Queueing Systems manuscript No.
(will be inserted by the editor)

December 6, 2010

Large-Time Asymptotics for the $G_t/M_t/s_t + GI_t$ Many-Server Fluid Queue with Abandonment

Yunan Liu · Ward Whitt

Received: date / Accepted: date

Abstract We previously introduced and analyzed the $G_t/M_t/s_t+GI_t$ many-server fluid queue with time-varying parameters, intended as an approximation for the corresponding stochastic queueing model when there are many servers and the system experiences periods of overload. In this paper we establish an asymptotic loss of memory (ALOM) property for that fluid model; i.e., we show that there is asymptotic independence from the initial conditions as time t evolves, under regularity conditions. We show that the difference in the performance functions dissipates over time exponentially fast, again under the regularity conditions. We apply ALOM to show that the stationary $G/M/s+GI$ fluid queue converges to steady state and the periodic $G_t/M_t/s_t + GI_t$ fluid queue converges to a periodic steady state as time evolves, for all finite initial conditions.

Keywords nonstationary queues · queues with time-varying arrivals · many-server queues · deterministic fluid model · customer abandonment · loss of memory · weakly ergodic · periodic steady state · transient behavior

Yunan Liu
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA
Tel.: +212-854-7255
Fax: +212-854-8103
E-mail: yl2342@columbia.edu

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA
Tel.: +212-854-7255
Fax: +212-854-8103
E-mail: ww2040@columbia.edu

1 Introduction

We seek a better understanding of large-scale multi-server queueing systems that evolve with time-varying arrival rate, numbers of servers and other model parameters. We are especially interested in large scale queueing systems that experience periods of significant overloading, typically alternating with underloaded periods. Toward that end, in [8,9] we introduced deterministic fluid models with time-varying parameters to approximate the performance of these queueing systems. In [8] we considered the $G_t/GI/s_t + GI$ multi-server fluid model, having time-varying arrival rate and staffing (number of servers), customer abandonment (the $+GI$) and non-exponential service and patience distributions (the two GI's); in [9] we considered the $(G_t/M_t/s_t+GI_t)^m/M_t$ open network of many-server fluid queues, having time-varying Markovian routing (the $/M_t$) among m queues with time-varying customer abandonment from each queue (the $+GI_t$) and time-varying Markovian service. The results in [8,9] extend previous results for the Markovian time-varying $M_t/M_t/s_t + M_t$ model in [11–13] and the non-Markovian stationary $G/GI/s + GI$ model in [18].

In this paper we focus on the impact of the initial conditions on the system performance as time evolves. To treat the general nonstationary setting, we show that, under regularity conditions, an initial difference in the state variables dissipates over time, i.e., the large-time behavior is asymptotically independent of the initial conditions; we call this the *asymptotic loss of memory* (ALOM) property. For non-stationary Markov processes, ALOM has been called *weak ergodicity* [6], Ch. V. We also quantify the rate of convergence, showing that it is exponentially fast, again under regularity conditions.

This ALOM property can be quite useful. First, we apply ALOM to establish the existence of a unique steady state in *stationary* fluid models (that have constant model parameters), and convergence to that steady state as time evolves. Although the existence and form of this steady state were established in [18], the convergence from transient system dynamics to this steady state (and the rate of the convergence) has never been shown before, to the best of our knowledge.

We also employ ALOM to establish the existence of a unique *periodic steady state* (PSS) in *periodic* fluid models (that have periodic model parameters), and convergence to this PSS as time evolves. This PSS can be very useful to determine system congestion in service systems with daily or weakly cycles. We use the algorithm developed in [8,9] to compute performance functions over initial intervals. Since convergence is exponentially fast, that directly yields the PSS performance, but we also develop an alternative direct algorithm to compute the PSS performance. The rapid (exponential rate of) convergence established for ALOM also supports approximating the transient performance in stationary and periodic models with associated steady-state performance.

The specific fluid model we consider here is $G_t/M_t/s_t+GI_t$. That model is placed on a firm mathematical foundation in §2 of [9]; it is a relatively minor modification of the corresponding $G_t/GI/s_t + GI$ fluid model introduced and analyzed in [8]. The performance of the $G_t/M_t/s_t+GI_t$ model is characterized in §§3-5 of [9], building on §§4-9 of [8]. Regularity conditions were developed under which all the standard performance functions are characterized. Moreover, an algorithm was developed to compute these performance functions. We will draw heavily upon this previous material.

The special case of the $G_t/M/s_t+GI$ fluid queue, where only the arrival rate and staffing function (number of servers) are time-varying, should be adequate for most applications. The most useful generalization then would be to allow GI service instead of M service. With GI service, the fluid content density in service, $b(t,x)$ (see (7) and (8) below) during an overloaded interval depends on the prior values of the rate fluid enters service, $\{b(s,0) : 0 \leq s \leq t\}$, (see equation (15) of [8]), and Theorem 2 of [8] shows that $b(t,0)$ is characterized as the solution of a fixed point equation ((18) in [8]). Here we exploit the fact that, with M_t service, the density of fluid in service $b(t,x)$ can be exhibited explicitly. We *conjecture* that ALOM extends to $G_t/GI/s_t+GI$ models with non-exponential service times, provided that all the regularity conditions in [8] are satisfied, including the service-time distribution having a density.

In fact, in [10] we provide a counterexample showing that ALOM does *not* extend beyond M_t service to *all* GI service. Indeed, we show in [10] that ALOM does not hold even in all stationary fluid models. That is done by considering the $GI/D/s+GI$ fluid model with deterministic service times. Of course, the deterministic service-time distribution does not satisfy the density condition in [8,18]. Nevertheless, the $G/D/s+GI$ fluid queue has the stationary performance given in [18] and Theorem 4 here. However, the performance does not converge to that stationary value when the system starts empty. Instead, it approaches a PSS. The same phenomenon occurs for two-point service-time distributions when one point is 0, but otherwise we *conjecture* that ALOM extends to all many-server fluid queues in which service-time distributions are neither deterministic nor exponential.

As in [2,11–13], the fluid models can be related to the queueing models they approximate via many-server heavy-traffic limits, but as in [8,9], we do not discuss such limits here. As in [18], we obtain important Markovian structure by considering two-parameter processes, such as $Q(t,y)$, recording the queue content at time t that has been there for a duration y; see (7) below. (For related many-server heavy-traffic limits, see [7,15].) Our use of deterministic fluid models to capture the first-order behavior of queueing systems is part of an established tradition [4,14].

The rest of the paper is organized as follows: In §2 we review the definition and performance formulas of the $G_t/M_t/s_t+GI_t$ fluid queue. In §3 we review comparison and Lipschitz continuity results from [9] that we will apply, and we establish a new boundedness lemma, Lemma 1. In §4 we establish ALOM. In §5 we show that the transient performance of the stationary $G/M/s+GI$ fluid queue converges to its steady state performance. In §6 we establish the existence of a unique PSS and convergence to it in the periodic $G_t/M_t/s_t+GI_t$ queue. We draw conclusions in §7. Additional supporting material appears in an appendix, including comparisons with simulations of corresponding stochastic queueing systems.

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

In this section we review the established results for the $G_t/M_t/s_t + GI_t$ fluid queue from [8,9]; see those sources for more detail.

2.1 Model Definition

There is a service facility with finite capacity and an associated waiting room or queue with unlimited capacity. Fluid is a deterministic, divisible and incompressible quantity that arrives over time. Fluid input flows directly into the service facility if there is free capacity available; otherwise it flows into the queue. Fluid leaves the queue and enters service in a first-come first-served (FCFS) manner whenever service capacity becomes available. There cannot be simultaneously free service capacity and positive queue content.

The staffing function (service capacity) s is an absolutely continuous positive function with

$$s(t) \equiv \int_0^t s'(y)\, dy, \quad t \geq 0. \tag{1}$$

We assume that the service capacity is exogenously specified and that it provides a hard constraint: the amount of fluid in service at time t cannot exceed $s(t)$. In general, there is no guarantee that some fluid that has entered service will not be later forced to leave without completing service, because we allow s to decrease. We directly assume that phenomenon does not occur; i.e., we directly assume that the given staffing function is *feasible*. However, Theorem 6 of [9] shows how to construct a minimum feasible staffing function greater than or equal to an initial infeasible staffing function.

The total fluid input over an interval $[0, t]$ is $\Lambda(t)$, where Λ is an absolutely continuous function with

$$\Lambda(t) \equiv \int_0^t \lambda(y)\, dy, \quad t \geq 0, \tag{2}$$

where λ is the arrival-rate function. If the total fluid content in service at time t is $B(t)$, then the total service completion rate at time t is

$$\sigma(t) \equiv B(t)\mu(t), \quad t \geq 0. \tag{3}$$

Let $S(t)$ be the total amount of fluid to complete service in the interval $[0, t]$; then

$$S(t) \equiv \int_0^t \sigma(y)\, dy = \int_0^t B(y)\mu(y)\, dy, \quad t \geq 0. \tag{4}$$

Service and abandonment occur deterministically in proportions. Since the service is M_t, the proportion of fluid in service at time t that will still be in service at time $t+x$ is

$$\bar{G}_t(x) = e^{-M(t,t+x)}, \quad \text{where} \quad M(t,t+x) \equiv \int_t^{t+x} \mu(y)\,dy, \tag{5}$$

for $t \geq 0$ and $x \geq 0$. The time-varying service-time cdf of a quantum of fluid that enters service at time t is $G_t \equiv 1 - \bar{G}_t(x)$. The cdf G_t has density $g_t(x) = \mu(t+x)\bar{G}_t(x)$ and hazard rate $h_{G_t}(x) = \mu(t+x)$, $x \geq 0$.

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

$$F_t(x) = \int_0^x f_t(y)\,dy, \quad x \geq 0, \quad \text{and} \quad \bar{F}_t(x) \equiv 1 - F_t(x), \quad x \geq 0. \tag{6}$$

Let $h_{F_t}(y) \equiv f_t(y)/\bar{F}_t(y)$ be the hazard rate associated with the patience (abandonment) cdf F_t. We assume that $f_t(y)$ is jointly measurable in t and y, so the same will be true for $F_t(y)$ and $h_{F_t}(y)$.

System performance is described by a pair of two-parameter deterministic functions $(\hat{B}, \hat{Q})$, where $\hat{B}(t,y)$ $(\hat{Q}(t,y))$ is the total quantity of fluid in service (in queue) at time t that has been so for a *duration* at most y, for $t \geq 0$ and $y \geq 0$. These functions will be absolutely continuous in the second parameter, so that

$$\hat{B}(t,y) \equiv \int_0^y b(t,x)\,dx \quad \text{and} \quad \hat{Q}(t,y) \equiv \int_0^y q(t,x)\,dx, \tag{7}$$

for $t \geq 0$ and $y \geq 0$. Performance is primarily characterized through the pair of two-parameter fluid content densities (b, q). Let $B(t) \equiv \hat{B}(t, \infty)$ and $Q(t) \equiv \hat{Q}(t,\infty)$ be the total fluid content in service and in queue, respectively. Let $X(t) \equiv B(t) + Q(t)$ be the total fluid content in the system at time t. Since service is assumed to be M_t, the performance will primarily depend on b via B. (We will not directly discuss $\hat{B}$.)

Since fluid in service (queue) that is not served (does not abandon or enter service) remains in service (queue), we see that the fluid content densities b and q must satisfy the equations

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}_{t-x}(x+u)}{\bar{G}_{t-x}(x)} = b(t,x)e^{-M(t,t+u)}, \tag{8}$$

$$q(t+u, x+u) = q(t,x)\frac{\bar{F}_{t-x}(x+u)}{\bar{F}_{t-x}(x)}, \quad 0 \leq x+u < w(t), \tag{9}$$

for $t \geq 0$, $x \geq 0$ and $u \geq 0$, where M is defined in (5) and $w(t)$ is the *boundary waiting time* (BWT) at time t,

$$w(t) \equiv \inf\{x > 0 : q(t,y) = 0 \quad \text{for all} \quad y > x\}. \tag{10}$$

(By Assumptions 4 and 5 below, we are never dividing by 0 in (8) and (9).) Since the service discipline is FCFS, fluid leaves the queue to enter service from the right boundary of $q(t,x)$.

Let $A(t)$ be the total amount of fluid to abandon in the interval $[0,t]$ and Let $E(t)$ be the amount of fluid to enter service in $[0,t]$. Clearly, we have the *flow conservation equations*: For each $t \geq 0$,

$$Q(t) = Q(0) + \Lambda(t) - A(t) - E(t) \quad \text{and} \quad B(t) = B(0) + E(t) - S(t). \tag{11}$$

The abandonment satisfies

$$A(t) \equiv \int_0^t \alpha(y)\,dy, \quad \alpha(t) \equiv \int_0^\infty q(t,y) h_{F_{t-y}}(y)\,dy \tag{12}$$

for $t \geq 0$, where $\alpha(t)$ is the abandonment rate at time t and $h_{F_t}(y)$ is the hazard rate associated with the patience cdf F_t. (Recall that F_t is defined for t extending into the past.) The flow into service satisfies

$$E(t) \equiv \int_0^t b(u,0)\,du, \quad t \geq 0, \tag{13}$$

where $b(t,0)$ is the rate fluid enters service at time t. If the system is OL, then the fluid to enter service is determined by the *rate that service capacity becomes available* at time t,

$$\eta(t) \equiv s'(t) + \sigma(t) = s'(t) + B(t)\mu(t), \quad t \geq 0, \tag{14}$$

Then $\eta(t)$ coincides with the *maximum possible rate that fluid can enter service* at time t,

$$\gamma(t) \equiv s'(t) + s(t)\mu(t). \tag{15}$$

To describe waiting times, let the BWT $w(t)$ be the delay experienced by the quantum of fluid at the head of the queue at time t, already given in (10), and let the *potential waiting time* (PWT) $v(t)$ be the virtual delay of a quantum of fluid arriving at time t under the assumption that the quantum has infinite patience. A proper definition of q, w and v is somewhat complicated, because w depends on q, while q depends on w, but that has been done in §7 in [8].

We specify the initial conditions via the initial fluid densities $b(0,x)$ and $q(0,x)$, $x \geq 0$. Then $\hat{B}(0,y)$ and $\hat{Q}(0,y)$ are defined via (7), while $B(0) \equiv \hat{B}(0,\infty)$ and $Q(0) \equiv \hat{Q}(0,\infty)$, as before. Let $w(0)$ be defined in terms of $q(0,\cdot)$ as in (10). In summary, the six-tuple $(\lambda(t), s(t), \mu(t), F_t(x), b(0,x), q(0,x))$ of functions of the variables t and x specifies the *model data*. The system performance is characterized by the six-tuple $(b(t,x), q(t,x), w(t), v(t), \alpha(t), \sigma(t))$.

2.2 Assumptions on the Model Data

We directly assume that the initial values are finite:

Assumption 1 (*finite initial content*) *$B(0) < \infty$, $Q(0) < \infty$ and $w(0) < \infty$.*

As in [8,9], we consider a *smooth model*. Let $\mathbb{C}_p$ be the space of piecewise continuous real-valued functions of a real variable, by which we mean that there are only finitely many discontinuities in each finite interval, and that left and right limits exist at each discontinuity point, where the whole function is right continuous. Thus, $\mathbb{C}_p$ is a subset of $\mathbb{D}$, the right-continuous functions with left limits.

Assumption 2 (*smoothness*) *$s', \lambda, f_t, f_\cdot(x), \mu, b(0,\cdot), q(0,\cdot)$ in $\mathbb{C}_p$ for each $x \geq 0$ and t, $-\infty < t < \infty$.*

To treat the BWT w, we need to impose a regularity condition on the arrival rate function and the initial queue density, as in Assumption 10 of [8]. Here and later we use the notation ↑ and ↓ to denote supremum and infimum, respectively, e.g.,

$$\lambda_t^{\uparrow} \equiv \sup_{0 \leq u \leq t} \{\lambda(u)\} \quad \text{and} \quad \lambda_t^{\downarrow} \equiv \inf_{0 \leq u \leq t} \{\lambda(u)\}. \tag{16}$$

These apply in the obvious way, e.g., $q^{\downarrow}(0,x)$ below denotes the infimum over the second variable over $[0,x]$ and $\lambda_{\infty}^{\uparrow}$ denotes the supremum over the positive halfline.

Assumption 3 (*positive arrival rate and initial queue density*) *For all $t \geq 0$, $\lambda_t^{\downarrow} > 0$ and $q^{\downarrow}(0, w(0)) > 0$ if $w(0) > 0$.*

Appendix E of [8] illustrates the more complicated behavior that can occur for the BWT w when $\lambda_t^{\downarrow} = 0$.

To ensure that the PWT v is finite, we assume bounds on the minimum staffing level and the minimum service rate, as in Assumptions 7 and 8 of [9].

Assumption 4 (*minimum staffing and service rate*) *$s_{\infty}^{\downarrow} > 0$ and $\mu_{\infty}^{\downarrow} > 0$.*

To treat the time-varying abandonment cdf F_t, we introduce bounds for the time-varying pdf f_t and complementary cdf $\bar{F}_t$, as in [9]. Let

$$f^{\uparrow} \equiv \sup \{f_t(x) : x \geq 0, \quad -\infty < t < \infty\} \tag{17}$$

and

$$\bar{F}^{\downarrow}(x) \equiv \inf \{\bar{F}_t(x) : -\infty \leq t < \infty\}. \tag{18}$$

Assumption 5 (*controlling the time-varying abandonment*) *$f^{\uparrow} < \infty$, where $f^{\uparrow}$ is defined in (17), and $\bar{F}^{\downarrow}(x) > 0$ for all $x > 0$, where $\bar{F}^{\downarrow}(x)$ is defined in (18).*

We analyze the fluid queue under the assumptions above by considering alternating intervals over which the system is either *underloaded* (UL) or *overloaded* (OL), where these intervals include what is usually regarded as critically loaded. In particular, an interval starting at time 0 with (i) $Q(0) > 0$ or (ii) $Q(0) = 0$, $B(0) = s(0)$ and $\lambda(0) > s'0) + \sigma(0)$ is OL. The OL interval ends at the *OL termination time*

$$T \equiv \inf \{u \geq 0 : Q(u) = 0 \quad \text{and} \quad \lambda(u) \leq s'(u) + \sigma(u)\}. \tag{19}$$

Case (ii) in which $Q(0) = 0$ and $B(0) = s(0)$ is often regarded as critically loaded, but because the arrival rate $\lambda(0)$ exceeds the rate that new service capacity becomes available, $s'(0)+\sigma(0)$, we must have the right limit $Q(0+) > 0$, so that there exists $\epsilon > 0$ such that $Q(u) > 0$ for all $u \in (0, 0 + \epsilon)$. Hence, we necessarily have $T > 0$.

An interval starting at time 0 with (i) $Q(0) < 0$ or (ii) $Q(0) = 0$, $B(0) = s(0)$ and $\lambda(0) \leq s'(0) + \sigma(0)$ is UL. The UL interval ends at *UL termination time*

$$T \equiv \inf \{u \geq 0 : B(u) = s(u) \quad \text{and} \quad \lambda(u) > s'(u) + \sigma(u)\}. \tag{20}$$

As before, case (ii) in which $Q(0) = 0$ and $B(0) = s(0)$ is often regarded as critically loaded, but because the arrival rate $\lambda(0)$ does not exceed the rate that new service capacity becomes available, $\eta(0) \equiv s'(0)+\sigma(0)$, we must have the right limit $Q(0+) = 0$. The UL 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 UL subintervals. However, unlike an overloaded interval, we cannot conclude that we necessarily have $T > 0$ for a UL interval. Moreover, even if $T > 0$ for each UL interval, we could have infinitely many switches between OL intervals and UL intervals in a finite interval. Thus we make assumptions to ensure that those pathological situations do not occur.

As discussed in [8], for engineering applications it is reasonable to directly assume that there are only finitely many switches between OL and UL intervals in each finite time interval, but it is unappealing mathematically. In §3 of [9] we provided sufficient conditions based directly on the model parameters for there to be only finitely many switches between OL intervals and UL intervals in each finite time interval. In particular, we showed that it suffices to impose regularity conditions on the function $\zeta(t) \equiv \lambda(t) - s'(t) - s(t)\mu(t)$, $t \geq 0$. Let $Z_{\zeta,T}$ be the subset of zeros of the function ζ in $[0, T]$ and let $|A|$ be the cardinality of a set A. Theorem 2 of [9] shows that the number of switches between overloaded and underloaded intervals is finite in each finite interval if $|Z_{\zeta,T}| < \infty$ for each $T > 0$.

Assumption 6 (*controlling the number of switches*) *For all $T > 0$, $|Z_{\zeta,T}| < \infty$.*

In §3 of [9] we also showed that a sufficient condition for $|Z_{\zeta,T}| < \infty$ for each $T > 0$ is for the functions λ, s and μ to be piecewise polynomials (with finitely many discontinuities in each finite interval). Assumption 6 is also easy to verify in other settings, as we illustrate here with sinusoidal functions. *We assume that all assumptions in this section are in force throughout the paper.*

2.3 The Performance Formulas

In [8,9] we showed how the system performance expressed via the basic functions (b, q, w, v) depends on the model data $(\lambda, s, \mu, F, b(0,\cdot), q(0,\cdot)$. From the basic performance four-tuple (b, q, w, v), we easily compute the associated vector of performance functions $(\hat{B}, \hat{Q}, B, Q, X, \sigma, S, \alpha, A, E)$ via the definitions in §2.1. We quickly review the main results for the basic functions (b, q, w, v); see [8,9] for more details.

For the fluid model with unlimited service capacity ($s(t) \equiv \infty$ for all $t \geq 0$), starting at time 0,

$$\begin{aligned} b(t,x) &= e^{-M(t-x,t)}\lambda(t-x)1_{\{x\leq t\}} + e^{-M(0,t)}b(0,x-t)1_{\{x>t\}}, \qquad (21) \\ B(t) &= \int_0^t e^{-M(t-x,t)}\lambda(t-x)\,dx + B(0)e^{-M(0,t)}, \quad t \geq 0. \end{aligned}$$

where M is defined in (5). If, instead, a finite-capacity system starts UL, then the same formulas apply over the interval $[0,T)$, where $T \equiv \inf\{t \geq 0 : B(t) > s(t)\}$, with $T = \infty$ if the infimum is never obtained.

For the fluid model in an OL interval, $B(t) = s(t)$ and

$$\begin{aligned} b(t,x) = (s'(t-x) + s(t-x)\mu(t-x))e^{-M(t-x,t)}1_{\{x\leq t\}} \\ +b(0,x-t)e^{-M(0,t)}1_{\{x>t\}}. \qquad (22) \end{aligned}$$

Let $\tilde{q}(t,x)$ be $q(t,x)$ during an OL interval $[0,T]$ under the assumption that no fluid enters service from queue. During an OL interval,

$$\tilde{q}(t,x) = \lambda(t-x)\bar{F}_{t-x}(x)1_{\{x\leq t\}} + q(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\}}; \qquad (23)$$

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

We characterize the BWT w appearing in the formula for q above by equating the quantity of new fluid admitted into service in the interval $[t, t+\delta)$ to the amount of fluid removed from the right boundary of $q(t,x)$ that does not abandon in the same interval $[t, t+\delta)$. By careful analysis (Theorem 3 of [8]), that leads to the nonlinear first-order ODE

$$w'(t) = \Psi(t, w(t)) \equiv 1 - \frac{\gamma(t)}{\tilde{q}(t, w(t))}, \qquad (24)$$

for γ in (15), where $w'(t)$ denotes the derivative. (By Assumptions 3, 4 and 5, we are not dividing by 0 in (23) and (24). More detail on the structure of w is given in [8]. Overall, w is continuously differentiable everywhere except for finitely many t.) We compute the end of an OL interval by letting it be the first time t that $w(t) = 0$ and $\lambda(t) \leq s'(t) + s(t)\mu(t)$. During an OL interval, the PWT v is finite and is the unique function in $\mathbb{D}$ satisfying the equation

$$v(t - w(t)) = w(t) \quad \text{for all} \quad t \geq 0. \tag{25}$$

These results yield an efficient algorithm to compute the basic performance four tuple (b, q, w, v). First, for each UL interval, we compute b directly via (21), terminating the first time we obtain $B(t) > s(t)$. Second, for each OL interval, we compute b via (22), $\tilde{q}$ via (23) and then the BWT w by solving the ODE (24). We consider terminating the OL interval when $w(t) = 0$. We actually do terminate the OL interval if also $\lambda(t) \leq s'(t) + s(t)\mu(t)$. The proof of Theorem 5 in [8] provides an elementary algorithm to compute v during an OL interval from (25) once w has been computed. Theorem 6 of [8] shows that v satisfies its own ODE under additional regularity conditions.

3 Structural Results

In this section we present three structural results that we will apply here, two from [9] and one new. We first review the important comparison and Lipschitz continuity results established in Theorems 7 and 8 of [9].

Our comparison result establishes an ordering of the performance functions given an assumed ordering for the model data functions.

Theorem 1 (fundamental comparison theorem) *Consider two fluid models with common staffing function s and service rate function μ. If $\lambda_1 \leq \lambda_2$, $B_1(0) \leq B_2(0)$, $q_1(0, \cdot) \leq q_2(0, \cdot)$ and $h_{F_{t,1}} \geq h_{F_{t,2}}$, then*

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

Our Lipschitz continuity result also applies to functions. For it, we use the uniform norm on real-valued functions on the interval $[0, T]$: $\|x\|_T \equiv \sup\{|x(t)| : 0 \leq t \leq T\}$.

Theorem 2 (Lipschitz continuity) *The functions mapping (i) $(\lambda, B(0))$ in $\mathbb{C}_p \times \mathbb{R}$ into (B, σ) in $\mathbb{C}_p^2$, (ii) $(\lambda, B(0), Q(0))$ in $\mathbb{C}_p \times \mathbb{R}^2$ into Q in $\mathbb{C}_p$, and (iii) $(\lambda, X(0))$ in $\mathbb{C}_p \times \mathbb{R}$ into X in $\mathbb{C}_p$, all over $[0, T]$, are Lipschitz continuous. In particular,*

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

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

$$\|B_1 - B_2\|_T \le T\|\lambda_1 - \lambda_2\|_T, \quad \|Q_1 - Q_2\|_T \le T\|\lambda_1 - \lambda_2\|_T,$$
$$\|X_1 - X_2\|_T \le 2T\|\lambda_1 - \lambda_2\|_T.$$

We now add a new structural result: boundedness. For this elementary boundedness result and other results to follow, we make a stronger assumption on the staffing and the rates in the model data, requiring that they be uniformly bounded above and below. Our conditions will involve the maximum rate fluid can enter service: γ in (15) as well as the two-parameter abandonment hazard rate $h_{F_t}(y) \equiv f_t(y)/\bar{F}_t(y)$, defined after (6). Let

$$h^{\uparrow}_{F_T} \equiv \sup_{-\infty<t\le T, x\ge 0} h_{F_t}(x), \quad h^{\downarrow}_{F_T} \equiv \inf_{-\infty<t\le T, x\ge 0} h_{F_t}(x),$$
$$\bar{F}^{\uparrow}(x) \equiv \sup_{-\infty<t<\infty} \bar{F}_t(x), \quad \bar{F}^{\downarrow}(x) \equiv \inf_{-\infty<t<\infty} \bar{F}_t(x).$$

Assumption 7 (*uniformly bounded staffing and rates*) *The staffing and the rates in the model data are uniformly bounded above and below, i.e.,*

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

Assumption 7 repeats Assumption 4 and strengthens Assumptions 3 and 5.

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

Assumption 8 (*abandonment cdf tail*) $\bar{F}^{\uparrow}(x) \to 0$ *as* $x \to \infty$.

We assume that these two additional assumptions are in force for the remainder of the paper. Our boundedness result also exploits the finite initial conditions, provided by Assumption 1.

Lemma 1 (*boundedness*) *Under the assumptions above, all performance functions are uniformly bounded. In particular,*

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

Proof Most are elementary; only $Q(t)$ and $w(t)$ require detailed argument. Flow conservation in (11) implies that $Q'(t) = \lambda(t) - \alpha(t) - \gamma(t) \le \lambda^{\uparrow}_{\infty} - \alpha(t)$. Since $\alpha(t) \ge h^{\downarrow}_{F_\infty} Q(t)$, we have $Q'(t) < 0$ whenever $Q(t) > \lambda^{\uparrow}_{\infty}/h^{\downarrow}_{F_\infty}$. The bound for $w(t)$ follows directly from (30) and the final part of the proof of Theorem 3 below, which does not use the present lemma. ■

4 Asymptotic loss of Memory (ALOM)

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

Example 1 (a sinusoidal $G_t/M/s+M$ example) Consider a $G_t/M/s+M$ fluid queue that has the sinusoidal arrival rate function

$$\lambda(t) = a + b \cdot \sin(c\,t), \tag{26}$$

with $a = c = 1$ and $b = 0.6$, exponential service distribution with rate $\mu = 1$, constant staffing function $s = 1$, and exponential abandonment time distribution with rate $\theta = 0.5$. Applying the algorithm in §8 of [8], we compute and compare the performance measures $w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$ with four different (ordered) initial conditions: the system is initially (i) empty with $Q(0) = B(0) = 0$ (the yellow solid lines), (ii) UL with $Q(0) = 0$, $B(0) = 0.5 < 1 = s$ (the dark dashed lines), (iii) OL with $Q(0) = 0.4$, $B(0) = 1 = s$ (the light-blue dashed lines) and (iv) OL with $Q(0) = 0.8$, $B(0) = 1 = s$ (the red dotted lines), as shown in Figure 1.

Figure 1 shows that the differences in these four cases converge to zero so fast that it looks as if the distance becomes 0 after finite time (but that actually never occurs), even though the initial conditions are dramatically different. Figure 1 also illustrates the comparison result in Theorem 1. ■

To state our ALOM result, we use Δ to denote absolute difference. Specifically, for real-valued functions X_i on $[0, \infty)$, $i = 1, 2$, and $0 < T \leq \infty$, let $\Delta X_{1,2}(t) \equiv \Delta X(t) \equiv |X_1(t) - X_2(t)|$, $t \geq 0$.

Theorem 3 (asymptotic loss of memory) *Consider two $G_t/M_t/s_t+GI_t$ fluid models with common arrival rate function λ, service rate function μ, staffing function s, and time-varying abandon-time cdf's F_t, but different initial conditions (satisfying Assumption 1). Then (a)*

$$\Delta X(T) \leq C_1 e^{-C(T)} \quad \textit{for} \quad C(T) \equiv T\,(\mu_T^{\downarrow} \wedge h_{F_T}^{\downarrow}), \tag{27}$$

where $C_1 \equiv C_1(B_1(0), B_2(0), q_1(0, \cdot), q_2(0, \cdot))$ is the constant

$$\begin{aligned} C_1 &\equiv \Delta B(0) + \int_0^\infty ([q_1(0,x) \vee q_2(0,x)] - [q_1(0,x) \wedge q_2(0,x)])\,dx \\ &\leq \Delta B(0) + Q_1(0) + Q_2(0). \end{aligned} \tag{28}$$

Moreover,

$$\Delta\alpha(T) \leq h_{F_T}^{\uparrow} C_1 e^{-C(T)} \quad \textit{and} \quad \Delta\sigma(T) \leq \mu_T^{\uparrow} C_1 e^{-C(T)} \tag{29}$$

for all $T > 0$. Hence, for $C_2 \equiv \mu_\infty^{\downarrow} \wedge h_{F_\infty}^{\downarrow} > 0$ and all $T > 0$,

$$\Delta X(T) \leq C_1 e^{-C_2 T}, \quad \Delta\alpha(T) \leq h_{F_\infty}^{\uparrow} C_1 e^{-C_2 T} \quad \textit{and} \quad \Delta\sigma(T) \leq \mu_\infty^{\uparrow} C_1 e^{-C_2 T}.$$

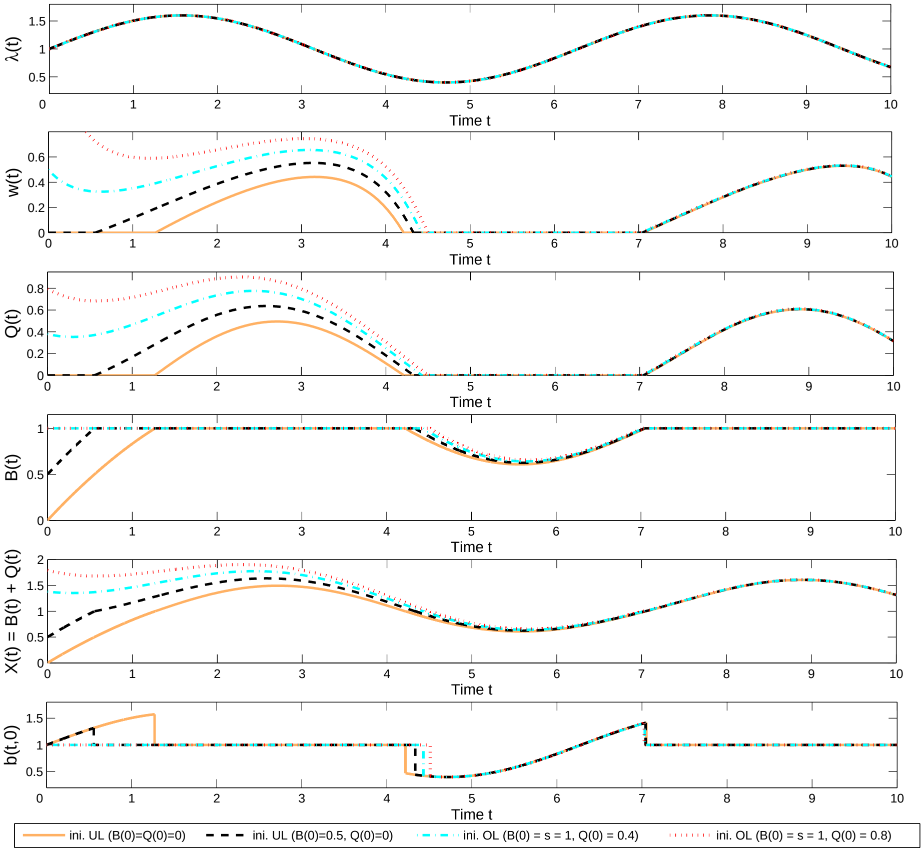

Fig. 1 The performance measures for the $G_t/M/s+M$ model in Example 1 with four different (ordered) initial conditions.

In addition, for each $T > 0$*,*

$$\begin{aligned} \Delta w(T) &\le \frac{\Delta X(T)}{\lambda_T^{\downarrow}\bar{F}^{\downarrow}(w_1(T)\vee w_2(T))} \\ &\le C_3\Delta X(T) \le (C_3C_1)e^{-C_2T}, \end{aligned} \qquad (30)$$

where

$$C_3 \equiv (\bar{F}^{\uparrow})^{-1}(s_\infty^{\downarrow}\mu_\infty^{\downarrow}/\lambda_\infty^{\uparrow}) \vee \left((w_1(0)\vee w_2(0)) + \frac{Q_1(0)+Q_2(0)}{s_\infty^{\downarrow}\mu_\infty^{\downarrow}}\right). \qquad (31)$$

(*b*) *If, in addition, the initial content is ordered by*

$$X_1(0) \le X_2(0) \quad \textit{and} \quad q_1(0,x) \le q_2(0,x) \quad \textit{for all} \quad x \ge 0, \qquad (32)$$

then $X_1(t) \le X_2(t)$ *for all* $t \ge 0$*,*

$$\Delta X'(T) \le 0 \quad \textit{and} \quad \Delta X(T) \le \frac{\Delta X(0)}{1+C(T)}, \quad T > 0, \qquad (33)$$

for $C(T)$ in (27)*, so that*

$$\begin{aligned} &\Delta X(T) \le e^{-C(T)}\Delta X(0), \\ &\Delta\alpha(T) \le h^{\uparrow}_{F_T}\Delta X(T) \quad \textit{and} \quad \Delta\sigma(T) \le \mu^{\uparrow}_{T}\Delta X(T). \end{aligned} \tag{34}$$

Proof We first show that (a) follows from (b). Without loss of generality, we have $X_1(0) \le X_2(0)$. Then $X_1(0) \le X_2(0)$ is equivalent to $B_1(0) \le B_2(0)$ and $Q_1(0) \le Q_2(0)$. In order to derive (a) from (b), construct another two systems, 3 and 4, with $q_3(0,x) \equiv q_1(0,x) \vee q_2(0,x)$, $B_3(0) \equiv B_1(0) \vee B_2(0)$, $q_4(0,x) \equiv q_1(0,x) \wedge q_2(0,x)$ and $B_4(0) \equiv B_2(0) \wedge B_2(0)$. With this construction, systems 3 and 4 are bonafide fluid models, with $X_4(t) \le X_1(t) \le X_3(t)$ and $X_4(t) \le X_2(t) \le X_3(t)$ for all t, which implies that $\Delta X_{1,2}(t) \le \Delta X_{3,4}(t)$ for all t. Since $\Delta X_{3,4}(0) \le C_1$ for C_1 in (28), (27) in (a) follows from (34) for $\Delta X_{3,4}(t)$. (The final bound on C_1 in (28) arises when the supports of $q_1(0,\cdot)$ and $q_2(0,\cdot)$ are disjoint sets, which actually is not allowed by Assumption 3, but can be approached.)

Now we prove (b). Observe that (34) follows (33) because dividing the interval $[0,T]$ into N subintervals yields

$$\Delta X(T) \le \left(\frac{1}{1+\frac{T}{N}\,(\mu^{\downarrow}_{T} \wedge h^{\downarrow}_{F_T})}\right)^{N} \Delta X(0).$$

Letting $N \to \infty$, we get (34).

We now prove (33). With the ordering assumed in (32), all functions in the two systems can be ordered according to Theorem 1. Hence, there are only three cases: (i) both systems are UL; (ii) both systems are OL; (iii) system 1 is UL and system 2 is OL. We treat the three cases separately and use mathematical induction to show (33).

In case (i) we have $B_1(0) \le B_2(0) \le s(0)$ and $Q_1(0) = Q_2(0) = 0$. Let T^* be the underload termination time of system 2. For $0 \le t < T^*$, neither system changes regime. Observe that $\Delta X(t) = \Delta B(t)$. Flow conservation implies that

$$B_i'(t) = \lambda(t) - \mu(t)\,B_i(t) \quad \text{for} \quad i = 1,2,$$

which yields

$$\Delta X'(s) = \Delta B'(s) = -\mu(s)\,\Delta B(s) \le -\mu^{\downarrow}_t\,\Delta B(t) = -\mu^{\downarrow}_t\,\Delta X(t), \qquad 0 \le s \le t,$$

where the inequality follows from $\mu(s) \ge \mu^{\downarrow}_t$ and $\Delta B(s) \ge \Delta B(t)$ since $\Delta B(s)$ has negative derivative. Therefore, we have

$$\Delta X(t) - \Delta X(0) \le -\mu^{\downarrow}_t\,t\,\Delta X(t)$$

and

$$\Delta X(t) \le \left(\frac{1}{1+\mu^{\downarrow}_t\,t}\right)\Delta X(0). \tag{35}$$

In case (ii) we have $B_1(0) = B_2(0) = s(0)$ and $q_1(0,\cdot) \le q_2(0,\cdot)$. Let T^* be the overload termination time of system 1. For $0 \le t < T^*$, neither system changes regime. Observe that $\Delta X(t) = \Delta Q(t)$. Theorem 1 implies that $q_1(t,\cdot) \le q_2(t,\cdot)$ and $w_1(t) \le w_2(t)$ for $\le t \le T^*$. Therefore, we have

$$\begin{aligned}
\alpha_2(t) - \alpha_1(t) \;&=\; \int_0^{w_2(t)} q_2(t,x)\, h_{F_{t-x}}(x)dx - \int_0^{w_1(t)} q_1(t,x)\, h_{F_{t-x}}(x)dx \\
&=\; \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_{F_{t-x}}(x)dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x)\, h_{F_{t-x}}(x)dx \\
&\ge\; h^{\downarrow}_{F_t} \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x))dx + h^{\downarrow}_{F_t} \int_{w_1(t)}^{w_2(t)} q_2(t,x)dx \\
&= h^{\downarrow}_{F_t}\,(Q_2(t) - Q_1(t)) = h^{\downarrow}_{F_t}\,\Delta Q(t).
\end{aligned} \tag{36}$$

Flow conservation implies that

$$Q_i'(t) = \lambda(t) - \alpha_i(t) - \gamma(t) \quad \text{for} \quad i = 1, 2,$$

which yields

$$\begin{aligned}
\Delta X'(s) &= \Delta Q'(s) = -(\alpha_2(s) - \alpha_1(s)) \\
&\le -h^{\downarrow}_{F_t}\,\Delta Q(s) \le -h^{\downarrow}_{F_t}\,\Delta Q(t) = -h^{\downarrow}_{t}\,\Delta X(t), \qquad 0 \le s \le t,
\end{aligned}$$

where the inequality follows from (36). Hence, reasoning as for (35) in case (i), we have

$$\Delta X(t) \le \left(\frac{1}{1 + h^{\downarrow}_{F_t}\, t} \right) \Delta X(0). \tag{37}$$

In case (iii) we have $B_1(0) \le s(0) = B_2(0)$ and $Q_1(0) = 0 \le Q_2(0)$. Let $T^* \equiv T_1 \wedge T_2$ where T_1 is the underload termination time of system 1 and T_2 is the overload termination time of system 2. For $0 \le t < T^*$, neither system changes regime. Observe that $\Delta X(t) = \Delta B(t) + \Delta Q(t) = s(t) - B_1(t) + Q_2(t)$. Flow conservation in (11) implies that the derivatives satisfy

$$\begin{aligned}
Q_2'(t) &= \lambda(t) - \alpha_2(t) - \gamma(t) \\
s'(t) &= \gamma(t) - \mu(t)\, s(t) \\
B_1'(t) &= \lambda(t) - \mu(t)\, B_1(t),
\end{aligned}$$

which implies that

$$\begin{aligned}
\Delta X'(t) &= s'(t) - B_1'(t) + Q_2'(t) \\
&= -\alpha_2(t) - \mu(t)\,(s(t) - B_1(t)).
\end{aligned} \tag{38}$$

Reasoning as in case (ii), we have

$$\alpha_2(t) \ge h^{\downarrow}_{F_t}\, Q_2(t) = h^{\downarrow}_{F_t}\, \Delta Q(t). \tag{39}$$

Therefore, (38) and (39) imply that

$$\begin{aligned}\Delta X'(s) &\le -h^{\downarrow}_{F_t}\, \Delta Q(s) - \mu^{\downarrow}_t \Delta B(s) \\ &\le -(h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)(\Delta Q(s) + \Delta B(s)) \\ &\le -(h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)\Delta X(s) \le -(h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)\Delta X(t), \qquad 0 < s \le t.\end{aligned}$$

Hence, reasoning as for (35) in case (i), we have

$$\Delta X(t) \le \left(\frac{1}{1 + (h^{\downarrow}_{F_t} \wedge \mu^{\downarrow}_t)\, t}\right) \Delta X(0). \tag{40}$$

Finally, combining (35), (37) and (40), the desired (33) follows by mathematical induction.

We directly have the second and third inequalities in (34), which implies (29) because $\Delta Q(T) \le \Delta X(T)$ and $\Delta B(T) \le \Delta X(T)$.

Finally, we treat $w(t)$. As above, it suffices to assume that we have the ordering in (32) of (b). Then (30) follows from

$$\begin{aligned}\Delta X(T) \ge \Delta Q(T) &= \int_{w_1(T)}^{w_2(T)} \lambda(T - x)\, \bar{F}_{T-x}(x) dx \\ &\ge \lambda^{\downarrow}_T \bar{F}^{\downarrow}(w_2(T))\, \Delta w(T).\end{aligned} \tag{41}$$

We now construct w^* such that $w_2(T) \le w^*$ for all T; in general, w^* will depend on $w_2(0)$. First note that at time $T_w \equiv Q_2(0)/\mu^{\downarrow}_{\infty} s^{\downarrow}_{\infty}$, all fluid that was in queue 2 at time 0 is gone (entered service or abandoned). Choose $\bar{w} > 0$ big enough such that $\bar{F}^{\uparrow}(\bar{w}) < s^{\downarrow}_{\infty}\mu^{\downarrow}_{\infty}/\lambda^{\uparrow}_{\infty}$. ODE (24) implies that for $t > T_w$,

$$\begin{aligned}w_2'(t) &= 1 - \frac{s(t)\, \mu(t)}{\lambda(t - w_2(t))\, \bar{F}_{t-w_2(t)}(w_2(t))} \\ &\le 1 - \frac{s^{\downarrow}_{\infty}\, \mu^{\downarrow}_{\infty}}{\lambda^{\uparrow}_{\infty} \bar{F}^{\uparrow}(\bar{w})} < 0,\end{aligned}$$

if $w_2(t) > \bar{w}$ for some t. Hence $\bar{w}$ is an upper bound for $w_2(t)$ if $w_2(T_w) < \bar{w}$. If $w_2(T_w) \ge \bar{w}$, it is easy to see that $w_2(t)$ decreases until it is below $\bar{w}$ because we can bound $w_2'(t)$. This argument implies that $w_2(t) \le w_2^* \equiv (\bar{w} \vee (w_2(0) + T_w))$ for all $t \ge 0$. The constant C_3 in (30) is obtained by inserting established bounds. ∎

For a real-valued function x on $[0, \infty)$, let $\|x\|_1 \equiv \int_0^{\infty} |x(t)|\, dt$.

Corollary 1 *Under the conditions of Theorem* 3 *(b),*

$$\begin{aligned}\|b_1(T, \cdot) - b_2(T, \cdot)\|_1 = \Delta B(T) \le \Delta X(T) \le \Delta X(0) e^{-C(T)}, \\ \|q_1(T, \cdot) - q_2(T, \cdot)\|_1 = \Delta Q(T) \le \Delta X(T) \le \Delta X(0) e^{-C(T)}.\end{aligned} \tag{42}$$

Hence, there is exponential rate of convergence under the conditions in (a).

Remark 1 (monotonicity of the difference of two queues) Theorem 3 shows that except for the densities q and b, the differences of all performance measures ($\varDelta X$, $\varDelta\alpha$, $\varDelta\sigma$, and $\varDelta w$) of the two queues go to 0 as $t \to \infty$. However, even in case (b), only $\varDelta X(t)$ goes to 0 monotonically. Note that $\varDelta\alpha(t) = 0$, $\varDelta w(t) = 0$ and $\varDelta\sigma(t) \geq 0$ when both queues are UL; $\varDelta\alpha(t) \geq 0$, $\varDelta w(t) \geq 0$ and $\varDelta\sigma(t) = 0$ when both queues are OL.

Remark 2 (Example 1 revisited) In Example 1 we have $C(T) = \mu \wedge \theta = 0.5$ in (27) of Theorem 3, $\lambda_\infty^\downarrow = 0.4 > 0$, $\lambda_\infty^\uparrow = 1.6 < \infty$, $\bar{F}^\downarrow(x) = e^{-\theta x} > 0$ and $\bar{F}^\uparrow(x) \to 0$ as $x \to \infty$. Moreover, $\zeta(t) = \lambda(t) - \mu\, s(t) - s'(t) = a - \mu\, s + b \cdot \sin(c\, t)$ is sinusoidal so that it has finitely many zeros in any bounded interval. Therefore, all conditions in Theorem 3 are satisfied, establishing the exponential rate of convergence seen in Figure 1.

5 The Stationary $G/M/s+GI$ Fluid Queue

In this section we focus on the stationary $G/M/s+GI$ fluid queue. The steady-state performance of the more general $GI/GI/s+GI$ fluid queue with GI service was characterized in [18], but the transient dynamics was only characterized completely in [8]. We first review the steady-state performance with GI service.

Theorem 4 (*steady state of the $G/GI/s+GI$ fluid queue, from* [18]) *The $G/GI/s+GI$ fluid model specified with model parameter (λ, s, μ, G, F) has a steady-state performance described by the vector $(b, q, B, Q, w, \sigma, \alpha)$, whose character depends on whether $\rho \equiv \lambda/s\mu \leq 1$ or $\rho > 1$.*
(a) UL and balanced cases: $\rho \leq 1$. If $\rho \leq 1$, then for $x \geq 0$

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

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

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

Complementing the proof of Theorem 4 in [18], we can apply [8] to give an alternative proof to show that the steady state given in Theorem 4 is indeed an invariant state, i.e., if the system is initially in this state, then it stays there forever.

Proof First consider (a) with $\rho \leq 1$. By (9) of [8], the initial rate that service is being completed with $b(0, x) = \lambda\bar{G}(x)$ is

$$\sigma(0) = \int_0^\infty b(0, x)h_G(x)\, dx = \int_0^\infty \lambda\bar{G}(x)\frac{g(x)}{\bar{G}(x)}\, dx = \lambda. \tag{43}$$

If $\rho < 1$, then $B(0) = s\rho < s$ and there initially is spare capacity. If $\rho = 1$, then $\lambda(0) = \lambda = \sigma$. In both cases, the system remains UL. Hence we can apply (13) in Proposition 2 of [8] to characterize the evolution of b. For suitably small $t > 0$, we get

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

which implies that the system stays UL with $b(t,x) = b(0,x)$, $B(t) = B(0)$ and $\sigma(t) = \sigma(0)$ for $t \ge 0$. For an alternative proof under the extra condition of differentiability, we can exploit the transport partial differential equation (PDE) from Appendix B of [8]. That tells us that $b(t,x)$ satisfies the PDE

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

which implies that

$$\begin{aligned} \frac{\partial b}{\partial t}(0,x) &= -\frac{\partial b}{\partial x}(0,x) - h_G(x)\, b(0,x) = -\frac{d(\lambda\, \bar{G}(x))}{dx} - h_G(x)\lambda\bar{G}(x) \\ &= \lambda\, g(x) - h_G(x)\bar{G}(x)\lambda = 0. \end{aligned}$$

Next consider case (b) with $\rho > 1$. We can apply (43) to see that the initial rate of service completion, starting with $b(0,x) = s\mu\bar{G}(x)$, is $\sigma(0) = s\mu$. Since $\rho > 1$, we necessarily have $\lambda(0) = \lambda > s\mu = \sigma(0)$. Hence, the system necessarily remains OL over a positive interval. Next we apply the fixed point equation for b during an overloaded interval. Assumption 8 in [8] is satisfied with this initial density $b(0,x)$ because

$$\tau(b,g,T) \equiv \sup_{0\le s\le T} \int_0^\infty \frac{b(0,y)g(s+y)}{\bar{G}(y)}\, dy = s\mu < \infty. \tag{44}$$

Next we observe that $b(0,x)$ satisfies the fixed point equation (18) of [8], i.e.,

$$b(t,0) = \hat{a}(t) + \int_0^t b(t-x,0)g(x)\, dx = s\mu\bar{G}(t) + \int_0^t b(t-x,0)g(x)\, dx, \tag{45}$$

yielding $s\mu = s\mu\bar{G}(t) + s\mu G(t) = s\mu$. Theorem 2 of [8] implies that $b(t,0) = s\mu$, $t \ge 0$, is the unique fixed point. Next Proposition 6 of [8] implies that the service density in queue satisfies

$$\begin{aligned} q(t,x) &= \lambda\bar{F}(x)1_{\{x\le t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\le w(t)\}} \\ &= \lambda\bar{F}(x)1_{\{0\le x\le w(t)\}}. \end{aligned} \tag{46}$$

It remains to show that $w'(0) = 0$, so that $w(t) = w(0) = F^{-1}(1 - (1/\rho))$. However, ODE (24) implies that

$$w'(0) = 1 - \frac{\gamma(0)}{q(0, w(0))} = 1 - \frac{\mu\, s}{\lambda\, \bar{F}(w(0))} = 1 - \frac{\mu\, s}{\lambda(1/\rho)} = 0,$$

where the third equality holds since $w(0) = w = F^{-1}(1 - 1/\rho)$. The last equality holds since $\rho = \lambda/s\mu$. Hence, $w(t) = w$ in (46), so that $q(t, x) = q(x)$ and all performance functions are constants for $0 \le t \le \delta$ for some small δ and thus for all $t \ge 0$. ∎

Now we apply Theorem 3 to show that the transient performance in the $G/M/s+GI$ fluid queue with exponential service converges to the steady state described in Theorem 4 for any given initial conditions. As a byproduct, this establishes uniqueness for the steady-state performance in Theorem 4 in the special case of M service. We give two convergence results, the first obtained by directly combining Theorems 3 and 4.

Theorem 5 (*direct implication of ALOM*) *For the stationary $G/M/s + GI$ fluid model, as $t \to \infty$,*

$$(\alpha(t), w(t), Q(t), \sigma(t), B(t)) \;\to\; (\alpha, w, Q, \sigma, B), \tag{47}$$

$$\|q(t, \cdot) - q(\cdot)\|_1 \to 0 \quad \textit{and} \quad \|b(t, \cdot) - b(\cdot)\|_1 \to 0, \tag{48}$$

where vector $(q(\cdot), \alpha, w, Q, b(\cdot), \sigma, B)$ is the steady-state performance in Theorem 4. Hence, the steady-state performance specified by Theorem 4 is unique.

Proof Consider two $G/M/s + GI$ fluid queues that have identical model parameters but different initial conditions. Let system 1 be initially in the steady state given in Theorem 4, let system 2 have arbitrary initial condition. Theorem 4 implies that system 1 stays in steady state for all $t \ge 0$. Therefore, the convergence in (47) and (48) follows from ALOM in Theorem 3.

We next establish a stronger convergence result, whose proof does not rely on the ALOM property in Theorem 3. We establish pointwise convergence of the fluid content densities b and q as $t \to \infty$ in addition to (47) and (48).

Theorem 6 (*more on convergence to steady state*) *Consider the stationary $G/M/s+GI$ fluid model. In addition to Assumption 1, assume that the initial service density satisfies*

$$\limsup_{x\to\infty} b(0, x) < \infty. \tag{49}$$

Then, in addition to the conclusions of Theorem 5*,*

$$(q(t, x), b(t, x)) \to (q(x), b(x)) \quad \textit{as} \quad t \to \infty,$$

for each $x \ge 0$, where the limit $(q(x), b(x))$ is the pair of steady-state fluid densities in Theorem 4. Moreover, there is at most one switch between the OL and UL (*including critically loaded*) *regimes during the convergence. More* *precisely, the number of switches depends on the the model parameter* $\rho \equiv \lambda/s\mu$ *and the initial conditions as shown in Table 1. If* $\rho > 1$*, there exists a* $T > 0$ *such that for* $t > T$*,* $w(t) \to w$ *monotonically, as* $t \to \infty$*. If, in addition,* $C \equiv f^{\downarrow}_{(Q(0)/s\mu)\vee w} > 0$ *where* $f^{\downarrow}_t \equiv \inf_{0 \le x \le t} f(x)$*, then*

$$\varDelta w(t) \equiv |w(t) - w| \le \frac{1}{1 + (t - T)C}\, \varDelta w(T), \quad \textit{for } t > T \tag{50}$$

so that

$$\varDelta w(t) \le e^{-(t-T)C} \varDelta w(T), \quad t > T. \tag{51}$$

traffic intensity	initial condition	number of switchings
$\rho > 1$	OL	0
	UL(CL)	1
$\rho < 1$	OL	1
	UL(CL)	0
$\rho = 1$	OL	0
	UL(CL)	0

Table 1 How the number of switches between OL and UL intervals depends on the model parameter ρ and the initial conditions, in the setting of Theorem 6.

Proof We only give the proof for the case in which the system is initially UL, i.e., $q(0, x) = w(0) = 0$ for any x and $B(0) = \int_0^\infty b(0, x)dx < s$. The other case in which the system is initially OL or critically loaded is treated in essentially the same way; the details are given in the appendix. For simplicity, we assume $\mu = s = 1$ and therefore $\rho = \lambda/s\mu = \lambda$.

(i) $\rho \le 1$. Since the service is exponential at the fixed rate $\mu = 1$ and the staffing is fixed at $s = 1$, the maximum output rate of the service facility is 1. Hence, the system always stay in the UL regime. Thus we can apply (21) to characterize the density in service. By Assumption (49),

$$\begin{aligned} b(t, x) &= \rho e^{-x} 1_{\{0 \le x \le t\}} + b(0, x - t)e^{-t} 1_{\{x > t\}} \\ &\to \rho e^{-x} \quad \textit{as } t \to \infty, \quad x \ge 0. \\ B(t) &= \int_0^t \rho e^{-x} dx + \int_t^\infty b(0, x - t)e^{-t} dx \\ &= \rho(1 - e^{-t}) + e^{-t} B(0), \\ &= \rho - (\rho - B(0))\, e^{-t} \to \rho, \quad \textit{as } t \to \infty, \end{aligned}$$

Moreover, $\sigma(t) = B(t) \to \rho$, as $t \to \infty$. If $\rho = 1$, then we obtain the monotone convergence

$$B(t) = 1 - (1 - B(0))\, e^{-t} \uparrow 1 \quad \text{as} \quad t \to \infty.$$

(ii) $\rho > 1$. As in case (i), the maximum output rate of the service facility is 1. Since $\rho > 1$, $\lambda > 1$, so that the the system necessarily will switch to the OL regime in finite time. From (21), we see the $b(t,x)$ and $B(t)$ initially evolve as

$$\begin{aligned} b(t,x) &= \rho e^{-x} 1_{\{x \le t\}} + e^{-t} b(0, x-t) 1_{\{x>t\}} \\ B(t) &= \rho - (\rho - B(0)) e^{-t}, \quad 0 \le t \le t_1. \end{aligned} \tag{52}$$

The total fluid content in service $B(t)$ increases in t until time t_1 at which we first have $B(t) = B(t_1) = 1$. After time t_1, since the arrival rate ρ is greater than the maximum departure rate which is 1, the system stays in the OL regime. After time t_1, we can apply (22) to describe the evolution of $b(t,x)$. In particular, for $t > t_1$ and for each $x \ge 0$,

$$b(t - t_1, x) = e^{-x} 1_{\{x \le t - t_1\}} + b(t_1, x - t + t_1) e^{-(t - t_1)} 1\{x > t - t_1\}, \tag{53}$$

where

$$b(t_1, x) = \rho e^{-x} 1_{\{x \le t_1\}} + e^{-t_1} b(0, x - t_1) 1_{\{x > t_1\}}, \tag{54}$$

so that, by assumption (49), the second term in (53) is asymptotically negligible as $t \to \infty$, implying that $b(t,x) \to e^{-x} = b(x)$ as $t \to \infty$.

Since we start UL, we first have a queue buildup at time t_1. By (23), we have

$$q(t,x) = \rho \bar{F}(x) 1_{\{x \le w(t) \wedge (t - t_1)\}}, \quad t > t_1, \tag{55}$$

where the BWT w satisfies the ODE

$$w'(t) = 1 - \frac{1}{\rho \bar{F}(w(t))} \equiv H(w(t)), \quad for\ t \ge t_1, \tag{56}$$

with initial condition $w(t_1) = 0$. It is easy to see that $q(t,x) \to q(x) = \rho \bar{F}(x) 1_{\{x \le w(t)\}}$ if $w(t) \to w$ as $t \to \infty$.

Let $w \equiv F^{-1}(1 - 1/\rho)$. Since the cdf F has a positive density, the function H is strictly decreasing and $H(w) = 0$. Therefore if $w(t_2) = w$ at some t_2, $w(t)$ will stay at w for all $t \ge t_2$, since $w'(t_2) = H(w) = 0$. Moreover, if $w(t) < w$, then $w'(t) = H(w(t)) > H(w) = 0$.

The function $w(t)$ starts at 0 at time t_1, and is increasing (has positive derivative) as long as $w(t) < w$. We also know that $w(t)$ will stay at w if it hits w, and $w(t)$ is continuous. Therefore, to show that $w(t) \to w$ as $t \to \infty$, it remains to show that for any $\epsilon > 0$, there exits a t_ϵ such that $w(t) > w - \epsilon$ for any $t > t_\epsilon$.

Because H is strictly decreasing in a neighborhood of w, we have $w'(t) = H(w(t)) \ge H(w - \epsilon) \equiv \delta(\epsilon) > H(w) = 0$, if $w(t) \le w - \epsilon$. Therefore, the derivative of $w(t)$ is not only positive, but also bounded by $\delta(\epsilon) > 0$. So $w(t)$ will hit $w - \epsilon$ at least linearly fast with slope $\delta(\epsilon)$, i.e., for any $t \ge (w - \epsilon)/\delta(\epsilon)$, we have $w(t) \ge w - \epsilon$. Therefore, we conclude that $w(t) \uparrow w$ as $t \uparrow \infty$. As a consequence, we get $q(t,x) \to q(x) = \rho \bar{F}(x) 1_{\{0 \le x \le w\}}$ as $t \to \infty$ from (55).

We now establish (50) and (51). To do so, we assume the system is initially OL with $w(0) = w_0$. From the above analysis, if $\rho > 1$, then the system stays OL for all $t \ge 0$, which implies that $\gamma(t) = \mu s = 1$ for all $t \ge 0$.

Hence, after $T \equiv Q(0)/\mu s = Q(0)$, all fluid that was in queue at $t = 0$ is gone (has entered service or abandoned). If $w(T) = w$, then the system is already in equilibrium. If $w(T) > w$ (the case $w(T) < w$ is similar), then the above analysis implies that $w'(t) \le 0$ for $t \ge T$ since H in (56) is decreasing. Therefore, the monotonicity of w follows. Integrating equation (56) yields, for $t \ge T$,

$$\begin{aligned} w(t) - w(T) &= t - T - \frac{1}{\rho}\int_T^t \frac{1}{\bar{F}(w(s))} ds \\ &\le t - T - \frac{1}{\rho}\int_T^t \frac{1}{\bar{F}(w(t))} ds = (t-T)\left(1 - \frac{1}{\rho\,\bar{F}(w(t))}\right) \\ &= -(t-T)\frac{\bar{F}(w) - \bar{F}(w(t))}{\bar{F}(w(t))} \\ &\le -(t-T)(w(t) - w) f^{\downarrow}_{w(t)} \le -(t-T)(w(t)-w) f^{\downarrow}_{w(0)+T}, \end{aligned}$$

where the first inequality holds because $w(s) \ge w(t)$ by the monotonicity of w, the third equality holds because $\bar{F}(w) = 1/\rho$, the second inequality holds because $w(t) \ge w$ and $\bar{F}(w(s)) \le 1$, the last inequality holds because $w(t) \le w(0) + T$ for $0 \le t \le T$ and w is monotone non-increasing for $t > T$. This immediately yields

$$\begin{aligned} \Delta w(t) = w(t) - w &\le -f^{\downarrow}_{w(0)+T}(t-T)\Delta w(t) + (w(T) - w) \\ &= -f^{\downarrow}_{w(0)+T}(t-T)\Delta w(t) + \Delta w(T), \end{aligned}$$

and

$$\Delta w(t) \le \frac{1}{1 + f^{\downarrow}_{w(0)+T}(t-T)}\,\Delta w(T).$$

Relation (51) follows from (50) by splitting interval $[T, t]$ into N disjoint subintervals with equal lengths. Mathematical induction implies that

$$\Delta w(t) \le \left(\frac{1}{1 + f^{\downarrow}_{w(0)+T}\left(\frac{t-T}{N}\right)}\right)^N \Delta w(T).$$

Letting $N \to \infty$ yields the desired (51). ∎

We next give explicit expressions of all performance functions in the $G/M/s+M$ fluid model, with exponential abandonment, when the system is initially empty.

Corollary 2 (*the $G/M/s+M$ fluid queue*) *Consider the $G/M/s+M$ fluid queue with model parameters λ, μ, s, θ, where $\theta > 0$ is the abandonment rate,*

starting empty.
(a) if $\rho \equiv \lambda/s\mu > 1$*, then*

$$\begin{aligned}
w(t) &= \frac{1}{\theta}\log\left(\frac{\rho}{1+(\rho-1)e^{-\theta(t-t_1)}}\right)1_{\{t\geq t_1\}} \uparrow \frac{1}{\theta}\log\rho, && (57)\\
q(t,x) &= \lambda\, e^{-\theta\, x}\, 1_{\{0\leq x\leq w(t),\, t\geq t_1\}} \uparrow \lambda\, e^{-\theta\, x}\, 1_{\{0\leq x\leq (\log\rho)/\theta\}}, && (58)\\
Q(t) &= \frac{\lambda}{\theta}\left(1-\frac{1}{\rho}\right)\left(1-e^{-\theta(t-t_1)}\right)1_{\{t\geq t_1\}} \uparrow \frac{\lambda}{\theta}\left(1-\frac{1}{\rho}\right), && (59)\\
\alpha(t) &= \theta\, Q(t) \uparrow \lambda\left(1-\frac{1}{\rho}\right), && (60)\\
b(t,x) &= \lambda\, e^{-\mu\, x}\, 1_{\{0\leq x\leq t,\, 0\leq t<t_1\}} + \mu\, s\, e^{-\mu\, x}\, 1_{\{0\leq x\leq t,\, t\geq t_1\}} \to \mu\, s\, e^{-\mu\, x}, && (61)\\
B(t) &= \rho\, s(1-e^{-\mu t})\cdot 1_{\{0\leq t<t_1\}} + s\cdot 1_{\{t\geq t_1\}} \uparrow s, && (62)\\
\sigma(t) &= \mu\, B(t) \uparrow \mu\, s, \quad \textit{as} \quad t\to\infty, \quad \textit{for} \quad x\geq 0, && (63)
\end{aligned}$$

where $t_1 \equiv -1/\mu\, \log(1-1/\rho)$*.*
(b) if $\rho \leq 1$*, then*

$$\begin{aligned}
q(t,x) &= Q(t) = \alpha(t) = w(t) = 0,\\
b(t,x) &= \mu\, s\, e^{-\mu\, x}\, 1_{\{0\leq x\leq t\}} \uparrow \mu\, s\, e^{-\mu\, x},\\
B(t) &= \rho\, s(1-e^{-\mu\, t}) \uparrow \rho\, s,\\
\sigma(t) &= \lambda(1-e^{-\mu\, t}) \uparrow \lambda.
\end{aligned}$$

Proof We only prove case (a) since (b) is similar. First, since the system is initially empty, flow conservation of the service facility implies

$$\lambda = B'(t) + \mu\, B(t), \quad B(0) = 0,$$

which has unique solution $B(t) = \rho\, s(1-e^{-\mu\, t})$ when t is small. The system switches to the OL regime at t_1 where $\rho\, s(1-e^{-\mu\, t_1}) = s$, and stays in that regime for all $t > t_1$. This yields (62), from which (63) and (61) follow. For $t \geq t_1$, we have the ODE for BWT

$$w'(t) = \frac{s\,\mu}{\lambda\, e^{\theta\, w(t)}}, \quad w(t_1) = 0,$$

which has unique solution (57), from which (58), (59) and (60) follow. ■

We give a numerical example illustrating Corollary 2 in §B.

Remark 3 (explicit results for queues in series) We can apply Corollary 2 to obtain explicit expressions for the performance functions with two or more queues in series, with exponential abandonment, because the arrival rate of each successive queue is the departure rate from the previous queue, and the departure rate from each queue is available explicitly.

6 Periodic Steady State (PSS) for Periodic Models

In this section we consider the special case of periodic fluid models. We provide conditions under which (i) there exists a unique periodic steady state (PSS) for a periodic fluid model and (ii) the time-varying performance converges to that PSS for all (finite) initial conditions.

6.1 Theory

Recall that a function of a nonnegative real variable, g, is *periodic* with *period* τ if $g(t+\tau) = g(t)$ for all $t \geq 0$, where τ is the least such value, required to be strictly positive. If the relation holds for arbitrary small τ, then the function is constant; we exclude that case. We say that a $G_t/M_t/s_t + GI_t$ fluid queue is a *periodic model* if the function mapping t into the vector $(\lambda(t), \mu(t), s(t), \{F_t(x) : x \geq 0\})$ in $\mathbb{R}^3 \times \mathbb{D}$ is periodic. If the four component functions are periodic, where there is a finite least common multiple of the periods, then the overall function is periodic with the overall period being that least common multiple of the component periods. (The condition is needed; e.g., $\sqrt{2}$ and 1 have no least common multiple.) Since the time-varying abandonment time cdf's $\{F_t(x) : x \geq 0\})$ are defined on the entire real line, we require that they be periodic on their entire domain.

We have not yet said anything about the initial conditions $\{b(0,x) : x \geq 0\}$ and $\{q(0,x) : x \geq 0\}$. If these initial conditions can be chosen so that the system performance of the periodic model with period τ, $\{\mathcal{P}(t) : t \geq 0\}$, where the system state vector $\mathcal{P}(t) \equiv (\{b(t,x) : x \geq 0\}, \{(q(t,x) : x \geq 0\}, B(t), Q(t), w(t), v(t), \sigma(t), \alpha(t))$, is a periodic function of t with period τ, then those initial conditions produce a *periodic steady state* (PSS) for the periodic model with period τ. The performance function $\mathcal{P}$ constitutes the PSS. See Figure 5 for an example. In order to discuss continuity and convergence in the domain of $\mathcal{P}$, we use norm

$$\begin{aligned}\|\mathcal{P}(t)\| \equiv |B(t)| + |Q(t)| + |\alpha(t)| + |\sigma(t)| + |w(t)| + |v(t)| \\ + \left|\int_0^\infty b(t,x)dx\right| + \left|\int_0^\infty q(t,x)dx\right|.\end{aligned}$$

A common case is a periodic model that does not start in a PSS. We then want to conclude that the performance converges to a PSS as time evolves for all finite initial conditions. We say that a function of a nonnegative real variable, g, is *asymptotically periodic* with period $\tau > 0$ if there exists a (finite) function g_∞ such that $g(n\tau+t) \to g_\infty(t)$ as $n \to \infty$ for all t with $0 \leq t \leq \tau$, for the given positive value of τ, but no smaller value; the limit g_∞ necessarily is a periodic function with period τ. This limit can be viewed as an application of the shift operator Ψ_τ on the function g: $\Psi_\tau(g)(t) \equiv g(\tau + t)$, $t \geq 0$. The function g is asymptotically periodic if and only if successive iterates of the shift operator converge, i.e., if $\Psi_\tau^{(n)}(g) \equiv \Psi_\tau(\Psi_\tau^{(n-1)}(g))$ converges as $n \to \infty$.

Theorem 7 (*PSS for the periodic fluid model*) *Consider a periodic fluid queue with period $\tau > 0$. If the conditions of Lemma* 1 *hold, then*

(*a*) *There exists a unique PSS $\mathcal{P}^*$ with period τ, but not with smaller period.*

(*b*) *For any finite initial conditions, the performance $\mathcal{P}$ is asymptotically periodic with period τ, i.e.,*

$$\Psi^{(n)}(\mathcal{P})(t) \equiv \mathcal{P}(n\tau + t) \to \mathcal{P}^*(t) \quad as \quad n \to \infty, \quad 0 \le t \le \tau. \tag{64}$$

Proof First suppose that the system starts empty. By Theorem 1, the shift operator Ψ_τ is a monotone operator on $\mathcal{P}(n\tau)$ for any n, because we can think of the performance $b(\tau, \cdot)$ and $q(\tau, \cdot)$ as alternative initial conditions for the model at time 0, since the model is periodic with period τ. Therefore, the sequence of system performance vectors $\mathcal{P}(0), \mathcal{P}(\tau), \mathcal{P}(2\tau), \ldots$ (at discrete time $0, \tau, 2\tau, \ldots$) is monotonically non-decreasing. By Lemma 1, the performance is bounded, so that there is a finite limit for $\mathcal{P}(n\tau)$ as $n \to \infty$. By Theorem 2, the operator is continuous as well, which implies that $\mathcal{P}(t + n\tau) = \Psi_t(\mathcal{P}(n\tau))$ is convergent for all $0 \le t \le \tau$ as $n \to \infty$. Hence the limit is a PSS. By Theorem 3, we have ALOM, which implies that we get the same limit for all initial conditions. ■

Theorem 3 shows that the rate of convergence to the PSS in Theorem 7 is exponentially fast as well, under regularity conditions.

6.2 An Example

Example 2 (an $G_t/M/s_t + M$ example with periodic arrival rate and staffing)

We now consider a variant of Example 1 that has sinusoidal staffing as well as a sinusoidal arrival rate. As before, we have the fluid queue with arrival rate function in (26) with $a = c = 1$, $b = 0.6$, constant service rate $\mu = 1$ and constant abandonment rate $\theta = 0.5$. However, now we also use the sinusoidal staffing function

$$s(t) = \bar{s} + u\, \sin(\gamma\, t). \tag{65}$$

Let $\bar{s} = a = c = \mu = 1$ $u = 0.3$ and $\gamma = 2$. Note the period of λ is $2\pi/c = 2\pi$, while the period of s is $2\pi/\gamma = \pi$. Hence the overall model has period $2/pi$. Figure 2 shows the results after applying the algorithm in §8 of [8] to compute the performance measures $w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$. Instead of plotting just one OL and UL interval in $[0, T]$ with $T = 10$ as we did in Example 1, here we plot four OL and UL intervals in $[0, T']$ with $T' = 23$.

Figure 2 shows that performance measures ($w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t, 0)$) converge very quickly to periodic limit functions, with period $\tau = \pi$. In Appendix 6 we compare the fluid approximation in this example to simulation results for a large-scale queueing system. As in [8], we see that the fluid model provides a useful approximation for the queueing systems. It is very accurate for very large queueing systems (with thousands of servers) and provides a good approximation for mean values for smaller queueing systems (with tens of

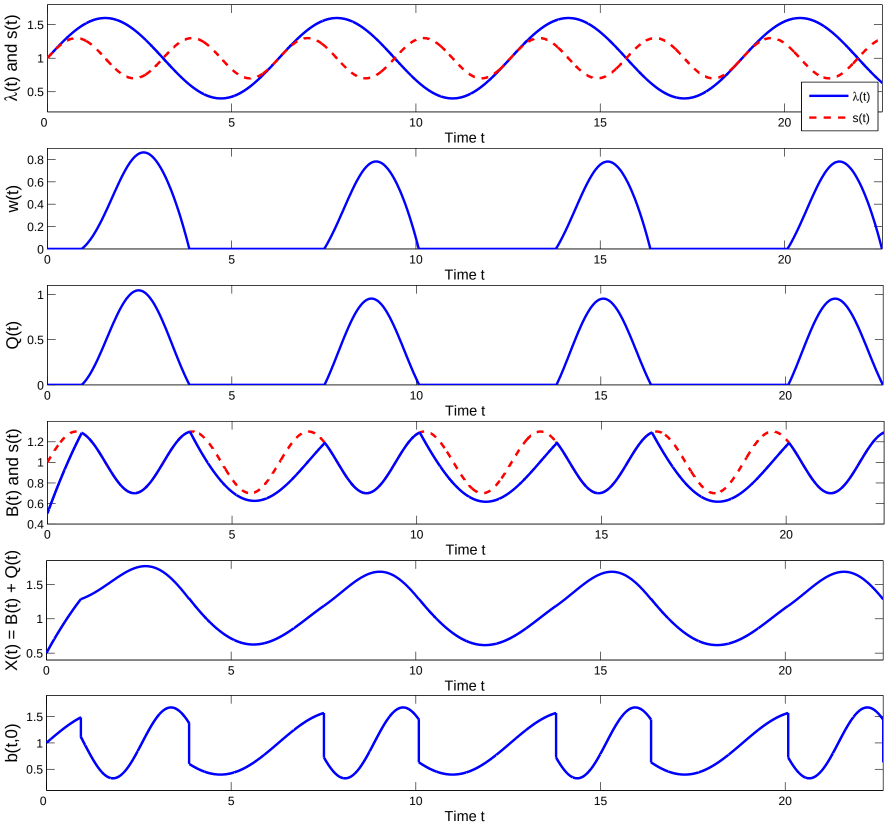

Fig. 2 Performance of the $G_t/M/s_t + M$ model with sinusoidal arrival and staffing, $\gamma = 2$.

servers). In the Appendix we also consider the performance when γ is changed from 2 to 0.5. Figure 4 there shows that the period of the PSS becomes $\tau = 4\pi$. ■

6.3 Direct Computation of PSS Performance

Given the rapid convergence, it usually is not difficult to compute the PSS by simply applying the algorithm with any convenient initial condition. However, the PSS can also be determined in another way. We can start by observing that there are only three cases for PSS: (i) the system is OL for all $0 \leq t \leq \tau$; (ii) the system is UL for all $0 \leq \tau$; or (iii) there is at least one switch between UL and OL regimes in $[0, \tau]$. We can simply check which of these cases prevails. For each of these scenarios, we can seek a fixed point in the performance at times τ and 0. That produces equations we can solve. One of these three cases will yield the PSS.

Consider case (i), in which the system is OL. It suffices to characterize its performance in one cycle $[0,\tau]$. We can write

$$B(t) = s(t) \quad \text{and} \quad Q(0) = \int_0^{w(0)} \lambda(t-x)\bar{F}_{t-x}(x)dx \quad \text{for} \quad w(0) > 0,$$

because in the PSS the system remains OL. Hence, we must have $q(t,0) = \lambda(t)$ and $q(t,x) = \lambda(t-x)\bar{F}_{t-x}(x)$. Note that $w_0 \equiv w(0)$ is the only unknown here. To solve for the PSS, we do a search of the initial w_0 such that during the cycle $[0,\tau]$, the system is always OL, i.e., $w(t) > 0$, and $w(\tau) = w_0$. The uniqueness of the PSS guarantees that there is at most one of such w_0. If the system switches to UL regime at some time, then we know this is not the right scenario for the PSS.

Next consider case (ii), in which the system is UL in the interval $[0,\tau]$. Since the system is UL, the fluid content in service $B(t)$ satisfies the ODE $\lambda(t) = B'(t) + \mu(t)\,B(t)$ with initial condition $B(0) = B_0 > 0$ which has a unique solution

$$B(t) = e^{-\int_0^t \mu(s)ds}\left(\int_0^t e^{\int_0^s \mu(u)du}\lambda(s)ds + B_0\right), \quad \text{for } 0 \le t \le \tau. \qquad (66)$$

Since we seek $B(\tau) = B_0$, it suffices to solve equation

$$B_0 = e^{-\int_0^\tau \mu(s)ds}\left(\int_0^\tau e^{\int_0^s \mu(u)du}\lambda(s)ds + B_0\right)$$

for B_0. Again, the uniqueness of PSS guarantees that there is at most one such $B_0 > 0$. If this equation does not have a solution, then we know this is not the right scenario for the PSS.

Finally, consider case (iii), in which the system switches at least twice between UL and OL regimes, as shown in Figure 2. Since system regime changes in the PSS, we consider the interval $[0,\tau]$ and assume that in PSS the system is critically loaded at $t = 0$ and becomes OL at $t+$, i.e., we can always let the beginning of the cycle of PSS be a regime switching point from UL to OL. We assume that the phase difference between the PSS cycle and the model functions is $0 \le t_0 \le \tau$. Hence, we start with the BWT ODE

$$w'(t) = 1 - \frac{\mu(t+t_0)\,s(t+t_0) + s'(t_0)}{\lambda(t+t_0-w(t))\bar{F}_{t+t_0-w(t)}(w(t))}, \quad \text{with } w(0) = 0,$$

and let $t_1 \equiv \inf\{t > 0 : w(t) = 0,\, \lambda(t+t_1) \le \mu(t)\,s(t) + s'(t)\}$. If $t_1 > \tau$ (e.g., $t_1 = \infty$), then we know this is not the right scenario. If $t_1 < \tau$, the system switches to the UL regime at t_1. Then, just as in (66), we have

$$B(t) = e^{-\int_{t_1}^t \mu(s+t_0)ds}\left(\int_{t_1}^t e^{\int_0^s \mu(u+t_0)du}\lambda(s+t_0)ds + B(t_1)\right),$$

with $B(t_1) = s(t_1+t_0)$. We let $t_2 \equiv \inf\{t > t_1 : B(t) > s(t+t_0)\}$. If $t_2 < \tau$, then the system switches back to OL regime after t_2. We repeat the above procedure until we get to time τ. If the initial phase difference variable t_0 is the right one, the system should again be critically loaded at τ. We do a search for t_0 in $[0, \tau]$.

Since analytic expressions are available for the $G/M/s+M$ fluid model as shown in Corollary 2, we show how explicit PSS performance functions can be calculated in the next example.

Example 3 (explicit PSS performance in special cases) Consider the $G_t/M/s+M$ fluid model in Example 1 that has sinusoidal arrival rate as in (26), exponential service distribution with rate μ, constant staffing s and exponential patience distribution with rate θ. We suppose that we are in case (iii) above, in which there is a switching point from UL to OL regimes, which we can take to be at the beginning of a cycle. We assume the arrival rate is $\tilde{\lambda}(t) \equiv \lambda(t+t_0)$ for some $0 \le t_0 \le \tau$. At some t_1 for $0 < t_1 < \tau \equiv 2\pi/c$, the system will switch to the UL regime. Hence, in order to characterize the complete performance in a cycle $[0, \tau]$, it remains to determine the values of t_0 and t_1 for $0 \le t_0 \le \tau$, $0 \le t_1 \le \tau$.

Since the system is critically loaded at $t = 0$, OL in $[0, t_1)$ and UL in $[t_1, \tau]$, we need two equations for two unknowns t_0 and t_1. First, the BWT ODE implies that $w(0) = 0$ and

$$w'(t) = 1 - \frac{\mu\, s}{\tilde{\lambda}(t - w(t))\, e^{-\theta\, w(t)}} = 1 - \frac{\mu\, s\, e^{\theta\, t}}{\tilde{\lambda}(t - w(t))\, e^{\theta (t - w(t))}}, \quad 0 \le t \le t_1,$$

which yields that

$$\mu\, s\, e^{\theta\, t} = \tilde{\lambda}(t - w(t))\, e^{\theta(t - w(t))}(1 - w'(t)) = \tilde{\lambda}(t - w(t))\, e^{\theta(t-w(t))} \frac{d(t - w(t))}{dt}.$$

Integrating both sides and let $v(t) \equiv t - w(t)$, we have

$$\int_0^t \mu\, s\, e^{\theta\, u} du = \int_0^{v(t)} \tilde{\lambda}(y) e^{\theta\, y} dy.$$

Plugging the sinusoidal arrival rate $\tilde{\lambda}(t) = \lambda(t + t_0)$ into the above equation yields that

$$\begin{aligned}\frac{\mu\, s}{\theta}(e^{\theta\, t} - 1) \;=\; & \frac{a}{\theta}(e^{\theta\, v(t)} - 1) + \frac{b}{1 + c^2/\theta^2} \left[\frac{1}{\theta} e^{\theta\, v(t)} \sin(c\, v(t) + c\, t_0) \right.\\ & \left. - \frac{c}{\theta^2}(e^{\theta\, v(t)} \cos(c\, v(t) + c\, t_0) - \cos(c\, t_0))\right].\end{aligned}$$

Since $v(t_1) = t_1 - w(t_1) = t_1$, letting $t = t_1$ in the above equation yields

$$\begin{aligned}\frac{\mu\, s}{\theta}(e^{\theta\, t_1} - 1) \;=\; & \frac{a}{\theta}(e^{\theta\, t_1} - 1) + \frac{b}{1 + c^2/\theta^2} \left[\frac{1}{\theta} e^{\theta\, t_1} \sin(c\, t_1 + c\, t_0) \right.\\ & \left. - \frac{c}{\theta^2}(e^{\theta\, t_1} \cos(c\, t_1 + c\, t_0) - \cos(c\, t_0))\right]. \qquad (67)\end{aligned}$$

Second, since the system is UL in $[t_1, \tau]$, we have

$$\lambda(t + t_0) = \tilde{\lambda}(t) = B'(t) + \mu\, B(t), \quad t_1 \le t \le \tau,$$

which implies that

$$B(t)e^{\mu\, t} - B(t_1)e^{\mu\, t_1} = \int_{t_1}^{t} \lambda(u + t_0)e^{\mu\, u} du.$$

Since the system becomes critically loaded again at t_1 and at the end of the cycle, i.e., $B(t_1) = B(\tau) = B(2\pi/c) = s$, plugging the sinusoidal arrival rate into the above equation yields

$$\begin{aligned} s(e^{-\mu\, 2\pi/c} - e^{-\mu\, t_1}) &= \frac{a}{\mu}(e^{-\mu\, 2\pi/c} - e^{-\mu\, t_1}) \\ &+ \frac{b}{1 + c^2/\mu^2}\left[\frac{1}{\mu}(e^{\mu\, 2\pi/c}\sin(2\pi + c\, t_0) - e^{\mu\, t_1}\sin(c\, t_0 + c\, t_1)) \right. \\ &\left. - \frac{c}{\mu^2}(e^{\mu\, 2\pi/c}\cos(2\pi + c\, t_0) - e^{\mu\, t_1}\cos(c\, t_0 + c\, t_1))\right]. \end{aligned} \tag{68}$$

Unfortunately, Equation (67) and (68) evidently do not have explicit solutions in general, but they can be solved quite easily numerically by performing a search over the two unknowns. However, we can continue analytically in a special case with convenient parameters: (a) $a = s\mu$ and (b) $\mu = \theta$.

Note that (a) says that the average traffic intensity is $\bar{\rho} = \bar{\lambda}/s\mu = a/s\mu = 1$ and (b) says that this model is equivalent to an infinite-server model, because $\theta = \mu$.

With these extra assumptions, equations (67) and (68) simplify to

$$\begin{aligned} \frac{c}{\theta}\cos(c\, t_0) &= -e^{\theta\, t_1}[\sin(c\, t_1 + c\, t_0) - \frac{c}{\theta}\cos(c\, t_1 + c\, t_0)], \\ e^{\mu\, 2\pi/c}[\sin(c\, t_0) - \frac{c}{\mu}\cos(c\, t_0)] &= e^{\mu\, t_1}[\sin(c\, t_1 + c\, t_0) - \frac{c}{\mu}\cos(c\, t_1 + c\, t_0)]. \end{aligned}$$

Adding these two equations yields

$$0 \le t_0 = \frac{1}{c}\arctan(1 - e^{-\mu\, 2\pi/c}) \le \pi/c. \tag{69}$$

Note that we need $\lambda(0) = a + b\sin(c\, t_0) \ge \mu\, s$ so that the system switches from UL to UL regime at $t = 0$. Similarly, we require $\lambda(t_0 + t_1) \le \mu\, s$, which implies that $\pi/c \le t_0 + t_1 \le 2\pi/c$. Hence, plugging (69) into the first equation above implies that t_1 is the solution to

$$\sin(ct_1 + \psi) = -\frac{(c/\theta)e^{e^{\mu\, 2\pi/c}}}{\sqrt{x^2 + y^2}}\, e^{-\theta\, t_1}, \tag{70}$$

where $\psi \equiv \arctan(x/y)$, $x \equiv e^{\mu\, 2\pi/c} - 1 - (c/\theta)e^{\mu\, 2\pi/c}$, $y \equiv e^{\mu\, 2\pi/c} + (c/\theta)(e^{\mu\, 2\pi/c} - 1)$.

Given t_0 and t_1, we can compute analytically all performance functions of this $G_t/M/s+M$ example in a cycle $[0,\tau]=[0,2\pi/c]$. For $0 \le t < t_1$, the system is OL with

$$\begin{aligned}
q(t,0) &= \tilde{\lambda}(t) = a + b\,\sin[c(t+t_0)], \\
q(t,x) &= \tilde{\lambda}(t-x)\,e^{-\theta\,x} = e^{-\theta\,x}(a + b\,\sin[c(t+t_0-x)]), \\
w(t) &= t - \Lambda^{-1}\left(\frac{\mu\,s}{\theta}(e^{\theta\,t}-1)\right), \\
Q(t) &= \int_0^{w(t)} q(t,x)dx = e^{-\theta\,t}\Lambda(t) - \frac{\mu\,s}{\theta}(1-e^{-\theta\,t}), \\
\alpha(t) &= \theta\,Q(t), \\
B(t) &= s, \quad \sigma(t) = \mu\,s, \\
b(t,x) &= \mu\,s\,e^{-\mu\,x}\,1_{\{x\in\cup_{k=0}^{\infty}((t+k\tau-t_2)^+,t+k\tau]\}} \\
&\quad +\lambda(t-x)\,e^{-\mu\,x}\,1_{\{x\in\cup_{k=0}^{\infty}(t+k\tau,t+(k+1)\tau-t_2]\}},
\end{aligned}$$

where $\Lambda(x) \equiv \int_0^x \lambda(y)\,e^{\theta\,y}dy$. For $t_1 \le t \le \tau$, the system is UL with

$$\begin{aligned}
q(t,x) &= Q(t) = w(t) = \alpha(t) = 0, \\
b(t,0) &= \tilde{\lambda}(t) = a + b\,\sin[c(t+t_0)], \\
b(t,x) &= \tilde{\lambda}(t-x)\,e^{-\mu\,x}\,1_{\{x\in\cup_{k=0}^{\infty}((t+(k-1)\tau)^+,t+k\tau-t_2]\}} \\
&\quad +\mu\,s\,e^{-\mu\,x}1_{\{x\in\cup_{k=0}^{\infty}(t-t_2+k\tau,t+k\tau]\}}, \\
B(t) &= s\,e^{-\mu(t-t_1)} + e^{-\mu\,t}\int_{t_1}^{t}\tilde{\lambda}(u)e^{\mu\,u}du, \\
\sigma(t) &= \mu\,B(t),
\end{aligned}$$

7 Conclusions

In this paper we supplemented [8,9,18] by studying the large-time asymptotic behavior of the $G_t/M_t/s_t+GI_t$ many-server fluid queue with time-varying model parameters. In §4 we established the asymptotic loss of memory (ALOM) property, concluding that the difference between performance functions evaluated at time t, with different initial conditions, dissipates exponentially fast as $t \to \infty$, under regularity conditions. In §5 we applied ALOM to establish convergence to steady state for the stationary model. In §5 we also went beyond ALOM to provide additional details; e.g., we showed that the system changes regimes (overloaded or underloaded) at most once. In §6 we applied ALOM, first, to establish the existence of a unique periodic steady state (PSS) and, second, to establish convergence to that PSS in the periodic model, where the period is the least common multiple of the periods of the model functions, assumed to be some finite value.

There are many directions for future research: First, it remains to establish ALOM properties for the $G_t/GI/s_t+GI$ fluid queue with non-exponential

(GI) service that was considered in [8] (under regularity conditions that exclude the counterexample in [10]) and the $(G_t/M_t/s_t + GI_t)^m/M_t$ network of fluid queues with proportional routing considered in [9]. Second, it remains to establish many-server heavy-traffic limits showing that appropriately scaled stochastic processes in many-server queues converge to the fluid queues, as discussed in [8,18]. It also remains to establish refined stochastic approximations as a consequence of many-server heavy-traffic limits. Third, it remains to establish corresponding ALOM (or weak ergodicity) and PSS properties for the corresponding stochastic queueing models and the refined stochastic approximation; see [3,5,6,19] and references therein. Fourth, it remains to exploit the deterministic fluid models to approximately solve important control problems for the stochastic systems and, fifth, it remains to apply the fluid models to analyze large-scale service systems, such as hospital emergency departments. We hope to contribute to these goals in the future.

Acknowledgements This research was supported by NSF grant CMMI 0948190.

References

1. Eick, S. G., W. A. Massey, W. Whitt.: $M_t/G/\infty$ queues with sinusoidal arrival rates. Management Sci. **39(2)**, 241-252 (1993)
2. Garnett, O., Mandelbaum, A., Reiman, M. I.: Designing a call center with impatient customers. Manufacturing Service Oper. Management. **4**, 208-227 (2002)
3. Granovsky, B. L. Zeifman, A.: Nonstationary queues: estimating the rate of convergence. Queueing Systems. **46**, 363-388 (2004)
4. Hall, R. W.: *Queueing Methods ofor Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ (1991)
5. Heyman, D., Whitt, W.: The Asymptotic Behavior of Queues with Time-Varying Arrival Rates. Journal of Applied Probability. **21**, 143-156 (1984)
6. Isaacson, D., Madsen, R.: Markov Chains: Theory and Applications. Wiley, New York (1976)
7. Krichagina, E. V., Puhalskii, A. A.: A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. Queueing Systems. **25**, 235-280 (1997)
8. Liu, Y., Whitt, W.: A fluid approximation for the $G_t/GI/s_t + GI$ queue. Columbia University, NY, NY (2010) http://www.columbia.edu/∼ww2040/allpapers.html
9. Liu, Y., Whitt, W.: A network of time-varying many-server fluid queues with customer abandonment. Columbia University, NY, NY (2010) http://www.columbia.edu/∼ww2040/allpapers.html
10. Liu, Y., Whitt, W.: The heavily loaded many-server queue with abandonment and deterministic service times, Columbia University, NY, NY (2010) http://www.columbia.edu/∼ww2040/allpapers.html
11. Mandelbaum, A., Massey, W. A., Reiman, M. I.: Strong approximations for Markovian service networks. Queueing Systems. **30**, 149-201 (1998)
12. Mandelbaum, A., Massey, W. A., Reiman, M. I., Rider, B.: Time varying multiserver queues with abandonments and retrials. Proceedings of the 16th International Teletraffic Congress, P. Key and D. Smith (des.) (1999)
13. Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A.: Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. Proceedings of the Thirty-Seventh Annual Allerton Conference on Communication, Control and Computing, Allerton, IL, 1095-1104 (1999)
14. Newell, G. F.: Applications of Queueing Theory. second ed., Chapman and Hall, London (1982)

15. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues. Queueing Systems. **65** 325–364 (2010)
16. Puhalskii, A. A.: The $M_t/M_t/k_t + M_t$ queue in heavy traffic. Mathematics Departure. University of Colorado at Denver (2008)
17. Whitt, W.: Stochastic-Process Limits, Springer, New York (2002)
18. Whitt, W.: Fluid models for multiserver queues with abandonments. Oper. Res. **54**, 37-54 (2006)
19. Wilie, H.: Periodic steady state of loss systems. Adv. Appl. Prob. **30**, 152-166 (1998)

APPENDIX

A Overview.

This appendix contains additional supplementary material. In §B we give a numerical example illustrating convergence to steady state for the stationary $G/M/s+M$ model starting empty. In §C we give the other half of the proof of Theorem 6, establishing pointwise convergence of the fluid densities $b(t,x)$ and $q(t,x)$ as $t \to \infty$ when the system is initially OL. In §D we give another example of periodic steady state (PSS) in a model with both sinusoidal arrival rate and staffing function, complementing Example 2. In §E we verify the explicit formulas for the PSS in Example 3. In §6, we compare the fluid approximation to results from simulations of corresponding stochastic queueing models, for the example considered in §B. These simulation results substantiate that (i) the theorems are correct, (ii) the numerical algorithm is effective and (iii) the fluid approximation for the stochastic queueing system is effective. The fluid model accurately describes single sample paths of very large queueing systems and accurately describes the mean values for smaller queueing systems, e.g., with 20 servers.

B Convergence to Steady State in the $G/M/s+M$ Fluid Queue

In this section we give a numerical example illustrating the convergence to steady state for a $G/M/s+M$ queue starting empty, as characterized by Corollary 2. Here we let $\mu = 1$, $\lambda = 1.5$, $s = 1$, $\theta = 0.5$. In Figure 3, we show how performance functions (the solid lines) converge to their steady states (the dashed lines), applying the algorithm described in §8 of [8]. Figure 3 shows that $w(t)$, $Q(t)$, $B(t)$ and $b(t,0)$ quickly converge to their steady state values.

C Proof of Theorem 6

Proof We now complete the proof of Theorem 6 by proving (47) and (48) when the system is initially OL, i.e., $q(0,x) \geq 0$ for some x, $w(0) \geq 0$, $Q(0) \geq 0$ and $B(0) = s$. As before, for simplicity, we assume $\mu = s = 1$ and therefore $\rho = \lambda/s\mu = \lambda$.

(i) $\rho < 1$. Since the service is exponential at the fixed rate $\mu = 1$ and the staffing is fixed at $s = 1$, the output rate of the service facility is 1. Hence, $Q'(t) = \lambda - \alpha(t) - b(t,0) < \lambda - b(t,0) < 1$ as long as the system is in the OL regime; moreover, the OL regime will end after some $0 < T < 1/(1-\rho)$. The system will switch to the UL regime at T (i.e., $Q(T) = w(T) = 0$, $B(T) = s = 1$) and will stay there for all $t > T$. Thus we can apply (21) to characterize the density in service. By Assumption (49), for $t \geq T$,

$$\begin{aligned} b(t,x) &= \rho e^{-x} 1_{\{0 \leq x \leq t-T\}} + b(T, x-t+T) e^{-(t-T)} 1_{\{x > t-T\}} \\ &= \rho e^{-x} 1_{\{0 \leq x \leq t-T\}} + b(0, x-t) e^{-t} 1_{\{x > t-T\}} \\ &\to \rho e^{-x} \quad as\ t \to \infty, \quad x \geq 0. \\ B(t) &= \int_0^{t-T} \rho e^{-x} dx + \int_{t-T}^{\infty} b(T, x-t+T) e^{-(t-T)} dx \\ &= \rho(1 - e^{-(t-T)}) + e^{-(t-T)} B(T) \to \rho, \quad as\ t \to \infty, \end{aligned}$$

Moreover, $\sigma(t) = B(t) \to \rho$, as $t \to \infty$.

(ii) $\rho \geq 1$. As in case (i), the maximum output rate of the service facility is 1. Since $\rho \geq 1$, $\lambda \geq 1$, so that the the system necessarily will stay in the OL or CL regime forever. Since $b(t,0) = \sigma(t) = 1$, all old fluid will leave the queue after $T \equiv Q(0)/b(t,0) = Q(0)$.

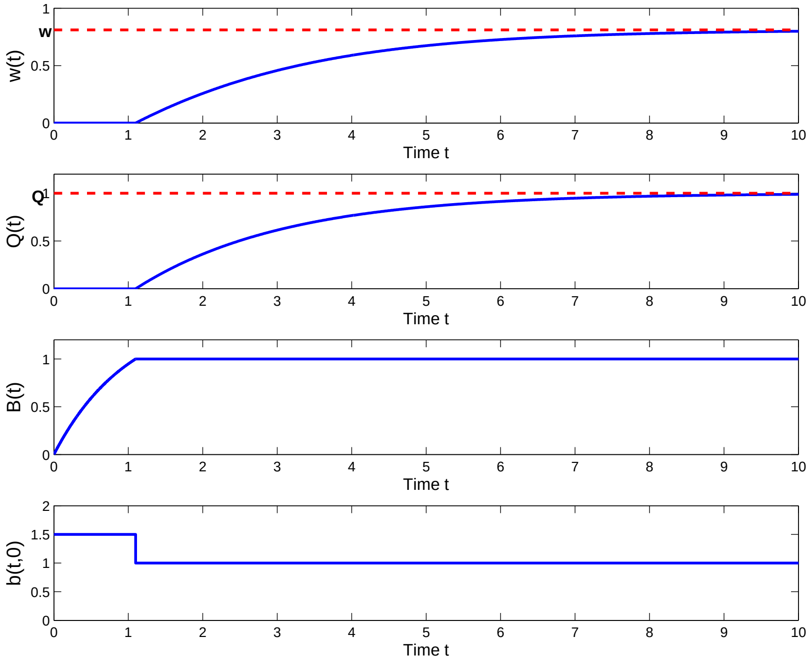

Fig. 3 Performance measures of the $G/M/s+M$ fluid queue converge to their steady states.

Therefore, for $t \le T$, we have $q(t,x) = \rho\bar{F}(x)1_{\{x\le w(t)\wedge(t-T)\}} \to q(x) = \rho\bar{F}(x)1_{\{x\le w\}}$ if $w(t) \to w$ as $t \to \infty$.

If $w(T) < w$, the same reasoning in part (ii) of the proof in the main paper implies that $w(t) \uparrow w$ monotonically after T. If $w(T) = w$, then from (56) we see that $w'(T) = 0$, which implies that the system is already in steady state and thus will stays there forever. If $w(T) > w$, it is easy to see that $w'(t) = H(w(t)) < H(w) = 0$ for $t \ge T$, where $H(\cdot)$ is defined in (56). Therefore, $w(t)$ is decreasing (has negative derivative) as long as $w(t) > w$. To show that $w(t) \to w$ as $t \to \infty$, it remains to show that for any $\epsilon > 0$, there exits a t_ϵ such that $w(t) < w + \epsilon$ for any $t > t_\epsilon$. Because H is strictly decreasing in a neighborhood of w, we have $w'(t) = H(w(t)) \le H(w+\epsilon) \equiv \delta(\epsilon) < H(w) = 0$, if $w(t) \ge w + \epsilon$. Therefore, the derivative of $w(t)$ is not only negative, but also bounded by $\delta(\epsilon) < 0$. So $w(t)$ will hit $w + \epsilon$ at least linearly fast with slope $\delta(\epsilon)$, i.e., for any $t \ge T + (w(T) - w - \epsilon)/|\delta(\epsilon)|$, we have $w(t) \le w + \epsilon$. Therefore, we conclude that $w(t) \downarrow w$ as $t \to \infty$. All the other results follow from the same reasoning as in the proof in the main paper.

D Another Example of Periodic Steady State

We complement Example 2 by considering another value for the parameter γ in the sinusoidal staffing function in (65). Here we let $\gamma = 0.5$ instead of 2.0. That makes the model period 4π instead of π. Figure 4) shows the performance functions.

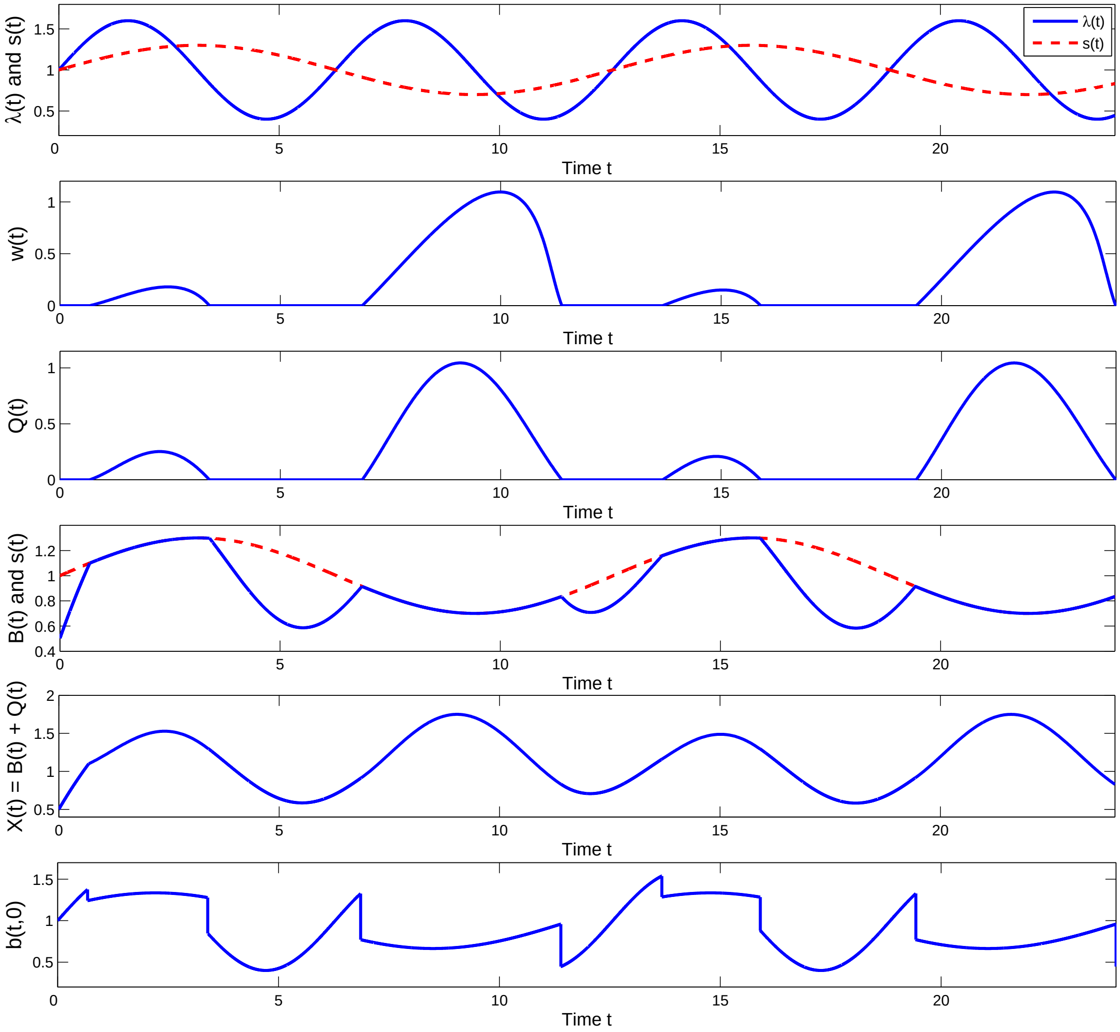

Fig. 4 Performance of the $G_t/M/s_t+M$ model with sinusoidal arrival and staffing, $\gamma = 0.5$.

E Verifying the Sinusoidal PSS

We now verify the PSS for Example 3. To verify t_0 and t_1 in (69) and (70), we let $a = s = \mu = c = \theta = 1$, $b = 0.6$. For these parameters, we get $t_0 = 0.78$ and $t_1 = 3.15$ from (69) and (70). We apply the algorithm in §8 of [8] and plot the performance measures $w(t)$, $Q(t)$, $B(t)$, $X(t)$ and $b(t,0)$ in Figure 5 for $0 \leq t \leq 3 \cdot 2\pi/c = 6\pi$ (three cycles) with the system initially critically loaded and arrival rate $\lambda(t) = a + b \cdot \sin(c(t + t_0))$ (see Plot 1 in Figure 5 for the phase difference: $6.28 - 5.50 = 0.78 = t_0$).

Figure 5 shows that the fluid performance immediately becomes stationary (a DSS cycle starts at time 0 and ends at 2π). Since the $M_t/M/s + M$ model here is equivalent to the $M_t/M/\infty$ model, we can also verify these analytical formulas by showing that they agree with previous ones derived for the $M_t/M/\infty$ model in [1].

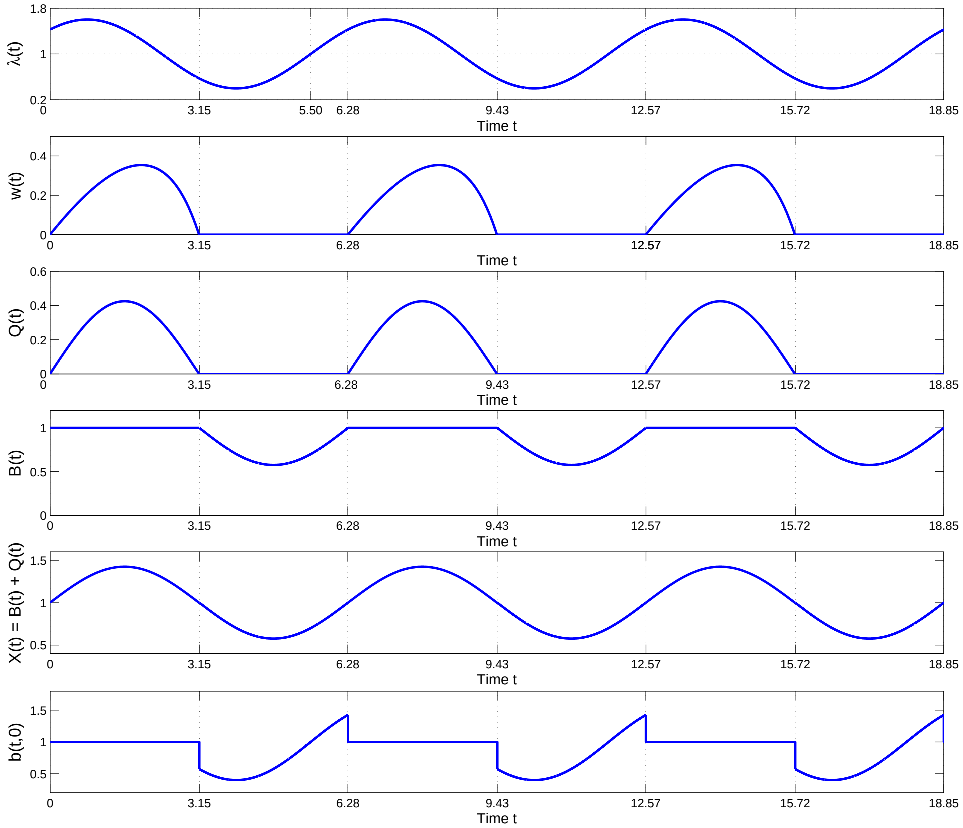

Fig. 5 The $G_t/M/s{+}M$ model in Example 3 is in PSS at time 0, with period $\tau = 2\pi = 6.28$. In each cycle $[n\tau, (n+1)\tau]$ of PSS, the system switches between UL and OL regimes twice at time $n\tau$ and $n\tau + 3.15$.

6 A Comparison with Simulation

In §D, we considered the $G_t/M/s_t + M$ fluid queue, which has a sinusoidal arrival rate $\lambda(t)$ as in (26) with $a = c = 1$, $b = 0.6$, sinusoidal staffing function $s(t)$ as in (65) with $\bar{s} = 1$, $u = 0.3$, $\gamma = 0.5$, exponential service and abandonment distributions with rate $\mu = 1$ and $\theta = 0.5$. We let the system be initially UL with $B(0) = 0.5 < s(0)$. We now compare the fluid approximation as shown in 4 with computer simulations of the associated $M_t/M/s_t + M$ queueing model.

This queueing model has the same service and abandonment rates, but scaled arrival rate and number of servers: $n\,\lambda(t)$ and $n\,s(t)$. There are $n\,B(0)$ customers in service at time 0. Let $W_n(t)$ be the elapsed waiting time of the customer at the head of the queue at t, $\tilde{Q}_n(t)$ be the number of customers in queue and $\tilde{B}_n$ be the number of customers in service. Applying the spatial scaling, we let $Q_n(t) \equiv \tilde{Q}_n(t)/n$ and $B_n(t) \equiv \tilde{B}_n(t)/n$. We let $X_n(t) \equiv Q_n(t) +$

$B_n(t)$ be the scaled total number of customers in the system at t. In Figure 6, 7 and 8, we compare the simulation results for the queue performance functions W_n, Q_n and B_n from a single simulation run to the associated fluid model counterparts w, Q and B, with $n = 30$, $n = 100$ and $n = 1000$. The blue solid lines represent the queueing model performance, while the red dashed lines represent the corresponding fluid performance. We observe that the bigger the scaling n is, the more accurate the fluid approximation becomes. When $n = 1000$, we have a large-scale queueing model (with arrival rate $1000 + 600\,\sin(t)$ and staffing $1000{+}300\,\sin(0.5\,t)$ servers) and we get close agreement for individual sample paths.

When n is smaller, there are bigger stochastic fluctuations as shown in Figures 6 and 7, but the mean values of the queueing functions still are quite well approximated by the fluid performance functions when the system is not nearly critically loaded. We illustrate by considering the cases $n = 100$ and $n = 30$ in Figures 9 and 10, where average sample paths of simulation estimates are compared with fluid approximations. In Figure 9, we average 20 sample paths for $n = 100$; in Figure 10, we average 200 sample paths for $n = 30$. We need more samples for smaller scaling n, because there are bigger fluctuations.

A careful examination of Figure 9 and 10 show that in both cases the total fluid content, $X(t)$, very accurately approximates the expected value of the scaled total number of customers, $X_n(t)$, in the queueing system. However, the fluid queue content $Q(t)$ and the fluid service content $B(t)$ do not approximate the mean values of their counterparts in the queueing system as well. In particular, the quality of these approximations degrades when the system is nearly critically loaded. That is understandable, because only positive fluctuations will be captured by the queue length, while only negative fluctuations will be captures by the number of busy servers.

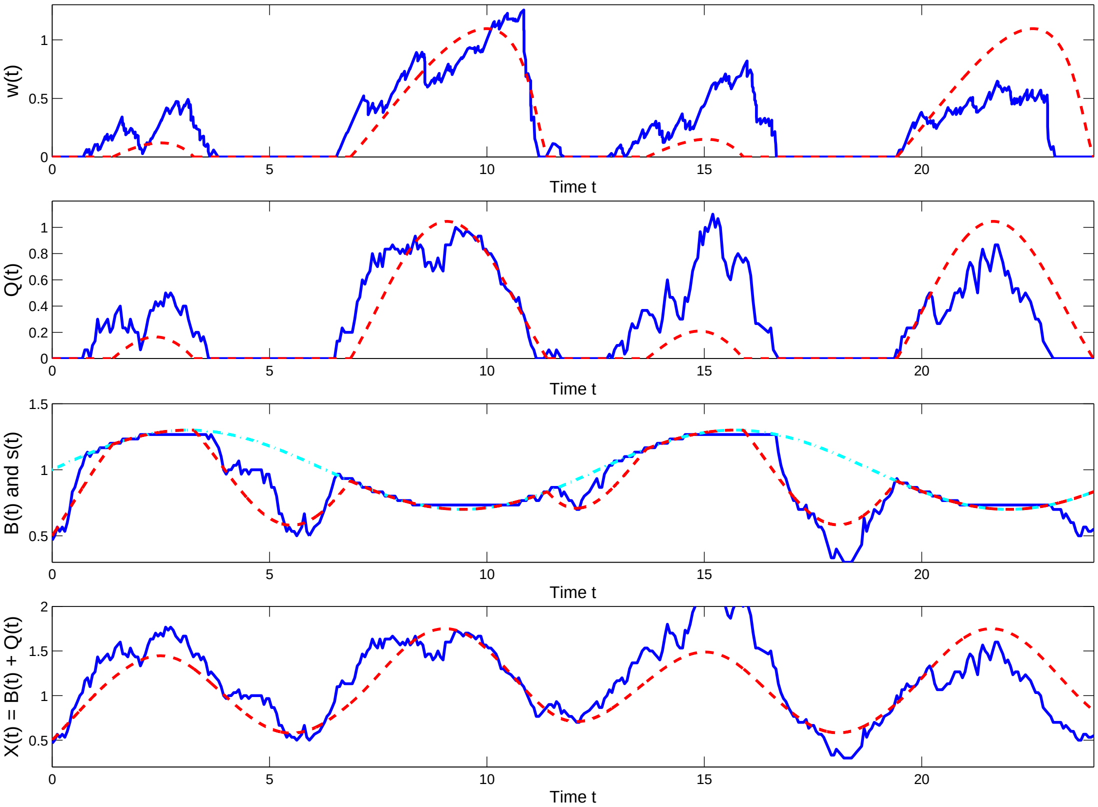

Fig. 6 Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 30$.

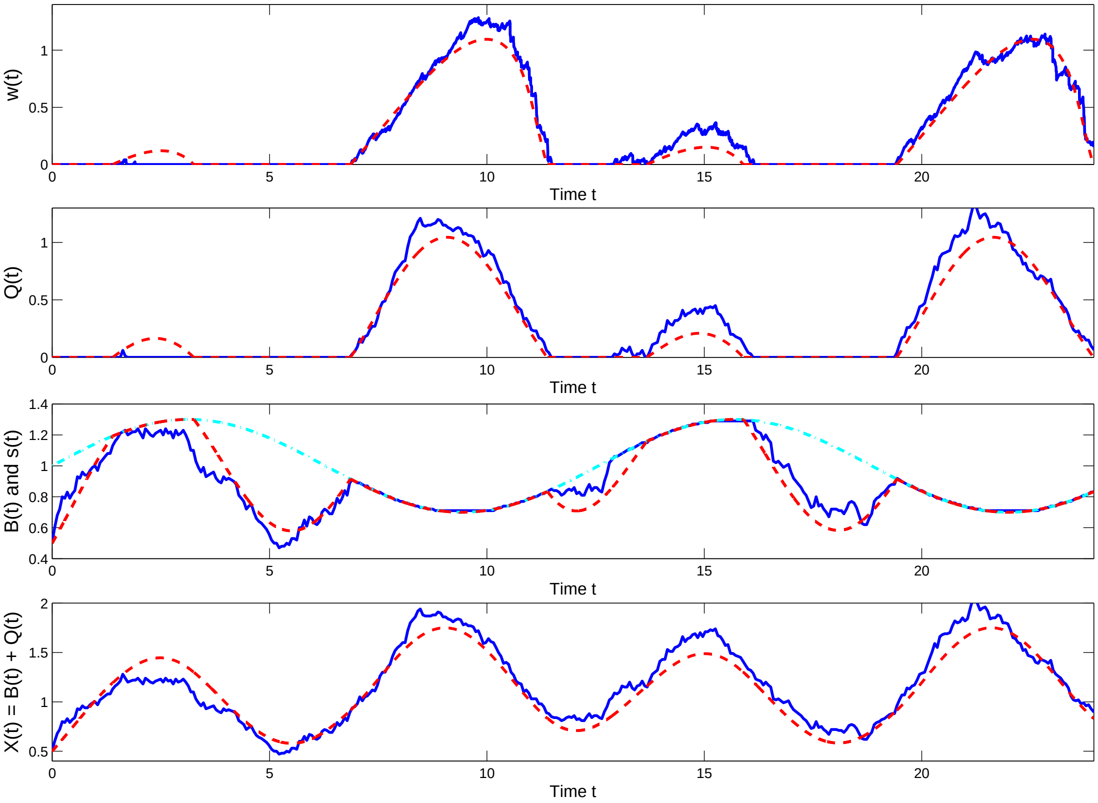

Fig. 7 Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.

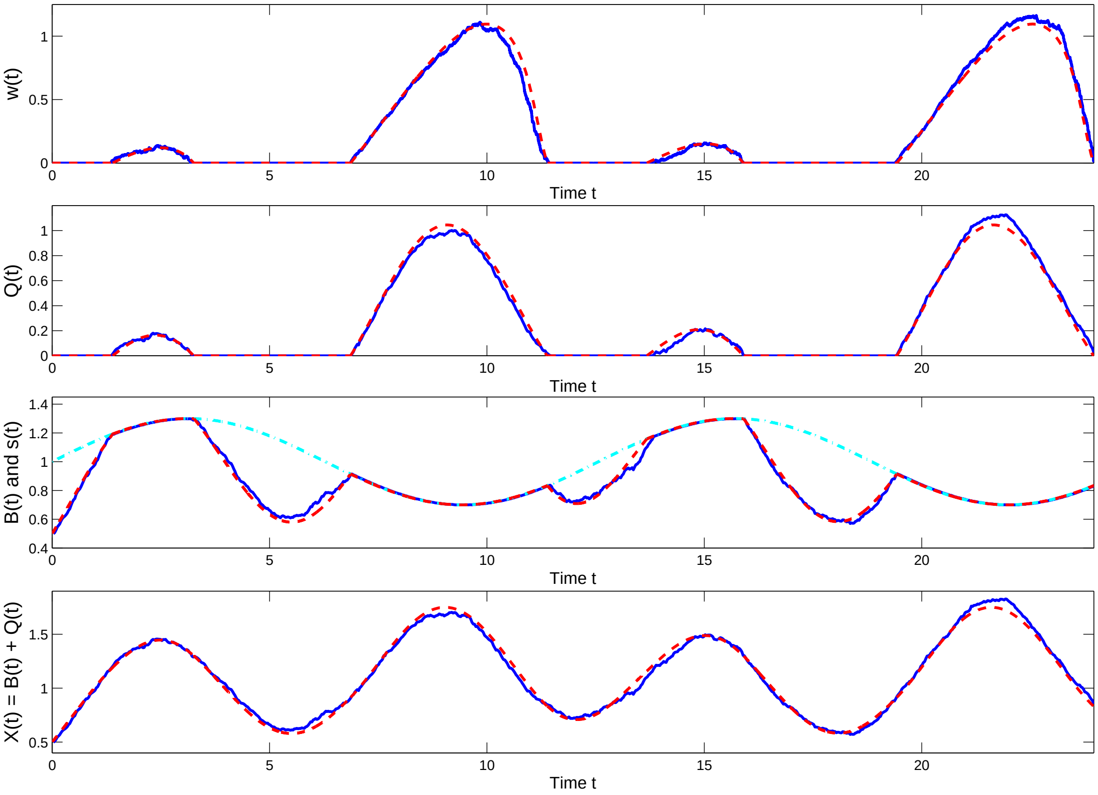

Fig. 8 Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 1000$.

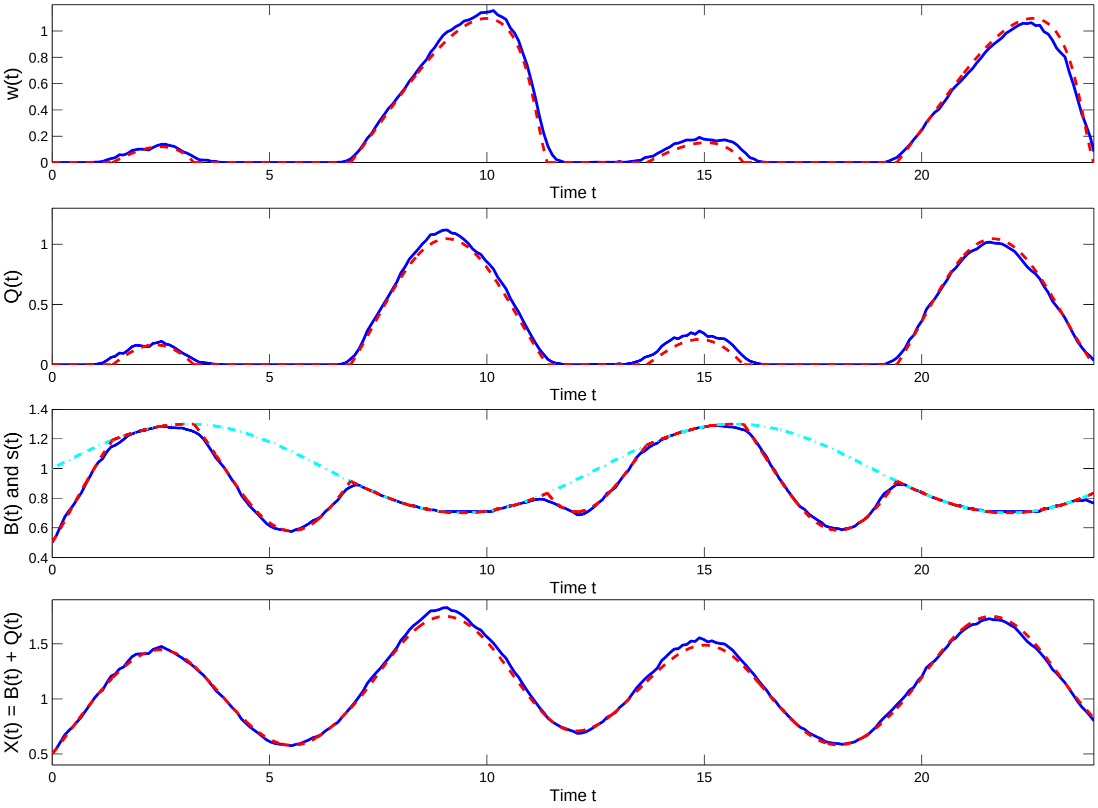

Fig. 9 Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: an average of 20 sample paths of the scaled queueing model based on $n = 100$.

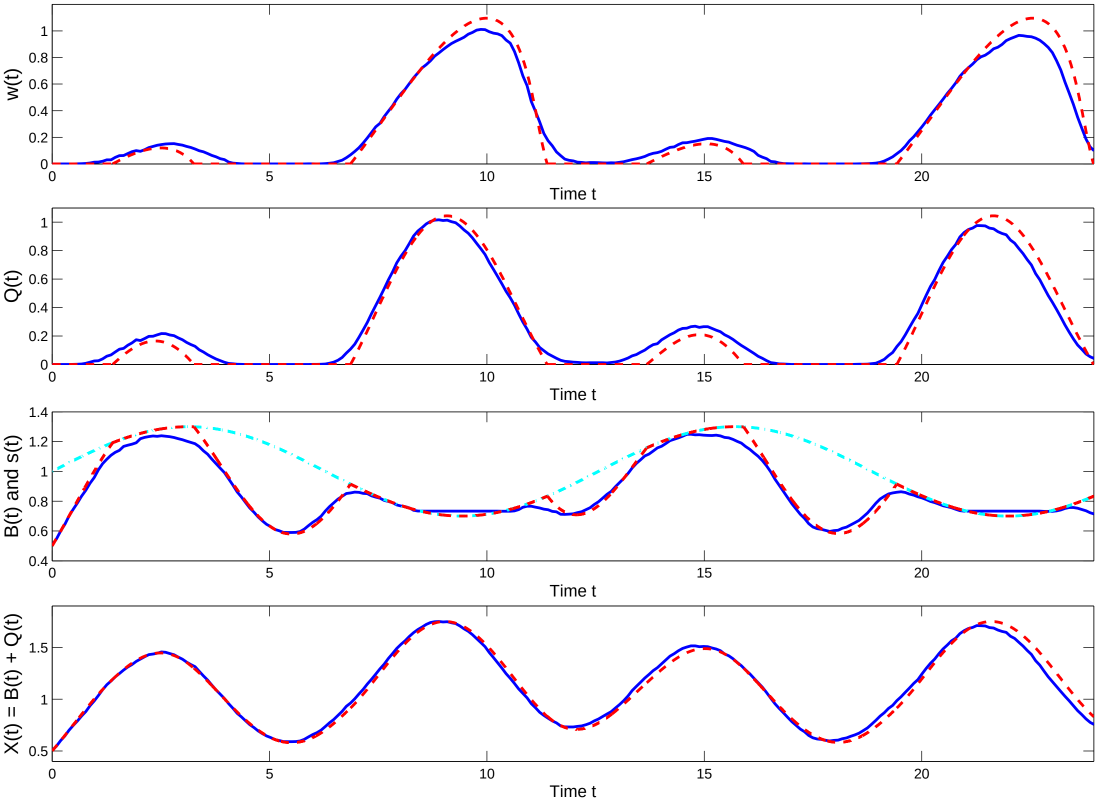

Fig. 10 Performance of the $G_t/M/s_t + M$ fluid model compared with simulation results: an average of 200 sample paths of the scaled queueing model based on $n = 30$.