A Fluid Model for the Many-Server $G_t/GI/s_t + GI$ Queue

Yunan Liu and Ward Whitt

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

August 13, 2010

Abstract

In a recent paper we introduced and analyzed a deterministic $GI_t/M/s_t + GI$ fluid model that can be used to show how queue lengths and waiting times depend on model parameters in a large-scale queueing 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 a non-exponential patience distribution. Our key assumptions were (i) that the scale is large (there are many servers), (ii) that the system alternates between overloaded intervals and underloaded intervals, and (iii) that the service-time distribution is exponential. Here we extend the analysis to a large class of non-exponential service distributions. To do so, we express the service content density in an overloaded interval as the solution of a fixed point equation. We apply the Banach contraction fixed point theorem to show that the equation has a unique solution and to develop an efficient algorithm. Given the service content density, all other performance measures can be calculated by previous methods. We also show how the staffing can be chosen to stabilize delays at any target value.

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; stabilizing delays.

1 Introduction

This paper is a sequel to [6], which developed and analyzed a deterministic fluid model approximating the $G_t/M/s_t + GI$ queueing model, having a general arrival process with time-varying

arrival rate (the initial G_t), independent and identically distributed (i.i.d.) service times with an exponential distribution (the M), a time-varying large number of servers (the s_t) and customer abandonment from queue with i.i.d. patience times with a general cdf F (the final $+GI$). The purpose of the present paper is to extend the model and the performance description to a large class of non-exponential service-time distributions.

From our earlier work in [11, 12], we know that the steady-state performance of large-scale stationary $G/GI/s + GI$ queueing models tends to be relatively insensitive to the service-time distribution beyond its mean. That is consistent with the well known (exact) insensitivity property of the $M/GI/s/0$ and $M/GI/\infty$ models. Thus one might think this extension is of little practical value. However, as we showed in [2] for many-server loss models, in §5 of [7] we show that the performance of the time-varying $G_t/GI/s_t + GI$ fluid model depends strongly on the service-time distribution beyond its mean.

Both this paper and [6] are extensions of [12], which developed a deterministic fluid model to approximate the steady-state performance of a stationary $G/GI/s + GI$ queueing model. In doing so, we provide for the first time a full description of the transient behavior, even for the stationary $G/GI/s + GI$ fluid model. The fundamental evolution equations, here in (2.3), are the same as in [12], but the performance depends on a boundary waiting time (BWT), which is characterized in [6] as the solution of an ordinary differential equation (ODE). We also go beyond [12] to determine the potential waiting time, i.e., the virtual waiting time of an arrival if that arrival would elect never to abandon.

For background on previous work on many-server queues with time-varying arrival rates, see [5]. There has been considerable work on large-scale Markovian models with time-varying arrivals; e.g., [8, 9]. The importance of customer abandonment and ways to treat it are discussed in [1, 3, 4, 13].

Most queueing models are stochastic, because a primary cause of congestion is random fluctuation 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 innovative aspect of our approach is to consider systems that alternate between overloaded intervals and underloaded intervals. With time-varying arrival rates, such alternating behavior commonly occurs when it is difficult for system managers 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. That leads to the alternating overloaded and underloaded intervals that we consider.

On the other hand, if staffing can be adjusted dynamically, then it may be possible to stabilize performance. For example, [3] developed a simulation-based algorithm to identify a staffing function that can stabilize delays in a large-scale service system with time-varying arrival rate. Even for that problem, we contribute by showing how the fluid model can also be used to stabilize delays; see §6. The fluid model is revealing because it also provides formulas for other performance functions with that staffing. These other performance measures are stabilized to some extent, but they are not stabilized completely. Hence, there necessarily are tradeoffs between different performance measures when we want to stabilize performance.

Here is how this paper is organized. We start in §2 by defining the $G_t/GI/s_t + GI$ fluid model and specifying key regularity conditions. In §3 we show how to extend the analysis in an overloaded interval to non-exponential service, exploiting the Banach contraction fixed point theorem. In §4 we discuss the minor adjustments needed to treat the other performance functions in an overloaded interval and characterize the departure and abandonment rate functions. In §5 we compare results of the numerical algorithm developed in §3 with simulation estimates of corresponding large-scale queueing models. In §6 we show how to construct a staffing function to stabilize delays at any fixed constant level v^*. Finally, in §7 we draw conclusions. Additional supporting material appears in an appendix.

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

We refer to §2 of [6] for a careful definition of the model; we provide a brief summary here.

The total input of fluid over the interval $[0, t]$ is $\Lambda(t) \equiv \int_0^t \lambda(u)\, du$, $t \geq 0$, where $\lambda \equiv \{\lambda(t) : t \geq 0\}$ is a time-dependent deterministic arrival-rate function. 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. Let $\bar{G}$ denote the associated complementary cdf (ccdf), defined by $\bar{G}(x) \equiv 1 - G(x)$. We assume that $\bar{G}(x) > 0$ and $\bar{F}(x) > 0$ for all x, but that condition on the service-time distribution can be relaxed; see Remark 3.2. We assume that the mean service time is 1, which is without loss of generality, because it simply corresponds to measuring time in units of mean service times. 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 $Q(t,y)$: $B(t,y)$ ($Q(t,y)$) is the quantity of fluid in service (queue) at time t that has been in service for time less than or equal to y. These functions will admit representations

$$Q(t,y) = \int_0^y q(t,x)\,dx \quad \text{and} \quad B(t,y) = \int_0^y b(t,x)\,dx, \quad y \geq 0, \tag{2.1}$$

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. The initial conditions are specified by the two functions $B(0,y)$ and $Q(0,y)$, which are defined as above, and also satisfy (2.1) with densities $b(0,x)$ and $q(0,x)$. We assume that $B(0) < \infty$ and $Q(0) < \infty$. 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 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; see [10]. Let $\mathbb{C}_p^1$ denote the set of differentiable functions with derivatives that belong to $\mathbb{C}_p$.

Assumption 2.1 (*smoothness*) *$s, \Lambda, F, G, B(0,\cdot), Q(0,\cdot)$ are differentiable functions with derivatives $s', \lambda, f, g, b(0,\cdot), q(0,\cdot)$ in $\mathbb{C}_p$.*

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. 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 in §6 of [6] we indicated how to detect the first violation and then construct the minimum feasible staffing function greater than or equal to the given staffing function.

We let the service discipline in the fluid model be first-come first-served (FCFS). As a consequence of FCFS service, at time t there will be a boundary of the waiting time (BWT) $w(t)$ such that

$$w(t) \equiv \inf\{x > 0, q(t,y) = 0 \ \textit{for all}\ y > x\}. \tag{2.2}$$

Clearly, first, $w(t) \geq 0$ and, second, $w(t) > 0$ if and only if $Q(t) > 0$. (Equation (2.2) is informal, because it is circular, with w depending on q, while q depends on w. The BWT w has been carefully defined and characterized in §4.2 of [6].

System behavior is primarily determined by the following two fundamental evolution equations.

Assumption 2.2 (*fundamental evolution equations*) *For $t \geq 0$, $x \geq 0$ and $u \geq 0$,*

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)}, \quad q(t+u, x+u) = q(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad 0 \leq x < w(t), \tag{2.3}$$

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 PWT is analyzed in §4.3 of [6].

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

$$Q(t) = Q(0) + \Lambda(t) - A(t) - E(t) \quad \text{and} \quad B(t) = B(0) + E(t) - S(t), \quad t \geq 0. \tag{2.4}$$

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

Then

$$A(t) \equiv \int_0^t \alpha(u)\, du, \quad \text{where} \quad \alpha(t) \equiv \int_0^\infty q(t,x) h_F(x) dx, \quad t \geq 0. \tag{2.5}$$

(2.6)

$$E(t) \equiv \int_0^t b(u,0)\, du, \quad t \geq 0. \tag{2.7}$$

$$S(t) \equiv \int_0^t \sigma(u)\, du, \quad \text{where} \quad \sigma(t) \equiv \int_0^\infty b(t,x) h_G(x) dx, \quad t \geq 0 \tag{2.8}$$

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

$$(i) \quad Q(t_1) > 0 \quad \text{or} \quad (ii) \quad Q(t_1) = 0, \quad B(t_1) = s(t_1) \quad \text{and} \quad \lambda(t_1) > s'(t_1) + \sigma(t_1), \tag{2.9}$$

and ends at the *overload termination time*

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

Case (ii) in (2.9) 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

$$(i) \quad Q(t_2) < 0 \quad \text{or} \quad (ii) \quad Q(t_2) = 0, \quad B(t_2) = s(t_2) \quad \text{and} \quad \lambda(t_2) \le s'(t_2) + \sigma(t_2), \tag{2.11}$$

and ends at *underload termination time*

$$T_2 \equiv \inf\{u \ge t_2 : B(u) = s(u) \quad \text{and} \quad \lambda(u) > s'(u) + \sigma(u)\}. \tag{2.12}$$

As before, case (ii) in (2.11) 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 2.3 (*finitely many switches between intervals in finite time*) *For each underloaded interval, $T_2 > t_2$ for t_2 in* (2.11) *and T_2 in* (2.12)*, 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 special case of exponential (M) service, i.e., $g(t) \equiv e^{-t}$, and the extension to time-varying Markovian service (M_t), we provide sufficient conditions for Assumption 2.3 to be satisfied in [7]. However, from a practical perspective, Assumption 2.3 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.

3 An Overloaded Interval with General GI Service

In §3 of [6] we already determined the performance of the $G_t/GI/s_t + GI$ fluid model during an underloaded interval. We now obtain results for the service content density $b \equiv b(t,x)$ in an overloaded interval. Once we have succeeded in obtaining the service content density b, we can calculate all the other performance functions by applying the results in §4 of [6].

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 $T^* < T$.

Since the system is assumed to be overloaded over an initial interval $[0,T)$, the rate into service is determined by the rate service capacity becomes available. Thus, by (2.8), we have

$$b(t,0) = s'(t) + \sigma(t) = s'(t) + \int_0^\infty b(t,x) h_G(x) dx, \quad 0 \le t < T. \tag{3.1}$$

However, this is an equation with b appearing on both sides; it remains to show that there exists a unique solution and find it. From (2.3), we can write down an expression for $b(t,x)$ during the overloaded interval:

$$b(t,x) = b(t-x,0)\bar{G}(x)1_{\{x \le t\}} + b(0,x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}1_{\{x>t\}}, \tag{3.2}$$

where $b(0,x-t)$ is part of the initial conditions, but where $b(t-x,0)$ is only specified recursively through the integral equation (3.1) (with time shift). We now substitute equation (3.2) into equation (3.1) to obtain the following equation for the function $b(t,0)$:

$$b(t,0) = \hat{a}(t) + \int_0^t b(t-x,0) g(x)\, dx, \tag{3.3}$$

where

$$\hat{a}(t) \equiv s'(t) + \int_0^\infty \frac{b(0,y) g(t+y)}{\bar{G}(y)}\, dy. \tag{3.4}$$

From (3.4), we see that $\hat{a} \in \mathbb{C}_p \subseteq \mathbb{D}$ provided that the integral in (3.4) is finite. From (C.6), it is evident that $b(t,0)$ is a fixed point of the operator $\mathcal{T} : \mathbb{D} \to \mathbb{D}$, where

$$\mathcal{T}(u)(t) \equiv \hat{a}(t) + \int_0^t u(t-x) g(x)\, dx. \tag{3.5}$$

We can show that there exists a unique solution to equation (C.6) by applying the Banach (contraction) fixed point theorem. We will use the complete (nonseparable) normed space $\mathbb{D}$ with the uniform norm over the interval $[0, T]$, i.e.,

$$\|u\|_T \equiv \sup_{0 \le t \le T} \{|u(t)|\}. \tag{3.6}$$

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

Assumption 3.1 (*tail of $b(0, \cdot)$*) *The tail of $b(0, \cdot)$ is bounded relative to the service-time pdf g via*

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

Assumption 3.1 restricts the class of allowed service cdf's in a rather complicated way. It is important to note that Assumption 3.1 is always satisfied in the case of principle interest: if there exists y_0 such that $b(0, y) = 0$ for all $y \ge y_0$. That case occurs whenever the overloaded interval of interest begins at time t, $0 \le t < T$, after the system has begun empty with $b(0, y) \equiv 0$ for all y; then necessarily $b(t, y) = 0$ for all $y > t$, by virtue of Assumption 2.2. Then

$$\tau \le B(0, T) g^{\uparrow}(2T) / \bar{G}(T) < \infty, \tag{3.7}$$

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

Nevertheless, other initial conditions are interesting. For example, for the stationary model, we might start with the stationary fluid content, which has the form we have $b(0, y) = \bar{G}(y)$, $y \ge 0$, because $\bar{G}$ is the stationary-excess or equilibrium-residual-lifetime density of the service-time distribution; see [12]. Thus we now present other sufficient conditions for Assumption 3.1.

Remark 3.1 (*sufficient conditions for the bound when $B(t) - B(0, y) > 0$ for all y*) Clearly, we need to control the initial content density $b(0, y)$ and/or the service pdf $g(y)$ in order for Assumption 3.1 to hold. An easy sufficient condition directly related to the stationary fluid content density for the stationary model is for there to exist a constant K such that $b(0, y) \le K \bar{G}(y)$ for all $y \ge 0$. Another easy sufficient condition for the bound in Assumption 3.1 is to have

$$\sup_{0 \le t < T} \{ \int_0^\infty b(0, y) h_G(y + t) \, dy \} < \infty. \tag{3.8}$$

In turn, three different sufficient conditions for (3.8) are:

$$
\begin{array}{ll}
(i) & \sup_{x\geq 0}\{h_G(x)\} < \infty \quad \text{(bounded hazard rate, using } B(0)<\infty\text{)}; \\
(ii) & \text{there exists } \beta > 0 \text{ and } K \text{ such that} \\
 & \quad \int_0^\infty b(0,y)e^{\beta y}\,dy < \infty \quad \text{and} \quad h_G(x) \leq Ke^{\beta}x \quad \text{for all} \quad x \geq 0. \\
(iii) & \limsup_{y\to\infty}\{b(0,y)/\bar{G}(y)\} < \infty \\
 & \quad \text{(using} \quad \sup_{0\leq y\leq t} b(0,y) < \infty \quad \text{and} \quad \sup_{0\leq y\leq t} h_G(0,y) < \infty \quad \text{for all} \quad t \geq 0) \qquad (3.9)
\end{array}
$$

Theorem 3.1 (service content in the overloaded case) *Consider an overloaded interval* $[0,T]$. *If Assumption* 3.1 *holds, then the operator* $\mathcal{T}$ *in* (3.5) *is a monotone contraction operator on* $\mathbb{D}$ *with contraction modulus* $G(T)$ *for the norm* $\|\cdot\|_T$ *defined in* (C.1), *so that a finite function* $b(t,0)$ *is uniquely characterized via equation* (C.6). *Hence, for any* $u \in \mathbb{D}$, *the fixed point can be approximated by the* n*-fold iteration* $\mathcal{T}^{(n)}$ *of the operator* $\mathcal{T}$ *applied to* u, *with*

$$
\|\mathcal{T}^{(n)}(u) - \hat{b}\|_T \leq \frac{G(T)^n}{1-G(T)}\|\mathcal{T}(u) - u\|_T \to 0 \quad as \quad n \to \infty \qquad (3.10)
$$

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

Proof. Clearly, Assumption 3.1 implies that $\|\hat{a}\|_T < \infty$, so that $\mathcal{T}$ maps $\mathbb{D}$ into $\mathbb{D}$. Moreover, the contraction property follows from

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

Remark 3.2 Note we require $G(T) < 1$ in the proof of Theorem C.1, which holds because we have assumed that $\bar{G}(x) > 0$ for all x. However, that requirement is actually not necessary, because we can always work in an interval $[0,\delta]$ as long as $G(\delta) < 1$ for some $\delta > 0$. We can show the uniqueness of $b(\cdot,0)$ for all $0 \leq t \leq T$ by recursively considering successive intervals of length δ.

So far, we can only conclude that the function $b(t,0) \in \mathbb{D}$. We can obtain additional smoothness properties by imposing additional smoothness conditions on the model elements s and g. We use these properties for $b(\cdot,0)$ to establish properties of the ODE to calculate the BWT w in §4 of [6].

Corollary 3.1 (smoothness of service content in the overloaded case) *If* s' *and* g *are continuous, then* $b(\cdot,0)$ *is continuous as well. In that case,* $b(\cdot,x)$ *and* $b(t,x)$ *are elements of* $\mathbb{C}_p$ *for each* $x \geq 0$ *and* $t \geq 0$.

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

We close this subsection by briefly discussing alternative algorithms to calculate b. If Assumption 3.1 holds, then a finite function b is uniquely characterized via equation (3.2), where

$$b(t,x) = \hat{b}(t,x)/h_G(x), \quad 0 \le x \le t < T, \tag{3.11}$$

with $\hat{b}$ being the unique solution of the equation

$$\hat{b}(t,x) \equiv \hat{a}(t,x) + g(x) \int_0^{t-x} \hat{b}(t-x,y)\, dy, \quad 0 \le x \le t < T, \tag{3.12}$$

where

$$\hat{a}(t,x) \equiv g(x)s'(t-x) + g(x) \int_0^\infty \frac{b(0,y)g(y+t-x)}{\bar{G}(y)}\, dy \in \mathcal{F}_T. \tag{3.13}$$

We establish the existence of a unique solution to equation (3.12) by applying the Banach fixed point theorem on an appropriate space of functions of two variables; see §C for more details.

Although this new fixed-point equation is more complicated, it can lead to a PDE characterization of b. This PDE representation follows directly by differentiating in the equation (3.12). (Convenient cancellation occurs.)

Theorem 3.2 (*PDE for* $\hat{b}$) *Under the assumptions of Theorems* C.1 *and* C.2*, wherever* $\hat{b}$ *has first partial derivatives with respect to* t *and* x*, it satisfies the PDE*

$$\hat{b}_t(t,x) + \hat{b}_x(t,x) = \hat{y}(t,x) + \hat{z}(x)\hat{b}(t,x), \quad 0 \le x \le t \le T, \tag{3.14}$$

where

$$\hat{y}(t,x) \equiv \hat{a}_t(t,x) + \hat{a}_x(t,x) - \frac{g'(x)}{g(x)}\hat{a}(t,x) \quad \textit{and} \quad \hat{z}(x) \equiv \frac{g'(x)}{g(x)} \tag{3.15}$$

for $\hat{a}(t,x)$ *in* (C.4)*. (The functions* $\hat{y}$ *and* $\hat{z}$ *in* (3.15) *are well defined by the assumptions in Theorem* C.2.) *Associated with the PDE is the boundary condition*

$$\hat{b}(t,t) = \hat{a}(t,t) = g(t)s'(0) + g(t) \int_0^\infty b(0,y)h_G(y)\, dy, \quad 0 \le t \le T, \tag{3.16}$$

which is finite by (3.8)*.*

4 The Other Performance Measures

Almost everything about the other performance functions in an overloaded interval follows directly from §4 of [6]. We briefly review the results here; see §4 of [6] for more details.

The queue content density q can be expressed in terms of w and the associated queue content density where no fluid enters service, $\tilde{q}$. In particular, q can be shown to have the form

$$q(t,x) \quad = \quad \tilde{q}(t-x,0)\bar{F}(x)1_{\{x\le w(t)\wedge t\}} + \tilde{q}(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\le w(t)\}}, \tag{4.1}$$

where first, $\tilde{q}$ has the same form as b in an underloaded interval, i.e.,

$$\tilde{q}(t,x) = \lambda(t-x)\bar{F}(x)1_{\{x\le t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\}}. \tag{4.2}$$

while the boundary waiting time (BWT) w satisfies the ODE

$$w'(t) = \Psi(t,w(t)) \equiv 1 - \frac{b(t,0)}{\tilde{q}(t,w(t))} \tag{4.3}$$

for any initial value $w(0)$; see Theorem 4.1 of [6]. Then the potential waiting time (PWT) v, i.e., the virtual waiting time of a quantum of fluid with infinite patience, can be found by solving the equation $v(t-w(t)) = w(t)$, given the BWT w; see Theorem 4.3 of [6].

We now remark on extra conditions needed. In order to ensure that the potential waiting time (PWT) v is finite, we need to supplement Assumption 4.3 of [6] with

Assumption 4.1 (*minimum service hazard rate*) *There exists a constant $h_{G,L}$ such that $h_G(x) \ge h_{G,L} > 0$ for all $x \ge 0$.*

Theorem 4.1 (finite PWT) *Under Assumptions 4.3 of* [6] *and* 4.1*, the rate of service completion is bounded below: $\sigma(t) \ge s_L h_{G,L}$ for all $t \ge 0$. As a consequence,*

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

The proof of Theorem 4.1 is essentially the same as the proof of Theorem 4.2 of [6]. The remaining results for the PWT in Theorems 4.3 and 4.4 of [6] hold again here provided that we make Assumption 4.1 and the other assumptions in §3.

We next discuss the departure function S in (2.8) and the abandonment function A in (2.6). These flows are performance measures of interest in their own right, but they are also important because they enable us to extend the model treated here directly to open networks of fluid queues, in which the departing fluid or abandoning fluid from one queue become input to another queue;see

[7]. In this section we show that the flows S and A inherit the structure of the original input Λ, so that the results in this paper extend to open networks of fluid queues.

The following results are elementary. The proofs and other properties are given in Appendix F.

Theorem 4.2 (the departure rate) *Assume that the conditions in Theorem* C.2 *hold. For* $t \geq 0$,

$$\sigma(t) = \int_0^t b(t-x,0)g(x)\,dx + \int_0^\infty \frac{b(0,y)g(t+y)}{\bar{G}(y)}\,dy,$$

where $b(t,0) = \lambda(t-u)$ *in an underloaded interval, but is the solution to the fixed point equation in Theorem C.1 during an overloaded interval. As a consequence,* $\sigma \in \mathbb{C}_p$.

Theorem 4.3 (abandonment rate) *Assume that the conditions in Theorem 4.1 of* [6] *hold, so that the BWT* w *is well defined and continuous. For* $t \geq 0$,

$$\begin{aligned}\alpha(t) &= \left(\int_0^{w(t)} \lambda(t-x)f(x)\,dx\right) 1_{\{w(t)\leq t\}} \\ &\quad + \left(\int_0^t \lambda(t-x)f(x)\,dx + \int_0^{w(t)-t} \frac{q(0,y)f(t+y)}{\bar{F}(y)}\,dy\right) 1_{\{w(t)>t\}}.\end{aligned}$$

As a consequence, $\alpha \in \mathbb{C}_p$.

5 An Example with Simulation Comparison

The fluid approximation and algorithm for the $G_t/M/s_t + GI$ model developed in [6] have been shown to be effective, see [6] for details. However, it can be restrictive to assume an exponential (M) service distribution in many real-life service systems. For instance, statistical analysis shows that service times in call centers are often lognormally distributed, e.g., see [1].

In this section we illustrate the algorithm for the $G_t/GI/s_t + GI$ fluid model described in §3 by applying it to an example. Moreover, we compare it to simulation estimates for associated large-scale queueing models. In particular, we hereby apply the algorithm to an $M_t/H_2/s + E_2$ example, that has a simusoidal arrival rate function

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

a two-phase hyperexponential (H2) (the H_2) service distribution, i.e., the service pdf

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

a constant service capacity s, and an Erlang-2 (E2) (the E_2) patience distribution, i.e., the abandonment pdf

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

The E2 abandonment time has squared coefficient of variation (SCV) $C^2 \equiv Var(A)/E[A]^2 = 1/2$; for the H2 service time, we let $p = 0.5(1-\sqrt{0.6})$, $\mu_1 = 2p\mu$, $\mu_2 = 2(1-p)\mu$, which produces $C^2 = 4$. We let $\theta = 0.5$, $\mu = a = c = s = 1$, $b = 0.6\,a = 0.6$.

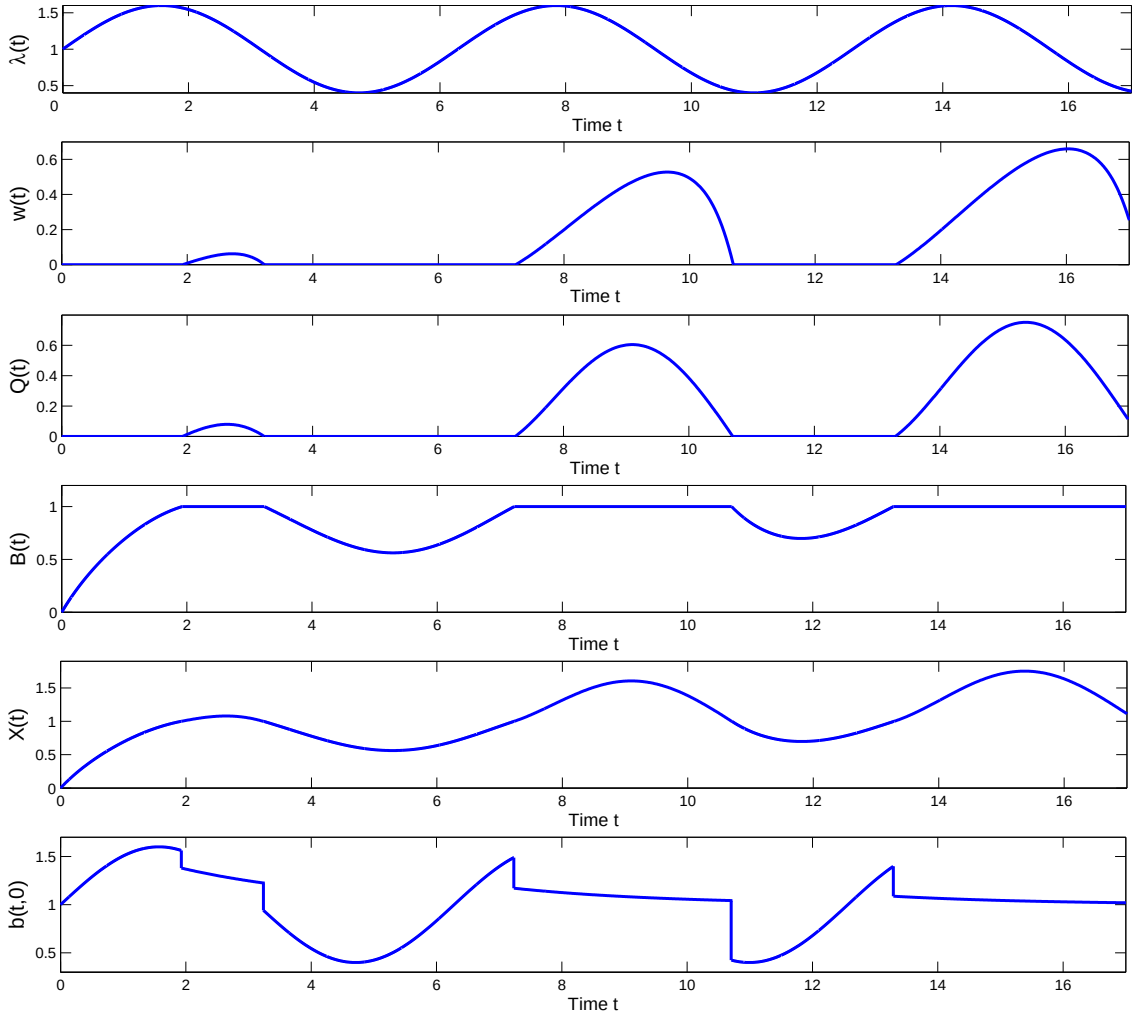

Figure 1: The performance functions of the $G_t/H_2/s + E_2$ fluid model with sinusoidal arrival-rate function: (i) arrival rate $\lambda(t)$; (ii) BWT $w(t)$; (iii) fluid in buffer $Q(t)$; (iv) fluid in service $B(t)$; (v) total fluid $X(t)$; (vi) rate into service $b(t,0)$.

We plot key fluid performance measures for $0 \leq t \leq T \equiv 17$, starting out empty, in Figure 1. Figure 1 clearly shows the alternating overloaded and underloaded intervals. All performance functions are continuous except for the rate-into-service function $b(t,0)$. In underloaded intervals,

$b(t,0) = \lambda(t)$; in overloaded intervals, $b(t,0)$ is the unique solution of the fixed-point equation (C.6). That is unlike the case of exponential service (as in [6]) where $b(t,0) = s'(t) + \mu s(t) = s = 1$. In the algorithm (as summarized in §B), we choose the error threshold parameter $\epsilon = 0.0001$. Although ϵ is small, it takes less than 20 iterations to meet the error threshold target.

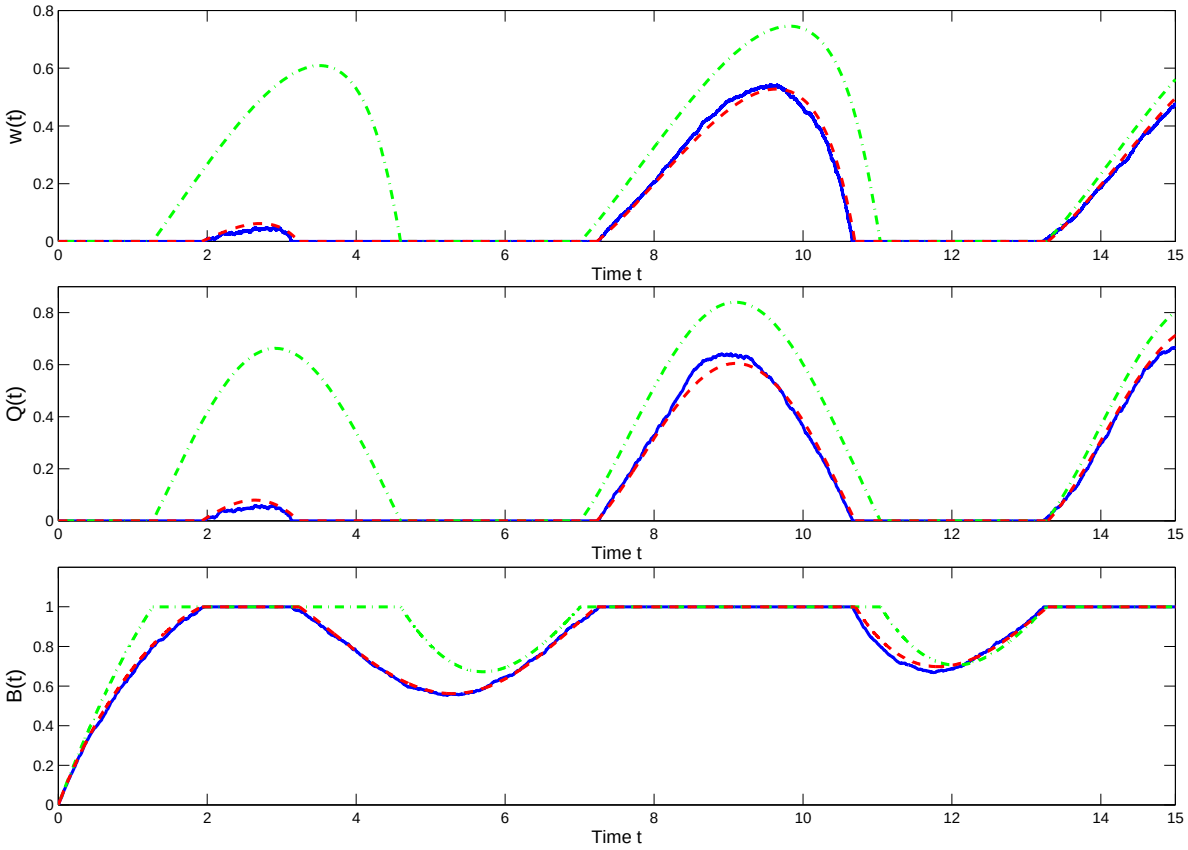

Figure 2: Simulation comparison of the $M_t/H_2/s + E_2$ fluid model: (i) simulation estimates of single sample paths of the scaled queueing model based on $n = 2000$ (blue solid lines), (ii) fluid functions computed by the algorithm in §3 (red dashed lines) and (iii) fluid functions assuming M service computed by the algorithm in [6] (green dashed lines).

We next compare this fluid approximation with computer simulations of the associated $M_t/H_2/s+E_2$ queueing system. We consider a sequence of queues under the standard fluid scaling (with factor n) discussed in §5 of [6], and compare the scaled queue-length processes and unscaled waiting time process

$$\left(\bar{B}_n \equiv \frac{\tilde{B}_n}{n}, \bar{Q}_n \equiv \frac{\tilde{Q}_n}{n}, W_n \right)$$

with the fluid performance functions (B, Q, w) computed from our fluid algorithm, where $\tilde{B}_n(t)$ ($\tilde{Q}_n(t)$) is the number of customers that are waiting in queue (service) at time t in queue n, and

$W_n(t)$ is the elapsed waiting time of the customer at the head of the queue at t. See §5 of [6] for details.

In Figure 2 we compare the simulation results for the queueing performance functions $(\bar{B}_n, \bar{Q}_n, W_n)$ with $n = 2000$ from a single simulation run (the blue solid lines) to the associated fluid model counterparts (B, Q, w) (the dashed red lines). Since n is large, we get close agreement for individual sample paths. Hence there is no need to show averages over multiple simulation runs. The green dashed lines show the corresponding fluid performance functions obtained by the algorithm in [6], assuming exponential service distribution with the same mean $1/\mu = 1$. As seen from Figure 2, the algorithm for the $G_t/M/s_t + GI$ model (although simple) is not effective for the model that has a general service distribution. (The fluid model with M service agrees closely with each simulation sample path when the queueing system has M service, as shown in [6].

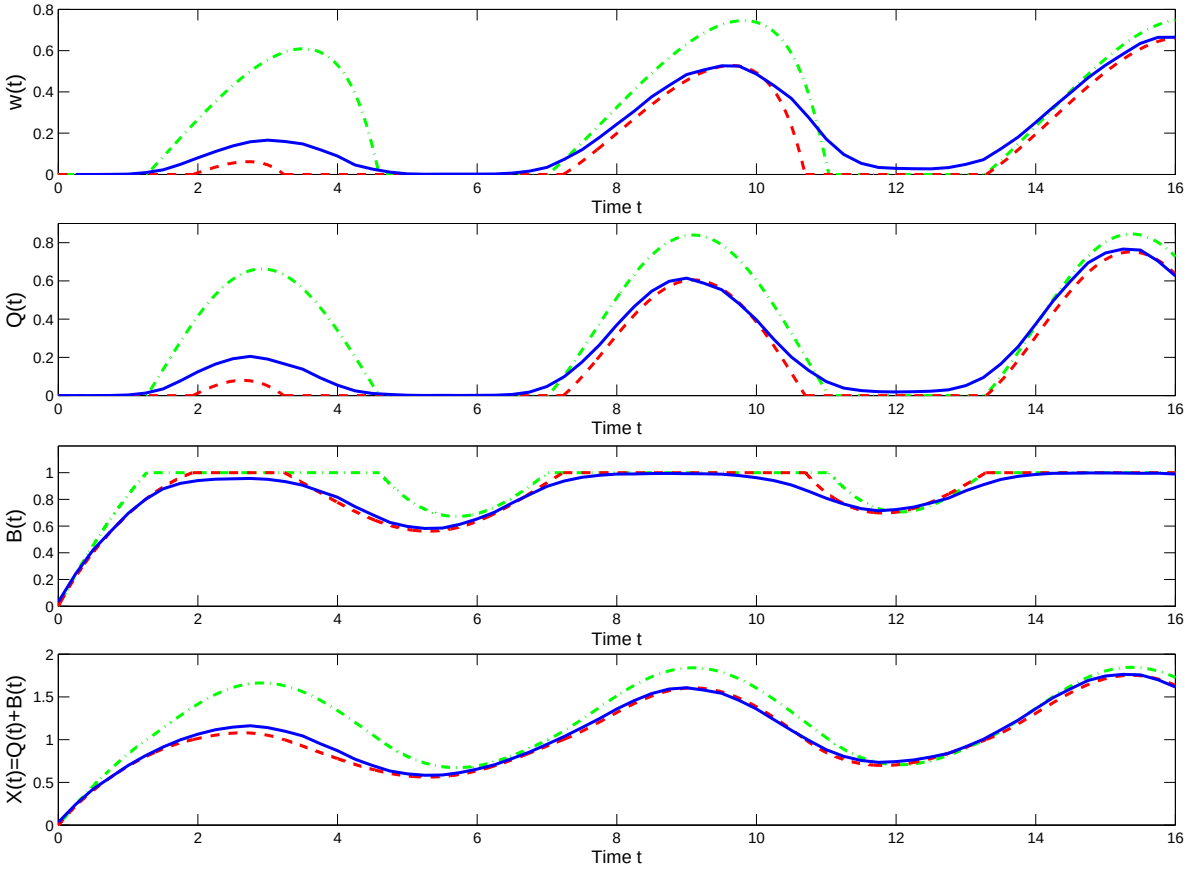

Figure 3: Simulation comparison of the $M_t/H_2/s + E_2$ fluid model: (i) simulation estimates of an average of 200 sample paths of the scaled queueing model based on $n = 30$ (blue solid lines), (ii) fluid functions computed by the algorithm in §3 (red dashed lines) and (iii) fluid functions assuming M service computed by the algorithm in [6] (green dashed lines).

When n is small, there are significant stochastic fluctuations, so that the sample paths from one simulation run are not close to the fluid performance functions. However, the mean values of these queueing performance functions still are quite well approximated by the fluid performance functions when the system is unambiguously overloaded and not nearly critically loaded. We illustrate by consider the case $n = 30$. As shown in Figure 3, the fluid model serves as a good approximation for the mean value functions, obtained by averaging the paths of 200 independent simulation runs. However, the fluid approximation is not good when the system is critically loaded or nearly critically loaded, because only positive fluctuations are captured while computing the mean value functions. However, the mean value process of the total number of customers is well approximated by its fluid approximation, as shown in the last plot of Figure 3. We also consider the case $n = 15$ in §D

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

So far, we have discussed the performance analysis of the $G_t/GI/s_t + GI$ fluid model with the staffing function s regarded as a given function. In this section, we assume that we are free to choose the staffing function s, and do so with the objective of stabilizing the potential waiting time v at some (constant) target $v^* > 0$. This delay stabilization problem was considered previously for many-server queueing models with time-varying arrival rates in [3].

As a consequence, of Theorem 4.3 of [6], we see that, in order to stabilize v at v^*, it suffices to stabilize w at v^*. By Theorem 4.1 of [6], we see that we will be able to do so if and only if we can find a staffing functions s for which the resulting performance satisfies the equation

$$0 = w'(t) = 1 - \frac{b(t,0)}{q(t,v^*)}, \quad t \geq 0 \tag{6.1}$$

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

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

With the initial conditions $q(t,0) = \lambda(t)$ and $q(0,x) = 0$, the queue instantly becomes overloaded at time 0, and we can apply Proposition 4.1 and Corollary 4.1 of [6] (or (2.3)) to obtain

$$q(t,x) = \bar{F}(x)\lambda(t-x)1_{\{0\leq x\leq t\}}, \quad 0 \leq t \leq v^*. \tag{6.2}$$

Combining (6.1) and (6.2), we obtain the transportation rate after $t = v^*$:

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

With the explicit expression of $b(t,0)$ and $b(0,x) \equiv 0$, $x \geq 0$, (2.3) implies that

$$b(t,x) = \bar{G}(x)\bar{F}(v^*)\lambda(t-x-v^*)1_{\{0\leq x\leq t-v^*\}}, \quad t \geq 0 \quad \text{and} \quad x \geq 0. \tag{6.3}$$

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

Theorem 6.1 *Consider the $G_t/GI/s_t + GI$ fluid model with a general arrival-rate function λ. Suppose the system is initially empty. For any specified constant $v^* > 0$, we can make the system overloaded such that the PWT is fixed at v^*, i.e., $v(t) = v^*$ for all $t \geq 0$, by (i) not allowing any input to enter service until time $t = v^*$, (ii) letting the service-capacity function be*

$$s(v^*,t) \equiv s^*(t) = \bar{F}(v^*)\int_0^{t-v^*} \bar{G}(x)\lambda(t-v^*-x)dx \cdot 1_{\{t>v^*\}} \tag{6.4}$$

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

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

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

Remark 6.1 (general initial conditions or no delay) Theorem 6.1 is based on starting empty. However, it is possible to stabilize delays with arbitrary initial conditions. We present the details in Appendix E. We can also achieve the minimum staffing level so that there is no delay at all by simply staffing at the fluid content $B(t)$ in the underloaded regime, as specified in §3 in [6]. These two variants may involve having an atom of initial fluid content enter service at time 0, so that we leave the smooth framework.

7 Conclusions

In §3, we showed how the results for the $G_t/M/s_t + GI$ fluid model in [6] can be extended to a large class of non-exponential service-time distributions, but we must solve a fixed point equation to find the service content density b. We applied the Banach contraction fixed point theorem to establish the existence of a unique solution to the fixed point equation and provide the basis for an efficient algorithm using iteration. Another application of Banach contraction fixed point theorem ensures that b has appropriate smoothness, under regularity conditions. After finding b, the rest of the performance functions are computed exactly as for exponential service.

We have demonstrated the importance of the extension from M service to G. In §5 we have compare simulation results of queueing systems to two versions of fluid approximations: (i) the $G_t/M/s_t + GI$ algorithm in [6] and (ii) the $G_t/GI/s_t + GI$ algorithm in §3. The simulation experiment shows that (ii) is effective as an approximation for mean values even when the scale is not too large. Moreover, the experiment shows that (i) is not effective and that the non-exponential service distribution plays an important role in the fluid dynamics.

In §6 we showed that, instead of taking the staffing function as given, we can choose a staffing function in order to stabilize the PWT v at any (constant) target $v^* > 0$. When that is done, $v(t) = v^*$ for all $t \geq 0$, but other performance measures are not constant. Expressions for all other performance functions were given.

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

References

[1] 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.

[2] J. L. Davis, W. A. Massey and W. Whitt. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Science* **41** (1995) 1107–1116.

[3] Z. Feldman, A. Mandelbaum, W. A. Massey and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54** (2008), 324–338.

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

[5] 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.

[6] Y. Liu and W. Whitt, A fluid model for a large-scale service system experiencing periods of overloading. working paper, Columbia University, New York, NY, 2010. Available at: http://www.columbia.edu/∼ww2040/allpapers.html

[7] Y. Liu and W. Whitt, A network of time-varying many-server fluid queues with customer abandonment. working paper, Columbia University, New York, NY, 2010. Available at: http://www.columbia.edu/∼ww2040/allpapers.html

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

[9] 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.

[10] W. Whitt, *Stochastic-Process Limits*, Springer, New York, 2002.

[11] W. Whitt, W. Engineering solution of a basic call-center model. *Management Sci.* **51** (2005) 221–235.

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

[13] 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 the Many-Server $G_t/GI/s_t + GI$ Queue

A Overview

This appendix contains material supplementing the main paper. In §B we summarize the algorithm that characterizes the fluid performance functions in an overloaded interval (based on §3). In §C we show that the two-parameter fluid content density b can indeed be directly represented as the solution of the equation (3.12); i.e., we show that the equation has a unique solution. In §D we present more simulation comparison for the example considered in 5. In §E we present additional material related to §6 on choosing staffing functions to stabilize delays. In particular, we show how to stabilize delays with general initial conditions. Additional material on the flows is provided here in §F.

B The Algorithm for an Overloaded Interval

We formally state the algorithm which computes all performance functions in an overloaded interval of the $G_t/GI/s_t + GI$ fluid model. Consider an interval $[0, T]$ and assume we know the system is overloaded at $t = 0$, i.e., $Q(0) > 0$ and $B(0) = s(0)$. Here we may not know when the overloaded interval ends in advance. The objective is to determine the overload termination time T_1 defined in (2.10) with $t_1 = 0$ along with the other performance functions. Hence, we determine $q(t, \cdot)$ and $b(t, \cdot)$ for $0 \le t \le T \wedge T_1$. If $T_1 < T$, the system simply switches to an underloaded interval; otherwise, the system stays overloaded in $[0, T]$.

The input functions are the model parameters F, G, $\lambda(t)$ and $s(t)$ for $0 \le t \le T$ and initial condition $q(0, \cdot)$, $b(0, \cdot)$ and $w(0)$. We require these conditions satisfy (i) $s(0) = B(0) = \int_0^\infty b(0, y)dy$ and (ii) $Q(0) = \int_0^{w(0)} q(0, y)dy > 0$. Exploring the fixed-point operator discussed in §3, we have the following algorithm:

Step 1: $u^{(0)}(t) \leftarrow 0, \quad a(t) \leftarrow s'(t) + \int_0^\infty b(0, y)\frac{g(t+y)}{G(y)}dy, \quad i \leftarrow 1$

Step 2: $u^{(i)}(t) \leftarrow a(t) + \int_0^t u^{(i-1)}(y)g(t - y)dy \quad$ for $\quad 0 \le t \le T$

Step 3: If $\|u^{(i)} - u^{(i-1)}\|_T > \epsilon$, then $i \leftarrow i + 1$ and go to Step 2;
otherwise $b(t, 0) \leftarrow u^{(i)}(t) \quad$ for $\quad 0 \le t \le T$

Step 4: Solve the BWT ODE and determine T_1.

Step 5: Compute $b(t, x)$ with (3.2) for $0 \le t \le T \wedge T_1$. End.

Note that $\epsilon > 0$ is the error threshold level that we can specify in advance. Here we let the contraction iteration in Step 2 ends when the uniform distance of the u functions in two consecutive iterations is small. We hereby assume that the given staffing function s is feasible. However, we can also easily modify the algorithm so that infeasibility can be detected. In Step 4, we claim that s is infeasible if we find $b(t, 0) \le 0$ for some $0 \le t \le T$.

C A Fixed-Point Equation for $b(t,x)$

To consider the two-parameter function $b \equiv b(t,x)$, we use the space $\mathcal{F}_{T,1}$ of measureable real-valued functions of the pair of real variables (t,x) over the "triangular" domain $0 \le x \le t \le T$, for which the norm

$$\|u\|_{T,1} \equiv \sup_{0\le t\le T} \int_0^t |u(t,x)|\,dx. \tag{C.1}$$

is finite. The norm $\|\cdot\|_{T,1}$ is an L_1 norm in one coordinate and an L_∞ norm in the other; it makes $\mathcal{F}_{T,1}$ a Banach space.

Theorem C.1 (service content in the overloaded case) *Consider an overloaded interval $[0,T)$. If Assumption* 3.1 *holds, then a finite function b is uniquely characterized via equation* (3.2)*, where*

$$b(t,x) = \hat{b}(t,x)/h_G(x), \quad 0 \le x \le t < T, \tag{C.2}$$

with $\hat{b}$ being the unique fixed point of the operator $\mathcal{T} : \mathcal{F}_{T,1} \to \mathcal{F}_{T,1}$ defined by

$$\mathcal{T}(u)(t,x) \equiv \hat{a}(t,x) + g(x)\int_0^{t-x} u(t-x,y)\,dy, \quad 0 \le x \le t < T, \tag{C.3}$$

where

$$\hat{a}(t,x) \equiv g(x)s'(t-x) + g(x)\int_0^\infty \frac{b(0,y)g(y+t-x)}{\bar{G}(y)}\,dy \in \mathcal{F}_T. \tag{C.4}$$

Moreover, the operator $\mathcal{T}$ is a monotone contraction operator on $\mathcal{F}_{T,1}$ with contraction modulus $G(T)$ for the norm $\|\cdot\|_{T,1}$ defined in (C.1)*, so that, for any $u \in \mathcal{F}_{T,1}$, the fixed point can be approximated by the n-fold iteration $\mathcal{T}^{(n)}$ of the operator $\mathcal{T}$ applied to u, with*

$$\|\mathcal{T}^{(n)}(u) - \hat{b}\|_{T,1} \le \frac{G(T)^n}{1-G(T)}\|\mathcal{T}(u) - u\|_{T,1} \to 0 \quad as \quad n \to \infty \tag{C.5}$$

and, if $u \le (\ge)\mathcal{T}(u)$, then $\mathcal{T}^{(n-1)}(u) \le (\ge)\mathcal{T}^{(n)}(u) \le (\ge)\hat{b}$ for all $n \ge 1$. Finally, $\hat{b}(t,t) = \hat{a}(t,t) = g(t)b(0,0)$.

Proof. First, we show that $\hat{b}$ in (C.2) is a fixed point of the operator $\mathcal{T}$, i.e., that $\mathcal{T}(\hat{b}) = \hat{b}$. To see that, multiply (3.2) through by $h_G(x)$, noting that (i) $h_G(x)\bar{G}(x) = g(x)$ and (ii) we are interested in the case $x \le t$. We get $\hat{b}(t,x) = b(t,x)h_G(x) = b(t-x,0)g(x)$. Next we successively apply (3.1), (2.3) and a change of variables to get

$$\begin{aligned}
\hat{b}(t,x) &= b(t-x,0)g(x) = s'(t-x)g(x) + g(x)\int_0^\infty b(t-x,y)h_G(y)\,dy \\
&= s'(t-x)g(x) + g(x)\int_{t-x}^\infty b(t-x,y)h_G(y)\,dy + g(x)\int_0^{t-x} b(t-x,y)h_G(y)\,dy \\
&= s'(t-x)g(x) + g(x)\int_{t-x}^\infty b(0,y-(t-x))\frac{\bar{G}(y)}{\bar{G}(y-(t-x))}h_G(y)\,dy + g(x)\int_0^{t-x}\hat{b}(t-x,y)\,dy \\
&= s'(t-x)g(x) + g(x)\int_0^\infty b(0,y)\frac{g(y+t-x)}{\bar{G}(y)}\,dy + g(x)\int_0^{t-x}\hat{b}(t-x,y)\,dy \\
&= \hat{a}(t,x) + g(x)\int_0^{t-x}\hat{b}(t-x,y)\,dy = \mathcal{T}(\hat{b})(t,x),
\end{aligned} \tag{C.6}$$

where $\hat{a}(t,x) = \hat{c}(t,x) + \hat{d}(t,x)$ with

$$\hat{c}(t,x) \equiv g(x)s'(t-x) \quad \text{and} \quad \hat{d}(t,x) \equiv g(x)\int_0^\infty b(0,y)\frac{g(y+t-x)}{\bar{G}(y)}\,dy.$$

We next show that $\|\hat{a}\|_{T,1} < \infty$. First, $\|\hat{c}\|_{T,1} \le G(T)\|s'\|_T < \infty$ because $s' \in \mathbb{C}_p \subset \mathbb{D}$. Because of the factor $g(x)$, $\|\hat{d}\|_{T,1}$ is bounded by the integral term. Taking the supremum over x and t with $0 \le x \le t \le T$ of the integral in the expression for $\hat{d}$ yields the term τ in Assumption 3.1, which we have assumed is bounded. Hence $\|\hat{d}\|_{T,1} < \infty$, and so $\|\hat{a}\|_{T,1} < \infty$.

Next note that $\mathcal{T}$ is indeed a contraction operator on $(\mathcal{F}_{T,1}, \|\cdot\|_{T,1})$, because

$$\|\mathcal{T}(u_1) - \mathcal{T}(u_2)\|_{T,1} \quad \le \quad \sup_{0\le t\le T}\int_0^t g(x)\left(\int_0^{t-x}|u_1-u_2|(t-x,y)\,dy\right)dx \le G(T)\|u_1-u_2\|_{T,1},$$

and we have assumed that $G(T) < 1$ for all T. The geometric rate of convergence in (C.5) is the standard conclusion from the Banach fixed point theorem, and the subsequent ordering follows from the monotonicity of $\mathcal{T}$. Finally, $\hat{b}(t,t) = \hat{a}(t,t)$ because the subset of u in $\mathcal{F}_{T,1}$ for which $u(t,t) = \hat{a}(t)$ is closed, and $\mathcal{T}$ maps that subset into itself, because $\mathcal{T}(u)(t,t) = \hat{a}(t,t)$, $0 \le t \le T$, for all u in $\mathcal{F}_{T,1}$. By (3.1), $\hat{a}(t,t) = g(t)b(0,0)$. ∎

We now provide conditions for $\hat{b}(\cdot,x)$ and $b(\cdot,x)$ to be in $\mathbb{C}_p$ for all $x \ge 0$. (We use these properties for $b(\cdot,0)$ to establish properties of the ODE to calculate the BWT w in §4 of [6].) We first introduce extra smoothness conditions.

Assumption C.1 (*extra smoothness for g and s*) *g and s' are differentiable with derivatives g' and s'' in $\mathbb{C}_p$.*

We next impose additional regularity conditions on the service-time pdf g. For that purpose, let $\|g\|_\infty$ be the uniform norm, i.e., $\|g\|_\infty \equiv \sup_{x\ge 0}\{|g(x)|\}$.

Assumption C.2 (*extra regularity for g*) *The service-time pdf g satisfies: $g(x) > 0$ for all x, $\|g\|_\infty < \infty$ and there exists K such that $g(x) \le g(0)e^{Kx}$ for all $x \ge 0$.*

We will use the last inequality in Assumption C.2 in its equivalent form: $|g'(x)| \le Kg(x)$ for all x. (To see the equivalence, Divide by $g(x)$, integrate and take the exponential.)

Theorem C.2 (smoothness of service content in the overloaded case) *If Assumptions* 3.1–C.2 *all hold, then $\hat{b}(\cdot,x)$ and $b(\cdot,x)$ are differentiable functions for each $x \ge 0$, almost everywhere equal to their partial derivatives with respect to t, for b in* (C.2) *and $\hat{b}$ in* (C.3)*. Hence, $\hat{b}(\cdot,x), b(\cdot,x) \in \mathbb{C}_p$ for all $x \ge 0$.*

Proof. We again apply the Banach fixed point theorem, but now on a subspace of $\mathcal{F}_{T,1}$ with a new norm. Consider the subspace of measurable real-valued functions u of the pair of real variables (t,x) over the same triangular domain $0 \le x \le t \le T$ that are differentiable with respect to the variable t, and equal almost everywhere to the integral of its partial derivative u_t, with finite norm $\|u\|_{T,2}$, where

$$\|u\|_{T,2} \equiv \sup_{0\le t\le T}\{\int_0^t (|u(t,x)| + |u_t(t,x)|)\,dx,\} \tag{C.7}$$

which is like the Sobolev norm on the Sobolev space $\mathcal{W}^{1,\infty}(0,t)$. The functions in $\mathcal{F}_{T,2}$ are Lipschitz continuous in the first variable t for each x in $0 \le x \le t \le T$. Reasoning as in the proof of Theorem C.1, we will show that $\|\hat{a}\|_{T,2} < \infty$, and then we will show that $\mathcal{T}$ maps $\mathcal{F}_{T,2}$ into itself.

Then,

$$\|\hat{a}\|_{T,2} \le \|\hat{a}\|_{T,1} + G(T)\left(\|s''\|_T + K \sup_{0\le s\le T}\int_0^\infty (b(0,y)g(s+y)/\bar{G}(y))\,dy\right) < \infty$$

by the proof of Theorem C.1 and the conditions in Assumptions 3.1, C.1 and C.2. (Since $\mathbb{C}_p \subset \mathbb{D}$, $\|s''\|_T < \infty$.) Next, $\|\mathcal{T}(u)\|_{T,2} \le \|\hat{a}\|_{T,2} + G(T)(\|u\|_{T,1} + \sup_{0\le t\le T}\{|u(t,t)\} + \|u_t\|_{T,1}) < \infty$. Then we see that $\mathcal{T}$ is again a contraction operator on $(\mathcal{F}_{T,2}, \|\cdot\|_{T,2})$ with modulus $G(T)$. We can ignore the term involving $|u_1(t,t) - u_2(t,t)|$, because, as noted at the end of Theorem C.1, we can restrict attention to the closed subspace $\mathcal{F}_{T,2}$ containing only u for which $u(t,t) = g(t)b(0,0)$; as a consequence, $u_1(t,t) = u_2(t,t)$ for all t. Hence, the fixed point $\hat{b}$ is an element of $\mathcal{F}_{T,2}$, and so has the claimed smoothness properties. ∎

D More Simulation Comparison of the Example in §5

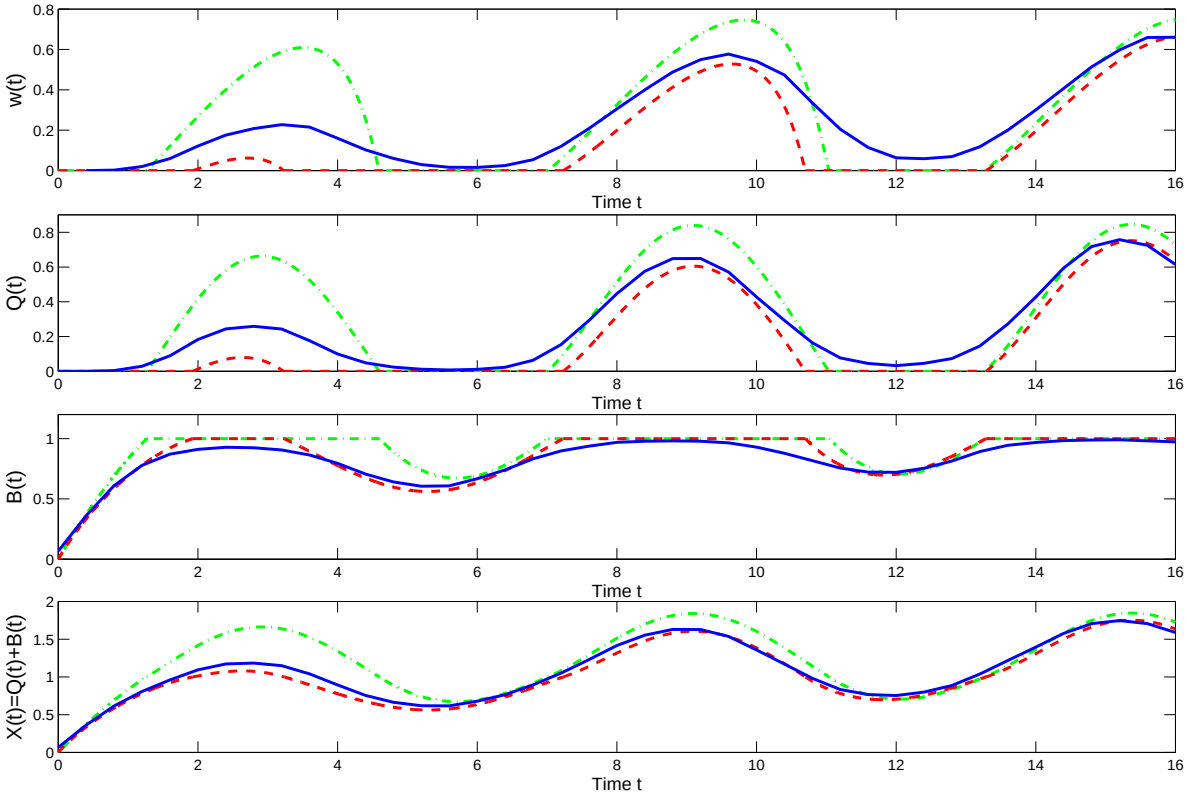

Figure 4: Simulation comparison of the $M_t/H_2/s + E_2$ fluid model: (i) simulation estimates of an average of 500 sample paths of the scaled queueing model based on $n = 15$ (blue solid lines), (ii) fluid functions computed by the algorithm in §3 (red dashed lines) and (iii) fluid functions assuming M service computed by the algorithm in [6] (green dashed lines).

Here we consider the $M_t/H_2/s + E_2$ example in §5 with smaller n. As shown in Figure 4, we plot the mean value functions, obtained by averaging the paths of 500 independent simulation runs, with $n = 15$. Although less accurate than the case $n = 30$, the fluid model serves as a much better approximation than the algorithm of M service.

E Stabilizing Delays with General Initial Conditions

Is §6 we showed how to choose a staffing function to stabilize the PWT v at any desired target v^*. However, Theorem 6.1 considered a special initial condition: the system is initially empty. We generalize Theorem 6.1 to arbitrary initial conditions in the next theorem.

Theorem E.1 *Consider the $G_t/GI/s_t+GI$ fluid model with a general arrival-rate function λ and initial conditions $w(0-) \equiv w_0 \geq 0$, $b(0-,x) \equiv \psi(x) \geq 0$ for $x \geq 0$, $q(0-,x) \equiv \phi(x) \geq 0$ for $0 \leq x \leq w_0$, $Q(0-) = \int_0^{w_0} q(0-,x)dx$, $s(0-) = B(0-) = \int_0^\infty b(0-,x)dx$. For any given $v^* \geq 0$, we can make the system overloaded such that the PWT is fixed at v^*, i.e., $v(t) = v^*$ for all $t \geq 0$, by letting the service-capacity function be*

$$\begin{aligned} s(t) &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0 \vee v^*} \phi(x)dx \\ &+ \bar{F}(v^*)\left(\int_{(t-v^*)^+}^{t-(v^*-w_0)^+} \frac{\phi(w_0 \wedge v^* - t + x)\,\bar{G}(x)}{\bar{F}(w_0 \wedge v^* - t + x)}dx\right) \cdot 1_{\{t \geq (v^*-w_0)^+\}} \\ &+ \bar{F}(v^*)\left(\int_0^{t-v^*} \lambda(t-x-v^*)\bar{G}(x)dx\right) \cdot 1_{\{t \geq v^*\}}. \end{aligned} \tag{E.1}$$

If we do so, then

$$
\begin{aligned}
w(t) &= v^* \cdot 1_{\{t \geq (v^*-w_0)^+\}}, \\
b(t,0) &= \delta_0(t) \int_{v^*}^{w_0 \vee v^*} \phi(x)dx + \frac{\phi(w_0 \wedge v^* - t)\, \bar{F}(v^*)}{\bar{F}(w_0 \wedge v^* - t)} \cdot 1_{\{(v^*-w_0)^+ \leq t < v^*\}} + \lambda(t-v^*)\bar{F}(v^*) \cdot 1_{\{t \geq v^*\}}, \\
B(t) &= s(t), \\
\sigma(t) &= \int_t^\infty \psi(x-t) \frac{g(x)}{\bar{G}(x-t)} dx + g(t) \int_{v^*}^{w_0 \vee v^*} \phi(x)dx \\
&+ \bar{F}(v^*) \left(\int_{(t-v^*)^+}^{t-(v^*-w_0)^+} \frac{\phi(w_0 \wedge v^* - t + x)\, g(x)}{\bar{F}(w_0 \wedge v^* - t + x)} dx \right) \cdot 1_{\{t \geq (v^*-w_0)^+\}} \\
&+ \bar{F}(v^*) \left(\int_0^{t-v^*} \lambda(t-x-v^*) g(x) dx \right) \cdot 1_{\{t \geq v^*\}}, \\
Q(t) &= \left(\int_t^{w_0 \wedge v^*} \frac{\phi(x-t)\, \bar{F}(x)}{\bar{F}(x-t)} dx + \int_0^t \lambda(t-x)\, \bar{F}(x) dx \right) \cdot 1_{\{0 \leq t \leq (v^*-w_0)^+\}} \\
&+ \left(\int_0^t \lambda(t-x)\, \bar{F}(x) dx + \int_t^{v^*} \frac{\phi(x-t)\, \bar{F}(x)}{\bar{F}(x-t)} dx \right) \cdot 1_{\{(v^*-w_0)^+ < t < v^*\}} \\
&+ \left(\int_0^{v^*} \lambda(t-x)\, \bar{F}(x) dx \right) \cdot 1_{\{t \geq v^*\}}, \\
\alpha(t) &= \left(\int_t^{w_0 \wedge v^*} \frac{\phi(x-t)\, f(x)}{\bar{F}(x-t)} dx + \int_0^t \lambda(t-x)\, f(x) dx \right) \cdot 1_{\{0 \leq t \leq (v^*-w_0)^+\}} \\
&+ \left(\int_0^t \lambda(t-x)\, f(x) dx + \int_t^{v^*} \frac{\phi(x-t)\, f(x)}{\bar{F}(x-t)} dx \right) \cdot 1_{\{(v^*-w_0)^+ < t < v^*\}} \\
&+ \left(\int_0^{v^*} \lambda(t-x)\, f(x) dx \right) \cdot 1_{\{t \geq v^*\}}
\end{aligned}
$$

where $\delta_y(t)$ is the direct-delta function at y, i.e., $\delta_y(t) = 0$ for $t \neq y$, $\int_a^b \delta_y(t)dt = 1$ if $a \leq y \leq b$.

Proof. (i) If the system is initially underloaded, i.e., $w(0-) = w_0 = 0$, $q(0-,x) = \phi(x) = 0$, $Q(0-) = 0$, $B(0-) \leq s(0-)$. This case is similar to Theorem 6.1 where the system is initially empty. Note the only difference is that there is fluid in the service facility. Let $B^o(t)$ be the fluid in service that has been in service at $0-$. Then we have

$$B^o(t) = \int_t^\infty b(t,x)dx = \int_t^\infty b(0-,x-t) \frac{\bar{G}(x)}{\bar{G}(x-t)} dx.$$

Again, we do not allow any input to enter service until time $t = v^*$, we can let the staffing function be

$$
\begin{aligned}
s(t) &= B^o(t) + s^*(t) \\
&= \int_t^\infty \frac{\psi(x-t)\, \bar{G}(x)}{\bar{G}(x-t)} dx + \bar{F}(v^*) \int_0^{t-v^*} \bar{G}(x) \lambda(t-v^*-x) dx \cdot 1_{\{t > v^*\}},
\end{aligned}
$$

where $s^*(t)$ is defined in (6.4). It is obvious that this expression coincides with (E.1) when $w_0 = q(0-,x) = \psi(x) = 0$. When we do this, the input rate to the service $b(t,0)$ is the same as in Theorem 6.1. The proof of other performance measures are similar.

(ii) If the system is initially overloaded, i.e., $w(0-) = w_0 > 0$, $q(0-,x) = \phi(x) \geq 0$, $Q(0-) = \int_0^{w_0} \phi(x)dx > 0$, $s(0-) = B(0-)$. There are two cases (a) $w_0 \geq v^*$, (b) $w_0 < v^*$.

(ii.a) If $w_0 > v^*$, then in order for $v(t) = v^*$. We let all fluid that has been in queue for $x > v^*$ enter service immediately at time 0. The quantity of fluid that enters service at 0 is $\int_{v^*}^{w_0} q(0-,x)dx = \int_{v^*}^{w_0} \phi(x)dx$. However, this will make $B(t)$ have an atom at 0. Similar argument to Theorem 6.1 implies that it suffices to match $b(t,0)$ with $q(t,v^*)$ for all $t \geq 0$. If $t \leq v^*$, $q(t,v^*) = q(0-,v^*-t)\bar{F}(v^*)/\bar{F}(v^*-t)$. If $t > v^*$, then all fluid that has been in queue at $0-$ has entered service, which implies that $q(t,v^*) = q(t-v^*,0)\bar{F}(v^*) = \lambda(t-v^*)\bar{F}(v^*)$. Therefore, we have

$$\begin{aligned} b(t,0) &= \delta_0(t)\int_{v^*}^{w_0} \phi(x)dx + q(t,v^*) \\ &= \delta_0(t)\int_{v^*}^{w_0} \phi(x)dx + \frac{\phi(v^*-t)\bar{F}(v^*)}{\bar{F}(v^*-t)} \cdot 1_{\{0\leq t<v^*\}} + \lambda(t-v^*)\bar{F}(v^*) \cdot 1_{\{t\geq v^*\}}. \end{aligned}$$

The service capacity and fluid content in service are

$$s(t) = B(t) = B^o(t) + \int_0^t b(t-x,0)\bar{G}(x)dx.$$

If $0 \leq t < v^*$, we have

$$\begin{aligned} s(t) &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \int_{v^*}^{w_0} \phi(x)dx \int_0^t \delta_0(t-x)\bar{G}(x)dx + \bar{F}(v^*)\int_0^t \frac{\phi(v^*-t+x)\bar{G}(x)}{\bar{F}(v^*-t+x)}dx, \\ &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0} \phi(x)dx + \bar{F}(v^*)\int_0^t \frac{\phi(v^*-t+x)\bar{G}(x)}{\bar{F}(v^*-t+x)}dx. \end{aligned}$$

If $t \geq v^*$, we have

$$\begin{aligned} s(t) &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0} \phi(x)dx \\ &+ \int_0^t \left(\frac{\phi(v^*-t+x)\bar{F}(v^*)}{\bar{F}(v^*-t+x)} \cdot 1_{\{0\leq t-x<v^*\}} + \lambda(t-x-v^*)\bar{F}(v^*) \cdot 1_{\{t-x\geq v^*\}} \right) \bar{G}(x)dx \\ &= \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx + \bar{G}(t)\int_{v^*}^{w_0} \phi(x)dx \\ &+ \bar{F}(v^*)\left(\int_{t-v^*}^t \frac{\phi(v^*-t+x)\bar{G}(x)}{\bar{F}(v^*-t+x)}dx + \int_0^{t-v^*} \lambda(t-x-v^*)\bar{G}(x)dx \right). \end{aligned}$$

It is easy to see that this expression coincides with (E.1).

(ii.b) If $w_0 \leq v^*$, then we do not allow any input to enter service until time $v^* - w_0$, which implies

$$b(t,0) = \frac{\phi(w_0-t)\,\bar{F}(v^*)}{\bar{F}(w_0-t)} \cdot 1_{\{v^*-w_0\leq t<v^*\}} + \lambda(t-v^*)\bar{F}(v^*) \cdot 1_{\{t\geq v^*\}}.$$

Therefore, if $0 \leq t \leq v^* - w_0$, no new fluid enters service,

$$s(t) = B^o(t) = \int_t^\infty \psi(x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}dx.$$

If $v^* - w_0 < t < v^*$,

$$\begin{aligned} s(t) &= B^o(t) + \int_0^t \frac{\phi(w_0 - t + x)\,\bar{F}(v^*)}{\bar{F}(w_0 - t + x)} \cdot 1_{\{v^* - w_0 \le t - x < v^*\}} \bar{G}(x) dx \\ &= \int_t^\infty \psi(x-t) \frac{\bar{G}(x)}{\bar{G}(x-t)} dx + \bar{F}(v^*) \int_0^{t-(v^*-w_0)} \frac{\phi(w_0 - t + x)\,\bar{G}(x)}{\bar{F}(w_0 - t + x)} dx. \end{aligned}$$

If $t \ge v^*$,

$$\begin{aligned} s(t) &= B^o(t) + \int_0^t \left(\frac{\phi(w_0 - t + x)\,\bar{F}(v^*)}{\bar{F}(w_0 - t + x)} \cdot 1_{\{v^* - w_0 \le t - x < v^*\}} + \lambda(t - x - v^*) \bar{F}(v^*) \cdot 1_{\{t - x \ge v^*\}} \right) \bar{G}(x) dx \\ &= \int_t^\infty \psi(x-t) \frac{\bar{G}(x)}{\bar{G}(x-t)} dx + \bar{F}(v^*) \left(\int_{t-v^*}^{t-(v^*-w_0)} \frac{\phi(w_0 - t + x)\,\bar{G}(x)}{\bar{F}(w_0 - t + x)} dx + \int_0^{t-v^*} \lambda(t - x - v^*) \bar{G}(x) dx \right). \end{aligned}$$

It is easy to see that this expression coincides with (E.1). The proof of other performance measures is similar.

F More on the Flows

We now elaborate on the discussion about the flows in §4; i.e., we discuss the departure process S in (2.8) and the abandonment process A in (2.6). Make the same assumptions as in §4, including the conditions in Theorem C.2 and Assumption 4.1.

Theorem F.1 (departure rate)

(i) For $t \ge 0$,

$$\sigma(t) = \int_0^\infty b(t,x) h_G(x)\,dx = \int_0^t b(t-x,0) g(x)\,dx + \int_0^\infty \frac{b(0,y) g(t+y)}{\bar{G}(y)}\,dy, \quad \text{(F.1)}$$

where $b(t,0) = \lambda(t-u)$ in an underloaded interval, but is the solution to the fixed point equation in Theorem C.1 during an overloaded interval.

(ii) $\sigma \in \mathbb{C}_p$, as assumed for λ in Assumption 2.1.

(iii) $\sigma(t) \ge B(t) h_{G,L} > 0$ for all $t \ge 0$, so that σ satisfies the requirement for λ in Assumption 4.2 of [6] over the interval $[\epsilon, t]$ for each $\epsilon > 0$.

(iv) If there exists a constant $h_{G,U}$ such that $h_G(x) \le h_{G,U} < \infty$ for all $x \ge 0$, then $\sigma(t) \le B(t) h_{G,U} \le s(t) h_{G,U}$ for all $t \ge 0$.

(v) If $b(t,0)$ is absolutely continuous with derivative $b'(u,0)$ in $\mathbb{C}_p$ on the interval $[0,t]$ (as occurs in the case of exponential service) and if

$$\tau_2(b,g,t) \equiv \sup_{0 \le s \le t} \int_0^\infty \frac{b(0,y)|g'(s+y)|}{\bar{G}(y)}\,dy < \infty, \quad \text{(F.2)}$$

then σ is absolutely continuous with derivative (a.e.)

$$\sigma'(t) = b(0,0) g(t) + \int_0^t b'(u,0) g(x)\,dx + \int_0^\infty \frac{b(0,y) g'(s+y)}{\bar{G}(y)}\,dy. \quad \text{(F.3)}$$

Proof. We prove the properties in turn:

(i) (representation (F.1)) Apply (2.8) and Assumption 2.2.

(ii)($\sigma \in \mathbb{C}_p$) By the finiteness of the initial conditions, Assumption 3.1 and the continuity of $b(\cdot, 0)$ from Theorem C.2, $\sigma(t) < \infty$. By Theorem C.2, $b(\cdot, 0)$ is in $\mathbb{C}_p$. By the Lebesgue dominated convergence theorem, the continuity of $b(t, 0)$ and $g(t + y)$ in the integrands of (F.1) is inherited by σ, so $\sigma \in \mathbb{C}_p$, as claimed.

(iii) (lower bound) By the initial relation in (F.1), we have $\sigma(t) \geq B(t)h_{G,L}$. Since $s(u) \geq s_L > 0$ for $0 \leq u \leq t$, $\lambda(t) \geq \lambda_{inf}(t) > 0$ and $\bar{G}(x) > 0$ for all x, we have $B(t) \geq t\lambda_{inf}(t)\bar{G}(t) \wedge s_L$ for all $t \geq 0$, which implies that there exist constants $\epsilon > 0$ and $\sigma_{\{inf,\eta,\epsilon\}}$ such that $\sigma(u) > \sigma_{\{inf,\eta,\epsilon\}} > 0$ for $0 < \epsilon \leq u \leq t$.

(iv) (upper bound) By the initial relation in (F.1), we have $\sigma(t) \leq B(t)h_{G,U}$, but we always have $B(t) \leq s(t)$.

(v)(derivative) We differentiate under the integral in (F.1) using Leibniz integral formula for differentiation under the integral, for which we require the finiteness of τ_2 in (F.2). ∎

The abandonment rate is somewhat more difficult. First, the abandonment is only positive during the overloaded intervals, so we assume that we are focusing on a single overloaded interval. Second, the abandonment depends on q, which in turn depends on w, which also is more complicated, requiring more conditions.

Theorem F.2 (abandonment rate) *Assume that the conditions in Theorem 4.1 of* [6] *hold, so that the BWT w is well defined and continuous.*

(i) For $t \geq 0$,

$$\begin{aligned}\alpha(t) &= \left(\int_0^{w(t)} \lambda(t-x)f(x)\,dx\right) 1_{\{w(t)\leq t\}} \\ &\quad + \left(\int_0^t \lambda(t-x)f(x)\,dx + \int_0^{w(t)-t} \frac{q(0,y)f(t+y)}{\bar{F}(y)}\,dy\right) 1_{\{w(t)>t\}}. \end{aligned} \quad \text{(F.4)}$$

(ii) $\alpha \in \mathbb{C}_p$, as assumed for λ in Assumption 2.1.

(iii) If Assumption 4.1 *holds, then $\alpha(t) \geq Q(t)h_{G,L}$ for all $t \geq 0$.*

(iv) If there exists a constant $h_{G,U}$ such that $h_G(x) \leq h_{G,U} < \infty$ for all $x \geq 0$, then $\sigma(t) \leq Q(t)h_{G,U}$, which is bounded over finite intervals, because Q is continuous.

(v) If $b(t, 0) > 0$ a.e., then α is absolutely continuous with derivative (a.e.)

$$\begin{aligned}\alpha'(t) &= \left(\lambda(t-w(t))f(w(t))w'(t) + \int_0^{w(t)} \lambda'(t-x)f(x)\,dx\right) 1_{\{w(t)\leq t\}} \\ &\quad + \left(\lambda(0)f(w(t)) + \int_0^t \lambda'(t-x)f(x)\,dx + \left(\frac{q(0,w(t)-t)f(w(t))}{\bar{F}(w(t)-t)}\right)(w'(t)-1)\right) \\ &\quad + \int_0^{w(t)-t} \frac{q(0,y)f'(s+y)}{\bar{F}(y)}\,dy\Bigg) 1_{\{w(t)>t\}}. \end{aligned} \quad \text{(F.5)}$$

Proof. We prove the properties in turn:

(i) (representation) Applying definition (2.6) and Assumption 2.2, we have

$$\alpha(t) = \int_0^\infty q(t,x)h_F(x)\,dx = \int_0^t q(t-x,0)f(x)\,dx + \int_0^\infty \frac{q(0,y)f(t+y)}{\bar{F}(y)}\,dy, \tag{F.6}$$

from which (F.4) follows.

(ii) ($\alpha \in \mathbb{C}_p$) Note that $\lambda, q(0,\cdot) \in \mathbb{C}_p$ by Assumption 2.1, $q(\cdot,0) \in \mathbb{C}_p$ by Corollary **??** and w is continuous by Theorem 4.1 of [6]. Hence, by the Lebesgue dominated convergence theorem, the continuity of $\lambda(t,0)$ and $f(t+y)$ as a function of t in the integrands of (F.1) is inherited by σ, so $\sigma \in \mathbb{C}_p$, as claimed.

(iii) (lower bound) By the initial relation in (F.4), we have $\alpha(t) \geq Q(t)h_{F,L}$.

(iv) (upper bound) By the initial relation in (F.4), we have $\alpha(t) \leq Q(t)h_{F,U}$.

(v) (derivative) We differentiate under the integral in (F.1) using Leibniz integral formula for differentiation under the integral. Since the integrands are bounded over the finite intervals, the integrals are finite. ∎