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.

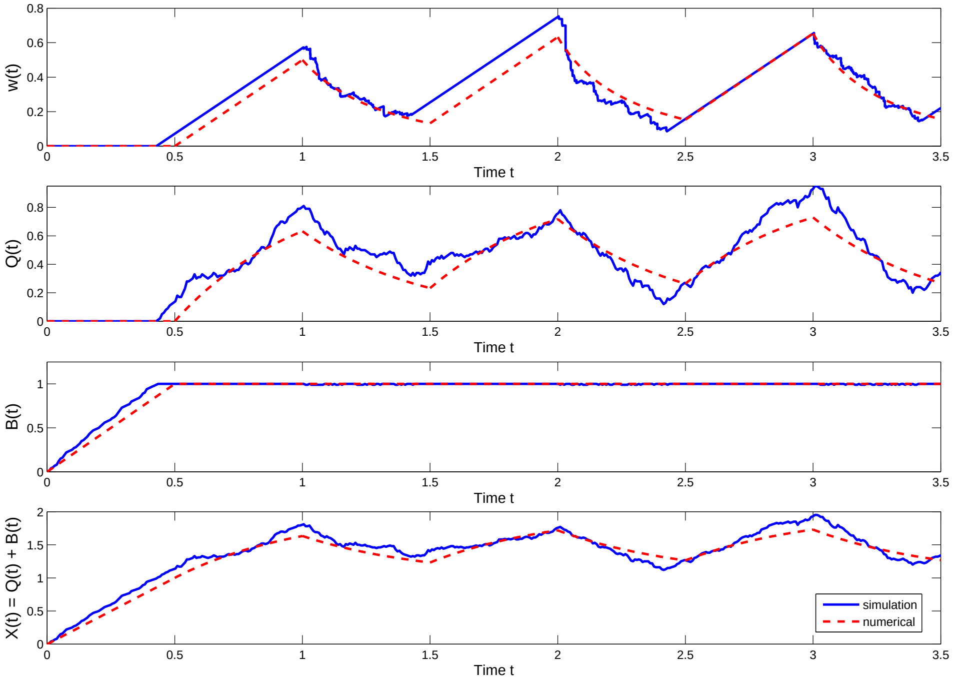

FIG 4. *Performance of the $G/D/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 100$.*

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

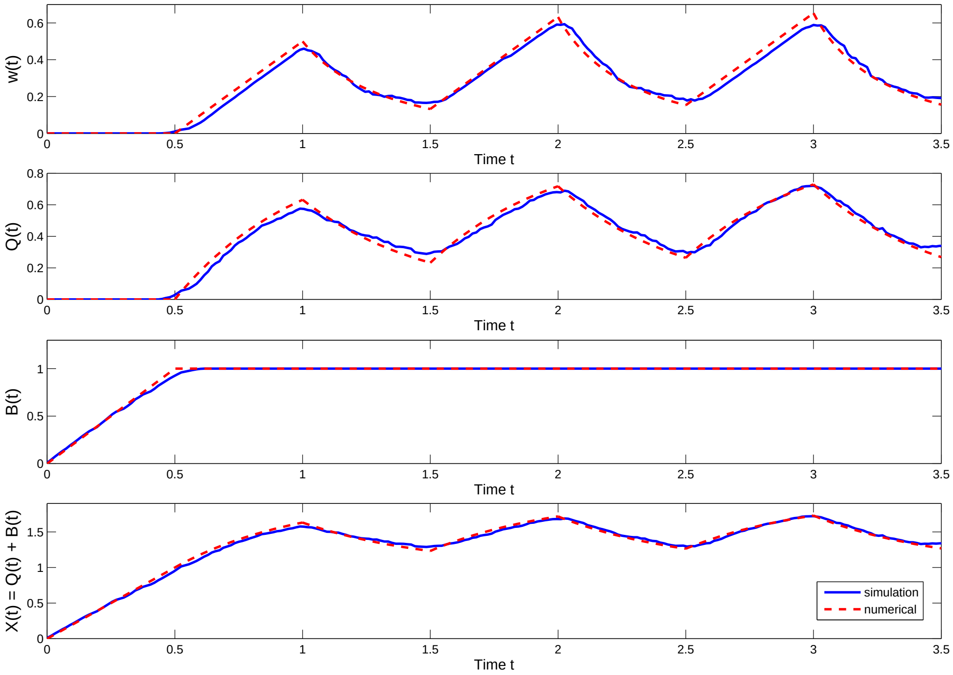

Fig 5. *Performance of the $G/D/s+M$ fluid model compared with simulation results: an average of 10 sample paths of the scaled queueing model based on $n=100$.*

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 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

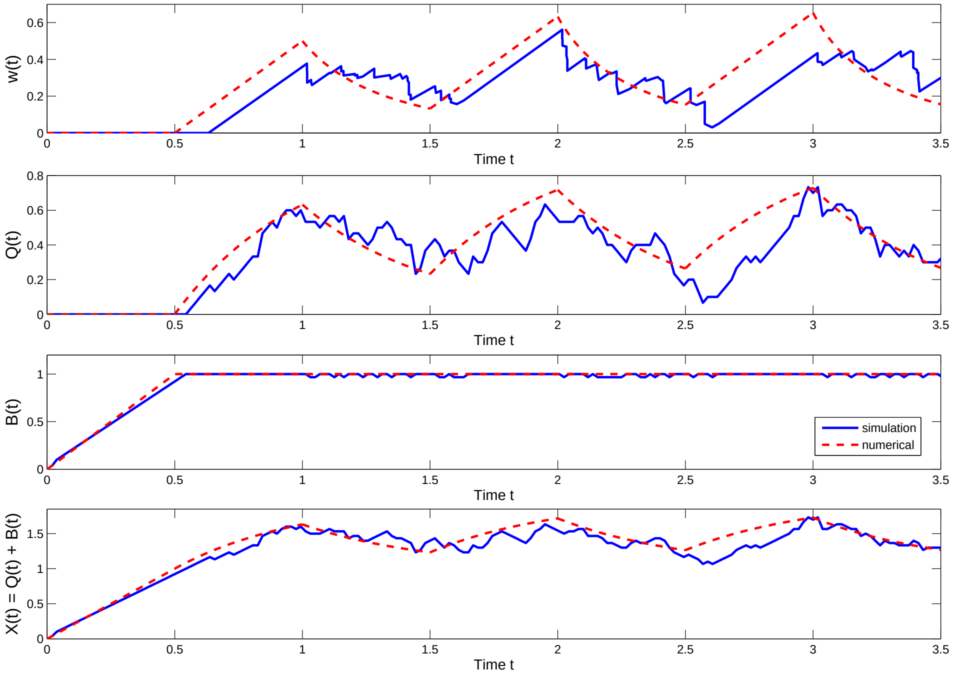

FIG 6. *Performance of the $G/D/s+M$ fluid model compared with simulation results: one sample path of the scaled queueing model for $n = 30$.*

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, 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

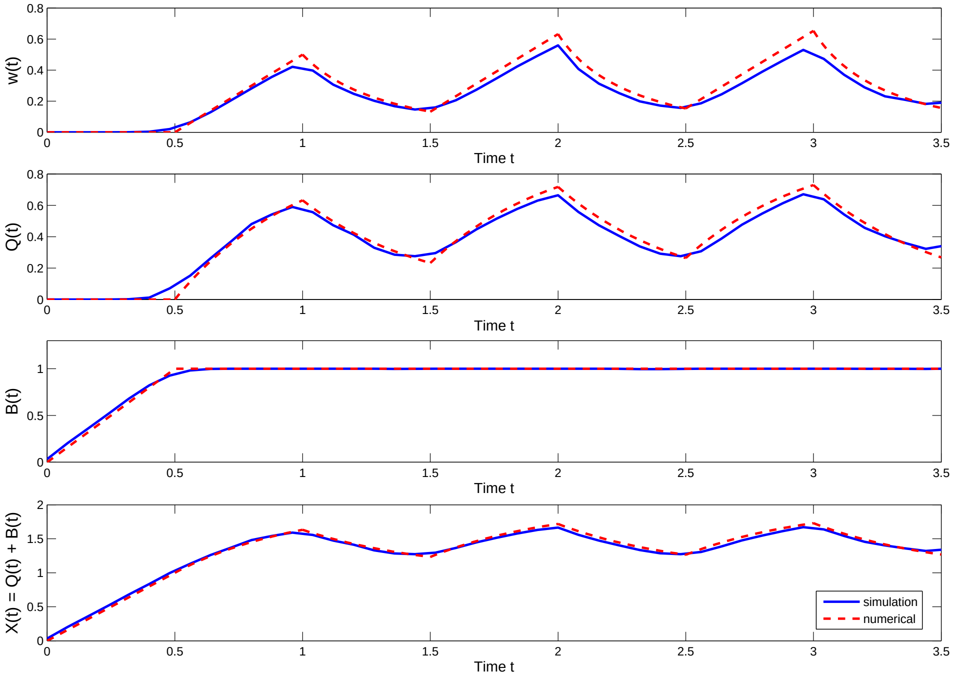

FIG 7. *Performance of the $G/D/s+M$ fluid model compared with simulation results: an average of 100 sample paths of the scaled queueing model based on $n=30$.*

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.

It remains to show $w_1(t) \leq w_2(t)$ for all $t \geq 0$. We will do a proof by contradiction. Hence suppose this inequality does not hold for some $t > 0$. Then continuity of w_1 and w_2 implies that there exists some $0 < t_1 < t$ such that $w_1(t_1) = w_2(t_1) \equiv \tilde{w}$. However, the ordering of $\tilde{q}_1$ and $\tilde{q}_2$ implies that $\tilde{q}_1(t_1, \tilde{w}) \leq \tilde{q}_1(t_1, \tilde{w})$. Hence the BWT ODE in Theorem 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

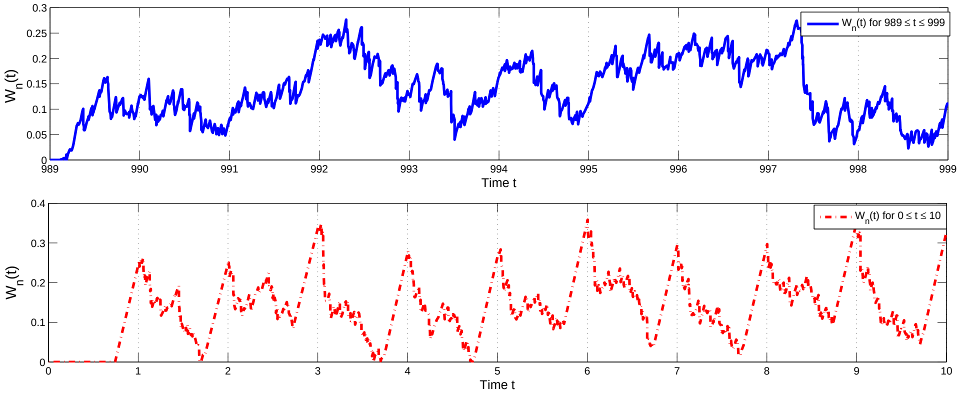

Fig 8. *Large-time periodic behavior of an overloaded $G/D/s+M$ queueing model: simulation estimates of the head-of-line waiting time W_n with $\lambda = 1.3$, $s = \mu = 1$, $\theta = 2$, $\rho = 1.3$, $n = 100$, $T = 1000$.*

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 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

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

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 $q_1(t,x) \leq q_2(t,x)$, $w_1(t) \leq w_2(t)$ and $\alpha_1(t) \leq \alpha_2(t)$ for all $t \geq 0$. Hence, we have

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

Flow conservation implies that

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

which yields

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

where the first inequality follows from (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

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

and

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

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

Finally, we treat $w(t)$. As above, it suffices to assume that we have the ordering in (7.8). We have $b(t,0) \geq b^{\downarrow}$ 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

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

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

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

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 $w_2(T^*) \geq \bar{w}$, 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 $w_2(t) \leq \bar{w} \vee (w_2(0) + T^*)$ for all $t \geq 0$. 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 $Q'(t) = \lambda - \alpha(t) - b(t,0) \leq \lambda - \alpha(t)$. Since $\alpha(t) \geq h_F^{\downarrow}\,Q(t)$, 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.

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 $b(0,x) = 1.5 \cdot 1_{\{0 \leq x \leq 1/2\}} + 0.5 \cdot 1_{\{1/2 \leq x \leq 1\}}$, $Q(0) = 0$ (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.

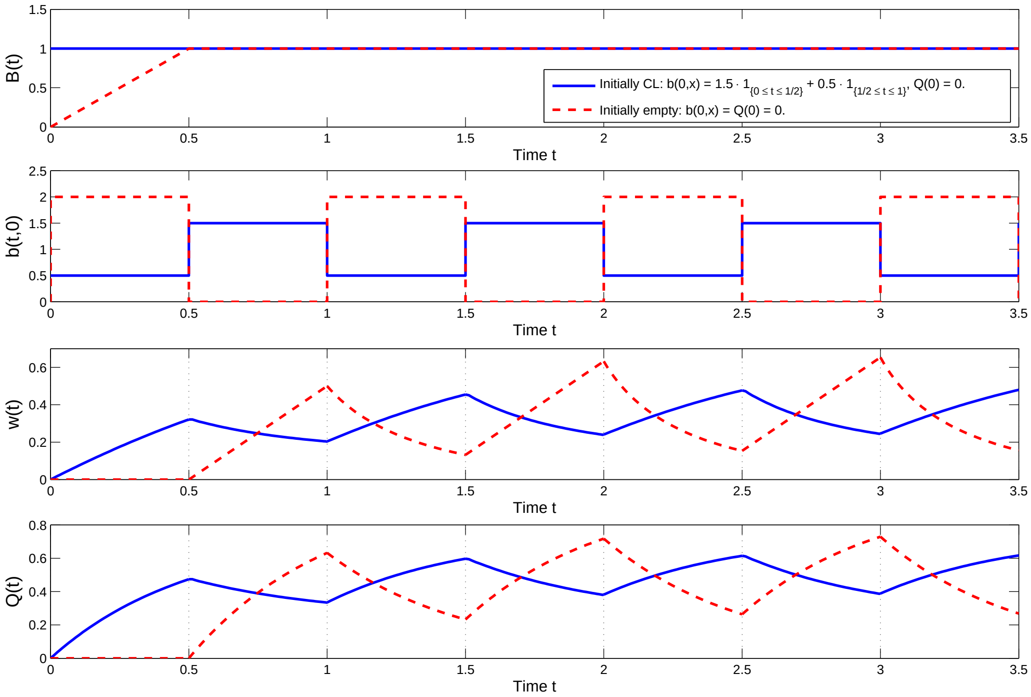

FIG 9. *A comparison of the PSS performance of the $G/D/s+M$ fluid queue with different initial conditions: (i) critically loaded with $b(0,x) = 1.5 \cdot 1_{\{0 \le x \le 1/2\}} + 0.5 \cdot 1_{\{1/2 \le x \le 1\}}$, $Q(0) = 0$ (the blue solid lines); (ii) starting empty (the red dashed lines).*

Let A be the generic abandonment time. A follows E_2 implies that $A = X_1 + X_2$ in distribution, where X_1 and X_2 are two iid exponential random variables. Moreover, $f(x) = \gamma^2\, x\, e^{-\gamma\, x}$, where γ is rate of X_1. If A follows H_2, then A is a mixture of two exponential random variables, i.e., $f(x) = p \cdot \theta_1\, e^{-\theta_1\, x} + (1-p) \cdot \theta_2\, e^{-\theta_2\, x}$, where θ_1 and θ_2 are the rates of these two exponential random variables, and $0 < p < 1$ is the sampling probability.

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

We let $\lambda = 2$, $\theta = 2$, $\mu = s = 1$. In Figure 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

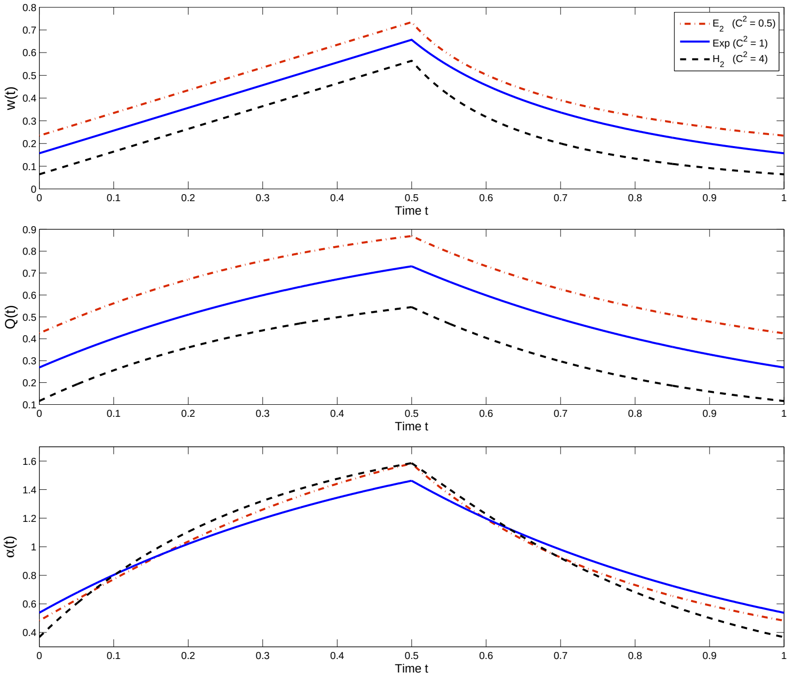

Fig 10. *A comparison of the PSS of the $G/D/s+GI$ fluid queues with different abandonment distributions: (i)E_2 (red dashed), (ii) M (blue solid) and (iii) H_2 (black dashed).*

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 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

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

TABLE 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.

of $w(t)$ in each cycle. Letting $v(t) \equiv t - w(t)$, we have for $1/\mu - s/\lambda \le t \le 1/\mu$,

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

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

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

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).

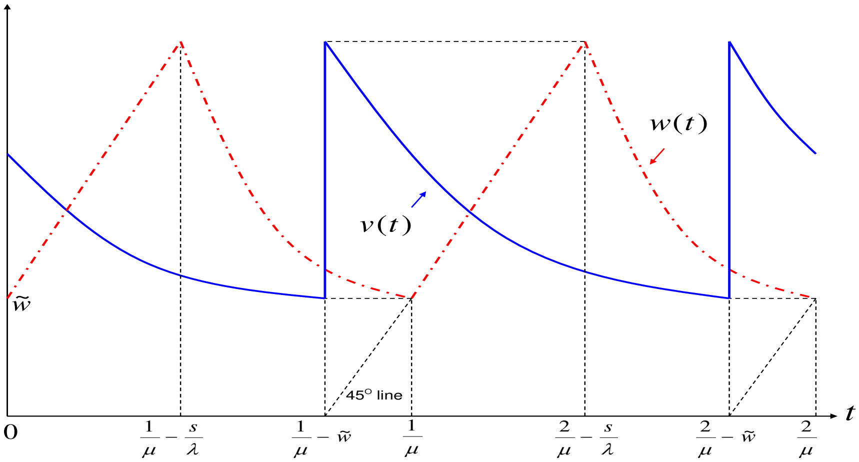

FIG 11. *PWT $v(t)$ and BWT $w(t)$ of the PSS of the $G/D/s+GI$ fluid queue.*

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 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

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

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

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

Letting $w(t) = 0$ yields that

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

that is the time at which the system becomes underloaded. Note that for $t_1^{(1)} < t \leq 1/2$, $\sigma(t) = 2 > 1.2 = b(t,0) = \lambda$, therefore, the fluid content in service decreases (linearly) with $B(t) = s - (\sigma(t) - b(t,0))\,(t - t_1^{(1)}) = 1 - 0.8(t - t_1^{(1)})$. For $t > 1/2$, $b(t,0) = \lambda = 1.2 > 0 = \sigma(t)$, $B(t)$ increases (liearly) with $B(t) = B(1/2) + (b(t,0) - \sigma(t))\,(t - 1/2) = 0.96 + 1.2(t - 1/2)$. So the system again becomes overloaded at $t_2^{(1)} = 0.53$ since $B(t_2^{(1)}) = 1 = s$. Moreover, $t_1^{(1)}$ and $t_2^{(1)}$ satisfy $1.2(t_2^{(1)} - 1/2) = 0.8(1/2 - t_1^{(1)})$. For $t_2 \le t \le 1$, 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.

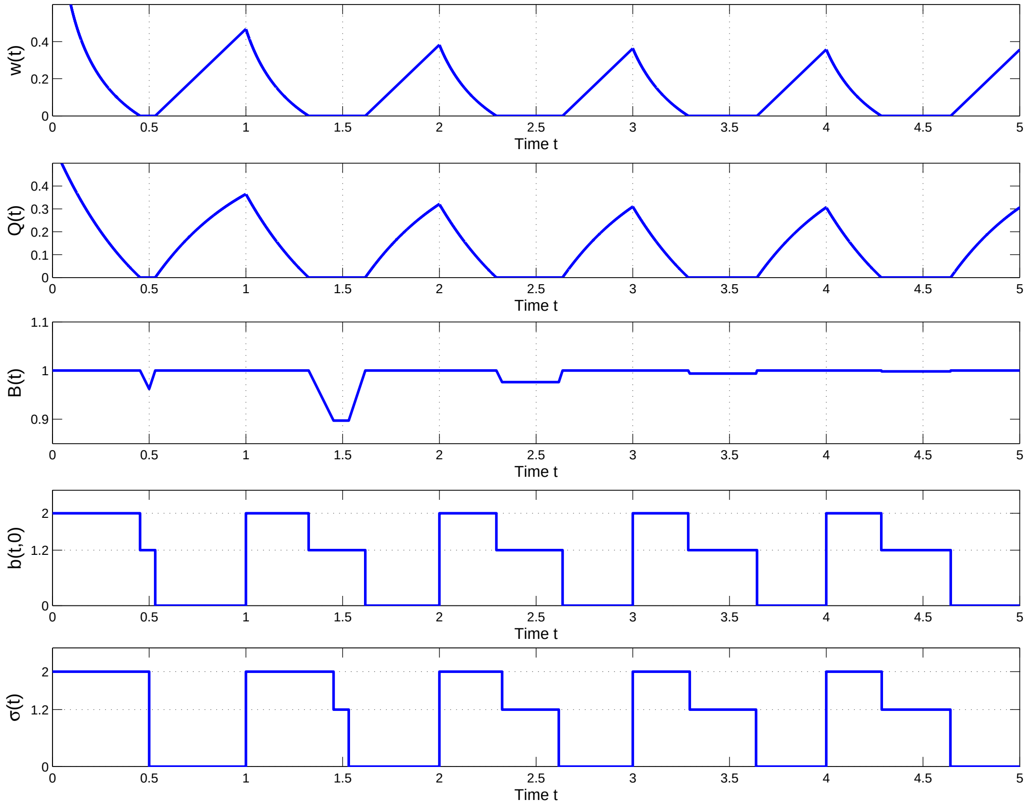

Fig 12. *The counterexample providing a fluid model that does not become (and stay) overloaded in finite time; it switches between overloaded and underloaded regimes infinitely often.*

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 $\sigma(t) = 2 \cdot 1_{\{0 \le t < t_1^{(1)}\}} + 1.2 \cdot 1_{\{t_1^{(1)} \le t \le t_2^{(1)}\}}$, 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

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

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

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 $t \geq T^*$ by changing the initial queue density. Moreover, we achieve this finite T^* in this example by *reducing* the initial fluid content in queue, not by increasing it.

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

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 $1 - (1/\lambda) \leq x \leq 1 - 1/2\lambda$ and

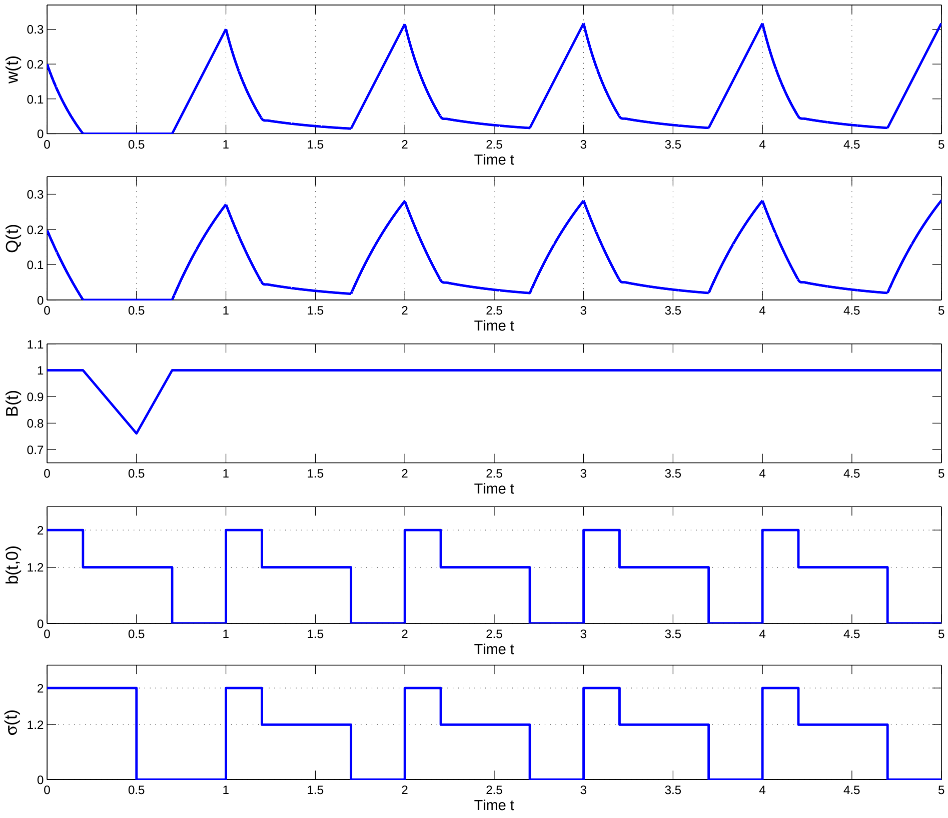

FIG 13. *The dynamics of the system performance of the example in Theorem 9.1 that has the same initial fluid density in service but $w(0) = 0.2$ instead of $w(0) = 2$.*

$b(0,x) = 0$ 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, $S_n(t) - S_n(1/2\lambda) = \lfloor (n+1)\lambda\,(t - 1/2\lambda) \rfloor \geq \lfloor n\,\lambda\,(t-1/2\lambda) \rfloor = N_n(t) - N_n(1/2\lambda)$ for $(n-1)/2n\lambda \leq t \leq (n-1)/n\lambda$. Therefore, the system is underloaded for $0 \leq t \leq 1/\lambda$. Hence, $T_n = T_n^* = 1/\lambda$ for all $n \geq 1$, 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 $0 < p \leq 1$ such that the mean service time is $1/\mu$. Suppose that Assumption 12 is satisfied. If $b(T^*, x) = s\mu$, $0 \leq x \leq 1/\mu$, then there exists a constant function $\mathcal{P}^*$ such that*

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

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 $0 \leq t \leq 1/p\mu$:*

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

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

where

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

(*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). □

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

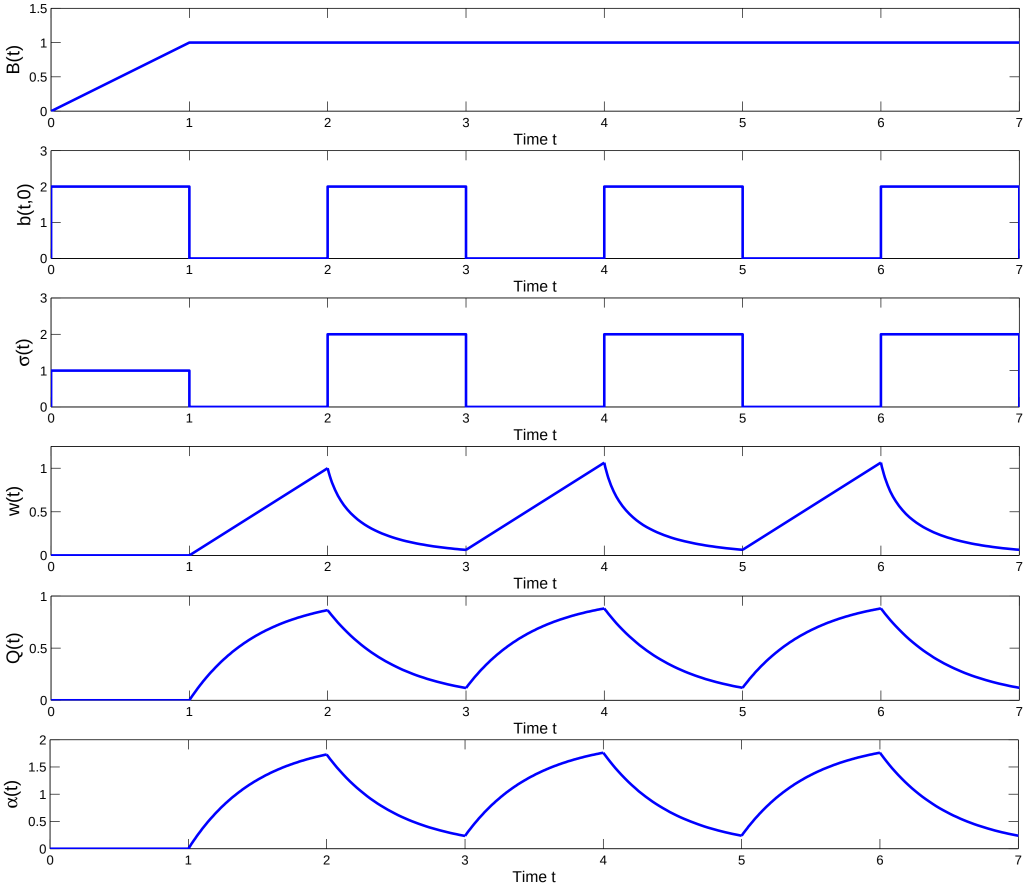

FIG 14. *Performance of the fluid model with the special two-point service distribution and* $s = \mu = 1$*,* $p = 1/2$*,* $\lambda = \theta = 2$*.*

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 $n\,s$ servers. We plot (i) the elapsed waiting time of the customer at the head of the line $W_n(t)$, (ii) the scaled number of customers waiting in queue $\bar{Q}_n(t) \equiv Q_n(t)/n$ and (iii) the scaled number of customers in service $\bar{B}_n(t) \equiv B_n(t)/n$. We plot single sample paths of these processes with $n = 1000$. Figure 15 shows that the fluid approximation is effective.

However, from simulation experiments of corresponding queueing mod-

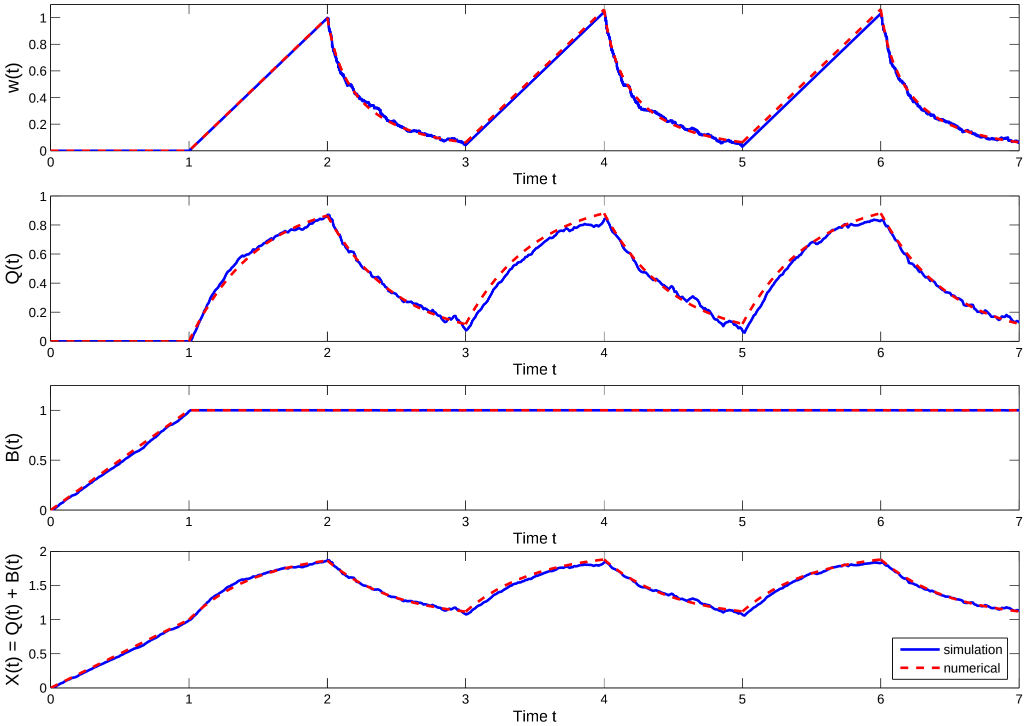

Fig 15. *A comparison of the fluid model with the special two-point service times with a simulation of a corresponding large-scale queue system.*

els, 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.

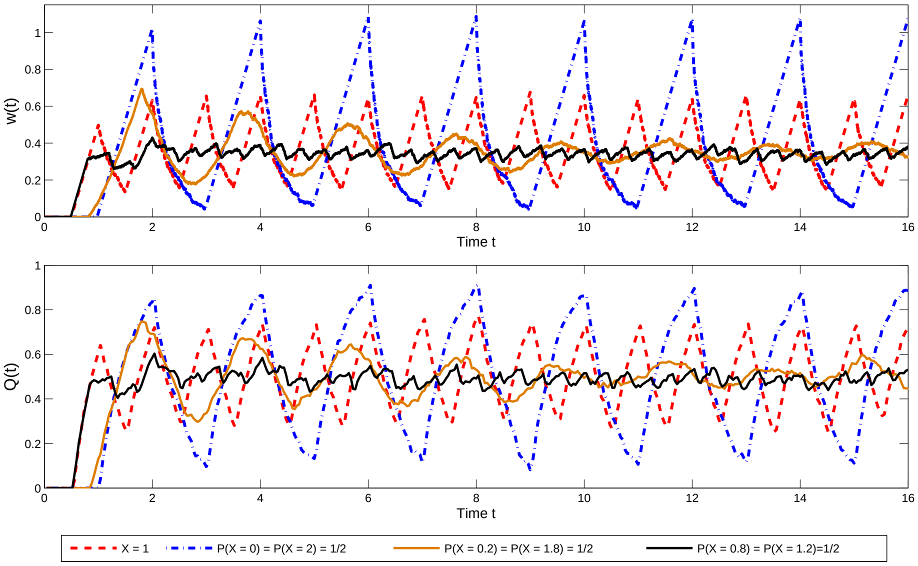

FIG 16. *A comparison of simulations of large-scale queue systems with two-point service-times distributions, all having mean* 1.

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 $\delta = 0.01$ in Figure 18, with $n = 100$, $T = 100$. Again, the periodic behavior

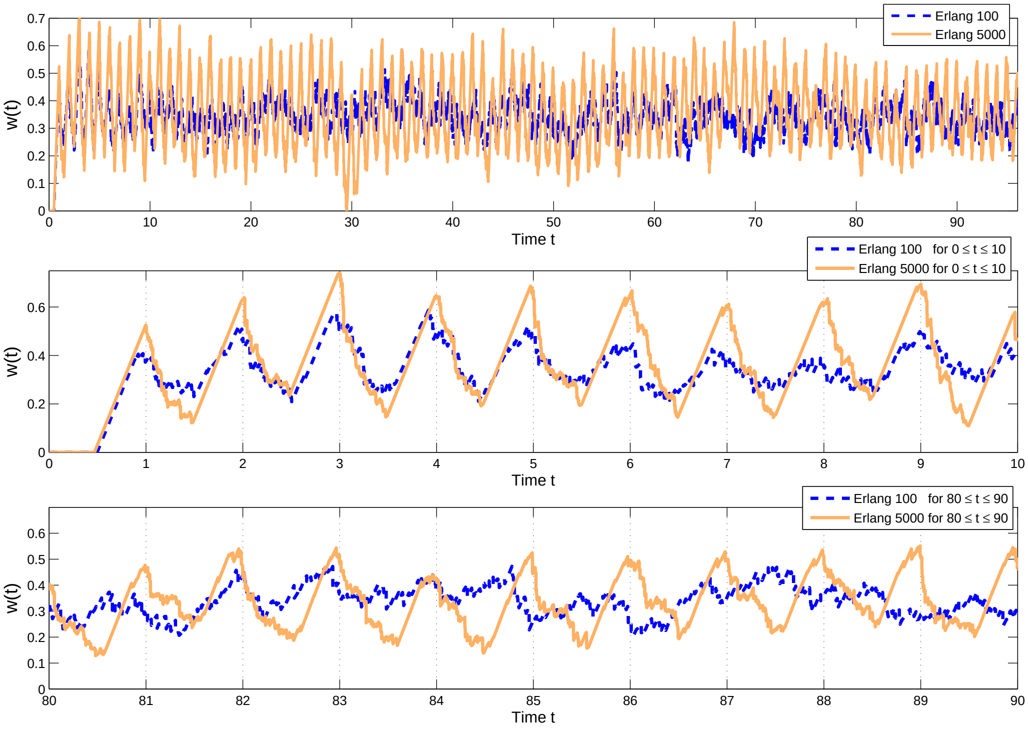

Fig 17. *Simulation estimates of the head-of-line waiting times W_n in an $G/E_N/s+M$ many-server queue with Erlang-N service, with $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $\rho = 2$, $n = 100$, $T = 100$ in two cases: (i) $N = 100$; (ii) $N = 5000$.*

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

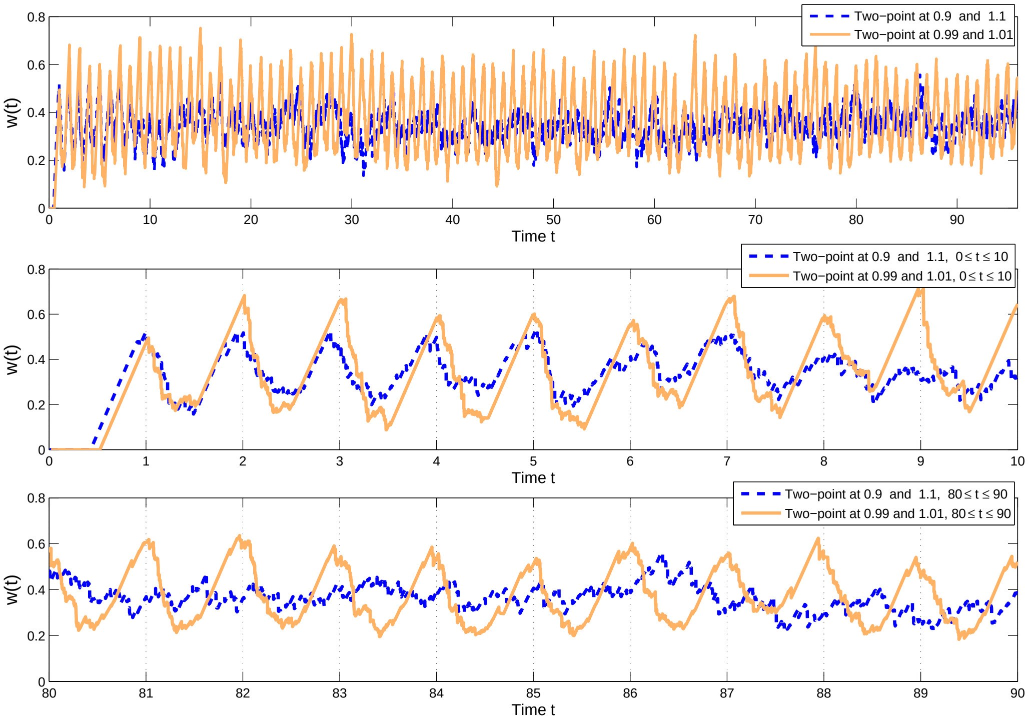

FIG 18. *Simulation estimates of the head-of-line waiting times* W_n *in a* $G/TP/s+M$ *many-server queue with a two-point (TP) service-time distribution taking values* $1/\mu \pm \delta$ *with 0.5 probability, with* $\lambda = 2$, $s = \mu = 1$, $\theta = 2$, $\rho = 2$, $n = 100$, $T = 100$ *in two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

$N \uparrow \infty$ in the E_N distribution, and then afterwards let $t \to \infty$, then we 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.

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