A Fluid Model for a Large-Scale Service System Experiencing Periods of Overloading

Yunan Liu and Ward Whitt

IEOR Department
Columbia University
{yl2342,ww2040}@columbia.edu

August 12, 2010

Abstract

Motivated by healthcare systems and customer contact centers, we introduce and analyze a deterministic fluid model that can be used to show how queue lengths and waiting times depend on model parameters in a large-scale service system that experiences periods of overloading. The main feature of the model is time-varying arrival rate and staffing, but the model also includes the realistic feature of customer abandonment with a non-exponential patience distribution. Our key assumptions are (i) that the scale is large (there are many servers) and (ii) that the system alternates between overloaded intervals and underloaded intervals. We develop algorithms to describe the time-dependent performance. For example, we determine, at each time, the content in queue that has been so for at most a specified duration, as a function of the two parameters: time and duration. We also determine the time-varying potential waiting time, i.e., the virtual waiting time of an arrival at a specified time, assuming that it will not abandon. We conduct simulations to confirm that the algorithm and the approximation are effective.

Keywords: Large-scale service systems; queues with time-varying arrivals; nonstationary queues; many-server queues; deterministic fluid model; fluid approximation; queues with abandonment; non-Markovian queues.

1 Introduction

Motivated by the need for tools to improve the performance of large-scale service systems, such as (customer) contact centers and hospitals, we introduce and analyze a deterministic fluid model that serves as an approximation for a many-server non-Markovian queueing model with time-varying

parameters; see [Aksin et al.(2007), Yom-Tov and Mandelbaum(2010)] and references therein for background on contact centers and healthcare systems, respectively. Large-scale service systems tend to be quite complicated, with multiple classes of customers and multiple pools of servers. Here we restrict attention to the basic model with a single class of customers handled by a single group of homogeneous servers, working in parallel. However, we develop methods that are sufficiently tractable that they can be extended to more complicated service networks; indeed, in a sequel to this paper we develop and analyze a corresponding network of fluid queues. Our fluid model follows a well established tradition in queueing theory, as in [Newell(1982)].

Specifically, this paper is an extension of [Whitt(2006)], which developed a deterministic fluid model to approximate the steady-state performance of a stationary $G/GI/s+GI$ queueing model, having a general stationary arrival process (the initial G), independent and identically distributed (i.i.d.) service times with a general cumulative distribution function (cdf) G (the first GI), a large number s of servers and customer abandonment from queue with i.i.d. patience times with a general cdf F (the final $+GI$). Abandonment is now recognized as an important feature, e.g., see [Garnett et al.(2002), Zeltyn and Mandelbaum(2005)], and patience times are typically non-exponential, e.g., see [Brown et al.(2005)]. Comparisons with simulation in Tables 1-3 there showed that the approximations can be very useful when the system is overloaded. Some degree of overloading is not uncommon, because the abandonment acts to keep the system stable.

Here we consider the analogous $G_t/M/s_t+GI$ fluid model, restricting attention to exponential service, but now including time-varying arrival rate and staffing (number of servers). We develop algorithms to calculate the performance functions. In doing so, we provide for the first time a full description of the transient behavior, even for the stationary $G/M/s+GI$ fluid model. The fundamental evolution equations, here in (2.5), are the same, but the performance depends on a boundary waiting time (BWT), which is characterized here as the solution of an ordinary differential equation (ODE); see Theorem 4.1. We also determine the potential waiting time, i.e., the virtual waiting time of an arrival if that arrival would elect never to abandon; see Theorems 4.3 and 4.4.

Our main goal here is to contribute to the techniques for analyzing service systems with the important and realistic feature of time-varying arrivals and staffing; see [Green et al.(2007)] for background. Again, there is precedent in [Whitt(2006)], because a discrete-time model with time-varying (and state-dependent) arrival rate was considered in §6 there. In contrast, here we develop a smooth model, which leads to the ODE.

Most queueing models are stochastic, because a primary cause of congestion is random fluctu-

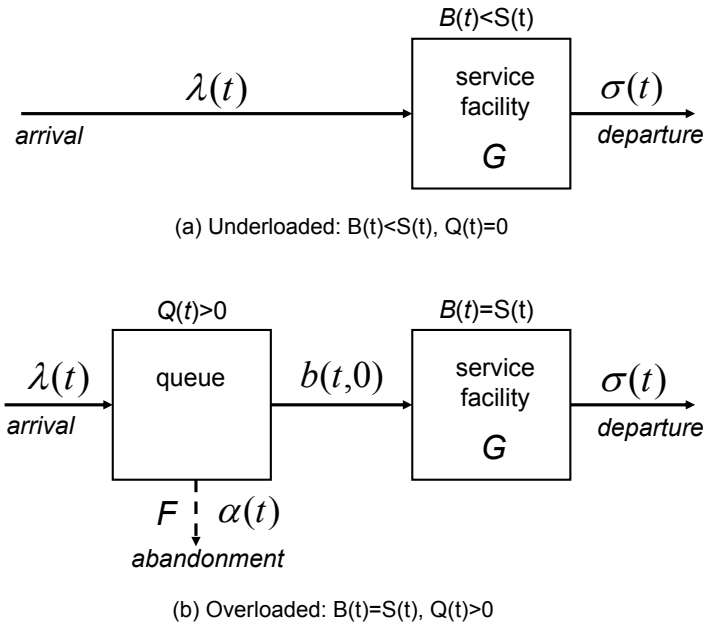

Figure 1: Fluid model in underloaded and overloaded intervals.

ation in arrivals and service. Our deterministic model can be useful when the *systematic* variation in the arrival rate dominates the *stochastic* variation in the arrivals and service. The analysis here applies to a system that is either overloaded or underloaded for an extensive period of time, but an innovation in our approach is to consider systems that alternate between overloaded intervals and underloaded intervals. The behavior in these two cases is depicted in Figure 1. The total fluid content in service (queue) is given by $B(t)$ ($Q(t)$). When the system is underloaded, the total system fluid content $X(t) \equiv B(t) + Q(t)$ is less than the service capacity $s(t)$, so that there is no fluid waiting and external input flows directly into service at time-varying rate $\lambda(t)$. When the system is overloaded, there is no spare service capacity ($B(t) = s(t)$), so that the input is buffered in a queue, where abandonment occurs. Service completion and abandonment occur at rate $\sigma(t)$ and $\alpha(t)$. Fluid flows from the queue into the service facility at rate $b(t, 0)$. We rigorously define these fluid functions in §2.

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 responds 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 deterministic fluid model may capture the essential performance.

We show that all fluid performance functions are well defined and that they can be computed with a simple algorithm, summarized in §4.4. Figure 2 shows key fluid performance functions of an $M_t/M/s + M$ example with sinusoidal arrival rate (see (5.1)) that we consider in §5. 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.

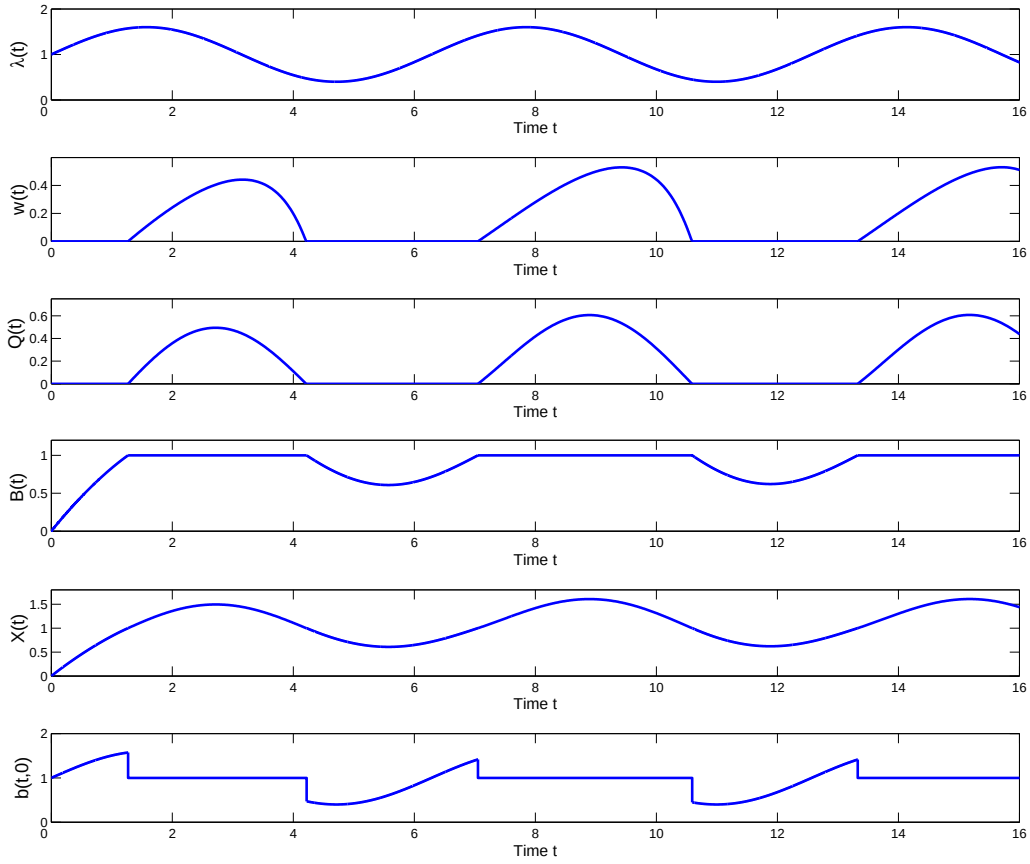

Figure 2: The performance functions of the $M_t/M/s + M$ fluid model with sinusoidal arrival-rate function: (i) arrival rate $\lambda(t)$; (ii) waiting time $w(t)$; (iii) fluid in buffer $Q(t)$; (iv) fluid in service $B(t)$; (v) total fluid $X(t)$; (vi) rate into service $b(t, 0)$.

It is of course 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 3. (In the plot, the queueing content processes are scaled by dividing by $n = 1000$, so that s remains at 1.)

Of course, most service systems have fewer servers. It is thus important that the fluid approx-

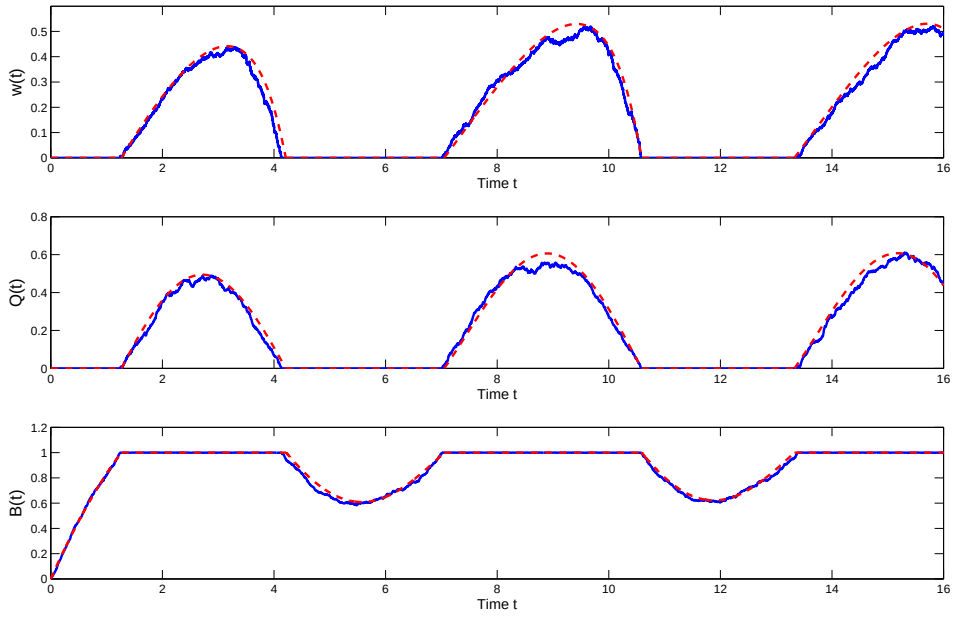

Figure 3: Performance of the $M_t/M/s+M$ fluid model (dashed lines) compared with simulation results (solid lines): one sample path of the scaled queueing model for $n = 1000$.

imation 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 4. See §5 for the details.

If staffing can be adjusted dynamically, then alternating overloaded and underloaded intervals may not occur. Then it may be better to use analysis techniques suitable for systems that are critically loaded or nearly critically loaded at all times. Figure 4 shows that there is a degradation in approximation accuracy in the critically loaded regions.

Here is how this paper is organized. We start in §2 by defining the $G_t/M/s_t+GI$ fluid model and specifying key regularity conditions. In §3 we characterize performance during an underloaded interval. (The only difficulties occur in overloaded intervals.) In §4 we describe the performance in an overloaded interval; in §4.4 we summarize the resulting algorithm. In §5 we compare results of the numerical algorithm developed for an exponential service-time distribution with simulations of corresponding large-scale queueing models. (The fluid model can be related to the queueing model directly, but it can be useful to make the connection via many-server heavy-traffic scaling. We

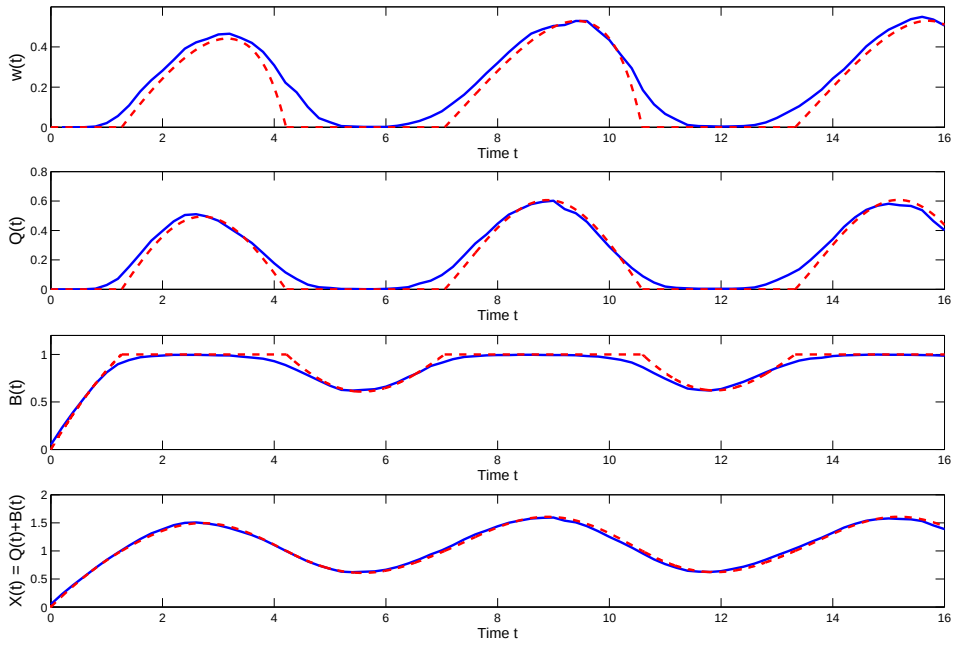

Figure 4: Performance of the $M_t/M/s+M$ fluid model (dashed lines) compared with simulation results (solid lines): an average of 200 sample paths of the scaled queueing model based on $n = 20$.

explain that connection in §5.) In §6 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 §7 we provide proofs of the main results, Theorems 4.1, 4.3, 4.4 and 6.1. Finally, in §8 we draw conclusions. Additional supporting material appears in an appendix.

2 The $G_t/M/s_t + GI$ Fluid Model

In this section we define the deterministic fluid model. There is a service facility and a waiting room or queue. 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 is also a time-dependent staffing (service capacity) function $s \equiv \{s(t) : t \geq 0\}$.

There are service-time and abandon-time cumulative distribution functions (cdf's) G and F, respectively, with probability density functions (pdf's) g and f, satisfying

$$G(x) = \int_0^x g(u)du \quad \text{and} \quad F(x) = \int_0^x f(u)du, \quad x \geq 0. \tag{2.1}$$

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. We consider the special case of *exponential service* $g(x) \equiv e^{-x}$, $x \geq 0$. 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

$$Q(t, y) = \int_0^y q(t, x)\, dx \quad \text{and} \quad B(t, y) = \int_0^y b(t, x)\, dx, \quad y \geq 0, \tag{2.2}$$

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 (2.2) 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 inital 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. Let $\mathbb{C}_p^1$ denote the set of differentiable functions with derivatives that belong to $\mathbb{C}_p$.

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 [Billingsley(1999)].) Together with Assumption 2.1, that implies the finite-content property in Assumption 2.1 holds for all t: $B(t) \le B(0) + \Lambda(t) < \infty$ and $Q(t) \le Q(0) + \Lambda(t) < \infty$ for all $t \ge 0$.

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 $t \ge 0$,*

$$(B(t) - s(t))Q(t) = 0 \quad \textit{and} \quad B(t) \le s(t). \tag{2.3}$$

In general, there is no guarantee that a staffing function s is feasible; i.e., having the property that no fluid that has entered service must leave without completing service, because we allow s to decrease. 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 §6.

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) $w(t)$ such that

$$w(t) \equiv \inf\{x > 0, q(t, y) = 0 \ for\ all\ y > x\}. \tag{2.4}$$

Clearly, first, $w(t) \geq 0$ and, second, $w(t) > 0$ if and only if $Q(t) > 0$. (Equation (2.4) 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 §4.2.)

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 leaves the queue from the right boundary of $q(t, x)$.

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

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

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 §4.

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

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

(2.8)

$$E(t) \equiv \int_0^t b(u,0)\, du, \quad t \geq 0. \tag{2.9}$$

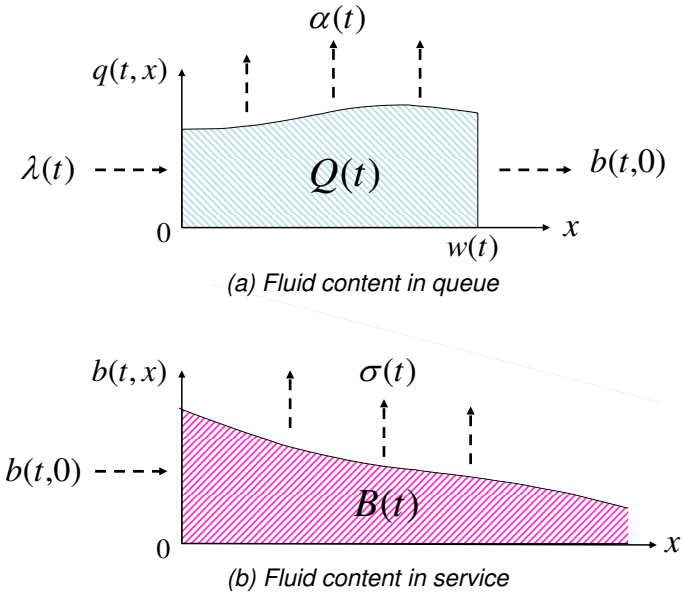

Figure 5: (a) The fluid in queue, (b) The fluid in service.

$$S(t) \equiv \int_0^t \sigma(u)\,du, \quad \text{where} \quad \sigma(t) \equiv \int_0^\infty b(t,x)\,dx, \quad t \geq 0 \tag{2.10}$$

We have now completed the definition of the $G_t/M/s_t + GI$ fluid model (with the exception of (w, q, v), for which more is given in §4.2 and 4.3; Figure 5 provides a pictorial summary. Our goal now is to fully characterize the six-tuple $(b, q, w, v, \sigma, \alpha)$ given the model parameters (λ, s, 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

$$(i) \quad Q(t_1) > 0 \quad \text{or} \quad (ii) \quad Q(t_1) = 0, \quad B(t_1) = s(t_1) \quad \text{and} \quad \lambda(t_1) > s'(t_1) + \sigma(t_1), \tag{2.11}$$

and ends at the *overload termination time*

$$T_1 \equiv \inf\{u \geq t_1 : Q(u) = 0 \quad \text{and} \quad \lambda(u) \leq s'(u) + \sigma(u)\}. \tag{2.12}$$

Case (ii) in (2.11) 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

$$(i) \quad Q(t_2) < 0 \quad \text{or} \quad (ii) \quad Q(t_2) = 0, \quad B(t_2) = s(t_2) \quad \text{and} \quad \lambda(t_2) \le s'(t_2) + \sigma(t_2), \tag{2.13}$$

and ends at *underload termination time*

$$T_2 \equiv \inf \{u \ge t_2 : B(u) = s(u) \quad \text{and} \quad \lambda(u) > s'(u) + \sigma(u)\}. \tag{2.14}$$

As before, case (ii) in (2.13) 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.

Assumption 2.7 (*finitely many switches between intervals in finite time*) *For each underloaded interval, $T_2 > t_2$ for t_2 in* (2.13) *and T_2 in* (2.14)*, 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 the both the present model and the extension to time-varying Markovian service (M_t), we provide sufficient conditions for Assumption 2.7 to be satisfied in a sequel to this paper. However, 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 paper. We will introduce additional regularity assumptions as needed, starting in §4. We now determine the performance, first considering an underloaded interval.

3 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. We assume that the interval starts at a time t_2 satisfying (2.13). We do not need to know in advance the termination time T_2 in (2.14) when the underloaded interval ends. Instead, we can assume that the system is underloaded over the full interval $[0, \infty)$ and then calculate T_2. Henceforth in this section, without loss of generality, we replace t_2 and T_2 by 0 and T, respectively. It suffices to assume that there is unlimited service capacity.

For an underloaded interval, we can also treat general GI service, so we state results in this section in the more general form. As in §2, we assume that the pdf g is in $\mathbb{C}_p$ and that $\bar{G}(x) > 0$ for all x. Thus the hazard rate $h_G(x) \equiv g(x)/\bar{G}(x)$ is well defined for all x. (We exploit this generality in the next section; see Proposition 4.2.) It is easy to read off the results for our exponential case of $g(x) \equiv e^{-x}$. If the $G_t/GI/s_t + GI$ fluid model is underloaded, then there is not queue, and so no abandonment. Then the model is is equivalent to the associated $G_t/GI/\infty$ fluid model. Because of the linear structure in the $M_t/GI/\infty$ queueing model, the service content in the $G_t/GI/\infty$ fluid model coincides exactly with the mean number of busy servers in the associated $M_t/GI/\infty$ queueing model; e.g., see [Eick et al.(1993), Massey and Whitt(1993)]. Thus, much is already known about this underloaded case. With that identification, this section is primarily a review. Remark 2.3 of [Massey and Whitt(1993)] observes that the formulas for the mean number of busy servers do not depend on the Poisson assumption, and thus apply in the full generality of this paper. See [Pang and Whitt(2010), Reed and Talreja(2009)] for both many-server heavy-traffic limits and descriptions of the fluid limit and its diffusion refinement for the two-parameter processes.

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

$$\begin{aligned} b(t,x) &= \bar{G}(x)\lambda(t-x)1_{\{x\leq t\}} + \frac{\bar{G}(x)}{\bar{G}(x-t)}b(0,x-t)1_{\{x>t\}} \leq \lambda(t-x) + b(0,x-t), \qquad (3.1) \\ B(t,y) &= \int_0^{t\wedge y} \bar{G}(x)\lambda(t-x)\,dx + \int_0^{(y-t)\vee 0} \frac{\bar{G}(x+t)}{\bar{G}(x)}b(0,x)\,dx, \\ B(t) &= \int_0^{t} \bar{G}(x)\lambda(t-x)\,dx + \int_0^{\infty} \frac{\bar{G}(x+t)}{\bar{G}(x)}b(0,x)\,dx \leq \Lambda(t) + B(0) < \infty, \quad 0 \leq t < T. \end{aligned}$$

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 5. Hence, Proposition 3.1 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 3.1 (*discontinuity at* $t = x$) From (3.1), 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 3.2 (*the generic scalar transport PDE*) If, in addition to the assumptions of Proposition *3.1*, 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 B.

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. It is easy to see that hazard rate order implies ordinary stochastic order via the representation

$$\bar{G}(x) = e^{-\int_0^x h_G(u)\, du}, \quad x \geq 0. \tag{3.2}$$

Proposition 3.2 (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 (3.1) after applying (3.2) to write $\bar{G}(x)/\bar{G}(x-t) = \exp\{-\int_{x-t}^{x} h_G(u)\,du\}$. ∎

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 [Eick et al.(1993), Massey and Whitt(1993)].) That allows us to directly construct stationary versions, including periodic versions, if that is warranted.

Proposition 3.3 (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 $\int_0^\infty \bar{G}(x)\lambda(t-x)\,dx, < \infty$, 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 [Eick et al.(1993), Massey and Whitt(1993)]; 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 [Pang and Whitt(2010)].

4 An Overloaded Interval with Exponential Service

We now consider the system during an overloaded interval, but now strongly exploiting the assumption of exponential (M) service. (In a sequel to this paper we will report corresponding results for a large class of non-exponential service distributions; then we exploit a fixed point theorem in order to calculate the service content density b.) We assume that the overloaded interval begins at a time t_1 satisfying (2.11). Again, we do not need to know the end time T_1 in (2.12) in advance, because we can calculate it while we are calculating the performance measures q and w. Since $T_1 > t_1$, there always exists an overloaded interval of positive length. In this section, without loss of generality, we let $t_1 = 0$ and $T_1 = T$.

4.1 The Service Content Density

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. For this reasoning, it is significant that we do not need to recalculate the service content density b over $[0, T^*)$ with $T^* < T$ or the new endpoint T^* if we later find that the overload interval ends at $T^* < T$.

From (2.5), we can write down an expression for $b(t, x)$ during the overloaded interval:

$$\begin{aligned} b(t,x) &= b(t-x,0)\bar{G}(x)1_{\{x\le t\}} + b(0,x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}1_{\{x>t\}}, \\ &= b(t-x,0)e^{-x}(x)1_{\{x\le t\}} + b(0,x-t)e^{-t}1_{\{x>t\}}, \end{aligned} \tag{4.1}$$

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 4.1 (the service content in an overloaded interval) *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*

$$\sigma(t) = B(t) = s(t) \quad \textit{and} \quad b(t,0) = s'(t) + s(t) \quad \textit{for all} \quad t, \tag{4.2}$$

depending only on the staffing function s. Then b is fully characterized by (4.1) and (4.2) during an overloaded interval. Also $b(t, \cdot), b(\cdot, x) \in \mathbb{C}_p$ for all x, $t < T$.

Proof. Apply (2.10).

4.2 The Queue Performance Functions

We now turn to the queue. To do so, it is convenient to initially ignore the flow into service. Hence, 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 treat $\tilde{q}(t, x)$ just as we treated b in §3, with the general patience cdf F playing the role of the general service-time cdf G; i.e., instead of (2.5), we can write

$$\tilde{q}(t+u, x+u) = \tilde{q}(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad x \ge 0, \tag{4.3}$$

to obtain the following proposition.

Proposition 4.2 (*queue content without transfer into service in the overloaded case*) *In the overloaded case,*

$$\tilde{q}(t,x) = \lambda(t-x)\bar{F}(x)1_{\{x\le t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\}}. \tag{4.4}$$

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

Remark 4.1 Just as we observed for b in an underloaded interval in Remark 3.2, in an overloaded interval $\tilde{q}$ satisfies a version of the generic scalar transport PDE.

Paralleling Proposition 3.2, we have the following comparison result, proved in the same way.

Proposition 4.3 (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 4.1 (*from $\tilde{q}$ to q*) *Given the BWT w,*

$$\begin{aligned} q(t,x) &= \tilde{q}(t-x,0)\bar{F}(x)1_{\{x\leq w(t)\wedge t\}} + \tilde{q}(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\leq w(t)\}} \\ &= q(t-x,0)\bar{F}(x)1_{\{x\leq w(t)\wedge t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\leq w(t)\}}. \end{aligned} \tag{4.5}$$

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

Proof. Combine Proposition 4.2 and (4.5) to deduce that $q(t,\cdot) \in \mathbb{C}_p$ for all t, x.

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.,

$$E(t+\delta) - E(t) \equiv \int_t^{t+\delta} b(u,0)\,du = I(t,w(t),\tilde{q},\delta) - A(t,t+\delta), \tag{4.6}$$

where

$$I \equiv I(t,w(t),\tilde{q},\delta) \equiv \int_{w(t)-\epsilon(t,\delta)}^{w(t)} \tilde{q}(t,x)dx \tag{4.7}$$

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 (4.6). We will show in Theorem 4.1 below that, under regularity conditions, the relation in (4.6) determines an ODE for w that has a unique solution. Hence, we will show that the relation (4.6) 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 4.1 (*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 4.2 (*positive arrival rate and initial queue density*) *For all* $t \geq 0$,

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

By equation (4.4), Assumption 4.2 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 (4.8) below. The BWT w can be discontinuous if these functions are 0 over subintervals; we give examples in Appendix E. 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 §7. Let $x(t+)$ and $x(t-)$ denote the right and left limits of a function x at t, respectively.

Theorem 4.1 (*the BWT ODE*) *Consider an overloaded interval* $[0,T)$. *If Assumptions* 4.1–4.2 *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*

$$w'(t+) = \Psi(t, w(t)) \equiv 1 - \frac{b(t+,0)}{\tilde{q}(t, w(t)-)} \tag{4.8}$$

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* (4.8) *and left differentiable everywhere* (*but not necessarily differentiable*) *with value*

$$w'(t-) \quad = \quad \tilde{\Psi}(t, w(t)) \equiv 1 - \frac{b(t-,0)}{\tilde{q}(t, w(t)+)}. \tag{4.9}$$

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

Remark 4.2 (*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 $0 \le x < w(t)$.

We give closed-form formulae for some special cases in the next corollary.

Corollary 4.2 *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 Λ *in* §2, $\Lambda^{-1}(x) \equiv \inf\{y > 0 : \Lambda(y) = x\}$, *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* ($\bar{F}(x) = e^{-\theta x}$, $x \ge 0$),

$$w(t) = t - \tilde{\Lambda}^{-1}(\int_0^t b(y,0)e^{\theta y}dy), \quad 0 \le t < \tilde{t}, \tag{4.10}$$

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* $\tilde{t} \equiv \inf\{t > 0 : \tilde{\Lambda}(t) = \int_{t_1}^t b(y,0)e^{\theta y}dy\}$.

4.3 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.

The PWT v can be defined as a first passage time, with abandonment after time t computed with the input turned off. 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 (2.8),

$$A_t(u) \equiv \int_t^{t+u} \alpha_t(s)\,ds \quad \text{and} \quad \alpha_t(s) \equiv \int_{s-t}^{\infty} q(s,x)h_F\,dx, \quad s \ge t, \tag{4.11}$$

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

With (4.11), we can define $v(t)$ as

$$v(t) \equiv \inf\{u \ge 0 : E(t+u) - E(t) + A_t(u) \ge Q(t)\}, \quad t \ge 0, \tag{4.12}$$

where $E(t)$ is the amount of fluid to enter service in the interval $[0,t]$, as in (2.9), i.e., $E(t) \equiv \int_0^t b(u,0)\,du$, $t \ge 0$. 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 4.4 (dissipation of initial fluid content) *For $t \geq 0$,*

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

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 4.1, 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 4.3 (*minimum staffing level*) *There exists a constant s_L such that $s(t) \geq s_L > 0$ for all $t \geq 0$.*

Theorem 4.2 (finite PWT) *Under Assumption* 4.3*, the rate of service completion is bounded below: $\sigma(t) \geq s_L$ for all $t \geq 0$. As a consequence,*

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

Given that the PWT v is indeed bounded above as in Theorem 4.2, 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 $w(t)$.

Theorem 4.3 (the PWT v and the BWT w) *Consider an overloaded interval with Assumptions* 4.1-4.2 *holding and $w(0) = 0$. If $v(t) < \infty$ for all $t \geq 0$ (for which Assumption* 4.3 *is a sufficient condition, by Theorem 4.2), then v is the unique function in $\mathbb{D}$ satisfying the equation*

$$v(t - w(t)) = w(t) \quad \textit{or, equivalently,} \quad v(t) = w(t + v(t)) \quad \textit{for all} \quad t \geq 0, \tag{4.13}$$

as depicted in Figure 6. *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.*

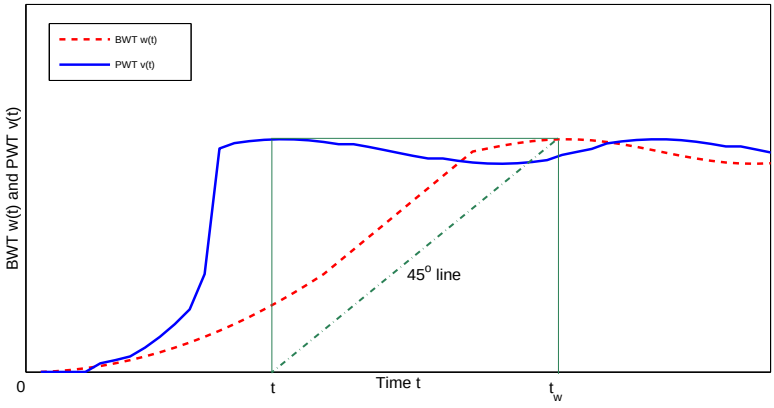

Figure 6: Potential waiting time $v(t)$ and boundary waiting time $w(t)$.

The proof of Theorem 4.3 directly gives an algorithm to compute the PWT v given the BWT w. Similarly, the second equation in (4.13) 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.

Theorem 4.4 (right derivative and ODE for v) *Under the conditions in Theorem* 4.3, *the right derivative of* v *always exists (except possibly infinite), with value*

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

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 4.3 (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

$b_\epsilon(t,0) \equiv b(t,0) + \epsilon$, $0 \le t \le T$, 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 $w' \ge w'_\epsilon$ and $v' \ge v'_\epsilon$, 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 $b(t,0) \ge b^* > 0$ 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

$$s'(t) \ge b^* - \sigma(t) = b^* - \int_0^\infty b(t,x)\,dx, \quad 0 \le t < T. \tag{4.14}$$

We now give closed-form formulae for some special cases. We omit the proof, which is similar to the proof of Corollary 4.2.

Corollary 4.3 *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 $\tilde{t} \equiv \inf\{t > 0 : \tilde{\Gamma}(t) = \int_{t_1}^t \lambda(y)e^{\theta y}dy\}$.

4.4 Overview of the Total Algorithm

We now summarize the full algorithm for the $G_t/M/s_t + GI$ model, which has been developed in this section. 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 (2.13). We can compute the single key performance measure b directly by applying Proposition 3.1. 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.

An overloaded interval is more complicated. The algorithm starts with initial conditions as in (2.11). The algorithm begins by calculating $\tilde{q}$ via Proposition 4.2 and b and $b(t,0)$ via Proposition

4.1. In order to determine w and q, we need to calculate $b(t,0)$, but $b(t,0) = s'(t) + s(t)$. We then calculate w by solving the ODE (4.8) and then the function v via the equation (4.13), as explained in the proof of Theorem 4.3. 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 (2.11)–(2.12)). If $r(t) > 0$, then we continue the overloaded interval. Otherwise, we shift to the next underloaded interval.

We have implemented and tested the algorithm. We will report results in the next section. We present additional details about the algorithm in Appendix D.

5 A Simulation Example

We have implemented the algorithm for the $G_t/M/s_t + GI$ fluid model described in §4.4 and compared it to simulation results for associated large-scale queueing models. In this section we describe the results of one such experiment. Here we apply the algorithm to the special case of an $M_t/M/s + M$ model, having time-varying arrival rate function. In Appendix F we present additional simulation results for three alternative models, Across these four examples we consider each of the model features: (i) time-varying arrival rate, (ii) time-varying staffing function, (iii) non-exponential abandonment-time cdf, and (iv) 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.

For this initial fluid example, we consider constant staffing s. We let the arrival rate function λ be sinusoidal, i.e.,

$$\lambda(t) \equiv a + b \cdot \sin(c \cdot t), \quad t \geq 0, \tag{5.1}$$

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 $s \equiv 1$.

The figures have already been presented in §1. In Figure 2, we plot key fluid performance measures for $0 \leq t \leq T$, where $T = 16$. Figure 2 shows the alternating overloaded and underloaded intervals. 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.

We next compare this fluid approximation with computer simulations of the associated $M_t/M/s+M$ queueing system. We connect the fluid model to the queueing model by exploiting the usual many-server heavy-traffic scaling, as in [Pang et al.(2007)]. We think of the given queueing system as the n^{th} queueing system in a sequence of queueing systems, where $n \to \infty$. Let the n^{th} queueing system have arrival process $N_n \equiv \{N_n(t) : t \geq 0\}$ and staffing function $s_n \equiv \{s_n(t) : t \geq 0\}$, but let the abandon-time cdf F and service-time cdf G be fixed independent of n; ($N_n(t)$ counts the number of arrivals in the interval $[0,t]$). Let $\bar{N}_n \equiv n^{-1}N_n$ and $\bar{s}_n \equiv n^{-1}s_n$. The standard scaling has

$$\bar{s}_n \to s \quad \text{in} \quad \mathbb{D} \quad \text{and} \quad \bar{N}_n \Rightarrow \Lambda \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \tag{5.2}$$

where s and Λ are the fluid model functions. Starting with the fluid model, we let $s_n(t) = \lceil ns(t) \rceil$, the least integer greater than $ns(t)$ and the arrival-rate function be $\lambda_n(t) = n\lambda(t)$.

The non-exponential patience distribution (and service distribution, when considered) leads to considering two-parameter processes in order to develop Markov process representations; see [Kang and Ramanan(2008), Kaspi and Ramanan(2007), Pang and Whitt(2010), Reed and Talreja(2009)] and references therein (extending limits for the usual one-parameter processes, as in [Mandelbaum et al.(1998), Puhalskii(2008), Pang et al.(2007)] and references therein). In the n^{th} queueing model, let $\tilde{B}_n(t,y)$ be the number of customers in service at time t that have been in service for time less than or equal to y, and let $\tilde{Q}_n(t,y)$ be the number of customers in queue at time t that have been in queue for time less than or equal to y, for $y \geq 0$. Then, we form the scaled processes

$$\bar{B}_n(t,y) \equiv \frac{\tilde{B}_n(t,y)}{n} \quad \text{and} \quad \bar{Q}_n(t,y) \equiv \frac{\tilde{Q}_n(t,y)}{n} \quad \text{for} \quad t \geq 0 \quad \text{and} \quad y \geq 0. \tag{5.3}$$

let $W_n(t)$ be the elapsed waiting time of the customer at the head of the queue at t, left unscaled.

We let $n \equiv 1000$ with the scaling in (5.2) and (5.3). 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 3 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.

When n is smaller, there are significant stochastic fluctuations, but the mean values of the queueing functions still are quite well approximated by the fluid performance functions when the system is unambiguously overloaded or overloaded, and not critically loaded or nearly critically loaded. We illustrate by considering the cases $n = 100$ and $n = 20$. In Appendix F we show the results for $n = 100$. When $n = 100$, there is definitely some stochastic fluctuation, but not too much. The fluid model provides a good rough approximation for one sample path for $n = 100$ and the average of 10 sample paths with $n = 100$ looks nearly as good as Figure 3, except at the critically loaded transition points between overloaded and underloaded intervals. The fluctuations for a single sample path are much greater for $n = 20$ (see Appendix F), but the fluid model provides a good approximation for the mean values when the system is unambiguously overloaded or overloaded, and not critically loaded or nearly critically loaded, as seen from the plot of the average of 200 independent sample paths for $n = 20$ in Figure 4. For the full distributions, we would want stochastic refinements, which remains a topic for future research.

6 Feasibility of the Staffing Function

As in §3, in this section we will consider the more general case of GI service, from which the desired conclusion for M service follows. So far, we have assumed that the staffing function s is feasible, yielding

$$b(t,0) \geq s'(t) + \sigma(t) = s'(t) + \int_0^\infty b(t,x) h_G(x)\, dx \geq 0 \quad \text{for all} \quad t \geq 0 \quad \text{such that} \quad B(t) = s(t). \tag{6.1}$$

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 $\lambda(t) \geq 0$; see (2.13). 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

$$t^* \equiv \inf\{t \in I : b(t,0) < 0\}. \tag{6.2}$$

Even though we require (6.1), 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

$$\mathcal{S}_{f,s}(t) \equiv \{\tilde{s} \in C_p^2(t) : b_{\tilde{s}}(u,0)1_{\{B_{\tilde{s}}(u)=\tilde{s}(u)\}} \geq 0, \quad 0 \leq u \leq t, \quad \text{and} \quad \tilde{s}(u) = s(u), \quad 0 \leq u \leq t^*\}, \tag{6.3}$$

for t^* in (6.2), where $b_{\tilde{s}}$ is the function b associated with the model with staffing function $\tilde{s}$.

Theorem 6.1 (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* (6.3) *for t^* in* (6.2) *such that*

$$s^* = \inf\{\tilde{s} \in \mathcal{S}_{f,s}(t^* + \delta)\}; \tag{6.4}$$

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,

$$s^*(t^* + u) \equiv \int_u^\infty b_s(t^*, x-u)\frac{\bar{G}(x)}{\bar{G}(x-u)}dx, \quad 0 \leq u \leq \delta. \tag{6.5}$$

Moreover, δ can be chosen so that

$$\delta = \inf\{u \geq 0 : s^*(t^* + u) = s(t^* + u)\}, \tag{6.6}$$

with $\delta \equiv \infty$ if the infimum in (6.6) *is not attained.*

Corollary 6.1 (minimum feasible staffing with exponential service times) *For the special case of exponential service times, i.e., with $\bar{G}(x) \equiv e^{-x}$,* (6.5) *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.

7 Proofs of the Main Results

Proof of Theorem 4.1. 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,

$$w(t+\delta) = w(t) + \delta - \epsilon(t,\delta), \tag{7.1}$$

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 7. (Note that $\delta > 0$ and $\epsilon(t,\delta) \ge 0$.) To determine $\epsilon(t,\delta)$, we apply (4.6), with (4.7). We will bound $\epsilon(t,\delta)$ in (7.3) below.

(c) (controlling the abandonment term) We will show that the abandonment term $A(t,t+\delta)$ in (4.6) 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

$$A(t,t+\delta) \le h_F^\uparrow \tilde{q}^\uparrow w^\uparrow \delta = C_1 \delta \tag{7.2}$$

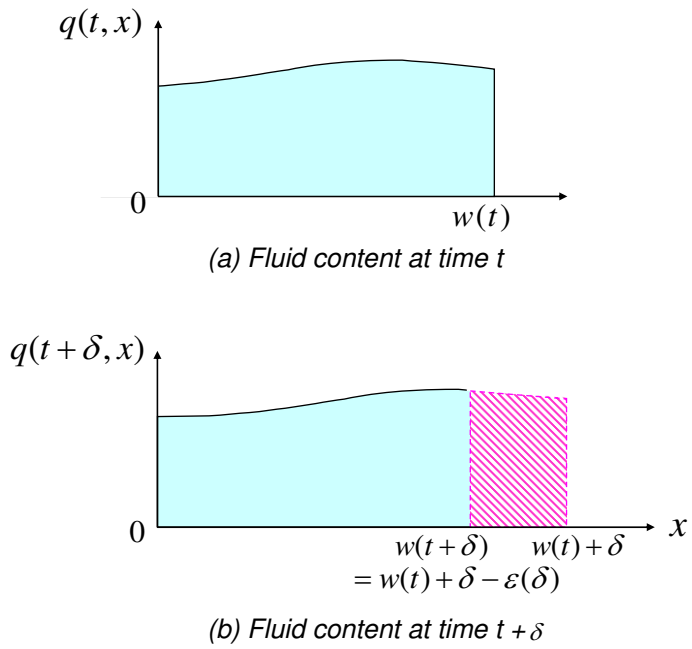

Figure 7: The boundary of the waiting time $w(t)$ under FCFS.

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 (7.1), we can show that w is Lipschitz continuous by showing that $\epsilon(t,\delta) \le C\delta$ for some constant C. Recall that $b(\cdot,0)$ is continuous by Theorem **??**. Hence, $\|b(\cdot,0)\|_T < \infty$, so that there exists a constant C_2 such that $E(t+\delta) - E(t) \le C_2\delta$ for $0 \le t \le t+\delta \le T$. Together with (7.2), that implies that the integral $I(t, w(t), \tilde{q}, \delta)$ is bounded above by $C\delta$ for $0 \le t \le t+\delta \le T$, where $C \equiv C_1 + C_2$. Since the integrand of I is bounded below by $c > 0$ by virtue of Assumption 4.2,

$$c\epsilon(t,\delta) \le I(t, w(t), \tilde{q}, \delta) \le (E(t+\delta) - E(t)) + A(t, t+\delta) \le C\delta \quad \text{for} \quad 0 \le t \le t+\delta \le T. \tag{7.3}$$

so that indeed

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

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 (7.2) 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 (4.6) and (4.7), that implies that a right derivative of w

exists at t with value in (4.8). 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 (4.9) is similar. We consider an interval $[t-\delta, t]$ that is overloaded. Similarly, we have

$$w(t) = w(t-\delta) + \delta - \epsilon(t-\delta, \delta), \tag{7.4}$$

and

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

where

$$J \equiv J(t, w(t), \tilde{q}) \equiv \int_{w(t)}^{w(t)+\epsilon(t-\delta,\delta)} \tilde{q}(t,x)dx, \tag{7.5}$$

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 (7.4) 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 (7.5), that implies that a left derivative of w exists at t with value in (4.9).

The stronger differentiability conclusion depends on the discontinuities of $\tilde{q}(t,x)$. From Proposition 4.2, 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 $w(t+u) \le w(t) + u$ for $0 \le t \le t+u \le T$, 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 (4.8) is continuous in each continuity interval. Hence, existence follows from Peano's theorem; see §2.6 of [Teschl(2000)]. We apply Assumption 4.1 to ensure that $w(0) < \infty$.

(g) (uniqueness of a solution) Under extra regularity conditions, the function Ψ in (4.8) 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 [Teschl(2000)]. 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 (4.8). By (4.5) in Corollary 4.1, $q(t, w(t)-)$ in the denominator or (4.8) takes one of two forms, depending on whether $w(t) \leq t$ or not. Our proof applies to both cases in the same way, so we only consider one case: we suppose that $w(t) \leq t$. Then $q(t, w(t)-) = \lambda((t - w(t))-)\bar{F}(w(t))$. Then ODE (4.8) implies that

$$\begin{aligned} \frac{b(t+,0)}{\bar{F}(w(t))} = \lambda((t-w(t))-)(1-w'(t)) &= \frac{d}{dt}\left(\int_{t_1}^{t-w(t)} \lambda(y)dy\right), \\ \text{so that} \quad \int_{t_1}^{t} \frac{b(y,0)}{\bar{F}(w(y))}dy &= \int_{t_1}^{t-w(t)} \lambda(y)dy, \quad t_1 \leq t \leq t_2. \end{aligned} \tag{7.6}$$

Now suppose there is another function $\tilde{w}$ that also satisfies ODE (4.8) with $\tilde{w}(t_1) = 0$. Then, by the same reasoning, we get

$$\int_{t_1}^{t} \frac{b(y,0)}{\bar{F}(\tilde{w}(y))}dy = \int_{t_1}^{t-\tilde{w}(t)} \lambda(y)dy, \quad t_1 \leq t \leq t_2. \tag{7.7}$$

Equations (7.6) and (7.7) imply that

$$\int_{t_1}^{t} b(y,0)\left(\frac{1}{\bar{F}(w(y))} - \frac{1}{\bar{F}(\tilde{w}(y))}\right)dy = \int_{t-\tilde{w}(t)}^{t-w(t)} \lambda(y)dy, \quad t_1 \leq t \leq t_2. \tag{7.8}$$

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 $\tilde{t} < t \leq \tilde{t}+\delta$. Then we have $1/\bar{F}(w(t)) < 1/\bar{F}(\tilde{w}(t))$ for all $\tilde{t} < t \leq \tilde{t} + \delta$ and $\tilde{t} + \delta - \tilde{w}(\tilde{t} + \delta) < \tilde{t} + \delta - w(\tilde{t} + \delta)$. Therefore, (7.8) 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 (4.8) must be unique. ∎

Proof of Theorem 4.3. To show that the two equations in (4.13) 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.

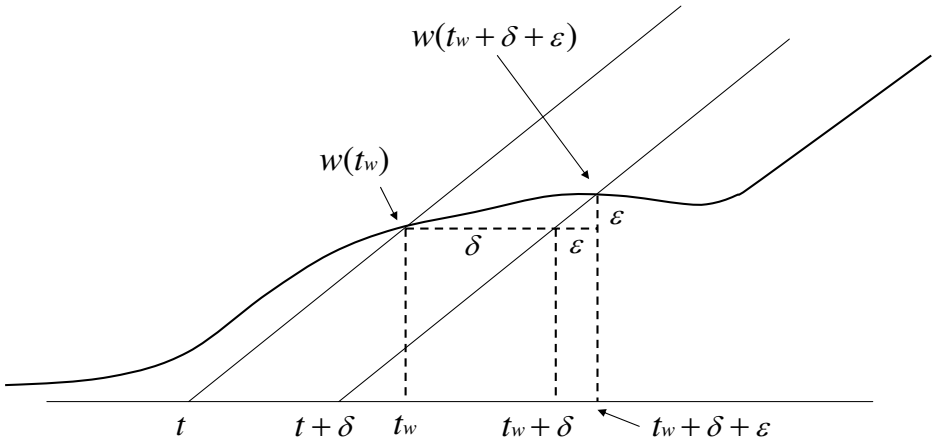

Figure 8: Potential waiting time $v(t)$ is right continuous and has limits from the left.

For a given w, we shall do three things: (i) construct v given the first equation in (4.13), (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 $(t, 0)$: $L(s) = s - t$, $s \geq t$. Let $v(t)$ be the largest t_w such that $L(t_w) = w(t_w)$, as shown in Figure 6. We 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 4.1, 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 4.1, $w'(t) \leq 1$. 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 6. 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 8. Our construction implies that

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

for some $t_w \le \hat{t} \le t_w + \delta + \epsilon$ 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 4.4. For $\delta > 0$, the second equation in (4.13) 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 (4.8), the fourth equality follows from the second equation in (4.13), 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 4.3, 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 4.1 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 (4.8) is discontinuous and the points where Φ is discontinuous. Now we have the relations (for the right derivatives everywhere)

$$v'(t) = \frac{w'(t + v(t))}{1 - w'(t + v(t))} \quad \text{and} \quad w'(t) = \frac{v'(t - w(t))}{v'(t - w(t)) + 1}, \quad t \geq 0, \tag{7.9}$$

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 (4.8) for w. ∎

8 Conclusions.

In §3 and §4 we characterized all the standard performance functions for the $G_t/M/s_t + GI$ fluid model, having time-varying arrival rate and staffing, exponential service and non-exponential abandonment. The relatively simple algorithm primarily requires solving the ODE in Theorem 4.1. The algorithm is summarized in §4.4. We characterized the model, as just reviewed, under the assumption that the staffing function s is feasible, but in Theorem 6.1 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.

The fluid model is intended to serve as an approximation for large-scale many-server queueing systems. We compared the fluid model to simulations of queueing systems in §5. The simulation results show that the fluid approximation 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. Additional simulation results are presented in the Appendix. The approximation tends to be more accurate when the system is either overloaded or underloaded, rather than critically loaded.

There are many directions for future research. It remains to develop results for corresponding networks of fluid models and fluid queues with non-exponential service; we have already made significant progress in both directions, exploiting fixed point equations. It remains to prove supporting

many-server heavy-traffic limits, as briefly discussed in §5, and to establish stochastic refinements. It remains to consider alternative approximations for systems that tend to be nearly critically loaded at all times. It remains to extend the model to include multiple classes of customers and multiple service pools, as in modern customer contact centers. It remains to apply these models to improve the performance of large-scale service systems.

Acknowledgment. This research was supported by NSF grant CMMI 0948190.

References

[Aksin et al.(2007)] Z. Aksin, M. Armony and V. Mehrotra, The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Mgmt.* **16** (2007) 665–688.

[Billingsley(1999)] Billingsley, P.: *Convergence of Probability Measures.*, second ed., Wiley, New York (1999).

[Brown et al.(2005)] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc* **100** (2000) 36–50.

[Eick et al.(1993)] S. G. Eick, W. A. Massey and W. Whitt, The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** (1993), 731–742.

[Garnett et al.(2002)] O. Garnett, A. Mandelbaum and M. I. Reiman. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** (2002), 208–227.

[Green et al.(2007)] L. V. Green, P. J. Kolesar and W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** (2007), 13–39.

[Kang and Ramanan(2008)] W. Kang, W. and K. Ramanan, Fluid limits of many-server queues with reneging. working paper, Carnegie Mellon University, Pittsbirgh, PA, 2008.

[Kaspi and Ramanan(2007)] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. working paper, Carnegie Mellon University, Pittsbirgh, PAA, 2007.

[Mandelbaum et al.(1998)] A. Mandelbaum, W. A. Massey and M. I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems* **30** (1998), 149–201.

[Mandelbaum and Zeltyn(2004)] A. Mandelbaum and S. Zeltyn. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+G$ queue. *OR Spectrum* **26** (2004), 377–411.

[Massey and Whitt(1993)] W. A. Massey and W. Whitt. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems* **13** (1993), 183–250.

[Newell(1982)] G. F. Newell. *Applications of Queueing Theory*, second ed., Chapman and Hall, London, 1982.

[Pang and Whitt(2010)] G. Pang and W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65** (2010) 235–275.

[Pang et al.(2007)] G. Pang, R. Talreja and W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** (2007), 193–267.

[Puhalskii(2008)] A. A. Puhalskii. The $M_t/M_t/k_t + M_t$ queue in heavy traffic. working paper, Mathematics Department, University of Colorado at Denver, 2008.

[Reed and Talreja(2009)] J. Reed and R., Talreja. Distribution-valued heavy-traffic limits for the G/GI/infty queue. working paper, New York University, New York, NY, 2009.

[Teschl(2000)] G. Teschl, *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/

[Whitt(2006)] W. Whitt. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54** (2006), 37–54.

[Yom-Tov and Mandelbaum(2010)] Yom-Tov, G. and A. Mandelbaum. The Erlang-R queue: time-varying QED queues with re-enttant customers in support of healthcare staffing. working paper, the Technion, Israel, 2010.

[Zeltyn and Mandelbaum(2005)] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems* **51** (2005), 361–402.

APPENDIX

to

A Fluid Model for a Large-Scale Service System Experiencing Periods of Overloading

A Overview.

This appendix contains material supplementing the main paper, mostly presented in the order of the related material in the main paper. First, §B explains why the service content density b satisfies the transport PDE in an underloaded interval, as noted in Remark 3.2. Two missing proofs for §4 on the $G_t/M/s_t + GI$ model in an overloaded interval and one proof in §6 are provided in §C.

In §D we supplement §4.4, which summarizes the algorithm, by providing more discussion of the algorithm. In particular, we provide a formal statement of the algorithm. We also specify the algorithm to adjust for an initially infeasible staffing function s and illustrate its performance. In §E we discuss the structure of the BWT function w. Theorem 4.1 requires the positivity $\lambda_{inf} > 0$ in Assumption 4.2. We now consider cases in which $\lambda(t) = 0$ for some $t \geq 0$. We show what can happen when the the zero set has zero Lebesque measure or positive Lebesque measure.

Finally, in §F we supplement §5 in the main paper by presenting additional comparisons of the fluid model to simulations of large-scale queueing systems.

B The Transport PDE in §3

In Remark 3.2 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. As in §3 we consider general GI service.

Proposition B.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:*

$$b_t(t,x) + b_x(t,x) \equiv \frac{\partial b}{\partial t}(t,x) + \frac{\partial b}{\partial x}(t,x) = -h_G(x)b(t,x). \tag{B.1}$$

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 (B.1) with initial conditions $\lambda(t)$ and $b(0,x)$, yields Proposition 3.1. To verify that, recall that the general solution to pde (B.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.

C Proofs in §4 and §6

Proof of Corollary 4.2. Since the proofs to (a) and (b) are similar, we will only prove (b). ODE (4.8) implies that

$$b(t,0)e^{\theta t} = \lambda(t-w(t))e^{\theta(t-w(t))}(1-w'(t)) \quad = \quad \frac{d}{dt}\left(\int_0^{t-w(t)} \lambda(y)e^{\theta y}dy\right),$$

which implies

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

and inverting function $\tilde{\Lambda}(\cdot)$ yields (4.10). Moreover,

$$\tilde{t} \equiv \inf\{t>0 : w(0)=0\} = \inf\{t>0 : \tilde{\Lambda}(t) = \int_{t_1}^t b(y,0)e^{\theta y}dy\}. \quad \blacksquare$$

Proof of Theorem 4.2. Recalling the definition of $\sigma(t)$ in (2.10), we obtain

$$\sigma(t) = \int_0^\infty b(t,x)h_G(x)\,dx = B(t).$$

However, in the overloaded interval, $B(t) = s(t)$ and $s(t) \geq s_{lbd}$ by Assumption 4.3. 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

$$E(t+u) - E(t) = \int_t^{t+u} b(v,0)\,dv = \int_t^{t+u} (s'(v) + \sigma(v))\,dv \geq s(t+u) - s(t) + s_L u.$$

By Assumption 4.3, $s(t+u) \geq s_L$. Starting from the definition (4.12), we apply the inequalities above to obtain

$$\begin{aligned} v(t) &\equiv \inf\{u \geq 0 : E(t+u) - E(t) + A_t(u) \geq Q(t)\} \\ &\leq \inf\{u \geq 0 : E(t+u) - E(t) \geq Q(t)\} \\ &\leq \inf\{u \geq 0 : (s_L u - s(t) + s_L)^+ \geq Q(t)\} \leq \frac{Q(t) + s(t) - s_L}{s_L} < \infty, \end{aligned}$$

where $Q(t) \leq Q(0) + \Lambda(t) < \infty$ for all t. $\blacksquare$

Proof of Theorem 6.1. First, since $b_s(\cdot,0)$ is continuous for our original s, the violation in (6.2) 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 (6.5) 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

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

where 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 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. ∎

D More on the Algorithm

D.1 The Algorithm for the $G_t/M/s_t + GI$ Model

We formally state the algorithm for the $G_t/M/s_t+GI$ fluid model. We assume the system does not alternate between overloaded and underloaded for infinitely many times in a finite time horizon. It might be possible to construct functions $\lambda(\cdot)$ and $s(\cdot)$ such that the lengths of intervals in which the system is overloaded and underloaded converge to 0, but we do not consider those cases.

For given model parameters F, G, $\lambda(t)$, $s(t)$ for $0 \leq t \leq T$, and initial condition $q(0, x) = r(x)$ for $x \geq 0$, $Q(0) = \int_0^\infty r(x)dx$, $w(0) = w_0 \geq 0$ and $B(0) \leq s(0)$ such that $Q(0)(s(0) - B(0)) = 0$, the algorithm is as follows:

- $n \leftarrow 1$, $t_0^u \leftarrow 0$, $t_0^d \leftarrow 0$, $F \leftarrow 1$
- IF $B(0) < s(0)$
  - $B_0 \leftarrow B(0)$
  - $t_n^u \leftarrow \inf\{t > t_{n-1}^d : B_0 e^{\mu t_{n-1}^d} + \int_{t_{n-1}}^t \lambda(y)e^{\mu y}dy \geq s(t)e^{\mu t}\}$
  - IF $t_n^u \geq T$, THEN $N \leftarrow n - 1$, $U \leftarrow 1$. END.
  - ELSE
    - $t^* \leftarrow t_n^u$, $w_0 \leftarrow 0$
    - $w(t)$ solves ODE (4.8) for $t \geq t^*$ with $w(t^*) = w_0$, $t_n^d \leftarrow \inf\{t > t_n^u : w(t) = 0\}$
    - IF $s'(t) + s(t)\mu < 0$ for some $t^* \leq t \leq t_n^d$, THEN $F \leftarrow 0$, END.
    - IF $t_n^d \geq T$, THEN $N \leftarrow n$, $U \leftarrow 0$. END.
    - ELSE
      - $B_0 \leftarrow s(t_n^d)$, $n \leftarrow n + 1$, go to line 4.
- ELSEIF $B(0) = s(0)$ and $Q(0) > 0$
  - $w(t)$ solves ODE (4.8) with $w(0) = w_0$, $\hat{t} \leftarrow \inf\{t > 0 : w(t) = t\}$

- IF $s'(t) + s(t)\mu < 0$ for some $0 \le t \le \hat{t}$, THEN $F \leftarrow 0$, END.
- $t^* \leftarrow \hat{t}$, $w_0 \leftarrow \hat{t}$, go to line 8.

- ELSE ($B(0) = s(0)$ and $Q(0) = 0$)
  - IF $\lambda(0) > \sigma(0)$, THEN $t^* \leftarrow 0$, $w_0 = 0$, go to line 8.
  - ELSE ($\lambda(0) > \sigma(0)$), $B_0 = s(0)$, go to line 4.

The output of this algorithm are integer variable N, indicator variables U and F, and a sequence of time points $\{t_n^u, t_n^d, n = 1, 2, \ldots, N\}$. $F = 0$ implies that the given service-capacity function $s(\cdot)$ is infeasible and the algorithm will end as soon as it detects this infeasibility. The service-capacity function $s(t)$ is feasible if the algorithm ends with $F = 1$. The value of indicator variable U indicates whether the system is underloaded ($U = 1$) or overloaded ($U = 0$) at time T. For instance, if the system is underloaded at time 0 and $U = 1$, then we know that the system is underloaded for $t \in [0, t_1^u] \cup [t_1^d, t_2^u] \cup \cdots \cup [t_{N-1}^d, t_N^u] \cup [t_N^d, T] \equiv T^U$, and the system is overload for $t \in [t_1^u, t_1^d] \cup [t_2^u, t_2^d] \cup \cdots \cup [t_N^u, t_N^d]$. Moreover, we have that $w(t)$ solves ODE (4.8) for $t \in [t_n^u, t_n^d]$ with $w(t_n^u) = 0$ for $1 \le n \le N$, and $w(t) = 0$ for all $t \in T^U$. Other cases are similar.

D.2 A Fluid Algorithm with Infeasible s.

Our main algorithm for the $G_t/M/s_t + GI$ 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 §6 to find the minimum feasible staffing function.

In the context of the $G_t/M/s_t + GI$ model, a sufficient condition for feasibility over $[0, T]$ is

$$s(t) + s'(t) \ge 0, \quad 0 \le t \le T. \tag{D.1}$$

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 $t_2 \equiv \inf\{t_1 < t \le T : s(t) = B(t)\} \wedge T = \inf\{t_1 < t \le T : s(t) = s(t_1)e^{-(t-t_1)}\} \wedge T$. Note that $b(t, 0) = 0$ for $t_1 \le t \le t_2$ 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 (4.8) 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 the main paper; 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 $s(t)$.

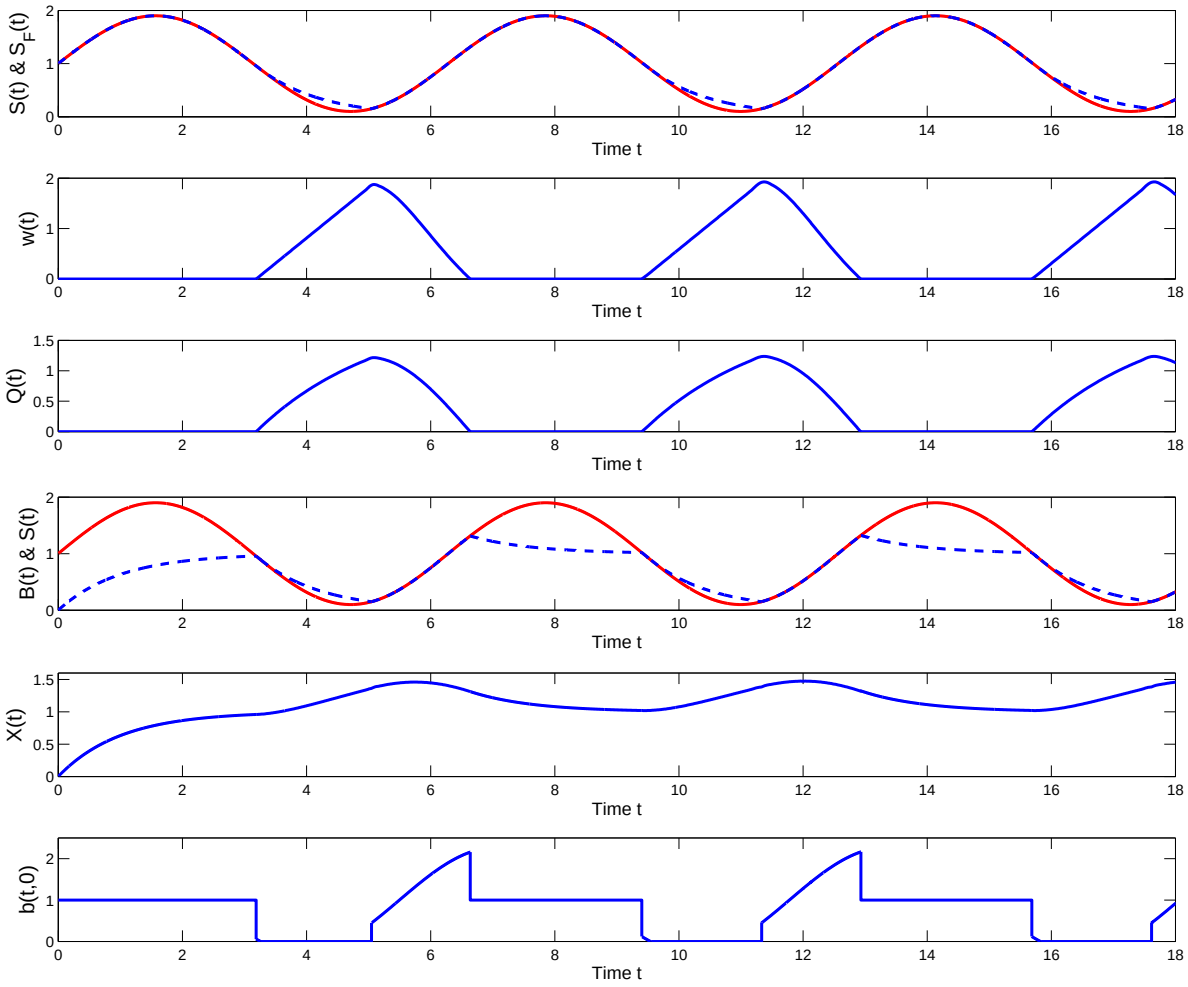

Figure 9: The $M/M/s_t + M$ fluid model with infeasible s.

To evaluate the performance of the modified algorithm, we use the example in §F.3, 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

$$s(t) \equiv \lambda + \bar{\lambda} \cdot \sin(c \cdot t). \tag{D.2}$$

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 §F.3. Now s has greater fluctuation and it is easy to see that condition (D.1) is no longer satisfied.

We plot the performance measures of the fluid model in Figure 9. Compared with Figure 19, 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 (above) s (the red solid curve), and $B(t)$

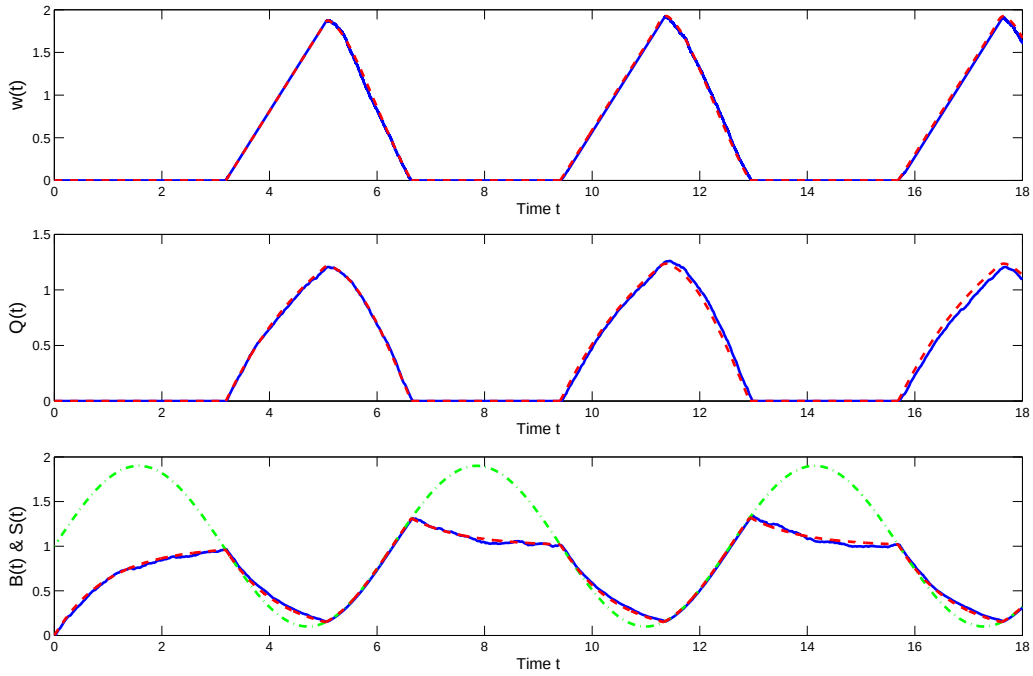

Figure 10: The $M/M/s_t + M$ fluid model with infeasible s compared with simulation.

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 10 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 $B(t)$ and $\hat{B}_n(t)$ are distinct from the given service-capacity function s (the dashed green curve) in I_{inf}.

E Structure of the Boundary Waiting Time $w(t)$.

Theorem 4.1 requires the positivity $\lambda_{inf} > 0$ in Assumption 4.2. We now consider cases in which $\lambda(t) = 0$ for some $t \geq 0$.

E.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 11(a). Again we assume that both $b(t,0)$ and $\lambda(t)$ are continuous for $0 \leq t \leq T$.

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 $x \geq 0$.

We can see from Figure 11(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 11(b) the boundary waiting time $w(t), 0 \leq t \leq T$ 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.

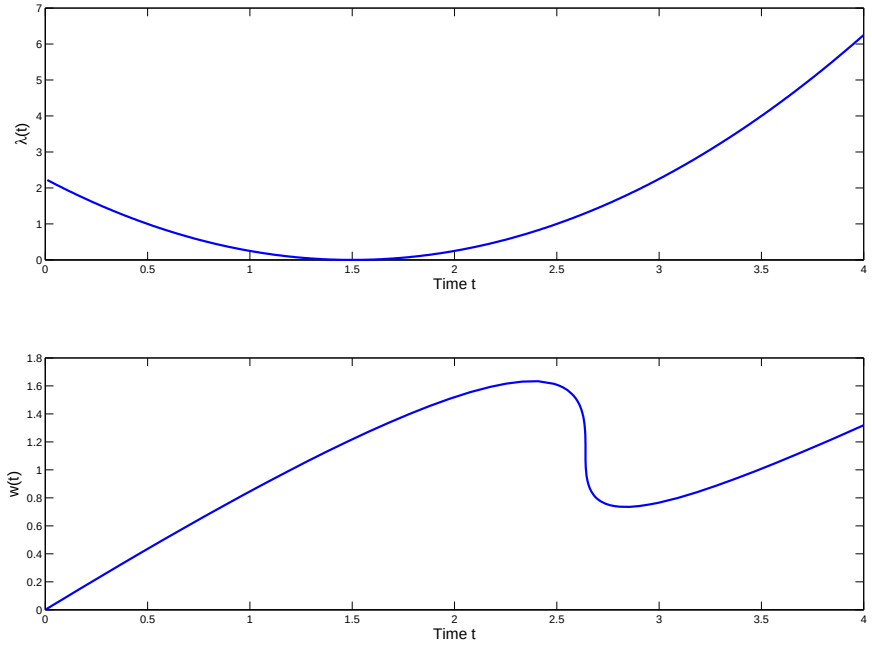

Figure 11: An example of boundary waiting time $w(t)$ with $\lambda(t) = 0$ once.

E.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 12.

Intuition tells us in this case $w(t)$ cannot be continuous for all $t \geq 0$, it will jump at some times. But when will $w(t)$ jump? What will be the heights of the jumps? To answer 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 12(a). Let the abandon-time distribution be exponential with rate 0.5.

In Figure 12(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

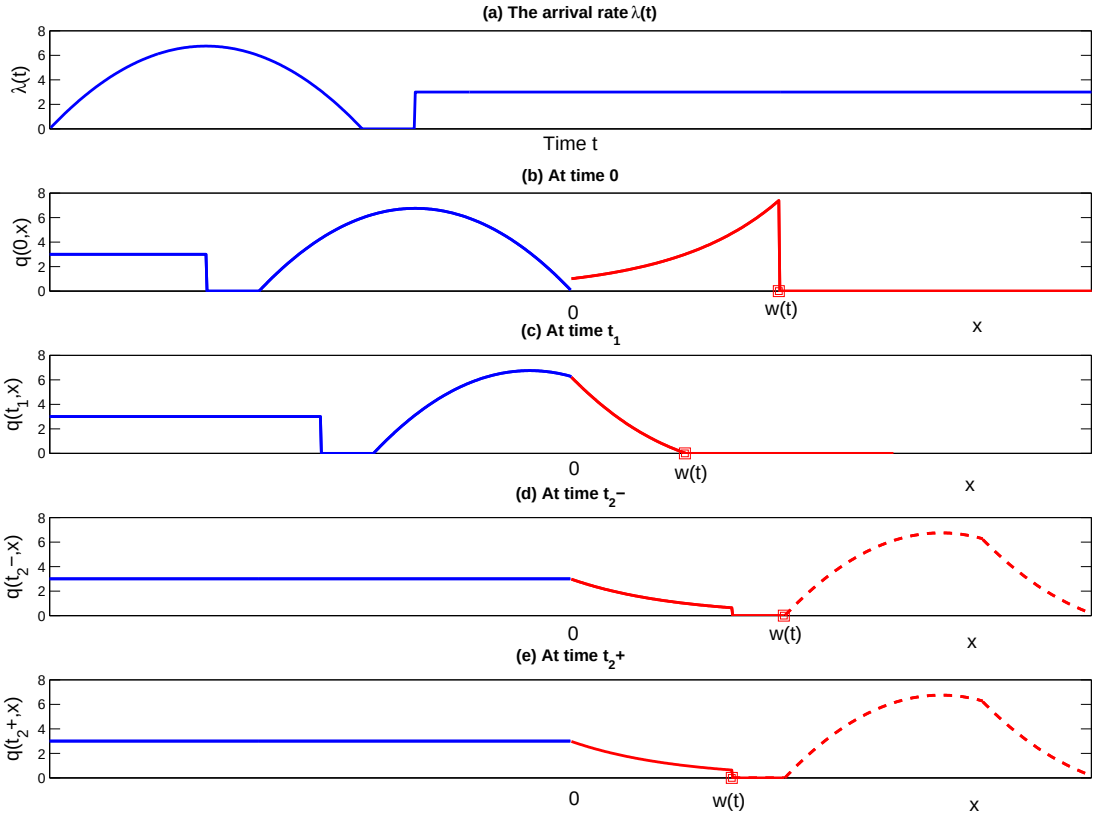

Figure 12: The dynamics of $q(t,x)$ of an example with $\lambda(t) = 0$ for $0 < t_1 \le t < t_2 < \infty$.

with the origin fixed, the blue line moves to the right at rate 1, because 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 12(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, $q(t_1,x) = \lambda(t_1 - x) \cdot e^{-0.5} \cdot 1_{\{0 \le x \le t_1\}}$ 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 13. 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 13. 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.

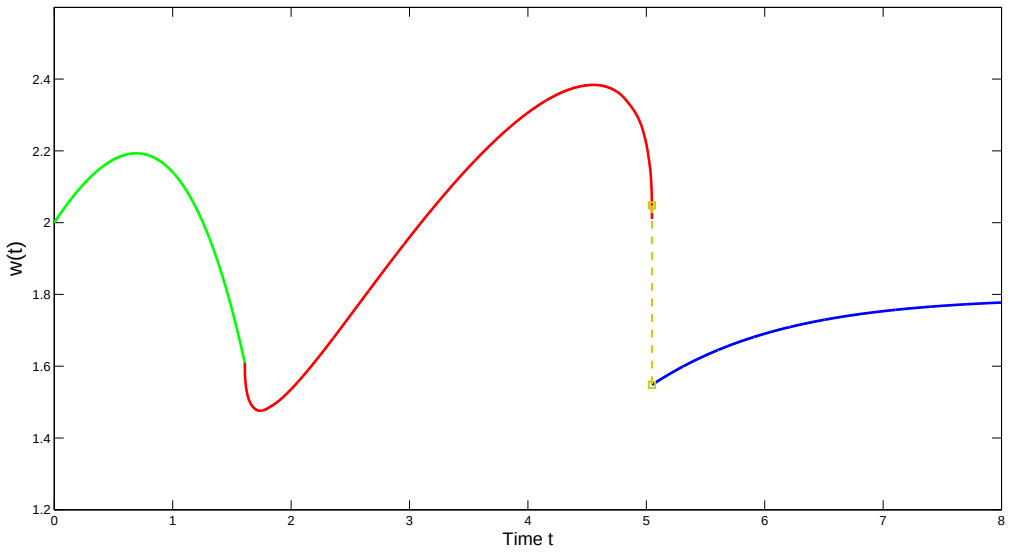

Figure 13: An example of the boundary waiting time $w(t)$ with $\lambda(t) = 0$ for $0 < t_1 \leq t < t_2 < \infty$.

F More Comparisons with Simulation.

In this section we present additional results comparing simulation results for large-scale queueing models to numerically calculated values of the corresponding approximating fluid model. These results complement those in §5.

F.1 Smaller Sample Sizes for the Example in Section 5.

The queueing model for the example in §5 was chosen with the very large scaling $n = 1000$ for two reasons: first, to demonstrate that many-server heavy-traffic limits should hold in this context and, second, to provide a good test of the numerical algorithm for the fluid model. In order to be useful as approximations for realistic large-scale queueing systems, the approximation should be reasonable for smaller scaling factors. We demonstrate that now.

In Figure 14 below we plot the analog of Figure 3 for the case of one sample path of the simulation with $n = 100$, for the same fluid model. In Figure 15 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 16 and the average over 200 sample paths in Figure 4. (The latter appears in the main paper.) 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 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).

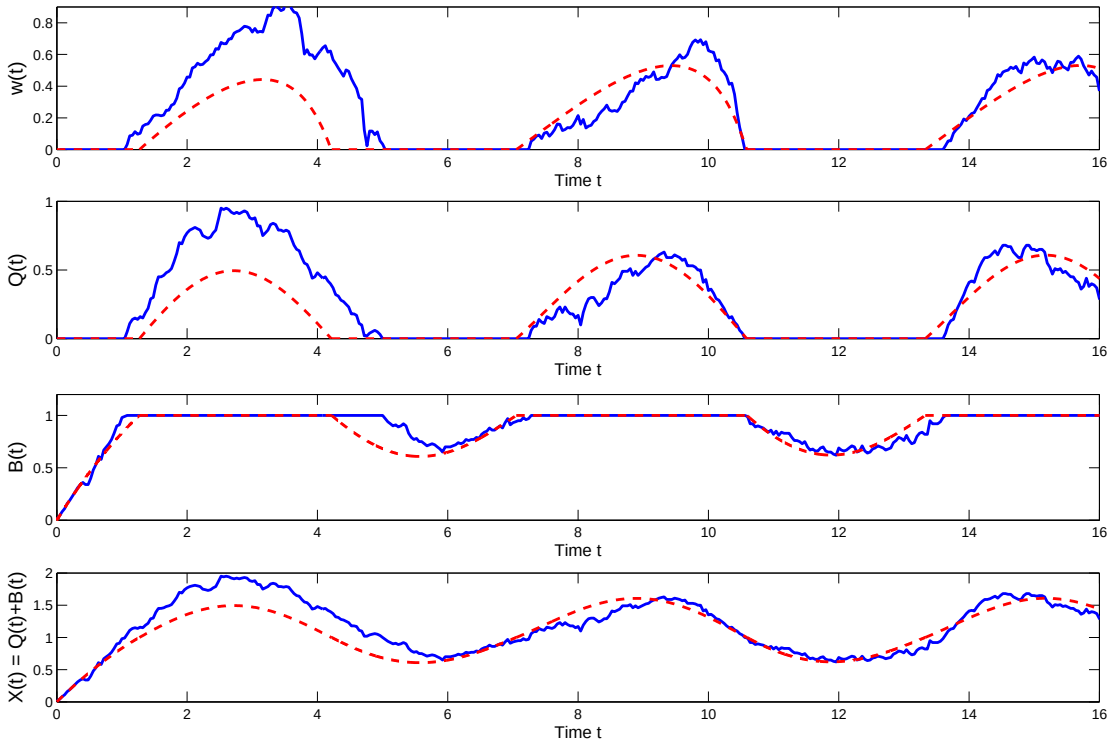

Figure 14: Performance of the $M_t/M/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.

F.2 Additional Simulation Results for the $M_t/M/s_t + M$ Fluid Model.

We now continue to discuss the simulation example in §5. In Figure 17 and 18, we plot the two-parameter functions $q(t, x)$ and $b(t, x)$ for $0 \le t \le 4$. The decays on the x dimension in Figure 17 and 18 are due to customer abandonment and service completion, the shape changes on the time dimension are according to the sinusoidal arrival rate. The discontinuity at $t_1 = 1.66$ of $b(t, x)$ is consistent with the discontinuity of $b(t, 0)$ in Figure 2, $q(t, x) = 0$ for $0 \le t \le t_1$ since the system is underloaded.

F.3 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 respectively, and a sinusoidal capacity function

$$s(t) \equiv \lambda + \bar{\lambda} \cdot \sin(c \cdot t). \tag{F.1}$$

In the previous example in §5, 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 feasibility of s is

$$s'(t) \ge -\mu s(t), \quad t \ge 0. \tag{F.2}$$

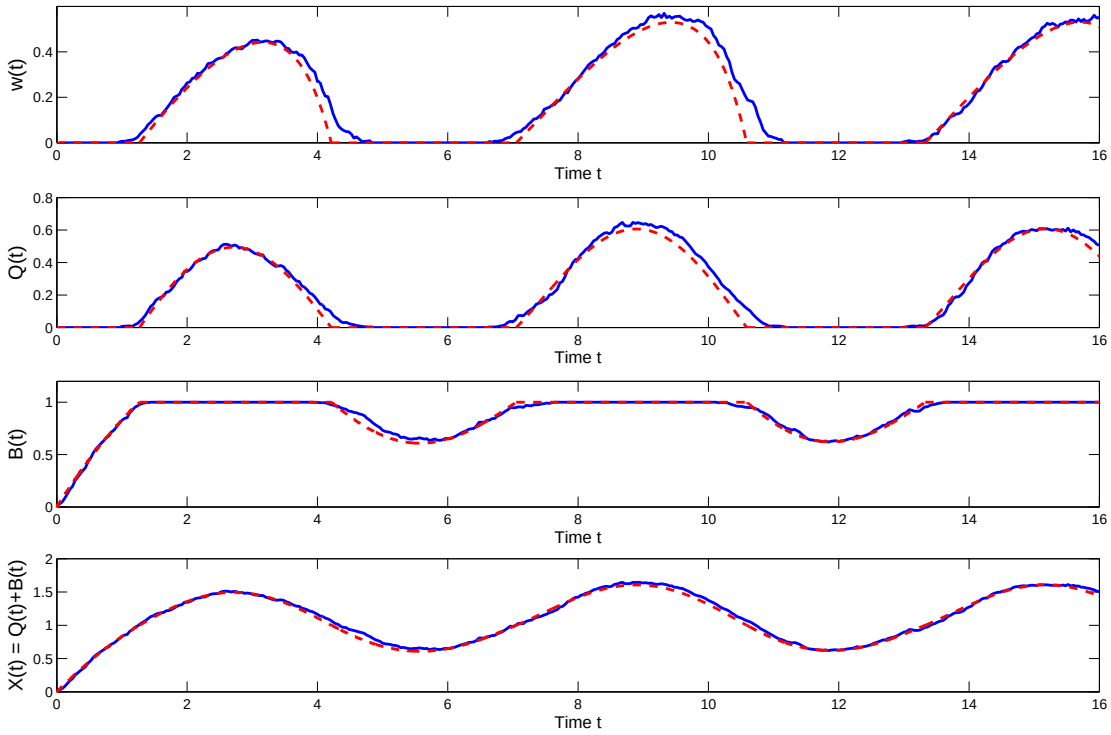

Figure 15: Performance of the $M_t/M/s+M$ fluid model compared with simulation results: an average of 10 sample paths of the scaled queueing model based on $n = 100$.

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 19 and compare them with simulation estimates in 20, analogs to Figure 2 and 3. In Figure 20, 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 not remove servers until they complete the service in progress. As in §5, we let $n = 1000$ for the sequence of scaled queueing models in §**??**. Thus we have $\lambda_n = a_n = 1000$, $b_n = 600$, $c_n = 1$.

F.4 Simulation Verifications 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., $f(x) = p \cdot \theta_1 e^{-\theta_1 x} + (1-p) \cdot \theta_2 e^{-\theta_2 x}$, 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.

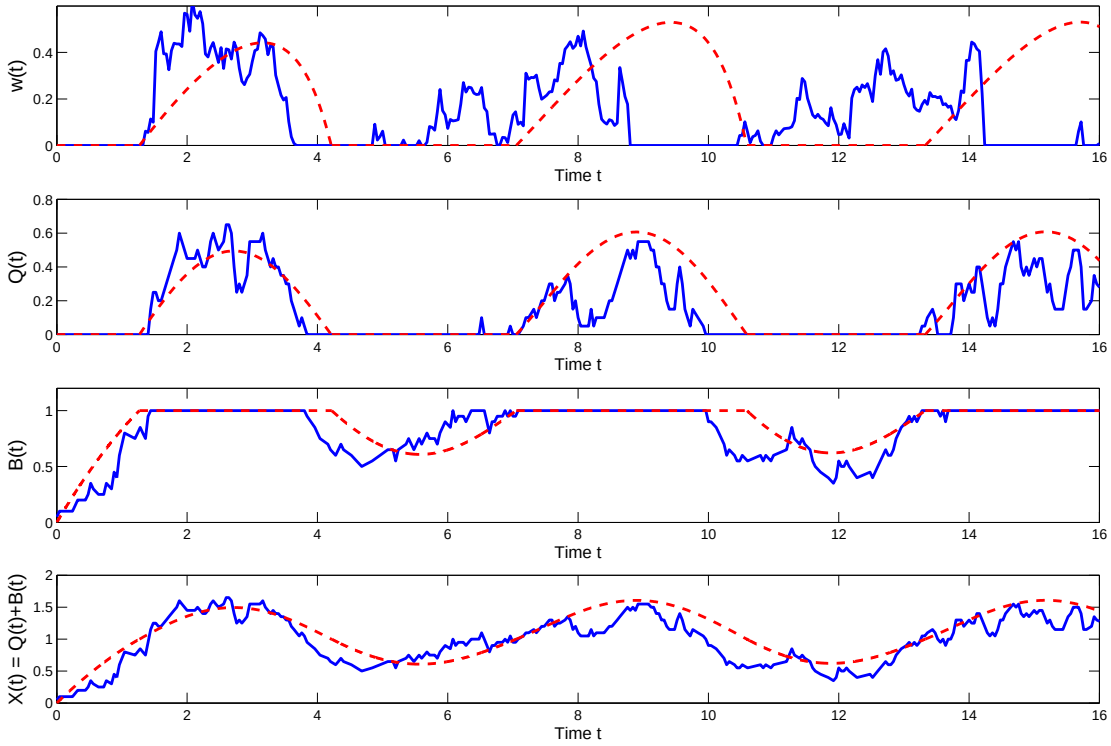

Figure 16: Performance of the $M_t/M/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 20$.

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 $C_{SCV} = 4$.

We still let the arrival-rate function λ be sinusoidal, as in (5.1). 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 21 and 23 respectively for $t \in [0, T]$ with $T = 16$. The performance measures shown in Figure 21 and 23 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 22 and 24 we compare the fluid approximations with simulation experiments. The queueing model has a nonhomogeneous Poisson arrival process with sinusoidal rate function as in (5.1), with $a = s = 2000$, $b = 0.6a = 1200$. In Figure 22 and 24, the blue solid lines of the simulation estimations of single sample paths applied with fluid scaling, and the red dashed lines are the fluid approximations. We conclude that the fluid approximation is remarkably accurate.

F.5 Simulation Verifications 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 $X_1, X_2, X_3, \ldots$ 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

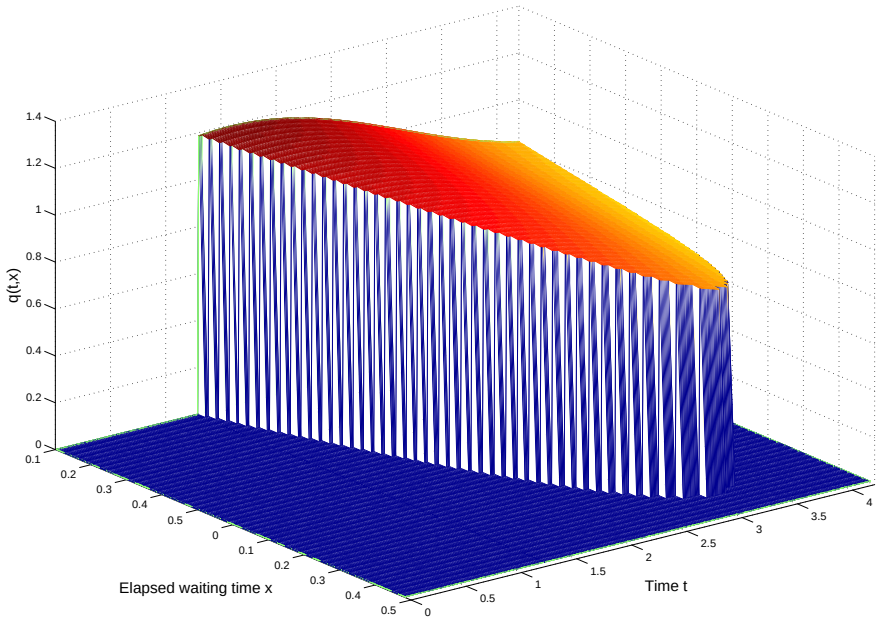

Figure 17: The two-parameter density function q for the $G_t/M/s+M$ fluid model with sinusoidal arrival-rate function.

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 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 25, we compare the fluid approximation with simulation experiments of the $G_t/M/s+M$ model. Here the only difference from Figure 3 is that the arrival process (G_t) is not Poisson but has the same sinusoidal rate function as (5.1).

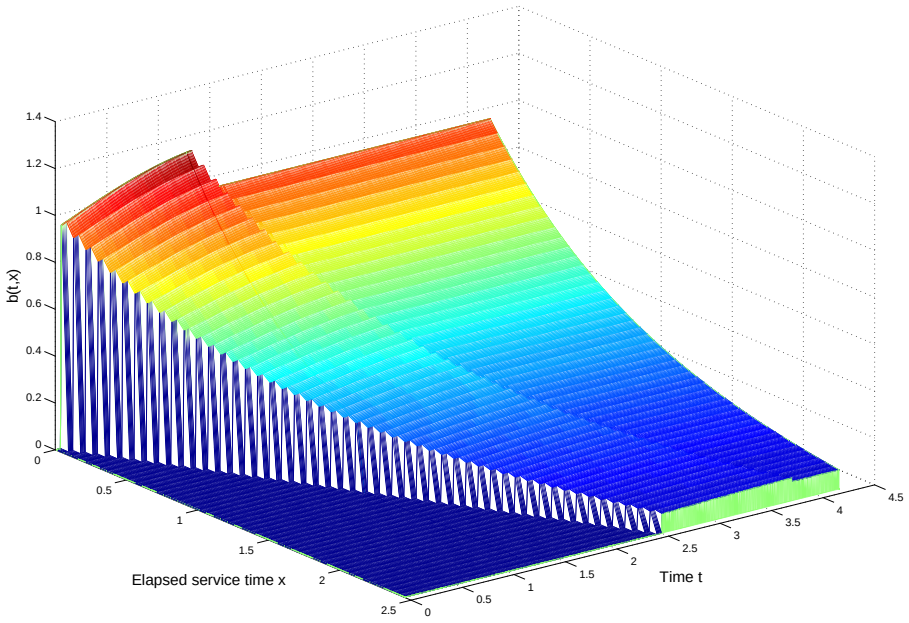

Figure 18: The two-parameter density function b for the $G_t/M/s + M$ fluid model with sinusoidal arrival-rate function.

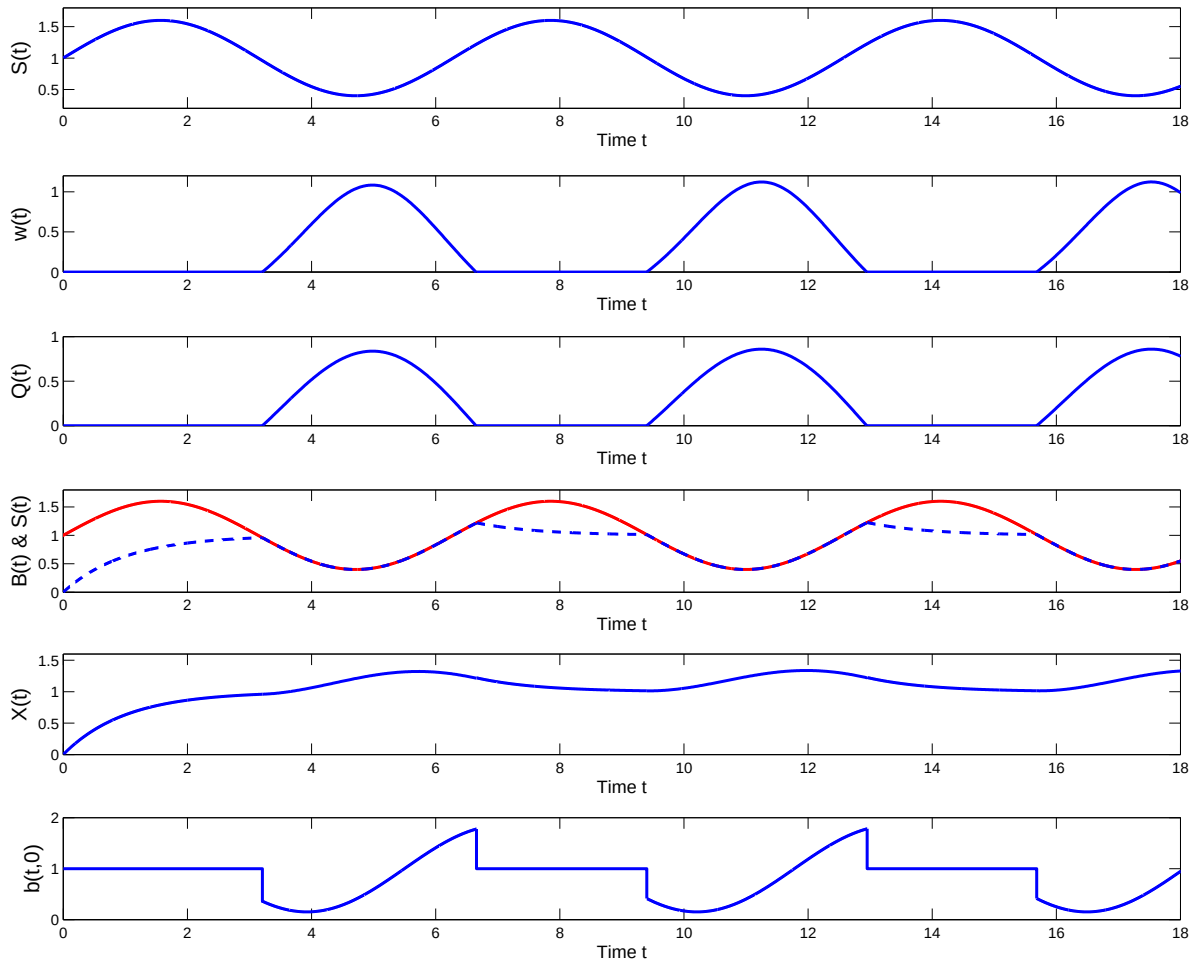

Figure 19: The $M/M/s_t + M$ fluid model with sinusoidal service-capacity function.

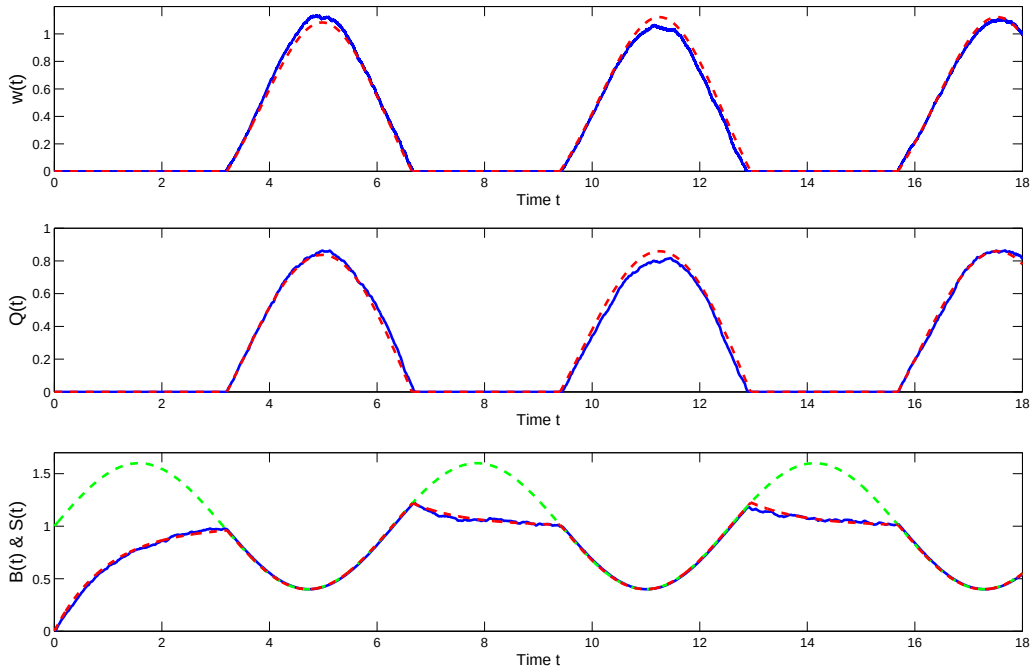

Figure 20: The $M/M/s_t + M$ fluid model compared with simulations of the queueing system.

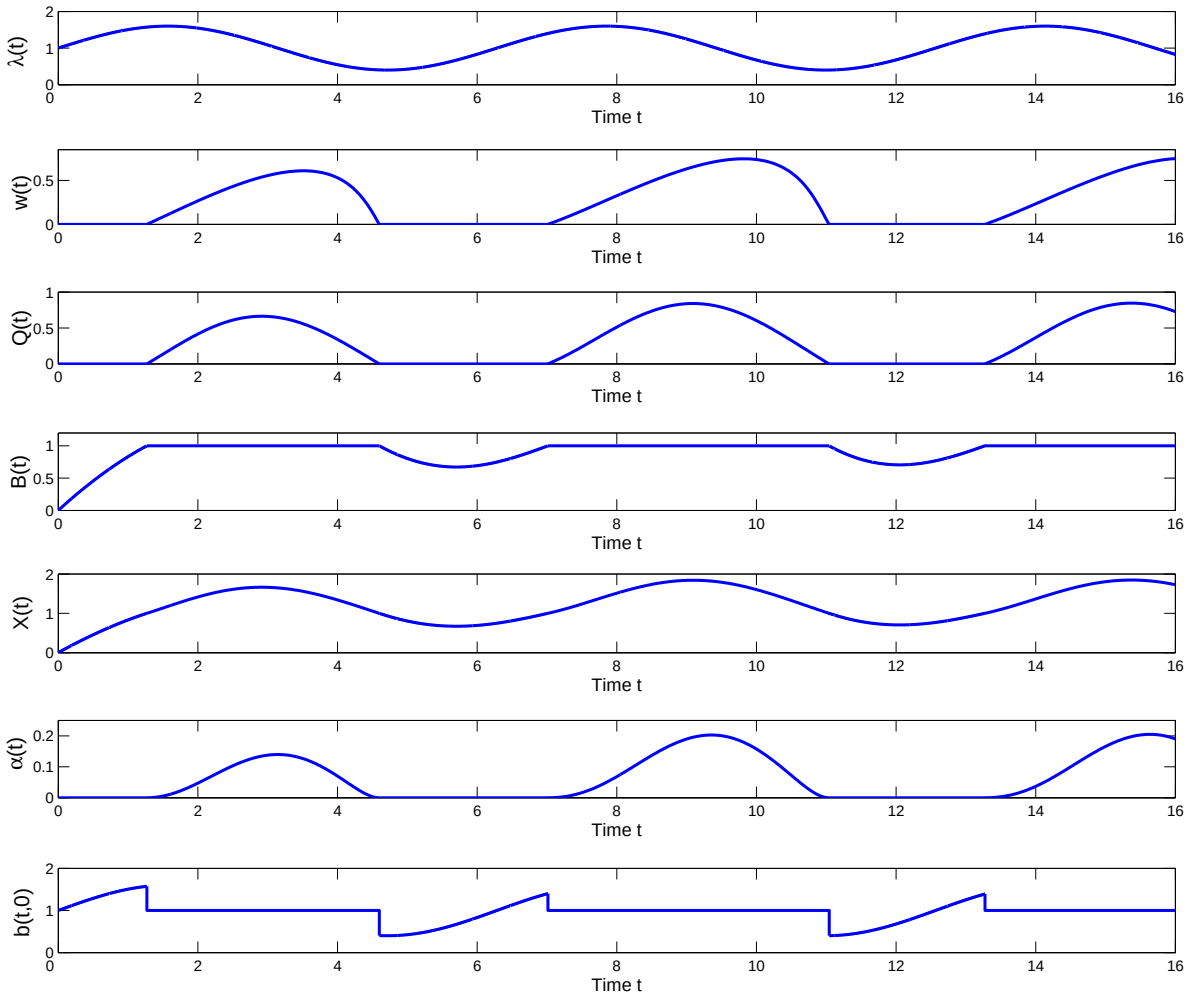

Figure 21: The $M_t/M/s + E2$ fluid model with sinusoidal arrival-rate function.

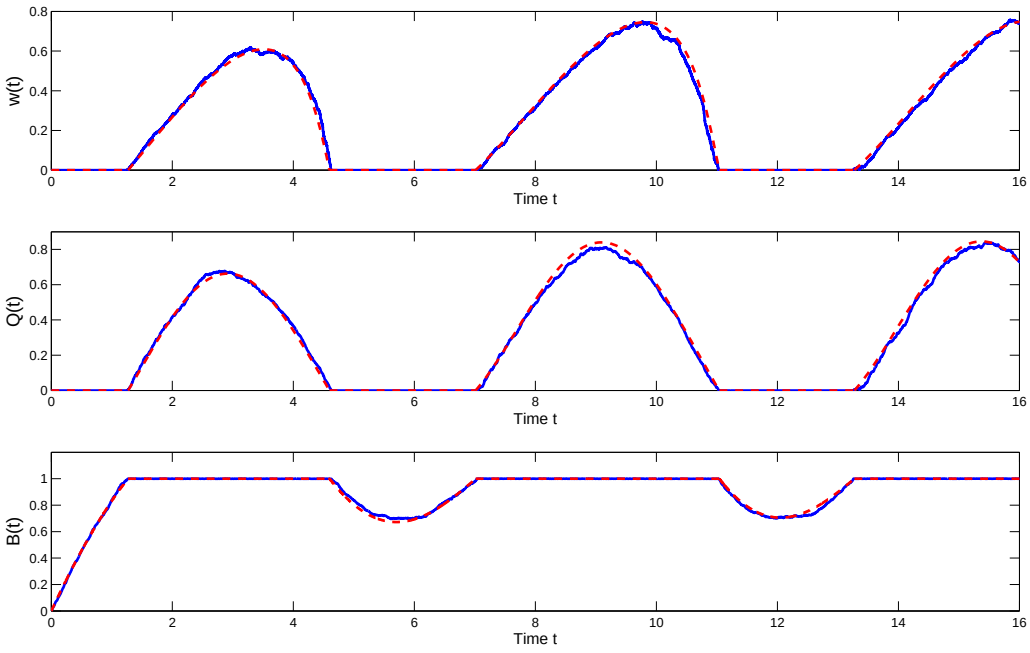

Figure 22: The $M_t/M/s + E2$ fluid model compared with simulations of the queueing system.

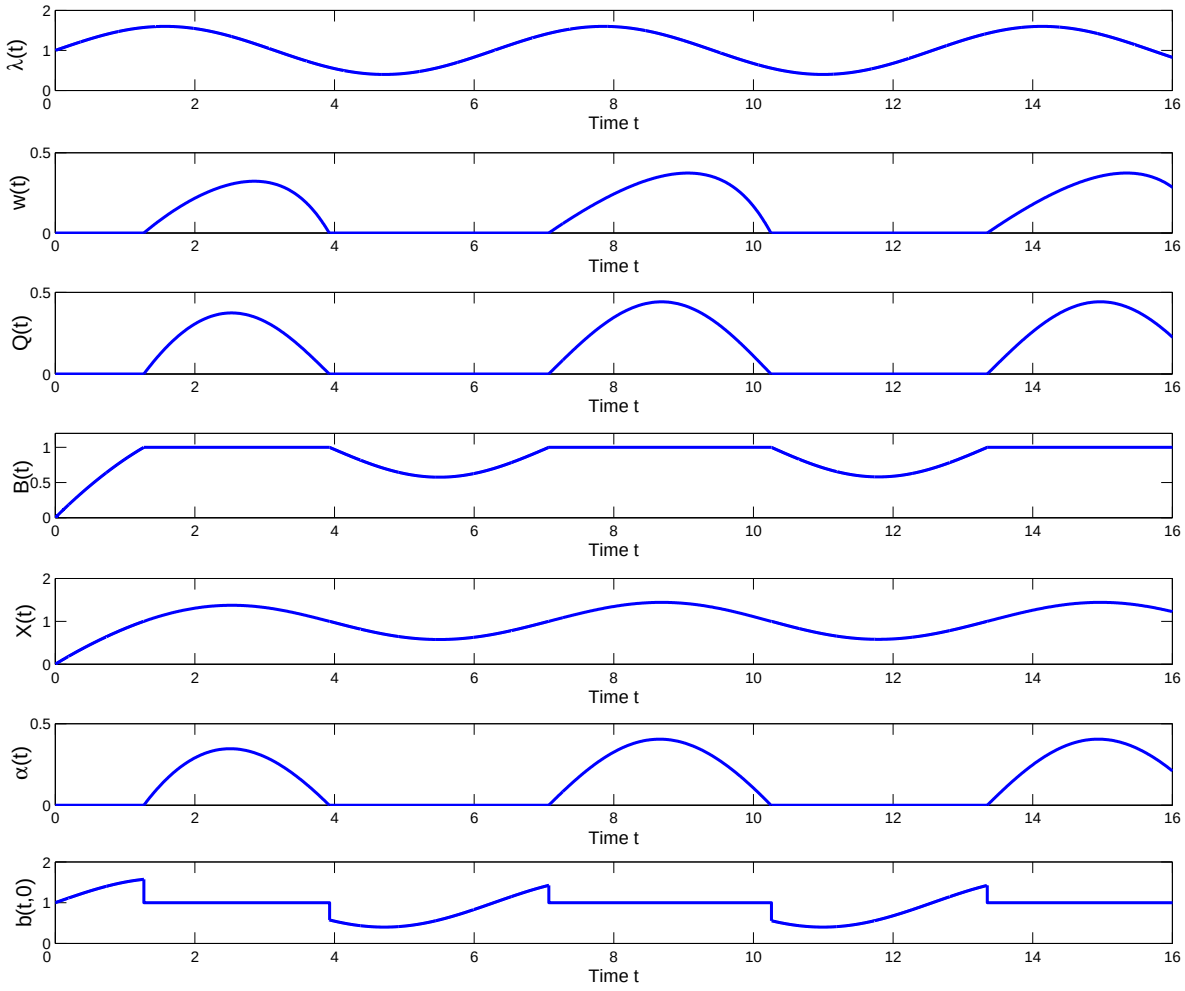

Figure 23: The $M_t/M/s + H2$ fluid model with sinusoidal arrival-rate function.

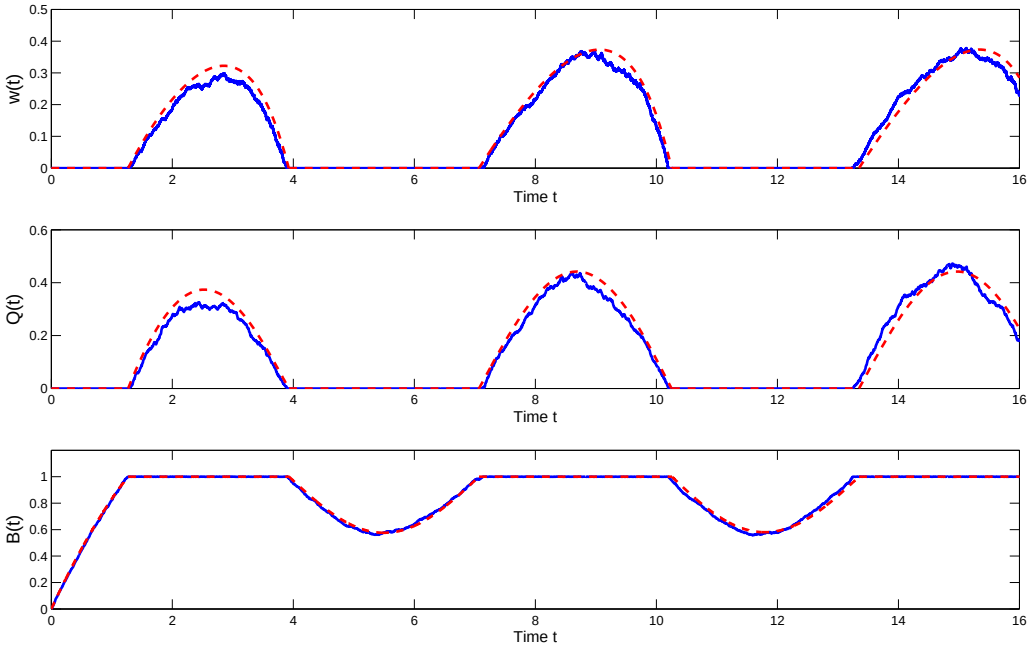

Figure 24: The $M_t/M/s + H2$ fluid model compared with simulations of the queueing system.

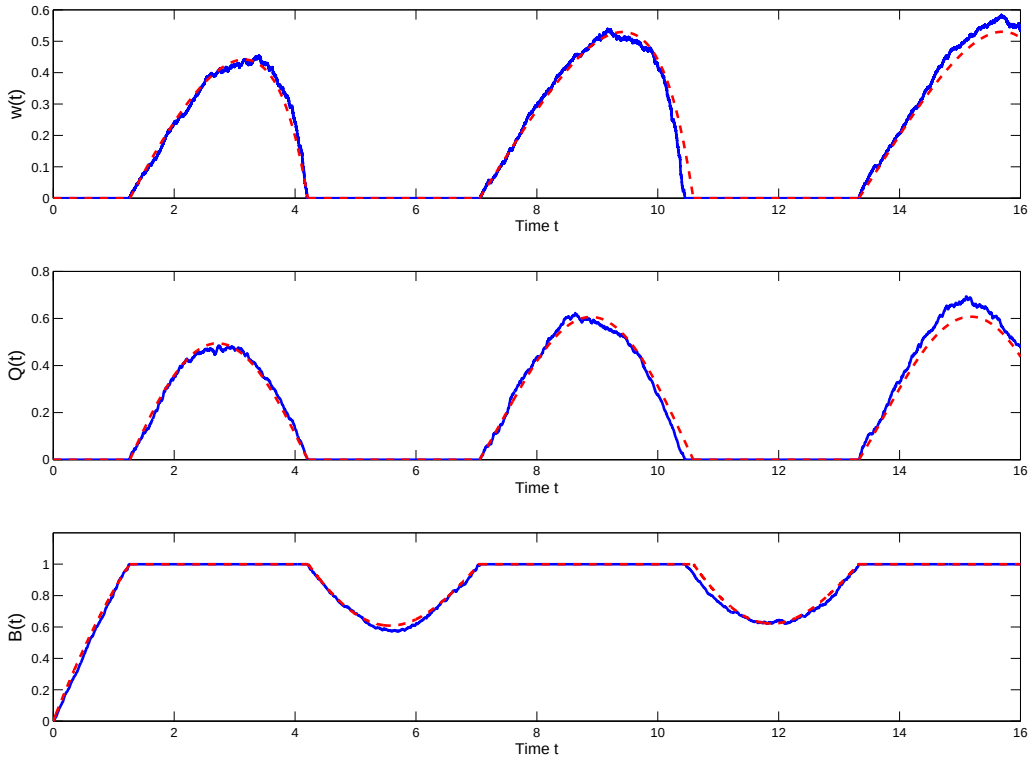

Figure 25: The $G_t/M/s+M$ fluid model compared with simulations of the queueing system.