = d_3, d_6 = d_5, d_1 = \lambda, d_5 = 1/(m_3 + m_6), d_3 = d_5$
2	$d_3 = d_4 = 0, d_2 = d_1, d_6 = \mu_6, d_1 = \lambda, d_5 = 0,$
3	$d_2 = d_1, d_4 = d_3, d_6 = d_5, d_1 = \lambda, d_3 = \lambda, d_5 = \mu_5(1 - \lambda m_2)$
4	$d_2 = d_1, d_4 = d_3, d_6 = d_5, d_1 = \lambda, d_5 = \mu_5(1 - \lambda m_2),$
5	$d_3 = d_4 = 0, d_2 = d_1, d_6 = \mu_6, d_1 = \lambda, d_5 = \mu_5(1 - \lambda m_2)$
6	$d_5 = d_6 = 0, d_2 = \mu_2, d_4 = d_3, d_3 = \mu_2, d_1 = \lambda,$
7	$d_5 = d_6 = 0, d_2 = \mu_2, d_4 = d_3, d_3 = \mu_3, d_1 = \lambda$
8	$d_3 = d_4 = d_5 = 0, d_2 = \mu_2, d_6 = \mu_6, d_1 = \lambda$
9	$d_2 = d_1, d_4 = d_3, d_6 = d_5, d_1 = 1/(m_1 + m_4), d_3 = d_1, d_5 = d_1$
10	$d_2 = d_1, d_4 = d_3, d_6 = d_5, d_3 = 1/(m_3 + m_6), d_5 = d_4, d_1 = \mu_1(1 - d_3 m_4)$
11	$d_3 = d_4 = 0, d_2 = d_1, d_6 = \mu_6, d_1 = \mu_1, d_5 = 0$
12	$d_2 = d_1, d_4 = d_3, d_6 = d_5, d_1 = 1/(m_1 + m_4), d_3 = d_1, d_5 = \mu_5(1 - d_1 m_2)$
13	$d_1 m_1 + d_3 m_4 = 1, d_1 m_2 + d_5 m_5 = 1, d_3 m_3 + d_5 m_6 = 1, d_2 = d_1, d_4 = d_3, d_6 = d_5$
14	$d_3 = d_4 = 0, d_2 = d_1, d_6 = \mu_6, d_1 = \mu_1, d_5 = \mu_5(1 - \mu_1 m_2)$
15	$d_5 = d_6 = 0, d_2 = \mu_2, d_4 = d_3, d_3 = \mu_2, d_1 = \mu_1(1 - \mu_2 m_4)$
16	$d_5 = d_6 = 0, d_2 = \mu_2, d_4 = d_3, d_3 = \mu_3, d_1 = \mu_1(1 - \mu_3 m_4)$
17	$d_3 = d_4 = d_5 = 0, d_2 = \mu_2, d_6 = \mu_6, d_1 = \mu_1$
18	$d_1 = d_2 = 0, d_4 = \mu_4, d_6 = d_5, d_5 = \mu_4, d_3 = 0,$
19	$d_1 = d_2 = 0, d_4 = \mu_4, d_6 = d_5, d_5 = \mu_4, d_3 = \mu_3(1 - \mu_4 m_6)$
20	$d_1 = d_2 = d_3 = 0, d_4 = \mu_4, d_6 = \mu_6, d_5 = \mu_4$
21	$d_1 = d_2 = 0, d_4 = \mu_4, d_6 = d_5, d_5 = \mu_5, d_3 = 0$
22	$d_1 = d_2 = 0, d_4 = \mu_4, d_6 = d_5, d_5 = \mu_5, d_3 = \mu_3(1 - \mu_5 m_6)$
23	$d_1 = d_2 = d_3 = 0, d_4 = \mu_4, d_6 = \mu_6, d_5 = \mu_5$
24	$d_1 = d_5 = d_6 = 0, d_2 = \mu_2, d_4 = \mu_4, d_3 = \mu_2$
25	$d_1 = d_5 = d_6 = 0, d_2 = \mu_2, d_4 = \mu_4, d_3 = \mu_3$
26	$d_1 = d_3 = d_5 = 0, d_2 = \mu_2, d_4 = \mu_4, d_6 = \mu_6$

Table 3: The departure rates for all 6 classes in all the states of the three-station fluid network under the static priority discipline $\pi_{\{4,2,6\}}$. Each state is characterized by giving the highest priority non-empty buffer (if any) at each station as indicated in Table 2.

Case	Departure Rate d_1	d_2	d_3	d_4	d_5	d_6	Busy Fraction A	B	C	feasible?
1	1.00	1.00	1.82	1.82	1.82	1.82	0.92	0.98	1.00	yes
2	1.00	1.00	0.00	0.00	0.00	2.22	0.10	0.80	1.00	yes
3	1.00	1.00	1.00	1.00	2.00	2.00	0.55	1.00	1.00	yes
4	1.00	1.00	1.00	1.00	2.00	2.00	0.55	1.00	1.00	yes
5	1.00	1.00	0.00	0.00	2.00	2.22	0.10	1.00	1.00	yes
6	1.00	1.25	1.25	1.25	0.00	0.00	0.66	1.00	0.13	yes
7	1.00	1.25	10.00	10.00	0.00	0.00	**4.60**	1.00	1.00	no
8	1.00	1.25	0.00	0.00	0.00	2.22	0.10	1.00	1.00	yes
9	1.82	1.82	1.82	1.82	1.82	1.82	1.00	**1.64**	1.00	no
10	1.82	1.82	1.82	1.82	1.82	1.82	1.00	**1.64**	1.00	no
11	10.00	10.00	0.00	0.00	0.00	2.22	1.00	**8.00**	1.00	no
12	1.82	1.82	1.82	1.82	**-4.55**	**-4.55**	1.00	1.00	-1.86	no
13	1.03	1.03	1.99	1.99	1.78	1.78	1.00	1.00	1.00	yes
14	10.00	10.00	0.00	0.00	**-70.00**	2.22	1.00	1.00	1.00	no
15	4.38	1.25	1.25	1.25	0.00	0.00	1.00	1.00	0.13	yes
16	**-35.00**	1.25	10.00	10.00	0.00	0.00	1.00	1.00	1.00	no
17	10.00	1.25	0.00	0.00	0.00	2.22	1.00	1.00	1.00	yes
18	0.00	0.00	0.00	2.22	2.22	2.22	1.00	0.22	1.00	yes
19	0.00	0.00	0.00	2.22	2.22	2.22	1.00	0.22	1.00	yes
20	0.00	0.00	0.00	2.22	2.22	2.22	1.00	0.22	1.00	yes
21	0.00	0.00	0.00	2.22	10.00	10.00	1.00	1.00	**4.50**	no
22	0.00	0.00	**-35.00**	2.22	10.00	10.00	1.00	1.00	1.00	no
23	0.00	0.00	0.00	2.22	10.00	2.22	1.00	1.00	1.00	yes
24	0.00	1.25	1.25	2.22	0.00	0.00	1.00	1.00	0.13	yes
25	0.00	1.25	10.00	2.22	0.00	0.00	1.00	1.00	1.00	yes
26	0.00	1.25	0.00	2.22	0.00	2.22	1.00	1.00	1.00	yes

Table 4: The departure rates for all 6 classes in all the states of the three-station fluid network with processing times $m = (0.1, 0.8.0.1, 0.45, 0.1, 0.45)$ under the static priority discipline $\pi_{\{4,2,6\}}$. Each state is characterized by giving the highest priority non-empty buffer (if any) at each station as indicated in Table 2. A state is feasible if the departure rates are non-negative and at most 100% of each server's time is allocated. Values preventing states from being feasible are indicated with boldfaced type.

Case	$\dot{Q}_1$	$\dot{Q}_2$	$\dot{Q}_3$	$\dot{Q}_4$	$\dot{Q}_5$	$\dot{Q}_6$	$\sum_k x_k \dot{Q}_k$
1	0.00	0.00	-0.82	0.00	0.00	0.00	-48.27
2	0.00	0.00	1.00	0.00	0.00	-2.22	-1.00
3	0.00	0.00	0.00	0.00	-1.00	0.00	-27.00
4	0.00	0.00	0.00	0.00	-1.00	0.00	-27.00
5	0.00	0.00	1.00	0.00	-2.00	-0.22	-1.00
6	0.00	-0.25	0.00	0.00	1.25	0.00	-1.00
8	0.00	-0.25	1.25	0.00	0.00	-2.22	-21.00
13	-0.03	0.00	-0.97	0.00	0.21	0.00	-55.05
15	-3.38	3.13	0.00	0.00	1.25	0.00	-1.00
17	-9.00	8.75	1.25	0.00	0.00	-2.22	-21.00
18	1.00	0.00	0.00	-2.22	0.00	0.00	-1.00
19	1.00	0.00	0.00	-2.22	0.00	0.00	-1.00
20	1.00	0.00	0.00	-2.22	0.00	0.00	-1.00
23	1.00	0.00	0.00	-2.22	-7.78	7.78	-1.00
24	1.00	-1.25	0.00	-0.97	2.22	0.00	-36.00
25	1.00	-1.25	-8.75	7.78	2.22	0.00	-1.00
26	1.00	-1.25	1.25	-2.22	2.22	-2.22	-101.00

Table 5: Rates of change in the buffer levels for the 17 feasible states in the three-station fluid network with processing times $m = (0.1, 0.8.0.1, 0.45, 0.1, 0.45)$ under the static priority discipline $\pi_{\{4,2,6\}}$. The last column computes $\sum_{k=1}^{6} \dot{Q}_k x_k$ where $x = (139, 139, 59, 63, 27, 27)$. This shows that the network is stable under the discipline $\pi_{\{4,2,6\}}$.

Chapter 7

Conclusions

While we expect that the results presented in the last few chapters will shed some light on the stability properties of complex multiclass queueing networks and the control and analysis of associated real-world systems, there is clearly much more to be done to gain a full understanding of this area. We now review some questions and directions for further research that arise from our investigations in the previous chapters. Certainly this is not meant to be a comprehensive review all the open issues related to stability, capacity, and scheduling. Our presentation somewhat follows the chronology of the main chapters in this dissertation.

In Chapter 4 we derived necessary and sufficient stability conditions for two-station fluid networks, provided that these networks were restricted to a class called *acyclic transfer mechanism networks* (ACTN's), essentially a class of networks that does not allow revisits to a class. In fact, our results could actually be extended to any fluid network which allows only a finite number revisits to a station, since such a network could be equivalently relabeled as an ACTN. Practically speaking, this should be a perfectly satisfactory class of networks to model real-life situations (since jobs rarely make an infinite number of revisits in the factory). However, this still leaves one with a theoretical yearning to extend the theory to the full class of two-station fluid OMQN's.

Unfortunately, the methods of Chapter 4 cannot be directly extended to this class of

networks. Although hidden in the analysis, our ability to transform the LP of Section 4.4 into a network flow problem relies on the fact that in any SBN, the classes can be labeled such that the routing matrix P is upper triangular. One possible way to overcome this difficulty is to model the OMQN as a SBN with an infinite number of classes. In this case, we can again write down a similar LP and transform it into a network flow problem with an infinite number of nodes. Of course, the problem then lies in showing that the capacity of this network is determined by only a finite number of cut conditions. While this route seems promising, it appears difficult to carry out. If this analysis were successful, it would essentially complete our understanding of the global stability properties of two-station fluid networks and fully complement the results of Bertsimas, et. al. [2].

A further point to be noted is that the results of Chapter 4 are valid only for fluid networks. For the corresponding class of queueing networks, our knowledge is still lacking. Dai and VandeVate's [15] results imply that our conditions are in fact necessary and sufficient for a discrete two-station ACTN **if** the stability conditions involve virtual stations only (no "pushstart conditions"). In this respect, the full connection between the stability of the stochastic and fluid models needs yet to be explored.

Chapter 5 provides necessary conditions for multi-station queueing and fluid MTN's to be globally stable. Actually, with some modifications of the notation and proofs, these results should easily carry through to the full class of OMQN's. Unfortunately, the results of Chapter 6, specifically Section 6.3, imply that the conditions we derived in general are not sufficient for stability. Despite this shortfall, it should be noted that it seems apparent that classes that form a pseudostation are important in determining the stability properties of multi-station networks, as is evidenced in the analysis of the three-station network of Chapter 6.

We can view the results of Chapter 6 as both a success and a discouragement. As such,

our analysis there has numerous implications for further study. Our course, one primary disappointment is that the techniques of Chapter 4, which work so well to analyze two-station networks, do not directly carry over to three-station networks in that the piecewise linear Lyapunov function we used is no longer able to sharply determine the global stability region for larger networks. However, as we saw in that chapter, the same Lyapunov function is able to sharply determine the *monotone* global stability region for our network. A question that now arises is whether or not such a Lyapunov function can yield the monotone global stability region for a class of multi-station networks. We have undertaken a preliminary investigation of this issue for d-station networks with a similar routing structure to the network in Chapter 6. However, it is disappointing that in fact we were not able to obtain the actual stability region of our three-station example. It appears that new techniques, perhaps new classes of Lyapunov functions, will be needed to investigate this matter. Initially, it was hoped that the new class of *linear* Lyapunov functions, suggested by Chen and Zhang [11] would at least be useful for determining the static buffer priority stability regions for larger networks. Our analysis of Chapter 6 indicates that the linear constraints arising from such Lyapunov functions are not sharp in determining the priority stability region and thus it is unclear if this class of Lyapunov functions will be useful in further stability analyses.

We recall from the results of Chapter 4 that it is the static buffer priority policies that essentially determine the global stability behavior of two-station fluid networks. The analysis in Chapter 6 indicates that this is no longer so for three-station networks. This now begs the question as to which "extremal" policies determine stability characteristics for three-station fluid networks. The instability proofs of Section 6.3 indicate that a new type of policy, which could be viewed as a piecewise static buffer priority policy (i.e. a policy which is static between emptying times of the buffers) plays an important role in the stability behavior for multi-station networks. Even if we could define such a class of policies

that affect stability, it is no longer clear (as is the case with static buffer priority policies) what the discrete analog of such policies would be. Furthermore, we must consider whether global stability is a useful concept for multi-station networks if it is determined by policies too complex to be likely used in practice.

Of course the directions suggested above are only the beginning. Greater understanding in general of the connection between the fluid model and the discrete model is definitely needed. In fact, Bramson [**?**] recently provided an example of a network with exponential service and interarrival times that is stable, but whose corresponding fluid network is unstable. So, even in the "simplest" stochastic networks, the stability issues appear to be quite complex. Another important area of active research is adding setup and batching considerations to the basic OMQN model. Certainly, in semiconductor wafer fabs, setup times between classes and the choice of a good switching policy have a large influence on system performance. As a result, it appears that the stability and efficiency of networks with setup considerations is a ripe area for research (see Dai and Jennings [25] for example). Naturally adding other features to the model, further analyzing dispatch policies, and continuing the investigation into the fluid/discrete connection all would provide enough opportunities to keep researchers in this area busy for a long time to come.

Bibliography

[1] R. K. Ahuja, T. K. Magnanti, and J. B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, Englewood Cliffs, N.J., 1993.

[2] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. Stability conditions for multiclass fluid queueing networks. *IEEE Transactions on Automatic Control*, 41:1618–1631, 1996.

[3] D. D. Botvich and A. A. Zamyatin. Ergodicity of conservative communication networks. Rapport de recherche 1772, INRIA, 1992.

[4] M. Bramson. Instability of FIFO queueing networks. *Annals of Applied Probability*, 4:414–431, 1994.

[5] M. Bramson. Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems: Theory and Applications*, 22:5–45, 1996.

[6] M. Bramson. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems: Theory and Applications*, 23:1–26, 1996.

[7] H. Chen. Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines. *Annals of Applied Probability*, 5:637–665, 1995.

[8] H. Chen and A. Mandelbaum. Hierarchical modeling of stochastic networks I: fluid models. In D. D. Yao, editor, *Probability Models in Manufacturing Systems*, chapter 2, pages 47–106. Springer, New York, 1994.

[9] H. Chen and D. Yao. Stable priority disciplines for multiclass networks. In K. S. Paul Glasserman and D. Yao, editors, *Proceedings of Workshop on Stochastic Networks: Stability and Rare Events*, Columbia University, New York, 1996. Springer-Verlag.

[10] H. Chen and H. Zhang. Stability of multiclass queueing networks under FIFO service discipline. *Mathematics of Operations Research*, 22:691–725, 1997.

[11] H. Chen and H. Zhang. Stability of multiclass queueing networks under priority service disciplines. *Operations Research*, 48:26–37, 2000.

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

[13] J. G. Dai. A fluid-limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability*, 6:751–757, 1996.

[14] J. G. Dai and J. VandeVate. Global stability of two-station queueing networks. In K. S. Paul Glasserman and D. Yao, editors, *Proceedings of Workshop on Stochastic Networks: Stability and Rare Events*, pages 1–26, Columbia University, New York, 1996. Springer-Verlag.

[15] J. G. Dai and J. VandeVate. Virtual stations and the capacity of two-station queueing networks. 1999. Under revision for *Operations Research*.

[16] J. G. Dai and J. VandeVate. The stability of two-station multitype fluid networks. *Operations Research*, 48:721–744, 2000.

[17] J. G. Dai and G. Weiss. Stability and instability of fluid models for re-entrant lines. *Mathematics of Operations Research*, 21:115–134, 1996.

[18] D. Down and S. Meyn. Piecewise linear test functions for stability and instability of queueing networks. *Queueing Systems: Theory and Applications*, 27:205–226, 1997.

[19] V. Dumas. *Approches fluides pour la stabilité et l'instabilité de réseaux de files d'attente stochastiques à plusieurs classes de clients*. PhD thesis, L'école Polytechnique, Paris, France, 1996.

[20] V. Dumas. Essential faces and stability conditions of multiclass networks with priorities. Rapport de recherche 3030, INRIA, 1996.

[21] V. Dumas. A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems: Theory and Applications*, 25:1–43, 1997.

[22] M. El-Taha and S. Stidham Jr. Sample-path stability conditions for multiserver input-output processes. *Journal of Applied Mathematics and Stochastic Analysis*, 7:437–456, 1994.

[23] J. M. Harrison and V. Nguyen. Some badly behaved closed queueing networks. In F. P. Kelly and R. J. Williams, editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 117–124, New York, 1995. Springer-Verlag.

[24] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

[25] O. B. Jennings. On the stability of queueing networks with setups. 2001. Submitted to *Queueing Systems*.

[26] F. P. Kelly. Networks of queues with customers of different types. *J. Appl. Probab.*, 12:542–554, 1975.

[27] P. R. Kumar. Re-entrant lines. *Queueing Systems: Theory and Applications*, 13:87–110, 1993.

[28] P. R. Kumar and T. I. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control*, AC-35:289–298, 1990.

[29] S. Kumar and P. R. Kumar. Fluctuation smoothing policies are stable for stochastic reentrant lines. *Discrete Event Dynamical Systems*, 6:361–370, 1996.

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

[31] S. P. Meyn. Transience of multiclass queueing networks via fluid limit models. *Annals of Applied Probability*, 5:946–957, 1995.

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

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

[34] W. Whitt. Large fluctuations in a deterministic multiclass network of queues. *Management Sciences*, 39:1020–1028, 1993.

Vita

John Hasenbein was born on April 25, 1968 in Milwaukee, Wisconsin. Shortly thereafter, he traveled to St. Louis, Missouri to attend college at Washington University. Surprisingly, he was later awarded a Bachelor of Science degree in Systems Science and Mathematics in May, 1991. After a few months of soul searching, he hitched a bus to Atlanta and enrolled in the doctoral program in the department of Industrial and Systems Engineering at the Georgia Institute of Technology, where he received his Master of Science degree in Operations Research in March, 1995. He completed his Ph.D. work in September of 1998 and is next headed to the University of Texas at Austin and other foreign lands to initiate his career in academia.