Many-Server Queues with Time-Varying Arrivals, Customer Abandonment and Non-Exponential Distributions

Yunan Liu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

Abstract

Many-Server Queues with Time-Varying Arrivals, Customer Abandonment and Non-Exponential Distributions

Yunan Liu

This thesis develops deterministic heavy-traffic fluid approximations for many-server stochastic queueing models. The queueing models, with many homogenous servers working independently in parallel, are intended to model large-scale service systems such as call centers and health care systems. Such models also have been employed to study communication, computing and manufacturing systems. The heavy-traffic approximations yield relatively simple formulas for quantities describing system performance, such as the expected number of customers waiting in the queue.

The new performance approximations are valuable because, in the generality considered, these complex systems are not amenable to exact mathematical analysis. Since the approximate performance measures can be computed quite rapidly, they usefully complement more cumbersome computer simulation. Thus these heavy-traffic approximations can be used to improve capacity planning and operational control.

More specifically, the heavy-traffic approximations here are for large-scale service systems, having many servers and a high arrival rate. The main focus is on systems that have time-varying arrival rates and staffing functions. The system is considered under the assumption that there are alternating periods of overloading and underloading, which commonly occurs when service providers are unable to adjust the staffing frequently enough to economically meet demand at all times.

The models also allow the realistic features of customer abandonment and non-exponential probability distributions for the service times and the times customers are willing to wait before abandoning. These features make the overall stochastic model non-Markovian and

thus very difficult to analyze directly. This thesis provides effective algorithms to compute approximate performance descriptions for these complex systems. These algorithms are based on ordinary differential equations and fixed point equations associated with contraction operators. Simulation experiments are conducted to verify that the approximations are effective.

This thesis consists of four pieces of work, each presented in one chapter. The first chapter (Chapter 2) develops the basic fluid approximation for a non-Markovian many-server queue with time-varying arrival rate and staffing. The second chapter (Chapter 3) extends the fluid approximation to systems with complex network structure and Markovian routing to other queues of customers after completing service from each queue. The extension to open networks of queues has important applications. For one example, in hospitals, patients usually move among different units such as emergency rooms, operating rooms, and intensive care units. For another example, in manufacturing systems, individual products visit different work stations one or more times. The open network fluid model has multiple queues each of which has a time-varying arrival rate and staffing function.

The third chapter (Chapter 4) studies the large-time asymptotic dynamics of a single fluid queue. When the model parameters are constant, convergence to the steady state as time evolves is established. When the arrival rates are periodic functions, such as in service systems with daily or seasonal cycles, the existence of a periodic steady state and the convergence to that periodic steady state as time evolves are established. Conditions are provided under which this convergence is exponentially fast.

The fourth chapter (Chapter 5) uses a fluid approximation to gain insight into nearly periodic behavior seen in overloaded stationary many-server queues with customer abandonment and nearly deterministic service times. Deterministic service times are of applied interest because computer-generated service times, such as automated messages, may well be deterministic, and computer-generated service is becoming more prevalent. With de-

terministic service times, if all the servers remain busy for a long interval of time, then the times customers enter service assumes a periodic behavior throughout that interval. In overloaded large-scale systems, these intervals tend to persist for a long time, producing nearly periodic behavior.

To gain insight, a heavy-traffic limit theorem is established showing that the fluid model arises as the many-server heavy-traffic limit of a sequence of appropriately scaled queueing models, all having these deterministic service times. Simulation experiments confirm that the transient behavior of the limiting fluid model provides a useful description of the transient performance of the queueing system. However, unlike the asymptotic loss of memory results in the previous chapter for service times with densities, the stationary fluid model with deterministic service times does not approach steady state as time evolves independent of the initial conditions. Since the queueing model with deterministic service times approaches a proper steady state as time evolves, this model with deterministic service times provides an example where the limit interchange (limiting steady state as time evolves and heavy traffic as scale increases) is not valid.

Contents

List of Figures

List of Tables

Acknowledgements

I have been fortunate enough to receive support, inspiration, and encouragement from many wonderful people during the course of my doctoral studies.

First and foremost, I would like to thank my wife, Sherry, for her unconditional love and full support. At the moment that I got to know Sherry in high school, my life was already changed. Since then she has changed my life for uncountably many times. One of the biggest ones was studying operations research at the IEOR department of Columbia University.

Sherry used to be a Ph.D. student of the Biomedical Engineering Department, the neighbor department of IEOR. In 2004, after taking the IEOR Ph.D. level Stochastic Modeling course taught by Professor Whitt, she introduced both the course and the IEOR department to me. In 2006, I became a master student of IEOR; in 2007, I was accepted as a Ph.D. student. This thesis would not have been possible without Sherry's love!

Next, I would like to thank my advisor, Professor Ward Whitt. My application of the IEOR Ph.D. program encountered significant difficulty because of my poor undergraduate record. When I was desperate, it was Professor Whitt who was always there, supporting me and believing in me. Without his encouragement and support, this thesis would not have been possible.

I could not have hoped for a better advisor. Professor Whitt has always been there ready to patiently listen whenever I had a new idea. He is a pioneer in the field of this thesis and his comprehensive knowledge of the literature never ceases to amaze me. It is his passion for scientific research that always makes me excited and motivated about our work. The skills that he has instilled in me, both mathematical and engineering, will be invaluable to me throughout my career. In addition, as a former marathon runner, Professor Whitt also

shared with me great advice and experience on long distance running. Working with him has been a tremendous honor and I will miss him greatly as I move on.

I would like to thank (in alphabetical order) Jose Blanchet, Carri Chan, Mariana Olvera-Cravioto, and Karl Sigman for agreeing to serve on my committee.

I would like to thank all IEOR faculty and staff members who made the past five years the most memorable in my life. In particular, I would like to thank Professor Goldfarb. Before I became an IEOR student in 2006, Professor Goldfarb let me audit his course: Deterministic Modeling, which built me with solid foundation for my later courses and research. Since I wasn't a Columbia student yet at that time, Professor Goldfarb put Sherry's UNI into the coursework system so that I can view and download all course materials. How creative and considerate he was!

I would like to thank all my fellow Ph.D. students at the IEOR Department for their friendship. They have exposed me to so many fascinating things and have enhanced my life in many respects during the years. In particular, I want to thank Zongjian Liu, Xianhua Peng, Ning Cai, Ruxian Wang, and Shiqian Ma, who helped and supported me greatly during my early Ph.D. years.

Long distance running and rock climbing became my major mind-balancing hobbies during my late Ph.D. years. They taught me the meaning of persistence, bravery, and confidence. I want to thank my wife Sherry for opening these doors to me, my climbing partner Jing Dong and my running partner Changyao Chen for their friendship, technical support, and mental encouragement.

Last but not least, I want to thank my parents Xin Wang and Sheng Liu for their constant love and support.

To my beloved wife

Sherry

who makes my life meaningful!

Chapter 1

Introduction

This research is motivated by the need for tools to improve the performance of large-scale service systems, such as telephone call centers, healthcare systems, judicial and penal systems, and both front-office and back-office operations in business systems; e.g., see [1, 79] and references therein for discussion of possible applications to customer contact centers and healthcare. Large-scale service systems tend to be quite complicated because they tend to have the following five features: (i) time varying arrival rates and staffing, (ii) abandonment from queue of impatient waiting customers, (iii) non-Markovian probability structures (stemming from non-exponential probability distributions), (iv) large scale (many servers and high arrival rates), and (v) complex network structure (multiple queues with flows from one to the other). This thesis proposes new mathematical models and tools to help analyze (and thus manage) the congestion in large-scale service systems. The models are deterministic fluid models. These fluid models serve as approximations for corresponding stochastic queueing models with all the complicating features above.

1.1 Time-Varying Model Data

It is important that our model assumptions capture realistic features of real service systems. One of them is the time variability of the model data, i.e., the arrival rate, the service and abandonment distributions, the number of servers and the routing probabilities. Among all these model elements, the most important is the arrival rate. The time-varying arrival rate in turn causes the staffing (the number of servers) to be time varying as well. We elaborate on these two forms of time variability below. However, other model parameters may be time varying as well. For instance, surgeons intend to schedule longer operations in the morning and shorter ones in the afternoon, which can result in an increasing service rate over the course of a day.

1.1.1 Time-Varying Arrival Rates

Unlike most textbook queueing models, real service systems typically have time-varying arrival rates, usually with significant variation over the day. For instance, the arrival rate of calls in a financial service call center might vary from 0 (during the late night) to 2000 over the course of a day, as shown in Figure 1.1, taken from [25]. Because of such time-varying arrivals, it is difficult to analyze the system performance. It is no longer possible to apply the steady-state analysis associated with queueing models having constant arrival rates, commonly found in textbooks.

Consequently, the standard tool for analyzing queues with time-varying arrival rates is computer simulation. However, in order to rapidly determine the performance consequences of different staffing plans, it is very helpful to have analytical models and methods for analyzing them. Almost all successful analytical methods employ approximations; see [26]. This thesis continues the effort to develop useful analytical approximations for analyzing queueing models with time-varying arrival rates.

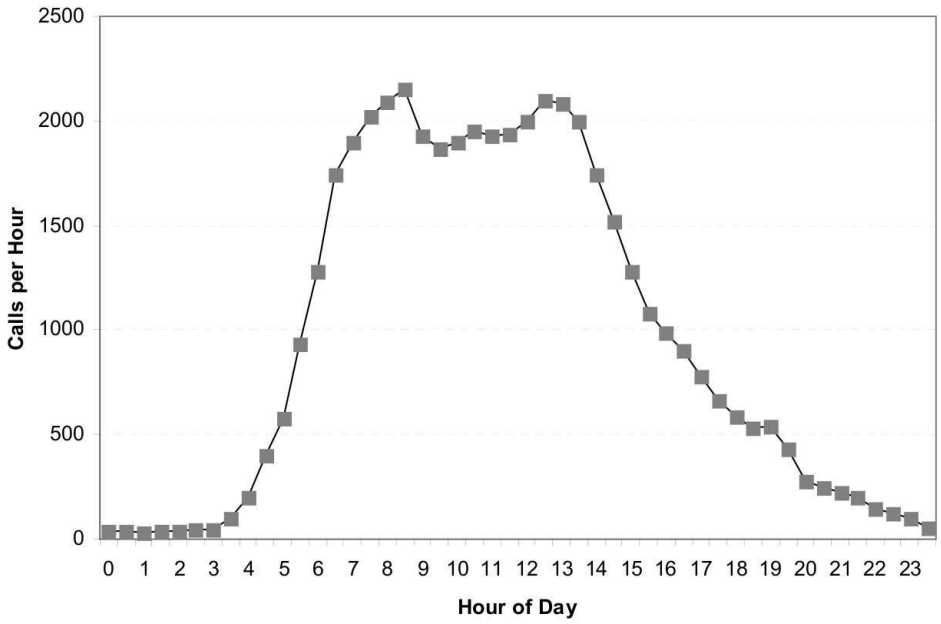

Figure 1.1: The arrival rate of incoming calls of a medium-size financial services call center.

When staffing is adequate and service times are short, as in many customer contact centers, it is often possible to apply stationary models to analyze many-server queueing models with time-varying arrival rates, using some variant of the pointwise-stationary approximation. The pointwise stationary approximation uses a different stationary model at each time, acting as if the arrival rate were constant with the instantaneous arrival rate at that time.

When staffing is occasionally inadequate or service times are longer, the pointwise stationary approximation can perform badly. Then other methods may be needed; see [26] for a review. To determine appropriate staffing levels and analyze performance in a many-server system with time-varying arrivals, infinite-server models often can be employed, as in [17, 50, 53] and references therein. However, the effectiveness of infinite-server models depends largely on the assumption that ultimately the system will be adequately staffed.

This thesis considers a different situation. This thesis focuses on systems that alternate between periods of overloading and underloading.

1.1.2 Time-Varying Staffing

In order to cope with the time variability of the arrival pattern, appropriate time-varying staffing functions are needed; see [26] for background. Therefore, it becomes necessary to go beyond the scope of models with constant staffing.

It is important to note that complications arise when we consider queueing systems with time-varying staffing. We need to carefully consider what happens when the service capacity is scheduled to decrease when all servers are busy. Do we require that customers in service stay in service with the same server until their service is complete? (The analysis here applies to the case in which we allow the service in progress to be handed off to another available server.) Even with such server-assignment switching, there are issues: Do we alter the prescribed staffing function to avoid forcing a customer out of service? If we adhere to the given staffing function, as assumed here, then some customers are necessarily forced out of service in the stochastic system. (That can be prevented in the idealistic deterministic fluid model; see Assumption 2.4.) In the stochastic system, when customers are forced out of service, which customers are forced out and what happens to them? Are these customers forced out of the system entirely? If so, is there service complete or do they retry? If customers are pushed back into the queue (as implicitly assumed in [46]), then where do they go in the queue, and what is their new abandonment behavior?

Under regularity conditions, these realistic features will be asymptotically negligible as the system scale grows (in a many-server heavy-traffic limit, discussed in §1.3.2), but these new considerations complicate the proofs of limit theorems. For the fluid model, we

directly assume feasibility of the staffing function, but we also show how to achieve it if it is not initially present; see §2.9.

1.1.3 Alternating Periods of Overloading and Underloading

As indicated above, this thesis focuses on systems that alternate between periods of overloading and underloading. In particular, this thesis develops heavy-traffic fluid approximations to approximate the performance of associated complex stochastic queueing models that experience alternating periods of overloading and underloading.

Of course, periods of overloading are not desired, because they cause large customer delays, producing significant customer dissatisfaction. But, also, periods of underloading (with many idle servers) are not desired, because they are inefficient, tending to produce large staffing costs. Nevertheless, many service systems commonly experience periods of overloading and underloading. That is so because, first, the arrival rate varies significantly over time and, second, system managers are unwilling or unable to change the number of servers dynamically in real time to efficiently meet demand at all times. For example, there may be constraints on the shifts. Consequently, service systems such as hospitals and call centers often alternate between periods of overloading and underloading. Therefore, there is an increasing need for better understanding of the performance of service systems that experience alternating intervals of overloading and underloading.

By considering alternating overloaded (OL) and underloaded (UL) intervals, we consider a new many-server heavy-traffic (MSHT) regime (discussed more in §1.3.1). The vast majority of the many papers on MSHT approximations focus on systems that are nearly critically loaded at all times. In other words, they focus on the so-called quality-and-efficiency driven (QED) regime. In contrast, this thesis does not consider the QED regime at all. The OL and UL intervals considered here correspond to the efficiency-driven (ED)

and quality-driven (QD) many-server heavy-traffic regimes, described in §1.3.1, as opposed to the more commonly studied quality-and-efficiency (QED) regime, also described in §1.3.1. Thus, this thesis focuses on MSHT approximations for systems that alternate between ED and QD MSHT regimes.

The structure of alternating OL asnd UL intervals is strongly exploited in this thesis. When the system is underloaded, i.e., when there are enough servers serving all customers, the system is identical to an infinite-server model; when the system is overloaded, i.e., there are customers waiting in the queue and all servers are busy, we decompose the system into two subsystems, the queue and the service facility, and separately treat the customers that are waiting in queue and those that are in service; see Chapter 2 for details.

1.2 Abandonment and non-Exponential Distributions

In addition to time-varying arrival rates and staffing, service systems often experience customer abandonment and have non-exponential distributions, which makes the major stochastic processes of interest, such as the number of customers waiting in queue, more difficult to analyze.

1.2.1 Customer Abandonment

In service systems, customers will often leave if they cannot begin service within a reasonable time after they arrive. For example, in call centers, customers abandon by hanging up if they are put on hold for a long time. In hospitals emergency rooms, patients often leave the waiting room before being seen by a doctor (i.e., abandon) because they have had to wait a long time; that is known as the “left without being seen” (LWBS) effect; see [79] for discussion. Moreover, the feature of customer abandonment is important to include in the

model, because even a small amount of customer abandonment can significantly alter the system performance; [20]. Thus Customer abandonment is now recognized as an important feature in service systems, e.g., see [20, 81].

The probability (or percentage) of customer abandonment is one of the most important performance criteria in service systems such as call centers, it provides direct feedback to the system managers on whether or not the offered service is worth its wait and to what extend customers are satisfied with the service. There are other commonly used measures such as the average waiting times and the probability (or percentage) of customer delay. However the different performance measures are all deeply connected, for instance, a nearly linear relationship between the average waiting time and the probability of abandonment was established in [49].

1.2.2 The Classical Erlang Models

Traditionally, the performance of service systems, such as telecommunication systems, has been analyzed by applying the classical Erlang models. The reference model is the Erlang C (or delay) model, denoted by $M/M/s$. In this model there is an external Poisson arrival process (the first M), independent and identically distributed (IID) exponential service times (the second M), s servers and an unlimited waiting room. The service times are assumed to be independent of the arrival process. When all servers are busy, new arrivals join a queue and wait for a free server. Customers are served in order of arrival by the first available server. The Erlang B (or loss) model is the variant that has no waiting room at all; then when all servers are busy, new arrivals are blocked and lost. The Erlang B model was especially appropriate for telephone equipment that had not provision for waiting.

Of special relevance for this thesis is the generalization of the Erlang C and B models to the Erlang A model, denoted by $M/M/s + M$. Just as in the Erlang C model, there

is an unlimited waiting room, so that when all servers are busy, new arrivals again join a queue and wait. However, the Erlang A model accounts for customers having only limited patience for waiting before entering service. The model assumes that each customer has a length of time (patience time) that the customer is willing to wait before beginning service. If the customer is unable to enter service before that time, then the customer leaves without receiving service. These patience times are assumed to be IID exponential random variables (the $+M$) with rate θ, independent of the arrival process and service times. The Erlang A model reduces to the Erlang C model when $\theta = 0$; the Erlang A model reduces to the Erlang B model when $\theta = \infty$. The performance in the Erlang A model approaches the performance in the Erlang C model as θ approaches 0; the performance in the Erlang A model approaches the performance in the Erlang B model as θ approaches ∞.

1.2.3 From Markov to Non-Markov Queueing Models

The Erlang models are relatively easy to analyze because the number of customers in the system at time t is a birth-and-death stochastic process, a relatively simple continuous-time Markov chain stochastic process. However, to obtain more realistic models it is important to go beyond these Markov models. In particular, statistical analysis shows that customers' service and patience times are typically not exponentially distributed in real service systems. For example, Brown et al. [7] found that the distribution of the duration of calls (service times) in call centers is close to the lognormal distribution, while the the hazard rate (the density divided by the complementary cdf) is far from constant (implying that the distribution of customer patience times is far from the exponential distribution), as can be seen from Figure 1.2 from [7].

It is thus important to determine to what extent the queueing models with exponential distributions provide useful performance description for systems where the exponential as-

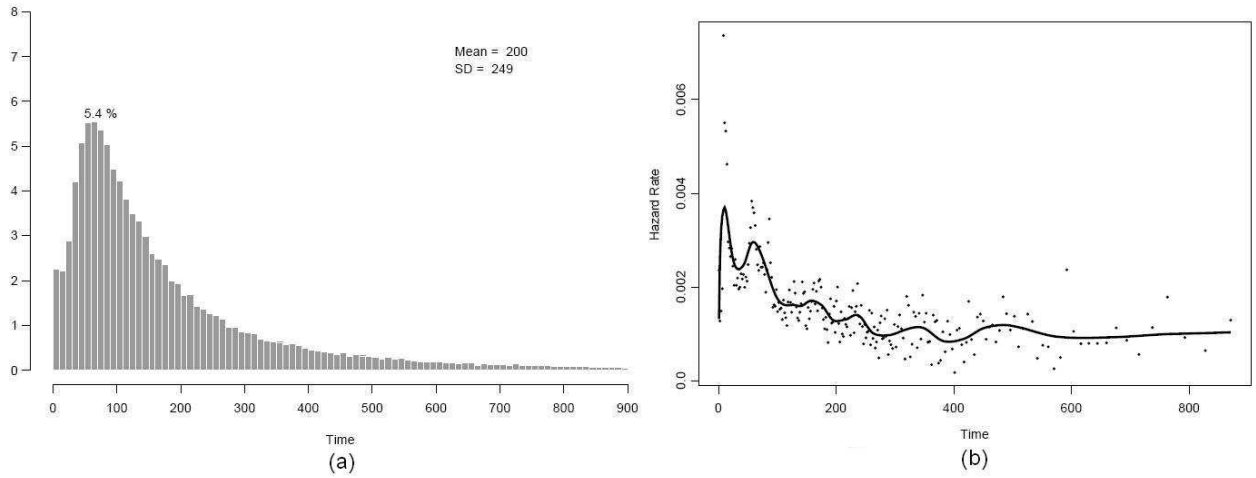

Figure 1.2: (a) A histogram of service times and (b) an estimate of the hazard rate rate of patience times in a medium-size call center.

sumptions are not nearly satisfied. Whitt [77] showed for the many-server $M/GI/s + GI$ model, that the steady-state system performance tends to be quite sensitive to the abandonment distribution beyond its mean, but relatively insensitive to the service-time distribution beyond its mean. However, in Chapter 2, we show that the service-time distribution beyond the mean can have a great impact to the transient performance.

Thus, there is growing interest in developing effective methods for analyzing models that allow the service-time and patience-time distributions to be IID random variables with general distributions (GI). Thus, there is a need to consider the $M/GI/n + GI$ model instead of the $M/M/n + M$ model. Unfortunately, however, the number of customers in the system at time t is no longer a Markov process. Analytic formulas are available, although complicated, for the steady-state performance of the $M/M/n+GI$ model, having a general abandonment distribution but still exponential service; see [49, 80, 81]. However, little in the way of explicit analytical results has been done more generally. Hence, even for the $M/GI/n + GI$ model, it is necessary to resort to approximations.

However, this thesis considers even more general models than the challenging $M/GI/n+GI$ model. In addition to non-exponential service and patience distributions, the queueing

models here allow non-Poisson arrivals and time-varying arrival rate and staffing. The base stochastic model in this thesis is the $G_t/GI/s_t + GI$ stochastic model, where the subscript t denotes time-varying. The G_t arrival process is a general (not necessarily Poisson) stochastic process with a time-varying arrival rate. However, the non-homogenous Poisson process, denoted by M_t, is the primary arrival process of interest.

Since the $M/GI/s+GI$ stochastic queueing model is not tractable by current methods, it is evident that the more general $G_t/GI/s_t + GI$ stochastic model is not tractable either. Thus we are motivated to look for approximations. Specifically, this thesis proposes and analyzes a deterministic fluid approximation for the $G_t/GI/s_t + GI$ stochastic model, which is called the $G_t/GI/s_t + GI$ fluid model. In this fluid model, the general time-varying G_t arrival process is characterized simply by the arrival rate function. However, the general service-time and patience-time cumulative distribution functions (cdf's) G and F, respectively, play important roles in the fluid model (beyond their mean values).

We obtain Markovian structure in the more complicated $G_t/GI/s_t + GI$ stochastic model and its fluid model counterpart by focusing on two-parameter stochastic processes. In particular, we consider the queue content (number of customers waiting in the queue) at time t that has been in queue for a *duration* at most y, denoted by $Q(t, y)$, and the service content (number of customers that are in service) at time t that has been in service for a *duration* at most y, denoted by $B(t, y)$, see (2.3). Here Q and B are functions of both t and y.

1.3 Many-Server Heavy-Traffic Fluid Approximations

Traditionally, the performance of service systems, such as telecommunication systems, have been analyzed by applying the classical Erlang models, which we reviewed above in §1.2. However, since real service systems typically do not have nearly Markovian proba-

bility structure, the generalization to non-Markovian models is important. However, once the Markovian assumption is relaxed, even for a little bit, exact analysis tends to become intractable. Therefore, heavy-traffic fluid and diffusion approximations become helpful; see [74] for a review.

Heavy-traffic involve a sequence of queueing systems in which the load is allowed to increase (become heavy). The congestion (e.g., the queue length) tends to grow in the heavy-traffic limit, but after appropriate scaling (e.g., multiplying by an appropriate asymptotically negligible quantity), there may be a nondegenerate limit, which can serve as an approximation for the pre-limit processes.

The fluid models studied here can be regarded as models of interest in their own right. However, their justification is enhanced by heavy-traffic limit theorems, which show that the fluid models arise as heavy-traffic limits for a sequence of queueing models. Thus we will approximate the expected total number of customers waiting in queue, $E[Q(t)]$, by the deterministic number in the corresponding fluid model. A heavy-traffic fluid limit provides theoretical support for the approximation by showing that the approximation is asymptotically correct as the scale increases. In a refined diffusion approximation, the diffusion term can be used to estimate the stochastic error or fluctuation around that mean trajectory.

There are two types of heavy-traffic regimes: the *conventional* heavy-traffic regime that focuses on queues with a single server or a fixed number of servers, and the *many-server* heavy-traffic regime that applies to queues with a large number of servers (where the number diverges to $+\infty$ in the limit). We review these two heavy-traffic regimes in §§1.3.1 and 1.3.2 below. Afterward, we discuss fluid models in §1.3.3.

1.3.1 The Conventional Heavy-Traffic Regime

The conventional heavy-traffic regime involves a sequence of queueing models with a fixed finite number of servers in which the associated sequence of traffic intensities is allowed to increase to the critical value for stability, 1; see [74] for an extensive account. The first conventional heavy-traffic limit (and approximation) was developed for the $GI/GI/1$ queue by Kingman [38]. The $GI/GI/1$ queue has a single server, IID interarrival times $\{A_i, i \geq 1\}$ with mean $1/\lambda$ and squared coefficient of variation (SCV, i.e., $Var(A)/E[A]^2$) c_A^2, and IID service times $\{S_i, i \geq 1\}$ with mean $1/\mu$ and SCV c_S^2. (Thus, finite variances is assumed.) The associated sequence of $GI/GI/1$ queues indexed by n is constructed by first letting $\lambda = \mu$ and then by making the nth queue have the same arrival process but modified service times $\{S_i^{(n)} \equiv \rho_n S_i, i \geq 1\}$, where $\rho_n \uparrow 1$ as $n \to \infty$ for $\rho_n \equiv E[S^{(n)}]/E[A] = \lambda/\mu^n$, with $\equiv$ denoting "equality by definition." The quantity ρ_n is the traffic intensity in model n.

Under those assumptions, Kingman [38] showed that

$$(1 - \rho_n)W^{(n)} \Rightarrow W \qquad \text{as} \quad n \to \infty, \tag{1.1}$$

where $W^{(n)}$ is the steady-state waiting time of the nth queue in that sequence, W is an exponential random variable with mean $(c_A^2 + c_S^2)/2\mu$ and $\Rightarrow$ denotes convergence in distribution. The limit in (1.1) can then be applied to generate the approximation

$$W^{(n)} \approx \frac{W}{(1 - \rho_n)} \qquad \text{for fixed} \quad n, \tag{1.2}$$

which tends to be increasingly accurate (in a relative sense) as n increases.

Borovkov [5] and Iglehart and Whitt [32] later extended the conventional heavy-traffic

limit for single-server queues to queues with multiple servers. Instead of establishing the limiting result for the steady-state queue length, Iglehart and Whitt [32] established the convergence of the entire queue-length process. Similar to Kingman [38], they considered a sequence of $GI/GI/s$ queues (as well as more general multichannel queues) indexed by n such that the nth queue has IID interarrival times $\{A_i/s, i \geq\}$, where A_i has mean $1/\mu$ and SCV c_A^2, and IID service times $\{\rho_n S_i, i \geq 1\}$, where S_i has mean $1/\mu$ and SCV c_S^2. (Again ρ_n is the traffic intensity in model n.) They let the traffic intensity approach 1 in the way that $\sqrt{n}(1-\rho_n) \to \beta$ as $n \to \infty$, where $0 < \beta < \infty$. They showed that

$$\frac{1}{\sqrt{n}} Q(nt) \Rightarrow \tilde{Q}(t) \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \quad \text{if} \quad \frac{1}{\sqrt{n}} Q(0) \Rightarrow \tilde{Q(0)},$$

where $\tilde{Q}$ is a reflected Brownian motion with a drift term $-\beta s\mu$ and a diffusion term $s\mu(c_A^2 + c_S^2)$, $\mathbb{D}$ is the space of real-valued functions that are right-continuous and have left limits.

1.3.2 Many-Server Heavy-Traffic Regimes

Unfortunately, however, the conventional heavy-traffic limits do not yield good approximation for large-scale service systems with many agents (servers). To develop better approximations for such systems, Halfin and Whitt [28] established a many-server heavy-traffic (MSHT) for the $GI/M/s$ model, including the Erlang C $M/M/s$ model. For the Erlang C model, there is a sequence of queues indexed by n such that the nth queue has a Poisson arrival process with rate λ_n, IID exponential service times with rate μ, and n servers. Halfin and Whitt [28] proposed the *quality-and-efficiency driven* (QED) regime (also known as

the *Halfin-Whitt regime*), which is characterized by

$$\sqrt{n}(1-\rho_n) \to \beta, \quad \text{for} \quad 0 < \beta < \infty, \tag{1.3}$$

where $\rho_n \equiv \lambda_n/n\mu$ is the traffic intensity of the nth queue. Since the number of servers in the nth queue grows to infinity, this regime is a many-server heavy-traffic (MSHT) regime.

Halfin and Whitt [28] showed that, under condition (1.3),

$$\sqrt{n}\left(\frac{Q_n}{n}-1\right) \Rightarrow \tilde{Q} \quad \text{in} \quad \mathbb{D}, \quad \text{if} \quad \sqrt{n}\left(\frac{Q_n(0)}{n}-1\right) \Rightarrow \tilde{Q}(0), \quad \text{as} \quad n \to \infty,$$

where $\tilde{Q}$ is a diffusion process with a drift term $m(x) = -\mu\beta\mathbf{1}_{\{y\geq 0\}} - \mu(x+\beta)\mathbf{1}_{\{y<0\}}$ and a diffusion term 2μ. In the QED MSHT regime, the steady-state probability of delay approaches a constant strictly between 0 and 1.

The results in Halfin and Whitt [28] were generalized to the model with phase-type service distributions by Puhalskii and Reiman [60] and to general GI service-time distributions by Reed [61] and Puhalskii and Reed [59]. Kaspi and Ramanan [36] proved a fluid limit for the measure-valued process tracking the ages of customers in the system.

The MSHT regime has also been generalized by incorporating customer abandonment. For the special case of the Erlang models, the original results of Halfin and Whitt [28] were extended from the Erlang C model to the Erlang A model by Garnett et al. [20]. With abandonment, the traffic intensity need not be less than 1 in order for a proper steady state to exist. Indeed, for the Erlang C model, a proper steady-state distribution exists for all traffic intensities.

Garnett et al. [20] introduced the names quality-driven (QD), quality-and-efficiency-driven (QED) and efficiency-driven (ED) for the three MSHT regimes. The QED regime

prevails when

$$\sqrt{n}(1-\rho_n) \to \beta, \quad \text{for} \quad -\infty < \beta < \infty, \tag{1.4}$$

In contrast, the QD regime arises when the limit in (1.4) is $+\infty$, whereas the ED regime arises when the limit in (1.4) is $-\infty$. Thus, if the traffic intensity remains fixed as $n \to \infty$, then the system is in the QD, QED or ED MSHT regime if and only if $\rho < 1$, $\rho = 1$ and $\rho > 1$, respectively. When $\rho > 1$, the probability of delay converges to 1, but the probability of abandonment converges to a constant strictly between 0 and 1. When $\rho < 1$, both probabilities converge to 0.

Most research has focused on the QED MSHT regime. However, as emphasized in [75], with abandonment, the ED regime is also of considerable practical importance. Indeed, the ED regime corresponds to the OL case considered here. The fluid model is of special interest only when the ED regime prevails at least part of the time. The MSHT fluid approximation for the general $G/GI/s+GI$ model with non-exponential service and abandonment distributions was established by Whitt [77]. In addition, a discrete version of the limiting convergence theorem was also provided in [77]. Further limits for this model have been obtained by Kang and Ramanan [37].

The paper by Whitt [77] was the original inspiration for this entire thesis. The initial goal of this research was to obtain corresponding results for the $G_t/GI/s_t+GI$ fluid limiting model with smooth model parameters, thus extending the discrete time results in §6 of [77] (which allowed time-varying arrival rates). In Chapter 2 we develop a complete analysis for the $G_t/GI/s_t+GI$ fluid model, which provides important new results for the $G/GI/s+GI$ model, thus complementing [77].

1.3.3 Deterministic Fluid Models

There is a long history of applying deterministic fluid models to approximate the performance of queueing systems, as can be seen from Newell [54]; also see Hall [29]. They tend to be especially useful when the congestion is primarily determined by differences in the total arrival rate and the maximum possible service rate, as occurs when the system experiences periods of substantial overloading. The fluid models can be applied directly, but additional insight can be obtained if they can be shown to arise as the limit of a sequence of queueing models. However, this thesis is not primarily concerned with establishing limits for sequences of scaled queueing processes associated with a sequence of queueing models. Instead, we are directly concerned with the fluid model itself. It is important to recognize that the fluid model can be considered directly as a legitimate model in its own right. By focusing on a continuous divisible quantity, which we call "fluid," our fluid model can be regarded as a storage or dam model, as in [57].

Even though the fluid model we consider can be directly regarded as a model of interest, it is helpful to see how the fluid model considered here arises in a limit of a sequence of queueing systems. In this thesis we focus on a fluid model that arises in the MSHT regime, as in [20,46,55,56]. As a consequence, the MSHT fluid approximations is more appropriate for large-scale systems, where the number of servers is 100 or more, but also may be useful for systems with fewer servers, such as $5 - 20$.

A theoretical basis for the MSHT limits for models with time varying arrival rates and staffing was established by Mandelbaum, Massey and Reiman [46]; see also [47, 48]. They established MSHT limits for the time-varying Markovian $M_t/M_t/s_t + M_t$ queueing model. Whitt [77] established a discrete-time generalization for the more general $G_t/GI/s_t + GI$ model considered here. Thus, even though we do not prove limit theorems here, the appropriate scaling is evident from these previous papers.

Our basic queueing model is the $G_t/GI/s_t + GI$ queue, which has a time-varying arrival rate (the G_t), a non-exponential service distribution (the GI), a time-varying staffing function (the s_t), and a non-exponential patience distribution (the $+GI$). We consider a sequence of systems indexed by n in which both the arrival rate and the number of servers increase linearly in n. Let $Q_n(t,y)$ ($B_n(t,y)$) be the number of customers in queue (in service) at t that has been in queue (in service) for at most y. We expect to see the convergence

$$\left(\frac{Q_n(t,y)}{n}, \frac{B_n(t,y)}{n}\right) \Rightarrow (Q(t,y), B(t,y)), \quad \text{as} \quad n \to \infty, \tag{1.5}$$

where Q and B are deterministic fluid functions. As $n \to \infty$, customers are shrunk down to atom of fluid, however their individual behavior remains unchanged. Paralleing (1.2), as a consequence of the limit (1.5), we propose the approximation

$$(Q_n(t,y), B_n(t,y)) \approx n(Q(t,y), B(t,y)), \quad \text{for fixed} \quad n, \tag{1.6}$$

where the accuracy of the approximation (again in a relative sense) improves as n increases.

For very large-scale service systems (with many servers at each queue and high arrival rates, i.e., n is large) such as large-scale call centers, the deterministic fluid values serve as good direct approximations for the stochastic queueing quantities, because the stochastic fluctuations around the mean values tend to be relatively small (essentially because of the law of large numbers (LLN)). When n is small, e.g., $n = 10$ such as in a hospital, this single sample path approximations become crude since stochastic fluctuations cannot be simply ignored. However, the fluid content can still be used to approximate the mean value of the corresponding stochastic process in the many-server queueing system. See computer simulation verifications in §2.2.

1.4 Network Structure

In this thesis, we primarily focus on a single fluid queue. However, queueing models with only one queue are not sufficient to represent all real service systems. In hospitals, for instance, patients move between different units, such as the emergency rooms, intensive care units, and operating rooms. In factories, each item may have to visit different work stations through its production line. In call centers, departing customers may want to call back later to require more service; this feedback provides network structure.

Thus we also consider fluid models with a general network structure. In particular, we consider open networks of fluid queues with proportional routing, which we denote as the $(G_t/GI/s_t + GI)^m/M_t$ fluid model. This model has m fluid queues, each with its own arrival process, service times and patience times. A proportion of the fluid completing service from each queue is routed to other queues or out of the network. Just as for the single fluid queue, this network of fluid queues is intended to serve as an approximation for the corresponding stochastic model. Each queue in the stochastic queuing model is a $G_t/GI/s_t + GI$ queueing model. In the stochastic queueing model the routing is assumed to be Markovian, with individual customers going to one of the other queues with specified probabilities, independent of the system history up to that time. See Chapter 3 for details.

This stochastic queueing network is a generalization of the open Jackson queueing network; see [10] for a review of the Jackson network. However, the non-Markovian structure and the time-varying arrival rates and staffing make the stochastic model extremely difficult to analyze. To find a balancing point between model applicability and mathematical tractability, we focus on the MSHT deterministic fluid approximation of this $(G_t/GI/s_t + GI)^m/M_t$ stochastic queueing network model. We provide efficient algorithms to compute the standard performance measures in a finite time interval, such as a day or a week.

1.5 Transient and Asymptotic Performance

In standard queueing models with constant parameters attention is usually focused on the long-run steady-state behavior. However, when the model parameters are time-varying, as in most real service systems, that is not possible. A steady state no longer exists. For systems with time-varying model elements, it is necessary to pay attention to the performance as a function of time. Therefore, it is important to carefully investigate the system performance in a relatively short time period, such as a day or a week. We intend to answer questions such as: How many total customers on average do we expect at 9am? How long does an arriving customer at 1pm have to wait before entering service? To prevent extensive overloading, how many agents do we need at 3pm?

To analyze the transient dynamics, fluid and diffusion approximations have been widely developed, see [46–48, 56, 58]. Consider a performance function, such as the total number of customers waiting in queue at t, $Q(t)$. the time-dependent fluid function characterizes its average (via sample path, not time) behavior as time evolves while the diffusion term describes the stochastic fluctuations around that average path. Complementing [46–48, 56, 58], we provide efficient algorithms to compute performance functions for the fluid approximations in any finite time interval.

In this thesis we also consider the steady-state behavior of the models when the model parameters are not time-varying. When the model data are not time-varying, the long-run average or steady-state performance is of primary interest. For example, in revenue-generating service systems such as call centers which take customer orders, these steady-state quantities can be very helpful in evaluating the average system costs and revenue, e.g., see [3]. Thus it is significant that we also characterize this steady-state performance for models without time-varying model elements.

From a theoretical perspective, it is also important to know that the system approaches

steady state as time evolves. In this thesis, we show that the stationary fluid model converges to steady state as time evolves. In addition, how fast the transient system performance functions converge to the steady state may also be of applied interest. When that convergence is rapid, we can more safely ignore any effect from the initial state and directly use the steady state quantities as approximations. We show, under regularity conditions, that the convergence to steady state in the fluid queues is exponentially fast.

Finally, there is another important case. Even though the arrival rate may be time-varying, the arrival rate may be periodic or nearly periodic. Indeed, this is common for service systems that experience daily or weekly cycles. For instance, call centers reveals similar arrival patterns on every Monday, which can be quite different from those on Sunday, so one week can be treated as a performance cycle, see [7].

In this periodic case, transient analysis is of course still important, but it is natural to expect that there would be a dynamic periodic steady state. In particular, we expect the successive cycles to be distributed the same in the long run. That is, there would be systematic time variation within each cycle, but the distribution of the performance over the successive cycles would tend be the same. In this thesis we establish the existence of a periodic steady state (PSS) for queues with periodic model parameters, and show convergence to that PSS as time evolves. See Chapters 4-5 for details.

1.6 Effective Algorithms

For engineering applications, it is essential that the performance descriptions can actually be efficiently computed. Thus it is significant that we develop efficient algorithms for the fluid models considered here. These algorithms are based on solving ordinary differential equations and solving fixed point equations associated with contraction operators. They are implemented in MatLab and solved on an ordinary personal computer.

Using these algorithms, we can predict, and thus control, the real-time performance of service systems. For example, we apply the algorithms to determine staffing functions that stabilize the performance at target levels of congestion, for an arbitrarily given arrival rate function; see §2.10. We now briefly describe the different algorithms developed in this thesis.

1.6.1 Algorithm for One Fluid Queue

We first decompose the system into two subsystems: (i) the *queue* where fluid is waiting in line and the *service facility* where fluid is in service. When the system is UL, the queue is empty so that external arrivals flow into the service facility directly; when the system is OL, external arrivals are buffered in the queue, the oldest fluid in queue moves into the service facility according to a first-come-first-serve (FCFS) discipline.

We next partition the desired time interval (such as a day) into disjoint OL and UL intervals and provide the OL-UL switching criterion. Given the initial system status, we recursively compute the performance measures in OL and UL intervals and locate those OL-UL switching time points, until the end of the time horizon is reached.

Comparing with a UL interval, an OL interval is more complicated. For general non-exponential service distributions, we have to solve a fixed-point equation (FPE) which admits a unique solution under general conditions; for exponential service distributions, the FPE simplifies to an easy ordinary differential equation (ODE). See Chapter 2 for details.

1.6.2 Algorithms for a Network of Fluid Queues

We next generalize the analysis from single-queue fluid models to networks. The major difficulty for the network model is that the total arrival rate at each queue of the network is not part of the model parameters. This rate is the sum of the external arrival rate (which is

part of the model data) and the rate of feedbacks from other queues. If we can obtain the total arrival rates for all queues, each of them can then be analyzed in an identical way as in Chapter 2 so that the single-queue fluid algorithm can be simply applied. Therefore, the main step of this network generalization is to obtain the total arrival rates to each queue.

We provide two algorithms for the fluid networks with exponential service distributions. In the first algorithm, we show that the vector of the total arrival rates is a fixed point in the multi-dimensional functional space. In addition, we show that this new FPE has a unique solution under general conditions so that we thus solve for the fixed point through a recursion-based algorithm. In the second algorithm, we determine the total arrival rates by solving a multi-dimensional ODE. The algorithm becomes more complicated for networks with non-exponential distributions because the single-queue fluid algorithm can no longer be applied, see Chapter 3 for details.

1.6.3 Algorithm for a Fluid Queue with Deterministic Service Times

The initial algorithm for one fluid queue is based on smooth model data, and thus does not apply to deterministic service times. However, when we analyze the fluid queue with deterministic service times in the last chapter, we modify the previous algorithm, so that it applies to models with deterministic service times. See Chapter 5 for details.

1.7 Simulation

In this thesis, computer simulation of the stochastic queueing models is employed extensively to test the accuracy of the deterministic fluid approximations of the corresponding expected values in the stochastic queueing model. Just as for the numerical algorithms, the simulations of the queueing models are run in MatLab on a personal computer.

For very large-scale models (that have a large arrival rate and a large number of servers), there is very little variability in the content stochastic processes in the queuieng model; i.e., sample paths from independent replications will tend to fall on top of each other. (The system can be said to be large when the arrival rate and the number of servers are around 1000, with the mean service time being 1.) Thus, it suffices to show that any one of these sample paths agrees closely with the numerical values computed by the algorithm for the fluid model. We show that simulation estimates of single sample paths of the performance measures, such as the time-dependent number of customers and waiting times, agree with the fluid approximating functions closely. This is consistent with expectations, because of the MSHT theoretical basis.

However, assuming a large arrival rate and a large number of servers is not reasonable for systems such as hospitals where the number of doctors and nurses can be 10 or even smaller. It is therefore important that our fluid approximation can be applied for small service systems. In this case we should not expect the deterministic fluid approximation to work well for each sample path because the stochastic fluctuations or errors cannot be simply ignored. However, the mean functions of these stochastic processes can still be well approximated by the fluid functions. See Chapters 2 and 3 for detailed examples.

We provide simulation verifications on both single-queue examples and network examples. All of these examples show that our fluid approximations are effective. Effectiveness increases as scale increases and the extent of overloading increases.

1.8 Organization of This Thesis

There are four chapters in the rest of this thesis; these are based on four completed papers [41–44], respectively.

In Chapter 2, we first restrict our attention to the deterministic $G_t/GI/s_t + GI$ fluid

model that approximates its corresponding many-server queueing model with a single class of customers handled by a single group of homogeneous servers, working in parallel. We determine the time-dependent performance functions, such as the fluid in queue and in service, the waiting time, the abandonment and service completion rate, etc. This model has a time-varying arrival rate and service capacity, abandonment from queue, and non-exponential service and patience distributions. Our key assumptions are that (i) the system alternates between OL and UL intervals, and (ii) the model functions are suitably smooth. The results show the impact of the time-varying parameters and the model distributions on the performance. Simulations confirm that the approximation and the algorithm are effective.

In Chapter 3, we extend our analysis in Chapter 2 to complex network queues, allowing time-dependent proportional routing among the queues. In particular, we consider the $(G_t/GI/s_t + GI)^m/M_t$ model. There are m queues, each with its own external fluid input, but in addition a proportion $P_{i,j}(t)$ of the fluid output from queue i at time t is routed immediately to queue j, and a proportion $P_{i,0}(t) \equiv 1 - \sum_{j=1}^{m} P_{i,j}(t) \leq 1$ is routed out of the network (departs having successfully completed all required service). This framework permits feedback, not only directly from i to i, but also indirectly from i to i after one or more transitions to other queues. We provide efficient algorithms computing all standard performance functions in a finite time interval. In addition, we characterize the steady-state behavior of the stationary version of this network fluid model.

In Chapter 4, we complement the analysis in Chapters 2 and 3 by investigating the large-time asymptotic behavior of the $G_t/M_t/s_t+GI_t$ fluid model with exponential service distributions. We establish an asymptotic loss of memory (ALOM) result which says that the impact of the initial condition dissipates as time evolves. Using this ALOM property, we develop the following two convergence results: For the $G/M/s + GI$ queue, i.e., when the model parameters are constant, we establish the convergence to the steady state (in the

infinite future). When the arrival rates are periodic functions (such as in service systems with daily or weakly cycles), we establish the existence of a periodic steady state (PSS) and the convergence to the PSS as time evolves. We also show that the convergence is exponentially fast under general regularity conditions.

In Chapter 5 we consider a stationary $GI/D/s+GI$ queueing model with a stationary general arrival process (the first GI), deterministic service times (the D), multiple servers (the s), and general abandonment times (the $+GI$). Under general conditions, the number of customers in this $GI/D/s+GI$ many-server queue at time t converges to a unique stationary distribution as $t \to \infty$. However, simulations show that the sample paths routinely exhibit nearly periodic behavior over long time intervals when the system is overloaded and s is large, provided that the system does not start in steady state. We provide insights into the transient behavior by studying the deterministic fluid model. The fluid model also has a unique stationary point, but that stationary point is not approached from any other initial state as $t \to \infty$. Instead, the fluid model performance approaches one of its uncountably many periodic steady states, depending on the initial conditions.

For this stationary $GI/D/s+GI$ queueing model, we also prove a MSHT limit, showing that the performance functions in the fluid mdoel are the limits of corresponding appropriately scaled performance functions in a sequence of the stochastic queueing models. As a result, we demonstrate the invalidity of the interchange of two limits: the steady state (obtained as $t \to \infty$) of the HT limiting process (obtained as $n \to \infty$) does not coincide with the HT limit (obtained as $n \to \infty$) of the steady state (obtained as $t \to \infty$) of the queueing processes.

Chapter 2

The $G_t/GI/s_t + GI_t$ Fluid Queue

We begin with the study of a single-queue fluid model that has time-varying arrival rate and staffing functions, general service and patience distributions. We provide an efficient algorithm for computing all standard performance measures. A key idea of this algorithm is to treat overloaded intervals and underloaded intervals separately.

2.1 Introduction

In this chapter, we study the $G_t/GI/s_t + GI$ deterministic fluid model. This model serves as an approximation for the corresponding many-server queueing model, that has a non-stationary general arrival process (the G_t), independent and identically distributed (IID) service times following a general distribution (the first GI), a time-varying staffing function (the s_t), and allows IID patience times following a general distribution (the $+GI$).

We have four important goals. First, we want to carefully define the $G_t/GI/s_t + GI$ fluid model. Second, we want to characterize its performance. Third, we want to develop

an effective algorithm for computing all the performance functions. Finally, we want to show that the resulting performance descriptions effectively approximate the performance of the corresponding large scale stochastic $G_t/GI/s_t + GI$ queueing systems. We do that by conducting simulation experiments. For very large systems, the fluid performance will closely match individual sample paths; for smaller systems, the fluid performance will closely match the mean values of the stochastic processes.

In order to recover important Markovian structure, in particular we focus on the two-parameter processes, $Q(t, y)$ and $B(t, y)$, denoting the number of customers in queue and in service for at most y at t. These quantities have interesting new features not evident from the $M_t/M_t/s_t + M_t$ fluid model.

By focusing on non-exponential service and patience distributions, we also extend [77], which developed a deterministic fluid model to approximate the steady-state performance of a *stationary* $G/GI/s+GI$ queueing model. Comparisons with simulation in Tables 1-3 of [77] show that the approximations can be very useful when the system is overloaded. Some degree of overloading is not uncommon, because even a small amount of abandonment acts to keep the system stable [3,20,76,77]. The accuracy of fluid models for capacity planning has been strongly supported by [3].

Here we consider the analogous $G_t/GI/s_t + GI$ fluid model, now including time-varying arrival rate and staffing (service capacity). We develop an algorithm to calculate all the standard performance functions. In doing so, we also provide important contributions even for the *stationary* $G/GI/s + GI$ fluid model introduced in [77]. Here we provide for the first time a full description of the transient behavior of the stationary $G/GI/s+GI$ fluid model. The fundamental evolution equations, here in (2.5), are the same as in (2.14) and (2.15) of [77], but the time-dependent performance when the system is overloaded actually depends on three features introduced for the first time here: First, for non-exponential service, the (two-parameter) fluid density in service $b(t, x)$ depends on the rate fluid enters

service, $b(t,0)$, which is characterized as the unique solution to a fixed point equation; see (4.20) and Theorem A.2. Second, the fluid density in queue, $q(t,x)$, depends on a *boundary waiting time* (BWT)

$$w(t) \equiv \inf\{y \geq 0 : q(t,x) = 0 \quad \text{for all} \quad x > y\}, \tag{2.1}$$

which is characterized here as the solution of an *ordinary differential equation* (ODE); see Theorem 2.3. Third, the *potential waiting time* (PWT) $v(t)$, i.e., the virtual waiting time of an arrival at time t if that arrival would elect never to abandon, is characterized as the unique solution of an equation involving the BWT $w(t)$ or by yet another ODE; see Theorems 2.5 and 2.6. To the best of our knowledge, none of this structure has been exposed previously.

Even though we have had to complete the story of the dynamics of the $G/GI/s+GI$ fluid model in this chapter, the steady-state description in [77] is evidently correct (which should not be surprising, since it was confirmed by simulations). For the special case of the $G/M/s+GI$ fluid model, in Chapter 4 we extend the results here to prove that the time-dependent performance converges to that steady-state performance as time evolves for any finite initial condition. Moreover, we provide bounds on the rate of convergence. In Chapter 4, we also establish convergence to a periodic steady state for periodic models and we establish asymptotic loss of memory (ALOM) for more general time-varying models.

We should also mention that a time-varying $G_t/GI/s+GI$ fluid model was already considered in §6 of [77], but that was done by considering an approximating discrete-time model, from which the new structure exposed here is not evident. In contrast, here we develop a *smooth model*; see Assumption 2.2. However, [77] provides important theoretical support because it establishes a MSHT limit for the discrete-time model, with the usual MSHT scaling, consistent with earlier asymptotic results in [20, 46–48]. Thus, we already know that the fluid model we consider arises as a MSHT limit of a sequence of scaled

queueing systems. Nevertheless, we intend to provide additional theoretical support for the fluid model with deterministic service distributions introduced here in Chapter 5 by showing that the fluid model arises as the MSHT limit of a sequence of queueing systems, under suitably regularity conditions. We do so by applying recent MSHT limits for infinite-server queues in [56]. The MSHT limit for the model with general service distributions is in progress, see [45]. The connection to infinite-server queues plays a critical role here as well; see §2.4, §2.5 and §2.7.1. The new limits in Chapter 5 are consistent with recent results in [36,37,56,62]. By uniquely characterizing the fluid limit here, the present chapter can be used as a step in the proof.

The results have significant relevance for applications. First, service systems typically have arrival rates that vary significantly over time, and the results dramatically reveal the consequence, e.g., showing how the peak congestion lags behind the peak arrival rate, as discussed for the $M_t/GI/\infty$ stochastic model in [14, 15]. Second, service systems often do have non-exponential service and patience distributions [7], and the results dramatically reveal the consequence. From [49,76,77,81], we know that the patience distribution beyond its mean has a significant impact. However, [76,77] show that the steady-state performance in the stationary $G/GI/s+GI$ model is relatively insensitive to the service-time cdf beyond its mean. In contrast, here we show that the service distribution beyond its mean can have a dramatic impact as well for the transient performance; see §2.2. Finally, the results in this chapter have already been applied in [31] to create new effective real-time delay predictors for arriving customers in a service system with time-varying arrivals.

The analysis here applies to a system that is either overloaded (OL) or underloaded (UL) for an extensive period of time, but an innovation in our approach is to consider systems that *alternate* between OL intervals and UL intervals. With time-varying arrival rates, such alternating behavior commonly occurs when it is difficult to dynamically adjust the staffing level in response to changes in demand. If the staffing cannot be changed rapidly enough,

then system managers must choose fixed or nearly fixed staffing levels that respond to several levels of demand. Then it may not be cost-effective to staff at a consistently high level in order to avoid overloading at any time. Then the fluid model introduced here may capture the essential performance.

We contend that the alternating OL and UL regime (corresponding to the MSHT ED and QD regimes [20]) can be very useful, but if staffing can be adjusted dynamically, then the system may be nearly critically loaded at all times. In that case, we anticipate that it would be better to use analysis techniques suitable for systems that are critically loaded or nearly critically loaded at all times (corresponding to the MSHT QED regime). However, it remains to develop a tractable QED approximation for the $G_t/GI/s_t + GI$ model. We think that the present model may even be useful in that setting as well, if skillfully applied.

Here is how the rest of this chapter is organized: We start in §2.2 by discussing an example, showing the results of the algorithm and how they compare to simulations of queueing systems. Next in §2.3 we carefully define the $G_t/GI/s_t + GI$ fluid model and specify key regularity conditions. In §2.4 we state important scale-proportionality results, which provide important simplification for UL intervals. In §2.5 we characterize performance during a UL interval.

In §2.6 we characterize the service content density during an OL interval. Subsections 2.6.1 and 2.6.2 are devoted to the special case of M service and non-M GI service, respectively. An explicit formula is available for M service; an iterative algorithm is developed for other cases. In §2.7 we characterize the queue performance functions: the queue content density $q(t, x)$, the BWT w and the PWT v. In §2.8 we summarize the resulting algorithm.

We have indicated in §2.3 that feasibility of the staffing function is an important issue when the staffing function can decrease during overloaded intervals. We directly assume feasibility, but in §2.9 we show how to detect the first violation of feasibility of a staffing

function and how to find the minimum feasible staffing function greater than or equal to the initial staffing function if that one is infeasible. In §2.10 we show how to construct a staffing function to stabilize delays at any fixed target value, contributing to prior work in [17, 35]. In §2.11 we provide three postponed longer proofs, the proofs of Theorems 2.3, 2.5 and 2.6. Finally, in §2.12 we draw conclusions. Additional supporting material appears in Appendix A.

2.2 An Example

We start with an example. We consider an $M_t/H_2/s + E_2$ fluid model with a sinusoidal arrival rate function: $\lambda(t) = 1 + 0.6\sin(t)$, mean service time $1/\mu = 1$, mean patience $1/\theta = 1$, and fixed service capacity $s = 1$. (We consider other examples in Appendix A.)

In choosing these values, we are not thinking of a single server and the corresponding arrival rate. Instead, we are planning to use the MSHT scaling, as discussed in [20, 46, 56, 77], when we connect the fluid model to associated queueing models. In the queueing model, we are thinking of the fluid staffing level and the arrival rate being scaled up by a factor n (e.g., $n = 20$ or $n = 100$), i.e., these models have $s_n = n\,s$ servers, arrival rate function $\lambda_n(t) = n\lambda(t)$, and the same service and patience distributions. The fluid model will serve as approximations for all such scaled queueing systems. Because of MSHT limits, we anticipate that the fluid model will yield better approximations as the scale factor n increases.

Specifically, we let the service distribution be a two-phase hyperexponential (H_2) with probability density function (pdf)

$$g(x) = p \cdot \mu_1 e^{-\mu_1 x} + (1-p) \cdot \mu_2 e^{-\mu_2 x}, \quad x \geq 0,$$

with parameters $p = 0.5(1 - \sqrt{0.6})$, $\mu_1 = 2p\mu$ and $\mu_2 = 2(1-p)\mu$, which produces squared coefficient of variation (variance divided by the square of the mean) $c^2 = 4$. We let the patience distribution be Erlang-2 (E_2) with pdf

$$f(x) = 4\theta^2 x e^{-2\theta x}, \quad x \geq 0.$$

The E_2 distribution has $c^2 = 1/2$.

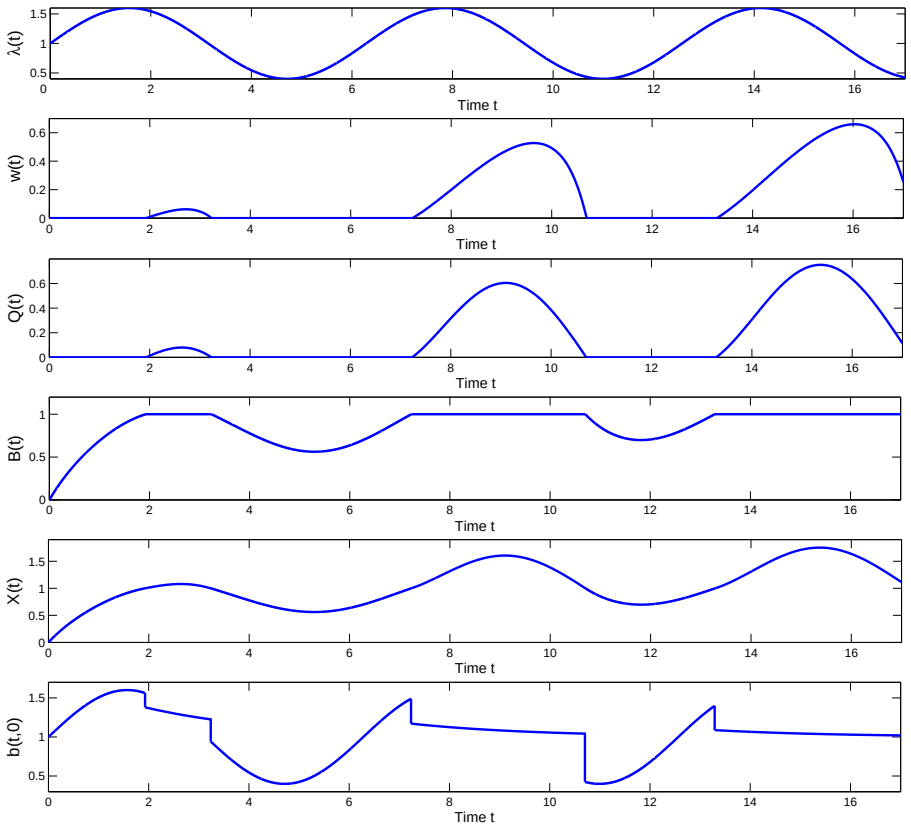

Figure 2.1: The performance functions of the $G_t/H_2/s + E_2$ fluid model with sinusoidal arrival-rate function: (i) arrival rate $\lambda(t)$; (ii) BWT $w(t)$; (iii) fluid waiting in queue $Q(t)$; (iv) fluid in service $B(t)$; (v) total fluid in system $X(t)$; (vi) rate into service $b(t, 0)$.

Figure 2.1 shows plots of several key performance functions for $0 \leq t \leq T \equiv 17$, starting out empty, together with the specified arrival rate $\lambda(t)$: the boundary waiting time (BWT) $w(t)$, the fluid content in queue $Q(t)$, the fluid content in service $B(t)$, the total

fluid content in system $X(t) \equiv Q(t) + B(t)$, and the rate fluid enters service $b(t, 0)$. All performance functions are continuous except for the rate-into-service function $b(t, 0)$. In underloaded intervals, $b(t, 0) = \lambda(t)$; in overloaded intervals, $b(t, 0)$ is the unique solution of the fixed-point equation (4.20).

It is important that the fluid model provide useful approximations for stochastic queueing models. We apply simulation to show that the fluid approximation indeed is effective for that purpose. For very large queueing systems, the stochastic system behaves like the fluid model, having relatively small stochastic fluctuations. That is illustrated for an $M_t/H_2/s + E_2$ queueing system with 2000 servers in Figure 2.2. In the plot, the queueing

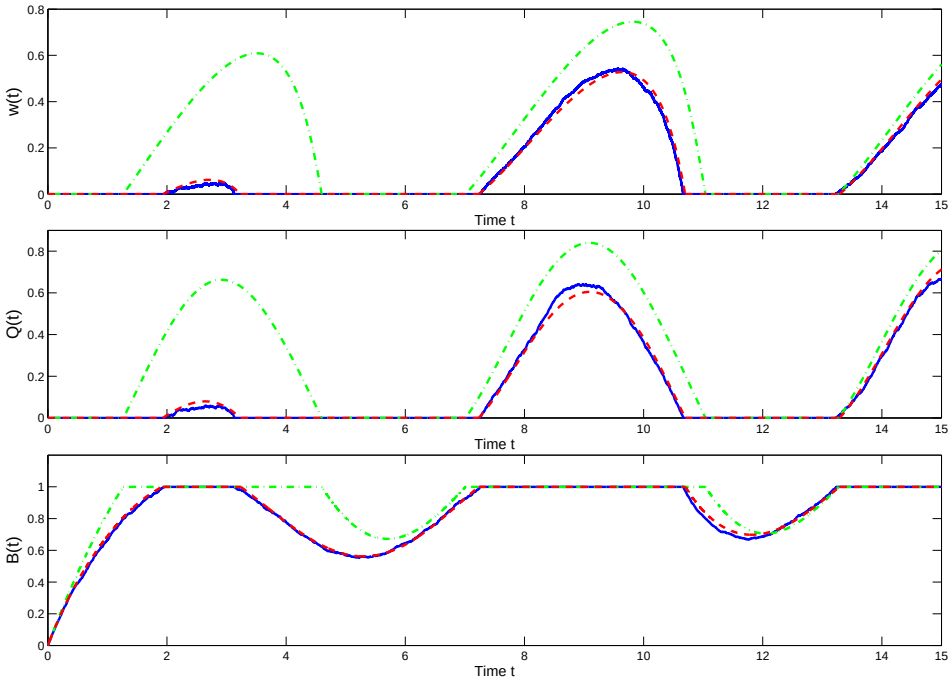

Figure 2.2: Simulation comparison for the $M_t/H_2/s + E_2$ fluid model: (i) single sample paths in the scaled queueing model based on $n = 2000$ (blue solid lines), (ii) fluid functions (red dashed lines) and (iii) fluid functions assuming M service (green dashed lines).

content processes are scaled by dividing by $n = 2000$, so that s remains at 1. For the actual queueing system, the quantities $\lambda(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$ should all be multiplied by $n = 2000$. See §2.4 for a discussion of scaling.

Figure 2.2 actually shows three plots. It also shows the fluid approximation for the corresponding $M_t/M/s + E_2$ model, having exponential service times with the same mean. For that alternative model, there is a more elementary algorithm, because it is not necessary to solve the fixed point equation for $b(t, 0)$ in order to calculate $b(t, x)$. Figure 2.2 shows two things: First, it shows that the simulation sample path for the $M_t/H_2/s + E_2$ model agrees closely with the fluid performance. Second, Figure 2.2 shows that the service distribution can make a big difference in the time-dependent performance. The performance of the fluid model changes significantly when we change the service distribution from H_2 to M (with the same mean); e.g., look at $Q(t)$ at time $t = 3$. (We do not show a simulation path for the $M_t/M/s + E_2$ model, but it agrees closely with its fluid model for $n = 2000$. See Appendix A.)

The impact of the service distribution may be surprising, because a major conclusion of [76, 77] was that the steady-state performance is relatively insensitive to the service distribution beyond its mean. However, there is precedent for this phenomenon: Davis et al. [13] showed that the performance in the time-varying $M_t/GI/s/0$ loss model depends quite strongly on the service distribution beyond its mean, even though the steady-state distribution of the stationary $M/GI/s/0$ loss model has the well known insensitivity property, concluding that the standard steady-state performance measures do not depend at all on the service distribution beyond its mean.

Figure 2.2 suggests that the periodic models approach a periodic steady state as time evolves; that is proved for the fluid model with M service in Chapter 4. (We conjecture that is also true with GI service under minimal regularity conditions, but it has not yet been proved.) Figure 2.2 also shows that the impact of the service cdf G beyond its mean evidently is far greater at the beginning when the system is starting up, and then dissipates considerably as the system approaches its periodic steady state. That is consistent with intuition, because with H_2 service, there will be more very short service times and unusually

long service times than would be the case of the exponential distribution. Hence, at the beginning starting empty, there are no old customers with long service times to compensate for many new customers with short service times in the H_2 case. As a consequence, the initial queue content is much less with H_2 than with M service. However, more supporting theory is needed.

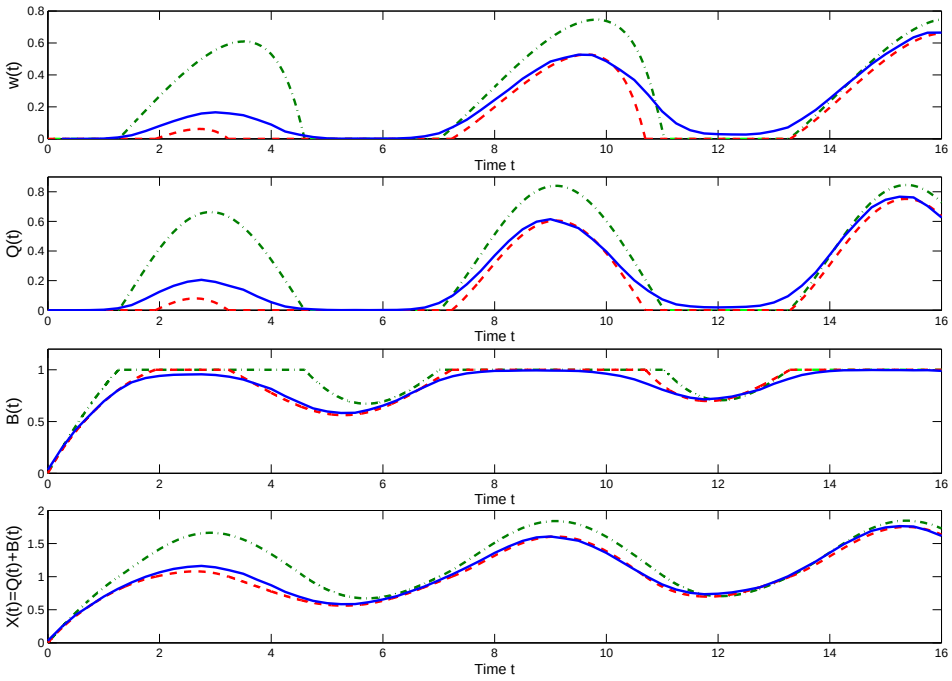

Figure 2.3: Simulation comparison for the $M_t/H_2/s + E_2$ fluid model: (i) the averages of 200 sample paths of the scaled queueing model based on $n = 30$ (blue solid lines), (ii) fluid functions (red dashed lines) and (iii) fluid functions assuming M service (green dashed lines).

Of course, most service systems have far fewer servers than the number $n = 2000$ we considered. It is thus important that the fluid approximation can still be useful with fewer servers. With fewer servers, the stochastic fluctuations in the queueing stochastic processes play an important role. In that case, the fluid model can still be very useful by providing a good approximation for the *mean values* of the queueing stochastic processes. That is illustrated from the plot of the average of the scaled performance measures of 200

independent sample paths when there are only 30 servers in Figure 2.3. We also consider the case $n = 15$ in Appendix A.

Work is in progress to establish MSHT limits and engineering refinements that will yield good approximations for the full distributions at each time t. A rough engineering approximation for $X(t)$ is to act as if it is normally distributed with variance equal to the determined mean; that is consistent with the exact Poisson distribution with the $M_t/GI/\infty$ model (and thus the stochastically equivalent $M_t/M/s_t + M$ model with $\theta = \mu$).

2.3 The $G_t/GI/s_t + GI$ Fluid Queue

In this section we define the deterministic $G_t/GI/s_t+GI$ fluid model and specify important regularity conditions. There is a service facility with finite capacity (staffing function) $s \equiv \{s(t) : t \geq 0\}$ that is set exogenously and enforced. There also is waiting space with unlimited capacity. There is a deterministic arrival process, with input directly entering the service facility if there is space available; otherwise the input flows into the waiting room. Fluid may leave the service facility only by completing service. However, fluid may leave the queue either by entering service or abandoning (leaving directly from the queue without receiving service). These flows are deterministic as well. The total input of fluid over the interval $[0, t]$ is $\Lambda(t) \equiv \int_0^t \lambda(u)\, du$, $t \geq 0$. We will be working with the time-dependent arrival-rate function $\lambda \equiv \{\lambda(t) : t \geq 0\}$.

There are service-time and abandon-time cdf's G and F, respectively, with pdf's g and f, satisfying

$$G(x) = \int_0^x g(u)du \quad \text{and} \quad F(x) = \int_0^x f(u)du, \quad x \geq 0. \tag{2.2}$$

Let $\bar{G}$ and $\bar{F}$ denote the associated complementary cdf's (ccdf's), defined by $\bar{G}(x) \equiv 1 -$

$G(x)$ and $\bar{F}(x) \equiv 1 - F(x)$. We assume that the the random service and abandon times are unbounded above, so that $\bar{G}(x) > 0$ and $\bar{F}(x) > 0$ for all x. We assume that the mean service time is 1; that choice is without loss of generality, because we can measure time in units of mean service times. In the fluid model, the cdf's act as *proportions*. A proportion $G(x)$ of any quantity of fluid completes service and departs within time x of the time it starts service; a proportion $F(x)$ of any quantity of fluid abandons and departs without receiving service within time x of the time it arrives, providing that it has remained waiting in queue, and has not already been admitted to service.

The key performance descriptors are the two-parameter functions $B(t, y)$ and $Q(t, y)$: $B(t, y)$ is the quantity of fluid in service at time t that has been in service for time less than or equal to y; $Q(t, y)$ is the quantity of fluid waiting in queue at time t that has been in queue for time less than or equal to y. These functions will admit representations

$$Q(t, y) = \int_0^y q(t, x)\, dx \quad \text{and} \quad B(t, y) = \int_0^y b(t, x)\, dx, \quad y \geq 0, \tag{2.3}$$

where the fluid densities b and q are non-negative integrable functions. Let $Q(t) \equiv Q(t, \infty)$ be the total fluid content in queue at time t, and let $B(t) \equiv B(t, \infty)$ be the total fluid content in service at time t. Let $X(t) \equiv B(t) + Q(t)$ be the total fluid content in the system at time t.

To fully specify the model, we also need to specify the initial conditions, describing the system state at time 0. The initial conditions are specified by the two functions $B(0, y)$ and $Q(0, y)$, which are defined as above, and also satisfy (2.3) with densities $b(0, x)$ and $q(0, x)$. Thus, the $G_t/GI/s_t + GI$ fluid *model data* consists of the six-tuple of functions $(\lambda, s, F, G, b(0, \cdot), q(0, \cdot))$.

We make several assumptions. The first is on the initial conditions.

Assumption 2.1 (*finite initial content*) $B(0) < \infty$ *and* $Q(0) < \infty$.

We develop a "smooth" model. For that purpose, let $\mathbb{C}_p$ be the set of *piecewise-continuous* real-valued functions, by which we mean that the function has only finitely many discontinuities in any finite interval, with left and right limits at each discontinuity point (within the interval); moreover, we assume that the function is right-continuous. Hence, $\mathbb{C}_p \subseteq \mathbb{D}$, where $\mathbb{D}$ is the space of right-continuous functions with left limits.

Assumption 2.2 (*smoothness*) $s, \Lambda, F, B(0,\cdot), Q(0,\cdot)$ *are differentiable functions with derivatives* $s', \lambda, f, b(0,\cdot), q(0,\cdot)$ *in* $\mathbb{C}_p$.

As a consequence of Assumption 2.2, $\Lambda(t) < \infty$ for all $t > 0$. (We use the assumption that $\mathbb{C}_p \subset \mathbb{D}$ here; see p. 122 of [4].) Together with Assumption 2.1, that implies the finite-content property in Assumption 2.1 holds for all t: $B(t) \leq B(0) + \Lambda(t) < \infty$ and $Q(t) \leq Q(0) + \Lambda(t) < \infty$ for all $t \geq 0$.

Whenever $Q(t) > 0$, we require there is no free capacity in service, i.e., $B(t) = s(t)$. Also, whenever $B(t) < s(t)$, then the queue is empty. These conditions are summarized in

Assumption 2.3 (*fluid dynamics constraints, FDC's*) *For all* $t \geq 0$,

$$(B(t) - s(t))Q(t) = 0 \quad \textit{and} \quad B(t) \leq s(t). \tag{2.4}$$

In general, there is no guarantee that a staffing function s is feasible; i.e., having the property that the staffing function is set exogenously and adhered to, without forcing any fluid that has entered service to leave without completing service, because we allow s to decrease. (The fluid is assumed to be incompressible.) We directly assume that the staffing

function we consider is feasible, but we also indicate how to detect the first violation and then construct the minimum feasible staffing function greater than or equal to the given staffing function; see §2.9.

Assumption 2.4 (*feasible staffing*) *The staffing function s is feasible, allowing all fluid that enters service to stay in service until service is completed; i.e., when s decreases, it never forces content out of service.*

We now consider the service discipline. We let the service discipline in the fluid model be first-come first-served (FCFS). We remark that there is much less motivation for considering other service disciplines, such as processor-sharing, with many servers than with few servers, because a few long service times can only make those few (of many) servers unavailable to other customers.

Assumption 2.5 (*FCFS service*) *Fluid enters service in order of arrival.*

As a consequence of Assumption 2.5, at time t there will be a boundary of the waiting time (BWT) as in (2.1). Clearly, first, $w(t) \geq 0$ and, second, $w(t) > 0$ if and only if $Q(t) > 0$. (Equation (2.1) is informal, because it is circular, with w depending on q, while q depends on w. We will carefully define and characterize the BWT w in §2.7.)

Based on the way the queueing system operates, we assume that q and b satisfy the following two fundamental evolution equations. Because of Assumption 2.5, fluid enters service from the queue from the right boundary of $q(t, x)$.

Assumption 2.6 (*fundamental evolution equations*) *For* $t \geq 0$, $x \geq 0$ *and* $u \geq 0$,

$$\begin{aligned} b(t+u, x+u) &= b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)}, \\ q(t+u, x+u) &= q(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad 0 \leq x < w(t) - u. \end{aligned} \tag{2.5}$$

The first equation in (2.5) says that the fluid in service that is not served remains in service (which requires that the staffing function be feasible, as in Assumption 2.4). The second equation in (2.5) says that the fluid waiting in queue that does not abandon and does not move into service, remains in queue.

Let $v(t)$ be the potential waiting time (PWT) at t, i.e., the virtual waiting time at t for an arriving quantum of fluid that has unlimited patience. The virtual waiting time at time t is the actual waiting time if there is positive input at time t; otherwise it is the waiting time of hypothetical input if it were to occur at time t. In order to simplify the analysis of the two waiting time functions w and v, we make extra assumptions: These extra assumptions will be introduced in §2.7.2 and §2.7.3.

We now turn to the flows. Let $A(t)$ be the total quantity of fluid to abandon in $[0,t]$; let $E(t)$ be the total quantity of fluid to enter service in $[0,t]$; and let $S(t)$ be the total quantity of fluid to complete service in $[0,t]$. Clearly we have the basic flow conservation equations

$$Q(t) = Q(0) + \Lambda(t) - A(t) - E(t) \quad \text{and} \quad B(t) = B(0) + E(t) - S(t), \quad t \geq 0. \tag{2.6}$$

These totals are determined by instantaneous rates. To define those rates, let $h_G(x) \equiv g(x)/\bar{G}(x) = 1$ and $h_F(x) \equiv f(x)/\bar{F}(x)$ be the hazard-rate functions of the service and abandonment time distributions, respectively. Then

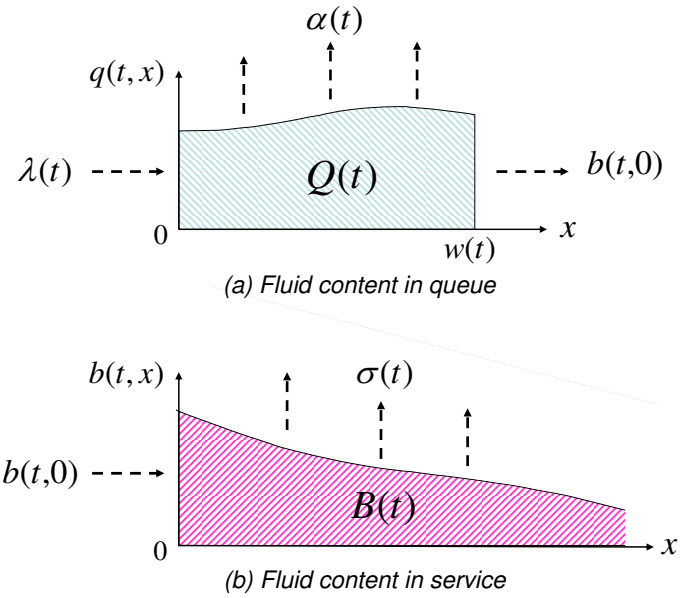

Figure 2.4: (a) The fluid in queue, (b) The fluid in service.

$$A(t) \equiv \int_0^t \alpha(u)\,du, \quad \text{where} \quad \alpha(t) \equiv \int_0^\infty q(t,x)h_F(x)dx, \quad t \geq 0. \tag{2.7}$$

$$E(t) \equiv \int_0^t b(u,0)\,du, \quad t \geq 0. \tag{2.8}$$

$$S(t) \equiv \int_0^t \sigma(u)\,du, \quad \text{where} \quad \sigma(t) \equiv \int_0^\infty b(t,x)h_G(x)\,dx, \quad t \geq 0 \tag{2.9}$$

We have now completed the definition of the $G_t/GI/s_t + GI$ fluid model (with the exception of (w,q,v), for which more is given in §2.7; Figure 2.4 provides a pictorial summary. Our goal now is to fully characterize the six-tuple (b,q,w,v,σ,α) given the model parameters (λ,s,G,F) and the initial conditions $\{(b(0,x),q(0,x)) : x \geq 0\}$, where

$q(0,x) > 0$ only if $Q(0) > 0$, which in turn, by Assumption 2.3, can hold only if $B(0) = s(0)$.

In doing so, we impose another regularity condition. We also assume that the system alternates between overloaded intervals and underloaded intervals, where these intervals include what is usually regarded as critically loaded. In particular, an *overloaded interval* starts at a time t_1 with (i) $Q(t_1) > 0$ or (ii) $Q(t_1) = 0$, $B(t_1) = s(t_1)$ and $\lambda(t_1) > s'(t_1) + \sigma(t_1)$, and ends at the *overload termination time*

$$T_1(t_1) \equiv \inf\{u \geq t_1 : Q(u) = 0 \quad \text{and} \quad \lambda(u) \leq s'(u) + \sigma(u)\}. \tag{2.10}$$

Case (ii) in which $Q(t_1) = 0$ and $B(t_1) = s(t_1)$ is often regarded as critically loaded, but because the arrival rate $\lambda(t_1)$ exceeds the rate that new service capacity becomes available, $s'(t_1) + \sigma(t_1)$, we must have the right limit $Q(t_1+) > 0$, so that there exists $\epsilon > 0$ such that $Q(u) > 0$ for all $u \in (t_1, t_1 + \epsilon)$. Hence, we necessarily have $T_1 > t_1$.

An *underloaded interval* starts at a time t_2 with (i) $Q(t_2) < 0$ or (ii) $Q(t_2) = 0$, $B(t_2) = s(t_2)$ and $\lambda(t_2) \leq s'(t_2) + \sigma(t_2)$, and ends at *underload termination time*

$$T_2(t_2) \equiv \inf\{u \geq t_2 : B(u) = s(u) \quad \text{and} \quad \lambda(u) > s'(u) + \sigma(u)\}. \tag{2.11}$$

As before, case (ii) in which $Q(t_2) = 0$ and $B(t_2) = s(t_1)$ is often regarded as critically loaded, but because the arrival rate $\lambda(t_2)$ does not exceed the rate that new service capacity becomes available, $s'(t_2) + \sigma(t_2)$, we must have the right limit $Q(t_2+) = 0$. The underloaded interval may contain subintervals that are conventionally regarded as critically loaded; i.e., we may have $Q(t) = 0$, $B(t) = s(t)$ and $\lambda(t) = s'(t) + \sigma(t)$. For the fluid models, such critically loaded subintervals can be treated the same as underloaded subintervals. However, unlike an overloaded interval, we cannot conclude that we necessarily

have $T_2 > t_2$ for an underloaded interval. Moreover, even if $T_2 > t_2$ for each underloaded interval, we could have infinitely many switches in a finite interval. We directly assume that those pathological situations do not occur. Let $\mathcal{R}$ denote the system regime, i.e., $\mathcal{R} =$ OL or UL. Let the interval termination time (starting at t_0

$$T_{\mathcal{R}}(t_0) \equiv T_1(t_0)\mathbf{1}_{\{\mathcal{R}(t_0\}} = \text{OL} + T_2(t_0)\mathbf{1}_{\{\mathcal{R}(t_0=\text{UL}\}}.$$

Assumption 2.7 (*finitely many switches between intervals in finite time*) *Each underloaded interval is of positive length, so that the positive half line* $[0, \infty)$ *can be partitioned into overloaded and underloaded intervals. Moreover, there are only finitely many switches between overloaded and underloaded intervals in each finite interval.*

For engineering applications, Assumption 2.7 is reasonable, but it is unappealing mathematically. We would like to have natural conditions on the model parameters under which the conclusion does hold. For the special case of M service and for the extension to time-varying Markovian service (M_t), we provide sufficient conditions for Assumption 2.7 to be satisfied in §3.3 of Chapter 3. From a practical perspective, Assumption 2.7 provides no restriction, because we can discover violations when calculating the performance descriptions, and remove any violation that we discover by negligibly modifying either the arrival rate function λ or the staffing function s in a neighborhood of the problem time t to remove the problem. That is most easily done with the arrival-rate function λ, because we only require that it be piecewise-continuous. For t in a short interval $[a, b]$, we can replace $\lambda(t)$ by $\lambda(t) \pm \epsilon$. This will introduce new discontinuity points at the end points a and b (if they were not already discontinuity points), but that leaves $\lambda \in \mathbb{C}_p$.

All assumptions above are in force throughout this chapter. We will introduce additional

regularity assumptions as needed, starting in §2.6. We now determine the performance, first considering an underloaded interval.

2.4 Scale Proportionality

To treat an underloaded interval in the next section, we will exploit an important scale proportionality property of the $M_t/GI/\infty$ stochastic queueing model; see Remark 5 of [14]. For each $c > 0$, let $B_c(t, y)$ be the number of customers in service in the $M_t/GI/\infty$ stochastic model at time t that have been so for a duration at most y when the system starts empty at time 0 and the arrival-rate function is $\lambda_c(t) \equiv c\lambda(t)$, for some given arrival-rate function λ and service cdf. The following is proved like Theorem 1 of [14], using the two-parameter framework, as in [56].

Proposition 2.1 (scale proportionality in the $M_t/GI/\infty$ stochastic model) *For all $c > 0$, $B_c(t, y)$ has a Poisson distribution with mean*

$$m_c(t, y) \equiv E[B_c(t, y)] = cm_1(t, y) = c \int_0^{t \wedge y} \lambda(t - x)\bar{G}(x)\, dx. \tag{2.12}$$

As a consequence of the SLLN for the Poisson distribution, we see that $c^{-1}B_c(t, y) \to m_1(t, y)$ as $c \to \infty$ for each t and y. In addition, we have the more general FWLLN in [56, 62], which implies that $c^{-1}B_c(t, y) \to m_1(t, y)$, regarded as functions of t and y. Hence, the mean function $m_1(t, y)$ in the $M_t/GI/\infty$ stochastic queueing model directly coincides with the limit of the scaled process; i.e.,

$$m_1(t, y) \equiv E[B_1(t, y)] = B(t, y),$$

where $B(t, y)$ is the fluid content in service at time t that have been so for a duration at most y in the $M_t/GI/\infty$ fluid model. Thus, aside from scale, the mean $m_c(t, y) \equiv E[B_c(t, y)]$ in the $M_t/GI/\infty$ stochastic model coincides with the corresponding fluid content in the deterministic fluid model.

Moreover, the conclusions above extend to the more general $G_t/GI/\infty$ models. First, the mean function in (2.12) above in the $G_t/GI/\infty$ stochastic model actually coincides with the mean function in the $M_t/GI/\infty$ stochastic model, provided that the arrival rate function is the same; this observation is made in Remark 2.3 of [50]. Second, the FWLLN in [56, 62] actually holds for the $G_t/GI/\infty$ stochastic model, provided that the arrival process satisfies a FWLLN. To summarize, the mean function in the $M_t/GI/\infty$ stochastic model coincides with the fluid content in the corresponding $G_t/GI/\infty$ fluid model, assuming appropriate scale.

This scale proportionality in the infinite-server stochastic model actually extends to the more general $G_t/GI/s_t + GI$ fluid model. The following scale proportionality result is a consequence of the results in this chapter.

Theorem 2.1 (scale proportionality in the $G_t/GI/s_t + GI$ fluid model)

If the vector $(b_c(t, x), q_c(t, x), w_c(t), v_c(t), \alpha_c(t), \sigma_c(t))$ *is the performance at time* t *associated with model data* $(c\lambda, cs, F, G, cb(0, \cdot), cq(0, \cdot))$*, then*

$$(b_c, q_c, \alpha_c, \sigma_c) = c(b_1, q_1, \alpha_1, \sigma_1) \quad \textit{and} \quad (w_c, v_c) = (w_1, v_1).$$

2.5 An Underloaded Interval

We will consider the system over successive intervals, during each of which it is either underloaded or overloaded, as defined above. We start with the easier case, in which the system is underloaded. Without loss of generality, we assume that an underloaded interval starts at time 0 and terminates at a time T, defined in (2.11). We do not need to know in advance the termination time T. Instead, we can assume that the system is underloaded over the full interval $[0, \infty)$ and then calculate T.

If the $G_t/GI/s_t + GI$ fluid model is underloaded, then there is no queue, and so no abandonment. Then the model is is equivalent to the associated $G_t/GI/\infty$ fluid model. We thus can obtain results for an underloaded interval directly from available results for the $M_t/GI/\infty$ queue in [14, 56] by invoking §2.4.

Since $b(t, 0) = \lambda(t)$ when the system is underloaded, we immediately obtain an expression for $b(t, x)$ from (2.5). Recall that we have assumed that $b(0, \cdot) \in \mathbb{C}_p$.

Proposition 2.2 (service content in an underloaded interval) *For the fluid model with unlimited service capacity* ($s(t) \equiv \infty$ *for all* $t \geq 0$),

$$\begin{aligned} b(t,x) &= \bar{G}(x)\lambda(t-x)1_{\{x\leq t\}} + \frac{\bar{G}(x)}{\bar{G}(x-t)}b(0,x-t)1_{\{x>t\}} \qquad (2.13)\\ B(t,y) &= \int_0^{t\wedge y} \bar{G}(x)\lambda(t-x)\,dx + \int_0^{(y-t)\vee 0} \frac{\bar{G}(x+t)}{\bar{G}(x)}b(0,x)\,dx,\\ B(t) &= \int_0^{t} \bar{G}(x)\lambda(t-x)\,dx + \int_0^{\infty} \frac{\bar{G}(x+t)}{\bar{G}(x)}b(0,x)\,dx\\ &\leq \Lambda(t) + B(0) < \infty, \quad 0 \leq t < T. \end{aligned}$$

If, instead, a finite-capacity system starts underloaded, then the same formulas apply over the interval $[0, T)$, *where the underload termination time is* $T \equiv \inf\{t \geq 0 : B(t) > s(t)\}$,

with $T = \infty$ if the infimum is never obtained. Hence, $b(t,\cdot), b(\cdot,x) \in \mathbb{C}_p$ for all $t \geq 0$ and $x \geq 0$, for t in the underloaded interval.

During an underloaded interval, $b(t,x)$ depends upon the pair (λ, G) and the initial condition $b(0,x)$. There is no queue, so (q, F, w, v) play no role. The different roles of the two regimes are summarized in Figure 2.4. Hence, Proposition 2.2 fully describes the performance during underloaded intervals. The final piecewise-continuity conclusion ensures that the piecewise-continuity property assumed for $b(0,\cdot)$ will pass on to subsequent intervals when we consider successive intervals.

Remark 2.1 $(\textit{discontinuity at } t = x)$ *From (2.13), we see that b inherits the smoothness of G, λ and $q(0,\cdot)$ except when $t = x$. That will be a persistent theme throughout our analysis. For general initial conditions, this discontinuity is fundamental, so we cannot expect greater smoothness. However, away from the set $\{(t,x) : t = x\}$, we can expect smoothness of the model parameters to be reflected in our performance descriptions.*

Remark 2.2 $(\textit{the generic scalar transport PDE})$ *If, in addition to the assumptions of Proposition* 2.2*, λ and $b(0,\cdot)$ are differentiable a.e. with respect to Lebesgue measure on $[0,\infty)$, then, for each t and x, $b(t,x)$ has first partial derivatives with respect to t and x a.e. with respect to Lebesgue measure on $[0,\infty)$. Moreover, b satisfies the following PDE a.e. with respect to Lebesgue measure on $[0,\infty) \times [0,\infty)$, a simple version of the generic scalar transport equation:*

$$b_t(t,x) + b_x(t,x) \equiv \frac{\partial b}{\partial t}(t,x) + \frac{\partial b}{\partial x}(t,x) = -h_G(x)b(t,x).$$

with boundary conditions $\{b(t,0) = \lambda(t) : t \geq 0\}$ *and* $\{b(0,x) : x \geq 0\}$*; see Appendix* §*A.2.*

We now give a monotonicity result comparing two underloaded fluid models. For this result, we exploit hazard rate order, writing $h_{G_1} \leq h_{G_2}$ if $h_{G_1}(x) \leq h_{G_2}(x)$ for all $x \geq 0$, for cdf's satisfying the assumptions in §2.3. It is easy to see that hazard rate order implies ordinary stochastic order via the representation

$$\bar{G}(x) = e^{-\int_0^x h_G(u)\,du}, \quad x \geq 0. \tag{2.14}$$

Proposition 2.3 (comparison result for b in an underloaded model) *Consider two underloaded fluid models. If* $\lambda_1 \leq \lambda_2$*,* $b_1(0,\cdot) \leq b_2(0,\cdot)$ *and* $h_{G_1} \geq h_{G_2}$ *as functions, then* $b_1 \leq b_2$*, i.e.,* $b_1(t,x) \leq b_2(t,x)$ *for all* $t \geq 0$ *and* $x \geq 0$*, and* $T_1 \leq T_2$*, where* T_i *is the underload termination time in model* i*.*

Proof. Apply (2.13) after applying (2.14) to write

$$\bar{G}(x)/\bar{G}(x-t) = \exp\{-\int_{x-t}^{x} h_G(u)\,du\}. \quad \blacksquare$$

The system could be in an underloaded period for an extended period of time. If so, it is often convenient to consider the system starting empty in the distant past. (That is done for the corresponding infinite-server queueing models in [14, 50].) That allows us to directly construct stationary versions, including periodic versions, if that is warranted.

Proposition 2.4 (starting empty in the distant past) *Suppose the system started empty in the distant past* (*at* $t = -\infty$) *and has been underloaded up to time* t*. If* $\int_0^\infty \bar{G}(x)\lambda(t-x)\,dx, <$

∞, then

$$\begin{aligned} b(t,x) &= \bar{G}(x)\lambda(t-x) \le \lambda(t-x), \quad B(t) = \int_0^\infty \bar{G}(x)\lambda(t-x)\,dx, \\ B(t,y) &= B(t) - \int_0^\infty \bar{G}(x+y)\lambda(t-x-y)\,dx = \int_0^y \bar{G}(x)\lambda(t-x)\,dx \end{aligned}$$

for $x \ge 0$ and $y \ge 0$. If the arrival-rate function λ is constant or periodic, then so are $b(t,\cdot)$, $B(t)$ and $B(t,\cdot)$.

As noted above, the expression for $B(t)$ coincides with the mean number of busy servers in the $M_t/GI/\infty$ model studied in [14, 50]; see these sources for additional structural results. The expressions for the two-parameter function $B(t,y)$ and $b(t,x)$ coincide with the corresponding mean values in [56].

2.6 The Service Content in An Overloaded Interval

Without loss of generality, we assume that the overloaded interval begins at time 0 and ends at time T satisfying (3.3). Again, we do not need to know the end time T in advance, because we can calculate it while we are calculating the performance measures q and w. We proceed under the assumption that the arrival rate is sufficiently large that the system is overloaded throughout a specified interval $[0,T)$ (up to, but not including, time T), and afterwards detect violations before time T, if there are any, and then reduce the interval, if necessary.

2.6.1 The Special Case of M Service

The service content density is easy to compute if the service distribution is exponential, so we consider that case first. From (2.5), we can write down an expression for $b(t,x)$ during the overloaded interval:

$$\begin{aligned} b(t,x) &= b(t-x,0)\bar{G}(x)1_{\{x\le t\}} + b(0,x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}1_{\{x>t\}}, && (2.15)\\ &= b(t-x,0)e^{-x}(x)1_{\{x\le t\}} + b(0,x-t)e^{-t}1_{\{x>t\}}, && (2.16)\end{aligned}$$

where $b(0,x-t)$ is part of the initial conditions, but where $b(t-x,0)$ remains to be specified.

Since the service is exponential, the output rate, $\sigma(t)$, and thus the rate fluid enters service, $b(t,0)$, depend only on the staffing function s, in particular, on the values $s(t)$ and $s'(t)$. (Recall that the mean service time has been fixed at 1.)

Proposition 2.5 (the service content in an overloaded interval) *When the service distribution is exponential, the departure (service completion) rate satisfies $\sigma(t) = B(t)$, $t \ge 0$, and, during each overloaded interval, the departure rate $\sigma(t)$ and rate fluid enters service $b(t,0)$ have the simple form*

$$\sigma(t) = B(t) = s(t) \quad \textit{and} \quad b(t,0) = s'(t) + s(t) \quad \textit{for all} \quad t, \qquad (2.17)$$

depending only on the staffing function s. Then b is fully characterized by (2.16) *and* (2.17) *during an overloaded interval. Also $b(t,\cdot), b(\cdot,x) \in \mathbb{C}_p$ for all x, $t < T$.*

Proof. Apply (2.9).

2.6.2 General GI Service

We start with the general expression for the service content density given in (2.16), but it requires the rate into service $b(t,0)$, which is part of what we are trying to determine. Since the system is assumed to be overloaded over an initial interval $[0,T)$, the rate into service is determined by the rate service capacity becomes available. Thus, by (2.9), we have

$$b(t,0) = s'(t) + \sigma(t) = s'(t) + \int_0^\infty b(t,x) h_G(x) dx, \quad 0 \le t < T. \tag{2.18}$$

We now substitute equation (2.16) into equation (2.18) to obtain the following equation for the function $b(t,0)$:

$$b(t,0) = \hat{a}(t) + \int_0^t b(t-x,0) g(x)\, dx, \tag{2.19}$$

where

$$\hat{a}(t) \equiv s'(t) + \int_0^\infty \frac{b(0,y) g(t+y)}{\bar{G}(y)}\, dy. \tag{2.20}$$

From (2.20), we see that $\hat{a} \in \mathbb{C}_p \subseteq \mathbb{D}$ provided that the integral in (2.20) is finite. From (4.20), it is evident that $b(t,0)$ is a fixed point of the operator $\mathcal{T} : \mathbb{D} \to \mathbb{D}$, where

$$\mathcal{T}(u)(t) \equiv \hat{a}(t) + \int_0^t u(t-x) g(x)\, dx. \tag{2.21}$$

Under regularity conditions, we can show that there exists a unique solution to equation (4.20) by applying the Banach (contraction) fixed point theorem. We will use the complete (nonseparable) normed space $\mathbb{D}$ with the uniform norm over the interval $[0,T]$, i.e.,

$$\|u\|_T \equiv \sup_{0 \le t \le T} \{|u(t)|\}. \tag{2.22}$$

We will require an additional bound on the tail of the initial service content density $b(0,\cdot)$. Recall that we have assumed that $\bar{G}(x) > 0$ for all x.

Assumption 2.8 $(\textit{tail of } b(0,\cdot))$ *The tail of* $b(0,\cdot)$ *is bounded relative to the service-time pdf* g *via*

$$\tau(b,g,T) \equiv \sup_{0\le s\le T} \int_0^\infty \frac{b(0,y)g(s+y)}{\bar{G}(y)}\,dy < \infty,$$

Assumption 2.8 warrants discussion, because it is unappealing. At first glance, it passes the requirement that the assumptions be on the model data, because the service density g, the associated cdf G and the initial fluid content in service $b(0,\cdot)$ are all part of the model data. However, in application we will be applying the algorithm recursively over several UL and OL intervals. We would thus not know in advance the function $b(0,\cdot)$ in all OL intervals after an initial one. It is thus important that we provide readily available sufficient conditions for Assumption 2.8 to hold; we do that after we state the theorem. For now, we point out that there is a simple practical condition implying Assumption 2.8 to hold: It suffices for the service hazard rate function h_G to be bounded. (See below.)

Theorem 2.2 (service content in the overloaded case) *Consider an overloaded interval* $[0,T]$. *If Assumption* 2.8 *holds, then the operator* $\mathcal{T}$ *in* (2.21) *is a monotone contraction operator on* $\mathbb{D}$ *with contraction modulus* $G(T)$ *for the norm* $\|\cdot\|_T$ *defined in* (A.8)*, so that a finite function* $b(t,0)$ *is uniquely characterized via equation* (4.20)*. Hence, for any* $u \in \mathbb{D}$*, the fixed point can be approximated by the* n*-fold iteration* $\mathcal{T}^{(n)}$ *of the operator* $\mathcal{T}$ *applied to* u*, with*

$$\|\mathcal{T}^{(n)}(u) - \hat{b}\|_T \le \frac{G(T)^n}{1-G(T)}\|\mathcal{T}(u) - u\|_T \to 0 \quad \textit{as} \quad n \to \infty \tag{2.23}$$

and, if $u \leq (\geq)\mathcal{T}(u)$*, then* $\mathcal{T}^{(n-1)}(u) \leq (\geq)\mathcal{T}^{(n)}(u) \leq (\geq)\hat{b}$ *for all* $n \geq 1$.

Proof. Clearly, Assumption 2.8 implies that $\|\hat{a}\|_T < \infty$, so that $\mathcal{T}$ maps $\mathbb{D}$ into $\mathbb{D}$. Moreover, the contraction property follows from

$$\begin{aligned} \|\mathcal{T}(u_1) - \mathcal{T}(u_2)\|_T &= \sup_{0\leq t\leq T} \{\int_0^t (u_1(t-x) - u_2(t-x))g(x)\} \\ &\leq \|u_1 - u_2\|_T \int_0^T g(x)\,dx = \|u_1 - u_2\|_T G(T). \quad \blacksquare \end{aligned}$$

Remark 2.3 *Note we require* $G(T) < 1$ *in the proof of Theorem A.2, which holds because we have assumed that* $\bar{G}(x) > 0$ *for all* x*. However, that requirement is actually not necessary, because we can always work in an interval* $[0, \delta]$ *as long as* $G(\delta) < 1$ *for some* $\delta > 0$*. We can show the uniqueness of* $b(\cdot, 0)$ *for all* $0 \leq t \leq T$ *by recursively considering successive intervals of length* δ*.*

We now return to Assumption 2.8, which restricts the class of allowed service cdf's in a rather complicated way. We will show that it suffices for the service hazard rate h_G to be bounded. But even that is often not necessary in practice. It is important to note that Assumption 2.8 is always satisfied in a case of principle interest: if there exists y_0 such that $b(0, y) = 0$ for all $y \geq y_0$. That case occurs whenever the system started empty at some (finite) time in the past. That case occurs if the overloaded interval of interest begins at time t, $0 \leq t < T$, after the system has begun empty with $b(0, y) \equiv 0$ for all y; then necessarily $b(t, y) = 0$ for all $y > t$, by virtue of Assumption 2.6. Then

$$\tau \leq B(0, T)g^{\uparrow}(2T)/\bar{G}(T) < \infty, \tag{2.24}$$

where $x^{\uparrow}(t) \equiv \sup\{x(s) : 0 \leq s \leq t\}$.

Nevertheless, other initial conditions are interesting. For example, for the stationary model, we might start with the stationary fluid content, which has the form we have $b(0,y) = \bar{G}(y)$, $y \geq 0$, because $\bar{G}$ is the stationary-excess or equilibrium-residual-lifetime density of the service-time distribution; see [77]. Thus we now present other sufficient conditions for Assumption 2.8.

Remark 2.4 (*sufficient conditions for the bound when* $B(t) - B(0,y) > 0$ *for all* y.) *Clearly, we need to control the initial content density* $b(0,y)$ *and/or the service pdf* $g(y)$ *in order for Assumption 2.8 to hold. An easy sufficient condition directly related to the stationary fluid content density for the stationary model is for there to exist a constant* K *such that* $b(0,y) \leq K\bar{G}(y)$ *for all* $y \geq 0$. *Another easy sufficient condition for the bound in Assumption 2.8 is to have*

$$\sup_{0 \leq t < T} \{ \int_0^\infty b(0,y) h_G(y+t)\, dy \} < \infty. \tag{2.25}$$

In turn, three different sufficient conditions for (2.25) are:

(i) $\sup_{x \geq 0} \{h_G(x)\} < \infty$ *(bounded hazard rate, using* $B(0) < \infty$*)*;

(ii) *there exists* $\beta > 0$ *and* K *such that*

$$\int_0^\infty b(0,y) e^{\beta y}\, dy < \infty \quad \textit{and} \quad h_G(x) \leq K e^{\beta} x \quad \textit{for all} \quad x \geq 0.$$

(iii) $\limsup_{y \to \infty} \{b(0,y)/\bar{G}(y)\} < \infty$

(*using* $\sup_{0 \leq y \leq t} b(0,y) < \infty$ *and* $\sup_{0 \leq y \leq t} h_G(0,y) < \infty$ *for all* $t \geq 0$)

So far, we can only conclude that the function $b(t,0) \in \mathbb{D}$. We can obtain additional smoothness properties by imposing additional smoothness conditions on the model elements s and g. We use these properties for $b(\cdot,0)$ to establish properties of the ODE to calculate the BWT w in §3.4 of Chapter 3.

Corollary 2.1 (smoothness of service content in the overloaded case) *If s' and g are continuous, then $b(\cdot,0)$ is continuous as well. In that case, $b(\cdot,x)$ and $b(t,x)$ are elements of $\mathbb{C}_p$ for each $x \geq 0$ and $t \geq 0$.*

Proof. Under the extra smoothness conditions, we can apply the contraction fixed point theorem on the closed subspace $\mathbb{C}$ of continuous functions in $\mathbb{D}$, with the same uniform norm. Then the fixed point is necessarily in $\mathbb{C}$ as well. ∎

We discuss alternative algorithms to calculate b in Appendix §A.3.

2.7 The Queue Performance Functions

We now turn to the queue during an overload interval. To do so, it is convenient to initially ignore the flow into service.

2.7.1 The Queue Content Ignoring Flow Into Service

Let $\tilde{q}(t,x)$ be $q(t,x)$ during the overload interval $[0,T)$ under the assumption that no fluid enters service from queue. We can once again invoke the connection to the $M_t/GI/\infty$ stochastic model, discussed in §2.4 to treat $\tilde{q}(t,x)$ just as we treated b in §2.5, because we can let the general patience cdf F play the role of the general service-time cdf G. Instead

of (2.5), we can write

$$\tilde{q}(t+u, x+u) = \tilde{q}(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad x \geq 0, \tag{2.26}$$

to obtain the following proposition.

Proposition 2.6 $(\textit{queue content without transfer into service in the overloaded case})$ *In the overloaded case,*

$$\tilde{q}(t,x) = \lambda(t-x)\bar{F}(x)1_{\{x \leq t\}} + q(0, x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\}}. \tag{2.27}$$

so that $\tilde{q}(t,\cdot)$ *and* $\tilde{q}(\cdot,x)$ *belong to* $\mathbb{C}_p$ *for each* t *and* x*.*

Remark 2.5 Just as we observed for b in an underloaded interval in Remark 2.2, in an overloaded interval $\tilde{q}$ satisfies a version of the generic scalar transport PDE.

Paralleling Proposition 2.3, we have the following comparison result, proved in the same way.

Proposition 2.7 $(\text{comparison result for } \tilde{q})$ *Consider two overloaded fluid models. If* $\lambda_1 \leq \lambda_2$*,* $q_1(0,\cdot) \leq q_2(0,\cdot)$ *and* $h_{F_1} \geq h_{F_2}$ *as functions, then* $\tilde{q}_1 \leq \tilde{q}_2$*, i.e.,* $\tilde{q}_1(t,x) \leq \tilde{q}_2(t,x)$ *for all* $t \geq 0$ *and* $x \geq 0$*.*

We now derive q and w. The proper definition and characterization of the BWT w is somewhat complicated. We easily get an expression for q provided that we can find w.

Corollary 2.2 *(from $\tilde{q}$ to q) Given the BWT w,*

$$\begin{aligned} q(t,x) &= \tilde{q}(t-x,0)\bar{F}(x)1_{\{x\le w(t)\wedge t\}} + \tilde{q}(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\le w(t)\}} \\ &= q(t-x,0)\bar{F}(x)1_{\{x\le w(t)\wedge t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\le w(t)\}}. \end{aligned} \quad (2.28)$$

Moreover, $q(t,\cdot)\in\mathbb{C}_p$ for all $t\ge 0$.

Proof. Combine Proposition 2.6 and (3.15) to deduce that $q(t,\cdot)\in\mathbb{C}_p$ for all t, x.

2.7.2 The Boundary Waiting Time w

It now remains to define and characterize the BWT w. We can *define* the BWT w by exploiting flow conservation, in particular, by exploiting the fact that two expressions for the amount of fluid to enter service over any interval $[t,t+\delta]$ coincide; i.e.,

$$E(t+\delta)-E(t)\equiv\int_t^{t+\delta} b(u,0)\,du = I(t,w(t),\tilde{q},\delta)-A(t,t+\delta), \quad (2.29)$$

where

$$I\equiv I(t,w(t),\tilde{q},\delta)\equiv\int_{w(t)-\epsilon(t,\delta)}^{w(t)}\tilde{q}(t,x)dx \quad (2.30)$$

is the amount of fluid removed from the right boundary of $\tilde{q}$, starting at $x=w(t)-\epsilon(t,\delta)$ and ending at $x=w(t)$, during the time interval $[t,t+\delta]$ (where $\epsilon(t,\delta)$ is yet to be determined) and $A(t,t+\delta)$ is the amount of the fluid content in I that abandons in the interval $[t,t+\delta]$. We *define* the BWT w by letting $\delta\downarrow 0$ in (2.29). We will show in Theorem 2.3 below that, under regularity conditions, the relation in (2.29) determines an ODE for w

that has a unique solution. Hence, we will show that the relation (2.29) serves to properly define w and characterize it.

We need two more regularity conditions. First, we assume that the initial value $w(0)$ for the interval we consider is finite. We will be representing w as the solution of an initial value problem involving an ODE, so this is needed.

Assumption 2.9 (*finite initial BWT*) $0 \leq w(0) < \infty$.

Second, we require that the functions $\lambda(t)$ and $q(0,x)$ be appropriately bounded away from 0.

Assumption 2.10 (*positive arrival rate and initial queue density*) *For all* $t \geq 0$,

$$\begin{aligned} \lambda_{\inf}(t) &\equiv \inf_{0\leq u\leq t}\{\lambda(u)\} > 0, \quad \textit{and} \\ q_{\inf}(0) &\equiv \inf_{0\leq u\leq w(0)}\{q(0,u)\} > 0 \quad \textit{if} \quad w(0) > 0. \end{aligned}$$

By equation (3.14), Assumption 2.10 for λ implies that $\tilde{q}(t,x) > \epsilon\bar{F}(x) > 0$ on $[0,T)$ for some positive ϵ. That is useful because $\tilde{q}(t,x)$ appears in the denominator in an expression for the derivative of w in (2.31) below. The BWT w can be discontinuous if these functions are 0 over subintervals; we give examples in Appendix A.5. We show that w can be discontinuous if $\lambda(t) = 0$ or $q(0,\cdot) = 0$ over a subinterval, while w can have an infinite derivative corresponding to zeros of these functions. However, we obtain the following positive result, proved in §2.11. Let $x(t+)$ and $x(t-)$ denote the right and left limits of a function x at t, respectively. We can obtain a more elementary statement and proof if we assume even more regularity conditions; see Appendix §A.4.

Theorem 2.3 (*the BWT ODE*) *Consider an overloaded interval* $[0,T)$. *If Assumptions*

2.9–2.10 *hold, then the BWT* w *is well defined being the unique solution of the initial value problem* (IVP) *on* $[0, T)$ *based on the ODE*

$$w'(t+) = \Psi(t, w(t)) \equiv 1 - \frac{b(t+, 0)}{\tilde{q}(t, w(t)-)} \tag{2.31}$$

and any initial value $w(0)$*. In addition,* w *is Lipschitz continuous on* $[0, T]$ *with* $w(t+u) \leq w(t) + u$ *for all* $t \geq 0$ *and* $u \geq 0$ *with* $t + u \leq T$*. Moreover,* w *is right differentiable everywhere with right derivative* $w'(t+)$ *given in* (2.31) *and left differentiable everywhere* $(but\ not\ necessarily\ differentiable)$ *with value*

$$w'(t-) \;\; = \;\; \tilde{\Psi}(t, w(t)) \equiv 1 - \frac{b(t-, 0)}{\tilde{q}(t, w(t)+)}. \tag{2.32}$$

Overall, w *is continuously differentiable everywhere except for finitely many* t*.*

Remark 2.6 $(different\ roles\ of\ b(t, 0)\ and\ F\ in\ shaping\ q)$ Our use of $\tilde{q}$ as an intermediate step in constructing q helps show the different roles played by $b(t, 0)$ and F in producing q. First, the abandonment (F) controls the shape of $\tilde{q}(t, x)$ and thus $q(t, x)$ only for $x < w(t)$. Second, the transportation rate $b(t, 0)$ controls only $w(t)$, the right boundary or the truncation of $\tilde{q}(t, x)$ on x; it does not affect $\tilde{q}(t, x)$ itself, and thus $q(t, x)$ for any $0 \leq x < w(t)$.

We give closed-form formulae for some special cases in the next corollary, proved in Appendix §A.4.

Corollary 2.3 *Suppose the system is overloaded for* $0 \le t < T$ *and* $w(0) = 0$.

(a). *For the* $G_t/M/s_t$ *fluid model without customer abandonment* ($\bar{F}(x) = 1$ *for* $x \ge 0$),

$$w(t) = t - \Lambda^{-1}(\int_0^t b(y,0)dy), \quad 0 \le t < \bar{t},$$

for $\Lambda^{-1}(x) \equiv \inf\{y > 0 : \Lambda(y) = x\}$, *and* $\bar{t} \equiv \inf\{t > 0 : \Lambda(t) = \int_0^t b(y,0)dy\}$.

(b). *For the* $G_t/M/s_t + M$ *fluid model, where the abandonment-time cdf is exponential* ($\bar{F}(x) = e^{-\theta x}$, $x \ge 0$),

$$w(t) = t - \tilde{\Lambda}^{-1}(\int_0^t b(y,0)e^{\theta y}dy), \quad 0 \le t < \tilde{t}, \tag{2.33}$$

where $\tilde{\Lambda}(t) \equiv \int_0^t \lambda(y)e^{\theta y}dy$, $\tilde{\Lambda}^{-1}(x) \equiv \inf\{y > 0 : \tilde{\Lambda}(y) = x\}$, *and* $\tilde{t} \equiv \inf\{t > 0 : \tilde{\Lambda}(t) = \int_{t_1}^t b(y,0)e^{\theta y}dy\}$.

2.7.3 The Potential Waiting Time

In the previous subsection, we characterized the dynamics of the BWT w. Now we want to connect w to the PWT v, the waiting time of an arriving quantum of fluid at time t that is infinitely patient.

As shown in [48], the PWT v can be defined as a first passage time, with abandonment after time t computed with the input turned off; also see [68]. Let $A_t(u)$ be the total fluid

abandoning in the interval $[t, t+u]$ in our fluid model, modified by having the input shut off after time t. Paralleling (2.7),

$$A_t(u) \equiv \int_t^{t+u} \alpha_t(s)\, ds \quad \text{and} \quad \alpha_t(s) \equiv \int_{s-t}^{\infty} q(s,x) h_F\, dx, \quad s \geq t, \tag{2.34}$$

where $\alpha_t(s)$ is the abandonment rate of the fluid that arrives before time t, at time s.

With (2.34), we can define $v(t)$ as

$$v(t) \equiv \inf\{u \geq 0 : E(t+u) - E(t) + A_t(u) \geq Q(t)\}, \quad t \geq 0, \tag{2.35}$$

where $E(t)$ is the amount of fluid to enter service in the interval $[0,t]$, as in (2.8), i.e., $E(t) \equiv \int_0^t b(u,0)\, du$, $t \geq 0$. However, in general, so far, we have not assumed enough to guarantee that the PWT v is finite. It is possible for fluid to arrive and never be served; we need to rule that out.

First, we show that any initial fluid content in the system eventually must leave. Let $B_0(t)$ be the portion of the initial fluid content in service, $B(0)$, that is still in service at time t; let $Q_0(t)$ be the portion of the initial fluid content in queue, $Q(0)$, that is still in queue at time t.

Proposition 2.8 (dissipation of initial fluid content) *For* $t \geq 0$,

$$\begin{aligned} B_0(t) &= \int_t^{\infty} b(0,y) \frac{\bar{G}(t+y)}{\bar{G}(y)}\, dy \to 0 \quad \textit{and} \\ Q_0(t) &\leq \tilde{Q}(0) = \int_t^{\infty} \tilde{q}(0,y) \frac{\bar{F}(t+y)}{\bar{F}(y)}\, dy \to 0 \quad \textit{as} \quad t \to \infty. \end{aligned}$$

Proof. The representation is immediate. It is elementary that $B_0(t) \leq B(0)$ and $\tilde{Q}_0(t) \leq \tilde{Q}(0) = Q(0)$. By Assumption 2.1, $B(0) < \infty$ and $Q(0) < \infty$. The convergence then follows from the Lebesgue dominated convergence theorem. ∎

However, the queue will not dissipate in finite time by abandonment alone, because $\bar{F}(x) > 0$ for all $x \geq 0$. Hence we need to have fluid enter service from the queue. Even if we invoke Assumption 2.9, and have $w(0) < \infty$, so that we have $w(t) \leq w(0) + t < \infty$ for all $t \geq 0$, we cannot guarantee that $v(0) < \infty$. Indeed, we would have $v(t) = \infty$ for all $t \geq 0$ if no fluid from queue were ever admitted into service. That in turn would be the case if we used the feasible staffing function $s(t) \equiv B_0(t)$, which is positive for all t when $B(0) > 0$, because $\bar{G}(x) > 0$ for all $x \geq 0$. In order to avoid such problems, we introduce two more regularity conditions:

Assumption 2.11 (*minimum staffing level*) *There exists a constant s_L such that $s(t) \geq s_L > 0$ for all $t \geq 0$.*

Assumption 2.12 (*minimum service hazard rate*) *There exists a constant $h_{G,L}$ such that $h_G(x) \geq h_{G,L} > 0$ for all $x \geq 0$.*

Theorem 2.4 (finite PWT) *Under Assumptions* 2.11 *and* 2.12*, the rate of service completion is bounded below: $\sigma(t) \geq s_L h_{G,L}$ for all $t \geq 0$. As a consequence,*

$$v(t) \leq \frac{Q(t) + s(t) - s_L}{s_L h_{G,L}} < \infty, \quad t \geq 0.$$

We give the proof in Appendix §A.4. Given that the PWT v is indeed bounded above as in Theorem 2.4, we can obtain it from our algorithm for w. The idea is simple: If, at time t, the elapsed waiting time of the quantum of fluid that is entering service is $w(t)$, then

this quantum of fluid arrived in queue $w(t)$ units of time ago. That implies that the PWT at $t - w(t)$ is $w(t)$. We prove the following in §2.11.

Theorem 2.5 (the PWT v and the BWT w) *Consider an overloaded interval with Assumptions* 2.9-2.10 *holding and* $w(0) = 0$. *If* $v(t) < \infty$ *for all* $t \geq 0$ (*for which Assumption* 2.11 *is a sufficient condition, by Theorem 2.4*), *then* v *is the unique function in* $\mathbb{D}$ *satisfying the equation*

$$v(t - w(t)) = w(t) \quad \textit{or, equivalently,} \quad v(t) = w(t + v(t)) \quad \textit{for all} \quad t \geq 0, \tag{2.36}$$

as depicted in Figure 2.5. *Moreover,* v *is discontinuous at* t *if and only if there exists* $\epsilon > 0$ *such that* $w(t + v(t) + \epsilon) = w(t + v(t)) + \epsilon$, *which in turn holds if and only if* $b(u, 0) = 0$ *for* $t + v(t) \leq u \leq t + v(t) + \epsilon$. *If* $b(\cdot, 0) > 0$ *a.e. with respect to Lebesgue measure, then* v *is continuous.*

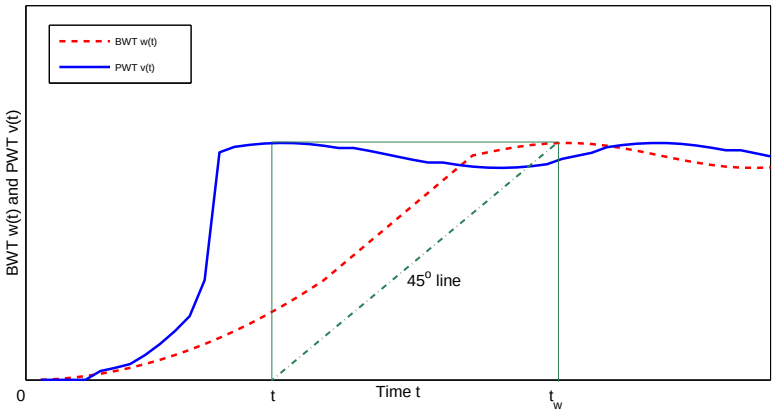

Figure 2.5: Potential waiting time $v(t)$ and boundary waiting time $w(t)$.

The proof of Theorem 2.5 directly gives an algorithm to compute the PWT v given the BWT w. Similarly, the second equation in (2.36) can provide an algorithm to construct

w given v. We now provide an alternative characterization of v via its own ODE, but this alternative characterization involves an extra condition. We give the proof in §2.11.

Theorem 2.6 (right derivative and ODE for v) *Under the conditions in Theorem* 2.5, *the right derivative of v always exists (except possibly infinite), with value*

$$\begin{aligned} v'(t+) &\equiv \lim_{\delta \downarrow 0} \frac{v(t+\delta) - v(t)}{\delta} = \Phi(t, v(t)) \equiv \frac{\tilde{q}(t + v(t), v(t)-)}{b((t + v(t))+, 0)} - 1 \\ &= \frac{\lambda(t+)\bar{F}(v(t))}{b((t + v(t))+, 0)} - 1 \geq -1. \end{aligned}$$

The right derivative at t is finite if and only if $b(t + v(t), 0) > 0$. If t is a continuity point of v, then the left derivative exists as well, with

$$v'(t-) = \tilde{\Phi}(t, v(t)) \equiv \frac{\tilde{q}(t + v(t), v(t)+)}{b((t + v(t))-, 0)} - 1 = \frac{\lambda(t-)\bar{F}(v(t))}{b((t + v(t))-, 0)} - 1 \geq -1.$$

If Φ is continuous at t, then v is differentiable at t, and v satisfies the first ODE. If, in addition, $b(t, 0) > 0$ for all t, then v is continuous. Then v is differentiable except at only finitely many t, and there exists a unique solution to the first ODE.

Remark 2.7 (algorithm for v and w) *In an algorithm, it is convenient to avoid the complications for w and v that occur when $b(t, 0) = 0$. To do so, we can introduce an ϵ-approximation, letting $b_\epsilon(t, 0) \equiv b(t, 0) + \epsilon$, $0 \leq t \leq T$, only to be used in the calculation of w and v. Let w_ϵ be w and v_ϵ be v with $b(t, 0)$ replaced by $b_\epsilon(t, 0)$. Since $w' \geq w'_\epsilon$ and $v' \geq v'_\epsilon$, we have $w_\epsilon \uparrow w$ and $v_\epsilon \uparrow v$ as $\epsilon \downarrow 0$.*

We could also enforce a lower bound for $b(t, 0)$ directly in our model by imposing a

constraint on our staffing. We could require that $b(t,0) \geq b^* > 0$ *for all* t *in order for the staffing function* s *to be feasible. Since* $b(t,0) = s'(t) + \sigma(t)$*, that translates into the staffing constraint*

$$s'(t) \geq b^* - \sigma(t) = b^* - \int_0^\infty b(t,x)\,dx, \quad 0 \leq t < T. \tag{2.37}$$

In Appendix A.4 we give closed-form formulae for the PWT v in some special cases, paralleling those for the BWT w given in Corollary 2.3.

2.8 Overview of the Total Algorithm

We now summarize the full algorithm for the $G_t/GI/s_t + GI$ fluid model. We alternately consider successive underloaded and overloaded intervals (under the assumption that any finite interval can be partitioned into finitely many of these, which can be verified in the computation). For each underloaded interval, we start with initial conditions as indicated in §2.3. We can compute the single key performance measure b directly by applying Proposition 2.2. We then end the underloaded interval the first time $B(t)$ exceeds $s(t)$. Since the queue is empty, the functions q, w and v do not appear.

2.8.1 An Overloaded Interval with M service

An overloaded interval is more complicated. There are two cases: (i) M service and (ii) non-M GI service. For M service, we do not need to solve the fixed point equation (4.20) for the rate fluid enters service from the queue, $b(t,0)$. With M service (at rate 1), we know that $b(t,0) = s'(t) + s(t)$, by Proposition 2.5. The algorithm starts with initial conditions as in §2.3. The algorithm begins by calculating $\tilde{q}$ via Proposition 2.6 and b and

$b(t,0)$ via Proposition 2.5. We then calculate w by solving the ODE (2.31) and then the function v via the equation (2.36), as explained in the proof of Theorem 2.5. We consider terminating the overloaded interval the first time that $w(t) = 0$. At that time we check to see if the interval actually remains overloaded, by looking at the net flow rate into the queue $r(t) \equiv \lambda(t) - s'(t) - \sigma(t)$ (see (3.3)). If $r(t) > 0$, then we continue the overloaded interval. Otherwise, we shift to the next underloaded interval.

2.8.2 An Overloaded Interval with GI service

With non-M service, we need to solve the fixed point equation (FPE) (4.20) for the rate fluid enters service from the queue, $b(t,0)$, in addition to the other steps with M service. We now formally state the algorithm to compute all performance functions in an overloaded interval of the $G_t/GI/s_t + GI$ fluid model. Consider an interval $[0,T]$ and assume that the system is overloaded at $t = 0$, i.e., $Q(0) > 0$ and $B(0) = s(0)$. However, we typically do not know when the overloaded interval ends in advance. The objective is to determine the overload termination time T_1 defined in (3.3) with $t_1 = 0$ along with the other performance functions. Hence, we determine $q(t,\cdot)$ and $b(t,\cdot)$ for $0 \le t \le T \wedge T_1$. If $T_1 < T$, the system simply switches to an underloaded interval; otherwise, the system stays overloaded in $[0,T]$.

Since the system performance is expressed via the basic density vector $\hat{\mathcal{P}}(t) \equiv (b(t,\cdot), q(t,\cdot))$ given the model data vector $\mathcal{D} \equiv \left(\lambda, s, \mu, F, \hat{\mathcal{P}}(0)\right)$, we want to compute the associated vector of all performance functions

$$\mathcal{P}(t) \equiv \left(\hat{\mathcal{P}}(t), w(t), v(t), B(t), Q(t), X(t), \sigma(t), S(t), \alpha(t), A(t), E(t)\right) \qquad (2.38)$$

via the definitions in §2.3. We require that $\mathcal{D}$ satisfies (i) $s(0) = B(0) = \int_0^\infty b(0,y)dy$ and

(ii) $Q(0) = \int_0^{w(0)} q(0,y)dy > 0$. Applying the fixed-point operator discussed in §2.6, we have the following algorithm:

Algorithm 1 : An FPE based algorithm for the $G_t/GI/s_t + GI$ Fluid Queue, with input $\mathcal{D} \equiv \left(\lambda, s, G, F, \hat{\mathcal{P}}(0)\right)$

1: Initialization: Update $\mathcal{R}$, let $t := 0$
2: **repeat**
3: **for** $k = 1, 2, \ldots, \lceil \frac{T-t}{\Delta T} \rceil$ **do**
4: **if** $\mathcal{R}$ = UL **then**
5: Compute $\mathcal{P}$ in interval $[t + (k-1)\Delta T, t + k\Delta T]$, using Proposition 2.2
6: **else**
7: $b^{(0)}(t, 0) := 0$ for $t \in [t + (k-1)\Delta T, t + k\Delta T]$
8: **for** $i = 1, 2, \ldots$ **do**
9: $b^{(i)} := \mathcal{T}(b^{(i-1)})$ for $\mathcal{T}$ defined in (2.21)
10: **if** $\|b^{(i)} - b^{(i-1)}\|_T < \epsilon$ **then**
11: $b := b^{(i)}$
12: BREAK inner for-loop
13: **end if**
14: **end for**
15: Compute $\mathcal{P}$ in interval $[t + (k-1)\Delta T, t + k\Delta T]$, using Proposition 2.6, Corollary 2.2, Theorem 2.3 and 2.6
16: **end if**
17: **if** $T_{\mathcal{R}}(t) < t + k\,\Delta T$ **then**
18: $t := T_{\mathcal{R}}(t)$
19: $\mathcal{R} := \{\text{OL}, \text{UL}\} \backslash \mathcal{R}$
20: BREAK outer for-loop
21: **end if**
22: **end for**
23: **until** $t \geq T$

Note that ϵ is the (small positive) error threshold level that we specify in advance. Here we let the contraction iteration in Step 2 end when the uniform distance between the u functions in two consecutive iterations is small.

The algorithm above requires that the given staffing function s be feasible. However, we can also easily modify the algorithm so that infeasibility can be detected. That extension is discussed in Appendix A.7.

2.9 Feasibility of the Staffing Function

So far, we have assumed that the staffing function s is feasible, yielding

$$b(t,0) \geq s'(t) + \sigma(t) = s'(t) + \int_0^\infty b(t,x) h_G(x)\, dx \geq 0 \tag{2.39}$$

for all $t \geq 0$ such that $B(t) = s(t)$. This requirement is automatically satisfied in underloaded intervals when $B(t) = s(t)$, because in that case we require that $s'(t) + \sigma(t) \geq \lambda(t)$ where necessarily $\lambda(t) \geq 0$. Feasibility is only a concern during overloaded intervals, and then only when the staffing function is decreasing, i.e., when $s'(t) < 0$.

The first violation is easy to detect: Let t^* be the time of first violation. Let I_n be the n^{th} overloaded subinterval in $[0,\infty)$ determined under the assumption that the original staffing function s is feasible. Let I be the union of these subintervals, i.e., the subset of $[0,\infty)$ during which the system is overloaded. Then

$$t^* \equiv \inf \{t \in I : b(t,0) < 0\}. \tag{2.40}$$

Even though we require (3.11), so far we have done nothing to prevent having $t^* < \infty$ (violation). Thus, we compute b and detect the first violation.

Correcting the staffing function is not difficult either (by which we mean replacing it with a higher feasible staffing function): We simply construct a new staffing function s^* consistent with turning off the input into the queue (setting $b(t,0) = 0$) starting at time t^* and lasting until the first time t after t^* at which $s^*(t) = s(t)$. (By the adjustment, we will have made $s^*(t^*+) > s(t^*+)$.) Since the system has operated differently during the time interval $[t^*, t]$, we must recalculate all the performance measures after time t, but we have now determined a feasible staffing function up to time $t > t^*$. By successive applications

of this correction method (adjusting the staffing function s and recalculating b), we can construct the minimum feasible staffing function overall.

To make this precise, let $\mathcal{S}_{f,s}(t)$ be the set of all feasible staffing functions for the system over the time interval $[0,t]$, $t > t^*$, that coincide with s over $[0,t^*]$; i.e., with $C_p^2(t)$ denoting the set of twice differentiable positive real-valued functions on $[0,t]$ with second derivatives in C_p, let

$$\begin{aligned}\mathcal{S}_{f,s}(t) \equiv \{\tilde{s} \in C_p^2(t) : b_{\tilde{s}}(u,0)1_{\{B_{\tilde{s}}(u)=\tilde{s}(u)\}} \geq 0, \quad 0 \leq u \leq t, \\ \text{and} \quad \tilde{s}(u) = s(u), \quad 0 \leq u \leq t^*\},\end{aligned} \tag{2.41}$$

for t^* in (3.12), where $b_{\tilde{s}}$ is the function b associated with the model with staffing function $\tilde{s}$.

Theorem 2.7 (minimum feasible staffing function) *Assume that $s \in \mathbb{C}_p^2$ and $b_{\tilde{s}}(\cdot,0)$ exists and is continuous for each $\tilde{s} \in \mathcal{S}_{f,s}(t))$. Then there exist $\delta > 0$ and $s^* \in \mathcal{S}_{f,s}(t^*+\delta)$ in* (B.13) *for t^* in* (3.12) *such that*

$$s^* = \inf\{\tilde{s} \in \mathcal{S}_{f,s}(t^*+\delta)\}; \tag{2.42}$$

i.e., $s^* \in \mathcal{S}_{f,s}(t^*+\delta)$ and $s^*(u) \leq \tilde{s}(u)$, $0 \leq u \leq t^*+\delta$, for all $\tilde{s} \in \mathcal{S}_{f,s}(t^*+\delta)$. In particular,

$$s^*(t^*+u) \equiv \int_u^\infty b_s(t^*, x-u)\frac{\bar{G}(x)}{\bar{G}(x-u)}dx, \quad 0 \leq u \leq \delta. \tag{2.43}$$

Moreover, δ can be chosen so that

$$\delta = \inf\{u \geq 0 : s^*(t^* + u) = s(t^* + u)\}, \tag{2.44}$$

with $\delta \equiv \infty$ if the infimum in (B.16) *is not attained.*

Proof First, since $b_s(\cdot, 0)$ is continuous for our original s, the violation in (3.12) must persist for a positive interval after t^*; that ensures that a strictly positive δ can be found.

We shall prove that $\tilde{s} \geq s^*$ over $[t^*, t^* + \delta]$ for s^* in (3.16) and any feasible function $\tilde{s}$, and we will show that s^* itself is feasible. For $0 \leq t \leq t^* + \delta$, suppose $\tilde{s}$ is feasible. Since the system is overloaded, system being in the overloaded regime implies that

$$\begin{aligned}
\tilde{s}(t^* + u) &= B_{\tilde{s}}(t^* + u) = \int_0^\infty b_{\tilde{s}}(t^* + u, x)dx \\
&= \int_0^u b_{\tilde{s}}(t^* + u - x, 0)\,\bar{G}(x)dx + \int_u^\infty b_s(t^*, x - u)\frac{\bar{G}(x)}{\bar{G}(x-u)}dx \\
&\geq \int_u^\infty b_s(t^*, x - u)\frac{\bar{G}(x)}{\bar{G}(x-u)}dx = s^*(t^* + u),
\end{aligned}$$

where equality on the second line holds because of the fundamental evolution equations in Assumption 2.6 and because $b_{\tilde{s}}(t^*, x) = b_s(t^*, x)$ for all x, and the inequality holds because $b_{\tilde{s}} \geq 0$. On the other hand, the equality holds when $b_{\tilde{s}}(t^* + u, 0) = 0$ for all u, which yields $B(t^* + u) = s^*(t + u)$. Therefore, the proof is complete. ∎

Corollary 2.4 (minimum feasible staffing with exponential service times) *For the special case of exponential service times, i.e., with $\bar{G}(x) \equiv e^{-x}$,* (3.16) *becomes simply $s^*(t^*+u) = B(t^*)e^{-u}$, $0 \leq u \leq \delta$.*

We have constructed a minimal feasible staffing function by requiring that the new

staffing function agree with the original one up until the time of the first violation. We have shown that assumption leads to a unique minimum feasible staffing function. However, it may be desirable to consider other approaches to feasibility, where we have the freedom to revise the staffing function before t^* as well as afterwards. It is natural to frame the issue as an optimization problem; e.g., as in productions smoothing, we might want to impose costs for for fluctuations of the staffing function as well high values. We leave such investigations for future work.

2.10 Staffing the $G_t/GI/s_t+GI$ Model to Stabilize Delays

So far, we have discussed the performance analysis of the $G_t/GI/s_t + GI$ fluid model with the staffing function s regarded as a given function. In this section, we assume that we are free to choose the staffing function s, and do so with the objective of stabilizing the potential waiting time v at some (constant) target $v^* > 0$. This delay stabilization problem is a variant of one considered previously for many-server queueing models with time-varying arrival rates in [17]. In [17], the goal was to stabilize the probability an arrival experiences any delay. in contrast, here we stabilize the delay of all fluid at precisely $v^* > 0$. Now everybody must wait, but only v^*.

As a consequence of Theorem 2.5, we see that, in order to stabilize v at v^*, it suffices to stabilize w at v^*. By Theorem 2.3, we see that we will be able to do so if and only if we can find a staffing functions s for which the resulting performance satisfies the equation

$$0 = w'(t) = 1 - \frac{b(t,0)}{q(t,v^*)}, \quad t \geq 0 \tag{2.45}$$

which implies that we must have $b(t,0) = q(t,v^*)$ when $w(t) = v^*$.

Suppose that the system is initially empty, i.e., $b(0,x) = q(0,x) = 0$ for all $x > 0$.

Thus, we do not start staffing the service facility until time v^*, so that no input enters service during $[0, v^*]$; i.e., we let $b(t, 0) = 0$ for $0 \le t \le v^*$, in order to let w increase from 0 to v^*. At time v^*, the input at time 0 is sent to the queue, after waiting precisely time v^*.

With the initial conditions $q(t, 0) = \lambda(t)$ and $q(0, x) = 0$, the queue instantly becomes overloaded at time 0, and we can apply Proposition 2.6 and Corollary 2.2 (or (2.5)) to obtain

$$q(t, x) = \bar{F}(x)\lambda(t - x)1_{\{0 \le x \le t\}}, \quad 0 \le t \le v^*. \tag{2.46}$$

Combining (2.45) and (2.46), we obtain the transportation rate after $t = v^*$:

$$b(t, 0) = q(t, v^*) = \bar{F}(v^*)\lambda(t - v^*)1_{\{t > v^*\}}.$$

With the explicit expression of $b(t, 0)$ and $b(0, x) \equiv 0$, $x \ge 0$, (2.5) implies that

$$b(t, x) = \bar{G}(x)\bar{F}(v^*)\lambda(t - x - v^*)1_{\{0 \le x \le t - v^*\}}, \quad t \ge 0 \quad \text{and} \quad x \ge 0. \tag{2.47}$$

Therefore, we can easily compute $B(t)$, $\sigma(t)$, $q(t, x)$, $Q(t)$ and $\alpha(t)$ for $t > v^*$. We have just proved the following theorem.

Theorem 2.8 *Consider the $G_t/GI/s_t+GI$ fluid model with a general arrival-rate function λ. Suppose the system is initially empty. For any specified constant $v^* > 0$, we can make the system overloaded such that the PWT is fixed at v^*, i.e., $v(t) = v^*$ for all $t \ge 0$, by (i) not allowing any input to enter service until time $t = v^*$, (ii) letting the service-capacity*

function be

$$s(v^*,t) \equiv s^*(t) = \bar{F}(v^*) \int_0^{t-v^*} \bar{G}(x)\lambda(t-v^*-x)dx \cdot 1_{\{t>v^*\}} \tag{2.48}$$

and (iii) *operating the queue in the usual FCFS manner after time* v^* *with* $b(t,0) > 0$*. If we do so, then* $w(t) = v^*$ *for* $t \geq v^*$ *and* $w(t) = t$ *for* $t \leq v^*$*,*

$$\begin{aligned}
B(t) &= s^*(t), \quad b(t,0) = \bar{F}(v^*)\lambda(t-v^*) \cdot 1_{\{t>v^*\}}, \\
Q(t) &= \int_0^t \bar{F}(x)\lambda(t-x)dx \cdot 1_{\{0\leq t\leq v^*\}} + \int_0^{v^*} \bar{F}(x)\lambda(t-x)dx \cdot 1_{\{t>v^*\}}, \\
\sigma(t) &= \bar{F}(v^*) \int_0^{t-v^*} \lambda(t-v^*-x)g(x)dx \cdot 1_{\{t>v^*\}}, \\
\alpha(t) &= \int_0^t \lambda(t-x)f(x)dx \cdot 1_{\{0\leq t\leq v^*\}} + \int_0^{v^*} \lambda(t-x)f(x)dx \cdot 1_{\{t>v^*\}}, \quad t \geq 0.
\end{aligned}$$

If λ *is a periodic function, then so are* $b(\cdot,x)$*,* $B(\cdot) = s^*(\cdot)$*,* σ*,* $q(\cdot,x)$*,* $Q(\cdot)$ *and* α *after time* v^**, with the same period.*

Remark 2.8 $($*connection to the QED regime when* $v^* = 0$ $)$ *All the analysis in this section can be extended to the delay target* $v^* = 0$*. In this case, the staffing function in Theorem 2.8 is just sufficient to guarantee that all fluid enters service immediately upon arrival (thus with 0 delay in the queue) and that the system is CL for all* t *(the service capacity is fully occupied, i.e.,* $B(t) = s(t)$*). This scenario corresponds to the heavy-traffic QED system regime.*

Remark 2.9 $($general initial conditions or no delay$)$ *Theorem 2.8 is based on starting empty. However, it is possible to stabilize delays with arbitrary initial conditions. We*

present the details in Appendix A.8. We can also achieve the minimum staffing level so that there is no delay at all by simply staffing at the fluid content $B(t)$ in the underloaded regime. These two variants may involve having an atom of initial fluid content enter service at time 0, so that we leave the smooth framework.

2.11 Proofs of the Main Results

Proof of Theorem 2.3. We establish the different results in turn:

(a) (rate of growth) Consider an interval $[t, t+\delta]$ that is overloaded. If no fluid enters service during this interval, i.e., if $b(s, 0)$=0 for $t \le s \le t+\delta$, then the waiting time of a quantum of fluid at the front of the queue will increase with rate 1, i.e., $w(t+\delta) = w(t)+\delta$, provided that quantum does not abandon. Hence, we have the claimed bound on the rate of growth: $w(t+u) \le w(t)+u$ for all $t \ge 0$ and $u \ge 0$ with $t+u \le T$. A more formal argument follows from (2.5) in Assumption 2.6.

(b) (characterization) However, we will have $w(t+\delta) < w(t)+\delta$ if $b(t, 0) > 0$ because the FCFS service discipline implies that the queue is being eaten away from the head. In other words, fluid is being transported from the queue to the service facility from the right boundary of $q(t, x)$. Therefore,

$$w(t+\delta) = w(t) + \delta - \epsilon(t, \delta), \tag{2.49}$$

where $\epsilon(t, \delta)$ is the amount of boundary waiting time $w(t)$ that is pushed back (eaten up) by $b(t, 0)$ from t to $t+\delta$, see Figure 2.6. (Note that $\delta > 0$ and $\epsilon(t, \delta) \ge 0$.) To determine $\epsilon(t, \delta)$, we apply (2.29), with (2.30). We will bound $\epsilon(t, \delta)$ in (2.51) below.

(c) (controlling the abandonment term) We will show that the abandonment term $A(t, t+$

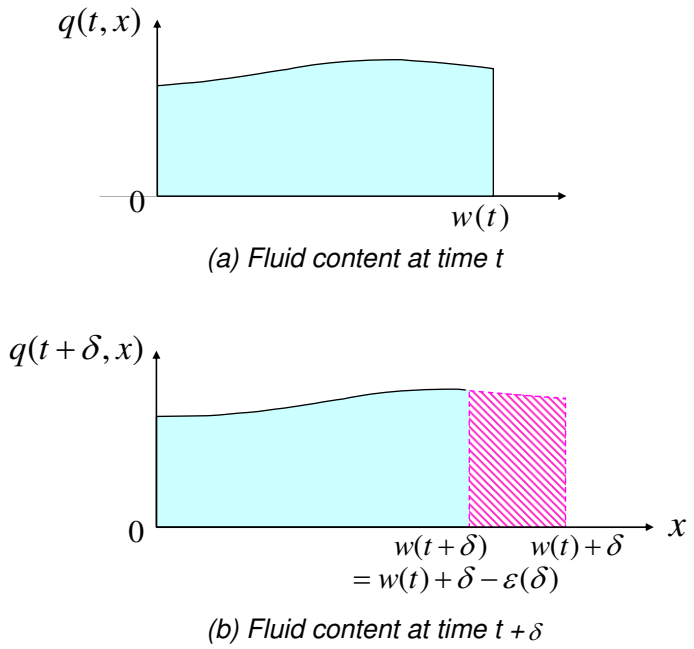

Figure 2.6: The boundary of the waiting time $w(t)$ under FCFS.

$\delta)$ in (2.29) is asymptotically negligible, so that it can be ignored when computing the derivative, but we use it to establish Lipschitz continuity. Even though $A(t, t+\delta)$ is somewhat complicated, we can easily bound it above. Moreover, we can do so uniformly in t over the entire interval $[0, T]$. First let $w^{\uparrow} \equiv \sup\{w(t) : 0 \le t \le T\}$. We necessarily have $w^{\uparrow} \le w(0) + T < \infty$ by virtue of the bound on the growth rate growth determined above. Next let $h_F^{\uparrow} \equiv \sup\{h_F(x) : 0 \le x \le w^{\uparrow}\}$ which necessarily is finite, since $f \in \mathbb{C}_p$ and $\bar{F}(w^{\uparrow}) > 0$; and let $\tilde{q}^{\uparrow} \equiv \sup\{\tilde{q}(t, x) : 0 \le x \le w^{\uparrow}\}$, which again necessarily is finite because $\tilde{q}(t, \cdot) \in \mathbb{C}_p$. We thus have the bound

$$A(t, t+\delta) \le h_F^{\uparrow} \tilde{q}^{\uparrow} w^{\uparrow} \delta = C_1 \delta \tag{2.50}$$

for $0 \le t \le t + \delta \le T$, where $C_1 \equiv h_F^{\uparrow} \tilde{q}^{\uparrow} w^{\uparrow}$.

(d) (Lipschitz continuity) By (2.49), we can show that w is Lipschitz continuous by

showing that $\epsilon(t,\delta) \leq C\delta$ for some constant C. Recall that $b(\cdot,0)$ is continuous by Theorem A.3. Hence, $\|b(\cdot,0)\|_T < \infty$, so that there exists a constant C_2 such that $E(t+\delta) - E(t) \leq C_2\delta$ for $0 \leq t \leq t+\delta \leq T$. Together with (2.50), that implies that the integral $I(t, w(t), \tilde{q}, \delta)$ is bounded above by $C\delta$ for $0 \leq t \leq t+\delta \leq T$, where $C \equiv C_1 + C_2$. Since the integrand of I is bounded below by $c > 0$ by virtue of Assumption 2.10,

$$c\epsilon(t,\delta) \leq I(t, w(t), \tilde{q}, \delta) \leq (E(t+\delta) - E(t)) + A(t, t+\delta) \leq C\delta \tag{2.51}$$

for $0 \leq t \leq t+\delta \leq T$, so that indeed

$$|w(t+\delta) - w(t)| \leq \delta + \epsilon(t,\delta) \leq (1 + (C/c))\delta \quad \text{for} \quad 0 \leq t \leq t+\delta \leq T.$$

as claimed.

(e) (the derivative) Since w is Lipschitz continuous, w necessarily is differentiable a.e., but we will establish a stronger result. Given that $\epsilon(t,\delta) = c\delta + o(\delta)$ as $\delta \downarrow 0$, from the first inequality in (2.50) we see that $A(t, t+\delta) = O(\delta^2) + o(\delta^2)$, so that the abandonment term can be ignored when we consider the derivative. Together with (2.29) and (2.30), that implies that a right derivative of w exists at t with value in (2.31). The convergence as $\delta \downarrow 0$ in the definition of that right derivative will be uniform over a neighborhood of t if $\tilde{q}(t,x)$ is continuous function of x at $x = w(t)$, but not otherwise.

To show (2.32) is similar. We consider an interval $[t-\delta, t]$ that is overloaded. Similarly, we have

$$w(t) = w(t-\delta) + \delta - \epsilon(t-\delta, \delta), \tag{2.52}$$

and

$$E(t) - E(t-\delta) \equiv \int_{t-\delta}^{t} b(u,0)\,du = J + K - A(t, t+\delta)),$$

where

$$J \equiv J(t, w(t), \tilde{q}) \equiv \int_{w(t)}^{w(t)+\epsilon(t-\delta,\delta)} \tilde{q}(t,x)dx, \tag{2.53}$$

and

$$\begin{aligned} K \equiv K(t, w(t), \tilde{q}) &\equiv I(t-\delta, w(t-\delta), \tilde{q}, \delta) - J(t, w(t), \tilde{q}) \\ &= \int_{w(t-\delta)-\epsilon(t-\delta,\delta)}^{w(t-\delta)} \tilde{q}(t-\delta, x)dx - \int_{w(t)}^{w(t)+\epsilon(t-\delta,\delta)} \tilde{q}(t,x)dx. \end{aligned}$$

A closer look at K implies

$$\begin{aligned} K &= \int_{w(t)-\delta}^{w(t)+\epsilon(t-\delta,\delta)-\delta} \tilde{q}(t-\delta,x)dx - \int_{w(t)}^{w(t)+\epsilon(t-\delta,\delta)} \tilde{q}(t-\delta, x-\delta)\frac{\bar{F}(x)}{\bar{F}(x-\delta)}dx \\ &= \int_{w(t)-\delta}^{w(t)+\epsilon(t-\delta,\delta)-\delta} \tilde{q}(t-\delta,x)dx - \int_{w(t)-\delta}^{w(t)+\epsilon(t-\delta,\delta)-\delta} \tilde{q}(t-\delta, y)\frac{\bar{F}(y+\delta)}{\bar{F}(y)}dy \\ &= \int_{w(t)-\delta}^{w(t)+\epsilon(t-\delta,\delta)-\delta} \tilde{q}(t-\delta,y)\left(1 - \frac{\bar{F}(y+\delta)}{\bar{F}(y)}\right)dy, \end{aligned}$$

where the first equality follows from (2.52) and fundamental evolution equations, the second equality holds by change of variable. It is easy to see that $K = o(\delta)$ as $\delta \downarrow 0$. Therefore, together with (2.53), that implies that a left derivative of w exists at t with value in (2.32).

The stronger differentiability conclusion depends on the discontinuities of $\tilde{q}(t,x)$. From Proposition 2.6, all discontinuity points lie on finitely many 45 degree lines in the upper right quadrant $[0,\infty) \times [0,\infty)$; i.e., in the set $\{(t,x) : x = t + c \quad \text{and} \quad c \in \mathcal{S}\}$ where $\mathcal{S}$ contains $c = 0$ and the finite set of discontinuities of λ for $c < 0$ and the finite subset of discontinuities of $q(0,\cdot)$ for $c > 0$. Since $w(t+u) \leq w(t) + u$ for $0 \leq t \leq t+u \leq T$, the trajectory of $\tilde{q}(t, w(t))$ crosses over each of these lines at most once. Moreover, it stays

on each line for at most a finite interval. If the trajectory immediately crosses over the line, then the crossing time t constitutes the sole discontinuity point for w' associated with that line. If the trajectory stays on the line for an interval, then the two endpoints constitute discontinuity points for w' associated with that line.

(f) (existence of a solution) The solution can be constructed by considering the successive intervals between discontinuity points and piecing together the solutions. The function Ψ in (2.31) is continuous in each continuity interval. Hence, existence follows from Peano's theorem; see §2.6 of [69]. We apply Assumption 2.9 to ensure that $w(0) < \infty$.

(g) (uniqueness of a solution) Under extra regularity conditions, the function Ψ in (2.31) will be locally Lipschitz on each continuity interval of w', so that each piece constructed in the existence argument above will be unique, by virtue of the classical Picard-Lindelöf theorem; e.g., Theorem 2.2 of [69]. Specifically, it suffices to assume that λ and $q(0,\cdot)$ (already assumed to be in $\mathbb{C}_p$) are differentiable on the subintervals where they are continuous with derivatives in $\mathbb{C}_p$ over these subintervals.

However, we can actually prove uniqueness without resorting to extra assumptions. To do so, we exploit the special structure of the ODE in (2.31). By (3.15) in Corollary 2.2, $q(t, w(t)-)$ in the denominator or (2.31) takes one of two forms, depending on whether $w(t) \le t$ or not. Our proof applies to both cases in the same way, so we only consider one case: we suppose that $w(t) \le t$. Then $q(t, w(t)-) = \lambda((t - w(t))-)\bar{F}(w(t))$. Then ODE (2.31) implies that

$$\frac{b(t+,0)}{\bar{F}(w(t))} = \lambda((t-w(t))-)(1-w'(t)) = \frac{d}{dt}\left(\int_{t_1}^{t-w(t)} \lambda(y)dy\right),$$

$$\text{so that} \quad \int_{t_1}^{t} \frac{b(y,0)}{\bar{F}(w(y))}dy = \int_{t_1}^{t-w(t)} \lambda(y)dy, \quad t_1 \le t \le t_2. \tag{2.54}$$

Now suppose there is another function $\tilde{w}$ that also satisfies ODE (2.31) with $\tilde{w}(t_1) = 0$.

Then, by the same reasoning, we get

$$\int_{t_1}^{t} \frac{b(y,0)}{\bar{F}(\tilde{w}(y))} dy = \int_{t_1}^{t-\tilde{w}(t)} \lambda(y) dy, \quad t_1 \le t \le t_2. \tag{2.55}$$

Equations (2.54) and (2.55) imply that

$$\int_{t_1}^{t} b(y,0) \left(\frac{1}{\bar{F}(w(y))} - \frac{1}{\bar{F}(\tilde{w}(y))} \right) dy = \int_{t-\tilde{w}(t)}^{t-w(t)} \lambda(y) dy, \quad t_1 \le t \le t_2. \tag{2.56}$$

Now suppose function w and $\tilde{w}$ are different. Since $w(t_1) = \tilde{w}(t_1) = 0$, let $\tilde{t} \equiv \inf\{t > t_1 : w(t) \neq \tilde{w}(t)\}$, which implies that $w'(\tilde{t}) \neq \tilde{w}'(\tilde{t})$. Without loss of generality suppose that $w'(\tilde{t}) < \tilde{w}'(\tilde{t})$, hence there exists a $\delta > 0$ such that $w(t) < \tilde{w}(t)$ for all $\tilde{t} < t \le \tilde{t} + \delta$. Then we have $1/\bar{F}(w(t)) < 1/\bar{F}(\tilde{w}(t))$ for all $\tilde{t} < t \le \tilde{t}+\delta$ and $\tilde{t}+\delta-\tilde{w}(\tilde{t}+\delta) < \tilde{t}+\delta-w(\tilde{t}+\delta)$. Therefore, (2.56) implies that

$$0 > \int_{\tilde{t}}^{\tilde{t}+\delta} b(y,0) \left(\frac{1}{\bar{F}(w(y))} - \frac{1}{\bar{F}(\tilde{w}(y))} \right) dy = \int_{\tilde{t}+\delta-\tilde{w}(\tilde{t}+\delta)}^{\tilde{t}+\delta-w(\tilde{t}+\delta)} \lambda(y) dy > 0,$$

which is a contradiction. Hence the solution to ODE (2.31) must be unique. ∎

Proof of Theorem 2.5. To show that the two equations in (2.36) are equivalent, make the change of variables $s \equiv t - w(t)$. Then the first equation gives $v(s) = w(t) = w(s + w(t)) = w(s + v(s))$, which is the second equation. The other direction is similar.

For a given w, we shall do three things: (i) construct v given the first equation in (2.36), (ii) show that this construction gives a function v that is right continuous and has limits from the left, and (iii) show that the construction in (i) is the unique one that satisfies (ii).

For an arbitrary t, we draw a 45-degree ray starting from point $(t, 0)$: $L(s) = s - t$, $s \ge t$. Let $v(t)$ be the largest t_w such that $L(t_w) = w(t_w)$, as shown in Figure 2.5. We

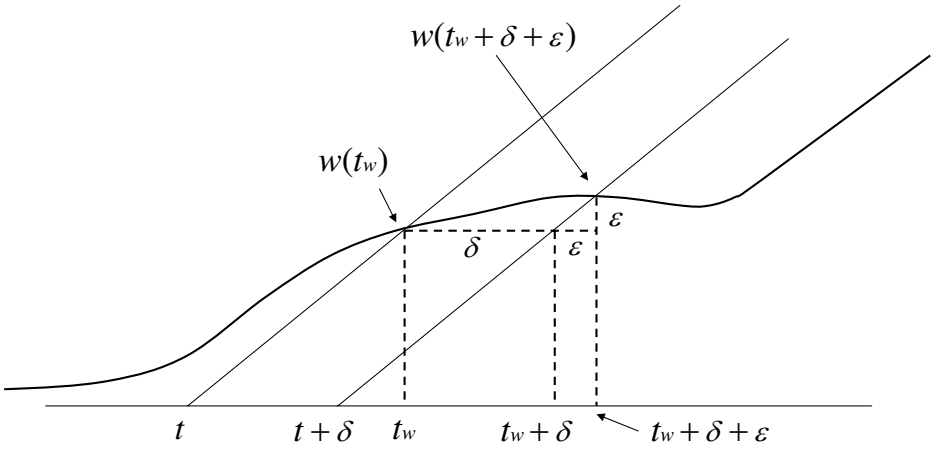

Figure 2.7: Potential waiting time $v(t)$ is right continuous and has limits from the left.

first show that there necessarily exists at least one time $t_w \geq t$ such that $L(t_w) = w(t_w)$. If $w(t) = 0$, then $t_w = t$ is a solution. Otherwise, we have $w(t) > 0 = L(t)$, and w starts above the line L at time t. By Theorem 2.3, w is a continuous function. In general, we could have $w(t) > L(t)$ for all t, but then we would have $v(t) = \infty$. Since $v(t) < \infty$, there necessarily is a time t_w such that $L(t_w) = w(t_w)$.

By Theorem 2.3, $w'(t) \leq 1$. Therefore, once $L(t_w) = w(t_w)$ for the first time, it either stays there or leaves, never to return. In other words, there are two cases: First, as always occurs if $w'(t_w) < 1$, there may be a unique $t_w \geq t$ such that $L(t_w) = w(t_w)$. Second, there may exist an interval $I \equiv [t_1, t_2]$ such that $L(t) = w(t)$ for $t \in I$, i.e., $L(t_1) = w(t_1)$ and $w'(t) = 1$ for $t \in I$; see Figure 2.5. In the first case, we let $v(t) \equiv t_w$; in the second case, we let $v(t) \equiv w(t_w)$ where $t_w \equiv \inf\{s > t_1 : L(s) \neq w(s)\}$. That completes our construction.

Next we show right continuity. For any $\epsilon > 0$, our construction shows that it is possible to choose $\delta > 0$ sufficiently small that $v(t + \delta) = w(t_w + \delta + \epsilon)$ such that $w(t_w + \delta + \epsilon) -$

$w(t_w) = \epsilon$, where $\epsilon \equiv \epsilon(t, \delta)$, as shown in Figure 2.7. Our construction implies that

$$\epsilon = w(t_w + \delta + \epsilon) - w(t_w) = w'(\hat{t})(\delta + \epsilon)$$

for some $t_w \leq \hat{t} \leq t_w + \delta + \epsilon$ and $w'(\hat{t}) < 1$, which implies that

$$\epsilon \equiv \epsilon(t, \delta) = \frac{w'(\hat{t})\,\delta}{1 - w'(\hat{t})} \to 0, \quad as\ \delta \to 0.$$

Therefore, as $\delta \to 0$,

$$v(t + \delta) - v(t) = w(t_w + \delta + \epsilon) - w(t_w) \to 0,$$

by the continuity of w. Therefore, v is right continuous. Similarly, we can show that v has limits from the left.

It is evident that, by this construction, we have ensured that v is right continuous with left limits and unique. Moreover, v is discontinuous at t if and only if we are in the second case with an interval of solutions. ■

Proof of Theorem 2.6. For $\delta > 0$, the second equation in (2.36) yields

$$\begin{aligned} \frac{v(t+\delta) - v(t)}{\delta} &= \left(\frac{w(t + \delta + v(t+\delta)) - w(t + v(t))}{v(t+\delta) - v(t) + \delta}\right)\left(\frac{v(t+\delta) - v(t) + \delta}{\delta}\right) \\ &= \left(\frac{w(t + v(t) + \epsilon(t,\delta)) - w(t + v(t))}{\epsilon(t,\delta)}\right)\left(\frac{v(t+\delta) - v(t)}{\delta} + 1\right), \end{aligned}$$

where $\epsilon(t,\delta) \equiv v(t+\delta) - v(t) + \delta$. Simple algebra implies that

$$\frac{v(t+\delta)-v(t)}{\delta} = \frac{1}{1-\frac{w(t+v(t)+\epsilon(t,\delta))-w(t+v(t))}{\epsilon(t,\delta)}} - 1.$$

Letting $\delta \downarrow 0$, we obtain

$$\begin{aligned} v'(t+) = \lim_{\delta\downarrow 0}\left(\frac{v(t+\delta)-v(t)}{\delta}\right) &= \frac{1}{1-\lim_{\delta\downarrow 0}\left(\frac{w(t+v(t)+\epsilon(t,\delta))-w(t+v(t))}{\epsilon(t,\delta)}\right)} - 1 \\ &= \frac{1}{1-w'((t+v(t))+)} - 1 \\ &= \frac{\tilde{q}(t+v(t), w(t+v(t))-)}{b((t+v(t))+,0)} - 1 \\ &= \frac{\tilde{q}(t+v(t), v(t)-)}{b((t+v(t),0)} - 1 \\ &= \frac{\lambda(t+)\bar{F}(v(t))}{b(t+v(t)+,0)} - 1, \end{aligned}$$

where the second equality holds since right continuity of v implies that $\epsilon(t,\delta) \to 0$ as $\delta \to 0$, the third equality follows from ODE (2.31), the fourth equality follows from the second equation in (2.36), the last equality holds because the system being overloaded at time $t + v(t)$ implies that $\tilde{q}(t+v(t), v(t)) = q(t,0)\bar{F}(v(t)) = \lambda(t)\bar{F}(v(t))$. The similar argument applies to the left derivative with $(v(t) - v(t-\delta))/\delta$ when t is a continuity point of v.

By Theorem 2.5, v is continuous under the extra condition that $b(t,0) > 0$ for all t. That clearly makes the right derivative finite for all t. Hence, v is differentiable wherever Φ is continuous. We can now exploit Theorem 2.3 and its proof. Since $b(t,0) > 0$ for all t, there will be a one-to-one correspondence between the finitely many points where Ψ in (2.31) is discontinuous and the points where Φ is discontinuous. Now we have the relations

(for the right derivatives everywhere)

$$v'(t) = \frac{w'(t + v(t))}{1 - w'(t + v(t))} \quad \text{and} \quad w'(t) = \frac{v'(t - w(t))}{v'(t - w(t)) + 1}, \quad t \geq 0, \tag{2.57}$$

with the denominators positive in both cases. Directly, we can establish existence and uniqueness of a solution to the ODE by the same reasoning as used for ODE (2.31) for w. ■

2.12 Conclusions

In this chapter we have characterized all the standard performance functions for the $G_t/GI/s_t+GI$ fluid model, having time-varying arrival rate and staffing, customer abandonment, and non-exponential service and patience distributions. Our results were obtained under two important regularity conditions: (i) Assumption 2.2, requiring that we have a smooth model, and (ii) Assumption 2.7, requiring that there be only finitely many switches between overloaded (OL) and underloaded (UL) intervals in finite time; see §2.3. There also is a restriction on the service distribution in Assumption 2.8 in order to guarantee that the fixed point equation (4.20) for the rate of flow from queue into service, $b(t, 0)$, has a unique solution that can be computed iteratively. It suffices for either (i) the service hazard function h_G to be bounded or (ii) the system to have started empty at some time in the (finite) past; see §2.6. Still other regularity conditions were imposed in §2.7.

For M service, the relatively simple algorithm primarily requires solving the ODE for the BWT w in Theorem 2.3 and the equation for the PWT v in Theorem 2.5. For non-exponential service, in addition we must solve the fixed point equation (4.20) for the flow rate into service $b(t, 0)$, which is needed to determine the full service content density $b(t, x)$. The algorithm is summarized in §2.8. We characterized the model, as just reviewed, under

the assumption that the staffing function s is feasible, but in Theorem 2.7 we also characterized the minimum feasible staffing function greater than or equal to any given staffing function, provided that it is not changed prior to the first infeasibility time. In §2.10 we showed that we can construct a staffing function to stabilize the potential waiting time v at any desired target $v^* > 0$.

The fluid model is well defined directly, but it is intended to serve as an approximation for large-scale many-server queueing systems. We performed extensive simulation experiments to confirm that the fluid model can provide a useful approximation for such stochastic queueing systems. One of these experiments is described in §2.2; others are described in Appendix A. The simulation results show that, first, the fluid approximation is essentially exact for very large queueing systems and, second, it can be effective as an approximation for mean values even when the scale is not too large; e.g., the number of servers might be only 20. The approximation tends to be more accurate when the system is either overloaded or underloaded, rather than critically loaded, as illustrated by Figure 2.3.

There are many directions for future research. First, it remains to provide conditions with GI service, paralleling our results for M_t service in this chapter, guaranteeing that there are only finitely many switches between OL and UL intervals in finite time, as we assumed in Assumption 2.7. Second, it remains to further explore Assumption 2.8 guaranteeing that the Banach contraction theorem can be applied to establish the existence of a unique service content density b in OL intervals, and develop an effective algorithm for calculating it. Third, it remains to consider alternative approaches to obtaining feasible staffing functions. The method in §2.9, detects any infeasibility of a candidate staffing function and removes the problem by increasing the staffing after the violation point. Alternative methods could modify the entire staffing function, aiming to achieve minimum cost subject to constraints. Fourth, it remains to establish existence, uniqueness and algo-

rithm results for the more general model in which many of the conditions imposed here are relaxed.

As explained in §2.1, there already is strong theoretical support for the fluid model here through previously established MSHT limits. Nevertheless, work is in progress to establish MSHT limits for the smooth fluid model here, paralleling the MSHT limit for the discrete-time model in §6 of [77]. A first goal is to obtain additional theoretical support; a second goal is to obtain a refined stochastic approximation, paralleling the results for Markovian models in [46–48]. It remains to develop alternative approximations and MSHT limits for $G_t/GI/s_t + GI$ systems that tend to be nearly critically loaded at all times, instead of switching back and forth between OL and UL intervals. Finally, it remains to extend the model to represented more complicated service systems with multiple service pools and multiple customer classes. A first step has been made for single-class networks of queues with time-varying Markovian routing in Chapter 3.

Chapter 3

A Network Generalization

We now extend our analysis in Chapter 2 to the case of a single class fluid network with a proportional routing and time-varying model parameters. We provide algorithms to compute time-dependent performance measures for all queues in a finite time interval. The key step of the algorithms is to characterize the total (or aggregated) arrival rates at each queue, which is based on solving a functional fixed-point equation. Computer simulation experiments verify the effectiveness of the approximation.

3.1 Introduction

The main feature of the model is time-varying arrival rates, which commonly occur in applications but which make performance analysis difficult; see [26] for background. The specific model is an open network of time-varying many-server fluid queues with proportional routing. There are m queues, each with its own external fluid input. In addition, a proportion $P_{i,j}(t)$ of the fluid output from queue i at time t is routed immediately to queue j, and a

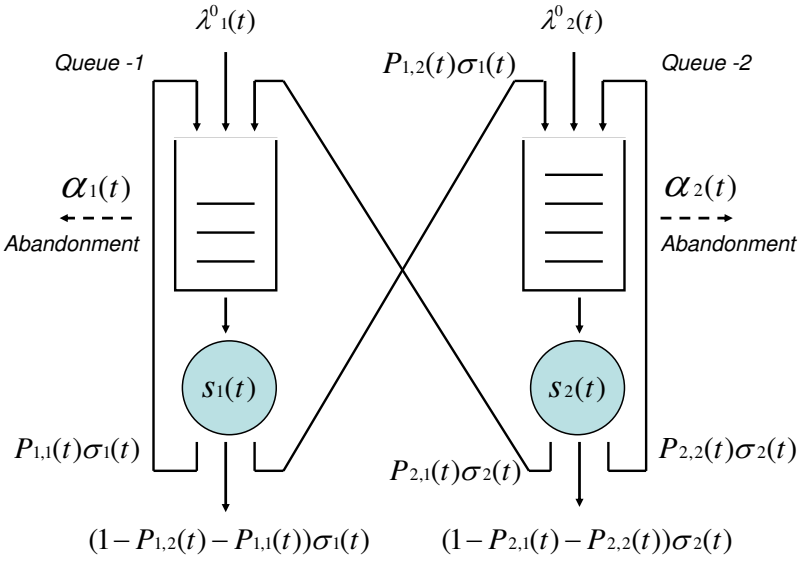

Figure 3.1: The open $(G_t/M_t/s_t + GI_t)^2/M_t$ fluid network.

proportion $P_{i,0}(t) \equiv 1 - \sum_{j=1}^{m} P_{i,j}(t) \leq 1$ is routed out of the network (departs having successfully completed all required service). This framework permits feedback, both directly and indirectly after one or more transitions to other queues, as shown in Figure 3.1 for the case $m = 2$. Following [50], we denote the model by $(G_t/M_t/s_t + GI_t)^m/M_t$, where the subscript t indicates time varying. The fluid model is intended to serve as an approximation for the corresponding many-server queueing system, having m queues, each with a general time-varying arrival process (the G_t), time-varying Markovian service (the first M_t), a time-varying (large) number of servers (the s_t), a general time-varying abandonment-time distribution (the $+GI_t$), and a Markovian routing (the last M_t) among queues. We later extend the M_t service to GI.

This $(G_t/M_t/s_t + GI_t)^m/M_t$ model is a generalization of the classical Jackson open network of queues in the following respects: (i) it allows customer abandonment while Jackson networks do not; (ii) it has time-varying model parameters while Jackson networks do not, and (iii) unlike Jackson networks that assume Poisson arrivals and exponential

service times, here the arrival process need not be renewal or Poisson and service times follow general distributions.

Since the new fluid model is tractable, we are providing the basis for creating a performance-analysis tool for large-scale service systems (allowing many queues and many servers at each queue) like the Queueing Network Analyzer (QNA), described in [73]; also see [8]. Algorithms based on performance formulas are appealing to supplement and complement computer simulation, because the models can be created and solved much more quickly. Thus they can be applied quickly in "what if" studies. They also can be efficiently embedded in optimization algorithms to systematically determine design and control parameters to meet performance objectives.

New methods are required because these large-scale service systems tend to be characterized by *many-server queues*, where a large number of homogeneous servers work in parallel. For a many-server fluid queue with time-varying Markovian service rate $\mu(t)$, when the system content is $X(t)$ and the staffing is $s(t)$, the total service completion rate at time t is $\min\{X(t), s(t)\}\mu(t)$. Unlike in single-server systems, when the many-server system is not overloaded, the service completion rate is *not* equal to the input rate, but is instead *proportional to the system content*, cf. [9].

This chapter extends earlier work. First, [77] described the steady-state fluid content in a stationary $G/GI/s + GI$ fluid model. Second, in Chapter 2 we developed an algorithm for describing the time-dependent behavior of the time-varying $G_t/GI/s_t + GI$ model, including the first full description of the transient behavior of the stationary $G/GI/s + GI$ fluid model. We make several important contributions here: First, for the case of exponential service times we extend the model from a single fluid queue to a network of fluid queues. Second, we treat time-varying service and abandonment. By focusing on M_t service instead of GI service, we are able to establish the existence of a unique (computable) performance description for both one fluid queue and the network generalization without

directly assuming that there are only finitely many switches between overloaded and underloaded intervals in any finite time interval. These results are based on monotonicity and Lipschitz continuity properties of the fluid queue model in §3.5, which are important in their own right. Finally, we characterize the steady state performance of the stationary network of fluid queues.

Here is how the rest of this chapter is organized: In §3.2 we introduce the $G_t/M_t/s_t + GI_t$ model of a single fluid queue. Even though we consider only a single queue there, the time-dependence in the service and abandonment prevents this model from being a special case of the model in Chapter 2. In §3.3 we show how the overloaded and underloaded times occur in alternating intervals of positive length, under regularity conditions, and we introduce a specific piecewise-polynomial framework for assuring that there are only finitely many switches in each finite time interval. In §3.4 we present the performance formulas for one queue. In §3.5 we extend the results to general piecewise-continuous arrival rate functions, thus providing an essential step for extending the analysis to networks. In §3.6 we define the $(G_t/M_t/s_t + GI_t)^m/M_t$ fluid network, that is a network generalization of the single fluid queue introduced in §3.2. In §3.7 We establish the existence of a unique vector of arrival rate functions at each queue and thus the performance in the network. We provide two algorithms to compute the system performance: (i) an algorithm based on solving a fixed-point equation (FPE) and (ii) an algorithm based on solving a multi-dimentional ordinary differential equation (ODE). In §3.8 we make an extension from the M_t service distribution to GI and provide an algorithm based on solving a functional FPE. In §3.9 we evaluate the performance of the algorithms developed in §§3.6 and 3.8 by considering Markovian and non-Markovian examples. In §3.10 we characterize the steady-state performance in the stationary $(G/GI/s + GI)^m/M$ fluid queue network. Finally, in §3.11 we

draw conclusions. In Appendix B we provide (i) some proofs, (ii) some remarks, and (iii) an illustrative comparison with simulation of a large-scale queueing system.

3.2 The $G_t/M_t/s_t + GI_t$ Fluid Queue

We define the $G_t/M_t/s_t + GI_t$ fluid model as an analog of the $G_t/GI/s_t + GI$ model described in §2.3. The notation largely follows §2.3, but some modification is needed. By M_t service, we mean that service is provided at the service facility at time-varying rate $\mu(t)$ per quantum of fluid in the service facility; i.e., if the total fluid content in service at time t is $B(t)$, then the total service completion rate at time t is

$$\sigma(t) \equiv B(t)\mu(t), \quad t \geq 0. \tag{3.1}$$

Let $S(t)$ be the total amount of fluid to complete service in the interval $[0, t]$; then $S(t) \equiv \int_0^t \sigma(y)\, dy$.

Fluid waiting in queue may abandon. Specifically, we assume that a proportion $F_t(x)$ of any fluid to enter the queue at time t will abandon by time $t + x$ if it has not yet entered service, where F_t is an absolutely continuous cumulative distribution function (cdf) for each t, $-\infty < t < +\infty$, with

$$F_t(x) = \int_0^x f_t(y)\, dy, \quad x \geq 0, \quad \text{and} \quad \bar{F}_t(x) \equiv 1 - F_t(x), \quad x \geq 0. \tag{3.2}$$

Let $h_{F_t}(y) \equiv f_t(y)/\bar{F}_t(y)$ be the hazard rate associated with the patience (abandonment) cdf F_t.

Let $\alpha(t)$ be the abandonment rate at time t. Since $q(t, x)$ is the density of fluid in queue

at time t that arrived at time $t - x$, the abandonment rate at time t is

$$\alpha(t) \equiv \int_0^\infty q(t,y) h_{F_{t-y}}(y)\, dy, \quad t \geq 0. \tag{3.3}$$

Let the following quantities be defined as in §2.3: the total input $\Lambda(t)$, staffing $s(t)$, total fluid abandoned $A(t)$, fluid in queue (service) that has been in queue (service) for at most x $B(t,x)$ ($Q(t,x)$), fluid density in queue (service) $q(t,x)$ ($b(t,x)$), and the boundary of waiting time $w(t)$, in the identical way as in §2.3.

Let $E(t)$ be the amount of fluid to enter service in $[0,t]$. We have

$$E(t) \equiv \int_0^t \gamma(u)\, du, \quad t \geq 0, \tag{3.4}$$

where $\gamma(t) \equiv b(t,0)$ is the rate fluid enters service at time t. Clearly, we have the *flow conservation equations*: For each $t \geq 0$,

$$Q(t) = Q(0) + \Lambda(t) - A(t) - E(t) \quad \text{and} \quad B(t) = B(0) + E(t) - S(t). \tag{3.5}$$

The rate fluid enters service depends on whether the system is underloaded or overloaded. If the system is underloaded, then the external input directly enters service; if the system is overloaded, then the fluid to enter service is determined by the rate, $\eta(t)$, that service capacity becomes available at time t. Service capacity becomes available due to service completion and any change in the staffing function. Hence the rate service becomes available is

$$\eta(t) \equiv s'(t) + \sigma(t) = s'(t) + B(t)\mu(t), \quad t \geq 0, \tag{3.6}$$

so that $\eta(t) = s'(t) + s(t)\mu(t)$ if the system is overloaded at time t.

We assume Assumptions 2.1-2.5 are satisfied. In addition, we make the following assumptions.

Since the service discipline is FCFS, fluid leaves the queue to enter service from the right boundary of $q(t,x)$. Since the service is M_t, the proportion of fluid in service at time t that will still be in service at time $t+x$ is

$$\bar{G}_t(x) = e^{-M(t,t+x)} \quad \text{where} \quad M(t,t+x) \equiv \int_t^{t+x} \mu(y)\,dy, \quad t \geq 0 \quad \text{and} \quad x \geq 0. \tag{3.7}$$

Note that G_t coincides with the time-varying service-time cdf of a quantum of fluid that enters service at time t. The cdf G_t has density $g_t(x) = \mu(t+x)\bar{G}_t(x)$ and hazard rate $h_{G_t}(x) = \mu(t+x)$, $x \geq 0$.

Paralleling to Assumption 2.6, we assume that q and b satisfy the following two fundamental evolution equations.

Assumption 3.1 $(\textit{fundamental evolution equations})$ *For $t \geq 0$, $x \geq 0$ and $u \geq 0$,*

$$q(t+u,x+u) = q(t,x)\frac{\bar{F}_{t-x}(x+u)}{\bar{F}_{t-x}(x)}, \quad 0 \leq x < w(t), \tag{3.8}$$

$$b(t+u,x+u) = b(t,x)\frac{\bar{G}_{t-x}(x+u)}{\bar{G}_{t-x}(x)} = b(t,x)e^{-M(t,t+u)}, \tag{3.9}$$

where M is defined in (3.7)*.*

In addition to Assumption 2.2, we have the following assumption for the model date.

Assumption 3.2 $(\textit{smoothness})$ *$s', \lambda, f_t, f_\cdot(x), \mu, b(0,\cdot), q(0,\cdot)$ in $\mathbb{C}_p$ for each x and t.*

As a consequence, $s, \Lambda, F_t, B(0,\cdot), Q(0,\cdot)$ are differentiable functions with derivatives in $\mathbb{C}_p$ for each t; we say that they are elements of $\mathbb{C}_p^1$.

In order to treat the BWT w, we need to impose a regularity condition on the arrival rate function and the initial queue density (when the initial queue content is positive, which never occurs after an underloaded interval). We make the following assumption.

Assumption 3.3 (*positive arrival rate and initial queue density*) *For all* $t \geq 0$,

$$\lambda_{\inf}(t) \equiv \inf_{0 \leq u \leq t} \{\lambda(u)\} > 0 \quad \textit{and} \quad q_{\inf}(0) \equiv \inf_{0 \leq u \leq w(0)} \{q(0,u)\} > 0 \quad \textit{if} \quad w(0) > 0.$$

In order to be sure that the PWT function v is finite, we make two more assumptions.

Assumption 3.4 (*minimum staffing level*) *There exists* s_L *such that* $s(t) \geq s_L > 0$ *for all* $t \geq 0$.

Assumption 3.5 (*minimum service rate*) *There exists* μ_L *such that* $\mu(t) \geq \mu_L > 0$ *for all* $t \geq 0$.

Finally to treat A with the time-varying abandonment cdf F_t, we first introduce bounds for the time-varying pdf f_t and complementary cdf $\bar{F}_t$. Let

$$f^{\uparrow} \equiv \sup \{f_t(x) : x \geq 0, \quad -\infty < t \leq T\} \quad \text{and} \quad \bar{F}^{\downarrow}(x) \equiv \inf \{\bar{F}_t(x) : -\infty \leq t \leq T\}. \tag{3.10}$$

Assumption 3.6 (*controlling the time-varying abandonment distribution*) $f^{\uparrow} < \infty$ *and* $\bar{F}^{\downarrow}(x) > 0$ *for all* $x > 0$, *where* $f^{\uparrow}$ *and* $\bar{F}^{\downarrow}(x)$ *is defined in* (3.10).

In summary, here we have made Assumptions of Chapter 2 (with minor modifications because of M_t service and GI_t abandonment instead of both being GI). We show how to

relax Assumption 2.7 there in the next section. Assumption 3.6 here is new, because of the time-varying abandonment.

3.3 Underloaded and Overloaded Intervals

In Assumption 2.7 of Chapter 2, we directly assumed that the system alternates between underloaded intervals and overloaded intervals, with there being only finitely many switches in any finite interval. In this chapter, we provide conditions under which that assumption can be guaranteed to hold, and then show how to treat the more general case as a limit of such systems. This extension is important to rigorously treat fluid queue networks. This extension is facilitated by having M_t service.

We initially classify the system state as overloaded or underloaded at time t as follows. Recall that the rate service capacity becomes available at time t is $\eta(t) \equiv s'(t) + \sigma(t)$, as in (3.6) above.

Definition 3.1 *The system is **overloaded** if either* $(i) \quad Q(t) > 0$ *or*

$(ii) \quad Q(t) = 0, \quad B(t) = s(t) \quad$ *and* $\quad \lambda(t) > \eta(t) = s'(t) + s(t)\mu(t)$*;*

*the system is **underloaded** if either* $(i) \quad B(t) < s(t)$ *or*

$(ii) \quad B(t) = s(t), \quad Q(t) = 0 \quad$ *and* $\quad \lambda(t) \le \eta(t) = s'(t) + s(t)\mu(t).$

At every time t, the system is thus either overloaded or underloaded.

We now define the set of switch times. For that purpose, let $\mathcal{O}(A)$ ($\mathcal{U}(A)$) be the set of overloaded (underloaded) times t in the subset A of a designated interval $[0, T]$. From Definition 3.1, $\mathcal{U}(A) = A - \mathcal{O}(A)$ for each subset A (the complement relative to A).

Definition 3.2 *The subset $\mathcal{S}$ be of **switch times** in $[0,T]$ is the subset of t for which*

$$\mathcal{U}(((t-\epsilon)\vee 0,(t+\epsilon)\wedge T)) \neq \emptyset \quad \textit{and} \quad \mathcal{O}(((t-\epsilon)\vee 0,(t+\epsilon)\wedge T)) \neq \emptyset \quad \textit{for all} \quad \epsilon > 0 \tag{3.10}$$

To neatly classify the switching times, we further classify some of the underloaded times.

Definition 3.3 *An underloaded time t is **isolated** if (i) either $[0,t)$ or (a,t) is an overloaded interval and (ii) either $(t,T]$ or (t,b) is an overloaded interval.*

We now reclassify all isolated underloaded points as overloaded points. When we reclassify each isolated underloaded point, we replace the two connecting overloaded intervals by the common overloaded interval; e.g., when t is an isolated underloaded time between overloaded intervals (a,t) and (t,b), we replace the two intervals by the single interval (a,b). In Appendix B.1 we show that this procedure is well defined. In the remainder of this section we present the key results allowing us to ensure that $\mathcal{S}$ is finite. We present the proofs in Appendix B.1. Our first structural result is

Theorem 3.1 (partition into intervals) *After all isolated underloaded times have been reclassified as overloaded and all overloaded intervals have been increased as specified above, the interval $[0,T]$ can be partitioned into at most countably many alternating overloaded and underloaded intervals (of positive length). The resulting switch points are the boundary points between overloaded intervals and underloaded intervals.*

Our analysis above has shown how to partition the interval $[0,T]$ into alternating over-

loaded and underloaded intervals of positive length. Then the switch points are clearly identified as the boundary points. It is then convenient to adopt the convention that all intervals be left closed and right open (e.g., of the form $[a, b)$), except at the interval endpoints 0 and T, so that the regime identification function $r(t) \equiv 1_{\{\mathcal{O}([0,T])\}}(t)$, where $1_{\{A\}}$ is the usual indicator function, is right continuous with left limits. This convention does not alter the switch points.

We now relate the subset $\mathcal{S}$ to the set of discontinuity points and the zero set of the function

$$\zeta(t) \equiv \lambda(t) - s'(t) - s(t)\mu(t), \quad t \geq 0. \tag{3.12}$$

Note that ζ depends only on the basic model functions λ, s and μ. Also note that $\zeta = \lambda - \eta$ in the overloaded case of Definition 3.1. Let $\mathcal{D}_\zeta$ be the set of discontinuities of ζ in (3.12) and let $\mathcal{Z}_\zeta \equiv \{t \in [0, T] : \zeta(t) = 0\}$ be the zero set.

Theorem 3.2 (relating switches to zeros and discontinuities of ζ) *For any interval $[0, T]$, the subsets $\mathcal{S}$, $\mathcal{Z}_\zeta$ and $\mathcal{D}_\zeta$ are closed subsets with $|\mathcal{S}| \leq |\mathcal{Z}_\zeta| + |\mathcal{D}_\zeta| - 1$. Moreover, the bound in is tight; i.e., there are examples for which the bound holds as an equality.*

We now introduce a convenient subset of functions in $\mathbb{C}_p$ to represent our model data λ, μ and s'. The class is sufficiently general that it can represent any function in $\mathbb{C}_p$ and, at the same time, it allows us to control the zeros of ζ, so that we know in advance that there are only finitely many switches between overloaded and underloaded intervals in any finite interval.

Let $\mathcal{P}_{m,n} \equiv \mathcal{P}_{T,m,n}$ be the space of *piecewise polynomials* on the interval $[0, T]$, where $[0, T]$ is partitioned into n subintervals, on each of which there is a polynomial of order at most m. We start with three elementary lemmas about $\mathcal{P}_{m,n}$. (We do not require that the overall function be continuous, but each function necessarily is in $\mathbb{C}_p$.) The first lemma

states that any function in C_p can be approximated uniformly by a function from $\mathcal{P}_{m,n}$, so that there is no practical loss of generality to restricting the model data to be in $\mathcal{P}_{m,n}$ instead of $\mathbb{C}_p$.

Lemma 3.1 (uniform approximation) *For any function $h \in \mathbb{C}_p$ over a finite interval $[0, T]$ and any $\epsilon > 0$, there exists a function $\tilde{h} \in \mathcal{P}_{m,n}$ for some positive integers m and n such that $\|h - \tilde{h}\|_T < \epsilon$.*

The second lemma states that we can go back and forth between the functions λ, s', μ and their integrals Λ, s, M in $\mathcal{P}_{m,n}$ conveniently; i.e., the integral or derivative of a polynomial is again a polynomial. In particular, we can analytically calculate the integral for M in definition (3.7), as needed for the fundamental evolution equation for b in (3.9).

Lemma 3.2 (representation of integrals) $\lambda, s', \mu \in \mathcal{P}_{m,n} \subset \mathbb{C}_p$ *if and only if* Λ, $M(t, t+\cdot)$, $M(u-\cdot, u)$, $s \in \mathcal{P}_{m+1,n} \cap C$.

The third lemma states that the function ζ inherits piecewise-polynomial structure assumed for the basic model functions λ, s', μ.

Lemma 3.3 (preservation of piecewise-polynomial structure) *If $\lambda \in \mathcal{P}_{m_1,n_1}$, $s' \in \mathcal{P}_{m_2,n_2}$, and $\mu \in \mathcal{P}_{m_3,n_3}$, then $\zeta \in \mathcal{P}_{m,n}$, where $n \leq n_1 + n_2 + n_3$ and $m \leq m_1 \vee m_2 \vee m_3(m_2+1)$.*

The following theorem serves as the basis for our analysis.

Theorem 3.3 (finitely many switches) *If $\zeta \in \mathcal{P}_{m,n}$ for ζ in* (3.12)*, then $|\mathcal{S}| \leq n(m+1)-1$.*

Hence, we can carry out the construction of the desired performance vector $(b, q, w, v, \sigma, \alpha)$ under the assumptions that the basic model functions (λ, s, μ) are such that there are only finitely many switches between overloaded intervals and underloaded intervals in any given

interval $[0, T]$. It suffices to have $\lambda, s', \mu \in \mathcal{P}_{m,n}$ for some m and n. The space $\mathcal{P}_{m,n}$ is useful for the theory, but it should not be needed in applications; see Remark B.3.

3.4 The Performance at One Queue

In this section we determine the performance functions under the assumption that there are only finitely many switches between overloaded and underloaded intervals. We have just seen that a sufficient condition for that is to have $\zeta \in \mathcal{P}_{m,n}$ for some m and n, for which a sufficient condition is to have $\lambda, s', \mu \in \mathcal{P}_{m,n}$ for some m and n. Here we can apply the previous results in §2.3, making proper adjustments to account for the change from GI service and abandonment to M_t service and GI_t abandonment.

An underloaded interval requires modification to account for M_t service. Since the rate fluid enters service is $\gamma(t) = b(t, 0) = \lambda(t)$ when the system is underloaded, we immediately obtain an expression for $b(t, x)$ from (3.9). Recall that we have assumed that $b(0, \cdot) \in \mathbb{C}_p$.

Proposition 3.1 (service content in the underloaded case) *For the fluid model with unlimited service capacity* ($s(t) \equiv \infty$ *for all* $t \geq 0$)*, starting at time* 0*,*

$$\begin{aligned} b(t, x) &= e^{-M(t-x,t)}\lambda(t-x)1_{\{x \leq t\}} + e^{-M(0,t)}b(0, x-t)1_{\{x>t\}}, \\ B(t) &= \int_0^t e^{-M(t-x,t)}\lambda(t-x)\,dx + B(0)e^{-M(0,t)}, \quad 0 \leq t < T, \end{aligned} \tag{3.13}$$

where M *is defined in* (3.7)*. If, instead, a finite-capacity system starts underloaded, then the same formulas apply over the interval* $[0, T)$*, where* $T \equiv \inf\{t \geq 0 : B(t) > s(t)\}$*,*

with $T = \infty$ if the infimum is never obtained. Hence, $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ for all $t \geq 0$ and $x \geq 0$, for t in the underloaded interval.

There is dramatic simplification in going from GI service to M_t service in an overloaded interval. Then we simply have $B(t) = s(t)$. The rate fluid enters service is equal to the rate service capacity becomes available: $\gamma(t) = \eta(t) = s'(t) + s(t)\mu(t)$. For an overloaded interval starting at time 0, we have

Proposition 3.2 $(\text{service content in the overloaded case})$ *For the fluid model in an overloaded interval, $B(t) = s(t)$ and*

$$b(t,x) \;\; = \;\; (s'(t-x) + s(t-x)\mu(t-x))e^{-M(t-x,t)}1_{\{x \leq t\}} + b(0,x-t)e^{-M(0,t)}1_{\{x>t\}},$$

where M is defined in (3.7)*. Hence, $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ for all $t \geq 0$ and $x \geq 0$ in an overloaded interval.*

Corollary 3.1 $(\text{overall smoothness for the service content})$ *If there are only finitely many switches between overloaded and underloaded intervals in $[0,T]$, then $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ for all t, $0 \leq t \leq T$, and $x \geq 0$.*

We treat q, w and v just as in Chapter 2, making adjustments for the time-varying abandonment cdf F_t. Let $\tilde{q}(t,x)$ be $q(t,x)$ during the overload interval $[0,T]$ under the assumption that no fluid enters service from queue. The next proposition is an analog of Proposition 2.6.

Proposition 3.3 $(\textit{queue content without transfer into service in the overloaded case})$ *Dur-*

ing an overloaded interval,

$$\tilde{q}(t,x) = \lambda(t-x)\bar{F}_{t-x}(x)1_{\{x\le t\}} + q(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\}}. \tag{3.14}$$

so that $\tilde{q}(t,\cdot)$ *and* $\tilde{q}(\cdot,x)$ *belong to* $\mathbb{C}_p$ *for each* t *and* x*.*

Since BWT w and PWT v are determined by two ODEs as in Theorem 2.3 and 2.6, we get an expression for q provided that we can find w, as an analog of Corollary 2.2.

Corollary 3.2 $($*from* $\tilde{q}$ *to* $q$$)$ *Given the BWT* w *in an overloaded interval,*

$$\begin{aligned} q(t,x) &= \tilde{q}(t-x,0)\bar{F}_{t-x}(x)1_{\{x\le w(t)\wedge t\}} + \tilde{q}(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\le w(t)\}} \\ &= \lambda(t-x)\bar{F}_{t-x}(x)1_{\{x\le w(t)\wedge t\}} + q(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\le w(t)\}}. \end{aligned} \tag{3.15}$$

Moreover, $q(t,\cdot) \in \mathbb{C}_p$ *for all* $t \ge 0$*.*

Corollary 3.3 $($*end of the overloaded interval*$)$ *We can compute the end of an overloaded interval as* $T \equiv \inf\{t \ge 0 : w(t) = 0 \quad \textit{and} \quad \lambda(t) \le s'(t) + s(t)\mu(t)\}$*.*

Corollary 3.4 $($*smoothness of* $q(t,\cdot)$$)$ *Under the assumptions of Theorem* 2.3*,* q *is given by* (3.15) *with* $q(\cdot,x) \in \mathbb{C}_p$ *for all* x*.* $($*We have already deduced that* $q(t,\cdot) \in \mathbb{C}_p$ *for all* t *in Corollary* 2.2.$)$

The Algorithm for One Queue. We now summarize the algorithm to compute the performance functions in the $G_t/M_t/s_t + GI_t$ model, assuming that there are only finitely many switches in each finite interval. We consider the basic density vector $\hat{\mathcal{P}}(t)$, given model data vector $\mathcal{D}$, and total performance vector $\mathcal{P}(t)$, all defined in §2.8.2 of Chapter 2.

Let $\mathcal{R}(t)$ denote the current system regime at t, i.e., $\mathcal{R}(t)$ =OL or UL. When $\mathcal{R}(t_0)$ =OL, the OL interval ends at the *OL termination time*

$$T_{OL}(t_0) \equiv \inf\{u \geq t_0 : Q(u) = 0 \quad \text{and} \quad \lambda(u) \leq s'(u) + \sigma(u)\}.$$

When $\mathcal{R}(t_0)$ =UL, the UL interval ends at the *UL termination time*

$$T_{UL}(t_0) \equiv \inf\{u \geq t_0 : B(u) = s(u) \quad \text{and} \quad \lambda(u) > s'(u) + \sigma(u)\}.$$

Therefore, the termination time of the current interval

$$T_{\mathcal{R}}(t_0) \equiv T_{OL}(t_0)\mathbf{1}_{\{\mathcal{R}(t_0)=OL\}} + T_{UL}(t_0)\mathbf{1}_{\{\mathcal{R}(t_0)=UL\}}.$$

An algorithm is summarized as below.

Algorithm 2 : A Fluid Algorithm for Single Queues (FASQ) for the $G_t/M_t/s_t + GI_t$ fluid model, with input $\mathcal{D} \equiv \left(\lambda, s, G, F, \hat{\mathcal{P}}(0)\right)$

```
Initialization: Update R, let t := 0
repeat
   for k = 0, 1, ..., ⌈(T−t)/ΔT⌉ do
      Given R, compute P in interval [t + (k − 1)ΔT, t + k ΔT] using Proposition 3.1
      ,3.2 and 3.3, Corollary 3.2, Theorem 2.3 and 2.6
      if T_R(t) < t + k ΔT then
         t := T_R(t)
         R := {OL, UL}\R
         BREAK for-loop
      end if
   end for
until t ≥ T
```

Feasibility of the staffing function. The construction above has been done under the assumption that the staffing function is feasible. As in §2.9 of Chapter 2, the algorithm

can detect violations of feasibility whenever they occur and can then produce the minimum feasible staffing function greater than or equal to the initial proposed staffing function. A violation is easy to detect; it necessarily occurs in an overloaded interval in $\mathcal{O}([0,T])$ at time $t^* \equiv \inf\{t \in \mathcal{O}([0,T]) : \gamma(t) < 0\}$. As in Chapter 2, let $\mathcal{S}_{f,s}$ be the set of feasible staffing functions over the interval $[0,t]$ for $t > t^*$.

Theorem 3.4 (minimum feasible staffing function) *There exist $\delta > 0$ and $s^* \in \mathcal{S}_{f,s}(t^*+\delta)$ such that $s^* = \inf\{\tilde{s} \in \mathcal{S}_{f,s}(t^*+\delta)\}$; i.e., $s^* \in \mathcal{S}_{f,s}(t^*+\delta)$ and $s^*(u) \le \tilde{s}(u)$, $0 \le u \le t^*+\delta$, for all $\tilde{s} \in \mathcal{S}_{f,s}(t^*+\delta)$. In particular,*

$$s^*(t^*+u) \equiv B(t^*) \cdot e^{-M(t^*,t^*+u)}, \quad 0 \le u \le \delta. \tag{3.16}$$

Moreover, δ can be chosen so that $\delta = \inf\{u \ge 0 : s^*(t^*+u) = s(t^*+u)\}$, with $\delta \equiv \infty$ if the infimum is not attained.

Corollary 3.5 (minimum feasible staffing with M service) *For M service, i.e., with exponential service times, so that $\bar{G}(x) \equiv e^{-\mu x}$,* (3.16) *becomes simply $s^*(t^*+u) = B(t^*)e^{-\mu u}$, $0 \le u \le \delta$.*

Theorem 3.4 shows how to construct a new staffing function that (i) agrees with the proposed staffing function s over its interval of feasibility $[0,t^*)$ and (ii) itself is feasible over the longer interval $[0,t^*+\delta)$ for some $\delta > 0$. To construct the minimum feasible staffing function over $[0,T]$, this algorithm may need to be applied several times.

3.5 General Arrival Rate Functions

In the previous two sections we have seen that we can get a nice clean theory if we assume that $\lambda, s', \mu \in \mathcal{P}_{m,n}$. In order to treat open networks of fluid queues, we would want the service completion rate σ, which becomes the part of the input rate at other queues, to be in $\mathcal{P}_{m,n}$ for some m and n as well, but σ does not inherit this property, because $\sigma(t) = B(t)\mu(t)$ and $B(t)$ has a complicated non-polynomial form in underloaded intervals, as shown in (3.13). We do have $\sigma \in \mathbb{C}_p$ by virtue of Corollary 3.1, but we need not have $\sigma \in \mathcal{P}_{m,n}$. Hence, we show how to treat the general case in which initially we only assume that $\lambda \in \mathbb{C}_p$.

We will treat the case of general $\lambda \in \mathbb{C}_p$ as the limit of a sequence of systems with $\lambda \in \mathcal{P}_{m,n}$. In particular, for arbitrary $\lambda \in \mathbb{C}_p$, we can represent it as the limit of a sequence of functions $\{\lambda_k : k \geq 1\}$, where $\lambda_k \in \mathcal{P}_{m_k,n_k}$ and $\lambda_k \geq 0$ for each k, and $\|\lambda_k - \lambda\|_T \to 0$ as $k \to \infty$, with $\|\cdot\|_T$ denoting the uniform norm over $[0, T]$. (Positivity is no problem because of Assumption 3.3 and the uniform convergence.) If we also assume that $s', \mu \in \mathcal{P}_{m,n}$ for some m, n, then we will necessarily have $\zeta_k \in \mathcal{P}_{m_k,n_k}$ for all k, with $m_k < \infty$ and $n_k < \infty$ for all k. We will also have $m_k \to \infty$ and $n_k \to \infty$ as $k \to \infty$ unless $\lambda \in \mathcal{P}_{m,n}$ for some m, n.

In this section we establish results that allow us to treat the case of general arrival rate functions $\lambda \in \mathbb{C}_p$, without requiring that $\lambda \in \mathcal{P}_{m,n}$ and without directly requiring that there be only finitely many switches between overloaded and underloaded intervals in the interval $[0, T]$. To do so, we establish monotonicity and Lipschitz continuity properties, which are of independent interest. We first establish these results assuming that $\zeta \in \mathcal{P}_{m,n}$, and then we show that they extend when we allow arbitrary $\lambda \in \mathbb{C}_p$. We thus start by assuming that $\zeta \in \mathcal{P}_{m,n}$. The proofs of the three theorems in this section are relatively straightforward, but long; they appear in Appendix B.3.

The M_t service allows us to extend the elementary comparison results in Propositions 2.3 and 2.7 of Chapter 2. Recall that order of functions (vectors) is defined as pointwise order for all arguments (coordinates). Let $X(t) \equiv B(t) + Q(t)$ be the total system fluid content. Let subscripts designate the model.

Theorem 3.5 (fundamental comparison theorem) *Consider two $G_t/M_t/s_t{+}GI_t$ fluid models with common staffing function s and service rate function μ. If $\zeta_1, \zeta_2 \in \mathcal{P}_{m,n}$ with $\lambda_1 \leq \lambda_2$, $B_1(0) \leq B_2(0)$, $q_1(0,\cdot) \leq q_2(0,\cdot)$ and $h_{F_{t,1}} \geq h_{F_{t,2}}$, then*

$$(B_1(\cdot), \tilde{q}_1, q_1, Q_1(\cdot), X_1, w_1, v_1, \sigma_1) \leq (B_2(\cdot), \tilde{q}_2, q_2, Q_2(\cdot), X_2, w_2, v_2, \sigma_2). \tag{3.17}$$

In addition to monotonicity, the model has additional basic Lipschitz continuity properties (beyond Proposition B.2).

Theorem 3.6 (more Lipschitz continuity) *Consider a $G_t/M_t/s_t + GI_t$ fluid model with $\lambda, s', \mu \in \mathcal{P}_{m,n}$ for some m, n. Then the functions mapping (i) $(\lambda, B(0))$ in $\mathcal{P}_{m,n} \times \mathbb{R}$ into (B, σ) in $\mathbb{C}_p^2$, (ii) $(\lambda, B(0), Q(0))$ in $\mathcal{P}_{m,n} \times \mathbb{R}^2$ into Q in $\mathbb{C}_p$, and (iii) $(\lambda, X(0))$ in $\mathcal{P}_{m,n} \times \mathbb{R}$ into X in $\mathbb{C}_p$, all over $[0,T]$, are Lipschitz continuous. In particular,*

$$\begin{aligned}
\|B_1 - B_2\|_T &\leq (1 \vee T)(\|\lambda_1 - \lambda_2\|_T \vee |B_1(0) - B_2(0)|), \\
\|\sigma_1 - \sigma_2\|_T &\leq \mu_T^{\uparrow} \|B_1 - B_2\|_T, \\
\|Q_1 - Q_2\|_T &\leq (1 \vee T)(\|\lambda_1 - \lambda_2\|_T \vee |B_1(0) - B_2(0)| \vee |Q_1(0) - Q_2(0)|), \\
\|X_1 - X_2\|_T &\leq 2(1 \vee T)(\|\lambda_1 - \lambda_2\|_T \vee |X_1(0) - X_1(0)|).
\end{aligned} \tag{3.18}$$

If $B_1(0) = B_2(0)$ *and* $Q_1(0) = Q_2(0)$ *(for* Q *and* X*), then*

$$\begin{aligned} \|B_1 - B_2\|_T &\leq T\|\lambda_1 - \lambda_2\|_T, \quad \|Q_1 - Q_2\|_T \leq T\|\lambda_1 - \lambda_2\|_T, \\ \|X_1 - X_2\|_T &\leq 2T\|\lambda_1 - \lambda_2\|_T. \end{aligned} \tag{3.19}$$

As a consequence of Theorems 3.3–3.6, we can regard the case of a general function λ as the limit of a sequence $\{\lambda_k : k \geq 1\}$, where $\zeta_k \in \mathcal{P}_{m_k,n_k}$ with $m_k \to \infty$ and $n_k \to \infty$ as $k \to \infty$. Hence, results for the k^{th} system can be "lifted" to the general case; i.e., Theorems 3.5–3.6 combine to imply the following general result.

Theorem 3.7 (lifting) *For a* $G_t/M_t/s_t + GI_t$ *fluid model with* $s', \mu \in \mathcal{P}_{m,n}$ *and* $\lambda \in \mathbb{C}_p$*, the system performance via* $(B, \tilde{q}, w)$*, for* $B \equiv \{B(t) : 0 \leq t \leq T\}$*, is well defined and the conclusions of* §3.3 *and Theorems* 3.5 *and* 3.6 *remain valid.*

3.6 The $(G_t/M_t/s_t + GI)^m/M_t$ Fluid Queue Network.

We now introduce the open network of $G_t/M_t/s_t + GI$ fluid queues, with time-dependent proportional routing. There are m queues, where each queue has model parameters as already defined in §3.2, with its own external fluid input, but in addition a proportion $P_{i,j}(t)$ of the fluid output from queue i at time t is routed immediately to queue j, and a proportion $P_{i,0}(t) \equiv 1 - \sum_{j=1}^{m} P_{i,j}(t) \leq 1$ is routed out of the network, as shown in Figure 3.1 for the case $m = 2$.

Assumption 3.7 (*proportional routing*) *The routing matrix function for proportional routing,* $P : [0, \infty) \to [0, 1]^{m^2}$*, is in* $\mathbb{C}_p$ *and* $\sum_{j=1}^{m} P_{i,j}(t) \leq 1$ *for each* $t \geq 0$ *and* i*,* $1 \leq i \leq m$*.*

It is elementary to treat the basic network operations of superposition and splitting: If two input streams are combined to form a single input (superposition), then the arrival rate functions are simply added. If one stream with arrival rate function λ is split, such that a proportion $p(t)$ of that stream goes into a new split stream at time t, then the arrival-rate function of the split stream is λ_p, where $\lambda_p(t) \equiv \lambda(t)p(t)$, $t \geq 0$; just like λ, the splitting proportion can be time-dependent. Similarly, if the departure flow from one queue becomes input to another, then the resulting arrival-rate function is σ; (We do not let the abandonment flow from one queue become input to another, but if we did, then the resulting arrival-rate function would be α.) However, converting departure rate or abandonment rate into new input rate is more complicated when feedback is allowed. We discuss that case now, for departures only.

As is usual with open queueing networks, there is an external exogenous arrival rate function to each queue (from outside the network) and there is a total arrival rate function to each queue (which we simply call the arrival rate function), taking into account the flow from other queues. Let the external arrival rate function into queue j be denoted by $\lambda_j^{(0)}$; let the arrival rate function into queue j be denoted by λ_j. The model data for the $G_t/M_t/s_t + GI_t$ fluid queues directly provides the external arrival rate functions $\lambda_j^{(0)}$ (with the superscript 0 now added), while the arrival rate function itself satisfies a system of *traffic rate equations*. In particular,

$$\lambda_j(t) = \lambda_j^{(0)}(t) + \sum_{i=1}^{m} \sigma_i(t) P_{i,j}(t), \quad \text{where} \tag{3.20}$$

$$\sigma_i(t) = B_i(t)\mu_i(t), \quad t \geq 0. \tag{3.21}$$

Equations (3.20) and (3.21) produce a system of equations, with λ_j depending upon σ_i for $1 \leq i \leq m$, while σ_i in turn depends on λ_i for each i, because B_i depends on λ_i. The formulas for B_i as a function of λ_i have been given in Propositions 3.1 and 3.2, provided

that we know whether the queue is overloaded or underloaded. That requirement is the major source of complexity.

Since (3.20) is a linear equation, it can be written in matrix notation as $\lambda = \lambda^{(0)} + \sigma\, P$ by omitting the argument t as below, provided that the product σP is interpreted as in (3.20). Moreover, we can combine (3.20) and (3.21) to express λ as the solution of a fixed point equation mapping $\mathbb{C}_p^m$ over $[0, T]$ into itself. To see this, note that $B_i(t)$ in (3.21) is a function of $\lambda_i(u)$, $0 \leq u < t$, and the model data (only needed for queue i). Hence the vector $B(t) \equiv (B_1(t), \ldots, B_m(t))$ is a function of λ over $[0, t)$ and the model data. Hence we can express (3.20) and (3.21) abstractly as

$$\lambda = \Psi(\lambda), \tag{3.22}$$

where $\Psi(x)(t)$ depends on its argument x only over $[0, t]$ for each $t \geq 0$. Here the function Ψ depends on all the model data $(\lambda_i^{(0)}, s_i, \mu_i, F_{i,\cdot}, b_i(0, \cdot), q_i(0, \cdot), P)$, $1 \leq i \leq m$.

3.7 Two Algorithms for the Network with M_t Service

In this section we establish two different algorithms to compute all standard performance measures for the $(G_t/M_t/s_t + GI_t)^m/M_t$ fluid network. The first algorithm is based on solving an FPE and the second is based on solving an ODE. In §3.8 we generalize our analysis to the network with GI service distributions.

3.7.1 An FPE Based Algorithm

This algorithm is based on solving the FPE (3.22). We first establish the following contraction property of the operator Ψ.

Theorem 3.8 (*contraction operator*) *If* $s_i', \mu_i \in \mathcal{P}_{m,n}$ *for* $1 \leq i \leq m$*, then the operator* Ψ *in* (3.22) *is a monotone contraction operator on the* m*-dimensional product space* $\mathbb{C}_p^m$ *over* $[0,T]$ *for all sufficiently small* $T > 0$*. Hence there exists a unique solution* λ *to the traffic rate equations* (3.20) *and* (3.21) *over* $[0,T]$ *for any fixed* $T > 0$*. For sufficiently short intervals, successive iterates* $\Psi^{(n)}(\tilde{\lambda})$ *converge uniformly, geometrically fast, to the fixed point for any initial point* $\tilde{\lambda} \in \mathbb{C}_p^m$*.*

Proof. We first show that Ψ actually maps $\mathbb{C}_p$ into itself. First, if $\lambda \in \mathbb{C}_p^m$, then $B \in \mathbb{C}_p^m$ by Corollary 3.1 and Theorem 3.7. By assumption $\mu \in \mathbb{C}_p^m$, so that $\sigma \in \mathbb{C}_p^m$, so the conclusion follows from (3.20) and (3.21). To show that Ψ is a contraction operator for sufficiently small $T > 0$, we use the norm $\|\lambda\|_T \equiv \sum_{i=1}^m \|\lambda_i\|_T$ for $\lambda \equiv (\lambda_1, \ldots, \lambda_m) \in (\mathbb{C}_p)^m$. For any $\lambda_1, \lambda_2 \in (\mathbb{C}_p)^m$, the traffic rate equations in (3.20) and (3.21) imply that

$$\begin{aligned}
\|\Psi(\lambda_1) - \Psi(\lambda_2)\|_T &\leq \sum_{j=1}^m \sup_{1\leq t\leq T} \sum_{i=1}^m \mu_i(t)|B_i^1(t) - B_i^2(t)|\, P_{i,j}(t) \\
&\leq m\,\mu_T^{\uparrow} \sum_{i=1}^m \sup_{0\leq t\leq T} |B_i^1(t) - B_i^2(t)| \\
&\leq m\,\mu_T^{\uparrow} T \sum_{i=1}^m \sup_{0\leq t\leq T} |\lambda_i^1(t) - \lambda_i^2(t)| \leq m\mu_T^{\uparrow} T\, \|\lambda_1 - \lambda_2\|_T,
\end{aligned}$$

where $m\mu_T^{\uparrow} T < 1$ for all sufficiently small $T > 0$. The second inequality holds since $P_{i,j}(t) \leq 1$. The crucial third inequality follows from (3.19) in Theorem 3.6. To establish uniqueness over $[0,T]$ for any fixed $T > 0$, we consider a succession of shorter intervals, over which the contraction property holds, and apply mathematical induction. Existence, uniqueness and geometric convergence are standard consequences of the Banach contraction fixed point theorem. Finally, monotonicity follows from Theorems 3.5 and 3.7 plus the traffic rate equations (3.20) and (3.21). ∎

Remark 3.1 *(starting at the external arrival rates) Theorem 3.8 implies that we can approach this system recursively. If we do so with initial vector* $\tilde{\lambda} = \lambda^{(0)}$*, the vector of external arrival rate functions, then the recursion has an important practical interpretation. Then the* k^{th} *iterate* $\lambda_j^{(k)}$ *is the arrival rate of fluid that has previously experienced* k *transitions in the fluid network. With this notation, we can write the recursive formulas*

$$\lambda_j^{(n)}(t) = \Psi^{(n)}(\lambda^{(0)})_j(t) = \lambda_j^{(0)}(t) + \sum_{i=1}^{m} \sigma_i^{(n-1)}(t) P_{i,j}(t), \quad n \geq 1, \tag{3.23}$$

$$\textit{where} \quad \sigma_i^{(n)}(t) = B_i^{(n)}(t)\mu_i(t) \quad n \geq 0. \tag{3.24}$$

Since we necessarily have $\lambda_i^{(1)} \geq \lambda_i^{(0)}$ *for each* i*, this recursion converges monotonically to the fixed point* λ*. By Theorems 3.5 and 3.7, all the performance measures increase toward their limiting values as well.*

The FPE based algorithm for the network of fluid queues. The algorithm consists of two successive steps: (i) solving the traffic-rate equations (3.20) and (3.21) and (ii) solving for the performance vector $(b, q, w, v, \sigma, \alpha)$ at each queue using the algorithm in §2.3. For step (i), we start with an initial vector of arrival rate functions, which can a rough estimate of the final arrival rate functions or the given external arrival rate functions. We then apply the performance formulas in §3.4 to determine the performance functions B_i and σ_i at each queue to determine a new vector of arrival rate functions. We then iteratively calculate successive vectors of arrival rate functions until the difference (measured in the supremum norm over a bounded interval) is suitably small. Then we apply step (ii).

Given a desired duration T of an interval $[0, T]$, we specify the following input data: (i) Model parameter input vector

$$\left(\lambda^{(0)}, s, G, F, \mathcal{P}(0)\right) \equiv \left(\lambda_i^{(0)}(t), s_i(t), G_i, F_i, \mathcal{P}_i(0), 1 \leq i \leq m,\, t \in [0, T]\right), \quad (3.25)$$

where the initial performance vector (at time 0) of queue i, $1 \leq i \leq m$

$$\mathcal{P}_i(0) \equiv (b_i(0, \cdot), q_i(0, \cdot), B_i(0), Q_i(0), w_i(0), v_i(0), \alpha_i(0), \sigma_i(0))\,;$$

and (ii) algorithm accuracy parameters: the error tolerance parameter (ETP) $\epsilon > 0$ and the step size $0 < \Delta T \leq T$. We next summarize the algorithm formally as the following.

Algorithm 3 : An FPE based algorithm for the $(G_t/M_t/s_t + GI_t)^m/M_t$ Fluid Network

1: Initialization: $\lambda^{(1)} := \lambda^{(0)}$, $0 \leq i \leq m$
2: **for** $k = 1, 2, \ldots$ **do**
3: **for** $i = 1, 2, \ldots, m$ **do**
4: Compute σ_i in $[0, T]$ using FASQ (Algorithm 2) with input $\left(\lambda_i^{(k)}, s_i, G_i, F_i, \hat{\mathcal{P}}_i(0)\right)$
5: **end for**
6: Let $\lambda^{(k+1)} := \lambda^{(0)} + P^T \cdot \sigma$ in $[0, T]$
7: **if** $\|\lambda^{(k+1)} - \lambda^{(k)}\|_T < \epsilon$ **then**
8: $\lambda := \lambda^{(k+1)}$
9: Break
10: **end if**
11: **end for**
12: Compute $\mathcal{P}_i$ for $1 \leq i \leq m$ using FASQ (Algorithm 2) with input $\left(\lambda_i, s_i, G_i, F_i, \hat{\mathcal{P}}_i(0)\right)$

Remark 3.2 $($*complexity of the FPE based algorithm with respect to the number of switching points $\mathcal{S}$ and the size of the system* $m)$ *The running time of this algorithm depends on the number of regime switchings (between UL and OL). Suppose the number of switchings*

for each queue in $[0,T]$ is $O(\mathcal{S})$, then in each iteration of the fixed-point recursion the complexity is $O(m\mathcal{S})$ because the single-queue fluid algorithm is called for m times to compute performance measures for all m queues. If the total number of iterations is n, then the total complexity is of order $O(nm\mathcal{S})$. Thus, the running time is linear both in $\mathcal{S}$ and in m.

We conclude this section by establishing a network generalization of the single queue comparison in Theorem 3.5. The proof appears in §B.4.

Theorem 3.9 (network comparison theorem) *Consider two $(G_t/M_t/s_t+GI_t)^m+M_t$ fluid queue networks with common staffing functions s_i, service rate functions μ_i, abandonment cdf's $F_{\cdot,i}$ and routing matrix function P for $1 \le i \le m$. If $\lambda_{1,i}^{(0)} \le \lambda_{2,i}^{(0)}$, $B_{1,i}(0) \le B_{2,i}(0)$, $q_{1,i}(0,\cdot) \le q_{2,i}(0,\cdot)$, $1 \le i \le m$, then the performance functions are ordered at each queue:*

$$\begin{aligned}&(\lambda_{1,i}, B_{1,i}, \sigma_{1,i}, \tilde{q}_{1,i}, q_{1,i}, Q_{1,i}, \alpha_{1,i}, X_{1,i}, w_{1,i}, v_{1,i})\\ &\quad \le (\lambda_{2,i}, B_{2,i}, \sigma_{2,i}, \tilde{q}_{2,i}, q_{2,i}, Q_{2,i}, \alpha_{2,i}, X_{2,i}, w_{2,i}, v_{2,i}) \quad \textit{for} \quad 1 \le i \le m. \end{aligned} \tag{3.26}$$

3.7.2 An ODE Based Algorithm

Now we consider an alternative algorithm for the $(G_t/M_t/s_t + GI_t)^m/M_t$ fluid queue network. Again, the key is to compute the total arrival rates for all queues and then treat them separately as single queues. This new algorithm is faster and easier to implement. In some special cases, analytic formulas are available.

Finding the total arrival rates: Instead of solving the FPE as in Chapter 2, we hereby solve an m-dimensional ODE. The key is to characterize and update the system regime in

different intervals and recursively advance in t. We describe the system regime at t with two sets: $\mathcal{U}(t)$ (the set of indices of queues that are UL) and $\mathcal{O}(t)$ (the set of indices of queues that are OL). In other words,

$$\begin{aligned}\mathcal{U}(t) &\equiv \{1 \le i \le m : B_i(t) \le s_i(t),\, Q_i(t) = 0\} \\ \mathcal{O}(t) &\equiv \{1 \le i \le m : B_i(t) = s_i(t),\, Q_i(t) > 0\}.\end{aligned}$$

Given $\mathcal{U}(t)$ and $\mathcal{O}(t)$, consider $1 \le i \le m$. (i) If Queue i is UL, i.e., $i \in \mathcal{U}(t)$, flow conservation implies that

$$B_i'(t) = \lambda_i^{(0)}(t) + \sum_{j \in \mathcal{U}(t)} \mu_j(t) P_{j,i}(t) B_j(t) + \sum_{k \in \mathcal{O}(t)} \mu_k(t) P_{k,i}(t) s_k(t) - \mu_i(t) B_i(t).$$

If $i \in \mathcal{O}(t)$, $B_i(t) = s_i(t)$. We partition the indices of queues so that $\mathbf{B}(t) \equiv [\mathbf{B}_{\mathcal{U}}(t), \mathbf{B}_{\mathcal{O}}(t)]$, $\lambda(t) \equiv [\lambda_{\mathcal{U}}(t), \lambda_{\mathcal{O}}(t)]$, $\lambda^{(0)}(t) \equiv [\lambda_{\mathcal{U}}^{(0)}(t), \lambda_{\mathcal{O}}^{(0)}(t)]$, $\mu(t) \equiv [\mu_{\mathcal{U}}(t), \mu_{\mathcal{O}}(t)]$, $\mathbf{s}(t) \equiv [\mathbf{s}_{\mathcal{U}}(t), \mathbf{s}_{\mathcal{O}}(t)]$, $\mathbf{\Gamma}_{\mathcal{U}}(t) \equiv diag(\mu_{\mathcal{U}}(t))$, $\mathbf{\Gamma}_{\mathcal{O}}(t) \equiv diag(\mu_{\mathcal{O}}(t))$,

$$\mathbf{P}(t) \equiv \begin{bmatrix} \mathbf{P}_{\mathcal{U}\mathcal{U}}(t) & \mathbf{P}_{\mathcal{U}\mathcal{O}}(t) \\ \mathbf{P}_{\mathcal{O}\mathcal{U}}(t) & \mathbf{P}_{\mathcal{O}\mathcal{O}}(t) \end{bmatrix},$$

where $\mathbf{P}_{\mathcal{U}\mathcal{U}}(t)$ ($\mathbf{P}_{\mathcal{O}\mathcal{U}}(t)$, $\mathbf{P}_{\mathcal{U}\mathcal{O}}(t)$, and $\mathbf{P}_{\mathcal{O}\mathcal{O}}(t)$) denotes the transition probability from a state in $\mathcal{U}$ ($\mathcal{O}$, $\mathcal{U}$, and $\mathcal{O}$) to a state in $\mathcal{U}$ ($\mathcal{U}$, $\mathcal{O}$, and $\mathcal{O}$) at time t. Let $\mathbf{P}_{\mathcal{O}\mathcal{U}}(t) = \mathbf{P}_{\mathcal{U}\mathcal{O}}(t) = \mathbf{P}_{\mathcal{O}\mathcal{O}}(t) = \mathbf{0}$ when $\mathbf{P}_{\mathcal{U}\mathcal{U}}(t) = \mathbf{P}(t)$ (i.e., all queues are UL) and let $\mathbf{P}_{\mathcal{O}\mathcal{U}}(t) = \mathbf{P}_{\mathcal{U}\mathcal{O}}(t) = \mathbf{P}_{\mathcal{U}\mathcal{U}}(t) = \mathbf{0}$ when $\mathbf{P}_{\mathcal{O}\mathcal{O}}(t) = \mathbf{P}(t)$ (i.e., all queues are OL) Therefore, in matrix notation,

we have

$$\mathbf{B}'_{\mathcal{U}}(t) = \mathbf{C}(t) \cdot \mathbf{B}_{\mathcal{U}}(t) + \mathbf{D}(t), \tag{3.27}$$

$$\mathbf{B}_{\mathcal{O}}(t) = \mathbf{s}_{\mathcal{O}}(t), \tag{3.28}$$

where

$$\begin{aligned} \mathbf{D}(t) &\equiv \lambda_{\mathcal{U},(0)}(t) + \Gamma_{\mathcal{O}}(t)\mathbf{P}^T_{\mathcal{O}\mathcal{U}}(t)\mathbf{s}_{\mathcal{U}}(t) \\ \mathbf{C}(t) &\equiv \mathbf{\Gamma}_{\mathcal{U}}(t)\left(\mathbf{P}^T_{\mathcal{U}\mathcal{U}}(t) - \mathbf{I}\right). \end{aligned}$$

If the service rates and the routing probability matrix are independent of time: $\mu_i(t) = \mu_i$ and $P_{i,j}(t) = P_{i,j}$, i.e., the model becomes the $(G_t/M/s_t + GI_t)^m/M$ network, then $\mathbf{\Gamma}_{\mathcal{U}} \equiv \mathbf{\Gamma}_{\mathcal{U}}(t) = diag(\mu_{\mathcal{U}})$, $\mathbf{C} \equiv \mathbf{C}(t) = \mathbf{\Gamma}_{\mathcal{U}}\left(\mathbf{P}^T_{\mathcal{U}\mathcal{U}} - \mathbf{I}\right)$, and (3.27) has the unique solution

$$\mathbf{B}_{\mathcal{U}}(t) = e^{-\mathbf{C}\,t}\left(\int_0^t e^{-\mathbf{C}\,u}\mathbf{D}(u)du + \mathbf{B}(0)\right).$$

In all cases, the total arrival rate

$$\lambda(t) = \lambda^{(0)}(t) + \mathbf{P}^T(t)\mathbf{\Gamma}(t) \cdot \mathbf{B}(t). \tag{3.29}$$

Regime termination criterion: It is also critical to determine when the system regime changes and to update $\mathcal{U}(t)$ and $\mathcal{O}(t)$. Since each queue can be either UL or OL, there are overall 2^m different regimes. We say that the system changes its regime if one of the queues changes its regime, i.e., from UL to OL or from UL to OL. We provide the following regime

termination time

$$\begin{aligned} T_{\mathcal{R}}(t_0) &\equiv T_1(t_0) \wedge T_2(t_0), \quad \text{where} \\ T_1(t_0) &\equiv \inf\{t \geq t_0 : \text{some } i \in \mathcal{O} \text{ s.t. } Q_i(t) = 0, \lambda_i(t) \leq \sigma_i(t)\}, \\ T_2(t_0) &\equiv \inf\{t \geq t_0 : \text{some } j \in \mathcal{U} \text{ s.t. } B_j(t) = s_j(t), \lambda_j(t) > \sigma_j(t)\}, \end{aligned} \tag{3.30}$$

t_0 is the starting time of the desired interval, the infimum of an empty set is understood to be infinity. When the system regimes changes, we update $\mathcal{U}(t)$ and $\mathcal{O}(t)$. Let k^* be the index of the queue that causes the regime switching. If $k^* \in \mathcal{O}(t-)$, i.e., $T = T_1$, let

$$\mathcal{O}(t) \leftarrow \mathcal{O}(t) \backslash \{k^*\} \quad \text{and} \quad \mathcal{U}(t) \leftarrow \mathcal{U}(t) \cup \{k^*\}; \tag{3.31}$$

if $k^* \in \mathcal{U}(t-)$, i.e., $T = T_2$, let

$$\mathcal{U}(t) \leftarrow \mathcal{U}(t) \backslash \{k^*\} \quad \text{and} \quad \mathcal{O}(t) \leftarrow \mathcal{O}(t) \cup \{k^*\}. \tag{3.32}$$

Given a desired duration T of an interval $[0, T]$, the vector of the model data defined as (3.25), and a step size $0 < \Delta T \leq T$, we summarize the algorithm formally as the following.

Remark 3.3 $($*complexity of the ODE based algorithm with respect to the number of switching points $\mathcal{S}$ and the size of the system m*$)$ *The running time of this algorithm again depends on the number of system regime switchings (between UL and OL). Suppose the number of switchings for each queue in $[0, T]$ is $O(\mathcal{S})$, then the number of system regime changes is at most the sum of the total number of regimes switches of all m queues in $[0, T]$ (assuming*

Algorithm 4 : An ODE based algorithm for the $(G_t/M_t/s_t + GI_t)^m/M_t$ Fluid Network

```
Initialization: t := 0
repeat
   for k = 0, 1, ..., ⌈(T−t)/ΔT⌉ do
      Compute λ(s) and B(s) for s ∈ [t + (k − 1)ΔT, t + k ΔT], using (3.27)-(3.29)
      Compute P(s) for s ∈ [t + (k − 1)ΔT, t + k ΔT] using Proposition 3.1 ,3.2 and
      3.3, Corollary 3.2, Theorem 2.3 and 2.6
      if T_R(t) < t + k ΔT for T_R(t) in (3.30) then
         t := T_R(t)
         Update U(t) and O(t) by (3.31)-(3.32)
         BREAK for-loop
      end if
   end for
until t ≥ T
```

no two queues change their regimes at the same time). Hence the complexity of the new algorithm is of order $O(m\mathcal{S})$. It is again linear both in $\mathcal{S}$ and in m.

3.8 An Extension to GI Service Distribution

In this section, we extend our analysis from the M service distribution to GI. Without the M service distribution, neither algorithms in §3.6 is applicable. Here we provide another algorithms that is based on solving a new FPE. Throughout this section, we make the following assumption.

Assumption 3.8 (*finitely many switches between intervals in finite time*) *Each interval is of positive length, so that the positive half line $[0, \infty)$ can be partitioned into 2^m intervals. Moreover, there are only finitely many switches between these intervals in each finite interval.*

The key is to obtain the total arrival rate $\lambda_i(t)$ for $1 \leq i \leq m$ and $0 \leq t \leq T$. Once $\lambda_i(t)$ is given, the algorithm developed in Chapter 2 can be applied to compute all other

performance measures. If queue j ($1 \leq j \leq m$) is UL, from Chapter 2 we have that

$$\begin{aligned} b_j(t,x) &= \bar{G}_j(x)\lambda_j(t-x)\mathbf{1}_{\{x\leq t\}} + \frac{\bar{G}_j(x)}{\bar{G}_j(x-t)}b_j(0,x-t)\mathbf{1}_{\{x>t\}}, \\ \sigma_j(t) &= \int_0^\infty b_j(t,x)h_{G,j}(x)dx \\ &= \int_0^t g_j(x)\lambda_j(t-x)dx + \int_0^\infty \frac{g_j(x+t)}{\bar{G}_j(x)}b_j(0,x)dx. \end{aligned} \tag{3.33}$$

If queue k ($1 \leq k \leq m$) is OL, from Chapter 2, then $\sigma_k(t) = b_k(t,0) - s_k'(t)$ and the rate into service (RIS) $b_k(t,0)$ satisfies the FPE

$$b_k(\cdot,0) = \mathcal{T}(b_k(\cdot,0)), \tag{3.34}$$

where

$$\begin{aligned} \mathcal{T}(y)(t) &\equiv \hat{a}_k(t) + \int_0^t y(t-x)g_k(x)dx, \\ \hat{a}_k(t) &\equiv s_k'(t) + \int_0^\infty \frac{b_k(0,y)g_k(t+y)}{\bar{G}_k(y)}dy. \end{aligned}$$

Moreover, we have showed in Chapter 2 that $\mathcal{T}$ is a contraction operator under mild conditions, which thus implies that (3.34) has a unique solution. Having $\sigma_k(t)$ and $b_k(t,0)$ computed for $k \in \mathcal{O}(t)$, the total arrival rate at queue i

$$\begin{aligned} \lambda_i(t) &= \lambda_i^{(0)}(t) + \sum_{k\in\mathcal{O}(t)} P_{k,i}(t)\sigma_k(t) + \sum_{j\in\mathcal{U}(t)} P_{j,i}(t)\sigma_j(t) \\ &= \hat{\gamma}_i(t) + \sum_{j\in\mathcal{U}(t)} P_{j,i}(t)\left(\int_0^t g_j(x)\lambda_j(t-x)dx\right), \end{aligned} \tag{3.35}$$

where

$$\hat{\gamma}_i(t) \equiv \lambda_i^{(0)}(t) + \sum_{k\in\mathcal{O}(t)} P_{k,i}(t)\sigma_k(t) + \sum_{j\in\mathcal{U}(t)} P_{j,i}(t)\int_0^\infty \frac{g_j(x+t)}{\bar{G}_j(x)} b_j(0,x)dx.$$

and the second equality holds by (3.33).

From (3.35), it is evident that λ satisfies a FPE, i.e.,

$$\lambda = \mathcal{J}(\lambda), \tag{3.36}$$

of the operator $\mathcal{J} : \mathbb{D}^m \to \mathbb{D}^m$, where

$$\mathcal{J}(u)_i(t) \equiv \hat{\gamma}_i(t) + \sum_{j\in\mathcal{U}(t)} P_{j,i}(t)\left(\int_0^t g_j(x)u_j(t-x)dx\right), \quad 1 \le i \le m. \tag{3.37}$$

Under regularity conditions, we can show that there exists a unique solution to equation (3.35) by applying the Banach contraction theorem. We will use the complete (nonseparable) normed space $\mathbb{D}^m$ with the uniform norm over the interval $[0, T]$, i.e.,

$$\|u\|_T \equiv \sum_{i=1}^m \sup_{0\le t\le T} |u_i(t)|. \tag{3.38}$$

Theorem 3.10 $(\textit{the aggregated arrival rate for } GI \textit{ service})$ *Assume the system regime does not change in a small interval* $[0, T]$. *The operator* $\mathcal{J}$ *in* (3.37) *is a monotone contraction operator on* $\mathbb{D}^n$ *with norm defined in* (3.38).

Proof. Assume that $T > 0$ is small enough so that the system regime does not change, i.e., $\mathcal{U}(t) = \mathcal{U}$ and $\mathcal{O}(t) = \mathcal{O}$ for $0 \leq t \leq T$.

$$\begin{aligned}
\|\mathcal{J}(u_1) - \mathcal{J}(u_2)\|_T &= \sum_{i=1}^{m} \sup_{0\leq t\leq T} \left| \sum_{j\in\mathcal{U}(t)} P_{j,i}(t) \left[\int_0^t g_j(x) \left(u_{1,j}(t-x) - u_{2,j}(t-x)\right) dx \right] \right| \\
&\leq \sum_{i=1}^{m} \sup_{0\leq t\leq T} \sum_{j\in\mathcal{U}} \|u_{1,j} - u_{2,j}\|_T P_{j,i}(t) G_j(t) \\
&\leq m \max_{1\leq j\leq m} G_j(T) \cdot \|u_1 - u_2\|_T \\
&\leq \tilde{C}(T) \|u_1 - u_2\|_T,
\end{aligned}$$

where

$$\tilde{C}(T) \equiv m \max_{1\leq j\leq m} G_j(T),$$

and the second inequality holds by the Lipschitz continuity assumption on $P_{i,j}(t)$. Note that we can make $\tilde{C}(T) < 1$ for small $T > 0$ since $G_i(t) \to 0$ as $t \to 0$ for all $1 \leq i \leq m$. □

Given a desired duration T of an interval $[0, T]$, the vector of the model data defined as (3.25), a step size $0 < \Delta T \leq T$, and an error tolerance parameter (ETP) $\epsilon > 0$, we summarize the algorithm formally as the following.

3.9 Examples

In this section we implement the algorithms in §§3.6-3.8 on a Markovian and non-Markovian fluid network models.

Algorithm 5 : An FPE based algorithm for the $(G_t/GI/s_t + GI_t)^m/M_t$ Fluid Network

Initialization: $t := 0$
repeat
  for $k = 0, 1, \ldots, \lceil \frac{T-t}{\Delta T} \rceil$ **do**
    for all $i \in \mathcal{O}(t)$ **do**
      - Compute $b_i(s, 0)$ solving FPE (3.34) with ETP ϵ, $s \in [t+(k-1)\Delta T, t+k\,\Delta T]$
      - Let $\sigma_i(s) := b_i(s, 0) - s_i'(s)$
    end for
    Compute $\lambda(s)$ using FPE (3.36) with ETP ϵ, $s \in [t + (k-1)\Delta T, t + k\,\Delta T]$
    Compute $\mathcal{P}(s)$ for $s \in [t + (k-1)\Delta T, t + k\,\Delta T]$ using using Proposition 2.6, Corollary 2.2, Theorem 2.3 and 2.6
    if $T_{\mathcal{R}}(t) < t + k\,\Delta T$ for $T_{\mathcal{R}}(t)$ in (3.30) **then**
      $t := T_{\mathcal{R}}$
      Update $\mathcal{U}(t)$ and $\mathcal{O}(t)$ by (3.31)-(3.32)
      BREAK for-loop
    end if
  end for
until $t \geq T$

3.9.1 An $(M_t/M/s_t + M)^2/M_t$ Marvovian Example

We first consider a Markovian $(M_t/M/s_t+M)^2/M_t$ example (a two-queue network), with sinusoidal external arrival rates

$$\lambda_i^{(0)}(t) = a_i + b_i\,\sin(c_i\,t + \phi_i), \quad i = 1, 2, \tag{3.39}$$

exponential service and patience distributions: $\bar{G}_i(x) = e^{-\mu_i\,x}$, $\bar{F}_i(x) = e^{-\theta_i\,x}$, $i = 1, 2$, constant staffing functions s_i, $i = 1, 2$, and a Markovian transition probability matrix

$$\mathbf{P}(t) = \begin{bmatrix} P_{1,1} & P_{1,2} \\ P_{2,1} & P_{2,2} \end{bmatrix} = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}. \tag{3.40}$$

Therefore, with probability $P_{1,0} = P_{2,0} = 0.5$, a customer leaves the system after finishing service at each queue. Let $a_1 = a_2 = 0.5$, $b_1 = 0.25$, $b_2 = 0.35$, $c_1 = c_2 = 1$, $\phi_1 = 0$,

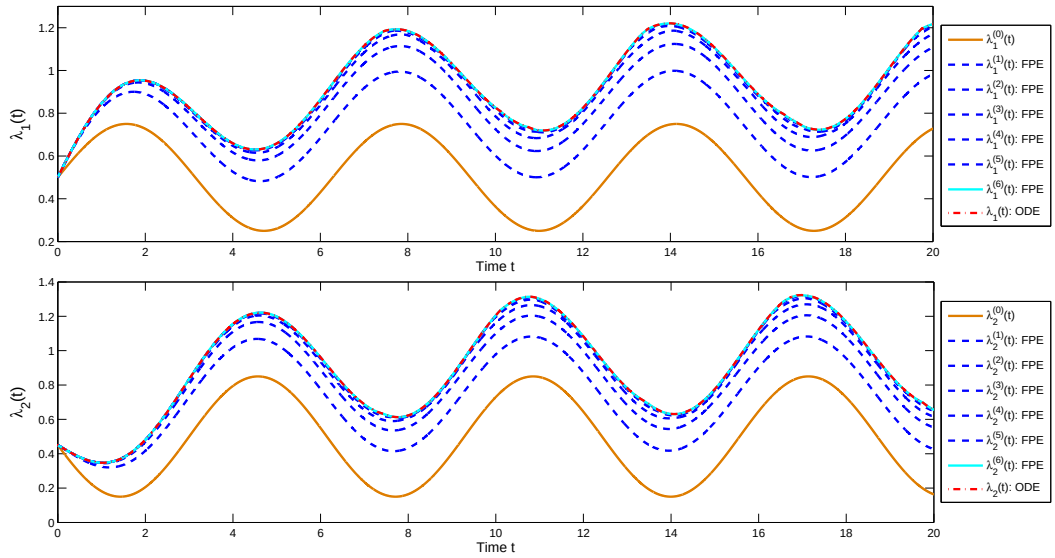

Figure 3.2: The convergence to the fixed point of the total arrival rate.

$\phi_2 = 1$, $\mu_1 = 1$, $\mu_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 0.3$, $s_1 = 1$, and $s_2 = 2$. We let the network be initially empty.

We first demonstrate how the FPE based algorithm works. Since it is key to obtain the total arrival rates $\lambda_1(t)$ and $\lambda_2(t)$ for $0 \leq t \leq T$, we first demonstrate how fast the fixed-point algorithm converges. We initially let $\lambda_i^{(1)}$ be $\lambda_i^{(0)}$, $i = 1, 2$. In Figure 3.2, we plot the total arrival rates in every iteration. The two functions at the bottom are $\lambda_1^{(0)}(t)$ and $\lambda_2^{(0)}(t)$; the functions at the top are the $\lambda_1(t)$ and $\lambda_2(t)$ (computed using the ODE based algorithm); the other functions are the intermediate values (computed using the FPE based algorithm). Recall that the FPE based algorithm terminate at step $N(\epsilon)$ and let $\lambda_i \equiv \lambda_i^{(N(\epsilon))}$, $i = 1, 2$, for

$$N(\epsilon) \equiv \inf \left\{ N \geq 0 : \mathcal{E}_T(N) \equiv \max_{j=1,2} \|\lambda_j^{(N)} - \lambda_j^{(N-1)}\|_T \leq \epsilon \right\},$$

where $\epsilon > 0$ is a pre-specified error tolerance parameter (ETP). For this example, we demonstrate how the number of iterations $N(\epsilon)$ and the terminating error $\mathcal{E}_T(N(\epsilon))$ depends

on the EPT ϵ in Table 1. Here the monotone convergence and the geometric convergence rate are explained by the monotone contraction property of the operator Ψ.

$\log_{10}(\epsilon)$	-1	-2	-3	-4	-5	-6	-7	-8	-9
$\mathcal{E}_T(N(\epsilon))$	0.81	0.007	9.2E-4	4.8E-5	4.9E-6	2.8E-7	5.2E-8	8.3E-9	1.4E-10
$N(\epsilon)$	3	6	8	11	13	15	16	17	19

Table 3.1: The number of iterations N of the FPE algorithm, depending on the ETP ϵ.

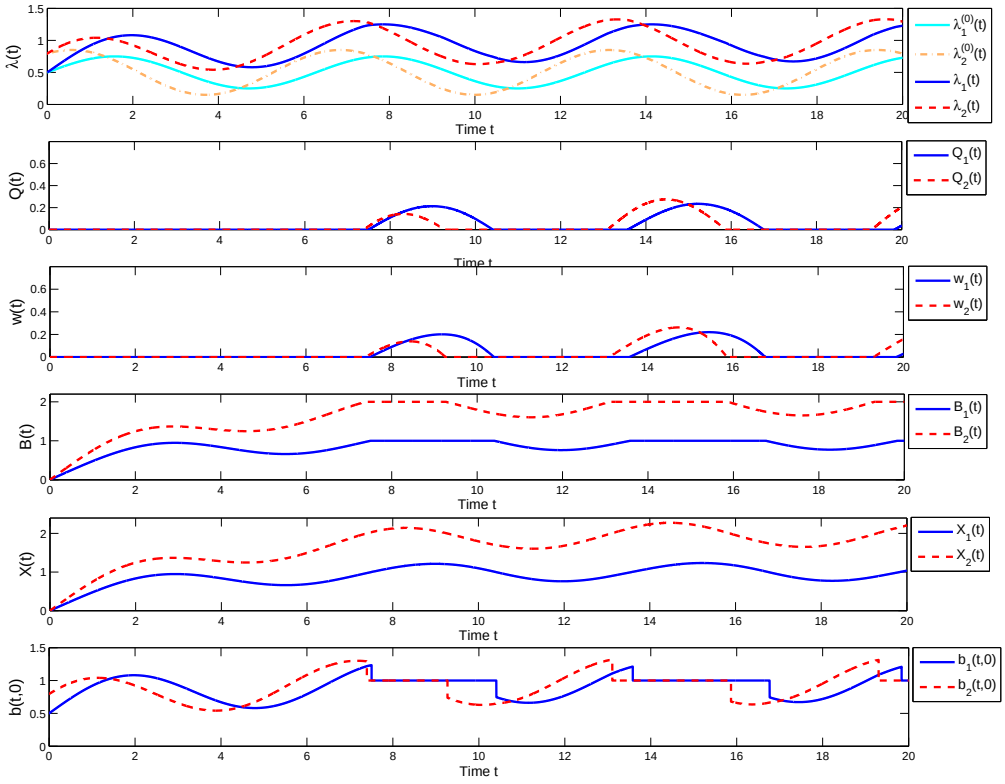

Figure 3.3: Computing the fluid performance functions for the $(M_t/M/s_t + M)^2/M_t$ network fluid model.

In Figure 3.3, we plot all standard performance measures of the fluid network using the FPE based algorithm, including λ_i, Q_i, w_i, B_i, X_i, and $b_i(\cdot, 0)$, $i = 1, 2$. In Figure 3.4, we compare the fluid approximations with results from a simulation experiment for a very large-scale queueing system. The queueing model has nonhomogeneous Poisson external

arrival processes with sinusoidal rate functions

$$\lambda_{n,i}^{(0)}(t) = n\lambda_i^{(0)}(t), \quad i = 1, 2,$$

with $n = 2000$. We compare the fluid model predictions to a single sample path of the queueing system (one simulation run). In Figure 3.4 the solid lines are the simulation estimations of single sample paths applied with fluid scaling, and the dashed lines are the fluid approximations. We conclude that the fluid approximation is remarkably accurate as an approximation when the scale of the queueing model is extremely large.

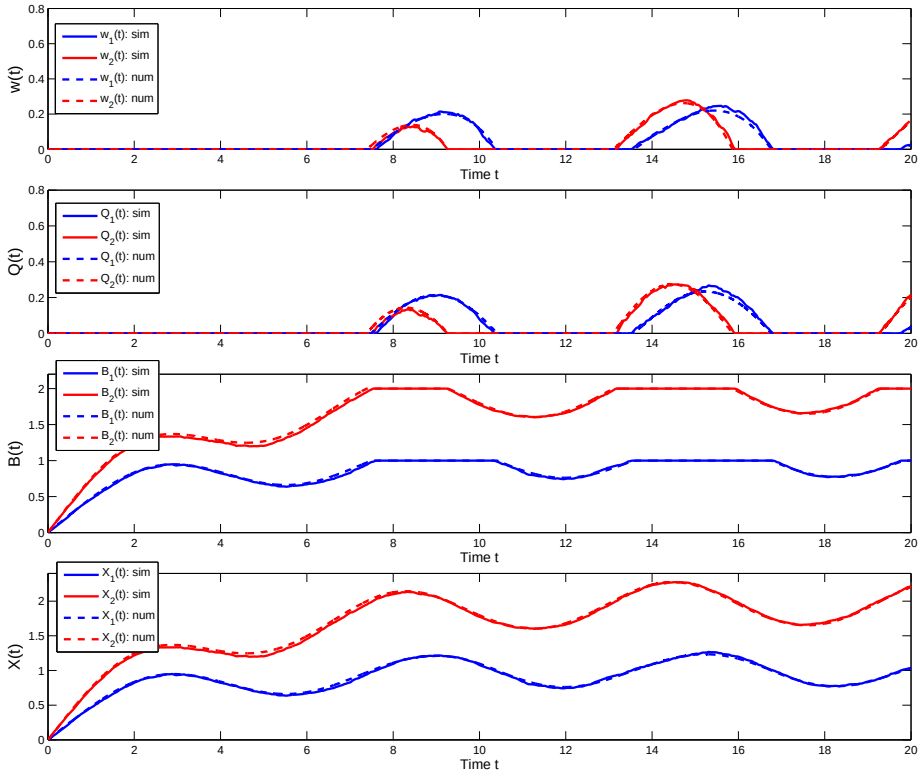

Figure 3.4: A comparison of the $(M_t/M/s_t + M)^2/M_t$ network fluid model with a simulation run of single sample paths, $n = 2000$.

When the scale of the queueing model is not large (i.e., n is small), single sample paths of the queueing functions do not necessarily agree with the fluid functions because of

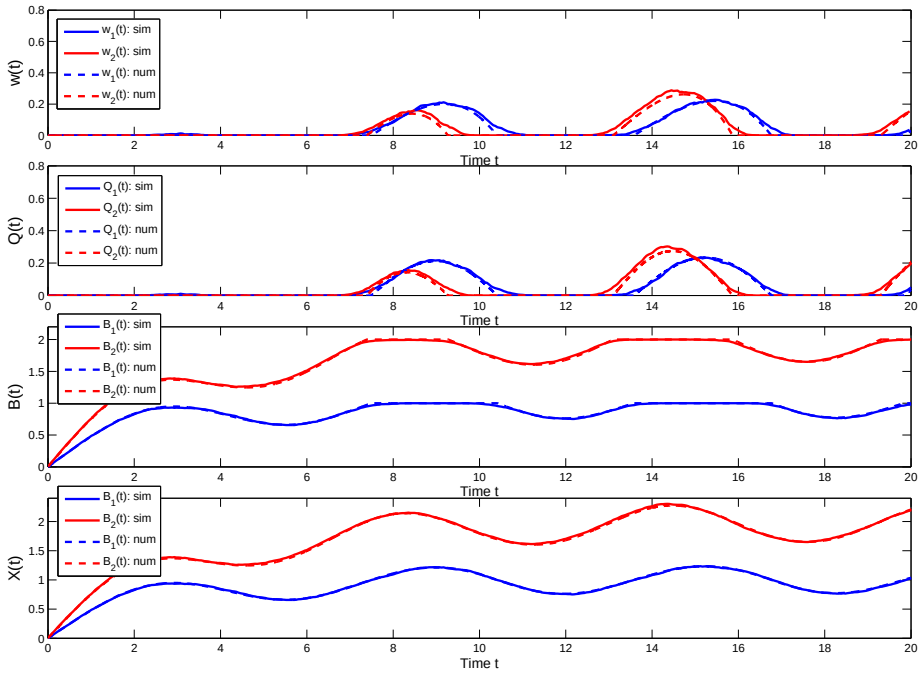

Figure 3.5: A comparison of the $(M_t/M/s_t + M)^2/M_t$ network fluid model with a simulation run averaging 50 independent sample paths, $n = 100$.

large stochastic fluctuations. However, the mean functions of these processes can be well approximated. In Figure 3.5 we estimate and means by averagine multiple independent sample paths and compare them with the fluid functions for the case $n = 100$. Therefore, the fluid approximation is still quite accurate when the system is not in a large scale.

3.9.2 A $(G_t/LN/s_t + E_2)^2/M_t$ non-Marvovian Example

We now evaluate the performance of the FPE based algorithm introduced in §3.8. We consider a non-Marvovian example: the $(G_t/LN/s_t + E_2)^2/M_t$ model with a Lognormal service distribution (the LN) and an Erlang-2 patience distribution (the E_2). Specifically, we let the service time at station i be $S_i \equiv e^{Z_i}$, where Z_i is a Normal random variable with

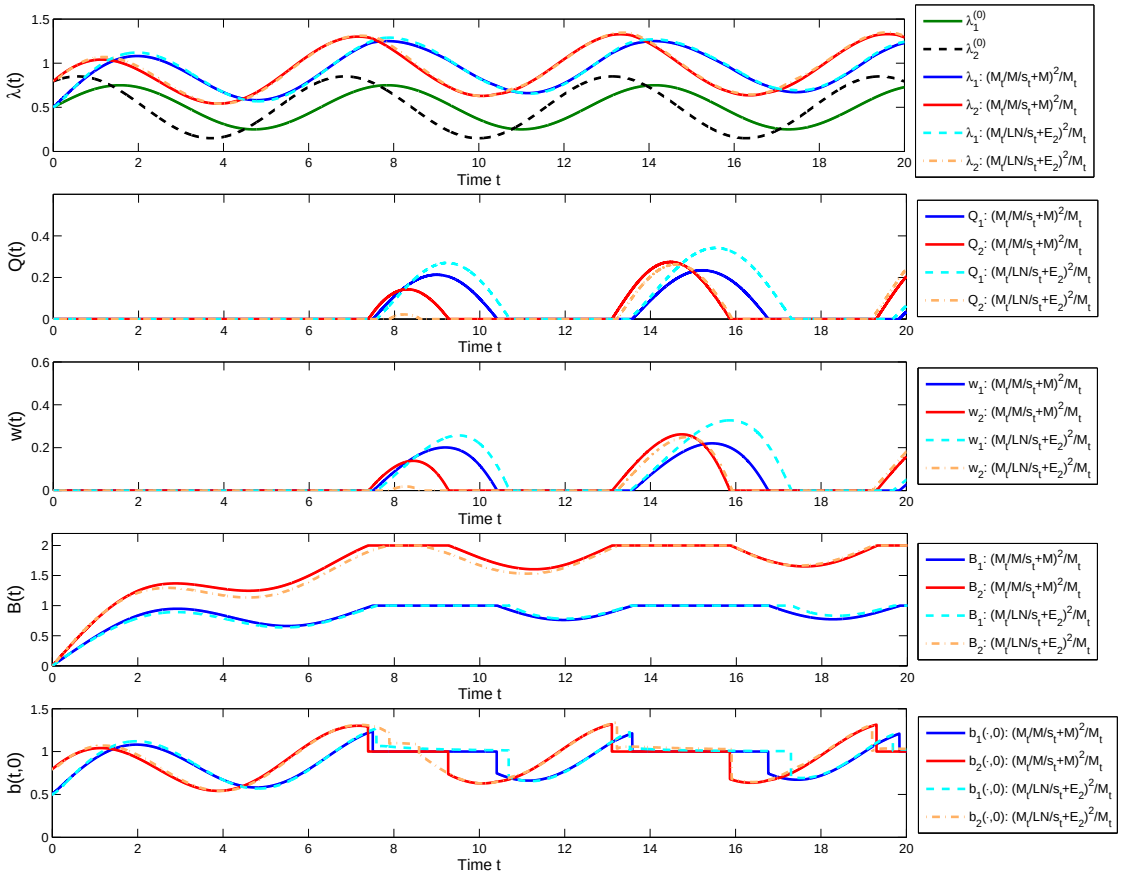

Figure 3.6: Computing the fluid performance functions for the $(M_t/LN/s_t + E_2)^2/M_t$ network fluid model.

mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$, i.e., $Z_i \sim N(\hat{\mu}_i, \hat{\sigma}_i^2)$, $i = 1, 2$. The service pdf is

$$g_i(x) = \frac{1}{x\hat{\sigma}_i\sqrt{2\pi}}\, e^{-\frac{(\log x - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}}, \quad x \geq 0, \quad i = 1, 2.$$

The mean service times and the variances are

$$\begin{aligned} \frac{1}{\mu_i} \equiv E[S_i] &= e^{\hat{\mu}_i + \frac{1}{2}\hat{\sigma}_i^2}, \\ \sigma_i^2 \equiv Var(S_i) &= (e^{\hat{\sigma}_i^2} - 1)\, e^{2\hat{\mu}_i + \hat{\sigma}_i^2}, \quad i = 1, 2. \end{aligned}$$

We let the patience distribution be Erlang-2 (E_2) with pdf

$$f_i(x) = 4\theta_i^2 x\, e^{-2\theta_i x}, \quad x \geq 0.$$

Let A_i be a generic patience time of a customer at queue i, we have $E[A_i] = 1/\theta_i$, $i = 1, 2$.

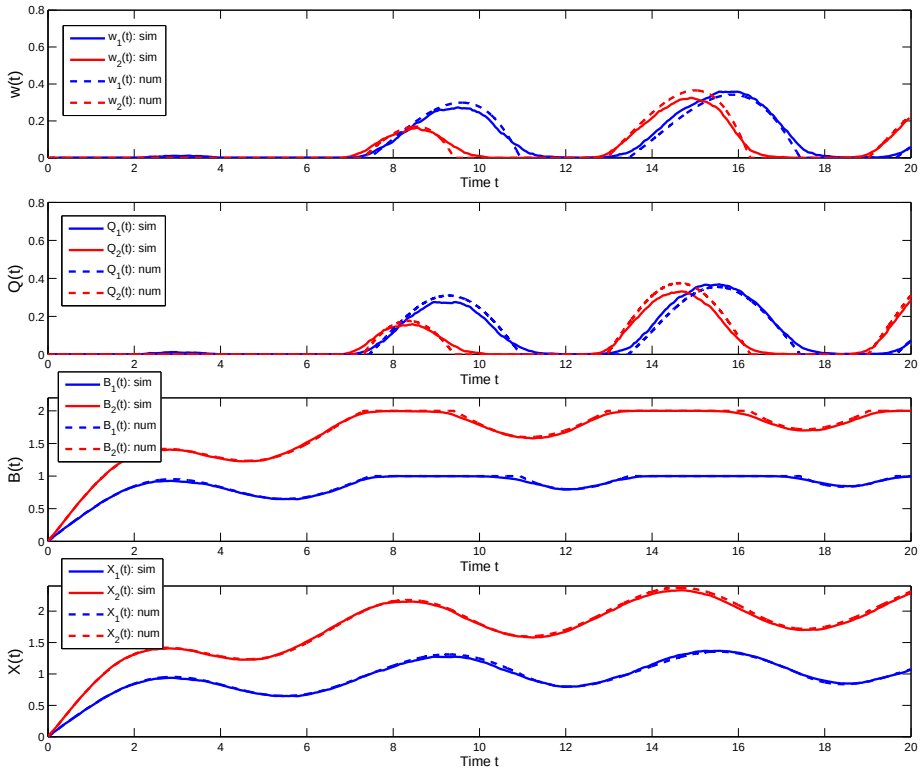

Figure 3.7: A comparison of the $(M_t/LN/s_t + E_2)^2/M_t$ network fluid model with a simulation run averaging 50 independent sample paths, $n = 100$.

The E_2 distribution has a squared coefficient of variation $c^2 \equiv Var(X)/E[X]^2 = 1/2$. We choose $\hat{\mu}_1 = -0.549$, $\hat{\sigma}_1 = 1.048$, $\hat{\mu}_2 = 0.144$, $\hat{\sigma}_2 = 1.048$ such that $\mu_1 = 1$, $\mu_2 = 0.5$, $\sigma_1^2 = 2$, $\sigma_2^2 = 8$. Thus, we have $c^2 = 2$ for the service distributions. We let $\theta_1 = 0.5$, $\theta_2 = 0.3$. In this way both the service rates (μ_1 and μ_2) and the patience rates (θ_1 and θ_2) remain the same as in the example in §3.9.1. For comparison purpose, we let the external

arrival rate $\lambda^{(0)}$ be sinusoidal (as in (3.39) and the Markovian routing matrix $\mathbf{P}$ be constant (as in (3.40)) with the same parameters there. We also let the system be initially empty.

We again plot the standard performance measures and compare them with simulation experiments in Figure 3.6 and 3.7 respectively, these two figures are analogs of Figure 3.3 and 3.5. In Figure 3.6, we plot and compare the fluid functions of the $(M_t/M/s_t+M)^2/M_t$ model (the solid lines: blue for Queue 1 and red for Queue2) and those of the $(M_t/LN/s_t+E_2)^2/M_t$ model (the dashed lines: lightblue for Queue 1 and lightbrown for Queue2). As we have described above, these two models have the same model parameters (including the service and patience rates μ and θ) except for the service and patience distributions. Figure 3.6 delivers an important message: unlike the stationary $G/GI/s+GI$ queue, the service and patience distributions beyond their means play an important role for the fluid network with time-varying model parameters; the transient system behavior can be significantly different if we change the service or patience distribution. Figure 3.7 verifies the effectiveness of the fluid approximations to the performance of the corresponding stochastic queue networks.

Finally, we end this section with a few remarks on the performance of these algorithms.

Remark 3.4 $(\textit{Performance of the algorithms with respect to } T, \Delta T, \textit{ and } \epsilon)$ (i) *The complexity of these algorithms is linear in the length of the interval* T. (ii) *The complexity is almost independent of the step size* ΔT. *The reason is intuitive: if* ΔT *is big, the algorithm reaches the end of the time horizon in less steps while the numerical computation in each interval of length* ΔT *takes more time; if* ΔT *is small, it takes more steps for the algorithm to advance in time while the numerical computation of each step becomes simpler.* (iii) *The way how the total number of iterations* $N(\epsilon)$ *of the FPE operator depends on the ETP* ϵ *is*

similar to the case of Table 1. Again, this is so because of the contraction property of these operators. In conclusion, the running time $\mathcal{T} = O(mT\log(1/\epsilon))$.

Remark 3.5 (*Comparison of the algorithms*) *On the running time of these three algorithms, the ordering is*

$$\textit{Algorithm 4} < \textit{Algorithm 3} < \textit{Algorithm 5}.$$

Consider the $(M_t/M/s_t + M)^2/M_t$ *example in §3.9.1 with* $T = 20$, $\Delta T = 0.5$, *and* $\epsilon = 10^{-5}$, *the running times are 44 (for Algorithm 4), 72 (Algorithm 3), and 118 (Algorithm 5) seconds. On the complexity of the implementation, the ordering is*

$$\textit{Algorithm 3} < \textit{Algorithm 4} < \textit{Algorithm 5}.$$

It is clear that when treating the $(G_t/M_t/s_t + GI)^m/M_t$ *model, Algorithm 4 runs with the least time and Algorithm 3 is the easiest to implement. However, to analyze the* $(G_t/GI/s_t + GI)^m/M_t$ *model, we have to use Algorithm 5 although it is the worst both in running time and in implementation complexity.*

3.10 The Stationary $(G/GI/s + GI)^m/M$ Fluid Network

This chapter is primarily devoted to the time-varying fluid queue network, but the corresponding stationary fluid queue network also is of interest. The stationary performance of a single $GI/GI/s + GI$ fluid queue was characterized in [77]. (The proof is completed by Chapter 2 because the transient dynamics are characterized there.) The corresponding stationary $(G/GI/s + GI)^m/M$ fluid queue network is actually quite elementary given [77].

In particular, the stationary performance of this model is determined by a fixed point equation for the (now constant) arrival rates. We start by reviewing that stationary distribution of the $GI/GI/s+GI$ fluid queue.

Theorem 3.11 (*steady state of the $G/GI/s+GI$ fluid queue, from* [77]). *The $G/GI/s+GI$ fluid model specified with model parameter vector (λ, s, μ, G, F) has a unique steady state described by the vector $(b, q, B, Q, w, \sigma, \alpha)$, whose character depends on whether $\rho \equiv \lambda/s\mu \leq 1$ or $\rho > 1$.*

(a) *Underloaded and balanced cases: $\rho \leq 1$. If $\rho \leq 1$, then for $x \geq 0$*

$$B = s\rho, \quad b(x) = \lambda \bar{G}(x), \quad \sigma = B\mu = \lambda, \quad Q = \alpha = w = q(x) = 0,$$

(b) *Overloaded case: $\rho > 1$. If $\rho > 1$, then for $x \geq 0$*

$$B = s, \quad b(x) = s\mu \bar{G}(x), \quad \sigma = s\mu, \quad \alpha = \lambda - s\mu = (\rho - 1)s\mu = \lambda \bar{F}(w),$$
$$w = F^{-1}\left(1 - \frac{1}{\rho}\right), \quad Q = \lambda \int_0^w \bar{F}(x)dx \quad \textit{and} \quad q(x) = \lambda \bar{F}(x) 1_{\{0 \leq x \leq w\}}.$$

We now turn to the arrival rates. As can be seen from Theorem 3.11 above, unlike for the time-varying model, for the stationary model we can easily handle GI service, because the total service content B is independent of the service-time distribution beyond its mean. The vector of constant arrival rates λ is determined by the system of fixed point equations

$$\lambda_j = \lambda_j^{(0)} + \sum_{i=1}^{m} (\lambda_i \wedge s_i\mu_i) P_{i,j}, \quad 1 \leq j \leq m, \tag{3.41}$$

where $\lambda, \lambda^{(0)}, s, \mu \in \mathbb{R}^m$ and P is an $m \times m$ stochastic matrix. We can write (3.41) more compactly as

$$\lambda = \Phi(\lambda) \equiv \lambda^{(0)} + (\lambda \wedge s\mu)P. \tag{3.42}$$

Equation (3.42) was already analyzed by [23] in the study of non-ergodic Jackson networks; also see [9] and p. 168 of [10]. However, the model here is different.

Theorem 3.12 *(fixed point equation for stationary arrival rates, from* [23]*) The arrival rates in the stationary $(G/GI/s + GI)^m/M$ fluid queue network satisfy equation* (3.41)*. Hence, if the stochastic matrix has spectral radius less than* 1 *(which holds if and only if $P^n \to 0$ as $n \to \infty$), then Φ in* (3.42) *is a monotone n-stage contraction operator on $\mathbb{R}^m$ with an appropriate norm, so that there exists a unique solution to the fixed point equation in* (3.41) *and* (3.42)*. The fixed point can be calculated by solving at most m different systems of m linear equations.*

Proof. Even for GI service, if fluid queue i is underloaded, then the stationary service content is $B_i = \lambda_i/\mu_i$ and the service completion rate is $\sigma_i = B_i\mu_i = \lambda_i$. On the other hand, if queue i is overloaded, then $B_i = s_i$ and the service completion rate is $s_i\mu_i$. In all cases, the service completion rate at queue i is $\lambda_i \wedge s_i\mu_i$. Since there is a unique solution to equation (3.41) or (3.42), that equation determines the stationary arrival rates at all queues and which queues are in fact overloaded. ∎

3.11 Conclusions

In section 3.2 we specified the single $G_t/M_t/s_t + GI_t$ fluid queue; it differs from Chapter 2 by having M_t service and GI_t abandonment instead of both being GI. The M_t service

eliminates the need to solve a fixed point equation to find the service content density b. In §3.3 and §3.4 we showed that a single fluid queue can be analyzed by assuming that the arrival rate function λ, the staffing function s and the service rate function μ are all piecewise polynomials. However, that did not permit an extension to networks because the departure rate function does not inherit that property. In §3.5 we used asymptotic methods to show how to analyze the single fluid queue without having to assume either (i) that the arrival rate function is piecewise polynomial or (ii) that there are only finitely many switches between overloaded and underloaded intervals in each finite interval. In §3.7 we provided (i) an FPE based algorithm and (ii) an ODE based algorithm to compute all standard performance functions for the $(G_t/M_t/s_t + GI_t)^m/M_t$ network in a finite time interval. In §3.8 we extend our analysis to the fluid network with GI service. We provided the theoretical basis and a new algorithm for the generalized model. In §3.9 we evaluated the performance of these algorithms described in §§3.6-3.8 with Markovian and non-Markovian examples. We conducted simulation experiments showing that the fluid model provides very accurate approximations for very large-scale many-server queueing systems. The approximations are also excellent for the *mean values* of the corresponding queueing random variables when the scale is quite small, e.g., when there are 100 servers or fewer In §3.10 we treated the stationary $(G/GI/s + GI)^m/M$ networks with constant model data and proportional routing. Theorem 3.8 established the existence of unique vector of arrival rate functions, allowing for feedback, and thus a corresponding unique performance description for the entire network. The performance functions at each queue are given in §3.4.

There are many directions for future research. It remains to establish supporting many-server heavy-traffic limits, including stochastic refinements. it remains to extend Theorem 3.8 to GI and GI_t service. It remains to develop alternative approximations for time-varying many-server queueing systems, where the staffing adjusts dynamically (appropri-

ately) to the time-varying demand, so that the system tends to be critically loaded at all times, as opposed to switching between overloaded intervals and underloaded intervals.

Chapter 4

Large-Time Asymptotics for the $G_t/M_t/s_t + GI_t$ Fluid Queue

We next focus on the fluid model with exponential service distribution. We allow all model parameters to be time dependent. Complementing Chapters 2-3 that investigated the transient dynamics in a finite interval, here we study the large-time asymptotic behavior of the fluid model. When the model parameters are periodic, we show that the performance functions converge to a periodic steady state (PSS); when the model is stationary with constant parameters, we establish the convergence to the conventional steady state.

4.1 Introduction

In Chapters 2-3 we investigated the deterministic $G_t/GI/s_t+GI$ and $(G_t/M_t/s_t+GI_t)^m/M_t$ fluid models with time-varying parameters. There we provided efficient algorithms to compute system performance formulas in finite time intervals. Complementing Chapters 2-3,

in this chapter we study the large time asymptotic behavior of the $G_t/M_t/s_t + GI_t$ fluid model. We focus on the impact of the initial conditions on the system performance as time evolves. To treat the general nonstationary setting, we show that, under regularity conditions, an initial difference in the state variables dissipates over time, i.e., the large-time behavior is asymptotically independent of the initial conditions; we call this the *asymptotic loss of memory* (ALOM) property. For non-stationary Markov processes, ALOM has been called *weak ergodicity* [33], Ch. V. We also quantify the rate of convergence (which is at the magnitude of the abandonment and service rates), showing that it is exponentially fast, again under regularity conditions. This fast convergence result also justifies the usefulness of approximating transient dynamics with steady-state performance.

This ALOM property can be quite useful. First, we apply ALOM to establish the existence of a unique steady state in *stationary* fluid models (that have constant model parameters), and convergence to that steady state as time evolves. Although the existence and form of this steady state were established in [77], the convergence from transient system dynamics to this steady state (and the rate of the convergence) has never been shown before to the best of our knowledge. We also employ ALOM to establish the existence of a unique *periodic steady state* (PSS) in *periodic* fluid models (that have periodic model parameters), and convergence to this PSS as time evolves. This PSS can be very useful to determine system congestion in service systems with daily or weakly cycles. We use the algorithm developed in Chapters 2-3 to compute performance functions over initial intervals. Since convergence is exponentially fast, that directly yields the PSS performance, but we also develop an alternative direct algorithm to compute the PSS performance.

The specific fluid model we consider here is $G_t/M_t/s_t + GI_t$. That model is placed on a firm mathematical foundation in §3.4 of Chapter 3; it is a relatively minor modification of the corresponding $G_t/GI/s_t + GI$ fluid model introduced and analyzed in Chapter 2. The performance of the $G_t/M_t/s_t + GI_t$ model is characterized in §§3.2-3.4 of Chapter

3, building on §§2.5-2.9 of Chapter 2. Regularity conditions were developed under which all the standard performance functions are characterized. Moreover, an algorithm was developed to compute these performance functions. We will draw heavily upon this previous material.

The special case of the $G_t/M/s_t + GI$ fluid queue, where only the arrival rate and staffing function (number of servers) are time-varying, should be adequate for most applications. The most useful generalization then would be to allow GI service instead of M service. With GI service, the fluid content density in service, $b(t, x)$ (see (2.3) and (3.9) below) during an overloaded interval depends on the prior values of the rate fluid enters service, $\{b(s, 0) : 0 \leq s \leq t\}$, (see equation (2.16) of Chapter 2), and Theorem A.2 of Chapter 2 shows that $b(t, 0)$ is characterized as the solution of a fixed point equation ((4.20) in Chapter 2). Here we exploit the fact that, with M_t service, the density of fluid in service $b(t, x)$ can be exhibited explicitly. We *conjecture* that ALOM extends to $G_t/GI/s_t + GI$ models with non-exponential service times, provided that all the regularity conditions in Chapter 2 are satisfied, including the service-time distribution having a density.

In fact, in Chapter 5 we provide a counterexample showing that ALOM does *not* extend beyond M_t service to *all* GI service. Indeed, we show in Chapter 5 that ALOM does not hold even in all stationary fluid models. That is done by considering the $GI/D/s + GI$ fluid model with deterministic service times. Of course, the deterministic service-time distribution does not satisfy the density condition in Chapter 2 and [77]. Nevertheless, the $G/D/s + GI$ fluid queue has the stationary performance given in [77] and Theorem 3.11 here. However, the performance does not converge to that stationary value when the system starts empty. Instead, it approaches a PSS. The same phenomenon occurs for two-point service-time distributions when one point is 0, but otherwise we *conjecture* that ALOM extends to all many-server fluid queues in which service-time distributions are neither deterministic nor exponential.

Here is how the rest of this chapter is organized: In §4.2 we review comparison and Lipschitz continuity results from Chapter 3 that we will apply, and we establish a new boundedness lemma, Lemma 4.1. In §4.3 we establish ALOM. In §4.4 we show that the transient performance of the stationary $G/M/s + GI$ fluid queue converges to its steady state performance. In §4.5 we establish the existence of a unique PSS and convergence to it in the periodic $G_t/M_t/s_t+GI_t$ queue. We draw conclusions in §4.6. Additional supporting material appears in Appendix C, including comparisons with simulations of corresponding stochastic queueing systems.

4.2 Structural Results

The model definition, assumptions, and performance formulas for the $G_t/M_t/s_t+GI_t$ fluid model are described in §§3.2-3.4 of Chapter 3. In this section we highlight three structural results that we will apply here to establish the ALOM result in §4.3, two from Chapter 3 and one new.

The first structural result is the *fundamental comparison result* established in Theorem 3.5 of Chapter 3. This result establishes an ordering of all performance functions in two fluid queues given an assumed ordering for the model data functions λ, h_F, $B(0)$, and $q(0, \cdot)$. See Theorem 3.5 for details.

The second is the *Lipschitz continuity result* established in Theorem 3.6 of Chapter 3. This result applies to the fluid content functions (e.g., B, Q, and X), it bounds their absolute uniform differences (in $[0, T]$) of two fluid queues by those of the two models' data functions λ, $B(0)$, $Q(0)$ and $X(0)$. See Theorem 3.6 for details.

We now add a new structural result: boundedness. For this elementary boundedness result and other results to follow, we make a stronger assumption on the staffing and the rates in the model data, requiring that they be uniformly bounded above and below. Our

conditions will involve the maximum rate fluid can enter service: γ in (5.25) as well as the two-parameter abandonment hazard rate $h_{F_t}(y) \equiv f_t(y)/\bar{F}_t(y)$, defined after (5.25). Let

$$\begin{aligned} h^{\uparrow}_{F_T} &\equiv \sup_{-\infty<t\leq T, x\geq 0} h_{F_t}(x), \quad h^{\downarrow}_{F_T} \equiv \inf_{-\infty<t\leq T, x\geq 0} h_{F_t}(x), \\ \bar{F}^{\uparrow}(x) &\equiv \sup_{-\infty<t<\infty} \bar{F}_t(x), \quad \bar{F}^{\downarrow}(x) \equiv \inf_{-\infty<t<\infty} \bar{F}_t(x). \end{aligned}$$

Assumption 4.1 (*uniformly bounded staffing and rates*) *The staffing and the rates in the model data are uniformly bounded above and below, i.e.,*

$$\begin{aligned} \lambda^{\uparrow}_{\infty} &< \infty, \quad \mu^{\uparrow}_{\infty} < \infty, \quad s^{\uparrow}_{\infty} < \infty, \quad \gamma^{\uparrow}_{\infty} < \infty, \quad h^{\uparrow}_{F_\infty} < \infty \\ \lambda^{\downarrow}_{\infty} &> 0, \quad \mu^{\downarrow}_{\infty} > 0, \quad s^{\downarrow}_{\infty} > 0, \quad \gamma^{\downarrow}_{\infty} > 0, \quad h^{\downarrow}_{F_\infty} > 0. \end{aligned}$$

Assumption 4.1 repeats Assumption 2.11 and strengthens Assumptions 2.10 and 3.6.

We also assume a further regularity condition on the abandonment cdf's.

Assumption 4.2 (*abandonment cdf tail*) $\bar{F}^{\uparrow}(x) \to 0$ *as* $x \to \infty$.

We assume that these two additional assumptions are in force for the remainder of the chapter. Our boundedness result also exploits the finite initial conditions, provided by Assumption 2.1.

Lemma 4.1 (*boundedness*) *Under the assumptions above, all performance functions are*

uniformly bounded. In particular,

$$\begin{aligned} B(t) &\leq s(t) \leq s^{\uparrow}_{\infty}, \quad b(t,x) \leq b(0,x) \vee \lambda^{\uparrow}_{\infty} \vee \gamma^{\uparrow}_{\infty}, \\ Q(t) &\leq \left(\frac{\lambda^{\uparrow}_{\infty}}{h^{\downarrow}_{F_\infty}}\right) \vee Q(0), \quad q(t,x) \leq q(0,x) \vee \lambda^{\uparrow}_{\infty}, \\ w(t) &\leq (\bar{F}^{\uparrow})^{-1}\left(\frac{\gamma^{\downarrow}_{\infty}}{\lambda^{\uparrow}_{\infty}}\right) \vee \left(\frac{Q(0)}{\gamma^{\downarrow}_{\infty}} + w(0)\right), \\ \alpha(t) &\leq \frac{h^{\uparrow}_{F_\infty}\,\lambda^{\uparrow}_{\infty}}{h^{\downarrow}_{F_\infty}}, \quad \textit{and} \quad \sigma(t) \leq \mu^{\uparrow}_{\infty}\, s^{\uparrow}_{\infty}. \end{aligned}$$

Proof. Most are elementary; only $Q(t)$ and $w(t)$ require detailed argument. Flow conservation in (3.5) implies that $Q'(t) = \lambda(t) - \alpha(t) - \gamma(t) \leq \lambda^{\uparrow}_{\infty} - \alpha(t)$. Since $\alpha(t) \geq h^{\downarrow}_{F_\infty}\, Q(t)$, we have $Q'(t) < 0$ whenever $Q(t) > \lambda^{\uparrow}_{\infty}/h^{\downarrow}_{F_\infty}$. The bound for $w(t)$ follows directly from (4.5) and the final part of the proof of Theorem 4.1 below, which does not use the present lemma. ■ □

4.3 Asymptotic loss of Memory (ALOM)

In this section we establish ALOM for the $G_t/M_t/s_t + GI_t$ fluid model. We start with an illustrative example.

Example 4.1 (*a sinusoidal* $G_t/M/s + M$ *example*) Consider a $G_t/M/s + M$ fluid queue that has the sinusoidal arrival rate function

$$\lambda(t) = a + b \cdot \sin(c\,t), \tag{4.1}$$

with $a = c = 1$ and $b = 0.6$, exponential service distribution with rate $\mu = 1$, constant staffing function $s = 1$, and exponential abandonment time distribution with rate $\theta = 0.5$. Applying the algorithm in Chapter 2, we compute and compare the performance measures $w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$ with four different (ordered) initial conditions: the system is initially (i) empty with $Q(0) = B(0) = 0$ (the yellow solid lines), (ii) UL with $Q(0) = 0$, $B(0) = 0.5 < 1 = s$ (the dark dashed lines), (iii) OL with $Q(0) = 0.4$, $B(0) = 1 = s$ (the light-blue dashed lines) and (iv) OL with $Q(0) = 0.8$, $B(0) = 1 = s$ (the red dotted lines), as shown in Figure 4.1.

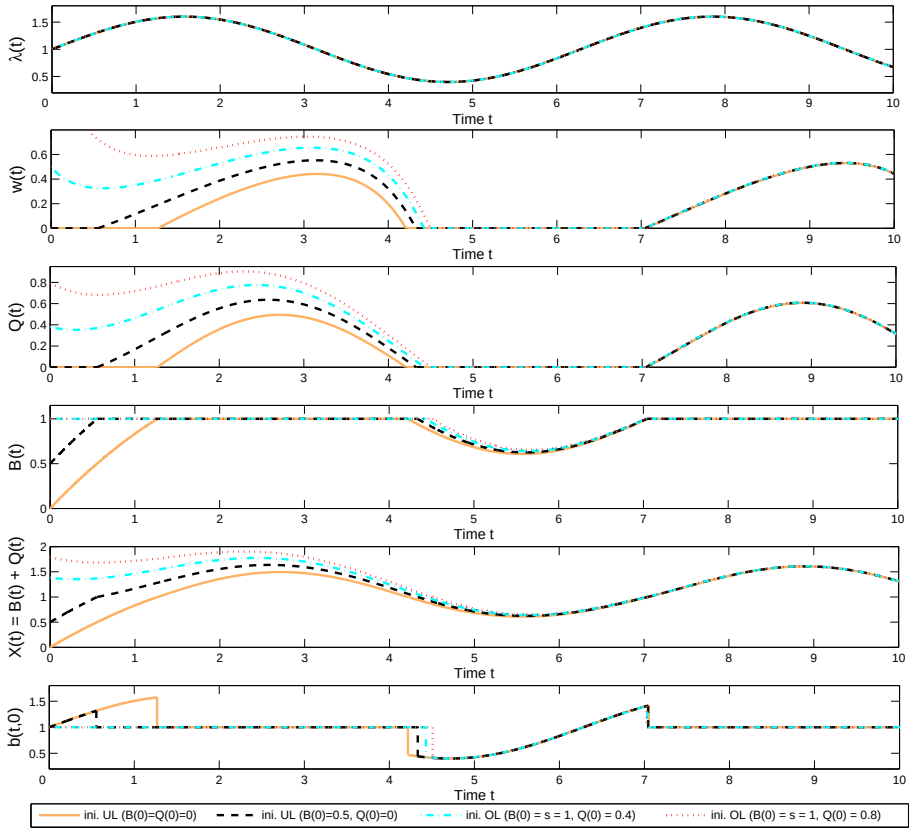

Figure 4.1: The performance measures for the $G_t/M/s + M$ model in Example 4.1 with four different (ordered) initial conditions.

Figure 4.1 shows that the differences in these four cases converge to zero so fast that it looks as if the distance becomes 0 after finite time (but that actually never occurs), even though the initial conditions are dramatically different. Figure 4.1 also illustrates the comparison result in Theorem 3.5. ∎

To state our ALOM result, we use Δ to denote absolute difference. Specifically, for real-valued functions X_i on $[0,\infty)$, $i = 1, 2$, and $0 < T \leq \infty$, let $\Delta X_{1,2}(t) \equiv \Delta X(t) \equiv |X_1(t) - X_2(t)|$, $t \geq 0$.

Theorem 4.1 (asymptotic loss of memory) *Consider two $G_t/M_t/s_t + GI_t$ fluid models with common arrival rate function λ, service rate function μ, staffing function s, and time-varying abandon-time cdf's F_t, but different initial conditions (satisfying Assumption 2.1). Then* (a)

$$\Delta X(T) \;\leq\; C_1 e^{-C(T)} \quad \textit{for} \quad C(T) \equiv T\,(\mu_T^{\downarrow} \wedge h_{F_T}^{\downarrow}), \tag{4.2}$$

where $C_1 \equiv C_1(B_1(0), B_2(0), q_1(0,\cdot), q_2(0,\cdot))$ is the constant

$$\begin{aligned} C_1 \;&\equiv\; \Delta B(0) + \int_0^\infty ([q_1(0,x) \vee q_2(0,x)] - [q_1(0,x) \wedge q_2(0,x)])\,dx \\ &\leq\; \Delta B(0) + Q_1(0) + Q_2(0). \end{aligned} \tag{4.3}$$

Moreover,

$$\Delta\alpha(T) \leq h_{F_T}^{\uparrow} C_1 e^{-C(T)} \quad \textit{and} \quad \Delta\sigma(T) \leq \mu_T^{\uparrow} C_1 e^{-C(T)} \tag{4.4}$$

for all $T > 0$*. Hence, for* $C_2 \equiv \mu_\infty^\downarrow \wedge h_{F_\infty}^\downarrow > 0$ *and all* $T > 0$*,*

$$\Delta X(T) \;\leq\; C_1 e^{-C_2 T}, \quad \Delta\alpha(T) \leq h_{F_\infty}^\uparrow C_1 e^{-C_2 T} \quad \textit{and} \quad \Delta\sigma(T) \leq \mu_\infty^\uparrow C_1 e^{-C_2 T}.$$

In addition, for each $T > 0$*,*

$$\begin{aligned} \Delta w(T) &\leq \frac{\Delta X(T)}{\lambda_T^\downarrow \bar{F}^\downarrow(w_1(T) \vee w_2(T))} \\ &\leq C_3 \Delta X(T) \leq (C_3 C_1) e^{-C_2 T}, \end{aligned} \tag{4.5}$$

where

$$C_3 \equiv (\bar{F}^\uparrow)^{-1}(s_\infty^\downarrow \mu_\infty^\downarrow / \lambda_\infty^\uparrow) \vee \left((w_1(0) \vee w_2(0)) + \frac{Q_1(0) + Q_2(0)}{s_\infty^\downarrow \mu_\infty^\downarrow} \right). \tag{4.6}$$

(b) *If, in addition, the initial content is ordered by*

$$X_1(0) \leq X_2(0) \quad \textit{and} \quad q_1(0,x) \leq q_2(0,x) \quad \textit{for all} \quad x \geq 0, \tag{4.7}$$

then $X_1(t) \leq X_2(t)$ *for all* $t \geq 0$*,*

$$\Delta X'(T) \;\leq\; 0 \quad \textit{and} \quad \Delta X(T) \leq \frac{\Delta X(0)}{1 + C(T)}, \quad T > 0, \tag{4.8}$$

for $C(T)$ in (4.2)*, so that*

$$\begin{aligned} \Delta X(T) &\leq e^{-C(T)}\Delta X(0), \\ \Delta\alpha(T) &\leq h^{\uparrow}_{F_T}\Delta X(T) \quad \textit{and} \quad \Delta\sigma(T) \leq \mu^{\uparrow}_{T}\Delta X(T). \end{aligned} \tag{4.9}$$

Proof. We first show that (a) follows from (b). Without loss of generality, we have $X_1(0) \leq X_2(0)$. Then $X_1(0) \leq X_2(0)$ is equivalent to $B_1(0) \leq B_2(0)$ and $Q_1(0) \leq Q_2(0)$. In order to derive (a) from (b), construct another two systems, 3 and 4, with $q_3(0,x) \equiv q_1(0,x) \vee q_2(0,x)$, $B_3(0) \equiv B_1(0) \vee B_2(0)$, $q_4(0,x) \equiv q_1(0,x) \wedge q_2(0,x)$ and $B_4(0) \equiv B_2(0) \wedge B_2(0)$. With this construction, systems 3 and 4 are bonafide fluid models, with $X_4(t) \leq X_1(t) \leq X_3(t)$ and $X_4(t) \leq X_2(t) \leq X_3(t)$ for all t, which implies that $\Delta X_{1,2}(t) \leq \Delta X_{3,4}(t)$ for all t. Since $\Delta X_{3,4}(0) \leq C_1$ for C_1 in (4.3), (4.2) in (a) follows from (4.9) for $\Delta X_{3,4}(t)$. (The final bound on C_1 in (4.3) arises when the supports of $q_1(0,\cdot)$ and $q_2(0,\cdot)$ are disjoint sets, which actually is not allowed by Assumption 2.10, but can be approached.)

Now we prove (b). Observe that (4.9) follows (4.8) because dividing the interval $[0,T]$ into N subintervals yields

$$\Delta X(T) \leq \left(\frac{1}{1+\frac{T}{N}\left(\mu^{\downarrow}_{T} \wedge h^{\downarrow}_{F_T}\right)}\right)^{N} \Delta X(0).$$

Letting $N \to \infty$, we get (4.9).

We now prove (4.8). With the ordering assumed in (4.7), all functions in the two systems can be ordered according to Theorem 3.5. Hence, there are only three cases: (i) both systems are UL; (ii) both systems are OL; (iii) system 1 is UL and system 2 is OL. We treat the three cases separately and use mathematical induction to show (4.8).

In case (i) we have $B_1(0) \leq B_2(0) \leq s(0)$ and $Q_1(0) = Q_2(0) = 0$. Let T^* be the underload termination time of system 2. For $0 \leq t < T^*$, neither system changes regime. Observe that $\Delta X(t) = \Delta B(t)$. Flow conservation implies that

$$B_i'(t) \;\; = \;\; \lambda(t) - \mu(t)\, B_i(t) \quad \text{for} \quad i = 1, 2,$$

which yields

$$\Delta X'(s) = \Delta B'(s) = -\mu(s)\, \Delta B(s) \leq -\mu_t^{\downarrow}\, \Delta B(t) = -\mu_t^{\downarrow}\, \Delta X(t), \qquad 0 \leq s \leq t,$$

where the inequality follows from $\mu(s) \geq \mu_t^{\downarrow}$ and $\Delta B(s) \geq \Delta B(t)$ since $\Delta B(s)$ has negative derivative. Therefore, we have

$$\Delta X(t) - \Delta X(0) \leq -\mu_t^{\downarrow}\, t\, \Delta X(t)$$

and

$$\Delta X(t) \leq \left(\frac{1}{1 + \mu_t^{\downarrow}\, t} \right) \Delta X(0). \tag{4.10}$$

In case (ii) we have $B_1(0) = B_2(0) = s(0)$ and $q_1(0, \cdot) \leq q_2(0, \cdot)$. Let T^* be the overload termination time of system 1. For $0 \leq t < T^*$, neither system changes regime. Observe that $\Delta X(t) = \Delta Q(t)$. Theorem 3.5 implies that $q_1(t, \cdot) \leq q_2(t, \cdot)$ and $w_1(t) \leq$

$w_2(t)$ for $\leq t \leq T^*$. Therefore, we have

$$\begin{aligned}
\alpha_2(t) - \alpha_1(t) &= \int_0^{w_2(t)} q_2(t,x)\, h_{F_{t-x}}(x)dx - \int_0^{w_1(t)} q_1(t,x)\, h_{F_{t-x}}(x)dx \\
&= \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_{F_{t-x}}(x)dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x)\, h_{F_{t-x}}(x)dx \\
&\geq h^{\downarrow}_{F_t} \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x))dx + h^{\downarrow}_{F_t} \int_{w_1(t)}^{w_2(t)} q_2(t,x)dx \\
&= h^{\downarrow}_{F_t} (Q_2(t) - Q_1(t)) = h^{\downarrow}_{F_t}\, \Delta Q(t). \qquad (4.11)
\end{aligned}$$

Flow conservation implies that

$$Q_i'(t) = \lambda(t) - \alpha_i(t) - \gamma(t) \quad \text{for} \quad i = 1, 2,$$

which yields

$$\begin{aligned}
\Delta X'(s) &= \Delta Q'(s) = -(\alpha_2(s) - \alpha_1(s)) \\
&\leq -h^{\downarrow}_{F_t}\, \Delta Q(s) \leq -h^{\downarrow}_{F_t}\, \Delta Q(t) = -h^{\downarrow}_t\, \Delta X(t), \quad 0 \leq s \leq t,
\end{aligned}$$

where the inequality follows from (4.11). Hence, reasoning as for (4.10) in case (i), we have

$$\Delta X(t) \leq \left(\frac{1}{1 + h^{\downarrow}_{F_t}\, t} \right) \Delta X(0). \qquad (4.12)$$

In case (iii) we have $B_1(0) \leq s(0) = B_2(0)$ and $Q_1(0) = 0 \leq Q_2(0)$. Let $T^* \equiv T_1 \wedge T_2$ where T_1 is the underload termination time of system 1 and T_2 is the overload termination time of system 2. For $0 \leq t < T^*$, neither system changes regime. Observe that $\Delta X(t) = \Delta B(t) + \Delta Q(t) = s(t) - B_1(t) + Q_2(t)$. Flow conservation in (3.5) implies

that the derivatives satisfy

$$\begin{aligned} Q_2'(t) &= \lambda(t) - \alpha_2(t) - \gamma(t) \\ s'(t) &= \gamma(t) - \mu(t)\, s(t) \\ B_1'(t) &= \lambda(t) - \mu(t)\, B_1(t), \end{aligned}$$

which implies that

$$\begin{aligned} \Delta X'(t) &= s'(t) - B_1'(t) + Q_2'(t) \\ &= -\alpha_2(t) - \mu(t)\,(s(t) - B_1(t)). \end{aligned} \tag{4.13}$$

Reasoning as in case (ii), we have

$$\alpha_2(t) \geq h^{\downarrow}_{F_t}\, Q_2(t) = h^{\downarrow}_{F_t}\, \Delta Q(t). \tag{4.14}$$

Therefore, (4.13) and (4.14) imply that

$$\begin{aligned} \Delta X'(s) &\leq -h^{\downarrow}_{F_t}\, \Delta Q(s) - \mu^{\downarrow}_t \Delta B(s) \\ &\leq -(h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)(\Delta Q(s) + \Delta B(s)) \\ &\leq -(h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)\Delta X(s) \leq -(h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)\Delta X(t), \quad 0 < s \leq t. \end{aligned}$$

Hence, reasoning as for (4.10) in case (i), we have

$$\Delta X(t) \leq \left(\frac{1}{1 + (h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)\, t} \right) \Delta X(0). \tag{4.15}$$

Finally, combining (4.10), (4.12) and (4.15), the desired (4.8) follows by mathematical induction.

We directly have the second and third inequalities in (4.9), which implies (4.4) because $\Delta Q(T) \le \Delta X(T)$ and $\Delta B(T) \le \Delta X(T)$.

Finally, we treat $w(t)$. As above, it suffices to assume that we have the ordering in (4.7) of (b). Then (4.5) follows from

$$\begin{aligned} \Delta X(T) &\ge \Delta Q(T) = \int_{w_1(T)}^{w_2(T)} \lambda(T-x)\,\bar{F}_{T-x}(x)dx \\ &\ge \lambda_T^{\downarrow}\bar{F}^{\downarrow}(w_2(T))\,\Delta w(T). \end{aligned} \tag{4.16}$$

We now construct w^* such that $w_2(T) \le w^*$ for all T; in general, w^* will depend on $w_2(0)$. First note that at time $T_w \equiv Q_2(0)/\mu_\infty^{\downarrow} s_\infty^{\downarrow}$, all fluid that was in queue 2 at time 0 is gone (entered service or abandoned). Choose $\bar{w} > 0$ big enough such that $\bar{F}^{\uparrow}(\bar{w}) < s_\infty^{\downarrow}\mu_\infty^{\downarrow}/\lambda_\infty^{\uparrow}$. ODE (2.31) implies that for $t > T_w$,

$$\begin{aligned} w_2'(t) &= 1 - \frac{s(t)\,\mu(t)}{\lambda(t-w_2(t))\,\bar{F}_{t-w_2(t)}(w_2(t))} \\ &\le 1 - \frac{s_\infty^{\downarrow}\,\mu_\infty^{\downarrow}}{\lambda_\infty^{\uparrow}\bar{F}^{\uparrow}(\bar{w})} < 0, \end{aligned}$$

if $w_2(t) > \bar{w}$ for some t. Hence $\bar{w}$ is an upper bound for $w_2(t)$ if $w_2(T_w) < \bar{w}$. If $w_2(T_w) \ge \bar{w}$, it is easy to see that $w_2(t)$ decreases until it is below $\bar{w}$ because we can bound $w_2'(t)$. This argument implies that $w_2(t) \le w_2^* \equiv (\bar{w} \vee (w_2(0) + T_w))$ for all $t \ge 0$. The constant C_3 in (4.5) is obtained by inserting established bounds. ∎ □

For a real-valued function x on $[0,\infty)$, let $\|x\|_1 \equiv \int_0^\infty |x(t)|\,dt$.

Corollary 4.1 *Under the conditions of Theorem* 4.1 (b)*,*

$$\|b_1(T,\cdot) - b_2(T,\cdot)\|_1 = \Delta B(T) \le \Delta X(T) \le \Delta X(0) e^{-C(T)},$$
$$\|q_1(T,\cdot) - q_2(T,\cdot)\|_1 = \Delta Q(T) \le \Delta X(T) \le \Delta X(0) e^{-C(T)}. \quad (4.17)$$

Hence, there is exponential rate of convergence under the conditions in Theorem 4.1 (a).

Remark 4.1 (*monotonicity of the difference of two queues*) *Theorem 4.1 shows that except for the densities* q *and* b*, the differences of all performance measures (*ΔX*,* $\Delta\alpha$*,* $\Delta\sigma$*, and* Δw*) of the two queues go to 0 as* $t \to \infty$*. However, even in case* (b)*, only* $\Delta X(t)$ *goes to 0 monotonically. Note that* $\Delta\alpha(t) = 0$*,* $\Delta w(t) = 0$ *and* $\Delta\sigma(t) \ge 0$ *when both queues are UL;* $\Delta\alpha(t) \ge 0$*,* $\Delta w(t) \ge 0$ *and* $\Delta\sigma(t) = 0$ *when both queues are OL.*

Remark 4.2 (*Example 4.1 revisited*) *In Example 4.1 we have* $C(T) = \mu \wedge \theta = 0.5$ *in (4.2) of Theorem 4.1,* $\lambda_\infty^\downarrow = 0.4 > 0$*,* $\lambda_\infty^\uparrow = 1.6 < \infty$*,* $\bar{F}^\downarrow(x) = e^{-\theta x} > 0$ *and* $\bar{F}^\uparrow(x) \to 0$ *as* $x \to \infty$*. Moreover,* $\zeta(t) = \lambda(t) - \mu\, s(t) - s'(t) = a - \mu\, s + b \cdot \sin(c\, t)$ *is sinusoidal so that it has finitely many zeros in any bounded interval. Therefore, all conditions in Theorem 4.1 are satisfied, establishing the exponential rate of convergence seen in Figure 4.1.*

4.4 The Stationary $G/M/s + GI$ Fluid Queue

In this section we focus on the stationary $G/M/s+GI$ fluid queue. The steady-state performance of the more general $GI/GI/s + GI$ fluid queue with GI service was characterized in [77], but the transient dynamics was only characterized completely in Chapter 2. See Theorem 3.11 of Chapter 2 and Theorem 4.4 in [77] for details. Complementing Theorem

4.4 in [77], our next result shows that the steady state given in Theorem 3.11 is indeed an invariant state, i.e., if the system is initially in this state, then it stays there forever.

Theorem 4.2 *(an invariant state for the $G/GI/s+GI$ fluid queue) Consider the $G/GI/s+GI$ fluid queue specified with model parameter (λ, s, μ, G, F). Then the steady state given in Theorem 3.11 is an invariant state. In other words, if the initial condition satisfies*

$$(b(0,\cdot), q(0,\cdot), w(0)) = (b(\cdot), q(\cdot), w),$$

that is the steady state given in Theorem 3.11, then the system stays in steady state, i.e., for all $t \geq 0$,

$$(b(t,\cdot), q(t,\cdot), B(t), Q(t), w(t), \alpha(t), \sigma(t)) = (b(\cdot), q(\cdot), B, Q, w, \alpha, \sigma),$$

that is given in Theorem 3.11.

Proof. First consider (a) with $\rho \leq 1$. By (2.9) of Chapter 2, the initial rate that service is being completed with $b(0,x) = \lambda\bar{G}(x)$ is

$$\sigma(0) = \int_0^\infty b(0,x)h_G(x)\,dx = \int_0^\infty \lambda\bar{G}(x)\frac{g(x)}{\bar{G}(x)}\,dx = \lambda. \tag{4.18}$$

If $\rho < 1$, then $B(0) = s\rho < s$ and there initially is spare capacity. If $\rho = 1$, then $\lambda(0) = \lambda = \sigma$. In both cases, the system remains UL. Hence we can apply (2.13) in Proposition 2.2 of Chapter 2 to characterize the evolution of b. For suitably small $t > 0$,

we get

$$\begin{aligned} b(t,x) &= b(t-x,0)\bar{G}(x)\,1_{\{0\le x\le t\}} + b(0,x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}\,1_{\{x>t\}} \\ &= \lambda\,\bar{G}(x)\,1_{\{0\le x\le t\}} + \lambda\,\bar{G}(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}\,1_{\{x>t\}} = \lambda\,\bar{G}(x) = b(0,x), \end{aligned}$$

which implies that the system stays UL with $b(t,x) = b(0,x)$, $B(t) = B(0)$ and $\sigma(t) = \sigma(0)$ for $t \ge 0$. For an alternative proof under the extra condition of differentiability, we can exploit the transport partial differential equation (PDE) from Appendix A.2 of Chapter 2. That tells us that $b(t,x)$ satisfies the PDE

$$\frac{\partial b}{\partial t}(t,x) + \frac{\partial b}{\partial x}(t,x) = -h_G(x)\,b(t,x),$$

which implies that

$$\begin{aligned} \frac{\partial b}{\partial t}(0,x) &= -\frac{\partial b}{\partial x}(0,x) - h_G(x)\,b(0,x) = -\frac{d(\lambda\,\bar{G}(x))}{dx} - h_G(x)\lambda\bar{G}(x) \\ &= \lambda\,g(x) - h_G(x)\bar{G}(x)\lambda = 0. \end{aligned}$$

Next consider case (b) with $\rho > 1$. We can apply (4.18) to see that the initial rate of service completion, starting with $b(0,x) = s\mu\bar{G}(x)$, is $\sigma(0) = s\mu$. Since $\rho > 1$, we necessarily have $\lambda(0) = \lambda > s\mu = \sigma(0)$. Hence, the system necessarily remains OL over a positive interval. Next we apply the fixed point equation for b during an overloaded interval. Assumption 2.8 in Chapter 2 is satisfied with this initial density $b(0,x)$ because

$$\tau(b,g,T) \equiv \sup_{0\le s\le T}\int_0^\infty \frac{b(0,y)g(s+y)}{\bar{G}(y)}\,dy = s\mu < \infty. \tag{4.19}$$

Next we observe that $b(0,x)$ satisfies the fixed point equation (4.20) of Chapter 2, i.e.,

$$b(t,0) = \hat{a}(t) + \int_0^t b(t-x,0)g(x)\,dx = s\mu\bar{G}(t) + \int_0^t b(t-x,0)g(x)\,dx, \qquad (4.20)$$

yielding $s\mu = s\mu\bar{G}(t) + s\mu G(t) = s\mu$. Theorem A.2 of Chapter 2 implies that $b(t,0) = s\mu$, $t \geq 0$, is the unique fixed point. Next Proposition 2.6 of Chapter 2 implies that the service density in queue satisfies

$$\begin{aligned} q(t,x) &= \lambda\bar{F}(x)1_{\{x\leq t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\leq w(t)\}} \\ &= \lambda\bar{F}(x)1_{\{0\leq x\leq w(t)\}}. \end{aligned} \qquad (4.21)$$

It remains to show that $w'(0) = 0$, so that $w(t) = w(0) = F^{-1}(1-(1/\rho))$. However, ODE (2.31) implies that

$$w'(0) = 1 - \frac{\gamma(0)}{q(0,w(0))} = 1 - \frac{\mu\, s}{\lambda\,\bar{F}(w(0))} = 1 - \frac{\mu\, s}{\lambda(1/\rho)} = 0,$$

where the third equality holds since $w(0) = w = F^{-1}(1-1/\rho)$. The last equality holds since $\rho = \lambda/s\mu$. Hence, $w(t) = w$ in (4.21), so that $q(t,x) = q(x)$ and all performance functions are constants for $0 \leq t \leq \delta$ for some small δ and thus for all $t \geq 0$. ■ □

Now we apply Theorem 4.1 to show that the transient performance in the $G/M/s+GI$ fluid queue with exponential service converges to the steady state described in Theorem 3.11 for any given initial conditions. As a byproduct, this establishes uniqueness for the steady-state performance in Theorem 3.11 in the special case of M service. We give two convergence results, the first obtained by directly combining Theorems 4.1 and 3.11.

Theorem 4.3 (*direct implication of ALOM*) *For the stationary $G/M/s+GI$ fluid model, as $t \to \infty$,*

$$(\alpha(t), w(t), Q(t), \sigma(t), B(t)) \quad \to \quad (\alpha, w, Q, \sigma, B), \tag{4.22}$$

$$\|q(t,\cdot) - q(\cdot)\|_1 \ \to 0 \qquad \textit{and} \quad \|b(t,\cdot) - b(\cdot)\|_1 \to 0, \tag{4.23}$$

where vector $(q(\cdot), \alpha, w, Q, b(\cdot), \sigma, B)$ is the steady-state performance in Theorem 3.11. Hence, the steady-state performance specified by Theorem 3.11 is unique.

Proof. Consider two $G/M/s+GI$ fluid queues that have identical model parameters but different initial conditions. Let system 1 be initially in the steady state given in Theorem 3.11, let system 2 have arbitrary initial condition. Theorem 3.11 implies that system 1 stays in steady state for all $t \geq 0$. Therefore, the convergence in (4.22) and (4.23) follows from ALOM in Theorem 4.1. □

We next establish a stronger convergence result, whose proof does not rely on the ALOM property in Theorem 4.1. We establish pointwise convergence of the fluid content densities b and q as $t \to \infty$ in addition to (4.22) and (4.23).

Theorem 4.4 (*more on convergence to steady state*) *Consider the stationary $G/M/s+GI$ fluid model. In addition to Assumption 2.1, assume that the initial service density satisfies*

$$\limsup_{x \to \infty} b(0,x) < \infty. \tag{4.24}$$

Then, in addition to the conclusions of Theorem 4.3*,*

$$(q(t,x), b(t,x)) \to (q(x), b(x)) \quad \textit{as} \quad t \to \infty,$$

for each $x \geq 0$*, where the limit* $(q(x), b(x))$ *is the pair of steady-state fluid densities in Theorem 3.11. Moreover, there is at most one switch between the OL and UL* (*including critically loaded*) *regimes during the convergence. More precisely, the number of switches depends on the the model parameter* $\rho \equiv \lambda/s\mu$ *and the initial conditions as shown in Table 1. If* $\rho > 1$*, there exists a* $T > 0$ *such that for* $t > T$*,* $w(t) \to w$ *monotonically, as* $t \to \infty$*. If, in addition,* $C \equiv f^{\downarrow}_{(Q(0)/s\mu)\vee w} > 0$ *where* $f^{\downarrow}_t \equiv \inf_{0\leq x\leq t} f(x)$*, then*

$$\Delta w(t) \equiv |w(t) - w| \leq \frac{1}{1 + (t-T)C}\, \Delta w(T), \quad \textit{for } t > T \tag{4.25}$$

so that

$$\Delta w(t) \leq e^{-(t-T)C} \Delta w(T), \quad t > T. \tag{4.26}$$

traffic intensity	initial condition	number of switchings
$\rho > 1$	OL	0
	UL(CL)	1
$\rho < 1$	OL	1
	UL(CL)	0
$\rho = 1$	OL	0
	UL(CL)	0

Table 4.1: How the number of switches between OL and UL intervals depends on the model parameter ρ and the initial conditions, in the setting of Theorem 4.4.

Proof. We only give the proof for the case in which the system is initially UL, i.e., $q(0,x) = w(0) = 0$ for any x and $B(0) = \int_0^\infty b(0,x)dx < s$. The other case in which the system is initially OL or critically loaded is treated in essentially the same way; the details are given in the appendix. For simplicity, we assume $\mu = s = 1$ and therefore $\rho = \lambda/s\mu = \lambda$.

(i) $\rho \leq 1$. Since the service is exponential at the fixed rate $\mu = 1$ and the staffing is fixed at $s = 1$, the maximum output rate of the service facility is 1. Hence, the system always stay in the UL regime. Thus we can apply (3.13) of Chapter 3 to characterize the density in service. By Assumption (4.24),

$$\begin{aligned} b(t,x) &= \rho e^{-x} 1_{\{0 \leq x \leq t\}} + b(0,x-t)e^{-t} 1_{\{x>t\}} \\ &\to \rho e^{-x} \quad as\ t \to \infty, \quad x \geq 0. \\ B(t) &= \int_0^t \rho e^{-x} dx + \int_t^\infty b(0,x-t)e^{-t} dx \\ &= \rho(1-e^{-t}) + e^{-t}B(0), \\ &= \rho - (\rho - B(0))\, e^{-t} \to \rho, \quad as\ t \to \infty, \end{aligned}$$

Moreover, $\sigma(t) = B(t) \to \rho$, as $t \to \infty$. If $\rho = 1$, then we obtain the monotone convergence

$$B(t) = 1 - (1 - B(0))\, e^{-t} \uparrow 1 \quad \text{as} \quad t \to \infty.$$

(ii) $\rho > 1$. As in case (i), the maximum output rate of the service facility is 1. Since $\rho > 1$, $\lambda > 1$, so that the the system necessarily will switch to the OL regime in finite time.

From (3.13), we see the $b(t,x)$ and $B(t)$ initially evolve as

$$\begin{aligned} b(t,x) &= \rho e^{-x} 1_{\{x \le t\}} + e^{-t} b(0, x-t) 1_{\{x > t\}} \\ B(t) &= \rho - (\rho - B(0)) e^{-t}, \quad 0 \le t \le t_1. \end{aligned} \tag{4.27}$$

The total fluid content in service $B(t)$ increases in t until time t_1 at which we first have $B(t) = B(t_1) = 1$. After time t_1, since the arrival rate ρ is greater than the maximum departure rate which is 1, the system stays in the OL regime. After time t_1, we can apply Proposition 3.2 of Chapter 3 to describe the evolution of $b(t,x)$. In particular, for $t > t_1$ and for each $x \ge 0$,

$$b(t - t_1, x) = e^{-x} 1_{\{x \le t - t_1\}} + b(t_1, x - t + t_1) e^{-(t - t_1)} 1\{x > t - t_1\}, \tag{4.28}$$

where

$$b(t_1, x) = \rho e^{-x} 1_{\{x \le t_1\}} + e^{-t_1} b(0, x - t_1) 1_{\{x > t_1\}}, \tag{4.29}$$

so that, by assumption (4.24), the second term in (4.28) is asymptotically negligible as $t \to \infty$, implying that $b(t,x) \to e^{-x} = b(x)$ as $t \to \infty$.

Since we start UL, we first have a queue buildup at time t_1. By (3.14), we have

$$q(t,x) = \rho \bar{F}(x) 1_{\{x \le w(t) \wedge (t - t_1)\}}, \quad t > t_1, \tag{4.30}$$

where the BWT w satisfies the ODE

$$w'(t) = 1 - \frac{1}{\rho \bar{F}(w(t))} \equiv H(w(t)), \quad for\ t \ge t_1, \tag{4.31}$$

with initial condition $w(t_1) = 0$. It is easy to see that $q(t,x) \to q(x) = \rho\bar{F}(x)1_{\{x \le w(t)\}}$ if $w(t) \to w$ as $t \to \infty$.

Let $w \equiv F^{-1}(1-1/\rho)$. Since the cdf F has a positive density, the function H is strictly decreasing and $H(w) = 0$. Therefore if $w(t_2) = w$ at some t_2, $w(t)$ will stay at w for all $t \ge t_2$, since $w'(t_2) = H(w) = 0$. Moreover, if $w(t) < w$, then $w'(t) = H(w(t)) > H(w) = 0$.

The function $w(t)$ starts at 0 at time t_1, and is increasing (has positive derivative) as long as $w(t) < w$. We also know that $w(t)$ will stay at w if it hits w, and $w(t)$ is continuous. Therefore, to show that $w(t) \to w$ as $t \to \infty$, it remains to show that for any $\epsilon > 0$, there exits a t_ϵ such that $w(t) > w - \epsilon$ for any $t > t_\epsilon$.

Because H is strictly decreasing in a neighborhood of w, we have $w'(t) = H(w(t)) \ge H(w-\epsilon) \equiv \delta(\epsilon) > H(w) = 0$, if $w(t) \le w - \epsilon$. Therefore, the derivative of $w(t)$ is not only positive, but also bounded by $\delta(\epsilon) > 0$. So $w(t)$ will hit $w-\epsilon$ at least linearly fast with slope $\delta(\epsilon)$, i.e., for any $t \ge (w-\epsilon)/\delta(\epsilon)$, we have $w(t) \ge w - \epsilon$. Therefore, we conclude that $w(t) \uparrow w$ as $t \uparrow \infty$. As a consequence, we get $q(t,x) \to q(x) = \rho\bar{F}(x)1_{\{0 \le x \le w\}}$ as $t \to \infty$ from (4.30).

We now establish (4.25) and (4.26). To do so, we assume the system is initially OL with $w(0) = w_0$. From the above analysis, if $\rho > 1$, then the system stays OL for all $t \ge 0$, which implies that $\gamma(t) = \mu\, s = 1$ for all $t \ge 0$. Hence, after $T \equiv Q(0)/\mu s = Q(0)$, all fluid that was in queue at $t = 0$ is gone (has entered service or abandoned). If $w(T) = w$, then the system is already in equilibrium. If $w(T) > w$ (the case $w(T) < w$ is similar), then the above analysis implies that $w'(t) \le 0$ for $t \ge T$ since H in (4.31) is decreasing.

Therefore, the monotonicity of w follows. Integrating equation (4.31) yields, for $t \geq T$,

$$\begin{aligned} w(t) - w(T) &= t - T - \frac{1}{\rho}\int_T^t \frac{1}{\bar{F}(w(s))} ds \\ &\leq t - T - \frac{1}{\rho}\int_T^t \frac{1}{\bar{F}(w(t))} ds = (t-T)\left(1 - \frac{1}{\rho\,\bar{F}(w(t))}\right) \\ &= -(t-T)\frac{\bar{F}(w) - \bar{F}(w(t))}{\bar{F}(w(t))} \\ &\leq -(t-T)(w(t)-w) f^{\downarrow}_{w(t)} \leq -(t-T)(w(t)-w) f^{\downarrow}_{w(0)+T}, \end{aligned}$$

where the first inequality holds because $w(s) \geq w(t)$ by the monotonicity of w, the third equality holds because $\bar{F}(w) = 1/\rho$, the second inequality holds because $w(t) \geq w$ and $\bar{F}(w(s)) \leq 1$, the last inequality holds because $w(t) \leq w(0) + T$ for $0 \leq t \leq T$ and w is monotone non-increasing for $t > T$. This immediately yields

$$\begin{aligned} \Delta w(t) = w(t) - w &\leq -f^{\downarrow}_{w(0)+T}(t-T)\Delta w(t) + (w(T) - w) \\ &= -f^{\downarrow}_{w(0)+T}(t-T)\Delta w(t) + \Delta w(T), \end{aligned}$$

and

$$\Delta w(t) \leq \frac{1}{1 + f^{\downarrow}_{w(0)+T}(t-T)}\,\Delta w(T).$$

Relation (4.26) follows from (4.25) by splitting interval $[T, t]$ into N disjoint subintervals with equal lengths. Mathematical induction implies that

$$\Delta w(t) \leq \left(\frac{1}{1 + f^{\downarrow}_{w(0)+T}\left(\frac{t-T}{N}\right)}\right)^N \Delta w(T).$$

Letting $N \to \infty$ yields the desired (4.26). ■ □

We next give explicit expressions of all performance functions in the $G/M/s+M$ fluid model, with exponential abandonment, when the system is initially empty.

Corollary 4.2 $(\textit{the } G/M/s+M \textit{ fluid queue})$ *Consider the* $G/M/s+M$ *fluid queue with model parameters* λ, μ, s, θ*, where* $\theta > 0$ *is the abandonment rate, starting empty.*

(a) *if* $\rho \equiv \lambda/s\mu > 1$*, then*

$$
\begin{aligned}
w(t) &= \frac{1}{\theta}\log\left(\frac{\rho}{1+(\rho-1)e^{-\theta(t-t_1)}}\right)1_{\{t\geq t_1\}} \uparrow \frac{1}{\theta}\log\rho, && (4.32)\\
q(t,x) &= \lambda\, e^{-\theta\, x}\, 1_{\{0\leq x\leq w(t),\, t\geq t_1\}} \uparrow \lambda\, e^{-\theta\, x}\, 1_{\{0\leq x\leq (\log\rho)/\theta\}}, && (4.33)\\
Q(t) &= \frac{\lambda}{\theta}\left(1-\frac{1}{\rho}\right)\left(1-e^{-\theta(t-t_1)}\right)1_{\{t\geq t_1\}} \uparrow \frac{\lambda}{\theta}\left(1-\frac{1}{\rho}\right), && (4.34)\\
\alpha(t) &= \theta\, Q(t) \uparrow \lambda\left(1-\frac{1}{\rho}\right), && (4.35)\\
b(t,x) &= \lambda\, e^{-\mu\, x}\, 1_{\{0\leq x\leq t,\, 0\leq t<t_1\}} + \mu\, s\, e^{-\mu\, x}\, 1_{\{0\leq x\leq t,\, t\geq t_1\}} \to \mu\, s\, e^{-\mu\, x}, && (4.36)\\
B(t) &= \rho\, s(1-e^{-\mu t})\cdot 1_{\{0\leq t<t_1\}} + s\cdot 1_{\{t\geq t_1\}} \uparrow s, && (4.37)\\
\sigma(t) &= \mu\, B(t) \uparrow \mu\, s, \quad \textit{as} \quad t\to\infty, \quad \textit{for} \quad x\geq 0, && (4.38)
\end{aligned}
$$

where $t_1 \equiv -1/\mu\, \log(1-1/\rho)$*.*

(b) *if* $\rho \leq 1$*, then*

$$
\begin{aligned}
q(t,x) &= Q(t) = \alpha(t) = w(t) = 0,\\
b(t,x) &= \mu\, s\, e^{-\mu\, x}\, 1_{\{0\leq x\leq t\}} \uparrow \mu\, s\, e^{-\mu\, x},\\
B(t) &= \rho\, s(1-e^{-\mu\, t}) \uparrow \rho\, s,\\
\sigma(t) &= \lambda(1-e^{-\mu\, t}) \uparrow \lambda.
\end{aligned}
$$

Proof. We only prove case (a) since (b) is similar. First, since the system is initially empty, flow conservation of the service facility implies

$$\lambda = B'(t) + \mu\, B(t), \quad B(0) = 0,$$

which has unique solution $B(t) = \rho\, s(1 - e^{-\mu\, t})$ when t is small. The system switches to the OL regime at t_1 where $\rho\, s(1 - e^{-\mu\, t_1}) = s$, and stays in that regime for all $t > t_1$. This yields (4.37), from which (4.38) and (4.36) follow. For $t \geq t_1$, we have the ODE for BWT

$$w'(t) = \frac{s\, \mu}{\lambda\, e^{\theta\, w(t)}}, \quad w(t_1) = 0,$$

which has unique solution (4.32), from which (4.33), (4.34) and (4.35) follow. ■ □

We give a numerical example illustrating Corollary 4.2 in Appendix C.2.

Remark 4.3 $(\textit{explicit results for queues in series})$ *We can apply Corollary 4.2 to obtain explicit expressions for the performance functions with two or more queues in series, with exponential abandonment, because the arrival rate of each successive queue is the departure rate from the previous queue, and the departure rate from each queue is available explicitly.*

4.5 Periodic Steady State (PSS) for Periodic Models

In this section we consider the special case of periodic fluid models. We provide conditions under which (i) there exists a unique periodic steady state (PSS) for a periodic fluid model and (ii) the time-varying performance converges to that PSS for all (finite) initial conditions.

4.5.1 Theory

Recall that a function of a nonnegative real variable, g, is *periodic* with *period* τ if $g(t+\tau) = g(t)$ for all $t \geq 0$, where τ is the least such value, required to be strictly positive. If the relation holds for arbitrary small τ, then the function is constant; we exclude that case. We say that a $G_t/M_t/s_t+GI_t$ fluid queue is a *periodic model* if the function mapping t into the vector $(\lambda(t), \mu(t), s(t), \{F_t(x) : x \geq 0\})$ in $\mathbb{R}^3 \times \mathbb{D}$ is periodic. If the four component functions are periodic, where there is a finite least common multiple of the periods, then the overall function is periodic with the overall period being that least common multiple of the component periods. Since the time-varying abandonment time cdf's $\{F_t(x) : x \geq 0\})$ are defined on the entire real line, we require that they be periodic on their entire domain.

We have not yet said anything about the initial conditions $\{b(0,x) : x \geq 0\}$ and $\{q(0,x) : x \geq 0\}$. If these initial conditions can be chosen so that the system performance of the periodic model with period τ, $\{\mathcal{P}(t) : t \geq 0\}$, where the system state vector

$$\mathcal{P}(t) \equiv (\{b(t,x) : x \geq 0\}, \{q(t,x) : x \geq 0\}, B(t), Q(t), w(t), v(t), \sigma(t), \alpha(t)) \,. \quad (4.39)$$

is a periodic function of t with period τ, then those initial conditions produce a *periodic steady state* (PSS) for the periodic model with period τ. The performance function $\mathcal{P}$ constitutes the PSS. See Figure C.3 for an example. In order to discuss continuity and convergence in the domain of $\mathcal{P}$, we use norm

$$\begin{aligned} \|\mathcal{P}(t)\| &\equiv \sup_{t \geq 0} \{|\mathcal{P}(t)|\}, \quad \text{where} \\ |\mathcal{P}(t)| &\equiv |B(t)| + |Q(t)| + |\alpha(t)| + |\sigma(t)| + |w(t)| + |v(t)| \\ &+ \left| \int_0^\infty b(t,x)dx \right| + \left| \int_0^\infty q(t,x)dx \right| . \end{aligned} \quad (4.40)$$

A common case is a periodic model that does not start in a PSS. We then want to conclude that the performance converges to a PSS as time evolves for all finite initial conditions. We say that a function of a nonnegative real variable, g, is *asymptotically periodic* with period $\tau > 0$ if there exists a (finite) function g_∞ such that $g(n\tau + t) \to g_\infty(t)$ as $n \to \infty$ for all t with $0 \le t \le \tau$, for the given positive value of τ, but no smaller value; the limit g_∞ necessarily is a periodic function with period τ. This limit can be viewed as an application of the shift operator Ψ_τ on the function g: $\Psi_\tau(g)(t) \equiv g(\tau + t)$, $t \ge 0$. The function g is asymptotically periodic if and only if successive iterates of the shift operator converge, i.e., if $\Psi_\tau^{(n)}(g) \equiv \Psi_\tau(\Psi_\tau^{(n-1)}(g))$ converges as $n \to \infty$.

Theorem 4.5 $(PSS\ for\ the\ periodic\ fluid\ model)$ *Consider a periodic fluid queue with period $\tau > 0$. If the conditions of Lemma* 4.1 *hold, then*

(a) *There exists a unique PSS $\mathcal{P}^*$ with period τ, but not with smaller period.*

(b) *For any finite initial conditions, the performance $\mathcal{P}$ is asymptotically periodic with period τ, i.e.,*

$$\Psi^{(n)}(\mathcal{P})(t) \equiv \mathcal{P}(n\tau + t) \to \mathcal{P}^*(t) \quad as \quad n \to \infty, \quad 0 \le t \le \tau. \tag{4.41}$$

Proof. First suppose that the system starts empty. By Theorem 3.5, the shift operator Ψ_τ is a monotone operator on $\mathcal{P}(n\tau)$ for any n, because we can think of the performance $b(\tau, \cdot)$ and $q(\tau, \cdot)$ as alternative initial conditions for the model at time 0, since the model is periodic with period τ. Therefore, the sequence of system performance vectors $\mathcal{P}(0), \mathcal{P}(\tau), \mathcal{P}(2\tau), \ldots$ (at discrete time $0, \tau, 2\tau, \ldots$) is monotonically non-decreasing. By Lemma 4.1, the performance is bounded, so that there is a finite limit for $\mathcal{P}(n\tau)$ as $n \to \infty$. By Theorem 3.6, the operator is continuous as well, which implies that

$\mathcal{P}(t + n\tau) = \Psi_t(\mathcal{P}(n\tau))$ is convergent for all $0 \leq t \leq \tau$ as $n \to \infty$. Hence the limit is a PSS. By Theorem 4.1, we have ALOM, which implies that we get the same limit for all initial conditions. ■ □

Theorem 4.1 shows that the rate of convergence to the PSS in Theorem 4.5 is exponentially fast as well, under regularity conditions.

4.5.2 An Example

Example 4.2 (*an $G_t/M/s_t + M$ example with periodic arrival rate and staffing*) We now consider a variant of Example 4.1 that has sinusoidal staffing as well as a sinusoidal arrival rate. As before, we have the fluid queue with arrival rate function in (4.1) with $a = c = 1$, $b = 0.6$, constant service rate $\mu = 1$ and constant abandonment rate $\theta = 0.5$. However, now we also use the sinusoidal staffing function

$$s(t) = \bar{s} + u \sin(\gamma t). \tag{4.42}$$

Let $\bar{s} = a = c = \mu = 1$ $u = 0.3$ and $\gamma = 2$. Note the period of λ is $2\pi/c = 2\pi$, while the period of s is $2\pi/\gamma = \pi$. Hence the overall model has period $2/pi$. Figure 4.2 shows the results after applying the algorithm in Chapter 2 to compute the performance measures $w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$. Instead of plotting just one OL and UL interval in $[0, T]$ with $T = 10$ as we did in Example 4.1, here we plot four OL and UL intervals in $[0, T']$ with $T' = 23$.

Figure 4.2 shows that performance measures ($w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$) converge very quickly to periodic limit functions, with period $\tau = \pi$. In Appendix C.6 we

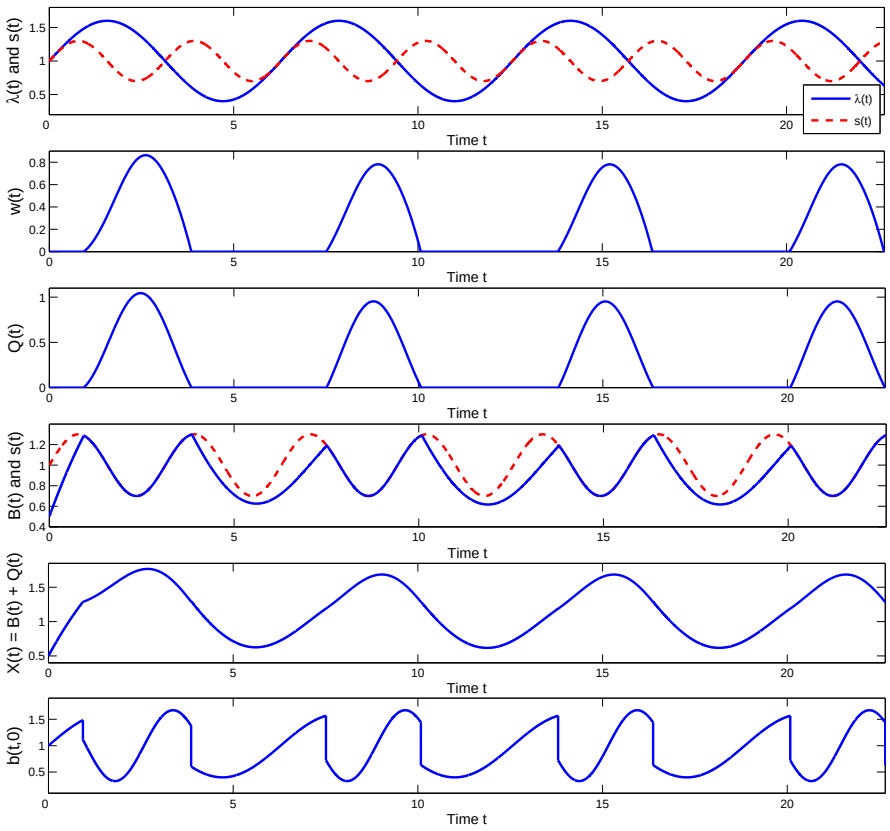

Figure 4.2: Performance of the $G_t/M/s_t + M$ model with sinusoidal arrival and staffing, $\gamma = 2$.

compare the fluid approximation in this example to simulation results for a large-scale queueing system. As in Chapter 2, we see that the fluid model provides a useful approximation for the queueing systems. It is very accurate for very large queueing systems (with thousands of servers) and provides a good approximation for mean values for smaller queueing systems (with tens of servers). In the Appendix we also consider the performance when γ is changed from 2 to 0.5. Figure C.2 there shows that the period of the PSS becomes $\tau = 4\pi$. ■

4.5.3 Direct Computation of PSS Performance

Given the rapid convergence, it usually is not difficult to compute the PSS by simply applying the algorithm with any convenient initial condition. However, the PSS can also be determined in another way. We can start by observing that there are only three cases for PSS: (i) the system is OL for all $0 \leq t \leq \tau$; (ii) the system is UL for all $0 \leq \tau$; or (iii) there is at least one switch between UL and OL regimes in $[0, \tau]$. We can simply check which of these cases prevails. For each of these scenarios, we can seek a fixed point in the performance at times τ and 0. That produces equations we can solve. One of these three cases will yield the PSS.

Consider case (i), in which the system is OL. It suffices to characterize its performance in one cycle $[0, \tau]$. We can write

$$B(t) = s(t) \quad \text{and} \quad Q(0) = \int_0^{w(0)} \lambda(t-x)\bar{F}_{t-x}(x)dx \quad \text{for} \quad w(0) > 0,$$

because in the PSS the system remains OL. Hence, we must have $q(t,0) = \lambda(t)$ and $q(t,x) = \lambda(t-x)\bar{F}_{t-x}(x)$. Note that $w_0 \equiv w(0)$ is the only unknown here. To solve for the PSS, we do a search of the initial w_0 such that during the cycle $[0, \tau]$, the system is always OL, i.e., $w(t) > 0$, and $w(\tau) = w_0$. The uniqueness of the PSS guarantees that there is at most one of such w_0. If the system switches to UL regime at some time, then we know this is not the right scenario for the PSS.

Next consider case (ii), in which the system is UL in the interval $[0, \tau]$. Since the system is UL, the fluid content in service $B(t)$ satisfies the ODE $\lambda(t) = B'(t) + \mu(t)\, B(t)$ with initial condition $B(0) = B_0 > 0$ which has a unique solution

$$B(t) = e^{-\int_0^t \mu(s)ds} \left(\int_0^t e^{\int_0^s \mu(u)du} \lambda(s)ds + B_0 \right), \quad \text{for } 0 \leq t \leq \tau. \tag{4.43}$$

Since we seek $B(\tau) = B_0$, it suffices to solve equation

$$B_0 = e^{-\int_0^\tau \mu(s)ds} \left(\int_0^\tau e^{\int_0^s \mu(u)du} \lambda(s)ds + B_0 \right)$$

for B_0. Again, the uniqueness of PSS guarantees that there is at most one such $B_0 > 0$. If this equation does not have a solution, then we know this is not the right scenario for the PSS.

Finally, consider case (iii), in which the system switches at least twice between UL and OL regimes, as shown in Figure 4.2. Since system regime changes in the PSS, we consider the interval $[0, \tau]$ and assume that in PSS the system is critically loaded at $t = 0$ and becomes OL at $t+$, i.e., we can always let the beginning of the cycle of PSS be a regime switching point from UL to OL. We assume that the phase difference between the PSS cycle and the model functions is $0 \leq t_0 \leq \tau$. Hence, we start with the BWT ODE

$$w'(t) = 1 - \frac{\mu(t + t_0)\, s(t + t_0) + s'(t_0)}{\lambda(t + t_0 - w(t))\bar{F}_{t+t_0-w(t)}(w(t))}, \quad \text{with } w(0) = 0,$$

and let $t_1 \equiv \inf\{t > 0 : w(t) = 0,\ \lambda(t + t_1) \leq \mu(t)\, s(t) + s'(t)\}$. If $t_1 > \tau$ (e.g., $t_1 = \infty$), then we know this is not the right scenario. If $t_1 < \tau$, the system switches to the UL regime at t_1. Then, just as in (4.43), we have

$$B(t) = e^{-\int_{t_1}^t \mu(s+t_0)ds} \left(\int_{t_1}^t e^{\int_0^s \mu(u+t_0)du} \lambda(s + t_0)ds + B(t_1) \right),$$

with $B(t_1) = s(t_1 + t_0)$. We let $t_2 \equiv \inf\{t > t_1 : B(t) > s(t + t_0)\}$. If $t_2 < \tau$, then the system switches back to OL regime after t_2. We repeat the above procedure until we get to time τ. If the initial phase difference variable t_0 is the right one, the system should again be critically loaded at τ. We do a search for t_0 in $[0, \tau]$.

Since analytic expressions are available for the $G/M/s+M$ fluid model as shown in Corollary 4.2, we show how explicit PSS performance functions can be calculated in the next example.

Example 4.3 (*explicit PSS performance in special cases*) Consider the $G_t/M/s+M$ fluid model in Example 4.1 that has sinusoidal arrival rate as in (4.1), exponential service distribution with rate μ, constant staffing s and exponential patience distribution with rate θ. We suppose that we are in case (iii) above, in which there is a switching point from UL to OL regimes, which we can take to be at the beginning of a cycle. We assume the arrival rate is $\tilde{\lambda}(t) \equiv \lambda(t+t_0)$ for some $0 \le t_0 \le \tau$. At some t_1 for $0 < t_1 < \tau \equiv 2\pi/c$, the system will switch to the UL regime. Hence, in order to characterize the complete performance in a cycle $[0,\tau]$, it remains to determine the values of t_0 and t_1 for $0 \le t_0 \le \tau$, $0 \le t_1 \le \tau$.

Since the system is critically loaded at $t=0$, OL in $[0,t_1)$ and UL in $[t_1,\tau]$, we need two equations for two unknowns t_0 and t_1. First, the BWT ODE implies that $w(0)=0$ and

$$w'(t) = 1 - \frac{\mu\, s}{\tilde{\lambda}(t-w(t))\, e^{-\theta\, w(t)}} = 1 - \frac{\mu\, s\, e^{\theta\, t}}{\tilde{\lambda}(t-w(t))\, e^{\theta(t-w(t))}}, \quad 0 \le t \le t_1,$$

which yields that

$$\mu\, s\, e^{\theta\, t} = \tilde{\lambda}(t-w(t))\, e^{\theta(t-w(t))}(1-w'(t)) = \tilde{\lambda}(t-w(t))\, e^{\theta(t-w(t))}\frac{d(t-w(t))}{dt}.$$

Integrating both sides and let $v(t) \equiv t - w(t)$, we have

$$\int_0^t \mu\, s\, e^{\theta\, u} du = \int_0^{v(t)} \tilde{\lambda}(y) e^{\theta\, y} dy.$$

Plugging the sinusoidal arrival rate $\tilde{\lambda}(t) = \lambda(t + t_0)$ into the above equation yields that

$$\begin{aligned}\frac{\mu\, s}{\theta}(e^{\theta\, t} - 1) &= \frac{a}{\theta}(e^{\theta\, v(t)} - 1) + \frac{b}{1 + c^2/\theta^2}\left[\frac{1}{\theta}e^{\theta\, v(t)}\sin(c\, v(t) + c\, t_0)\right. \\ &\quad \left. -\frac{c}{\theta^2}(e^{\theta\, v(t)}\cos(c\, v(t) + c\, t_0) - \cos(c\, t_0))\right].\end{aligned}$$

Since $v(t_1) = t_1 - w(t_1) = t_1$, letting $t = t_1$ in the above equation yields

$$\begin{aligned}\frac{\mu\, s}{\theta}(e^{\theta\, t_1} - 1) &= \frac{a}{\theta}(e^{\theta\, t_1} - 1) + \frac{b}{1 + c^2/\theta^2}\left[\frac{1}{\theta}e^{\theta\, t_1}\sin(c\, t_1 + c\, t_0)\right. \\ &\quad \left. -\frac{c}{\theta^2}(e^{\theta\, t_1}\cos(c\, t_1 + c\, t_0) - \cos(c\, t_0))\right]. \qquad (4.44)\end{aligned}$$

Second, since the system is UL in $[t_1, \tau]$, we have

$$\lambda(t + t_0) = \tilde{\lambda}(t) = B'(t) + \mu\, B(t), \quad t_1 \le t \le \tau,$$

which implies that

$$B(t)e^{\mu\, t} - B(t_1)e^{\mu\, t_1} = \int_{t_1}^{t} \lambda(u + t_0)e^{\mu\, u}du.$$

Since the system becomes critically loaded again at t_1 and at the end of the cycle, i.e., $B(t_1) = B(\tau) = B(2\pi/c) = s$, plugging the sinusoidal arrival rate into the above equation

yields

$$
\begin{aligned}
s(e^{-\mu\,2\pi/c} - e^{-\mu\,t_1}) &= \frac{a}{\mu}(e^{-\mu\,2\pi/c} - e^{-\mu\,t_1}) \\
&+ \frac{b}{1+c^2/\mu^2}\left[\frac{1}{\mu}(e^{\mu\,2\pi/c}\sin(2\pi + c\,t_0) - e^{\mu\,t_1}\sin(c\,t_0 + c\,t_1))\right. \\
&\left. - \frac{c}{\mu^2}(e^{\mu\,2\pi/c}\cos(2\pi + c\,t_0) - e^{\mu\,t_1}\cos(c\,t_0 + c\,t_1))\right]. \qquad (4.45)
\end{aligned}
$$

Unfortunately, Equation (4.44) and (4.45) evidently do not have explicit solutions in general, but they can be solved quite easily numerically by performing a search over the two unknowns. However, we can continue analytically in a special case with convenient parameters: (a) $a = s\mu$ and (b) $\mu = \theta$.

Note that (a) says that the average traffic intensity is $\bar{\rho} = \bar{\lambda}/s\mu = a/s\mu = 1$ and (b) says that this model is equivalent to an infinite-server model, because $\theta = \mu$.

With these extra assumptions, equations (4.44) and (4.45) simplify to

$$
\begin{aligned}
\frac{c}{\theta}\cos(c\,t_0) &= -e^{\theta\,t_1}[\sin(c\,t_1 + c\,t_0) - \frac{c}{\theta}\cos(c\,t_1 + c\,t_0)], \\
e^{\mu\,2\pi/c}[\sin(c\,t_0) - \frac{c}{\mu}\cos(c\,t_0)] &= e^{\mu\,t_1}[\sin(c\,t_1 + c\,t_0) - \frac{c}{\mu}\cos(c\,t_1 + c\,t_0)].
\end{aligned}
$$

Adding these two equations yields

$$
0 \le t_0 = \frac{1}{c}\arctan(1 - e^{-\mu\,2\pi/c}) \le \pi/c. \qquad (4.46)
$$

Note that we need $\lambda(0) = a + b\sin(c\,t_0) \ge \mu\,s$ so that the system switches from UL to UL regime at $t = 0$. Similarly, we require $\lambda(t_0 + t_1) \le \mu\,s$, which implies that $\pi/c \le t_0 + t_1 \le$

$2\pi/c$. Hence, plugging (4.46) into the first equation above implies that t_1 is the solution to

$$\sin(ct_1 + \psi) = -\frac{(c/\theta)e^{e^{\mu\, 2\pi/c}}}{\sqrt{x^2+y^2}}\, e^{-\theta\, t_1}, \tag{4.47}$$

where $\psi \equiv \arctan(x/y)$, $x \equiv e^{\mu\, 2\pi/c} - 1 - (c/\theta)e^{\mu\, 2\pi/c}$, $y \equiv e^{\mu\, 2\pi/c} + (c/\theta)(e^{\mu\, 2\pi/c} - 1)$.

Given t_0 and t_1, we can compute analytically all performance functions of this $G_t/M/s+M$ example in a cycle $[0, \tau] = [0, 2\pi/c]$. For $0 \le t < t_1$, the system is OL with

$$\begin{aligned}
q(t,0) &= \tilde{\lambda}(t) = a + b\, \sin[c(t+t_0)], \\
q(t,x) &= \tilde{\lambda}(t-x)\, e^{-\theta\, x} = e^{-\theta\, x}(a + b\, \sin[c(t+t_0-x)]), \\
w(t) &= t - \Lambda^{-1}\left(\frac{\mu\, s}{\theta}(e^{\theta\, t} - 1)\right), \\
Q(t) &= \int_0^{w(t)} q(t,x)dx = e^{-\theta\, t}\Lambda(t) - \frac{\mu\, s}{\theta}(1 - e^{-\theta\, t}), \\
\alpha(t) &= \theta\, Q(t), \\
B(t) &= s, \quad \sigma(t) = \mu\, s, \\
b(t,x) &= \mu\, s\, e^{-\mu\, x}\, 1_{\{x \in \cup_{k=0}^{\infty}((t+k\tau-t_2)^+,\, t+k\tau]\}} \\
&\quad + \lambda(t-x)\, e^{-\mu\, x}\, 1_{\{x \in \cup_{k=0}^{\infty}(t+k\tau,\, t+(k+1)\tau-t_2]\}},
\end{aligned}$$

where $\Lambda(x) \equiv \int_0^x \lambda(y)\, e^{\theta\, y} dy$. For $t_1 \leq t \leq \tau$, the system is UL with

$$\begin{aligned}
aq(t,x) &= Q(t) = w(t) = \alpha(t) = 0, \\
b(t,0) &= \tilde{\lambda}(t) = a + b\, \sin[c(t+t_0)], \\
b(t,x) &= \tilde{\lambda}(t-x)\, e^{-\mu\, x}\, 1_{\{x \in \cup_{k=0}^{\infty}((t+(k-1)\tau)^+,\, t+k\tau-t_2]\}} \\
&\quad + \mu\, s\, e^{-\mu\, x} 1_{\{x \in \cup_{k=0}^{\infty}(t-t_2+k\tau,\, t+k\tau]\}}, \\
B(t) &= s\, e^{-\mu(t-t_1)} + e^{-\mu\, t} \int_{t_1}^{t} \tilde{\lambda}(u) e^{\mu\, u} du, \\
\sigma(t) &= \mu\, B(t),
\end{aligned}$$

4.6 Conclusions

In this chapter we supplemented Chapters 2 and 3 and [77] by studying the large-time asymptotic behavior of the $G_t/M_t/s_t + GI_t$ many-server fluid queue with time-varying model parameters. In §4.3 we established the asymptotic loss of memory (ALOM) property, concluding that the difference between performance functions evaluated at time t, with different initial conditions, dissipates exponentially fast as $t \to \infty$, under regularity conditions. In §4.4 we applied ALOM to establish convergence to steady state for the stationary model. In §4.4 we also went beyond ALOM to provide additional details; e.g., we showed that the system changes regimes (overloaded or underloaded) at most once. In §4.5 we applied ALOM, first, to establish the existence of a unique periodic steady state (PSS) and, second, to establish convergence to that PSS in the periodic model, where the period is the least common multiple of the periods of the model functions, assumed to be some finite value.

There are many directions for future research: First, it remains to establish ALOM

properties for the $G_t/GI/s_t + GI$ fluid queue with non-exponential (GI) service that was considered in Chapter 2 (under regularity conditions that exclude the counterexample in Chapter 5) and the $(G_t/M_t/s_t + GI_t)^m/M_t$ network of fluid queues with proportional routing considered in Chapter 3. Second, it remains to establish many-server heavy-traffic limits showing that appropriately scaled stochastic processes in many-server queues converge to the fluid queues, as discussed in Chapter 2 and [77]. It also remains to establish refined stochastic approximations as a consequence of many-server heavy-traffic limits. Third, it remains to establish corresponding ALOM (or weak ergodicity) and PSS properties for the corresponding stochastic queueing models and the refined stochastic approximation; see [24, 30, 33, 78] and references therein. Fourth, it remains to exploit the deterministic fluid models to approximately solve important control problems for the stochastic systems and, fifth, it remains to apply the fluid models to analyze large-scale service systems, such as hospital emergency departments. We hope to contribute to these goals in the future.

Chapter 5

The Overloaded $G/D/s + GI$ Queue

We next focus on many-server queues with deterministic service times. In particular, we investigate the many-server $G/D/s + GI$ model with a stationary arrival process, deterministic service times, and general abandonment times. In addition, we study its associated fluid model to gain insights and establish an MSHT convergence theorem to that fluid model. Our main observation is that the system reveals nearly periodic behavior due to the assumption of deterministic service times. When the model is overloaded, we also demonstrate the invalidity of the interchange of two limits: the steady state (obtained as $t \to \infty$) of the limiting fluid model (obtained as $n \to \infty$) does not coincide with the fluid limit (obtained as $n \to \infty$) of the steady state (obtained as $t \to \infty$) of the queueing processes.

5.1 Introduction

In this chapter we continue to investigate the performance of overloaded many-server queueing systems with customer abandonment, extending earlier work in [75, 77] and Chapters 2-4; we focus on the special case of deterministic service times. By overloaded, we mean that $\rho > 1$, where ρ is the traffic intensity.

It was shown in [77] that the steady-state performance of the overloaded $G/GI/s+GI$ queueing model when s is large is well approximated by the steady-state performance of an associated deterministic $G/GI/s+GI$ fluid model (when the two models are connected by many-server heavy-traffic (MSHT) scaling; see §2 of [77] and §5.3 here). Supporting MSHT limits were established in [**?**, **?**]. In Chapter 2, as a special case of a more general fluid model with time-varying parameters, we fully specified that $G/GI/s+GI$ fluid model and described its transient performance. In Chapter 4 we showed for the special case of the $G/M/s + GI$ fluid model that the time-dependent performance functions converge to the steady state values as time evolves. It remains to establish convergence to steady state for the $G/GI/s+GI$ fluid model with other service distributions, even though the steady-state performance is available from Theorem 3.1 of [77] and Theorem 3.11 of Chapter 3. In this chapter we show that convergence to steady state in the fluid model does not occur for all service distributions; some conditions are needed.

We began investigating convergence to steady state for overloaded fluid models with non-exponential service distributions by considering the special case of deterministic service times, even though the deterministic distribution does not satisfy the smoothness conditions imposed on the model elements in [77] and Chapters 2-4. We began considering the case of deterministic service times primarily because it is relatively easy to analyze. However, deterministic service times are also of applied interest, because computer-generated service times, such as automated messages, may well be deterministic, and computer-

generated service is becoming more prevalent. Many message systems can handle multiple requests in parallel, justifying the many-server model.

We started by considering a specific example: a $G/D/s+M$ fluid model having arrival rate λ, deterministic service times equal to $1/\mu$, service capacity s and an exponential abandonment cdf F with mean $1/\theta$. (The model is specified in detail later in the chapter, starting in §5.4.) We let the other parameters be $\lambda = 2$ and $\mu = s = 1$, making the system overloaded with traffic intensity $\rho \equiv \lambda/s\mu = 2 > 1$, so that the model is overloaded.

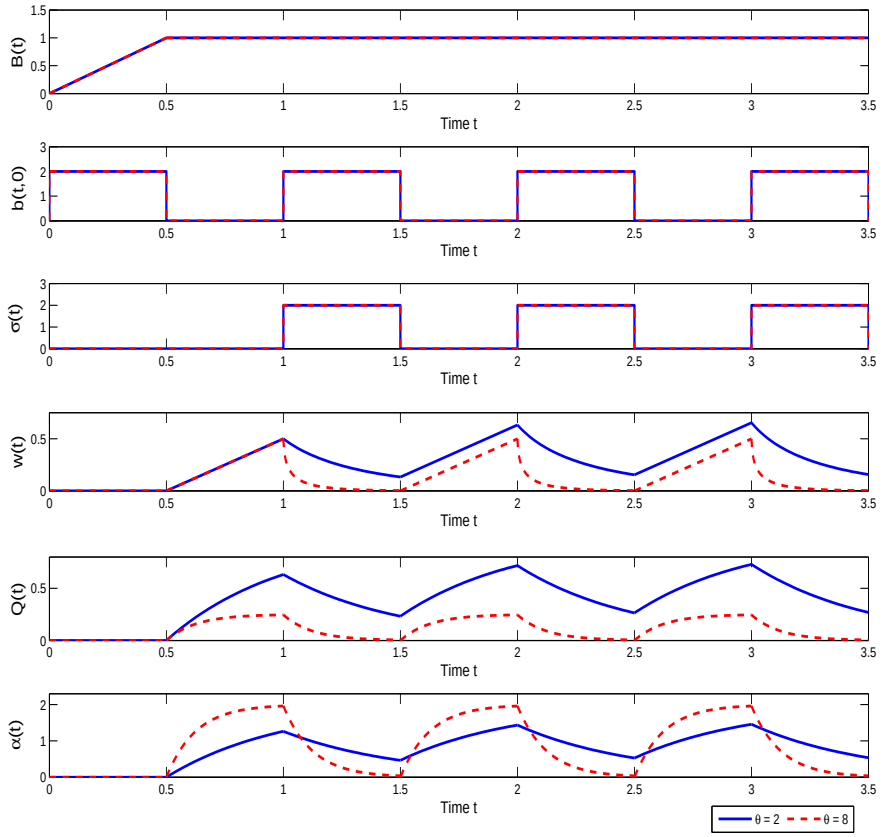

Figure 5.1: The $G/D/s + M$ fluid model with $s = \mu = 1$, $\lambda = 2$.

Figure 5.1 shows six performance functions evolving over time for the $G/D/s + M$ fluid model starting empty. The performance functions shown are the total fluid content in service, $B(t)$, the rate that fluid enters service, $b(t, 0)$, the departure rate, $\sigma(t)$, the elapsed

waiting time for the quantum of fluid at the head of the queue, $w(t)$, the total fluid content waiting in queue, $Q(t)$, and the abandonment rate $\alpha(t)$ over the initial time interval $[0, 3.5]$. There are two plots for the final three performance functions, the solid line for abandonment rate $\theta = 2$ and the dashed line for abandonment rate $\theta = 8$.

We had initially expected to see convergence to the stationary point of this fluid model (which we later show is well defined), because the fluid model is an approximation for the $M/D/s + M$ stochastic model, but instead we see that the performance becomes periodic with period equal to the service-time distribution after time $t = 1.0$. At first, we thought that the periodic performance was due to the special choice of the parameters, but that is not the case. Theorem 5.11 shows that the overloaded $G/D/s + GI$ fluid model starting empty exhibits periodic performance after a finite time for all arrival rates λ, service times $1/\mu$ and staffing levels s with $\rho \equiv \lambda/s\mu > 1$, for all abandonment-time cdf's F.

In fact, the functions displayed in Figure 5.1 are easy to understand. Since the system starts empty and the service capacity is $s = 1$, the arriving fluid flows directly into service at rate $b(t, 0) = \lambda = 2$ over the interval $[0, 0.5]$. Hence, the total fluid content in service, $B(t)$ grows linearly at rate 2 over the interval $[0, 0.5]$, reaching the capacity $s = 1$ at time $t = 0.5$, where it stays thereafter. The fluid that entered service in $[0, 0.5]$ completes service exactly $1/\mu = 1$ time units later. Hence there is service completion at rate $\sigma(t) = 2$ over the interval $[1, 1.5]$. Since new fluid cannot enter service until there is free capacity, new fluid enters service only at time 1. Hence, we have $b(t, 0) = 0$ during the interval $[0.5, 1]$ and then $b(t, 0) = 2$ again in the interval $[1, 1.5]$, which leads to the periodic behavior. Since no arriving fluid can enter service in the interval $[0.5, 1]$, the queue content grows during the interval $[0.5, 1]$. It does not grow linearly because some portion of the fluid entering the queue is lost due to fluid abandonment. For this example, we see that all functions exhibit periodic behavior beginning at time $t = 1$. Explicit expressions for the performance functions for the $G/D/s + M$ fluid model starting empty are given in Corollary 5.8.

Having seen how pervasive is this periodic behavior in the fluid model, we were led to seriously doubt the value of the fluid model as an approximation for the stochastic queueing system. For the special case of the $M/D/s + M$ stochastic model, it is evident that the stochastic model has a unique stationary performance and that the performance converges to that stationary performance as time evolves. Indeed, in §5.2 here we prove that the stochastic process $X \equiv \{X(t) : t \geq 0\}$ representing the number of customers in the more general $GI/D/s + GI$ queueing model is a regenerative stochastic process that converges to a unique stationary distribution as time evolves, provided only that the interarrival-time cdf G is nonlattice, has a finite mean $1/\lambda$ and is unbounded above, while the abandonment-time cdf F has finite mean $1/\theta$.

However, when we conducted simulations of the stochastic $GI/D/s + GI$ model, we found that the sample paths actually agree closely with the deterministic fluid model, exhibiting periodic performance over the horizon of our simulation runs. For example, we simulated a many-server $M/D/s_n + M$ stochastic queueing system with Poisson arrival process approximated by the $G/D/s + M$ fluid model, for which the periodic performance is shown in Figure 5.1. We obtain the related stochastic model by exploiting MSHT scaling, i.e., by letting the arrival rate be $\lambda_n \equiv n\lambda = 2n$ and the number of servers be $s_n \equiv \lceil ns \rceil = n$, where $\lceil x \rceil$ is the least integer greater than or equal to x, while leaving the service times and abandonment rate unchanged as $1/\mu = 1$ and θ, respectively. We expect to have a good approximation when n is large.

Figure 5.2 compares the fluid approximation (the dashed lines) with simulation estimates (the solid lines) for the large-scale $M/D/s + M$ queueing system with $n = 1000$. We plot (i) the elapsed waiting time of the customer at the head of the line $W_n(t)$, (ii) the scaled number of customers waiting in queue $\bar{Q}_n(t) \equiv Q_n(t)/n$ and (iii) the scaled number of customers in service $\bar{B}_n(t) \equiv B_n(t)/n$. We plot single sample paths of these processes. For this large value of n, there is little variability in the simulation sample paths. Each sim-

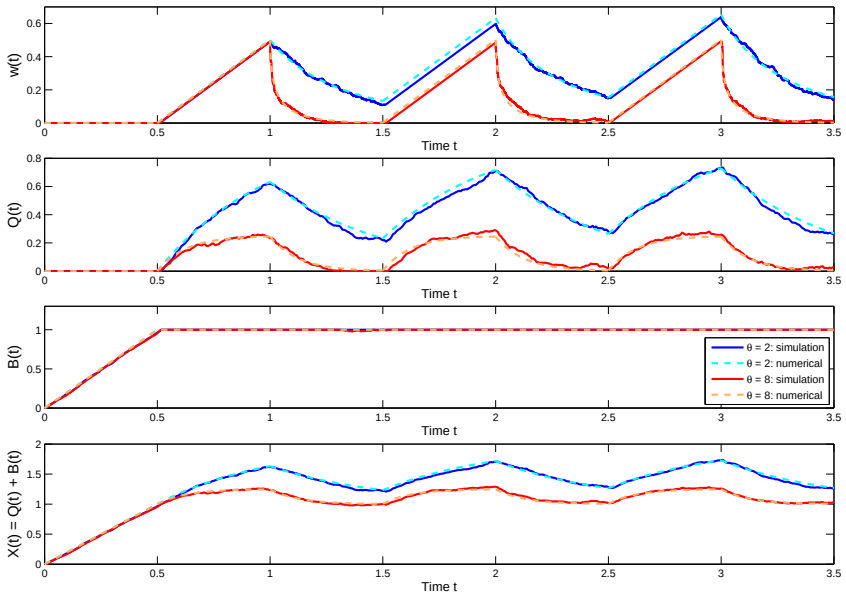

Figure 5.2: A comparison of the $G/D/s + M$ fluid model with a simulation (of single sample paths) of the corresponding $M/D/s + M$ stochastic model with $n = 1000$.

ulated sample path falls right on top of the the approximation. (The two different plots are two different cases of the abandonment rate θ.) Figure 5.2 shows that the fluid approximation is effective in describing the performance of the stochastic system. The deterministic periodic character is exhibited by the waiting times, which rise linearly at the end of each interval $[k, k + 1]$, reaching a peak at the integer endpoint.

However, Figure 5.2 only compares the performance over a relatively short initial interval of length 3.5, corresponding to 3.5 service times. At first, we thought that we only need look at a somewhat longer time interval. However, repeated simulations show that the same periodic behavior is seen in the stochastic system over time intervals of length 1000. That is illustrated by Figure 5.3, which shows simulation estimates of the elapsed waiting $W_n(t)$ for large time $T = 1000$ (instead of small $T = 3.5$ in Figure 5.2) of the same $M/D/s+M$ model with the same parameters ($\lambda = 2$, $s = \mu = 1$, $\theta = 2$) and initial conditions (initially empty), but with a smaller fluid scaling $n = 100$. The two plots in Figure 5.3 compare the behavior of a single sample path of $W_n(t)$ at the end ($[989, 999]$, the blue solid curve) and

at the beginning ($[0, 10]$, the red dashed curve). Figure 5.3 shows that the periodic behavior of $W_n(t)$ remains at time 1000 for $n = 100$. (The process $\bar{Q}_n$ behaves the same as W_n.)

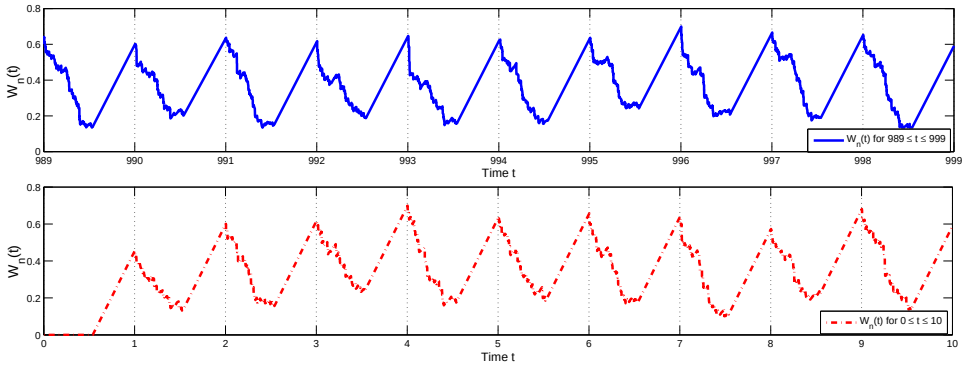

Figure 5.3: Large-time periodic behavior of an overloaded $G/D/s + M$ queueing model: simulation estimates of the head-of-line waiting time W_n with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $n = 100$, $T = 1000$.

Of course, the regenerative theory is not wrong. The stochastic system will eventually approach its stationary distribution if we consider a sufficiently long time. In fact, we do see the periodic pattern broken by 1000 service times in typical simulation sample paths if we decrease the system load ρ and the scale n sufficiently. For example, Figure D.5 in the appendix shows that occurs if we replace $\rho = 2$ by $\rho = 1.3$ (by changing λ). By time $T = 1000$, the periodic behavior of W_n is gone.

In §5.3 we will establish a many-server heavy-traffic limit showing that a sequence of scaled stochastic processes indexed by n converges to the deterministic fluid model as $n \to \infty$, under regularity conditions. Since we are considering overloaded models with $\rho > 1$, this is a many-server heavy-traffic limit for the $G/D/s+GI$ model in the efficiency driven (ED) regime [20], as in [75].

It is customary to apply HT approximations to approximate the steady-state performance of queueing systems. HT approximations for the steady-state performance of queueing processes are supported by results showing that two iterated limits coincide. For MSHT

fluid limits, we want

$$\lim_{t\to\infty}\lim_{n\to\infty} n^{-1}X_n(t) = \lim_{n\to\infty}\lim_{t\to\infty} n^{-1}X_n(t), \tag{5.1}$$

where $X_n(t)$ is a stochastic process or vector of stochastic processes characterizing performance in model n. On the left in (5.1), we have the steady-state (obtained as $t \to \infty$) of the HT limiting process (obtained as $n \to \infty$); on the right, we have the HT limit (obtained as $n \to \infty$) of the steady state (obtained as $t \to \infty$) of the queueing process. Such limit-interchange results have recently been obtained in [19, 27]. For MSHT approximations, such results were obtained for exponential service times in [20, 28].

Here we do not have that nice state of affairs. Indeed, after establishing the MSHT limit as $n \to \infty$, we show that the subsequent limit as $t \to \infty$ fails to hold because of the periodicity. Moreover, the form of that periodic behavior depends on the initial conditions. Even the average over a periodic cycle depends on the initial conditions; see Remark 5.5. We will show that the fluid performance is stationary if and only if the fluid model starts in its unique stationary point; see Theorem 5.14.

Here we directly consider only the iterated limit on the left in (5.1), but we can deduce that the two iterated limits do not tell the same story. In §5.2 we show that there exists regenerative structure implying that the $GI/D/s_n + GI$ stochastic model converges to a steady state as $t \to \infty$ for each n and each finite initial condition. Moreover, we can do so for two-parameter processes that yield a Markov process. For each n, we can then initialize with the stationary distribution of the Markov process, so that we obtain a stationary process (as a function of t) for each n. Now, if we consider the limit of the sequence of scaled stationary distributions as $n \to \infty$, if we obtain convergence, then we necessarily obtain convergence to a stationary process. If such a limit corresponds to the deterministic fluid function, then it necessarily must be the unique stationary point of the fluid model.

(We conjecture that the sequence of scaled steady-state queueing processes does indeed converge to the unique stationary point of the fluid model.)

However, a major conclusion from our analysis is that, for the many-server $G/D/s + GI$ stochastic queueing model, we should not focus on the steady-state behavior of the queueing model at all. After much analysis of this kind, we conclude that the periodic phenomenon associated with deterministic service is genuine for the stochastic model as well as the fluid model. Moreover, we conclude that, when there are many servers with deterministic service times and $\rho > 1$, the approximating fluid model is likely to better describe the time-dependent performance of the stochastic system than is the stationary distribution of the stochastic system. The present chapter might better deserve the title of [72].

In retrospect, we should perhaps have anticipated this nearly periodic behavior of the overloaded $G/D/s + GI$ queueing model. First, when the $G/GI/s + GI$ queueing model is overloaded and s is large, all the servers remain busy for long intervals of time; that is evident from the steady-state performance of the fluid model in [77]. With deterministic service times, when the servers remain busy, the times at which customers complete service and thus enter service in the intervals $[t + (k-1)/\mu, t + k/\mu]$ for integer k will be independent of k. That gives rise to the observed periodic behavior.

Once the periodic phenomenon is recognized, it can be controlled if it is considered undesirable. For example, the periodic behavior of an overloaded system starting empty leads to corresponding periodic behavior in the output flow, as illustrated by the plot of $\sigma(t)$ in Figure 5.1. Such fluctuations in the output may be deemed undesirable. For example, if that output became input at a following queue, then the fluctuations could cause congestion at the subsequent queue.

A simple way to avoid periodic output is to restrict the flow rate into service, allowing flow into service to be at most at rate $s\mu$ at all times. That can be done while still respecting

the first-come first-served service discipline. Starting empty, this control imposes extra delay on some of the initial input, but the output rate will soon become constant at $s\mu$.

There should be broader implications of this work, but one has to be careful about generalizing, because closely related models behave quite differently. In contrast to the overloaded $M/D/s+M$ and $GI/D/s+GI$ models considered here, the associated infinite-server $M/D/\infty$ and $GI/D/\infty$ models are remarkably well behaved, as shown by [22]. Indeed, the number of customers in the $M/D/\infty$ system reaches steady state in finite time, after just one service time. Similarly, the MSHT fluid and diffusion approximations in the $GI/D/\infty$ model reach steady state after one service time. Having finitely many servers that are busy all the time is an important part of the story in this chapter.

Closer to the model we consider is the $G/D/s$ model without customer abandonment in the QED MSHT regime. For this model, Reed [61] observed that the limiting $G/D/s$ fluid model can exhibit periodic behavior with a special initial condition in his Example 1 at the end of §4, but the implications of that example for the queueing model were not explored. The $G/D/s$ queueing model is considered further in [63, 64]. There the $G/D/s$ queueing model for large s is identified as an example of a *nearly deterministic queue*. That work establishes MSHT limits in which the traffic intensity approaches its critical value from below, extending earlier work in [34]. The papers [63, 64] also consider the limiting behavior as $n \to \infty$ in the $G_n/G_n/1$ model in which the interarrival-time and service-time distributions are n-fold convolutions of a given base distribution, generalizing the construction of the Erlang E_k distribution from k-fold convolutions of the exponential distribution. As n increases, the $G_n/G_n/1$ model approaches the $D/D/1$ model. Interesting limiting behavior is obtained by letting the traffic intensity increase as n increases.

Of course, in the stochastic $GI/D/s$ and $GI/D/s+GI$ queueing models, only the service times are directly deterministic; the interarrival-time and abandonment-time distributions may be far from deterministic. However, when n is large and the arrival rate is

large, the essential behavior of the arrival process and the abandonment becomes deterministic, primarily because of the law of large numbers (LLN). That can be explained by heavy-traffic limits, such as for non-Markovian infinite-server queues [6, 22, 39, 56, 62]. (If the system is underloaded, then the limits in [22] apply directly.) We elaborate throughout the chapter.

Finally, we mention that oscillating behavior and bi-stability have been found in other queueing systems [16,21,78]. Another recent example of the invalidity of limit interchange is [65].

Here is how the rest of this chapter is organized: In §5.2 we establish the regenerative structure in the $GI/D/s{+}GI$ stochastic model and show that the mean busy cycle increases rapidly in s. In §5.3 we establish a MSHT limit showing that a sequence of the queueing models indexed by the number of servers converges to the fluid model. In §5.4 we carefully specify the limiting $G/D/s+GI$ fluid model. In §5.5 we derive the performance formulas for the $G/D/s+GI$ fluid model, part of which are variants of those of the $G_t/GI/s_t+GI$ fluid model developed in Chapter 2. In §5.6 we focus on the case in which there exists a finite time T^* after which the system remains overloaded (has no idle capacity). In §5.7 we present key structural properties of the $G/D/s+GI$ fluid queue assuming the queue is overloaded for all $t \geq 0$. In §5.8 we analyze the periodic steady state of the $G/D/s+GI$ fluid model assuming the queue is overloaded after finite time. In §5.9 we discuss the asymptotic behavior of the $G/D/s+GI$ fluid queue with general initial conditions. In §5.10 we present three postponed longer proofs, namely, the proofs for Theorems 5.1, 5.2 and 5.5. Finally, in §5.11 we draw conclusions.

5.2 Regenerative Structure in the $GI/D/s + GI$ Model

It is well known that a regenerative process $X \equiv \{X(t) : t \geq 0\}$ with sample paths in the function space $\mathbb{D}$ of right-continuous functions with left limits in which a generic cycle T has a distribution that is nonlattice with finite mean has a proper limiting steady-state distribution. In particular, $X(t) \Rightarrow X(\infty)$ as $t \to \infty$, where $\Rightarrow$ denotes convergence in distribution, i.e., for any continuous and bounded real-valued function h,

$$E[h(X(t)] \to E[h(X(\infty)] = \frac{E_0[\int_0^T h(X(s))\, ds]}{E[T]} \quad \text{as} \quad t \to \infty, \tag{5.2}$$

where E_0 denotes the expectation conditional on a regeneration point at time 0 and T denotes the end of the first cycle; see Theorem VI.1.2 of [2]. The importance of the sample path regularity was observed in [51]. That regularity condition allows the process to take values in a general Polish topological space [74], but the condition is needed even with the usual real-valued processes. That sample-path regularity is easily seen to be satisfied in our queueing model.

Consider the $GI/D/s + GI$ model, having interarrival times distributed as U with cdf G, deterministic service times of length $1/\mu$ and abandonment times distributed as A with cdf F. Let the interarrival times and abandonment times be mutually independent. Let $X(t)$ represent the number of customers in the $GI/D/s + GI$ system at time t. Let a busy cycle be the interval between successive epochs at which an arrival comes to find an empty system. If the system starts with an arrival to an empty system at time 0, then the first busy cycle begins at time 0. Each busy cycle begins with a busy period and then is followed by an idle period. We prove the following in §5.10.

Theorem 5.1 *Consider the stochastic $GI/D/s + GI$ model in which an interarrival time*

U has a nonlattice cdf G with finite mean $E[U] \equiv 1/\lambda$ and support unbounded above, i.e., $G(x) < 1$ for all $x > 0$, and an abandonment A that has cdf F with finite mean $E[A] \equiv 1/\theta$ and has support unbounded above and below, i.e., $0 < F(x) < 1$ for all $x > 0$. Then the busy cycles for the $GI/D/s + GI$ system constitute an embedded renewal process for the stochastic process X for which a generic busy cycle T has a nonlattice distribution with $E[T] < \infty$, so that the the stochastic process X representing the number of customers in the system has a proper limiting steady-state distribution, as in (A.15)*, for all proper initial conditions. In addition, the mean $E[T]$ is bounded below by*

$$E[T] \geq \frac{G(1/\mu)}{\bar{G}(1/\mu)} E[U|U \leq 1/\mu] + 1/\mu. \tag{5.3}$$

Theorem 5.1 provides both good news and bad news: The good news is that there exists regenerative structure, so that a proper steady-state distribution for the stochastic process X exists under general conditions. The bad news for large-scale systems (explained below) is that the mean return time to 0 typically grows at least exponentially in s. Of course, that does not directly prove that the process converges to steady state slowly, but it lends support to that notion.

We can formalize this growth in n by considering a limit involving a sequence of models indexed by n. We scale time in the arrival process while changing n to keep the traffic intensity $\rho \equiv \lambda/n\mu$ fixed. The following corollary shows that $E[T^{(n)}]$ is at least $O(e^{cn})$ as $n \to \infty$, where c is some constant with $0 < c < \infty$ when the arrival process is Poisson or in a renewal process when the interarrival-time cdf has an exponential tail.

Corollary 5.1 *Consider a sequence of $GI/D/s_n + GI$ models indexed by n satisfying the conditions of Theorem* 5.1 *with generic interarrival times $U^{(n)} \equiv U^{(1)}/n$, while the service*

times and abandonment cdf's are independent of n*. Then*

$$\liminf_{n\to\infty} \{\lambda n \bar{G}^{(1)}(n/\mu) E[T^{(n)}]\} \geq 1, \tag{5.4}$$

so that $E[T^{(n)}] \to \infty$ *as* $n \to \infty$*. If, in addition, the arrival processes are Poisson with* $E[U^{(1)}] = 1/\lambda$*, then*

$$\liminf_{n\to\infty} \{\lambda n e^{-n\lambda/\mu} E[T^{(n)}]\} \geq 1. \tag{5.5}$$

Proof. First, as $n \to \infty$, $nE[U^{(n)}|U^{(n)} \leq 1/\mu] = E[U^{(1)}|U^{(1)} \leq n/\mu] \to 1/\lambda$, and $G^{(n)}(1/\mu) \equiv P(U^{(n)} \leq 1/\mu) = G^{(1)}(n/\mu) \to 1$. Also, the first moment condition $E[U^{(1)}] < \infty$ implies that $y\bar{G}^{(1)}(y/\mu) \to 0$ as $y \to \infty$; e.g., see the proof of Lemma 1 on p. 150 of [18]. Therefore, (A.16) in Theorem 5.1 implies (5.4), which in turn implies, first, that $E[T^{(n)}] \to \infty$ as $n \to \infty$ and, second, (5.5). □

The situation is quite intuitive. If indeed n is large and $\rho > 1$, then we will necessarily have $\lambda >> \mu$ and, since it is natural in applications to have θ be the same order as μ, it is natural to also have $\lambda >> \theta$. In that case only rarely will the queue be empty and even more rarely will the entire system be empty, so that the regeneration we are relying on to have a nice steady state is then a rare event.

As noted toward the end of §5.1, periodic behavior in the $G/D/s+GI$ stochastic model will occur over some time interval whenever *all* servers remain busy over that time interval. In §5.6 we provide conditions under which there exists a finite time T^* after which the fluid model remains overloaded (has no idle capacity). We can also conclude that there will be a strictly positive queue. Combined with the MSHT limit in the next section, we can deduce that, under regularity conditions, there will be long finite intervals over which no server

is idle in the queueing model. There is no contradiction with Theorem 5.1; here the limit interchange in (5.1) does not hold.

5.3 A Many-Server Heavy-Traffic Limit

In this section we establish a many-server heavy-traffic limit, showing that a sequence of $G/D/s_n + GI$ stochastic queueing models indexed by n converges to the $G/D/s + GI$ fluid model considered in §5.4 and §5.5 in the customary many-server heavy-traffic regime, under regularity conditions.

The sequence of models is indexed by the number of servers n. We let the arrival rate in model n be λ_n and the number of servers be s_n, where

$$\bar{\lambda}_n \equiv \frac{\lambda_n}{n} \to \lambda \quad \text{and} \quad \bar{s}_n \equiv \frac{s_n}{n} \to s \quad \text{as} \quad n \to \infty. \tag{5.6}$$

We let the deterministic service times take value $1/\mu$ and the abandonment times have cdf F, independent of n. We assume limits for the arrival process and the initial conditions. In particular, we assume that the sequence of stochastic processes satisfies a *functional weak law of large numbers* (FWLLN). For that purpose, let $\mathbb{D}$ be the usual function space of real-valued functions with limits from the left, endowed with one of the Skorohod topologies, which reduces to uniform convergence on bounded intervals when the limit is a continuous function [74]. Let $\Rightarrow$ denote convergence in distribution.

Let $B_n(t,x)$ ($\hat{Q}_n(t,x)$) be the number of customers in service (queue) at time t in model n that have been so for a duration less than or equal to x. Since model n has n servers, $0 \le B_n(t,\infty) = B_n(t,1/\mu) \le n$, $n \ge 1$. Let $Q_n(t) \equiv \hat{Q}_n(t,\infty)$ be the total number of customers in queue. Let $A_n(t)$, $S_n(t)$ and $E_n(t)$ be the numbers of customers to abandon, depart after completing service, and enter service, respectively, in $[0,t]$ in model n. In full

generality, we will establish a limit for the time-scaled process

$$(\bar{B}_n(t,x), \bar{S}_n(t), \bar{E}_n(t)) \equiv n^{-1}(B_n(t,x), S_n(t), E_n(t)), \tag{5.7}$$

which characterizes the performance of the service facility. Under the additional assumption of exponential abandonment, we will also establish a limit for the time scaled process

$$(\bar{Q}_n(t), \bar{A}_n(t)) \equiv n^{-1}(Q_n(t), A_n(t)). \tag{5.8}$$

Let $N_n(t)$ be the number of arrivals in the interval $[0,t]$ in model n.

Assumption 5.1 (*FWLLN for the arrival process*) *As* $n \to \infty$,

$$n^{-1}N_n \Rightarrow \Lambda \quad \textit{in} \quad \mathbb{D} \quad \textit{as} \quad n \to \infty, \quad \textit{where} \quad \Lambda(t) \equiv \lambda t, \quad t \geq 0, \tag{5.9}$$

for a positive constant λ.

The FWLLN in Assumption 5.1 is implied by either a functional central limit theorem (FCLT) or a functional strong law of large numbers (FSLLN). Most applications are covered by simple time scaling of a fixed stationary counting process, i.e., when $N_n(t) \equiv N(nt)$, $t \geq 0$, $n \geq 1$. An FSLLN holds for the time-scaled renewal counting process (GI) considered in §5.2, provided only that the interrenewal time has finite mean $1/\lambda$.

We now make assumptions about the initial conditions. We restrict attention to starting with the queue empty, but we allow customers to start in service, imposing some additional restrictions in the theorem.

Assumption 5.2 (*an initially empty queue*) *For each* $n \geq 1$, $Q_n(0) = 0$.

We also assume a FWLLN for the initial fluid content in service.

Assumption 5.3 (*FWLLN for the initial conditions*) *As* $n \to \infty$,

$$\bar{B}_n(0, \cdot) \Rightarrow B(0, \cdot) \quad \textit{in} \quad \mathbb{D}, \tag{5.10}$$

where

$$B(0, x) \equiv \int_0^x b(0, u)\, du, \quad x \geq 0, \tag{5.11}$$

for a deterministic function $b(0, \cdot)$ *on* $[0, \infty)$ *in* $\mathbb{C}_p$ *with* $b(0, x) \geq 0$ *for all* x *and* $B(0, 1/\mu) = B(0, \infty) \leq 1$.

We are now ready to state the many-server heavy-traffic limit. For that purpose, let $\mathbb{D}_{\mathbb{D}}$ be the space of $\mathbb{D}$-valued functions in $\mathbb{D}$, as in [**?**]. The limit below will be continuous, so the topology on $\mathbb{D}_{\mathbb{D}}$ is equivalent to uniform convergence over the compact sets $[0, t] \times [0, 1/\mu]$ for $t > 0$. Let a superscript k on a topological space, as with D^k, indicate the associated k-fold product space, endowed with the product topology.

Let T_n be the first time that all servers are busy in the stochastic queueing model, i.e.,

$$T_n \equiv \inf \{t \geq 0 : B_n(t, 1/\mu) = n\}, \quad n \geq 1. \tag{5.12}$$

Let T_n^* be the first time after which all servers remain busy forever, i.e.

$$T_n^* \equiv \inf \{t \geq 0 : B_n(u, 1/\mu) = n \quad \text{for all} \quad u \geq t\}, \tag{5.13}$$

with $T_n^* \equiv \infty$ if there exists no such time. Similarly, let t^* be the time that the limiting fluid model first has no idle service capacity, defined in (5.33), and let T^* be the time after which the limiting fluid model never has any idle capacity, defined in (5.31). The conditions in (5.14) and (5.16) below will imply that the limiting fluid model never has any idle capacity after time t^*, i.e., $T^* = t^* < \infty$; see §5.6.

Theorem 5.2 (*many-server heavy-traffic FWLLN*) *Suppose that Assumptions* 5.1–5.3 *hold with* $\lambda > \mu$,

$$b(0,x) \le \lambda, \quad 1/\mu - t^* \le x \le 1/\mu, \tag{5.14}$$

and, if $t^* > 0$,

$$b(0, 1/\mu - t^*) < \lambda \quad \textit{and} \quad b(0, 1/\mu - t) \quad \textit{continuous at} \quad t = t^*. \tag{5.15}$$

Then

$$(\bar{B}_n, \bar{E}_n, \bar{S}_n) \Rightarrow (B, E, S) \in \mathbb{D}_{\mathbb{D}} \times \mathbb{D}^2, \tag{5.16}$$

where

$$B(t,y) \equiv \int_0^y b(t,x)\,dx, \quad 0 \le y \le 1/\mu, \tag{5.17}$$

with $b(t,x)$ *given in* (5.28) *for* $0 \le t \le t^*$, b *periodic as a function of its first argument for* $t > t^*$ *with period* $1/\mu$ *and, for* $t \ge t^*$, $b(t - t^*, x)$ *given in* (5.29). *In addition,*

$$S(t) \equiv \int_0^t \sigma(y)\,dy \quad \textit{where} \quad \sigma(k/\mu + t) \equiv b(k/\mu, 1/\mu - t), \quad 0 \le t \le 1/\mu, \tag{5.18}$$

for integer k with $k \geq 0$,

$$E(t) \equiv \int_0^t b(y,0)dy \quad \textit{where} \quad b(t,0) = \lambda\, 1_{\{0 \leq t \leq t^*\}} + \sigma(t)\, 1_{\{t > t^*\}}. \tag{5.19}$$

If $B(0,1/\mu) < 1$, then $T_n \Rightarrow t^* = T^*$ as $n \to \infty$. If, in addition, the abandonment distribution is exponential, i.e., if $\bar{F}(x) = e^{-\theta x}$, then

$$(\bar{Q}_n, \bar{A}_n) \Rightarrow (Q, A) \in \mathbb{D}^2, \tag{5.20}$$

where $Q(t) = A(t) = 0$ for $0 \leq t \leq t^*$ and

$$\begin{aligned} Q(t) &= \int_0^{t-t^*} \bar{F}(t - t^* - s)\gamma(s)\, ds, & (5.21)\\ &= \int_0^{w(t)} \lambda\, \bar{F}(x)dx, \quad t \geq t^*, & (5.22)\\ A(t) &= \Lambda(t) - \int_0^{t-t^*} b(s,0)\, ds - Q(t), \quad t \geq t^*, & (5.23) \end{aligned}$$

where w satisfies ODE (2.32) with $w(t^*) = 0$, $\gamma(t) \equiv \lambda - b(t,0)$.

We now observe that in general we need not have either $T_n \Rightarrow t^*$ or $T_n^* \Rightarrow T^*$.

Example 5.1 (*counterexample on first passage times*) Suppose that $\lambda > \mu = 1$. Let $b(0,x) = \lambda$, $1 - (1/\lambda) \leq t \leq 1$, and $b(0,x) = 0$, $0 \leq x < 1 - (1/\lambda)$, so that $b(t,0) = \lambda$, $0 \leq t < 1/\lambda$, and $b(t,0) = 0$, $1/\lambda \leq t < 1$, $B(t,1/\mu) = 1$ for all $t \geq 0$ and $T^* = t^* = 0$.

For $n \geq 1$, let $\{B_n(0,y) : 0 \leq y \leq 1\}$ be deterministic. To be a legitimate sample path for a queueing system, $B_n(0,y)$ must be nondecreasing and integer-valued as well as satisfy

$0 \leq B_n(0,y) \leq n$. Thus, let $B_n(0,y) \equiv \lfloor B_n^f(0,y) \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $\bar{B}_n^f(0,y) \equiv n^{-1} B_n^f(0,y) \equiv \int_0^y b_n(0,x)\,dx$, where $b_n(0,x) = ((n+1)/n)\lambda$, $1 - ((n-1)/n\lambda) \leq t \leq 1$, and $b_n(0,x) = 0$, $0 \leq x < 1 - ((n-1)/n\lambda)$. First, observe that $\bar{B}_n^f(0,1/\mu) = (n^2-1)/n^2 < 1$ for all $n \geq 1$. Second, observe that we have $0 \leq \bar{B}_n^f(0,y) - \bar{B}_n(0,y) \leq 1/n$ for all y and n. Hence, $\bar{B}_n(0,1/\mu) \leq \bar{B}_n^f(0,1/\mu) < 1$ for all $n \geq 1$. Nevertheless, $\bar{B}_n(0,\cdot) \to B(0,\cdot)$ as $n \to \infty$. On the other hand, consider a deterministic arrival process with rate $n\,\lambda$, i.e., with $N_n(t) \equiv \lfloor n\,\lambda\,t \rfloor$, $t \geq 0$, $n \geq 1$. Then $S_n(t) = \lfloor (n+1)\lambda\,t \rfloor \geq N_n(t)$ for $0 \leq t \leq (n-1)/n\lambda$. Since $B_n(0,1/\mu) < n$, the system is underloaded for $0 \leq t < 1/\lambda$. However, $N_n(1/\lambda) = n$. Hence, $T_n = T_n^* = 1/\lambda$ for all $n \geq 1$, in contrast to $t^* = T^* = 0$. A similar example can be constructed if $B(0,1/\mu) < 1$ and condition (5.15) is not imposed; see Appendix D.8.

5.4 The $G/D/s + GI$ Fluid Queue

We now study the $G/D/s + GI$ fluid queue. The corresponding $G_t/GI/s_t + GI$ model, having time-varying arrival rate (G_t), time-varying staffing (s_t) and a general service-time distribution (GI) was studied in Chapter 2. Here we restrict attention to constant arrival rate λ and constant staffing s, although the model can easily be extended to allow these functions to be time-varying.

Paralleling to §5.4, we define the total input $\Lambda(t)$, departure rate $\sigma(t)$, total output $S(t)$, total fluid abandoned $A(t)$, fluid in queue (service) that has been in queue (service) for at most x $B(t,x)$ ($Q(t,x)$), total quanty of fluid $X(t)$, fluid density in queue (service) $q(t,x)$ ($b(t,x)$), and the boundary of waiting time $w(t)$, in the identical way as in §5.4. This model has constant staffing $s(t) = s$.

We assume Assumptions 2.1-2.5 are satisfied. In addition, we make the following assumptions.

Because the service time is deterministic, each quantum of fluid that enters service stays in service for time $1/\mu$ before leaving the system. The total service completion rate at time t is the density of fluid that has been in service for $1/\mu$. That is also the rate into service $1/\mu$ time units before, i.e.,

$$\sigma(t) \equiv b(t, 1/\mu) = b(t - 1/\mu, 0), \quad t \geq 0. \tag{5.24}$$

Let $E(t)$ be the amount of fluid to enter service in $[0, t]$; then

$$E(t) \equiv \int_0^t b(u, 0)\, du, \quad t \geq 0, \tag{5.25}$$

where $b(t, 0)$ is the rate fluid enters service at time t. The rate fluid enters service depends on whether the system is underloaded or overloaded. If the system is underloaded, then the external input directly enters service; if the system is overloaded, then the fluid to enter service is determined by the rate that service capacity becomes available at time t, which is the departure rate $\sigma(t)$, because the total fluid content in service $B(t) = s$ does not change at t.

Since the service discipline is FCFS, fluid leaves the queue to enter service from the right boundary of $q(t, x)$. The fluid content densities q and b satisfy the following two fundamental evolution equations. (Recall that the service-time ccdf is $\bar{G}(x) = 1_{\{0 \leq x \leq 1/\mu\}}$.)

Paralleling (2.6), we have the following fundamental evolution equations.

Assumption 5.4 (*fundamental evolution equations*) *For* $t \geq 0$, $x \geq 0$ *and* $u \geq 0$,

$$q(t+u, x+u) = q(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad 0 \leq x < w(t), \tag{5.26}$$

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)} = b(t,x)\,1_{\{x+u \leq 1/\mu\}}. \tag{5.27}$$

We assume that all assumptions in this section are in force throughout the chapter.

5.5 Performance of the $G/D/s + GI$ Fluid Queue

In Chapter 2 we showed how the system performance expressed via the basic functions (b, q, w, v) depends on the model data $(\lambda, s, \mu, F, b(0,\cdot), q(0,\cdot))$, for the time-varying fluid models, i.e., for $G_t/GI/s_t + GI$ and $G_t/M_t/s_t + GI_t$. From the basic performance four-tuple (b, q, w, v), we easily compute the associated vector of performance functions $(\hat{B}, \hat{Q}, B, Q, X, \sigma, S, \alpha, A, E)$ via the definitions in §5.4. We now establish similar results for the basic functions (b, q, w, v) of the $G/D/s + GI$ model.

The service content density b is elementary within each interval that the system is either entirely underloaded or entirely overloaded. The complications occur when there are changes from one regime to the other. We state basic results in this section and others in the next section. The results here provide the basis for an effective algorithm, assuming that there are only finitely many changes between underloaded and overloaded regimes in each interval $[0, T]$, for which we give a sufficient condition at the end of this section.

Theorem 5.3 (service content in the underloaded case) *For the* $G/D/s + GI$ *fluid model*

with unlimited service capacity $(s \equiv \infty)$*, starting at time* 0*,*

$$\begin{aligned} b(t,x) &= b(0,x-t)\cdot 1_{\{0\le t<x\le 1/\mu\}} \quad + \lambda\cdot 1_{\{0\le x\le 1/\mu,\, x\le t\}}, \\ B(t) &= \left(\lambda\, t + \int_t^{1/\mu} b(0,x-t)\,dx\right) 1_{\{0\le t\le \frac{1}{\mu}\}} + \frac{\lambda}{\mu} 1_{\{t>\frac{1}{\mu}\}}. \end{aligned} \tag{5.28}$$

If, instead, a finite-capacity system starts underloaded, then the same formulas apply over the interval $[0,T)$*, where* $T \equiv \inf\{t \ge 0 : B(t) > s\}$*, with* $T = \infty$ *if the infimum is never obtained. Hence,* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* $t \ge 0$ *and* $x \ge 0$*, for* t *in the underloaded interval.*

Proof. To show the first relation, note that $b(t,x) = 0$ for all $x > 1/\mu$ because the service time is exactly $1/\mu$. If $0 \le t \le 1/\mu$, $b(t,x) = b(0,x-t)$ for $t < x \le 1/\mu$ and $b(t,x) = \lambda$ for $0 \le x \le t$. If $t > 1/\mu$, then all fluid that was in service at time 0 is gone, hence $b(t,x) = \lambda$ if $0 \le x \le 1/\mu$. Simply integrating the first relation gives the second. □

Corollary 5.2 (reaches steady state at time $1/\mu$) *If the system is entirely underloaded, then the performance reaches steady state by time* $1/\mu$ *with* $\sigma(t) = b(t,x) = \lambda$*,* $0 \le x \le 1/\mu$ *and* $t \ge 1/\mu$*.*

The periodic behavior observed in the overloaded numerical examples is mostly explained by the following theorem and the subsequent Corollary 5.3.

Theorem 5.4 (service content in the overloaded case) *For the* $G/D/s+GI$ *fluid model in*

an overloaded interval, $B(t) = s$ *and*

$$\begin{aligned} b(t,x) &= b(0,x-t)\cdot 1_{\{0\le t<x\le 1/\mu\}} \\ &\quad + b\left(0, \frac{1}{\mu} - (t-x) + \frac{\lfloor (t-x)\mu \rfloor}{\mu}\right)\cdot 1_{\{0\le x\le 1/\mu,\, x\le t\}}, \end{aligned} \tag{5.29}$$

where $\lfloor x \rfloor$ *is the integer part of a real number* x*. Hence,* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* $t \ge 0$ *and* $x \ge 0$ *in an overloaded interval.*

Proof. Note $b(t,x) = 0$ for all $x > 1/\mu$. If $0 \le t \le 1/\mu$, $b(t,x) = b(0,x-t)$ for $t < x \le 1/\mu$; $b(t,x) = b(t-x,0) = \sigma(t-x) = b(0, 1/\mu - (t-x))$ for $0 \le x \le t$. If $t > 1/\mu$, then $t - x > 0$. Let $N \equiv \lfloor (t-x)\mu \rfloor$, we have $0 \le t - x - N/\mu \le 1/\mu$. Hence $b(t,x) = b(t-x,0) = \sigma(t-x) = \sigma(t-x-N/\mu) = b(0, 1/\mu - (t-x-N/\mu))$. Moreover, simple calculation by integrating (5.29) over x verifies that indeed $B(t) = \int_0^{1/\mu} b(t,x)dx = s$. □

Corollary 5.3 (periodic performance in service starts at time 0) *If* $B(t) = s$ *for all* $t \ge 0$*, then the density* b *is either stationary or in a PSS starting at time* 0*. It is stationary if* $b(0,x) = s\mu$*,* $0 \le x \le 1/\mu$*. Otherwise it is in a PSS with*

$$b\left(\frac{k}{\mu} + t, x\right) = b(t,x), \quad \sigma\left(\frac{k}{\mu} + t\right) = \sigma(t),$$

for $0 \le x \le 1/\mu$*,* $0 \le t \le 1/\mu$ *and* $k \ge 0$*.*

Corollary 5.4 (overall smoothness for the service content) *If the system changes regimes only finitely often in the interval* $[0,T]$*, then* $b(t,\cdot)$*,* $b(\cdot,x)$*,* $B \in \mathbb{C}_p$ *for all* t*,* $0 \le t \le T$*, and* $x \ge 0$*.*

The $G/D/s+GI$ model differs from the $G_t/GI/s_t+GI$ model in Chapter 2 in the service facility, but not in the queue. Therefore, the dynamics of q, w and v are the same. Their dynamics are described by Proposition 2.6, Corollary 2.2, Theorem 2.3, 2.5 and 2.6. Similarly, the regime termination criterion are characterized by that in Chapter 2.

We now provide a sufficient condition for there to be only finitely many switches between overloaded and underloaded intervals in any bounded interval $[0,T]$. To do so, we use a function involving the model elements λ and $b(0,x)$, $0 \le x \le 1/\mu$. In particular, let

$$\zeta(x) \equiv \sigma(x) - \lambda = b(0, 1/\mu - x) - \lambda.$$

Let $\mathcal{D}_\zeta$ be the set of discontinuities of ζ in $[0,1/\mu]$, let $\bar{\mathcal{Z}}_\zeta \equiv \{x \in [0,1/\mu] : \zeta(x) = 0\}$ be the zero set of ζ, and let $\mathcal{Z}_\zeta$, be a subset of $\bar{\mathcal{Z}}_\zeta$, defined by

$$\mathcal{Z}_\zeta \equiv \{x \in \bar{\mathcal{Z}}_\zeta : \nexists\ \epsilon > 0 \text{ such that } \zeta(y) = 0 \text{ for all } y \in (x-\epsilon, x+\epsilon)\}$$

The subset $\mathcal{Z}_\zeta$ excludes those points $x \in [0,1/\mu]$ such that $\zeta(x) = 0$ for $x \in (a,b)$.

Let $\mathcal{S}_T$ be the total number of regime-switching (between overloaded and underloaded) points in $[0,T]$ as in Chapters 2-3. For any set A, let $|A|$ be the cardinality of A.

Theorem 5.5 (*relating switches to zeros and discontinuities of* ζ) *For any interval* $[0,T]$ *with* $T \ge 1/\mu$,

$$|\mathcal{S}_T| \le \lceil T\mu \rceil (|\mathcal{Z}_\zeta| + |\mathcal{D}_\zeta| + 1), \tag{5.30}$$

where $\lceil x \rceil$ *is least integer greater than or equal to* x.

Remark 5.1 (*tightness of the bound in Theorem 5.5*) To show that the bound in Theorem 5.5 is tight, consider a $G/D/s + GI$ fluid queue in $[0, T] = [0, 2/3\mu]$ that is initially critically loaded, i.e., $B(0) = s$ and $Q(0) = 0$, with $b(0, x) = 2\mu\, s \cdot 1_{\{1/2\mu \le x \le 2/3\mu\}}$ and $\lambda = 1.5\mu\, s$. We know $\sigma(t) = b(0, 1/\mu - t) = 2\mu\, s \cdot 1_{\{0 \le t \le 1/2\mu\}}$. Hence, $B'(t) = \lambda - \sigma(t) = -0.5\mu\, s \cdot 1_{\{0 \le t \le 1/2\mu\}} + 1.5\mu\, s \cdot 1_{\{1/2\mu \le t \le 2/3\mu\}}$, which implies that $B(t) = (s - 0.5\mu\, s\, t) \cdot 1_{\{0 \le t \le 1/2\mu\}} + 1.5\mu\, s\, t \cdot 1_{\{1/2\mu \le t \le 2/3\mu\}}$. Therefore the system is underloaded in $[0, 2/3\mu]$ and becomes critically loaded again at $t = 2/3\mu$. In this case the bound in Theorem 5.5 is tight because $N = \lfloor 2/3 \rfloor + 1 = 1$, $|\mathcal{D}_\zeta| = 1$, $|\mathcal{Z}_\zeta| = 0$ and $|\mathcal{S}_T| = 2$, where the two switching points are 0 and $2/3\mu$.

Assumption 5.5 (*controlling the number of switches*) *For $\mu > 0$, $|\mathcal{Z}_\zeta| < \infty$, so that there are only finitely many switches between overloaded and overloaded intervals in any bounded subinterval.*

We assume that Assumption 5.5 is in force throughout the chapter.

Remark 5.2 (*an algorithm*) These results yield an efficient algorithm to compute the basic performance four tuple (b, q, w, v). First, we can compute $b(t, x)$ directly via Theorems 5.3 and 5.4. We compute $\tilde{q}$ directly from Proposition 2.6. We then compute the BWT w by solving the ODE in Theorem 2.3. The proof of Theorem 2.5 in Chapter 2 provides an elementary algorithm to compute v once w has been computed. Theorem 6 of Chapter 2 shows that v satisfies its own ODE under additional regularity conditions. Theorem 5.3 and 5.3 specify how to switch between alternating overloaded and underloaded inter-

vals. Assumption 5.5 ensures that the total number of switches between underloaded and overloaded intervals is finite.

5.6 The Fluid Model Eventually Always Overloaded

For the rest of this chapter, we assume that the fluid arrival rate λ exceeds the maximum possible long-run average service rate $s\mu$, so that $\rho \equiv \lambda/s\mu > 1$.

Assumption 5.6 $(\rho > 1)$ $\lambda > s\mu$.

We say that the service capacity (and thus the system) is overloaded at time t if $B(t) = s$. In this section we describe the fluid density in service, b, in the $G/D/n+GI$ fluid model assuming that there exists a finite time after which the system stays overloaded; let T^* be the first such time, i.e.,

$$T^* \equiv \inf \{t \geq 0 : B(u) = s \quad \text{for all} \quad u \geq t\}, \tag{5.31}$$

with $T^* \equiv \infty$ if there exists no such time.

We also provide a sufficient condition for T^* to be finite. We show that the service density b reaches a PSS at time T^*. In the next two sections we use this assumption to show that the queue performance (e.g. $Q(t)$ and $\alpha(t)$) converges to a PSS after time T^*. (These auxiliary performance functions typically do not reach PSS in finite time.)

Assumption 5.7 $(a\ time\ after\ which\ the\ system\ remains\ overloaded)$ *For* T^* *defined in* (5.31), $T^* < \infty$.

Assumption 5.7 is very useful because it identifies the time at which the service fluid density b reaches a PSS. The following is a consequence of Theorem 5.4 and Corollary 5.3.

Corollary 5.5 $(a\ PSS\ for\ b\ starting\ at\ T^*)$ *Under Assumption* 5.7*, the service fluid density* b *either reaches steady state or a PSS at time* T^**; i.e.,*

$$b((n/\mu)+t,x) = b(t,x), \quad n \geq 1, \quad t \geq T^*, \quad 0 \leq x \leq 1/\mu.$$

A steady state is achieved if and only if $b(T^*,x) = s\mu$*,* $0 \leq x \leq 1/\mu$*.*

In applications it is not necessary to identify T^*; it suffices to identify *any* time t with $t \geq T^*$. Corollary 5.5 implies that b is in a PSS starting at any time $t \geq T^*$. We now provide a sufficient condition for Assumption 5.7. To do so, let t^* be the time that the service facility *first* becomes full; i.e.,

$$t^* \equiv \inf\left\{t \geq 0 : \lambda t + B(0) - \int_0^t \sigma(x)dx = s\right\}. \tag{5.32}$$

If the system is initially overloaded, then $t^* = 0$. Necessarily $t^* < 1/\mu$, because no new input during the interval $[0, 1/\mu]$ can depart in that interval and $\lambda/\mu > s$, since $\rho \equiv \lambda/s\mu > 1$. Define a class of initial service densities

$$\mathcal{B}^*_{s,\lambda} \equiv \left\{b(0,\cdot) : B(0) = \int_0^{1/\mu} b(0,x) = s,\ b(0,x-t^*) \leq \lambda,\ t^* \leq x \leq 1/\mu\right\}.$$

Theorem 5.6 $(a\ sufficient\ condition\ for\ Assumption\ 5.7)$ *If* $b(0,\cdot) \in \mathcal{B}^*_{s,\lambda}$*, then Assumption* 5.7 *is satisfied with* $T^* = t^*$ *for* T^* *in* (5.31) *and* t^* *in* (5.32)*.*

Proof. If $t^* = 0$, i.e., $B(0) = s$ and $b(0, x) \leq \lambda$, $0 \leq x \leq 1/\mu$, then new fluid will arrive in the system at least as fast as the fluid is departing, throughout the interval $[0, 1/\mu]$. Hence, a full service facility is maintained throughout the interval $[0, 1/\mu]$. Hence fluid enters service immediately replacing all departing fluid. (This fluid will enter from the head of the queue if the queue is not empty, but that is not important for b.) Thus, the service facility remains full forever.

If $t^* > 0$, then $B(0) < s$, so that new fluid will enter service from outside at rate λ until the service facility becomes full at t^*. We have

$$t^* = \inf \{t \geq 0 : \lambda t + B(0, 1/\mu - t) = s\}, \tag{5.33}$$

following from (5.32) and Theorem 5.3. Since $b(0, x) \leq \lambda$ for $t^* \leq x \leq 1/\mu$, the system then reaches the first case starting at t^*, so we can apply the previous analysis to this case. □

Note that the condition of Theorem 5.6 is satisfied in the common case in which the system starts out empty. In §5.8 we will describe the system performance in detail in that special case. Also note that we can apply Theorem 5.6 to the state of the system at any finite time t, not just at time 0. In particular, we can apply the algorithm in Remark 5.2 over some finite interval $[0, t]$ and then check to see if the conditions of Theorem 5.6 are satisfied at time t.

5.7 Structural Results for the Queue Performance

In this section we focus on the performance related to the queue in an overloaded $G/D/s+GI$ fluid model with $\rho > 1$, thus showing how we can exploit Assumptions 5.6 and 5.7 in the previous section. In this section we assume that the fluid queue is overloaded for

all $t \geq 0$. We present four structural results: (i) comparison, (ii) Lipschitz continuity, (iii) asymptotic loss of memory (ALOM) and (iv) uniform boundedness. The proofs of Theorems 5.7-5.10 are also given in Appendix D.3.)

Our comparison result establishes an ordering of the performance functions given an assumed ordering for the model data functions.

Theorem 5.7 (*comparison of fluid content in queue for the overloaded $G/D/s + GI$ model*) *Consider two $G/D/s + GI$ fluid models with common staffing function s, service time $1/\mu$, abandonment cdf F and initial fluid density in service $b(0, \cdot)$. Assume both queues are overloaded for all $t \geq 0$ ($B_1(t) = B_2(t) = s$). If $q_1(0, \cdot) \leq q_2(0, \cdot)$ and $\lambda_1 \leq \lambda_2$, then*

$$(Q_1, q_1, \alpha_1, w_1, v_1) \leq (Q_2, q_2, \alpha_2, w_2, v_2).$$

For an integrable real-valued function x on $[0, \infty)$, let $\|x\|_1 \equiv \int_0^\infty |x(t)|dt$. Also, let

$$b^\downarrow \equiv \inf_{0 \leq x \leq 1/\mu} b(0, x), \quad b^\uparrow \equiv \sup_{0 \leq x \leq 1/\mu} b(0, x),$$
$$h_F^\downarrow \equiv \inf_{0 \leq x < \infty} h_F(x), \quad h_F^\uparrow \equiv \sup_{0 \leq x < \infty} h_F(x).$$

Our Lipschitz continuity result also applies to functions. For it, we use the uniform norm on real-valued functions on the interval $[0, T]$: $\|x\|_T \equiv \sup\{|x(t)| : 0 \leq t \leq T\}$.

Theorem 5.8 (*Lipschitz continuity of fluid content in queue for the overloaded $G/D/s + GI$ model*) *Consider a $G/D/s + GI$ fluid model with arrival rate λ, staffing function s, service time $1/\mu$, abandonment cdf F. Assume the queue is overloaded for all $t \geq 0$.*

Then the function mapping $(\lambda, Q(0))$ *in* $\mathbb{R}^2$ *into* (Q, α) *in* $\mathcal{C}_p^2$ *all over* $[0, T]$ *is Lipschitz continuous. In particular,*

$$\begin{aligned}
\|Q_1 - Q_2\|_T &\leq T|\lambda_1 - \lambda_2| + |Q_1(0) - Q_2(0)| \\
&\leq (1 \vee T)(|\lambda_1 - \lambda_2| \vee |Q_1(0) - Q_2(0)|), \qquad (5.34) \\
\|\alpha_1 - \alpha_2\|_T &\leq h_F^{\uparrow} \|Q_1 - Q_2\|_T, \qquad (5.35) \\
\|q_1 - q_2\|_{T,1} &\equiv \left\| \int_0^\infty q_1(\cdot, x)dx - \int_0^\infty q_2(\cdot, x)dx \right\|_T \\
&\leq T|\lambda_1 - \lambda_2| + \|q_1(0, \cdot) - q_2(0, \cdot)\|_1. \qquad (5.36)
\end{aligned}$$

Theorem 5.9 $(ALOM$ *of fluid content in queue for the overloaded* $G/D/s + GI$ *model)* *Consider two initially overloaded* $G/D/s + GI$ *fluid models (*$B_1(0) = B_2(0) = s$*). Suppose these two models have common arrival rate* λ*, staffing function* s*, service time* $1/\mu$*, abandonment cdf* F*, initial fluid densities in service* $b(0, x)$*, but different initial fluid densities in queue* $q_i(0, \cdot)$*.*

(a) *If both queues are overloaded for all* $t \geq 0$*, then*

$$\begin{aligned}
\Delta Q(T) &= \|q_1(T, \cdot) - q_2(T, \cdot)\|_1 \leq C_1 e^{-h_F^{\downarrow} T}, \qquad (5.37) \\
\Delta\alpha(T) &\leq h_F^{\uparrow} C_1 e^{-h_F^{\downarrow} T},
\end{aligned}$$

where $C_1 \equiv C_1(q_1(0,\cdot), q_2(0,\cdot))$ *is the constant*

$$\begin{aligned} C_1 &\equiv \int_0^\infty ([q_1(0,x) \vee q_2(0,x)] - [q_1(0,x) \wedge q_2(0,x)])dx \\ &\leq Q_1(0) + Q_2(0). \end{aligned} \tag{5.38}$$

In addition, if $b^\downarrow > 0$*, then for* $T > T^*$*,*

$$\begin{aligned} \Delta w(T) &\leq \frac{\Delta Q(T)}{\lambda \bar{F}(w_2(T) \vee w_1(T))} \\ &\leq C_2 \Delta Q(t) \leq (C_2\, C_1) e^{-h_F^\downarrow T}, \end{aligned} \tag{5.39}$$

where

$$\begin{aligned} T^* &\equiv \frac{Q_1(0) + Q_2(0)}{b^\downarrow}, \\ C_2 &\equiv \bar{F}\left[\frac{b^\downarrow}{\lambda} \vee \left(w_1(0) \vee w_2(0) + \frac{Q_1(0) + Q_2(0)}{b^\downarrow}\right)\right]^{-1}. \end{aligned} \tag{5.40}$$

(b) *If, in addition, the initial densities in queue are ordered by*

$$q_1(0,x) \leq q_2(0,x) \quad \textit{for all } x \geq 0, \tag{5.41}$$

then $Q_1(t) \leq Q_2(t)$ *for all* $t \geq 0$*,*

$$\Delta Q'(T) \leq 0 \quad \textit{and} \quad \Delta Q(T) \leq \frac{\Delta Q(0)}{1 + h_F^\downarrow T}, \quad T > 0, \tag{5.42}$$

so that

$$\Delta Q(T) \leq e^{-h_F^{\downarrow} T}\, \Delta Q(0), \quad \Delta\alpha(T) \;\leq\; h_F^{\downarrow} \Delta Q(T). \tag{5.43}$$

For the following boundedness result, we make a stronger assumption on the initial fluid density and the abandonment hazard rate in the model data, requiring that they be uniformly bounded above and below.

Assumption 5.8 (*uniformly bounded initial fluid density and hazard rate*) *The staffing and the rates in the model data are uniformly bounded above and below, i.e.,*

$$0 < b^{\downarrow} \leq b^{\uparrow} < \infty, \quad 0 < h_F^{\downarrow} \leq h_F^{\uparrow} < \infty.$$

Assumption 5.8 strengthens Assumptions 2.1 and 3.6. We assume that this additional assumption is in force for the remainder of the chapter.

Theorem 5.10 (*boundedness*) *Consider the $G/D/s + GI$ fluid queue that is overloaded for all $t \geq 0$. Under Assumption* 5.8 *and the previous the assumptions, all performance functions are uniformly bounded. In particular,*

$$\begin{aligned}
B(t) &= s, \quad b(t,x) \leq b(0,x) \vee b^{\uparrow}, \\
Q(t) &\leq \left(\frac{\lambda}{h_F^{\downarrow}}\right) \vee Q(0), \quad q(t,x) \leq q(0,x) \vee \lambda, \\
w(t) &\leq \bar{F}^{-1}\left(\frac{b^{\downarrow}}{\lambda}\right) \vee \left(\frac{Q(0)}{\gamma^{\downarrow}} + w(0)\right), \\
\alpha(t) &\leq \frac{h_F^{\uparrow}\lambda}{h_F^{\downarrow}}, \quad \textit{and} \quad \sigma(t) = b(t,0) \leq b^{\uparrow}.
\end{aligned}$$

5.8 The Full Performance Under Assumption 5.7

In §5.6 we saw that the fluid density in service, b, reaches steady state or a PSS at time T^* if the system remains overloaded after time T^*, as stipulated in Assumption 5.7. We now exploit the structural results in the previous section to describe the full queue performance, given Assumption 5.7. In the next section we show that Assumption 5.7 is not always satisfied.

As in §4.5 of Chapter 4, we consider the performance vector at time t $\mathcal{P}(t)$ defined by (4.39). If the initial condition $\mathcal{P}(0)$ can be chosen so that $\{\mathcal{P}(t) : t \geq 0\}$ is a periodic function of t with period τ, then this initial condition produces a PSS. If not, we want to show that the performance converges to a PSS $\mathcal{P}^*$ as time evolves. We follow our discussion on PSS as in §4.5 of Chapter 4. To discuss continuity and convergence in the domain of $\mathcal{P}$, we use norm $\|\mathcal{P}(t)\|$ defined by (4.40) in §4.5.

We primarily want to establish convergence to a PSS, but we also treat the case of stationary performance, which arises when $b(T^*, x) = s\mu$, $0 \leq x \leq 1/\mu$. Given that stationary b, the remaining stationary performance can be obtained by the reasoning in Theorem 4.4 of Chapter 4. The remaining stationary performance measures are

$$\begin{aligned} B &= s, \quad \alpha = \lambda - s\mu, \quad w = \bar{F}^{-1}(s\mu/\lambda), \\ Q &= \lambda \int_0^w \bar{F}(x)\, dx, \quad \text{and} \quad q(x) = \lambda\, \bar{F}(x), \quad 0 \leq x \leq w. \end{aligned} \tag{5.44}$$

Theorem 5.11 (*PSS for the overloaded $G/D/s + GI$ fluid model*) *Suppose that Assumption* 5.7 *is satisfied in the $G/D/s + GI$ fluid model with $\rho > 1$. If $b(T^*, x) = s\mu$,*

$0 \le x \le 1/\mu$, *then there exists a constant function* $\mathcal{P}^*$ *as in* (5.44) *such that*

$$\|\Psi_\tau^{(n)}(\mathcal{P}) - \mathcal{P}^*\| \to 0 \quad \textit{as} \quad n \to \infty. \tag{5.45}$$

for all $\tau > 0$. *Otherwise, the fluid performance* $\mathcal{P}$ *is asymptotically periodic with period* $1/\mu$, *i.e., there exists a periodic function* $\mathcal{P}^*$ *with period* $1/\mu$ *such that* (5.45) *holds for* $\tau \equiv 1/\mu$.

Proof. We can treat the two cases together by the same argument; we only discuss the second case. We must show that $\|\mathcal{P}((n/\mu) + \cdot) - \mathcal{P}^*(\cdot)\| \to 0$ as $n \to \infty$. However, since $\mathcal{P}^*$ is periodic and $\Psi_{1/\mu}^{(n)}(\mathcal{P})$ involves the shift operator, it suffices to prove that $\|\mathcal{P}((n/\mu) + \cdot) - \mathcal{P}^*(\cdot)\|_{1/\mu} \to 0$ as $n \to \infty$, where the supremum in the norm is over the finite interval $[0, 1/\mu]$, i.e., for $\|\mathcal{P}\|_{1/\mu} \equiv \sup\{|\mathcal{P}(t)| : 0 \le t \le 1/\mu\}$. That in turn is a form of the norm in Theorem 3.6.

If $T^* > 0$, we can simply move the origin to T^*. Therefore, it remains to consider the case where the system is initially overloaded, and remains so thereafter. In that case, $b(t, x)$ and $\sigma(t) = b(t, 0)$ are periodic with period $1/\mu$ starting from $t = 0$, by Theorem 5.4 and Corollary 5.3.

Next, suppose that $q(0, x) = 0$ for $x \ge 0$, i.e., the system is initially critically loaded. By Theorem 5.7, the shift operator $\Psi_{1/\mu}$ is a monotone operator on $\mathcal{P}((n/\mu) + \cdot)$ for any n, because we can think of the performance $q(1/\mu, \cdot)$ as alternative initial conditions for the model at time 0, since the model is periodic with period $1/\mu$ (λ and s are constant, $b(t, 0)$ is periodic with period $1/\mu$ by Theorem 5.4 and Corollary 5.3. Therefore, the sequence of system performance functions $\mathcal{P}(0 + \cdot), \mathcal{P}((1/\mu) + \cdot), \mathcal{P}((2/\mu) + \cdot), \ldots$ (at discrete time $0, 1/\mu, 2/\mu, \ldots$) is monotonically non-decreasing. Since the performance is also bounded, by Theorem 5.10, there is a finite limit for the sequence $\{\mathcal{P}((n/\mu) + \cdot)\}$ as $n \to \infty$. By

Theorem 3.6, the operator is continuous as well, which implies that $\Psi_{1/\mu}^{(n)}(\mathcal{P})$ is convergent in the specified norm as $n \to \infty$. Hence the limit is a PSS. By the ALOM property in Theorem 5.9, we get the same limit for all other initial fluid densities in queue $q(0,\cdot)$. $\square$

Remark 5.3 (*computation*) Given the rapid convergence, it usually is not difficult to compute the PSS associated with any given initial condition by simply applying the algorithm with that initial condition. We can then verify that the condition in Theorem 5.6 is satisfied after some finite time, so that we know T^* and we know the PSS for the fluid density in service b. We then can observe the convergence of the other performance measures. However, the PSS for the remaining performance functions can also be determined in another way, given T^* and b. First, if the abandonment distribution is exponential, then analytic expressions are available, see Corollary 5.8. Second, for the case of non-exponential abandonment, consider a cycle $[0, 1/\mu]$ of the PSS. For each candidate $\tilde{w} \geq 0$, we numerically solve the ODE (2.31) in $[0, 1/\mu]$ with $w(0) = \tilde{w}$ and $b(t, 0) = b(T^*, 1/\mu - t)$ and check if $w(1/\mu) = \tilde{w}$. Since $\tilde{w} \geq 0$ is our only unknown variable, we shall do a search for $\tilde{w} \geq 0$. Theorem 5.11 guarantees the existence and uniqueness of such a $\tilde{w} \geq 0$.

Remark 5.4 (*different initial conditions*) Theorems 5.6 and 5.11 provide sufficient conditions for Assumption 5.7 to hold, and for the performance function to converge to a PSS. That PSS depends strongly on the fluid density in service, b at the time T^* after which the system remains overloaded. In Appendix D.4 we show that very different PSS's can result by considering two different initial conditions for the example in §5.1.

We now describe the time-average performance over a periodic cycle. Some average

performance measures are independent of the initial conditions, and thus agree with the stationary performance, whereas others are not.

Corollary 5.6 $(\textit{average performance over a cycle})$ *Suppose that Assumption* 5.7 *holds for a* $G/D/s+GI$ *fluid queue and consider the PSS beginning at* T^**. The average abandonment rate* $\bar{\alpha}$ *and departure rate* $\bar{\sigma}$ *over a cycle* $[0,\tau] \equiv [0, 1/\mu]$ *of the PSS are*

$$\bar{\alpha} \equiv \frac{1}{\tau}\int_0^\tau \alpha(t)dt \;=\; \alpha^* \equiv \lambda - \mu\, s \tag{5.46}$$

$$\bar{\sigma} \equiv \frac{1}{\tau}\int_0^\tau \sigma(t)dt \;=\; \sigma^* \equiv \mu\, s, \tag{5.47}$$

If, in addition, the abandonment distribution is exponential, then

$$\bar{Q} \equiv \frac{1}{\tau}\int_0^\tau Q(t)dt \;=\; Q^* \equiv \int_0^{w^*} \lambda\, e^{-\theta\, x}\, dx. \tag{5.48}$$

where α^**,* σ^**,* Q^* *and* $w^* \equiv \bar{F}^{-1}(1/\rho)$ *are the stationary abandonment and departure rates, queue length and BWT given in* (5.44)*.*

Proof. First, (5.48) follows from (5.46) when $\bar{F}(x) = e^{-\theta\, x}$, because $\alpha(t) = \theta\, Q(t)$, which implies

$$\bar{Q} = \frac{1}{\theta}\bar{\alpha} = \frac{1}{\theta}(\lambda - \mu\, s),$$

which is equal to the right hand side of (5.48), as can be verified by simple calculation. Since the system is overloaded for all $t \geq T^*$, then $b(t,x)$ and $\sigma(t)$ are periodic for all $t \geq T^*$, by Theorem 5.4 and Corollary 5.3. Therefore, consider a cycle $[0, 1/\mu]$ of the PSS,

we must have $b(t,0) = \sigma(t) = b(T', 1/\mu - t)$ for some $T' \geq T^*$. Hence, (5.47) follows because $\int_0^{1/\mu} b(T', 1/\mu - t)dt = B(T') = s$.

To show (5.46), flow conservation of the queue implies that

$$Q'(t) = \lambda - \alpha(t) - b(t,0) = \lambda - \alpha(t) - \sigma(t), \quad \text{for } 0 \leq t \leq 1/\mu.$$

Integrating both sides from 0 to $1/\mu$ yields that

$$0 = Q(1/\mu) - Q(0) = \lambda\,\tau - \int_0^\tau \alpha(t)dt - \int_0^\tau \sigma(t)dt = \lambda\,\tau - \int_0^\tau \alpha(t)dt - \mu\,s\,\tau,$$

which implies (5.46). □

Remark 5.5 (*average of other performance functions*) Except for $\bar{\alpha}$ and $\bar{\sigma}$, the average of other performance functions in PSS typically does not agree with the corresponding stationary values. We illustrate with an example in Appendix D.5, considering Erlang and hyperexponential abandonment cdf's. In our numerical examples we found that the average BWT $\bar{w}$ is consistently greater than the stationary value w^*. In contrast the average $\bar{Q}$ is greater (less) than or equal to the stationary value Q^* when the abandonment-time cdf F is more (less) variable than exponential. It remains to establish supporting theorems.

A common case occurs when the system is initially empty. Obviously this initial condition belongs to class $\mathcal{B}^*_{s,\lambda}$. We next establish results for this special case.

Corollary 5.7 (*PSS for the initially empty $G/D/s{+}GI$ fluid model*) *Consider the $G/D/s{+}GI$ fluid model with $\rho > 1$. If the system is initially empty, then the performance $\mathcal{P}$ is asymptotically periodic and converges to a unique PSS $\mathcal{P}^*$ with period $\tau = 1/\mu$. In partic-*

ular, $B(t) = s$*,* $b(t,x)$ *and* $\sigma(t)$ *are periodic after* s/λ*,*

$$\begin{aligned} b(t+k/\mu, x) &= \begin{cases} \lambda \cdot 1_{\{0\le x\le t-1/\mu+s/\lambda\}\cup\{t\le x\le 1/\mu\}}, & \text{if } \frac{s}{\lambda} < t \le \frac{1}{\mu}, \\ \lambda \cdot 1_{\{t\le x\le t+s/\lambda\}}, & \text{if } \frac{1}{\mu} < t \le \frac{1}{\mu} + \frac{s}{\lambda}. \end{cases} \\ \sigma(t+k/\mu) &= b(t+k/\mu, 0) = \lambda\, 1_{\{1/\mu<t\le 1/\mu+s/\lambda\}}, \quad \text{for } k \ge 0. \end{aligned}$$

Performance functions in queue converge to a PSS with the following structure:

$$\begin{aligned} q(t+k/\mu, x) &\to \lambda\, \bar{F}(x) \cdot 1_{\{0\le x\le w^*(t)\}}, \\ Q(t+k/\mu) &\to \int_0^{w^*(t)} \lambda\, \bar{F}(x) dx, \\ \alpha(t+k/\mu) &\to \int_0^{w^*(t)} \lambda\, f(x) dx, \\ w(t+k/\mu) &\to w^*(t), \quad \text{as } k \to \infty, \end{aligned} \tag{5.49}$$

where $w^*(t) = \tilde{w} + t$ *(linear) for* $s/\lambda \le t \le 1/\mu$ *for some* $\tilde{w} \ge 0$*;* $w^*(t)$ *solves ODE* $w'(t) = 1 - 1/\bar{F}(w(t))$ *for* $1/\mu \le t \le 1/\mu + s/\lambda$ *with* $w(s/\lambda + 1/\mu) = \tilde{w}$*.*

Proof. Since the system is initially empty, it becomes overloaded at time $t^* = s/\lambda < 1/\mu$ and stays overloaded for all $t \ge t^*$ by Theorem 5.6. Hence, the formulas for b follow from Theorem 5.4 and Corollary 5.3. The convergence of other performance functions follows from (5.49). Therefore, it remains to show (5.49). Since $\sigma(t) = b(t,0) = 0$ for $(k-1)/\mu + s/\lambda < t \le k/\mu$, the BWT ODE (2.31) in Theorem 2.3 implies that $w'(t) = 1$ so that $w(t)$ is linear with slope 1 for $(k-1)/\mu + s/\lambda < t \le k/\mu$. □

We now give explicit expressions for the PSS of the $G/D/s+M$ fluid queue that has exponential abandonment and is initially empty. We give the proof in Appendix D.6.

Corollary 5.8 $(\textit{explicit expression for the PSS of the } G/D/s+M \textit{ fluid queue starting empty})$ *Consider the* $G/D/s+M$ *fluid queue starting out empty, with arrival rate* λ*, service time* $1/\mu$*, staffing* s*, exponential abandonment with rate* θ *and* $\rho \equiv \lambda/s\mu > 1$*. The system becomes overloaded and remains so at time* $t^* = T^* = s/\lambda$*. In the PSS (starting at time* 0*) the system is overloaded with performance functions given in two parts* $([0, 1/\mu - s/\lambda]$ *and* $(1/\mu - s/\lambda, 1/\mu])$ *of a cycle* $0 \le t \le 1/\mu$*:*

(a) *In the first part of the PSS cycle, for* $0 \le t \le 1/\mu - s/\lambda$*,*

$$\begin{aligned} w(t) &= t + \tilde{w}, && (5.50)\\ Q(t) &= \frac{\lambda}{\theta}\left[1 - \left(\frac{1 - e^{-\theta\, s/\lambda}}{1 - e^{-\theta/\mu}}\right) e^{-\theta\, t}\right], && (5.51)\\ b(t,x) &= \lambda \cdot 1_{\{t \le x \le t + s/\lambda\}}, \\ \sigma(t) &= b(t,0) = 0, \end{aligned}$$

where

$$\tilde{w} \equiv w(0) = w(1/\mu) = \frac{1}{\theta} \log\left(\frac{1 - e^{-\theta/\mu}}{1 - e^{-\theta\, s/\lambda}}\right) \ge 0. \qquad (5.52)$$

(b) *In the second part of the PSS cycle, for* $1/\mu - s/\lambda < t \le 1/\mu$*,*

$$
\begin{aligned}
w(t) &= -\frac{1}{\theta}\log\left(1+\left(\frac{1-e^{\theta(1/\mu-s/\lambda)}}{1-e^{-\theta/\mu}}\right)\cdot e^{-\theta\, t}\right), && (5.53)\\
Q(t) &= \frac{\lambda}{\theta}\left(\frac{e^{\theta(1/\mu-s/\lambda)}-1}{1-e^{-\theta/\mu}}\right)e^{-\theta\, t}, && (5.54)\\
b(t,x) &= \lambda\cdot 1_{\{0\le x\le t-1/\mu+s/\lambda\}\cup\{t\le x\le 1/\mu\}},\\
\sigma(t) &= b(t,0)=\lambda.
\end{aligned}
$$

In addition, for $0 \le t \le 1/\mu$*,*

$$B(t) = s, \quad q(t,x) = \lambda\,\bar{F}(x)\cdot 1_{\{0\le x\le w(t)\}}, \quad \alpha(t) = \theta\, Q(t),$$

(c) *If we consider a cycle* $[1/\mu - \tilde{w}, 2/\mu - \tilde{w}]$*, then the PWT*

$$v(t) = \frac{1}{\theta}\log\left(1+\left(e^{\theta/\mu}\frac{e^{\theta(1/\mu-s/\lambda)}-1}{1-e^{-\theta/\mu}}\right)\cdot e^{-\theta\, t}\right), \qquad (5.55)$$

for $1/\mu - \tilde{w} \le t < 2/\mu - \tilde{w}$ *and* v *jumps at* $2/\mu - \tilde{w}$ *to*

$$v(2/\mu - \tilde{w}) = v(1/\mu - \tilde{w}) = \tilde{w} + 1/\mu - s/\lambda.$$

Remark 5.6 Since we have an explicit expression for $Q(t)$, in which it is an exponential function in both (a) and (b), simple calculation directly verifies (5.48) in Corollary 5.6.

5.9 General Initial Conditions

In §5.7 and §5.8, we provided a quite complete description of system performance if there exists a finite time T^* such that the system is overloaded for all $t \geq T^*$. Moreover, Theorem 5.6 provides widely applicable conditions for the time T^* to coincide with t^*, the first time t that $B(t) = s$, which necessarily is less than or equal to $1/\mu$. More generally, Theorem 5.6 can be applied to show that the time T^* exists subsequently after applying the numerical algorithm to compute the performance over an initial interval, because we can check to see if the conditions in Theorem 5.6 hold after some finite time.

Nevertheless, we now show that in general there need not exist a finite time such that the system remains overloaded thereafter, i.e., T^* can be ∞. We have seen that the system necessarily becomes overloaded for a first time t^* with $t^* < 1/\mu$. However, with $\rho > 1$, it is possible for the the system to switch between overloaded and underloaded regimes infinitely often.

Theorem 5.12 *There need not exist a finite time T^* such that $B(t) = s$ for all $t \geq T^*$.*

Proof. We provide an explicit counterexample. We consider a $G/D/s + M$ fluid queue with $\lambda = 1.2$, $\mu = s = 1$, $\theta = 2$. Let the queue be initially overloaded with

$$\begin{aligned} b(0,x) &= 2 \cdot 1_{\{1/2 \leq x \leq 1\}} \quad \text{so that } B(0) = s = 1, \\ w(0) &= 2 \quad \text{and} \quad q(0,x) = \lambda\, e^{-\theta\, x} \cdot 1_{\{0 \leq x \leq w(0)\}} = 2\, e^{-2\, x} \cdot 1_{\{0 \leq x \leq 2\}}. \end{aligned}$$

We can apply mathematical induction to show that $B(n) = s$ and $B(n + 1/2) < B(n + 3/2) < s$ for all $n \geq 1$. We elaborate in Appendix D.7. □

Remark 5.7 (*The influence of $q(0,x)$*) It is important to note that the initial queue fluid density $q(0, \cdot)$ plays an important role, both in the counterexample above and in the system

performance more generally. For $t \geq T^*$, $q(t,\cdot)$ plays only a minor role, because then we have ALOM for the queue performance, by virtue of Theorem 5.9. However, the initial queue fluid density $q(0,\cdot)$ plays an important role in determining if $T^* < \infty$ and the form of the PSS. In §D.7 we consider the above example with the same initial fluid density in service but different initial fluid in queue ($w(0) = 0.2$ instead of $w(0) = 2$). There we show that this different value for $w(0)$ (initial fluid in queue) completely changes both the transient evolution of performance functions and the structure of the PSS.

We now obtain additional results for general initial conditions. To do so, let $\Lambda^{(n)}$ be the set of time points at which the rate of fluid entering service is equal to the arrival rate in the n^{th} cycle $[(n-1)/\mu, n/\mu]$, i.e.,

$$\Lambda^{(n)} \equiv \{t \in [0, 1/\mu] : b(t + (n-1)/\mu, 0) = \lambda\}. \tag{5.56}$$

For the example in the proof of Theorem 5.12, $\Lambda^{(n)} = [t_1^{(n)}, t_2^{(n)}]$ (see Appendix D.7). Since $t_1^{(n)}$ is strictly decreasing and $t_2^{(n)}$ is strictly increasing, we have $\Lambda^{(n)} \subseteq \Lambda^{(n+1)}$. In general $\Lambda^{(n)}$ may not be a single closed interval as in this case, nevertheless the monotonicity still holds in general.

Theorem 5.13 (*monotone convergence of the sets* $\Lambda^{(n)}$)

(a) *The sequence* $\{\Lambda^{(n)} : n \geq 1\}$ *is monotonically increasing, i.e.,*

$$\Lambda^{(n)} \subseteq \Lambda^{(n+1)} \quad \textit{for all } n \geq 1.$$

(b) *The sequence* $\{\Lambda^{(n)} : n \geq 1\}$ *converges to a bounded set, i.e.,*

$$\cup_{n=1}^{\infty} \Lambda^{(n)} \equiv \Lambda^{\infty} \subseteq [0, 1/\mu].$$

Proof. The convergence in (b) directly follows from (a) because $\Lambda^{(n)} \subseteq [0, 1/\mu]$ and is thus bounded for all $n \geq 1$. To show (a), consider any $t \in \Lambda^{(n)}$, we have $b(t+(n-1)/\mu, 0) = \lambda$, which implies that $\sigma(t + n/\mu) = b(t + (n-1)/\mu, 0) = \lambda$. If the system is overloaded at time $t + n/\mu$, then $b(t + n/\mu, 0) = \sigma(t + n/\mu) = \lambda$ by flow conservation of fluid in service; if the system is underloaded at time $t + n/\mu$, then we again have $b(t + n/\mu, 0) = \lambda$ because external arrival flows into service directly. Therefore, $b(t + n/\mu, 0) = \lambda$ implies that $t \in \Lambda^{(n+1)}$. □

We now show that convergence to the stationary point of the fluid density in service occurs *only if* the initial fluid density is that stationary point.

Theorem 5.14 $($*convergence to the unique stationary point*$)$ *The only initial fluid density in service* $b(0, \cdot)$ *for which* $b(t, x) \to b^*(x) \equiv s\mu$, $0 \leq x \leq 1/\mu$, *as* $t \to \infty$ *is the stationary point* b^* *itself.*

Proof. First the conclusion is clearly true whenever $B(t) = s$ for all $t \geq 0$, because the density $b((n/\mu), x) = b(0, x)$, $0 \leq x \leq 1/\mu$ for all $n \geq 1$. We shall show that for any $b(0, x)$ that is different from the steady state, i.e., $\max_{0 \leq x \leq 1/\mu} |b(0, x) - \mu\, s| > 0$, there exists a $0 \leq t \leq 1/\mu$ such that $b(t+n/\mu, 0) \neq \mu\, s$ for all $n \geq 0$ so that $b(t+n/\mu, 0) \nrightarrow \mu\, s$. In this case there must exist a $0 \leq t \leq 1/\mu$ such that $\mu\, s \neq b(0, t) = b(1/\mu - t, 0)$. If the system is overloaded at time $n/\mu - t$ for all $n \geq 1$, then $b(n/\mu - t, 0) = b(1/\mu - t, 0) \neq \mu\, s$ for all $n \geq 1$, by Theorem 5.4 and Corollary 5.3. If the system is underloaded at time $n'/\mu - t$ for some $n' \geq 1$, then we must have $b(n'/\mu - t, 0) = \lambda$, which implies that

$b(n/\mu - t, 0) = \lambda$ for all $n \geq n'$, following from Theorem 5.13 (because set $\Lambda^{(n)}$ is increasing). Therefore, we conclude $b(n/\mu - t, 0) \nrightarrow \mu\, s$ as $n \to \infty$. In particular, $|b(n/\mu - t, 0) - \mu\, s| \geq |b(0, t) - \mu\, s| \wedge (\lambda - \mu\, s)$. □

We now establish convergence of $b(t, \cdot)$ to a PSS for general initial conditions.

Theorem 5.15 (*PSS in service*) *Consider the $G/D/s + GI$ fluid queue with arbitrary initial condition $b(0, \cdot)$. For $0 \leq t \leq 1/\mu$, as $n \to \infty$,*

$$\begin{aligned} b(t + n/\mu, 0) &\to b^\infty(t, 0) \equiv \lambda \cdot 1_{\{t \in \Lambda^\infty\}} + b(0, 1 - t) \cdot 1_{\{t \notin \Lambda^\infty\}}, \\ b(t + n/\mu, x) &\to b^\infty(t - x, 0) \cdot 1_{\{0 \leq x \leq t\}} + b^\infty(t - x + 1/\mu, 0) \cdot 1_{\{t < x \leq 1/\mu\}}, \\ \sigma(t + n/\mu) &\to b^\infty(t, 0). \end{aligned}$$

Proof. First, it is easy to see that the third relation follows from the second(letting $x = 1/\mu$) and the second follows from the first. To establish the first relation, consider $0 \leq t \leq 1/\mu$. If the system is overloaded at $t + n/\mu, 0$ for all $n \geq 0$, then $b(t + n/\mu, 0) = b(0, 1 - t)$ for all $n \geq 0$ and thus converges to $b(0, 1 - t)$ as $n \to \infty$, following from Theorem 5.4 and Corollary 5.3. If the system is underloaded at $t + n'/\mu, 0$ for some $n' \geq 0$, then $b(t + n'/\mu, 0) = \lambda$, which implies $b(t + n/\mu, 0) = \lambda$ for all $n \geq n'$, by Theorem 5.13. □

We now show that the system is fully overloaded in each PSS, even if the PSS is only approached in the limit. For the proof, define the sets in which the system is overloaded (including critically loaded) and underloaded in a cycle of the PSS as

$$\mathcal{O}^\infty \equiv \{0 \leq t \leq 1/\mu : B(t) = s\} \quad \text{and} \quad \mathcal{U}^\infty \equiv \{0 \leq t \leq 1/\mu : B(t) < s\}.$$

Theorem 5.16 *(overloaded in each PSS) Each PSS for the $G/D/s + GI$ fluid model is overloaded everywhere, i.e., in a cycle $[0, 1/\mu]$, $\mathcal{O} = [0, 1/\mu]$ and $\mathcal{U} = \phi$.*

Proof. First, it is easy to see that $\mathcal{O}$ cannot be $\emptyset$, because $\rho > 1$. Suppose there exists a $0 \leq t \leq 1/\mu$ such that the system is underloaded at t, then there must exists a switching time $0 \leq t' \leq 1/\mu$ at which the system switches from overloaded to underloaded regime, which implies that $b(t, 0) = \lambda < \sigma(t)$. This will make $\sigma(t + 1/\mu) = b(t, 0) = \lambda \neq \sigma(t)$. Hence, this contradicts with our assumption that the system is initially in PSS. □

5.10 Proofs

In this section we present three postponed longer proofs.

Proof of Theorem 5.1 The busy cycle is a random sum of i.i.d. interarrival times, and so necessarily has a nonlattice distribution because the interarrival time cdf is nonlattice; see Proposition X.3.2 of [2]. Hence it suffices to focus on the mean busy cycle. We stochastically bound a busy cycle of the $GI/D/n + GI$ system above and below by quantities that are easier to analyze.

We start with the upper bound. For the upper bound, we use a coupling construction to produce sample-path stochastic order, as in [2, 40, 71]. We construct both systems on a common probability space so that the sample paths are ordered w.p.1 while each process separately has its own proper distribution. We give both systems the same arrival process (the same sample paths). For the upper bound, let $Y(t)$ be the number of customers in the queue of the associated system in which no servers are working. The stochastic process Y behaves as the number in system in a $GI/GI/\infty$ model with interarrival-time cdf G and service-time cdf F (our abandonment cdf). Then $n + Y$ is our candidate sample path upper bound for X. Start both X and Y with an arrival to an empty system at time 0. Continue

the sample path construction by assigning all customers that enter the queue in the original "X model" abandonment times equal to the service times assigned to the corresponding arrival in the bounding "Y model," both according to cdf F. As a consequence, whenever a customer completes service in the bounding Y model, the matching customer in the original X model customer will either have entered service or abandoned in the original X model. Hence the sample-path order is maintained. Since the abandonment times are i.i.d., this assignment rule does not alter the distribution of the processes.

The key now is to observe that the busy cycles in both the X model and the Y model (not counting the n) will end after one more interarrival time beyond the beginning of a busy cycle of the Y process if the interarrival-time and service-time pair (U, A) at the beginning of the Y busy cycle satisfies $U > 2/\mu > A$, which is an event, say C, with positive probability

$$p \equiv P(C) \equiv P(U > 2/\mu > A) = P(U > 2/\mu)P(A < 2/\mu) > 0, \qquad (5.57)$$

by the assumptions $G(x) < 1$ and $F(x) > 0$ for all x. In addition, $p < 1$ since $P(A < 2/\mu) < 1$ because we have assumed that $F(x) < 1$ for all x. For the Y model, given the event C, the one customer in the system at the start of the busy cycle will depart at time A, which is less than the time of the next arrival, U. Hence, given event C, the Y busy cycle is U. On the other hand, for the X model, at this same epoch, there are at most $n + 1$ customers in the system, with at most one in queue. Given event C, by time $1/\mu$, all customers initially in service will have completed service and departed. Again given event C, by time $2/\mu$, any initially waiting customer will have entered service and completed service if the customer did not abandon first. However, given event C, we also have $A \leq 2/\mu$, so that the customer also would have abandoned. (We only need the A part of the event C for the Y model.) Thus if event C occurs at the beginning of a busy cycle in

the Y model, then the current busy cycle ends in both models after the time U (which has been conditioned to be greater than $2/\mu$).

Thus the busy cycle T_X for the X model is bounded above by the random sum of N model-Y busy cycles, $T_{Y,i}$, until the event C first occurs at the beginning of a busy cycle, plus the single special U. For the Y models, these successive trials are i.i.d. because of the regenerative structure. The key fact we now exploit is the fact that a busy cycle T_Y of the Y process always has finite mean. For that, we can apply Corollary XII.2.5 of [2] or Theorem 2.2 of [70]. We can express the finite mean $E[T_Y]$ as

$$\begin{aligned} E[T_Y] &= pE[T_Y|C] + (1-p)E[T_Y|C^c] \\ &= pE[U|U > 2/\mu] + (1-p)E[T_Y|C^c]. \end{aligned} \tag{5.58}$$

Since, $E[U] < \infty$, necessarily $E[U|U > 2/\mu] < \infty$, so that

$$E[T_Y|C^c] \leq \frac{E[T_Y] - pE[U|U > 2/\mu]}{1-p} \leq \frac{E[T_Y]}{1-p} < \infty. \tag{5.59}$$

(Here we use the fact that $p < 1$.)

Finally, we can combine the results above to conclude that an X busy cycle T_X is stochastically bounded by a geometric random sum of i.i.d random variables, each distributed as $[T_Y|C^c]$, plus one more random variable distributed as $[U|U > 2/\mu]$. Hence, we have the bound

$$E[T_X] \leq \frac{E[T_Y|C^c]}{p} + E[U|U > 2/\mu] \leq \frac{E[T_Y]}{p(1-p)} + \frac{E[U]}{P(U > 2/\mu)} < \infty. \tag{5.60}$$

(Here we use the fact that $0 < p < 1$.)

We now consider the lower bound. We obtain a simple lower bound by observing that

the original (X) system cannot empty until at least one interarrival time exceeds the service time $1/\mu$ of that arrival. Let $N' \equiv \{n \geq 1 : U_n > 1/\mu\}$, a geometric random variable with parameter $p' \equiv P(U > 1/\mu) \equiv \bar{G}(1/\mu)$. Thus the cycle time T_X is stochastically bounded below by a sum of $N - 1$ i.i.d. interarrival times that are less than $1/\mu$ plus the last interarrival time that is greater than $1/\mu$. Hence the expected cycle time must be bounded below by

$$\begin{aligned} E[T_X] &\geq \sum_{i=1}^{N'-1} E[U|U \leq 1/\mu] + E[U|U > 1/\mu] \\ &= \frac{1-p'}{p'} E[U|U \leq 1/\mu] + 1/\mu. \end{aligned}$$

Proof of Theorem 5.2 We first establish the limit for $(\bar{B}_n, \bar{E}_n, \bar{S}_n)$ in (5.16). Since the service times are deterministic with constant value $1/\mu$, the departures (service completions) in the interval $[0, 1/\mu]$ are completely determined by the initial age distribution in service, i.e., $S(t) = B(0, 1/\mu) - B(0, 1/\mu - t)$ and $S_n(t) = B_n(0, 1/\mu) - B_n(0, 1/\mu - t)$, $n \geq 1$. By Assumption 5.3, $\bar{B}_n(0, \cdot) \Rightarrow \bar{B}(0, \cdot)$ Hence we necessarily have $\bar{S}_n \Rightarrow \bar{S}$ in $D([0, 1/\mu])$, where $\bar{S}$ is nondecreasing and continuous.

For the next step, we first do the proof in the case $B(0, 1/\mu) = 1$, i.e., $t^* = T^* = 0$; afterwards we reduce the other case to this one. By Assumption 5.1, we have $\bar{N}_n \Rightarrow \Lambda$. By condition (5.14), asymptotically, the instantaneous arrival rate is greater than or equal to the instantaneous service completion rate. Hence, the fluid entering service during $[0, 1/\mu]$ is asymptotically equivalent to the fluid completing service; i.e., we have $\|\bar{E}_n - \bar{S}_n\|_{1/\mu} \Rightarrow 0$ as $n \to \infty$, where $\|x\|_c$ denotes the uniform norm over the interval $[0, c]$. By the convergence-together theorem, Theorem 11.4.7 of [74], $\bar{E}_n \Rightarrow \bar{E}$ in $D([0, 1/\mu])$.

However, we can write $b(1/\mu, x) = b(1/\mu - x, 0)$, $0 \leq x \leq 1/\mu$, so that $B(1/\mu, x) = E(1/\mu) - E(1/\mu - x)$, $0 \leq x \leq 1/\mu$, and, similarly, $B_n(1/\mu, x) = E_n(1/\mu) - E_n(1/\mu - x)$,

$0 \leq x \leq 1/\mu$. Thus, by above, we get $B_n(1/\mu, \cdot) \Rightarrow B(1/\mu, \cdot)$ in $D([0, 1/\mu])$. We then see that the properties in Assumption 5.3 hold again at time $t = 1/\mu$. Hence we can apply mathematical indiction to conclude that $(\bar{S}_n, \bar{E}_n) \Rightarrow (\bar{S}, \bar{E})$ in $\mathbb{D}^2$ as $n \to \infty$. Since we can represent the two parameter process $\bar{B}_n$ in terms of $\bar{E}_n$, we get $\bar{B}_n \Rightarrow \bar{B}$ in $\mathbb{D}_{\mathbb{D}}$ as well. Since all limits are deterministic, all the limits are joint by Theorem 11.4.5 of [74]. That establishes (5.16) when $B(0, 1/\mu) = 1$.

We now consider the case in which $B(0, 1/\mu) < 1$. For the rest of the proof, let $V(t) \equiv B(t, 1/\mu)$ and $V_n(t) \equiv B_n(t, 1/\mu)$ with $\bar{V}_n(t) \equiv n^{-1}V_n(t)$. In this case, the limiting fluid model is underloaded until time $t^* = T^*$ in (5.33). Moreover, in this case (unlike Example 5.1) we can establish that $T_n \Rightarrow t^*$ as $n \to \infty$, exploiting condition (5.15).

We first show that, for any $\delta > 0$, $P(T_n > t^* - \delta) \to 1$ as $n \to \infty$. Since V is continuous, the definition of t^* implies that, for any $\delta > 0$, there exists $\epsilon > 0$ such that $\|V\|_{t^*-\delta} < 1 - \epsilon$. Now observe that, for all t, $\bar{V}_n(t) \leq \bar{V}_n^u(t) \equiv \bar{V}_n(0) + \bar{N}_n(t) - \bar{S}_n(t)$. However, $\|\bar{V}_n^u - V\|_t \Rightarrow 0$ for all $t > 0$, where $V(t) = V(0) + \lambda t - S(t)$ with $V(t) < 1$ for all $t < t^*$. Hence, for any $\delta > 0$ and $\epsilon > 0$, $P(\|\bar{V}_n^u - V\|_{t^*-\delta} > \epsilon) \to 0$ as $n \to \infty$. If $\|V\|_{t^*-\delta} < 1 - \epsilon$ and $\|\bar{V}_n^u - V\|_{t^*-\delta} \leq \epsilon$, then $\bar{V}_n(t) \leq \bar{V}_n^u(t) < 1$ for all t, $0 \leq t \leq t^* - \delta$, which implies that $T_n \geq t^* - \delta$. Hence, we have shown that, for any $\delta > 0$, $P(T_n > t^* - \delta) \to 1$ as $n \to \infty$.

We now show that, for any $\delta > 0$, $P(T_n > t^* + \delta) \to 0$ as $n \to \infty$. Given that we have just shown that $P(T_n > t^* - \delta) \to 1$ as $n \to \infty$, we necessarily also have $\|\bar{E}_n - \bar{N}_n\|_{t^*-\delta} \Rightarrow 0$, so that $\|\bar{V}_n - \bar{V}_n^u\|_{t^*-\delta} \Rightarrow 0$ for $\bar{V}_n^u$ defined above, so that $\|\bar{V}_n - V\|_{t^*-\delta} \Rightarrow 0$ as well for any $\delta > 0$. Moreover, since both $\bar{V}_n$ and V are bounded below by 0 and above by 1, we can obtain $\|\bar{V}_n - V\|_{t^*} \Rightarrow 0$, which implies that $\bar{V}_n(t^*) \Rightarrow \bar{V}(t^*) = 1$. as $n \to \infty$.

Since the limiting fluid model becomes overloaded at time t^*, we can apply condition (5.15) to conclude that there must exist $\delta > 0$ and $\eta > 0$ such that $\lambda\delta > S(t^*+\delta) - S(t^*) + \eta$.

Given that δ and η, define the following events:

$$\begin{aligned} C_{0,n} &\equiv \{T_n > t^* + \delta\} \\ C_{1,n} &\equiv \{\bar{V}_n(t^*) < 1 - \eta/4\} \\ C_{2,n} &\equiv \{S_n(t^* + \delta) - S_n(t^*) > \lambda\delta - \eta/2\} \\ C_{3,n} &\equiv \{N_n(t^* + \delta) - N_n(t^*) < \lambda\delta - \eta/4\}. \end{aligned} \tag{5.61}$$

Then observe that $C_{0,n} \subseteq C_{1,n} \cup C_{2,n} \cup C_{3,n}$, so that $P(C_{0,n}) \leq P(C_{1,n}) + P(C_{2,n}) + P(C_{3,n})$. However, $P(C_{i,n}) \to 0$ as $n \to \infty$ for each i, $1 \leq i \leq 3$. Hence, $P(T_n > t^* + \delta) \to 0$ as $n \to \infty$. Combining the two results, we obtain $T_n \Rightarrow t^*$ as $n \to \infty$.

We now continue to establish (5.16) in the case $V(0) \equiv B(0, 1/\mu) < 1$. The asymptotic behavior prior to time t^* is easy, because $E_n(t) = N_n(t)$ for $0 \leq t \leq T_n$, where $T_n \Rightarrow t^*$ as $n \to \infty$. Hence, we have $E_n \Rightarrow E$ in $D([0, t^*])$ as $n \to \infty$. For the rest of the proof, we shift t^* to the origin and apply the first part of the proof for the case $t^* = 0$.

It now remains to establish the limit (5.20) for $(\bar{Q}_n, \bar{A}_n)$, for which it suffices to consider the system after time t^*, when the system is full, but the queue is empty. Henceforth we assume that the system is full initially with an empty queue. For this remaining step, we can proceed under the assumption that, asymptotically, the service facility is always full with an asymptotic rate of fluid entering service and departing of

$$b((k-1)/\mu + t, 0) = \sigma(k/\mu + t) = b(k/\mu, 1/\mu - t) = b(0, 1/\mu - t), \quad 0 \leq t \leq 1/\mu.$$

Now we will focus only on the queue and regard the queue as a $G/GI/\infty$ model with service times equal to the original abandonment times and a new arrival process. Service completions in the $G/GI/\infty$ model are to be interpreted as abandonments, while the total

number of customers in the $G/GI/\infty$ system is to be interpreted as the number in queue. The arrival process for the $G/GI/\infty$ system in model n is $N_n(t) - E_n(t)$, where $E_n(t)$ is the number of customers to enter service in $[0, t]$.

Note that this representation fails to faithfully capture the original FCFS service discipline, because new arrivals go to the end of the queue, whereas customers enter service from the front of the queue. Instead, this representation applies directly to the last-come first-served (LCFS) discipline. However, that is where the exponential abandonment assumption comes in. With exponential abandonment, the number in queue $Q_n(t)$ is independent of the service discipline.

Given the $G/GI/\infty$ representation, we are able to directly apply FWLLN's established in [56]. Alternatively, we could apply [62]. Since E_n is asymptotically equivalent to the service completion process S_n, this new arrival process satisfies a FWLLN, having limit $\Lambda - S$, which in general is not a linear function. However, since $b(0, x) \leq \lambda$ for all x, $0 \leq x \leq 1/\mu$, we also have $\sigma(t) \leq \lambda$ for all $t \geq 0$, so it has a nonnegative rate. Hence we can prove (5.20) with (5.21) and (5.23) by applying Theorems 3.1 and 7.1 of [56]. To do so, we exploit the fact that the limit of the arrival process there is allowed to be nonlinear.

Finally, we complete the proof by showing (5.22) holds. We first exploit (5.21), which implies that

$$\begin{aligned} Q(t) &= \int_0^t e^{-\theta\,(t-s)}(\lambda - b(s,0))ds \\ &= \frac{\lambda}{\theta}(1 - e^{-\theta\,t}) - e^{-\theta\,t}\int_0^t b(s,0)\,e^{\theta\,s}ds. \end{aligned} \tag{5.62}$$

On the other hand, the ODE (2.32) implies that

$$w'(t) = 1 - \frac{b(t,0)}{\lambda\,e^{-\theta\,w(t)}}, \quad w(0) = 0,$$

which has a unique solution

$$w(t) = t - \frac{1}{\theta} \log \left(\frac{\theta}{\lambda} \int_0^t b(s,0) e^{\theta s} ds + 1 \right). \tag{5.63}$$

Combining (5.22)and (5.63), we obtain (5.62).

Proof of Theorem 5.5 First consider the interval $[0, 1/\mu]$. The departure rate is $\sigma(t) = b(t, 1/\mu) = b(0, 1/\mu - t)$ for $0 \le t \le 1/\mu$. Since the staffing function is constant s, it is necessary to have $\lambda > \sigma(t)$ ($\lambda < \sigma(t)$) if the system switches from underloaded (overloaded) to overloaded (underloaded) at t. Consider an underloaded interval $[a, b] \subset [0, 1/\mu]$ where a and b are switching points, we must have $\zeta(a) > 0 > \zeta(b)$, which implies that ζ changes its sign in (a, b) at least once from positive to negative. The sign changing can be achieved in two cases: (i) crossing level 0 continuously from above to below, or (ii) jumping from above 0 to below. Therefore, ζ has at least a zero in case (i) and a discontinuity in case (ii) in interval (a, b). Similar reasoning works for an overloaded interval. This reasoning applies to all overloaded and underloaded subintervals that begin and end in the interior $(0, 1/\mu)$ of the interval $[0, 1/\mu]$. In addition, there are the two intervals with the interval endpoints. Thus the number of switches exceeds the number of internal intervals by at most 1. Let $\mathcal{S}_{[0,1/\mu]}$ be the total number of switching points in $[0, 1/\mu]$. We have just shown that we must have $|\mathcal{S}_{[0,1/\mu]}| \le |\mathcal{D}_\zeta| + |\mathcal{Z}_\zeta| + 1$.

We are done if $T = 1/\mu$; hence assume that $T > 1/\mu$. We continue for $\lceil T\mu \rceil$ cycles of length $1/\mu$. Next we consider the next interval $[1/\mu, 2/\mu]$. We will show that the number of switching points can be no greater than in the first interval of length $1/\mu$ just considered. Recall that the departure rate is $\sigma(t) = b(t, 1/\mu) = b(t - 1/\mu, 0)$. Let $\zeta_2(t) \equiv \sigma(t + 1/\mu) - \lambda = b(t, 0) - \lambda$ for $0 \le t \le 1/\mu$. Therefore, $|\mathcal{S}_{[1/\mu,2/\mu]}|$, the number of switching points

in $[1/\mu, 2/\mu]$, is totally determined by the number of zeros and discontinuities of ζ_2, by the same argument as above.

We now show that $|\mathcal{Z}_{\zeta_2}| \leq |\mathcal{Z}_\zeta|$. To do so, we first observe that we have $b(t,0) = \sigma(t)$ when the system is overloaded. Hence the functions $\zeta(t)$ and $\zeta_2(t)$ differ only when the system is underloaded during $[0, 1/\mu]$. Consider an underloaded interval $[a,b] \subset [0,1/\mu]$ where a and b are switching points, which implies that $\sigma(a) > \lambda > \sigma(b)$ ($\zeta(a) > 0 > \zeta(b)$). Since the system is underloaded in $[a,b]$, we must have $b(t,0) = \lambda$. In case (i), ζ changes its sign in (a,b) with (at least) an zero at some $y \in \mathcal{Z}_\zeta \cap (a,b)$. However, ζ_2 has no such zeros in $\mathcal{Z}_{\zeta_2} \cap (a,b)$ because $\zeta_2(y) = 0$ for $a < y < b$ (which yields that $\mathcal{Z}_{\zeta_2} \cap (a,b) = \phi$), we have $|\mathcal{Z}_{\zeta_2} \cap (a,b)| = 0 \leq |\mathcal{Z}_\zeta \cap (a,b)|$, which implies that $|\mathcal{Z}_{\zeta_2}| \leq |\mathcal{Z}_\zeta|$ counting all underloaded intervals in $[0,1/\mu]$ that are in case (i).

In case (ii), ζ changes its sign in (a,b) with (at least) a jump from positive to negative. However ζ_2 has at most two discontinuity points (at a and b) in (a,b) (because $\zeta_2(y) = 0$ for $a < y < b$). Although the number of discontinuities of ζ_2 in $[a,b]$ may outnumber the discontinuities of ζ by at most 1, these two jumps ($\zeta_2(a-) > \lambda$ to $\zeta_2(a) = \lambda$ and $\zeta_2(b-) = \lambda$ to $\zeta_2(b) < \lambda$) can at most contribute to one sign change in (a,b). In other words, ζ_2 may have more discontinuities than ζ, but those extra ones are redundant. Hence, $|\mathcal{S}_{[1/\mu,2/\mu]}| \leq |\mathcal{D}_\zeta| + |\mathcal{Z}_{\zeta_2}| \leq |\mathcal{D}_\zeta| + |\mathcal{Z}_\zeta|$. The desired bound in (5.30) is obtained by induction on interval $[n/\mu, (n+1)/\mu]$, continuing until $N \equiv \lceil T\mu \rceil$.

5.11 Conclusions

We considered the heavily loaded many-server queue with customer abandonment and deterministic service times, i.e., the stochastic $GI/D/n + GI$ model. Even though the arrival rate exceeds the maximum possible service rate, the customer abandonment keeps the system stable. In §5.2 we showed that the busy cycles in the stochastic $GI/D/n + GI$

queueing model constitute regeneration times, so that stochastic processes describing the performance, such as the number of customers in the system, converge to proper steady state distributions as time evolves for any proper initial condition.

In §5.3 we showed that a sequence of $G/D/n + GI$ queueing systems with $\rho \equiv \lambda/\mu > 1$ indexed by n satisfies a many-server heavy-traffic limit in the efficiency-driven (ED) regime, converging to a deterministic fluid model, provided that the arrival processes and initial conditions obey functional weak laws of large numbers. In general, Theorem 5.2 only establishes a limit for the performance measures describing the service facility, e.g., $B_n(t, y)$, but those fluid limits capture the essential periodic character. A many-server heavy-traffic limit for the queue-length and abandonment processes was also obtained under the assumption of exponential abandonment.

Like the stochastic system, we found that the limiting fluid model has a unique stationary point. However, unlike the stochastic model, Theorem 5.14 shows that the fluid model never converges to that stationary point unless it starts in that stationary point. Instead, the fluid model tends to exhibit periodic behavior. Moreover, the specific form of the periodic behavior depends critically on the initial conditions. As a consequence, the asymptotic loss of memory (ALOM) property established for the $G_t/M_t/s_t + GI_t$ model in Chapter 4 does not nearly hold with deterministic service times.

Moreover, as illustrated in §5.1, simulations of the stochastic system show that the time-dependent behavior of the stochastic system is well described by the fluid model for large n. Indeed, the fluid model tends to provide a better description of the performance in the queueing model than the steady-state distribution of the queueing model, amplifying [73].

The rest of the chapter was devoted to a careful study of the limiting fluid model. We obtained quite complete results for the case in which there exists a finite time T^* after which the system remains overloaded. Theorem 5.6 provides general conditions for this to be true. That condition is in terms of the initial density of fluid in service $b(0, \cdot)$, but

can also be applied at later times after applying the algorithm in Remark 5.2 over some initial interval. However, §5.9 shows that, in general, such a finite time need not exist. Nevertheless, Theorem 5.15 shows that the fluid density in service b converges to a PSS,

In summary, the fluid content in service evolves in three different ways, depending on the initial conditions:

1. The fluid in service is in steady state for all $t \geq 0$ if it is initialized with $b(0, x) = \mu\, s$ for $0 \leq x \leq 1/\mu$.

2. The system first becomes overloaded at $t^* < 1/\mu$ and remains overloaded after time T^*, $t^* \leq T^* < \infty$, in which case $b(t, \cdot)$ is in a PSS determined by $b(T^*, \cdot)$.

3. The system first becomes overloaded at $t^* < 1/\mu$, but switches between overloaded and underloaded infinitely often. Then the fluid density b converges to an overloaded PSS.

In cases (ii) and (iii), if instead we initialize by redefining $b(0, \cdot)$, letting it have the PSS version, then the system is initially overloaded and the fluid density in service is periodic with period $1/\mu$ for all $t \geq 0$. The remaining queue performance then converges to a PSS as well. In case (i), the associated queue performance converges to the unique stationary point as well. In cases (ii) and (iii), if we start with the PSS for b, then the queue performance converges to a PSS as well. In case (iii) it remains to determine if the queue performance converges to the PSS associated with the limiting PSS for b when we use the given initial conditions; we conjecture that it does.

It is natural to wonder what happens with other service-time distributions. In Appendix D.9 we show that the same periodic behavior is exhibited by the corresponding model with a two-point service-time distribution, provided that one of the points is at the origin (in the same spirit as the corresponding special hyperexponential distribution in in [76]).

However, in Appendix D.10 we present results from simulation experiments showing that the periodic phenomenon ceases to hold for other two-point distributions and, more generally, if the service-time is only nearly deterministic. When the service-time distribution is nearly deterministic, the performance is similar to the performance with D service and the same initial conditions over suitably short time intervals, but convergence to stationary performance is evident as t increases.

We concentrated on the stationary $G/D/n + GI$ fluid model, but some of the results can be extended. First, as in Chapters 2-4, we can analyze, and obtain an algorithm for, the $G_t/D/s_t + GI$ fluid model in which the arrival rate and the number of servers are allowed to be time varying. In particular, §5.4, §5.5 and §5.7 extend to this case. In general, we lose the periodic structure, on which most of this chapter focuses, but that periodic structure is retained as well if the arrival rate function λ and the staffing function s are also periodic with the same period $1/\mu$. (However, the periodic structure is less surprising in that case.) Moreover, the structural properties of the queue established in §5.7 also extend to GI service, provided that the fluid density in service b is given. Of course, determining b is more complicated for GI service that is neither D nor M. Theorem A.2 of Chapter 2 shows that it is necessary to solve a complicated fixed point equation in order to determine b in those cases.

As stated in §5.1, we began this study in an effort to understand if ALOM holds for the $G/GI/s+GI$ and $G_t/GI_t/s_t+GI_t$ fluid models when the service-time distribution is neither M_t nor M. That question remains after we stipulate that the service distribution also is neither D nor the two-point distribution with one mass at 0. We conjecture that ALOM does hold for the fluid model under that extra condition and the regularity conditions imposed in Chapter 4.

Bibliography

[1] Aksin, Z., Armony, M., Mehrotra, V. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Mgmt.* **16** 665-688 (2007)

[2] Asmussen, S. *Applied Probability and Queues*, second edition, Springer, New York (2003)

[3] Bassamboo, A., Ranhawa, R. S. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Res.* **58** 1398-1413 (2010)

[4] Billingsley, P. *Convergence of Probability Measures.*, second ed., Wiley, New York (1999).

[5] Borovkov, A. A. Some limit theorems in the theory of mass service, II. *Theor. Probability Appl.* **10** 375-400 (1965)

[6] Borovkov, A. A. On limit laws for service processes in multi-channel systems (in Russian). *Siberian Math J.* **8** 746-763 (1967)

[7] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc* **100** 36-50 (2005)

[8] Buzacott, J. A., Shanthikumar, J. G. *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ (1992)

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

[10] Chen, H., Yao, D. D. *Fundamentals of Queueing Networks*, Springer, New York (2001)

[11] Choudhury, G. L., Mandelbaum, A., Reiman, M. I., Whitt, W. Fluid and diffusion limits for queues in slowly changing environments. *Stochastic Models* **13** 121-146 (1997)

[12] Courtois, P. J. *Decomposability*, Academic Press, New York (1977)

[13] Davis, J. L., Massey, W. A., Whitt, W. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Science* **41** 1107-1116 (1995)

[14] Eick, S. G., Massey W. A., Whitt, W. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731-742 (1993)

[15] Eick, S. G., Massey, W. A., Whitt, W. $M_t/G/\infty$ queues with sinusoidal arrival rates. Management Sci. **39(2)** 241-252 (1993)

[16] Erramilli, A., Forys, L. J. Oscillations and chaos in a flow model of a switching system. *IEEE J. Sel. Areas Commun.* **9** 171-178 (1991)

[17] Feldman, Z., Mandelbaum, A., Massey, W. A., Whitt, W. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54** 324-338 (2008)

[18] Feller, W. *An Introduction to Probability Theory and its Applications*, second edition, Wiley, New York (1971)

[19] Gamarnik, D., Zeevi, A. Validity of heavy-traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Prob.* **16** 56-90 (2006)

[20] Garnett, O., Mandelbaum, A., Reiman, M. I. Designing a call center with impatient customers. Manufacturing Service Oper. Management. **4** 208-227 (2002)

[21] Gibbens, R. J., Hunt, P. J., Kelly, F. P. Bistability in communication networks. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammersely*, G. Grimmett and D. Welsh (eds.), Oxford University Press, 113-127 (1990)

[22] Glynn, P. W., Whitt, W. A new view of the heavy-traffic limit for infinite-server queues. *Adv. Appl. Prob.* **23** 188-209 (1991)

[23] Goodman, J. B., Massey, W. A. The non-ergodic Jackson network. *Journal of Applied Probability* **21** 860-869 (1984)

[24] Granovsky, B. L. Zeifman, A. Nonstationary queues: estimating the rate of convergence. *Queueing Systems.* **46** 363-388 (2004)

[25] Green, L. V., Kolesar, P. J., Soares, J. Improving the SIPP approach for staffing service systems that have cyclic demand. *Operations Research* **49(4)** 549-564 (2001)

[26] Green, L. V., Kolesar, P. J., Whitt, W. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** 13-39 (2007)

[27] Gurvich, I. Validity of heavy-traffic steady-state approximations in multiclass queueing networks: sufficient conditions involving state-space collapse. Working paper, Northwestern University (2009)

[28] Halfin, S., Whitt, W. Heavy-traffic limits for queues with many exponential servers. *Operations Res.* **29** 567-588 (1981)

[29] Hall, R. W. *Queueing Methods for Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ (1991)

[30] Heyman, D., Whitt, W. The Asymptotic Behavior of Queues with Time-Varying Arrival Rates. *Journal of Applied Probability.* **21** 143-156 (1984)

[31] Ibrahim, R. E., Whitt, W. Wait-time predictors for customer service systems with time-varyng demand and capacity. *Oper. Res.* Forthcoming (2011) http://www.columbia.edu/∼ww2040/allpapers.html

[32] Iglehart, D. L., Whitt, W. Multiple channel queues in heavy traffic, II: sequences, networks, and batches. *Adv. Appl. Probab.* **2** 355-369 (1970)

[33] Isaacson, D., Madsen, R. *Markov Chains: Theory and Applications.* Wiley, New York (1976)

[34] Jelenkovic, P., Mandelbaum, A., Momcilovic, P. Heavy-traffic limits for queues with many deterministic servers. *Queueing Systems* **47** 53-69 (2004)

[35] Jennings, O. B., Mandelbaum, A., Massey, W. A., Whitt, W. Server staffing to meet time-varying demand. Management Sci. **42** 1383-1394 (1996)

[36] Kang, W., Ramanan, K. Fluid limits of many-server queues with reneging. *Annals of Applied Probability* **20(6)** 2204-2260 (2010)

[37] Kaspi, H., Ramanan, K. Law of large numbers limits for many-server queues. *Annals of Applied Probability* **21(1)** 33-114 (2011)

[38] Kingman, J. F. C. On queues in heavy traffic. *J. of the Royal Statistical Society* **24** 383-392 (1962)

[39] Krichagina, E. V., Puhalskii, A. A. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* **25** 235-280 (1997)

[40] Lindvall, T. *Lectures on the Coupling method*, Wiley, New York (1992)

[41] Liu, Y., Whitt, W. The $G_t/GI/s_t+GI$ many-server fluid queue. Columbia University, NY, NY (2011)
(http://www.columbia.edu/∼ww2040/allpapers.html)

[42] Liu, Y., Whitt, W. A network of time-varying many-server fluid queues with customer abandonment. Oper. Res. Forthcoming (2011)

[43] Liu, Y., Whitt, W. Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment, *Queueing Systems* **67(2)** 145-182 (2011)

[44] Liu, Y., Whitt, W. Nearly periodic behavior in the overloaded $G/D/s + GI$ queue. *Stochastic Systems* **1** (2011)

[45] Liu, Y., Whitt, W. A many-server heavy-traffic limit for the $G_t/GI/s_t+GI$ queueing model, in preparation.

[46] Mandelbaum, A., Massey, W. A., Reiman, M. I. Strong approximations for Markovian service networks. Queueing Systems. **30** 149-201 (1998)

[47] Mandelbaum, A., Massey, W. A., Reiman, M. I., Rider, B. Time varying multiserver queues with abandonments and retrials. *Proceedings of the 16th International Teletraffic Congress*, P. Key and D. Smith (des.) (1999)

[48] Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. *Proceedings of the Thirty-Seventh Annual Allerton Conference on Communication, Control and Computing*, Allerton, IL, 1095-1104 (1999)

[49] Mandelbaum, A., Zeltyn, S. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+G$ queue. *OR Spectrum* **26** 377-411 (2004)

[50] Massey, W. A., Whitt, W. Networks of infinite-server queues with nonstationary Poisson input. Queueing Systems **13** 183-250 (1993)

[51] Miller, D. Existence of limits in regenerative stochastic processes. *Ann. Math. Statist.* **43** 1273-1280 (1972)

[52] Nelson, B., Taaffe, M. The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II–the multiclass network. *INFORMS Journal on Computing* **16** 275-283 (2004)

[53] Nelson, B., Taaffe, M. The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II–the multiclass network. *INFORMS Journal on Computing* **16** 275-283 (2004)

[54] Newell, G. F. Applications of Queueing Theory. second ed., Chapman and Hall, London (1982)

[55] Pang, G., Talreja, R., Whitt, W. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 193-267 (2007)

[56] Pang, G., Whitt, W. Two-parameter heavy-traffic limits for infinite-server queues. Queueing Systems **65** 325-364 (2010)

[57] Prabhu, N. U. *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Coomunication*, second ed., Springer, NY (1998)

[58] Puhalskii, A. A. The $M_t/M_t/k_t+M_t$ queue in heavy traffic. Mathematics Departure. University of Colorado at Denver (2008)

[59] Puhalskii, A. A., Reed, J. On many-server queues in heavy traffic. *Ann. of Appl. Probab.* **20(1)** 129-195 (2010)

[60] Puhalskii, A. A., Reiman, M. I. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. in Appl. Probab.* **32(2)** 564-595 (2000)

[61] Reed, J. The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Prob.* **19** 2211-2269 (2009)

[62] Reed, J., Talreja, R. Distribution-valued heavy-traffic limits for the $G/GI/\infty$ queue. New York University, New York, NY (2009)

[63] Sigman, K., Whitt, W. Heavy-traffic limits for nearly deterministic queues. *J. Appl. Prob.*, forthcoming (2011)
http://www.columbia.edu/∼ww2040/allpapers.html

[64] Sigman, K., Whitt, W. Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Queueing Systems*. forthcoming (2011)
http://www.columbia.edu/∼ww2040/allpapers.html

[65] Stolyar, A.L., Yudovina, E. Systems with large flexible server pools: Instability of "natural" load balancing. *Bell Labs Technical Memo* (2010)
http://arxiv.org/abs/1012.4140

[66] Tak'acs, L. On a probability problem arising in the theory of counters. *Proc. Camb. Phil. Soc.* **52** 488-498 (1956)

[67] Tak'acs, L. *Introduction to the Theory of Queues*, Oxford University Press, New York (1962)

[68] Talreja, R., Whitt, W. Heavy-traffic limits for waiting times in many-server queues with abandonments. *Ann. Appl. Prob.* **19** 2137-2175 (2009)

[69] Teschl, G. *Ordinary Differential Equations and Dynamical Systems*, Lecture Notes, University of Vienna, Austria (2000)
Available at: http://www.mat.univie.ac.at/∼gerald/ftp/book-ode/

[70] Whitt, W. Embedded renewal processes in the $GI/G/s$ queue. *J. Appl. Prob.* **9** 650-658 (1972)

[71] Whitt, W. Comparing counting processes and queues. *Adv. Appl. Prob.* **13** 207-22 (1981)

[72] Whitt, W. Untold horrors of the waiting room. What the equilibrium distribution will never tell about the queue-length process. *Management Science* **29** 395-408 (1983)

[73] Whitt, W. The queueing network analyzer. *Bell System Technical Journal* **62** 2779-2815 (1983) (http://www.columbia.edu/∼ww2040/allpapers.html)

[74] Whitt, W. *Stochastic-Process Limits*, Springer, New York (2002)

[75] Whitt, W. Efficiency-Driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50** 1449-1461 (2004)

[76] Whitt, W. Engineering solution of a basic call-center model. *Management Sci.* **51** 221-235 (2005)

[77] Whitt, W. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54** 37-54 (2006)

[78] Willie, H. Periodic steady state of loss systems. *Adv. Appl. Prob.* **30** 152-166 (1998)

[79] Yom-Tov, G., Mandelbaum, A. The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. working paper, the Technion, Israel, (2010)

[80] Zeltyn S. Call centers with impatient customers: exact analysis and many-server asymptotics of the $M/M/N + G$ queue. PhD thesis (2004) Available at: http://iew3.technion.ac.il/serveng/References/MMNG.thesis.pdf

[81] Zeltyn, S., Mandelbaum, A. Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems* **51** 361-402 (2005)

Appendix A

Appendix for Chapter 2

A.1 Overview.

This appendix contains material supplementing Chapter 2. We start with results for the fluid model and conclude with simulation experiments.

First, §A.2 explains why the service content density $b(t, x)$ satisfies the transport PDE in an underloaded (UL) interval, as noted in Remark 2.2. In §A.3 we supplement §2.6 by presenting alternative algorithms for the service content density b during an overloaded (OL) interval. This leads to a another PDE for $b(t, x)$ under extra smoothness assumptions.

In §A.4 we present additional results for the BWT w and the PWT v during an OL interval, thus supplementing §2.7. We begin by providing a more elementary proof of Theorem 2.3 for the ODE for the BWT w under additional smoothness regularity conditions. Then we prove Corollary 2.3, which provides explicit formulas for the BWT in special cases. We also state an analog of Corollary 2.3 for the PWT v. We also prove Theorem 2.4, which

established conditions for the PWT v to be finite. In §A.5 we discuss the structure of the BWT function w. Theorem 2.3 requires the positivity $\lambda_{inf} > 0$ in Assumption 2.10. We now consider cases in which $\lambda(t) = 0$ for some $t \geq 0$. We show that the BWT w can have more complicated structure when the the zero set has zero Lebesque measure or positive Lebesque measure.

In §A.6 we say more about the flows, i.e., the service-completion-rate function, σ, and the abandonment-rate function, α, defined in (2.7) and (2.9). In §A.7 we supplement §2.8, which summarizes the algorithm, by providing more discussion of the algorithm. In particular, we specify the algorithm to adjust for an initially infeasible staffing function s and illustrate its performance. In §A.8 we present additional material related to §2.10 on choosing staffing functions to stabilize delays. In particular, we show how to stabilize delays with general initial conditions. (In §2.10 we assumed that the system starts empty.)

Finally, in §C.6 we supplement §2.2 in Chapter 2 by presenting additional comparisons of the fluid model to simulations of large-scale queueing systems. These additional simulations confirm the observations in §2.2: First, for very large queueing systems, with thousands of servers, the individual sample paths of the scaled queueing processes have negligible stochastic fluctuations and agree closely with the computed fluid model performance functions. Second, for smaller queueing systems, e.g., with about 20 servers, the fluid model performance functions still provide remarkably accurate approximations for the mean values of the queueing processes.

A.2 The Transport PDE for b in a UL Interval

In Remark 2.2 we observed that the service content density b satisfies a version of the generic scalar transport equation in the underloaded case. We provide more details here.

The same reasoning applies to the queue content density $\tilde{q}(t,x)$, during an overloaded interval, ignoring flow into service; see §2.7.1.

Proposition A.1 (transport pde) *In the underloaded region, if $b(0,\cdot)$ is differentiable in x, then the service content function b is differentiable for $t \neq x$ and satisfies the the following pde, a simple version of the generic scalar transport equation:*

$$b_t(t,x) + b_x(t,x) \equiv \frac{\partial b}{\partial t}(t,x) + \frac{\partial b}{\partial x}(t,x) = -h_G(x)b(t,x). \tag{A.1}$$

Proof. Since λ and $b(0,\cdot)$ are both differentiable, then it is easy to see that $b(t,x)$ is differentiable for $t \neq x$. If we let $p(u) \equiv b(t+u, x+u)$, we have that

$$\begin{aligned} b_t(t,x) + b_x(t,x) = p'(0) &= \lim_{u\to 0}\left(\frac{p(u)-p(0)}{u}\right) \\ &= \lim_{u\to 0}\left(\frac{b(t+u,x+u)-b(t,x)}{u}\right) \\ &= \lim_{u\to 0}\left(\frac{\bar{G}(x+u)-\bar{G}(x)}{u}\right)\left(\frac{b(t,x)}{\bar{G}(x)}\right) \\ &= -\frac{g(x)}{\bar{G}(x)}b(t,x) = -h_G(x)b(t,x), \end{aligned}$$

where we apply the chain rule of calculus and the fundamental evolution equation for b in (2.5).

Solving pde (A.1) with initial conditions $\lambda(t)$ and $b(0,x)$, yields Proposition 2.2. To verify that, recall that the general solution to pde (A.1) is $b(t,x) = e^{-\int_0^x h_G(u)du}\phi(t-x) = \bar{G}(x)l(t-x)$, where function ϕ is any differentiable function. Here we have $\phi(t) = \lambda(t)1_{\{t\geq 0\}}$. By the initial condition, $b(0,x) = \phi(-x)\bar{G}(x)$ when $x \geq 0$. Therefore we see that the claim is valid.

A.3 Alternative Algorithms for b in an OL Interval

We now discuss alternative algorithms to calculate the service content density b in an overloaded interval.

If Assumption 2.8 holds, then a finite function b is uniquely characterized via equation (2.16), where

$$b(t,x) = \hat{b}(t,x)/h_G(x), \quad 0 \le x \le t < T, \tag{A.2}$$

with $\hat{b}$ being the unique solution of the equation

$$\hat{b}(t,x) \equiv \hat{a}(t,x) + g(x)\int_0^{t-x} \hat{b}(t-x,y)\,dy, \quad 0 \le x \le t < T, \tag{A.3}$$

where

$$\hat{a}(t,x) \equiv g(x)s'(t-x) + g(x)\int_0^{\infty} \frac{b(0,y)g(y+t-x)}{\bar{G}(y)}\,dy \in \mathcal{F}_T. \tag{A.4}$$

We can establish the existence of a unique solution to equation (A.3) by applying the Banach fixed point theorem on an appropriate space of functions of *two* variables.

Although this new fixed-point equation is more complicated, it can lead to a PDE characterization of b. This PDE representation follows directly by differentiating in the equation (A.3). (Convenient cancelation occurs.)

Theorem A.1 $(PDE\ for\ \hat{b})$ *Under the assumptions of Theorems* A.2 *and* A.3*, wherever* $\hat{b}$ *has first partial derivatives with respect to* t *and* x*, it satisfies the PDE*

$$\hat{b}_t(t,x) + \hat{b}_x(t,x) = \hat{y}(t,x) + \hat{z}(x)\hat{b}(t,x), \quad 0 \le x \le t \le T, \tag{A.5}$$

where

$$\hat{y}(t,x) \equiv \hat{a}_t(t,x) + \hat{a}_x(t,x) - \frac{g'(x)}{g(x)}\hat{a}(t,x) \quad \textit{and} \quad \hat{z}(x) \equiv \frac{g'(x)}{g(x)} \tag{A.6}$$

for $\hat{a}(t,x)$ *in* (A.11). (*The functions* $\hat{y}$ *and* $\hat{z}$ *in* (A.6) *are well defined by the assumptions in Theorem* A.3.) *Associated with the PDE is the boundary condition*

$$\hat{b}(t,t) = \hat{a}(t,t) = g(t)s'(0) + g(t)\int_0^\infty b(0,y)h_G(y)\,dy, \quad 0 \le t \le T, \tag{A.7}$$

which is finite by (2.25).

We now continue with the two-parameter functions $b \equiv b(t,x)$. To apply the Banach fixed point theorem in this setting, we use the space $\mathcal{F}_{T,1}$ of measurable real-valued functions of the pair of real variables (t,x) over the "triangular" domain $0 \le x \le t \le T$, for which the norm

$$\|u\|_{T,1} \equiv \sup_{0\le t\le T}\int_0^t |u(t,x)|\,dx. \tag{A.8}$$

is finite. The norm $\|\cdot\|_{T,1}$ is an L_1 norm in one coordinate and an L_∞ norm in the other; it makes $\mathcal{F}_{T,1}$ a Banach space.

Theorem A.2 (service content in the overloaded case) *Consider an overloaded interval* $[0,T)$. *If Assumption* 2.8 *holds, then a finite function* b *is uniquely characterized via equation* (2.16)*, where*

$$b(t,x) = \hat{b}(t,x)/h_G(x), \quad 0 \le x \le t < T, \tag{A.9}$$

with $\hat{b}$ being the unique fixed point of the operator $\mathcal{T} : \mathcal{F}_{T,1} \to \mathcal{F}_{T,1}$ defined by

$$\mathcal{T}(u)(t,x) \equiv \hat{a}(t,x) + g(x) \int_0^{t-x} u(t-x,y)\, dy, \quad 0 \le x \le t < T, \tag{A.10}$$

where

$$\hat{a}(t,x) \equiv g(x)s'(t-x) + g(x) \int_0^{\infty} \frac{b(0,y)g(y+t-x)}{\bar{G}(y)}\, dy \in \mathcal{F}_T. \tag{A.11}$$

Moreover, the operator $\mathcal{T}$ is a monotone contraction operator on $\mathcal{F}_{T,1}$ with contraction modulus $G(T)$ for the norm $\|\cdot\|_{T,1}$ defined in (A.8)*, so that, for any $u \in \mathcal{F}_{T,1}$, the fixed point can be approximated by the n-fold iteration $\mathcal{T}^{(n)}$ of the operator $\mathcal{T}$ applied to u, with*

$$\|\mathcal{T}^{(n)}(u) - \hat{b}\|_{T,1} \le \frac{G(T)^n}{1-G(T)} \|\mathcal{T}(u) - u\|_{T,1} \to 0 \quad \textit{as} \quad n \to \infty \tag{A.12}$$

and, if $u \le (\ge)\mathcal{T}(u)$, then $\mathcal{T}^{(n-1)}(u) \le (\ge)\mathcal{T}^{(n)}(u) \le (\ge)\hat{b}$ for all $n \ge 1$. Finally, $\hat{b}(t,t) = \hat{a}(t,t) = g(t)b(0,0)$.

Proof. First, we show that $\hat{b}$ in (A.9) is a fixed point of the operator $\mathcal{T}$, i.e., that $\mathcal{T}(\hat{b}) = \hat{b}$. To see that, multiply (2.16) through by $h_G(x)$, noting that (i) $h_G(x)\bar{G}(x) = g(x)$ and (ii) we are interested in the case $x \le t$. We get $\hat{b}(t,x) = b(t,x)h_G(x) = b(t-x,0)g(x)$. Next

we successively apply (2.18), (2.5) and a change of variables to get

$$\begin{aligned}
\hat{b}(t,x) &= b(t-x,0)g(x) = s'(t-x)g(x) + g(x)\int_0^\infty b(t-x,y)h_G(y)\,dy \\
&= s'(t-x)g(x) + g(x)\int_{t-x}^\infty b(t-x,y)h_G(y)\,dy + g(x)\int_0^{t-x} b(t-x,y)h_G(y)\,dy \\
&= s'(t-x)g(x) + g(x)\int_{t-x}^\infty b(0,y-(t-x))\frac{\bar{G}(y)}{\bar{G}(y-(t-x))}h_G(y)\,dy \\
&\qquad +g(x)\int_0^{t-x}\hat{b}(t-x,y)\,dy \\
&= s'(t-x)g(x) + g(x)\int_0^\infty b(0,y)\frac{g(y+t-x)}{\bar{G}(y)}\,dy + g(x)\int_0^{t-x}\hat{b}(t-x,y)\,dy \\
&= \hat{a}(t,x) + g(x)\int_0^{t-x}\hat{b}(t-x,y)\,dy = \mathcal{T}(\hat{b})(t,x), \qquad \text{(A.13)}
\end{aligned}$$

where $\hat{a}(t,x) = \hat{c}(t,x) + \hat{d}(t,x)$ with

$$\hat{c}(t,x) \equiv g(x)s'(t-x) \quad \text{and} \quad \hat{d}(t,x) \equiv g(x)\int_0^\infty b(0,y)\frac{g(y+t-x)}{\bar{G}(y)}\,dy.$$

We next show that $\|\hat{a}\|_{T,1} < \infty$. First, $\|\hat{c}\|_{T,1} \le G(T)\|s'\|_T < \infty$ because $s' \in \mathbb{C}_p \subset \mathbb{D}$. Because of the factor $g(x)$, $\|\hat{d}\|_{T,1}$ is bounded by the integral term. Taking the supremum over x and t with $0 \le x \le t \le T$ of the integral in the expression for $\hat{d}$ yields the term τ in Assumption 2.8, which we have assumed is bounded. Hence $\|\hat{d}\|_{T,1} < \infty$, and so $\|\hat{a}\|_{T,1} < \infty$.

Next note that $\mathcal{T}$ is indeed a contraction operator on $(\mathcal{F}_{T,1}, \|\cdot\|_{T,1})$, because

$$\|\mathcal{T}(u_1) - \mathcal{T}(u_2)\|_{T,1} \le \sup_{0\le t\le T}\int_0^t g(x)\left(\int_0^{t-x}|u_1-u_2|(t-x,y)\,dy\right)dx \le G(T)\|u_1-u_2\|_{T,1},$$

and we have assumed that $G(T) < 1$ for all T. The geometric rate of convergence in (A.12) is the standard conclusion from the Banach fixed point theorem, and the subsequent

ordering follows from the monotonicity of $\mathcal{T}$. Finally, $\hat{b}(t,t) = \hat{a}(t,t)$ because the subset of u in $\mathcal{F}_{T,1}$ for which $u(t,t) = \hat{a}(t)$ is closed, and $\mathcal{T}$ maps that subset into itself, because $\mathcal{T}(u)(t,t) = \hat{a}(t,t)$, $0 \le t \le T$, for all u in $\mathcal{F}_{T,1}$. By (2.18), $\hat{a}(t,t) = g(t)b(0,0)$. ∎

We now provide conditions for $\hat{b}(\cdot,x)$ and $b(\cdot,x)$ to be in $\mathbb{C}_p$ for all $x \ge 0$. (We use these properties for $b(\cdot,0)$ to establish properties of the ODE to calculate the BWT w in §2.3 of Chapter 2.) We first introduce extra smoothness conditions.

Assumption A.1 (*extra smoothness for g and s*) *g and s' are differentiable with derivatives g' and s'' in $\mathbb{C}_p$.*

We next impose additional regularity conditions on the service-time pdf g. For that purpose, let $\|g\|_\infty$ be the uniform norm, i.e., $\|g\|_\infty \equiv \sup_{x \ge 0} \{|g(x)|\}$.

Assumption A.2 (*extra regularity for g*) *The service-time pdf g satisfies: $g(x) > 0$ for all x, $\|g\|_\infty < \infty$ and there exists K such that $g(x) \le g(0)e^{Kx}$ for all $x \ge 0$.*

We will use the last inequality in Assumption A.2 in its equivalent form: $|g'(x)| \le Kg(x)$ for all x. (To see the equivalence, Divide by $g(x)$, integrate and take the exponential.)

Theorem A.3 (smoothness of service content in the overloaded case) *If Assumptions* 2.8–A.2 *all hold, then $\hat{b}(\cdot,x)$ and $b(\cdot,x)$ are differentiable functions for each $x \ge 0$, almost everywhere equal to their partial derivatives with respect to t, for b in* (A.9) *and $\hat{b}$ in* (A.10). *Hence, $\hat{b}(\cdot,x), b(\cdot,x) \in \mathbb{C}_p$ for all $x \ge 0$.*

Proof. We again apply the Banach fixed point theorem, but now on a subspace of $\mathcal{F}_{T,1}$ with a new norm. Consider the subspace of measurable real-valued functions u of the pair of real variables (t,x) over the same triangular domain $0 \le x \le t \le T$ that are

differentiable with respect to the variable t, and equal almost everywhere to the integral of its partial derivative u_t, with finite norm $\|u\|_{T,2}$, where

$$\|u\|_{T,2} \equiv \sup_{0\le t\le T} \{ \int_0^t (|u(t,x)| + |u_t(t,x)|)\, dx, \} \tag{A.14}$$

which is like the Sobolev norm on the Sobolev space $\mathcal{W}^{1,\infty}(0,t)$. The functions in $\mathcal{F}_{T,2}$ are Lipschitz continuous in the first variable t for each x in $0 \le x \le t \le T$. Reasoning as in the proof of Theorem A.2, we will show that $\|\hat{a}\|_{T,2} < \infty$, and then we will show that $\mathcal{T}$ maps $\mathcal{F}_{T,2}$ into itself.

Then,

$$\|\hat{a}\|_{T,2} \le \|\hat{a}\|_{T,1} + G(T) \left(\|s''\|_T + K \sup_{0\le s\le T} \int_0^\infty (b(0,y)g(s+y)/\bar{G}(y))\, dy \right) < \infty$$

by the proof of Theorem A.2 and the conditions in Assumptions 2.8, A.1 and A.2. (Since $\mathbb{C}_p \subset \mathbb{D}$, $\|s''\|_T < \infty$.) Next, $\|\mathcal{T}(u)\|_{T,2} \le \|\hat{a}\|_{T,2} + G(T)(\|u\|_{T,1} + \sup_{0\le t\le T} \{|u(t,t)\} + \|u_t\|_{T,1}) < \infty$. Then we see that $\mathcal{T}$ is again a contraction operator on $(\mathcal{F}_{T,2}, \|\cdot\|_{T,2})$ with modulus $G(T)$. We can ignore the term involving $|u_1(t,t) - u_2(t,t)|$, because, as noted at the end of Theorem A.2, we can restrict attention to the closed subspace $\mathcal{F}_{T,2}$ containing only u for which $u(t,t) = g(t)b(0,0)$; as a consequence, $u_1(t,t) = u_2(t,t)$ for all t. Hence, the fixed point $\hat{b}$ is an element of $\mathcal{F}_{T,2}$, and so has the claimed smoothness properties. ∎

A.4 More on the Performance in Overloaded Intervals

We now present additional material on the queue performance functions during an OL interval.

A.4.1 More on the BWT w

Alternate Proof of Theorem 2.3: the ODE for the BWT w. If we assume additional smoothness, then we can obtain a simple direct proof of Theorem 2.3. In particular, we can obtain the expression for the ODE describing the evolution of the BWT $w(t)$ by differentiating in the basic flow conservation equation in (2.6). Consider an overloaded interval that starts out with the queue empty, so that $Q(0) = 0$. Then, when we differentiate with respect to t in (2.6), we get

$$\frac{d}{dt}Q(t) \equiv Q'(t) = \lambda(t) - \alpha(t) - b(t, 0), \tag{A.15}$$

where, from (2.7) and Corollary 2.2, by a change of variable,

$$\alpha(t) \equiv \int_0^\infty q(t, x)h_F(x)\,dx = \int_0^{w(t)} \lambda(t-x)f(x)\,dx = \int_{t-w(t)}^{t} \lambda(x)f(t-x)\,dx. \tag{A.16}$$

and,

$$Q(t) = \int_0^{w(t)} \lambda(t-x)\bar{F}(x)\,dx = \int_{t-w(t)}^{t} \lambda(x)\bar{F}(t-x)\,dx. \tag{A.17}$$

Then, assuming that w is differentiable (as well as $\bar{F}$), we can differentiate under the integral in (A.17) to get

$$Q'(t) = \lambda(t) - \tilde{q}(t, w(t))(1 - w'(t)) + \int_{t-w(t)}^{t} \lambda(x)f(t-x)\,dx. \tag{A.18}$$

We remark that the standard conditions to justify differentiation under the integral, i.e., differentiation of

$$I(t) \equiv \int_{a(t)}^{b(t)} h(t, x)\,dx \tag{A.19}$$

is to have (i) the partial derivative of $h(t,x)$ with respect to t be well defined, (ii) $h(t,x)$ and $\partial h(t,x)/\partial t$ both be continuous in the two variables t and x in some region including $\{(t,x) : a(t) \leq x \leq b(t), t_1 \leq t \leq t_2\}$, and (iii) a and b to have continuous derivatives in the region $\{t : t_1 \leq t \leq t_2\}$. Under these conditions,

$$I'(t) = h(t,b(t))b'(t) - h(t,a(t))a'(t) + \int_{a(t)}^{b(t)} \frac{\partial h(t,x)}{\partial t}\, dx. \tag{A.20}$$

Equation (A.18) is an application of (A.20) to (A.17).

Inserting (A.18) into (A.15) and making appropriate cancelations ($\lambda(t)$ and $\alpha(t)$ appear on both sides), we get

$$b(t,0) = \tilde{q}(t,w(t))(1-w'(t)), \tag{A.21}$$

which yields

$$w'(t) = 1 - \frac{b(t,0)}{\tilde{q}(t,w(t))}. \tag{A.22}$$

The more complicated analysis in our main proof is needed because we do not have all the smoothness conditions. ■

Proof of Corollary 2.3: explicit expressions for the BWT w. Since the proofs to (a) and (b) are similar, we will only prove (b). ODE (2.31) implies that

$$b(t,0)e^{\theta t} = \lambda(t-w(t))e^{\theta(t-w(t))}(1-w'(t)) \;=\; \frac{d}{dt}\left(\int_0^{t-w(t)} \lambda(y)e^{\theta y}dy\right),$$

which implies

$$\tilde{\Lambda}(t-w(t)) \;=\; \int_0^t b(y,0)e^{\theta y}dy,$$

and inverting function $\tilde{\Lambda}(\cdot)$ yields (2.33). Moreover,

$$\tilde{t} \equiv \inf\{t > 0 : w(0) = 0\} = \inf\{t > 0 : \tilde{\Lambda}(t) = \int_{t_1}^{t} b(y,0)e^{\theta y}dy\}. \quad \blacksquare$$

A.4.2 More on the PWT v

We now give closed-form formulae for the PWT v in some special cases, paralleling those for the BWT w in Corollary 2.3. We omit the proof, which is similar to the proof of Corollary 2.3, which is given in the next subsection.

Corollary A.1 *Suppose $v(0) = 0$, the system is overloaded for $0 < t < \delta$, $b(t,0) > 0$.*

(a). If there is no abandonment, i.e., if the model is $G_t/M/s_t$, then

$$v(t) = \Gamma^{-1}(\int_0^t \lambda(y)dy) - t,$$

for $0 \le t < \bar{t}$, where $\Gamma(t) \equiv \int_0^t b(y,0)dy$, $\Gamma^{-1}(x) \equiv \inf\{y > 0 : \Gamma(y) = x\}$, and $\bar{t} \equiv \inf\{t > 0 : \Gamma(t) = \int_0^t \lambda(y)dy\}$.

(b). If the abandonment-time distribution is exponential ($\bar{F}(x) = e^{-\theta x}$ for $x \ge 0$), i.e., if the model is $G_t/M/s_t + M$, then

$$v(t) = \tilde{\Gamma}^{-1}(\int_0^t \lambda(y)e^{\theta y}dy) - t,$$

for $0 \le t < \tilde{t}$, where $\tilde{\Gamma}(t) \equiv \int_0^t b(y,0)e^{\theta y}dy$, $\tilde{\Gamma}^{-1}(x) \equiv \inf\{y > 0 : \tilde{\Gamma}(y) = x\}$, and $\tilde{t} \equiv \inf\{t > 0 : \tilde{\Gamma}(t) = \int_{t_1}^t \lambda(y)e^{\theta y}dy\}$.

Proof of Theorem 2.4: finiteness of PWT v. **Proof.** Recalling the definition of $\sigma(t)$ in (2.9), and using Assumption 2.12, we obtain

$$\sigma(t) = \int_0^\infty b(t,x)h_G(x)\,dx \geq \int_0^\infty b(t,x)h_{G,L}\,dx = B(t)h_{G,L}.$$

However, in the overloaded interval, $B(t) = s(t)$ and $s(t) \geq s_{lbd}$ by Assumption 2.11. Hence we have the claimed lower bound on $\sigma(t)$. We use that lower bound to bound $E(t+u) - E(t)$ below. Note that

$$E(t+u) - E(t) = \int_t^{t+u} b(v,0)\,dv = \int_t^{t+u} (s'(v) + \sigma(v))\,dv \geq s(t+u) - s(t) + s_L h_{G,L} u.$$

By Assumption 2.11, $s(t+u) \geq s_L$. Starting from the definition (2.35), we apply the inequalities above to obtain

$$\begin{aligned} v(t) &\equiv \inf\{u \geq 0 : E(t+u) - E(t) + A_t(u) \geq Q(t)\} \\ &\leq \inf\{u \geq 0 : E(t+u) - E(t) \geq Q(t)\} \\ &\leq \inf\{u \geq 0 : (s_L h_{G,L} u - s(t) + s_L)^+ \geq Q(t)\} \leq \frac{Q(t) + s(t) - s_L}{s_L h_{G,L}} < \infty, \end{aligned}$$

where $Q(t) \leq Q(0) + \Lambda(t) < \infty$ for all t. □

A.5 Structure of the Boundary Waiting Time w.

Theorem 2.3 requires the positivity $\lambda_{inf} > 0$ in Assumption 2.10. We now consider cases in which $\lambda(t) = 0$ for some $t \geq 0$. That leads to more complicated behavior for the BWT function w.

A.5.1 The Zero Set of $\lambda(\cdot)$ Has Zero Lebesgue Measure.

First, suppose that $\lambda(t_0) = 0$ for some $t_0 > 0$ but the zero set of $\lambda(\cdot)$ has zero Lebesgue measure, i.e., $\int_0^T 1_{\{}\lambda(t) = 0\}dt = 0$, see Figure A.1(a). Again we assume that both $b(t, 0)$ and $\lambda(t)$ are continuous for $0 \le t \le T$.

We only consider the overloaded case (the underloaded case is not interesting since $w(t) = 0$). For simplicity, suppose the system is initially critically loaded, i.e., $B(0) = S(0)$, $w(0) = 0$, $Q(0) = 0$, and $\lambda(0) > \sigma(0)$, then the system becomes overloaded in the next moment.

We give a vivid example. Let the system be initially critically loaded and suppose $b(t, 0) = 1$ as long as the system is overloaded. For instance, this can be achieved if $S(t) = 1$ and the service-time distribution is exponential with rate 1. Let the arrival-rate function $\lambda(t) = t^2 - 3t + 9/4$ and the abandon-time distribution be exponential with rate 0.5, i.e., $\bar{F}(x) = 0.5 \cdot e^{-0.5x}$ for $x \ge 0$.

We can see from Figure A.1(a) that $\lambda(3/2) = 0$ and $\int_0^T 1_{\{\lambda(t)=0\}}dt = 0$ for all $T > 0$. Because $\lambda(0) = 9/4 > b(0) = 1$ the system becomes overloaded after time 0. We plot in Figure A.1(b) the boundary waiting time $w(t), 0 \le t \le T$ with $T = 3$. One can see that the derivative of $w(t)$ reaches $-\infty$ once, and this corresponds to the fact that $\lambda(t)$ touches 0 once but does not stay at 0.

A.5.2 The Zero Set of $\lambda(\cdot)$ Has Positive Lebesgue Measure.

In a more general setup of the arrival process, $\lambda(t)$ can stay at 0 for a while meaning that the arrival process is turned off. For instance, it is natural that the arrival process may look like the first picture in Figure A.2.

Intuition tells us in this case $w(t)$ cannot be continuous for all $t \ge 0$, it will jump at some times. But when will $w(t)$ jump? What will be the heights of the jumps? To answer

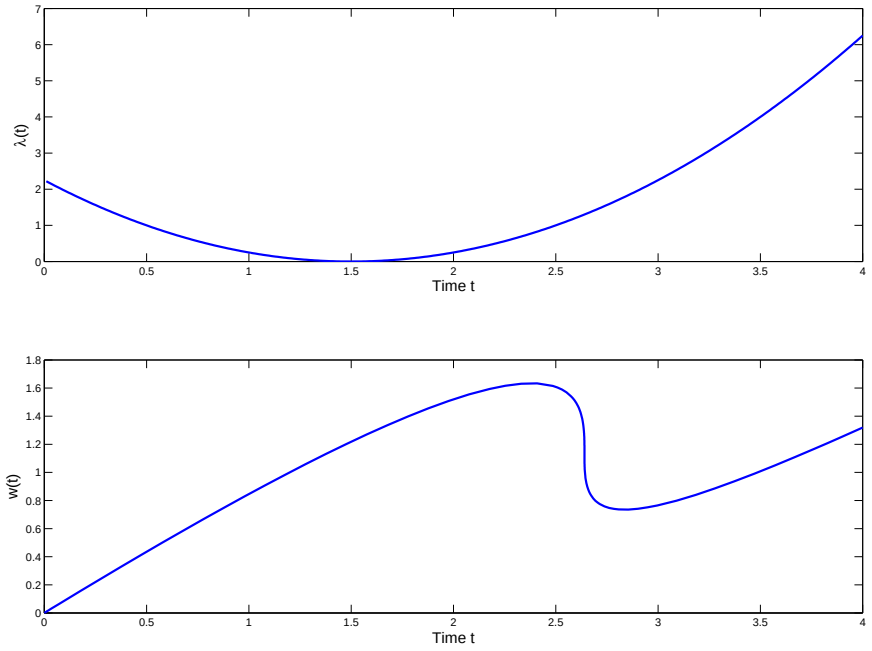

Figure A.1: An example of boundary waiting time $w(t)$ with $\lambda(t) = 0$ once.

these questions, we simply assume that $\lambda(t) = 0$ for $0 < \hat{t}_1 \leq t < \hat{t}_2 < \infty$. The case that $\lambda(t) = 0$ for t in finite disjoint intervals can be easily generalized. Note that $\lambda(t)$ being left-continuous or right-continous does not matter because it is just a rate function.

Again, we consider a vivid example. Suppose the system is initially overloaded with $w(0) = 2$ and $q(0, x) = e^{0.5x} 1_{\{0 \leq x \leq w(0)\}}$. We choose $\lambda(t)$ large enough such that the system stays overloaded for $t \geq 0$ and fix $b(t, 0) = 0.5$. Let $\lambda(t) = (9t - 3t^2) \cdot 1_{\{0 \leq t < 3\}} + 3 \cdot 1_{\{t \geq 3.5\}}$. In other words, $\lambda(t)$ is quadratic for $t \in [0, 3)$, stays at 0 for $t \in [3, 3.5)$, and is constant 3 for $t \geq 3.5$, see Figure A.2(a). Let the abandon-time distribution be exponential with rate 0.5.

In Figure A.2(b), the red line is $q(t, x)$ at $t = 0$, which is a function of x. The blue line on the negative half-line is the arrival-rate function $\lambda(t)$ reflected with respect to the y axis. Imagine that with the origin fixed, the blue line moves to the right at rate 1, because

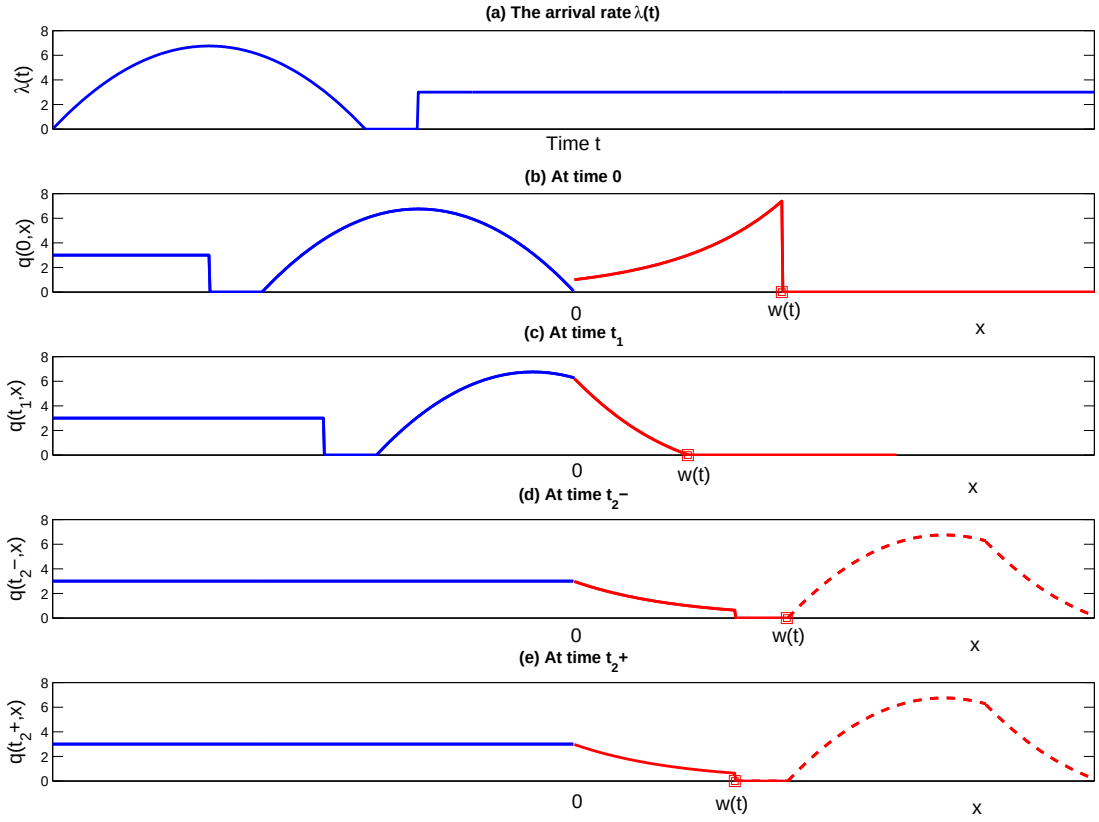

Figure A.2: The dynamics of $q(t,x)$ of an example with $\lambda(t) = 0$ for $0 < t_1 \leq t < t_2 < \infty$.

new fluid keeps arriving to the system after time 0. The right boundary of the red line is the boundary waiting time $w(t)$ at each t, which is being controlled by the ratio between $b(t,0) = 1$ and $q(t,w(t))$. So one can see that the right boundary of the red line is moving at rate $1 - b(t,0)/q(t,w(t))$ since fluid at the front of the queue is being transported into service (eaten away) by $b(t,0)$.

As time evolves, for the part of the reflected arrival-rate function that exceeds the origin (that is pushed onto the positive half-line), the height decreases with time because of abandonment. In Figure A.2(c), all fluid that was in queue at time 0 is just gone at time t_1, and $w(t_1) = t_1$ because the blue line travelled by t_1 to the right. At time t_1,

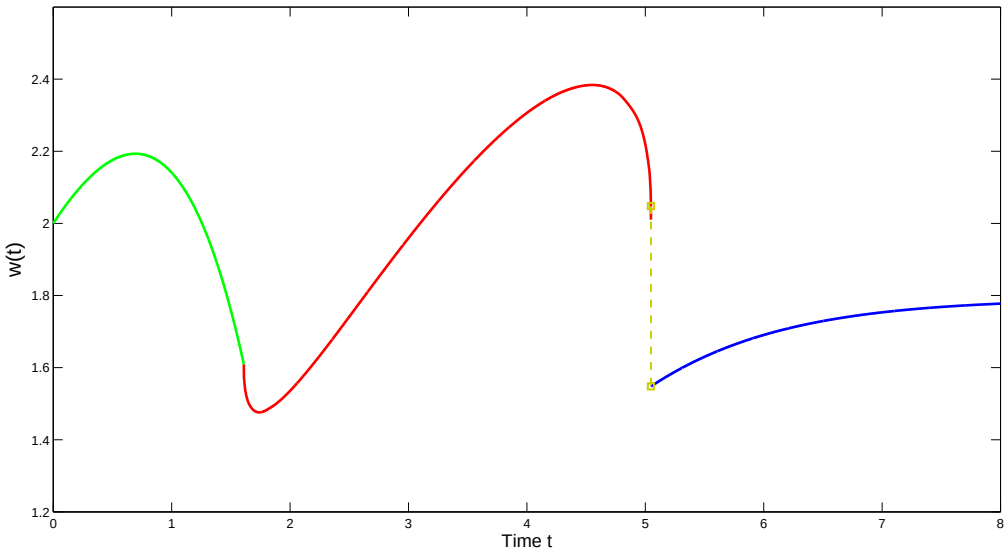

Figure A.3: An example of the boundary waiting time $w(t)$ with $\lambda(t) = 0$ for $0 < t_1 \leq t < t_2 < \infty$.

$q(t_1, x) = \lambda(t_1 - x) \cdot e^{-0.5} \cdot 1_{\{0 \leq x \leq t_1\}}$ which is the red line, and $q(t_1, w(t_1)) = q(t_1, t_1) = 0$ implies that $w'(t_1) = -\infty$, see Figure A.3. Although $w'(t)$ has a discoutinuity at t_1, $w(t)$ itself is continuous at t_1.

At time t_2- which is the moment right before the quadratic part of $\lambda(t)$ is eaten away, the boundary waiting time $w(t_2-) = t_2 - 3$, where 3 is the length of the quadratic part of $\lambda(t)$. Then at time t_2+, $w(t)$ jumps from $w(t_2-) = t_2 - 3$ to $w(t_2+) = t_2 - 3.5$, because there is an interval of length 0.5 in which $\lambda(t) = 0$, see Figure A.3. At t_2 the left derivative $w(t_2-) = \infty$ because $q(t_2-, w(t_2-)) = 0$.

This example shows that discontinuities of λ yield discontinuities of w', and λ staying at 0 over in interval yields discontinuities of w.

A.6 More on the Flows

We next discuss the departure function S in (2.9) and the abandonment function A in (2.7). These flows are performance measures of interest in their own right, but they are also important because they enable us to extend the model treated here directly to open networks of fluid queues, in which the departing fluid or abandoning fluid from one queue become input to another queue; see Chapter 2.

A.6.1 Main Results

We show that the flows S and A inherit the structure of the original input Λ, so that the results in Chapter 2 extend to open networks of fluid queues. The following results are elementary. The proofs and other properties are given in the following subsection.

Theorem A.4 (the departure rate) *Assume that the conditions in Theorem* A.3 *hold. For* $t \geq 0$,

$$\sigma(t) = \int_0^t b(t-x,0)g(x)\,dx + \int_0^\infty \frac{b(0,y)g(t+y)}{\bar{G}(y)}\,dy,$$

where $b(t,0) = \lambda(t-u)$ *in an underloaded interval, but is the solution to the fixed point equation in Theorem A.2 during an overloaded interval. As a consequence,* $\sigma \in \mathbb{C}_p$.

Theorem A.5 (abandonment rate) *Assume that the conditions in Theorem 2.3 of* Chapter 2 *hold, so that the BWT w is well defined and continuous. For $t \geq 0$,*

$$\begin{aligned}\alpha(t) &= \left(\int_0^{w(t)} \lambda(t-x)f(x)\,dx\right) 1_{\{w(t)\leq t\}} \\ &\quad + \left(\int_0^{t} \lambda(t-x)f(x)\,dx + \int_0^{w(t)-t} \frac{q(0,y)f(t+y)}{\bar{F}(y)}\,dy\right) 1_{\{w(t)>t\}}.\end{aligned}$$

As a consequence, $\alpha \in \mathbb{C}_p$.

A.6.2 Elaboration on the Flows

We now elaborate on the discussion about the flows in the previous subsection; i.e., we discuss the departure process S in (2.9) and the abandonment process A in (2.7). Make the same assumptions as above including the conditions in Theorem A.3 and Assumption 2.12.

Theorem A.6 (departure rate)

1. *For $t \geq 0$,*

$$\sigma(t) = \int_0^\infty b(t,x)h_G(x)\,dx = \int_0^t b(t-x,0)g(x)\,dx + \int_0^\infty \frac{b(0,y)g(t+y)}{\bar{G}(y)}\,dy, \tag{A.23}$$

   where $b(t,0) = \lambda(t-u)$ in an underloaded interval, but is the solution to the fixed point equation in Theorem A.2 during an overloaded interval.

2. *$\sigma \in \mathbb{C}_p$, as assumed for λ in Assumption* 2.2.

3. *$\sigma(t) \geq B(t)h_{G,L} > 0$ for all $t \geq 0$, so that σ satisfies the requirement for λ in Assumption 2.10 over the interval $[\epsilon, t]$ for each $\epsilon > 0$.*

4. *If there exists a constant* $h_{G,U}$ *such that* $h_G(x) \leq h_{G,U} < \infty$ *for all* $x \geq 0$*, then* $\sigma(t) \leq B(t)h_{G,U} \leq s(t)h_{G,U}$ *for all* $t \geq 0$*.*

5. *If* $b(t,0)$ *is absolutely continuous with derivative* $b'(u,0)$ *in* $\mathbb{C}_p$ *on the interval* $[0,t]$ *(as occurs in the case of exponential service) and if*

$$\tau_2(b,g,t) \equiv \sup_{0\leq s\leq t} \int_0^\infty \frac{b(0,y)|g'(s+y)|}{\bar{G}(y)}\, dy < \infty, \tag{A.24}$$

then σ *is absolutely continuous with derivative (a.e.)*

$$\sigma'(t) = b(0,0)g(t) + \int_0^t b'(u,0)g(x)\, dx + \int_0^\infty \frac{b(0,y)g'(s+y)}{\bar{G}(y)}\, dy. \tag{A.25}$$

Proof. We prove the properties in turn:

(i) (representation (A.23)) Apply (2.9) and Assumption 2.6.

(ii)($\sigma \in \mathbb{C}_p$) By the finiteness of the initial conditions, Assumption 2.8 and the continuity of $b(\cdot,0)$ from Theorem A.3, $\sigma(t) < \infty$. By Theorem A.3, $b(\cdot,0)$ is in $\mathbb{C}_p$. By the Lebesgue dominated convergence theorem, the continuity of $b(t,0)$ and $g(t+y)$ in the integrands of (A.23) is inherited by σ, so $\sigma \in \mathbb{C}_p$, as claimed.

(iii) (lower bound) By the initial relation in (A.23), we have $\sigma(t) \geq B(t)h_{G,L}$. Since $s(u) \geq s_L > 0$ for $0 \leq u \leq t$, $\lambda(t) \geq \lambda_{inf}(t) > 0$ and $\bar{G}(x) > 0$ for all x, we have $B(t) \geq t\lambda_{inf}(t)\bar{G}(t) \wedge s_L$ for all $t \geq 0$, which implies that there exist constants $\epsilon > 0$ and $\sigma_{\{inf,\eta,\epsilon\}}$ such that $\sigma(u) > \sigma_{\{inf,\eta,\epsilon\}} > 0$ for $0 < \epsilon \leq u \leq t$.

(iv) (upper bound) By the initial relation in (A.23), we have $\sigma(t) \leq B(t)h_{G,U}$, but we always have $B(t) \leq s(t)$.

(v)(derivative) We differentiate under the integral in (A.23) using Leibniz integral for-

mula for differentiation under the integral, for which we require the finiteness of τ_2 in (A.24). ■

The abandonment rate is somewhat more difficult. First, the abandonment is only positive during the overloaded intervals, so we assume that we are focusing on a single overloaded interval. Second, the abandonment depends on q, which in turn depends on w, which also is more complicated, requiring more conditions.

Theorem A.7 (abandonment rate) *Assume that the conditions in Theorem 2.3 hold, so that the BWT w is well defined and continuous.*

1. *For $t \geq 0$,*

$$\begin{aligned} \alpha(t) &= \left(\int_0^{w(t)} \lambda(t-x) f(x)\, dx \right) 1_{\{w(t) \leq t\}} \\ &\quad + \left(\int_0^{t} \lambda(t-x) f(x)\, dx + \int_0^{w(t)-t} \frac{q(0,y) f(t+y)}{\bar{F}(y)}\, dy \right) 1_{\{w(t) > t\}} \end{aligned} \tag{A.26}$$

2. *$\alpha \in \mathbb{C}_p$, as assumed for λ in Assumption* 2.2.

3. *If Assumption* 2.12 *holds, then $\alpha(t) \geq Q(t) h_{G,L}$ for all $t \geq 0$.*

4. *If there exists a constant $h_{G,U}$ such that $h_G(x) \leq h_{G,U} < \infty$ for all $x \geq 0$, then $\sigma(t) \leq Q(t) h_{G,U}$, which is bounded over finite intervals, because Q is continuous.*

5. *If $b(t,0) > 0$ a.e., then α is absolutely continuous with derivative (a.e.)*

$$\begin{aligned}\alpha'(t) &= \left(\lambda(t-w(t))f(w(t))w'(t) + \int_0^{w(t)} \lambda'(t-x)f(x)\,dx\right) 1_{\{w(t)\le t\}} \\ &\quad + \left(\lambda(0)f(w(t)) + \int_0^t \lambda'(t-x)f(x)\,dx + \left(\frac{q(0,w(t)-t)f(w(t))}{\bar{F}(w(t)-t)}\right)(w'(t)-1)\right) \\ &\quad + \int_0^{w(t)-t} \frac{q(0,y)f'(s+y)}{\bar{F}(y)}\,dy\Bigg) 1_{\{w(t)>t\}}. \qquad \text{(A.27)}\end{aligned}$$

Proof. We prove the properties in turn:

(i) (representation) Applying definition (2.7) and Assumption 2.6, we have

$$\alpha(t) = \int_0^\infty q(t,x)h_F(x)\,dx = \int_0^t q(t-x,0)f(x)\,dx + \int_0^\infty \frac{q(0,y)f(t+y)}{\bar{F}(y)}\,dy, \quad \text{(A.28)}$$

from which (A.26) follows.

(ii) ($\alpha \in \mathbb{C}_p$) Note that $\lambda, q(0,\cdot) \in \mathbb{C}_p$ by Assumption 2.2, $q(\cdot,0) \in \mathbb{C}_p$ by Theorem 2.3 and Corollary 2.2 and w is continuous by Theorem 2.3. Hence, by the Lebesgue dominated convergence theorem, the continuity of $\lambda(t,0)$ and $f(t+y)$ as a function of t in the integrands of (A.23) is inherited by σ, so $\sigma \in \mathbb{C}_p$, as claimed.

(iii) (lower bound) By the initial relation in (A.26), we have $\alpha(t) \ge Q(t)h_{F,L}$.

(iv) (upper bound) By the initial relation in (A.26), we have $\alpha(t) \le Q(t)h_{F,U}$.

(v) (derivative) We differentiate under the integral in (A.23) using Leibniz integral formula for differentiation under the integral. Since the integrands are bounded over the finite intervals, the integrals are finite. ∎

A.7 A Fluid Algorithm with Infeasible s.

Our main algorithm in §2.8 for the $G_t/GI/s_t + GI$ fluid model assumes that the staffing function s is feasible. That algorithm is designed to stop whenever the given staffing function s is detected to be infeasible. Now we want to apply the results in §2.9 to find the minimum feasible staffing function.

We illustrate how to do so for the $G_t/M/s_t + GI$ model; §2.9 shows how to do the same for more general GI service. In the context of the $G_t/M/s_t + GI$ model, a sufficient condition for feasibility over $[0, T]$ is

$$s(t) + s'(t) \geq 0, \quad 0 \leq t \leq T. \tag{A.29}$$

Here we want to generalize our algorithm. Suppose the target staffing function s is not feasible for all t. Instead of stopping the algorithm, we want (i) to produce a 'best' modified capacity function $s_f(t)$ and (ii) to finish the algorithm with our new target $s_f(t)$.

We only need to modify our initial algorithm when the system is in the overloaded regime. Flow conservation of the service facility says that $b(t, 0) = B'(t) + \mu\, B(t)$ which is equal to $s'(t) + s(t)$ if $s(t)$ were feasible. However, if we want to make $B(t)$ decrease as fast as possible, the best we can do is to set $b(t, 0) = 0$ and let fluid deplete with only its service completion. Therefore, when s becomes infeasible at t_1, i.e., $s'(t_1+) + s(t_1+)$ becomes negative, $B(t)$ will satisfy ODE $B'(t) = -\,B(t)$ for $t \in [t_1, t_1 + \delta]$ with $B(t_1) = s(t_1)$, which implies that $B(t) = s(t_1)e^{-(t-t_1)}$.

We let $t_2 \equiv \inf\{t_1 < t \leq T : s(t) = B(t)\} \wedge T = \inf\{t_1 < t \leq T : s(t) = s(t_1)e^{-(t-t_1)}\} \wedge T$. Note that $b(t, 0) = 0$ for $t_1 \leq t \leq t_2$ guarantees that the queue does not empty out before t_2 so that the system does not switch from overloaded to underloaded regime before t_2. This is so because with $b(t, 0) = 0$, abandonment becomes the only

source that deplete the queue, and the abandonment rate $\alpha(t)$ goes to 0 as $Q(t)$ goes to 0. For instance, if the abandonment distribution is exponential with rate θ, then $\alpha(t) = \theta\, Q(t)$.

If $t_2 = T$, the system stays overloaded until T and we are done. Otherwise, we let $t_3 \equiv \inf\{t_2 < t \le T : s'(t) + S(t) < 0\} \wedge T$, $b(t,0) = s'(t) + \mu\, s(t)$ for $t_2 \le t \le t_3$. Just as in the original algorithm, we solve ODE (2.31) with $w(t_2) = 0$ for $t_2 \le t \le t_3$. If $t_U \equiv \{t > t_2 : w(t) = 0\} < t_3$, then the system switches from overloaded to underloaded regime and we continue with the old algorithm in Chapter 2; otherwise, s becomes infeasible once again at t_3 while the system is overloaded, and we shall repeat the above argument, and as before, we run the algorithm dynamically until we proceed to time T.

It is not hard to see that under the above construction, we successfully obtain the interval I_{inf} in which s is infeasible and a modified service-capacity function $s_f(t) = B(t)\, 1_{t\in I_{inf}} + s(t)\, 1_{t\in[0,T]/I_{inf}}$. Also, $s_f(t)$ is the closest feasible function to the given target $s(t)$.

Example of the Algorithm. To evaluate the performance of the modified algorithm, we use the example in §A.9.2, i.e., we consider the Markovian $M/M/s_t + M$ model that has a Poisson arrival process with a constant rate λ, exponential service and abandonment distributions with rates μ and θ respectively, and a sinusoidal capacity function

$$s(t) \equiv \lambda + \bar{\lambda} \cdot \sin(c \cdot t). \tag{A.30}$$

We still let $\lambda = 1$, $c = 1$, $\mu = 1$, $\theta = 0.5$. To make s infeasible, we let $\bar{\lambda} = 0.9\lambda = 0.9$ instead of $0.6\lambda = 0.6$ in §A.9.2. Now s has greater fluctuation and it is easy to see that condition (A.29) is no longer satisfied.

We plot the performance measures of the fluid model in Figure A.4. Compared with Figure A.12, we see that $I_{inf} \equiv [3.27, 5.05] \cup [9.55, 11.33] \cup [15.84, 17.62]$ is the interval in which s becomes infeasible. For $t \in I_{inf}$, $s_f(t)$ (the blue dashed curve) is different from

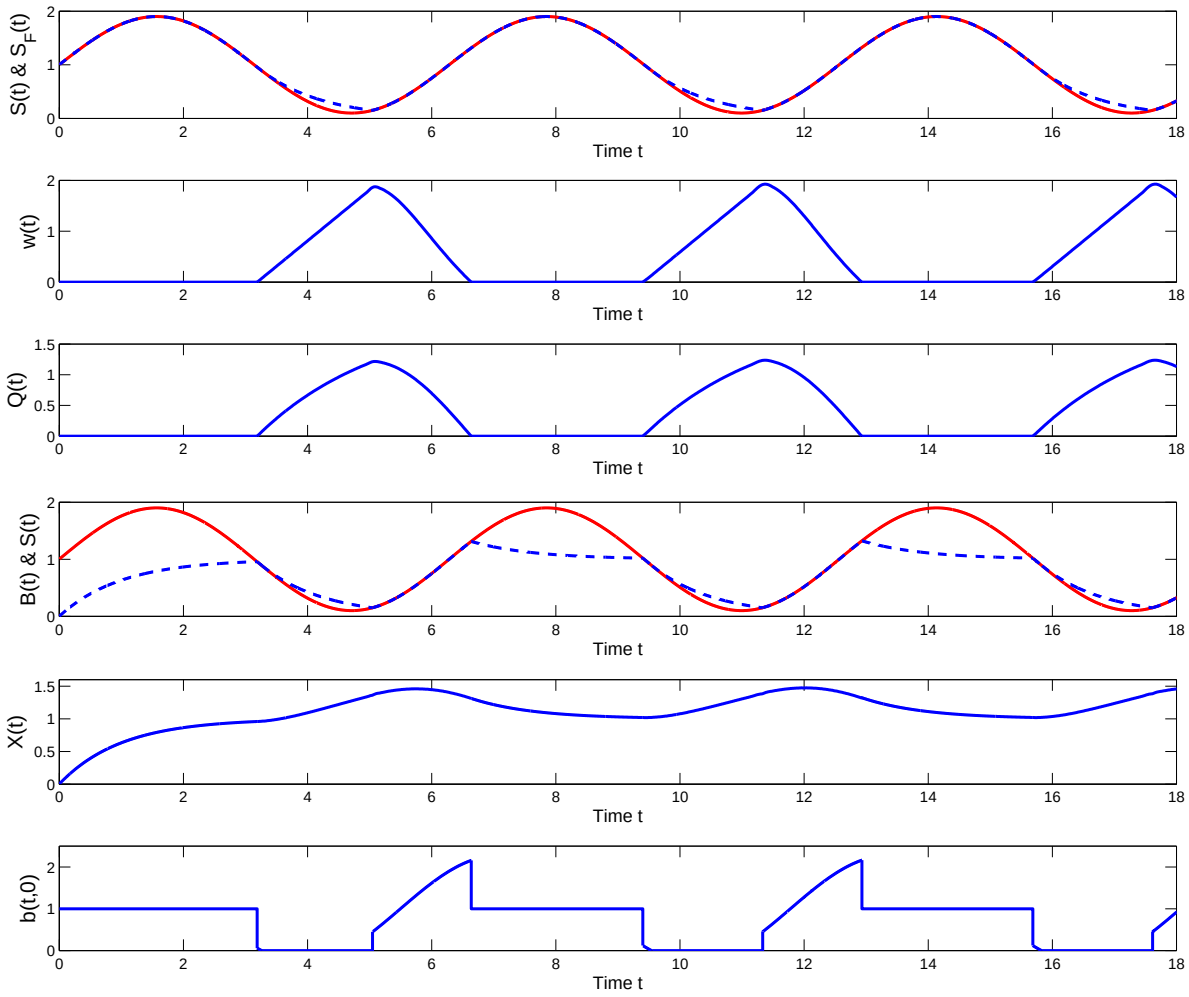

Figure A.4: The $M/M/s_t + M$ fluid model with infeasible s.

(above) s (the red solid curve), and $B(t)$ follows s_f instead of s since $B(t)$ cannot decrease as fast as $s(t)$. Moreover, since $b(t, 0) = 0$ for $t \in I_{inf}$, $w(t)$ increases with slope 1. In other words, since the system stops transporting fluid from the queue into service, whatever is waiting at the head of the queue keeps waiting there. However, $Q(t)$ does not increase with rate 1 because abandonment still occurs.

Figure A.5 shows that $w(t)$, $Q(t)$ and $B(t)$ obtained from our modified algorithm (the red dashed curves) agrees with single sample paths of simulation estimates of $w_n(t)$, $\hat{Q}_n(t)$ and $\hat{B}_n(t)$ (the blue solid curves), where we still set the fluid scaling factor n = 1000. Both

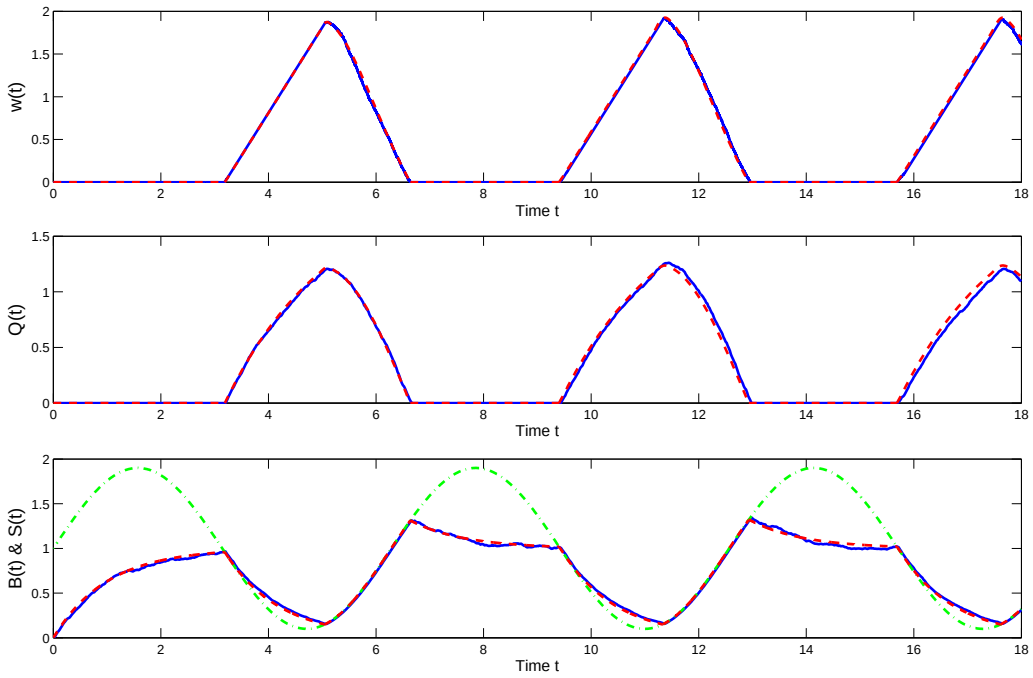

Figure A.5: The $M/M/s_t + M$ fluid model with infeasible s compared with simulation.

$B(t)$ and $\hat{B}_n(t)$ are distinct from the given service-capacity function s (the dashed green curve) in I_{inf}.

A.8 Stabilizing Delays with General Initial Conditions

Is §2.10 we showed how to choose a staffing function to stabilize the PWT v at any desired target v^*. However, Theorem 2.8 considered a special initial condition: the system is initially empty. We generalize Theorem 2.8 to arbitrary initial conditions in the next theorem.

Theorem A.8 *Consider the $G_t/GI/s_t + GI$ fluid model with a general arrival-rate function λ and initial conditions $w(0-) \equiv w_0 \geq 0$, $b(0-,x) \equiv \psi(x) \geq 0$ for $x \geq 0$, $q(0-,x) \equiv \phi(x) \geq 0$ for $0 \leq x \leq w_0$, $Q(0-) = \int_0^{w_0} q(0-,x)dx$, $s(0-) = B(0-) = \int_0^\infty b(0-,x)dx$.*

For any given $v^* \geq 0$, we can make the system overloaded such that the PWT is fixed at v^*, i.e., $v(t) = v^*$ for all $t \geq 0$, by letting the service-capacity function be

$$\begin{aligned} s(t) &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0 \vee v^*} \phi(x)dx \\ &+ \bar{F}(v^*)\left(\int_{(t-v^*)^+}^{t-(v^*-w_0)^+} \frac{\phi(w_0 \wedge v^* - t + x)\,\bar{G}(x)}{\bar{F}(w_0 \wedge v^* - t + x)}dx\right) \cdot 1_{\{t \geq (v^*-w_0)^+\}} \quad \text{(A.31)} \\ &+ \bar{F}(v^*)\left(\int_0^{t-v^*} \lambda(t-x-v^*)\bar{G}(x)dx\right) \cdot 1_{\{t \geq v^*\}}. \end{aligned}$$

If we do so, then

$$
\begin{aligned}
w(t) &= v^* \cdot 1_{\{t \geq (v^*-w_0)^+\}}, \\
b(t,0) &= \delta_0(t) \int_{v^*}^{w_0 \vee v^*} \phi(x)dx + \frac{\phi(w_0 \wedge v^* - t)\, \bar{F}(v^*)}{\bar{F}(w_0 \wedge v^* - t)} \cdot 1_{\{(v^*-w_0)^+ \leq t < v^*\}} + \lambda(t-v^*)\bar{F}(v^*) \cdot 1_{\{t \geq v^*\}}, \\
B(t) &= s(t), \\
\sigma(t) &= \int_t^\infty \psi(x-t)\frac{g(x)}{\bar{G}(x-t)}dx + g(t)\int_{v^*}^{w_0 \vee v^*} \phi(x)dx \\
&+ \bar{F}(v^*)\left(\int_{(t-v^*)^+}^{t-(v^*-w_0)^+} \frac{\phi(w_0 \wedge v^* - t + x)\, g(x)}{\bar{F}(w_0 \wedge v^* - t + x)}dx\right) \cdot 1_{\{t \geq (v^*-w_0)^+\}} \\
&+ \bar{F}(v^*)\left(\int_0^{t-v^*} \lambda(t-x-v^*)g(x)dx\right) \cdot 1_{\{t \geq v^*\}}, \\
Q(t) &= \left(\int_t^{w_0 \wedge v^*} \frac{\phi(x-t)\, \bar{F}(x)}{\bar{F}(x-t)}dx + \int_0^t \lambda(t-x)\, \bar{F}(x)dx\right) \cdot 1_{\{0 \leq t \leq (v^*-w_0)^+\}} \\
&+ \left(\int_0^t \lambda(t-x)\, \bar{F}(x)dx + \int_t^{v^*} \frac{\phi(x-t)\, \bar{F}(x)}{\bar{F}(x-t)}dx\right) \cdot 1_{\{(v^*-w_0)^+ < t < v^*\}} \\
&+ \left(\int_0^{v^*} \lambda(t-x)\, \bar{F}(x)dx\right) \cdot 1_{\{t \geq v^*\}}, \\
\alpha(t) &= \left(\int_t^{w_0 \wedge v^*} \frac{\phi(x-t)\, f(x)}{\bar{F}(x-t)}dx + \int_0^t \lambda(t-x)\, f(x)dx\right) \cdot 1_{\{0 \leq t \leq (v^*-w_0)^+\}} \\
&+ \left(\int_0^t \lambda(t-x)\, f(x)dx + \int_t^{v^*} \frac{\phi(x-t)\, f(x)}{\bar{F}(x-t)}dx\right) \cdot 1_{\{(v^*-w_0)^+ < t < v^*\}} \\
&+ \left(\int_0^{v^*} \lambda(t-x)\, f(x)dx\right) \cdot 1_{\{t \geq v^*\}}
\end{aligned}
$$

where $\delta_y(t)$ is the direct-delta function at y, i.e., $\delta_y(t) = 0$ for $t \neq y$, $\int_a^b \delta_y(t)dt = 1$ if $a \leq y \leq b$.

Proof. (i) If the system is initially underloaded, i.e., $w(0-) = w_0 = 0$, $q(0-,x) = \phi(x) = 0$, $Q(0-) = 0$, $B(0-) \leq s(0-)$. This case is similar to Theorem 2.8 where the system is initially empty. Note the only difference is that there is fluid in the service facility.

Let $B^o(t)$ be the fluid in service that has been in service at $0-$. Then we have

$$B^o(t) = \int_t^\infty b(t,x)dx = \int_t^\infty b(0-, x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx.$$

Again, we do not allow any input to enter service until time $t = v^*$, we can let the staffing function be

$$\begin{aligned} s(t) &= B^o(t) + s^*(t) \\ &= \int_t^\infty \frac{\psi(x-t)\,\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{F}(v^*)\int_0^{t-v^*} \bar{G}(x)\lambda(t-v^*-x)dx \cdot 1_{\{t>v^*\}}, \end{aligned}$$

where $s^*(t)$ is defined in (2.48). It is obvious that this expression coincides with (A.31) when $w_0 = q(0-,x) = \psi(x) = 0$. When we do this, the input rate to the service $b(t,0)$ is the same as in Theorem 2.8. The proof of other performance measures are similar.

(ii) If the system is initially overloaded, i.e., $w(0-) = w_0 > 0$, $q(0-,x) = \phi(x) \geq 0$, $Q(0-) = \int_0^{w_0}\phi(x)dx > 0$, $s(0-) = B(0-)$. There are two cases (a) $w_0 \geq v^*$, (b) $w_0 < v^*$.

(ii.a) If $w_0 > v^*$, then in order for $v(t) = v^*$. We let all fluid that has been in queue for $x > v^*$ enter service immediately at time 0. The quantity of fluid that enters service at 0 is $\int_{v^*}^{w_0} q(0-,x)dx = \int_{v^*}^{w_0}\phi(x)dx$. However, this will make $B(t)$ have an atom at 0. Similar argument to Theorem 2.8 implies that it suffices to match $b(t,0)$ with $q(t,v^*)$ for all $t \geq 0$. If $t \leq v^*$, $q(t,v^*) = q(0-,v^*-t)\bar{F}(v^*)/\bar{F}(v^*-t)$. If $t > v^*$, then all fluid that has been in queue at $0-$ has entered service, which implies that $q(t,v^*) = q(t-v^*,0)\bar{F}(v^*) =$

$\lambda(t-v^*)\bar{F}(v^*)$. Therefore, we have

$$\begin{aligned} b(t,0) &= \delta_0(t)\int_{v^*}^{w_0}\phi(x)dx + q(t,v^*) \\ &= \delta_0(t)\int_{v^*}^{w_0}\phi(x)dx + \frac{\phi(v^*-t)\bar{F}(v^*)}{\bar{F}(v^*-t)}\cdot 1_{\{0\le t<v^*\}} + \lambda(t-v^*)\bar{F}(v^*)\cdot 1_{\{t\ge v^*\}}. \end{aligned}$$

The service capacity and fluid content in service are

$$s(t) = B(t) = B^o(t) + \int_0^t b(t-x,0)\bar{G}(x)dx.$$

If $0 \le t < v^*$, we have

$$\begin{aligned} s(t) &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \int_{v^*}^{w_0}\phi(x)dx\int_0^t \delta_0(t-x)\bar{G}(x)dx \\ &\quad +\bar{F}(v^*)\int_0^t \frac{\phi(v^*-t+x)\bar{G}(x)}{\bar{F}(v^*-t+x)}dx, \\ &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0}\phi(x)dx + \bar{F}(v^*)\int_0^t \frac{\phi(v^*-t+x)\bar{G}(x)}{\bar{F}(v^*-t+x)}dx. \end{aligned}$$

If $t \ge v^*$, we have

$$\begin{aligned} s(t) &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0}\phi(x)dx \\ &+ \int_0^t \left(\frac{\phi(v^*-t+x)\bar{F}(v^*)}{\bar{F}(v^*-t+x)}\cdot 1_{\{0\le t-x<v^*\}} + \lambda(t-x-v^*)\bar{F}(v^*)\cdot 1_{\{t-x\ge v^*\}}\right)\bar{G}(x)dx \\ &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0}\phi(x)dx \\ &+ \bar{F}(v^*)\left(\int_{t-v^*}^t \frac{\phi(v^*-t+x)\bar{G}(x)}{\bar{F}(v^*-t+x)}dx + \int_0^{t-v^*}\lambda(t-x-v^*)\bar{G}(x)dx\right). \end{aligned}$$

It is easy to see that this expression coincides with (A.31).

(ii.b) If $w_0 \le v^*$, then we do not allow any input to enter service until time $v^* - w_0$,

which implies

$$b(t,0) = \frac{\phi(w_0 - t)\,\bar{F}(v^*)}{\bar{F}(w_0 - t)} \cdot 1_{\{v^* - w_0 \le t < v^*\}} + \lambda(t - v^*)\bar{F}(v^*) \cdot 1_{\{t \ge v^*\}}.$$

Therefore, if $0 \le t \le v^* - w_0$, no new fluid enters service,

$$s(t) = B^o(t) = \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx.$$

If $v^* - w_0 < t < v^*$,

$$\begin{aligned} s(t) &= B^o(t) + \int_0^t \frac{\phi(w_0 - t + x)\,\bar{F}(v^*)}{\bar{F}(w_0 - t + x)} \cdot 1_{\{v^* - w_0 \le t - x < v^*\}} \bar{G}(x)dx \\ &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{F}(v^*)\int_0^{t-(v^*-w_0)} \frac{\phi(w_0 - t + x)\,\bar{G}(x)}{\bar{F}(w_0 - t + x)}dx. \end{aligned}$$

If $t \ge v^*$,

$$\begin{aligned} s(t) &= B^o(t) \\ &\quad + \int_0^t \left(\frac{\phi(w_0 - t + x)\,\bar{F}(v^*)}{\bar{F}(w_0 - t + x)} \cdot 1_{\{v^* - w_0 \le t - x < v^*\}} + \lambda(t - x - v^*)\bar{F}(v^*) \cdot 1_{\{t - x \ge v^*\}} \right) \bar{G}(x)dx \\ &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx \\ &\quad + \bar{F}(v^*)\left(\int_{t-v^*}^{t-(v^*-w_0)} \frac{\phi(w_0 - t + x)\,\bar{G}(x)}{\bar{F}(w_0 - t + x)}dx + \int_0^{t-v^*} \lambda(t - x - v^*)\bar{G}(x)dx \right). \end{aligned}$$

It is easy to see that this expression coincides with (A.31). The proof of other performance measures is similar.

A.9 Comparisons with Simulation

In this section we present additional results evaluating the fluid model approximations by comparing them to simulation results for large-scale queueing models. These results complement those for the $M_t/H_2/s + E_2$ example in §2.2.

We start by applying our algorithm to the special "base" case of an $M_t/M/s + M$ model, having only a time-varying arrival rate function. For this special case, we could also have applied [46–48]. In §A.9.2 we present additional simulation results for allowing the alternative features: (i) time-varying staffing function, (ii) non-exponential abandonment-time cdf, and (iii) non-Poisson arrival process. (The fluid model does not change when we change the arrival process from M_t, to G_t, but the queueing system does.)

In §2.2 we already considered the $M_t/H_2/s + E_2$ model, which has both time-varying arrival rate and non-exponential service and patience distributions. We consider other examples in §A.9.3.

A.9.1 A Base Example

We start by applying our algorithm to the base case of an $M_t/M/s + M$ model, having only a time-varying arrival rate function.

For the initial $M_t/M/s + M$ model fluid example, we consider constant staffing s. We let the arrival rate function λ be sinusoidal, i.e.,

$$\lambda(t) \equiv a + b \cdot \sin(c \cdot t), \quad t \geq 0, \tag{A.32}$$

where we let $b \equiv 0.6a$, $c \equiv 1$ and $a \equiv s$. By making the average input rate a coincide with the fixed staffing level s, we ensure that the system will alternate between overloaded and underloaded. We let the service rate be $\mu \equiv 1$ and the abandonment rate $\theta \equiv 0.5$; i.e.,

$G(x) \equiv 1 - e^{-x}$ and $F(x) = 1 - e^{-\theta x} = 1 - e^{-0.5x}$ for $x \geq 0$. Without loss of generality, for the fluid model we let $s \equiv 1$.

Figure A.6 shows key fluid performance functions of this $M_t/M/s + M$ example. In Figure A.6, we plot key fluid performance measures for $0 \leq t \leq T$, where $T = 16$. It is easy to see that the system alternates between underloaded (when $Q(t) = 0$ and $B(t) < s(t) = 1$) and overloaded (when $Q(t) > 0$ and $B(t) = s(t) = 1$) intervals.

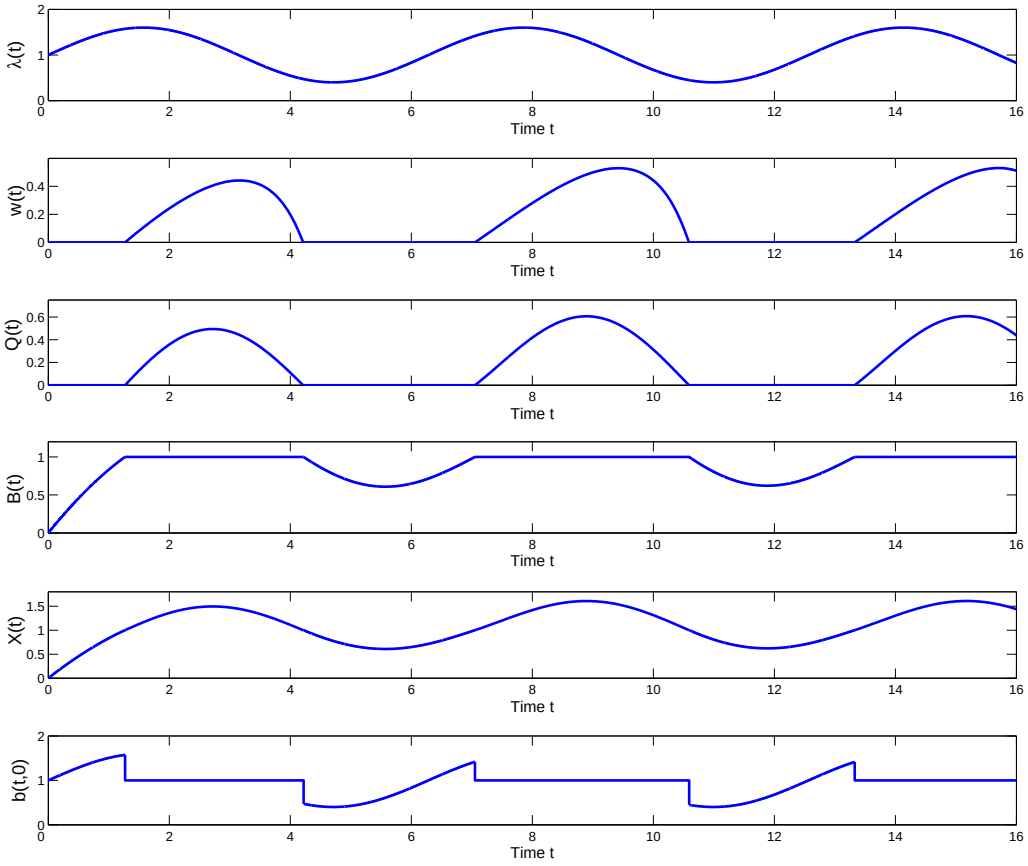

Figure A.6: The performance functions of the $M_t/M/s + M$ fluid model with sinusoidal arrival-rate function: (i) arrival rate $\lambda(t)$; (ii) waiting time $w(t)$; (iii) fluid in buffer $Q(t)$; (iv) fluid in service $B(t)$; (v) total fluid $X(t)$; (vi) rate into service $b(t, 0)$.

As discussed in §2.2, it is important that the fluid model provide useful approximations for stochastic queueing models. We apply simulation to show that the fluid approximation indeed is effective for that purpose. For very large queueing systems, the stochastic system

behaves like the fluid model, having relatively small stochastic fluctuations. That is illustrated for the same example for a queueing system with 1000 servers in Figure A.7. (In the plot, the queueing content processes are scaled by dividing by $n = 1000$, so that s remains at 1.)

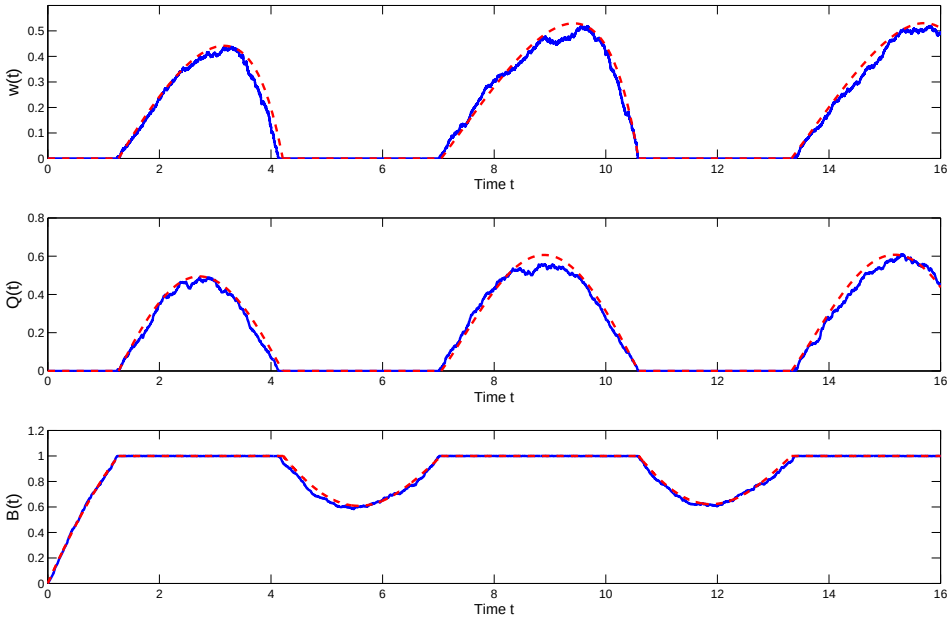

Figure A.7: Performance of the $M_t/M/s + M$ fluid model (dashed lines) compared with simulation results (solid lines): one sample path of the scaled queueing model for $n = 1000$.

We did not plot the abandonment rate α and the service-completion rate σ, because in the exponential case they are simple functions of the performance measures shown: $\alpha(t) = \theta Q(t) = 0.5Q(t)$ and $\sigma(t) = \mu B(t) = B(t)$. All performance functions are continuous except for the transportation-rate function $b(\cdot, 0)$, which has discontinuities when the system alternates between underloaded and overloaded: $b(t, 0) = \lambda(t)$ when the system is underloaded; $b(t, 0) = s = 1$ when the system is overloaded.

With the MSHT scaling, we let $n \equiv 1000$. Since, $s = 1$, that makes $s_n = a_n = 1000$, which of course is very large. The other parameters of the queueing model are the same

as for the fluid model, e.g., $b_n = 0.6a_n = 600$. In Figure A.7 we compare the simulation results for the queueing performance functions W_n, $\bar{Q}_n$ and $\bar{B}_n$ from a single simulation run to the associated fluid model counterparts w, Q and B. The blue solid lines represent the queueing model performance, while the red dashed lines represent the corresponding fluid performance. Since n is so large, we get close agreement for individual sample paths; we are not displaying averages over multiple simulation runs.

Of course, most service systems have fewer servers. It is thus important that the fluid approximation can still be useful with fewer servers. With fewer servers, the stochastic fluctuations in the queueing stochastic processes play an important role. In that case, the fluid model can still be very useful by providing a good approximation for the *mean values* of the queueing stochastic processes. That is illustrated from the plot of the average of the scaled performance measures of 200 independent sample paths when there are only 20 servers in Figure A.8.

In Figure A.9 below we plot the analog of Figure A.7 for the case of one sample path of the simulation with $n = 100$, for the same fluid model. In Figure A.10 below we plot the average of 10 sample paths. We see that the fluid approximation provides only a rough approximation for a single sample path, but it is remarkably accurate for the average over 10 sample paths. The accuracy is especially high in this example, because the extent of the overloads and underloads are quite large.

The quality of the approximation does degrade as n decreases, for the given fluid model. To illustrate, we plot a single sample path for $n = 20$ in Figure A.11 and the average over 200 sample paths in Figure A.8. (The latter appears in Chapter 2.) The stochastic fluctuations are so much greater for a single sample path that we need to average over more sample paths to get a good estimate. For $n = 20$, the fluid model clearly yields a good approximation only for the mean values, but the mean is remarkably well approximated for $n = 20$. The approximation for the mean values in Figure A.8 are so good that it is evident

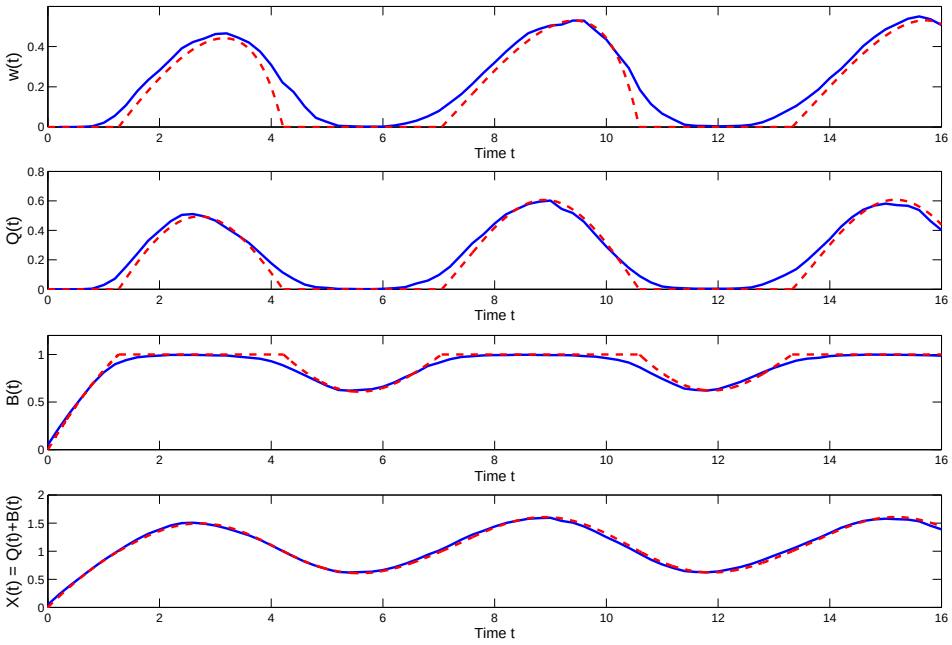

Figure A.8: Performance of the $M_t/M/s+M$ fluid model (dashed lines) compared with simulation results (solid lines): an average of 200 sample paths of the scaled queueing model based on $n = 20$.

that the fluid model approximations can provide useful approximations for the mean values for much smaller n (and thus the number of servers, s_n).

A.9.2 Variants of the Base Model

We now consider three variants of the base model in order to illustrate consider: (i) time-varying staffing, (ii) non-exponential abandonment and (iii) a non-Poisson arrival process.

Time-Varying Staffing Levels

We now consider a Markovian $M/M/s_t+M$ model that has a Poisson arrival process with a constant rate λ, exponential service and abandonment distributions with rates μ and θ

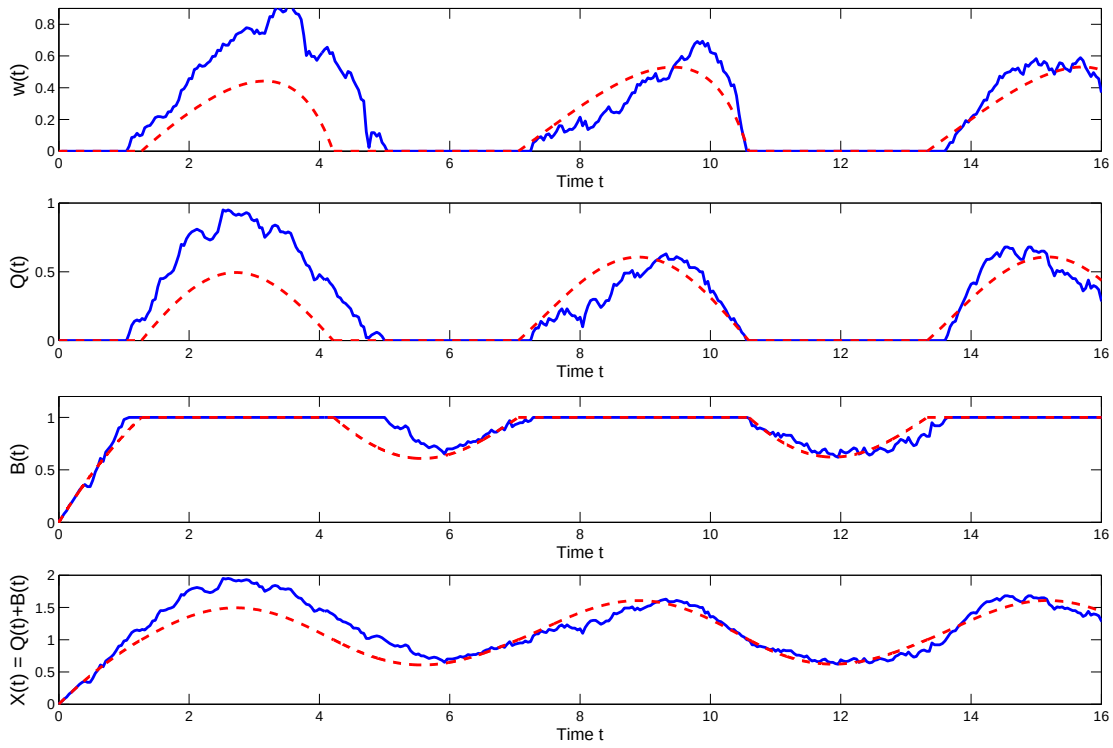

Figure A.9: Performance of the $M_t/M/s + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.

respectively, and a sinusoidal capacity function

$$s(t) \equiv \lambda + \bar{\lambda} \cdot \sin(c \cdot t). \tag{A.33}$$

In the previous base example in §A.9.1, we fixed the capacity function and varied the arrival rate around it; now we fix the arrival rate λ and vary $s(t)$ around λ. We let $\lambda = 1$, $\bar{\lambda} = 0.6\lambda = 0.6$, $c = 1$, $\mu = 1$ and $\theta = 0.5$.

Before implementing the algorithm, we first verify that this capacity function s is feasible. With exponential service distribution, we know that a sufficient condition for the

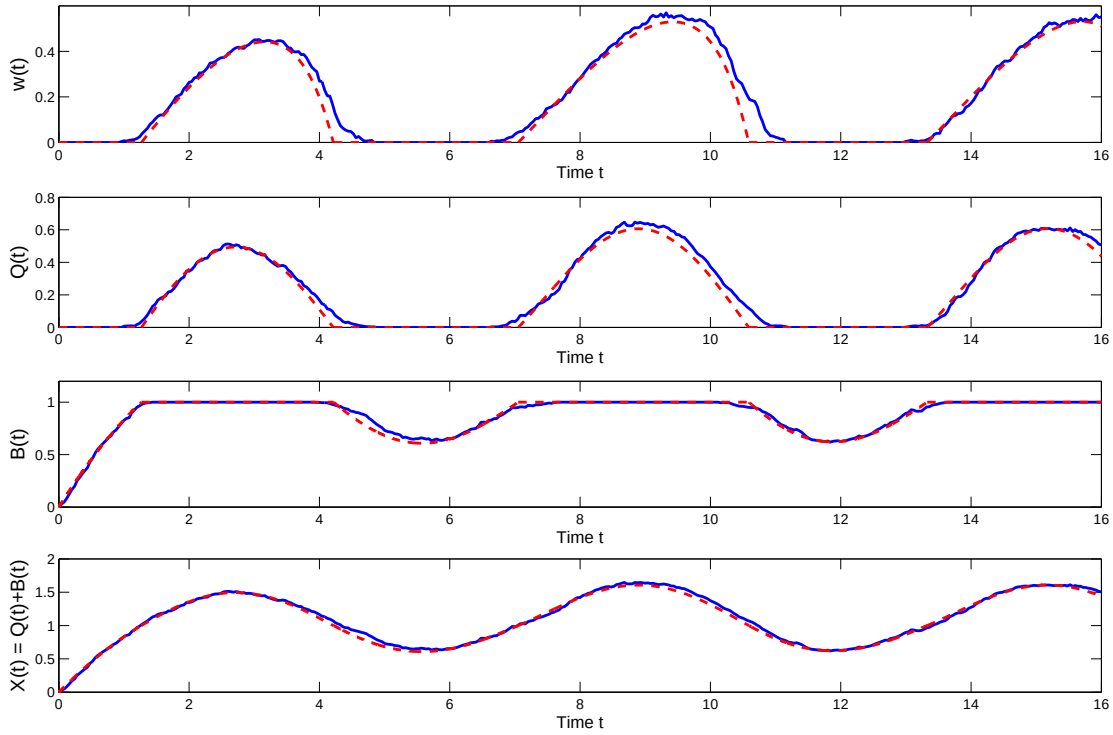

Figure A.10: Performance of the $M_t/M/s + M$ fluid model compared with simulation results: an average of 10 sample paths of the scaled queueing model based on $n = 100$.

feasibility of s is

$$s'(t) \geq -\mu s(t), \quad t \geq 0. \tag{A.34}$$

In this example, we require $c\cos(ct) \geq -\mu\lambda - \mu\bar{\lambda}\sin(ct)$ which is equivalent to $\sin(ct + \bar{\theta}) \geq -(\mu/\sqrt{c^2+\mu^2})(\lambda/\bar{\lambda})$ where $\bar{\theta} \equiv \arctan(c/\mu)$. It is easy to check that this equality holds with $\lambda = 1$, $\bar{\lambda}$, $\mu = 1$ and $c = 1$.

We plot the performance measures of the $M/M/s_t + M$ fluid model in Figure A.12 and compare them with simulation estimates in A.13, analogs to Figure A.6 and A.7. In Figure A.13, our simulations add real system constraints. First the staffing levels must be integer-valued, so they must be rounded. Second, when the staffing levels decrease, we do

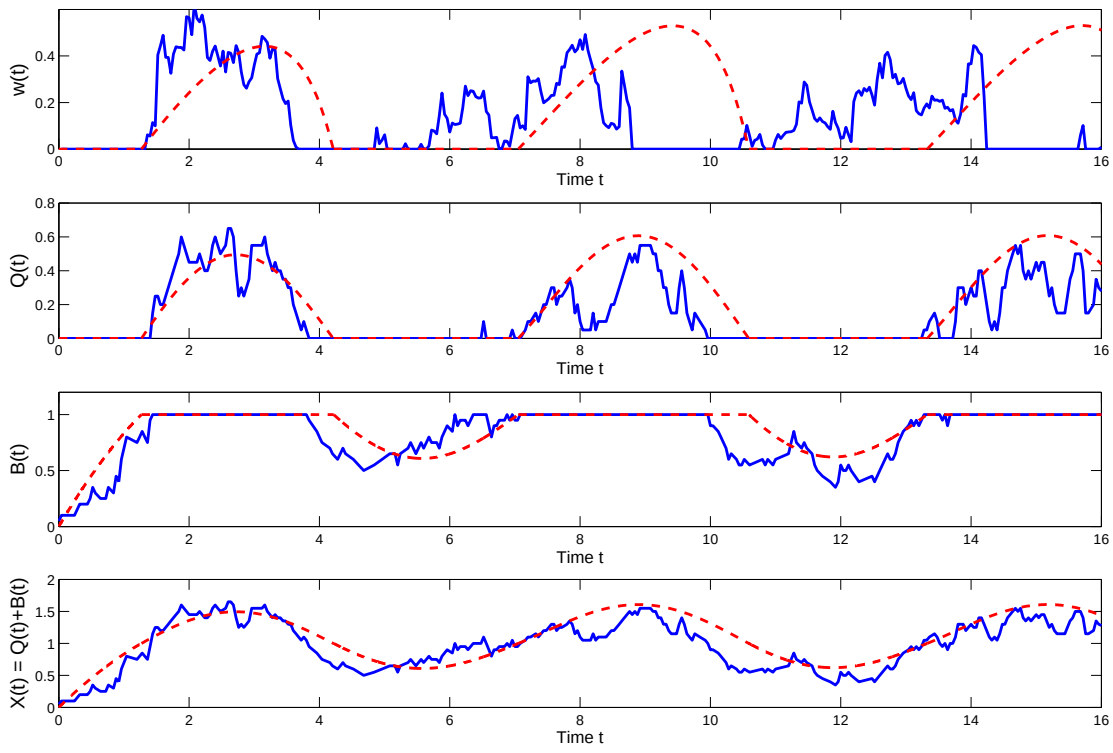

Figure A.11: Performance of the $M_t/M/s + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 20$.

not remove servers until they complete the service in progress. As in §C.6, we let $n = 1000$ for the sequence of scaled queueing models. Thus we have $\lambda_n = a_n = 1000$, $b_n = 600$, $c_n = 1$.

Simulation Comparisons for the $M_t/M/s_t + GI$ Fluid Model.

For the general abandon-time distribution, we considered two cases: Erlang-2 (E2) and Hyperexponential-2 (H2). Let A be the generic abandonment time. A follows E2 implies that $A = X_1 + X_2$ in distribution, where X_1 and X_2 are two iid exponential random variables. Moreover, $f(x) = \gamma^2 x e^{-\gamma x}$, where γ is rate of X_1.

If A follows H2, then A is a composition of two exponential random variables, i.e.,

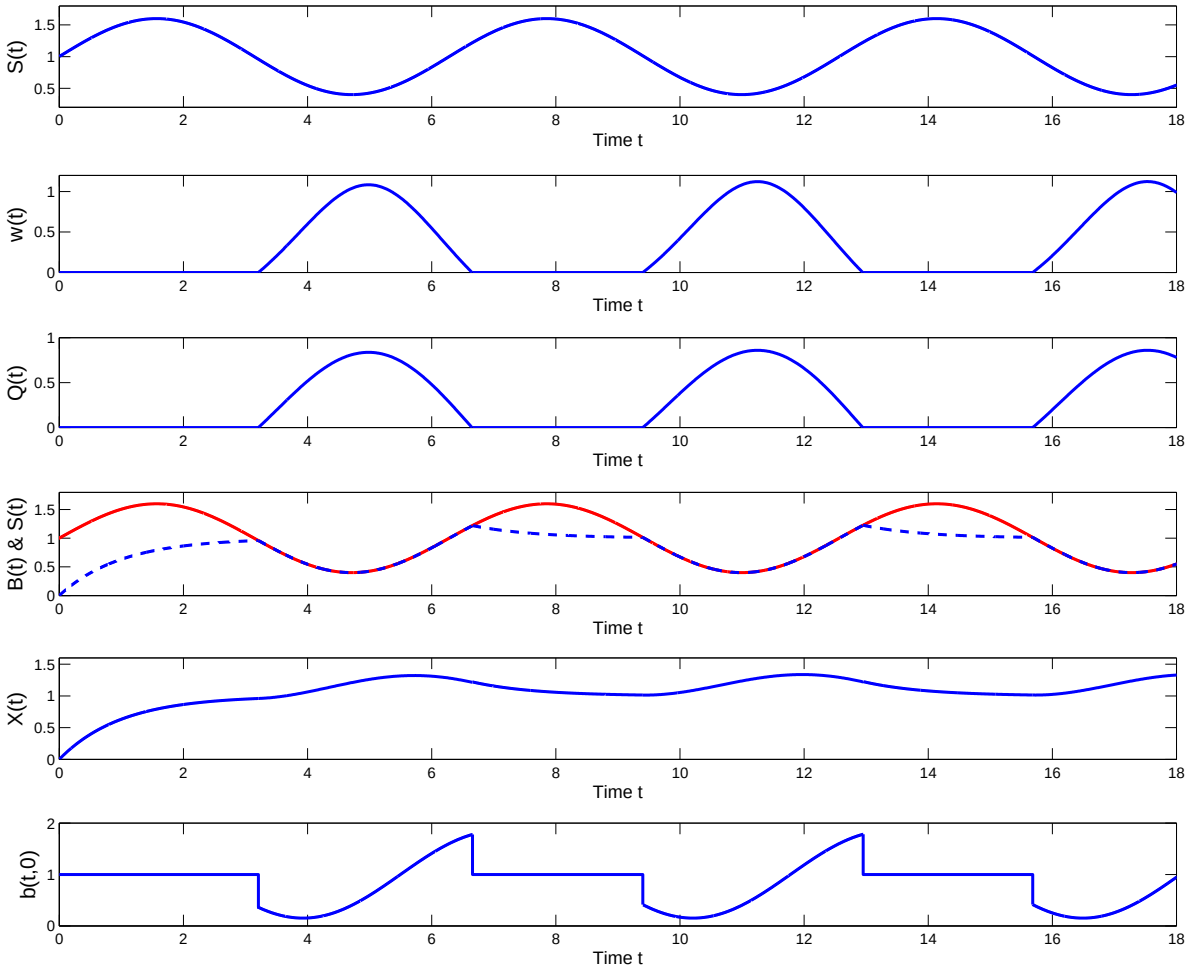

Figure A.12: The $M/M/s_t + M$ fluid model with sinusoidal service-capacity function.

$f(x) = p \cdot \theta_1 e^{-\theta_1 x} + (1-p) \cdot \theta_2 e^{-\theta_2 x}$, where θ_1 and θ_2 are the rates of these two exponential random variables, and $0 < p < 1$ is the sampling probability.

If we fix the mean of A, i.e., let $E[A] = 1/\theta$, E2 has squared coefficient of variation (SCV) $C_{SCV} \equiv Var(A)/E[A]^2$ less than 1; H2 has C_{SCV} greater than 1 if p, θ_1 and θ_2 are appropriately chosen.

For E2, we let $f(x) \equiv 4\theta^2 x e^{-2\theta x}$ such that $C_{SCV} = 1/2$. For H2, we let $f(x) = p \cdot \theta_1 e^{-\theta_1 x} + (1-p) \cdot \theta_2 e^{-\theta_2 x}$ with $p = 0.5(1 - \sqrt{0.6})$, $\theta_1 = 2p\theta$, $\theta_2 = 2(1-p)\theta$, such that $C_{SCV} = 4$.

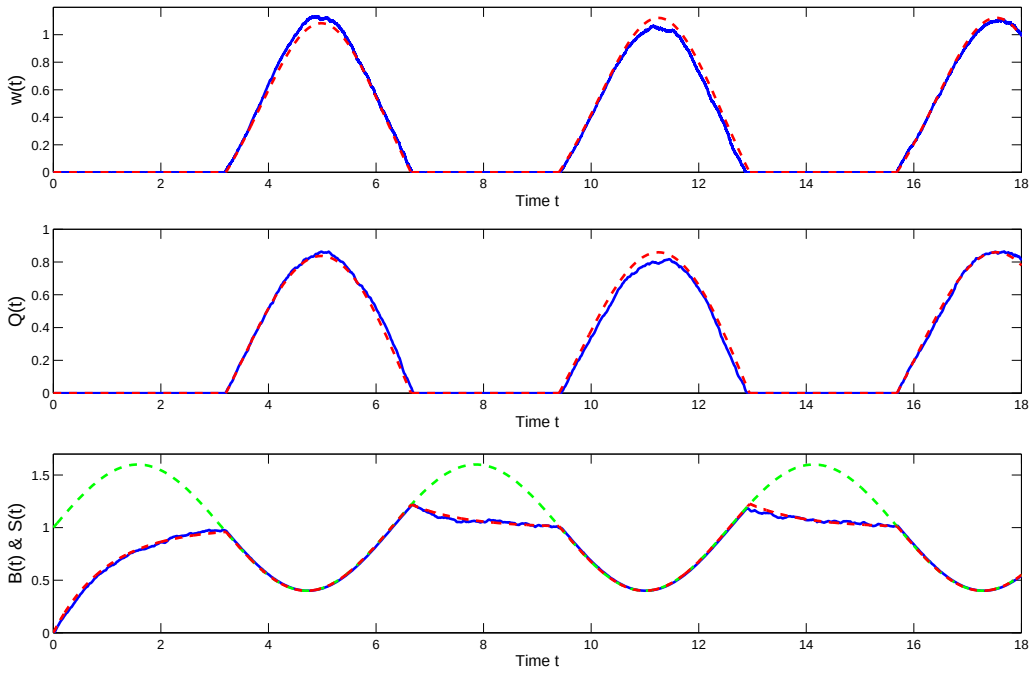

Figure A.13: The $M/M/s_t + M$ fluid model compared with simulations of the queueing system.

We still let the arrival-rate function λ be sinusoidal, as in (A.32). We let $a = 1$, $b = 0.6 * a = 0.6$, $c = 1$. We let the service-capacity function be constant $s = 1$. Let $\theta = 0.5$ and $\mu = 1$. We plot the dynamics of the $M_t/M/s + E2$ and $M_t/M/s + H2$ fluid models in Figure A.14 and A.16 respectively for $t \in [0, T]$ with $T = 16$. The performance measures shown in Figure A.14 and A.16 are the boundary waiting time $w(t)$, the fluid in queue $Q(t)$, the fluid in service $B(t)$, the total fluid in the system $X(t)$, the abandonment rate $\alpha(t)$, and the transportation rate $b(t, 0)$. We omit the departure rate $\sigma(t) = \mu B(t)$ because of the exponential service times.

In Figure A.15 and A.17 we compare the fluid approximations with simulation experiments. The queueing model has a nonhomogeneous Poisson arrival process with sinusoidal rate function as in (A.32), with $a = s = 2000$, $b = 0.6a = 1200$. In Figure A.15 and A.17, the blue solid lines of the simulation estimations of single sample paths applied with fluid

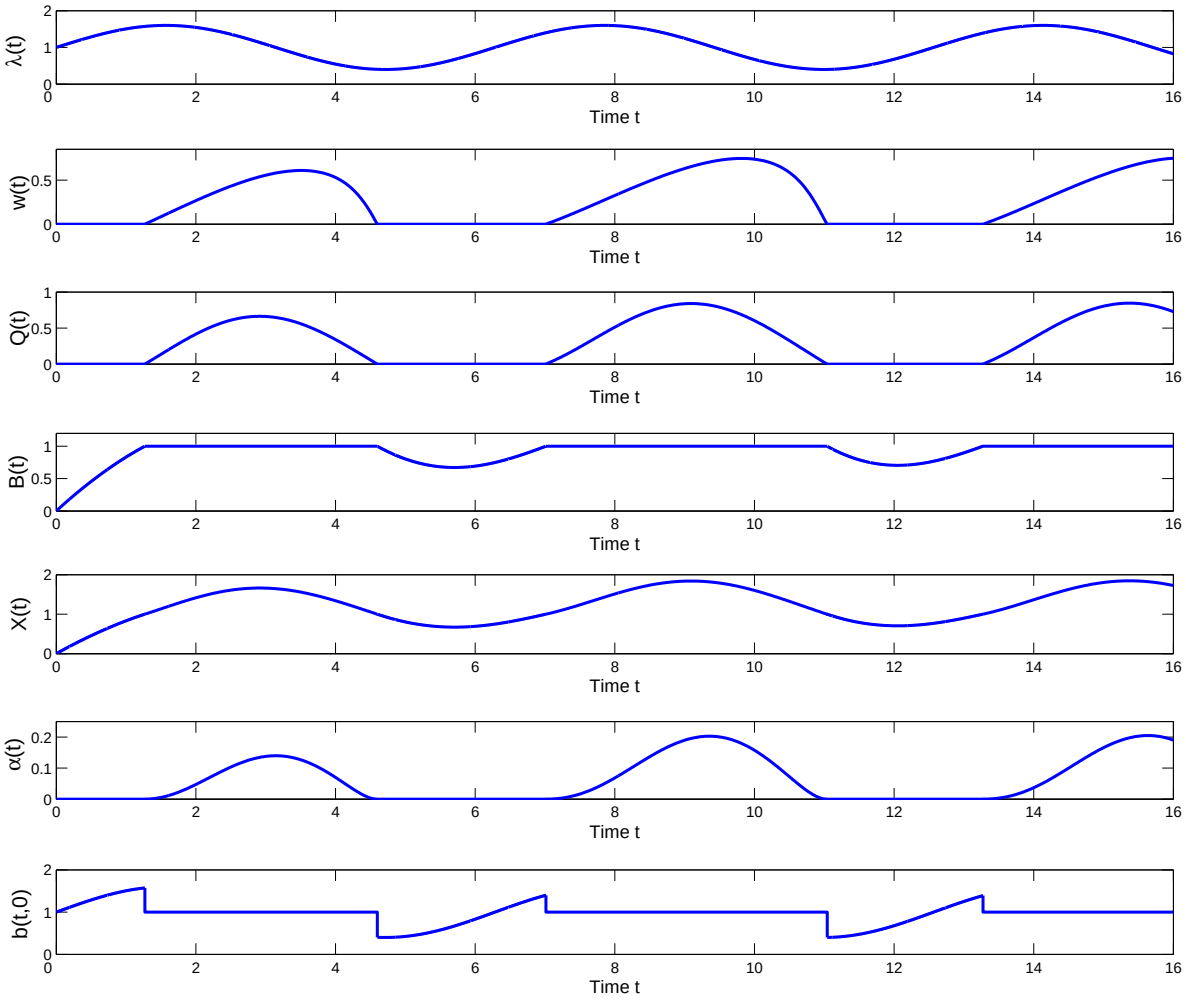

Figure A.14: The $M_t/M/s + E2$ fluid model with sinusoidal arrival-rate function.

scaling, and the red dashed lines are the fluid approximations. We conclude that the fluid approximation is remarkably accurate.

Simulation Comparisons for the $G_t/M/s_t + M$ Fluid Model.

We first explain how to construct a non-Poisson arrival process that has a well-defined rate function.

Let $\mathbf{M} \equiv \{M(t) : t \geq 0\}$ be a delayed renewal process. In other words, let $X_1, X_2, X_3, \ldots$

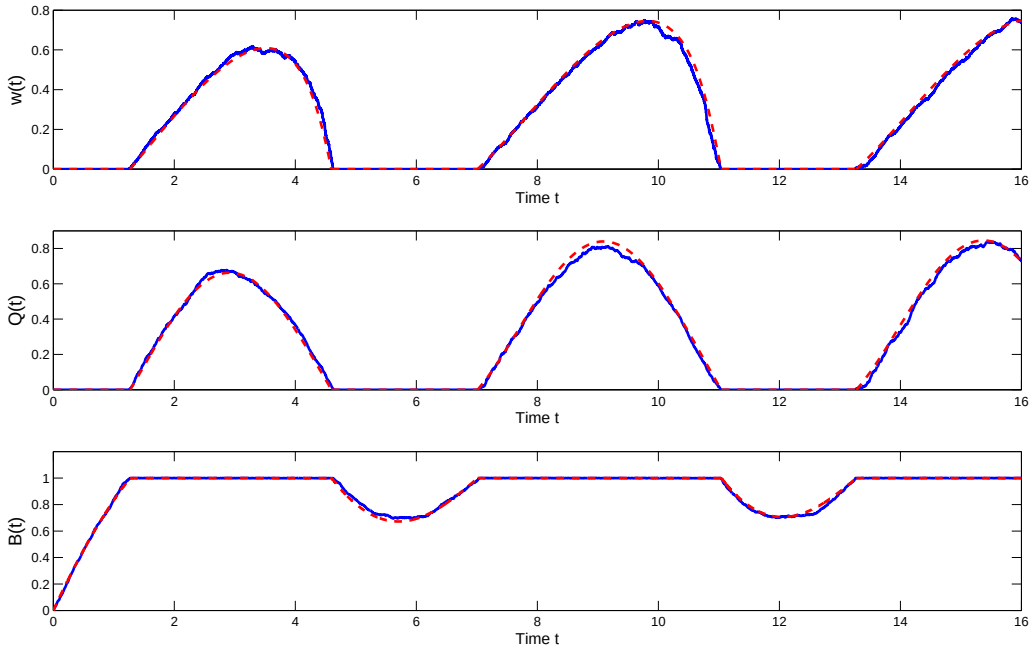

Figure A.15: The $M_t/M/s + E2$ fluid model compared with simulations of the queueing system.

be independent random variables with finite means, such that X_1 follows cdf H, X_n follows cdf G for $n \geq 2$. Let $S_n \equiv \sum_{k=1}^{n} X_k$ and define $M(t) \equiv \sup\{n \geq 0 : S_n \leq t\}$.

In particular, if we let $H(x) = G_e(x) \equiv 1/m_X \int_0^t \bar{G}(u)du$ for $m_X \equiv E[X_2]$, which is the equilibrium distribution of G, then $\mathbf{M}$ becomes an equilibrium renewal process and we have $E[M(t)] = t/m_X$ for any $t \geq 0$. We call $\mathbf{M}$ standard equilibrium renewal process (SERP) if $m_X = 1$.

For a given rate function $\lambda(t)$, let $\Lambda(t) \equiv \int_0^t \lambda(u)du$. We assume that $\lambda(t) > 0$ for $t \geq 0$, hence $\Lambda(t)$ is a strictly increasing function. For a given SERP $\mathbf{M}$, we construct a process that has rate function $\lambda(t)$ by performing a change of time with respect to this function $\Lambda(t)$. We define $\mathbf{N} \equiv \{N(t) \equiv M(\Lambda(t)) : t \geq 0\}$. Since $E[N(t)] = \Lambda(t)$ for $t \geq 0$, process $\mathbf{N}$ has a well-defined rate function.

Since the cdf G is not necessarily exponential, $\mathbf{N}$ is just in general a non-Markovian

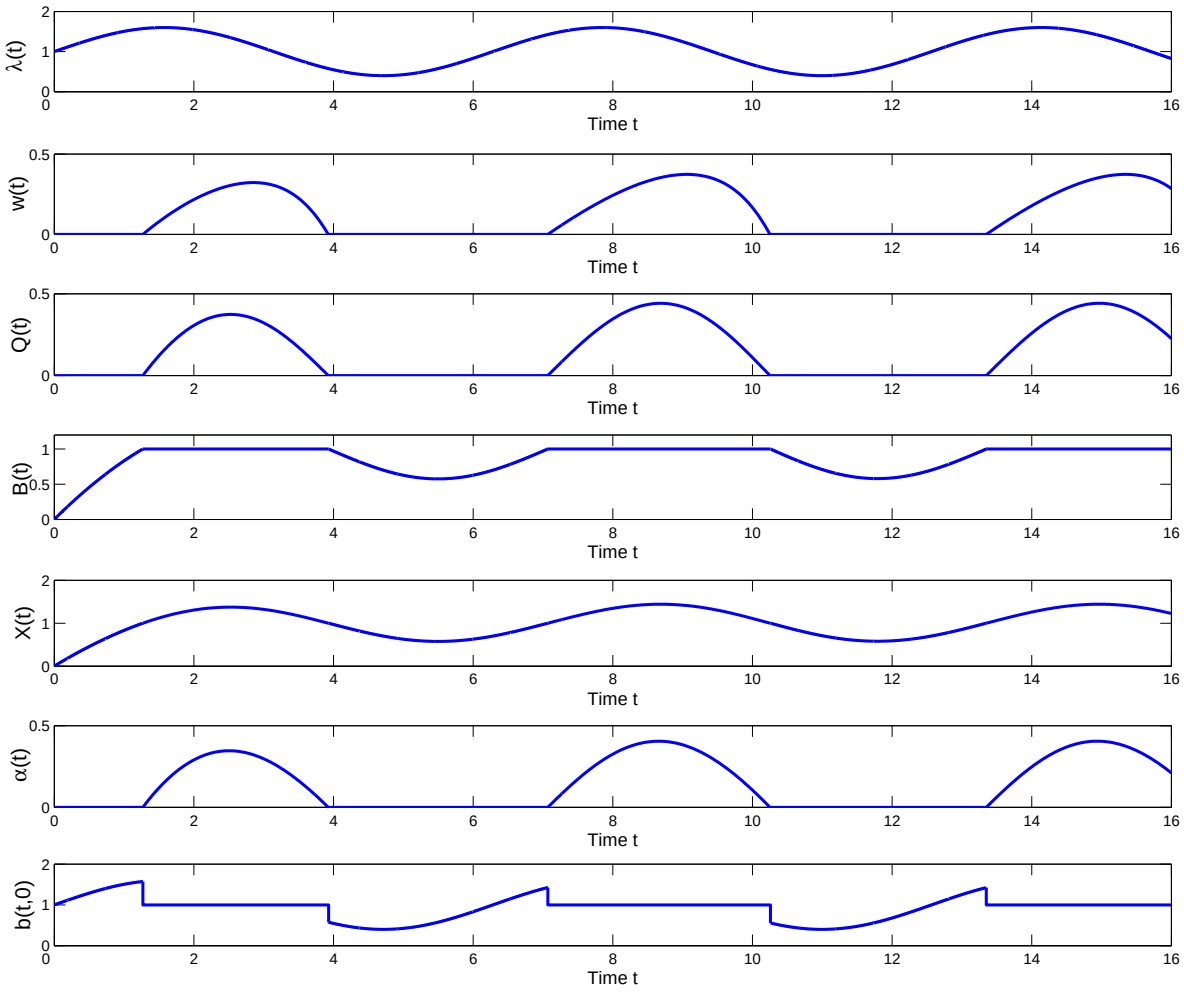

Figure A.16: The $M_t/M/s + H2$ fluid model with sinusoidal arrival-rate function.

arrival counting process that has time-dependent rate function $\lambda(t)$. Now we explain how to simulate the point process associated with $\mathbf{N}$, i.e., to simulate the times of arrivals of $\mathbf{M}$. For a given sample path of the SERP $\mathbf{M}$, let $S_n = s_n$ for $n \geq 0$, we want to determine the arrival times t_n's, where t_n is the time at which the nth arrival occurs. It is easy to see that $t_n = \Lambda^{-1}(s_n)$ for $n \geq 0$, where $\Lambda^{-1}(\cdot)$ is unique since $\Lambda(\cdot)$ is strictly increasing. Therefore, to obtain a sample path of $\mathbf{N}$, we simulate a sample path of $\mathbf{M}$ and do a change of time.

In Figure A.18, we compare the fluid approximation with simulation experiments of the

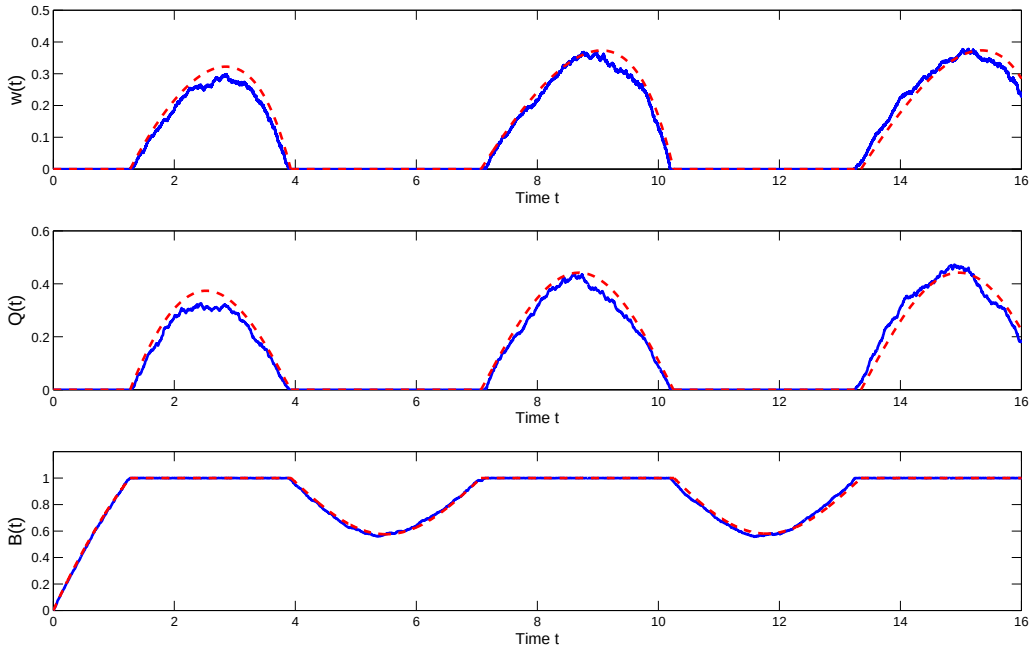

Figure A.17: The $M_t/M/s + H2$ fluid model compared with simulations of the queueing system.

$G_t/M/s + M$ model. Here the only difference from Figure A.7 is that the arrival process (G_t) is not Poisson but has the same sinusoidal rate function as (A.32).

A.9.3 More Comparisons for the Example in §2.2 with GI Service

Here we consider the $M_t/H_2/s + E_2$ example in §C.6 with smaller n. As shown in Figure A.19, we plot the mean value functions, obtained by averaging the paths of 500 independent simulation runs, with $n = 15$. Although less accurate than the case $n = 30$, the fluid model serves as a much better approximation than the algorithm of M service.

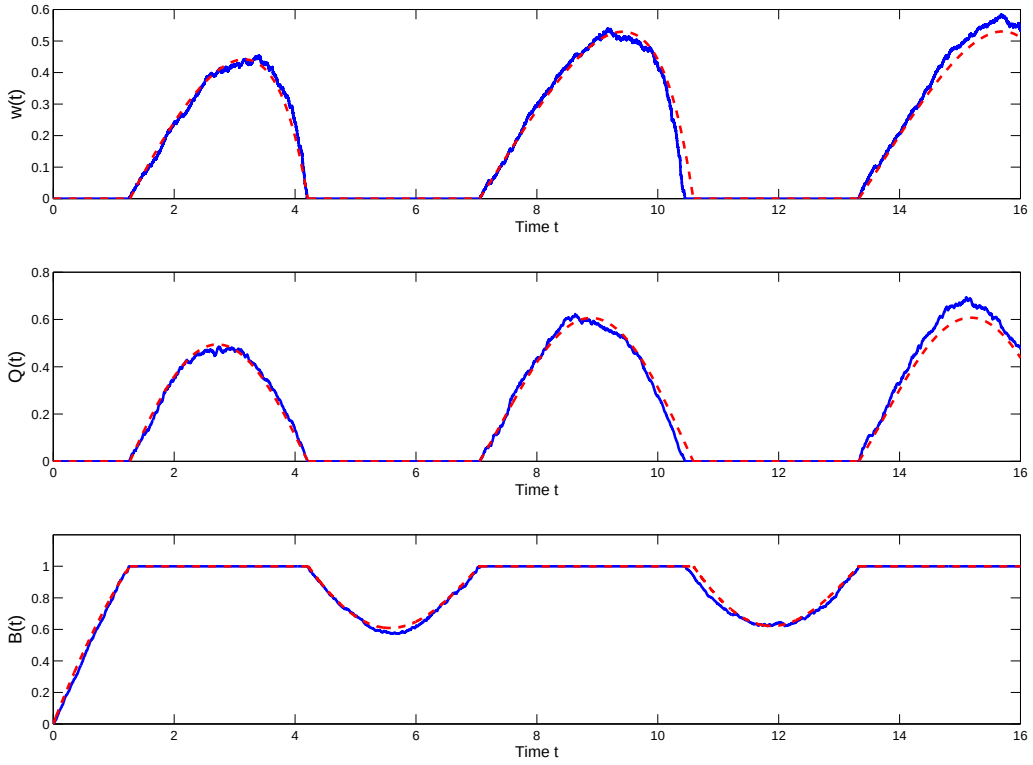

Figure A.18: The $G_t/M/s + M$ fluid model compared with simulations of the queueing system.

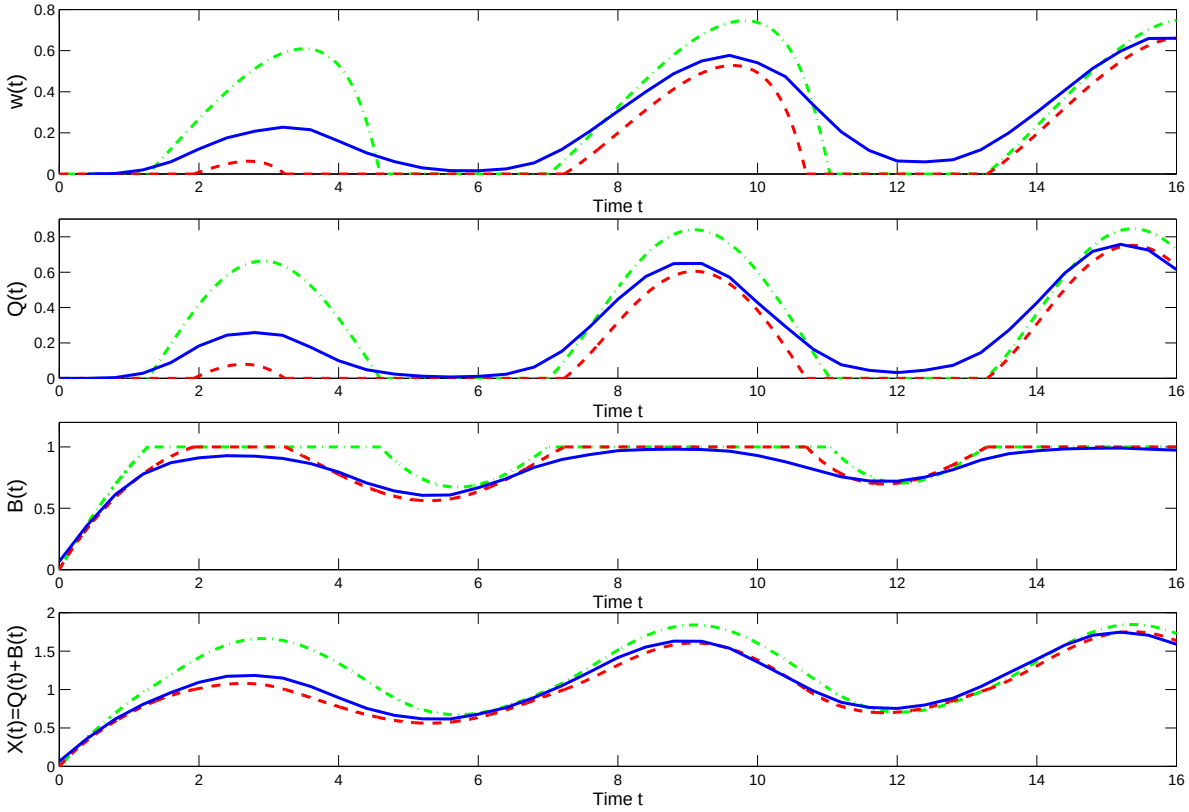

Figure A.19: Simulation comparison for the $M_t/H_2/s + E_2$ fluid model: (i) simulation estimates of an average of 500 sample paths of the scaled queueing model based on $n = 15$ (blue solid lines), (ii) fluid functions for H_2 service (red dashed lines) and (iii) fluid functions assuming M service (green dashed lines).

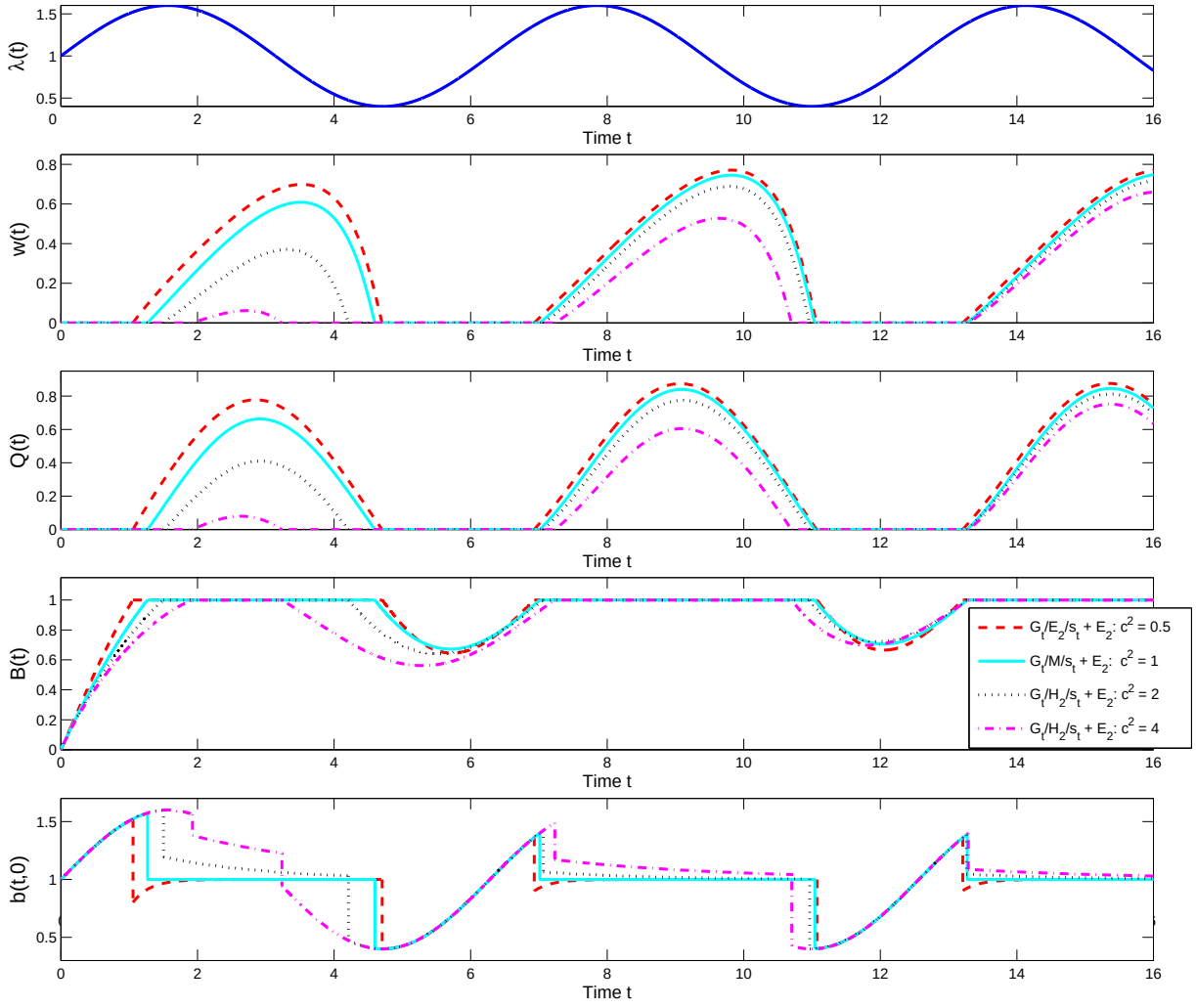

Figure A.20: Fluid dynamics of the $G_t/GI/s + E_2$ model with fixed mean service time and $E2$ patience distribution. The service distributions are: (i) $E2$ ($CVS = 0.5$); (ii) M ($CVS = 1$); (iii) $H2$ ($CVS = 2$) and (iv) $H2$ ($CVS = 4$).

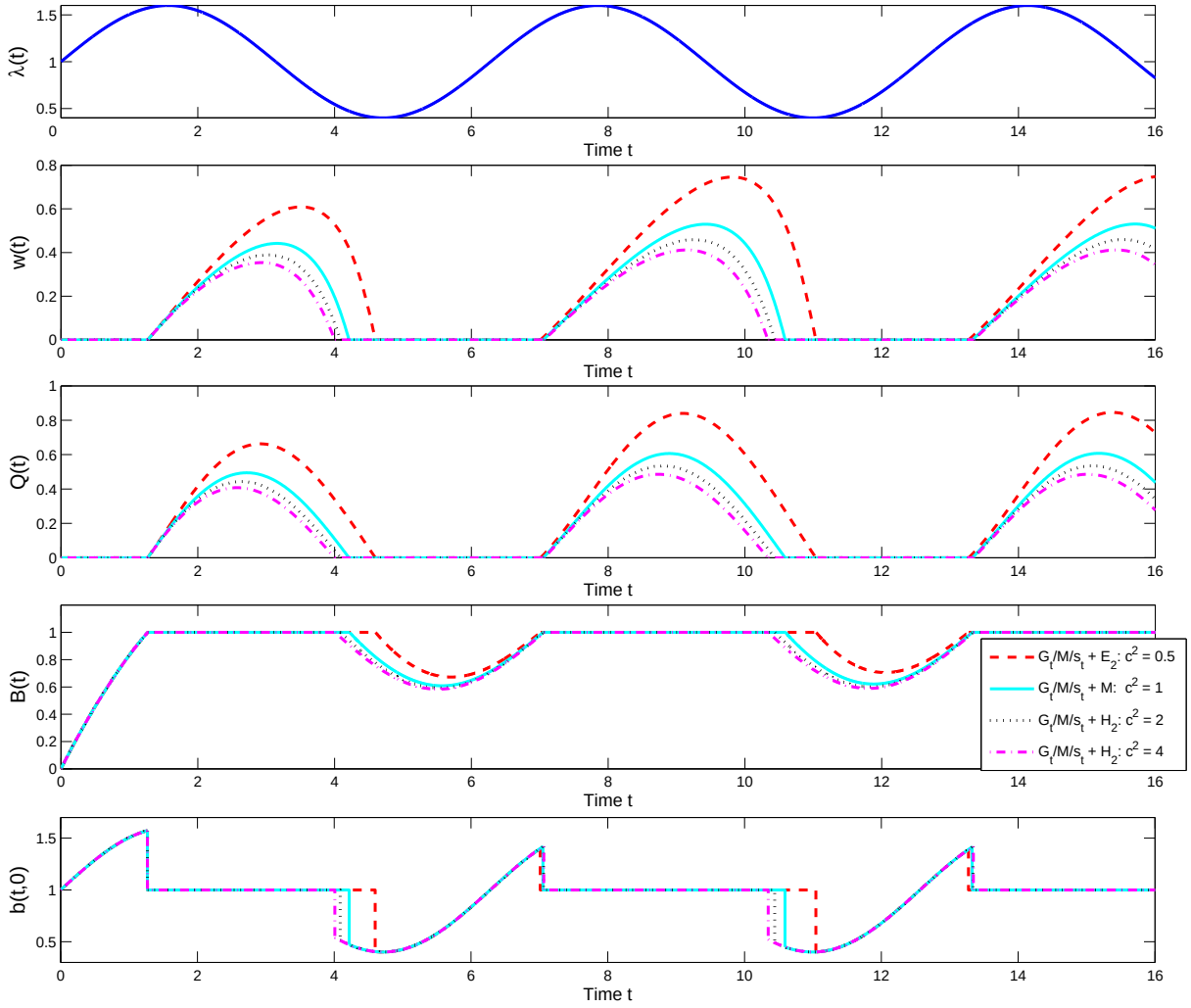

Figure A.21: Fluid dynamics of the $G_t/M/s + GI$ model with fixed mean patience time and M service distribution. The patience distributions are: (i) $E2$ ($CVS = 0.5$); (ii) M ($CVS = 1$); (iii) $H2$ ($CVS = 2$) and (iv) $H2$ ($CVS = 4$).

Appendix B

Appendix for Chapter 3

This e-companion has six sections, presenting supporting material primarily in the order that it relates to Chapter 3. In §B.1 we present the proofs for §3.3. In §B.2 we present proofs for §3.4. In §B.3 we present proofs for §3.5. In §B.4 we present one proof for §3.6. In §B.5, we make remarks about: (i) characterizing the isolated underloaded points in §3.3, (ii) representation of the fluid content B in an underloaded interval via an ODE, and (iii) the applied significance of the space of piecewise polynomials $\mathcal{P}_{m,n}$.

B.1 Proofs for §3.3.

We need some basic regularity properties of Q and B, which will be valid with the assumptions in §3.2. For that purpose, we exploit two basic *flow-conservation equations*: (i) the queue content at time t equals the initial queue content plus input minus output to either abandonment or entering service, and (ii) the service content at time t equals the initial service content plus input minus output. However, the input enters the queue only when

the system is overloaded; otherwise it directly enters service. Thus we have the following elementary bounds and the subsequent Lipschitz continuity.

Proposition B.1 (elementary bounds) $Q(t) + A(t) + E(t) \leq Q(0) + \Lambda(t) < \infty$ *and*

$$B(t) + S(t) = B(0) + E(t) \leq B(0) + Q(0) + \Lambda(t) < \infty,$$

so that Q*,* E*,* A*,* B *and* S *are all bounded for* $0 \leq t \leq T$*.*

Proof. The relations follow from flow conservation. The first relation is an inequality instead of an equality because input enters the queue instead of the service facility only when the system is overloaded. ∎

Proposition B.2 (Lipschitz continuity) *The functions* S*,* E*,* B*,* A *and* Q *are Lipschitz continuous.*

Proof. For a nonnegative real-valued function f on $[0, \infty)$, let $f_t^{\uparrow} \equiv \sup_{0 \leq y \leq t} f(y)$. To treat S, recall that S is the integral of σ, where

$$\sigma(t) = B(t)\mu(t) \leq s(t)\mu(t), \quad \text{so that} \quad \sigma(t) \leq s_t^{\uparrow}\mu_t^{\uparrow}, \quad t \geq 0, \tag{B.1}$$

and

$$|S(t+u) - S(t)| = \int_t^{t+u} \sigma(y)\, dy \leq s_T^{\uparrow}\mu_T^{\uparrow}u, \quad 0 \leq t \leq t+u \leq T. \tag{B.2}$$

To treat E, recall that it is the integral of the rate fluid enters service, where the rate fluid enters service is either $\gamma(t) = \lambda(t)$ if the system is underloaded or $\gamma(t) = s'(t) + \sigma(t) =$

$s'(t) + s(t)\mu(t)$ if the system is overloaded. Hence,

$$|E(t+u) - E(t)| \le \gamma_T^{\uparrow} u, \quad 0 \le t \le t+u \le T, \tag{B.3}$$

where $\gamma_T^{\uparrow} \equiv \lambda_T^{\uparrow} \vee (|s'|_T^{\uparrow} + s_T^{\uparrow}\mu_T^{\uparrow}) < \infty$. By the second equation in Proposition B.1,

$$B(t+u) - B(t) = (E(t+u) - E(t)) - (S(t+u) - S(t)), \tag{B.4}$$

so that

$$|B(t+u) - B(t)| \le |E(t+u) - E(t)| + |S(t+u) - S(t)| \le (e_T^{\uparrow} + s_T^{\uparrow}\mu_T^{\uparrow})u \tag{B.5}$$

for $0 \le t \le t+u \le T$.

Next we combine (3.3) with (3.8) to get

$$\alpha(t) = \int_0^{t \wedge w(t)} \lambda(t-x) f_{t-x}(x)\, dx + \int_{w(t)\wedge t}^{t} \frac{q(0, x-t) f_{t-x}(x)}{\bar{F}_{t-x}(x-t)}\, dx, \tag{B.6}$$

so that, by applying Assumption 3.6, we get

$$\alpha(t) \le \alpha_t^{\uparrow} \equiv f^{\uparrow}\Lambda(t) + \frac{f^{\uparrow}}{\bar{F}^{\downarrow}(w(0))} Q(0) < \infty \tag{B.7}$$

and

$$|A(t+u) - A(t)| \le \int_t^{t+u} \alpha(y)\, dy \le \alpha_T^{\uparrow} u, \quad 0 \le t \le t+u \le T. \tag{B.8}$$

Finally, by the first relation in Proposition B.1,

$$\begin{aligned} |Q(t+u) - Q(t)| &\leq |\Lambda(t+u) - \lambda(t)| + |E(t+u) - E(t)| + |A(t+u) - A(t)| \\ &\leq (\lambda_T^{\uparrow} + \gamma_T^{\uparrow} + \alpha_T^{\uparrow})u, \quad 0 \leq t \leq t+u \leq T. \ \blacksquare \end{aligned} \tag{B.9}$$

We now apply Proposition B.2 to relate $\mathcal{S}$ to the zeros of $X - s$, where $X(t) \equiv Q(t) + B(t)$.

Lemma B.1 (zeros of $X - s$) $\mathcal{S} \subseteq Z_{X-s}$.

Proof. Since Q and B are continuous by Proposition B.2 and s is continuous by assumption, $X - s$ is continuous. Since $X - s$ is continuous, if $X(t) - s(t) \neq 0$, then t cannot be an element of $\mathcal{S}$. ∎

We now characterize the overloaded times.

Lemma B.2 (overloaded intervals) *With the possible exception of* 0 *and* T*, all overloaded times appear in intervals of positive length. Hence, underloaded sets consist of either single isolated points or intervals.*

Proof. If $t \in \mathcal{O}([0,T])$, then either (i) $X(t) - s(t) > 0$ or (ii) $X(t) - s(t) = 0$ and $\zeta(t) > 0$. In case (i), since $X - s$ is continuous by Proposition B.2, there exists a neighborhood of t that is overloaded. In case (ii), since $\zeta(t) > 0$, we will have $X(t) - s(t) > 0$ in an interval $(t, t+\epsilon)$ for some positive ϵ. Since overloaded sets are necessarily intervals by Lemma B.2, each underloaded set must fall between two overloaded intervals. ∎

Proof of Theorem 3.1. We apply the results above. Since there can be at most countably many overloaded intervals of positive length in $[0,T]$, the isolated points are well defined

and countably infinite. Since the isolated points are at most countably infinite, we can order them and reclassify them one by one. With that construction, we reduce the number of disjoint overloaded intervals by one at each step. Finally, all underloaded times appear in intervals too. ∎

We now relate the zeros of ζ in (3.12) to the overloaded and underloaded intervals.

Lemma B.3 (zeros and intervals) *For each interval in the partition of* $[0, T]$ *into underloaded and overloaded intervals, there exists at least one zero or discontinuity point of* ζ.

Proof. First, consider the closure of an overloaded interval $[a, b]$. If ζ has one of its finitely many discontinuity points in $[a, b]$, then we are done. Suppose that ζ is continuous on the closed interval $[a, b]$. Necessarily, we have $X(a) - s(a) = X(b) - s(b) = 0$, $\zeta(a + \epsilon) > 0$ for all suitably small $\epsilon > 0$ and $\zeta(b) \leq 0$. First, we could have $\zeta(b) = 0$ and we are done. If instead $\zeta(U(t)) < 0$, then there must exist t^* with $a < t^* < b$ such that $\zeta(t^*) = 0$ by the intermediate value theorem. The reasoning is essentially the same in the closure of an underloaded interval, say $[a, b]$. If ζ has one of its finitely many discontinuity points in $[a, b]$, then we are again done. Suppose that ζ is continuous on the closed interval $[a, b]$. If either $\zeta(a) = 0$ or $\zeta(b) = 0$, then we are done. Hence we must have $\zeta(a) < 0$. Since b is a switch point and ζ is continuous at b, we must have $\zeta(b) > 0$. As before, there must exist t^* with $a < t^* < b$ such that $\zeta(t^*) = 0$ by the intermediate value theorem. ∎

Proof of Theorem 3.2 Since the interval $[0, T]$ can be partitioned into at most countably many intervals that alternate between overloaded and underloaded after reclassifying isolated underloaded points as overloaded, the switch points can be placed in one-to-one correspondence with the internal boundary points (excluding 0 and T). Hence the number

of switch points is equal to $n-1$, if the number of intervals in the paritition is n for some $n < \infty$. Otherwise both sets are countably infinite. Next, Lemma B.3 implies that there is either a discontinuity point or a zero in every overloaded and underloaded interval. Since the number of intervals is 1 greater than the number of switches, we obtain the conclusion. To see that the bound is tight, consider the common case in which ζ is differentiable on $[0,T]$ and $\zeta(t) \neq 0$ at all switch times. Then ζ has a zero where it attains its maximum in each overloaded interval, while ζ has a zero where it attains its minimum in each overloaded interval. To have the bound an equality, let ζ have no other zeros. ■

Proof of Theorem 3.3. First, any discontinuity points of ζ must be contained in the set of n interval boundary points. Hence, $\mathcal{D}_\zeta \leq n$. On each of the n subintervals, ζ is a polynomial of order at most m. By the fundamental theorem of algebra, on each of these intervals the zero set is either a finite set of cardinality at most m or it is the entire subinterval. If $\zeta = 0$ throughout the interval, then there can be at most a single switch in the interval, where $(Q(t), B(t))$ becomes $(0, s(t))$, after which it will remain there throughout the subinterval. In other words, the first subinterval is overloaded and the second is underloaded, so this interval produces at most a single switch. We can thus treat this interval just like any of the others; we can act as if it produces at most m zeros. Hence, $\mathcal{D}_\zeta \leq n$ and $Z_\zeta \leq mn$. Finally, Theorem 3.2 implies that $|\mathcal{S}| \leq mn + n - 1$, as claimed. ■

Proof of Lemma 3.1. The Weierstrass approximation theorem implies that continuous functions can be approximated uniformly over bounded intervals by polynomials. That uniform approximation extends to $\mathbb{C}_p$ provided that the boundary points of the polynomial pieces of the function in $\mathcal{P}_{m,n}$ includes the finitely many discontinuity points of the function in $\mathbb{C}_p$. ■

B.2 Proofs for §3.4.

B.2.1 Proof of Uniqueness in Theorem 2.3.

When the abandonment cdf's F_t are independent of t, the proof of uniqueness of the solution to the ODE (2.31) in Theorem 2.3 is the same as the proof of the corresponding part of Theorem 5.3 in Chapter 2. However, that argument does not extend directly to time-varying abandonment cdf's. Hence we give a different proof under different conditions. In particular, in Theorem 2.3 for time-varying abandonment cdf's we imposed additional regularity conditions. With those extra regularity conditions, we can apply the classical Picard-Lindelöf theorem for the uniqueuenss of a solution to the ODE $w'(t) = \Psi(t, w(t))$, which requires that $\Psi(t, x)$ be locally Lipschitz in the argument x uniformly in the argument t; e.g., Theorem 2.2 of [69].

One regularity condition added in Theorem 2.3 was for the rate fluid enters service to be bounded below. We will show how to guarantee that condition in the next section. Given that the rate fluid enters service is indeed bounded below, i.e., given that $\gamma(t) \geq e_L > 0$ for all $t \in [0, T]$, from (2.31), there exists a constant $w_L > 0$ such that $w'(t) \leq 1 - w_L < 1$ for all $t \in [0, T]$. Since $w(0) < \infty$, by assumption, and $w(t) \leq w(0) + t$ for all t, we have $w(t) \leq w(0) + T$ for $0 \leq t \leq T$. Together with the fact that $\lambda, q(0, \cdot) \in C_p$, that implies that the denominator in (2.31) is bounded above.

Since $w'(t) \leq 1 - w_L < 1$ for all t, for each x we will have $t - w(t) = x$ for at most one value of t. Since $\lambda, q(0, \cdot)$ have been assumed to have bounded derivatives where they are continuous, and since the partial derivative $\partial F_t(x)/\partial t$ of the time-varying abandonment cdf F_t as been assumed to be bounded, the mapping Ψ in (2.31) is Lipschitz continuous in the argument x except at only finitely many x, uniformly in t. Hence, we can deduce

uniqueness of the solution of the ODE in (2.31) under these extra regularity conditions by applying the Picard-Lindelöf theorem.

We now elaborate on the details. Here we have

$$\Psi(t,x) \equiv 1 - \frac{\gamma(t)}{\tilde{q}(t,x)} = 1 - \frac{\mu(t)s(t) + s'(t)}{\tilde{q}(t,x)}, \tag{B.10}$$

where $\tilde{q}(t,x)$ is given in (3.14). Consider the region $0 \le x_1 \le t$, $0 \le x_2 \le t$. In this region we have

$$\begin{aligned}
& |\Psi(t,x_1) - \Psi(t,x_2)| \\
= \quad & \frac{\mu(t)s(t) + s'(t)}{\lambda(t-x_1)\lambda(t-x_2)\bar{F}_{t-x_1}(x_1)\bar{F}_{t-x_2}(x_2)} |\lambda(t-x_1)\bar{F}_{t-x_1}(x_1) - \lambda(t-x_2)\bar{F}_{t-x_2}(x_2)| \\
\le \quad & \frac{\mu^{\uparrow}s^{\uparrow} + s'^{\uparrow}}{(\lambda^{\downarrow})^2(\bar{F}^{\downarrow})^2} |\lambda(t-x_1)\bar{F}_{t-x_1}(x_1) - \lambda(t-x_2)\bar{F}_{t-x_1}(x_1) \\
+ \quad & \lambda(t-x_2)\bar{F}_{t-x_1}(x_1) - \lambda(t-x_2)\bar{F}_{t-x_2}(x_2)| \\
\le \quad & \frac{\mu^{\uparrow}s^{\uparrow} + s'^{\uparrow}}{(\lambda^{\downarrow})^2(\bar{F}^{\downarrow})^2} (|\lambda(t-x_1) - \lambda(t-x_2)| + \lambda(t-x_2)|\bar{F}_{t-x_1}(x_1) - \bar{F}_{t-x_2}(x_2)|) \\
\le \quad & \frac{\mu^{\uparrow}s^{\uparrow} + s'^{\uparrow}}{(\lambda^{\downarrow})^2(\bar{F}^{\downarrow})^2} (\lambda'^{\uparrow}|x_1 - x_2| + \lambda^{\uparrow}|\bar{F}_{t-x_1}(x_1) - \bar{F}_{t-x_1}(x_2) + \bar{F}_{t-x_1}(x_2) - \bar{F}_{t-x_2}(x_2)|) \\
\le \quad & \frac{\mu^{\uparrow}s^{\uparrow} + s'^{\uparrow}}{(\lambda^{\downarrow})^2(\bar{F}^{\downarrow})^2} (\lambda'^{\uparrow}|x_1 - x_2| + \lambda^{\uparrow}\frac{\partial \bar{F}^{\uparrow}}{\partial t}|x_1 - x_2| + \lambda^{\uparrow}g^{\uparrow}|x_1 - x_2|) \\
\equiv \quad & C\,|x_1 - x_2|,
\end{aligned}$$

where $C \equiv \frac{\mu^{\uparrow}s^{\uparrow} + s'^{\uparrow}}{(\lambda^{\downarrow})^2(\bar{F}^{\downarrow})^2}(\lambda'^{\uparrow} + \lambda^{\uparrow}\frac{\partial \bar{F}}{\partial t}^{\uparrow} + \lambda^{\uparrow}g^{\uparrow})$. The case $x_1, x_2 > t$ is similar. Hence the regularity conditions given in Theorem 2.3 are sufficient for Ψ to be locally Lipschitz in x uniformly in t.

B.2.2 e_L-Feasibility of the Staffing Function s.

We have two goals in this section: first, to prove Theorem 2.7, showing how to construct the minimum feasible staffing function greater than or equal to any proposed staffing function s and, second, to determine the minimum feasible staffing function such that the rate fluid enters service at time t, $\gamma(t)$, is bounded below. We use this stronger notion of feasibility to provided conditions for the ODE in (2.31) in Theorem 2.3 to have a unique solution. We treat both problems at once by introducing the notion of e_L-feasibility: A staffing function s is said to be e_L-feasible if $\gamma(t) \geq e_L \geq 0$ for all $t \in [0, T]$.

So far, we have assumed that the staffing function s is e_L-feasible (as one condition in Theorem 2.3) or simply feasible (e_L-feasible for $e_L \equiv 0$), yielding

$$\gamma(t) \geq s'(t) + \sigma(t) = s'(t) + \int_0^\infty b(t,x) h_G(x)\, dx \geq e_L \geq 0 \quad \text{when} \quad B(t) = s(t). \tag{B.11}$$

This requirement is automatically satisfied in underloaded intervals when $B(t) = s(t)$, provided that $\lambda_{inf}(T) \geq e_L$ for λ_{inf} in Assumption 2.10, because in that case we require that $s'(t) + \sigma(t) \geq \lambda(t)$ where necessarily $\lambda(t) \geq e_L$; see Definition 3.1; e_L-Feasibility is only a concern during overloaded intervals, and then only when the staffing function is decreasing, i.e., when $s'(t) < 0$.

A violation is easy to detect; it necessarily occurs in an overloaded interval in $\mathcal{O}([0, T])$ at time $t^* \equiv \inf\{t \in \mathcal{O}([0, T]) : \gamma(t) < e_L\}$. Paralleling Chapter 2, let $\mathcal{S}_{f,s,e_L}$ be the set of e_L-feasible staffing functions over the interval $[0, t]$ for $t > t^*$. Then

$$t^* \equiv t^*(e_L) \equiv \inf\{t \in I : \gamma(t) < e_L\}. \tag{B.12}$$

Even though we require (B.11), so far we have done nothing to prevent having $t^* < \infty$ (violation). Thus, we compute γ and detect the first violation.

Correcting the staffing function is not difficult either (by which we mean replacing it with a higher feasible staffing function): We simply construct a new staffing function s^* consistent with reducing the input into the queue to its minimum allowed level (setting $\gamma(t) = e_L \geq 0$) starting at time t^* and lasting until the first time t after t^* at which $s^*(t) = s(t)$. (By the adjustment, we will have made $s^*(t^*+) > s(t^*+)$.) Since the system has operated differently during the time interval $[t^*, t]$, we must recalculate all the performance measures after time t, but we have now determined a feasible staffing function up to time $t > t^*$. By successive applications of this correction method (adjusting the staffing function s and recalculating b), we can construct the minimum feasible staffing function overall.

To make this precise, let $\mathcal{S}_{f,s,e_L}(t)$ be the set of all e_L-feasible staffing functions for the system over the time interval $[0, t]$, $t > t^*$, that coincide with s over $[0, t^*]$; i.e., let

$$\mathcal{S}_{f,s,e_L}(t) \equiv \{\tilde{s} \in C_p^1(t) : \gamma_{\tilde{s}}(u)1_{\{B_{\tilde{s}}(u)=\tilde{s}(u)\}} \geq e_L, \quad 0 \leq u \leq t, \quad \tilde{s}(u) = s(u), \quad 0 \leq u \leq t^*\}, \tag{B.13}$$

for t^* in (B.12), where $\gamma_{\tilde{s}}$ and $B_{\tilde{s}}$ are the functions γ and B associated with the model with staffing function $\tilde{s}$.

Theorem B.1 (minimum e_L-feasible staffing function) *For each e_L such that $0 \leq e_L \leq \lambda_{inf}(T)$ for $\lambda_{inf}(T)$ in Assumption* 2.10*, there exist $\delta \equiv \delta(e_L)$ and $s^* \in \mathcal{S}_{f,s,e_L}(t^* + \delta)$ in* (B.13) *for t^* in* (B.12) *such that*

$$s^* \equiv s^*(e_L) = \inf\{\tilde{s} \in \mathcal{S}_{f,s,e_L}(t^* + \delta)\}; \tag{B.14}$$

i.e., $s^* \in \mathcal{S}_{f,s,e_L}(t^*+\delta)$ *and* $s^*(u) \le \tilde{s}(u)$, $0 \le u \le t^*+\delta$, *for all* $\tilde{s} \in \mathcal{S}_{f,s,e_L}(t^*+\delta)$. *In particular,*

$$s^*(t^*+u) = e_L \int_0^u e^{-M(t^*+u-x,t^*+u)}\,dx + B(t^*)\,e^{-M(t^*,t^*+u)}. \tag{B.15}$$

Moreover, δ *can be chosen so that*

$$\delta = \inf\{u \ge 0 : s^*(t^*+u) = s(t^*+u)\}, \tag{B.16}$$

with $\delta \equiv \infty$ *if the infimum in* (B.16) *is not attained.*

Proof. First, since γ_s is continuous for our original s, the violation in (B.12) must persist for a positive interval after t^*; that ensures that a strictly positive δ can be found. We shall prove that $\tilde{s} \ge s^*$ over $[t^*, t^*+\delta]$ for s^* in (B.15) and any feasible function $\tilde{s}$, and we will show that s^* itself is feasible. For $0 \le t \le t^*+\delta$, suppose $\tilde{s}$ is feasible. Since the system is overloaded, system being in the overloaded regime implies that

$$\begin{aligned}
\tilde{s}(t^*+u) &= B_{\tilde{s}}(t^*+u) = \int_0^\infty b_{\tilde{s}}(t^*+u,x)dx \\
&= \int_0^u \gamma_{\tilde{s}}(t^*+u-x)\,\bar{G}_{t^*+u-x}(x)dx + \int_u^\infty b_{\tilde{s}}(t^*,x-u)\frac{\bar{G}_{t^*+u-x}(x)}{\bar{G}_{t^*+u-x}(x-u)}dx \\
&= \int_0^u \gamma_{\tilde{s}}(t^*+u-x)\,e^{-M(t^*+u-x,t^*+u)}\,dx + \int_u^\infty b_s(t^*,x-u)\,e^{-M(t^*,t^*+u)}dx \\
&\ge e_L \int_0^u e^{-M(t^*+u-x,t^*+u)}\,dx + e^{-M(t^*,t^*+u)} \int_0^\infty b_s(t^*,y)dy = s^*(t^*+u).
\end{aligned}$$

where the second equality holds because of the fundamental evolution equations in Assumption 2.6, the third equality holds because $b_{\tilde{s}}(t^*,x) = b_s(t^*,x)$ for all x, and the in-

equality holds because $\gamma_{\tilde{s}} \geq e_L$. On the other hand, the equality holds when $\gamma_{\tilde{s}}(t^*+u) = e_L$ for all u, which yields $B(t^*+u) = s^*(t+u)$. Therefore, the proof is complete. ∎

Corollary B.1 (minimum e_L-feasible staffing with exponential service times) *For the special case of exponential service times, i.e., with $\bar{G}(x) \equiv e^{-\mu x}$, independent of t,* (B.15) *becomes simply $s^*(t^*+u) = e_L(1-e^{-\mu u})/\mu + B(t^*)e^{-\mu u}$, $0 \leq u \leq \delta$.*

B.3 Proofs for §3.5.

B.3.1 Proof of Theorem 3.5.

First, the assumption that $\zeta_1, \zeta_2 \in \mathcal{P}_{m,n}$ assures that there are only finitely many switches between overloaded intervals and underloaded intervals in both systems. That leads to three cases: (i) when both systems are underloaded, (ii) when the upper system is overloaded and the lower system is underloaded, and (iii) when both systems are overloaded. We apply mathematical induction over the successive alternating intervals of these three kinds. (The switch points are the union of the two separate sets of switch points.) We ensure that the initial conditions for each succeeding interval satisfy the initial ordering assumed in the theorem. If we start in an interval where both systems are underloaded, then the ordering holds while both systems are underloaded by virtue of the explicit representation in Proposition 3.1. Consequently, the underload termination times are ordered as well, by Proposition 3.1. The ordering $B_1(t) \leq B_2(t)$ necessarily remains valid when the upper system is overloaded and the lower system is underloaded, because then we have $B_1(t) \leq s(t) = B_2(t)$. For an interval where both systems are overloaded, it suffices to consider the two systems starting the first time both systems are overloaded. At that time, the initial

conditions necessarily will be ordered properly, because the system to become overloaded later has $Q_1(t) = 0$. At this initial time, $B_1(t) = B_2(t) = s(t)$.

The M_t service assumption comes to the fore in an interval where both systems are overloaded. Here we use the fact that σ and $\gamma(t) = b(t, 0)$ depend only upon s and μ during the overloaded interval, and so are the same for the two systems, because the functions s and μ have been assumed to be fixed. The rate of service completion is $\sigma(t) = s'(t) + s(t)\mu(t)$. When the two systems are both overloaded over a common interval $[t, t + u]$, the total fluid to enter service from queue, $E(t + u) - E(t)$ is therefore the same in the two systems.

When both systems are overloaded, we have the ordering $\tilde{q}_1 \leq \tilde{q}_2$ directly from Proposition 3.3, just as in Proposition 2.6 of Chapter 2, exploiting the representation

$$\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)} = e^{-\int_{x-t}^{x} h_{F_{t-x}}(y)\,dy}.$$

Hence, to show that $q_1 \leq q_2$, it suffices to show that $w_1 \leq w_2$, which would imply that that the overload termination times are ordered as well.

Suppose we start at t_1 with $w_1(t_1) \leq w_2(t_1)$. Suppose that $w_1(t) > w_2(t)$ at some $t > t_1$. The continuity of w_1 and w_2 implies that there exists some $t_1 < t_2 < t$ such that $w_1(t_2) = w_2(t_2) \equiv \tilde{w}$. However, the ordering of $\tilde{q}_1$ and $\tilde{q}_2$ implies that $\tilde{q}_1(t_2, \tilde{w}) \leq \tilde{q}_2(t_2, \tilde{w})$. Therefore, ODE (2.31) implies that $w_1'(t_2) \leq w_2'(t_2)$. This contradicts with our assumption that there exists a t such that $w_1(t) > w_2(t)$.

Now we turn to v. The equation (2.36) in Theorem 2.5 implies that the ordering of w is inherited by v. That is made clear by applying the proof of Theorem 2.5, which shows that $v(t)$ is determined by the intersection of the function w with the linear function $L_t(u) \equiv t + u$. Clearly, if we increase the w function, then that intersection point increases as well. ∎

B.3.2 Proof of Theorem 3.6.

We directly prove (3.18); the corresponding results in (3.19) will be obtained along the way. To show (i), consider two models with common model data except for $\lambda, B(0)$, where $\lambda_1, \lambda_2, s', \mu \in \mathcal{P}_{m,n}$ for some m, n. Without loss of generality, by Theorem 3.5, it suffices to assume that $\lambda_1 \leq \lambda_2$ and $B_1(0) \leq B_2(0)$. If that is not initially the case, consider $\tilde{\lambda}_1 \equiv \lambda_1 \wedge \lambda_2$, $\tilde{\lambda}_2 \equiv \lambda_1 \vee \lambda_2$, $\tilde{B}_1(0) \equiv B_1(0) \wedge B_2(0)$ and $\tilde{B}_2(0) \equiv B_1(0) \vee B_2(0)$ to get $\tilde{\lambda}_1 \leq \tilde{\lambda}_2$ and $\tilde{B}_1(0) \leq \tilde{B}_2(0)$ with $\|\tilde{\lambda}_1 - \tilde{\lambda}_2\|_T = \|\lambda_1 - \lambda_2\|_T$ and $|\tilde{B}_1(0) - \tilde{B}_2(0)| = |B_1(0) - B_2(0)|$.

When both systems are overloaded, we have $B_1(t) = B_2(t) = s(t)$. Hence, the overall story depends on what happens when (a) both systems are underloaded, and (b) system 1 is underloaded and system 2 is overloaded.

For simplicity, suppose that the two systems both start underloaded at time 0 with $B_1(0) \leq B_2(0)$, $\lambda_1 \leq \lambda_2$. If both systems remain underloaded over the interval $[0, t_1]$, then by Proposition 3.1 we have

$$\begin{aligned} |B_1(t) - B_2(t)| &\leq \|\lambda_1 - \lambda_2\|_T \int_0^t e^{-M(x)}\, dx + |B_1(0) - B_2(0)| \\ &\leq t \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)|, \quad 0 \leq t \leq t_1. \end{aligned} \tag{B.17}$$

Suppose system 2 becomes overloaded at $t_1 > 0$ while system 1 remains underloaded. For $t > t_1$, we have $B_1(t) \leq B_2(t) = s(t) \leq X_2(t) \equiv B_2(t) + s(t)$. Hence we have $0 \leq |B_2(t) - B_1(t)| = B_2(t) - B_1(t) \leq X_2(t) - B_1(t)$. Flow conservations of both systems implies that $B_1'(t) = \lambda_1(t) - \mu(t)\, B_1(t)$ and $X_2'(t) = \lambda_2(t) - \alpha_2(t) - \mu(t)\, s(t)$. Therefore,

$$X_2'(t) - B_1'(t) = \lambda_2(t) - \lambda_1(t) - \alpha_2(t) - \mu(t)\,(s(t) - B_1(t)) \leq \lambda_2(t) - \lambda_1(t),$$

which implies that

$$\begin{aligned} |B_1(t) - B_2(t)| &\leq |B_1(t_1) - B_2(t_1)| + (t - t_1) \cdot \|\lambda_1 - \lambda_2\|_T \\ &\leq t_1 \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)| + (t - t_1) \cdot \|\lambda_1 - \lambda_2\|_T \\ &\leq t \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)|, \end{aligned} \tag{B.18}$$

where the second inequality follows from (B.17) with $t = t_1$.

If we then later start a second underloaded interval for both systems at time t_2, where $0 < t_1 < t_2 < T$, then we will have inequality (B.17) holding at time t_2. Thus proceeding forward, applying (B.17) with initial values $B_i(t_2)$, during the following underloaded interval we have for $t > t_2$

$$\begin{aligned} |B_1(t) - B_2(t)| &\leq \|\lambda_1 - \lambda_2\|_T \int_{t_2}^{t} e^{-M(x)}\, dx + |B_1(t_2) - B_2(t_2)| \\ &\leq (t - t_2) \cdot \|\lambda_1 - \lambda_2\|_T + t_2 \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)| \\ &\leq t \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)| \\ &\leq (1 \vee t)(\|\lambda_1 - \lambda_2\|_T \vee |B_1(0) - B_2(0)|). \end{aligned} \tag{B.19}$$

where the second inequality follows from (B.18) with $t = t_2$. Applying mathematical induction over successive underloaded subintervals of $[0, T]$, using the second to last inequality, we obtain the first relation in (3.18), from which the desired conclusion follows.

To show (ii), when both systems are underloaded, we have $Q_1(t) = Q_2(t) = 0$. Hence, the overall story depends on what happens when (a) both systems are overloaded, and (b) system 1 is underloaded and system 2 is overloaded.

When both systems are overloaded, flow conservation implies that

$$Q_i'(t) = \lambda_i(t) - \alpha_i(t) - \gamma_i(t) = \lambda_i(t) - \alpha_i(t) - \mu(t)\, s(t) - s'(t).$$

Hence, we have

$$Q_2'(t) - Q_1'(t) = \lambda_2(t) - \lambda_1(t) - (\alpha_2(t) - \alpha_1(t)) \leq \lambda_2(t) - \lambda_1(t),$$

where the inequality simply follows from Theorem 3.5 when the two systems have common abandon-time distribution. This yields

$$|Q_1(t) - Q_2(t)| = Q_2(t) - Q_1(t) \leq |Q_1(0) - Q_2(0)| + t\, \|\lambda_1 - \lambda_2\|_T. \qquad \text{(B.20)}$$

When system 2 is overloaded and system 1 is underloaded. For simplicity, assume at time 0 the two system have initial conditions $B_2(0) = s(0) > B_1(0)$, $Q_2(0) \geq 0 = Q_1(0)$. Let $T^* \equiv T_1 \wedge T_2$, where T_1 denotes the underload termination time of system 1 and T_2 denotes the overload termination time of system 2. Hence we know that both systems will not change regimes for $0 \leq t \leq T^*$. For $0 \leq t \leq T^*$, we have

$$\begin{aligned} Q_2'(t) &= \lambda_2(t) - \alpha_2(t) - \gamma_2(t) \leq \lambda_2(t) - \gamma_2(t) \\ &\leq (\lambda_2(t) - \lambda_1(t)) + (\lambda_1(t) - \gamma_2(t)) \\ &\leq (\lambda_2(t) - \lambda_1(t)) + (\lambda_1(t) - \mu(t)\, s(t) - s'(t)), \end{aligned}$$

which implies that

$$\begin{aligned}
& |Q_2(t) - Q_1(t)| = Q_2(t) \\
\leq\ & Q_2(0) + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + \int_0^t \lambda_1(u) - \mu(u)\,s(u) - s'(u)du \\
\leq\ & Q_2(0) + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + \int_0^t \lambda_1(u) - \mu(u)\,B_1(u)du - (s(t) - s(0)) \\
\leq\ & Q_2(0) + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + \int_0^t B_1'(u)du - s(t) + s(0) \\
\leq\ & Q_2(0) + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + (s(0) - B_1(0)) - (s(t) - B_1(t)) \\
\leq\ & |Q_2(0) - Q_1(0)| + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + |B_2(0) - B_1(0)|, \qquad \text{(B.21)}
\end{aligned}$$

where the second inequality holds because $B_1(t) \leq s(t)$, the third inequality holds since $B_1'(t) = \lambda_1(t) - \mu(t)\,B_1(t)$, and the last inequality holds since $Q_1(0) = 0$, $B_2(0) = s(0)$ and $B_1(t) \leq s(t)$. Again, the desired conclusion follows by mathematical induction.

Finally, to show (iii), (B.18), (B.19), (B.20), (B.21) imply that

$$\begin{aligned}
|X_1(t) - X_2(t)| &\leq |B_1(t) - B_2(t)| + |Q_1(t) - Q_2(t)| \\
&\leq 2t\,\|\lambda_1 - \lambda_2\| + 2\,|B_1(0) - B_2(0)| + |Q_1(0) - Q_2(0)| \\
&\leq 2(1 \vee t)(\|\lambda_1 - \lambda_2\|_T \vee |X_1(0) - X_2(0)|),
\end{aligned}$$

where the third inequality holds because $|X_1(0) - X_2(0)| = |B_1(0) - B_2(0)| + |Q_1(0) - Q_2(0)|$ in all regimes. ∎

B.3.3 Proof of Theorem 3.7.

Given $\lambda \in \mathbb{C}_p$, we choose an increasing sequence $\{\lambda_k : k \geq 1\}$ with $\lambda_k \in \mathcal{P}_{m_k,n_k}$ for each $k \geq 1$ such that $\|\lambda_k - \lambda\|_T \to 0$ as $k \to \infty$. For each $k \geq 1$, we can apply all

the results above. By Theorem 3.6, we can define the pair (B, σ) in $\mathbb{C}_p^2$ as the limit of the sequence $\{(B_k, \sigma_k)$ in $\mathbb{C}_p^2$ with the maximum/uniform norm. There is such a limit, because the sequence is necessarily Cauchy and the space is a complete metric space. Given the limit, the convergence holds in the space by Theorem 3.6.

To show that the monotonicity extends, we start with $\lambda_1 \leq \lambda_2$. We then construct sequences $\{\lambda_{i,k} : k \geq 1\}$ for $i = 1, 2$ with $\lambda_{1,k} \leq \lambda_{2,k}$ for each k and $\|\lambda_{i,k} - \lambda_i\|_T \to 0$ as $k \to \infty$. We apply Theorem 3.5 for each k. Since the ordering is preserved in the limit, the conclusion of Theorem 3.5 holds for the limiting pair by Lebesgue monotone convergence. We use a similar argument to show that the Lipschitz continuity properties in Theorem 3.6 extend as well: Starting with $\|\lambda_1 - \lambda_2\|_T = c$, for any $\epsilon > 0$, we construct sequences $\{\lambda_{i,k} : k \geq 1\}$ for $i = 1, 2$ with $\|\lambda_{1,k} - \lambda_{2,k}\| \leq c + \epsilon$ for each k and $\|\lambda_{i,k} - \lambda_i\|_T \to 0$ as $k \to \infty$ for $i = 1, 2$. We then can apply Theorem 3.6 for each $k \geq 1$, and get the conclusion there with modification by ϵ. However, since ϵ is arbitrary, we get the preservation of the Lipschitz property to the limit. ∎

B.4 One proof for §3.6.

Proof of Theorem 3.9. We recursively apply the monotone contraction operator Ψ in Theorem 3.8, starting with $\sigma_{j,i}^{(0)} = 0$, so that $\lambda_{1,i}^{(1)} \leq \lambda_{2,i}^{(1)}$ for all i, because $\lambda_{j,i}^{(1)} = \lambda_{j,i}^{(0)}$, $j = 1, 2$ and the external arrival rate functions have been assumed to be ordered: $\lambda_{1,i}^{(0)} \leq \lambda_{2,i}^{(0)}$. By Theorem 3.5 applied to each queue separately, using the assumed ordering $B_{1,i}(0) \leq B_{2,i}(0)$ for all i, we have first $B_{1,i}^{(1)} \leq B_{2,i}^{(1)}$ and then $\sigma_{1,i}^{(1)} \leq \sigma_{2,i}^{(1)}$. By (3.23), we then have $\lambda_{1,i}^{(2)} \leq \lambda_{2,i}^{(2)}$. We then get the order holding for all n by applying mathematical induction. However, $\lambda_{1,i}^{(n)} \to \lambda_{1,i}$ as $n \to \infty$. Since the order is preserved in the convergence, we deduce that $\lambda_{1,i} \leq \lambda_{2,i}$ for $1 \leq i \leq m$. Finally, we can apply Theorem 3.5 to each queue separately to get the remaining orderings. ∎

B.5 Remarks

Remark B.1 $(\textit{characterization of isolated points})$

Definition 3.3 implies that t is an isolated point only if $Q(t) = 0$, $B(t) = s(t)$. Moreover, if t is a discontinuity point of ζ, then $\zeta(t-\delta) < 0$ and $\zeta(t) > 0$ for some $\delta > 0$; if t is a continuity point of ζ, then $\zeta(t-\delta) < 0$, $\zeta(t) = 0$ and $\zeta(t+\delta) < 0$ for some $\delta > 0$.

Remark B.2 $(\textit{an ODE for } B \textit{ in an underloaded interval})$

In an underloaded interval, the total fluid content in service $B(t)$ can also be characterized via the ODE

$$B'(t) = \lambda(t) - \mu(t)B(t), \quad t \geq 0. \tag{B.22}$$

The formula in Proposition 3.1 provides the solution to the initial value problem determined by this ODE with initial condition $B(0)$.

Remark B.3 *(applied significance of $\mathcal{P}_{mn}$) We have provided a full algorithm when $\lambda, s', \mu \in \mathcal{P}_{m,n}$. An algorithm for $\lambda \in \mathbb{C}_p$ can be developed by considering a sequence of successive approximations in $\mathcal{P}_{m_n,n}$, but we see no motivation for doing so. We have introduced the space $\mathcal{P}_{m,n}$ of piecewise polynomials as a device to establish mathematical results. In applications, it should suffice to use* any *convenient representations of the functions λ and s, and* assume *that there are only finitely many switches in any finite interval. While running the algorithm, that assumption can be verified, and the model can be modified if too many switches occur. However, if we start from data, then we could choose to let the functions be in $\mathcal{P}_{m,n}$ without loss of generality. Lemma 3.2 shows that it is convenient to work in*

the space $\mathcal{P}_{m,n}$, because we can obtain closed form expressions for integrals. Moreover, if we want to bound the number of switches in advance, then we can bound the parameters m and n, with the understanding that there is a tradeoff between the quality of fit and the maximum number of switches.

Appendix C

Appendix for Chapter 4

C.1 Overview.

This appendix contains additional supplementary material. In §C.2 we give a numerical example illustrating convergence to steady state for the stationary $G/M/s + M$ model starting empty. In §C.3 we give the other half of the proof of Theorem 4.4, establishing pointwise convergence of the fluid densities $b(t, x)$ and $q(t, x)$ as $t \to \infty$ when the system is initially OL. In §C.4 we give another example of periodic steady state (PSS) in a model with both sinusoidal arrival rate and staffing function, complementing Example 4.2. In §C.5 we verify the explicit formulas for the PSS in Example 4.3. In §C.6, we compare the fluid approximation to results from simulations of corresponding stochastic queueing models, for the example considered in §C.2. These simulation results substantiate that (i) the theorems are correct, (ii) the numerical algorithm is effective and (iii) the fluid approximation for the stochastic queueing system is effective. The fluid model accurately

describes single sample paths of very large queueing systems and accurately describes the mean values for smaller queueing systems, e.g., with 20 servers.

C.2 Convergence to Steady State in the $G/M/s+M$ Model

In this section we give a numerical example illustrating the convergence to steady state for a $G/M/s+M$ queue starting empty, as characterized by Corollary 4.2. Here we let $\mu = 1$, $\lambda = 1.5$, $s = 1$, $\theta = 0.5$. In Figure C.1, we show how performance functions (the solid lines) converge to their steady states (the dashed lines), applying the algorithm described in Chapter 2. Figure C.1 shows that $w(t)$, $Q(t)$, $B(t)$ and $b(t,0)$ quickly converge to their steady state values.

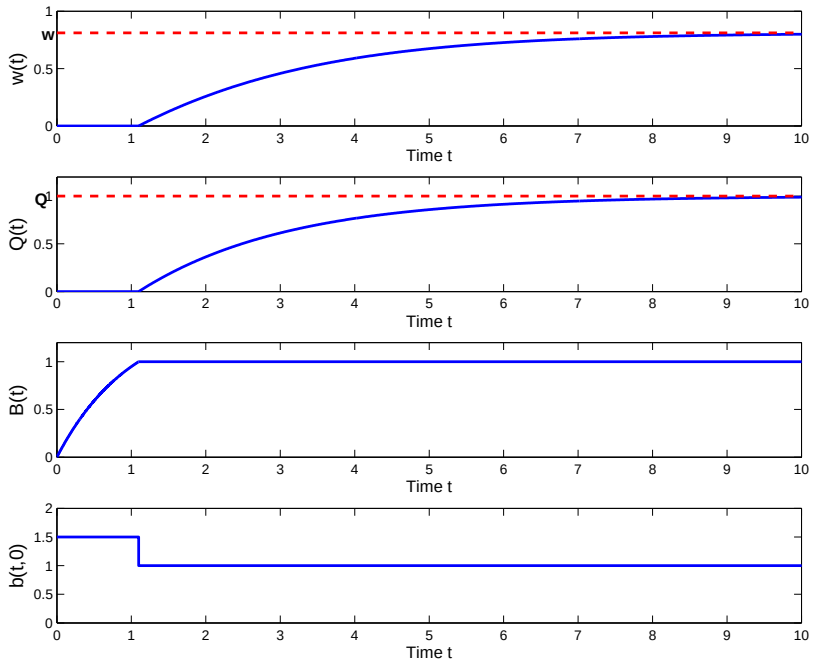

Figure C.1: Performance measures of the $G/M/s+M$ fluid queue converge to their steady states.

C.3 Proof of Theorem 4.4

Proof. We now complete the proof of Theorem 4.4 by proving (4.22) and (4.23) when the system is initially OL, i.e., $q(0,x) \geq 0$ for some x, $w(0) \geq 0$, $Q(0) \geq 0$ and $B(0) = s$. As before, for simplicity, we assume $\mu = s = 1$ and therefore $\rho = \lambda/s\mu = \lambda$.

(i) $\rho < 1$. Since the service is exponential at the fixed rate $\mu = 1$ and the staffing is fixed at $s = 1$, the output rate of the service facility is 1. Hence, $Q'(t) = \lambda - \alpha(t) - b(t,0) < \lambda - b(t,0) < 1$ as long as the system is in the OL regime; moreover, the OL regime will end after some $0 < T < 1/(1-\rho)$. The system will switch to the UL regime at T (i.e., $Q(T) = w(T) = 0$, $B(T) = s = 1$) and will stay there for all $t > T$. Thus we can apply (2.13) to characterize the density in service. By Assumption (4.24), for $t \geq T$,

$$\begin{aligned}
b(t,x) &= \rho e^{-x} 1_{\{0 \leq x \leq t-T\}} + b(T, x-t+T) e^{-(t-T)} 1_{\{x > t-T\}} \\
&= \rho e^{-x} 1_{\{0 \leq x \leq t-T\}} + b(0, x-t) e^{-t} 1_{\{x > t-T\}} \\
&\to \rho e^{-x} \quad as\ t \to \infty, \quad x \geq 0. \\
B(t) &= \int_0^{t-T} \rho e^{-x} dx + \int_{t-T}^{\infty} b(T, x-t+T) e^{-(t-T)} dx \\
&= \rho(1 - e^{-(t-T)}) + e^{-(t-T)} B(T) \to \rho, \quad as\ t \to \infty,
\end{aligned}$$

Moreover, $\sigma(t) = B(t) \to \rho$, as $t \to \infty$.

(ii) $\rho \geq 1$. As in case (i), the maximum output rate of the service facility is 1. Since $\rho \geq 1$, $\lambda \geq 1$, so that the the system necessarily will stay in the OL or CL regime forever. Since $b(t,0) = \sigma(t) = 1$, all old fluid will leave the queue after $T \equiv Q(0)/b(t,0) = Q(0)$. Therefore, for $t \leq T$, we have $q(t,x) = \rho \bar{F}(x) 1_{\{x \leq w(t) \wedge (t-T)\}} \to q(x) = \rho \bar{F}(x) 1_{\{x \leq w\}}$ if $w(t) \to w$ as $t \to \infty$.

If $w(T) < w$, the same reasoning in part (ii) of the proof in Chapter 4 implies that $w(t) \uparrow w$ monotonically after T. If $w(T) = w$, then from (4.31) we see that $w'(T) = 0$,

which implies that the system is already in steady state and thus will stays there forever. If $w(T) > w$, it is easy to see that $w'(t) = H(w(t)) < H(w) = 0$ for $t \geq T$, where $H(\cdot)$ is defined in (4.31). Therefore, $w(t)$ is decreasing (has negative derivative) as long as $w(t) > w$. To show that $w(t) \to w$ as $t \to \infty$, it remains to show that for any $\epsilon > 0$, there exits a t_ϵ such that $w(t) < w + \epsilon$ for any $t > t_\epsilon$. Because H is strictly decreasing in a neighborhood of w, we have $w'(t) = H(w(t)) \leq H(w + \epsilon) \equiv \delta(\epsilon) < H(w) = 0$, if $w(t) \geq w + \epsilon$. Therefore, the derivative of $w(t)$ is not only negative, but also bounded by $\delta(\epsilon) < 0$. So $w(t)$ will hit $w + \epsilon$ at least linearly fast with slope $\delta(\epsilon)$, i.e., for any $t \geq T + (w(T) - w - \epsilon)/|\delta(\epsilon)|$, we have $w(t) \leq w + \epsilon$. Therefore, we conclude that $w(t) \downarrow w$ as $t \to \infty$. All the other results follow from the same reasoning as in the proof in Chapter 4. □

C.4 Another Example of Periodic Steady State

We complement Example 4.2 by considering another value for the parameter γ in the sinusoidal staffing function in (4.42). Here we let $\gamma = 0.5$ instead of 2.0. That makes the model period 4π instead of π. Figure C.2) shows the performance functions.

C.5 Verifying the Sinusoidal PSS

We now verify the PSS for Example 4.3. To verify t_0 and t_1 in (4.46) and (4.47), we let $a = s = \mu = c = \theta = 1$, $b = 0.6$. For these parameters, we get $t_0 = 0.78$ and $t_1 = 3.15$ from (4.46) and (4.47). We apply the algorithm in Chapter 2 and plot the performance measures $w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$ in Figure C.3 for $0 \leq t \leq 3 \cdot 2\pi/c = 6\pi$ (three cycles) with the system initially critically loaded and arrival rate $\lambda(t) = a + b \cdot \sin(c(t + t_0))$ (see Plot 1 in Figure C.3 for the phase difference: $6.28 - 5.50 = 0.78 = t_0$).

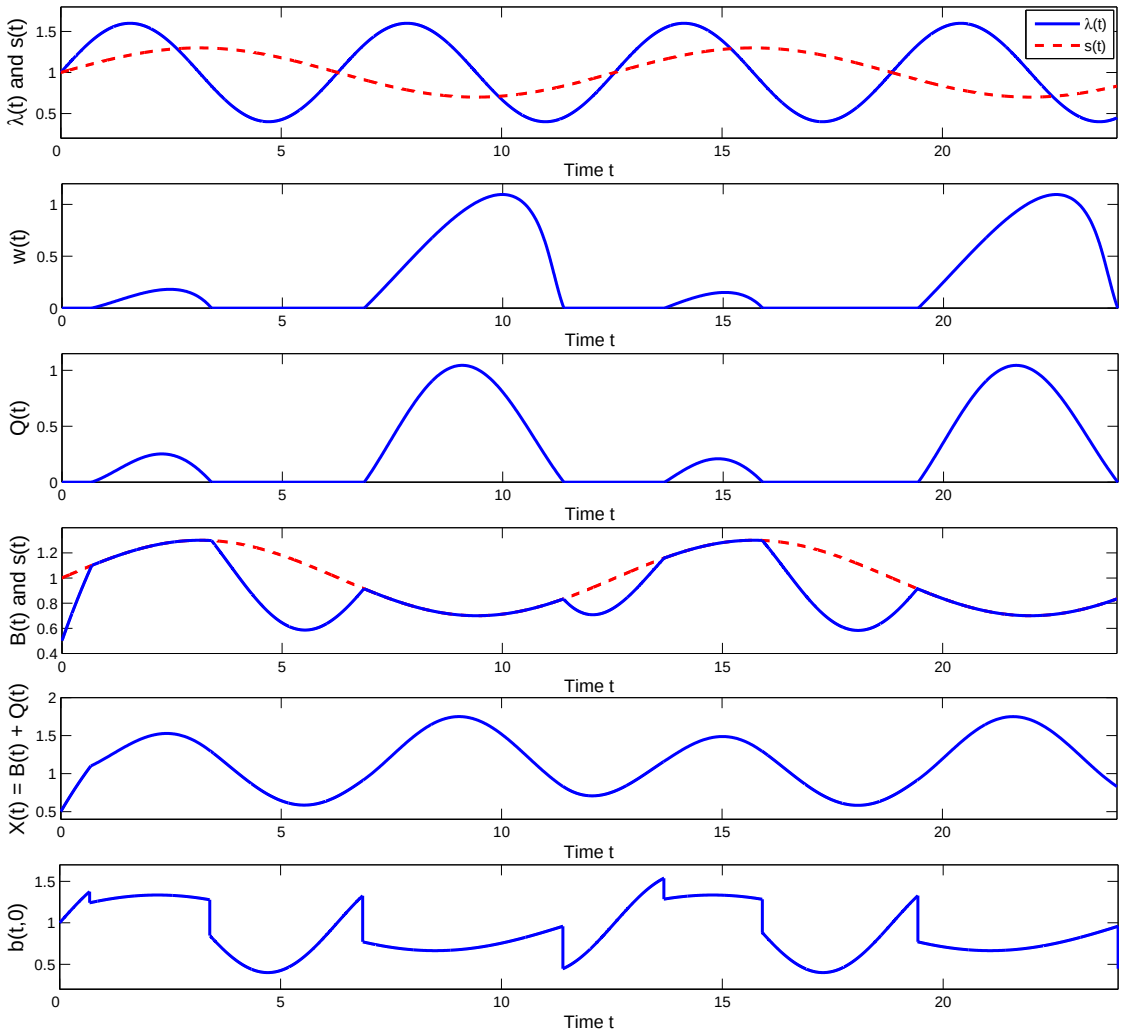

Figure C.2: Performance of the $G_t/M/s_t + M$ model with sinusoidal arrival and staffing, $\gamma = 0.5$.

Figure C.3 shows that the fluid performance immediately becomes stationary (a DSS cycle starts at time 0 and ends at 2π). Since the $M_t/M/s + M$ model here is equivalent to the $M_t/M/\infty$ model, we can also verify these analytical formulas by showing that they agree with previous ones derived for the $M_t/M/\infty$ model in [15].

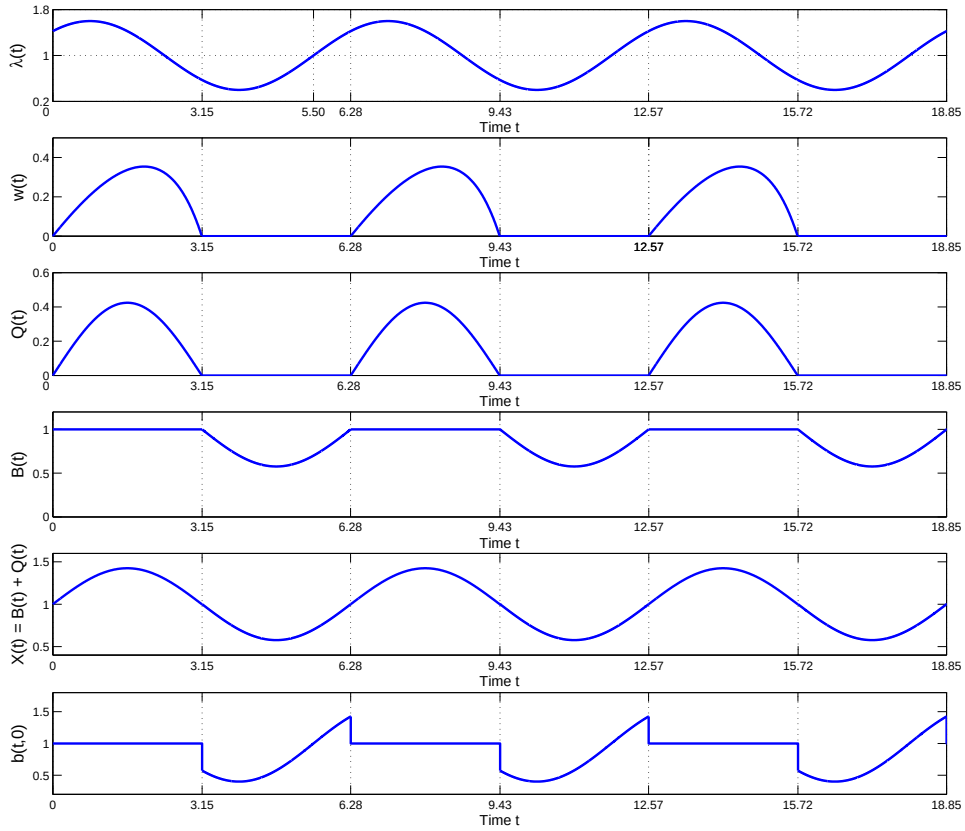

Figure C.3: The $G_t/M/s + M$ model in Example 4.3 is in PSS at time 0, with period $\tau = 2\pi = 6.28$. In each cycle $[n\tau, (n+1)\tau]$ of PSS, the system switches between UL and OL regimes twice at time $n\tau$ and $n\tau + 3.15$.

C.6 A Comparison with Simulation

In §C.4, we considered the $G_t/M/s_t + M$ fluid queue, which has a sinusoidal arrival rate $\lambda(t)$ as in (4.1) with $a = c = 1$, $b = 0.6$, sinusoidal staffing function $s(t)$ as in (4.42) with $\bar{s} = 1$, $u = 0.3$, $\gamma = 0.5$, exponential service and abandonment distributions with rate $\mu = 1$ and $\theta = 0.5$. We let the system be initially UL with $B(0) = 0.5 < s(0)$. We now compare the fluid approximation as shown in C.2 with computer simulations of the associated $M_t/M/s_t + M$ queueing model.

This queueing model has the same service and abandonment rates, but scaled arrival rate and number of servers: $n\,\lambda(t)$ and $n\,s(t)$. There are $n\,B(0)$ customers in service at

time 0. Let $W_n(t)$ be the elapsed waiting time of the customer at the head of the queue at t, $\tilde{Q}_n(t)$ be the number of customers in queue and $\tilde{B}_n$ be the number of customers in service. Applying the spatial scaling, we let $Q_n(t) \equiv \tilde{Q}_n(t)/n$ and $B_n(t) \equiv \tilde{B}_n(t)/n$. We let $X_n(t) \equiv Q_n(t)+B_n(t)$ be the scaled total number of customers in the system at t. In Figure C.4, C.5 and C.6, we compare the simulation results for the queue performance functions W_n, Q_n and B_n from a single simulation run to the associated fluid model counterparts w, Q and B, with $n = 30$, $n = 100$ and $n = 1000$. The blue solid lines represent the queueing model performance, while the red dashed lines represent the corresponding fluid performance. We observe that the bigger the scaling n is, the more accurate the fluid approximation becomes. When $n = 1000$, we have a large-scale queueing model (with arrival rate $1000 + 600\ \sin(t)$ and staffing $1000 + 300\ \sin(0.5\,t)$ servers) and we get close agreement for individual sample paths.

When n is smaller, there are bigger stochastic fluctuations as shown in Figures C.4 and C.5, but the mean values of the queueing functions still are quite well approximated by the fluid performance functions when the system is not nearly critically loaded. We illustrate by considering the cases $n = 100$ and $n = 30$ in Figures C.7 and C.8, where average sample paths of simulation estimates are compared with fluid approximations. In Figure C.7, we average 20 sample paths for $n = 100$; in Figure C.8, we average 200 sample paths for $n = 30$. We need more samples for smaller scaling n, because there are bigger fluctuations.

A careful examination of Figure C.7 and C.8 show that in both cases the total fluid content, $X(t)$, very accurately approximates the expected value of the scaled total number of customers, $X_n(t)$, in the queueing system. However, the fluid queue content $Q(t)$ and the fluid service content $B(t)$ do not approximate the mean values of their counterparts in the queueing system as well. In particular, the quality of these approximations degrades when the system is nearly critically loaded. That is understandable, because only positive

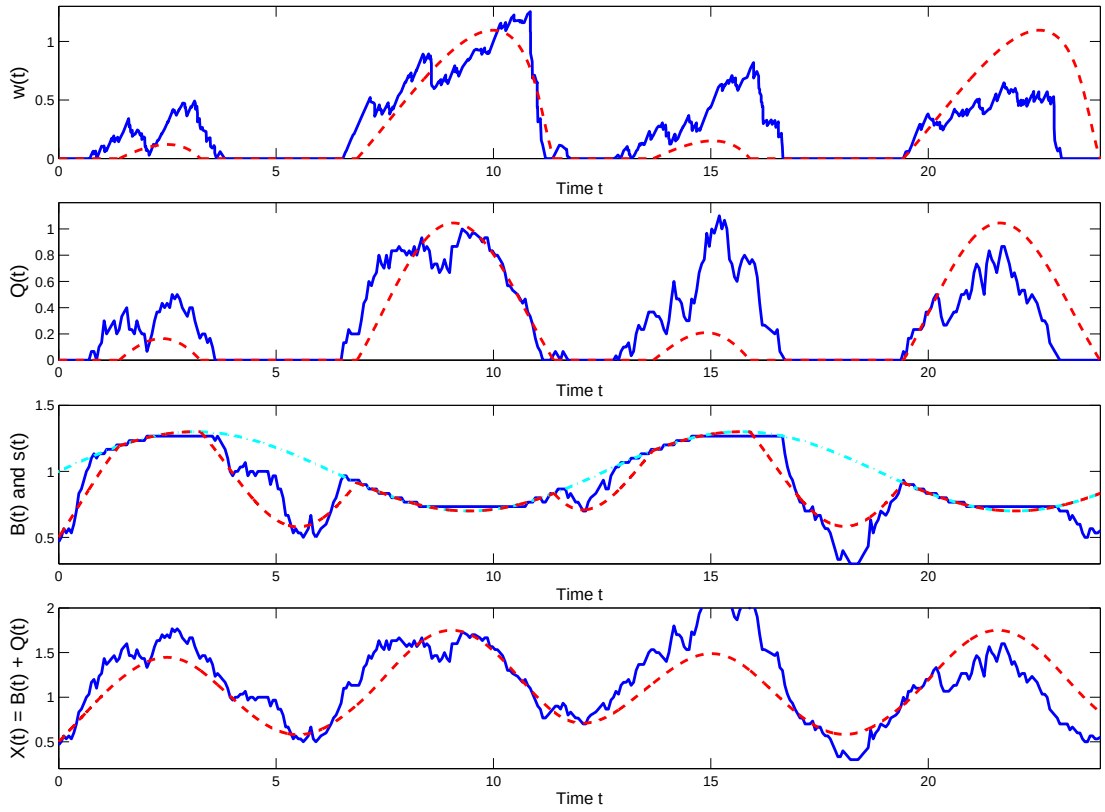

Figure C.4: Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 30$.

fluctuations will be captured by the queue length, while only negative fluctuations will be captures by the number of busy servers.

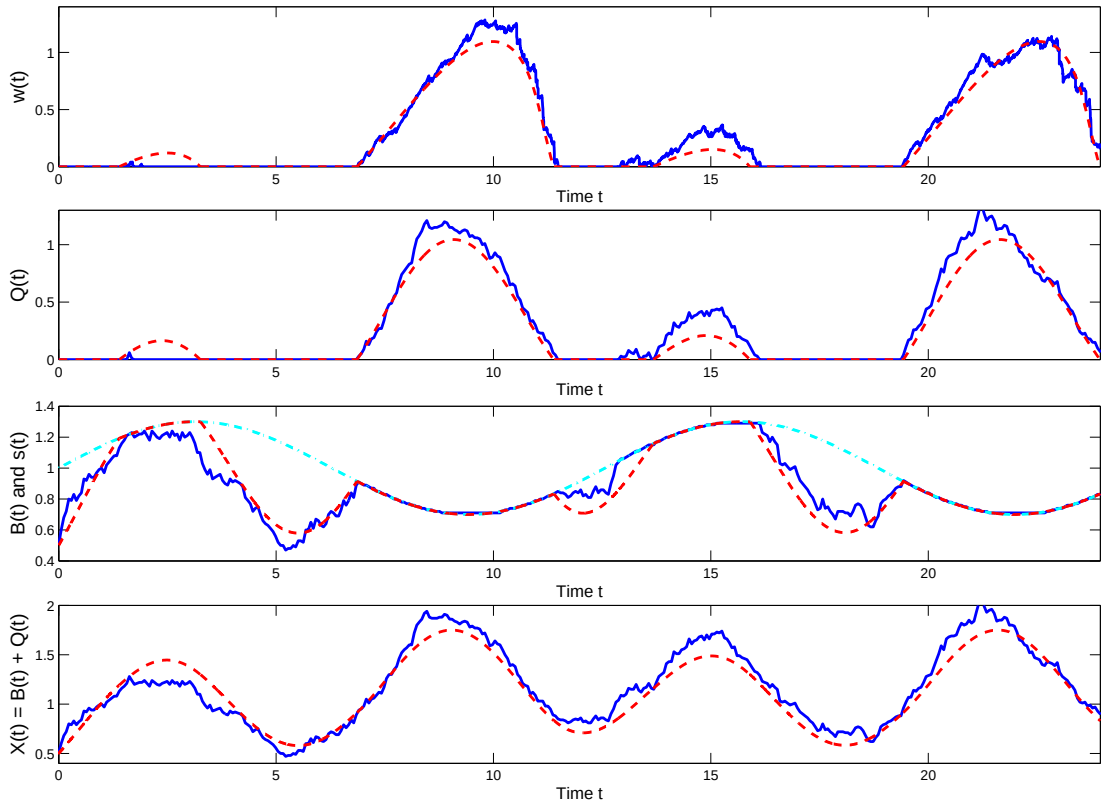

Figure C.5: Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.

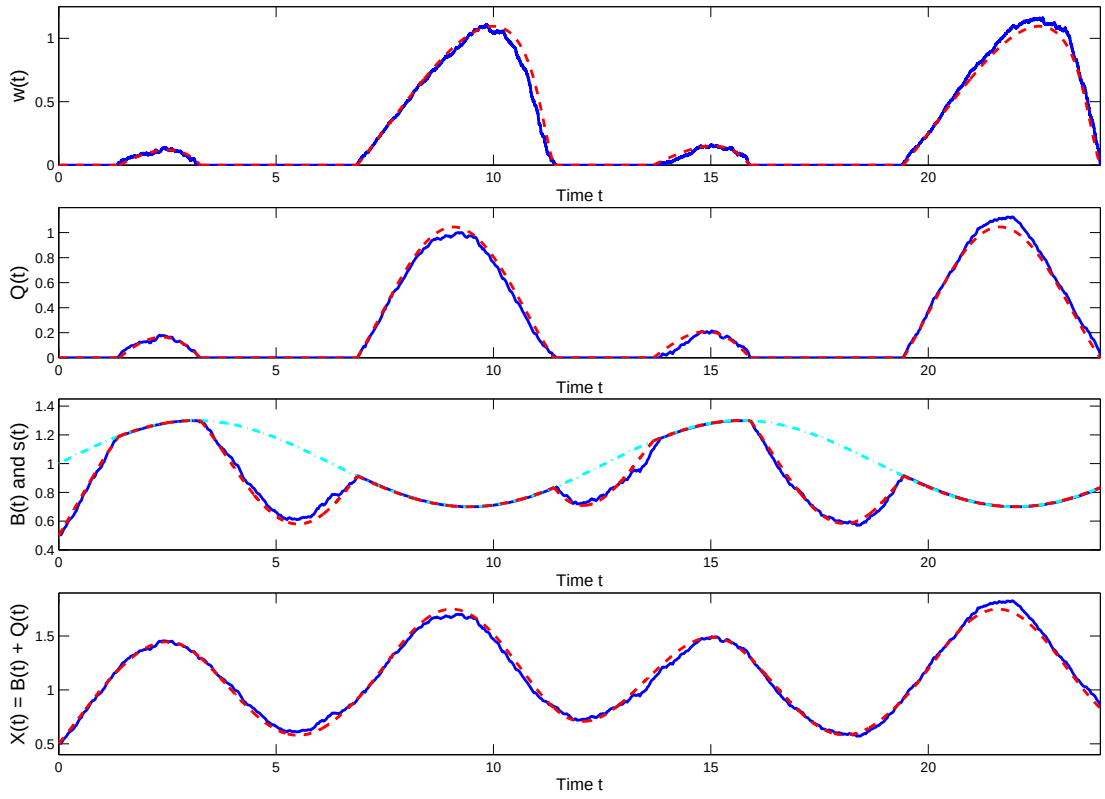

Figure C.6: Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 1000$.

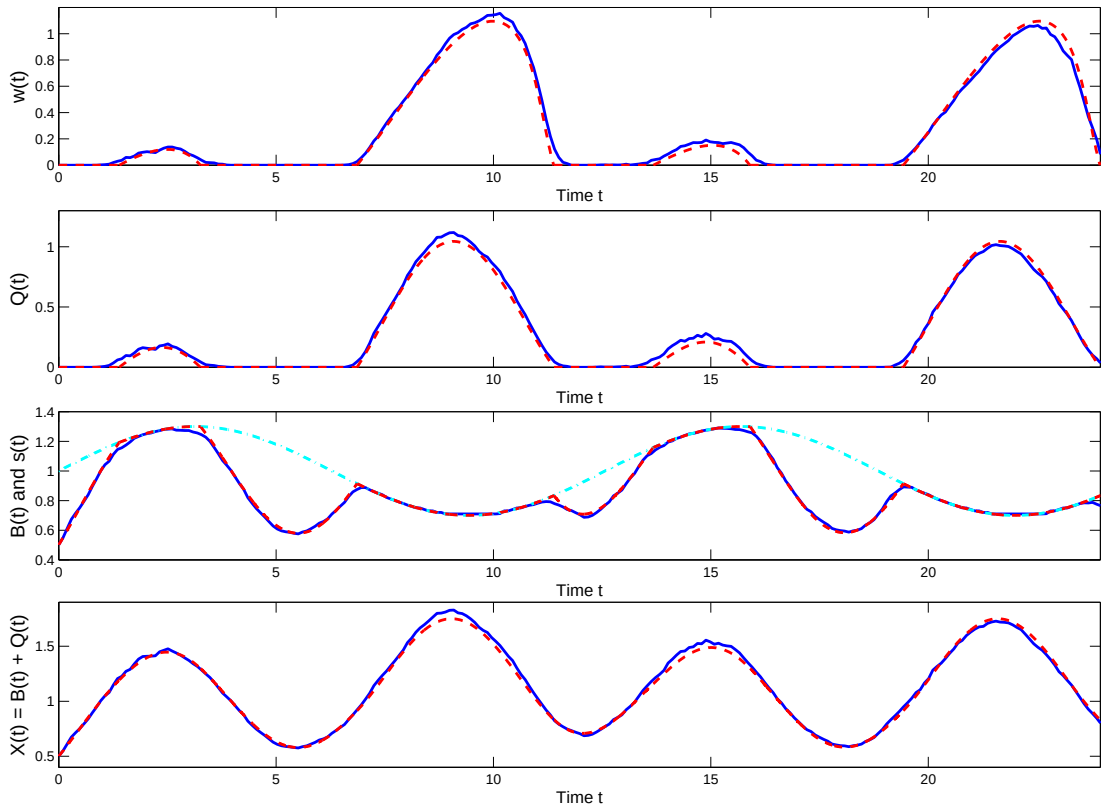

Figure C.7: Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: an average of 20 sample paths of the scaled queueing model based on $n = 100$.

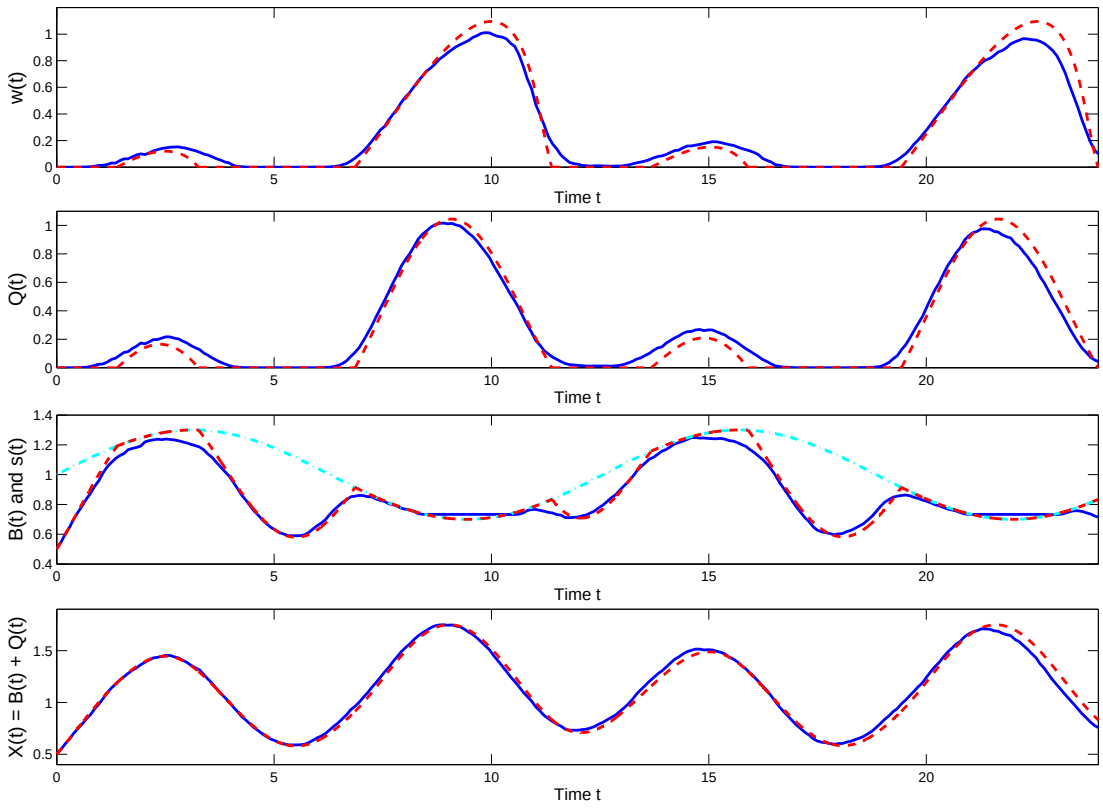

Figure C.8: Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: an average of 200 sample paths of the scaled queueing model based on $n = 30$.

Appendix D

Appendix for Chapter 5

D.1 Overview.

This appendix contains additional supplementary material, which is presented in order of the material to which it relates. First, in §D.2 we present additional simulation results for the example in §5.1. Specifically, we report results of simulations with smaller scaling n but averaged over multiple sample paths, to show the quality of the fluid model as an approximation for mean values in the queueing system. We also consider an example with smaller traffic intensity ρ for the example in §5.1 to show that the periodic behavior is eventually broken.

In §D.3 we give proofs of Theorems 5.7-5.10 in §5.7. In §D.4 we return to the example in §5.1 and show that different initial conditions can yield very different PSS's. In §D.5 we apply the algorithm in Remark 5.2 to numerically evaluate the average performance over a cycle with non-exponential abandonment distributions. These examples show that the

average boundary waiting time over a cycle tends to be strictly greater than the stationary value, whereas the average queue length over a cycle can be either strictly greater or strictly less than the stationary queue content in the fluid model. In §D.6 we provide a proof of Corollary 5.7, giving explicit expressions for the performance in the $G/D/s + M$ fluid model with an exponential abandonment cdf. In §D.7 we provide a proof of Theorem 5.12 showing that there need not exist a finite time T^* after which the system remains overloaded. To do so, we show that the given example switches back and forth between overloaded and overloaded infinitely often, with two switches in each cycle. In §D.8, we give another counterexample with $B(0) < 1$ that is an analog of Example 5.1 in §5.3.

We then start to consider other service distributions. In §D.9 we provide the same PSS results for fluid models that have two-point service distributions with one of the points at 0. Simulation verification is also given there. In §D.10 we provide results of simulation experiments for queues that have nearly deterministic service times. The simulation results shows that the behavior for D service is not exhibited for other two-point distributions. This supports (but of course does not prove) our conjecture that ALOM holds in all other $GI/GI/s + GI$ models and even in the more general $G_t/GI/s_t + GI$ models.

D.2 More on the Example in §5.1

D.2.1 Smaller Scaling n

We used a very large scaling, in particular $n = 1000$, for the queueing model in the example in §5.1. We used a very large n for two reasons: first, to demonstrate that the fluid model becomes accurate in the limit as $n \to \infty$ and, second, to provide a good test of the numerical algorithm for the fluid model. However, in order to be useful as approximations

for realistic large-scale queueing systems, the approximation also should be reasonable for smaller scaling factors. We demonstrate that now.

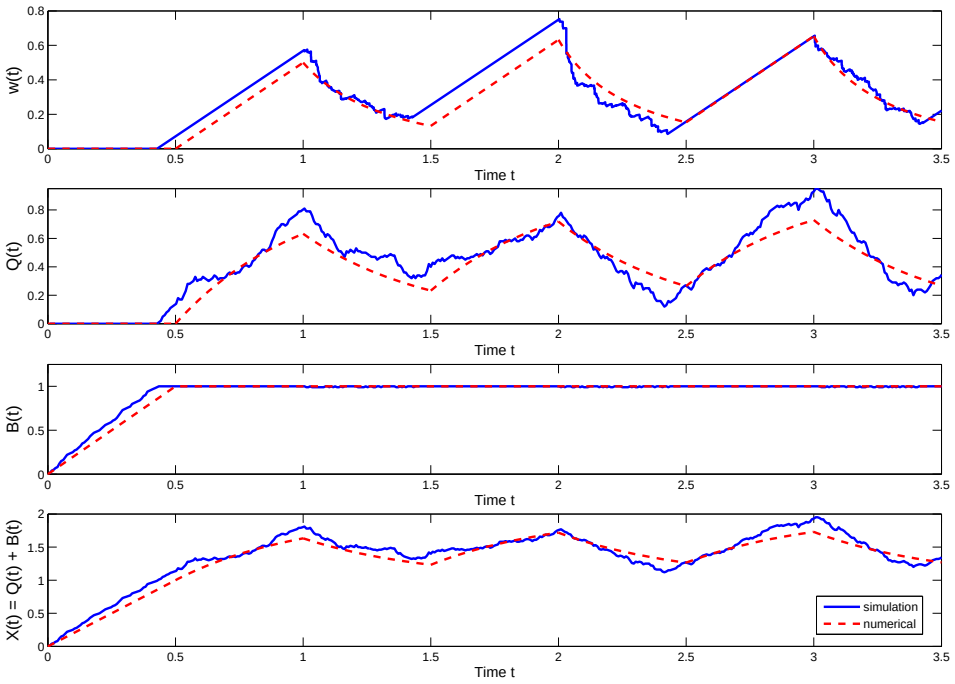

Figure D.1: Performance of the $G/D/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.

We consider the same base $M/D/n + M$ fluid model here as in §5.1, but we only consider the case $\theta = 2$. The other parameters remain unchanged: $\lambda = 2$, $\mu = s = 1$. However, we consider different values of the scaling factor n for the associated stochastic queueing model, which coincides with the number of servers (since we set $s = 1$).

Figure D.1 below provides the analog of Figure 5.2 for the case of one sample path of the simulation with $n = 100$, for the same fluid model. Figure D.2 below gives the average of 10 sample paths for the same model. We see that the fluid approximation provides only a rough approximation for a single sample path when $n = 100$ instead of $n = 1000$, but it is remarkably accurate for the average over 10 sample paths. The accuracy is especially high in this example, because the extent of the overloads and underloads are quite large.

The quality of the approximation does degrade as n decreases, for the given fluid model.

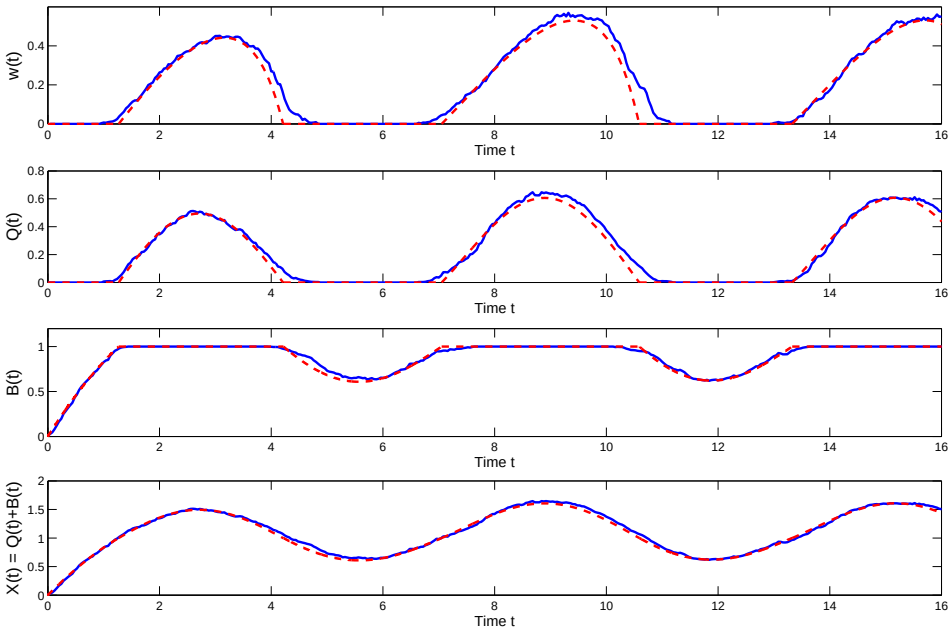

Figure D.2: Performance of the $G/D/s+M$ fluid model compared with simulation results: an average of 10 sample paths of the scaled queueing model based on $n = 100$.

To illustrate, we plot a single sample path for $n = 30$ in Figure D.3 and the average over 100 sample paths in Figure D.4. The stochastic fluctuations are so much greater for a single sample path that we need to average over more sample paths to get a good estimate of the mean values. For $n = 30$, the fluid model clearly yields a good approximation only for the mean values, but the mean is remarkably well approximated for $n = 30$. The approximation for the mean values in Figure D.4 are so good that it is evident that the fluid model approximations can provide useful approximations for the mean values for much smaller n (and thus s).

D.2.2 Smaller Traffic Intensity ρ

For the initial heavily loaded example with $\rho \equiv \lambda/s\mu = 2$ and scaling $n = 1000$ discussed in §5.1 we were not able to detect a break in the periodic behavior in simulations. For

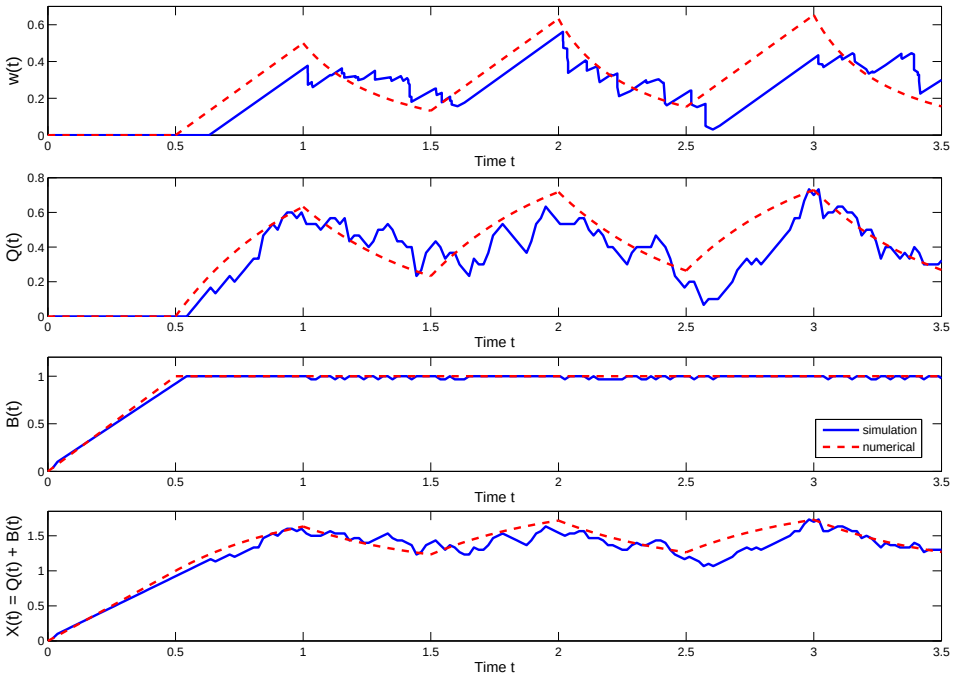

Figure D.3: Performance of the $G/D/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 30$.

example, Figure 5.3 shows that the periodic behavior of $W_n(t)$, the head-of-line waiting time at t, remains even for large T ($T = 1000$). However, we found that a break in the periodic behavior can be observed if we considered less heavily loaded examples.

To illustrate, we now consider the same $M/D/n + M$ queue in §5.1 with the same parameters ($\mu = 1$, $\theta = 2$, $n = 100$) except for a smaller λ, now letting $\lambda = 1.3\,n$, so that the system has a lower traffic intensity, $\rho = \lambda/n\mu = 1.3$ instead of $\rho = 2$ as in §5.1. We repeat the same simulation experiment with $\rho = 1.3$ and plot W_n in Figure D.5. Figure D.5 shows essentially the same periodic behavior over the initial interval $[0, 10]$, but it shows that the periodic behavior is gone by $T = 1000$.

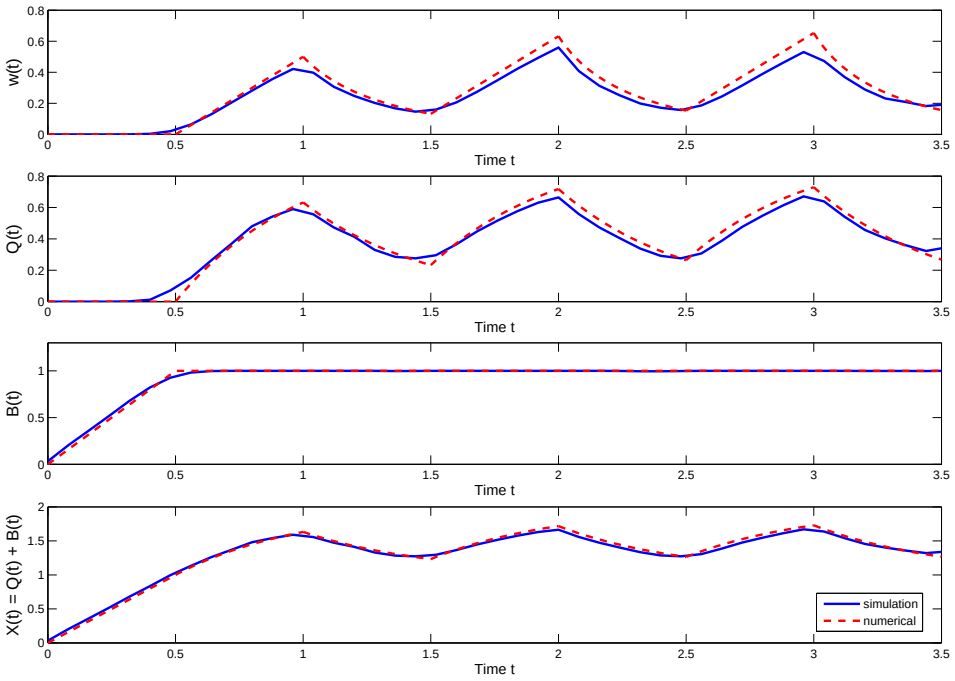

Figure D.4: Performance of the $G/D/s+M$ fluid model compared with simulation results: an average of 100 sample paths of the scaled queueing model based on $n = 30$.

D.3 Proofs for §5.7

We omitted the proofs for the four theorems in §5.7 because they follow from the proofs of corresponding results in Chapter 4. Nevertheless, we provide the details here.

D.3.1 Proof of Theorem 5.7

Proof. Since both queues are overloaded for all $t \geq 0$ and they have the same initial fluid densities in service, we have $b_1(t,0) = b_2(t,0) = \sigma_1(t) = \sigma_2(t)$ by Theorem 3.2. For the fluid content in queue, we have $\tilde{q}_1(t,x) \leq \tilde{q}_2(t,x)$ for all x by Proposition 2.6 because the two queues share the same F.

It remains to show $w_1(t) \leq w_2(t)$ for all $t \geq 0$. We will do a proof by contradiction. Hence suppose this inequality does not hold for some $t > 0$. Then continuity of w_1 and w_2 implies that there exists some $0 < t_1 < t$ such that $w_1(t_1) = w_2(t_1) \equiv \tilde{w}$. However, the

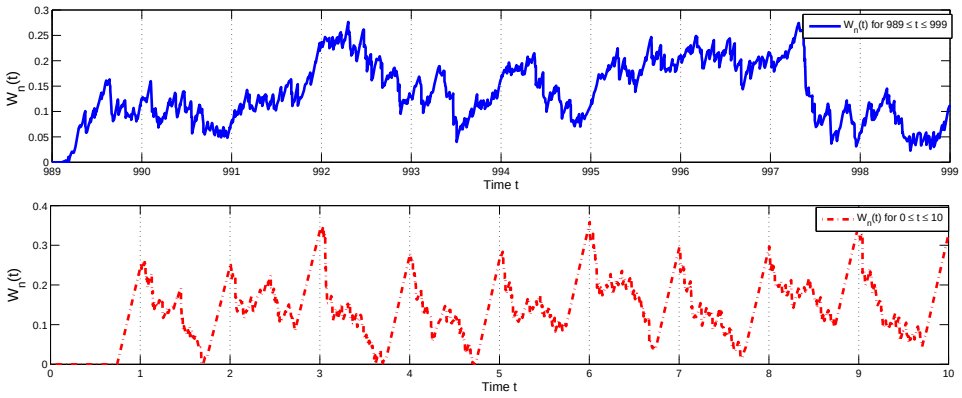

Figure D.5: Large-time periodic behavior of an overloaded $G/D/s+M$ queueing model: simulation estimates of the head-of-line waiting time W_n with $\lambda = 1.3$, $s = \mu = 1$, $\theta = 2$, $\rho = 1.3$, $n = 100$, $T = 1000$.

ordering of $\tilde{q}_1$ and $\tilde{q}_2$ implies that $\tilde{q}_1(t_1, \tilde{w}) \leq \tilde{q}_1(t_1, \tilde{w})$. Hence the BWT ODE in Theorem 2.3 of Chapter 2 implies that $w_1'(t_1) = w_2'(t_1)$ because $b_1(t, 0) = b_2(t, 0)$. Therefore, this contradicts our assumption that there exists a t such that $w_1(t) > w_2(t)$. Hence that establishes the desired ordering.

The ordering of Q and α follow directly from the ordering of q and w since

$$\begin{aligned} Q_1(t) &= \int_0^{w_1(t)} q_1(t,x)dx \leq \int_0^{w_2(t)} q_2(t,x)dx = Q_2(t), \\ \alpha_1(t) &= \int_0^{w_1(t)} q_1(t,x)h_F(x)dx \leq \int_0^{w_2(t)} q_2(t,x)h_F dx = \alpha_2(t). \end{aligned}$$

Now we turn to v. The equation (27) in Theorem 5 implies that the ordering of w is inherited by v. That is made clear by applying the proof of Theorem 5, which shows that $v(t)$ is determined by the intersection of the function w with the linear function $L_t(u) = t + u$. Clearly, if we increase the w function, then that intersection point increases as well. □

D.3.2 Proof of Theorem 3.6

Proof. Without loss of generality, by Theorem 5.7, it suffices to assume that $\lambda_1 \leq \lambda_2$ and $q_1(0,\cdot) \leq q_2(0,\cdot)$. If that is not initially the case, consider another two systems, system 3 and 4 with $\lambda_3 \equiv \lambda_1 \wedge \lambda_2$, $q_3(0,x) \equiv q_1(0,x) \wedge q_2(0,x)$, $\lambda_4 \equiv \lambda_1 \vee \lambda_2$, $q_4(0,x) \equiv q_1(0,x) \vee q_2(0,x)$. Therefore, it is easy to see that $|\lambda_1 - \lambda_2| = |\lambda_3 - \lambda_4|$ and $|Q_1(0) - Q_2(0)| \leq |Q_3(0) - Q_4(0)|$.

Since both queues are overloaded and $b_1(t,0) = b_2(t,0)$, flow conservation of fluid in queue implies that for $i = 1, 2$,

$$Q_i'(t) = \lambda_i - \alpha_i(t) - b_i(t,0).$$

Hence, we have

$$Q_2'(t) - Q_1'(t) = \lambda_2 - \lambda_1 - (\alpha_2 - \alpha_1) \leq \lambda_2 - \lambda_1, \tag{D.1}$$

where the inequality follows from Theorem 5.7. This yields

$$|Q_1(t) - Q_2(t)| = Q_2(t) - Q_1(t) \leq |Q_1(0) - Q_2(0)| + t\,|\lambda_1 - \lambda_2|.$$

Obviously, (5.36) directly follows from (5.34). To show (5.35), we have

$$\begin{aligned}
|\alpha_1(t) - \alpha_2(t)| &= \alpha_2(t) - \alpha_1(t) \\
&= \int_0^{w_2(t)} q_2(t,x) h_F(x) dx - \int_0^{w_1(t)} q_1(t,x) h_F(x) dx \\
&= \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_F(x) dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x) h_F(x) dx \\
&\le h_F^{\uparrow} \left(\int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_F(x) dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x) h_F(x) dx \right) \\
&= h_F^{\uparrow}(Q_2 - Q_1) = h_F^{\uparrow} |Q_2 - Q_1|,
\end{aligned}$$

where the first and last equality, and the inequality all follows from Theorem 5.7. □

D.3.3 Proof of Theorem 5.9

Proof. We first show that (a) follows from (b). Without loss of generality, we assume $Q_1(0) \le Q_2(0)$. We construct another two systems, 3 and 4, with $q_3(0,x) \equiv q_1(0,x) \wedge q_2(0,x)$ and $q_4(0,x) \equiv q_1(0,x) \vee q_2(0,x)$. With this construction, systems 3 and 4 are bona fide fluid models, with $Q_3(t) \le Q_1(t) \le Q_4(t)$ and $Q_3(t) \le Q_2(t) \le Q_4(t)$ for all t, by Theorem 5.7. This implies that $\Delta Q_{1,2}(t) \le \Delta Q_{3,4}(t)$ for all t. Since $\delta Q_{3,4}(t)(0) \le C_1$ for C_1 in (5.38), (5.37) in (a) follows from (5.43) for $\Delta Q_{3,4}(t)$. (The final bound on C_1 in (5.38) arises when the supports of $q_1(0,\cdot)$ and $q_2(0,\cdot)$ are disjoint sets.)

Now we prove (b). Observe that the first inequality in (5.43) follows (5.42) because dividing the interval $[0,T]$ into N subintervals yields

$$\Delta Q(T) \le \left(\frac{1}{1 + h_F^{\downarrow} \frac{T}{N}} \right)^N \Delta Q(0).$$

Letting $N \to \infty$, we get (5.42).

We now prove (5.42). Since both queues are overloaded for all $t \geq 0$ and they have the same initial fluid densities in service, we have $b_1(t,0) = b_2(t,0) = \sigma_1(t) = \sigma_2(t)$, following from Theorem 3.2. Since $q_1(0,x) \leq q_2(0,x)$, we have $q_1(t,x) \leq q_2(t,x)$, $w_1(t) \leq w_2(t)$ and $\alpha_1(t) \leq \alpha_2(t)$ for all $t \geq 0$. Hence, we have

$$\begin{aligned}
\alpha_2(t) - \alpha_1(t) &= \int_0^{w_2(t)} q_2(t,x) h_F(x) dx - \int_0^{w_1(t)} q_1(t,x) h_F(x) dx \\
&= \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_F(x) dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x) h_F(x) dx \\
&\geq h_F^{\downarrow} \left(\int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x)(x) dx \right) \\
&= h_F^{\downarrow} \left(Q_2(t) - Q_1(t) \right) = h_F^{\downarrow} \, \Delta Q(t). \qquad \text{(D.2)}
\end{aligned}$$

Flow conservation implies that

$$Q_i'(t) = \lambda - \alpha_i(t) - b_i(t,0) \quad \text{for } i = 1, 2,$$

which yields

$$\Delta Q'(s) = -(\alpha_2(s) - \alpha_1(s)) \leq -h_F^{\downarrow} \, \Delta Q(s) \leq -h_F^{\downarrow} \, \Delta Q(t), \quad 0 \leq s \leq t,$$

where the first inequality follows from (D.2) and the second inequality holds since $\Delta Q(t)$ has negative derivative. Therefore, integrating both sides with respect to s from 0 to t, we have

$$\Delta Q(t) - \Delta Q(0) \leq -h_F^{\downarrow} \, t \, \Delta Q(t)$$

and

$$\Delta Q(t) \le \left(\frac{1}{1 + h_F^{\downarrow} t} \right) \Delta Q(0).$$

To show the second inequality in (5.43), repeat the reasoning in (D.2) and use the face $h_F(x) \le h_F^{\uparrow}$ instead of $h_F(x) \ge h_F^{\downarrow}$.

Finally, we treat $w(t)$. As above, it suffices to assume that we have the ordering in (5.41). We have $b(t,0) \ge b^{\downarrow}$ following from Proposition 5.4 and Corollary 5.3. First note that at time $T^* = (Q_1(0) + Q_2(0))/b^{\downarrow}$, all fluid that was in queue 1 and 2 at time 0 is gone (entered service or abandoned). Then (5.39) follows from

$$\Delta Q(T) = \int_{w_1(T)}^{w_2(T)} \lambda \, \bar{F}(x) dx \le \lambda \, \bar{F}(w_2(T)) \, \Delta w(T), \quad T \ge T^*.$$

Choose $\bar{w} > 0$ big enough such that $\bar{F}(\bar{w}) < b^{\downarrow}/\lambda$. The BWT ODE implies that for $t > T^*$,

$$w_2'(t) = 1 - \frac{b(t,0)}{\lambda \, \bar{F}(w_2(t))} \le 1 - \frac{b^{\downarrow}}{\lambda \, \bar{F}(\bar{w})} < 0,$$

if $w_2(t) > \bar{w}$ for some t. Hence $\bar{w}$ is an upper bound for $w_2(t)$ if $w_2(T^*) < \bar{w}$. If $w_2(T^*) \ge \bar{w}$, it is easy to see that $w_2(t)$ decreases until it is below $\bar{w}$ because we can bound $w_2'(t)$. This argument implies that $w_2(t) \le \bar{w} \vee (w_2(0) + T^*)$ for all $t \ge 0$. The constant C_2 in (5.40) is obtained by inserting established bounds. □

D.3.4 Proof of Theorem 4.1

Proof. Most are elementary; only $Q(t)$ and $w(t)$ require detailed argument. Flow conservation implies that $Q'(t) = \lambda - \alpha(t) - b(t,0) \le \lambda - \alpha(t)$. Since $\alpha(t) \ge h_F^{\downarrow} Q(t)$, we have

$Q'(t) < 0$ whenever $Q(t) > \lambda/h_F^{\downarrow}$. The bound for $w(t)$ follows directly from (5.39) and the proof of Theorem 5.9. □

D.4 Different Initial Conditions

Theorems 5.6 and 5.11 provide sufficient conditions for Assumption 5.7 to hold, and for the performance function to converge to a PSS. That PSS depends strongly on the fluid density in service, b at the time T^* after which the system remains overloaded. We now illustrate that different initial conditions can yield very different PSS's.

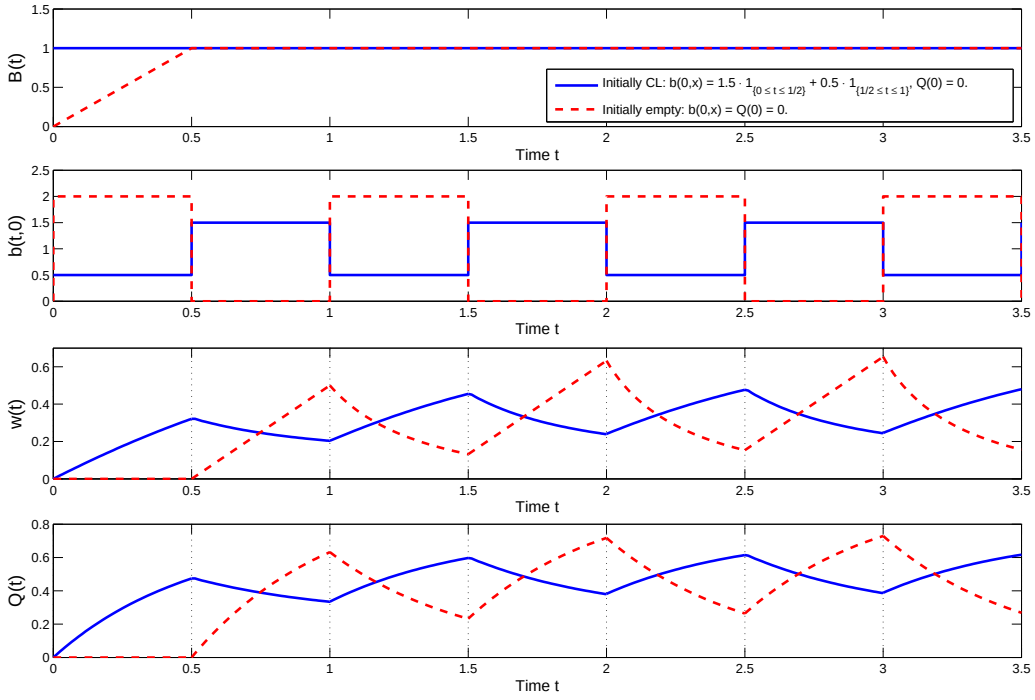

Figure D.6: A comparison of the PSS performance of the $G/D/s + M$ fluid queue with different initial conditions: (i) critically loaded with $b(0, x) = 1.5 \cdot 1_{\{0 \le x \le 1/2\}} + 0.5 \cdot 1_{\{1/2 \le x \le 1\}}$, $Q(0) = 0$ (the blue solid lines); (ii) starting empty (the red dashed lines).

We again consider the $G/D/s + M$ example in §5.1 with $\lambda = 2$, $\mu = s = 1$, $\theta = 2$. In Figure D.6, we apply the algorithm in Remark 5.2 and plot the performance functions $B(t)$, $b(t, 0)$, $w(t)$ and $Q(t)$ in interval $[0, 3.5]$ for two different initial conditions: (i) The

system is initially critically loaded (CL) with $b(0,x) = 1.5 \cdot 1_{\{0 \le x \le 1/2\}} + 0.5 \cdot 1_{\{1/2 \le x \le 1\}}$, $Q(0) = 0$ (the blue solid lines); (ii) The system is initially empty (the red dashed lines). Both cases yield a PSS with period $1/\mu = 1$, but the performance in these two cases differs greatly.

D.5 The Average Performance Over a Cycle

In Remark 5.5 we noted that, unlike $\bar{\alpha}$ and $\bar{\sigma}$, the averages of other performance functions in a PSS typically do not agree with the steady-state values. We investigate $\bar{Q}$ and $\bar{w} \equiv \tau^{-1} \int_0^\tau w(t)\, dt$ now.

We consider an initially empty $G/D/s + GI$ fluid model with three types of abandonment distributions: (i) Erlang-2 (E_2), (ii) exponential (M) and (iii) Hyperexponential-2 (H_2). We first review these distributions.

Let A be the generic abandonment time. A follows E_2 implies that $A = X_1 + X_2$ in distribution, where X_1 and X_2 are two iid exponential random variables. Moreover, $f(x) = \gamma^2\, x\, e^{-\gamma\, x}$, where γ is rate of X_1. If A follows H_2, then A is a mixture of two exponential random variables, i.e., $f(x) = p \cdot \theta_1\, e^{-\theta_1\, x} + (1-p) \cdot \theta_2\, e^{-\theta_2\, x}$, where θ_1 and θ_2 are the rates of these two exponential random variables, and $0 < p < 1$ is the sampling probability.

We fix the mean of A, letting $E[A] = 1/\theta$. An E_2 distribution has squared coefficient of variation (SCV) $C^2 \equiv Var(A)/E[A]^2 = 1/2$, which is less than 1. On the other hand, all H_2 distributions have C^2 greater than 1. For E_2, we let $\gamma = 2\,\theta$. For H_2, we let $p = 0.5(1 - \sqrt{0.6})$, $\theta_1 = 2p\,\theta$, $\theta_2 = 2(1-p)\,\theta$, so that $C^2 = 4$.

We let $\lambda = 2$, $\theta = 2$, $\mu = s = 1$. In Figure D.7, we plot w, Q and α in one cycle $[0, 1/\mu]$ of PSS for these three abandonment distributions, by applying the algorithm described in Remark 5.2. (Here we start the system empty and compute these performance functions

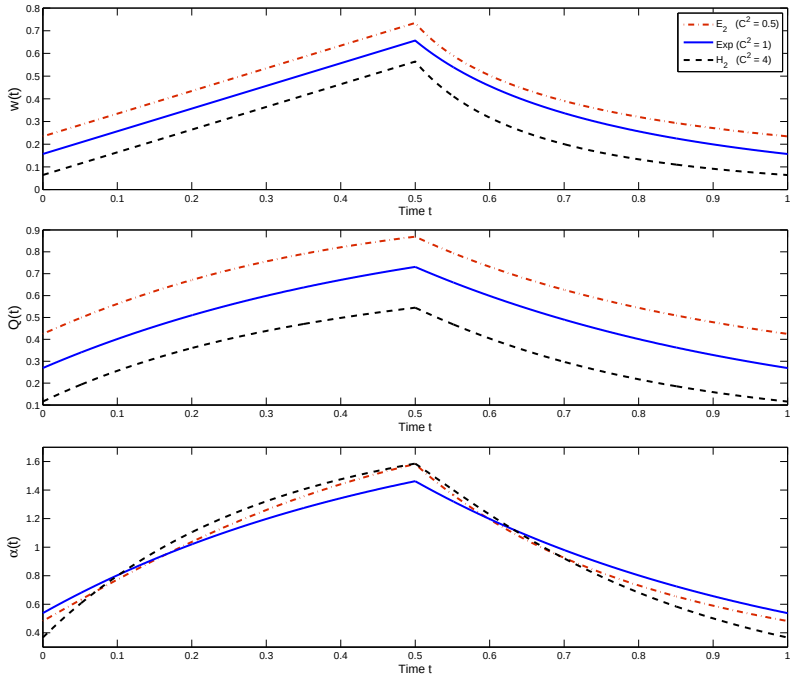

Figure D.7: A comparison of the PSS of the $G/D/s+GI$ fluid queues with different abandonment distributions: (i)E_2 (red dashed), (ii) M (blue solid) and (iii) H_2 (black dashed).

in N cycles for N large.) In Table 1, we compute and compare $\bar{w}$, $\bar{Q}$ and $\bar{\alpha}$, the average of w, Q and α in one cycle to w^*, Q^* and α^*, their steady-state values. We have three observations: (i) As proved in Corollary 5.6, $\bar{\alpha}$ indeed agrees with α^* (except for a small computation error from numerical integration); (ii) $\bar{Q} \neq Q^*$ in general, in particular, $\bar{Q} < Q^*$ for E_2 abandonment and $\bar{Q} > Q^*$ for H_2 abandonment; (iii) $\bar{w} \geq w^*$, i.e., customers' average waiting is longer in PSS than in the steady state.

D.6 The Case of Exponential Abandonment

In this section we prove Corollary 5.7, giving explicit formulas in the case of exponential abandonment. We give two different proofs.

abandonment dist.	E_2 ($C^2 = 0.5$)	M ($C^2 = 1$)	H_2 ($C^2 = 4$)
$\bar{\alpha}$ (PSS average)	1.001	1	1.001
α^* (steady state)	1	1	1
$\bar{w}$ (PSS average)	0.437	0.367	0.260
w^* (steady state)	0.420	0.347	0.226
$\bar{Q}$ (PSS average)	0.649	0.5	0.330
Q^* (steady state)	0.657	0.5	0.324

Table D.1: A comparison of the average performance of PSS of the $G/D/s + GI$ fluid queue with (i) E_2, (ii) M and (iii) H_2 abandonment distribution to the steady-state values.

D.6.1 First Proof of Corollary 5.7

First, since $b(t,x)$ and $\sigma(t)$ are periodic functions and $Q(t)$ and $\alpha(t)$ can be written as expressions in terms of $w(t)$, it remains to derive the dynamics of $w(t)$.

In a cycle $[0, 1/\mu]$, $w(t) = \tilde{w} + t$ for $0 \le t \le 1/\mu - s/\lambda$ and $w(t)$ solves ODE $w'(t) = 1 - 1/\bar{F}(w(t)) = 1 - 1/e^{-\theta w(t)}$ with $w(1/\mu - s/\lambda) = \tilde{w} + 1/\mu - s/\lambda$ for $1/\mu - s/\lambda \le t \le 1/\mu$, where $\tilde{w} \ge 0$ is both the starting and the ending value of $w(t)$ in each cycle. Letting $v(t) \equiv t - w(t)$, we have for $1/\mu - s/\lambda \le t \le 1/\mu$,

$$e^{\theta t} = (1 - w'(t))e^{\theta(t-w(t))} = v'(t)e^{\theta v(t)}.$$

For $1/\mu - s/\lambda \le t \le 1/\mu$, integrating both sides from $1/\mu - s/\lambda$ to t yields

$$\begin{aligned} e^{\theta t} - e^{\theta(1/\mu - s/\lambda)} &= \theta \int_{1/\mu - s/\lambda}^{t} e^{\theta u} du = \theta \int_{v(1/\mu - s/\lambda)}^{v(t)} e^{\theta u} du \\ &= e^{\theta(t-w(t))} - e^{\theta(1/\mu - s/\lambda - w(1/\mu - s/\lambda))}. \end{aligned} \tag{D.3}$$

Because $w(1/\mu - s/\lambda) = \tilde{w} + 1/\mu - s/\lambda$ and $w(1/\mu) = \tilde{w}$, letting $t = 1/\mu$ in (D.3) yields (5.52), from which (5.50) follows. Solving the ODE yields (5.53).

Finally, to show (c), we consider a cycle $[1/\mu - \tilde{w}, 2/\mu - \tilde{w}]$ instead of $[0, 1/\mu]$. First,

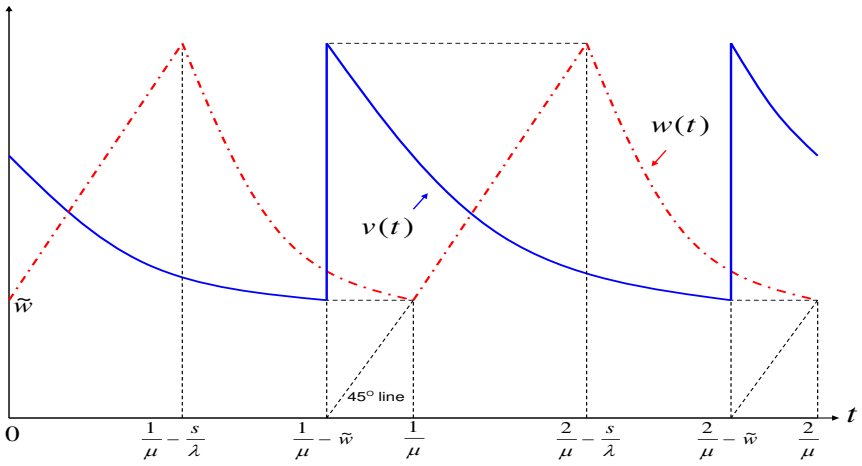

Figure D.8: PWT $v(t)$ and BWT $w(t)$ of the PSS of the $G/D/s+GI$ fluid queue.

the PWT $v(t)$ is periodic with the same period $1/\mu$. Moreover, it is continuous over $[1/\mu - \tilde{w}, 2/\mu - \tilde{w})$ and it has a discontinuity at $t = 2/\mu - \tilde{w}$, as shown in Figure D.8, following from Theorem 2.5. Also see Theorem 2.3 and 2.6 in Chapter 2 for details. Following Theorem 2.6 in Chapter 2, $v(t)$ satisfies the ODE

$$\begin{aligned} v'(t) &= \frac{\lambda \bar{F}(v(t))}{b(t+v(t),0)} - 1 = \frac{\lambda\, e^{-\theta\, v(t)}}{\lambda} - 1 \\ &= e^{-\theta\, v(t)} - 1, \quad \frac{1}{\mu} - \tilde{w} \le t < \frac{2}{\mu} - \tilde{w}, \end{aligned} \tag{D.4}$$

where the second equality holds because $b(t,0) = \lambda$ for $2/\mu - s/\lambda \le t \le 2/\mu$ and $t+v(t) \ge 2/\mu - s/\lambda$ (obviously from Figure D.8). Since $v(1/\mu - \tilde{w}) = \tilde{w} + 1/\mu - s/\lambda \equiv v_0$, solving ODE (D.4) with $(1/\mu - \tilde{w}) = v_0$ yields (5.55).

D.6.2 Second Proof of Corollary 5.8

We can provide an alternative proof of Corollary 5.8 by focusing on $Q(t)$. Since $\sigma(t) = b(t,0) = 0$, $Q(t)$ satisfies an ODE for $0 \le t \le 1/\mu - s/\lambda$ with

$$Q'(t) = \lambda - \theta\, Q(t),$$

which has a unique solution

$$Q(t) = \frac{\lambda}{\theta}\left(1 - e^{-\theta\, t}\right) + Q(0)\, e^{-\theta\, t}. \tag{D.5}$$

Since $\sigma(t) = b(t,0) = \lambda$ for $1/\mu - s/\lambda < t \le 1/\mu$, $Q(t)$ satisfies another ODE

$$Q'(t) = \lambda - \theta\, Q(t) - b(t,0) = -\theta\, Q(t),$$

which has a unique solution

$$Q(t) = Q^*\, e^{-\theta\, t}, \tag{D.6}$$

where

$$Q^* \equiv Q\left(\frac{1}{\mu} - \frac{s}{\lambda}\right) = \frac{\lambda}{\theta}\left(1 - e^{-\theta\left(\frac{1}{\mu} - \frac{s}{\lambda}\right)}\right) + Q(0)\, e^{-\theta\left(\frac{1}{\mu} - \frac{s}{\lambda}\right)}$$

is the ending value of $Q(t)$ in $[0, 1/\mu - s/\lambda]$; i.e., let $t = 1/\mu - s/\lambda$ in (D.5). Since $Q(t)$ is periodic in the PSS with period $1/\mu$, we must have $\tilde{Q} \equiv Q(0) = Q(1/\mu)$. Equating $Q(0)$

to $Q(t)$ in (D.6) with $t = 1/\mu$ yields

$$\tilde{Q} = \frac{\lambda}{\theta}\left(\frac{e^{-\theta\, s/\lambda} - e^{-\theta/\mu}}{1 - e^{-\theta/\mu}}\right). \tag{D.7}$$

Plugging $Q(0) = \tilde{Q}$ in (D.7) into (D.5) and (D.6) yields (5.51) and (5.54). To show (5.52), we let

$$\tilde{Q} = \int_0^{\tilde{w}} \lambda\, e^{-\theta\, x} dx = \frac{\lambda}{\theta}\left(1 - e^{-\theta\, \tilde{w}}\right), \tag{D.8}$$

which yields (5.52).

D.7 On Theorem 9.1

Recall that Theorem 5.12 concludes that there need not exist a finite time T^* after which the system remains overloaded; i.e., there need not exist $T^* < \infty$ such that $B(t) = s$ for all $t \geq T^*$. The proof involves a concrete counterexample. We now show that the counterexample indeed has the claimed property.

D.7.1 Proof of Theorem 5.12

We start by giving a feel for the performance by applying the numerical algorithm in Remark 5.2. We plot the performance functions $w(t)$, $Q(t)$, $B(t)$, $b(t, 0)$ and $\sigma(t)$ for $0 \leq t \leq 5$ in Figure D.9. Figure D.9 clearly shows that $B(n) = s$ for all n and that $B(n + (1/2))$ increases towards s.

However, from the picture alone, we cannot be sure that $B(n + (1/2)) < s$ for all n. To justify that, we need to consider the behavior more carefully. To show that the system

alternates between overloaded and underloaded infinitely often, we consider successive intervals $[n, n+1]$ for $n \geq 0$. First, in the first unit $[0,1]$, we have $b(t,0) = \sigma(t) = b(0, 1-x) = 2 \cdot 1_{\{0 \leq x \leq 1/2\}}$. Since $b(t,0) = \sigma(t)$ whenever the system is overloaded and the system is initially overloaded, the BWT $w(t)$ satisfies the ODE

$$w'(t) = 1 - \frac{b(t,0)}{\lambda \bar{F}(w(t))} = 1 - \frac{2}{1.2\, e^{-2\,w(t)}} 1_{\{0 \leq t \leq 1/2\}}, \tag{D.9}$$

with $w(0) = 2$, which has a unique solution

$$w(t) = t - \frac{1}{2} \log \left(\frac{e^{2\,t} - 1}{0.6} + e^{-2\,w(0)} \right) \quad \text{for } 0 \leq t \leq 1/2.$$

Letting $w(t) = 0$ yields that

$$t_1^{(1)} = \frac{1}{2} \log \left(\frac{1 - 0.6\, e^{-2\,w(0)}}{0.4} \right) = 0.453 < 1/2, \tag{D.10}$$

that is the time at which the system becomes underloaded. Note that for $t_1^{(1)} < t \leq 1/2$, $\sigma(t) = 2 > 1.2 = b(t,0) = \lambda$, therefore, the fluid content in service decreases (linearly) with $B(t) = s - (\sigma(t) - b(t,0))\,(t - t_1^{(1)}) = 1 - 0.8(t - t_1^{(1)})$. For $t > 1/2$, $b(t,0) = \lambda = 1.2 > 0 = \sigma(t)$, $B(t)$ increases (liearly) with $B(t) = B(1/2) + (b(t,0) - \sigma(t))\,(t - 1/2) = 0.96 + 1.2(t - 1/2)$. So the system again becomes overloaded at $t_2^{(1)} = 0.53$ since $B(t_2^{(1)}) = 1 = s$. Moreover, $t_1^{(1)}$ and $t_2^{(1)}$ satisfy $1.2(t_2^{(1)} - 1/2) = 0.8(1/2 - t_1^{(1)})$. For $t_2 \leq t \leq 1$, by ODE (D.9), $w(t) = t - t_2^{(1)}$, which implies that $w(1) = 1 - t_2^{(1)} = 0.47 < 2 = w(0)$. In summary, the system is overloaded in $[0, t_1^{(1)}] \cup [t_2^{(1)}, 1]$ and (strictly) underloaded in $(t_1^{(1)}, t_2^{(1)})$, $b^{(1)}(t,0) \equiv b(t,0) = 2 \cdot 1_{\{0 \leq t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \leq t \leq 1/2\}}$ and $w^{(1)}(0) \equiv w(0) > w(1) \equiv w^{(1)}(1)$, with $0 < t_1^{(1)} < 1/2 < t_2^{(1)} < 1$. See Figure D.9.

Now consider the next unit interval $[1,2]$. We can simply shift the origin to time 1

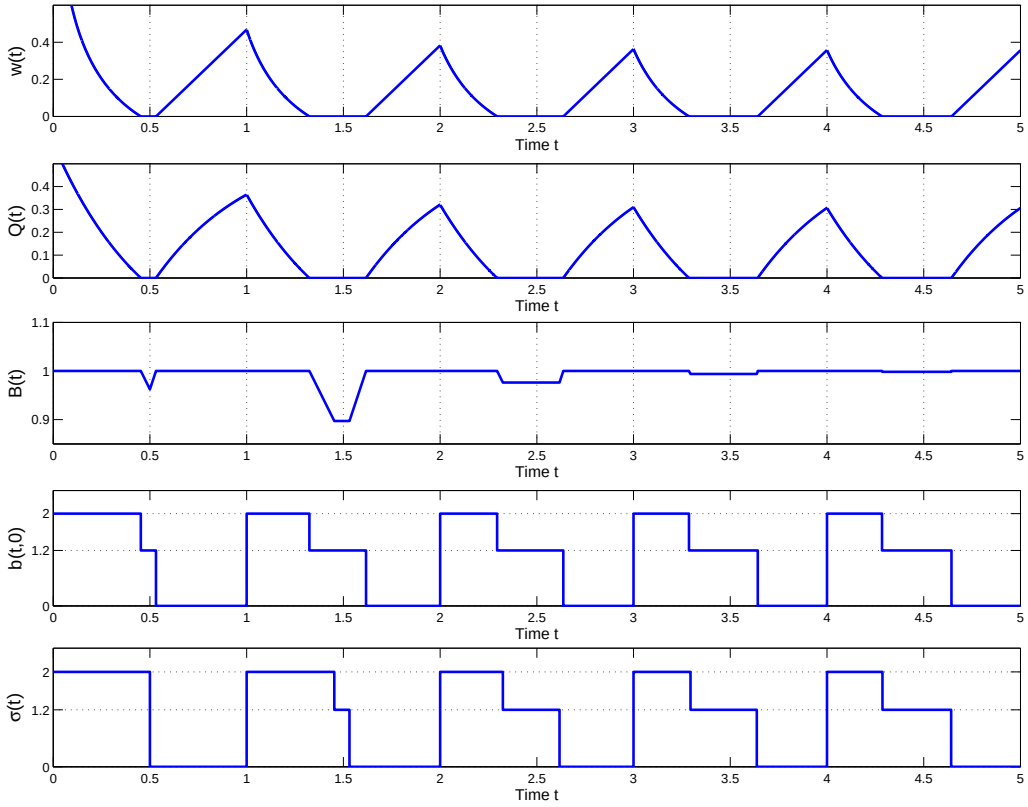

Figure D.9: The counterexample providing a fluid model that does not become (and stay) overloaded in finite time; it switches between overloaded and underloaded regimes infinitely often.

and again consider the interval $[0, 1]$. Therefore the system is initially overloaded with $w^{(2)}(0) \equiv w(0) = w^{(1)}(1) < w^{(0)}(0)$, $\sigma(t) = b^{(1)}(t, 0) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le t_2^{(1)}\}}$ (which is the rate into service in the previous interval). We want to show that the same structure of all performance functions are preserved in the second unit interval. The switching time (from overloaded to underloaded) is a strict monotone function of $w(0)$, by (D.10), therefore the system becomes underloaded at $t_1^{(2)}$ such that $t_1^{(2)} < t_1^{(1)}$ since $w(0) =$

$w^{(1)}(1) < w^{(1)}(0)$. Because $\sigma(t) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le t_2^{(1)}\}}$, we have

$$\begin{aligned} B(t) &= 1_{\{t \in [0, t_1^{(2)}) \cup (t_2^{(2)}, 1]\}} + [1 - 0.8(t - t_1^{(2)})] 1_{\{t_1^{(2)} \le t < t_1^{(1)}\}} \\ &\quad + [1 - 0.8(t_1^{(1)} - t_1^{(2)})] 1_{\{t_1^{(1)} \le t \le 1/2\}} \\ &\quad + [1 - 0.8(t_1^{(1)} - t_1^{(2)}) + 1.2(t - t_2^{(1)})] 1_{\{t_2^{(1)} \le t \le t_2^{(2)}\}}, \end{aligned}$$

where $t_2^{(2)}$ satisfies $1.2(t_2^{(2)} - t_2^{(1)}) = 0.8(t_1^{(1)} - t_1^{(2)})$ so that $t_2^{(2)} > t_2^{(1)}$, which implies that the system is overloaded for $t_2^{(2)} \le t \le 1$ and $w^{(2)}(1) \equiv w(1) = 1 - t_2^{(2)} < w(0) = w^{(1)}(1) = w^{(2)}(0)$. In summary, in the second interval, the system is overloaded in $[0, t_1^{(2)}] \cup [t_2^{(2)}, 1]$ and (strictly) underloaded in $(t_1^{(2)}, t_2^{(2)})$, $b^{(2)}(t, 0) \equiv b(t, 0) = 2 \cdot 1_{\{0 \le t < t_1^{(2)}\}} + 1.2 \cdot 1_{\{t_1^{(2)} \le t \le t_2^{(2)}\}}$, $\sigma^{(2)}(t) \equiv \sigma(t) = b^{(1)}(t, 0) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le t_2^{(1)}\}}$ and $w^{(2)}(0) \equiv w(0) > w(1) \equiv w^{(2)}(1)$, with $0 < t_1^{(2)} < t_1^{(1)} \le t_2^{(1)} < t_2^{(2)} < 1$. See Figure D.9.

Using an inductive argument, we can show that in the nth unit interval $[n-1, n]$, the same structure is preserved. In particular, if we move the origin to time $n-1$ (i.e., consider $[0, 1]$ instead of $[n-1, n]$), then

$$\begin{aligned} \text{the system} \quad \text{is} \quad & \begin{cases} \text{overloaded}, & \text{for } t \in [0, t_1^{(n)}] \cup [t_2^{(n)}, 1], \\ \text{(strictly) underloaded}, & \text{for } t \in (t_1^{(n)}, t_2^{(n)}). \end{cases} \\ b^{(n)}(t, 0) &\equiv b(t, 0) = 2 \cdot 1_{\{0 \le t < t_1^{(n)}\}} + 1.2 \cdot 1_{\{t_1^{(n)} \le t \le t_2^{(n)}\}}, \\ \sigma^{(n)}(t) &\equiv \sigma(t) = b^{(n-1)}(t, 0) = 2 \cdot 1_{\{0 \le t < t_1^{(n-1)}\}} + 1.2 \cdot 1_{\{t_1^{(n-1)} \le t \le t_2^{(n-1)}\}}, \\ w^{(n)}(0) &\equiv w(0) > w(1) \equiv w^{(n)}(1), \end{aligned}$$

with $0 \le t_1^{(n)} < t_1^{(n-1)} \le t_2^{(n-1)} < t_2^{(n)} \le 1$. Therefore, the bounded sequence $t_1^{(1)}, t_1^{(2)}, \ldots$ is strictly decreasing and the bounded sequence $t_2^{(1)}, t_2^{(2)}, \ldots$ is strictly increasing so that we must have $t_1^{(n)} \downarrow t_1^{\infty} \ge 0$ and $t_2^{(n)} \uparrow t_2^{\infty} \le 1$. We next show that $t_1^{\infty} > 0$ and $t_2^{\infty} < 1$.

Suppose $t_1^\infty = 0$, then $w^\infty(0) = w^\infty(1) = 0$, which implies that $t_2^\infty = 1$ (the monotonicity structure is preserved in the limit). Therefore, the system is underloaded or critically loaded in $[0, 1]$. However, since we have $\rho = \lambda/s\mu = 1.2 > 1$, this cannot happen. Hence a contradiction.

D.7.2 More On Theorem 5.12

The example in the proof of Theorem 5.12 discussed above in §D.7.1 also can illustrate the important role played by the initial queue density $q(0, \cdot)$ on the asymptotic performance. Indeed, we can ensure that a time $T^* < \infty$ exists such that $B(t) = s$ for all $t \geq T^*$ by changing the initial queue density. Moreover, we achieve this finite T^* in this example by *reducing* the initial fluid content in queue, not by increasing it.

We consider the same example as before, as discussed in §D.7.1, with the same initial fluid density in service but $w(0) = 0.2$ (instead of $w(0) = 2$). Figure D.10 is the analog of Figure D.9. As shown in Figure D.10, the system becomes overloaded in the second cycle and stays overloaded thereafter. Moreover, the structure of the PSS is entirely different (in this case there is no critically loaded interval as in Figure D.9).

As concluded in §5.6 - 5.8, the initial fluid density in queue $q(0, x)$ does not play a role in determining the system's asymptotic behavior if the system is overloaded for all $t \geq 0$, by the ALOM property in Theorem 5.9. In this example, however, $q(0, x)$ is also critical, because it determines the behavior of b as well.

By a minor modification of the reasoning used in §D.7.1, we can show that the system is overloaded for all $t \geq 1/\mu$. Let $0 \leq t_1 \leq 1/\mu$ be the time at which the system switches from overloaded to underloaded intervals in $[0, 1/\mu]$. First, we can establish a similar (strict) monotonicity result. With $w(0) = 0.2$, we can show that $w(1) \approx 0.3 > w(0)$, which implies that $Q(1/\mu + t_1) > 0$. Since $\sigma(t + 1/\mu) = b(t, 0)$ for $0 \leq t \leq 1/\mu$, we have

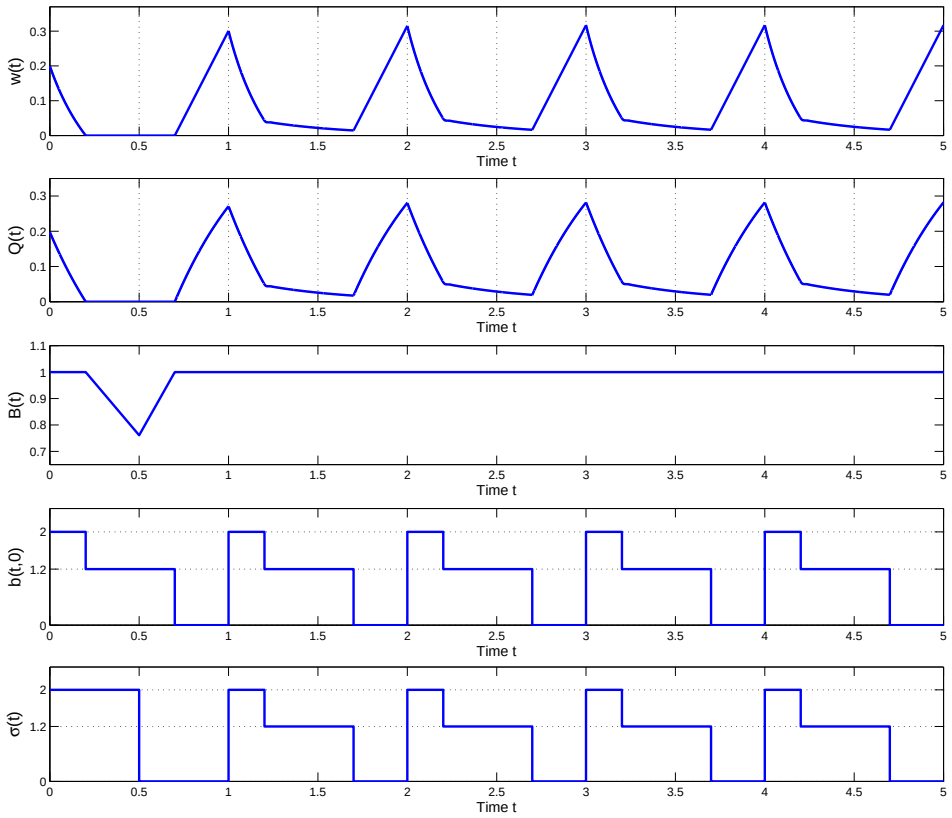

Figure D.10: The dynamics of the system performance of the example in Theorem 5.12 that has the same initial fluid density in service but $w(0) = 0.2$ instead of $w(0) = 2$.

$b(t + 1/\mu, 0) = b(t, 0)$. Therefore, the system is overloaded in $[1/\mu, 2/\mu]$. Using an inductive argument, we can show that $w(n+1) > w(n)$ and $\sigma(t+n/\mu) = b(t+n/\mu, 0) = b(t, 0)$ so that the system is overloaded in $[n, n + 1]$ for all $n \geq 1$.

D.8 More on First Passage Times

As an analog of Example 5.1 in §5.3, below we give another counterexample for first passage times with $B(0) < 1$.

Example D.1 (*counterexample on first passage times with* $B(0) < 1$) Suppose that $\lambda > \mu = 1$. Let $b(0, x) = \lambda$ for $1 - (1/\lambda) \leq x \leq 1 - 1/2\lambda$ and $b(0, x) = 0$ otherwise, so that

$B(0) = 1/2$, $b(t,0) = \lambda$, $0 \le t < 1/\lambda$, and $b(t,0) = 0$, $1/\lambda \le t < 1$, $B(t) = 1/2 + \lambda\, t$ for $0 \le t \le 1/2\lambda$ and $B(t) = 1$ for $t > 1/2\lambda$. Therefore, $T^* = t^* = 1/2\lambda$.

For $n \ge 1$, let $\{B_n(0,y) : 0 \le y \le 1\}$ be deterministic. To be a legitimate sample path for a queueing system, $B_n(0,y)$ must be nondecreasing and integer-valued as well as satisfy $0 \le B_n(0,y) \le n$. Thus, let $B_n(0,y) \equiv \lfloor B_n^f(0,y) \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $\bar{B}_n^f(0,y) \equiv n^{-1} B_n^f(0,y) \equiv \int_0^y b_n(0,x)\, dx$, where $b_n(0,x) = ((n+1)/n)\lambda$, $1-((n-1)/n\lambda) \le x \le 1-((n-1)/2n\lambda)$, and $b_n(0,x) = 0$ otherwise. First, observe that $\bar{B}_n^f(0,1/\mu) = (n^2-1)/2n^2 < 1/2$ for all $n \ge 1$. Second, observe that we have $0 \le \bar{B}_n^f(0,y) - \bar{B}_n(0,y) \le 1/n$ for all y and n. Hence, $\bar{B}_n(0,1/\mu) \le \bar{B}_n^f(0,1/\mu) < 1/2$ for all $n \ge 1$. Nevertheless, $\bar{B}_n(0,\cdot) \to B(0,\cdot)$ as $n \to \infty$. On the other hand, consider a deterministic arrival process with rate $n\,\lambda$. Then $B_n(1/2\lambda) = B_n(0) + N_n(1/2\lambda) = \lfloor (n^2-1)/2n^2 \rfloor + \lfloor (n-1)/2 \rfloor = n-1 < n$ (note there is no departure in $[1, 1/2\lambda]$). Also, $S_n(t) - S_n(1/2\lambda) = \lfloor (n+1)\lambda\,(t-1/2\lambda) \rfloor \ge \lfloor n\,\lambda\,(t-1/2\lambda) \rfloor = N_n(t) - N_n(1/2\lambda)$ for $(n-1)/2n\lambda \le t \le (n-1)/n\lambda$. Therefore, the system is underloaded for $0 \le t \le 1/\lambda$. Hence, $T_n = T_n^* = 1/\lambda$ for all $n \ge 1$, in contrast to $t^* = T^* = 1/2\lambda$.

D.9 A Two-Point Service Distribution

We next generalize the PSS result of the $G/D/s + GI$ fluid queue discussed in §5.8 to the $G/GI/s + GI$ model with a special two-point service-time distribution, in particular, to a two-point distribution where one of the two points is 0. We also give an analog of Corollary 5.8 where analytic expressions for the PSS functions are available when the system is

initially empty and the abandonment distribution is exponential. The proofs are similar to the proofs of Theorem 5.11 and Corollary 5.8.

Corollary D.1 ($\mathit{PSS\ for\ the\ overloaded\ } G/D/s+GI\ \mathit{fluid\ model}$) *Consider the stationary* $G/GI/s+GI$ *fluid model with parameter* (λ, μ, p, s, F) *where* $\rho \equiv \lambda/s\mu > 1$ *and the service distribution* G *is a two-point distribution with* $P(X = 1/p\mu) = p$ *and* $P(X = 0) = 1 - p$ *for* $0 < p \leq 1$ *such that the mean service time is* $1/\mu$*. Suppose that Assumption* 5.7 *is satisfied. If* $b(T^*, x) = s\mu$*,* $0 \leq x \leq 1/\mu$*, then there exists a constant function* $\mathcal{P}^*$ *such that*

$$\|\Psi_\tau^{(n)}(\mathcal{P}) - \mathcal{P}^*\| \to 0 \quad \textit{as} \quad n \to \infty. \tag{D.11}$$

for all $\tau > 0$*. Otherwise, the fluid performance* $\mathcal{P}$ *is asymptotically periodic with period* $1/\mu$*, i.e., there exists a periodic function* $\mathcal{P}^*$ *with period* $1/\mu$ *such that* (D.11) *holds for* $\tau \equiv 1/\mu$*.*

Corollary D.2 ($\mathit{explicit\ expressions\ for\ the\ PSS\ with\ the\ special\ two\text{-}point\ service\ times}$) *Consider the* $G/D/s+M$ *fluid queue with two-point service distribution given in Corollary D.1. If* $\rho \equiv \lambda/s\mu > 1$ *and the system is initially empty, then the system is overloaded in the PSS with performance functions given in two parts (*$[0, 1/p\mu - s/p\lambda]$ *and* $(1/p\mu - s/p\lambda, 1/p\mu]$*) of a cycle* $0 \leq t \leq 1/p\mu$*:*

(a) *In the first part of the PSS cycle, (i.e., for* $0 \le t \le 1/p\mu - s/p\lambda$*),*

$$\begin{aligned} w(t) &= t + \tilde{w}, \\ Q(t) &= \frac{\lambda}{\theta}\left[1 - \left(\frac{1 - e^{-\theta\, s/p\lambda}}{1 - e^{-\theta/p\mu}}\right) e^{-\theta\, t}\right], \\ b(t,x) &= \lambda \cdot 1_{\{t \le x \le t + s/p\lambda\}}, \\ \sigma(t) &= b(t,0) = 0, \end{aligned}$$

where

$$\tilde{w} = \frac{1}{\theta} \log\left(\frac{1 - e^{-\theta/p\mu}}{1 - e^{-\theta\, s/p\lambda}}\right) \ge 0, \tag{D.12}$$

(b) *In the second part of the PSS cycle, (i.e., for* $1/p\mu - s/p\lambda < t \le 1/p\mu$*),*

$$\begin{aligned} w(t) &= -\frac{1}{\theta} \log\left(1 + \left(\frac{1 - e^{\theta(1/\mu - s/\lambda)/p}}{1 - e^{-\theta/p\mu}}\right) \cdot e^{-\theta\, t}\right), \\ Q(t) &= \frac{\lambda}{\theta}\left(\frac{e^{\theta(1/\mu - s/\lambda)/p} - 1}{1 - e^{-\theta/p\mu}}\right) e^{-\theta\, t} \\ b(t,x) &= \lambda \cdot 1_{\{0 \le x \le t - 1/p\mu + s/p\lambda\} \cup \{t \le x \le 1/p\mu\}}, \\ \sigma(t) &= b(t,0) = \lambda. \end{aligned}$$

Moreover, for $0 \le t \le 1/p\mu$*,*

$$B(t) = s, \quad q(t,x) = \lambda \cdot 1_{\{0 \le x \le w(t)\}}, \quad \alpha(t) = \theta\, Q(t).$$

Proof. In a cycle $[0, 1/p\lambda]$, $w(t) = \tilde{w} + t$ for $0 \le t \le 1/p\mu - s/p\lambda$ and $w(t)$ solves ODE $w'(t) = 1 - 1/e^{-\theta w(t)}$ with $w(1/p\mu - s/p\lambda) = \tilde{w} + 1/p\mu - s/p\lambda$ for $1/p\mu - s/p\lambda \le t \le 1/p\lambda$, where $\tilde{w} \ge 0$ is both the starting and the ending value of $w(t)$ in each cycle. Similar to the proof of Corollary 5.8, solving this ODE in $[1/p\mu - s/p\lambda, 1/p\mu]$ and set $w(1/p\mu) = \tilde{w}$ yields (D.12). □

Remark D.1 *Theorem 5.11 and Corollary 5.8 in Chapter 5 arise as special cases of Corollary D.1 and D.2 when* $p = 1$.

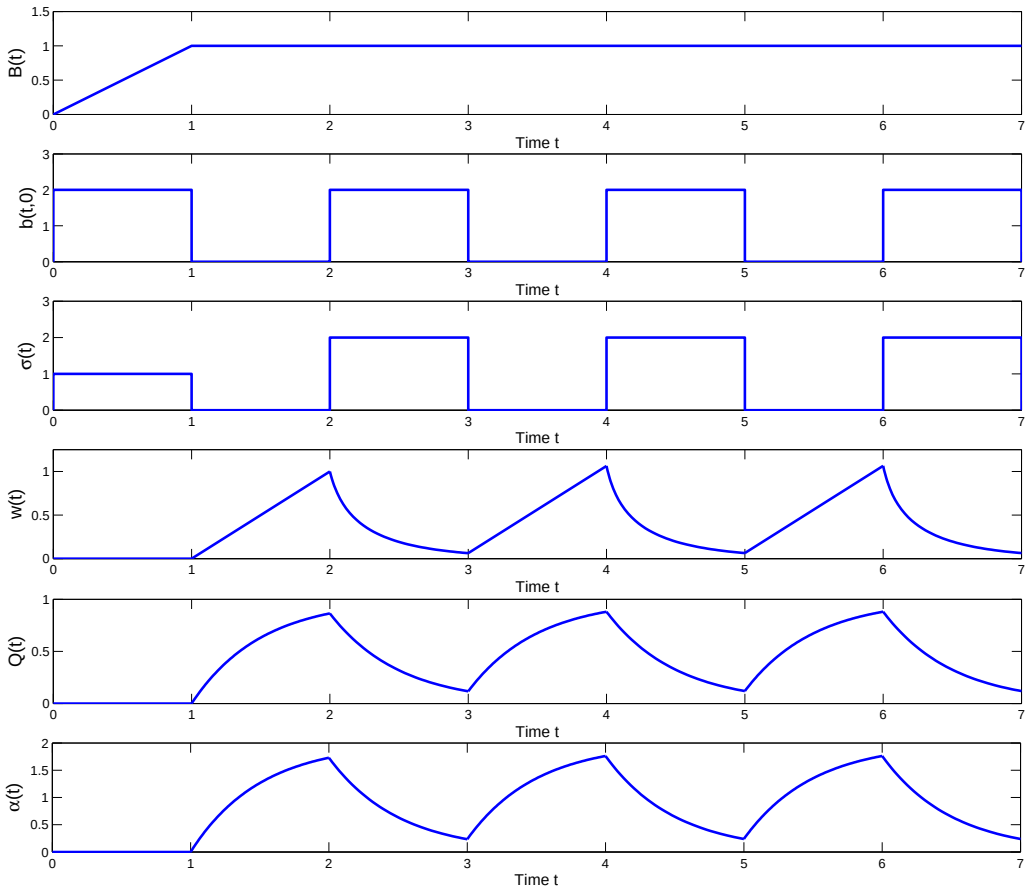

Figure D.11: Performance of the fluid model with the special two-point service distribution and $s = \mu = 1$, $p = 1/2$, $\lambda = \theta = 2$.

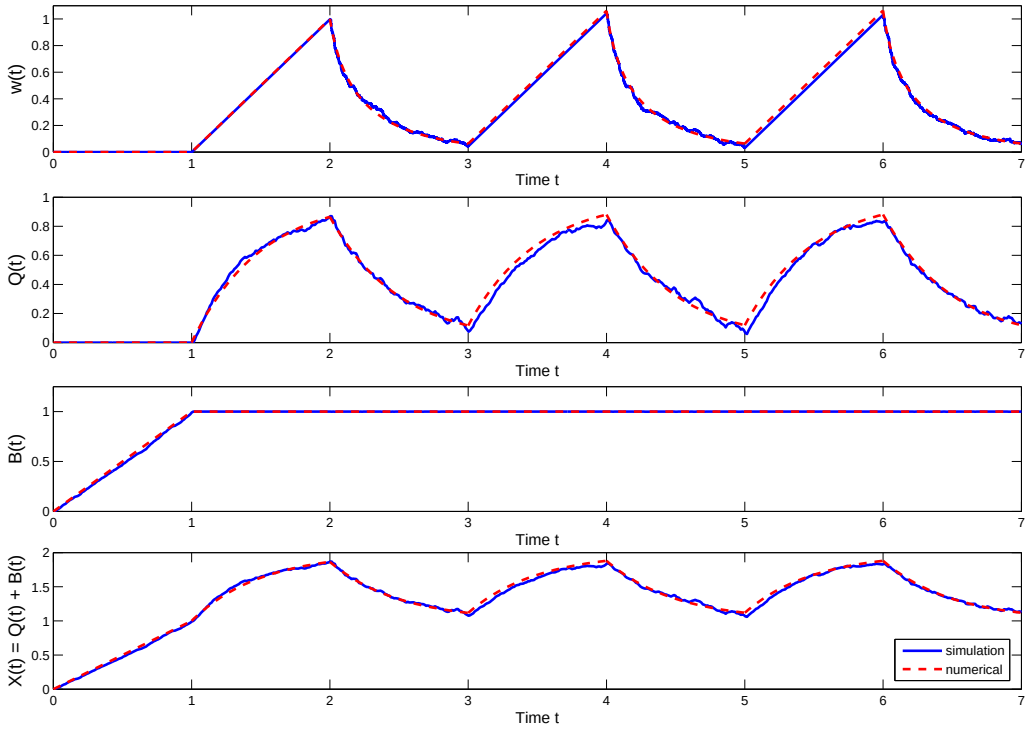

Figure D.12: A comparison of the fluid model with the special two-point service times with a simulation of a corresponding large-scale queue system.

We next compare the fluid performance with simulation estimations of large-scale queueing systems. We consider the overloaded ($\rho > 1$) $G/GI/s + M$ example with two-point service distribution such that $P(X = 1/p\mu) = p$ and $P(X = 0) = 1 - p$. Let the system be initially empty. We plot the system performance ($Q(t)$, $B(t)$, $w(t)$, $b(t, 0)$, $\alpha(t)$, $\sigma(t)$) in Figure D.11. We let $\lambda = \theta = 2$, $p = 1/2$ and $s = \mu = 1$. We have $\tilde{w} \approx 0.0635$ when $\theta = 2$ from (D.12), which can be verified by Figure D.11.

In Figure D.12 we compare our fluid approximation (the dashed red lines) with simulation estimates (the solid blue lines) of a large-scale $G/GI/s + M$ queueing system that has arrival rate $n\,\lambda$ and $n\,s$ servers. We plot (i) the elapsed waiting time of the customer at the head of the line $W_n(t)$, (ii) the scaled number of customers waiting in queue $\bar{Q}_n(t) \equiv Q_n(t)/n$ and (iii) the scaled number of customers in service $\bar{B}_n(t) \equiv B_n(t)/n$.

We plot single sample paths of these processes with $n = 1000$. Figure D.12 shows that the fluid approximation is effective.

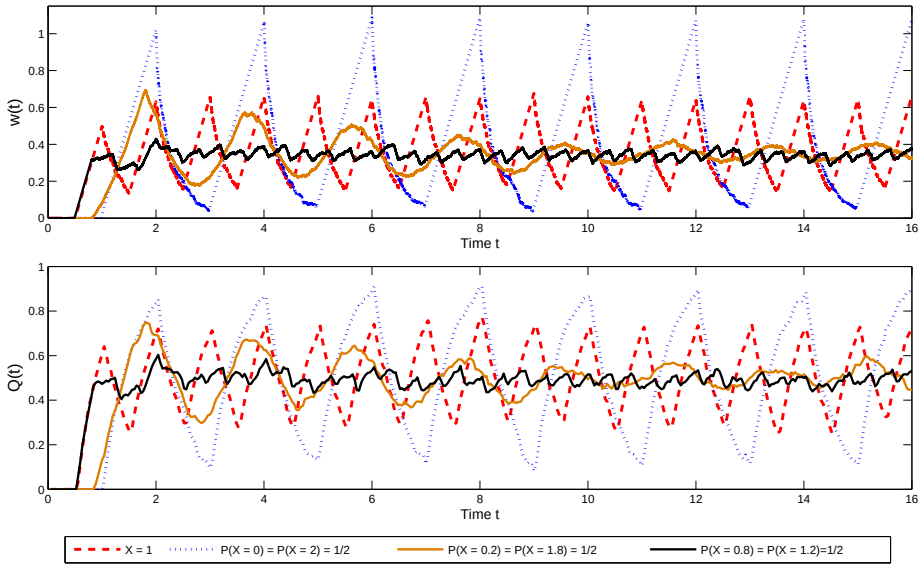

Figure D.13: A comparison of simulations of large-scale queue systems with two-point service-times distributions, all having mean 1.

However, from simulation experiments of corresponding queueing models, we conclude that the fluid model with other kinds of two-point service distributions must not converge to a PSS.

To illustrate, in Figure D.13, we plot single sample paths of processes W_n and Q_n of four two-point distributions: (a) $P(S = 1) = 1$ (red dashed curves), (b) $P(S = 0) = P(S = 2) = 1/2$ (blue dashed curves), (c) $P(S = 0.2) = P(S = 1.8) = 1/2$ (yellow solid curves) and (d) $P(S = 0.8) = P(S = 1.2) = 1/2$ (black solid curves), with $n = 1000$ in interval $[0, 16]$. The traffic intensity is $\rho = \lambda/n\mu = 2$ here. Figure D.13 shows that the periodic structure is preserved only for case (a) and (b), where he have established periodic behavior of the associated fluid model. Cases (c) and (d) involve two-point distributions, but the periodic structure fades away very quickly and the fluctuations decrease substan-

tially. Thus we conclude that the corresponding fluid models must not have asymptotically periodic structure.

D.10 Nearly Deterministic Service Times

It is natural to wonder to what extent our results for deterministic service times apply to other service-time distributions that are nearly deterministic, but not fully deterministic. We investigated this question by conducting simulation experiments of corresponding queueing systems with nearly deterministic service times.

For the experiments reported here, as before, we consider the $M/GI/n+M$ queueing model with $\lambda = 2$, $\mu = 1$ and $\theta = 2$, but now we let the service-time distribution be nearly deterministic. For all examples, $E[S] = 1/\mu = 1$ and we make $Var[S]$ small, where S is a generic service time.

In our examples now we consider two kinds of service-time distributions, both of which have small variance: (i) Erlang-N and (ii) a two-point distribution, taking the values $1/\mu \pm \delta$ with probability $1/2$. For the Erlang-N service times, the variance (and C^2) is $Var(S) = 1/N$. We plot single sample paths of process W_n with $N = 100$ and $N = 5000$ in Figure D.14, with smaller n ($n = 100$) and larger T ($T = 100$). The periodic behavior is preserved for the case $N = 5000$ but not for $N = 100$.

For the two-point distribution at $1/\mu \pm \delta$ with 1/2 probability, the variance $Var(S) = \delta^2$. We plot single sample path of process W_n with $\delta = 0.1$ and $\delta = 0.01$ in Figure D.15, with $n = 100$, $T = 100$. Again, the periodic behavior is preserved for the case $\delta = 0.01$ but not for $\delta = 0.1$.

From these experiments, we conclude, first, that over suitably short finite intervals, both the large-scale many-server queueing systems and the approximating fluid models with nearly deterministic service-time distributions should behave much like the fluid model

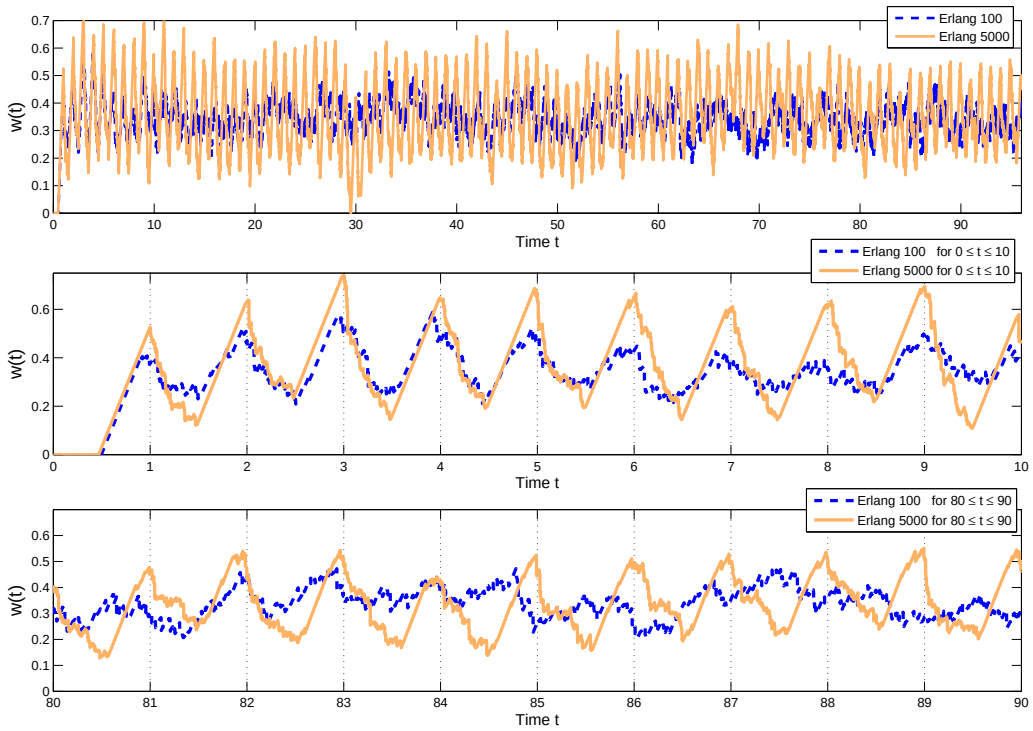

Figure D.14: Simulation estimates of the head-of-line waiting times W_n in an $G/E_N/s+M$ many-server queue with Erlang-N service, with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $\rho = 2$, $n = 100$, $T = 100$ in two cases: (i) $N = 100$; (ii) $N = 5000$.

with deterministic service times and, second, that the asymptotic behavior of the approximating fluid model will not be periodic. We conclude that a small amount of variability in the service time distribution will eventually break up the periodic behavior (provided of course we do not have the special two-point distribution considered in the previous section).

More generally, we conclude that the quality of the approximation provided by the fluid model with D service over finite time intervals $[0, T]$ should improve as the service-time distribution becomes more nearly deterministic, e.g., as the variance $Var(S)$ decreases. We conjecture that again the order of the limits cannot be interchanged: If we first let $Var(S) \downarrow 0$, e.g., by letting $N \uparrow \infty$ in the E_N distribution, and then afterwards let $t \to \infty$, then we have the asymptotic PSS established in Chapter 5. On the other hand, if we first let $T \to \infty$ for any fixed N in the Erlang E_N distribution, and then let $N \uparrow \infty$, then our

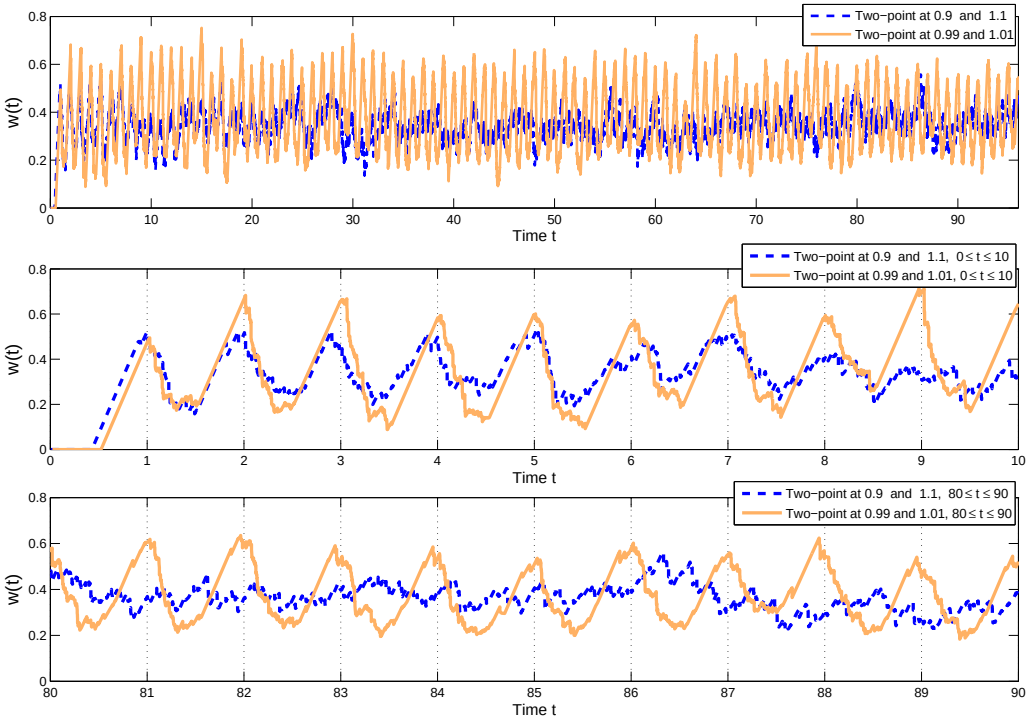

Figure D.15: Simulation estimates of the head-of-line waiting times W_n in a $G/TP/s+M$ many-server queue with a two-point (TP) service-time distribution taking values $1/\mu \pm \delta$ with 0.5 probability, with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $\rho = 2$, $n = 100$, $T = 100$ in two cases: (i) $\delta = 0.1$; (ii) $\delta = 0.01$.

simulation experiments lead us to conjecture that the performance converges to the unique steady state of the fluid model.

Even more generally, we conclude that when s system tends to behave in a deterministic or nearly deterministic way, that the transient behavior over suitably short time intervals may not be well captured by long-run stationary or steady-state descriptions.