OPERATIONS RESEARCH
DOI 10.1287/opre.1110.0942ec

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment" by Yunan Liu and Ward Whitt, *Operations Research*, DOI 10.1287/opre.1110.0942.

E-Companion

This e-companion has six sections, presenting supporting material primarily in the order that it relates to the main paper. In §EC.1 we present the proofs for §3. In §EC.2 we present proofs for §4. In §EC.3 we present proofs for §5. In §EC.4 we present one proof for §6. In §EC.5, we make remarks about: (i) characterizing the isolated underloaded points in §3, (ii) representation of the fluid content B in an underloaded interval via an ODE, and (iii) the applied significance of the space of piecewise polynomials $\mathcal{P}_{m,n}$. In §EC.6 we compare the fluid model performance predictions to simulation results for a large-scale queueing system.

EC.1. Proofs for Section 3.

We need some basic regularity properties of Q and B, which will be valid with the assumptions in §2. For that purpose, we exploit two basic *flow-conservation equations*: (i) the queue content at time t equals the initial queue content plus input minus output to either abandonment or entering service, and (ii) the service content at time t equals the initial service content plus input minus output. However, the input enters the queue only when the system is overloaded; otherwise it directly enters service. Thus we have the following elementary bounds and the subsequent Lipschitz continuity.

PROPOSITION EC.1. (elementary bounds) $Q(t) + A(t) + E(t) \le Q(0) + \Lambda(t) < \infty$ *and*

$$B(t) + S(t) = B(0) + E(t) \le B(0) + Q(0) + \Lambda(t) < \infty,$$

so that Q, E, A, B and S are all bounded for $0 \le t \le T$.

Proof. The relations follow from flow conservation. The first relation is an inequality instead of an equality because input enters the queue instead of the service facility only when the system is overloaded. ∎

PROPOSITION EC.2. (Lipschitz continuity) *The functions S, E, B, A and Q are Lipschitz continuous.*

Proof. For a nonnegative real-valued function f on $[0,\infty)$, let $f_t^{\uparrow} \equiv \sup_{0\le y\le t} f(y)$. To treat S, recall that S is the integral of σ, where

$$\sigma(t) = B(t)\mu(t) \le s(t)\mu(t), \quad \text{so that} \quad \sigma(t) \le s_t^{\uparrow}\mu_t^{\uparrow}, \quad t \ge 0, \tag{EC.1}$$

and

$$|S(t+u) - S(t)| = \int_t^{t+u} \sigma(y)\, dy \le s_T^{\uparrow}\mu_T^{\uparrow} u, \quad 0 \le t \le t+u \le T. \tag{EC.2}$$

To treat E, recall that it is the integral of the rate fluid enters service, where the rate fluid enters service is either $\gamma(t) = \lambda(t)$ if the system is underloaded or $\gamma(t) = s'(t) + \sigma(t) = s'(t) + s(t)\mu(t)$ if the system is overloaded. Hence,

$$|E(t+u) - E(t)| \le \gamma_T^{\uparrow} u, \quad 0 \le t \le t+u \le T, \tag{EC.3}$$

where $\gamma_T^{\uparrow} \equiv \lambda_T^{\uparrow} \vee (|s'|_T^{\uparrow} + s_T^{\uparrow}\mu_T^{\uparrow}) < \infty$. By the second equation in Proposition EC.1,

$$B(t+u) - B(t) = (E(t+u) - E(t)) - (S(t+u) - S(t)), \tag{EC.4}$$

so that

$$|B(t+u) - B(t)| \le |E(t+u) - E(t)| + |S(t+u) - S(t)| \le (e_T^{\uparrow} + s_T^{\uparrow}\mu_T^{\uparrow})u \tag{EC.5}$$

for $0 \le t \le t+u \le T$.

Next we combine (4) with (9) to get

$$\alpha(t) = \int_0^{t\wedge w(t)} \lambda(t-x) f_{t-x}(x)\, dx + \int_{w(t)\wedge t}^{t} \frac{q(0, x-t) f_{t-x}(x)}{\bar{F}_{t-x}(x-t)}\, dx, \tag{EC.6}$$

so that, by applying Assumption 9, we get

$$\alpha(t) \le \alpha_t^{\uparrow} \equiv f^{\uparrow}\Lambda(t) + \frac{f^{\uparrow}}{\bar{F}^{\downarrow}(w(0))} Q(0) < \infty \tag{EC.7}$$

and

$$|A(t+u) - A(t)| \le \int_t^{t+u} \alpha(y)\, dy \le \alpha_T^{\uparrow} u, \quad 0 \le t \le t+u \le T. \tag{EC.8}$$

Finally, by the first relation in Proposition EC.1,

$$|Q(t+u) - Q(t)| \le |\Lambda(t+u) - \lambda(t)| + |E(t+u) - E(t)| + |A(t+u) - A(t)|$$

$$\leq (\lambda_T^{\uparrow} + \gamma_T^{\uparrow} + \alpha_T^{\uparrow})u, \quad 0 \leq t \leq t+u \leq T. \quad \blacksquare \tag{EC.9}$$

We now apply Proposition EC.2 to relate $\mathcal{S}$ to the zeros of $X - s$, where $X(t) \equiv Q(t) + B(t)$.

LEMMA EC.1. (zeros of $X - s$) $\mathcal{S} \subseteq Z_{X-s}$.

Proof. Since Q and B are continuous by Proposition EC.2 and s is continuous by assumption, $X - s$ is continuous. Since $X - s$ is continuous, if $X(t) - s(t) \neq 0$, then t cannot be an element of $\mathcal{S}$. $\blacksquare$

We now characterize the overloaded times.

LEMMA EC.2. (overloaded intervals) *With the possible exception of* 0 *and* T*, all overloaded times appear in intervals of positive length. Hence, underloaded sets consist of either single isolated points or intervals.*

Proof. If $t \in \mathcal{O}([0,T])$, then either (i) $X(t) - s(t) > 0$ or (ii) $X(t) - s(t) = 0$ and $\zeta(t) > 0$. In case (i), since $X - s$ is continuous by Proposition EC.2, there exists a neighborhood of t that is overloaded. In case (ii), since $\zeta(t) > 0$, we will have $X(t) - s(t) > 0$ in an interval $(t, t+\epsilon)$ for some positive ϵ. Since overloaded sets are necessarily intervals by Lemma EC.2, each underloaded set must fall between two overloaded intervals. $\blacksquare$

Proof of Theorem 1. We apply the results above. Since there can be at most countably many overloaded intervals of positive length in $[0,T]$, the isolated points are well defined and countably infinite. Since the isolated points are at most countably infinite, we can order them and reclassify them one by one. With that construction, we reduce the number of disjoint overloaded intervals by one at each step. Finally, all underloaded times appear in intervals too. $\blacksquare$

We now relate the zeros of ζ in (13) to the overloaded and underloaded intervals.

LEMMA EC.3. (zeros and intervals) *For each interval in the partition of* $[0,T]$ *into underloaded and overloaded intervals, there exists at least one zero or discontinuity point of* ζ*.*

Proof. First, consider the closure of an overloaded interval $[a,b]$. If ζ has one of its finitely many discontinuity points in $[a,b]$, then we are done. Suppose that ζ is continuous on the closed interval $[a,b]$. Necessarily, we have $X(a)-s(a)=X(b)-s(b)=0$, $\zeta(a+\epsilon)>0$ for all suitably small $\epsilon>0$ and $\zeta(b)\le 0$. First, we could have $\zeta(b)=0$ and we are done. If instead $\zeta(U(t))<0$, then there must exist t^* with $a<t^*<b$ such that $\zeta(t^*)=0$ by the intermediate value theorem. The reasoning is essentially the same in the closure of an underloaded interval, say $[a,b]$. If ζ has one of its finitely many discontinuity points in $[a,b]$, then we are again done. Suppose that ζ is continuous on the closed interval $[a,b]$. If either $\zeta(a)=0$ or $\zeta(b)=0$, then we are done. Hence we must have $\zeta(a)<0$. Since b is a switch point and ζ is continuous at b, we must have $\zeta(b)>0$. As before, there must exist t^* with $a<t^*<b$ such that $\zeta(t^*)=0$ by the intermediate value theorem. ∎

Proof of Theorem 2 Since the interval $[0,T]$ can be partitioned into at most countably many intervals that alternate between overloaded and underloaded after reclassifying isolated underloaded points as overloaded, the switch points can be placed in one-to-one correspondence with the internal boundary points (excluding 0 and T). Hence the number of switch points is equal to $n-1$, if the number of intervals in the paritition is n for some $n<\infty$. Otherwise both sets are countably infinite. Next, Lemma EC.3 implies that there is either a discontinuity point or a zero in every overloaded and underloaded interval. Since the number of intervals is 1 greater than the number of switches, we obtain the conclusion. To see that the bound is tight, consider the common case in which ζ is differentiable on $[0,T]$ and $\zeta(t)\neq 0$ at all switch times. Then ζ has a zero where it attains its maximum in each overloaded interval, while ζ has a zero where it attains its minimum in each overloaded interval. To have the bound an equality, let ζ have no other zeros. ∎

Proof of Theorem 3. First, any discontinuity points of ζ must be contained in the set of n interval boundary points. Hence, $\mathcal{D}_\zeta\le n$. On each of the n subintervals, ζ is a polynomial of order at most m. By the fundamental theorem of algebra, on each of these intervals the zero set is either a finite set of cardinality at most m or it is the entire subinterval. If $\zeta=0$ throughout the interval, then there can be at most a single switch in the interval, where $(Q(t),B(t))$ becomes $(0,s(t))$, after which it will remain there throughout the subinterval. In other words, the first subinterval is overloaded

and the second is underloaded, so this interval produces at most a single switch. We can thus treat this interval just like any of the others; we can act as if it produces at most m zeros. Hence, $\mathcal{D}_\zeta \le n$ and $Z_\zeta \le mn$. Finally, Theorem 2 implies that $|\mathcal{S}| \le mn + n - 1$, as claimed. ∎

Proof of Lemma 1. The Weierstrass approximation theorem implies that continuous functions can be approximated uniformly over bounded intervals by polynomials. That uniform approximation extends to $\mathbb{C}_p$ provided that the boundary points of the polynomial pieces of the function in $\mathcal{P}_{m,n}$ includes the finitely many discontinuity points of the function in $\mathbb{C}_p$. ∎

EC.2. Proofs for §4.

EC.2.1. Proof of Uniqueness in Theorem 4.

When the abandonment cdf's F_t are independent of t, the proof of uniqueness of the solution to the ODE (18) in Theorem 4 is the same as the proof of the corresponding part of Theorem 5.3 in Liu and Whitt (2010). However, that argument does not extend directly to time-varying abandonment cdf's. Hence we give a different proof under different conditions. In particular, in Theorem 4 for time-varying abandonment cdf's we imposed additional regularity conditions. With those extra regularity conditions, we can apply the classical Picard-Lindelöf theorem for the uniqueuenss of a solution to the ODE $w'(t) = \Psi(t, w(t))$, which requires that $\Psi(t,x)$ be locally Lipschitz in the argument x uniformly in the argument t; e.g., Theorem 2.2 of Teschl (2000).

One regularity condition added in Theorem 4 was for the rate fluid enters service to be bounded below. We will show how to guarantee that condition in the next section. Given that the rate fluid enters service is indeed bounded below, i.e., given that $\gamma(t) \ge e_L > 0$ for all $t \in [0,T]$, from (18), there exists a constant $w_L > 0$ such that $w'(t) \le 1 - w_L < 1$ for all $t \in [0,T]$. Since $w(0) < \infty$, by assumption, and $w(t) \le w(0) + t$ for all t, we have $w(t) \le w(0) + T$ for $0 \le t \le T$. Together with the fact that $\lambda, q(0,\cdot) \in C_p$, that implies that the denominator in (18) is bounded above.

Since $w'(t) \le 1 - w_L < 1$ for all t, for each x we will have $t - w(t) = x$ for at most one value of t. Since $\lambda, q(0,\cdot)$ have been assumed to have bounded derivatives where they are continuous, and since the partial derivative $\partial F_t(x)/\partial t$ of the time-varying abandonment cdf F_t as been assumed

to be bounded, the mapping Ψ in (18) is Lipschitz continuous in the argument x except at only finitely many x, uniformly in t. Hence, we can deduce uniqueness of the solution of the ODE in (18) under these extra regularity conditions by applying the Picard-Lindelöf theorem.

We now elaborate on the details. Here we have

$$\Psi(t,x) \equiv 1 - \frac{\gamma(t)}{\tilde{q}(t,x)} = 1 - \frac{\mu(t)s(t) + s'(t)}{\tilde{q}(t,x)}, \tag{EC.10}$$

where $\tilde{q}(t,x)$ is given in (15). Consider the region $0 \le x_1 \le t$, $0 \le x_2 \le t$. In this region we have

$$\begin{aligned}
|\Psi(t,x_1) - \Psi(t,x_2)| &= \frac{\mu(t)s(t) + s'(t)}{\lambda(t-x_1)\lambda(t-x_2)\bar{F}_{t-x_1}(x_1)\bar{F}_{t-x_2}(x_2)} |\lambda(t-x_1)\bar{F}_{t-x_1}(x_1) - \lambda(t-x_2)\bar{F}_{t-x_2}(x_2)| \\
&\le \frac{\mu^\uparrow s^\uparrow + s'^\uparrow}{(\lambda^\downarrow)^2(\bar{F}^\downarrow)^2} |\lambda(t-x_1)\bar{F}_{t-x_1}(x_1) - \lambda(t-x_2)\bar{F}_{t-x_1}(x_1) \\
&\quad + \lambda(t-x_2)\bar{F}_{t-x_1}(x_1) - \lambda(t-x_2)\bar{F}_{t-x_2}(x_2)| \\
&\le \frac{\mu^\uparrow s^\uparrow + s'^\uparrow}{(\lambda^\downarrow)^2(\bar{F}^\downarrow)^2} (|\lambda(t-x_1) - \lambda(t-x_2)| + \lambda(t-x_2)|\bar{F}_{t-x_1}(x_1) - \bar{F}_{t-x_2}(x_2)|) \\
&\le \frac{\mu^\uparrow s^\uparrow + s'^\uparrow}{(\lambda^\downarrow)^2(\bar{F}^\downarrow)^2} (\lambda'^\uparrow |x_1 - x_2| + \lambda^\uparrow |\bar{F}_{t-x_1}(x_1) - \bar{F}_{t-x_1}(x_2) + \bar{F}_{t-x_1}(x_2) - \bar{F}_{t-x_2}(x_2)|) \\
&\le \frac{\mu^\uparrow s^\uparrow + s'^\uparrow}{(\lambda^\downarrow)^2(\bar{F}^\downarrow)^2} (\lambda'^\uparrow |x_1 - x_2| + \lambda^\uparrow \frac{\partial \bar{F}}{\partial t}^\uparrow |x_1 - x_2| + \lambda^\uparrow g^\uparrow |x_1 - x_2|) \\
&\equiv C\,|x_1 - x_2|,
\end{aligned}$$

where $C \equiv \frac{\mu^\uparrow s^\uparrow + s'^\uparrow}{(\lambda^\downarrow)^2(\bar{F}^\downarrow)^2}(\lambda'^\uparrow + \lambda^\uparrow \frac{\partial \bar{F}}{\partial t}^\uparrow + \lambda^\uparrow g^\uparrow)$. The case $x_1, x_2 > t$ is similar. Hence the regularity conditions given in Theorem 4 are sufficient for Ψ to be locally Lipschitz in x uniformly in t.

EC.2.2. e_L-Feasibility of the Staffing Function s.

We have two goals in this section: first, to prove Theorem 6, showing how to construct the minimum feasible staffing function greater than or equal to any proposed staffing function s and, second, to determine the minimum feasible staffing function such that the rate fluid enters service at time t, $\gamma(t)$, is bounded below. We use this stronger notion of feasibility to provided conditions for the ODE in (18) in Theorem 4 to have a unique solution. We treat both problems at once by introducing the notion of e_L-feasibility: A staffing function s is said to be e_L-feasible if $\gamma(t) \ge e_L \ge 0$ for all $t \in [0,T]$.

So far, we have assumed that the staffing function s is e_L-feasible (as one condition in Theorem 4) or simply feasible (e_L-feasible for $e_L \equiv 0$), yielding

$$\gamma(t) \geq s'(t) + \sigma(t) = s'(t) + \int_0^\infty b(t,x) h_G(x)\, dx \geq e_L \geq 0 \quad \text{when} \quad B(t) = s(t). \tag{EC.11}$$

This requirement is automatically satisfied in underloaded intervals when $B(t) = s(t)$, provided that $\lambda_{inf}(T) \geq e_L$ for λ_{inf} in Assumption 6, because in that case we require that $s'(t) + \sigma(t) \geq \lambda(t)$ where necessarily $\lambda(t) \geq e_L$; see Definition 1; e_L-Feasibility is only a concern during overloaded intervals, and then only when the staffing function is decreasing, i.e., when $s'(t) < 0$.

A violation is easy to detect; it necessarily occurs in an overloaded interval in $\mathcal{O}([0,T])$ at time $t^* \equiv \inf\{t \in \mathcal{O}([0,T]) : \gamma(t) < e_L\}$. Paralleling Liu and Whitt (2010), let $\mathcal{S}_{f,s,e_L}$ be the set of e_L-feasible staffing functions over the interval $[0,t]$ for $t > t^*$. Then

$$t^* \equiv t^*(e_L) \equiv \inf\{t \in I : \gamma(t) < e_L\}. \tag{EC.12}$$

Even though we require (EC.11), so far we have done nothing to prevent having $t^* < \infty$ (violation). Thus, we compute γ and detect the first violation.

Correcting the staffing function is not difficult either (by which we mean replacing it with a higher feasible staffing function): We simply construct a new staffing function s^* consistent with reducing the input into the queue to its minimum allowed level (setting $\gamma(t) = e_L \geq 0$) starting at time t^* and lasting until the first time t after t^* at which $s^*(t) = s(t)$. (By the adjustment, we will have made $s^*(t^*+) > s(t^*+)$.) Since the system has operated differently during the time interval $[t^*, t]$, we must recalculate all the performance measures after time t, but we have now determined a feasible staffing function up to time $t > t^*$. By successive applications of this correction method (adjusting the staffing function s and recalculating b), we can construct the minimum feasible staffing function overall.

To make this precise, let $\mathcal{S}_{f,s,e_L}(t)$ be the set of all e_L-feasible staffing functions for the system over the time interval $[0,t]$, $t > t^*$, that coincide with s over $[0,t^*]$; i.e., let

$$\mathcal{S}_{f,s,e_L}(t) \equiv \{\tilde{s} \in C_p^1(t) : \gamma_{\tilde{s}}(u) 1_{\{B_{\tilde{s}}(u) = \tilde{s}(u)\}} \geq e_L, \quad 0 \leq u \leq t, \quad \tilde{s}(u) = s(u), \quad 0 \leq u \leq t^*\}, \tag{EC.13}$$

for t^* in (EC.12), where $\gamma_{\tilde{s}}$ and $B_{\tilde{s}}$ are the functions γ and B associated with the model with staffing function $\tilde{s}$.

THEOREM EC.1. (minimum e_L-feasible staffing function) *For each e_L such that $0 \le e_L \le \lambda_{inf}(T)$ for $\lambda_{inf}(T)$ in Assumption 6, there exist $\delta \equiv \delta(e_L)$ and $s^* \in \mathcal{S}_{f,s,e_L}(t^*+\delta)$ in* (EC.13) *for t^* in* (EC.12) *such that*

$$s^* \equiv s^*(e_L) = \inf\{\tilde{s} \in \mathcal{S}_{f,s,e_L}(t^*+\delta)\}; \tag{EC.14}$$

i.e., $s^* \in \mathcal{S}_{f,s,e_L}(t^*+\delta)$ and $s^*(u) \le \tilde{s}(u)$, $0 \le u \le t^*+\delta$, for all $\tilde{s} \in \mathcal{S}_{f,s,e_L}(t^*+\delta)$. In particular,

$$s^*(t^*+u) = e_L \int_0^u e^{-M(t^*+u-x,t^*+u)}\,dx + B(t^*)\,e^{-M(t^*,t^*+u)}. \tag{EC.15}$$

Moreover, δ can be chosen so that

$$\delta = \inf\{u \ge 0 : s^*(t^*+u) = s(t^*+u)\}, \tag{EC.16}$$

with $\delta \equiv \infty$ if the infimum in (EC.16) *is not attained.*

Proof. First, since γ_s is continuous for our original s, the violation in (EC.12) must persist for a positive interval after t^*; that ensures that a strictly positive δ can be found. We shall prove that $\tilde{s} \ge s^*$ over $[t^*, t^*+\delta]$ for s^* in (EC.15) and any feasible function $\tilde{s}$, and we will show that s^* itself is feasible. For $0 \le t \le t^*+\delta$, suppose $\tilde{s}$ is feasible. Since the system is overloaded, system being in the overloaded regime implies that

$$\begin{aligned}
\tilde{s}(t^*+u) &= B_{\tilde{s}}(t^*+u) = \int_0^\infty b_{\tilde{s}}(t^*+u,x)dx \\
&= \int_0^u \gamma_{\tilde{s}}(t^*+u-x)\,\bar{G}_{t^*+u-x}(x)dx + \int_u^\infty b_{\tilde{s}}(t^*,x-u)\frac{\bar{G}_{t^*+u-x}(x)}{\bar{G}_{t^*+u-x}(x-u)}dx \\
&= \int_0^u \gamma_{\tilde{s}}(t^*+u-x)\,e^{-M(t^*+u-x,t^*+u)}\,dx + \int_u^\infty b_s(t^*,x-u)\,e^{-M(t^*,t^*+u)}dx \\
&\ge e_L \int_0^u e^{-M(t^*+u-x,t^*+u)}\,dx + e^{-M(t^*,t^*+u)} \int_0^\infty b_s(t^*,y)dy = s^*(t^*+u).
\end{aligned}$$

where the second equality holds because of the fundamental evolution equations in Assumption 4, the third equality holds because $b_{\tilde{s}}(t^*,x) = b_s(t^*,x)$ for all x, and the inequality holds because $\gamma_{\tilde{s}} \ge e_L$. On the other hand, the equality holds when $\gamma_{\tilde{s}}(t^*+u) = e_L$ for all u, which yields $B(t^*+u) = s^*(t+u)$. Therefore, the proof is complete. ∎

COROLLARY EC.1. (minimum e_L-feasible staffing with exponential service times) *For the special case of exponential service times, i.e., with* $\bar{G}(x) \equiv e^{-\mu x}$, *independent of* t, (EC.15) *becomes simply* $s^*(t^*+u) = e_L(1-e^{-\mu u})/\mu + B(t^*)e^{-\mu u}$, $0 \le u \le \delta$.

EC.3. Proofs for §5.

EC.3.1. Proof of Theorem 7.

First, the assumption that $\zeta_1, \zeta_2 \in \mathcal{P}_{m,n}$ assures that there are only finitely many switches between overloaded intervals and underloaded intervals in both systems. That leads to three cases: (i) when both systems are underloaded, (ii) when the upper system is overloaded and the lower system is underloaded, and (iii) when both systems are overloaded. We apply mathematical induction over the successive alternating intervals of these three kinds. (The switch points are the union of the two separate sets of switch points.) We ensure that the initial conditions for each succeeding interval satisfy the initial ordering assumed in the theorem. If we start in an interval where both systems are underloaded, then the ordering holds while both systems are underloaded by virtue of the explicit representation in Proposition 1. Consequently, the underload termination times are ordered as well, by Proposition 1. The ordering $B_1(t) \le B_2(t)$ necessarily remains valid when the upper system is overloaded and the lower system is underloaded, because then we have $B_1(t) \le s(t) = B_2(t)$. For an interval where both systems are overloaded, it suffices to consider the two systems starting the first time both systems are overloaded. At that time, the initial conditions necessarily will be ordered properly, because the system to become overloaded later has $Q_1(t) = 0$. At this initial time, $B_1(t) = B_2(t) = s(t)$.

The M_t service assumption comes to the fore in an interval where both systems are overloaded. Here we use the fact that σ and $\gamma(t) = b(t,0)$ depend only upon s and μ during the overloaded interval, and so are the same for the two systems, because the functions s and μ have been assumed to be fixed. The rate of service completion is $\sigma(t) = s'(t) + s(t)\mu(t)$. When the two systems are both overloaded over a common interval $[t, t+u]$, the total fluid to enter service from queue, $E(t+u) - E(t)$ is therefore the same in the two systems.

When both systems are overloaded, we have the ordering $\tilde{q}_1 \le \tilde{q}_2$ directly from Proposition 3, just as in Proposition 5.3 of Liu and Whitt (2010), exploiting the representation

$$\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)} = e^{-\int_{x-t}^{x} h_{F_{t-x}}(y)\,dy}.$$

Hence, to show that $q_1 \le q_2$, it suffices to show that $w_1 \le w_2$, which would imply that that the overload termination times are ordered as well.

Suppose we start at t_1 with $w_1(t_1) \le w_2(t_1)$. Suppose that $w_1(t) > w_2(t)$ at some $t > t_1$. The continuity of w_1 and w_2 implies that there exists some $t_1 < t_2 < t$ such that $w_1(t_2) = w_2(t_2) \equiv \tilde{w}$. However, the ordering of $\tilde{q}_1$ and $\tilde{q}_2$ implies that $\tilde{q}_1(t_2, \tilde{w}) \le \tilde{q}_2(t_2, \tilde{w})$. Therefore, ODE (18) implies that $w_1'(t_2) \le w_2'(t_2)$. This contradicts with our assumption that there exists a t such that $w_1(t) > w_2(t)$.

Now we turn to v. The equation (20) 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 $L_t(u) \equiv t + u$. Clearly, if we increase the w function, then that intersection point increases as well. ∎

EC.3.2. Proof of Theorem 8.

We directly prove (23); the corresponding results in (24) will be obtained along the way. To show (i), consider two models with common model data except for $\lambda, B(0)$, where $\lambda_1, \lambda_2, s', \mu \in \mathcal{P}_{m,n}$ for some m, n. Without loss of generality, by Theorem 7, it suffices to assume that $\lambda_1 \le \lambda_2$ and $B_1(0) \le B_2(0)$. If that is not initially the case, consider $\tilde{\lambda}_1 \equiv \lambda_1 \wedge \lambda_2$, $\tilde{\lambda}_2 \equiv \lambda_1 \vee \lambda_2$, $\tilde{B}_1(0) \equiv B_1(0) \wedge B_2(0)$ and $\tilde{B}_2(0) \equiv B_1(0) \vee B_2(0)$ to get $\tilde{\lambda}_1 \le \tilde{\lambda}_2$ and $\tilde{B}_1(0) \le \tilde{B}_2(0)$ with $\|\tilde{\lambda}_1 - \tilde{\lambda}_2\|_T = \|\lambda_1 - \lambda_2\|_T$ and $|\tilde{B}_1(0) - \tilde{B}_2(0)| = |B_1(0) - B_2(0)|$.

When both systems are overloaded, we have $B_1(t) = B_2(t) = s(t)$. Hence, the overall story depends on what happens when (a) both systems are underloaded, and (b) system 1 is underloaded and system 2 is overloaded.

For simplicity, suppose that the two systems both start underloaded at time 0 with $B_1(0) \le B_2(0)$, $\lambda_1 \le \lambda_2$. If both systems remain underloaded over the interval $[0, t_1]$, then by Proposition 1 we have

$$\begin{aligned} |B_1(t) - B_2(t)| &\le \|\lambda_1 - \lambda_2\|_T \int_0^t e^{-M(x)}\, dx + |B_1(0) - B_2(0)| \\ &\le t \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)|, \quad 0 \le t \le t_1. \end{aligned} \tag{EC.17}$$

Suppose system 2 becomes overloaded at $t_1 > 0$ while system 1 remains underloaded. For $t > t_1$, we have $B_1(t) \le B_2(t) = s(t) \le X_2(t) \equiv B_2(t) + s(t)$. Hence we have $0 \le |B_2(t) - B_1(t)| = B_2(t) - B_1(t) \le X_2(t) - B_1(t)$. Flow conservations of both systems implies that $B_1'(t) = \lambda_1(t) - \mu(t)\, B_1(t)$ and $X_2'(t) = \lambda_2(t) - \alpha_2(t) - \mu(t)\, s(t)$. Therefore,

$$X_2'(t) - B_1'(t) = \lambda_2(t) - \lambda_1(t) - \alpha_2(t) - \mu(t)\,(s(t) - B_1(t)) \le \lambda_2(t) - \lambda_1(t),$$

which implies that

$$\begin{aligned} |B_1(t) - B_2(t)| &\le |B_1(t_1) - B_2(t_1)| + (t - t_1) \cdot \|\lambda_1 - \lambda_2\|_T \\ &\le t_1 \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)| + (t - t_1) \cdot \|\lambda_1 - \lambda_2\|_T \\ &\le t \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)|, \end{aligned} \tag{EC.18}$$

where the second inequality follows from (EC.17) with $t = t_1$.

If we then later start a second underloaded interval for both systems at time t_2, where $0 < t_1 < t_2 < T$, then we will have inequality (EC.17) holding at time t_2. Thus proceeding forward, applying (EC.17) with initial values $B_i(t_2)$, during the following underloaded interval we have for $t > t_2$

$$\begin{aligned} |B_1(t) - B_2(t)| &\le \|\lambda_1 - \lambda_2\|_T \int_{t_2}^t e^{-M(x)}\, dx + |B_1(t_2) - B_2(t_2)| \\ &\le (t - t_2) \cdot \|\lambda_1 - \lambda_2\|_T + t_2 \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)| \\ &\le t \cdot \|\lambda_1 - \lambda_2\|_T + |B_1(0) - B_2(0)| \\ &\le (1 \vee t)(\|\lambda_1 - \lambda_2\|_T \vee |B_1(0) - B_2(0)|). \end{aligned} \tag{EC.19}$$

where the second inequality follows from (EC.18) with $t = t_2$. Applying mathematical induction over successive underloaded subintervals of $[0, T]$, using the second to last inequality, we obtain the first relation in (23), from which the desired conclusion follows.

To show (ii), when both systems are underloaded, we have $Q_1(t) = Q_2(t) = 0$. Hence, the overall story depends on what happens when (a) both systems are overloaded, and (b) system 1 is underloaded and system 2 is overloaded.

When both systems are overloaded, flow conservation implies that

$$Q_i'(t) = \lambda_i(t) - \alpha_i(t) - \gamma_i(t) = \lambda_i(t) - \alpha_i(t) - \mu(t)\, s(t) - s'(t).$$

Hence, we have

$$Q_2'(t) - Q_1'(t) = \lambda_2(t) - \lambda_1(t) - (\alpha_2(t) - \alpha_1(t)) \le \lambda_2(t) - \lambda_1(t),$$

where the inequality simply follows from Theorem 7 when the two systems have common abandon-time distribution. This yields

$$|Q_1(t) - Q_2(t)| = Q_2(t) - Q_1(t) \le |Q_1(0) - Q_2(0)| + t\, \|\lambda_1 - \lambda_2\|_T. \tag{EC.20}$$

When system 2 is overloaded and system 1 is underloaded. For simplicity, assume at time 0 the two system have initial conditions $B_2(0) = s(0) > B_1(0)$, $Q_2(0) \ge 0 = Q_1(0)$. Let $T^* \equiv T_1 \wedge T_2$, where T_1 denotes the underload termination time of system 1 and T_2 denotes the overload termination time of system 2. Hence we know that both systems will not change regimes for $0 \le t \le T^*$. For $0 \le t \le T^*$, we have

$$\begin{aligned} Q_2'(t) &= \lambda_2(t) - \alpha_2(t) - \gamma_2(t) \le \lambda_2(t) - \gamma_2(t) \\ &\le (\lambda_2(t) - \lambda_1(t)) + (\lambda_1(t) - \gamma_2(t)) \\ &\le (\lambda_2(t) - \lambda_1(t)) + (\lambda_1(t) - \mu(t)\, s(t) - s'(t)), \end{aligned}$$

which implies that

$$\begin{aligned} &|Q_2(t) - Q_1(t)| = Q_2(t) \\ &\le Q_2(0) + t\, \|\lambda_2(t) - \lambda_1(t)\|_T + \int_0^t \lambda_1(u) - \mu(u)\, s(u) - s'(u) du \\ &\le Q_2(0) + t\, \|\lambda_2(t) - \lambda_1(t)\|_T + \int_0^t \lambda_1(u) - \mu(u)\, B_1(u) du - (s(t) - s(0)) \end{aligned}$$

$$\begin{aligned}
&\le Q_2(0) + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + \int_0^t B_1'(u)du - s(t) + s(0) \\
&\le Q_2(0) + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + (s(0) - B_1(0)) - (s(t) - B_1(t)) \\
&\le |Q_2(0) - Q_1(0)| + t\,\|\lambda_2(t) - \lambda_1(t)\|_T + |B_2(0) - B_1(0)|, \qquad \text{(EC.21)}
\end{aligned}$$

where the second inequality holds because $B_1(t) \le s(t)$, the third inequality holds since $B_1'(t) = \lambda_1(t) - \mu(t)\,B_1(t)$, and the last inequality holds since $Q_1(0) = 0$, $B_2(0) = s(0)$ and $B_1(t) \le s(t)$. Again, the desired conclusion follows by mathematical induction.

Finally, to show *(iii)*, (EC.18), (EC.19), (EC.20), (EC.21) imply that

$$\begin{aligned}
|X_1(t) - X_2(t)| &\le |B_1(t) - B_2(t)| + |Q_1(t) - Q_2(t)| \\
&\le 2t\,\|\lambda_1 - \lambda_2\| + 2\,|B_1(0) - B_2(0)| + |Q_1(0) - Q_2(0)| \\
&\le 2(1 \vee t)(\|\lambda_1 - \lambda_2\|_T \vee |X_1(0) - X_2(0)|),
\end{aligned}$$

where the third inequality holds because $|X_1(0) - X_2(0)| = |B_1(0) - B_2(0)| + |Q_1(0) - Q_2(0)|$ in all regimes. ∎

EC.3.3. Proof of Theorem 9.

Given $\lambda \in \mathbb{C}_p$, we choose an increasing sequence $\{\lambda_k : k \ge 1\}$ with $\lambda_k \in \mathcal{P}_{m_k,n_k}$ for each $k \ge 1$ such that $\|\lambda_k - \lambda\|_T \to 0$ as $k \to \infty$. For each $k \ge 1$, we can apply all the results above. By Theorem 8, we can define the pair (B, σ) in $\mathbb{C}_p^2$ as the limit of the sequence $\{(B_k, \sigma_k)$ in $\mathbb{C}_p^2$ with the maximum/uniform norm. There is such a limit, because the sequence is necessarily Cauchy and the space is a complete metric space. Given the limit, the convergence holds in the space by Theorem 8.

To show that the monotonicity extends, we start with $\lambda_1 \le \lambda_2$. We then construct sequences $\{\lambda_{i,k} : k \ge 1\}$ for $i = 1, 2$ with $\lambda_{1,k} \le \lambda_{2,k}$ for each k and $\|\lambda_{i,k} - \lambda_i\|_T \to 0$ as $k \to \infty$. We apply Theorem 7 for each k. Since the ordering is preserved in the limit, the conclusion of Theorem 7 holds for the limiting pair by Lebesgue monotone convergence. We use a similar argument to show that the Lipschitz continuity properties in Theorem 8 extend as well: Starting with $\|\lambda_1 - \lambda_2\|_T = c$,

for any $\epsilon > 0$, we construct sequences $\{\lambda_{i,k} : k \geq 1\}$ for $i = 1, 2$ with $\|\lambda_{1,k} - \lambda_{2,k}\| \leq c + \epsilon$ for each k and $\|\lambda_{i,k} - \lambda_i\|_T \to 0$ as $k \to \infty$ for $i = 1, 2$. We then can apply Theorem 8 for each $k \geq 1$, and get the conclusion there with modification by ϵ. However, since ϵ is arbitrary, we get the preservation of the Lipschitz property to the limit. ∎

EC.4. one proof for §6.

Proof of Theorem 11. We recursively apply the monotone contraction operator Ψ in Theorem 10, starting with $\sigma_{j,i}^{(0)} = 0$, so that $\lambda_{1,i}^{(1)} \leq \lambda_{2,i}^{(1)}$ for all i, because $\lambda_{j,i}^{(1)} = \lambda_{j,i}^{(0)}$, $j = 1, 2$ and the external arrival rate functions have been assumed to be ordered: $\lambda_{1,i}^{(0)} \leq \lambda_{2,i}^{(0)}$. By Theorem 7 applied to each queue separately, using the assumed ordering $B_{1,i}(0) \leq B_{2,i}(0)$ for all i, we have first $B_{1,i}^{(1)} \leq B_{2,i}^{(1)}$ and then $\sigma_{1,i}^{(1)} \leq \sigma_{2,i}^{(1)}$. By (28), we then have $\lambda_{1,i}^{(2)} \leq \lambda_{2,i}^{(2)}$. We then get the order holding for all n by applying mathematical induction. However, $\lambda_{1,i}^{(n)} \to \lambda_{1,i}$ as $n \to \infty$. Since the order is preserved in the convergence, we deduce that $\lambda_{1,i} \leq \lambda_{2,i}$ for $1 \leq i \leq m$. Finally, we can apply Theorem 7 to each queue separately to get the remaining orderings. ∎

EC.5. Remarks

Remark EC.1. (characterization of isolated points)

Definition 3 implies that t is an isolated point only if $Q(t) = 0$, $B(t) = s(t)$. Moreover, if t is a discontinuity point of ζ, then $\zeta(t - \delta) < 0$ and $\zeta(t) > 0$ for some $\delta > 0$; if t is a continuity point of ζ, then $\zeta(t - \delta) < 0$, $\zeta(t) = 0$ and $\zeta(t + \delta) < 0$ for some $\delta > 0$.

Remark EC.2. (an ODE for B in an underloaded interval)

In an underloaded interval, the total fluid content in service $B(t)$ can also be characterized via the ODE

$$B'(t) = \lambda(t) - \mu(t)B(t), \quad t \geq 0. \tag{EC.22}$$

The formula in Proposition 1 provides the solution to the initial value problem determined by this ODE with initial condition $B(0)$.

REMARK EC.3. (applied significance of $\mathcal{P}_{mn}$) We have provided a full algorithm when $\lambda, s', \mu \in \mathcal{P}_{m,n}$. An algorithm for $\lambda \in \mathbb{C}_p$ can be developed by considering a sequence of successive approximations in $\mathcal{P}_{m_n,n}$, but we see no motivation for doing so. We have introduced the space $\mathcal{P}_{m,n}$ of piecewise polynomials as a device to establish mathematical results. In applications, it should suffice to use *any* convenient representations of the functions λ and s, and *assume* that there are only finitely many switches in any finite interval. While running the algorithm, that assumption can be verified, and the model can be modified if too many switches occur. However, if we start from data, then we could choose to let the functions be in $\mathcal{P}_{m,n}$ without loss of generality. Lemma 2 shows that it is convenient to work in the space $\mathcal{P}_{m,n}$, because we can obtain closed form expressions for integrals. Moreover, if we want to bound the number of switches in advance, then we can bound the parameters m and n, with the understanding that there is a tradeoff between the quality of fit and the maximum number of switches.

EC.6. Simulation Verification for the $M_t/M/s+GI$ Model

In this section we illustrate the single-queue algorithm for a relatively simple case, the $M_t/M/s+GI$ fluid queue model, in which only the arrival rate is time varying and only the abandonment cdf F is non-exponential. We let the arrival rate function λ be sinusoidal, i.e.,

$$\lambda(t) \equiv a + b \cdot \sin(c \cdot t), \quad t \geq 0, \tag{EC.23}$$

where we let $b \equiv 0.6a$, $c \equiv 1$ and $a \equiv s$. By making the average input rate a coincide with the fixed staffing level s, we ensure that the system will alternate between overloaded and underloaded. We let the service rate be $\mu \equiv 1$ and the abandonment rate $\theta \equiv 0.5$; i.e., $G(x) \equiv 1 - e^{-x}$ for $x \geq 0$. Without loss of generality, for the fluid model we let $s \equiv 1$.

For the general abandon-time cdf F, we considered two cases: Erlang-2 (E_2) and hyperexponential-2 (H_2). We fix the mean at $1/\theta$. An E_2 random variable is the sum of two i.i.d. exponential random variables, so that there are no additional parameters. An H_2 cdf is the mixture

of two exponential cdf's, and so has two additional parameters beyond its mean. An H_2 pdf is of the form

$$f(x) = p \cdot \theta_1 e^{-\theta_1 x} + (1-p) \cdot \theta_2 e^{-\theta_2 x}, \quad x \geq 0,$$

We let $p = 0.5(1-\sqrt{0.6})$, $\theta_1 = 2p\theta$, $\theta_2 = 2(1-p)\theta$, which produces "balanced means" and squared coefficient of variation (SCV, variance divided by the square of the mean) $SCV \equiv c^2 = 4$.

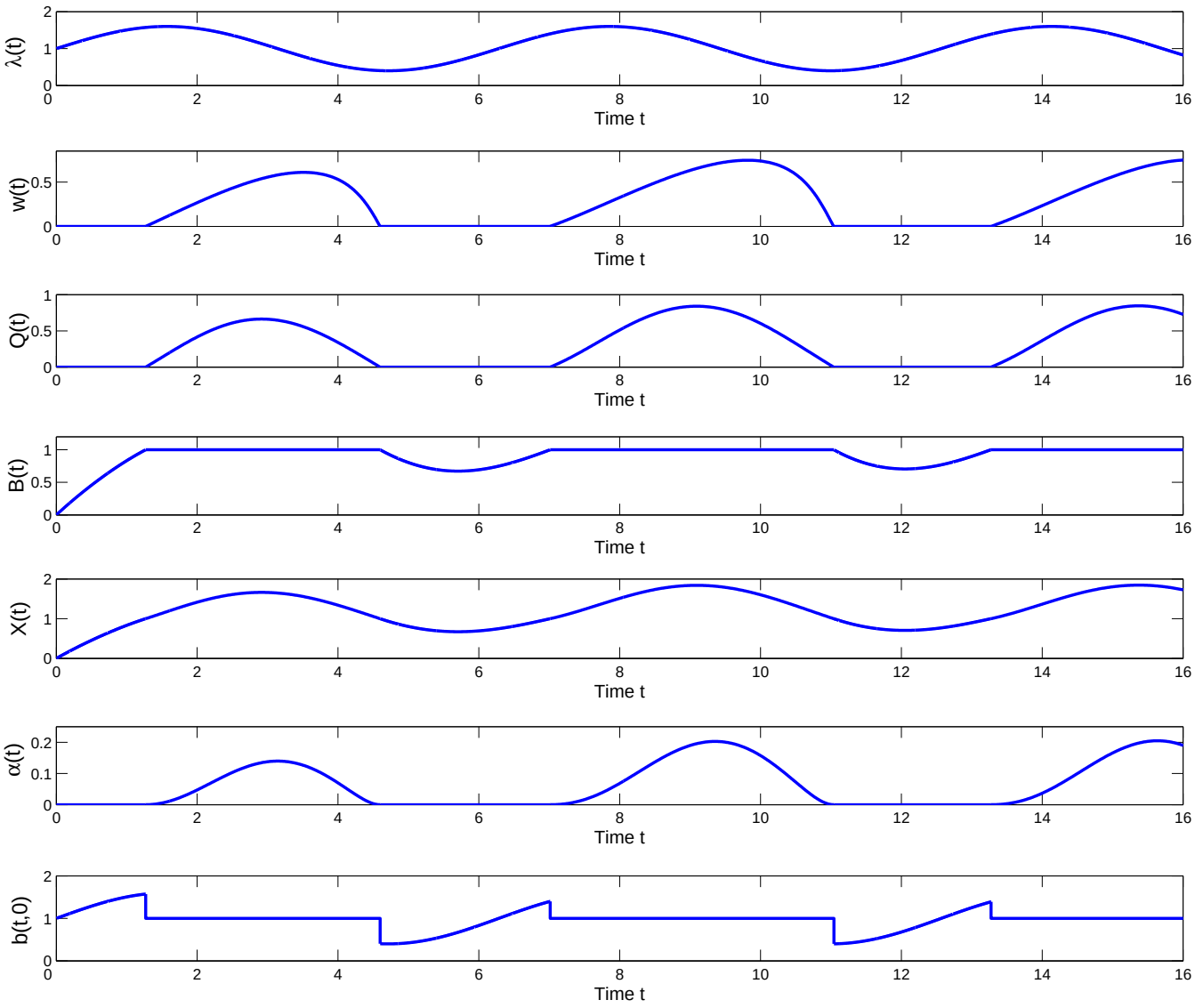

Figure EC.1 Performance for the $M_t/M/s+E_2$ fluid model with sinusoidal arrival-rate function.

We only show the results for E_2 abandonment; the results for H_2 are similar. The fluid perfor-

mance functions for E_2 abandonment are shown in Figure EC.1 for $t \in [0,T]$ with $T = 16$. The performance functions shown in Figure EC.1 are the boundary waiting time $w(t)$, the fluid in queue $Q(t)$, the fluid in service $B(t)$, the total fluid in the system $X(t)$, the abandonment rate $\alpha(t)$, and the rate fluid enters service (transportation rate) $\gamma(t) \equiv b(t,0)$. We omit the departure rate $\sigma(t) = \mu B(t)$ because of the exponential service times.

In Figure EC.2 we compare the fluid approximations with results from a simulation experiment for a very large-scale queueing system. The queueing model has a nonhomogeneous Poisson arrival process with sinusoidal rate function as in (EC.23), with $a = s = 2000$, $b = 0.6a = 1200$. We compare the fluid model predictions to a single sample path of the queueing system (one simulation run). In Figure EC.2 the blue solid lines of the simulation estimations of single sample paths applied with fluid scaling, and the red dashed lines are the fluid approximations. We conclude that the fluid approximation is remarkably accurate as an approximation when the scale of the queueing model is extremely large.

As discussed in Liu and Whitt (2010), the accuracy of the fluid approximations for large-scale queueing systems can be explained by a many-server heavy-traffic limit. As discussed in §9 of Liu and Whitt (2010), for smaller systems the queueing system has much greater stochastic fluctuations. In those cases, the fluid model performance functions quite accurately describe the mean values of the time-varying queue performance when the system experiences significant periods of overload; e.g., see Figure 7 there.

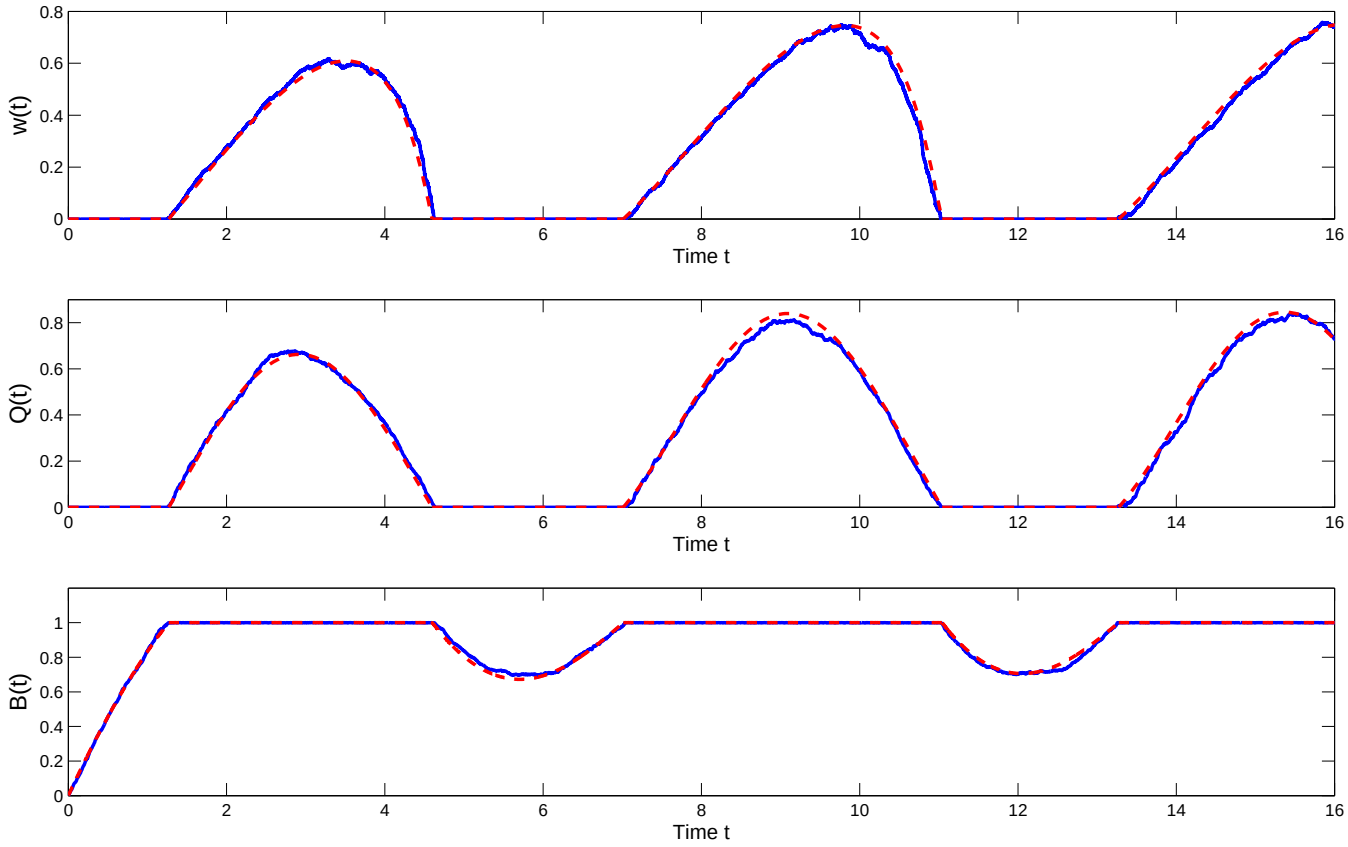

Figure EC.2 A comparison of the $M_t/M/s+E_2$ fluid model with a simulation of the large-scale queueing system.