J. G. "Jim" Dai

Stability of Fluid and Stochastic Processing Networks

January 3, 1999

ISSN 1398-5981 MPS-misc 1999-9

Web Sites
http://www.isye.gatech.edu/faculty/dai/

Preface

These notes were written for a concentrated course on "Queueing Network Theory" given in November 1988 at the Centre for Mathematical Physics and Stochastics (MaPhySto, `http://www.maphysto.dk/`) at the University of Aarhus, Denmark. The main subject is stability of fluid models and its connection with stability of queueing networks. There are many active research areas related to fluid models. Two such topics are notably missing in these notes. One is the optimal draining of a fluid network and its connection with dynamic control of the corresponding queueing network. The other concerns stability and state space collapse for a "critical" or "balanced" fluid model, and their connection with the Brownian approximation of the corresponding queueing network. Interested readers can find references for these topics at the end of Chapter 2.

These notes eventually will become two chapters in the book "Brownian models of Stochastic Processing Networks" being written by Jim Dai, Michael Harrison and Ruth Williams. A tentative table of contents of that book is included as an appendix. Please be aware that these notes are still preliminary. Some of the proofs are not complete. They will be completed and polished in the book. If you have comments or suggestions on these notes, please send them to dai@isye.gatech.edu.

I would like to acknowledge the financial support from MaPhySto and the Georgia Tech Foundation that made possible my sabbatical leave in Fall 1998. I would like to thank John Hasenbein from University of Texas at Austin for suggesting numerous improvements on the early drafts of these notes. Thanks also go to Soren Asmussen and Ole Barndorff-Nielsen for arranging my visit to Aarhus and to Wenjiang Jiang and Oddbjorg Wethelund for making our stay at Aarhus most enjoyable.

Jim Dai
December 28, 1998
Palo Alto, California

Contents

Chapter 1

Open Multiclass Networks

In this chapter, we introduce a class of stochastic processing networks called *multiclass queueing networks*. These networks may be used to model computer systems, telecommunication networks, and complex manufacturing systems like wafer fabrication facilities.

1.1 Informal Description of the Basic Model

For our purposes, there are J service stations in the network. Each station has a single server with unlimited waiting space. There are $K \geq J$ *classes* of jobs or customers. Each job arrives at the network from the outside, receives services at a number of stations and eventually leaves the network. Throughout the lifetime of the job in the network, the job belongs to one of the K job classes. The job changes classes as it moves through the network. It changes classes each time a service is completed on the job. All jobs within a class are served at a unique station. There can be *multi*ple job *class*es served at a station. The term *multiclass queueing network* means that there is at least one station serving more than one job class. Otherwise, the network is called a *single-class network*. An ordered sequence of classes that a job visits is called a *route*. We assume that jobs initially arriving at class k all follow that same route, at least probabilistically. Each job is assumed to eventually leave the network. Such a network is called an *open* queueing network.

An example of a multiclass queueing network is a manufacturing system having 3 stations producing 2 *types* of products as pictured in Figure 1.1. Type A jobs need 3 stages of processing, with stage i being processed at station i. Type B jobs need 6 stages of processing, with stages 1, 2 and 3 being processed at stations 1, 2 and 3, respectively, and stages 4, 5 and 6 being processed again at stations 1, 2 and 3, respectively. One can define a combination of *type* and *stage* as a *class* so that jobs within the same type and the same stage of processing are in the same class. For example, we can designate type A jobs in stages 1, 2 and 3 as in classes 1, 2 and 3, respectively, and type B jobs in stage i, as in

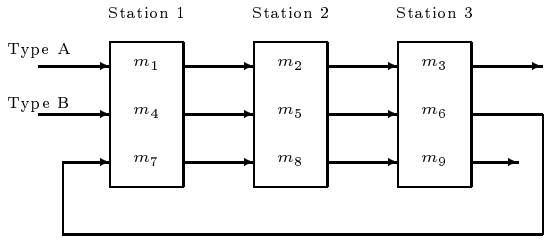

Figure 1.1: A 2-type 9-class network

class $3+i$, $i = 1, \ldots, 6$. For this network, $J = 3$ and $K = 9$. A job initially enters the network either as a class 1 or class 4 job. Jobs initially entering class 1 visit classes 1, 2 and 3 and then exit the network. Jobs initially entering class 4 visit classes 4 through 9 and then exit the network.

When there are multiple job classes served at a station, the server needs a policy to select the next job to work on. Such a policy is called a *service discipline*. For example, first-in-first-out (FIFO) is a popular service discipline.

1.2 Open Multiclass Queueing Networks

In this section, we define open multiclass queueing networks under a variety of service disciplines.

1.2.1 Primitive Cumulatives

The network under study has J single-server stations and K job classes. Stations are labelled $j = 1, \ldots, J$, and classes by $k = 1, \ldots, K$. Class k jobs are served at a unique station. We use $\mathcal{C}(j)$ to denote the set of classes belonging to station j, and $s(k)$ to denote the station to which class k belongs; when j and k appear together, we implicitly set $j = s(k)$.

For each class k, there are three *cumulative processes* $E_k = \{E_k(t), t \geq 0\}$, $V_k = \{V_k(n) : n = 1, 2, \ldots\}$ and $\Phi^k = \{\Phi^k(n) : n = 1, 2, \ldots\}$. For each time $t \geq 0$, $E_k(t)$ counts the number of *external* arrivals to class k in $[0, t]$. For each positive integer n, $V_k(n)$ is the total service requirement (in terms of the server's time) for the first n class k jobs. (When a preemption service discipline is used, preempt-resume is assumed. See service discipline section for more discussion.) For each positive integer n, $\Phi^k(n)$ is a K-dimensional vector taking values in $\mathbb{Z}_+^K$. For each class ℓ, $\Phi_\ell^k(n)$ is the total number of jobs going to class ℓ among the first n jobs finishing services at class k. By convention, we assume

$$E_k(0) = 0, \quad V_k(0) = 0 \quad \text{and} \quad \Phi^k(0) = 0.$$

For each time $t \geq 0$, we extend the definitions of $V_k(t)$ and $\Phi^k(t)$ as

$$V_k(t) = V_k(\lfloor t \rfloor) \quad \text{and} \quad \Phi^k(t) = \Phi^k(\lfloor t \rfloor),$$

where $\lfloor t \rfloor$ denotes the largest integer less than or equal to t. Note that the processes V_k and Φ^k are right continuous having left limits in time. We also assume that E_k is right continuous having left limits. As we have seen in Figure 1.1, not all classes have external arrivals. We use $\mathcal{E}$ to denote the set of classes that have external arrivals. By convention, for $k \notin \mathcal{E}$, $E_k(t) = 0$ for all $t \geq 0$. We denote $E = \{E(t), t \geq 0\}$, $V = \{V(t), t \geq 0\}$, $\Phi = \{\Phi(t), t \geq 0\}$ with $E(t) = (E_1(t), \ldots, E_K(t))'$, $V(t) = (V_1(t), \ldots, V_K(t))'$ and $\Phi(t) = (\Phi^k_\ell(t), k, \ell = 1, \ldots, K)$. (All vectors are envisioned as column vectors and prime denotes transpose.) The triple (E, V, Φ) is called the *primitive cumulatives*. It is a part of the specification of the network. We assume that the *strong law of large numbers* holds for the primitive cumulatives, namely, with probability one,

$$\lim_{t\to\infty} E(t)/t = \alpha, \quad \lim_{n\to\infty} V(n)/n = m \quad \text{and} \quad \lim_{n\to\infty} \Phi(n)/n = P, \tag{1.1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and $m = (m_1, \ldots, m_K)'$ are K-dimensional vectors, and $P = (P_{k\ell})$ is a $K \times K$ matrix. For each class k, m_k is the mean service time for class k jobs and α_k is the external arrival rate to class k. For classes k and ℓ, $P_{k\ell}$ is the long-run fraction of jobs departing class k which become class ℓ jobs. It is also called the routing probability from class k to class ℓ. The $K \times K$ matrix $P = (P_{k\ell})$ is the *routing matrix* of the network. We assume that the network is *open*, i.e., the matrix

$$Q \stackrel{\text{def}}{=} I + P' + (P')^2 + \ldots$$

is finite, which is equivalent to the fact that $(I - P')$ is invertible and $Q = (I - P')^{-1}$.

Assumption (1.1) is the minimal assumption we impose on the network. It allows basic terms like arrival rates, average service times and routing probabilities to be properly defined. Proposition 1.2.4 below strengthens assumption (1.1) to a *functional* strong law of large numbers. Before stating the proposition, we need to introduce the Skorohod path space.

For an integer $d \geq 1$, recall that $\mathbb{D}^d[0, \infty)$ is a set of functions $x : [0, \infty) \to \mathbb{R}^d$ that are right continuous on $[0, \infty)$ having the left limits on $(0, \infty)$. For $t > 0$, we use $x(t-)$ to denote $\lim_{s\uparrow t} x(s)$. By convention, $x(0-) = x(0)$.

We endow the function space $\mathbb{D}^d[0, \infty)$ with the Skorohod J_1-topology, or simply the Skorohod topology. We will not introduce a metric on $\mathbb{D}^d[0, \infty)$ that induces the Skorohod topology here. Rather we provide an equivalent definition for a sequence of functions in $\mathbb{D}^d[0, \infty)$ to converge. We use Λ to denote the set of strictly increasing, continuous functions $x : \mathbb{R}_+ \to \mathbb{R}_+$ such that $x(0) = 0$ and $\lim_{t\to\infty} x(t) = \infty$.

Definition 1.2.1. A sequence $\{x^n\} \subset \mathbb{D}^d[0, \infty)$ is said to converge to $x \in \mathbb{D}^d[0, \infty)$ in the Skorohod topology if for each $t > 0$, there exists $\{\gamma^n\} \subset \Lambda$

(possibly depending on t) such that

$$\lim_{n\to\infty} \sup_{0\le s\le t} |\gamma^n(s) - s| = 0, \tag{1.2}$$

$$\lim_{n\to\infty} \sup_{0\le s\le t} |x^n(\gamma^n(s)) - x(s)| = 0. \tag{1.3}$$

Definition 1.2.2. For a sequence of functions $\{x^n\} \subset \mathbb{D}^d[0,\infty)$, the sequence is said to converge uniformly on compact intervals (u.o.c.) to $x \in \mathbb{D}^d[0,\infty)$ as $n \to \infty$, denoted by $x^n \to x$ u.o.c., if for each $t > 0$,

$$\lim_{n\to\infty} \sup_{0\le s\le t} |x^n(s) - x(s)| = 0.$$

Clearly, if $x^n \to x$ u.o.c., x^n converges to x in the Skorohod topology. The converse is not true in general. Consider, for example, $x^n(t) = 1$ for $t \in [1/2 + 1/n, \infty)$ and $x^n(t) = 0$ for $t \in [0, 1/2+1/n)$. It can be shown that $x^n \to x$ in the Skorohod topology, where $x(t) = 1$ for $t \in [1/2, \infty)$ and $x(t) = 0$ for $t \in [0, 1/2)$. Clearly, x^n does not converge uniformly on compact sets to x. When the limit point x is continuous, the two notions of convergence are equivalent. This fact will repeatedly be used in the remainder of this chapter. Let $\mathbb{C}^d[0,\infty)$ be the set of continuous functions $x : [0,\infty \to \mathbb{R}^d$.

The following lemma is needed to prove Proposition 1.2.4.

Lemma 1.2.3. *Let $\{f_n\}$ be a sequence of nondecreasing functions on $\mathbb{R}_+$ and f be a continuous function on $\mathbb{R}_+$. Assume that $f_n(t) \to f(t)$ for all rational $t \ge 0$. Then $f_n \to f$ u.o.c.*

Proof. First, because f_n is nondecreasing and f is continuous, one can easily check that $f_n(t) \to f(t)$ for every $t \in \mathbb{R}_+$. Next, suppose that f_n does not converge to f uniformly on compact sets. Then there exist $\epsilon > 0$, $t > 0$ and $\{t_{n_l}\}$ such that $t_{n_l} \le t$ and

$$|f_{n_l}(t_{n_l}) - f(t_{n_l})| \ge \epsilon \quad \text{for all } l. \tag{1.4}$$

Because $\{t_{n_l}\}$ is bounded, we may assume that $t_{n_l} \to t_0 \le t$. Thus for any $\delta > 0$, t_{n_l} eventually is less than $t_0 + \delta$. Hence for l large enough,

$$\begin{aligned} f_{n_l}(t_{n_l}) - f(t_{n_l}) &\le f_{n_l}(t_0+\delta) - f(t_{n_l}) \\ &= f_{n_l}(t_0+\delta) - f(t_0+\delta) + f(t_0+\delta) - f(t_0) + f(t_0) - f(t_{n_l}). \end{aligned}$$

Therefore,

$$\limsup_{l\to\infty} (f_{n_l}(t_{n_l}) - f(t_{n_l})) \le f(t_0+\delta) - f(t_0).$$

Because f is continuous and δ is arbitrary, we have

$$\limsup_{l\to\infty} (f_{n_l}(t_{n_l}) - f(t_{n_l})) \le 0.$$

When $t_0 > 0$, one can similarly prove that

$$\liminf_{l\to\infty} (f_{n_l}(t_{n_l}) - f(t_{n_l})) \ge 0.$$

When $t_0 = 0$,

$$\liminf_{l\to\infty} (f_{n_l}(t_{n_l}) - f(t_{n_l})) \geq \lim_{l\to\infty} (f_{n_l}(0) - f(t_{n_l})) = 0.$$

Thus we have

$$\lim_{l\to\infty} (f_{n_l}(t_{n_l}) - f(t_{n_l})) = 0,$$

which contradicts (1.4). Hence the lemma is proved. □

For each class k, $r > 0$ and $t \geq 0$, define

$$\bar{E}^r(t) = \frac{1}{r}E(rt), \quad \bar{V}^r(t) = \frac{1}{r}V(rt), \quad \text{and} \quad \bar{\Phi}^{k,r}(t) = \frac{1}{r}\Phi^k(rt).$$

For each $r > 0$, the processes $\bar{E}^r$, $\bar{V}^r$ and $\bar{\Phi}^{k,r}$ take values in the Skorohod path space $\mathbb{D}^K[0,\infty)$. With a slight abuse of notation, we define functions $\alpha_k(\cdot)$, $m_k(\cdot)$ and $P_{k\ell}(\cdot)$ by

$$\alpha_k(t) = \alpha_k t, \quad m_k(t) = m_k t \quad \text{and} \quad P_{k\ell}(t) = P_{k\ell}\, t \quad \text{for } t \geq 0.$$

The following functional strong law of large numbers follows from Lemma 1.2.3 immediately.

Proposition 1.2.4. *Assume that (1.1) holds. With probability one, as $r \to \infty$,*

$$\bar{E}^r_k(\cdot) \to \alpha_k(\cdot) \quad u.o.c., \quad k = 1, \ldots, K, \tag{1.5}$$

$$\bar{V}^r_k(\cdot) \to m_k(\cdot) \quad u.o.c., \quad k = 1, \ldots, K, \tag{1.6}$$

$$\bar{\Phi}^{k,r}_\ell(\cdot) = P_{k\ell}(\cdot) \quad u.o.c., \quad k, \ell = 1, \ldots, K. \tag{1.7}$$

Often the network assumptions are more naturally imposed on incremental random variables. For this purpose, we let $u_k = \{u_k(i), i \geq 1\}$ and $v_k = \{v_k(i), i \geq 1\}$ be two sequence of nonnegative random variables, and $\phi^k = \{\phi^k(i), i \geq 1\}$ be a sequence of random vectors. For each i, $u_k(i)$ is the *interarrival time* between the $(i-1)$th and the ith *externally* arriving job to class k, and $v_k(i)$ is the *service time* for the ith class k job. We assume that the *routing vector* $\phi^k(i)$ takes values in $\{e_0, e_1, \ldots, e_K\}$, where e_0 is the K-dimensional vector of all 0's and for $\ell = 1, \ldots, K$, e_ℓ is the K-dimensional vector with the ℓth component 1 and all other components 0. When $\phi^k(i) = e_\ell$, the ith job departing class k becomes a class ℓ job, $\ell = 1, \ldots, K$. When $\phi^k(i) = e_0$, the ith class k job departs from the network.

The triple (u, v, ϕ) is called the *primitive increments*. Given primitive increments (u, v, ϕ), one can uniquely define primitive cumulatives (E, V, Φ) via the following equations: for each class k, each time $t \geq 0$ and each positive integer n,

$$E_k(t) = \max\{i : u_k(1) + \ldots + u_k(i) \leq t\}, \tag{1.8}$$

$$V_k(n) = \sum_{i=1}^{n} v_k(i), \tag{1.9}$$

$$\Phi^k(n) = \sum_{i=1}^{n} \phi^k(i). \tag{1.10}$$

When the primitive increments are independent, iid sequences, the network is called a *network with iid increments*. In this case, E_k is a *renewal process*, V_k and Φ^k are *random walks*. For a network with iid increments, assumption (1.1) is satisfied. If, furthermore, all interarrival and service time distributions are exponential, the network is called a *network with exponential increments* or simply an *exponential network*. When $K = J$, an exponential network is a *Jackson network*. For this reason, when $K = J$, we call a network with iid increments a *Jackson type network* or a generalized Jackson network. A *re-entrant line* is a multiclass queueing network in which exactly one class, say class 1, has external arrivals and transitions among classes are deterministic. Re-Entrant lines can be used to model many production systems including semiconductor wafer fabrication facilities.

1.2.2 Service Disciplines

A service discipline dictates the order in which jobs are served at each station. A service discipline is *non-idling* (or work conserving) if a server is never idle when there are jobs waiting to be served at its station. Examples of non-idling service disciplines include first-in-first-out (FIFO) and last-in-first-out (LIFO).

We restrict our disciplines to non-idling *head-of-the-line* (HL) disciplines. Under an HL service discipline, each class has at most one job (the leading job) receiving service at any given time. Jobs within a class are served on FIFO basis. Each class receives a proportion (possibly zero) of the associated server's time, where this proportion may be random but it is kept constant between changes in the arrival or departure processes. Furthermore, these proportions should depend in a measurable way on the "state" of the queueing network, and they should not anticipate (external) interarrival times, service times or routing vectors for future arrivals. The FIFO discipline is an HL discipline, whereas the LIFO discipline is *not*. We allow different stations to employ different service disciplines.

As said before, a popular service discipline is FIFO. Under FIFO jobs among *all* classes at a station are ranked according the *current* arrival time to the station. A family of service disciplines that has been studied extensively in the literature are *static buffer priority* (SBP) disciplines. Under a SBP discipline, classes at each station have a fixed ranking. For simplicity, we assume that the ranking is strict, i.e., there is no tie in the ranking. When the server switches attention from one job to another, the new job is taken from the head-of-the-line of the highest ranking non-empty class at the server's station. We consider *preempt-resume* static buffer priority service. Under this service, if a job arrives at a station having a higher rank than a job currently being served, the service of the current job is interrupted until service of all jobs with higher ranks is completed, at which time the interrupted service continues from where it left off (preempt then resume).

Many families of service disciplines fall under the name *processor sharing* (PS). Under such a discipline, several jobs may *simultaneously* share the server's service capacity. The portion of service capacity that each job receives may

differ from job to job. Different specifications of the proportions or *weights* give rise to different PS disciplines. When a PS service discipline is employed, each server's capacity is assumed to be infinitely divisible. This, of course, is a mathematical idealization that is rarely met by realistic models. However, such service disciplines approximate more "implementable PS" disciplines, like round-robin polling disciplines. PS disciplines are popular in telecommunication networks. They allow "fair" access to resources (servers). Short jobs can be completed even in the presence of long jobs.

The first family of PS disciplines are the generalized head-of-the-line processor sharing (GHLPS) disciplines. There is a *proportion vector* $\beta = (\beta_1, \ldots, \beta_K) > 0$ associated with a GHLPS discipline. The server simultaneously serves the jobs at the head-of-the-line of each (non-empty) class with service effort to class k in proportion to β_k. If there are no empty classes, the proportion of effort that each class receives is static (independent of the network dynamics). When some classes are empty, efforts to these empty classes are redistributed to the non-empty classes. When $\beta = (1, \ldots, 1)$, the GHLPS discipline is called the head-of-the-line processor sharing (HLPS) discipline.

The second family of PS disciplines are the generalized head-of-the-line *proportional* processor sharing (GHLPPS) disciplines. Under a GHLPPS service discipline with *weight vector* $\beta = (\beta_1, \ldots, \beta_K) > 0$, the server simultaneously serves the jobs at the head-of-the-line of each (non-empty) class. The server allocates its effort to each class in proportion to β_k multiplying the number of jobs in that class. Thus, the higher the jobcount in a class is, the more service effort that the leading job in the class receives. When the weight vector $\beta = (1, \ldots, 1)$, the GHLPPS discipline becomes the head-of-the-line proportional processor sharing (HLPPS) service discipline.

The earliest PS discipline studied allows *every* job in the network simultaneously receives service with equal proportions among all jobs. Such a discipline is *not* an HL discipline.

1.3 Performance Processes

The following descriptive processes Z, D, W, Y will be used to measure the performance of our queueing network. The processes $Z = \{Z(t), t \geq 0\}$ and $D = \{D(t), t \geq 0\}$ are K-dimensional with $Z_k(t)$ denoting the number of class k jobs that are in queue or being served at station $s(k)$ at time t and $D_k(t)$ denoting the number of departures from class k in $[0, t]$. They are called the *jobcount process* and *departure process*, respectively. The other two processes, $W = \{W(t), t \geq 0\}$ and $Y = \{Y(t), t \geq 0\}$, are J-dimensional. For each station j, $W_j(t)$ denotes the amount of work for server j (measured in units of remaining service time) embodied in those jobs who are at station j at time t. If no more arrivals (external or internal) are allowed to station j after time t, server j has to work an additional $W_j(t)$ units of time to finish her work. The process W is called the (immediate) *workload process*. For each station j, $Y_j(t)$ denotes the total amount of time that the sever at station j has been idle in the time

interval $[0,t]$. Y is called the *cumulative idletime process*. The jobcount and workload processes measure congestion and delay in the network. The idletime process measures utilization of the resources (servers) in the network.

For a class k, if there is a constant $\widetilde{\lambda}_k$ such that

$$\mathbb{P}\Big\{\lim_{t\to\infty} D_k(t)/t = \widetilde{\lambda}_k\Big\} = 1, \tag{1.11}$$

we call $\widetilde{\lambda}_k$ the *throughput* from class k. For a re-entrant line, $\widetilde{\lambda}_K$, if it exists, is the throughput of the system. Obviously, the throughput is less than or equal to the input rate α_1. If, for a station j, there is a constant $\widetilde{\rho}_j$ such that

$$\mathbb{P}\Big\{\lim_{t\to\infty} Y_j(t)/t = 1-\widetilde{\rho}_j\Big\} = 1, \tag{1.12}$$

we call $\widetilde{\rho}_j$ the *utilization* of server j and $1-\widetilde{\rho}_j$ the *idle rate* of server j. Performance measures like throughput and utilization are *first order* measures. Often they depend on the first moments of primitive increments. Fluid models introduced in Chapter 2 are relevant to study first order performance measures.

For a class k, if there exists a constant $\mathbb{E}[Z_k(\infty)]$ such that

$$\mathbb{P}\left\{\lim_{t\to\infty} t^{-1}\int_0^t Z_k(s)\,ds = \mathbb{E}[Z_k(\infty)]\right\} = 1, \tag{1.13}$$

$\mathbb{E}[Z_k(\infty)]$ is called the (long-run) *average jobcount* in buffer k. Similarly, for a station j, if there exists a constant $\mathbb{E}[W_j(\infty)]$ such that

$$\mathbb{P}\left\{\lim_{t\to\infty} t^{-1}\int_0^t W_j(s)\,ds = \mathbb{E}[W_j(\infty)]\right\} = 1, \tag{1.14}$$

$\mathbb{E}[W_j(\infty)]$ is called the (long-run) *average workload* at station j. It is often called (average) *virtual waiting time* because under FIFO, a job arriving at station j at time t would have to wait $W_j(t)$ units of time before being served. One can define the actual average waiting time and relate it with the average jobcount via *Little's formula*. For example, for a re-entrant line, let $S(i)$ be the total time in the system for the ith job. If there exists a constant $\mathbb{E}[S(\infty)]$ such that

$$\mathbb{P}\Big\{\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} S(i) = \mathbb{E}[S(\infty)]\Big\} = 1, \tag{1.15}$$

$\mathbb{E}[S(\infty)]$ is called the average time in system. Notice that the average in (1.15) is over the number of jobs whereas the average in (1.13) is over the time intervals. The Little's law alluded earlier asserts that $\mathbb{E}[S(\infty)]$ exists if and only if $\mathbb{E}[|Z(\infty)|]$ exists, and when they exist, they are related by

$$\mathbb{E}[|Z(\infty)|) = \alpha_1\mathbb{E}[S(\infty)], \tag{1.16}$$

where $\mathbb{E}[|Z(\infty)|]$ is a number determined by

$$\mathbb{P}\Big\{\lim_{t\to\infty} t^{-1}\int_0^t |Z(s)|\,ds = \mathbb{E}[|Z(\infty)|]\Big\} = 1,$$

and $|Z(s)| = \sum_k Z_k(s)$. Performance measures like $\mathbb{E}[Z_k(\infty)]$ are *second order* measures. They often heavily depend on the variability of the primitive increments. Fluid models are relevant to the *existence* of these performance measures. Brownian models introduced in a later chapter can often be used to predict second order performance measures when the network is heavily loaded.

1.4 Traffic Equations

To investigate open multiclass queueing networks, one employs the solution λ_ℓ, $\ell = 1, \ldots, K$, of the *traffic equations*

$$\lambda_\ell = \alpha_\ell + \sum_{k=1}^{K} \lambda_k P_{k\ell}. \tag{1.17}$$

or equivalently, in vector form, of $\lambda = \alpha + P'\lambda$. Since P is transient, the unique solution in (1.17) of λ is then given by $\lambda = Q\alpha$. The term λ_k is referred to as the *nominal total arrival rate* to class k due the external and internal arrivals. The qualifier *nominal* is used here because the traffic equation (1.17) implicitly assumes that for each class k there is a long run average rate λ_k of flow into and out of that class and that this does not exceed the maximal mean service rate $\mu_k = 1/m_k$ for class k.

Employing m and λ, one defines the *traffic intensity* ρ_j for the jth server as

$$\rho_j = \sum_{k\in\mathcal{C}(j)} m_k\lambda_k, \tag{1.18}$$

with ρ being the corresponding vector. We interpret ρ_j as the nominal average amount of work brought to server j per unit of time. When $\rho_j \le 1$, it is also called the nominal fraction of time that server j is busy. When Brownian models are considered, we are interested in the network when the system is in *heavy traffic* whose precise definition will be given in a later chapter. Roughly speaking, the system is in heavy traffic when ρ_j is close to one for each station j.

When

$$\rho_j < 1, \quad j = 1, \ldots, J, \tag{1.19}$$

is satisfied, we say that the *usual traffic condition* is satisfied. The issue of whether the nominal total arrival rate λ_k is actually a long run average departure rate or throughput $\widetilde{\lambda}_k$ is related to the stability of the network. We leave this subject for Chapter 2.

1.5 Dynamics of Queueing Networks

In this section, we define *queueing network equations* and *queueing network processes*.

To specify the dynamics of the queueing network, we introduce two additional K-dimensional processes $A = \{A(t), t \geq 0\}$ and $T = \{T(t), t \geq 0\}$ with $A_k(t)$ denoting the total number of arrivals to class k (external and internal arrivals) in $[0, t]$ and $T_k(t)$ denoting the amount of time that server $s(k)$ has spent serving class k jobs in $[0, t]$. One can check that A, D, T, W, Y and Z satisfies the following *queueing network equations:*

$$A(t) = E(t) + \sum_k \Phi^k(D_k(t)), \tag{1.20}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{1.21}$$

$$W(t) = CV(A(t) + Z(0)) - CT(t), \tag{1.22}$$

$$CT(t) + Y(t) = et, \tag{1.23}$$

$$Y_j(t) \text{ can only increase when } W_j(t) = 0, \quad j = 1, \ldots, J, \tag{1.24}$$

for all $t \geq 0$. Here, C is the *constituency matrix* defined as

$$C_{jk} = \begin{cases} 1 & \text{if } k \in \mathcal{C}(j), \\ 0 & \text{otherwise,} \end{cases} \tag{1.25}$$

e denotes the J-vector of all 1's. We note that T and Y are continuous, and that A, D, W and Z are right continuous with left limits. All of the variables are nonnegative in each component, with A, D and T and Y being nondecreasing. By assumption, one has

$$A(0) = D(0) = T(0) = 0 \text{ and } Y(0) = 0. \tag{1.26}$$

In (1.24), we mean that $Y_j(t_2) > Y_j(t_1)$ implies $W_j(t) = 0$ for some $t \in [t_1, t_2]$, which reflects the non-idling property. Since Y is continuous, this can also be written as

$$\int_0^\infty W_j(t)\, dY_j = 0, \quad j = 1, \ldots, J. \tag{1.27}$$

HL queueing networks satisfy

$$V(D(t)) \leq T(t) \leq V(D(t) + e) \tag{1.28}$$

in addition to (1.20)-(1.24), where the inequalities are componentwise and e denotes the K-vector of all 1's. In fact, for each class k, $V_k(D_k(t)) \leq T_k(t)$ for $t \geq 0$ holds for any service discipline. Because the $(D_k(t) + 1)$th job has not departed from class k yet by time t, and only the leading job receives service, we have $T_k(t) \leq V_k(D_k(t) + 1)$.

From our perspective, the 6-tuple

$$\mathbb{X}(t) = (A(t), D(t), T(t), W(t), Y(t), Z(t)), \quad t \geq 0, \tag{1.29}$$

contains all of the essential information on the evolution of the system. We refer to $\mathbb{X}$ as the *queueing network process*, or in the HL setting, as the *HL queueing network process*.

We have presented (1.20)-(1.24) and (1.28) that $\mathbb{X}$ must satisfy. Additional equations which are satisfied by $\mathbb{X}$ will be introduced when specific service disciplines are given.

1.5.1 FIFO Queueing Networks

We recall that FIFO queueing networks are those networks where jobs are served in the order of their arrival at a station. This property can be written as

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t), \quad k = 1, \ldots, K, \tag{1.30}$$

for all $t \geq 0$. Together, (1.20)-(1.24), (1.28)and (1.30) form the *FIFO queueing network equations*. The corresponding 6-tuple $\mathbb{X}$ will be referred to as the *FIFO queueing network process*. One can check that the behaviors of (E, V, Φ) and

$$\{D_k(t) \text{ for } t \leq W_j(0), \quad k = 1, \ldots, K\} \tag{1.31}$$

determines $\mathbb{X}(t)$ for all $t \geq 0$ through the FIFO queueing network equations. Thus, the quantity in (1.31) serves the role of the *initial data* for these equations.

1.5.2 SBP Queueing Networks

Recall that under a SBP discipline, classes at each station are assigned a fixed ranking, with jobs from higher ranking classes being served first. For each class k, we denote by $Z_k^+(t)$ the total number of jobs at time t in classes whose priorities are at least as great as k, and by $T_k^+(t)$ the cumulative time that server $s(k)$ has spent on classes whose priorities are at least as great as k. Since the discipline is assumed to be preempt-resume, the SBP property is given by

$$t - T_k^+(t) \text{ can only increase when } Z_k^+(t) = 0, \quad k = 1, \ldots, K, \tag{1.32}$$

for all $t \geq 0$. Similar to (1.27), one can instead write this as

$$\int_0^\infty Z_k^+(t)\, d(t - T_k^+(t)) = 0, \quad k = 1, \ldots, K. \tag{1.33}$$

Together, (1.20)-(1.24), (1.28) and (1.33) form the *SBP queueing network equations*. The corresponding 6-tuple $\mathbb{X}$ will be referred to as the *SBP queueing network process*. One can check that the values taken by (E, V, Φ) and $Z(0)$ determine $\mathbb{X}(t)$ for all $t \geq 0$ through the SBP queueing network equations; $Z(0)$ therefore serves the role of the *initial data* for these equations.

1.5.3 GHLPS Queueing Networks

We recall that under a GHLPS discipline with proportion vector $\beta = (\beta_\ell)$, all nonempty classes present at a station are served simultaneously, with the fraction of time spent serving a non-empty class, say k, being proportional to β_k. All service goes into the first job of each class with the job departing from the station when his service requirement is attained.

The GHLPS property can be written as

$$T_k(t) = \int_0^t \frac{\beta_k 1_{\{Z_k(s)>0\}}}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell 1_{\{Z_\ell(s)>0\}}} \, ds \tag{1.34}$$

for all $t \geq 0$ and each class k, where, for a set B, 1_B is the indicator function of B. (By convention $0/0 = 0$.) The term $\beta_k 1_{\{Z_k(s)>0\}} / \sum_{\ell \in \mathcal{C}(j)} \beta_\ell 1_{\{Z_\ell(s)>0\}}$ is the proportion of effort received from server $s(k)$ at time s. Together, (1.20)-(1.24), (1.28) and (1.34) form the *GHLPS queueing network equations*. The corresponding 6-tuple $\mathbb{X}$ will be referred to as the *GHLPS queueing network process*. One can check that the values taken by (E, V, Φ) and $Z(0)$ determine $\mathbb{X}(t)$ for all $t \geq 0$ through the GHLPS queueing network equations; $Z(0)$ therefore serves the role of the *initial data* for these equations.

1.5.4 GHLPPS Queueing Networks

We recall that under a GHLPPS discipline with weight vector $\beta = (\beta_\ell)$, all nonempty classes present at a station are served simultaneously, with the fraction of time spent serving a class, say k, being proportional to β_k times the number of jobs in the class. All service goes into the first job of each class to arrive at the station, with the job departing from the station when his service requirement is attained.

The GHLPPS property can be written as

$$T_k(t) = \int_0^t \frac{\beta_k Z_k(s)}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(s)} \, ds \tag{1.35}$$

for all $t \geq 0$. The term

$$\frac{\beta_k Z_k(s)}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(s)}$$

is the proportion of effort received from server $s(k)$ at time s. Together, (1.20)-(1.24), (1.28) and (1.35) form the *GHLPPS queueing network equations*. The corresponding 6-tuple $\mathbb{X}$ will be referred to as the *GHLPPS queueing network process*. One can check that the values taken by (E, V, Φ) and $Z(0)$ determine $\mathbb{X}(t)$ for all $t \geq 0$ through the GHLPPS queueing network equations; $Z(0)$ therefore serves the role of the *initial data* for these equations.

1.6 Steady-State Distributions for FIFO Kelly Networks

An exponential network is said to be a *FIFO Kelly network* if the service discipline is FIFO and for each station, the mean processing times for each class at the station are the same. For an exponential FIFO network, knowing $Z(t)$ is *not* enough to predict the future evolution of the network. Therefore, $Z = \{Z(t), t \geq 0\}$ is *not* a Markov process. The *state* of the network is the order in which jobs are lined up at each station. Let $N_j(t)$ be the total number of jobs at station j at time t. Let

$$X_j(t) = (k_{j,1}, k_{j,2}, k_{j,N_j(t)}),$$

where $k_{j,i}$ is the class number of the ith job at station j at time t. (When $N_j(t) = 0$, $X_j(t)$ is the empty list.) Let $X(t) = (X_1(t), \ldots, X_J(t))$. It can be checked that $X = \{X(t), t \geq 0\}$ is a continuous time Markov chain living in the state space $\prod_{j=1}^{J} \left(\mathbb{Z}_{\mathcal{C}(j)}^{\infty}\right)^{J}$, where $\mathbb{Z}_{\mathcal{C}(j)}^{\infty}$ is the space of finitely terminating sequences of integers in $\mathcal{C}(j)$. The state space is infinite dimensional. The process X is irreducible. When the exponential network is FIFO Kelly, the recurrence condition for X is surprisingly simple. Furthermore, the stationary distribution of X exhibits a *product form*. Let $X(\infty) = (X_1(\infty), \ldots, X_J(\infty))$ be the random variable having the stationary distribution of X. The stationary distribution is of product form if

$$\mathbb{P}\{X_1(\infty) = x_1, \ldots, X_J(\infty) = x_J\} = \mathbb{P}\{X_1(\infty) = x_1\} \cdots \mathbb{P}\{X_J(\infty) = x_J\}$$

for any state $x = (x_1, \ldots, x_J)$.

Theorem 1.6.1. *For a FIFO Kelly network, if the usual traffic condition (1.19) is satisfied, the continuous time Markov chain X is positive recurrent. Further, the stationary distribution of X is of product form with*

$$\mathbb{P}\{X_j(\infty) = x_j\} = (1 - \rho_j) \prod_{k \in \mathcal{C}(j)} (\lambda_k m_k)^{z_k}, \tag{1.36}$$

where z_k is the number of jobs in class k in state x_j.

Proof. The theorem can be verified by checking the balance equations for the stationary distribution of a continuous time Markov chain. □

To interpret (1.36), recall that for an $M/M/1$ queue with a Poisson arrival process with rate α and exponential service times with rate μ, the traffic intensity ρ is simply α/μ. The one-dimensional jobcount process $Z = \{Z(t), t \geq 0\}$ is a birth-death process with reflection at origin. When $\rho < 1$, one can check that the stationary distribution of Z is geometric. Namely,

$$\mathbb{P}\{Z(\infty) = i\} = (1 - \rho)\rho^i \quad \text{for } i = 0, 1, \ldots .$$

Let $N_j(\infty)$ and $Z_k(\infty)$ be the number of jobs at station j and in class k, respectively, in state $X_j(\infty)$. First,

$$\mathbb{P}\{N_j(\infty) = n_j\} = (1-\rho_j)\rho_j^{n_j} \quad \text{for } n_j = 0, 1, \ldots .$$

Therefore, in steady state, the total number of jobs at station j has the same distribution as an $M/M/1$ queue with traffic intensity ρ_j. Next,

$$\mathbb{P}\{Z_k(\infty) = z_k, k \in \mathcal{C}(j) | N_j(\infty) = n_j\} = \frac{n_j!}{\prod_{k\in\mathcal{C}(j)}(z_k!)} \prod_{k\in\mathcal{C}(j)} \left(\lambda_k m_k/\rho_j\right)^{z_k}$$

for any integer $z_k \geq 0$ with $\sum_{k\in\mathcal{C}(j)} z_k = n_j$. Thus, given $N_j(\infty) = n_j$, $Z_k(\infty)$, $k \in \mathcal{C}(j)$, has multinomial distribution with parameters $\lambda_k m_k/\rho_j$, $k \in \mathcal{C}(j)$. Last,

$$\mathbb{P}\{X_j(\infty) = x_j | Z_k(\infty) = z_k, k \in \mathcal{C}(j)\} = \frac{\prod_{k\in\mathcal{C}(j)}(z_k!)}{n_j!},$$

where $n_j = \sum_k z_k$. Therefore, given $Z_k(\infty) = z_k$ for $k \in \mathcal{C}(j)$, $X_j(\infty)$ is equally likely to take any feasible sequence.

1.7 Problems, Notes and Complements

Multiclass queueing networks with dynamic control capability were introduced by Harrison [28]. Many of the queueing network equations were also first introduced in that paper. Brownian models of FIFO queueing networks were studied in Harrison and Nguyen [31, 32] in which the FIFO queueing network equations were introduced, although it is difficult to determine who first introduced the FIFO equation (1.30). Bramson [6] introduced the HLPPS network and its queueing network equations. General HL networks were introduced by Bramson [7]. Section 1.6 contains classic materials; see for example, Kelly [36] or Baskett, Chandy, Muntz and Palacios [2].

Chapter 2

Fluid Networks and Stability Analysis

In this chapter we study fluid models and their role in the study of stability of queueing networks.

2.1 Introduction

Consider the 2-station 5-class re-entrant line in Figure 2.1. For a production manager of this line, he might be interested in finding the maximum input rate possible so that the system still "functions normally" or "is stable". The last terms are admittedly vague. They will be made precise later on. The maximum input rate is related to the maximum *throughput* or maximum *production rate* of this system. Whatever our notion of the stability is, it is intuitively clear that the maximum input rate α_1 is constrained by the traffic condition:

$$\rho_1 = \alpha_1(m_1 + m_3 + m_5) \leq 1 \quad \text{and} \quad \rho_2 = \alpha_1(m_2 + m_4) \leq 1,$$

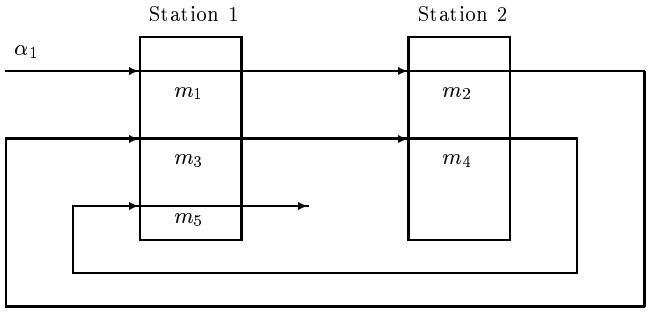

Figure 2.1: A 2-station 5-class re-entrant line

or

$$\alpha_1 \leq \min\Big\{1/(m_1+m_3+m_5), \quad 1/(m_2+m_4)\Big\}. \tag{2.1}$$

Here, as before, m_k is the mean processing time for class k jobs. Equation (2.1) says that the maximum input rate, and hence the maximum throughput, of the system is constrained by the service speed of the slower server or the *bottleneck station*. The fact that the maximum throughput can be calculated using such static models is hardly surprising and these static models are used routinely in many production systems.

To be concrete, let us assume that the network has iid exponential increments. The service discipline is the SBP discipline

$$\pi = \{(5,3,1),(2,4)\} \tag{2.2}$$

that gives the highest priority to class 5, the next priority to class 3 and the lowest priority to class 1 at station 1, and the highest priority to class 2 and the lowest priority to class 4 at station 2. Then, the 5-dimensional jobcount process $Z = \{Z(t), t \geq 0\}$ is a continuous time Markov chain with state space $\mathbb{Z}_+^5$. Note that state 0 (empty system) can be reached from any other state when a long interarrival time occurs. Therefore, the Markov chain is irreducible among the set of states that are reachable from state 0. It is easy to specify the *transition rates* for the Markov chain. For example, if the current state is $(6,3,1,4,2)$, the Markov chain can jump to $(7,3,1,4,2)$ due to an external arrival with rate α_1, to $(6,2,2,4,2)$ due to a service completion of a class 2 job with rate $\mu_2 = 1/m_2$, and to $(6,3,1,4,1)$ due to a service completion of a class 5 job with rate $\mu_5 = 1/m_5$. For this Markov chain, we have the following theorem whose proof is delayed to later sections of this chapter.

Theorem 2.1.1. *(a) If*

$$\rho_1 < 1, \tag{2.3}$$

$$\rho_2 < 1, \tag{2.4}$$

$$\rho_{\rm v} \stackrel{\rm def}{=} \alpha_1(m_2+m_5) < 1, \tag{2.5}$$

then Z is positive recurrent, and hence possesses a unique stationary distribution.
(b) Conversely, if one of the following conditions is satisfied:

$$\rho_1 > 1,$$
$$\rho_2 > 1,$$
$$\rho_{\rm v} > 1,$$

then with probability one, $|Z(t)| \to \infty$ as $t \to \infty$.

For the moment, let us call a system stable if Z is positive recurrent. Theorem 2.1.1 asserts that the stability region constrained by (2.3)-(2.5) is sharp in

some sense. Notice that condition (2.5) is unconventional. It links the average processing time m_2 in class 2 at station 2 with the average processing time m_5 in class 5 at station 1. Condition (2.5) is a so-called *virtual station* condition to be explained shortly in this section.

Assume that $\alpha_1 = 1$, $m_1 = m_3 = m_4 = 0.1$ and $m_2 = m_5 = 0.6$. Then, $\rho_1 = 0.8$, $\rho_2 = 0.7$, and $\rho_{\rm v} = 1.2$. A computer simulation was performed on this network. The simulation starts the system empty and terminates at time 1600. Plotted in Figure 2.2 is one sample of jobcounts at stations 1 and 2 over the time period. Notice that the jobcount at each station shows an almost periodic pattern except that the magnitude of each period gets higher and higher as the period moves on. Also, in many of the time intervals, when one station has a lot of jobs, the other station is empty. This *mutual blocking* by the two servers somehow reduces the service capacity of both servers. To see the utilization of each server, a longer simulation run was performed. Table 2.1 shows the utilizations of both servers averaged at times when the number of jobs that leave the system is 100, 1,000, 10,000 and 100,000, respectively. The

Number of jobs departed	100	1K	10K	100K
Utilization 1	0.65	0.60	0.61	0.65
Utilization 2	0.59	0.68	0.67	0.61

Table 2.1: Utilizations at two stations

utilizations of both servers are significantly below the nominal utilizations of 0.80 and 0.70 while the total number of jobs in the system grows higher and higher.

A naive calculation using (2.1) gives a maximum possible throughput of $1/0.8 = 1.25$ jobs per unit of time. However, Theorem 2.1.1 asserts that the maximum throughput possible is $1/1.2 = 0.833$, a 50% relative difference. The theorem shows that production capacity calculated by using static models can be misleading.

To understand constraint (2.5), we need the following lemma.

Lemma 2.1.2. *Assume that* $Z_2(0)Z_5(0) = 0$. *With probability one,*

$$Z_2(t)Z_5(t) = 0 \quad \textit{for all } t \geq 0. \tag{2.6}$$

Proof. Since interarrival times and service times are exponentially distributed, with probability one, there will be no simultaneous arrivals at any given time. Consider a sample path for which there are no simultaneous arrivals. Assume, on the contrary, that (2.6) does not hold for the sample path. Let τ be the first time that $Z_2(\tau)Z_5(\tau) > 0$ or equivalently, $Z_2(\tau) > 0$ and $Z_5(\tau) > 0$. Since τ is the first such time, we have $Z_2(\tau-)Z_5(\tau-) = 0$. Thus, either $Z_2(\tau-)$ or $Z_5(\tau-)$ is 0. Let us assume that $Z_2(\tau-) = 0$. In this case, at time τ there is an arrival to class 2. Since $Z_5(\tau) > 0$ and there are no simultaneous arrivals, $Z_5(\tau-) > 0$. Because class 5 has preempt-resume priority over classes 3 and 1,

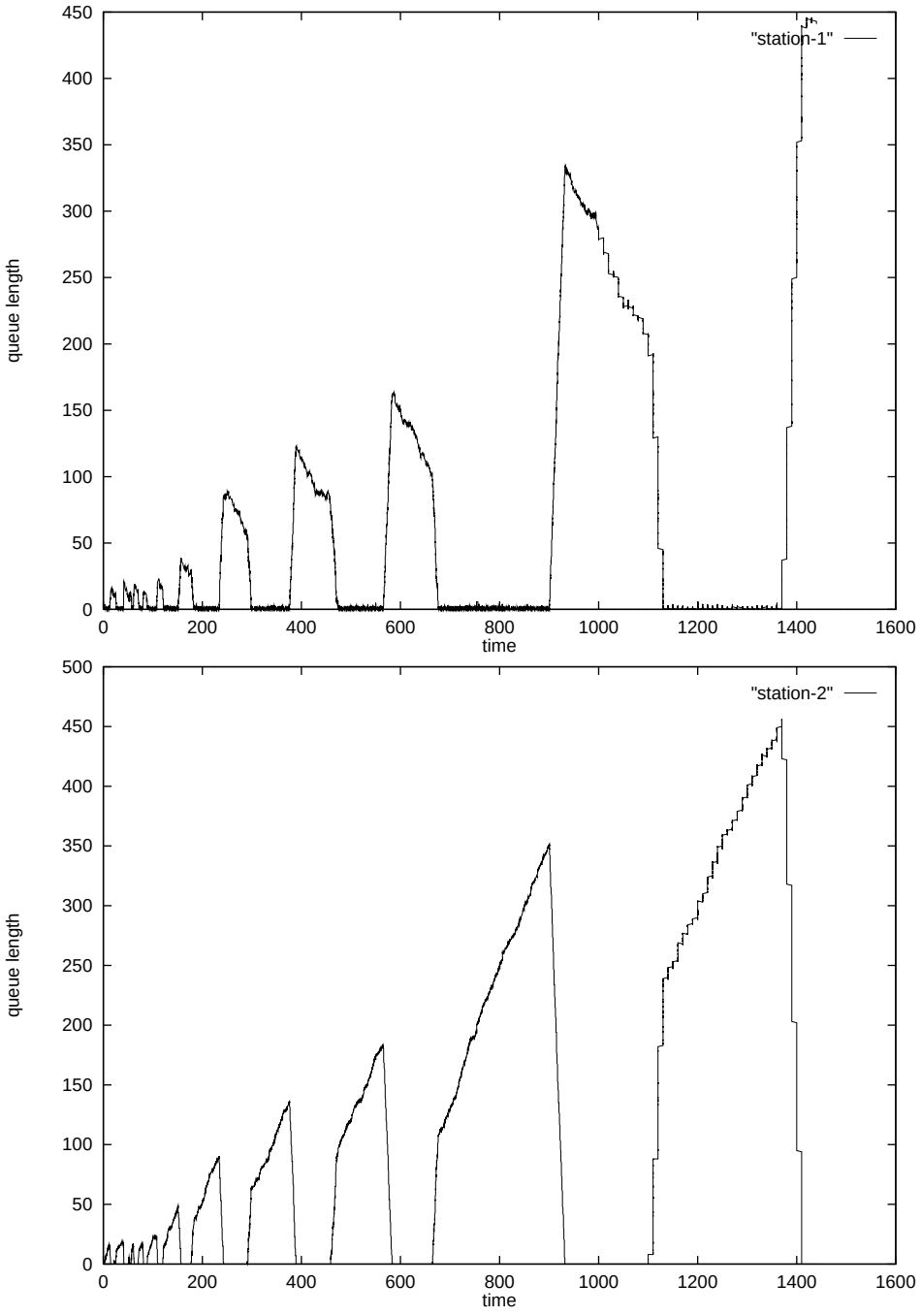

Figure 2.2: Jobcounts at two stations

there is a $\delta > 0$ such that server 1 has been working on class 5 jobs during the time interval $(\tau - \delta, \tau]$. Thus, it is impossible for a class 1 job to enter class 2 at time τ, leading to a contradiction. Assuming $Z_5(\tau-) = 0$ leads to a similar contradiction. □

Given a proper policy of handling simultaneous arrivals to classes, one can in fact show that (2.6) holds for *every* sample path. Lemma 2.1.2 shows that at any given time t, either buffer 2 or buffer 5 is empty. (They can both be empty.) One can imagine that there is a *virtual server* either serving class 2 or class 5 jobs, but never both. In this sense, classes 2 and 5 form a virtual station. Thus, the usual traffic condition applies to the virtual station, which leads to constraint (2.5). This *explains*, but does not prove, the necessity part of the virtual station condition. A rigorous proof of the necessity of the theorem is given in Corollaries 2.5.4 and 2.5.5. It turns out that both the sufficiency and necessity proofs rely heavily on the *fluid model* of the network.

2.2 Fluid Model Equations

The formal deterministic analog of the queueing network process $\mathbb{X}$ has components which satisfy the equations

$$A(t) = \alpha t + P'D(t), \tag{2.7}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{2.8}$$

$$W(t) = CM(A(t) + Z(0)) - CT(t), \tag{2.9}$$

$$CT(t) + Y(t) = et, \tag{2.10}$$

$$Y_j(t) \text{ can only increase when } W_j(t)=0, \quad j = 1, \ldots, J, \tag{2.11}$$

for all $t \geq 0$. In the HL setting, one includes

$$T(t) = MD(t), \tag{2.12}$$

where $M = \text{diag}(m)$.

Equation (2.11) means that whenever $W_j(t) > 0$, $Y_j(\cdot)$ is "flat" at t. More precisely, for each $t > 0$, whenever $W_j(t) > 0$, there exists a $\delta > 0$ such that $Y_j(t + \delta) = Y_j(t - \delta)$, or equivalently, $Y_j(\cdot)$ is constant on $s \in (t - \delta, t + \delta)$. One obtains (2.7)-(2.12) from (1.20)-(1.24) and (1.28) by replacing E, V and Φ by their respective asymptotic means α, M and P. The display (2.7)-(2.11) are known as *fluid model equations*; their solutions, written as

$$\mathbb{X}(t) = (A(t), D(t), T(t), W(t), Y(t), Z(t)), \quad t \geq 0,$$

will be referred to as *fluid model solutions*. When (2.12) is included with (2.7)-(2.11), we refer to the corresponding quantities as *HL fluid model equations* and *HL fluid model solutions*. Unless specified otherwise, all fluid models are HL fluid models in this book.

We have intentionally reused symbols A, D, T, W, Y, Z and $\mathbb{X}$ for the fluid model description. The re-usage emphasizes the parallel between *stochastic discrete* queueing networks and their corresponding *deterministic continuous* fluid models. It should be clear from the context whether these symbols are associated with queueing networks or fluid models. Occasionally, when it is necessary, we add a bar to the fluid quantities. Therefore,

$$\bar{\mathbb{X}} = (\bar{A}, \bar{D}, \bar{T}, \bar{W}, \bar{Y}, \bar{Z})$$

will sometime denote a fluid model solution. When convenient, we will employ the same vocabulary for the fluid model analogs of queueing network quantities, such as the workload W. We prefer to call Z in the fluid model the *fluid level* or buffer level process.

We will assume that the components T and Y of a fluid model solution $\mathbb{X}$ are nondecreasing. One can check that A and D are also nondecreasing with

$$A(0) = D(0) = T(0) = 0 \quad \text{and} \quad Y(0) = 0.$$

It follows from (2.8), (2.9) and (2.12) that

$$W(t) = CMZ(t) \quad \text{for all } t \geq 0 \tag{2.13}$$

Using (2.7)-(2.12), it is easy to show that each component of $\mathbb{X}$ is Lipschitz continuous. That is, for some $N > 0$ (depending on (α, m, P)),

$$|f(t_2) - f(t_1)| \leq N|t_2 - t_1| \quad \text{for all } t_1, t_2 \geq 0,$$

if f is any of the above functions. (When dealing with vectors, we always employ the L_1 norm, although this is merely a matter of convenience.) In particular, each component of $\mathbb{X}$ is *absolutely continuous*, and hence differentiable almost everywhere with respect to Lebesgue measure on $[0, \infty)$. A time $t > 0$ is said to be a *regular point* for the fluid model solution $\mathbb{X}$ if $\mathbb{X}$ is differentiable at time t. Whenever the derivative of a component of $\mathbb{X}$ at t is involved, we always assume that t is a regular point for the fluid model solution $\mathbb{X}$. We use $\dot{f}(t)$ to denote the derivative of f at t.

Again, for each service discipline, there are additional equations for the fluid model solution $\mathbb{X}$ to satisfy. The *FIFO fluid model equations* consist of (2.7)-(2.12) and

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t), \quad k = 1, \ldots, K, \tag{2.14}$$

for all $t \geq 0$. The *initial data* are given by

$$\{D_k(t) \text{ for } t \leq W_j(0), \quad k = 1, \ldots, K\}. \tag{2.15}$$

By (2.10)-(2.12),

$$\sum_{k \in \mathcal{C}(j)} m_k D_k(t) = t \quad \text{for } t \leq W_j(0),$$

which serves as a consistency condition for the initial data. Most fluid model solutions are *not* unique even though their queueing network counterparts often determine the evolution of the queueing network uniquely.

The *SBP fluid model equations* consist of (2.7)-(2.12), together with

$$\int_0^\infty Z_k^+(t)\, d(t - T_k^+(t)) = 0, \quad k = 1, \ldots, K, \tag{2.16}$$

which is equivalent to

$$\dot{T}_k^+(t) = 1 \quad \text{when } Z_k^+(t) > 0, \quad k = 1, \ldots, K, \tag{2.17}$$

for all regular t's. (In this setting, (2.11) is redundant, since it is equivalent to (2.17) when k is the lowest ranked class at its station.) The corresponding 6-tuples $\mathbb{X}$ are the *SBP fluid model solutions*. Here, $Z(0)$ serves the role of the *initial data* for these equations.

The *GHLPS fluid model equations* consist of (2.7)-(2.12), together with

$$\dot{T}_k(t) = \frac{\beta_k}{\sum_{\ell \in \mathcal{C}(j): Z_\ell(t) > 0} \beta_\ell} \Big(1 - \sum_{\ell \in \mathcal{C}(j): Z_\ell(t) = 0} \dot{A}_\ell(t) m_\ell\Big) \tag{2.18}$$

when $Z_k(t) > 0$. Note that server $s(k)$ spends the fraction $\dot{A}_\ell(t) m_\ell$ of its effort to keep buffer ℓ empty. Thus,

$$1 - \sum_{\ell \in \mathcal{C}(j): Z_\ell(t) = 0} \dot{A}_\ell(t) m_\ell$$

is the remaining fraction of the server's capacity available to nonzero buffers. The equality (2.18) states that among nonzero buffers, they receive the server's effort that is proportional to β_k. (When a station is empty, $\dot{T}_k(t)$ may still be positive, and so (1.34) need not hold for the fluid model.) Here, $Z(0)$ serves the role of the *initial data* for the GHLPS fluid model equations. An additional inequality which holds for the GHLPS fluid model is

$$\dot{T}_k(t) \geq \frac{\beta_k}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell} \quad \text{when } Z_k(t) > 0. \tag{2.19}$$

It turns out that in some situations it is more productive to work with (2.19) than (2.18); see Corollary 2.4.11.

The *GHLPPS fluid model equations* consist of (2.7)-(2.12), together with

$$\dot{T}_k(t) = Z_k^\beta(t) \quad \text{when} \quad \sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(t) > 0, \quad k = 1, \ldots, K, \tag{2.20}$$

where

$$Z_k^\beta(t) = \frac{\beta_k Z_k(t)}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(t)} \quad \text{when} \quad \sum_{\ell \in \mathcal{C}(j)} \beta_\ell Z_\ell(t) > 0.$$

The equality (2.20) states that when a station j is nonempty, the server allocation rates $\dot{T}_k(t)$ exist and are proportional to the weighted fluid level of each class k present. (When a station is empty, $\dot{T}_k(t)$ may still be positive, and so (1.35) need not hold for the fluid model.) Here, $Z(0)$ serves the role of the *initial data* for the GHLPPS fluid model equations.

We end this section with a proposition that holds for *any* HL fluid model.

Proposition 2.2.1. *Let $\mathbb{X}$ be a fluid model solution satisfying (2.7)-(2.12). For each station j,*

$$\tau_j(t) = \inf\{s \geq t : W_j(s) = 0\}.$$

Assume that for some station j, $\rho_j < 1$. There exists a constant c that depends on α, m and P such that

$$\tau_j(t) \leq c\, W^M(t),$$

where

$$W^M(t) = \max_j W_j(t).$$

Proof. Let $\mathbb{X}$ be a fluid model solution satisfying (2.7)-(2.12). Then, for any $s > t$,

$$CM(I - P')^{-1}Z(s) = CM(I - P')^{-1}Z(t) + (\rho - e)(s - t) + Y(s) - Y(t).$$

Let $f_j(s)$ be the jth component of $CM(I-P')^{-1}Z(s)$. It is the amount of work for server j embodied in the fluids that are currently *anywhere* in the network. If no future arrivals are allowed, server j has to work *at least* $f_j(s)$ units of time before she can go home. Since $W_j(s) > 0$ for $s \in (t, t + \tau_j(t))$, $Y(s) = Y(t)$ for $s \in (t, t + \tau_j(t))$. Therefore, by the continuity of the fluid model solution,

$$f_j(t + \tau_j(t)) = f_j(t) + (\rho_j - 1)\tau_j(t).$$

Since $f_j(t + \tau_j(t)) \geq 0$, we have

$$\tau_j(t) \leq f_j(t)/(1 - \rho_j) \leq c\, W^M(t).$$

□

2.3 Fluid Limits

For an HL queueing network process $\mathbb{X}$ and $r > 0$, define the *fluid scaling* of X via

$$\bar{\mathbb{X}}^r(t) = r^{-1}\mathbb{X}(rt).$$

Sometimes, the queueing network process $\mathbb{X}$ may depend on r (see Section 2.6.2). For example, r can be the initial number of jobs in the queueing network. In this case,

$$\bar{\mathbb{X}}^r(t) = r^{-1}\mathbb{X}^r(rt),$$

where $\mathbb{X}^r$ is the queueing network process associated with the rth queueing network. Recall that the queueing network process $\mathbb{X}$ is random in general. However, given a realization of the primitive increments, denoted by $(u(\omega), v(\omega), \phi(\omega))$, the evolution of $\mathbb{X}$ is completely determined (modulo tie break policies dealing with simultaneous arrivals or departures). To explicitly state the dependency on the randomness ω, we sometimes use $\mathbb{X}(\cdot, \omega)$ to denote the realization of $\mathbb{X}$ when the sample path is ω. For each sample path ω, $\mathbb{X} \in \mathbb{D}^{4K+2J}[0, \infty)$.

Theorem 2.3.1. *Assume that (1.1) holds for a sample ω. If*

$$|Z^r(0, \omega)|/r \quad \textit{is bounded as } r \to \infty,$$

then $\bar{\mathbb{X}}^r(\cdot, \omega)$ is pre-compact as $r \to \infty$ in the Skorohod path space $\mathbb{D}^{4K+2J}[0, \infty)$ endowed with the u.o.c. topology.

Proof. Note that for any $r > 0$, any sample ω and any class k,

$$\bar{T}_k^r(t, \omega) - \bar{T}_k^r(s, \omega) \leq (t - s) \quad \text{for any } t \geq s \geq 0.$$

Therefore, the family $\{\bar{T}^r(\cdot, \omega) : r > 0\}$ is equi-continuous in $\mathbb{D}^K[0, \infty)$ under the u.o.c. topology. Also $\bar{T}^r(0, \omega) = 0$ for all $r > 0$. Therefore, $\{\bar{T}^r(\cdot, \omega) : r > 0\}$ is tight. Now, for the ω satisfying (1.1), the functional strong law of large numbers (1.5) holds. The lemma follows from the pre-compactness of $|Z^r(0, \omega)|/r$ as $r \to \infty$, the head-of-the-line assumption (1.28) and the functional strong law of large numbers for the primitive cumulatives (E, V, Φ). □

When $\bar{\mathbb{X}}^r(\cdot, \omega)$ is tight as $r \to \infty$, for each sequence $r_n \to \infty$, there is a subsequence $r_{n_k} \to \infty$ such that

$$\bar{\mathbb{X}}^r(\cdot, \omega) \to \bar{\mathbb{X}} \quad \text{u.o.c.}$$

for some $\bar{\mathbb{X}} \in \mathbb{D}^{4K+2J}[0, \infty)$. The process $\bar{\mathbb{X}}$ is called a *fluid limit*. We let $\mathcal{X}(\omega)$ be the set of fluid limits associated with sample path ω. Whenever a fluid limit $\bar{X}$ is concerned, it is always assumed that $\bar{X} \in \mathcal{X}(\omega)$ for some ω satisfying the strong law of large numbers (1.1). Under an HL discipline, the collection of all fluid limits is sometimes called the *fluid limit model* of the discipline.

Each fluid limit must satisfy fluid model equations (2.7)-(2.12). In fact, (2.12) follows from (1.28). Equations (2.7)-(2.10) follow from (1.20)-(1.23). We now show that (2.11) is satisfied for a fluid limit $\bar{\mathbb{X}}$. Let $t > 0$. Assume that $\bar{W}_j(t) > 0$. By the continuity of $\bar{\mathbb{X}}$, there exists a $\delta > 0$ such that $\epsilon \equiv \min_{s \in (t-\delta, t+\delta)} \bar{W}_j(s) > 0$. Since $\bar{\mathbb{X}}$ is a fluid limit, there exists a sample path ω and a sequence $r_n \to \infty$ such that

$$\left(\bar{W}^{r_n}(\cdot, \omega), \bar{Y}^{r_n}(\cdot, \omega)\right) \to (\bar{W}, \bar{Y}) \quad \text{u.o.c.}$$

as $n \to \infty$. In particular, there exists integer N such that

$$\inf_{s \in (t-\delta, t+\delta)} \bar{W}_j^{r_n}(s, \omega) \geq \epsilon/2$$

for $n \geq N$. Therefore, $W_j^{r_n}(s,\omega) > 0$ for $s \in (r_n(t-\delta), r_n(t+\delta))$ and $n \geq N$. Thus, when $n \geq N$, under a non-idling service discipline, $Y_j^{r_n}(s,\omega)$ is flat for $s \in (r_n(t-\delta), r_n(t+\delta))$ or equivalently, $\bar{Y}_j^{r_n}(s,\omega)$ is flat for $s \in (t-\delta, t+\delta)$. Letting $n \to \infty$, we have that $\bar{Y}_j(s)$ is flat for $s \in (t-\delta, t+\delta)$, and thus prove (2.11).

For a FIFO queueing network process $\mathbb{X}$, one can check that each fluid limit satisfies the FIFO fluid equation (2.14). Likewise, for a SBP queueing network process $\mathbb{X}$, one can check that each fluid limit satisfies the SBP fluid equation (2.16) using a proof similar to the proof of (2.11). For a GHLPS queueing network process, one can show that each fluid limit satisfies (2.18) and (2.19). In fact, **a proof is needed here**. Similarly, one can show that for a GHLPPS queueing network process, each fluid limit satisfies (2.20). Hence we have the following theorem.

Theorem 2.3.2. *For each of the HL queueing network considered in Chapter 1, each fluid limit is a fluid model solution.*

One wonders that, for a given HL service discipline, what additional fluid model equations should be added? A sensible approach is to add whatever fluid model equations that are satisfied by fluid limits. Therefore, Theorem 2.3.2 should be generally true. In fact, for a given HL discipline, Theorem 2.3.2 should be the guiding principle in adding fluid model equations for the discipline.

2.4 Calculus for Fluid Models

Let us fix an HL discipline. In this section, whenever a fluid model is mentioned, it is associated with the service discipline.

Definition 2.4.1. The fluid model is *stable* if there exists a $\delta > 0$ such that for each fluid solution $\mathbb{X}$ with $|Z(0)| \leq 1$, $Z(t) = 0$ for $t \geq \delta$.

Definition 2.4.2. The fluid model is *weakly stable* if for each fluid solution $\mathbb{X}$ with $Z(0) = 0$, $Z(t) = 0$ for $t \geq 0$.

Definition 2.4.3. A fluid model solution $\mathbb{X}$ is *unstable* if there exists a sequence $\{t_n\}$ with $t_n \to \infty$ such that $Z(t_n) > 0$ for each n. The fluid model is *unstable* if there exists an unstable fluid model solution.

Definition 2.4.4. The fluid model is *weakly unstable* if there exists a $\delta > 0$ such that for *every* fluid solution $\mathbb{X}$ with $Z(0) = 0$, $Z(\delta) \neq 0$.

Notice that the fluid model being unstable is equivalent to the fluid model being not stable. However, the fluid model being weakly unstable is often much stronger than the fluid model being not weakly stable. The gap between them is subtle, important and somewhat annoying. There is a considerable amount of current current research activity devoted to the subject. The term "weakly unstable" is *not* standard. A more descriptive, but cumbersome term perhaps is "strongly not-weakly-stable".

To show the instability of the fluid model, it is enough to construct *one* unstable fluid model solution. To show stability of the fluid model, one needs to show that *all* fluid model solutions drain to zero uniformly fast. Since it is difficult to follow the exact dynamics of each fluid model solution, we need to develop a *calculus* for the fluid model.

Recall that each component of a fluid model solution $\mathbb{X}$ is absolutely continuous and hence differentiable almost everywhere in $[0, \infty)$ and a time t is regular for $\mathbb{X}$ if each component of $\mathbb{X}$ is differentiable at t. Whenever the derivative of a component at t is considered, t is assumed to be a regular point. Our first lemma is a general result from analysis.

Lemma 2.4.5. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be an absolutely continuous function, and $\epsilon > 0$. Assume that whenever $f(t) > 0$ and $f(t)$ is differentiable at t, $\dot{f}(t) \leq -\epsilon$. Then*

$$f(t) = 0 \quad \textit{for} \quad t \geq f(0)/\epsilon.$$

Let $a_k(t) = \dot{A}_k(t)$ and $d_k(t) = \dot{D}_k(t)$. In the following proposition

$$\mathbb{X} = (A, D, T, W, Y, Z)$$

is a fluid model solution satisfying (2.7)-(2.12), and t is a regular point.

Lemma 2.4.6. *(a) If $Z_k(t) = 0$, then $a_k(t) = d_k(t)$; (b) if $W_j(t) > 0$,*

$$\sum_{k \in \mathcal{C}(j)} m_k d_k(t) = 1;$$

and (c) $a_k(t) = \alpha_k + \sum_\ell P_{\ell k} d_\ell(t)$.

Proof. If $Z_k(t) = 0$, Z_k attains a minimum at t. Since Z_k is differentiable at t, $\dot{Z}_k(t) = 0$ or equivalently, $a_k(t) = d_k(t)$. This proves part (a).

For part (b), when $W_j(t) > 0$, it follows from (2.11) that $\dot{Y}_j(t) = 0$. Hence by (2.10) that

$$\sum_{k \in \mathcal{C}(j)} \dot{T}_k(t) = 1.$$

From (2.12), $m_k d_k(t) = \dot{T}_k(t)$, and thus part (b) is proved. Part (c) follows from (2.7) immediately. □

A function $g : \mathbb{R}_+^K \to \mathbb{R}_+$ is said to be *locally Lipschitz continuous* if for any compact set B, there exists a constant $\kappa(B)$ such that for any $x, y \in B$,

$$|g(x) - g(y)| \leq \kappa(B)|x - y|.$$

Lemma 2.4.7. *Let $g : \mathbb{R}_+^K \to \mathbb{R}_+$ be a locally Lipschitz function such that $g(x) = 0$ if and only if $x = 0$. Let $\epsilon > 0$ be a constant. For each fluid model solution $\mathbb{X}$, let $f(t) = g(Z(t))$. Assume that whenever $Z(t) \neq 0$ and t is regular for both $\mathbb{X}$ and f,*

$$\dot{f}(t) \leq -\epsilon.$$

then the fluid model is stable.

Proof. Let $\mathbb{X}$ be a fluid model solution with $|Z(0)| \leq 1$. Since Z is Lipschitz (with Lipschitz constant N), for any $t > 0$,

$$|Z(s)| \leq |Z(0)| + |Z(s) - Z(0)| \leq 1 + Nt \quad \text{for } 0 \leq s \leq t.$$

Therefore, for any $0 \leq u \leq s \leq t$

$$|f(s) - f(u)| \leq \kappa(B)|Z(s) - Z(u)| \leq \kappa(B)N(s-u),$$

where $B = \{z \in \mathbb{R}_+^K : |z| \leq 1 + Nt\}$. Thus, f is locally Lipschitz, and hence absolutely continuous. Since $f(t) = 0$ iff $Z(t) = 0$, it follows Lemma 2.4.5 that $f(t) = 0$ and $Z(t) = 0$ for $t \geq f(0)/\epsilon$. Let

$$\delta = \max_{z \in R_+^K : |z| \leq 1} g(z).$$

Then $f(0) = g(Z(0)) \leq \delta$. Thus, $Z(t) = 0$ for $t \geq \delta/\epsilon$, proving the stability of the fluid model. □

Definition 2.4.8. The function g in Lemma 2.4.7 is called a *Lyapunov* function.

One nice thing about a Lyapunov function is that one need not know exactly how the dynamics of each fluid model solution behaves, yet all fluid model solutions are "driven" down to zero by the Lyapunov function uniformly fast.

To show the power of Lemmas 2.4.6 and 2.4.7, we present the following example.

Example. For a re-entrant line operating under the last-buffer-first-serve (LBFS) service discipline, if the usual traffic condition (1.19) is satisfied, the fluid model is stable.

Proof. Let $\mathbb{X}$ be a LBFS fluid model solution with $|Z(0)| \leq 1$. The LBFS discipline is a SBP discipline which gives higher priority to higher numbered classes. Let

$$f(t) = \sum_{k=1}^{K} Z_k(t) = f(0) + \alpha_1 t - D_K(t)$$

be the total amount of fluid in the system at time t. Notice that $f(t) = g(Z(t))$ for the linear function $g(z) = |z|$. Assume that $Z(t) \neq 0$. Let k be the last buffer such that $Z_k(t) > 0$ at (regular) time t. By parts (a) and (c) of Lemma 2.4.6, we have $d_k(t) = d_{k+1}(t) = \ldots = d_K(t)$. Thus,

$$\dot{f}(t) = \alpha_1 - d_k(t).$$

Since $Z_k(t) > 0$, by Part (b) of Lemma 2.4.6,

$$\sum_{\ell \in \mathcal{C}(j)} m_\ell d_\ell(t) = 1,$$

where $j = s(k)$. Using (2.17), we have $d_\ell(t) = 0$ for each $\ell < k$ with $\sigma(\ell) = s(k)$. Thus,

$$\sum_{\ell \in \mathcal{C}(j)} m_\ell d_\ell(t) = d_k(t) \sum_{\ell \in \mathcal{C}(j): \ell \geq k} m_\ell,$$

and

$$d_k(t) = \frac{1}{\sum_{\ell \in \mathcal{C}(j): \ell \geq k} m_\ell}.$$

Thus,

$$\begin{aligned} \dot{f}(t) &= \alpha_1 - \frac{1}{\sum_{\ell \in \mathcal{C}(j): \ell \geq k} m_\ell} \leq \alpha_1 - \frac{1}{\sum_{\ell \in \mathcal{C}(j)} m_\ell} \\ &= \alpha_1(1 - 1/\rho_j) = -(\alpha_1/\rho_j)(1 - \rho_j). \end{aligned}$$

Let

$$\epsilon = \min_j (\alpha_1/\rho_j)(1 - \rho_j) > 0.$$

Then f is a Lyapunov function satisfying the conditions in Lemma 2.4.7. □

We end this section by proving the following theorem.

Theorem 2.4.9. *Let $\epsilon > 0$. Assume that $\mathbb{X}$ is a fluid model solution satisfying (2.7)-(2.8) and*

$$\dot{D}_k(t) \geq \lambda_k + \epsilon \quad \textit{whenever } Z_k(t) > 0. \tag{2.21}$$

Then $Z(t) = 0$ for $t \geq |(I - P')^{-1} Z(0)|/\epsilon$.

Proof. Let

$$f(t) = e'(I - P')^{-1} Z(t).$$

In the queueing network analogs, the kth component of $(I - P')^{-1} Z(t)$ is the total expected number of class k services required by the jobs currently *anywhere* in the network. Thus $f(t)$ is the total expected number of services generated by the jobs currently anywhere in the network.

From (2.7)-(2.8), we have

$$f(t) = f(0) + e'\lambda t - e' D(t).$$

Thus, for $Z(t) \neq 0$,

$$\begin{aligned} \dot{f}(t) &= \sum_{k=1}^{K} (\lambda_k - \dot{D}_k(t)) \\ &= \sum_{k: Z_k(t) \neq 0} (\lambda_k - \dot{D}_k(t)) + \sum_{k: Z_k(t) = 0} (\lambda_k - \dot{D}_k(t)) \\ &\leq -\epsilon |\{k : Z_k(t) \neq 0\}| + \sum_{k: Z_k(t) = 0} (\lambda_k - \dot{D}_k(t)) \qquad (2.22) \\ &\leq -\epsilon, \qquad (2.23) \end{aligned}$$

where, for a finite set $\mathcal{S}$, $|\mathcal{S}|$ is the cardinality if $\mathcal{S}$. Note that (2.22) follows from (2.21), and (2.23) follows from $Z(t) \neq 0$ and the following claim:

$$\dot{D}_k(t) \geq \lambda_k \quad \text{when } Z_k(t) = 0. \tag{2.24}$$

Assuming (2.24) for the moment, then, f is a Lyapunov function satisfying the conditions in Lemma 2.4.7. Thus, $Z(t) = 0$ for $t \geq |(I - P')^{-1}Z(0)|/\epsilon$.

To prove (2.24), we introduce the following convention. For a vector $x \in \mathbb{R}^K$ and a set $\mathbf{k} \subset \{1, \ldots, K\}$, we let $x_{\mathbf{k}}$ denote the sub-vector $(x_k), k \in \mathbf{k}$. For a $K \times K$ matrix B, we use $B_{\mathbf{k},\mathbf{k}'}$ to denote the submatrix $B_{k,\ell}$ with $k \in \mathbf{k}$ and $\ell \in \mathbf{k}'$. Now, let $\mathbf{k} = \{k : Z_k(t) = 0\}$. Since t is a fixed regular point, we omit the dependency on t in the definition of $\mathbf{k}$. We let $\mathbf{k}^c$ denote the complement of $\mathbf{k}$. Since

$$\dot{Z}(t) = \alpha + (P' - I)\dot{D}(t),$$

we have

$$\dot{Z}_{\mathbf{k}}(t) = \alpha_{\mathbf{k}} + ((P')_{\mathbf{k},\mathbf{k}} - I)\dot{D}_{\mathbf{k}}(t) + (P')_{\mathbf{k},\mathbf{k}^c}\dot{D}_{\mathbf{k}^c}(t).$$

By Part (a) of Lemma 2.4.6, $\dot{Z}_{\mathbf{k}}(t) = 0$. Thus,

$$\begin{aligned}
\dot{D}_{\mathbf{k}}(t) &= (I - (P')_{\mathbf{k},\mathbf{k}})^{-1}\alpha_{\mathbf{k}} + (I - (P')_{\mathbf{k},\mathbf{k}})^{-1}(P')_{\mathbf{k},\mathbf{k}^c}\dot{D}_{\mathbf{k}^c}(t) \\
&\geq (I - (P')_{\mathbf{k},\mathbf{k}})^{-1}\alpha_{\mathbf{k}} + (I - (P')_{\mathbf{k},\mathbf{k}})^{-1}(P')_{\mathbf{k},\mathbf{k}^c}\lambda_{\mathbf{k}^c} \\
&= \lambda_{\mathbf{k}},
\end{aligned}$$

where we have used the facts that each entry of $(I - (P')_{\mathbf{k},\mathbf{k}})^{-1}(P')_{\mathbf{k},\mathbf{k}^c}$ is non-negative and $\dot{D}_{\mathbf{k}^c}(t) \geq \lambda_{\mathbf{k}^c}$ to get the inequality, and the traffic equation (1.17) to get the last equality.

□

The following corollary holds for Jackson type networks.

Corrollary 2.4.10. *For $K = J$, if the usual traffic condition (1.19) holds, the fluid model is stable.*

Proof. Let

$$\epsilon = \min\{\mu_j - \lambda_j : j = 1, \ldots, J\},$$

where $\mu_j = 1/m_i$. If $Z_j(t) > 0$, by (2.11) and (2.12), $\dot{D}_j(t) = \mu_j \geq \lambda_j + \epsilon$. Thus, the corollary follows from Theorem 2.4.9. □

Corrollary 2.4.11. *Assume the usual traffic condition (1.19) holds. Let $\beta = (\lambda_1 m_1, \ldots, \lambda_K m_K)$. The GHLPS fluid model with the proportion vector β is stable.*

Proof. Condition (2.21) follows from (1.19) and the additional GHLPS fluid model equation (2.19). □

2.5 Instability of Fluid and Queueing Networks

The significance of the weak instability of a fluid model lies in the following theorem.

Theorem 2.5.1. *If the fluid model is weakly unstable, the queueing network is unstable in the sense that, with probability one,*

$$|Z(t)| \to \infty \quad as \quad t \to \infty,$$

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

Proof. Let ω be fixed so that (1.1) is satisfied. Let $\mathcal{X}_\omega$ be the set of fluid limits for the sample ω. If the fluid model is weakly unstable, there is a $\delta > 0$ such that for each $\bar{\mathbb{X}} \in \mathcal{X}_\omega$, $\bar{Z}(\delta) \neq 0$. We claim that

$$\liminf_{r\to\infty} |Z(r\delta, \omega)/r| > 0, \tag{2.25}$$

which, of course, implies that $\lim_{t\to\infty} |Z(t,\omega)| = \infty$. To prove (2.25), suppose that

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

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

$$|Z(r_n\delta, \omega)/r_n| \to 0. \tag{2.26}$$

Because $\{\mathbb{X}(r\cdot, \omega)/r\}$ is pre-compact as $r \to \infty$, there exists a subsequence $\{r_{n_m}\} \subset \{r_n\}$ such that $\mathbb{X}(r_{n_m}\cdot, \omega)/r_{n_m}$ converges u.o.c to a limit $\bar{\mathbb{X}}(\cdot) \in \mathcal{X}_\omega$. Hence,

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

which contradicts (2.26). □

Note that the theorem still holds if the weak instability assumption is weakened so that for almost all ω, there exists $\delta = \delta(\omega) > 0$ such that for each $\bar{\mathbb{X}} \in \mathcal{X}_\omega$, $\bar{Z}(\delta) \neq 0$. This remark applies to Corollary 2.5.2 as well.

Stolyar [48] proved that the set of fluid limits $\mathcal{X}_\omega$ is pre-compact or tight. Therefore, the weak instability implies the stronger condition

$$\inf_{\bar{\mathbb{X}}\in\mathcal{X}_\omega} |\bar{Z}(\delta)| > 0, \tag{2.27}$$

where $\delta > 0$ is as in Definition 2.4.4.

Corrollary 2.5.2. *Assume that the fluid model is weakly unstable with* $\delta > 0$ *as in Definition 2.4.4. For each* ω *satisfying (1.1) and each* $\epsilon > 0$ *with*

$$\epsilon < \inf_{\bar{\mathbb{X}}\in\mathcal{X}_\omega} |\bar{Z}(\delta)|,$$

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

$$|Z(t,\omega)| \geq \frac{\epsilon}{\delta} t \quad for\ t \geq M(\omega).$$

The following proposition provides one way to check the weak instability of a fluid model.

Proposition 2.5.3. *If there exists a set B of job classes and positive constants a_k, $k \in B$, such that*

$$\sum_{k \in B} a_k \lambda_k m_k > 1 \tag{2.28}$$

and for each fluid model solution $\mathbb{X}$,

$$\sum_{k \in B} a_k \dot{T}_k(t) \leq 1, \tag{2.29}$$

then the fluid model is weakly unstable.

Proof. It follows from (2.7), (2.8) and (2.12) that

$$Z(t) = Z(0) + \alpha t - (I - P')M^{-1}T(t). \tag{2.30}$$

Next, let

$$L(t) = (I - P')^{-1} Z(t).$$

The kth component of $L(t)$ is the total amount of fluid *anywhere* in the network at time t that will eventually go through class k. Thus, the kth component of $ML(t)$ is the amount of time that server $s(k)$ spends on class k to clear out this amount fluid. It follows from (2.30) that

$$ML(t) = ML(0) + M\lambda t - T(t).$$

Let

$$f(t) = \sum_{k \in B} a_k m_k L_k(t).$$

Then

$$f(t) = f(0) + \sum_{k \in B} a_k m_k \lambda_k t - \sum_{k \in B} a_k T_k(t).$$

By assumptions (2.28) and (2.29), we have

$$\dot{f}(t) \geq \big(\sum_{k \in B} a_k m_k \lambda_k - 1 \big) > 0.$$

Thus, $f(t) \geq (\sum_{k \in B} a_k m_k \lambda_k - 1)t$ for any $t \geq 0$, proving the weak instability of the fluid model. □

Corrollary 2.5.4. *Assume that $\rho_j > 1$ for some station j. The fluid model under any HL discipline is weakly unstable.*

Proof. The proof follows from Proposition 2.5.3 by taking $B = \mathcal{C}(j)$ and $a_k = 1$ for $k \in B$. □

Corrollary 2.5.5. *For the 2-station 5-class re-entrant line pictured in Figure 2.1 operating under the SBP discipline (2.2), if* $\rho_{\rm v} > 1$, *with probability one, the total number of jobs in the system goes to infinity.*

Proof. Assume that the queueing network is initially empty. By Lemma 2.1.2,

$$t - (T_2(t) + T_5(t))$$

is nondecreasing. This property carries over to each fluid limit. Thus, each fluid limit $\bar{\mathbb{X}}$ satisfies

$$\dot{\bar{T}}_2(t) + \dot{\bar{T}}_5(t) \leq 1, \tag{2.31}$$

in addition to (2.7)-(2.12) and (2.17). Let $B = \{2, 5\}$ and $a_2 = a_5 = 1$. It follows from Proposition 2.5.3 that the fluid model, augmented with (2.31), is weakly unstable. Thus, with probability one, the total number of jobs goes to infinity. □

2.6 Stability of Queueing Networks

Different notions of stability for a queueing network exist. *Rate stability* is perhaps the easiest one to study. It requires very weak assumptions on the primitive cumulatives. When the network has iid increments, one can consider *positive Harris recurrence* of the network.

2.6.1 Rate Stability

Definition 2.6.1. The queueing network is said to be *rate stable* if for each fixed initial data, with probability one,

$$\lim_{t\to\infty} D_k(t)/t = \lambda_k \quad \text{for } k = 1, \ldots, K. \tag{2.32}$$

Therefore, the queueing network is rate stable if the throughput rate or departure rate from a class is equal to the nominal total arrival rate to the class.

Theorem 2.6.2. *The queueing network is rate stable if and only if for each fixed initial data, with probability one, the fluid limit* $\bar{\mathbb{X}}$ *is uniquely given by*

$$\begin{aligned} \bar{A}(t) &= \lambda t, & \bar{D}(t) &= \lambda t, \\ \bar{T}(t) &= M\lambda t, & \bar{W}(t) &= 0, \\ \bar{Y}(t) &= (e - \rho)t, & \bar{Z}(t) &= 0. \end{aligned}$$

Proof. Assume that the queueing network is rate stable. Let $\bar{\mathbb{X}}$ be a fluid limit taken with the initial data fixed. Then, $\bar{Z}(0) = 0$. Since, with probability one, each fluid limit is a fluid solution, the fluid limit satisfies (2.7)-(2.12). Because the queueing network is rate stable, it follows from Lemma 1.2.3 that $\bar{D}(t) = \lambda t$

for $t \geq 0$. From (2.7) and the traffic equation (1.17), $\bar{A}(t) = \lambda t$. Since $\bar{Z}(0) = 0$, from (2.8), we have $\bar{Z}(t) = 0$ for $t \geq 0$. Equation (2.12) gives $\bar{T}(t) = M\lambda t$ for $t \geq 0$ and (2.13) gives $\bar{W}(t) = 0$ for $t \geq 0$. Finally, $\bar{Y}(t) = (e - \rho)t$ for each $t \geq 0$ follows from (2.10).

Conversely, assume that, with probability one, the fluid limit is unique. Let us fix a sample ω such that the fluid limit along the sample is unique and the strong law of large numbers (1.1) holds. Thus,

$$r^{-1}D_k(r\cdot,\omega) \to \lambda_k \cdot \quad \text{u.o.c. as } r \to \infty.$$

In particular, we have

$$r^{-1}D_k(r,\omega) \to \lambda_k \quad \text{as } r \to \infty.$$

Thus, the queueing network is rate stable. □

Corrollary 2.6.3. *Assume that a queueing network is rate stable. Then, with probability one,*

$$\lim_{t\to\infty} T_k(t)/t = \lambda_k m_k, \quad k = 1, \ldots, K, \tag{2.33}$$

$$\lim_{t\to\infty} Y_j(t)/t = 1 - \rho_j, \quad j = 1, \ldots, J. \tag{2.34}$$

As a consequence,

$$\rho_j \leq 1 \quad j = 1, \ldots, J.$$

Thus, if the queueing network is rate stable, the fraction of time that server $s(k)$ spends on class k jobs is equal to $\lambda_k m_k$, the nominal amount of class k work brought to the server per unit of time. Conversely, if one can allocate servers' effort such that (2.33) holds, one can show from (2.7)-(2.12) that the fluid limit is unique and hence the queueing network is rate stable. This criterion should give us a guiding principle to construct stable service disciplines.

Corrollary 2.6.4. *If the fluid model is weakly stable, the queueing network is rate stable.*

Proof. The proof follows from Theorem 2.3.2. □

Corrollary 2.6.5. *If the fluid model is weakly unstable, the queueing network is* not *rate stable.*

Proof. Suppose that the queueing network is rate stable. By Theorem 2.3.2, for each fluid limit $\mathbb{X}$, the fluid level is always zero. Since the fluid limit is a fluid solution, this contradicts the definition of weak instability of the fluid model. □

2.6.2 Positive Harris Recurrence

In this section, we consider HL queueing networks with iid increments. We now define a Markov process $X = \{X(t), t \geq 0\}$ which describes the dynamics of a queueing network. The Markov property requires that given $X(t)$ at time t, the future evolution of the queueing network in (t, ∞) can be determined, at least in probabilistic distributions. This requires us to carefully define the *states* or the values that $X(t)$ may take. The state of the process at any time is given by a point

$$x \in \mathbb{R}_+^K \times \left(\mathbb{Z}_K\right)^\infty \times \mathbb{R}_+^{|\mathcal{E}|} \times \mathbb{R}_+^K = \mathbb{Z}_+^\infty \times \mathbb{R}_+^{2K+|\mathcal{E}|}, \tag{2.35}$$

where $\mathbb{Z}_K = \{1, \ldots, K\}$, $\mathbb{Z}_K^\infty$ is the set of finitely terminating sequences in $\mathbb{Z}_K$, and $|\mathcal{E}|$ is the cardinality of $\mathcal{E}$. The component in the first $\mathbb{R}_+^K$, denoted by $\beta = (\beta_1, \ldots, \beta_K)$, determines the proportion of service effort that each class receives from its server. We require that

$$\sum_{k \in \mathcal{C}(j)} \beta_k = 1 \quad \text{for } j = 1, \ldots, J.$$

The component in $(\mathbb{Z}_K)^\infty$, denoted by $\vec{k} = (k_1, \ldots, k_{|z|})$, determines the order of all jobs in the network, where $z = (z_1, \ldots, z_K)$ is the number of jobs in each class and $|z| = \sum_k z_k$ is the total number of jobs in the state. The component k_i is the class number of the ith job. The component in $\mathbb{R}_+^{|\mathcal{E}|}$, denoted by $u = (u_k), k \in \mathcal{E}$, determines the residual external interarrival times. Finally, the component in the last $\mathbb{R}_+^K$, denoted by $v = (v_1, \ldots, v_K)$, determines the residual service times for the leading job of each class.

For notational convenience, we have maintained the *global* order list $\vec{k}$ in our description of a state. It is indeed sufficient to just keep track of the order of the jobs at each station. Under certain service disciplines, like SBP or HLPS disciplines, there is no need to keep track of the order of jobs at all. In these cases, it is enough to keep track of the jobcount z instead of $\vec{k}$ and the state space is a subset of

$$\mathbb{Z}_+^K \times \mathbb{R}^{2K+|\mathcal{E}|},$$

which is finitely dimensional. Often, only one class at a station receives service at a time. In this case, there is no need to keep track of the entire vector v of residual service times. However, in this case, one needs to keep track of which class is currently being served. Again, for notational convenience, we simply keep track of the entire vector v. If interarrival and service times are exponentially distributed, we can drop components u and v in the state description. The metric on the state space is determined by a "norm" $|\cdot|$, which is defined to be

$$|x| = |z| + |u| + |v|,$$

where $|v| = \sum_k v_k$ and $|u|$ is defined similarly.

We assume that the server's efforts β do not change between arrivals and departures. After an arrival or a departure at t, we may assign new service rates

$\beta(t)$ which depend measurably on $(\vec{k}, u, v)$ at time t. All the service disciplines considered in Chapter 1 satisfy this assumption on β.

We assume that the process $X = \{X(t), t \geq 0\}$ is right continuous. One can check that X is Markov for each of the HL disciplines discussed. Notice that as time t goes on, all components of u and some components of v decrease deterministically while the remainder of the state remains constant. When one of these residual processes reaches zero (corresponding to the time of an external arrival or a departure), a jump occurs for β and $\vec{k}$ and the residual time is reset to a new random variable. Hence X is a *piecewise deterministic Markov* (PDM) process that conforms to Assumption 3.1 of Davis [22]. It follows from Davis [22, page 362] that

Proposition 2.6.6. *The process $X = \{X(t), t \geq 0\}$ is a strong Markov process.*

We use X^x to denote the process with $X(0) = x$. Let $\mathbb{S}$ be the state space with the Borel σ-field $\mathfrak{B}_{\mathbb{S}}$. Let $P^t(x, \cdot)$ be the transition probability of X. That is

$$P^t(x, B) = \mathbb{P}\{X^x(t) \in B\} \quad \text{for } B \in \mathfrak{B}_{\mathbb{S}}.$$

A nonzero measure π on $(\mathbb{S}, \mathfrak{B}_{\mathbb{S}})$ is *invariant* for X if π is σ-finite, and for each $t \geq 0$,

$$\pi(B) = \int_S P^t(x, B)\, \pi(dx), \quad \text{for all } B \in \mathfrak{B}_{\mathbb{S}}.$$

An invariant measure π is said to be unique if the only invariant measures for X are positive scalar multiples of π. For $B \subset \mathbb{S}$, let $\tau_B^x = \inf\{t \geq 0 : X^x(t) \in B\}$.

Definition 2.6.7. The process X is *Harris recurrent* if there exists some σ-finite measure ν on $(\mathbb{S}, \mathfrak{B}_{\mathbb{S}})$, such that whenever $\nu(B) > 0$ and $B \in \mathfrak{B}_{\mathbb{S}}$,

$$\mathbb{P}\{\tau_B^x < \infty\} = 1 \quad \text{for } x \in \mathbb{S}.$$

Harris recurrence implies an apparently stronger condition that

$$\mathbb{P}\left\{\int_0^\infty 1_B(X^x(s))\, ds = \infty\right\} = 1 \quad \text{for } x \in \mathbb{S} \tag{2.36}$$

whenever $\nu'(B) > 0$ for some σ-finite measure ν'. Thus, for a Harris recurrent chain X, with probability one, it spends an infinite amount of time in all "nonzero" sets. It is known that if X is Harris recurrent then an essentially unique invariant measure π exists, see for example Getoor [27].

Definition 2.6.8. For a Harris recurrent Markov process X, if the invariant measure is finite, then it may be normalized to a probability measure; in this case X is called *positive Harris recurrent*.

Assume that $X = \{X(t), t \geq 0\}$ is positive Harris recurrent with unique stationary probability distribution π. For any measurable function f on $(\mathbb{S}, \mathfrak{B}_{\mathbb{S}})$, let

$$\pi(f) = \int_{\mathbb{S}} f(x)\, \pi(dx)$$

whenever the integral makes sense. Positive Harris recurrence implies the following ergodic property: for every measurable f on $\mathbb{S}$ with $\pi(|f|) < \infty$,

$$\mathbb{P}\left\{\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X^x(s))\,ds = \pi(f)\right\} = 1 \quad \text{for each } x \in \mathbb{S}.$$

Take $f(x)$ to be the number of class k jobs when the state is in x. Because f is nonnegative, positive Harris recurrence for X implies that for each job class k,

$$\mathbb{P}\left\{\lim_{t\to\infty} \frac{1}{t}\int_0^t Z_k(s)\,ds = \pi(f)\right\} = 1 \quad \text{for each } x \in \mathbb{S}.$$

Note that in general $\pi(f)$ may be infinite.

Definition 2.6.9. A distribution ν on $[0,\infty)$ is said to be *unbounded* if

$$\nu([t,\infty)) > 0 \quad \text{for each } t > 0; \tag{2.37}$$

the distribution is said to be *spread out* if there exist some positive integer i and some nonnegative function $q : \mathbb{R}_+ \to \mathbb{R}_+$ with $\int_0^\infty q(t)\,dt > 0$ such that

$$\nu^{*i}(dt) \geq q(t)\,dt, \tag{2.38}$$

where ν^{*i} is the ith convolution of ν.

Theorem 2.6.10. *Consider an HL open queueing network with iid increments. Assume that all distributions have finite mean and each interarrival time distribution is unbounded and spread out. If the corresponding fluid model is stable, then X is positive Harris recurrent.*

We outline some of the key steps in the proof of the theorem before we discuss some extensions to the theorem.

Suppose that a is a probability distribution on $\mathbb{R}_+$. Define the Markov transition function K_a as

$$K_a(x,\cdot) \equiv \int_0^\infty P^t(x,\cdot)\,a(dt).$$

A non-empty set $B \in \mathfrak{B}_{\mathbb{S}}$ is called ν_a-*petite* if ν_a is a non-trivial measure on $(\mathbb{S}, \mathfrak{B}_{\mathbb{S}})$ and a is a probability distribution on $(0,\infty)$ satisfying,

$$K_a(x,\cdot) \geq \nu_a(\cdot)$$

for all $x \in B$. The distribution a is called the *sampling distribution* for the petite set B. A petite set B has the property that all sets A are "equally accessible" from any $x \in B$. It provides some form of "irreducibility" in the *continuous* state space for the Markov process to have a *unique* invariant measure.

The following is a result for general Markov process from Meyn and Tweedie [43, Theorem 4.1].

Lemma 2.6.11. *Let B be a closed petite set, suppose that $\mathbb{P}(\tau_B^x < \infty) = 1$ for $x \in \mathbb{S}$, and that for some $\delta > 0$*

$$\sup_{x \in B} \mathbb{E}\left[\tau_B^x(\delta)\right] < \infty, \tag{2.39}$$

where $\tau_B^x(\delta) = \inf\{t \geq \delta : X^x(t) \in B\}$. Then, X is positive Harris recurrent.

Lemma 2.6.12. *Under the unboundedness and spread out assumptions (2.37)-(2.38) on interarrival time distributions,*

$$B = \{|x| \leq \kappa\} \text{ is a closed petite set for any } \kappa > 0. \tag{2.40}$$

This is the only place where the unboundedness and spread out assumptions are used. When these assumptions are unnecessarily restrictive, one can replace them by condition (2.40).

Theorem 2.6.13. *If there exists a $\delta > 0$ such that*

$$\lim_{|x| \to \infty} \frac{1}{|x|} \mathbb{E}\,|X^x(|x|\delta)| = 0, \tag{2.41}$$

then (2.39) holds for $B = \{x \in \mathbb{S} : |x| \leq \kappa\}$ with some $\kappa > 0$. In particular, X is positive Harris recurrent.

Proof. Let $0 < \epsilon < 1$, for example, $1/2$. From (2.41) there exists $\kappa \geq 1$ such that

$$\frac{1}{|x|} \mathbb{E}\,|X^x(|x|\delta)| \leq 1 - \epsilon \tag{2.42}$$

for all x such that $|x| > \kappa$. Let $B = \{x \in \mathbb{S} : |x| \leq \kappa\}$. It follows that, for some constant $b > 0$,

$$\mathbb{E}\,|X^x(|x|\delta)| \leq (1 - \epsilon)|x| + b1_B(x) \tag{2.43}$$

for all x. Let

$$n(x) = \begin{cases} |x|\delta & \text{if } x \notin B \\ \delta & \text{if } x \in B. \end{cases} \tag{2.44}$$

Because $\kappa \geq 1$, $n(x) \geq \delta$ for all $x \in \mathbb{S}$, it follows from (2.43) that

$$\mathbb{E}|X^x(n(x))| \leq (1 - \epsilon)|x| + b1_B(x) \leq |x| - \frac{\epsilon}{\delta} n(x) + \widetilde{b}1_B(x)$$

for some $\widetilde{b} > 0$ and all $x \in \mathbb{S}$. Proceeding exactly the same as in the proof of Theorem 2.1(ii) of Meyn and Tweedie [44], we have for each $x \in \mathbb{S}$,

$$\mathbb{E}[\tau_B^x(\delta)] \leq \frac{\delta}{\epsilon}\left(|x| + \widetilde{b}\right) < \infty$$

and

$$\sup_{x \in B} \mathbb{E}\left[\tau_B^x(\delta)\right] \leq \frac{\delta}{\epsilon}\left(\sup_{x \in B} |x| + \widetilde{b}\right) = \frac{\delta}{\epsilon}(\kappa + \widetilde{b}) < \infty.$$

Thus, $\mathbb{P}(\tau_B^x < \infty) = 1$ for each $x \in \mathbb{S}$ and it follows from Lemmas 2.6.11 and 2.6.12 that X is positive Harris recurrent. □

An outline of the proof of Theorem 2.6.10. Assume that the fluid model is stable. Let δ be the constant in the stability definition (Definition 2.4.1). Then, with probability one, each fluid limit $\bar{\mathbb{X}}$ of

$$\left\{ |x|^{-1} \mathbb{X}^x(|x|\cdot) \right\}$$

taken along $u = v = 0$ as $|x| \to \infty$ has the property that $\bar{Z}(\delta) = 0$. Since the limit point is unique (equal to 0), this leads to

$$\mathbb{P}\left\{ \lim_{|x|\to\infty} |x|^{-1} Z^x(|x|\delta) = 0 \right\} = 1$$

with x restricted to $u = v = 0$. Some uniform integrability analysis leads to

$$\lim_{|x|\to\infty} |x|^{-1} \mathbb{E}\Big[|Z^x(|x|\delta)| \Big] = 0$$

with x restricted to $u = v = 0$. Some careful analysis leads to (2.41) with $|x| \to \infty$ along arbitrary initial states. For more details of the proof, see Dai [14, Theorem 4.2] and Bramson [7]. □

2.7 Non-Uniqueness of Fluid Solutions

Consider the 2-station 5-class network pictured in Figure 2.1 operating under the SBP discipline π in (2.2). For concreteness, we assume that $\alpha_1 = 1$, $m_1 = m_3 = m_4 = 0.1$ and $m_2 = m_5 = 0.6$. We will construct two fluid solutions to fluid equations (2.7)-(2.12) and (2.17), one unstable and the other stable.

Both solutions start with one unit of fluid in buffer 1, namely, $Z(0) = (1, 0, 0, 0, 0)$. To specify a fluid solution, it is enough to specify $d_k(t)$ for all $t \geq 0$ and $k = 1, \ldots, 5$. One can then construct Z and T via

$$T_k(t) = m_k \int_0^t d_k(s)\, ds, \tag{2.45}$$

$$Z_k(t) = Z_k(0) + \int_0^t (d_{k-1}(s) - d_k(s))\, ds \tag{2.46}$$

for $k = 1, \ldots, 5$, where, by convention, $d_0(t) = t$ for $t \geq 0$. Clearly, one needs to show that the resulting (Z, T) satisfies (2.7)-(2.12) and (2.17).

We first construct an unstable solution. Let t_1 be the first time that buffer 1 empties. In $[0, t_1]$, buffer 1 is draining, buffer 2 and buffer 3 remain empty, buffer 4 is accumulating and buffer 5 remains empty. Server 2 is fully working on buffer 2, and thus $d_2(t) = \mu_2 = 5/3$. Server 1 spends just enough effort to keep buffer 3 empty and the remaining effort is spent on buffer 1. Since $d_3(t) = d_2(t) = 5/3$, server 1 spends a fraction $(5/3)(0.1) = 1/6$ of its effort on buffer 3. The remaining fraction, $5/6$, of its effort is spent on buffer 1, pumping

fluid at the rate of $d_1(t) = \mu_1(5/6) = 25/3$. During this period, both servers are 100% busy. At time

$$t_1 = \frac{1}{25/3 - 1} = \frac{3}{22},$$

buffer 1 empties. Let t_2 be the first time (after t_1) that buffer 2 empties. In $[t_1, t_2]$, all buffers at station 1 remain empty, buffer 2 is draining, and buffer 4 is accumulating. Thus, $d_1(t) = \alpha_1$ to keep buffer 1 empty. Server 2 is working fully on buffer 2, pumping fluid at the rate of $d_2 = \mu_2 = 5/3$. To keep buffer 3 empty, $d_3(t) = d_2(t) = 5/3$. Thus, the fraction of time that server 1 is busy is

$$d_1(t)m_1 + d_3(t)m_3 = 0.1 + 5/3(0.1) = \frac{4}{15}.$$

At the end of this time interval, all the fluid (the initial amount plus the newly arrived fluid) has moved into buffer 4. Thus, the state of the fluid network at time t_2 is

$$Z(t_2) = (0, 0, 0, 1 + t_2, 0).$$

During the entire interval $[0, t_2]$, the departure rate from buffer 3 is $d_3(t) = \mu_2 = 5/3$ and the arrival rate to buffer 1 is $\alpha_1 = 1$. Hence, the pipe of buffers 1-3 empties at time

$$t_2 = \frac{1}{\mu_2 - 1} = \frac{3}{2}$$

and

$$Z_4(t_2) = 1 + t_2 = \frac{1}{1 - m_2} = \frac{5}{2}.$$

At time t_2, server 2 begins to serve the buffer 4 fluid that accumulates in buffer 5, keeping server 1 100% busy. Let t_3 be the first time (after t_2) that buffer 4 empties. During $[t_2, t_3]$, buffers 1 and 5 are accumulating, buffer 4 is draining and buffers 2 and 3 remain empty. Thus, $d_4(t) = \mu_4 = 10$, $d_5(t) = \mu_5 = 5/3$ and $t_3 - t_2 = (5/2)/10 = 1/4$. Both servers are 100% busy during this interval. At t_3, buffer 5 continues to drain. Let t_4 is the first time (after t_3) that buffer 5 empties. During $[t_3, t_4]$, all buffers at station 2 and buffer 3 remain empty. Buffer 1 accumulates while buffer 5 is draining. During this interval, server 2 is 100% idle while server 1 is 100% busy. Since, during $[t_2, t_4]$, $d_5(t) = \mu_5 = 5/3$, we have

$$t_4 - t_2 = \frac{1}{1 - m_2}(1/\mu_5) = \frac{m_5}{1 - m_2} = 3/2$$

or $t_4 = 3$. At time t_4, all buffers are empty except that buffer 1 has accumulated 3/2 units of fluid, namely,

$$Z(3) = (3/2, 0, 0, 0, 0),$$

which is similar to the initial state except that buffer 1 has 3/2 units of fluid instead of 1 unit. Using mathematical induction, one can construct a fluid solution $Z = \{Z(t), t \geq 0\}$ such that for each integer $n \geq 1$,

$$Z(s_n) = \left(\left(\frac{3}{2}\right)^n, 0, 0, 0, 0\right),$$

where

$$s_n = 6\Big(\Big(\frac{3}{2}\Big)^n - 1\Big).$$

Since $|Z(s_n)| \to \infty$ as $n \to \infty$, Z is an unstable fluid solution.

The fluid solution Z diverges *linearly* in the sense that there exists a $c > 0$ such that $|Z(t)| \geq ct$ for $t \geq 0$. In fact, from the construction in each cycle $[s_{n-1}, s_n]$, buffers 2 and 5 have never been served simultaneously, namely,

$$d_2(t)d_5(t) = 0, \quad t \geq 0.$$

Therefore,

$$m_2 D_2(t) + m_5 D_5(t) \leq t, \quad t \geq 0.$$

Recall that $L_k(t) = \sum_{\ell \leq k} Z_\ell(t)$. One can check that

$$\begin{aligned} m_2 L_2(t) + m_5 L_5(t) &= m_2 L_2(0) + m_5 L_5(0) \\ &\quad + (m_2 + m_5)t - (m_2 D_2(t) + m_5 D_5(t)) \\ &\geq (0.2)t \end{aligned}$$

for $t \geq 0$. Hence, $|Z(t)| \geq ct$ for some $c > 0$.

The second fluid solution is much easier to describe. We start with initial state,

$$Z(0) = (1, 0, 0, 0, 0).$$

We keep all buffers empty except buffer 1. Thus, $d_1(t) = d_2(t) = d_3(t) = d_4(t) = d_5(t)$ for all $t \geq 0$. Let t_1 be the first time that buffer 1 empties. Since

$$d_1(t)m_1 + d_3(t)m_3 + d_5(t)m_5 = 1$$

for $t \in [0, t_1]$, we have

$$d_1(t) = 1/(0.8) = 1.25 \quad \text{for } t \in [0, t_1].$$

Thus, $t_1 = 1/(1.25 - 1) = 4$ and $Z(4) = 0$. In $[4, \infty)$, let $d_1(t) = \ldots = d_5(t) = 1$, we have $Z(t) = 0$ for $t \geq 4$. One can check that the corresponding Z and T constructed via (2.46) and (2.45) is a fluid solution satisfying (2.7)-(2.12) and (2.17). Notice that (2.17) is trivially satisfied by the fluid solution. In fact, this fluid solution is a fluid solution under the HLPPS discipline and many other non-idling disciplines.

In our presentation of the fluid solutions, we have used the terms "draining" and "accumulating" in an intuitive sense. A formal mathematical verification involves (a) specifying $d_k(t)$ in each time interval, $k = 1, \ldots, 5$, (b) constructing Z and T via (2.46) and (2.45) and (c) verifying that Z and T satisfy (2.7)-(2.12) and (2.17). Since this verification is tedious and intuitively obvious, we have chosen to omit the details. Readers are encouraged to fill in the details themselves.

Can both of th above fluid solutions be fluid limits, obtained from the queueing network under a fluid limit procedure? The answer is no. By Lemma 2.1.2, we know that for each fluid limit, we have

$$m_2 D_2(t) + m_5 D_5(t) \leq t, \quad t \geq 0. \tag{2.47}$$

But, for the second solution, we have

$$m_2 d_2(t) + m_5 d_5(t) = \frac{1.2}{0.8} = \frac{3}{2} > 1 \quad \text{for } t \in [0, t_1].$$

Thus,

$$m_2 D_2(t) + m_5 D_5(t) \geq (3/2)t$$

for $t \in [0, t_1]$ contradicting (2.47).

We have presented two fluid solutions for the fluid network. Clearly, one can construct infinitely many fluid solutions from these two solutions by alternately "running" stable and unstable solutions. In a general queueing network, it is difficult to rule out those fluid solutions that cannot be obtained from fluid limits. This fact makes formulating a practical, sharp converse to Theorem 2.6.10 difficult, if not impossible.

2.8 Stability of Fluid Models

Sections 2.5 and 2.6 have established strong connections between the stability of a stochastic queueing network and that of the corresponding deterministic fluid model. Studying the stability of the fluid model is certainly a great simplification. It is by no means trivial, however, at least in the multiclass setting.

Definition 2.8.1. For a fluid model with given routing matrix P and a non-idling HL discipline, *the stability region* of the discipline is the set of parameters α and $m > 0$ for which the fluid model under the discipline is stable.

For a discipline π, we use $\mathcal{D}_\pi$ to denote the stability region of π. When the discipline is clear from the context, we simply use $\mathcal{D}$ to denote the stability region. To avoid trivial complications,, we require $m > 0$ when $m \in \mathcal{D}_\pi$.

Definition 2.8.2. For a fluid model with given routing matrix P, *the global stability region* of the fluid model is the set of parameters α and $m > 0$ for which the fluid model is stable under *any* non-idling HL discipline.

We use $\mathcal{D}_\infty$ to denote the global stability region of a fluid model. Clearly,

$$\mathcal{D} \subset \mathcal{D}_\pi \tag{2.48}$$

for any discipline π.

Researchers have had limited success in determining the stability or global stability region for a given network. Most of this work uses some form of Lyapunov function. Determining stability regions is currently a very active area. Since the area is quite new, it is quite likely that motivated newcomers can make significant contributions.

2.8.1 FIFO Fluid Model of Kelly Type

We recall that a FIFO fluid model is of *Kelly type* if for each station j, the mean processing times for each class at the station are the same. With a slight abuse of notation, we use m_j to denote the mean processing time at station j. Throughout this section, we assume $\mathbb{X}$ is a FIFO fluid model solution of Kelly type.

Theorem 2.8.3. *Assume that the usual traffic condition (1.19) is satisfied. The FIFO fluid model of Kelly type is stable, i.e, assuming $m_k = m_\ell$ whenever $s(k) = s(\ell)$,*

$$\mathcal{D}_{\text{FIFO}} = \{(\alpha, m) : m > 0,\ \rho_j < 1, \quad j = 1, \ldots, j\}.$$

The main tool used in proving Theorem 2.8.3 is an *entropy* type Lyapunov function. For the fluid solution $\mathbb{X}$, such a Lyapunov function is a *functional* of the paths D and W, not just a *function* of $Z(t)$ for a fixed time t. Entropy type Lyapunov functions have proven to be quite effective in demonstrating the stability of FIFO and HLPPS fluid models. It is conceivable that other type of Lyapunov functions may be equally effective for these fluid models.

To define the Lyapunov functions used for FIFO fluid models of Kelly type, let

$$h(x) = x \log(x) \quad \text{for} \quad x \geq 0 \tag{2.49}$$

be the *entropy function*. Then h is a continuous function in $[0, \infty)$ with $h(0) = h(1) = 0$. Taking derivatives,

$$dh(x)/dx = \log x + 1 \quad \text{for} \quad x > 0, \tag{2.50}$$

$$d^2h(x)/dx^2 = 1/x > 0 \quad \text{for} \quad x > 0. \tag{2.51}$$

Thus, h is a convex function on $(0, \infty)$. Note that $h(x) < 0$ for $0 < x < 1$.

For the FIFO fluid model solution $\mathbb{X}$, we define

$$f(t) = \sum_k \int_t^{t+W_j(t)} \lambda_k h(\dot{D}_k(s)/\lambda_k)\, ds. \tag{2.52}$$

Recall that the symbol j denotes a station whereas k denotes a class. When they appear together, j is implicitly assumed to be $s(k)$. Note that we hope that Z eventually reaches an equilibrium where the flow rate out of buffer k is equal to the nominal total arrival rate λ_k to buffer k. Therefore, in equilibrium, $\dot{D}_k(s) = \lambda_k$ or $h(\dot{D}_k(s)/\lambda_k) = 0$. Thus, $f(t)$, in some sense, measures the imbalance between the current state at time t and the equilibrium state.

Our primary purpose in introducing the entropy Lyapunov function (2.52) is to show the stability of the FIFO fluid model of *Kelly type* when the usual traffic condition holds. Our first observation is that f is well defined and is a Lipschitz function because D_k is Lipschitz and hence $h(\dot{D}_k(s)/\lambda_k)$ is bounded for almost all s under Lebesgue measure on $(0, \infty)$.

Next, we show in the following lemma that $f(t) \geq 0$ for $t \geq 0$. Clearly, $f(t) = 0$ when $W(t) = 0$.

Lemma 2.8.4. *Let $\mathbb{X}$ be a FIFO fluid model solution of Kelly type with initial data satisfying (2.15). Assume that the usual traffic condition (1.19) is satisfied, Then $f(t) > 0$ when $W(t) \neq 0$.*

Proof. Assume that $W(t) \neq 0$. Let

$$\lambda_j^{\Sigma} = \sum_{k \in \mathcal{C}(j)} \lambda_k.$$

Then

$$f(t) = \sum_{j: W_j(t) > 0} \lambda_j^{\Sigma} \int_t^{t+W_j(t)} \sum_{k \in \mathcal{C}(j)} (\lambda_k / \lambda_k^{\Sigma}) h(\dot{D}_k(s)/\lambda_k)\, ds.$$

By Jensen's inequality, $f(t)$ is at least

$$\sum_{j: W_j(t) > 0} \lambda_j^{\Sigma} \int_t^{t+W_j(t)} h\left(\lambda_k^{\Sigma} \sum_{k \in \mathcal{C}(j)} \dot{D}_k(s)\right) ds. \tag{2.53}$$

When $W_j(t) > 0$, it is clear that $W_j(s) > 0$ for $s \in (t, t + W_j(t))$. By Part (b) of Lemma 2.4.6 and the initial condition (2.15) that

$$m_j \sum_{k \in \mathcal{C}(j)} \dot{D}_k(s) = 1.$$

Thus,

$$h\left(\lambda_k^{\Sigma} \sum_{k \in \mathcal{C}(j)} \dot{D}_k(s)\right) = h(1/\rho_j) > 0$$

for almost all s on $(t, t + W_j(t))$, and hence

$$f(t) \geq \sum_j \lambda_j^{\Sigma} h(1/\rho_j) W_j(t) > 0. \tag{2.54}$$

□

Now, we show that f is nonincreasing.

Lemma 2.8.5. *Under the assumptions in Lemma 2.8.4, for each point t that is regular for both the fluid model solution $\mathbb{X}$ and f, $\dot{f}(t) \leq 0$.*

Proof. Taking the derivative with respect to f,

$$\begin{aligned}
\dot{f}(t) &= \sum_k \Big[(1+\dot{W}_j(t))\lambda_k h(\dot{D}_k(t+W_j(t))/\lambda_k) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big] \\
&= \sum_k \Big[(1+\dot{W}_j(t))\lambda_k h\big((\dot{A}_k(t)/\lambda_k)/(1+\dot{W}_j(t))\big) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big] \\
&= \sum_k \Big[\dot{A}_k \log\big((\dot{A}_k(t)/\lambda_k)/(1+\dot{W}_j(t))\big) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big] \\
&= \sum_k \Big[\lambda_k h(\dot{A}_k(t)/\lambda_k) - \dot{A}_k(t)\log(1+\dot{W}_j(t)) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big] \\
&= \sum_k \Big[\lambda_k h(\dot{A}_k(t)/\lambda_k) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big] \\
&\quad - \sum_j \left(\sum_{k\in\mathcal{C}(j)} \dot{A}_k(t)\right) \log(1+\dot{W}_j(t)),
\end{aligned}$$

where, in the second equality, we have used the FIFO equation (2.14) to obtain

$$(1+\dot{W}_j(t))\dot{D}_k(t+W_j(t)) = \dot{A}_k(t).$$

From (2.11), $\dot{Y}_j(t) > 0$ implies $W_j(t) = 0$ and hence $\dot{W}_j(t) = 0$. Thus,

$$\left(\sum_{k\in\mathcal{C}(j)} \dot{A}_k(t)\right) \log(1+\dot{W}_j(t)) = \left(\sum_{k\in\mathcal{C}(j)} \dot{A}_k(t) + \mu_j\dot{Y}_j(t)\right) \log(1+\dot{W}_j(t)).$$

If follows from (2.9) that

$$\left(\sum_{k\in\mathcal{C}(j)} \dot{A}_k(t) + \mu_j\dot{Y}_j(t)\right) \log(1+\dot{W}_j(t)) = \mu_j h(1+\dot{W}_j(t)).$$

We are going to show that

$$\sum_j \mu_j h(1+\dot{W}_j(t)) \geq \sum_k \dot{Z}_k(t), \tag{2.55}$$

$$\sum_k \Big(\lambda_k h(\dot{A}_k(t)/\lambda_k) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big) \leq \sum_k \dot{Z}_k(t). \tag{2.56}$$

It is clear that (2.55) and (2.56) imply $\dot{f}(t) \leq 0$.

To show (2.55), recall that h is convex with $h(1) = 0$ and $\dot{h}(1) = 1$. Thus,

$$\mu_j h(1+\dot{W}_j(t)) \geq \mu_j \dot{W}_j(t),$$

and using (2.13), we have (2.55). To show (2.56), we notice that

$$\lambda_k h(\dot{A}_k(t)/\lambda_k) = \lambda_k h\left(\lambda_k^{-1}\left[\alpha_k + \sum_\ell (\lambda_\ell P_{\ell k})(\dot{D}_\ell(t)/\lambda_\ell)\right]\right).$$

By the traffic equation (1.17),

$$\lambda_k^{-1}\Big[\alpha_k + \sum_\ell (\lambda_\ell P_{\ell k})\Big] = 1.$$

So by Jensen's inequality and $h(1) = 0$, $\lambda_k h(\dot{A}_k(t)/\lambda_k)$ is less than or equal to

$$\alpha_k h(1) + \sum_\ell \lambda_\ell P_{\ell k} h(\dot{D}_\ell(t)/\lambda_\ell) = \sum_\ell \lambda_\ell P_{\ell k} h(\dot{D}_\ell(t)/\lambda_\ell).$$

Thus, for all k,

$$\lambda_k h(\dot{A}_k(t)/\lambda_k) \le \sum_\ell \lambda_\ell P_{\ell k} h(\dot{D}_\ell(t)/\lambda_\ell).$$

Summing over k, we have

$$\sum_k \Big(\lambda_k h(\dot{A}_k(t)/\lambda_k) - \lambda_k h(\dot{D}_k(t)/\lambda_k)\Big) \le -\sum_k (1 - \sum_\ell P_{k\ell})\lambda_k h(\dot{D}_k(t)/\lambda_k).$$

Using the convexity of $h(x)$ again, the right-hand side is less than equal to

$$\begin{aligned} -\sum_k (1 - \sum_\ell P_{k\ell})\lambda_k (\dot{D}_k(t)/\lambda_k - 1) &= -\sum_k (1 - \sum_\ell P_{k\ell})(\dot{D}_k(t) - \lambda_k) \\ &= \sum_k \Big(\alpha_k - (1 - \sum_\ell P_{k\ell})\dot{D}_k(t)\Big) \\ &= \sum_k \dot{Z}_k(t). \end{aligned}$$

□

Now we are ready to prove Theorem 2.8.3.

Proof of Theorem 2.8.3. First, we observe that (2.55) can be strengthened to

$$\sum_j \mu_j h(1 + \dot{W}_j(t)) \ge \sum_k \dot{Z}_k(t) + c\sum_j (\dot{W}_j(t))^2,$$

for some constant c. This follows from the fact that

$$h(1 + x) \ge x + c(N)x^2 \quad \text{for } x \in [-1, N],$$

for some constant $c = c(N)$, where N is the Lipschitz constant for W. Thus, following the proof of Lemma 2.8.5,

$$\dot{f}(t) \le -c\sum_j (\dot{W}_j(t))^2.$$

By Chebyshev's inequality,

$$f(t) - f(t') \ge c\int_t^{t'} (\dot{W}_j(s))^2\, ds \ge \frac{c}{t' - t}(W_j(t') - W_j(t))^2$$

for $0 \leq t < t'$ and any j. Suppose that $W_j(t) \neq 0$ for given t and j. Setting $t' = t + \tau_j(t)$, where $t + \tau_j(t)$ is the first time s after t such that $W_j(s) = 0$, then

$$f(t) - f(t + \tau_j(t)) \geq \frac{c\,(W_j(t))^2}{\tau_j(t)}.$$

By Proposition 2.2.1 there exists a constant c_1 such that

$$\tau_j(t) \leq c_1 W^M(t) \equiv \max_j W_j(t).$$

Choosing j so that $W_j(t) = W^M(t)$, it follows from the monotonicity of f that

$$f(t) - f(t + c_1 W^M(t)) \geq (c/c_1) W^M(t) \quad \text{for } t \geq 0. \tag{2.57}$$

One can iterate (2.57) along times

$$t_{i+1} = t_i + c_1 W^M(t_i), \tag{2.58}$$

$i = 0, 1, \ldots$, where $t_0 = 0$. This gives

$$f(0) - f(t_i) \geq c_2 t_i \quad \text{for } i = 0, 1, \ldots,$$

where $c_2 = (c/c_1^2)$. Since $f(t_i) \geq 0$, we have

$$t_\infty \equiv \lim_{i \to \infty} t_i \leq f(0)/c_2.$$

Taking limits in (2.58), we have $W(t_\infty) = 0$ which implies $f(t_\infty) = 0$. Thus,

$$W(t) = 0 \quad \text{for } t \geq t_\infty.$$

Assume that $|Z(0)| \leq 1$. We have $f(0) \leq c_3$ for some constant c_3. Therefore,

$$W(t) = 0 \quad \text{for } t \geq c_3/c_2,$$

proving the stability of the fluid model. □

2.8.2 Piecewise Linear Lyapunov Functions

For any fluid model solution $\mathbb{X}$ satisfying (2.7)-(2.12), as before, let

$$L(t) = (I - P')^{-1} Z(t)$$

be the vector of potential fluid levels for each class at time t. Using (2.7), (2.8) and (2.12), we have

$$L_k(t) = L_k(0) + \lambda_k t - \mu_k T_k(t) \quad \text{for } k = 1, \ldots, K. \tag{2.59}$$

For a given $x = (x_k) > 0$, let

$$f_j(t) = \sum_{k \in \mathcal{C}(j)} x_k L_k(t). \tag{2.60}$$

be the generalized potential workload for station j at time t. If $x = m$, $m_k L_k(t)$ is the potential amount of class k work for server $j = s(k)$ at time t. Thus, when $x = m$, $f_j(t)$ is the potential workload or total workload for station j at time t. For any $x, z \in \mathbb{R}^K$, let

$$h(x, z) = C \operatorname{diag}(x)(I - P)^{-1} z, \tag{2.61}$$

where $\operatorname{diag}(x)$ is the $K \times K$ diagonal matrix with diagonal entries given by x_k, $k = 1, \ldots, K$. For a fixed z, $h(x, z)$ is a linear function of x. For a fixed x, $h(x, z)$ is a linear function of z. One can check that $f_j(t) = h_j(x, Z(t))$ for $j = 1, \ldots, J$. By (2.59),

$$f_j(t) = f_j(0) + \sum_{k \in \mathcal{C}(j)} x_k(\lambda_k t - \mu_k T_k(t)). \tag{2.62}$$

Define

$$f(t) = \max_{j=1,\ldots,J} f_j(t). \tag{2.63}$$

Since each f_j is a linear function of $Z(t)$, f is a piecewise linear Lyapunov function of $Z(t)$. For $(x_k) > 0$, $f(t) = 0$ if and only if $Z(t) = 0$. One can check that $f(t)$ is a Lipschitz function of t, and is thus absolutely continuous. In this section, t is said to be regular if both $\mathbb{X}$ and f are differentiable at t. As before, whenever a derivative is used at time t, t is assumed to be a regular point.

Lemma 2.8.6. *For a regular point t, if*

$$f_j(t) = \max_{i=1,\ldots,J} f_i(t),$$

$\dot{f}(t) = \dot{f}_j(t)$.

It is crucial that t is a regular point (for both f and $\mathbb{X}$). Even if t is a regular point for $\mathbb{X}$ and hence for f_j, $j = 1, \ldots, J$, it is not necessarily regular for f.

Proposition 2.8.7. *Suppose that there exist $(x_k) > 0$ and $\epsilon > 0$ such that for each j*

$$W_j(t) > 0 \quad \textit{implies } \dot{f}_j(t) \leq -\epsilon \tag{2.64}$$

and

$$W_j(t) = 0 \quad \textit{implies } f_j(t) \leq \max_{i \neq j} f_i(t). \tag{2.65}$$

Then $Z(t) = 0$ for $t \geq f(0)/\epsilon$.

Proof. Assume that $f(t) > 0$. Then $f(t) = f_j(t)$ for some j. By assumption (2.65), we can choose j such that $W_j(t) > 0$. By Lemma 2.8.6, $\dot{f}(t) = \dot{f}_j(t)$. Since $W_j(t) > 0$, assumption (2.64) yields $\dot{f}(t) = \dot{f}_j(t) \leq -\epsilon$. The lemma follows from Lemma 2.4.5. □

Condition (2.64) leads to natural linear inequalities on (x_k). In fact, we have the following lemma.

Lemma 2.8.8. *Suppose that there exist $(x_k) > 0$ such that for each class k with $j = s(k)$*

$$\sum_{\ell \in \mathcal{C}(j)} \lambda_\ell x_\ell < x_k \mu_k. \tag{2.66}$$

Then condition (2.64) holds with some $\epsilon > 0$ for any fluid model solution $\mathbb{X}$ satisfying (2.7)-(2.12).

Proof. Assume that there exist $(x_k) > 0$ satisfying linear constraints (2.66). Let

$$\epsilon = \min_k \Big(x_k \mu_k - \sum_{\ell \in \mathcal{C}(j)} \lambda_\ell x_\ell \Big) > 0.$$

If $W_j(t) > 0$, by condition (2.11),

$$\sum_{k \in \mathcal{C}(j)} \dot{T}_k(t) = 1. \tag{2.67}$$

From (2.62),

$$\dot{f}_j(t) = \sum_{k \in \mathcal{C}(j)} (x_k \lambda_k - x_k \mu_k \dot{T}_k(t)). \tag{2.68}$$

Since $\dot{T}_k(t) \geq 0$, the lemma follows from (2.67) and (2.68). □

Condition (2.65) often generates non-linear constraints on (x_k). However, when $J = 2$, we again have linear inequalities on (x_k).

Lemma 2.8.9. *Assume that $J = 2$ and*

$$h_1(x, e_k) \leq h_2(x, e_k) \quad \text{for each } k \in \mathcal{C}(2), \tag{2.69}$$

$$h_2(x, e_k) \leq h_1(x, e_k) \quad \text{for each } k \in \mathcal{C}(1), \tag{2.70}$$

where h is defined in (2.61) and e_k is the K-dimensional vector with the kth component 1 and all other components 0. Then (2.65) is satisfied.

Proof. For a fixed x, $h(x, z)$ is a linear function of z. Assume that $W_1(t) = 0$. Then

$$Z(t) = \sum_{k \in \mathcal{C}(2)} Z_k(t) e_k.$$

Therefore, by (2.69),

$$\begin{aligned} f_1(t) &= h_1(x, Z(t)) = \sum_{k \in \mathcal{C}(2)} Z_k(t) h_1(x, e_k) \\ &\leq \sum_{k \in \mathcal{C}(2)} Z_k(t) h_2(x, e_k) = h_2(x, Z(t)) \\ &= f_2(t). \end{aligned}$$

This proves (2.65) for $j = 1$. Similarly, one can prove that (2.65) holds for $j = 2$. □

Note that for a fixed z, $h(x, z)$ is a linear function of x. Thus, (2.69)-(2.70) are *linear* inequalities on (x_k). Summarizing Lemmas 2.8.8 and 2.8.9, we have the following theorem.

Theorem 2.8.10. *For $J = 2$, assume that there exist $(x_k) > 0$ satisfying (2.66) and (2.69)-(2.70). Then the fluid model is globally stable.*

Consider the following linear program (LP):

$$\max \epsilon \tag{2.71}$$

with constraints

$$\sum_{\ell \in \mathcal{C}(j)} \lambda_\ell x_\ell + \epsilon \le x_k \mu_k, \quad k = 1, \ldots, K, \tag{2.72}$$

$$h_1(x, e_k) \le h_2(x, e_k) \quad \text{for each } k \in \mathcal{C}(2), \tag{2.73}$$

$$h_2(x, e_k) \le h_1(x, e_k) \quad \text{for each } k \in \mathcal{C}(1), \tag{2.74}$$

$$x_k \ge 0, \quad k = 1, \ldots, K, \tag{2.75}$$

$$0 \le \epsilon \le 1. \tag{2.76}$$

Since the constraints (2.72)-(2.75) are homogeneous in x and ϵ, the LP (2.71) has objective value either 0 or 1.

Corrollary 2.8.11. *If LP (2.71) has objective value 1, the fluid model is globally stable.*

Proof. One can check that there exist $(x_k) > 0$ satisfying (2.66) and (2.69)-(2.70) if and only if there exist $(x_k) \ge 0$ and $\epsilon > 0$ satisfying (2.72)-(2.75). □

LP (2.71) has $K + 1$ variables and $2K$ constraints. For given input parameter (α, m, P), the LP can be solved numerically by efficient algorithms.

Example. Consider the 2-station 5-class network in Figure 2.1. Constraint (2.66) gives rise to

$$\alpha_1(x_1 + x_3 + x_5) < \mu_1 x_1, \tag{2.77}$$

$$\alpha_1(x_1 + x_3 + x_5) < \mu_3 x_3, \tag{2.78}$$

$$\alpha_1(x_1 + x_3 + x_5) < \mu_5 x_5, \tag{2.79}$$

$$\alpha_1(x_2 + x_4) < \mu_2 x_2, \tag{2.80}$$

$$\alpha_1(x_2 + x_4) < \mu_4 x_4. \tag{2.81}$$

Constraint (2.69) takes the form

$$x_3 + x_5 \le x_2 + x_4, \tag{2.82}$$

$$x_5 \le x_4, \tag{2.83}$$

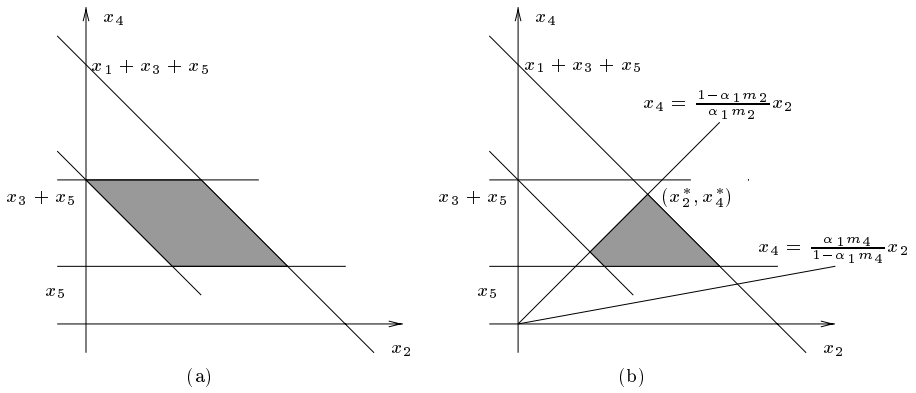

Figure 2.3: (a) The region of (x_2, x_4) constrained by (2.88)-(2.89). (b) The region in (a) intersects with region (2.87).

and constraint (2.70) takes the form

$$x_2 + x_4 \leq x_1 + x_3 + x_5, \tag{2.84}$$
$$x_4 \leq x_3 + x_5, \tag{2.85}$$
$$0 \leq x_5. \tag{2.86}$$

Since we require $(x_k) > 0$, constraint (2.86) is redundant. To be consistent with the notation for general networks, we retain the redundant constraint.

Proposition 2.8.12. *There exist $(x_k) > 0$ satisfying (2.77)-(2.86) if and only (2.3)-(2.5) hold.*

Proof. Assume that there exist $(x_k) > 0$ satisfying (2.77)-(2.86). Conditions (2.77)-(2.79) imply (2.3), and conditions (2.80)-(2.81) imply (2.4). Assume that (2.3) and (2.4) hold. Note that (2.80)-(2.81) are equivalent to

$$x_2 \frac{\alpha_1 m_4}{1 - \alpha_1 m_4} < x_4 < \frac{1 - \alpha_1 m_2}{\alpha_1 m_2} x_2, \tag{2.87}$$

and for a given $(x_1, x_3, x_5) > 0$, (x_2, x_4) satisfying (2.82)-(2.86) is equivalent to the fact that (x_2, x_4) belongs to the parallelogram bounded by

$$x_5 \leq x_4 \leq x_3 + x_5, \tag{2.88}$$
$$x_3 + x_5 \leq x_2 + x_4 \leq x_1 + x_3 + x_5; \tag{2.89}$$

see the shaded region in Part (a) of Figure 2.3. Note that the line

$$x_4 = \frac{(1 - \alpha_1 m_2)}{\alpha_1 m_2} x_2$$

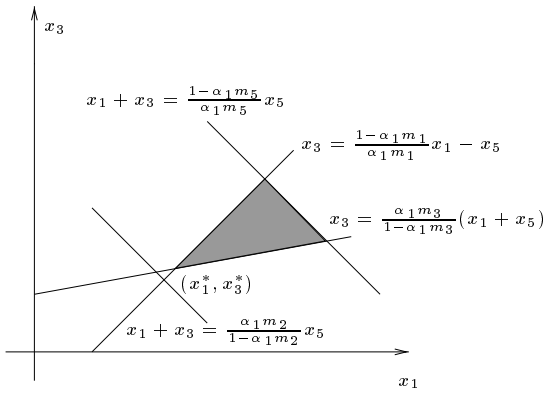

Figure 2.4: The region of (x_1, x_3) constrained by (2.77)-(2.79) and (2.90)

intersects the line

$$x_2 + x_4 = x_1 + x_3 + x_5,$$

with (x_1, x_3, x_5) fixed, at

$$(x_2^*, x_4^*) = \Big(\alpha_1 m_2(x_1 + x_3 + x_5),\ (1 - \alpha_1 m_2)(x_1 + x_3 + x_5)\Big).$$

Therefore, the region (2.87) and the parallelogram have a nonempty intersection, (the shaded region in Part (b) of Figure 2.3), if and only if $x_4^* > x_5$ or equivalently

$$(1 - \alpha_1 m_2)(x_1 + x_3 + x_5) > x_5. \tag{2.90}$$

Finally, conditions (2.79) and (2.90) imply (2.5).

Conversely, assume that (2.3)-(2.5) hold. We would like to show that there exist $(x_k) > 0$ satisfying (2.77)-(2.86). The argument in the preceding paragraph shows that it is enough to find $(x_1, x_3, x_5) > 0$ satisfying (2.77)-(2.79) and (2.90). For a fixed $x_5 > 0$, (2.77)-(2.78) are equivalent to

$$\frac{\alpha_1 m_3}{1 - \alpha_1 m_3}(x_1 + x_5) < x_4 < \frac{1 - \alpha_1 m_1}{\alpha_1 m_1} x_1 - x_5. \tag{2.91}$$

The region is nonempty in $(x_1, x_3) > 0$ and the two boundaries intersect at

$$(x_1^*, x_3^*) = \Big(\frac{\alpha_1 m_1}{1 - \alpha_1(m_1 + m_3)} x_5, \quad \frac{\alpha_1 m_3}{1 - \alpha_1(m_1 + m_3)} x_5\Big).$$

From (2.3),

$$x_1^* + x_3^* = \frac{\alpha_1(m_1 + m_3)}{1 - \alpha_1(m_1 + m_3)} x_5 < \frac{1 - \alpha_1 m_5}{\alpha_1 m_5} x_5.$$

Therefore, for any fixed $x_5 > 0$, the region of (x_1, x_3) constrained by (2.77)-(2.79) is nonempty; see the shaded region in Figure 2.4. Since (2.5) holds, this region has a nonempty intersection with the region constrained by (2.90). □

2.8.3 Two-station Multi-type Networks

In this section, we show that the piecewise linear Lyapunov functions introduced in Section 2.8.2 are sharp in determining the global stability of 2-station *multi-type* fluid models. A multiclass queueing network is said to be multi-type if $P_{k\ell}$ is either 1 or 0, namely, the routing is *deterministic*. In a multi-type queueing network, there can be more than one external arrival source. When there is a single external arrival source, the multi-type queueing network is a *re-entrant* line.

For a multi-type fluid model, the linear constraints (2.66) and (2.69)-(2.70) can be solved explicitly. The resulting region in terms of (α, m, P) happens to *characterize* the global stability of the fluid model. The global stability conditions for a two-station fluid network fall into two categories: the nominal workload conditions (1.19) that arise because classes at the same station must share the server's time; and the *virtual workload conditions*, generalizing condition $\alpha_1(m_2 + m_5) < 1$ for the fluid network in Figure 2.1 that arise due to the interactions between virtual stations and push starts (to be described shortly).

Two intuitively appealing phenomena give rise to the virtual workload conditions. The intuition behind the first of these phenomena is best described in the context of queueing networks. The second phenomenon is most easily understood in the context of fluid networks.

Consider the 2-station 5-class queueing network in Figure 2.1. If we give highest priority to class 5 at Station 1 and to class 2 at Station 2, these two classes can only be served simultaneously during a transient initial period, see Lemma 2.1.2. Thus, these two classes form a "virtual station" and, although they are served at different stations, the workload per unit of time at these two classes cannot exceed 1. This virtual station gives rise to the virtual workload condition:

$$\alpha_1(m_2 + m_5) < 1,$$

which we refer to as a "virtual station condition". These conditions also apply to fluid networks.

The fluid network of Figure 2.5 illustrates the second phenomenon giving rise to virtual workload conditions. Assume that the nominal workload conditions (1.19) hold. If we give highest priority to class 1 at Station 1 and to class 2 at Station 2 in this network, the fluid levels in these two buffers will reach zero and remain zero thereafter. For the sake of our discussion, we assume that these two buffers are always empty. Then, the server at Station 1 will *constantly* devote a fraction $\alpha_1 m_1$ of its time to class 1 to keep the buffer empty, and hence have only a fraction $1 - \alpha_1 m_1$ of its time left for the other classes at Station 1. Similarly, the server at Station 2 will constantly devote a fraction $\alpha_1 m_2$ of its time to class 2 and have only a fraction $1 - \alpha_1 m_2$ of its time left for the other classes at Station 2. Note that in a queueing network we cannot anticipate a constant, uninterrupted devotion of time to these classes, but we can in a fluid network. The fact that the servers are slowed due to their efforts on classes 1 and 2 "magnifies" the time required to serve each unit of fluid in the remaining

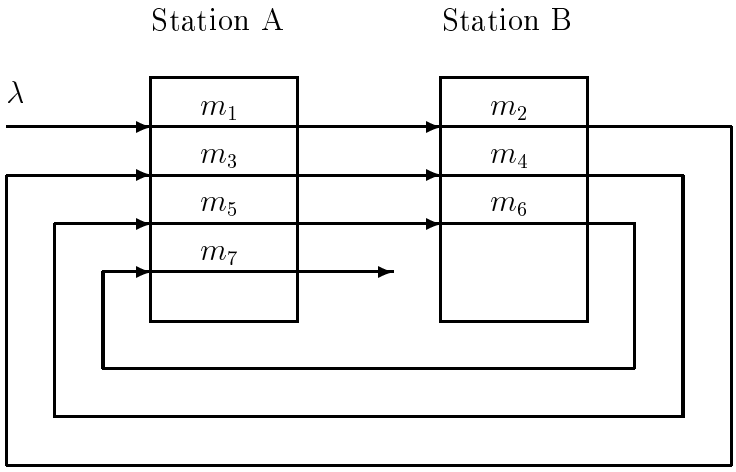

Figure 2.5: A seven class network

classes. In particular, the server at Station 1 will require $m_7/(1-\alpha_1 m_1)$ units of time to complete one unit of class 7 fluid and the server at Station 2 will require $m_4/(1-\alpha_1 m_2)$ units of time to complete one unit of class 4 fluid. Since buffers 1 and 2 remain empty, fluid passes through them as quickly as it arrives, and hence arrives at buffer 3 at rate α_1. Thus, push starting the first two classes magnifies the virtual station condition:

$$\alpha_1(m_4 + m_7) < 1$$

in the induced network (consisting of classes 3-7) to give the virtual workload condition:

$$\frac{\alpha_1 m_4}{1-\alpha_1 m_2} + \frac{\alpha_1 m_7}{1-\alpha_1 m_1} < 1,$$

which ensures that the virtual station can divide its time between serving the two classes. We refer to this condition as a "push start condition".

Together, these two phenomena explain all the virtual workload conditions of two-station fluid networks. Although these ideas are intuitively appealing, formalizing them is more involved. For the reminder of this section, only re-entrant fluid models are considered.

We formalize the conditions under which classes at different stations cannot receive service simultaneously in the following way. The first notion in our characterization of virtual stations is the idea of an *excursion* or set of consecutive classes at the same station.

We use symbol e to denote the eth excursion. Let $c[e]$ denote the set of classes in e. We partition $c[e]$ into the *last class* and all the rest, which we call *first classes* of the excursion. We let $\ell[e]$ denote the last class and $f[e]$ the set of first classes in $c[e]$. If an excursion consists of only one class, that class is the last class and the excursion has no first classes.

Definition 2.8.13. A set of excursions is said to be *separating* if it contains no consecutive excursions. A separating set is said to be *strict* if it does not contain the first excursion.

Therefore, a set S of excursions is separating if and only if whenever $e \in S$, $e-1 \notin S$ and $e+1 \notin S$.

The set of excursions at Station 1, for example, is separating. Likewise, the set of excursions at Station 2 is separating. We refer to these two separating sets as *trivial* separating sets.

Each non-trivial separating set S of excursions induces a virtual station $V(S)$ or maximal collection of classes with the property that if we give highest priority to these classes the two servers can simultaneously serve classes of $V(S)$ only during a transient initial period.

Definition 2.8.14. Each separating set S of excursions induces a collection $V(S)$ consisting of the classes in excursions of S together with the first classes of excursions whose immediate predecessor is not in S. Thus,

$$V(S) = (\cup_{e \in S} c[e]) \bigcup (\cup_{e \notin S} f[e+1]) .$$

When S is strictly separating we refer to $V(S)$ as a *virtual station*.

A virtual station V, then, is a set of classes satisfying:

1. No class of the first excursion is in V, i.e., $c[1] \cap V = \emptyset$.

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

3. The last class of an excursion (except a last excursion) is in V if and only if no class of the next excursion is in V, i.e., if e is not the last excursion, $\ell[e] \in V$ if and only if $c[e+1] \cap V = \emptyset$.

In the network of Figure 2.1, the separating set $S = \{2, 5\}$ of excursions gives rise to the virtual station $V(S)$ consisting of classes 2 and 5 (there are no first classes in excursion 3). This is the only virtual station that is not itself a subset of the classes at a station. The following proposition justifies usage of the term virtual station. The proof of the proposition is similar to the proof of Lemma 2.1.2.

Proposition 2.8.15. *Consider a multiclass queueing network. Assume that* $\prod_{k \in V(S)} Z_k(0) = 0$. *If the classes in* $V(S)$ *have higher priority than the classes that are not in* $V(S)$, *then*

$$\prod_{k \in V(S)} Z_k(t) = 0 \quad \textit{for all } t \geq 0. \tag{2.92}$$

For each excursion e and a separating set S, let

$$V_j^e(S) = \{k \in V(S) : k > \ell[e] \text{ and } s(k) = j\}, \quad j = 1, 2.$$

Theorem 2.8.16. *The following are equivalent. (a) A two-station fluid network is globally stable; (b) the parameter $(\alpha_1, m) > 0$ satisfies the usual traffic condition (1.19) and*

$$\frac{\alpha_1 \sum_{k \in V_1^e(S)} m_k}{1 - \alpha_1 \sum_{k < \ell[e], s(k)=1} m_k} + \frac{\alpha_1 \sum_{k \in V_2^e(S)} m_k}{1 - \alpha_1 \sum_{k < \ell[e], s(k)=2} m_k} < 1 \tag{2.93}$$

for each excursion e and each separating set S. (c) there exist $(x_k) > 0$ satisfying the linear constraints (2.66) and (2.69)-(2.70).

We refer to the conditions (2.93) as the virtual workload conditions. When the summations in the denominators are empty, we refer to the virtual workload condition (2.93) as a virtual station condition. Otherwise, the condition involves push starting classes $\{k : k < \ell[e]\}$ and we refer to it as a push start condition. For example, the virtual workload conditions of the fluid network in Figure 2.5 are:

$$\begin{aligned}
\alpha_1(m_2 + m_5 + m_7) &< 1, \\
\alpha_1(m_2 + m_4 + m_7) &< 1, \\
\frac{\alpha_1 m_3}{1 - \alpha_1 m_1} + \alpha_1 m_6 &< 1, \\
\frac{\alpha_1 m_4}{1 - \alpha_1 m_2} + \frac{\alpha_1 m_7}{1 - \alpha_1 m_1} &< 1.
\end{aligned}$$

Proof of Theorem 2.8.16. The equivalence of (b) and (c) was established in Dai and Vande Vate [20]. Their proof, given for any two-station multi-type fluid model, involves converting the linear constraints (2.66) and (2.69)-(2.70) into a parametric *network flow problem*. Proposition 2.8.12 offers a direct argument for the 2-station 5-class fluid model without using the network flow problem. Theorem 2.8.10 asserts that (c) implies (a). To show that (a) implies (b), we assume that either $\rho_j \geq 1$ for some j or (2.93) is violated for some excursion e and separating set S. If $\rho_j \geq 1$ for some j, the proof of Corollary 2.5.4 leads to the instability of the fluid model. Now assume that

$$\frac{\alpha_1 \sum_{k \in V_1^e(S)} m_k}{1 - \alpha_1 \sum_{k < \ell[e], s(k)=1} m_k} + \frac{\alpha_1 \sum_{k \in V_2^e(S)} m_k}{1 - \alpha_1 \sum_{k < \ell[e], s(k)=2} m_k} \geq 1$$

for some excursion e and separating set S. Then, Dai and Vande Vate [20] shows that there exists a SBP discipline under which the fluid model is unstable. □

In fact, Dai and Vande Vate [20] proved the following theorem.

Theorem 2.8.17. *If there exist an excursion e and a separating set S such that*

$$\frac{\alpha_1 \sum_{k\in V_1^e(S)} m_k}{1-\alpha_1 \sum_{k<\ell[e],s(k)=1} m_k} + \frac{\alpha_1 \sum_{k\in V_2^e(S)} m_k}{1-\alpha_1 \sum_{k<\ell[e],s(k)=2} m_k} > 1, \tag{2.94}$$

then there exists a fluid solution $\mathbb{X}$ under a SBP discipline such that $|Z(t)| \to \infty$ as $t \to \infty$.

2.9 Stabilizing Queueing Networks

When the usual traffic condition (1.19) is satisfied, we offer two techniques to stabilize open queueing networks.

2.9.1 The Leaky-Bucket-Controlled Network

If one has no control over the service discipline at each station in a network, one can add "leaky bucket" control stations to stabilize the network.

Example. Consider the 2-station 5-class re-entrant line pictured in Figure 2.1. Assume that the traffic conditions (2.3) and (2.4) are satisfied, namely,

$$\rho_1 = \alpha_1(m_1+m_3+m_5) < 1 \quad \text{and} \quad \rho_2 = \alpha_1(m_2+m_4) < 1.$$

We construct a new network, called the *leaky-bucket-controlled network*, by adding 5 new control stations, one station for each original job class. As pictured in Figure 2.6, stations 3-7 are leaky bucket control stations that serve job classes 6-10. As indicated in the figure, jobs in an original class k have to come from a control station $k+2$ that serves class $k+5$, $k = 1, \ldots, 5$. In the controlled network, jobs arrive to class 6 from the outside at rate α_1.

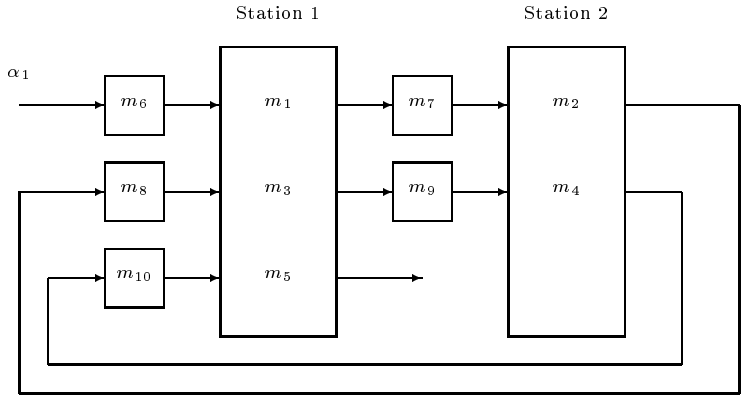

Figure 2.6: The 2-station 5-class re-entrant line under a "leaky bucket" control scheme

Proposition 2.9.1. *Consider the 2-station 5-class re-entrant line. Assume that $\rho_1 < 1$ and $\rho_2 < 1$, and the leaky-bucket-controlled network is constructed as in Figure 2.6. Let*

$$m_6 = m_8 = m_{10} = \alpha_1^{-1}(1+\rho_1)/2 \quad \textit{and} \quad m_7 = m_9 = \alpha_1^{-1}(1+\rho_2)/2. \quad (2.95)$$

Then the controlled fluid network is stable under any non-idling service discipline.

Proof. Let $\mathbb{X}$ be a fluid solution to the controlled fluid network with $|Z(0)| \leq 1$ and $d_k(t) = \dot{D}_k(t)$ be the departure rate from class k at time t. The input load to station 1 at time t is

$$d_6(t)m_1 + d_8(t)m_3 + d_{10}(t)m_5.$$

Since $d_k(t) \leq \mu_k = 1/m_k$ for any class k, the input load to station 1 is bounded above by the *peak load*

$$\widehat{\rho}_1 \equiv \alpha_1 \left(\frac{2}{1+\rho_1}\right)(m_1 + m_3 + m_5) = \frac{2\rho_1}{1+\rho_1} < 1.$$

Let

$$\begin{aligned} f_1(t) &= m_1 Z_1(t) + m_3 Z_3(t) + m_5 Z_5(t) \quad \text{and} \\ f_2(t) &= m_2 Z_2(t) + m_4 Z_4(t) \end{aligned}$$

be the immediate workload at stations 1 and 2, respectively. One can check that whenever $f_1(t) > 0$, $\dot{f}_1(t) \leq \widehat{\rho}_1 - 1$. Thus, $f_1(t) = 0$ for $t \geq f_1(0)/(1-\widehat{\rho}_1)$. Similarly, $f_2(t) = 0$ for $t \geq f_2(0)/(1-\widehat{\rho}_2)$, where $\widehat{\rho}_2 = 2\rho_2/(1+\rho_2)$. Therefore, there exists a $\delta > 0$ such that for any fluid solution $\mathbb{X}$ to the controlled fluid network with $|Z(0)| \leq 1$, the fluid levels at both stations 1 and 2 reach zero at time δ and remain zero afterwards.

Since

$$\begin{aligned} \rho_3 = \rho_5 = \rho_7 &= (1+\rho_1)/2 < 1, \\ \rho_4 = \rho_6 &= (1+\rho_2)/2 < 1, \end{aligned}$$

there exists an $\epsilon > 0$ such that for any class k, $k = 6, \ldots, 10$, $1/m_k \geq \alpha_1 + \epsilon$. Thus, for $k = 6, \ldots, 10$, whenever $Z_k(t) > 0$, $\dot{D}_k(t) = 1/m_k \geq \alpha_1 + \epsilon$. Invoking Theorem 2.4.9 in the time interval $[\delta, \infty)$, we have $Z(t) = 0$ for

$$t \geq \delta + |(I - \widetilde{P}')^{-1} Z(\delta)|/\epsilon,$$

where $\widetilde{P}$ is the routing matrix for the controlled network. Since $Z_k(\delta) \leq Z_k(0) + \delta\mu_{k-1}$, $k = 1, \ldots, 10$, and $(I - \widetilde{P}')^{-1}$ is a non-negative matrix, there exists a δ_1 such that for any fluid solution $\mathbb{X}$ to the controlled fluid network with $|Z(0)| \leq 1$, $Z(t) = 0$ for $t \geq \delta_1$. □

For the 2-station 5-class queueing network in Figure 2.1 operating under the SBP discipline π in (2.2), assume that $\alpha_1 = 1$, $m_1 = m_3 = m_4 = 0.1$ and $m_2 = m_5 = 0.6$. It follows from Theorem 2.1.1 that the queueing network is unstable. Yet, the corresponding leaky-bucket-controlled fluid network, and hence the leaky-bucket-controlled queueing network, is stable if $m_6 = m_8 = m_{10} = 0.9$ and $m_7 = m_9 = 0.85$. In fact, the leaky-bucket-controlled fluid network is globally stable or stable under any non-idling service discipline. Note that if we set the service times at the control stations to be 0, the controlled network is reduced to the original network that is unstable. Therefore, speeding up service rates may turn the leaky-bucket-controlled 7-station network that is globally stable into an unstable network.

Consider the following service variations in the controlled queueing network. Instead of actually serving jobs at the control stations, each server at its control station generates "tokens". If there is a token at class k ($k = 1, \ldots, 5$), the leading class k job leaves the class instantaneously for the next class on its route. The tokens are generated *autonomously*, regardless of whether there is a job waiting at the station or not. The class k tokens are generated at the rate of μ_k, say, deterministically, $k = 6, \ldots, 10$, and they are placed in a class k token buffer with buffer size s_k. Such a control scheme, with the use of control tokens, is called *leaky bucket control* in the telecommunications literature. The leaky bucket control scheme is used to smooth out the burstiness of arrival traffic in high speed telecommunications networks.

One can think of leaky bucket control stations as logical stations. It is not necessary to perform actual services at the control stations. Nor is it necessary to create new storage at the control stations. The role of the control stations is merely to add artificial delays when jobs are ready to move from one class to another class in the original network. Thus, it can be "inexpensive" to implement the leaky bucket control scheme in a queueing network.

Consider now a general multiclass queueing network with J service stations and K jobs classes. Recall that λ is the vector of nominal total arrival rates in the network. The corresponding leaky-bucket-controlled network is constructed as follows. Preceding each class k, add a leaky bucket control station $J + k$ that serves class $K + k$ jobs. The mean service time at this station is $\lambda_k^{-1}(1 + \rho_j)/2$. Jobs leaving class $K + k$ always go next to class k, and jobs leaving class k go next to class $K + \ell$ with probability $P_{k\ell}$ and exit the network with probability $1 - \sum_\ell P_{k\ell}$. Let $\widetilde{\lambda}$ be the vector of nominal total arrival rates in the controlled network. One can check that

$$\widetilde{\lambda}_{K+k} = \widetilde{\lambda}_k = \lambda_k.$$

Theorem 2.9.2. *Assume that the usual traffic condition (1.19) is satisfied for the original multiclass queueing network. Assume that the processing rate for jobs in class $K + k$ at station $J + k$ is*

$$\mu_{K+k} = \lambda_k^{-1}\left(\frac{1 + \rho_j}{2}\right),$$

where $j = s(k)$. Then the controlled fluid network is stable.

Proof. For $j = 1, \ldots, J$, the peak input load to station j is

$$\widehat{\rho}_j = \sum_{k \in \mathcal{C}(j)} \mu_{K+k}^{-1} m_k = \frac{2\rho_j}{1+\rho_j} < 1.$$

The rest of the proof follows that of Proposition 2.9.1. □

When the routing is deterministic, one can envision the leaky-bucket-controlled queueing network as the original queueing network with *input and output buffers*. For each class k, there are two buffers with unlimited buffer sizes associated with the class. One buffer is the input buffer for class k, holding jobs that are yet to be processed at class k, and the other buffer is the output buffer for class k, holding jobs that have been processed at class k, but have not left station $s(k)$. Jobs leave the output buffer k *one at a time* with travel time, say, deterministically, $m_{K+\ell} = \lambda_\ell^{-1}(1+\rho_j)/2$, where ℓ is the next class that class k jobs visit and $j = s(\ell)$. A job begins to travel only when the preceding job reaches its destination.

2.9.2 Generalized Round-Robin Discipline

If one has control over the service discipline used at each station in a network, there are many service disciplines that stabilize the network. One such simple discipline is a so-called *generalized round-robin* (GRR) discipline.

To explain the GRR discipline, recall that $\mathcal{C}(j)$ is the set of classes at station j. For concreteness, we assume that the set is ordered in an increasing order according to the class index. (Any other order works as well.) Let $\beta = (\beta_1, \ldots, \beta_K)$ be a vector of *positive integers*. Under the GRR discipline with weight vector β, for a class $k \in \mathcal{C}(j)$, once server j starts to serve class k jobs, server j serves the first β_k jobs in class k, then serves the first β_ℓ jobs in class ℓ and so on, where ℓ is the class that follows class k. We make the convention that when k is the last class in $\mathcal{C}(j)$, class ℓ is the first class in $\mathcal{C}(j)$. We assume that empty buffers are always skipped. When the number of jobs served in a busy cycle at class k is fewer than β_k, the server immediately leaves the empty class at the end of the busy cycle (i.e., it does not wait for new arrivals).

Theorem 2.9.3. *Let $\beta = (\beta_1, \ldots, \beta_K)$ be a vector of positive integers. Assume that for each class k,*

$$\frac{\beta_k m_k}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell m_\ell} > \lambda_k m_k. \tag{2.96}$$

Then the fluid model operating under the GRR discipline with weight vector β is stable.

Proof. Let $\bar{\mathbb{X}}$ be a fluid limit of the queueing network. One can show that for each regular $t > 0$,

$$\dot{\bar{T}}_k(t) \geq \frac{\beta_k m_k}{\sum_{\ell \in \mathcal{C}(j)} \beta_\ell m_\ell} \quad \text{when } \bar{Z}_k(t) > 0. \tag{2.97}$$

It follows from (2.96) and Theorem 2.4.9 that the fluid network is stable. □

2.10 Problems, Notes and Complements

Kumar and Seidman [37] presented the first example of a *deterministic* queueing network that is unstable under a non-idling discipline when the traffic intensity at each station is less than 1. Their service discipline is somewhat complicated. Shortly thereafter, Lu and Kumar [39] found that a SBP discipline is unstable in the so called Lu-Kumar network, which is a slight variation of the network considered by Kumar and Seidman [37]. Rybko and Stolyar [46] studied a *stochastic* version of the Kumar-Seidman network, and found that the SBP discipline used in Lu and Kumar [39] is unstable in the stochastic setting. These pioneering works on unstable networks are no doubt important. However, some of them, including the paper by Kumar and Seidman, did not initially receive the attention that they deserved. This is, perhaps, due to the fact that the service disciplines studied in these papers were perceived as somewhat "artificial", although they are nonetheless *non-idling*.

In 1994, Bramson [4] presented an exponential queueing network in which the FIFO discipline is unstable, even when the usual traffic conditions hold. Around the same time, Seidman [47] independently found a deterministic queueing network in which the FIFO discipline is unstable. Perhaps because FIFO is a discipline that is commonly used in practice, the unstable FIFO network examples finally inspired a wave of research that has led to the main body of this chapter. The unstable example in Theorem 2.1.1 is taken from Dai and Vande Vate [19].

Fluid models or fluid approximations of queueing networks have been studied for a long time, and were extensively studied by Chen and Mandelbaum [11]. The parallel between the stability of the Kumar-Seidman queueing network and its corresponding fluid model was first drawn by Rybko and Stolyar [46]. The rigorous connection between the positive Harris recurrence of a stochastic queueing network and its corresponding fluid model, in a form similar to Theorem 2.6.10, was established by Dai [14]. Stolyar [48] independently established a similar connection. The work of Dai was mainly inspired by a paper by Dupuis and Williams [25] that provides the connection between the positive recurrence of a high-dimensional reflecting Brownian motion and the stability of a corresponding deterministic Skorohod problem that is closely related to fluid models. Theorem 2.6.10 is from [14] with important refinements and generalizations from Chen [9] and Bramson [7]. The latter paper also first introduced the formal notion of an HL discipline. Dai and Meyn [18] explored issues beyond the positive Harris recurrence of a queueing network. They established sufficient conditions, in terms of fluid model stability and moment assumptions on the primitive increments, for sample performance measures to converge to the steady-state performance measures.

The connection between the stability of a queueing network and the stability of its corresponding *fluid model* is currently not complete, and will perhaps

remain so indefinitely. It is conceivable that a complete connection can be made between the stability of a queueing network and the stability of its corresponding *fluid limit model*. But a theorem of this generality would be of rather limited use because it is difficult to understand the behavior of every fluid limit. Nevertheless, studying fluid limits can be fruitful by leading to additional fluid equations like (2.31). Theorem 2.5.1, a partial converse to Theorem 2.6.10, is taken from Dai [16]. A weaker result was proved earlier in Meyn [42]. A recent generalization of Meyn's result is reported in Puhalskii and Rybko [45].

The virtual station phenomenon like the one in (2.6) was first observed by Harrison and Nguyen [33], and was later independently discovered by Dumas [24] who used the term "non-essential faces" to describe the phenomenon. Dai and Vande Vate [19, 20] gave a general definition of a virtual station and connected it to the global stability of 2-station fluid and queueing networks. Section 2.8.3 is taken from [20].

Pathwise stability of a queueing network is summarized in a recent book by El-Taha and Stidham [26]. Chen [9] made the connection between the pathwise stability and the weak stability of the corresponding fluid limit model. The calculus for fluid models was first developed in Dai [15] and Dai and Weiss [21]. The often used Theorem 2.4.9 is taken from Bramson [7] who attributes the current proof to Vincent Dumas.

Lyapunov functions play a key role in proving the stability of a fluid model. Entropy type Lyapunov functions were introduced by Bramson to prove the stability of FIFO Kelly type fluid networks [5] and HLPPS fluid networks [6]. Theorem 2.8.3 is due to Bramson [5]. Piecewise linear Lyapunov functions were first used by Botvich and Zamyatim [3] to study the (global) stability region of the Kumar-Seidman network. They were generalized independently by Dai and Weiss [21] and Down and Meyn [23]. Section 2.8.2 on the formulation of piecewise linear Lyapunov functions is taken from Dai and Weiss [21]. For two-station multi-type networks, the linear constraints from these Lyapunov functions were solved explicitly in Dai and Vande Vate [20]. Extensions and limitations of piecewise linear Lyapunov functions were investigated in Dai, Hasenbein and Vande Vate [17]. Linear Lyapunov functions have been promoted by Chen [10] for reflecting Brownian motions and by Chen and Zhang [12] for fluid networks. They have been shown to be sharp in characterizing the stability region for some networks operating under SBP disciplines. However, they cannot detect the stability region of a 3-station network operating under a SBP discipline [17]. The theory of using Lyapunov functions to show *instability* of a queueing/fluid network is less developed. Proposition 2.5.3 can only be used in some cases.

There are many service disciplines which can stabilize a queueing network. Harrison's discrete-review policies or BIGSTEP approach [29] were shown by Magalaras [41] to stabilize networks. Magalaras [40] further showed that these policies are asymptotically optimal under a fluid scaling. The leaky bucket control in Section 2.9.1 is taken from Bramson [7]. Humes [34] developed a similar control scheme earlier to stabilize deterministic queueing networks. The generalized round-robin discipline introduced in Section 2.9.2 is taken from Jen-

nings [35]. Jennings' work generalizes the work of Kumar and Seidman [37] from deterministic queueing networks to stochastic queueing networks. Actually, the networks considered in [37, 35] allow setup time to be incurred when a server switches from one class to another class. The complete proof of Theorem 2.9.3 can be found in [35].

Many important, active subjects related to fluid models have been left out of this chapter. One such subject is to find the "cheapest" way to empty or drain a fluid network; see Avram, Bertsimas and Ricard [1] and Weiss [49, 50]. Chen and Meyn [13] related the optimal draining of a fluid network to the optimal control of the corresponding multiclass queueing network under a long-run average cost structure. Furthermore, they suggested that knowing the optimal value function for the controlled fluid network, either exactly or approximately, can greatly speed up the value iteration procedure to find an optimal policy in a Markov decision process formation of the optimal control of the multiclass queueing network.

In a later chapter, we will introduce the notion of critical stability with state space collapse for a fluid model. Bramson [8] showed that critical stability of a fluid model implies a multiplicative state space collapse for the corresponding queueing network. Together with Williams [51], these two papers provide a powerful framework to prove heavy traffic limit theorems for multiclass queueing networks. Fluid model also plays a key role in proving asymptotic optimality of some service disciplines in some queueing networks under diffusion scaling; see Harrison [30] and Kumar [38].

Appendix A

Table of contents: Brownian Models of Stochastic Processing Networks

Bibliography

[1] Avram, F., Bertsimas, D., and Ricard, M. Fluid models of sequencing problems in open queueing networks : an optimal control approach. In Kelly, F. and Williams, R. J., editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 199–237, New York, 1995. Springer-Verlag.

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

[3] Botvich, D. D. and Zamyatin, A. A. Ergodicity of conservative communication networks. Rapport de recherche 1772, INRIA, 1992.

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

[5] Bramson, M. Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems: Theory and Applications* **22**, 5–45 (1996).

[6] Bramson, M. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems: Theory and Applications* **23**, 1–26 (1997).

[7] Bramson, M. Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems: Theory and Applications* **28**, 7–31 (1998).

[8] Bramson, M. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory and Applications* **30**, 89–148 (1998).

[9] Chen, H. Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines. *Annals of Applied Probability* **5**, 637–665 (1995).

[10] Chen, H. A sufficient condition for the positive recurrence of a semimartingale reflecting Brownian motion in an orthant. *Annals of Applied Probability* **6**, 758–765 (1996).

[11] Chen, H. and Mandelbaum, A. Discrete flow networks: Bottlenecks analysis and fluid approximations. *Mathematics of Operations Research* **16**, 408–446 (1991).

[12] Chen, H. and Zhang, H. Stability of multiclass queueing networks under priority service disciplines. *Operations Research* (1998). To appear.

[13] Chen, R. R. and Meyn, S. P. MDP. *Queueing Systems: Theory and Applications* (1998).

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

[15] Dai, J. G. Stability of open multiclass queueing networks via fluid models. In Kelly, F. and Williams, R. J., editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 71–90, New York, 1995. Springer-Verlag.

[16] Dai, J. G. A fluid-limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability* **6**, 751–757 (1996).

[17] Dai, J. G., Hasenbein, J., and Vande Vate, J. H. Stability of a three-station fluid network. (1998). Under revision for *QUESTA*.

[18] Dai, J. G. and Meyn, S. P. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control* **40**, 1889–1904 (1995).

[19] Dai, J. G. and VandeVate, J. Global stability of two-station queueing networks. In Paul Glasserman, K. S. and Yao, D., editors, *Proceedings of Workshop on Stochastic Networks: Stability and Rare Events*, volume 117 of *Lecture Notes in Statistics*, pages 1–26. Columbia University, New York, Springer-Verlag, 1996.

[20] Dai, J. G. and VandeVate, J. The stability of two-station multi-type fluid networks. *Operations Research* (1998). To appear.

[21] Dai, J. G. and Weiss, G. Stability and instability of fluid models for re-entrant lines. *Mathematics of Operations Research* **21**, 115–134 (1996).

[22] Davis, M. H. A. Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of Royal Statist. Soc. series B* **46**, 353–388 (1984).

[23] Down, D. and Meyn, S. P. Piecewise linear test functions for stability and instability of queueing networks. *Queueing Systems: Theory and Applications* **27**, 205–226 (1997).

[24] Dumas, V. Essential faces and stability conditions of multiclass networks with priorities. Rapport de recherche 3030, INRIA, 1996.

[25] Dupuis, P. and Williams, R. J. Lyapunov functions for semimartingale reflecting Brownian motions. *Annals of Probability* **22**, 680–702 (1994).

[26] El-Taha, M. and Stidham Jr., S. *Sample-Path Analysis of Queueing Systems.* Kluwer, 1999.

[27] Getoor, R. K. Transience and recurrence of Markov processes. In Azéma, J. and Yor, M., editors, *Séminaire de Probabilités XIV*, pages 397–409. Springer-Verlag, New York, 1979.

[28] Harrison, J. M. Brownian models of queueing networks with heterogeneous customer populations. *Proceedings of the IMA Workshop on Stochastic Differential Systems* (1988). Springer-Verlag.

[29] Harrison, J. M. The BIGSTEP approach to flow management in stochastic processing networks. In F. P. Kelly, S. Z. and Ziedins, I., editors, *Stochastic Networks: Theory and Applications*, volume 4 of *Lecture Note Series*, pages 57–90. Royal Statistical Society, Oxford University Press, 1996.

[30] Harrison, J. M. Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies. *Annals of Applied Probability* **8**, 822–848 (1998).

[31] Harrison, J. M. and Nguyen, V. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems: Theory and Applications* **6**, 1–32 (1990).

[32] Harrison, J. M. and Nguyen, V. Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems: Theory and Applications* **13**, 5–40 (1993).

[33] Harrison, J. M. and Nguyen, V. Some badly behaved closed queueing networks. In Kelly, F. P. and Williams, R. J., editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 117–124, New York, 1995. Springer-Verlag.

[34] Humes Jr., C. A regulator stabilization technique: Kumar-Seidman revisited. *IEEE Transactions on Automatic Control* **39**, 191–196 (1994).

[35] Jennings, O. B. *Generalized Round-Robin Service Disciplines in Stochastic Networks with Setup: Stability Analysis and Diffusion Approximation.* PhD thesis, School of ISyE, Georgia Institute of Technology, 1999.

[36] Kelly, F. P. *Reversibility and Stochastic Networks.* Wiley, 1979.

[37] Kumar, P. R. and Seidman, T. I. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control* **AC-35**, 289–298 (1990).

[38] Kumar, S. Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality. *Annals of Applied Probability* (1998). Submitted.

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

[40] Maglaras, C. Discrete-review policies for scheduling stochastic networks: fluid asymptotic optimality. *Annals of Applied Probability* (1998). Submitted.

[41] Maglaras, C. Dynamic scheduling in multiclass queueing networks: stability under discrete-review policies. *Queueing Systems: Theory and Applications* (1998). Submitted.

[42] Meyn, S. P. Transience of multiclass queueing networks via fluid limit models. *Annals of Applied Probability* **5**, 946–957 (1995).

[43] Meyn, S. P. and Tweedie, R. L. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes. *Adv. Appl. Probab.* **25**, 518–548 (1993).

[44] Meyn, S. P. and Tweedie, R. L. State-dependent criteria for convergence of Markov chains. *Annals of Applied Probability* **24**, 542–574 (1994).

[45] Pulhaskii and Rybko, A. Instability. *A Russian Journal* (1998).

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

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

[48] Stolyar, A. L. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes and Related Fields* , 491–512 (1995).

[49] Weiss, G. On optimal draining of re-entrant fluid lines. In Kelly, F. P. and Williams, R. J., editors, *Stochastic Networks*, volume 71 of *The IMA volumes in mathematics and its applications*, pages 91–103, New York, 1995. Springer-Verlag.

[50] Weiss, G. Optimal draining of fluid re-entrant lines: some solved examples. In F. P. Kelly, S. Z. and Ziedins, I., editors, *Stochastic Networks: Theory and Applications*, volume 4 of *Lecture Note Series*, pages 19–34. Royal Statistical Society, Oxford University Press, 1996.

[51] Williams, R. J. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems: Theory and Applications* **30**, 27–88 (1998).