Stochastic Systems
2011, Vol. 1, No. 2, 340–410
DOI: 10.1214/10-SSY024

NEARLY PERIODIC BEHAVIOR IN THE OVERLOADED $G/D/s+GI$ QUEUE

By Yunan Liu* and Ward Whitt*

Columbia University

Under general conditions, the number of customers in a $GI/D/s+GI$ many-server queue at time t converges to a unique stationary distribution as $t \to \infty$. However, simulations show that the sample paths routinely exhibit nearly periodic behavior over long time intervals when the system is overloaded and s is large, provided that the system does not start in steady state. Moreover, the precise periodic behavior observed depends critically on the initial conditions. We provide insight into the transient behavior by studying the deterministic fluid model, which arises as the many-server heavy-traffic limit. The limiting fluid model also has a unique stationary point, but that stationary point is not approached from any other initial state as $t \to \infty$. Instead, the fluid model performance approaches one of its uncountably many periodic steady states, depending on the initial conditions. Simulation experiments confirm that the time-dependent performance of the stochastic queueing model is well approximated by the fluid model. Like the fluid model, the behavior of the queueing system can be highly sensitive to the initial conditions over long intervals of time.

1. Introduction. This paper continues to investigate the performance of overloaded many-server queueing systems with customer abandonment, extending earlier work in [19–21, 35, 37]; we focus on the special case of deterministic service times. By overloaded, we mean that $\rho > 1$, where ρ is the traffic intensity. With customer abandonment, overloaded systems are practically meaningful because the abandonment acts to keep the system stable. Fluid models can be remarkably effective in determining approximately optimal staffing levels [2].

1.1. *Convergence to steady state in approximating fluid models.* In [37] we showed that the steady-state performance of the overloaded $G/GI/s+GI$ queueing model when s is large is well approximated by the steady-state

Received December 2010.

*Department of Industrial Engineering and Operations Research, Columbia University.

AMS 2000 subject classifications: Primary 60K25; secondary 60F17, 90B22, 37C55.

Keywords and phrases: Many-server queues, overloaded queues, deterministic service times, customer abandonment, heavy traffic, interchanging limits, deterministic fluid approximation, periodic steady state, multiple equilibria, transient behavior.

performance of an associated deterministic $G/GI/s+GI$ fluid model (when the two models are connected by many-server heavy-traffic (MS-HT) scaling; see §2 of [37] and §3 here). Supporting MS-HT limits were established in [15, 16]. In [19], as a special case of a more general fluid model with time-varying parameters, we fully specified that $G/GI/s+GI$ fluid model and described its transient performance. In [21] we showed for the special case of the $G/M/s+GI$ fluid model that the time-dependent performance functions converge to the steady state values as time evolves. It remains to establish convergence to steady state for the $G/GI/s+GI$ fluid model with other service distributions, even though the steady-state performance is available from Theorem 3.1 of [37] and Theorem 6 of [21]. In this paper we show that convergence to steady state in the fluid model does not occur for all service distributions; some conditions are needed.

1.2. *A fluid model with deterministic service times.* We began investigating convergence to steady state for overloaded fluid models with non-exponential service distributions by considering the special case of deterministic service times, even though the deterministic distribution does not satisfy the smoothness conditions imposed on the model elements in [19–21, 37]. We began considering the case of deterministic service times primarily because it is relatively easy to analyze. However, deterministic service times are also of applied interest, because computer-generated service times, such as automated messages, may well be deterministic, and computer-generated service is becoming more prevalent. Many message systems can handle multiple requests in parallel, justifying the many-server model.

We started by considering a specific example: a $G/D/s+M$ fluid model having arrival rate λ, deterministic service times equal to $1/\mu$, service capacity s and an exponential abandonment cdf F with mean $1/\theta$. (The model is specified in detail later in the paper, starting in §4.) We let the other parameters be $\lambda = 2$ and $\mu = s = 1$, making the system overloaded with traffic intensity $\rho \equiv \lambda/s\mu = 2 > 1$, so that the model is overloaded.

Figure 1 shows six performance functions evolving over time for the $G/D/s+M$ fluid model starting empty. The performance functions shown are the total fluid content in service, $B(t)$, the rate that fluid enters service, $b(t,0)$, the departure rate, $\sigma(t)$, the elapsed waiting time for the quantum of fluid at the head of the queue, $w(t)$, the total fluid content waiting in queue, $Q(t)$, and the abandonment rate $\alpha(t)$ over the initial time interval $[0, 3.5]$. There are two plots for the final three performance functions, the solid line for abandonment rate $\theta = 2$ and the dashed line for abandonment rate $\theta = 8$.

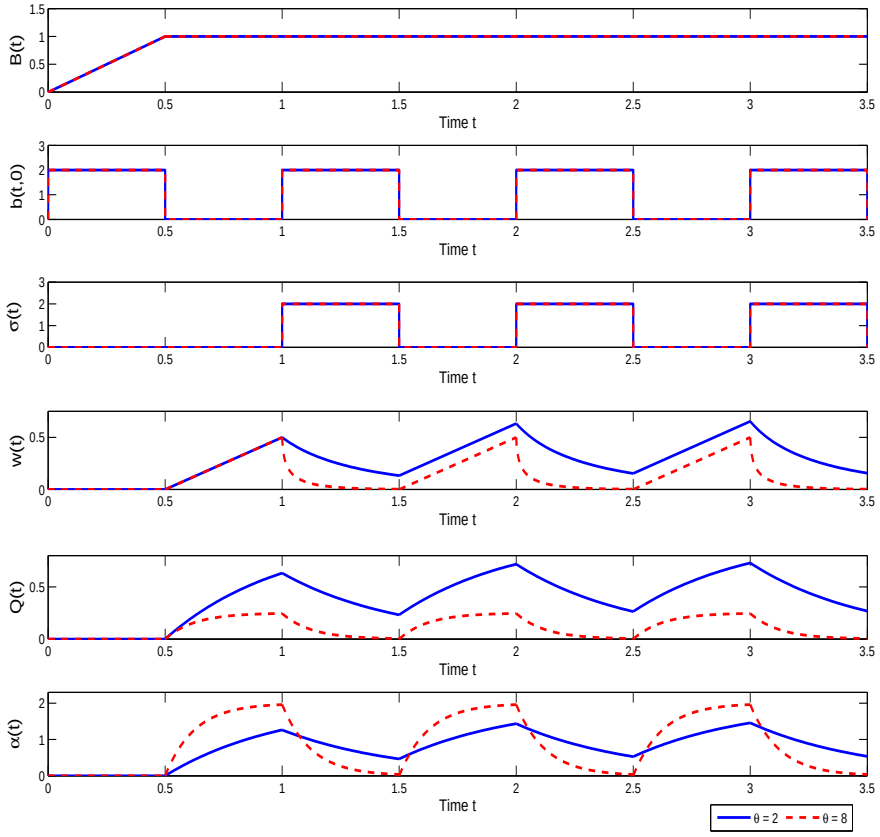

FIG 1. *The $G/D/s+M$ fluid model with $s=\mu=1$, $\lambda=2$.*

We had initially expected to see convergence to the stationary point of this fluid model (which we later show is well defined), because the fluid model is an approximation for the $M/D/s+M$ stochastic model, but instead we see that the performance becomes periodic with period equal to the service-time distribution after time $t=1.0$. At first, we thought that the periodic performance was due to the special choice of the parameters, but that is not the case. Theorem 8.1 shows that the overloaded $G/D/s+GI$ fluid model starting empty exhibits periodic performance after a finite time for all arrival rates λ, service times $1/\mu$ and staffing levels s with $\rho \equiv \lambda/s\mu > 1$, for all abandonment-time cdf's F.

In fact, the functions displayed in Figure 1 are easy to understand. Since the system starts empty and the service capacity is $s=1$, the arriving fluid

flows directly into service at rate $b(t,0) = \lambda = 2$ over the interval $[0, 0.5]$. Hence, the total fluid content in service, $B(t)$ grows linearly at rate 2 over the interval $[0, 0.5]$, reaching the capacity $s = 1$ at time $t = 0.5$, where it stays thereafter. The fluid that entered service in $[0, 0.5]$ completes service exactly $1/\mu = 1$ time units later. Hence there is service completion at rate $\sigma(t) = 2$ over the interval $[1, 1.5]$. Since new fluid cannot enter service until there is free capacity, new fluid enters service only at time 1. Hence, we have $b(t,0) = 0$ during the interval $[0.5, 1]$ and then $b(t,0) = 2$ again in the interval $[1, 1.5]$, which leads to the periodic behavior. Since no arriving fluid can enter service in the interval $[0.5, 1]$, the queue content grows during the interval $[0.5, 1]$. It does not grow linearly because some portion of the fluid entering the queue is lost due to fluid abandonment. For this example, we see that all functions exhibit periodic behavior beginning at time $t = 1$. Explicit expressions for the performance functions for the $G/D/s + M$ fluid model starting empty are given in Corollary 8.3.

1.3. *Simulations of the associated $M/D/s+M$ queueing model.* Having seen how pervasive is this periodic behavior in the fluid model, we were led to seriously doubt the value of the fluid model as an approximation for the stochastic queueing system. For the special case of the $M/D/s + M$ stochastic model, it is evident that the stochastic model has a unique stationary performance and that the performance converges to that stationary performance as time evolves. Indeed, in §2 here we prove that the stochastic process $X \equiv \{X(t) : t \geq 0\}$ representing the number of customers in the more general $GI/D/s + GI$ queueing model is a regenerative stochastic process that converges to a unique stationary distribution as time evolves, provided only that the interarrival-time cdf G is nonlattice, has a finite mean $1/\lambda$ and is unbounded above, while the abandonment-time cdf F has finite mean $1/\theta$.

However, when we conducted simulations of the stochastic $GI/D/s + GI$ model, we found that the sample paths actually agree closely with the deterministic fluid model, exhibiting periodic performance over the horizon of our simulation runs. For example, we simulated a many-server $M/D/s_n + M$ stochastic queueing system with Poisson arrival process approximated by the $G/D/s + M$ fluid model, for which the periodic performance is shown in Figure 1. We obtain the related stochastic model by exploiting MS-HT scaling, i.e., by letting the arrival rate be $\lambda_n \equiv n\lambda = 2n$ and the number of servers be $s_n \equiv \lceil ns \rceil = n$, where $\lceil x \rceil$ is the least integer greater than or equal to x, while leaving the service times and abandonment rate unchanged as $1/\mu = 1$ and θ, respectively. We expect to have a good approximation when n is large.

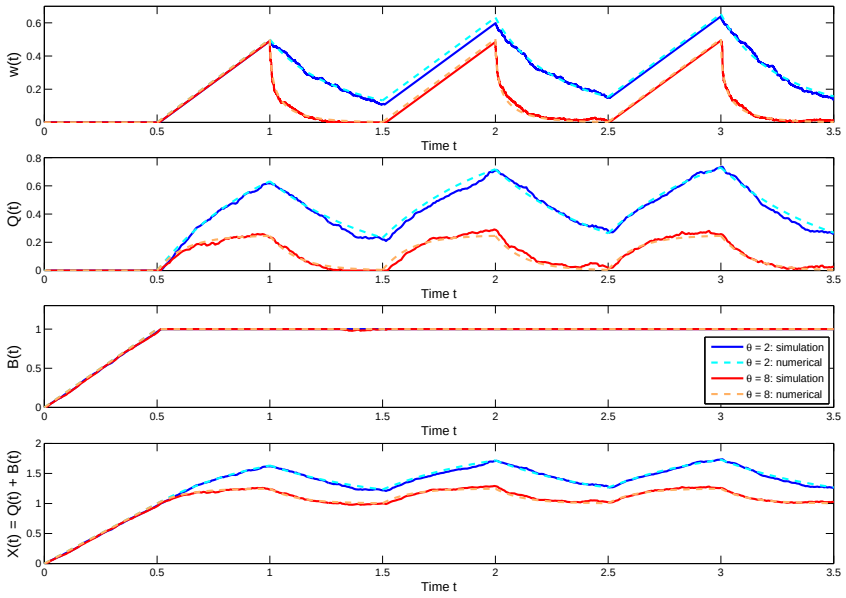

FIG 2. *A comparison of the $G/D/s+M$ fluid model with a simulation (of single sample paths) of the corresponding $M/D/s+M$ stochastic model with $n=1000$.*

Figure 2 compares the fluid approximation (the dashed lines) with simulation estimates (the solid lines) for the large-scale $M/D/s+M$ queueing system with $n=1000$. We plot (i) the elapsed waiting time of the customer at the head of the line $W_n(t)$, (ii) the scaled number of customers waiting in queue $\bar{Q}_n(t) \equiv Q_n(t)/n$ and (iii) the scaled number of customers in service $\bar{B}_n(t) \equiv B_n(t)/n$. We plot single sample paths of these processes. For this large value of n, there is little variability in the simulation sample paths. Each simulated sample path falls right on top of the the approximation. (The two different plots are two different cases of the abandonment rate θ.) Figure 2 shows that the fluid approximation is effective in describing the performance of the stochastic system. The deterministic periodic character is exhibited by the waiting times, which rise linearly at the end of each interval $[k, k+1]$, reaching a peak at the integer endpoint.

However, Figure 2 only compares the performance over a relatively short initial interval of length 3.5, corresponding to 3.5 service times. At first, we thought that we only need to look at a somewhat longer time interval. However, repeated simulations show that the same periodic behavior is seen in the stochastic system over time intervals of length 1000. That is illustrated by Figure 3, which shows simulation estimates of the elapsed waiting $W_n(t)$

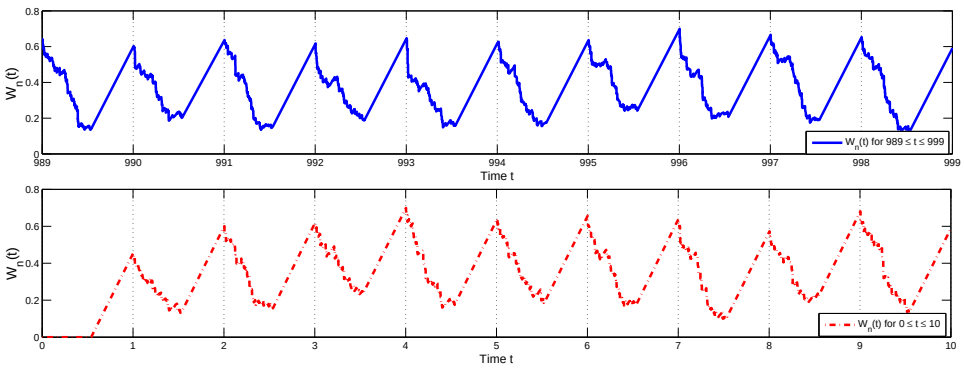

FIG 3. *Large-time periodic behavior of an overloaded $G/D/s+M$ queueing model: simulation estimates of the head-of-line waiting time W_n with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $n = 100$, $T = 1000$.*

for large time $T = 1000$ (instead of small $T = 3.5$ in Figure 2) of the same $M/D/s+M$ model with the same parameters ($\lambda = 2$, $s = \mu = 1$, $\theta = 2$) and initial conditions (initially empty), but with a smaller fluid scaling $n = 100$. The two plots in Figure 3 compare the behavior of a single sample path of $W_n(t)$ at the end ($[989, 999]$, the blue solid curve) and at the beginning ($[0, 10]$, the red dashed curve). Figure 3 shows that the periodic behavior of $W_n(t)$ remains at time 1000 for $n = 100$. (The process $\bar{Q}_n$ behaves the same as W_n.)

Of course, the regenerative theory is not wrong. The stochastic system will eventually approach its stationary distribution if we consider a sufficiently long time. In fact, we do see the periodic pattern broken by 1000 service times in typical simulation sample paths if we decrease the system load ρ and the scale n sufficiently. For example, Figure 8 in the appendix shows that occurs if we replace $\rho = 2$ by $\rho = 1.3$ (by changing λ). By time $T = 1000$, the periodic behavior of W_n is gone.

1.4. *The order of two limits.* In §3 we will establish a MS-HT limit showing that a sequence of scaled stochastic processes indexed by n converges to the deterministic fluid model as $n \to \infty$, under regularity conditions. Since we are considering overloaded models with $\rho > 1$, this is a MS-HT limit for the $G/D/s + GI$ model in the efficiency driven (ED) regime [9], as in [35].

It is customary to apply HT approximations to approximate the steady-state performance of queueing systems. HT approximations for the steady-state performance of queueing processes are supported by results showing

that two iterated limits coincide. For MSHT fluid limits, we want

$$\lim_{t\to\infty}\lim_{n\to\infty} n^{-1}X_n(t) = \lim_{n\to\infty}\lim_{t\to\infty} n^{-1}X_n(t), \tag{1.1}$$

where $X_n(t)$ is a stochastic process or vector of stochastic processes characterizing performance in model n. On the left in (1.1), we have the steady-state (obtained as $t \to \infty$) of the HT limiting process (obtained as $n \to \infty$); on the right, we have the HT limit (obtained as $n \to \infty$) of the steady state (obtained as $t \to \infty$) of the queueing process. Such limit-interchange results have recently been obtained in [8, 12]. For MS-HT approximations, such results were obtained for exponential service times in [9, 13].

Here we do not have that nice state of affairs. Indeed, after establishing the MS-HT limit as $n \to \infty$, we show that the subsequent limit as $t \to \infty$ fails to hold because of the periodicity. Moreover, the form of that periodic behavior depends on the initial conditions. Even the average over a periodic cycle depends on the initial conditions; see Remark 8.3. We will show that the fluid performance is stationary if and only if the fluid model starts in its unique stationary point; see Theorem 9.3.

Here we directly consider only the iterated limit on the left in (1.1), but we can deduce that the two iterated limits do not tell the same story. In §2 we show that there exists regenerative structure implying that the $GI/D/s_n + GI$ stochastic model converges to a steady state as $t \to \infty$ for each n and each finite initial condition. Moreover, we can do so for two-parameter processes that yield a Markov process. For each n, we can then initialize with the stationary distribution of the Markov process, so that we obtain a stationary process (as a function of t) for each n. Now, if we consider the limit of the sequence of scaled stationary distributions as $n \to \infty$, if we obtain convergence, then we necessarily obtain convergence to a stationary process. If such a limit corresponds to the deterministic fluid function, then it necessarily must be the unique stationary point of the fluid model. (We conjecture that the sequence of scaled steady-state queueing processes does indeed converge to the unique stationary point of the fluid model.)

However, a major conclusion from our analysis is that, for the many-server $G/D/s + GI$ stochastic queueing model, we should not focus on the steady-state behavior of the queueing model at all. After much analysis of this kind, we conclude that the periodic phenomenon associated with deterministic service is genuine for the stochastic model as well as the fluid model. Moreover, we conclude that, when there are many servers with deterministic service times and $\rho > 1$, the approximating fluid model is likely to better describe the time-dependent performance of the stochastic system than is the

stationary distribution of the stochastic system. The present paper might better deserve the title of [33].

1.5. *A simple explanation.* In retrospect, we should perhaps have anticipated this nearly periodic behavior of the overloaded $G/D/s+GI$ queueing model. First, when the $G/GI/s+GI$ queueing model is overloaded and s is large, all the servers remain busy for long intervals of time; that is evident from the steady-state performance of the fluid model in [37]. With deterministic service times, when the servers remain busy, the times at which customers complete service and thus enter service in the intervals $[t+(k-1)/\mu, t+k/\mu]$ for integer k will be independent of k. That gives rise to the observed periodic behavior.

1.6. *A simple control.* Once the periodic phenomenon is recognized, it can be controlled if it is considered undesirable. For example, the periodic behavior of an overloaded system starting empty leads to corresponding periodic behavior in the output flow, as illustrated by the plot of $\sigma(t)$ in Figure 1. Such fluctuations in the output may be deemed undesirable. For example, if that output became input at a following queue, then the fluctuations could cause congestion at the subsequent queue.

A simple way to avoid periodic output is to restrict the flow rate into service, allowing flow into service to be at most at rate $s\mu$ at all times. That can be done while still respecting the first-come first-served service discipline. Starting empty, this control imposes extra delay on some of the initial input, but the output rate will soon become constant at $s\mu$.

1.7. *Other models.* There should be broader implications of this work, but one has to be careful about generalizing, because closely related models behave quite differently. In contrast to the overloaded $M/D/s+M$ and $GI/D/s+GI$ models considered here, the associated infinite-server $M/D/\infty$ and $GI/D/\infty$ models are remarkably well behaved, as shown by [11]. Indeed, the number of customers in the $M/D/\infty$ system reaches steady state in finite time, after just one service time. Similarly, the MS-HT fluid and diffusion approximations in the $GI/D/\infty$ model reach steady state after one service time. Having finitely many servers that are busy all the time is an important part of the story in this paper.

Closer to the model we consider is the $G/D/s$ model without customer abandonment in the QED MS-HT regime. For this model, Reed [24] observed that the limiting $G/D/s$ fluid model can exhibit periodic behavior with a special initial condition in his Example 1 at the end of §4, but the implications of that example for the queueing model were not explored. The

$G/D/s$ queueing model is considered further in [26, 27]. There the $G/D/s$ queueing model for large s is identified as an example of a *nearly deterministic queue*. That work establishes MS-HT limits in which the traffic intensity approaches its critical value from below, extending earlier work in [14]. The papers [26, 27] also consider the limiting behavior as $n \to \infty$ in the $G_n/G_n/1$ model in which the interarrival-time and service-time distributions are n-fold convolutions of a given base distribution, generalizing the construction of the Erlang E_k distribution from k-fold convolutions of the exponential distribution. As n increases, the $G_n/G_n/1$ model approaches the $D/D/1$ model. Interesting limiting behavior is obtained by letting the traffic intensity increase as n increases.

Of course, in the stochastic $GI/D/s$ and $GI/D/s+GI$ queueing models, only the service times are directly deterministic; the interarrival-time and abandonment-time distributions may be far from deterministic. However, when n is large and the arrival rate is large, the essential behavior of the arrival process and the abandonment becomes deterministic, primarily because of the law of large numbers (LLN). That can be explained by heavy-traffic limits, such as for non-Markovian infinite-server queues [3, 11, 17, 23, 25]. (If the system is underloaded, then the limits in [11] apply directly.) We elaborate throughout the paper.

Finally, we mention that oscillating behavior and bi-stability have been found in other queueing systems [6, 10, 38]. Another recent example of the invalidity of limit interchange is [28].

1.8. *Organization of the paper.* In §2 we establish the regenerative structure in the $GI/D/s+GI$ stochastic model and show that the mean busy cycle increases rapidly in s. In §3 we establish a MS-HT limit showing that a sequence of the queueing models indexed by the number of servers converges to the fluid model. In §4 we carefully specify the limiting $G/D/s+GI$ fluid model. In §5 we derive the performance formulas for the $G/D/s+GI$ fluid model, some of which are variants of those of the $G_t/GI/s_t+GI$ fluid model developed in [19]. In §6 we focus on the case in which there exists a finite time T^* after which the system remains overloaded (has no idle capacity). In §7 we present key structural properties of the $G/D/s+GI$ fluid queue assuming the queue is overloaded for all $t \geq 0$. In §8 we analyze the periodic steady state of the $G/D/s+GI$ fluid model assuming the queue is overloaded after finite time. In §9 we discuss the asymptotic behavior of the $G/D/s+GI$ fluid queue with general initial conditions. In §10 we present three postponed longer proofs, namely, the proofs for Theorems 2.1, 3.1 and 5.5. Finally, in §11 we draw conclusions. Additional supporting material appears in an appendix available on the authors' web pages.

2. Regenerative structure in the stochastic $GI/D/s+GI$ model. It is well known that a regenerative process $X \equiv \{X(t) : t \geq 0\}$ with sample paths in the function space $\mathbb{D}$ of right-continuous functions with left limits in which a generic cycle T has a distribution that is nonlattice with finite mean has a proper limiting steady-state distribution. In particular, $X(t) \Rightarrow X(\infty)$ as $t \to \infty$, where here and throughout the paper $\Rightarrow$ denotes convergence in distribution, i.e., for any continuous and bounded real-valued function h,

$$E[h(X(t)] \to E[h(X(\infty)] = \frac{E_0[\int_0^T h(X(s))\,ds]}{E[T]} \quad \text{as} \quad t \to \infty, \tag{2.1}$$

where E_0 denotes the expectation conditional on a regeneration point at time 0 and T denotes the end of the first cycle; see Theorem VI.1.2 of [1]. The importance of the sample path regularity was observed in [22]. That regularity condition allows the process to take values in a general Polish topological space [34], but the condition is needed even with the usual real-valued processes. That sample-path regularity is easily seen to be satisfied in our queueing model.

Consider the $GI/D/s + GI$ model, having interarrival times distributed as U with cdf G, deterministic service times of length $1/\mu$ and abandonment times distributed as A with cdf F. Let the interarrival times and abandonment times be mutually independent. Let $X(t)$ represent the number of customers in the $GI/D/s+GI$ system at time t. Let a busy cycle be the interval between successive epochs at which an arrival comes to find an empty system. If the system starts with an arrival to an empty system at time 0, then the first busy cycle begins at time 0. Each busy cycle begins with a busy period and then is followed by an idle period. We prove the following in §10.

THEOREM 2.1. *Consider the stochastic $GI/D/s+GI$ model in which an interarrival time U has a nonlattice cdf G with finite mean $E[U] \equiv 1/\lambda$ and support unbounded above, i.e., $G(x) < 1$ for all $x > 0$, and an abandonment A that has cdf F with finite mean $E[A] \equiv 1/\theta$ and has support unbounded above and below, i.e., $0 < F(x) < 1$ for all $x > 0$. Then the busy cycles for the $GI/D/s + GI$ system constitute an embedded renewal process for the stochastic process X for which a generic busy cycle T has a nonlattice distribution with $E[T] < \infty$, so that the the stochastic process X representing the number of customers in the system has a proper limiting steady-state distribution, as in (2.1), for all proper initial conditions. In addition, the mean $E[T]$ is bounded below by*

$$E[T] \geq \frac{G(1/\mu)}{\bar{G}(1/\mu)} E[U|U \leq 1/\mu] + 1/\mu. \tag{2.2}$$

Theorem 2.1 provides both good news and bad news: The good news is that there exists regenerative structure, so that a proper steady-state distribution for the stochastic process X exists under general conditions. The bad news for large-scale systems (explained below) is that the mean return time to 0 typically grows at least exponentially in s. Of course, that does not directly prove that the process converges to steady state slowly, but it lends support to that notion.

We can formalize this growth in n by considering a limit involving a sequence of models indexed by n. We scale time in the arrival process while changing n to keep the traffic intensity $\rho \equiv \lambda/n\mu$ fixed. The following corollary shows that $E[T^{(n)}]$ is at least $O(e^{cn})$ as $n \to \infty$, where c is some constant with $0 < c < \infty$ when the arrival process is Poisson or in a renewal process when the interarrival-time cdf has an exponential tail.

COROLLARY 2.1. *Consider a sequence of $GI/D/s_n + GI$ models indexed by n satisfying the conditions of Theorem 2.1 with generic interarrival times $U^{(n)} \equiv U^{(1)}/n$, while the service times and abandonment cdf's are independent of n. Then*

$$\liminf_{n\to\infty} \{\lambda n \bar{G}^{(1)}(n/\mu) E[T^{(n)}]\} \geq 1, \tag{2.3}$$

so that $E[T^{(n)}] \to \infty$ as $n \to \infty$. If, in addition, the arrival processes are Poisson with $E[U^{(1)}] = 1/\lambda$, then

$$\liminf_{n\to\infty} \{\lambda n e^{-n\lambda/\mu} E[T^{(n)}]\} \geq 1. \tag{2.4}$$

PROOF. First, as $n \to \infty$, $nE[U^{(n)}|U^{(n)} \leq 1/\mu] = E[U^{(1)}|U^{(1)} \leq n/\mu] \to 1/\lambda$, and $G^{(n)}(1/\mu) \equiv P(U^{(n)} \leq 1/\mu) = G^{(1)}(n/\mu) \to 1$. Also, the first moment condition $E[U^{(1)}] < \infty$ implies that $y\bar{G}^{(1)}(y/\mu) \to 0$ as $y \to \infty$; e.g., see the proof of Lemma 1 on p. 150 of [7]. Therefore, (2.2) in Theorem 2.1 implies (2.3), which in turn implies, first, that $E[T^{(n)}] \to \infty$ as $n \to \infty$ and, second, (2.4). □

The situation is quite intuitive. If indeed n is large and $\rho > 1$, then we will necessarily have $\lambda >> \mu$ and, since it is natural in applications to have θ be the same order as μ, it is natural to also have $\lambda >> \theta$. In that case only rarely will the queue be empty and even more rarely will the entire system be empty, so that the regeneration we are relying on to have a nice steady state is then a rare event.

As noted toward the end of §1, periodic behavior in the $G/D/s + GI$ stochastic model will occur over some time interval whenever *all* servers

remain busy over that time interval. In §6 we provide conditions under which there exists a finite time T^* after which the fluid model remains overloaded (has no idle capacity). We can also conclude that there will be a strictly positive queue. Combined with the MS-HT limit in the next section, we can deduce that, under regularity conditions, there will be long finite intervals over which no server is idle in the queueing model. There is no contradiction with Theorem 2.1; here the limit interchange in (1.1) does not hold.

3. A many-server heavy-traffic limit. In this section we establish a many-server heavy-traffic limit, showing that a sequence of $G/D/s_n + GI$ stochastic queueing models indexed by n converges to the $G/D/s + GI$ fluid model considered in §4 and §5 in the customary many-server heavy-traffic regime, under regularity conditions.

The sequence of models is indexed by the number of servers n. We let the arrival rate in model n be λ_n and the number of servers be s_n, where

$$\bar{\lambda}_n \equiv \frac{\lambda_n}{n} \to \lambda \quad \text{and} \quad \bar{s}_n \equiv \frac{s_n}{n} \to s \quad \text{as} \quad n \to \infty. \tag{3.1}$$

We let the deterministic service times take value $1/\mu$ and the abandonment times have cdf F, independent of n. We assume limits for the arrival process and the initial conditions. In particular, we assume that the sequence of stochastic processes satisfies a *functional weak law of large numbers* (FWLLN). For that purpose, let $\mathbb{D}$ be the usual function space of real-valued functions with limits from the left, endowed with one of the Skorohod topologies, which reduces to uniform convergence on bounded intervals when the limit is a continuous function [34].

Let $B_n(t,x)$ $(\hat{Q}_n(t,x))$ be the number of customers in service (queue) at time t in model n that have been so for a duration less than or equal to x. Since model n has n servers, $0 \le B_n(t,\infty) = B_n(t,1/\mu) \le n$, $n \ge 1$. Let $Q_n(t) \equiv \hat{Q}_n(t,\infty)$ be the total number of customers in queue. Let $A_n(t)$, $S_n(t)$ and $E_n(t)$ be the numbers of customers to abandon, depart after completing service, and enter service, respectively, in $[0,t]$ in model n. In full generality, we will establish a limit for the time-scaled process

$$(\bar{B}_n(t,x), \bar{S}_n(t), \bar{E}_n(t)) \equiv n^{-1}(B_n(t,x), S_n(t), E_n(t)), \tag{3.2}$$

which characterizes the performance of the service facility. Under the additional assumption of exponential abandonment, we will also establish a limit for the time scaled process

$$(\bar{Q}_n(t), \bar{A}_n(t)) \equiv n^{-1}(Q_n(t), A_n(t)). \tag{3.3}$$

Let $N_n(t)$ be the number of arrivals in the interval $[0,t]$ in model n.

ASSUMPTION 1 (FWLLN for the arrival process). *As* $n \to \infty$,

$$n^{-1}N_n \Rightarrow \Lambda \quad \textit{in} \quad \mathbb{D} \quad \textit{as} \quad n \to \infty, \quad \textit{where} \quad \Lambda(t) \equiv \lambda t, \quad t \geq 0, \tag{3.4}$$

for a positive constant λ.

The FWLLN in Assumption 1 is implied by either a functional central limit theorem (FCLT) or a functional strong law of large numbers (FSLLN). Most applications are covered by simple time scaling of a fixed stationary counting process, i.e., when $N_n(t) \equiv N(nt)$, $t \geq 0$, $n \geq 1$. An FSLLN holds for the time-scaled renewal counting process (GI) considered in §2, provided only that the interrenewal time has finite mean $1/\lambda$.

We now make assumptions about the initial conditions. We restrict attention to starting with the queue empty, but we allow customers to start in service, imposing some additional restrictions in the theorem.

ASSUMPTION 2 (an initially empty queue). *For each* $n \geq 1$, $Q_n(0) = 0$.

We also assume a FWLLN for the initial fluid content in service.

ASSUMPTION 3 (FWLLN for the initial conditions). *As* $n \to \infty$,

$$\bar{B}_n(0,\cdot) \Rightarrow B(0,\cdot) \quad \textit{in} \quad \mathbb{D}, \tag{3.5}$$

where

$$B(0,x) \equiv \int_0^x b(0,u)\,du, \quad x \geq 0, \tag{3.6}$$

for a deterministic function $b(0,\cdot)$ *on* $[0,\infty)$ *in* $\mathbb{C}_p$ *with* $b(0,x) \geq 0$ *for all* x *and* $B(0,1/\mu) = B(0,\infty) \leq 1$.

We are now ready to state the many-server heavy-traffic limit. For that purpose, let $\mathbb{D}_{\mathbb{D}}$ be the space of $\mathbb{D}$-valued functions in $\mathbb{D}$, as in [23]. The limit below will be continuous, so the topology on $\mathbb{D}_{\mathbb{D}}$ is equivalent to uniform convergence over the compact sets $[0,t] \times [0,1/\mu]$ for $t > 0$. Let a superscript k on a topological space, as with D^k, indicate the associated k-fold product space, endowed with the product topology.

Let T_n be the first time that all servers are busy in the stochastic queueing model, i.e.,

$$T_n \equiv \inf\{t \geq 0 : B_n(t,1/\mu) = n\}, \quad n \geq 1. \tag{3.7}$$

Let T_n^* be the first time after which all servers remain busy forever, i.e.

$$T_n^* \equiv \inf\{t \geq 0 : B_n(u, 1/\mu) = n \quad \text{for all} \quad u \geq t\}, \tag{3.8}$$

with $T_n^* \equiv \infty$ if there exists no such time. Similarly, let t^* be the time that the limiting fluid model first has no idle service capacity, defined in (6.3), and let T^* be the time after which the limiting fluid model never has any idle capacity, defined in (6.1). The conditions in (3.9) and (3.11) below will imply that the limiting fluid model never has any idle capacity after time t^*, i.e., $T^* = t^* < \infty$; see §6.

THEOREM 3.1 (many-server heavy-traffic FWLLN). *Suppose that Assumptions 1–3 hold with* $\lambda > \mu$,

$$b(0, x) \leq \lambda, \quad 1/\mu - t^* \leq x \leq 1/\mu, \tag{3.9}$$

and, if $t^* > 0$,

$$b(0, 1/\mu - t^*) < \lambda \quad \textit{and} \quad b(0, 1/\mu - t) \quad \textit{continuous at} \quad t = t^*. \tag{3.10}$$

Then

$$(\bar{B}_n, \bar{E}_n, \bar{S}_n) \Rightarrow (B, E, S) \in \mathbb{D}_{\mathbb{D}} \times \mathbb{D}^2, \tag{3.11}$$

where

$$B(t, y) \equiv \int_0^y b(t, x)\, dx, \quad 0 \leq y \leq 1/\mu, \tag{3.12}$$

with $b(t, x)$ *given in* (5.1) *for* $0 \leq t \leq t^*$, b *periodic as a function of its first argument for* $t > t^*$ *with period* $1/\mu$ *and, for* $t \geq t^*$, $b(t - t^*, x)$ *given in* (5.2). *In addition,*

$$S(t) \equiv \int_0^t \sigma(y)\, dy \quad \textit{where} \quad \sigma(k/\mu + t) \equiv b(k/\mu, 1/\mu - t), \quad 0 \leq t \leq 1/\mu, \tag{3.13}$$

for integer k *with* $k \geq 0$,

$$E(t) \equiv \int_0^t b(y, 0) dy \quad \textit{where} \quad b(t, 0) = \lambda\, 1_{\{0 \leq t \leq t^*\}} + \sigma(t)\, 1_{\{t > t^*\}}. \tag{3.14}$$

If $B(0, 1/\mu) < 1$, *then* $T_n \Rightarrow t^* = T^*$ *as* $n \to \infty$. *If, in addition, the abandonment distribution is exponential, i.e., if* $\bar{F}(x) = e^{-\theta x}$, *then*

$$(\bar{Q}_n, \bar{A}_n) \Rightarrow (Q, A) \in \mathbb{D}^2, \tag{3.15}$$

where $Q(t) = A(t) = 0$ *for* $0 \le t \le t^*$ *and*

$$Q(t) = \int_0^{t-t^*} \bar{F}(t - t^* - s)\gamma(s)\, ds, \tag{3.16}$$

$$= \int_0^{w(t)} \lambda\, \bar{F}(x) dx, \quad t \ge t^*, \tag{3.17}$$

$$A(t) = \Lambda(t) - \int_0^{t-t^*} b(s,0)\, ds - Q(t), \quad t \ge t^*, \tag{3.18}$$

where w *satisfies ODE* (5.7) *with* $w(t^*) = 0$, $\gamma(t) \equiv \lambda - b(t,0)$.

We now observe that in general we need not have either $T_n \Rightarrow t^*$ or $T_n^* \Rightarrow T^*$.

EXAMPLE 3.1 (counterexample on first passage times). Suppose that $\lambda > \mu = 1$. Let $b(0,x) = \lambda$, $1 - (1/\lambda) \le t \le 1$, and $b(0,x) = 0$, $0 \le x < 1 - (1/\lambda)$, so that $b(t,0) = \lambda$, $0 \le t < 1/\lambda$, and $b(t,0) = 0$, $1/\lambda \le t < 1$, $B(t, 1/\mu) = 1$ for all $t \ge 0$ and $T^* = t^* = 0$.

For $n \ge 1$, let $\{B_n(0,y) : 0 \le y \le 1\}$ be deterministic. To be a legitimate sample path for a queueing system, $B_n(0,y)$ must be nondecreasing and integer-valued as well as satisfy $0 \le B_n(0,y) \le n$. Thus, let $B_n(0,y) \equiv \lfloor B_n^f(0,y) \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $\bar{B}_n^f(0,y) \equiv n^{-1} B_n^f(0,y) \equiv \int_0^y b_n(0,x)\, dx$, where $b_n(0,x) = ((n+1)/n)\lambda$, $1 - ((n-1)/n\lambda) \le t \le 1$, and $b_n(0,x) = 0$, $0 \le x < 1 - ((n-1)/n\lambda)$. First, observe that $\bar{B}_n^f(0, 1/\mu) = (n^2 - 1)/n^2 < 1$ for all $n \ge 1$. Second, observe that we have $0 \le \bar{B}_n^f(0,y) - \bar{B}_n(0,y) \le 1/n$ for all y and n. Hence, $\bar{B}_n(0, 1/\mu) \le \bar{B}_n^f(0, 1/\mu) < 1$ for all $n \ge 1$. Nevertheless, $\bar{B}_n(0,\cdot) \to B(0,\cdot)$ as $n \to \infty$. On the other hand, consider a deterministic arrival process with rate $n\lambda$, i.e., with $N_n(t) \equiv \lfloor n\lambda t \rfloor$, $t \ge 0$, $n \ge 1$. Then $S_n(t) = \lfloor (n+1)\lambda t \rfloor \ge N_n(t)$ for $0 \le t \le (n-1)/n\lambda$. Since $B_n(0, 1/\mu) < n$, the system is underloaded for $0 \le t < 1/\lambda$. However, $N_n(1/\lambda) = n$. Hence, $T_n = T_n^* = 1/\lambda$ for all $n \ge 1$, in contrast to $t^* = T^* = 0$. A similar example can be constructed if $B(0, 1/\mu) < 1$ and condition (3.10) is not imposed; see Appendix H.

4. The $G/D/s + GI$ fluid queue. We now study the $G/D/s + GI$ fluid queue. The corresponding $G_t/GI/s_t + GI$ and $G_t/M_t/s_t + GI_t$ models, having time-varying arrival rate (G_t), time-varying staffing (s_t) and a smooth general service-time distribution (GI) or time-varying Markov service (M_t) were studied in [19–21]. Here we restrict attention to constant arrival rate λ and constant staffing s, although the model can easily be extended to

allow these functions to be time-varying, as in [19–21]. In the $G_t/GI/s_t + GI$ model, the service distribution was assumed to have a pdf g; in the $G_t/M_t/s_t + GI_t$ model, the pdf was time-varying exponential, i.e., $g_t(x) = \mu(t+x)\, e^{\int_t^{t+x} \mu(y)dy}$. Hence, strictly speaking, the $G/D/s + GI$ fluid queue considered here was not considered before.

Fluid is a deterministic divisible quantity that arrives over time. Fluid input flows directly into a service facility with fixed capacity s if there is free capacity available; otherwise it flows into the queue. The total fluid input over an interval $[0,t]$ is $\Lambda(t) = \lambda t$, where λ is a positive constant.

System performance will be 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 of at most y, for $t \geq 0$ and $y \geq 0$. These functions will be absolutely continuous in the second parameter, so that

$$\hat{B}(t,y) \equiv \int_0^y b(t,x)\, dx \quad \text{and} \quad \hat{Q}(t,y) \equiv \int_0^y q(t,x)\, dx, \tag{4.1}$$

for $t \geq 0$ and $y \geq 0$. We will be characterizing performance primarily 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.

The system has unlimited waiting room and the FCFS service discipline. Whenever $Q(t) > 0$, we require that there be no free capacity in service; whenever $B(t) < s$, we require that the queue be empty. These requirements are both covered by the following.

Assumption 4 (fluid dynamics constraints, FDC's). *For all* $t \geq 0$,

$$(B(t) - s)Q(t) = 0 \quad \textit{and} \quad B(t) \leq s. \tag{4.2}$$

Because the service time is deterministic, each quantum of fluid that enters service stays in service for time $1/\mu$ before leaving the system. The total service completion rate at time t is the density of fluid that has been in service for $1/\mu$. That is also the rate into service $1/\mu$ time units before, i.e.,

$$\sigma(t) \equiv b(t, 1/\mu) = b(t - 1/\mu, 0), \quad t \geq 0. \tag{4.3}$$

The model allows for abandonment of fluid waiting in the queue. In particular, a proportion $F(x)$ of any fluid to enter the queue will abandon before

waiting x time units in queue it has not yet entered service, where F is an absolutely continuous cumulative distribution function (cdf), with

$$F(x) = \int_0^x f(y)\,dy, \quad x \geq 0, \quad \text{and} \quad \bar{F}(x) \equiv 1 - F(x), \quad x \geq 0. \tag{4.4}$$

Let $h_F(y) \equiv f(y)/\bar{F}(y)$ be the hazard rate associated with the patience (abandonment) cdf F.

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

$$A(t) \equiv \int_0^t \alpha(y)\,dy, \quad t \geq 0, \tag{4.5}$$

where $\alpha(t)$ is the abandonment rate at time t. Since $q(t, x)$ is the density of fluid in queue at time t that arrived at time $t - x$, the abandonment rate at time t is

$$\alpha(t) \equiv \int_0^\infty q(t, y)h_F(y)\,dy, \quad t \geq 0, \tag{4.6}$$

where $h_F(y)$ is the hazard rate associated with the patience cdf F.

Let $E(t)$ be the amount of fluid to enter service in $[0, t]$; then

$$E(t) \equiv \int_0^t b(u, 0)\,du, \quad t \geq 0, \tag{4.7}$$

where $b(t, 0)$ is the rate fluid enters service at time t. The rate fluid enters service depends on whether the system is underloaded or overloaded. If the system is underloaded, then the external input directly enters service; if the system is overloaded, then the fluid to enter service is determined by the rate that service capacity becomes available at time t, which is the departure rate $\sigma(t)$, because the total fluid content in service $B(t) = s$ does not change at t.

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 (4.1), 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 $q(0, \cdot)$ by

$$w(0) \equiv \inf\{x > 0 : q(0, y) = 0 \quad \text{for all} \quad y > x\}. \tag{4.8}$$

Assumption 5 (finite initial values). *$B(0) < \infty$, $Q(0) < \infty$, $w(0) < \infty$, $b(0, x) < \infty$ and $q(0, x) < \infty$ for all $x \geq 0$.*

In summary, the six-tuple $(\lambda, s, \mu, F., b(0, \cdot), q(0, \cdot))$ specifies the *model data*.

To describe waiting times, let the *boundary waiting time* (BWT) $w(t)$ be the delay experienced by the quantum of fluid at the head of the queue at time t 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. Informally, as in (4.8),

$$w(t) \equiv \inf\{x > 0 : q(t, y) = 0 \quad \text{for all} \quad y > x\}. \tag{4.9}$$

A proper definition of q, w and v is somewhat complicated, but that has already been done in §7 of [19]; we review in the next subsection.

Since the service discipline is FCFS, fluid leaves the queue to enter service from the right boundary of $q(t, x)$. The fluid content densities q and b satisfy the following two fundamental evolution equations. (Recall that the service-time ccdf is $\bar{G}(x) = 1_{\{0 \le x \le 1/\mu\}}$.)

ASSUMPTION 6 (fundamental evolution equations). *For $t \ge 0$, $x \ge 0$ and $u \ge 0$,*

$$q(t + u, x + u) = q(t, x)\frac{\bar{F}(x + u)}{\bar{F}(x)}, \quad 0 \le x < w(t), \tag{4.10}$$

$$b(t + u, x + u) = b(t, x)\frac{\bar{G}(x + u)}{\bar{G}(x)} = b(t, x)\, 1_{\{x+u \le 1/\mu\}}. \tag{4.11}$$

In addition, we impose regularity conditions on the model data. Some we impose now, to be in force throughout the paper, but others we impose as needed. As in [19, 20], we develop 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. Hence, $\mathbb{C}_p \subset \mathbb{D}$, where $\mathbb{D}$ is the usual function space of right continuous functions with left limits; see [34].

ASSUMPTION 7 (smoothness). *$f, b(0, \cdot), q(0, \cdot)$ in $\mathbb{C}_p$ for each $x \ge 0$ and t.*

As in §7.2 of [19], we need to impose a regularity condition on the arrival rate function and the initial queue density in order to treat the BWT w. Here and later we use the notation $\uparrow$ and $\downarrow$ to denote supremum and infimum, respectively, e.g.,

$$q^{\uparrow}(0, x) \equiv \sup_{0 \le u \le x}\{q(0, u)\} \quad \text{and} \quad q^{\downarrow}(0, x) \equiv \inf_{0 \le u \le x}\{q(0, u)\}. \tag{4.12}$$

ASSUMPTION 8 (positive arrival rate and initial queue density). *For all* $x \geq 0$, $\lambda > 0$ *and* $q^{\downarrow}(0,x) > 0$.

As in [19], we introduce bounds for the pdf f. Let

$$f^{\uparrow} \equiv \sup\{f(x) : x \geq 0\}. \tag{4.13}$$

ASSUMPTION 9 (controlling the abandonment). $f^{\uparrow} < \infty$, *where* $f^{\uparrow}$ *is defined in* (4.13), *and* $\bar{F}(x) > 0$ *for all* $x > 0$.

We assume that all assumptions in this section are in force throughout the paper.

5. Performance of the $G/D/s + GI$ fluid queue. In [19, 20] we showed how the system performance expressed via the basic functions (b,q,w,v) depends on the model data $(\lambda, s, \mu, F, b(0,\cdot), q(0,\cdot))$, for the time-varying fluid models, i.e., for $G_t/GI/s_t + GI$ and $G_t/M_t/s_t + GI_t$. From the basic performance four-tuple (b,q,w,v), we easily compute the associated vector of performance functions $(\hat{B}, \hat{Q}, B, Q, X, \sigma, S, \alpha, A, E)$ via the definitions in §4. We now establish similar results for the basic functions (b,q,w,v) of the $G/D/s + GI$ model.

The service content density b is elementary within each interval that the system is either entirely underloaded or entirely overloaded. The complications occur when there are changes from one regime to the other. We state basic results in this section and others in the next section. The results here provide the basis for an effective algorithm, assuming that there are only finitely many changes between underloaded and overloaded regimes in each interval $[0,T]$, for which we give a sufficient condition at the end of this section.

THEOREM 5.1 (service content in the underloaded case). *For the* $G/D/s+GI$ *fluid model with unlimited service capacity* ($s \equiv \infty$), *starting at time* 0,

$$\begin{aligned} b(t,x) &= b(0,x-t)\cdot 1_{\{0\leq t<x\leq 1/\mu\}} + \lambda\cdot 1_{\{0\leq x\leq 1/\mu,\, x\leq t\}}, \\ B(t) &= \left(\lambda t + \int_t^{1/\mu} b(0,x-t)\,dx\right) 1_{\{0\leq t\leq \frac{1}{\mu}\}} + \frac{\lambda}{\mu} 1_{\{t>\frac{1}{\mu}\}}. \end{aligned} \tag{5.1}$$

If, instead, a finite-capacity system starts underloaded, then the same formulas apply over the interval $[0,T)$, *where* $T \equiv \inf\{t \geq 0 : B(t) > s\}$, *with* $T = \infty$ *if the infimum is never obtained. Hence,* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* $t \geq 0$ *and* $x \geq 0$, *for* t *in the underloaded interval.*

PROOF. To show the first relation, note that $b(t,x) = 0$ for all $x > 1/\mu$ because the service time is exactly $1/\mu$. If $0 \le t \le 1/\mu$, $b(t,x) = b(0, x-t)$ for $t < x \le 1/\mu$ and $b(t,x) = \lambda$ for $0 \le x \le t$. If $t > 1/\mu$, then all fluid that was in service at time 0 is gone, hence $b(t,x) = \lambda$ if $0 \le x \le 1/\mu$. Simply integrating the first relation gives the second. □

COROLLARY 5.1 (reaches steady state at time $1/\mu$). *If the system is entirely underloaded, then the performance reaches steady state by time* $1/\mu$ *with* $\sigma(t) = b(t,x) = \lambda$, $0 \le x \le 1/\mu$ *and* $t \ge 1/\mu$.

The periodic behavior observed in the overloaded numerical examples is mostly explained by the following theorem and the subsequent Corollary 5.2.

THEOREM 5.2 (service content in the overloaded case). *For the* $G/D/s+GI$ *fluid model in an overloaded interval,* $B(t) = s$ *and*

$$
\begin{aligned}
b(t,x) \quad &= \quad b(0, x-t) \cdot 1_{\{0 \le t < x \le 1/\mu\}} \\
&\quad + b\left(0, \frac{1}{\mu} - (t-x) + \frac{\lfloor (t-x)\mu \rfloor}{\mu}\right) \cdot 1_{\{0 \le x \le 1/\mu,\, x \le t\}},
\end{aligned} \tag{5.2}
$$

where $\lfloor x \rfloor$ *is the integer part of a real number* x. *Hence,* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* $t \ge 0$ *and* $x \ge 0$ *in an overloaded interval.*

PROOF. Note $b(t,x) = 0$ for all $x > 1/\mu$. If $0 \le t \le 1/\mu$, $b(t,x) = b(0, x-t)$ for $t < x \le 1/\mu$; $b(t,x) = b(t-x, 0) = \sigma(t-x) = b(0, 1/\mu - (t-x))$ for $0 \le x \le t$. If $t > 1/\mu$, then $t - x > 0$. Let $N \equiv \lfloor (t-x)\mu \rfloor$, we have $0 \le t - x - N/\mu \le 1/\mu$. Hence $b(t,x) = b(t-x, 0) = \sigma(t-x) = \sigma(t - x - N/\mu) = b(0, 1/\mu - (t - x - N/\mu))$. Moreover, simple calculation by integrating (5.2) over x verifies that indeed $B(t) = \int_0^{1/\mu} b(t,x)dx = s$. □

COROLLARY 5.2 (periodic performance in service starts at time 0). *If* $B(t) = s$ *for all* $t \ge 0$, *then the density* b *is either stationary or in a PSS starting at time* 0. *It is stationary if* $b(0,x) = s\mu$, $0 \le x \le 1/\mu$. *Otherwise it is in a PSS with*

$$
b\left(\frac{k}{\mu} + t, x\right) \quad = \quad b(t,x), \quad \sigma\left(\frac{k}{\mu} + t\right) = \sigma(t),
$$

for $0 \le x \le 1/\mu$, $0 \le t \le 1/\mu$ *and* $k \ge 0$.

COROLLARY 5.3 (overall smoothness for the service content). *If the system changes regimes only finitely often in the interval* $[0,T]$, *then* $b(t,\cdot)$, $b(\cdot,x)$, $B \in \mathbb{C}_p$ *for all* t, $0 \le t \le T$, *and* $x \ge 0$.

The $G/D/s+GI$ model differs from the $G_t/GI/s_t+GI$ model in [19] in the service facility, but not in the queue. Therefore, the dynamics of q, w and v are the same. We next review these results from [19]. Let $\tilde{q}(t,x)$ be $q(t,x)$ during the overload interval $[0,T]$ under the assumption that no fluid enters service from queue.

PROPOSITION 5.1 (queue content without transfer into service in the overloaded case [19]). *During an overloaded interval,*

$$\tilde{q}(t,x) = \lambda\bar{F}(x)1_{\{x\le t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\}}. \tag{5.3}$$

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

COROLLARY 5.4 (from $\tilde{q}$ to q [19]). *Given the BWT w in an overloaded interval,*

$$\begin{aligned} q(t,x) &= \tilde{q}(t-x,0)\bar{F}(x)1_{\{x\le w(t)\wedge t\}} + \tilde{q}(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\le w(t)\}} \\ &= \lambda\bar{F}(x)1_{\{x\le w(t)\wedge t\}} + q(0,x-t)\frac{\bar{F}(x)}{\bar{F}(x-t)}1_{\{t<x\le w(t)\}}. \end{aligned} \tag{5.4}$$

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

We define the BWT w by stipulating that two expressions for the amount of fluid to enter service over any interval $[t,t+\delta]$, namely,

$$E(t+\delta)-E(t) \equiv \int_t^{t+\delta} b(u,0)\,du = I(t,\delta) - A(t,t+\delta), \tag{5.5}$$

where $I\equiv I(t,\delta)$ is the amount of fluid removed from the right boundary of $\tilde{q}$ during the time interval $[t,t+\delta]$ and $A(t,t+\delta)$ is the amount of the fluid content in I that abandons in the interval $[t,t+\delta]$. We then show that, if (5.5) holds, then w satisfies an ODE.

THEOREM 5.3 (the BWT ODE [19]). *Consider an overloaded interval $[0,T)$. The BWT w is well defined by the relation (5.5), being Lipschitz continuous on $[0,T]$ with $w(t+u)\le w(t)+u$ for all $t\ge 0$ and $u\ge 0$ with $t+u\le T$. Moreover, w is right differentiable everywhere with right derivative*

$$w'(t+) = \Psi(t,w(t)) \equiv 1 - \frac{\gamma(t+)}{\tilde{q}(t,w(t)-)}, \tag{5.6}$$

where $\gamma(t) \equiv b(t,0)$, $t \geq 0$, and left differentiable everywhere (but not necessarily differentiable) with value

$$w'(t-) \quad = \quad \tilde{\Psi}(t, w(t)) \equiv 1 - \frac{\gamma(t-)}{\tilde{q}(t, w(t)+)}. \tag{5.7}$$

Overall, w is continuously differentiable everywhere except for finitely many t. The BWT w is characterized as the unique solution of the initial value problem (IVP) on $[0,T)$ based on the ODE (5.6) and any initial value $w(0)$.

COROLLARY 5.5 (end of the overloaded interval [19]). *We can compute the end of an overloaded interval as $T \equiv \inf\{t \geq 0 : w(t) = 0$ and $\lambda(t) \leq \gamma(t)\}$.*

COROLLARY 5.6 (smoothness of $q(t,\cdot)$ [19]). *Under the assumptions of Theorem 5.3, q is given by (5.4) with $q(\cdot, x) \in \mathbb{C}_p$ for all x. (We have already deduced that $q(t,\cdot) \in \mathbb{C}_p$ for all t in Corollary 5.4.)*

THEOREM 5.4 (v and w [19]). *Consider an overloaded interval. Then v is finite and v is the unique function in $\mathbb{D}$ satisfying the equation*

$$v(t - w(t)) = w(t) \quad \text{or, equivalently,} \quad v(t) = w(t + v(t)) \tag{5.8}$$

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

We now provide a sufficient condition for there to be only finitely many switches between overloaded and underloaded intervals in any bounded interval $[0,T]$. To do so, we use a function involving the model elements λ and $b(0,x)$, $0 \leq x \leq 1/\mu$. In particular, let

$$\zeta(x) \equiv \sigma(x) - \lambda = b(0, 1/\mu - x) - \lambda.$$

Let $\mathcal{D}_\zeta$ be the set of discontinuities of ζ in $[0,1/\mu]$, let $\bar{\mathcal{Z}}_\zeta \equiv \{x \in [0,1/\mu] : \zeta(x) = 0\}$ be the zero set of ζ, and let $\mathcal{Z}_\zeta$, be a subset of $\bar{\mathcal{Z}}_\zeta$, defined by

$$\mathcal{Z}_\zeta \equiv \{x \in \bar{\mathcal{Z}}_\zeta : \nexists\, \epsilon > 0 \text{ such that } \zeta(y) = 0 \text{ for all } y \in (x - \epsilon, x + \epsilon)\}$$

The subset $\mathcal{Z}_\zeta$ excludes those points $x \in [0,1/\mu]$ such that $\zeta(x) = 0$ for $x \in (a,b)$.

Let $\mathcal{S}_T$ be the total number of regime-switching (between overloaded and underloaded) points in $[0,T]$ as in [19, 20]. For any set A, let $|A|$ be the cardinality of A.

THEOREM 5.5 (relating switches to zeros and discontinuities of ζ). *For any interval* $[0,T]$ *with* $T \geq 1/\mu$,

$$|\mathcal{S}_T| \leq \lceil T\mu \rceil (|\mathcal{Z}_\zeta| + |\mathcal{D}_\zeta| + 1), \tag{5.9}$$

where $\lceil x \rceil$ *is least integer greater than or equal to* x.

REMARK 5.1 (tightness of the bound in Theorem 5.5). To show that the bound in Theorem 5.5 is tight, consider a $G/D/s + GI$ fluid queue in $[0,T] = [0, 2/3\mu]$ that is initially critically loaded, i.e., $B(0) = s$ and $Q(0) = 0$, with $b(0,x) = 2\mu\, s \cdot 1_{\{1/2\mu \leq x \leq 2/3\mu\}}$ and $\lambda = 1.5\mu\, s$. We know $\sigma(t) = b(0, 1/\mu - t) = 2\mu\, s \cdot 1_{\{0 \leq t \leq 1/2\mu\}}$. Hence, $B'(t) = \lambda - \sigma(t) = -0.5\mu\, s \cdot 1_{\{0 \leq t \leq 1/2\mu\}} + 1.5\mu\, s \cdot 1_{\{1/2\mu \leq t \leq 2/3\mu\}}$, which implies that $B(t) = (s - 0.5\mu\, s\, t) \cdot 1_{\{0 \leq t \leq 1/2\mu\}} + 1.5\mu\, s\, t \cdot 1_{\{1/2\mu \leq t \leq 2/3\mu\}}$. Therefore the system is underloaded in $[0, 2/3\mu]$ and becomes critically loaded again at $t = 2/3\mu$. In this case the bound in Theorem 5.5 is tight because $N = \lfloor 2/3 \rfloor + 1 = 1$, $|\mathcal{D}_\zeta| = 1$, $|\mathcal{Z}_\zeta| = 0$ and $|\mathcal{S}_T| = 2$, where the two switching points are 0 and $2/3\mu$.

ASSUMPTION 10 (controlling the number of switches). *For* $\mu > 0$, $|\mathcal{Z}_\zeta| < \infty$, *so that there are only finitely many switches between overloaded and overloaded intervals in any bounded subinterval.*

We assume that Assumption 10 is in force throughout the paper.

REMARK 5.2 (an algorithm). These results yield an efficient algorithm to compute the basic performance four tuple (b, q, w, v). First, we can compute $b(t,x)$ directly via Theorems 5.1 and 5.2. We compute $\tilde{q}$ directly from Proposition 5.1. We then compute the BWT w by solving the ODE in Theorem 5.3. The proof of Theorem 5.4 in [19] provides an elementary algorithm to compute v once w has been computed. Theorem 6 of [19] shows that v satisfies its own ODE under additional regularity conditions. Theorem 5.1 and Corollary 5.5 specify how to switch between alternating overloaded and underloaded intervals. Assumption 10 ensures that the total number of switches between underloaded and overloaded intervals is finite.

6. The fluid model eventually always overloaded. For the rest of this paper, we assume that the fluid arrival rate λ exceeds the maximum possible long-run average service rate $s\mu$, so that $\rho \equiv \lambda/s\mu > 1$.

ASSUMPTION 11 ($\rho > 1$). $\lambda > s\mu$.

We say that the service capacity (and thus the system) is overloaded at time t if $B(t) = s$. In this section we describe the fluid density in service, b, in the $G/D/n + GI$ fluid model assuming that there exists a finite time after which the system stays overloaded; let T^* be the first such time, i.e.,

$$T^* \equiv \inf\{t \geq 0 : B(u) = s \quad \text{for all} \quad u \geq t\}, \tag{6.1}$$

with $T^* \equiv \infty$ if there exists no such time.

We also provide a sufficient condition for T^* to be finite. We show that the service density b reaches a PSS at time T^*. In the next two sections we use this assumption to show that the queue performance (e.g. $Q(t)$ and $\alpha(t)$) converges to a PSS after time T^*. (These auxiliary performance functions typically do not reach PSS in finite time.)

ASSUMPTION 12 (a time after which the system remains overloaded). *For T^* defined in (6.1), $T^* < \infty$.*

Assumption 12 is very useful because it identifies the time at which the service fluid density b reaches a PSS. The following is a consequence of Theorem 5.2 and Corollary 5.2.

COROLLARY 6.1 (a PSS for b starting at T^*). *Under Assumption 12, the service fluid density b either reaches steady state or a PSS at time T^*; i.e.,*

$$b((n/\mu) + t, x) = b(t, x), \quad n \geq 1, \quad t \geq T^*, \quad 0 \leq x \leq 1/\mu.$$

A steady state is achieved if and only if $b(T^*, x) = s\mu$, $0 \leq x \leq 1/\mu$.

In applications it is not necessary to identify T^*; it suffices to identify *any* time t with $t \geq T^*$. Corollary 6.1 implies that b is in a PSS starting at any time $t \geq T^*$. We now provide a sufficient condition for Assumption 12. To do so, let t^* be the time that the service facility *first* becomes full; i.e.,

$$t^* \equiv \inf\left\{t \geq 0 : \lambda t + B(0) - \int_0^t \sigma(x)dx = s\right\}. \tag{6.2}$$

If the system is initially overloaded, then $t^* = 0$. Necessarily $t^* < 1/\mu$, because no new input during the interval $[0, 1/\mu]$ can depart in that interval and $\lambda/\mu > s$, since $\rho \equiv \lambda/s\mu > 1$. Define a class of initial service densities

$$\mathcal{B}^*_{s,\lambda} \equiv \left\{b(0,\cdot) : B(0) = \int_0^{1/\mu} b(0,x) = s,\, b(0, x - t^*) \leq \lambda,\, t^* \leq x \leq 1/\mu\right\}.$$

THEOREM 6.1 (a sufficient condition for Assumption 12). *If* $b(0,\cdot) \in \mathcal{B}^*_{s,\lambda}$, *then Assumption 12 is satisfied with* $T^* = t^*$ *for* T^* *in* (6.1) *and* t^* *in* (6.2).

PROOF. If $t^* = 0$, i.e., $B(0) = s$ and $b(0,x) \le \lambda$, $0 \le x \le 1/\mu$, then new fluid will arrive in the system at least as fast as the fluid is departing, throughout the interval $[0, 1/\mu]$. Hence, a full service facility is maintained throughout the interval $[0, 1/\mu]$. Hence fluid enters service immediately replacing all departing fluid. (This fluid will enter from the head of the queue if the queue is not empty, but that is not important for b.) Thus, the service facility remains full forever.

If $t^* > 0$, then $B(0) < s$, so that new fluid will enter service from outside at rate λ until the service facility becomes full at t^*. We have

$$t^* = \inf\{t \ge 0 : \lambda t + B(0, 1/\mu - t) = s\}, \tag{6.3}$$

following from (6.2) and Theorem 5.1. Since $b(0,x) \le \lambda$ for $t^* \le x \le 1/\mu$, the system then reaches the first case starting at t^*, so we can apply the previous analysis to this case. □

Note that the condition of Theorem 6.1 is satisfied in the common case in which the system starts out empty. In §8 we will describe the system performance in detail in that special case. Also note that we can apply Theorem 6.1 to the state of the system at any finite time t, not just at time 0. In particular, we can apply the algorithm in Remark 5.2 over some finite interval $[0,t]$ and then check to see if the conditions of Theorem 6.1 are satisfied at time t.

7. Structural results for the queue performance. In this section we focus on the performance related to the queue in an overloaded $G/D/s+GI$ fluid model with $\rho > 1$, thus showing how we can exploit Assumptions 11 and 12 in the previous section. In this section we assume that the fluid queue is overloaded for all $t \ge 0$. We present four structural results: (i) comparison, (ii) Lipschitz continuity, (iii) asymptotic loss of memory (ALOM) and (iv) uniform boundedness. We omit the proofs of Theorem 7.1–7.4 below because these results follow directly from the proofs of Theorems 3-5 and Lemma 1 of [21]. (The statements of Theorems 3-5 and Lemma 1 of [21] do not directly imply the statements of Theorem 7.1–7.4 here, because the service-time distribution was assumed to have a finite density in [21], but the proofs apply without change once we have determined the density b. Detailed proofs of Theorems 7.1–7.4 are also given in Appendix C.)

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

THEOREM 7.1 (comparison of fluid content in queue for the overloaded $G/D/s+GI$ model). *Consider two $G/D/s+GI$ fluid models with common staffing function s, service time $1/\mu$, abandonment cdf F and initial fluid density in service $b(0,\cdot)$. Assume both queues are overloaded for all $t \geq 0$ $(B_1(t) = B_2(t) = s)$. If $q_1(0,\cdot) \leq q_2(0,\cdot)$ and $\lambda_1 \leq \lambda_2$, then*

$$(Q_1, q_1, \alpha_1, w_1, v_1) \leq (Q_2, q_2, \alpha_2, w_2, v_2).$$

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

$$b^\downarrow \equiv \inf_{0\leq x\leq 1/\mu} b(0,x), \quad b^\uparrow \equiv \sup_{0\leq x\leq 1/\mu} b(0,x),$$
$$h_F^\downarrow \equiv \inf_{0\leq x<\infty} h_F(x), \quad h_F^\uparrow \equiv \sup_{0\leq x<\infty} h_F(x).$$

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

THEOREM 7.2 (Lipschitz continuity of fluid content in queue for the overloaded $G/D/s+GI$ model). *Consider a $G/D/s+GI$ fluid model with arrival rate λ, staffing function s, service time $1/\mu$, abandonment cdf F. Assume the queue is overloaded for all $t \geq 0$. Then the function mapping $(\lambda, Q(0))$ in $\mathbb{R}^2$ into (Q,α) in $\mathcal{C}_p^2$ all over $[0,T]$ is Lipschitz continuous. In particular,*

$$\begin{aligned} \|Q_1 - Q_2\|_T &\leq T|\lambda_1 - \lambda_2| + |Q_1(0) - Q_2(0)| \\ &\leq (1 \vee T)(|\lambda_1 - \lambda_2| \vee |Q_1(0) - Q_2(0)|), \qquad (7.1) \\ \|\alpha_1 - \alpha_2\|_T &\leq h_F^\uparrow \|Q_1 - Q_2\|_T, \qquad (7.2) \\ \|q_1 - q_2\|_{T,1} &\equiv \left\| \int_0^\infty q_1(\cdot,x)dx - \int_0^\infty q_2(\cdot,x)dx \right\|_T \\ &\leq T|\lambda_1 - \lambda_2| + \|q_1(0,\cdot) - q_2(0,\cdot)\|_1. \qquad (7.3) \end{aligned}$$

THEOREM 7.3 (ALOM of fluid content in queue for the overloaded $G/D/s+GI$ model). *Consider two initially overloaded $G/D/s+GI$ fluid models $(B_1(0) = B_2(0) = s)$. Suppose these two models have common arrival rate λ, staffing function s, service time $1/\mu$, abandonment cdf F, initial fluid densities in service $b(0,x)$, but different initial fluid densities in queue $q_i(0,\cdot)$.*

(*a*) *If both queues are overloaded for all* $t \geq 0$, *then*

$$\begin{aligned} \Delta Q(T) &= \|q_1(T,\cdot) - q_2(T,\cdot)\|_1 \leq C_1 e^{-h_F^{\downarrow} T}, \\ \Delta\alpha(T) &\leq h_F^{\uparrow} C_1 \, e^{-h_F^{\downarrow} T}, \end{aligned} \tag{7.4}$$

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

$$\begin{aligned} C_1 &\equiv \int_0^{\infty} ([q_1(0,x) \vee q_2(0,x)] - [q_1(0,x) \wedge q_2(0,x)])dx \\ &\leq Q_1(0) + Q_2(0). \end{aligned} \tag{7.5}$$

In addition, if $b^{\downarrow} > 0$, *then for* $T > T^*$,

$$\begin{aligned} \Delta w(T) &\leq \frac{\Delta Q(T)}{\lambda \bar{F}(w_2(T) \vee w_1(T))} \\ &\leq C_2 \Delta Q(t) \leq (C_2 \, C_1) e^{-h_F^{\downarrow} T}, \end{aligned} \tag{7.6}$$

where

$$\begin{aligned} T^* &\equiv \frac{Q_1(0) + Q_2(0)}{b^{\downarrow}}, \\ C_2 &\equiv \bar{F}\left[\frac{b^{\downarrow}}{\lambda} \vee \left(w_1(0) \vee w_2(0) + \frac{Q_1(0) + Q_2(0)}{b^{\downarrow}}\right)\right]^{-1}. \end{aligned} \tag{7.7}$$

(*b*) *If, in addition, the initial densities in queue are ordered by*

$$q_1(0,x) \leq q_2(0,x) \quad \textit{for all } x \geq 0, \tag{7.8}$$

then $Q_1(t) \leq Q_2(t)$ *for all* $t \geq 0$,

$$\Delta Q'(T) \leq 0 \quad \textit{and} \quad \Delta Q(T) \leq \frac{\Delta Q(0)}{1 + h_F^{\downarrow} T}, \quad T > 0, \tag{7.9}$$

so that

$$\Delta Q(T) \leq e^{-h_F^{\downarrow} T} \, \Delta Q(0), \quad \Delta\alpha(T) \leq h_F^{\downarrow} \Delta Q(T). \tag{7.10}$$

For the following boundedness result, we make a stronger assumption on the initial fluid density and the abandonment hazard rate in the model data, requiring that they be uniformly bounded above and below.

ASSUMPTION 13 (uniformly bounded initial fluid density and hazard rate). *The staffing and the rates in the model data are uniformly bounded above and below, i.e.,*

$$0 < b^{\downarrow} \le b^{\uparrow} < \infty, \quad 0 < h_F^{\downarrow} \le h_F^{\uparrow} < \infty.$$

Assumption 13 strengthens Assumptions 5 and 9. We assume that this additional assumption is in force for the remainder of the paper.

THEOREM 7.4 (boundedness). *Consider the $G/D/s + GI$ fluid queue that is overloaded for all $t \ge 0$. Under Assumption 13 and the previous the assumptions, all performance functions are uniformly bounded. In particular,*

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

8. The full performance under Assumption 12. In §6 we saw that the fluid density in service, b, reaches steady state or a PSS at time T^* if the system remains overloaded after time T^*, as stipulated in Assumption 12. We now exploit the structural results in the previous section to describe the full queue performance, given Assumption 12. In the next section we show that Assumption 12 is not always satisfied.

As in §7 of [21], let the performance vector at time t be

$$\mathcal{P}(t) \equiv (\{b(t,x) : x \ge 0\}, \{q(t,x) : x \ge 0\}, B(t), Q(t), w(t), v(t), \sigma(t), \alpha(t)).$$

If the initial condition $\mathcal{P}(0)$ can be chosen so that $\{\mathcal{P}(t) : t \ge 0\}$ is a periodic function of t with period τ, then this initial condition produces a PSS. If not, we want to show that the performance converges to a PSS $\mathcal{P}^*$ as time evolves. We say a function 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 $t \in [0, \tau]$. The limit can be viewed as an application of the time-shift operator Ψ_τ on the function g, i.e., $\Psi_\tau(g)(t) \equiv g(\tau + t)$. 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 $n \to \infty$. To discuss continuity and convergence in the domain of $\mathcal{P}$, we use norm

$$\begin{aligned} \|\mathcal{P}(t)\| &\equiv \sup_{t\ge 0}\{|\mathcal{P}(t)|\}, \quad \text{where} \\ |\mathcal{P}(t)| &\equiv |B(t)| + |Q(t)| + |\alpha(t)| + |\sigma(t)| + |w(t)| + |v(t)| \\ &\quad + \left|\int_0^{1/\mu} b(t,x)dx\right| + \left|\int_0^{\infty} q(t,x)dx\right|. \end{aligned} \tag{8.1}$$

We primarily want to establish convergence to a PSS, but we also treat the case of stationary performance, which arises when $b(T^*, x) = s\mu$, $0 \le x \le 1/\mu$. Given that stationary b, the remaining stationary performance can be obtained by the reasoning in Theorem 6 of [21]. The remaining stationary performance measures are

$$\begin{aligned} B &= s, \quad \alpha = \lambda - s\mu, \quad w = \bar{F}^{-1}(s\mu/\lambda), \\ Q &= \lambda \int_0^w \bar{F}(x)\,dx, \quad \text{and} \quad q(x) = \lambda\,\bar{F}(x), \quad 0 \le x \le w. \end{aligned} \tag{8.2}$$

THEOREM 8.1 (PSS for the overloaded $G/D/s + GI$ fluid model). *Suppose that Assumption 12 is satisfied in the $G/D/s + GI$ fluid model with $\rho > 1$. If $b(T^*, x) = s\mu$, $0 \le x \le 1/\mu$, then there exists a constant function $\mathcal{P}^*$ as in (8.2) such that*

$$\|\Psi_\tau^{(n)}(\mathcal{P}) - \mathcal{P}^*\| \to 0 \quad \text{as} \quad n \to \infty. \tag{8.3}$$

for all $\tau > 0$. Otherwise, the fluid performance $\mathcal{P}$ is asymptotically periodic with period $1/\mu$, i.e., there exists a periodic function $\mathcal{P}^*$ with period $1/\mu$ such that (8.3) holds for $\tau \equiv 1/\mu$.

PROOF. We can treat the two cases together by the same argument; we only discuss the second case. We must show that $\|\mathcal{P}((n/\mu)+\cdot) - \mathcal{P}^*(\cdot)\| \to 0$ as $n \to \infty$. However, since $\mathcal{P}^*$ is periodic and $\Psi_{1/\mu}^{(n)}(\mathcal{P})$ involves the shift operator, it suffices to prove that $\|\mathcal{P}((n/\mu)+\cdot) - \mathcal{P}^*(\cdot)\|_{1/\mu} \to 0$ as $n \to \infty$, where the supremum in the norm is over the finite interval $[0, 1/\mu]$, i.e., for $\|\mathcal{P}\|_{1/\mu} \equiv \sup\{|\mathcal{P}(t)| : 0 \le t \le 1/\mu\}$. That in turn is a form of the norm in Theorem 7.2.

If $T^* > 0$, we can simply move the origin to T^*. Therefore, it remains to consider the case where the system is initially overloaded, and remains so thereafter. In that case, $b(t, x)$ and $\sigma(t) = b(t, 0)$ are periodic with period $1/\mu$ starting from $t = 0$, by Theorem 5.2 and Corollary 5.2.

Next, suppose that $q(0,x) = 0$ for $x \geq 0$, i.e., the system is initially critically loaded. By Theorem 7.1, the shift operator $\Psi_{1/\mu}$ is a monotone operator on $\mathcal{P}((n/\mu)+\cdot)$ for any n, because we can think of the performance $q(1/\mu,\cdot)$ as alternative initial conditions for the model at time 0, since the model is periodic with period $1/\mu$ (λ and s are constant, $b(t,0)$ is periodic with period $1/\mu$ by Theorem 5.2 and Corollary 5.2. Therefore, the sequence of system performance functions $\mathcal{P}(0+\cdot), \mathcal{P}((1/\mu)+\cdot), \mathcal{P}((2/\mu)+\cdot), \ldots$ (at discrete time $0, 1/\mu, 2/\mu, \ldots$) is monotonically non-decreasing. Since the performance is also bounded, by Theorem 7.4, there is a finite limit for the sequence $\{\mathcal{P}((n/\mu)+\cdot)\}$ as $n \to \infty$. By Theorem 7.2, the operator is continuous as well, which implies that $\Psi^{(n)}_{1/\mu}(\mathcal{P})$ is convergent in the specified norm as $n \to \infty$. Hence the limit is a PSS. By the ALOM property in Theorem 7.3, we get the same limit for all other initial fluid densities in queue $q(0,\cdot)$. □

REMARK 8.1 (computation). Given the rapid convergence, it usually is not difficult to compute the PSS associated with any given initial condition by simply applying the algorithm with that initial condition. We can then verify that the condition in Theorem 6.1 is satisfied after some finite time, so that we know T^* and we know the PSS for the fluid density in service b. We then can observe the convergence of the other performance measures. However, the PSS for the remaining performance functions can also be determined in another way, given T^* and b. First, if the abandonment distribution is exponential, then analytic expressions are available, see Corollary 8.3. Second, for the case of non-exponential abandonment, consider a cycle $[0, 1/\mu]$ of the PSS. For each candidate $\tilde{w} \geq 0$, we numerically solve the ODE (5.6) in $[0, 1/\mu]$ with $w(0) = \tilde{w}$ and $b(t,0) = b(T^*, 1/\mu - t)$ and check if $w(1/\mu) = \tilde{w}$. Since $\tilde{w} \geq 0$ is our only unknown variable, we shall do a search for $\tilde{w} \geq 0$. Theorem 8.1 guarantees the existence and uniqueness of such a $\tilde{w} \geq 0$.

REMARK 8.2 (different initial conditions). Theorems 6.1 and 8.1 provide sufficient conditions for Assumption 12 to hold, and for the performance function to converge to a PSS. That PSS depends strongly on the fluid density in service, b at the time T^* after which the system remains overloaded. In Appendix D we show that very different PSS's can result by considering two different initial conditions for the example in §1.

We now describe the time-average performance over a periodic cycle. Some average performance measures are independent of the initial conditions, and thus agree with the stationary performance, whereas others are not.

COROLLARY 8.1 (average performance over a cycle). *Suppose that Assumption 12 holds for a $G/D/s + GI$ fluid queue and consider the PSS beginning at T^*. The average abandonment rate $\bar{\alpha}$ and departure rate $\bar{\sigma}$ over a cycle $[0,\tau] \equiv [0, 1/\mu]$ of the PSS are*

$$\bar{\alpha} \equiv \frac{1}{\tau}\int_0^\tau \alpha(t)dt \quad = \quad \alpha^* \equiv \lambda - \mu\, s \tag{8.4}$$

$$\bar{\sigma} \equiv \frac{1}{\tau}\int_0^\tau \sigma(t)dt \quad = \quad \sigma^* \equiv \mu\, s, \tag{8.5}$$

If, in addition, the abandonment distribution is exponential, then

$$\bar{Q} \equiv \frac{1}{\tau}\int_0^\tau Q(t)dt \quad = \quad Q^* \equiv \int_0^{w^*} \lambda\, e^{-\theta\, x}\, dx. \tag{8.6}$$

where α^*, σ^*, Q^* and $w^* \equiv \bar{F}^{-1}(1/\rho)$ are the stationary abandonment and departure rates, queue length and BWT given in (8.2).

PROOF. First, (8.6) follows from (8.4) when $\bar{F}(x) = e^{-\theta\, x}$, because $\alpha(t) = \theta\, Q(t)$, which implies

$$\bar{Q} = \frac{1}{\theta}\bar{\alpha} = \frac{1}{\theta}(\lambda - \mu\, s),$$

which is equal to the right hand side of (8.6), as can be verified by simple calculation. Since the system is overloaded for all $t \geq T^*$, then $b(t,x)$ and $\sigma(t)$ are periodic for all $t \geq T^*$, by Theorem 5.2 and Corollary 5.2. Therefore, consider a cycle $[0, 1/\mu]$ of the PSS, we must have $b(t,0) = \sigma(t) = b(T', 1/\mu - t)$ for some $T' \geq T^*$. Hence, (8.5) follows because $\int_0^{1/\mu} b(T', 1/\mu - t)dt = B(T') = s$.

To show (8.4), flow conservation of the queue implies that

$$Q'(t) = \lambda - \alpha(t) - b(t,0) = \lambda - \alpha(t) - \sigma(t), \quad \text{for } 0 \leq t \leq 1/\mu.$$

Integrating both sides from 0 to $1/\mu$ yields that

$$0 = Q(1/\mu) - Q(0) = \lambda\,\tau - \int_0^\tau \alpha(t)dt - \int_0^\tau \sigma(t)dt = \lambda\,\tau - \int_0^\tau \alpha(t)dt - \mu\, s\,\tau,$$

which implies (8.4). □

REMARK 8.3 (average of other performance functions). Except for $\bar{\alpha}$ and $\bar{\sigma}$, the average of other performance functions in PSS typically does not agree with the corresponding stationary values. We illustrate with an example in

Appendix E, considering Erlang and hyperexponential abandonment cdf's. In our numerical examples we found that the average BWT $\bar{w}$ is consistently greater than the stationary value w^*. In contrast the average $\bar{Q}$ is greater (less) than or equal to the stationary value Q^* when the abandonment-time cdf F is more (less) variable than exponential. It remains to establish supporting theorems.

A common case occurs when the system is initially empty. Obviously this initial condition belongs to class $\mathcal{B}^*_{s,\lambda}$. We next establish results for this special case.

COROLLARY 8.2 (PSS for the initially empty $G/D/s+GI$ fluid model). *Consider the $G/D/s+GI$ fluid model with $\rho > 1$. If the system is initially empty, then the performance $\mathcal{P}$ is asymptotically periodic and converges to a unique PSS $\mathcal{P}^*$ with period $\tau = 1/\mu$. In particular, $B(t) = s$, $b(t,x)$ and $\sigma(t)$ are periodic after s/λ,*

$$\begin{aligned} b(t+k/\mu, x) &= \begin{cases} \lambda \cdot 1_{\{0\le x\le t-1/\mu+s/\lambda\}\cup\{t\le x\le 1/\mu\}}, & \text{if } \frac{s}{\lambda} < t \le \frac{1}{\mu}, \\ \lambda \cdot 1_{\{t\le x\le t+s/\lambda\}}, & \text{if } \frac{1}{\mu} < t \le \frac{1}{\mu}+\frac{s}{\lambda}. \end{cases} \\ \sigma(t+k/\mu) &= b(t+k/\mu, 0) = \lambda\, 1_{\{1/\mu<t\le 1/\mu+s/\lambda\}}, \quad \text{for } k \ge 0. \end{aligned}$$

Performance functions in queue converge to a PSS with the following structure:

$$\begin{aligned} q(t+k/\mu, x) &\to \lambda\, \bar{F}(x) \cdot 1_{\{0\le x\le w^*(t)\}}, \\ Q(t+k/\mu) &\to \int_0^{w^*(t)} \lambda\, \bar{F}(x)dx, \\ \alpha(t+k/\mu) &\to \int_0^{w^*(t)} \lambda\, f(x)dx, \\ w(t+k/\mu) &\to w^*(t), \quad \text{as } k \to \infty, \end{aligned} \tag{8.7}$$

where $w^*(t) = \tilde{w}+t$ (linear) for $s/\lambda \le t \le 1/\mu$ for some $\tilde{w} \ge 0$; $w^*(t)$ solves ODE $w'(t) = 1-1/\bar{F}(w(t))$ for $1/\mu \le t \le 1/\mu+s/\lambda$ with $w(s/\lambda+1/\mu) = \tilde{w}$.

PROOF. Since the system is initially empty, it becomes overloaded at time $t^* = s/\lambda < 1/\mu$ and stays overloaded for all $t \ge t^*$ by Theorem 6.1. Hence, the formulas for b follow from Theorem 5.2 and Corollary 5.2. The convergence of other performance functions follows from (8.7). Therefore, it remains to show (8.7). Since $\sigma(t) = b(t,0) = 0$ for $(k-1)/\mu+s/\lambda < t \le k/\mu$, the BWT ODE (5.6) in Theorem 5.3 implies that $w'(t) = 1$ so that $w(t)$ is linear with slope 1 for $(k-1)/\mu + s/\lambda < t \le k/\mu$. □

We now give explicit expressions for the PSS of the $G/D/s+M$ fluid queue that has exponential abandonment and is initially empty. We give the proof in Appendix F.

COROLLARY 8.3 (explicit expression for the PSS of the $G/D/s+M$ fluid queue starting empty). *Consider the $G/D/s+M$ fluid queue starting out empty, with arrival rate λ, service time $1/\mu$, staffing s, exponential abandonment with rate θ and $\rho \equiv \lambda/s\mu > 1$. The system becomes overloaded and remains so at time $t^* = T^* = s/\lambda$. In the PSS (starting at time 0) the system is overloaded with performance functions given in two parts ($[0, 1/\mu - s/\lambda]$ and $(1/\mu - s/\lambda, 1/\mu]$) of a cycle $0 \le t \le 1/\mu$:*

(a) *In the first part of the PSS cycle, for $0 \le t \le 1/\mu - s/\lambda$,*

$$w(t) = t + \tilde{w}, \tag{8.8}$$

$$Q(t) = \frac{\lambda}{\theta}\left[1 - \left(\frac{1 - e^{-\theta\, s/\lambda}}{1 - e^{-\theta/\mu}}\right) e^{-\theta\, t}\right], \tag{8.9}$$

$$b(t,x) = \lambda \cdot 1_{\{t \le x \le t + s/\lambda\}},$$

$$\sigma(t) = b(t,0) = 0,$$

where

$$\tilde{w} \equiv w(0) = w(1/\mu) = \frac{1}{\theta}\log\left(\frac{1 - e^{-\theta/\mu}}{1 - e^{-\theta\, s/\lambda}}\right) \ge 0. \tag{8.10}$$

(b) *In the second part of the PSS cycle, for $1/\mu - s/\lambda < t \le 1/\mu$,*

$$w(t) = -\frac{1}{\theta}\log\left(1 + \left(\frac{1 - e^{\theta(1/\mu - s/\lambda)}}{1 - e^{-\theta/\mu}}\right) \cdot e^{-\theta\, t}\right), \tag{8.11}$$

$$Q(t) = \frac{\lambda}{\theta}\left(\frac{e^{\theta(1/\mu - s/\lambda)} - 1}{1 - e^{-\theta/\mu}}\right) e^{-\theta\, t}, \tag{8.12}$$

$$b(t,x) = \lambda \cdot 1_{\{0 \le x \le t - 1/\mu + s/\lambda\} \cup \{t \le x \le 1/\mu\}},$$

$$\sigma(t) = b(t,0) = \lambda.$$

In addition, for $0 \le t \le 1/\mu$,

$$B(t) = s, \quad q(t,x) = \lambda\, \bar{F}(x) \cdot 1_{\{0 \le x \le w(t)\}}, \quad \alpha(t) = \theta\, Q(t),$$

(c) *If we consider a cycle $[1/\mu - \tilde{w}, 2/\mu - \tilde{w}]$, then the PWT*

$$v(t) = \frac{1}{\theta}\log\left(1 + \left(e^{\theta/\mu}\frac{e^{\theta(1/\mu - s/\lambda)} - 1}{1 - e^{-\theta/\mu}}\right) \cdot e^{-\theta\, t}\right), \tag{8.13}$$

for $1/\mu - \tilde{w} \le t < 2/\mu - \tilde{w}$ *and* v *jumps at* $2/\mu - \tilde{w}$ *to*

$$v(2/\mu - \tilde{w}) = v(1/\mu - \tilde{w}) = \tilde{w} + 1/\mu - s/\lambda.$$

REMARK 8.4. Since we have an explicit expression for $Q(t)$, in which it is an exponential function in both (a) and (b), simple calculation directly verifies (8.6) in Corollary 8.1.

9. General initial conditions. In §7 and §8, we provided a quite complete description of system performance if there exists a finite time T^* such that the system is overloaded for all $t \ge T^*$. Moreover, Theorem 6.1 provides widely applicable conditions for the time T^* to coincide with t^*, the first time t that $B(t) = s$, which necessarily is less than or equal to $1/\mu$. More generally, Theorem 6.1 can be applied to show that the time T^* exists subsequently after applying the numerical algorithm to compute the performance over an initial interval, because we can check to see if the conditions in Theorem 6.1 hold after some finite time.

Nevertheless, we now show that in general there need not exist a finite time such that the system remains overloaded thereafter, i.e., T^* can be ∞. We have seen that the system necessarily becomes overloaded for a first time t^* with $t^* < 1/\mu$. However, with $\rho > 1$, it is possible for the the system to switch between overloaded and underloaded regimes infinitely often.

THEOREM 9.1. *There need not exist a finite time* T^* *such that* $B(t) = s$ *for all* $t \ge T^*$.

PROOF. We provide an explicit counterexample. We consider a $G/D/s+M$ fluid queue with $\lambda = 1.2$, $\mu = s = 1$, $\theta = 2$. Let the queue be initially overloaded with

$$\begin{aligned} b(0,x) &= 2 \cdot 1_{\{1/2 \le x \le 1\}} \quad \text{so that } B(0) = s = 1, \\ w(0) &= 2 \quad \text{and} \quad q(0,x) = \lambda\, e^{-\theta\, x} \cdot 1_{\{0 \le x \le w(0)\}} = 2\, e^{-2\, x} \cdot 1_{\{0 \le x \le 2\}}. \end{aligned}$$

We can apply mathematical induction to show that $B(n) = s$ and $B(n + 1/2) < B(n + 3/2) < s$ for all $n \ge 1$. We elaborate in Appendix G. □

REMARK 9.1 (The influence of $q(0,x)$). It is important to note that the initial queue fluid density $q(0,\cdot)$ plays an important role, both in the counterexample above and in the system performance more generally. For $t \ge T^*$, $q(t,\cdot)$ plays only a minor role, because then we have ALOM for the queue performance, by virtue of Theorem 7.3. However, the initial queue fluid density $q(0,\cdot)$ plays an important role in determining if $T^* < \infty$ and the

form of the PSS. In §G we consider the above example with the same initial fluid density in service but different initial fluid in queue ($w(0) = 0.2$ instead of $w(0) = 2$). There we show that this different value for $w(0)$ (initial fluid in queue) completely changes both the transient evolution of performance functions and the structure of the PSS.

We now obtain additional results for general initial conditions. To do so, let $\Lambda^{(n)}$ be the set of time points at which the rate of fluid entering service is equal to the arrival rate in the n^{th} cycle $[(n-1)/\mu, n/\mu]$, i.e.,

$$\Lambda^{(n)} \equiv \{t \in [0, 1/\mu] : b(t + (n-1)/\mu, 0) = \lambda\}. \tag{9.1}$$

For the example in the proof of Theorem 9.1, $\Lambda^{(n)} = [t_1^{(n)}, t_2^{(n)}]$ (see Appendix G). Since $t_1^{(n)}$ is strictly decreasing and $t_2^{(n)}$ is strictly increasing, we have $\Lambda^{(n)} \subseteq \Lambda^{(n+1)}$. In general $\Lambda^{(n)}$ may not be a single closed interval as in this case, nevertheless the monotonicity still holds in general.

THEOREM 9.2 (monotone convergence of the sets $\Lambda^{(n)}$).

(*a*) *The sequence* $\{\Lambda^{(n)} : n \geq 1\}$ *is monotonically increasing, i.e.,*

$$\Lambda^{(n)} \subseteq \Lambda^{(n+1)} \quad \textit{for all } n \geq 1.$$

(*b*) *The sequence* $\{\Lambda^{(n)} : n \geq 1\}$ *converges to a bounded set, i.e.,*

$$\cup_{n=1}^{\infty} \Lambda^{(n)} \equiv \Lambda^{\infty} \subseteq [0, 1/\mu].$$

PROOF. The convergence in (b) directly follows from (a) because $\Lambda^{(n)} \subseteq [0, 1/\mu]$ and is thus bounded for all $n \geq 1$. To show (a), consider any $t \in \Lambda^{(n)}$, we have $b(t + (n-1)/\mu, 0) = \lambda$, which implies that $\sigma(t + n/\mu) = b(t + (n-1)/\mu, 0) = \lambda$. If the system is overloaded at time $t + n/\mu$, then $b(t + n/\mu, 0) = \sigma(t + n/\mu) = \lambda$ by flow conservation of fluid in service; if the system is underloaded at time $t + n/\mu$, then we again have $b(t + n/\mu, 0) = \lambda$ because external arrival flows into service directly. Therefore, $b(t + n/\mu, 0) = \lambda$ implies that $t \in \Lambda^{(n+1)}$. □

We now show that convergence to the stationary point of the fluid density in service occurs *only if* the initial fluid density is that stationary point.

THEOREM 9.3 (convergence to the unique stationary point). *The only initial fluid density in service* $b(0, \cdot)$ *for which* $b(t, x) \to b^*(x) \equiv s\mu$, $0 \leq x \leq 1/\mu$, *as* $t \to \infty$ *is the stationary point* b^* *itself.*

PROOF. First the conclusion is clearly true whenever $B(t) = s$ for all $t \geq 0$, because the density $b((n/\mu), x) = b(0, x)$, $0 \leq x \leq 1/\mu$ for all $n \geq 1$. We shall show that for any $b(0, x)$ that is different from the steady state, i.e., $\max_{0 \leq x \leq 1/\mu} |b(0, x) - \mu s| > 0$, there exists a $0 \leq t \leq 1/\mu$ such that $b(t + n/\mu, 0) \neq \mu s$ for all $n \geq 0$ so that $b(t + n/\mu, 0) \nrightarrow \mu s$. In this case there must exist a $0 \leq t \leq 1/\mu$ such that $\mu s \neq b(0, t) = b(1/\mu - t, 0)$. If the system is overloaded at time $n/\mu - t$ for all $n \geq 1$, then $b(n/\mu - t, 0) = b(1/\mu - t, 0) \neq \mu s$ for all $n \geq 1$, by Theorem 5.2 and Corollary 5.2. If the system is underloaded at time $n'/\mu - t$ for some $n' \geq 1$, then we must have $b(n'/\mu - t, 0) = \lambda$, which implies that $b(n/\mu - t, 0) = \lambda$ for all $n \geq n'$, following from Theorem 9.2 (because set $\Lambda^{(n)}$ is increasing). Therefore, we conclude $b(n/\mu - t, 0) \nrightarrow \mu s$ as $n \to \infty$. In particular, $|b(n/\mu - t, 0) - \mu s| \geq |b(0, t) - \mu s| \wedge (\lambda - \mu s)$. □

We now establish convergence of $b(t, \cdot)$ to a PSS for general initial conditions.

THEOREM 9.4 (PSS in service). *Consider the $G/D/s + GI$ fluid queue with arbitrary initial condition $b(0, \cdot)$. For $0 \leq t \leq 1/\mu$, as $n \to \infty$,*

$$\begin{aligned} b(t + n/\mu, 0) &\to b^\infty(t, 0) \equiv \lambda \cdot 1_{\{t \in \Lambda^\infty\}} + b(0, 1 - t) \cdot 1_{\{t \notin \Lambda^\infty\}}, \\ b(t + n/\mu, x) &\to b^\infty(t - x, 0) \cdot 1_{\{0 \leq x \leq t\}} + b^\infty(t - x + 1/\mu, 0) \cdot 1_{\{t < x \leq 1/\mu\}}, \\ \sigma(t + n/\mu) &\to b^\infty(t, 0). \end{aligned}$$

PROOF. First, it is easy to see that the third relation follows from the second(letting $x = 1/\mu$) and the second follows from the first. To establish the first relation, consider $0 \leq t \leq 1/\mu$. If the system is overloaded at $t + n/\mu, 0$ for all $n \geq 0$, then $b(t + n/\mu, 0) = b(0, 1 - t)$ for all $n \geq 0$ and thus converges to $b(0, 1 - t)$ as $n \to \infty$, following from Theorem 5.2 and Corollary 5.2. If the system is underloaded at $t + n'/\mu, 0$ for some $n' \geq 0$, then $b(t + n'/\mu, 0) = \lambda$, which implies $b(t + n/\mu, 0) = \lambda$ for all $n \geq n'$, by Theorem 9.2. □

We now show that the system is fully overloaded in each PSS, even if the PSS is only approached in the limit. For the proof, define the sets in which the system is overloaded (including critically loaded) and underloaded in a cycle of the PSS as

$$\mathcal{O}^\infty \equiv \{0 \leq t \leq 1/\mu : B(t) = s\} \quad \text{and} \quad \mathcal{U}^\infty \equiv \{0 \leq t \leq 1/\mu : B(t) < s\}.$$

THEOREM 9.5 (overloaded in each PSS). *Each PSS for the $G/D/s + GI$ fluid model is overloaded everywhere, i.e., in a cycle $[0, 1/\mu]$, $\mathcal{O} = [0, 1/\mu]$ and $\mathcal{U} = \phi$.*

PROOF. First, it is easy to see that $\mathcal{O}$ cannot be $\emptyset$, because $\rho > 1$. Suppose there exists a $0 \le t \le 1/\mu$ such that the system is underloaded at t, then there must exists a switching time $0 \le t' \le 1/\mu$ at which the system switches from overloaded to underloaded regime, which implies that $b(t,0) = \lambda < \sigma(t)$. This will make $\sigma(t + 1/\mu) = b(t,0) = \lambda \ne \sigma(t)$. Hence, this contradicts with our assumption that the system is initially in PSS. □

10. Proofs. In this section we present three postponed longer proofs.

Proof of Theorem 2.1. The busy cycle is a random sum of i.i.d. interarrival times, and so necessarily has a nonlattice distribution because the interarrival time cdf is nonlattice; see Proposition X.3.2 of [1]. Hence it suffices to focus on the mean busy cycle. We stochastically bound a busy cycle of the $GI/D/n + GI$ system above and below by quantities that are easier to analyze.

We start with the upper bound. For the upper bound, we use a coupling construction to produce sample-path stochastic order, as in [1, 18, 32]. We construct both systems on a common probability space so that the sample paths are ordered w.p.1 while each process separately has its own proper distribution. We give both systems the same arrival process (the same sample paths). For the upper bound, let $Y(t)$ be the number of customers in the queue of the associated system in which no servers are working. The stochastic process Y behaves as the number in system in a $GI/GI/\infty$ model with interarrival-time cdf G and service-time cdf F (our abandonment cdf). Then $n + Y$ is our candidate sample path upper bound for X. Start both X and Y with an arrival to an empty system at time 0. Continue the sample path construction by assigning all customers that enter the queue in the original "X model" abandonment times equal to the service times assigned to the corresponding arrival in the bounding "Y model," both according to cdf F. As a consequence, whenever a customer completes service in the bounding Y model, the matching customer in the original X model customer will either have entered service or abandoned in the original X model. Hence the sample-path order is maintained. Since the abandonment times are i.i.d., this assignment rule does not alter the distribution of the processes.

The key now is to observe that the busy cycles in both the X model and the Y model (not counting the n) will end after one more interarrival time beyond the beginning of a busy cycle of the Y process if the interarrival-time and service-time pair (U, A) at the beginning of the Y busy cycle satisfies $U > 2/\mu > A$, which is an event, say C, with positive probability

$$p \equiv P(C) \equiv P(U > 2/\mu > A) = P(U > 2/\mu)P(A < 2/\mu) > 0, \tag{10.1}$$

by the assumptions $G(x) < 1$ and $F(x) > 0$ for all x. In addition, $p < 1$ since $P(A < 2/\mu) < 1$ because we have assumed that $F(x) < 1$ for all x. For the Y model, given the event C, the one customer in the system at the start of the busy cycle will depart at time A, which is less than the time of the next arrival, U. Hence, given event C, the Y busy cycle is U. On the other hand, for the X model, at this same epoch, there are at most $n + 1$ customers in the system, with at most one in queue. Given event C, by time $1/\mu$, all customers initially in service will have completed service and departed. Again given event C, by time $2/\mu$, any initially waiting customer will have entered service and completed service if the customer did not abandon first. However, given event C, we also have $A \leq 2/\mu$, so that the customer also would have abandoned. (We only need the A part of the event C for the Y model.) Thus if event C occurs at the beginning of a busy cycle in the Y model, then the current busy cycle ends in both models after the time U (which has been conditioned to be greater than $2/\mu$).

Thus the busy cycle T_X for the X model is bounded above by the random sum of N model-Y busy cycles, $T_{Y,i}$, until the event C first occurs at the beginning of a busy cycle, plus the single special U. For the Y models, these successive trials are i.i.d. because of the regenerative structure. The key fact we now exploit is the fact that a busy cycle T_Y of the Y process always has finite mean. For that, we can apply Corollary XII.2.5 of [1] or Theorem 2.2 of [31]. We can express the finite mean $E[T_Y]$ as

$$\begin{aligned} E[T_Y] &= pE[T_Y|C] + (1-p)E[T_Y|C^c] \\ &= pE[U|U > 2/\mu] + (1-p)E[T_Y|C^c]. \end{aligned} \tag{10.2}$$

Since, $E[U] < \infty$, necessarily $E[U|U > 2/\mu] < \infty$, so that

$$E[T_Y|C^c] \leq \frac{E[T_Y] - pE[U|U > 2/\mu]}{1-p} \leq \frac{E[T_Y]}{1-p} < \infty. \tag{10.3}$$

(Here we use the fact that $p < 1$.)

Finally, we can combine the results above to conclude that an X busy cycle T_X is stochastically bounded by a geometric random sum of i.i.d random variables, each distributed as $[T_Y|C^c]$, plus one more random variable distributed as $[U|U > 2/\mu]$. Hence, we have the bound

$$E[T_X] \leq \frac{E[T_Y|C^c]}{p} + E[U|U > 2/\mu] \leq \frac{E[T_Y]}{p(1-p)} + \frac{E[U]}{P(U > 2/\mu)} < \infty. \tag{10.4}$$

(Here we use the fact that $0 < p < 1$.)

We now consider the lower bound. We obtain a simple lower bound by observing that the original (X) system cannot empty until at least one

interarrival time exceeds the service time $1/\mu$ of that arrival. Let $N' \equiv \{n \geq 1 : U_n > 1/\mu\}$, a geometric random variable with parameter $p' \equiv P(U > 1/\mu) \equiv \bar{G}(1/\mu)$. Thus the cycle time T_X is stochastically bounded below by a sum of $N-1$ i.i.d. interarrival times that are less than $1/\mu$ plus the last interarrival time that is greater than $1/\mu$. Hence the expected cycle time must be bounded below by

$$\begin{aligned} E[T_X] &\geq \sum_{i=1}^{N'-1} E[U|U \leq 1/\mu] + E[U|U > 1/\mu] \\ &= \frac{1-p'}{p'} E[U|U \leq 1/\mu] + 1/\mu. \end{aligned}$$

Proof of Theorem 3.1. We first establish the limit for $(\bar{B}_n, \bar{E}_n, \bar{S}_n)$ in (3.11). Since the service times are deterministic with constant value $1/\mu$, the departures (service completions) in the interval $[0, 1/\mu]$ are completely determined by the initial age distribution in service, i.e., $S(t) = B(0, 1/\mu) - B(0, 1/\mu - t)$ and $S_n(t) = B_n(0, 1/\mu) - B_n(0, 1/\mu - t)$, $n \geq 1$. By Assumption 3, $\bar{B}_n(0, \cdot) \Rightarrow \bar{B}(0, \cdot)$ Hence we necessarily have $\bar{S}_n \Rightarrow \bar{S}$ in $D([0, 1/\mu])$, where $\bar{S}$ is nondecreasing and continuous.

For the next step, we first do the proof in the case $B(0, 1/\mu) = 1$, i.e., $t^* = T^* = 0$; afterwards we reduce the other case to this one. By Assumption 1, we have $\bar{N}_n \Rightarrow \Lambda$. By condition (3.9), asymptotically, the instantaneous arrival rate is greater than or equal to the instantaneous service completion rate. Hence, the fluid entering service during $[0, 1/\mu]$ is asymptotically equivalent to the fluid completing service; i.e., we have $\|\bar{E}_n - \bar{S}_n\|_{1/\mu} \Rightarrow 0$ as $n \to \infty$, where $\|x\|_c$ denotes the uniform norm over the interval $[0, c]$. By the convergence-together theorem, Theorem 11.4.7 of [34], $\bar{E}_n \Rightarrow \bar{E}$ in $D([0, 1/\mu])$.

However, we can write $b(1/\mu, x) = b(1/\mu - x, 0)$, $0 \leq x \leq 1/\mu$, so that $B(1/\mu, x) = E(1/\mu) - E(1/\mu - x)$, $0 \leq x \leq 1/\mu$, and, similarly, $B_n(1/\mu, x) = E_n(1/\mu) - E_n(1/\mu - x)$, $0 \leq x \leq 1/\mu$. Thus, by above, we get $B_n(1/\mu, \cdot) \Rightarrow B(1/\mu, \cdot)$ in $D([0, 1/\mu])$. We then see that the properties in Assumption 3 hold again at time $t = 1/\mu$. Hence we can apply mathematical indiction to conclude that $(\bar{S}_n, \bar{E}_n) \Rightarrow (\bar{S}, \bar{E})$ in $\mathbb{D}^2$ as $n \to \infty$. Since we can represent the two parameter process $\bar{B}_n$ in terms of $\bar{E}_n$, we get $\bar{B}_n \Rightarrow \bar{B}$ in $\mathbb{D}_{\mathbb{D}}$ as well. Since all limits are deterministic, all the limits are joint by Theorem 11.4.5 of [34]. That establishes (3.11) when $B(0, 1/\mu) = 1$.

We now consider the case in which $B(0, 1/\mu) < 1$. For the rest of the proof, let $V(t) \equiv B(t, 1/\mu)$ and $V_n(t) \equiv B_n(t, 1/\mu)$ with $\bar{V}_n(t) \equiv n^{-1} V_n(t)$. In this case, the limiting fluid model is underloaded until time $t^* = T^*$

in (6.3). Moreover, in this case (unlike Example 3.1) we can establish that $T_n \Rightarrow t^*$ as $n \to \infty$, exploiting condition (3.10).

We first show that, for any $\delta > 0$, $P(T_n > t^* - \delta) \to 1$ as $n \to \infty$. Since V is continuous, the definition of t^* implies that, for any $\delta > 0$, there exists $\epsilon > 0$ such that $\|V\|_{t^*-\delta} < 1 - \epsilon$. Now observe that, for all t, $\bar{V}_n(t) \leq \bar{V}_n^u(t) \equiv \bar{V}_n(0) + \bar{N}_n(t) - \bar{S}_n(t)$. However, $\|\bar{V}_n^u - V\|_t \Rightarrow 0$ for all $t > 0$, where $V(t) = V(0) + \lambda t - S(t)$ with $V(t) < 1$ for all $t < t^*$. Hence, for any $\delta > 0$ and $\epsilon > 0$, $P(\|\bar{V}_n^u - V\|_{t^*-\delta} > \epsilon) \to 0$ as $n \to \infty$. If $\|V\|_{t^*-\delta} < 1-\epsilon$ and $\|\bar{V}_n^u - V\|_{t^*-\delta} \leq \epsilon$, then $\bar{V}_n(t) \leq \bar{V}_n^u(t) < 1$ for all t, $0 \leq t \leq t^* - \delta$, which implies that $T_n \geq t^* - \delta$. Hence, we have shown that, for any $\delta > 0$, $P(T_n > t^* - \delta) \to 1$ as $n \to \infty$.

We now show that, for any $\delta > 0$, $P(T_n > t^* + \delta) \to 0$ as $n \to \infty$. Given that we have just shown that $P(T_n > t^* - \delta) \to 1$ as $n \to \infty$, we necessarily also have $\|\bar{E}_n - \bar{N}_n\|_{t^*-\delta} \Rightarrow 0$, so that $\|\bar{V}_n - \bar{V}_n^u\|_{t^*-\delta} \Rightarrow 0$ for $\bar{V}_n^u$ defined above, so that $\|\bar{V}_n - V\|_{t^*-\delta} \Rightarrow 0$ as well for any $\delta > 0$. Moreover, since both $\bar{V}_n$ and V are bounded below by 0 and above by 1, we can obtain $\|\bar{V}_n - V\|_{t^*} \Rightarrow 0$, which implies that $\bar{V}_n(t^*) \Rightarrow \bar{V}(t^*) = 1$. as $n \to \infty$.

Since the limiting fluid model becomes overloaded at time t^*, we can apply condition (3.10) to conclude that there must exist $\delta > 0$ and $\eta > 0$ such that $\lambda\delta > S(t^* + \delta) - S(t^*) + \eta$. Given that δ and η, define the following events:

$$\begin{aligned} C_{0,n} &\equiv \{T_n > t^* + \delta\} \\ C_{1,n} &\equiv \{\bar{V}_n(t^*) < 1 - \eta/4\} \\ C_{2,n} &\equiv \{S_n(t^* + \delta) - S_n(t^*) > \lambda\delta - \eta/2\} \\ C_{3,n} &\equiv \{N_n(t^* + \delta) - N_n(t^*) < \lambda\delta - \eta/4\}. \end{aligned} \tag{10.5}$$

Then observe that $C_{0,n} \subseteq C_{1,n} \cup C_{2,n} \cup C_{3,n}$, so that $P(C_{0,n}) \leq P(C_{1,n}) + P(C_{2,n}) + P(C_{3,n})$. However, $P(C_{i,n}) \to 0$ as $n \to \infty$ for each i, $1 \leq i \leq 3$. Hence, $P(T_n > t^* + \delta) \to 0$ as $n \to \infty$. Combining the two results, we obtain $T_n \Rightarrow t^*$ as $n \to \infty$.

We now continue to establish (3.11) in the case $V(0) \equiv B(0, 1/\mu) < 1$. The asymptotic behavior prior to time t^* is easy, because $E_n(t) = N_n(t)$ for $0 \leq t \leq T_n$, where $T_n \Rightarrow t^*$ as $n \to \infty$. Hence, we have $E_n \Rightarrow E$ in $D([0, t^*])$ as $n \to \infty$. For the rest of the proof, we shift t^* to the origin and apply the first part of the proof for the case $t^* = 0$.

It now remains to establish the limit (3.15) for $(\bar{Q}_n, \bar{A}_n)$, for which it suffices to consider the system after time t^*, when the system is full, but the queue is empty. Henceforth we assume that the system is full initially with an empty queue. For this remaining step, we can proceed under the assumption that, asymptotically, the service facility is always full with an

asymptotic rate of fluid entering service and departing of

$$b((k-1)/\mu+t,0) = \sigma(k/\mu+t) = b(k/\mu, 1/\mu-t) = b(0,1/\mu-t), \quad 0 \le t \le 1/\mu.$$

Now we will focus only on the queue and regard the queue as a $G/GI/\infty$ model with service times equal to the original abandonment times and a new arrival process. Service completions in the $G/GI/\infty$ model are to be interpreted as abandonments, while the total number of customers in the $G/GI/\infty$ system is to be interpreted as the number in queue. The arrival process for the $G/GI/\infty$ system in model n is $N_n(t) - E_n(t)$, where $E_n(t)$ is the number of customers to enter service in $[0,t]$.

Note that this representation fails to faithfully capture the original FCFS service discipline, because new arrivals go to the end of the queue, whereas customers enter service from the front of the queue. Instead, this representation applies directly to the last-come first-served (LCFS) discipline. However, that is where the exponential abandonment assumption comes in. With exponential abandonment, the number in queue $Q_n(t)$ is independent of the service discipline.

Given the $G/GI/\infty$ representation, we are able to directly apply FWLLN's established in [23]. Alternatively, we could apply [25]. Since E_n is asymptotically equivalent to the service completion process S_n, this new arrival process satisfies a FWLLN, having limit $\Lambda - S$, which in general is not a linear function. However, since $b(0,x) \le \lambda$ for all x, $0 \le x \le 1/\mu$, we also have $\sigma(t) \le \lambda$ for all $t \ge 0$, so it has a nonnegative rate. Hence we can prove (3.15) with (3.16) and (3.18) by applying Theorems 3.1 and 7.1 of [23]. To do so, we exploit the fact that the limit of the arrival process there is allowed to be nonlinear.

Finally, we complete the proof by showing (3.17) holds. We first exploit (3.16), which implies that

$$\begin{aligned} Q(t) &= \int_0^t e^{-\theta\,(t-s)}(\lambda - b(s,0))ds \\ &= \frac{\lambda}{\theta}(1 - e^{-\theta\,t}) - e^{-\theta\,t}\int_0^t b(s,0)\,e^{\theta\,s}ds. \end{aligned} \tag{10.6}$$

On the other hand, the ODE (5.7) implies that

$$w'(t) = 1 - \frac{b(t,0)}{\lambda\,e^{-\theta\,w(t)}}, \quad w(0) = 0,$$

which has a unique solution

$$w(t) = t - \frac{1}{\theta}\log\left(\frac{\theta}{\lambda}\int_0^t b(s,0)e^{\theta\,s}ds + 1\right). \tag{10.7}$$

Combining (3.17)and (10.7), we obtain (10.6).

Proof of Theorem 5.5. First consider the interval $[0, 1/\mu]$. The departure rate is $\sigma(t) = b(t, 1/\mu) = b(0, 1/\mu - t)$ for $0 \le t \le 1/\mu$. Since the staffing function is constant s, it is necessary to have $\lambda > \sigma(t)$ ($\lambda < \sigma(t)$) if the system switches from underloaded (overloaded) to overloaded (underloaded) at t. Consider an underloaded interval $[a, b] \subset [0, 1/\mu]$ where a and b are switching points, we must have $\zeta(a) > 0 > \zeta(b)$, which implies that ζ changes its sign in (a, b) at least once from positive to negative. The sign changing can be achieved in two cases: (i) crossing level 0 continuously from above to below, or (ii) jumping from above 0 to below. Therefore, ζ has at least a zero in case (i) and a discontinuity in case (ii) in interval (a, b). Similar reasoning works for an overloaded interval. This reasoning applies to all overloaded and underloaded subintervals that begin and end in the interior $(0, 1/\mu)$ of the interval $[0, 1/\mu]$. In addition, there are the two intervals with the interval endpoints. Thus the number of switches exceeds the number of internal intervals by at most 1. Let $\mathcal{S}_{[0,1/\mu]}$ be the total number of switching points in $[0, 1/\mu]$. We have just shown that we must have $|\mathcal{S}_{[0,1/\mu]}| \le |\mathcal{D}_\zeta| + |\mathcal{Z}_\zeta| + 1$.

We are done if $T = 1/\mu$; hence assume that $T > 1/\mu$. We continue for $\lceil T\mu \rceil$ cycles of length $1/\mu$. Next we consider the next interval $[1/\mu, 2/\mu]$. We will show that the number of switching points can be no greater than in the first interval of length $1/\mu$ just considered. Recall that the departure rate is $\sigma(t) = b(t, 1/\mu) = b(t - 1/\mu, 0)$. Let $\zeta_2(t) \equiv \sigma(t + 1/\mu) - \lambda = b(t, 0) - \lambda$ for $0 \le t \le 1/\mu$. Therefore, $|\mathcal{S}_{[1/\mu,2/\mu]}|$, the number of switching points in $[1/\mu, 2/\mu]$, is totally determined by the number of zeros and discontinuities of ζ_2, by the same argument as above.

We now show that $|\mathcal{Z}_{\zeta_2}| \le |\mathcal{Z}_\zeta|$. To do so, we first observe that we have $b(t, 0) = \sigma(t)$ when the system is overloaded. Hence the functions $\zeta(t)$ and $\zeta_2(t)$ differ only when the system is underloaded during $[0, 1/\mu]$. Consider an underloaded interval $[a, b] \subset [0, 1/\mu]$ where a and b are switching points, which implies that $\sigma(a) > \lambda > \sigma(b)$ ($\zeta(a) > 0 > \zeta(b)$). Since the system is underloaded in $[a, b]$, we must have $b(t, 0) = \lambda$. In case (i), ζ changes its sign in (a, b) with (at least) an zero at some $y \in \mathcal{Z}_\zeta \cap (a, b)$. However, ζ_2 has no such zeros in $\mathcal{Z}_{\zeta_2} \cap (a, b)$ because $\zeta_2(y) = 0$ for $a < y < b$ (which yields that $\mathcal{Z}_{\zeta_2} \cap (a, b) = \phi$), we have $|\mathcal{Z}_{\zeta_2} \cap (a, b)| = 0 \le |\mathcal{Z}_\zeta \cap (a, b)|$, which implies that $|\mathcal{Z}_{\zeta_2}| \le |\mathcal{Z}_\zeta|$ counting all underloaded intervals in $[0, 1/\mu]$ that are in case (i).

In case (ii), ζ changes its sign in (a, b) with (at least) a jump from positive to negative. However ζ_2 has at most two discontinuity points (at a and b) in (a, b) (because $\zeta_2(y) = 0$ for $a < y < b$). Although the number of discontinuities of ζ_2 in $[a, b]$ may outnumber the discontinuities of ζ by at most 1, these two jumps ($\zeta_2(a-) > \lambda$ to $\zeta_2(a) = \lambda$ and $\zeta_2(b-) = \lambda$ to

$\zeta_2(b) < \lambda$) can at most contribute to one sign change in (a, b). In other words, ζ_2 may have more discontinuities than ζ, but those extra ones are redundant. Hence, $|\mathcal{S}_{[1/\mu,2/\mu]}| \le |\mathcal{D}_\zeta| + |\mathcal{Z}_{\zeta_2}| \le |\mathcal{D}_\zeta| + |\mathcal{Z}_\zeta|$. The desired bound in (5.9) is obtained by induction on interval $[n/\mu, (n+1)/\mu]$, continuing until $N \equiv \lceil T\mu \rceil$.

11. Conclusions. We considered the heavily loaded many-server queue with customer abandonment and deterministic service times, i.e., the stochastic $GI/D/n+GI$ model. Even though the arrival rate exceeds the maximum possible service rate, the customer abandonment keeps the system stable. In §2 we showed that the busy cycles in the stochastic $GI/D/n+GI$ queueing model constitute regeneration times, so that stochastic processes describing the performance, such as the number of customers in the system, converge to proper steady state distributions as time evolves for any proper initial condition.

In §3 we showed that a sequence of $G/D/n+GI$ queueing systems with $\rho \equiv \lambda/\mu > 1$ indexed by n satisfies a many-server heavy-traffic limit in the efficiency-driven (ED) regime, converging to a deterministic fluid model, provided that the arrival processes and initial conditions obey functional weak laws of large numbers. In general, Theorem 3.1 only establishes a limit for the performance measures describing the service facility, e.g., $B_n(t, y)$, but those fluid limits capture the essential periodic character. A many-server heavy-traffic limit for the queue-length and abandonment processes was also obtained under the assumption of exponential abandonment.

Like the stochastic system, we found that the limiting fluid model has a unique stationary point. However, unlike the stochastic model, Theorem 9.3 shows that the fluid model never converges to that stationary point unless it starts in that stationary point. Instead, the fluid model tends to exhibit periodic behavior. Moreover, the specific form of the periodic behavior depends critically on the initial conditions. As a consequence, the asymptotic loss of memory (ALOM) property established for the $G_t/M_t/s_t+GI_t$ model in [21] does not nearly hold with deterministic service times.

Moreover, as illustrated in §1, simulations of the stochastic system show that the time-dependent behavior of the stochastic system is well described by the fluid model for large n. Indeed, the fluid model tends to provide a better description of the performance in the queueing model than the steady-state distribution of the queueing model, amplifying [33].

The rest of the paper was devoted to a careful study of the limiting fluid model. We obtained quite complete results for the case in which there exists a finite time T^* after which the system remains overloaded. Theorem 6.1 provides general conditions for this to be true. That condition is in terms

of the initial density of fluid in service $b(0,\cdot)$, but can also be applied at later times after applying the algorithm in Remark 5.2 over some initial interval. However, §9 shows that, in general, such a finite time need not exist. Nevertheless, Theorem 9.4 shows that the fluid density in service b converges to a PSS,

In summary, the fluid content in service evolves in three different ways, depending on the initial conditions:

1. The fluid in service is in steady state for all $t \geq 0$ if it is initialized with $b(0,x) = \mu\, s$ for $0 \leq x \leq 1/\mu$.
2. The system first becomes overloaded at $t^* < 1/\mu$ and remains overloaded after time T^*, $t^* \leq T^* < \infty$, in which case $b(t,\cdot)$ is in a PSS determined by $b(T^*,\cdot)$.
3. The system first becomes overloaded at $t^* < 1/\mu$, but switches between overloaded and underloaded infinitely often. Then the fluid density b converges to an overloaded PSS.

In cases (ii) and (iii), if instead we initialize by redefining $b(0,\cdot)$, letting it have the PSS version, then the system is initially overloaded and the fluid density in service is periodic with period $1/\mu$ for all $t \geq 0$. The remaining queue performance then converges to a PSS as well. In case (i), the associated queue performance converges to the unique stationary point as well. In cases (ii) and (iii), if we start with the PSS for b, then the queue performance converges to a PSS as well. In case (iii) it remains to determine if the queue performance converges to the PSS associated with the limiting PSS for b when we use the given initial conditions; we conjecture that it does.

It is natural to wonder what happens with other service-time distributions. In Appendix I we show that the same periodic behavior is exhibited by the corresponding model with a two-point service-time distribution, provided that one of the points is at the origin (in the same spirit as the corresponding special hyperexponential distribution in in [36]). However, in Appendix J we present results from simulation experiments showing that the periodic phenomenon ceases to hold for other two-point distributions and, more generally, if the service-time is only nearly deterministic. When the service-time distribution is nearly deterministic, the performance is similar to the performance with D service and the same initial conditions over suitably short time intervals, but convergence to stationary performance is evident as t increases.

We concentrated on the stationary $G/D/n + GI$ fluid model, but some of the results can be extended. First, as in [19–21], we can analyze, and obtain an algorithm for, the $G_t/D/s_t + GI$ fluid model in which the arrival rate

and the number of servers are allowed to be time varying. In particular, §4, §5 and §7 extend to this case. In general, we lose the periodic structure, on which most of this paper focuses, but that periodic structure is retained as well if the arrival rate function λ and the staffing function s are also periodic with the same period $1/\mu$. (However, the periodic structure is less surprising in that case.) Moreover, the structural properties of the queue established in §7 also extend to GI service, provided that the fluid density in service b is given. Of course, determining b is more complicated for GI service that is neither D nor M. Theorem 5.1 of [19] shows that it is necessary to solve a complicated fixed point equation in order to determine b in those cases.

As stated in §1, we began this study in an effort to understand if ALOM holds for the $G/GI/s + GI$ and $G_t/GI_t/s_t + GI_t$ fluid models when the service-time distribution is neither M_t nor M. That question remains after we stipulate that the service distribution also is neither D nor the two-point distribution with one mass at 0. We conjecture that ALOM does hold for the fluid model under that extra condition and the regularity conditions imposed in [21].

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

Appendix

APPENDIX A: OVERVIEW

This appendix contains additional supplementary material, which is presented in order of the material to which it relates. First, in §B we present additional simulation results for the example in §1. Specifically, we report results of simulations with smaller scaling n but averaged over multiple sample paths, to show the quality of the fluid model as an approximation for mean values in the queueing system. We also consider an example with smaller traffic intensity ρ for the example in §1 to show that the periodic behavior is eventually broken.

In §C we give proofs of Theorems 7.1–7.4 in §7. In §D we return to the example in §1 and show that different initial conditions can yield very different PSS's. In §E we apply the algorithm in Remark 5.2 to numerically evaluate the average performance over a cycle with non-exponential abandonment distributions. These examples show that the average boundary waiting time over a cycle tends to be strictly greater than the stationary value, whereas the average queue length over a cycle can be either strictly greater or strictly less than the stationary queue content in the fluid model. In §F we provide a proof of Corollary 8.2, giving explicit expressions for the performance in

the $G/D/s+M$ fluid model with an exponential abandonment cdf. In §G we provide a proof of Theorem 9.1 showing that there need not exist a finite time T^* after which the system remains overloaded. To do so, we show that the given example switches back and forth between overloaded and overloaded infinitely often, with two switches in each cycle. In §H, we give another counterexample with $B(0) < 1$ that is an analog of Example 3.1 in §3.

We then start to consider other service distributions. In §I we provide the same PSS results for fluid models that have two-point service distributions with one of the points at 0. Simulation verification is also given there. In §J we provide results of simulation experiments for queues that have nearly deterministic service times. The simulation results shows that the behavior for D service is not exhibited for other two-point distributions. This supports (but of course does not prove) our conjecture that ALOM holds in all other $GI/GI/s+GI$ models and even in the more general $G_t/GI/s_t+GI$ models.

APPENDIX B: MORE ON THE EXAMPLE IN SECTION 1

B.1. Smaller scaling n. We used a very large scaling, in particular $n = 1000$, for the queueing model in the example in §1. We used a very large n for two reasons: first, to demonstrate that the fluid model becomes accurate in the limit as $n \to \infty$ and, second, to provide a good test of the numerical algorithm for the fluid model. However, in order to be useful as approximations for realistic large-scale queueing systems, the approximation also should be reasonable for smaller scaling factors. We demonstrate that now.

We consider the same base $M/D/n+M$ fluid model here as in §1, but we only consider the case $\theta = 2$. The other parameters remain unchanged: $\lambda = 2$, $\mu = s = 1$. However, we consider different values of the scaling factor n for the associated stochastic queueing model, which coincides with the number of servers (since we set $s = 1$).

Figure 4 below provides the analog of Figure 2 for the case of one sample path of the simulation with $n = 100$, for the same fluid model. Figure 5 below gives the average of 10 sample paths for the same model. We see that the fluid approximation provides only a rough approximation for a single sample path when $n = 100$ instead of $n = 1000$, but it is remarkably accurate for the average over 10 sample paths. The accuracy is especially high in this example, because the extent of the overloads and underloads are quite large.

The quality of the approximation does degrade as n decreases, for the given fluid model. To illustrate, we plot a single sample path for $n = 30$ in

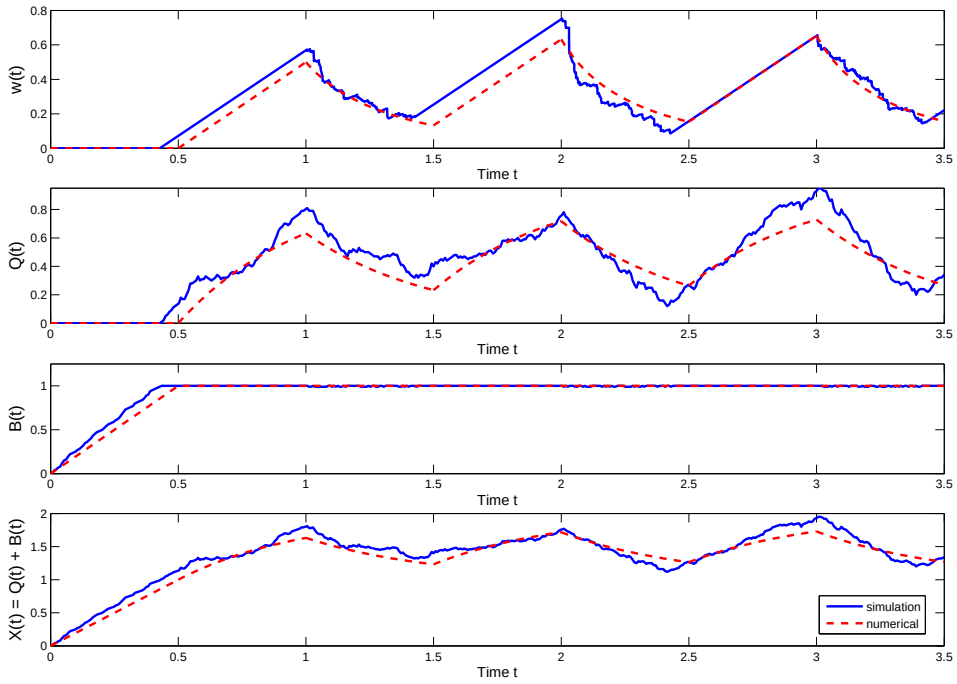

FIG 4. *Performance of the $G/D/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.*

Figure 6 and the average over 100 sample paths in Figure 7. The stochastic fluctuations are so much greater for a single sample path that we need to average over more sample paths to get a good estimate of the mean values. For $n = 30$, the fluid model clearly yields a good approximation only for the mean values, but the mean is remarkably well approximated for $n = 30$. The approximation for the mean values in Figure 7 are so good that it is evident that the fluid model approximations can provide useful approximations for the mean values for much smaller n (and thus s).

B.2. Smaller traffic intensity ρ. For the initial heavily loaded example with $\rho \equiv \lambda/s\mu = 2$ and scaling $n = 1000$ discussed in §1 we were not able to detect a break in the periodic behavior in simulations. For example, Figure 3 shows that the periodic behavior of $W_n(t)$, the head-of-line waiting time at t, remains even for large T ($T = 1000$). However, we found that a break in the periodic behavior can be observed if we considered less heavily loaded examples.

To illustrate, we now consider the same $M/D/n+M$ queue in §1 with the same parameters ($\mu = 1$, $\theta = 2$, $n = 100$) except for a smaller λ,

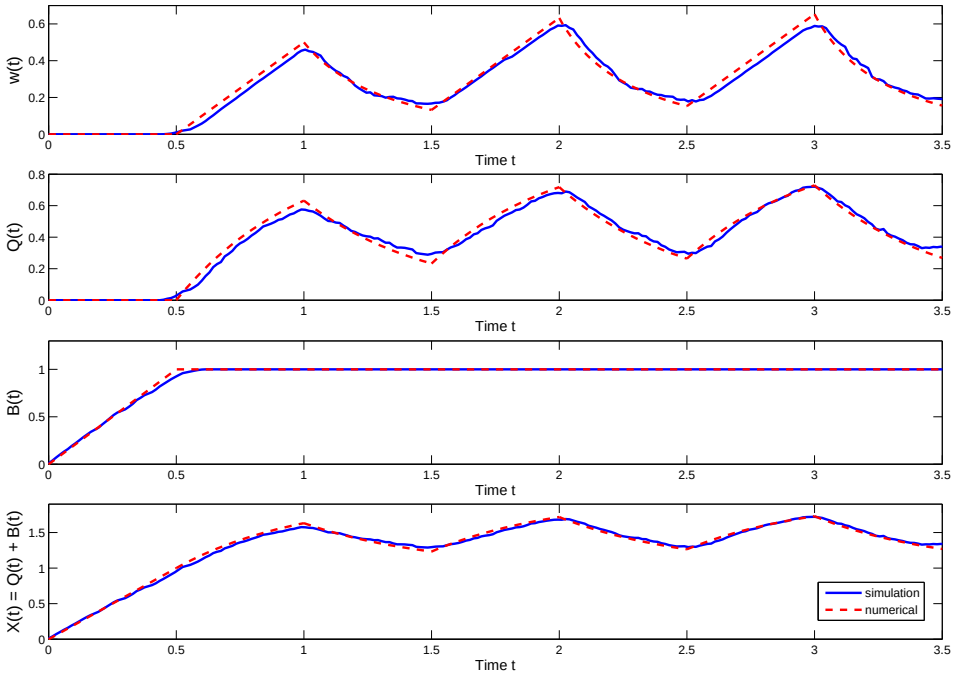

FIG 5. *Performance of the $G/D/s+M$ fluid model compared with simulation results: an average of 10 sample paths of the scaled queueing model based on $n = 100$.*

now letting $\lambda = 1.3\,n$, so that the system has a lower traffic intensity, $\rho = \lambda/n\mu = 1.3$ instead of $\rho = 2$ as in §1. We repeat the same simulation experiment with $\rho = 1.3$ and plot W_n in Figure 8. Figure 8 shows essentially the same periodic behavior over the initial interval $[0, 10]$, but it shows that the periodic behavior is gone by $T = 1000$.

APPENDIX C: PROOFS FOR §7

We omitted the proofs for the four theorems in §7 because they follow from the proofs of corresponding results in [21]. Nevertheless, we provide the details here.

C.1. Proof of Theorem 7.1.

PROOF. Since both queues are overloaded for all $t \geq 0$ and they have the same initial fluid densities in service, we have $b_1(t, 0) = b_2(t, 0) = \sigma_1(t) = \sigma_2(t)$ by Theorem 5.2. For the fluid content in queue, we have $\tilde{q}_1(t, x) \leq \tilde{q}_2(t, x)$ for all x by Proposition 5.1 because the two queues share the same F.

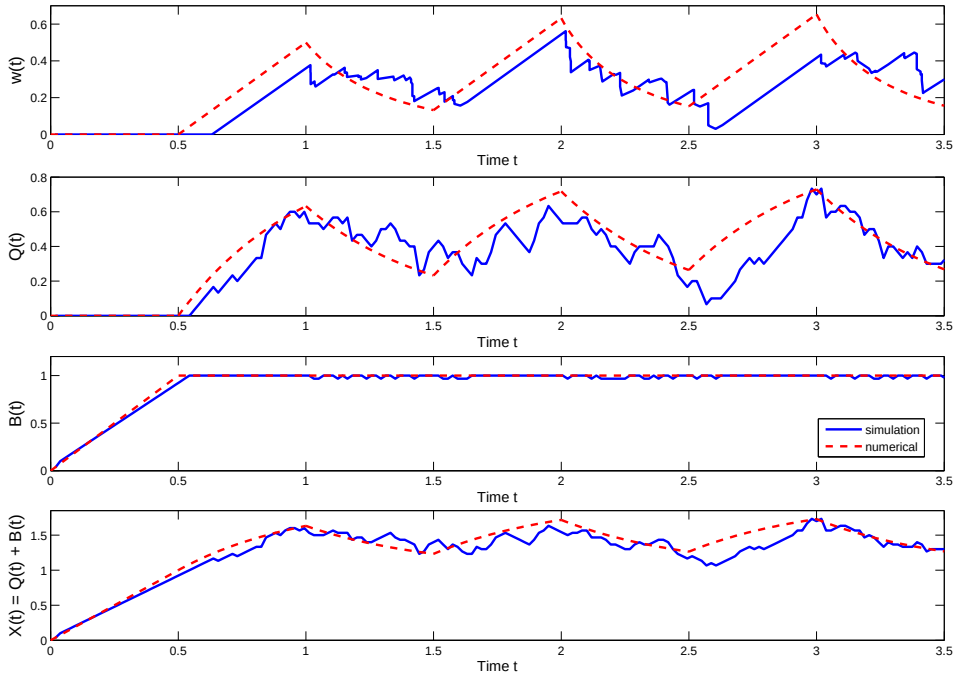

FIG 6. *Performance of the $G/D/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 30$.*

It remains to show $w_1(t) \le w_2(t)$ for all $t \ge 0$. We will do a proof by contradiction. Hence suppose this inequality does not hold for some $t > 0$. Then continuity of w_1 and w_2 implies that there exists some $0 < t_1 < t$ such that $w_1(t_1) = w_2(t_1) \equiv \tilde{w}$. However, the ordering of $\tilde{q}_1$ and $\tilde{q}_2$ implies that $\tilde{q}_1(t_1, \tilde{w}) \le \tilde{q}_1(t_1, \tilde{w})$. Hence the BWT ODE in Theorem 5.3 of [19] implies that $w_1'(t_1) = w_2'(t_1)$ because $b_1(t, 0) = b_2(t, 0)$. Therefore, this contradicts our assumption that there exists a t such that $w_1(t) > w_2(t)$. Hence that establishes the desired ordering.

The ordering of Q and α follow directly from the ordering of q and w since

$$\begin{aligned} Q_1(t) &= \int_0^{w_1(t)} q_1(t,x)dx \le \int_0^{w_2(t)} q_2(t,x)dx = Q_2(t), \\ \alpha_1(t) &= \int_0^{w_1(t)} q_1(t,x)h_F(x)dx \le \int_0^{w_2(t)} q_2(t,x)h_F dx = \alpha_2(t). \end{aligned}$$

Now we turn to v. The equation (27) in Theorem 5 implies that the ordering of w is inherited by v. That is made clear by applying the proof of Theorem

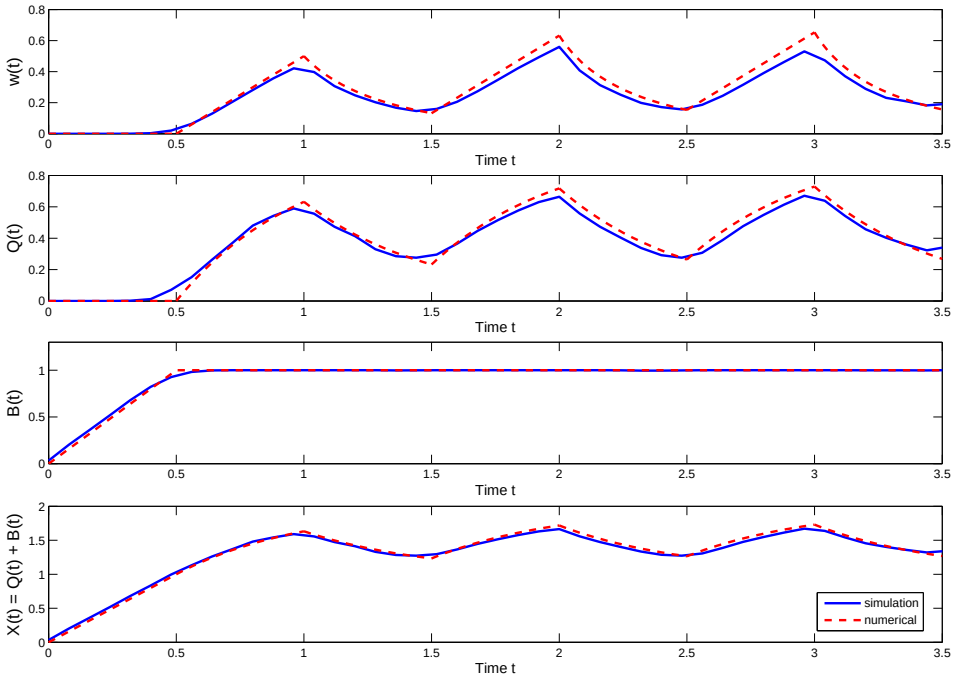

FIG 7. *Performance of the $G/D/s+M$ fluid model compared with simulation results: an average of 100 sample paths of the scaled queueing model based on $n = 30$.*

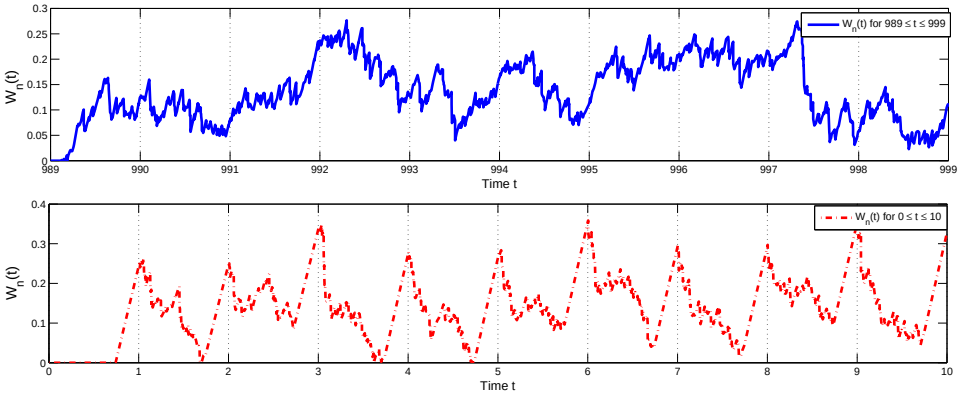

FIG 8. *Large-time periodic behavior of an overloaded $G/D/s+M$ queueing model: simulation estimates of the head-of-line waiting time W_n with $\lambda = 1.3$, $s = \mu = 1$, $\theta = 2$, $\rho = 1.3$, $n = 100$, $T = 1000$.*

5, which shows that $v(t)$ is determined by the intersection of the function w with the linear function $L_t(u) = t+u$. Clearly, if we increase the w function, then that intersection point increases as well. □

C.2. Proof of Theorem 7.2.

PROOF. Without loss of generality, by Theorem 7.1, it suffices to assume that $\lambda_1 \le \lambda_2$ and $q_1(0,\cdot) \le q_2(0,\cdot)$. If that is not initially the case, consider another two systems, system 3 and 4 with $\lambda_3 \equiv \lambda_1 \wedge \lambda_2$, $q_3(0,x) \equiv q_1(0,x) \wedge q_2(0,x)$, $\lambda_4 \equiv \lambda_1 \vee \lambda_2$, $q_4(0,x) \equiv q_1(0,x) \vee q_2(0,x)$. Therefore, it is easy to see that $|\lambda_1 - \lambda_2| = |\lambda_3 - \lambda_4|$ and $|Q_1(0) - Q_2(0)| \le |Q_3(0) - Q_4(0)|$.

Since both queues are overloaded and $b_1(t,0) = b_2(t,0)$, flow conservation of fluid in queue implies that for $i = 1, 2$,

$$Q_i'(t) = \lambda_i - \alpha_i(t) - b_i(t,0).$$

Hence, we have

$$Q_2'(t) - Q_1'(t) = \lambda_2 - \lambda_1 - (\alpha_2 - \alpha_1) \le \lambda_2 - \lambda_1, \tag{C.1}$$

where the inequality follows from Theorem 7.1. This yields

$$|Q_1(t) - Q_2(t)| = Q_2(t) - Q_1(t) \le |Q_1(0) - Q_2(0)| + t\,|\lambda_1 - \lambda_2|.$$

Obviously, (7.3) directly follows from (7.1). To show (7.2), we have

$$\begin{aligned}
|\alpha_1(t) - \alpha_2(t)| &= \alpha_2(t) - \alpha_1(t) \\
&= \int_0^{w_2(t)} q_2(t,x) h_F(x)dx - \int_0^{w_1(t)} q_1(t,x) h_F(x)dx \\
&= \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_F(x)dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x) h_F(x)dx \\
&\le h_F^{\uparrow} \left(\int_0^{w_1(t)} (q_2(t,x) - q_1(t,x)) h_F(x)dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x) h_F(x)dx \right) \\
&= h_F^{\uparrow}(Q_2 - Q_1) = h_F^{\uparrow} |Q_2 - Q_1|,
\end{aligned}$$

where the first and last equality, and the inequality all follows from Theorem 7.1. □

C.3. Proof of Theorem 7.3.

PROOF. We first show that (a) follows from (b). Without loss of generality, we assume $Q_1(0) \le Q_2(0)$. We construct another two systems, 3 and 4, with $q_3(0,x) \equiv q_1(0,x) \wedge q_2(0,x)$ and $q_4(0,x) \equiv q_1(0,x) \vee q_2(0,x)$. With this construction, systems 3 and 4 are bona fide fluid models, with

$Q_3(t) \le Q_1(t) \le Q_4(t)$ and $Q_3(t) \le Q_2(t) \le Q_4(t)$ for all t, by Theorem 7.1. This implies that $\Delta Q_{1,2}(t) \le \Delta Q_{3,4}(t)$ for all t. Since $\delta Q_{3,4}(t)(0) \le C_1$ for C_1 in (7.5), (7.4) in (a) follows from (7.10) for $\Delta Q_{3,4}(t)$. (The final bound on C_1 in (7.5) arises when the supports of $q_1(0,\cdot)$ and $q_2(0,\cdot)$ are disjoint sets.)

Now we prove (b). Observe that the first inequality in (7.10) follows (7.9) because dividing the interval $[0,T]$ into N subintervals yields

$$\Delta Q(T) \le \left(\frac{1}{1+h_F^{\downarrow}\frac{T}{N}}\right)^N \Delta Q(0).$$

Letting $N \to \infty$, we get (7.9).

We now prove (7.9). Since both queues are overloaded for all $t \ge 0$ and they have the same initial fluid densities in service, we have $b_1(t,0) = b_2(t,0) = \sigma_1(t) = \sigma_2(t)$, following from Theorem 5.2. Since $q_1(0,x) \le q_2(0,x)$, we have $q_1(t,x) \le q_2(t,x)$, $w_1(t) \le w_2(t)$ and $\alpha_1(t) \le \alpha_2(t)$ for all $t \ge 0$. Hence, we have

$$\begin{aligned}
\alpha_2(t) - \alpha_1(t) &= \int_0^{w_2(t)} q_2(t,x)h_F(x)dx - \int_0^{w_1(t)} q_1(t,x)h_F(x)dx \\
&= \int_0^{w_1(t)} (q_2(t,x) - q_1(t,x))h_F(x)dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x)h_F(x)dx \\
&\ge h_F^{\downarrow}\left(\int_0^{w_1(t)} (q_2(t,x) - q_1(t,x))dx + \int_{w_1(t)}^{w_2(t)} q_2(t,x)(x)dx\right) \\
&= h_F^{\downarrow}\left(Q_2(t) - Q_1(t)\right) = h_F^{\downarrow}\,\Delta Q(t).
\end{aligned} \tag{C.2}$$

Flow conservation implies that

$$Q_i'(t) = \lambda - \alpha_i(t) - b_i(t,0) \quad \text{for } i = 1,2,$$

which yields

$$\Delta Q'(s) = -(\alpha_2(s) - \alpha_1(s)) \le -h_F^{\downarrow}\,\Delta Q(s) \le -h_F^{\downarrow}\,\Delta Q(t), \quad 0 \le s \le t,$$

where the first inequality follows from (C.2) and the second inequality holds since $\Delta Q(t)$ has negative derivative. Therefore, integrating both sides with respect to s from 0 to t, we have

$$\Delta Q(t) - \Delta Q(0) \le -h_F^{\downarrow}\,t\,\Delta Q(t)$$

and

$$\Delta Q(t) \le \left(\frac{1}{1 + h_F^{\downarrow} t}\right) \Delta Q(0).$$

To show the second inequality in (7.10), repeat the reasoning in (C.2) and use the face $h_F(x) \le h_F^{\uparrow}$ instead of $h_F(x) \ge h_F^{\downarrow}$.

Finally, we treat $w(t)$. As above, it suffices to assume that we have the ordering in (7.8). We have $b(t,0) \ge b^{\downarrow}$ following from Proposition 5.2 and Corollary 5.2. First note that at time $T^* = (Q_1(0)+Q_2(0))/b^{\downarrow}$, all fluid that was in queue 1 and 2 at time 0 is gone (entered service or abandoned). Then (7.6) follows from

$$\Delta Q(T) = \int_{w_1(T)}^{w_2(T)} \lambda \bar{F}(x) dx \le \lambda \bar{F}(w_2(T)) \Delta w(T), \quad T \ge T^*.$$

Choose $\bar{w} > 0$ big enough such that $\bar{F}(\bar{w}) < b^{\downarrow}/\lambda$. The BWT ODE implies that for $t > T^*$,

$$w_2'(t) = 1 - \frac{b(t,0)}{\lambda \bar{F}(w_2(t))} \le 1 - \frac{b^{\downarrow}}{\lambda \bar{F}(\bar{w})} < 0,$$

if $w_2(t) > \bar{w}$ for some t. Hence $\bar{w}$ is an upper bound for $w_2(t)$ if $w_2(T^*) < \bar{w}$. If $w_2(T^*) \ge \bar{w}$, it is easy to see that $w_2(t)$ decreases until it is below $\bar{w}$ because we can bound $w_2'(t)$. This argument implies that $w_2(t) \le \bar{w} \vee (w_2(0) + T^*)$ for all $t \ge 0$. The constant C_2 in (7.7) is obtained by inserting established bounds. □

C.4. Proof of Theorem 7.4.

Proof. Most are elementary; only $Q(t)$ and $w(t)$ require detailed argument. Flow conservation implies that $Q'(t) = \lambda - \alpha(t) - b(t,0) \le \lambda - \alpha(t)$. Since $\alpha(t) \ge h_F^{\downarrow} Q(t)$, we have $Q'(t) < 0$ whenever $Q(t) > \lambda/h_F^{\downarrow}$. The bound for $w(t)$ follows directly from (7.6) and the proof of Theorem 7.3. □

APPENDIX D: DIFFERENT INITIAL CONDITIONS

Theorems 6.1 and 8.1 provide sufficient conditions for Assumption 12 to hold, and for the performance function to converge to a PSS. That PSS depends strongly on the fluid density in service, b at the time T^* after which the system remains overloaded. We now illustrate that different initial conditions can yield very different PSS's.

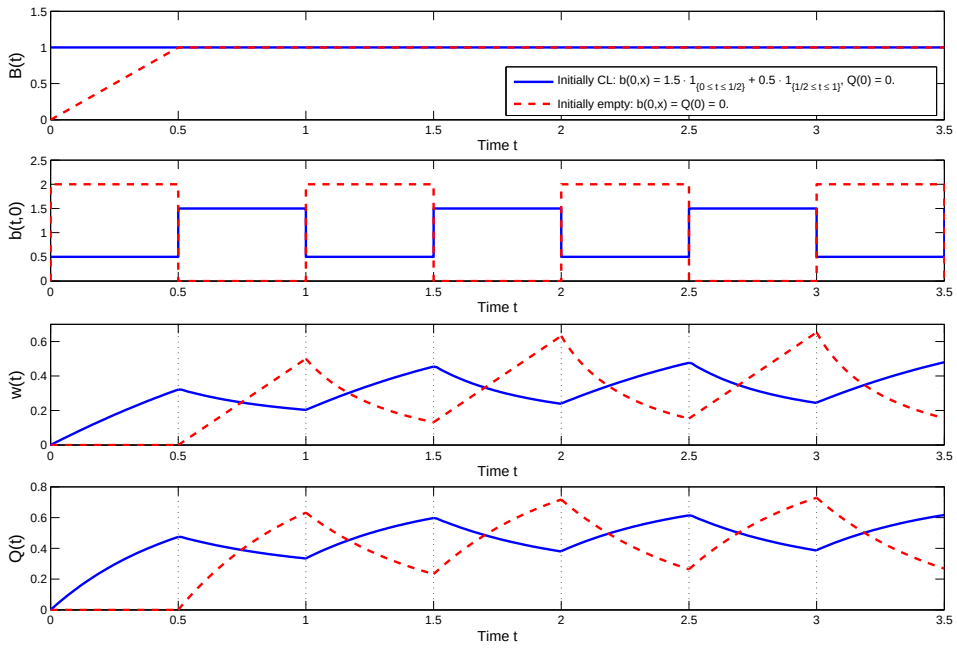

FIG 9. *A comparison of the PSS performance of the $G/D/s+M$ fluid queue with different initial conditions: (i) critically loaded with $b(0,x) = 1.5 \cdot 1_{\{0\le x\le 1/2\}} + 0.5 \cdot 1_{\{1/2\le x\le 1\}}$, $Q(0) = 0$ (the blue solid lines); (ii) starting empty (the red dashed lines).*

We again consider the $G/D/s+M$ example in §1 with $\lambda = 2$, $\mu = s = 1$, $\theta = 2$. In Figure 9, we apply the algorithm in Remark 5.2 and plot the performance functions $B(t)$, $b(t,0)$, $w(t)$ and $Q(t)$ in interval $[0, 3.5]$ for two different initial conditions: (i) The system is initially critically loaded (CL) with $b(0,x) = 1.5 \cdot 1_{\{0\le x\le 1/2\}} + 0.5 \cdot 1_{\{1/2\le x\le 1\}}$, $Q(0) = 0$ (the blue solid lines); (ii) The system is initially empty (the red dashed lines). Both cases yield a PSS with period $1/\mu = 1$, but the performance in these two cases differs greatly.

APPENDIX E: THE AVERAGE PERFORMANCE OVER A CYCLE

In Remark 8.3 we noted that, unlike $\bar{\alpha}$ and $\bar{\sigma}$, the averages of other performance functions in a PSS typically do not agree with the steady-state values. We investigate $\bar{Q}$ and $\bar{w} \equiv \tau^{-1}\int_0^\tau w(t)\,dt$ now.

We consider an initially empty $G/D/s+GI$ fluid model with three types of abandonment distributions: (i) Erlang-2 (E_2), (ii) exponential (M) and (iii) Hyperexponential-2 (H_2). We first review these distributions.

Let A be the generic abandonment time. A follows E_2 implies that $A = X_1 + X_2$ in distribution, where X_1 and X_2 are two iid exponential random

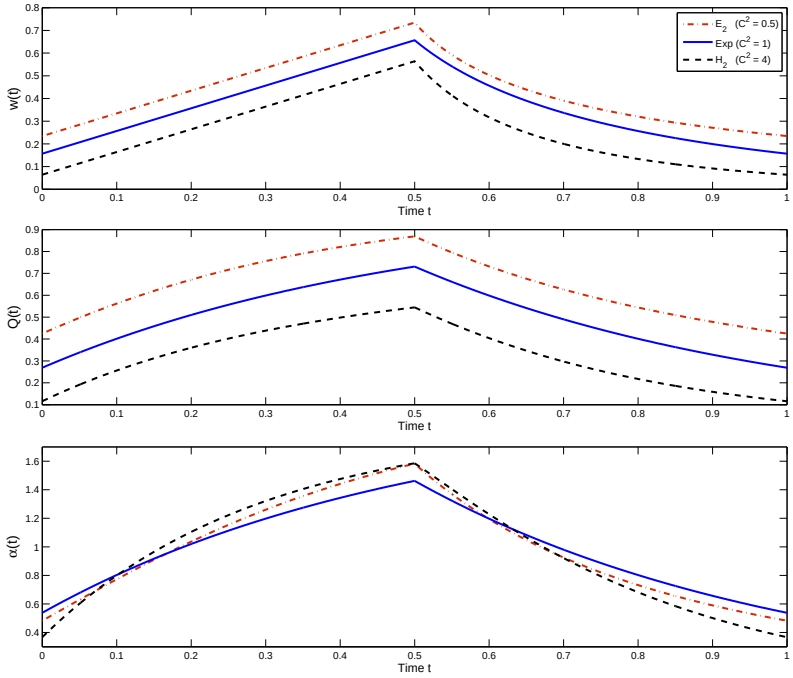

FIG 10. *A comparison of the PSS of the $G/D/s+GI$ fluid queues with different abandonment distributions: (i)E_2 (red dashed), (ii) M (blue solid) and (iii) H_2 (black dashed).*

variables. Moreover, $f(x) = \gamma^2 x e^{-\gamma x}$, where γ is rate of X_1. If A follows H_2, then A is a mixture of two exponential random variables, i.e., $f(x) = p \cdot \theta_1 e^{-\theta_1 x} + (1-p) \cdot \theta_2 e^{-\theta_2 x}$, where θ_1 and θ_2 are the rates of these two exponential random variables, and $0 < p < 1$ is the sampling probability.

We fix the mean of A, letting $E[A] = 1/\theta$. An E_2 distribution has squared coefficient of variation (SCV) $C^2 \equiv Var(A)/E[A]^2 = 1/2$, which is less than 1. On the other hand, all H_2 distributions have C^2 greater than 1. For E_2, we let $\gamma = 2\theta$. For H_2, we let $p = 0.5(1 - \sqrt{0.6})$, $\theta_1 = 2p\theta$, $\theta_2 = 2(1-p)\theta$, so that $C^2 = 4$.

We let $\lambda = 2$, $\theta = 2$, $\mu = s = 1$. In Figure 10, we plot w, Q and α in one cycle $[0, 1/\mu]$ of PSS for these three abandonment distributions, by applying the algorithm described in Remark 5.2. (Here we start the system empty and compute these performance functions in N cycles for N large.) In Table 1, we compute and compare $\bar{w}$, $\bar{Q}$ and $\bar{\alpha}$, the average of w, Q and α in one cycle to w^*, Q^* and α^*, their steady-state values. We have three observations: (i) As proved in Corollary 8.1, $\bar{\alpha}$ indeed agrees with α^* (except for a small computation error from numerical integration); (ii) $\bar{Q} \neq Q^*$ in

TABLE 1
A comparison of the average performance of PSS of the $G/D/s+GI$ fluid queue with (i) E_2, (ii) M and (iii) H_2 abandonment distribution to the steady-state values

abandonment dist.	E_2 $(C^2=0.5)$	M $(C^2=1)$	H_2 $(C^2=4)$
$\bar{\alpha}$ (PSS average)	1.001	1	1.001
α^* (steady state)	1	1	1
$\bar{w}$ (PSS average)	0.437	0.367	0.260
w^* (steady state)	0.420	0.347	0.226
$\bar{Q}$ (PSS average)	0.649	0.5	0.330
Q^* (steady state)	0.657	0.5	0.324

general, in particular, $\bar{Q} < Q^*$ for E_2 abandonment and $\bar{Q} > Q^*$ for H_2 abandonment; (iii) $\bar{w} \geq w^*$, i.e., customers' average waiting is longer in PSS than in the steady state.

APPENDIX F: THE CASE OF EXPONENTIAL ABANDONMENT

In this section we prove Corollary 8.2, giving explicit formulas in the case of exponential abandonment. We give two different proofs.

F.1. First Proof of Corollary 8.2. First, since $b(t,x)$ and $\sigma(t)$ are periodic functions and $Q(t)$ and $\alpha(t)$ can be written as expressions in terms of $w(t)$, it remains to derive the dynamics of $w(t)$.

In a cycle $[0, 1/\mu]$, $w(t) = \tilde{w}+t$ for $0 \leq t \leq 1/\mu - s/\lambda$ and $w(t)$ solves ODE $w'(t) = 1-1/\bar{F}(w(t)) = 1-1/e^{-\theta w(t)}$ with $w(1/\mu - s/\lambda) = \tilde{w}+1/\mu-s/\lambda$ for $1/\mu - s/\lambda \leq t \leq 1/\mu$, where $\tilde{w} \geq 0$ is both the starting and the ending value of $w(t)$ in each cycle. Letting $v(t) \equiv t-w(t)$, we have for $1/\mu - s/\lambda \leq t \leq 1/\mu$,

$$e^{\theta t} = (1-w'(t))e^{\theta(t-w(t))} = v'(t)e^{\theta v(t)}.$$

For $1/\mu - s/\lambda \leq t \leq 1/\mu$, integrating both sides from $1/\mu - s/\lambda$ to t yields

$$\begin{aligned} e^{\theta t} - e^{\theta(1/\mu - s/\lambda)} &= \theta \int_{1/\mu - s/\lambda}^{t} e^{\theta u} du = \theta \int_{v(1/\mu - s/\lambda)}^{v(t)} e^{\theta u} du \\ &= e^{\theta(t-w(t))} - e^{\theta(1/\mu - s/\lambda - w(1/\mu - s/\lambda))}. \end{aligned} \tag{F.1}$$

Because $w(1/\mu - s/\lambda) = \tilde{w} + 1/\mu - s/\lambda$ and $w(1/\mu) = \tilde{w}$, letting $t = 1/\mu$ in (F.1) yields (8.10), from which (8.8) follows. Solving the ODE yields (8.11).

Finally, to show (c), we consider a cycle $[1/\mu - \tilde{w}, 2/\mu - \tilde{w}]$ instead of $[0, 1/\mu]$. First, the PWT $v(t)$ is periodic with the same period $1/\mu$. Moreover, it is continuous over $[1/\mu - \tilde{w}, 2/\mu - \tilde{w})$ and it has a discontinuity at $t = 2/\mu - \tilde{w}$, as shown in Figure 11, following from Theorem 5.4. Also see Theorem 5

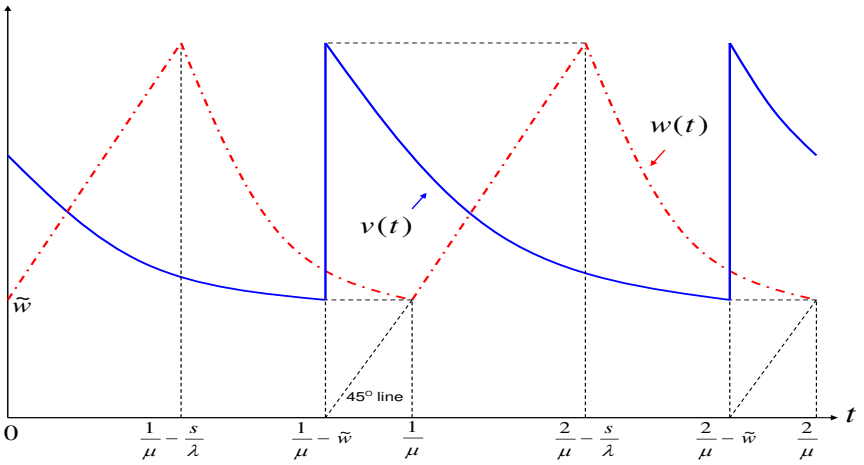

FIG 11. *PWT $v(t)$ and BWT $w(t)$ of the PSS of the $G/D/s+GI$ fluid queue.*

and 6 in [19] for details. Following Theorem 6 in [19], $v(t)$ satisfies the ODE

$$\begin{aligned} v'(t) &= \frac{\lambda \bar{F}(v(t))}{b(t+v(t),0)} - 1 = \frac{\lambda e^{-\theta v(t)}}{\lambda} - 1 \\ &= e^{-\theta v(t)} - 1, \quad \frac{1}{\mu} - \tilde{w} \le t < \frac{2}{\mu} - \tilde{w}, \end{aligned} \tag{F.2}$$

where the second equality holds because $b(t,0) = \lambda$ for $2/\mu - s/\lambda \le t \le 2/\mu$ and $t + v(t) \ge 2/\mu - s/\lambda$ (obviously from Figure 11). Since $v(1/\mu - \tilde{w}) = \tilde{w} + 1/\mu - s/\lambda \equiv v_0$, solving ODE (F.2) with $(1/\mu - \tilde{w}) = v_0$ yields (8.13).

F.2. Second Proof of Corollary 8.3. We can provide an alternative proof of Corollary 8.3 by focusing on $Q(t)$. Since $\sigma(t) = b(t,0) = 0$, $Q(t)$ satisfies an ODE for $0 \le t \le 1/\mu - s/\lambda$ with

$$Q'(t) = \lambda - \theta\, Q(t),$$

which has a unique solution

$$Q(t) = \frac{\lambda}{\theta}\left(1 - e^{-\theta t}\right) + Q(0)\, e^{-\theta t}. \tag{F.3}$$

Since $\sigma(t) = b(t,0) = \lambda$ for $1/\mu - s/\lambda < t \le 1/\mu$, $Q(t)$ satisfies another ODE

$$Q'(t) = \lambda - \theta\, Q(t) - b(t,0) = -\theta\, Q(t),$$

which has a unique solution

$$Q(t) = Q^* e^{-\theta t}, \tag{F.4}$$

where

$$Q^* \equiv Q\left(\frac{1}{\mu} - \frac{s}{\lambda}\right) = \frac{\lambda}{\theta}\left(1 - e^{-\theta\left(\frac{1}{\mu}-\frac{s}{\lambda}\right)}\right) + Q(0)\, e^{-\theta\left(\frac{1}{\mu}-\frac{s}{\lambda}\right)}$$

is the ending value of $Q(t)$ in $[0, 1/\mu - s/\lambda]$; i.e., let $t = 1/\mu - s/\lambda$ in (F.3). Since $Q(t)$ is periodic in the PSS with period $1/\mu$, we must have $\tilde{Q} \equiv Q(0) = Q(1/\mu)$. Equating $Q(0)$ to $Q(t)$ in (F.4) with $t = 1/\mu$ yields

$$\tilde{Q} = \frac{\lambda}{\theta}\left(\frac{e^{-\theta\, s/\lambda} - e^{-\theta/\mu}}{1 - e^{-\theta/\mu}}\right). \tag{F.5}$$

Plugging $Q(0) = \tilde{Q}$ in (F.5) into (F.3) and (F.4) yields (8.9) and (8.12). To show (8.10), we let

$$\tilde{Q} = \int_0^{\tilde{w}} \lambda\, e^{-\theta\, x} dx = \frac{\lambda}{\theta}\left(1 - e^{-\theta\, \tilde{w}}\right), \tag{F.6}$$

which yields (8.10).

APPENDIX G: ON THEOREM 9.1

Recall that Theorem 9.1 concludes that there need not exist a finite time T^* after which the system remains overloaded; i.e., there need not exist $T^* < \infty$ such that $B(t) = s$ for all $t \geq T^*$. The proof involves a concrete counterexample. We now show that the counterexample indeed has the claimed property.

G.1. Proof of Theorem 9.1. We start by giving a feel for the performance by applying the numerical algorithm in Remark 5.2. We plot the performance functions $w(t)$, $Q(t)$, $B(t)$, $b(t, 0)$ and $\sigma(t)$ for $0 \leq t \leq 5$ in Figure 12. Figure 12 clearly shows that $B(n) = s$ for all n and that $B(n+(1/2))$ increases towards s.

However, from the picture alone, we cannot be sure that $B(n+(1/2)) < s$ for all n. To justify that, we need to consider the behavior more carefully. To show that the system alternates between overloaded and underloaded infinitely often, we consider successive intervals $[n, n+1]$ for $n \geq 0$. First, in the first unit $[0, 1]$, we have $b(t, 0) = \sigma(t) = b(0, 1-x) = 2 \cdot 1_{\{0 \leq x \leq 1/2\}}$. Since

$b(t,0) = \sigma(t)$ whenever the system is overloaded and the system is initially overloaded, the BWT $w(t)$ satisfies the ODE

$$w'(t) = 1 - \frac{b(t,0)}{\lambda \bar{F}(w(t))} = 1 - \frac{2}{1.2\, e^{-2\,w(t)}} 1_{\{0 \le t \le 1/2\}}, \tag{G.1}$$

with $w(0) = 2$, which has a unique solution

$$w(t) = t - \frac{1}{2} \log \left(\frac{e^{2\,t} - 1}{0.6} + e^{-2\,w(0)} \right) \quad \text{for } 0 \le t \le 1/2.$$

Letting $w(t) = 0$ yields that

$$t_1^{(1)} = \frac{1}{2} \log \left(\frac{1 - 0.6\, e^{-2\,w(0)}}{0.4} \right) = 0.453 < 1/2, \tag{G.2}$$

that is the time at which the system becomes underloaded. Note that for $t_1^{(1)} < t \le 1/2$, $\sigma(t) = 2 > 1.2 = b(t,0) = \lambda$, therefore, the fluid content in service decreases (linearly) with $B(t) = s - (\sigma(t) - b(t,0))\,(t - t_1^{(1)}) = 1 - 0.8(t - t_1^{(1)})$. For $t > 1/2$, $b(t,0) = \lambda = 1.2 > 0 = \sigma(t)$, $B(t)$ increases (liearly) with $B(t) = B(1/2) + (b(t,0) - \sigma(t))\,(t - 1/2) = 0.96 + 1.2(t - 1/2)$. So the system again becomes overloaded at $t_2^{(1)} = 0.53$ since $B(t_2^{(1)}) = 1 = s$. Moreover, $t_1^{(1)}$ and $t_2^{(1)}$ satisfy $1.2(t_2^{(1)} - 1/2) = 0.8(1/2 - t_1^{(1)})$. For $t_2 \le t \le 1$, by ODE (G.1), $w(t) = t - t_2^{(1)}$, which implies that $w(1) = 1 - t_2^{(1)} = 0.47 < 2 = w(0)$. In summary, the system is overloaded in $[0, t_1^{(1)}] \cup [t_2^{(1)}, 1]$ and (strictly) underloaded in $(t_1^{(1)}, t_2^{(1)})$, $b^{(1)}(t,0) \equiv b(t,0) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le 1/2\}}$ and $w^{(1)}(0) \equiv w(0) > w(1) \equiv w^{(1)}(1)$, with $0 < t_1^{(1)} < 1/2 < t_2^{(1)} < 1$. See Figure 12.

Now consider the next unit interval $[1,2]$. We can simply shift the origin to time 1 and again consider the interval $[0,1]$. Therefore the system is initially overloaded with $w^{(2)}(0) \equiv w(0) = w^{(1)}(1) < w^{(0)}(0)$, $\sigma(t) = b^{(1)}(t,0) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le t_2^{(1)}\}}$ (which is the rate into service in the previous interval). We want to show that the same structure of all performance functions are preserved in the second unit interval. The switching time (from overloaded to underloaded) is a strict monotone function of $w(0)$, by (G.2), therefore the system becomes underloaded at $t_1^{(2)}$ such that $t_1^{(2)} < t_1^{(1)}$ since $w(0) = w^{(1)}(1) < w^{(1)}(0)$. Because $\sigma(t) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le t_2^{(1)}\}}$,

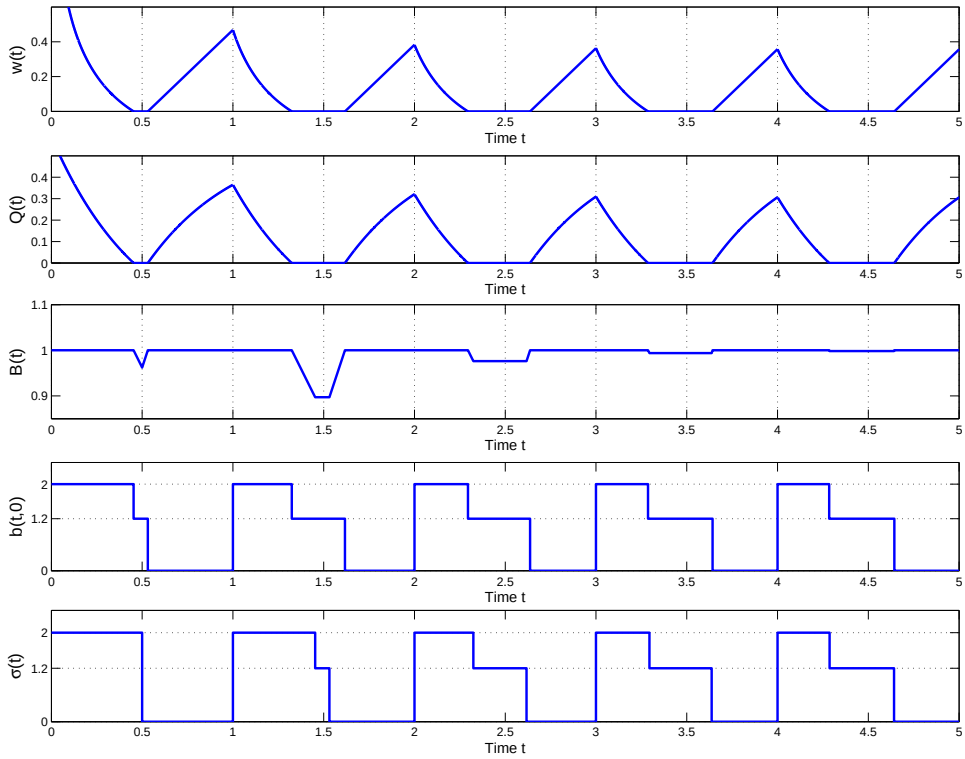

FIG 12. *The counterexample providing a fluid model that does not become (and stay) overloaded in finite time; it switches between overloaded and underloaded regimes infinitely often.*

we have

$$\begin{aligned} B(t) &= 1_{\{t\in[0,t_1^{(2)})\cup(t_2^{(2)},1]\}} + [1-0.8(t-t_1^{(2)})]1_{\{t_1^{(2)}\le t<t_1^{(1)}\}} \\ &\quad +[1-0.8(t_1^{(1)}-t_1^{(2)})]1_{\{t_1^{(1)}\le t\le 1/2\}} \\ &\quad +[1-0.8(t_1^{(1)}-t_1^{(2)})+1.2(t-t_2^{(1)})]1_{\{t_2^{(1)}\le t\le t_2^{(2)}\}}, \end{aligned}$$

where $t_2^{(2)}$ satisfies $1.2(t_2^{(2)}-t_2^{(1)}) = 0.8(t_1^{(1)}-t_1^{(2)})$ so that $t_2^{(2)} > t_2^{(1)}$, which implies that the system is overloaded for $t_2^{(2)} \le t \le 1$ and $w^{(2)}(1) \equiv w(1) = 1 - t_2^{(2)} < w(0) = w^{(1)}(1) = w^{(2)}(0)$. In summary, in the second interval, the system is overloaded in $[0, t_1^{(2)}] \cup [t_2^{(2)}, 1]$ and (strictly) underloaded in $(t_1^{(2)}, t_2^{(2)})$, $b^{(2)}(t,0) \equiv b(t,0) = 2 \cdot 1_{\{0\le t<t_1^{(2)}\}} + 1.2 \cdot 1_{\{t_1^{(2)}\le t\le t_2^{(2)}\}}$, $\sigma^{(2)}(t) \equiv \sigma(t) = b^{(1)}(t,0) = 2 \cdot 1_{\{0\le t<t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)}\le t\le t_2^{(1)}\}}$ and $w^{(2)}(0) \equiv w(0) > w(1) \equiv w^{(2)}(1)$, with $0 < t_1^{(2)} < t_1^{(1)} \le t_2^{(1)} < t_2^{(2)} < 1$. See Figure 12.

Using an inductive argument, we can show that in the nth unit interval $[n-1, n]$, the same structure is preserved. In particular, if we move the origin to time $n-1$ (i.e., consider $[0,1]$ instead of $[n-1,n]$), then

$$\text{the system} \quad \text{is} \quad \begin{cases} \text{overloaded}, & \text{for } t \in [0, t_1^{(n)}] \cup [t_2^{(n)}, 1], \\ \text{(strictly) underloaded}, & \text{for } t \in (t_1^{(n)}, t_2^{(n)}). \end{cases}$$

$$\begin{aligned} b^{(n)}(t,0) &\equiv b(t,0) = 2 \cdot 1_{\{0 \le t < t_1^{(n)}\}} + 1.2 \cdot 1_{\{t_1^{(n)} \le t \le t_2^{(n)}\}}, \\ \sigma^{(n)}(t) &\equiv \sigma(t) = b^{(n-1)}(t,0) = 2 \cdot 1_{\{0 \le t < t_1^{(n-1)}\}} + 1.2 \cdot 1_{\{t_1^{(n-1)} \le t \le t_2^{(n-1)}\}}, \\ w^{(n)}(0) &\equiv w(0) > w(1) \equiv w^{(n)}(1), \end{aligned}$$

with $0 \le t_1^{(n)} < t_1^{(n-1)} \le t_2^{(n-1)} < t_2^{(n)} \le 1$. Therefore, the bounded sequence $t_1^{(1)}, t_1^{(2)}, \ldots$ is strictly decreasing and the bounded sequence $t_2^{(1)}, t_2^{(2)}, \ldots$ is strictly increasing so that we must have $t_1^{(n)} \downarrow t_1^{\infty} \ge 0$ and $t_2^{(n)} \uparrow t_2^{\infty} \le 1$. We next show that $t_1^{\infty} > 0$ and $t_2^{\infty} < 1$. Suppose $t_1^{\infty} = 0$, then $w^{\infty}(0) = w^{\infty}(1) = 0$, which implies that $t_2^{\infty} = 1$ (the monotonicity structure is preserved in the limit). Therefore, the system is underloaded or critically loaded in $[0,1]$. However, since we have $\rho = \lambda/s\mu = 1.2 > 1$, this cannot happen. Hence a contradiction.

G.2. More on Theorem 9.1. The example in the proof of Theorem 9.1 discussed above in §G.1 also can illustrate the important role played by the initial queue density $q(0,\cdot)$ on the asymptotic performance. Indeed, we can ensure that a time $T^* < \infty$ exists such that $B(t) = s$ for all $t \ge T^*$ by changing the initial queue density. Moreover, we achieve this finite T^* in this example by *reducing* the initial fluid content in queue, not by increasing it.

We consider the same example as before, as discussed in §G.1, with the same initial fluid density in service but $w(0) = 0.2$ (instead of $w(0) = 2$). Figure 13 is the analog of Figure 12. As shown in Figure 13, the system becomes overloaded in the second cycle and stays overloaded thereafter. Moreover, the structure of the PSS is entirely different (in this case there is no critically loaded interval as in Figure 12).

As concluded in §6–8, the initial fluid density in queue $q(0,x)$ does not play a role in determining the system's asymptotic behavior if the system is overloaded for all $t \ge 0$, by the ALOM property in Theorem 7.3. In this example, however, $q(0,x)$ is also critical, because it determines the behavior of b as well.

By a minor modification of the reasoning used in §G.1, we can show that the system is overloaded for all $t \ge 1/\mu$. Let $0 \le t_1 \le 1/\mu$ be the

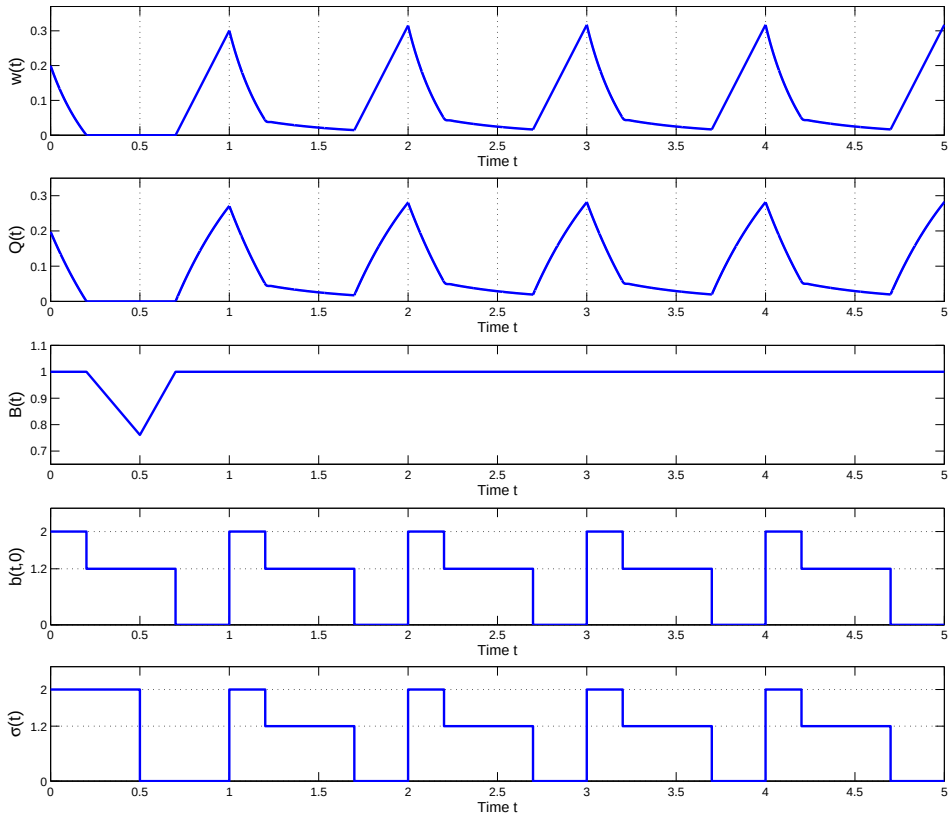

FIG 13. *The dynamics of the system performance of the example in Theorem 9.1 that has the same initial fluid density in service but* $w(0) = 0.2$ *instead of* $w(0) = 2$.

time at which the system switches from overloaded to underloaded intervals in $[0, 1/\mu]$. First, we can establish a similar (strict) monotonicity result. With $w(0) = 0.2$, we can show that $w(1) \approx 0.3 > w(0)$, which implies that $Q(1/\mu + t_1) > 0$. Since $\sigma(t + 1/\mu) = b(t, 0)$ for $0 \le t \le 1/\mu$, we have $b(t+1/\mu, 0) = b(t, 0)$. Therefore, the system is overloaded in $[1/\mu, 2/\mu]$. Using an inductive argument, we can show that $w(n+1) > w(n)$ and $\sigma(t+n/\mu) = b(t + n/\mu, 0) = b(t, 0)$ so that the system is overloaded in $[n, n + 1]$ for all $n \ge 1$.

APPENDIX H: MORE ON FIRST PASSAGE TIMES

As an analog of Example 3.1 in §3, below we give another counterexample for first passage times with $B(0) < 1$.

EXAMPLE H.1 (counterexample on first passage times with $B(0) < 1$). Suppose that $\lambda > \mu = 1$. Let $b(0, x) = \lambda$ for $1 - (1/\lambda) \le x \le 1 - 1/2\lambda$ and

$b(0,x) = 0$ otherwise, so that $B(0) = 1/2$, $b(t,0) = \lambda$, $0 \le t < 1/\lambda$, and $b(t,0) = 0$, $1/\lambda \le t < 1$, $B(t) = 1/2 + \lambda t$ for $0 \le t \le 1/2\lambda$ and $B(t) = 1$ for $t > 1/2\lambda$. Therefore, $T^* = t^* = 1/2\lambda$.

For $n \ge 1$, let $\{B_n(0,y) : 0 \le y \le 1\}$ be deterministic. To be a legitimate sample path for a queueing system, $B_n(0,y)$ must be nondecreasing and integer-valued as well as satisfy $0 \le B_n(0,y) \le n$. Thus, let $B_n(0,y) \equiv \lfloor B_n^f(0,y) \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $\bar{B}_n^f(0,y) \equiv n^{-1} B_n^f(0,y) \equiv \int_0^y b_n(0,x)\,dx$, where $b_n(0,x) = ((n+1)/n)\lambda$, $1 - ((n-1)/n\lambda) \le x \le 1 - ((n-1)/2n\lambda)$, and $b_n(0,x) = 0$ otherwise. First, observe that $\bar{B}_n^f(0,1/\mu) = (n^2-1)/2n^2 < 1/2$ for all $n \ge 1$. Second, observe that we have $0 \le \bar{B}_n^f(0,y) - \bar{B}_n(0,y) \le 1/n$ for all y and n. Hence, $\bar{B}_n(0,1/\mu) \le \bar{B}_n^f(0,1/\mu) < 1/2$ for all $n \ge 1$. Nevertheless, $\bar{B}_n(0,\cdot) \to B(0,\cdot)$ as $n \to \infty$. On the other hand, consider a deterministic arrival process with rate $n\,\lambda$. Then $B_n(1/2\lambda) = B_n(0) + N_n(1/2\lambda) = \lfloor (n^2-1)/2n^2 \rfloor + \lfloor (n-1)/2 \rfloor = n - 1 < n$ (note there is no departure in $[1, 1/2\lambda]$). Also, $S_n(t) - S_n(1/2\lambda) = \lfloor (n+1)\lambda\,(t - 1/2\lambda) \rfloor \ge \lfloor n\,\lambda\,(t - 1/2\lambda) \rfloor = N_n(t) - N_n(1/2\lambda)$ for $(n-1)/2n\lambda \le t \le (n-1)/n\lambda$. Therefore, the system is underloaded for $0 \le t \le 1/\lambda$. Hence, $T_n = T_n^* = 1/\lambda$ for all $n \ge 1$, in contrast to $t^* = T^* = 1/2\lambda$.

APPENDIX I: A TWO-POINT SERVICE DISTRIBUTION

We next generalize the PSS result of the $G/D/s + GI$ fluid queue discussed in §8 to the $G/GI/s+GI$ model with a special two-point service-time distribution, in particular, to a two-point distribution where one of the two points is 0. We also give an analog of Corollary 8.3 where analytic expressions for the PSS functions are available when the system is initially empty and the abandonment distribution is exponential. The proofs are similar to the proofs of Theorem 8.1 and Corollary 8.3.

COROLLARY I.1 (PSS for the overloaded $G/D/s+GI$ fluid model). *Consider the stationary $G/GI/s + GI$ fluid model with parameter (λ, μ, p, s, F) where $\rho \equiv \lambda/s\mu > 1$ and the service distribution G is a two-point distribution with $P(X = 1/p\mu) = p$ and $P(X = 0) = 1 - p$ for $0 < p \le 1$ such that the mean service time is $1/\mu$. Suppose that Assumption 12 is satisfied. If $b(T^*, x) = s\mu$, $0 \le x \le 1/\mu$, then there exists a constant function $\mathcal{P}^*$ such that*

$$\|\Psi_\tau^{(n)}(\mathcal{P}) - \mathcal{P}^*\| \to 0 \quad \text{as} \quad n \to \infty. \tag{I.1}$$

for all $\tau > 0$. Otherwise, the fluid performance $\mathcal{P}$ is asymptotically periodic

with period $1/\mu$, i.e., there exists a periodic function $\mathcal{P}^*$ with period $1/\mu$ such that (I.1) holds for $\tau \equiv 1/\mu$.

COROLLARY I.2 (explicit expressions for the PSS with the special two–point service times). *Consider the $G/D/s+M$ fluid queue with two-point service distribution given in Corollary I.1. If $\rho \equiv \lambda/s\mu > 1$ and the system is initially empty, then the system is overloaded in the PSS with performance functions given in two parts ($[0, 1/p\mu - s/p\lambda]$ and $(1/p\mu - s/p\lambda, 1/p\mu]$) of a cycle $0 \le t \le 1/p\mu$:*

(a) *In the first part of the PSS cycle, (i.e., for $0 \le t \le 1/p\mu - s/p\lambda$),*

$$\begin{aligned} w(t) &= t + \tilde{w}, \\ Q(t) &= \frac{\lambda}{\theta}\left[1 - \left(\frac{1 - e^{-\theta\, s/p\lambda}}{1 - e^{-\theta/p\mu}}\right) e^{-\theta\, t}\right], \\ b(t,x) &= \lambda \cdot 1_{\{t \le x \le t + s/p\lambda\}}, \\ \sigma(t) &= b(t,0) = 0, \end{aligned}$$

where

$$\tilde{w} = \frac{1}{\theta} \log\left(\frac{1 - e^{-\theta/p\mu}}{1 - e^{-\theta\, s/p\lambda}}\right) \ge 0, \tag{I.2}$$

(b) *In the second part of the PSS cycle, (i.e., for $1/p\mu - s/p\lambda < t \le 1/p\mu$),*

$$\begin{aligned} w(t) &= -\frac{1}{\theta} \log\left(1 + \left(\frac{1 - e^{\theta(1/\mu - s/\lambda)/p}}{1 - e^{-\theta/p\mu}}\right) \cdot e^{-\theta\, t}\right), \\ Q(t) &= \frac{\lambda}{\theta}\left(\frac{e^{\theta(1/\mu - s/\lambda)/p} - 1}{1 - e^{-\theta/p\mu}}\right) e^{-\theta\, t} \\ b(t,x) &= \lambda \cdot 1_{\{0 \le x \le t - 1/p\mu + s/p\lambda\} \cup \{t \le x \le 1/p\mu\}}, \\ \sigma(t) &= b(t,0) = \lambda. \end{aligned}$$

Moreover, for $0 \le t \le 1/p\mu$,

$$B(t) = s, \quad q(t,x) = \lambda \cdot 1_{\{0 \le x \le w(t)\}}, \quad \alpha(t) = \theta\, Q(t).$$

PROOF. In a cycle $[0, 1/p\lambda]$, $w(t) = \tilde{w} + t$ for $0 \le t \le 1/p\mu - s/p\lambda$ and $w(t)$ solves ODE $w'(t) = 1 - 1/e^{-\theta w(t)}$ with $w(1/p\mu - s/p\lambda) = \tilde{w} + 1/p\mu - s/p\lambda$ for $1/p\mu - s/p\lambda \le t \le 1/p\lambda$, where $\tilde{w} \ge 0$ is both the starting and the ending value of $w(t)$ in each cycle. Similar to the proof of Corollary 8.3, solving this ODE in $[1/p\mu - s/p\lambda, 1/p\mu]$ and set $w(1/p\mu) = \tilde{w}$ yields (I.2). □

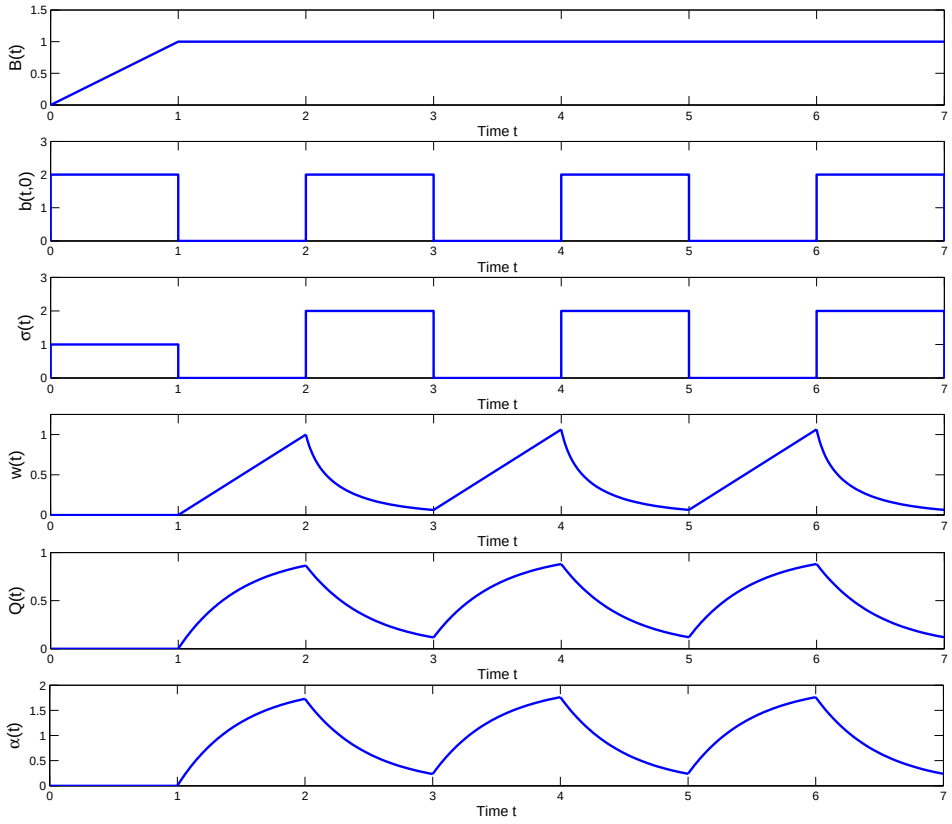

FIG 14. *Performance of the fluid model with the special two-point service distribution and* $s = \mu = 1$, $p = 1/2$, $\lambda = \theta = 2$.

REMARK I.1. *Theorem 8.1 and Corollary 8.3 in the main paper arise as special cases of Corollary I.1 and I.2 when* $p = 1$.

We next compare the fluid performance with simulation estimations of large-scale queueing systems. We consider the overloaded ($\rho > 1$) $G/GI/s+M$ example with two-point service distribution such that $P(X = 1/p\mu) = p$ and $P(X = 0) = 1 - p$. Let the system be initially empty. We plot the system performance ($Q(t)$, $B(t)$, $w(t)$, $b(t, 0)$, $\alpha(t)$, $\sigma(t)$) in Figure 14. We let $\lambda = \theta = 2$, $p = 1/2$ and $s = \mu = 1$. We have $\tilde{w} \approx 0.0635$ when $\theta = 2$ from (I.2), which can be verified by Figure 14.

In Figure 15 we compare our fluid approximation (the dashed red lines) with simulation estimates (the solid blue lines) of a large-scale $G/GI/s+M$ queueing system that has arrival rate $n\lambda$ and ns servers. We plot (i) the elapsed waiting time of the customer at the head of the line $W_n(t)$, (ii) the

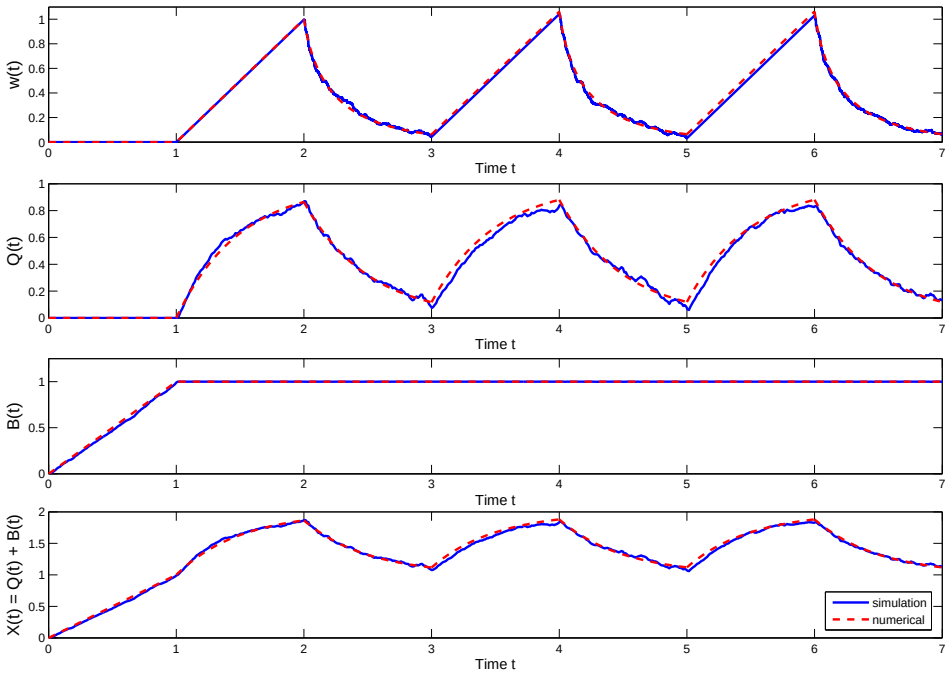

FIG 15. *A comparison of the fluid model with the special two-point service times with a simulation of a corresponding large-scale queue system.*

scaled number of customers waiting in queue $\bar{Q}_n(t) \equiv Q_n(t)/n$ and (iii) the scaled number of customers in service $\bar{B}_n(t) \equiv B_n(t)/n$. We plot single sample paths of these processes with $n = 1000$. Figure 15 shows that the fluid approximation is effective.

However, from simulation experiments of corresponding queueing models, we conclude that the fluid model with other kinds of two-point service distributions must not converge to a PSS.

To illustrate, in Figure 16, we plot single sample paths of processes W_n and Q_n of four two-point distributions: (a) $P(S = 1) = 1$ (red dashed curves), (b) $P(S = 0) = P(S = 2) = 1/2$ (blue dashed curves), (c) $P(S = 0.2) = P(S = 1.8) = 1/2$ (yellow solid curves) and (d) $P(S = 0.8) = P(S = 1.2) = 1/2$ (black solid curves), with $n = 1000$ in interval $[0, 16]$. The traffic intensity is $\rho = \lambda/n\mu = 2$ here. Figure 16 shows that the periodic structure is preserved only for case (a) and (b), where he have established periodic behavior of the associated fluid model. Cases (c) and (d) involve two-point distributions, but the periodic structure fades away very quickly and the fluctuations decrease substantially. Thus we conclude that the corresponding fluid models must not have asymptotically periodic structure.

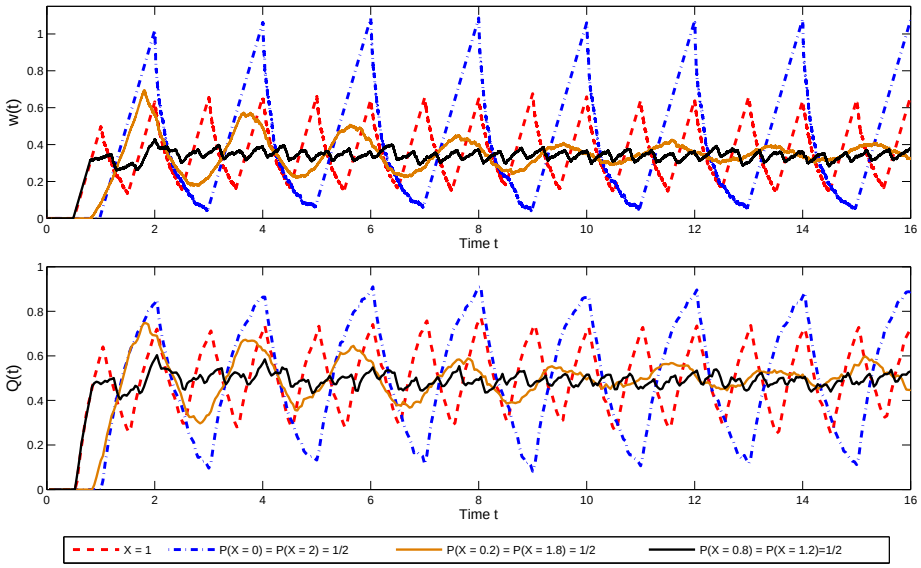

FIG 16. *A comparison of simulations of large-scale queue systems with two-point service-times distributions, all having mean* 1.

APPENDIX J: NEARLY DETERMINISTIC SERVICE TIMES

It is natural to wonder to what extent our results for deterministic service times apply to other service-time distributions that are nearly deterministic, but not fully deterministic. We investigated this question by conducting simulation experiments of corresponding queueing systems with nearly deterministic service times.

For the experiments reported here, as before, we consider the $M/GI/n+M$ queueing model with $\lambda = 2$, $\mu = 1$ and $\theta = 2$, but now we let the service-time distribution be nearly deterministic. For all examples, $E[S] = 1/\mu = 1$ and we make $Var[S]$ small, where S is a generic service time.

In our examples now we consider two kinds of service-time distributions, both of which have small variance: (i) Erlang-N and (ii) a two-point distribution, taking the values $1/\mu \pm \delta$ with probability 1/2. For the Erlang-N service times, the variance (and C^2) is $Var(S) = 1/N$. We plot single sample paths of process W_n with $N = 100$ and $N = 5000$ in Figure 17, with smaller n ($n = 100$) and larger T ($T = 100$). The periodic behavior is preserved for the case $N = 5000$ but not for $N = 100$.

For the two-point distribution at $1/\mu \pm \delta$ with 1/2 probability, the variance $Var(S) = \delta^2$. We plot single sample path of process W_n with $\delta = 0.1$ and

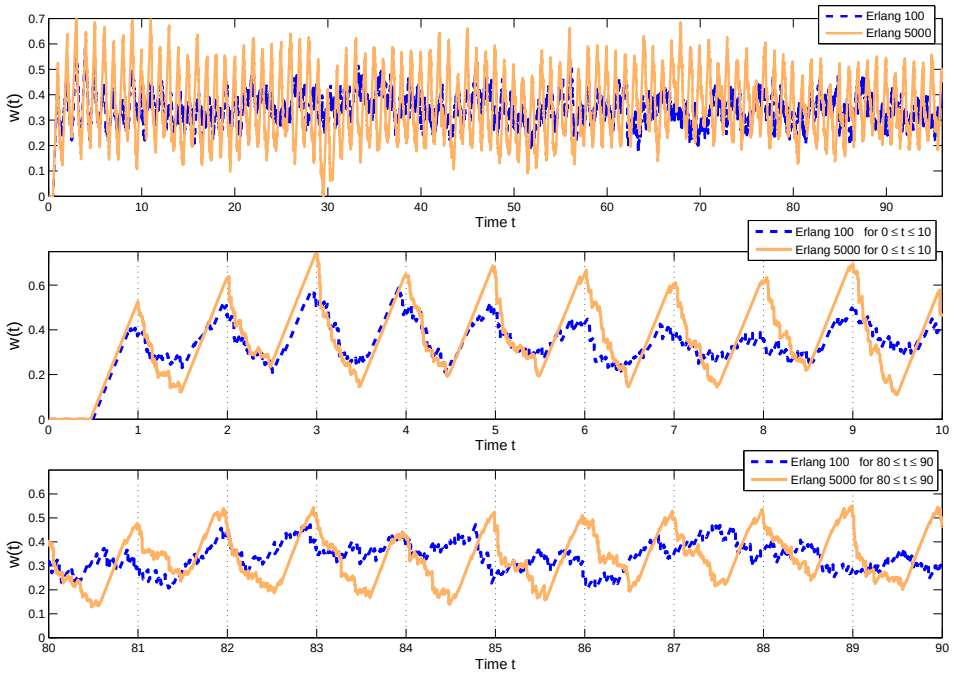

FIG 17. *Simulation estimates of the head-of-line waiting times W_n in an $G/E_N/s + M$ many-server queue with Erlang-N service, with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $\rho = 2$, $n = 100$, $T = 100$ in two cases: (i) $N = 100$; (ii) $N = 5000$.*

$\delta = 0.01$ in Figure 18, with $n = 100$, $T = 100$. Again, the periodic behavior is preserved for the case $\delta = 0.01$ but not for $\delta = 0.1$.

From these experiments, we conclude, first, that over suitably short finite intervals, both the large-scale many-server queueing systems and the approximating fluid models with nearly deterministic service-time distributions should behave much like the fluid model with deterministic service times and, second, that the asymptotic behavior of the approximating fluid model will not be periodic. We conclude that a small amount of variability in the service time distribution will eventually break up the periodic behavior (provided of course we do not have the special two-point distribution considered in the previous section).

More generally, we conclude that the quality of the approximation provided by the fluid model with D service over finite time intervals $[0, T]$ should improve as the service-time distribution becomes more nearly deterministic, e.g., as the variance $Var(S)$ decreases. We conjecture that again the order of the limits cannot be interchanged: If we first let $Var(S) \downarrow 0$, e.g., by letting $N \uparrow \infty$ in the E_N distribution, and then afterwards let $t \to \infty$, then we

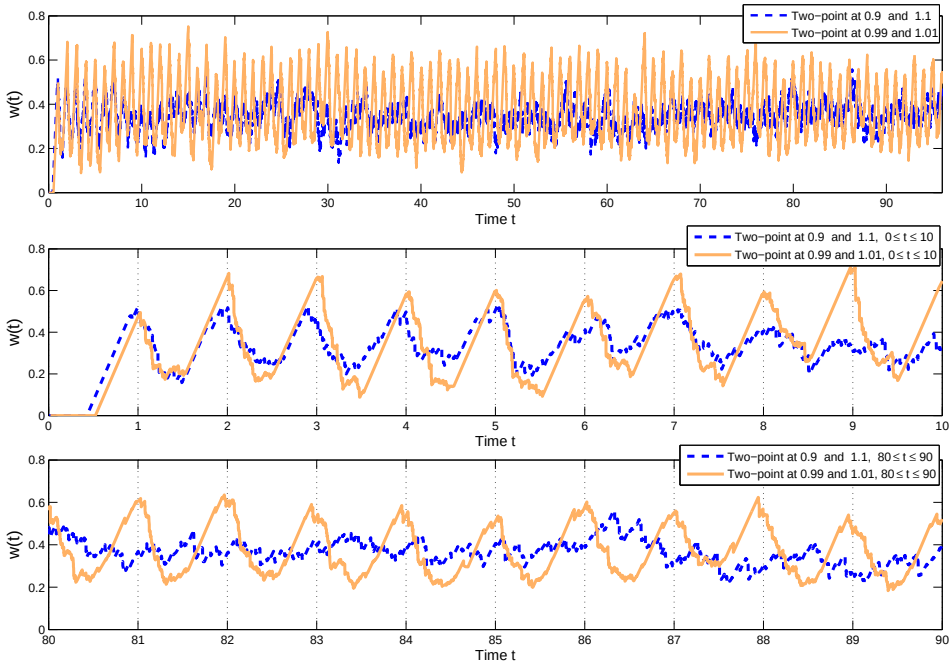

FIG 18. *Simulation estimates of the head-of-line waiting times W_n in a $G/TP/s + M$ many-server queue with a two-point (TP) service-time distribution taking values $1/\mu \pm \delta$ with 0.5 probability, with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $\rho = 2$, $n = 100$, $T = 100$ in two cases: (i) $\delta = 0.1$; (ii) $\delta = 0.01$.*

have the asymptotic PSS established in this paper. On the other hand, if we first let $T \to \infty$ for any fixed N in the Erlang E_N distribution, and then let $N \uparrow \infty$, then our simulation experiments lead us to conjecture that the performance converges to the unique steady state of the fluid model.

Even more generally, we conclude that when s system tends to behave in a deterministic or nearly deterministic way, that the transient behavior over suitably short time intervals may not be well captured by long-run stationary or steady-state descriptions.

References.

[1] ASMUSSEN, S. (2003). *Applied Probability and Queues*, second edition, Springer, New York. MR1978607

[2] BASSAMBOO, A., RANHAWA, R. S. (2010). On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Res.* **58** 1398–1413. MR2560543

[3] BOROVKOV, A. A. (1967). On limit laws for service processes in multi-channel systems (in Russian). *Siberian Math J.* **8** 746–763. MR0222973

[4] CHOUDHURY, G. L., MANDELBAUM, A., REIMAN, M. I., WHITT, W. (1997). Fluid

and diffusion limits for queues in slowly changing environments. *Stochastic Models* **13** 121–146. MR1430932

[5] Courtois, P. J. (1977). *Decomposability*, Academic Press, New York. MR0479702

[6] Erramilli, A., Forys, L. J. (1991). Oscillations and chaos in a flow model of a switching system. *IEEE J. Sel. Areas Commun.* **9** 171–178.

[7] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, second edition, Wiley, New York. MR0270403

[8] Gamarnik, D., Zeevi, A. (2006). Validity of heavy-traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Prob.* **16** 56–90. MR2209336

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

[10] Gibbens, R. J., Hunt, P. J., Kelly, F. P. (1990). Bistability in communication networks. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammersely*, G. Grimmett and D. Welsh (eds.), Oxford University Press, 113–127. MR1064558

[11] Glynn, P. W., Whitt, W. (1991). A new view of the heavy-traffic limit for infinite-server queues. *Adv. Appl. Prob.* **23** 188–209. MR1091098

[12] Gurvich, I. (2009). Validity of heavy-traffic steady-state approximations in multi-class queueing networks: sufficient conditions involving state-space collapse. Working paper, Northwestern University.

[13] Halfin, S., Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Res.* **29** 567–588. MR0629195

[14] Jelenkovic, P., Mandelbaum, A., Momcilovic, P. (2004). Heavy-traffic limits for queues with many deterministic servers. *Queueing Systems* **47** 53–69. MR2074672

[15] Kang, W., Ramanan, K. (2010). Fluid limits of many-server queues with reneging. *Ann. Appl. Prob.* **20** 2204–2260. MR2759733

[16] Kaspi, H., Ramanan, K. (2011). Law of large numbers limits for many-server queues. *Ann. Appl. Prob.* **21** 33–114. MR2759196

[17] Krichagina, E. V., Puhalskii, A. A. (1997). A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* **25** 235–280. MR1458591

[18] Lindvall, T. (1992). *Lectures on the Coupling method*, Wiley, New York. MR1180522

[19] Liu, Y., Whitt, W. (2010a). The $G_t/GI/s+GI$ many-server fluid queue. Columbia University, NY, NY, 2010. http://www.columbia.edu/~ww2040/allpapers.html

[20] Liu, Y., Whitt, W. (2010b). A network of time-varying many-server fluid queues with customer abandonment. *Operations Res.*, forthcoming. http://www.columbia.edu/∼ww2040/allpapers.html

[21] Liu, Y., Whitt, W. (2011). Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with customer abandonment. *Queueing Systems* **67** 145–182. MR2771198

[22] Miller, D. (1972). Existence of limits in regenerative stochastic processes. *Ann. Math. Statist.* **43** 1273–1280. MR0312592

[23] Pang, G., Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65** 325–364. MR2671058

[24] Reed, J. (2009). The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Prob.* **19** 2211-2269. MR2588244

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

[26] Sigman, K., Whitt, W. (2011a). Heavy-traffic limits for nearly deterministic queues. *J. Appl. Prob.*, forthcoming. http://www.columbia.edu/~ww2040/allpapers.html

[27] Sigman, K., Whitt, W. (2011b) Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Queueing Systems*, forthcoming. http://www.columbia.edu/~ww2040/allpapers.html
[28] Stolyar, A.L., Yudovina, E. Systems with large flexible server pools: Instability of "natural" load balancing. Bell Labs Technical Memo, December 2010. http://arxiv.org/abs/1012.4140
[29] Tak'acs, L. (1956). On a probability problem arising in the theory of counters. *Proc. Camb. Phil. Soc.* **52** 488–498. MR0081585
[30] Tak'acs, L. (1962). *Introduction to the Theory of Queues*, Oxford University Press, New York. MR0133880
[31] Whitt, W. (1972). Embedded renewal processes in the $GI/G/s$ queue. *J. Appl. Prob.* **9** 650–658. MR0341670
[32] Whitt, W. (1981). Comparing counting processes and queues. *Adv. Appl. Prob.* **13** 207–22. MR0595895
[33] Whitt, W. (1983). Untold horrors of the waiting room. What the equilibrium distribution will never tell about the queue-length process. *Management Science* **29** 395–408. MR0704592
[34] Whitt, W. (2002). *Stochastic-Process Limits*, Springer, New York. MR1876437
[35] Whitt, W. (2004). Efficiency-Driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50** 1449–1461.
[36] Whitt, W. (2005). Heavy-traffic limits for the $G/H2*/n/m$ queue. *Math. Oper. Res.* **30** 1–27. MR2125135
[37] Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research* **54** 37–54. MR2201245
[38] Willie, H. (1998). Periodic steady state of loss systems. *Adv. Appl. Prob.* **30** 152–166. MR1618825

Department of Industrial Engineering
and Operations Research,
Columbia University New York
New York 10027-6699,
E-mail: yl2342@columbia.edu