OPERATIONS RESEARCH
Vol. 59, No. 4, July–August 2011, pp. 835–846
ISSN 0030-364X | EISSN 1526-5463 | 11 | 5904 | 0835

http://dx.doi.org/10.1287/opre.1110.0942

A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment

Yunan Liu, Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
{yl2342@columbia.edu, ww2040@columbia.edu}

To describe the congestion in large-scale service systems, we introduce and analyze a non-Markovian open network of many-server fluid queues with customer abandonment, proportional routing, and time-varying model elements. Proportions of the fluid completing service from each queue are immediately routed to the other queues, with the fluid not routed to one of the queues being immediately routed out of the network. The fluid queue network serves as an approximation for the corresponding non-Markovian open network of many-server queues with Markovian routing, where all model elements may be time varying. We establish the existence of a unique vector of (net) arrival rate functions at each queue and the associated time-varying performance. In doing so, we provide the basis for an efficient algorithm, even for networks with many queues.

Subject classifications: queues; time-varying arrivals; queueing networks; many-server queues; deterministic fluid model; customer abandonment; non-Markovian queues.
Area of review: Stochastic Models.
History: Received February 2010; revision received July 2010; accepted October 2010.

1. Introduction

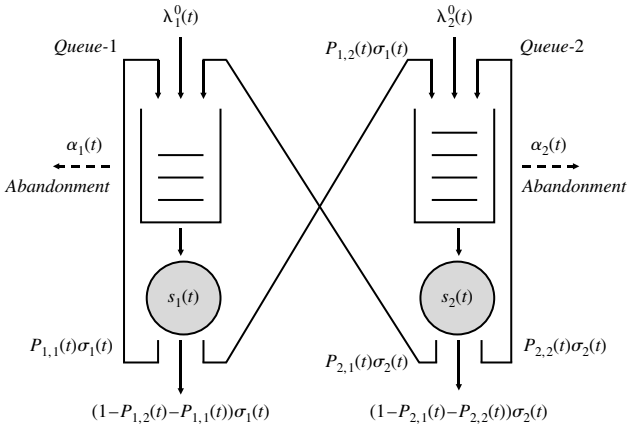
We introduce a new mathematical model intended to help analyze (and thus manage) the congestion in large-scale service systems, such as in health-care, judicial, and penal systems, and both front-office and back-office operations in business systems; e.g., see Aksin et al. (2007), Yom-Tov and Mandelbaum (2010), and references therein for discussion of possible applications to customer contact centers and health care. The model also should have other applications, because the model is both general and tractable.

The main feature of the model is time-varying arrival rates that commonly occur in applications, but which make performance analysis difficult; see Green et al. (2007) for background. The specific model is an open network of time-varying many-server fluid queues with proportional routing. There are m queues, each with its own external fluid input. In addition, a proportion $P_{i,j}(t)$ of the fluid output from queue i at time t is routed immediately to queue j, and a proportion $P_{i,0}(t) \equiv 1 - \sum_{j=1}^{m} P_{i,j}(t) \leqslant 1$ is routed out of the network (departs having successfully completed all required service). This framework permits feedback, both directly and indirectly after one or more transitions to other queues, as shown in Figure 1 for the case $m = 2$. Following Massey and Whitt (1993), we denote the model by $(G_t/M_t/s_t + GI_t)^m/M_t$, where the subscript t indicates time varying. The fluid model is intended to serve as an approximation for the corresponding many-server queueing system having m queues, each with a general time-varying arrival process (the G_t), time-varying Markovian service (the first M_t), a time-varying (large) number of servers (the s_t), and a general time-varying abandonment-time distribution (the $+GI_t$).

Strong support for the fluid approximation for the stochastic queueing system can be based on many-server heavy-traffic limits, as in Garnett et al. (2002), Mandelbaum et al. (1998), Pang et al. (2007), and Pang and Whitt (2010), but we do not establish such limits here. The fluid content is intended to approximate the mean value of the corresponding stochastic process in the many-server queueing system. For very large-scale service systems (with many servers at each queue and high arrival rates), the stochastic fluctuations about the mean values tend to be relatively small (essentially because of the law of large numbers), so that the deterministic fluid values serve as good direct approximations for the stochastic queueing quantities. The quality of approximations can be verified by simulation, as we illustrate in §EC.6. An electronic companion is part of the online version that can be found at http://or.journal.informs.org/. We also propose a simple heuristic stochastic refinement to estimate the full distribution at each time, beyond the mean values, in §8.

Because the model is tractable, we are providing the basis for creating a performance-analysis tool for large-scale service systems (allowing many queues and many servers at each queue) like the Queueing Network Analyzer (QNA) described in Whitt (1983); also see Buzacott and Shanthikumar (1992). Algorithms based on performance formulas are appealing to supplement and complement computer simulation, because the models can be created and solved much

Figure 1. The open $(G_t/M_t/s_t + GI_t)^2/M_t$ fluid network.



more quickly. Thus, they can be applied quickly in "what if" studies. They also can be efficiently embedded in optimization algorithms to systematically determine design and control parameters to meet performance objectives.

New methods are required because these large-scale service systems tend to be characterized by *many-server queues*, where a large number of homogeneous servers work in parallel. For a many-server fluid queue with time-varying Markovian service rate $\mu(t)$, when the system content is $X(t)$ and the staffing is $s(t)$, the total service completion rate at time t is $\min\{X(t), s(t)\}\mu(t)$. Unlike in single-server systems, when the many-server system is not overloaded, the service completion rate is *not* equal to the input rate, but is instead *proportional to the system content*, cf. Chen and Mandelbaum (1991).

When staffing is adequate in many-server systems, waiting times tend to be much shorter than service times. With few servers, congestion can be caused by only a few customers occasionally having exceptionally long service times. In contrast, congestion in many-server systems tends to be caused more by the cumulative impact of many customers and/or many servers. That cumulative impact often tends to be realized through a *time-varying arrival rate* and a *time-varying staffing function*.

When staffing is adequate and service times are short, as in many customer contact centers, it is often possible to apply stationary models to analyze many-server queueing models with time-varying arrival rates, using some variant of the pointwise-stationary approximation, but when staffing is occasionally inadequate or service times are longer, then other methods may be needed; see Green et al. (2007) for a review. To determine appropriate staffing levels and analyze performance in a many-server system with time-varying arrivals, we can often employ infinite-server models, as in Massey and Whitt (1993), Nelson and Taaffe (2004), Feldman et al. (2008), and references therein. However, the effectiveness of infinite-server models depends largely on the assumption that ultimately the system will be adequately staffed.

Many large-scale service systems inevitably experience periods of significant overloading, in which queues build up and customers experience significant delays. Indeed, with significant time variation of arrivals, periods of overloading often occur when it is difficult to dynamically adjust the staffing, and it is not cost effective to staff at high levels at all times. We directly address this feature by considering systems that experience alternating intervals of overload and underload. The proposed fluid models are in the spirit of early work by Newell (1982), but different in detail.

This paper extends our earlier work. First, in Whitt (2006) we described the steady-state fluid content in a stationary $G/GI/s + GI$ fluid model. Second, in Liu and Whitt (2010) we developed an algorithm for describing the time-dependent behavior of the time-varying $G_t/GI/s_t + GI$ model, including the first full description of the transient behavior of the stationary $G/GI/s + GI$ fluid model. We make several important contributions here: First, for the case of exponential service times, we extend the model from a single fluid queue to a network of fluid queues. Second, we treat time-varying service and abandonment. By focusing on M_t service instead of GI service, we are able to establish the existence of a unique (computable) performance description for both one fluid queue and the network generalization without directly assuming that there are only finitely many switches between overloaded and underloaded intervals in any finite time interval. These results are based on monotonicity and Lipschitz continuity properties of the fluid queue model in §5, which are important in their own right. Finally, we characterize the steady-state performance of the stationary network of fluid queues.

This paper is organized as follows: In §2 we introduce the $G_t/M_t/s_t + GI_t$ model of a single fluid queue. In §3 we show how the overloaded and underloaded times occur in alternating intervals of positive length, under regularity conditions, and we introduce a specific piecewise-polynomial framework for assuring that there are only finitely many switches in each finite time interval. In §4 we present the performance formulas for one queue. In §5 we extend the results to general piecewise-continuous arrival rate functions, thus providing an essential step for extending the analysis to networks. In §6 we define the network generalization and establish the existence of a unique vector of arrival rate functions at each queue and thus the performance in the network. In §7 we characterize the steady-state performance in the stationary $(G/GI/s + GI)^m/M$ fluid queue network. In §8 we propose a heuristic stochastic refinement. Finally, in §9 we draw conclusions. In the e-companion we provide (i) some proofs, (ii) some remarks, and (iii) an illustrative comparison with simulation of a large-scale queueing system.

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

Fluid is a deterministic divisible quantity that enters the system from outside. The total fluid input over an interval

$[0, t]$ is $\Lambda(t)$, where Λ is an absolutely continuous function with $\Lambda(t) \equiv \int_0^t \lambda(y)\,dy$, $t \geqslant 0$. Fluid input flows directly into the service facility if the system is underloaded; otherwise it flows into the queue.

By M_t service, we mean that service is provided at the service facility at time-varying rate $\mu(t)$ per quantum of fluid in the service facility; i.e., if the total fluid content in service at time t is $B(t)$, then the total service completion rate at time t is

$$\sigma(t) \equiv B(t)\mu(t), \quad t \geqslant 0. \tag{1}$$

Let $S(t)$ be the total amount of fluid to complete service in the interval $[0, t]$; then $S(t) \equiv \int_0^t \sigma(y)\,dy$.

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

$$F_t(x) = \int_0^x f_t(y)\,dy, \quad x \geqslant 0, \quad \text{and}$$
$$\bar{F}_t(x) \equiv 1 - F_t(x), \quad x \geqslant 0. \tag{2}$$

Let $h_{F_t}(y) \equiv f_t(y)/\bar{F}_t(y)$ be the hazard rate associated with the patience (abandonment) cdf F_t.

Let the staffing function (service capacity) s be an absolutely continuous function with $s(t) \equiv \int_0^t s'(y)\,dy$, $t \geqslant 0$. Because s is allowed to decrease, there is no guarantee that a staffing function s is feasible; i.e., having the property that no fluid that has entered service must leave without completing service. We directly assume that the staffing function we consider is feasible, but we also indicate how to detect the first violation and then construct the minimum feasible staffing function greater than or equal to the given staffing function; see Theorem 6.

Assumption 1 (Feasible Staffing). *The staffing function s is feasible, allowing all fluid that enters service to stay in service until service is completed*; *i.e., when s decreases, it never forces content out of service.*

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

$$\hat{B}(t, y) \equiv \int_0^y b(t, x)\,dx \quad \text{and} \quad \hat{Q}(t, y) \equiv \int_0^y q(t, x)\,dx, \tag{3}$$

for $t \geqslant 0$ and $y \geqslant 0$. 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. Because service is assumed to be M_t, the performance will primarily depend on b via B. (We will not directly discuss $\hat{B}$.)

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, i.e., $B(t) = s(t)$. Also, whenever $B(t) < s(t)$, then the queue must be empty. These constraints are summarized in the following assumption.

Assumption 2 (Fluid Dynamics Constraints, FDCs). *For all $t \geqslant 0$, $(B(t) - s(t))Q(t) = 0$ and $B(t) \leqslant s(t)$.*

Let $A(t)$ be the total amount of fluid to abandon in the interval $[0, t]$; then $A(t) \equiv \int_0^t \alpha(y)\,dy$, $t \geqslant 0$, where $\alpha(t)$ is the abandonment rate at time t. Because $q(t, x)$ is the density of fluid in queue at time t that arrived at time $t - x$, the abandonment rate at time t is

$$\alpha(t) \equiv \int_0^\infty q(t, y) h_{F_{t-y}}(y)\,dy, \quad t \geqslant 0, \tag{4}$$

where $h_{F_t}(y)$ is the hazard rate associated with the patience cdf F_t. (Recall that F_t is defined for t extending into the past.) Hence,

$$A(t) = \int_0^t \left(\int_0^\infty q(u, y) h_{F_{u-y}}(y)\,dy, \right) du, \quad t \geqslant 0. \tag{5}$$

Let $E(t)$ be the amount of fluid to enter service in $[0, t]$. We have $E(t) \equiv \int_0^t \gamma(u)\,du$, $t \geqslant 0$, where $\gamma(t) \equiv b(t, 0)$ is the rate fluid enters service at time t. The rate that 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, $\eta(t)$, that service capacity becomes available at time t. Service capacity becomes available due to service completion and any change in the staffing function. Hence, the rate that service becomes available is

$$\eta(t) \equiv s'(t) + \sigma(t) = s'(t) + B(t)\mu(t), \quad t \geqslant 0, \tag{6}$$

so that $\eta(t) = s'(t) + s(t)\mu(t)$ if the system is overloaded at time t.

We will also be interested in waiting-time functions. 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,

$$w(t) \equiv \inf\{x > 0\colon q(t, y) = 0 \text{ for all } y > x\}. \tag{7}$$

A proper definition of q, w, and v is somewhat complicated, but that has already been done in §5.2 and §5.3 in Liu and Whitt (2010), to which we refer.

We need to specify the initial conditions. That is done via the initial fluid densities $b(0, x)$ and $q(0, x)$, $x \geqslant 0$; then $\hat{B}(0, y)$ and $\hat{Q}(0, y)$ are defined via (**??**), whereas $B(0) \equiv \hat{B}(0, \infty)$ and $Q(0) \equiv \hat{Q}(0, \infty)$, as defined before. Let $w(0)$ be defined in terms of $q(0, \cdot)$ as in (7).

Assumption 3 (Finite Initial Values). $B(0) < \infty$, $Q(0) < \infty$ *and* $w(0) < \infty$.

In summary, the basic model data are in the six-tuple $(\lambda, s, \mu, F_\cdot, b(0,\cdot), q(0,\cdot))$.

Because the service discipline is FCFS, fluid leaves the queue to enter service from the right boundary of $q(t,x)$. Because the service is M_t, the proportion of fluid in service at time t that will still be in service at time $t+x$ is

$$\bar{G}_t(x) = e^{-M(t,t+x)} \quad \text{where} \quad M(t,t+x) \equiv \int_t^{t+x} \mu(y)\,dy, \quad t \geqslant 0 \text{ and } x \geqslant 0. \tag{8}$$

Note that G_t coincides with the time-varying service-time cdf of a quantum of fluid that enters service at time t. The cdf G_t has density $g_t(x) = \mu(t+x)\bar{G}_t(x)$ and hazard rate $h_{G_t}(x) = \mu(t+x)$, $x \geqslant 0$.

Based on the way the queueing system operates, we assume that q and b satisfy the following two fundamental evolution equations.

Assumption 4 (Fundamental Evolution Equations). *For* $t \geqslant 0$, $x \geqslant 0$, *and* $u \geqslant 0$,

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

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}_{t-x}(x+u)}{\bar{G}_{t-x}(x)} = b(t,x)e^{-M(t,t+u)}, \tag{10}$$

where M *is defined in* (8).

We now turn to the regularity conditions we impose on the model data. We develop a "smooth" model. For that purpose, 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 Whitt (2002).

Assumption 5 (Smoothness). s', λ, f_t, $f_\cdot(x)$, μ, $b(0,\cdot)$, $q(0,\cdot)$ *in* $\mathbb{C}_p$ *for each* x *and* t.

As a consequence, $s, \Lambda, F_t, B(0,\cdot), Q(0,\cdot)$ are differentiable functions with derivatives in $\mathbb{C}_p$ for each t; we say that they are elements of $\mathbb{C}_p^1$.

In order to treat the BWT w, we need to impose a regularity condition on the arrival rate function and the initial queue density (when the initial queue content is positive, which never occurs after an underloaded interval). We make the following assumption.

Assumption 6 (Positive Arrival Rate and Initial Queue Density). *For all* $t \geqslant 0$,

$$\lambda_{\inf}(t) \equiv \inf_{0 \leqslant u \leqslant t}\{\lambda(u)\} > 0 \quad \textit{and}$$

$$q_{\inf}(0) \equiv \inf_{0 \leqslant u \leqslant w(0)}\{q(0,u)\} > 0 \quad \textit{if } w(0) > 0.$$

In order to be sure that the PWT function v is finite, we make two more assumptions.

Assumption 7 (Minimum Staffing Level). *There exists* s_L *such that* $s(t) \geqslant s_L > 0$ *for all* $t \geqslant 0$.

Assumption 8 (Minimum Service Rate). *There exists* μ_L *such that* $\mu(t) \geqslant \mu_L > 0$ *for all* $t \geqslant 0$.

Finally, to treat A with the time-varying abandonment cdf F_t, we first introduce bounds for the time-varying pdf f_t and complementary cdf $\bar{F}_t$. Let

$$f^{\uparrow} \equiv \sup\{f_t(x):\ x \geqslant 0,\ -\infty < t \leqslant T\} \quad \text{and}$$

$$\bar{F}^{\downarrow}(x) \equiv \inf\{\bar{F}_t(x):\ -\infty \leqslant t \leqslant T\}. \tag{11}$$

Assumption 9 (Controlling the Time-Varying Abandonment Distribution). $f^{\uparrow} < \infty$ *and* $\bar{F}^{\downarrow}(x) > 0$ *for all* $x > 0$, *where* $f^{\uparrow}$ *and* $\bar{F}^{\downarrow}(x)$ *is defined in* (11).

In summary, here we have made Assumptions 3.1–3.6 and 5.4–5.7 of Liu and Whitt (2010) (with minor modifications because of M_t service and GI_t abandonment instead of both being GI). Assumption 3 above combines Assumptions 3.4 and 5.4 there. We show how to relax Assumption 3.7 there in the next section. We no longer need Assumptions 5.1–5.3 because we do not need to solve the fixed-point equation for b in Theorem 5.1 of Liu and Whitt (2010). Assumption 9 here is new because of the time-varying abandonment.

3. Underloaded and Overloaded Intervals

In Assumption 3.7 of Liu and Whitt (2010), we directly assumed that the system alternates between underloaded intervals and overloaded intervals, with there being only finitely many switches in any finite interval. In this paper, we provide conditions under which that assumption can be guaranteed to hold, and then show how to treat the more general case as a limit of such systems. This extension is important to rigorously treat fluid queue networks. This extension is facilitated by having M_t service.

We initially classify the system state as overloaded or underloaded at time t as follows. Recall that the rate service capacity becomes available at time t is $\eta(t) \equiv s'(t) + \sigma(t)$, as in (6) above.

Definition 1. The system is *overloaded* if either (i) $Q(t) > 0$ or (ii) $Q(t) = 0$, $B(t) = s(t)$, and $\lambda(t) > \eta(t) = s'(t) + s(t)\mu(t)$; the system is *underloaded* if either (i) $B(t) < s(t)$ or (ii) $B(t) = s(t)$, $Q(t) = 0$, and $\lambda(t) \leqslant \eta(t) = s'(t) + s(t)\mu(t)$.

At every time t, the system is thus either overloaded or underloaded.

We now define the set of switch times. For that purpose, let $\mathcal{O}(A)$ ($\mathcal{U}(A)$) be the set of overloaded (underloaded) times t in the subset A of a designated interval $[0, T]$. From Definition 1, $\mathcal{U}(A) = A - \mathcal{O}(A)$ for each subset A (the complement relative to A).

Definition 2. The subset $\mathcal{S}$ of *switch times* in $[0, T]$ is the subset of t for which

$$\mathcal{U}(((t-\epsilon)\vee 0, (t+\epsilon)\wedge T)) \neq \varnothing \quad \text{and}$$

$$\mathcal{O}(((t-\epsilon)\vee 0, (t+\epsilon)\wedge T)) \neq \varnothing \quad \text{for all } \epsilon > 0. \tag{12}$$

To neatly classify the switching times, we further classify some of the underloaded times.

Definition 3. An underloaded time t is *isolated* if (i) either $[0, t)$ or (a, t) is an overloaded interval and (ii) either $(t, T]$ or (t, b) is an overloaded interval.

We now reclassify all isolated underloaded points as overloaded points. When we reclassify each isolated underloaded point, we replace the two connecting overloaded intervals by the common overloaded interval; e.g., when t is an isolated underloaded time between overloaded intervals (a, t) and (t, b), we replace the two intervals by the single interval (a, b). In §EC.1 we show that this procedure is well defined. In the remainder of this section we present the key results allowing us to ensure that $\mathcal{S}$ is finite. We present the proofs in §EC.1. Our first structural result is the following:

Theorem 1 (Partition Into Intervals). *After all isolated underloaded times have been reclassified as overloaded and all overloaded intervals have been increased as specified above, the interval* $[0, T]$ *can be partitioned into at most countably many alternating overloaded and underloaded intervals* (*of positive length*). *The resulting switch points are the boundary points between overloaded intervals and underloaded intervals.*

Our analysis above has shown how to partition the interval $[0, T]$ into alternating overloaded and underloaded intervals of positive length. Then the switch points are clearly identified as the boundary points. It is then convenient to adopt the convention that all intervals be left closed and right open (e.g., of the form $[a, b)$), except at the interval endpoints 0 and T, so that the regime identification function $r(t) \equiv 1_{\{\mathcal{O}([0,T])\}}(t)$, where $1_{\{A\}}$ is the usual indicator function, is right continuous with left limits. This convention does not alter the switch points.

We now relate the subset $\mathcal{S}$ to the set of discontinuity points and the zero set of the function

$$\zeta(t) \equiv \lambda(t) - s'(t) - s(t)\mu(t), \quad t \geqslant 0. \tag{13}$$

Note that ζ depends only on the basic model functions λ, s, and μ. Also note that $\zeta = \lambda - \eta$ in the overloaded case of Definition 1. Let $\mathcal{D}_\zeta$ be the set of discontinuities of ζ in (13) and let $\mathcal{Z}_\zeta \equiv \{t \in [0, T]\colon \zeta(t) = 0\}$ be the zero set.

Theorem 2 (Relating Switches to Zeros and Discontinuities of ζ). *For any interval* $[0, T]$, *the subsets* $\mathcal{S}$, $\mathcal{Z}_\zeta$, *and* $\mathcal{D}_\zeta$ *are closed subsets with* $|\mathcal{S}| \leqslant |\mathcal{Z}_\zeta| + |\mathcal{D}_\zeta| - 1$. *Moreover, the bound is tight*; *i.e., there are examples for which the bound holds as an equality.*

We now introduce a convenient subset of functions in $\mathbb{C}_p$ to represent our model data λ, μ, and s'. The class is sufficiently general that it can represent any function in $\mathbb{C}_p$ and, at the same time, it allows us to control the zeros of ζ, so that we know in advance that there are only finitely many switches between overloaded and underloaded intervals in any finite interval.

Let $\mathcal{P}_{m,n} \equiv \mathcal{P}_{T,m,n}$ be the space of *piecewise-polynomials* on the interval $[0, T]$, where $[0, T]$ is partitioned into n subintervals, on each of which there is a polynomial of order at most m. We start with three elementary lemmas about $\mathcal{P}_{m,n}$. (We do not require that the overall function be continuous, but each function necessarily is in $\mathbb{C}_p$.) The first lemma states that any function in C_p can be approximated uniformly by a function from $\mathcal{P}_{m,n}$, so that there is no practical loss of generality to restricting the model data to be in $\mathcal{P}_{m,n}$ instead of $\mathbb{C}_p$.

Lemma 1 (Uniform Approximation). *For any function* $h \in \mathbb{C}_p$ *over a finite interval* $[0, T]$ *and any* $\epsilon > 0$, *there exists a function* $\tilde{h} \in \mathcal{P}_{m,n}$ *for some positive integers* m *and* n *such that* $\|h - \tilde{h}\|_T < \epsilon$.

The second lemma states that we can go back and forth between the functions λ, s', μ and their integrals Λ, s, M in $\mathcal{P}_{m,n}$ conveniently; i.e., the integral or derivative of a polynomial is again a polynomial. In particular, we can analytically calculate the integral for M in definition (8), as needed for the fundamental evolution equation for b in (10).

Lemma 2 (Representation of Integrals). $\lambda, s', \mu \in \mathcal{P}_{m,n} \subset \mathbb{C}_p$ *if and only if* Λ, $M(t, t+\cdot)$, $M(u-\cdot, u)$, $s \in \mathcal{P}_{m+1,n} \cap C$.

The third lemma states that the function ζ inherits its piecewise-polynomial structure assumed for the basic model functions λ, s', μ.

Lemma 3 (Preservation of Piecewise-Polynomial Structure). *If* $\lambda \in \mathcal{P}_{m_1,n_1}$, $s' \in \mathcal{P}_{m_2,n_2}$, *and* $\mu \in \mathcal{P}_{m_3,n_3}$, *then* $\zeta \in \mathcal{P}_{m,n}$, *where* $n \leqslant n_1 + n_2 + n_3$ *and* $m \leqslant m_1 \vee m_2 \vee m_3(m_2 + 1)$.

The following theorem serves as the basis for our analysis.

Theorem 3 (Finitely Many Switches). *If* $\zeta \in \mathcal{P}_{m,n}$ *for* ζ *in* (13), *then* $|\mathcal{S}| \leqslant n(m+1) - 1$.

Hence, we can carry out the construction of the desired performance vector $(b, q, w, v, \sigma, \alpha)$ under the assumptions that the basic model functions (λ, s, μ) are such that there are only finitely many switches between overloaded intervals and underloaded intervals in any given interval $[0, T]$. It suffices to have $\lambda, s', \mu \in \mathcal{P}_{m,n}$ for some m and n. The space $\mathcal{P}_{m,n}$ is useful for the theory, but it should not be needed in applications; see Remark EC.3.

4. The Performance at One Queue

In this section we determine the performance functions under the assumption that there are only finitely many switches between overloaded and underloaded intervals. We have just seen that a sufficient condition for that is to have $\zeta \in \mathcal{P}_{m,n}$ for some m and n, for which a sufficient condition is to have $\lambda, s', \mu \in \mathcal{P}_{m,n}$ for some m and n. Here we can apply the previous results in Liu and Whitt (2010), making proper adjustments to account for the change from GI service and abandonment to M_t service and GI_t abandonment.

An underloaded interval requires modification to account for M_t service. Because the rate that fluid enters service is $\gamma(t) = b(t,0) = \lambda(t)$ when the system is underloaded, we immediately obtain an expression for $b(t,x)$ from (10). Recall that we have assumed that $b(0,\cdot) \in \mathbb{C}_p$.

Proposition 1 (Service Content in the Underloaded Case). *For the fluid model with unlimited service capacity* ($s(t) \equiv \infty$ *for all* $t \geqslant 0$), *starting at time* 0,

$$b(t,x) = e^{-M(t-x,t)}\lambda(t-x)1_{\{x\leqslant t\}} + e^{-M(0,t)}b(0,x-t)1_{\{x>t\}},$$

$$B(t) = \int_0^t e^{-M(t-x,t)}\lambda(t-x)\,dx + B(0)e^{-M(0,t)},$$

$$0 \leqslant t < T, \quad (14)$$

where M *is defined in* (8). *If, instead, a finite-capacity system starts underloaded, then the same formulas apply over the interval* $[0,T)$, *where* $T \equiv \inf\{t \geqslant 0\colon B(t) > s(t)\}$, *with* $T = \infty$ *if the infimum is never obtained. Hence,* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* $t \geqslant 0$ *and* $x \geqslant 0$, *for* t *in the underloaded interval.*

There is dramatic simplification in going from GI service to M_t service in an overloaded interval. Then we simply have $B(t) = s(t)$. The rate that fluid enters service is equal to the rate that service capacity becomes available: $\gamma(t) = \eta(t) = s'(t) + s(t)\mu(t)$. For an overloaded interval starting at time 0, we have

Proposition 2 (Service Content in the Overloaded Case). *For the fluid model in an overloaded interval,* $B(t) = s(t)$ *and*

$$b(t,x) = (s'(t-x) + s(t-x)\mu(t-x))e^{-M(t-x,t)}1_{\{x\leqslant t\}} + b(0,x-t)e^{-M(0,t)}1_{\{x>t\}},$$

where M *is defined in* (8). *Hence,* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* $t \geqslant 0$ *and* $x \geqslant 0$ *in an overloaded interval.*

Corollary 1 (Overall Smoothness for the Service Content). *If there are only finitely many switches between overloaded and underloaded intervals in* $[0,T]$, *then* $b(t,\cdot), b(\cdot,x), B \in \mathbb{C}_p$ *for all* t, $0 \leqslant t \leqslant T$, *and* $x \geqslant 0$.

We treat q, w, and v just as in §5.2 and §5.3 in Liu and Whitt (2010), making adjustments for the time-varying abandonment cdf F_t. 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 3 (Queue Content Without Transfer Into Service in the Overloaded Case). *During an overloaded interval,*

$$\tilde{q}(t,x) = \lambda(t-x)\bar{F}_{t-x}(x)1_{\{x\leqslant t\}} + q(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\}}, \quad (15)$$

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

We get an expression for q provided that we can find w.

Corollary 2 (From $\tilde{q}$ to q). *Given the BWT* w *in an overloaded interval,*

$$\begin{aligned} q(t,x) &= \tilde{q}(t-x,0)\bar{F}_{t-x}(x)1_{\{x\leqslant w(t)\wedge t\}} + \tilde{q}(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\leqslant w(t)\}} \\ &= \lambda(t-x)\bar{F}_{t-x}(x)1_{\{x\leqslant w(t)\wedge t\}} + q(0,x-t)\frac{\bar{F}_{t-x}(x)}{\bar{F}_{t-x}(x-t)}1_{\{t<x\leqslant w(t)\}}. \end{aligned} \quad (16)$$

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

It now remains to define and characterize the BWT w. We can *define* the BWT w by postulating that two expressions for the amount of fluid to enter service over any interval $[t, t+\delta]$, namely,

$$\begin{aligned} E(t+\delta) - E(t) &\equiv \int_t^{t+\delta} b(u,0)\,du \\ &= I(t, w(t), \tilde{q}) - A(t, t+\delta), \end{aligned} \quad (17)$$

where $I \equiv I(t, w(t), \tilde{q})$ 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 (17) holds, then w satisfies an ordinary differential equation (ODE). However, our previous proof of uniqueuenss for the solution of that ODE does not extend directly to time-varying abandonment cdfs. Hence, we assume that either (i) the abandonment cdf F_t is independent of t or (ii) extra conditions hold, allowing us to apply the classical Picard-Lindelöf theorem, Theorem 2.2 of Teschl (2000); see §EC.2. One extra requirement is that the rate that fluid enters service is bounded below; we also show how to obtain that in EC.2. The proofs of Theorems 4 and 6 are in §EC.2.

Theorem 4 (The BWT ODE). *Consider an overloaded interval* $[0,T)$. *The BWT* w *is well defined by relation* (17), *being Lipschitz continuous on* $[0,T]$ *with* $w(t+u) \leqslant w(t) + u$ *for all* $t \geqslant 0$ *and* $u \geqslant 0$ *with* $t + u \leqslant T$. *Moreover,* w *is right differentiable everywhere with right derivative*

$$w'(t+) = \Upsilon(t, w(t)) \equiv 1 - \frac{\gamma(t+)}{\tilde{q}(t, w(t)-)}, \quad (18)$$

where $\gamma(t) = s'(t) + s(t)\mu(t)$, $t \geqslant 0$, *and left differentiable everywhere* (*but not necessarily differentiable*) *with value*

$$w'(t-) = \tilde{\Upsilon}(t, w(t)) \equiv 1 - \frac{\gamma(t-)}{\tilde{q}(t, w(t)+)}. \tag{19}$$

Overall, w *is continuously differentiable everywhere except for finitely many* t. *If either* (i) *the abandonment cdfs* F_t *are independent of* t *or* (ii) *the partial derivative* $\partial F_t(x)/\partial t$ *is bounded over* $[0, T] \times [0, c]$ *for all* c, *and* λ, $q(0, \cdot)$ *have bounded derivatives in the intervals where they are continuous, and there exists a constant* $e_L > 0$ *such that* $\gamma(t) \geqslant e_L$ *for* $0 \leqslant t \leqslant T$, *then* w *is characterized as the unique solution of the initial value problem* (*IVP*) *on* $[0, T)$ *based on the ODE* (18) *and any initial value* $w(0)$.

Corollary 3 (End of the Overloaded Interval). *We can compute the end of an overloaded interval as* $T \equiv \inf\{t \geqslant 0: w(t) = 0 \text{ and } \lambda(t) \leqslant s'(t) + s(t)\mu(t)\}$.

Corollary 4 (Smoothness of $q(t, \cdot)$). *Under the assumptions of Theorem* 4, q *is given by* (16) *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* 2.)

Theorem 5 (The PWT v and the BWT w). *In an overloaded interval, the PWT* v *is finite and is the unique function in* $\mathbb{D}$ *satisfying the equation*

$$v(t - w(t)) = w(t) \quad \textit{or, equivalently,}$$
$$v(t) = w(t + v(t)) \quad \textit{for all } t \geqslant 0, \tag{20}$$

where w *is the BWT. 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* $t + v(t) \leqslant u \leqslant t + v(t) + \epsilon$. *If* $b(\cdot, 0) > 0$ *a.e. with respect to Lebesgue measure, then* v *is continuous.*

As shown in Liu and Whitt (2010), the proof of Theorem 5 provides an elementary algorithm to compute v once w has been computed. Theorem 5.6 of Liu and Whitt (2010) shows that v satisfies its own ODE under additional regularity conditions.

The Algorithm for One Queue. We now summarize the algorithm to compute the performance function $(b, q, w, v, \sigma, \alpha)$ in the $G_t/M_t/s_t + GI_t$ model, assuming that there are only finitely many switches in each finite interval. During each underloaded interval, we compute b and B, and determine the end of the interval, by applying Proposition 1. During each overloaded interval, we compute these by applying Proposition 2. During each overloaded interval, we successively compute $\tilde{q}$, the BWT w, q and the PWT v, respectively, from Proposition 3, Theorem 4, Corollary 2, and Theorem 5. While computing w, we determine the end of the overloaded interval by applying Corollary 3. We compute the service completion rate σ from (1) and the abandonment rate α from (4).

Feasibility of the Staffing Function. The construction above has been done under the assumption that the staffing function is feasible. As in §6.2 of Liu and Whitt (2010), the algorithm can detect violations of feasibility whenever they occur and can then produce the minimum feasible staffing function greater than or equal to the initial proposed staffing function. A violation is easy to detect; it necessarily occurs in an overloaded interval in $\mathcal{O}([0, T])$ at time $t^* \equiv \inf\{t \in \mathcal{O}([0, T]): \gamma(t) < 0\}$. As in Liu and Whitt (2010), let $\mathcal{S}_{f,s}$ be the set of feasible staffing functions over the interval $[0, t]$ for $t > t^*$.

Theorem 6 (Minimum Feasible Staffing Function). *There exist* $\delta > 0$ *and* $s^* \in \mathcal{S}_{f,s}(t^* + \delta)$ *such that* $s^* = \inf\{\tilde{s} \in \mathcal{S}_{f,s}(t^* + \delta)\}$; *i.e.,* $s^* \in \mathcal{S}_{f,s}(t^* + \delta)$ *and* $s^*(u) \leqslant \tilde{s}(u)$, $0 \leqslant u \leqslant t^* + \delta$, *for all* $\tilde{s} \in \mathcal{S}_{f,s}(t^* + \delta)$. *In particular,*

$$s^*(t^* + u) \equiv B(t^*) \cdot e^{-M(t^*, t^* + u)}, \quad 0 \leqslant u \leqslant \delta. \tag{21}$$

Moreover, δ *can be chosen so that* $\delta = \inf\{u \geqslant 0: s^*(t^* + u) = s(t^* + u)\}$, *with* $\delta \equiv \infty$ *if the infimum is not attained.*

Corollary 5 (Minimum Feasible Staffing with M Service). *For* M *service, i.e., with exponential service times, so that* $\bar{G}(x) \equiv e^{-\mu x}$, (21) *becomes simply* $s^*(t^* + u) = B(t^*)e^{-\mu u}$, $0 \leqslant u \leqslant \delta$.

Theorem 6 shows how to construct a new staffing function that (i) agrees with the proposed staffing function s over its interval of feasibility $[0, t^*)$ and (ii) itself is feasible over the longer interval $[0, t^* + \delta)$ for some $\delta > 0$. To construct the minimum feasible staffing function over $[0, T]$, this algorithm may need to be applied several times.

5. General Arrival Rate Functions

In the previous two sections we have seen that we can get a nice clean theory if we assume that $\lambda, s', \mu \in \mathcal{P}_{m,n}$. In order to treat open networks of fluid queues, we would want the service completion rate σ, which becomes the part of the input rate at other queues, to be in $\mathcal{P}_{m,n}$ for some m and n as well, but σ does not inherit this property, because $\sigma(t) = B(t)\mu(t)$ and $B(t)$ has a complicated nonpolynomial form in underloaded intervals, as shown in (14). We do have $\sigma \in \mathbb{C}_p$ by virtue of Corollary 1, but we need not have $\sigma \in \mathcal{P}_{m,n}$. Hence, we show how to treat the general case in which initially we only assume that $\lambda \in \mathbb{C}_p$.

We will treat the case of general $\lambda \in \mathbb{C}_p$ as the limit of a sequence of systems with $\lambda \in \mathcal{P}_{m,n}$. In particular, for arbitrary $\lambda \in \mathbb{C}_p$, we can represent it as the limit of a sequence of functions $\{\lambda_k: k \geqslant 1\}$, where $\lambda_k \in \mathcal{P}_{m_k, n_k}$ and $\lambda_k \geqslant 0$ for each k, and $\|\lambda_k - \lambda\|_T \to 0$ as $k \to \infty$, with $\|\cdot\|_T$ denoting the uniform norm over $[0, T]$. (Positivity is no problem because of Assumption 6 and the uniform convergence.) If we also assume that $s', \mu \in \mathcal{P}_{m,n}$ for some m, n, then we will necessarily have $\zeta_k \in \mathcal{P}_{m_k, n_k}$ for all k, with $m_k < \infty$ and

$n_k < \infty$ for all k. We will also have $m_k \to \infty$ and $n_k \to \infty$ as $k \to \infty$ unless $\lambda \in \mathcal{P}_{m,n}$ for some m, n.

In this section we establish results that allow us to treat the case of general arrival rate functions $\lambda \in \mathbb{C}_p$, without requiring that $\lambda \in \mathcal{P}_{m,n}$ and without directly requiring that there be only finitely many switches between overloaded and underloaded intervals in the interval $[0, T]$. To do so, we establish monotonicity and Lipschitz continuity properties, which are of independent interest. We first establish these results assuming that $\zeta \in \mathcal{P}_{m,n}$, and then we show that they extend when we allow arbitrary $\lambda \in \mathbb{C}_p$. We thus start by assuming that $\zeta \in \mathcal{P}_{m,n}$. The proofs of the three theorems in this section are relatively straightforward, but long; they appear in §EC.3.

The M_t service allows us to extend the elementary comparison results in Propositions 4.2 and 5.3 of Liu and Whitt (2010). Recall that order of functions (vectors) is defined as pointwise order for all arguments (coordinates). Let $X(t) \equiv B(t) + Q(t)$ be the total system fluid content. Let subscripts designate the model.

THEOREM 7 (FUNDAMENTAL COMPARISON THEOREM). *Consider two $G_t/M_t/s_t + GI_t$ fluid models with common staffing function s and service rate function μ. If $\zeta_1, \zeta_2 \in \mathcal{P}_{m,n}$ with $\lambda_1 \leqslant \lambda_2$, $B_1(0) \leqslant B_2(0)$, $q_1(0,\cdot) \leqslant q_2(0,\cdot)$ and $h_{F_{t,1}} \geqslant h_{F_{t,2}}$, then*

$$(B_1(\cdot), \tilde{q}_1, q_1, Q_1(\cdot), X_1, w_1, v_1, \sigma_1) \leqslant (B_2(\cdot), \tilde{q}_2, q_2, Q_2(\cdot), X_2, w_2, v_2, \sigma_2). \tag{22}$$

In addition to monotonicity, the model has additional basic Lipschitz continuity properties (beyond Proposition EC.2).

THEOREM 8 (MORE LIPSCHITZ CONTINUITY). *Consider a $G_t/M_t/s_t + GI_t$ fluid model with $\lambda, s', \mu \in \mathcal{P}_{m,n}$ for some m, n. Then the functions mapping* (i) $(\lambda, B(0))$ *in* $\mathcal{P}_{m,n} \times \mathbb{R}$ *into* (B, σ) *in* $\mathbb{C}_p^2$, (ii) $(\lambda, B(0), Q(0))$ *in* $\mathcal{P}_{m,n} \times \mathbb{R}^2$ *into* Q *in* $\mathbb{C}_p$, *and* (iii) $(\lambda, X(0))$ *in* $\mathcal{P}_{m,n} \times \mathbb{R}$ *into* X *in* $\mathbb{C}_p$, *all over* $[0, T]$, *are Lipschitz continuous. In particular,*

$$\begin{aligned}
&\|B_1 - B_2\|_T \leqslant (1 \vee T)(\|\lambda_1 - \lambda_2\|_T \vee |B_1(0) - B_2(0)|),\\
&\|\sigma_1 - \sigma_2\|_T \leqslant \mu_T^{\uparrow} \|B_1 - B_2\|_T,\\
&\|Q_1 - Q_2\|_T \leqslant (1 \vee T)(\|\lambda_1 - \lambda_2\|_T \vee |B_1(0) - B_2(0)| \vee |Q_1(0) - Q_2(0)|),\\
&\|X_1 - X_2\|_T \leqslant 2(1 \vee T)(\|\lambda_1 - \lambda_2\|_T \vee |X_1(0) - X_1(0)|).
\end{aligned} \tag{23}$$

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

$$\|B_1 - B_2\|_T \leqslant T\|\lambda_1 - \lambda_2\|_T, \quad \|Q_1 - Q_2\|_T \leqslant T\|\lambda_1 - \lambda_2\|_T, \quad \|X_1 - X_2\|_T \leqslant 2T\|\lambda_1 - \lambda_2\|_T. \tag{24}$$

As a consequence of Theorems 3–8, we can regard the case of a general function λ as the limit of a sequence $\{\lambda_k : k \geqslant 1\}$, where $\zeta_k \in \mathcal{P}_{m_k, n_k}$ with $m_k \to \infty$ and $n_k \to \infty$ as $k \to \infty$. Hence, results for the kth system can be "lifted" to the general case; i.e., Theorems 7–8 combine to imply the following general result.

THEOREM 9 (LIFTING). *For a $G_t/M_t/s_t + GI_t$ fluid model with $s', \mu \in \mathcal{P}_{m,n}$ and $\lambda \in \mathbb{C}_p$, the system performance via $(B, \tilde{q}, w)$, for $B \equiv \{B(t) : 0 \leqslant t \leqslant T\}$, is well defined, and the conclusions of §3 and Theorems 7 and 8 remain valid.*

6. The $(G_t/M_t/s_t + GI)^m/M_t$ Fluid Queue Network

We now introduce the open network of $G_t/M_t/s_t + GI$ fluid queues, with time-dependent proportional routing. There are m queues, where each queue has model parameters as already defined in §2, with its own external fluid input, but in addition a proportion $P_{i,j}(t)$ of the fluid output from queue i at time t is routed immediately to queue j, and a proportion $P_{i,0}(t) \equiv 1 - \sum_{j=1}^m P_{i,j}(t) \leqslant 1$ is routed out of the network, as shown in Figure 1 for the case $m = 2$.

ASSUMPTION 10 (PROPORTIONAL ROUTING). *The routing matrix function for proportional routing, P: $[0, \infty) \to [0, 1]^{m^2}$, is in $\mathbb{C}_p$ and $\sum_{j=1}^m P_{i,j}(t) \leqslant 1$ for each $t \geqslant 0$ and i, $1 \leqslant i \leqslant m$.*

It is elementary to treat the basic network operations of superposition and splitting: If two input streams are combined to form a single input (superposition), then the arrival rate functions are simply added. If one stream with arrival rate function λ is split, such that a proportion $p(t)$ of that stream goes into a new split stream at time t, then the arrival-rate function of the split stream is λ_p, where $\lambda_p(t) \equiv \lambda(t)p(t)$, $t \geqslant 0$; just like λ, the splitting proportion can be time dependent. Similarly, if the departure flow from one queue becomes input to another, then the resulting arrival rate function is σ. (We do not let the abandonment flow from one queue become input to another, but if we did, then the resulting arrival-rate function would be α.) However, converting departure rate or abandonment rate into new input rate is more complicated when feedback is allowed. We discuss that case now, for departures only.

As is usual with open queueing networks, there is an external exogenous arrival rate function to each queue (from outside the network) and there is a total arrival rate function to each queue (which we simply call the arrival rate function), taking into account the flow from other queues. Let the external arrival rate function into queue j be denoted by $\lambda_j^{(0)}$; let the arrival rate function into queue j be denoted by λ_j. The model data for the $G_t/M_t/s_t + GI_t$ fluid queues directly provides the external arrival rate functions $\lambda_j^{(0)}$ (with the superscript 0 now added), whereas the arrival rate function itself satisfies a system of *traffic rate equations*. In particular,

$$\lambda_j(t) = \lambda_j^{(0)}(t) + \sum_{i=1}^m \sigma_i(t) P_{i,j}(t), \quad \text{where} \tag{25}$$

$$\sigma_i(t) = B_i(t)\mu_i(t), \quad t \geqslant 0. \tag{26}$$

Equations (25) and (26) produce a system of equations, with λ_j depending upon σ_i for $1 \leqslant i \leqslant m$, whereas σ_i in turn depends on λ_i for each i, because B_i depends on λ_i.

The formulas for B_i as a function of λ_i have been given in Propositions 1 and 2, provided that we know whether the queue is overloaded or underloaded. That requirement is the major source of complexity.

Because (25) is a linear equation, it can be written in matrix notation as $\lambda = \lambda^{(0)} + \sigma P$ by omitting the argument t as below, provided that the product σP is interpreted as in (25). Moreover, we can combine (25) and (26) to express λ as the solution of a fixed-point equation mapping $\mathbb{C}_p^m$ over $[0, T]$ into itself. To see this, note that $B_i(t)$ in (26) is a function of $\lambda_i(u)$, $0 \leqslant u < t$, and the model data (only needed for queue i). Hence, the vector $B(t) \equiv (B_1(t), \ldots, B_m(t))$ is a function of λ over $[0, t)$ and the model data. Hence, we can express (25) and (26) abstractly as

$$\lambda = \Psi(\lambda), \tag{27}$$

where $\Psi(x)(t)$ depends on its argument x only over $[0, t]$ for each $t \geqslant 0$. Here the function Ψ depends on all the model data $(\lambda_i^{(0)}, s_i, \mu_i, F_{i,\cdot}, b_i(0,\cdot), q_i(0,\cdot), P)$, $1 \leqslant i \leqslant m$.

THEOREM 10 (CONTRACTION OPERATOR). *If $s_i', \mu_i \in \mathscr{P}_{m,n}$ for $1 \leqslant i \leqslant m$, then the operator Ψ in (27) is a monotone contraction operator on the m-dimensional product space $\mathbb{C}_p^m$ over $[0, T]$ for all sufficiently small $T > 0$. Hence, there exists a unique solution λ to the traffic rate Equations (25) and (26) over $[0, T]$ for any fixed $T > 0$. For sufficiently short intervals, successive iterates $\Psi^{(n)}(\tilde{\lambda})$ converge uniformly, geometrically fast, to the fixed point for any initial point $\tilde{\lambda} \in \mathbb{C}_p^m$.*

PROOF. We first show that Ψ actually maps $\mathbb{C}_p$ into itself. First, if $\lambda \in \mathbb{C}_p^m$, then $B \in \mathbb{C}_p^m$ by Corollary 1 and Theorem 9. By assumption $\mu \in \mathbb{C}_p^m$, so that $\sigma \in \mathbb{C}_p^m$, so the conclusion follows from (25) and (26). To show that Ψ is a contraction operator for sufficiently small $T > 0$, we use the norm $\|\lambda\|_T \equiv \sum_{i=1}^m \|\lambda_i\|_T$ for $\lambda \equiv (\lambda_1, \ldots, \lambda_m) \in (\mathbb{C}_p)^m$. For any $\lambda_1, \lambda_2 \in (\mathbb{C}_p)^m$, the traffic rate equations in (25) and (26) imply that

$$\begin{aligned}
&\|\Psi(\lambda_1) - \Psi(\lambda_2)\|_T \\
&\quad \leqslant \sum_{j=1}^m \sup_{1 \leqslant t \leqslant T} \sum_{i=1}^m \mu_i(t) |B_i^1(t) - B_i^2(t)| \, P_{i,j}(t) \\
&\quad \leqslant m \, \mu_T^{\uparrow} \sum_{i=1}^m \sup_{0 \leqslant t \leqslant T} |B_i^1(t) - B_i^2(t)| \\
&\quad \leqslant m \, \mu_T^{\uparrow} T \sum_{i=1}^m \sup_{0 \leqslant t \leqslant T} |\lambda_i^1(t) - \lambda_i^2(t)| \leqslant m \mu_T^{\uparrow} T \, \|\lambda_1 - \lambda_2\|_T,
\end{aligned}$$

where $m\mu_T^{\uparrow} T < 1$ for all sufficiently small $T > 0$. The second inequality holds because $P_{i,j}(t) \leqslant 1$. The crucial third inequality follows from (24) in Theorem 8. To establish uniqueness over $[0, T]$ for any fixed $T > 0$, we consider a succession of shorter intervals over which the contraction property holds, and apply mathematical induction. Existence, uniqueness, and geometric convergence are standard consequences of the Banach contraction fixed-point theorem. Finally, monotonicity follows from Theorems 7 and 9 plus the traffic rate equations (25) and (26). □

REMARK 1 (STARTING AT THE EXTERNAL ARRIVAL RATES). Theorem 10 implies that we can approach this system recursively. If we do so with initial vector $\tilde{\lambda} = \lambda^{(0)}$, the vector of external arrival rate functions, then the recursion has an important practical interpretation. Then the kth iterate $\lambda_j^{(k)}$ is the arrival rate of fluid that has previously experienced k transitions in the fluid network. With this notation, we can write the recursive formulas

$$\begin{aligned}
\lambda_j^{(n)}(t) &= \Psi^{(n)}(\lambda^{(0)})_j(t) \\
&= \lambda_j^{(0)}(t) + \sum_{i=1}^m \sigma_i^{(n-1)}(t) P_{i,j}(t), \quad n \geqslant 1,
\end{aligned} \tag{28}$$

where

$$\sigma_i^{(n)}(t) = B_i^{(n)}(t) \mu_i(t) \quad n \geqslant 0. \tag{29}$$

Because we necessarily have $\lambda_i^{(1)} \geqslant \lambda_i^{(0)}$ for each i, this recursion converges monotonically to the fixed point λ. By Theorems 7 and 9, all the performance measures increase toward their limiting values as well.

The Algorithm for the Network of Fluid Queues. The algorithm consists of two successive steps: (i) solving the traffic-rate Equations (25) and (26) (or (27)) and (ii) solving for the performance vector $(b, q, w, v, \sigma, \alpha)$ at each queue using the algorithm in §4. For step (i), we start with an initial vector of arrival rate functions, which can be a rough estimate of the final arrival rate functions or the given external arrival rate functions as suggested in Remark 1. We then apply Propositions 1 and 2, Corollary 2, and (1) to determine the performance functions B_i and σ_i at each queue to determine a new vector of arrival rate functions. We then iteratively calculate successive vectors of arrival rate functions until the difference (measured in the supremum norm over a bounded interval) is suitably small. Then we apply step (ii).

REMARK 2 (AN m-DIMENSIONAL ODE). Algorithmically, there is an alternative approach to $(G_t/M_t/s_t)^m/M_t$ fluid queue networks. Instead of applying nm iterations of the single-queue algorithm to achieve n iterations of the operator Ψ, we can characterize the vector of arrival rates λ as the solution of one m-dimensional ODE. We obtain this ODE by differentiating with respect to t in the traffic rate equations in (25) and (26). We intend to discuss this approach in a subsequent paper. It yields useful explicit expressions for the special cases of one fluid queue with immediate proportional feedback and a network of two fluid queues, as depicted in Figure 1, plus an algorithm for the general case.

REMARK 3 (THE $(G_t/GI/s_t + GI)^m/M_t$ FLUID QUEUE NETWORK). Analogs of what we have done in this section apply to the $(G_t/GI/s_t + GI)^m/M_t$ generalization of the $G_t/GI/s_t + GI$ fluid queue considered in Liu and Whitt (2010); we only need to replace Equations (26) and (29) with the more complicated expressions given for the service completion rate σ given in Theorem 6.1 of Liu and Whitt (2010). In particular, (26) should be replaced by

$$\sigma_i(t) = \int_0^\infty b_i(t,x)h_{G_i}(x)\,dx = \int_0^t b_i(t-x,0)g_i(x)\,dx + \int_0^\infty \frac{b_i(0,y)g_i(t+y)}{\bar{G}_i(y)}\,dy,$$

whereas (29) should be replaced by

$$\sigma_i^{(n)}(t) = \int_0^\infty b_i^{(n)}(t,x)h_{G_i}(x)\,dx = \int_0^t b_i^{(n)}(t-x,0)g_i(x)\,dx + \int_0^\infty \frac{b_i^{(n)}(0,y)g_i(t+y)}{\bar{G}_i(y)}\,dy, \quad n \geqslant 0.$$

However, the service content densities at each queue, b_i, in general are characterized only as the solution of a fixed point equation. Moreover, it remains to establish an analog of Theorem 10. The space $\mathscr{P}_{mn}$ no longer helps immediately. So far, we must *assume* that there are finitely many switches between overloaded and underloaded intervals in any finite interval, and assume that there exists a unique solution to the new equations. However, from a practical perspective, the $(G_t/GI/s_t + GI)^m/M_t$ and even the more general $(G_t/GI_t/s_t + GI_t)^m/M_t$ model can be analyzed in the same way.

We conclude this section by establishing a network generalization of the single queue comparison in Theorem 7. The proof appears in §EC.4.

THEOREM 11 (NETWORK COMPARISON THEOREM). *Consider two $(G_t/M_t/s_t + GI_t)^m + M_t$ fluid queue networks with common staffing functions s_i, service rate functions μ_i, abandonment cdfs $F_{\cdot,i}$, and routing matrix function P for $1 \leqslant i \leqslant m$. If $\lambda_{1,i}^{(0)} \leqslant \lambda_{2,i}^{(0)}$, $B_{1,i}(0) \leqslant B_{2,i}(0)$, $q_{1,i}(0,\cdot) \leqslant q_{2,i}(0,\cdot)$, and $1 \leqslant i \leqslant m$, then the performance functions are ordered at each queue*:

$$(\lambda_{1,i}, B_{1,i}, \sigma_{1,i}, \tilde{q}_{1,i}, q_{1,i}, Q_{1,i}, \alpha_{1,i}, X_{1,i}, w_{1,i}, v_{1,i}) \leqslant (\lambda_{2,i}, B_{2,i}, \sigma_{2,i}, \tilde{q}_{2,i}, q_{2,i}, Q_{2,i}, \alpha_{2,i}, X_{2,i}, w_{2,i}, v_{2,i}) \quad \text{for } 1 \leqslant i \leqslant m. \tag{30}$$

7. The Stationary $(G/GI/s + GI)^m/M$ Fluid Queue Network

This paper is primarily devoted to the time-varying fluid queue network, but the corresponding stationary fluid queue network also is of interest. The stationary performance of a single $GI/GI/s + GI$ fluid queue was characterized in Whitt (2006). (The proof is completed by Liu and Whitt 2010 because the transient dynamics are characterized there.) The corresponding stationary $(G/GI/s + GI)^m/M$ fluid queue network is actually quite elementary, given Whitt (2006). In particular, the stationary performance of this model is determined by a fixed-point equation for the (now constant) arrival rates. We start by reviewing that stationary distribution of the $GI/GI/s + GI$ fluid queue.

THEOREM 12 (STEADY STATE OF THE $G/GI/s + GI$ FLUID QUEUE). *The $G/GI/s + GI$ fluid model specified with model parameter vector (λ, s, μ, G, F) has a unique steady state described by the vector $(b, q, B, Q, w, \sigma, \alpha)$, whose character depends on whether $\rho \equiv \lambda/s\mu \leqslant 1$ or $\rho > 1$.*

(a) *Underloaded and balanced cases*: $\rho \leqslant 1$. *If $\rho \leqslant 1$, then for $x \geqslant 0$*

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

(b) *Overloaded case*: $\rho > 1$. *If $\rho > 1$, then for $x \geqslant 0$*

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

We now turn to the arrival rates. As can be seen from Theorem 12 above, unlike for the time-varying model, for the stationary model we can easily handle GI service, because the total service content B is independent of the service-time distribution beyond its mean. The vector of constant arrival rates λ is determined by the system of fixed-point equations

$$\lambda_j = \lambda_j^{(0)} + \sum_{i=1}^m (\lambda_i \wedge s_i\mu_i)P_{i,j}, \quad 1 \leqslant j \leqslant m, \tag{31}$$

where $\lambda, \lambda^{(0)}, s, \mu \in \mathbb{R}^m$, and P is an $m \times m$ stochastic matrix. We can write (31) more compactly as

$$\lambda = \Phi(\lambda) \equiv \lambda^{(0)} + (\lambda \wedge s\mu)P. \tag{32}$$

Equation (32) has already been analyzed by Goodman and Massey (1984) in the study of nonergodic Jackson networks; also see Chen and Mandelbaum (1991) and p. 168 of Chen and Yao (2001). However, the model here is different.

THEOREM 13 (FIXED-POINT EQUATION FOR STATIONARY ARRIVAL RATES, FROM GOODMAN AND MASSEY 1984). *The arrival rates in the stationary $(G/GI/s + GI)^m/M$ fluid queue network satisfy Equation* (31). *Hence, if the stochastic matrix has spectral radius less than* 1 (*which holds if*

and only if $P^n \to 0$ *as* $n \to \infty$*), then* Φ *in* (32) *is a monotone n-stage contraction operator on* $\mathbb{R}^m$ *with an appropriate norm, so that there exists a unique solution to the fixed-point equation in* (31) *and* (32)*. The fixed point can be calculated by solving at most m different systems of m linear equations.*

Proof. Even for GI service, if fluid queue i is underloaded, then the stationary service content is $B_i = \lambda_i/\mu_i$ and the service completion rate is $\sigma_i = B_i\mu_i = \lambda_i$. On the other hand, if queue i is overloaded, then $B_i = s_i$ and the service completion rate is $s_i\mu_i$. In all cases, the service completion rate at queue i is $\lambda_i \wedge s_i\mu_i$. Because there is a unique solution to Equation (31) or (32), that equation determines the stationary arrival rates at all queues and which queues are in fact overloaded. □

8. Heuristic Stochastic Refinement for Many-Server Queues

As illustrated by Figure EC.2, the fluid model performance functions are remarkably effective in approximating the performance of large-scale many-server queueing systems. That is to be expected because of many-server heavy-traffic limits, as we mentioned in §1. In §9 of Liu and Whitt (2010) we show that the fluid performance functions are useful more generally to describe the mean values of smaller-scale many-server queueing systems, e.g., with only 20 servers or even fewer, provided that they experience significant overloading at some times. That should be very helpful, but it is also of interest to better understand the stochastic fluctuations about those mean values in the queueing system.

For some of the stochastic processes in the $G_t/M/s_t + GI$ queueing model, where the service and abandonment are not time varying, we can invoke existing heavy-traffic limits for infinite-server queues. In particular, in the $G_t/M/s_t + GI$ queueing system, the stochastic process $\hat{B}(t, y)$ recording the number of customers in service at time t that have been so for time at most y is the same as in the $G_t/M/\infty$ model during each underloaded interval. Similarly, during each overloaded interval, the stochastic process $\tilde{Q}(t, y)$ recording the number of customers in queue at time t that have been so for time at most y, not allowing customers to enter service (parallelling Proposition 3), is the same as in the $G_t/GI/\infty$ model, with the abandonment cdf F playing the role of the service-time cdf. Thus, many-server heavy-traffic limits in Pang and Whitt (2010) apply to them, yielding Gaussian approximations.

More generally, we suggest a practical heuristic approximation that is in the spirit of those infinite-server results. The idea is very simple: We simply approximate the distribution of the total number of customers in the system, $X(t)$, by a *Poisson* distribution, taking the computed value from the fluid queue model as its mean. This simple Poisson approximation approach is in fact *exact* in the special case of the $M_t/GI/s_t + GI_t$ and $M_t/M_t/s_t + GI_t$ models if they are always underloaded, starting out empty at time 0 or in the distant past. As discussed in Liu and Whitt (2010), in that case the model reduces to the $M_t/GI/\infty$ or $M_t/M_t/\infty$ fluid model, for which the fluid values of $B(t) = X(t)$ coincide with the mean values in the stochastic model. In addition, $X(t)$ has a Poisson distribution for the stochastic infinite-server model. Finally, unless the mean is very small, we approximate the Poisson distribution by a normal distribution. For the underloaded system, this proposal coincides with §9 of Massey and Whitt (1993).

Given the approximation for $X(t)$ in the queueing system, we approximate the random variables $Q(t)$ and $B(t)$ using $Q(t) = (X(t) - s(t))^+$ and $B(t) = X(t) \wedge s(t)$, which leads to "one-sided" normal approximations, which in regions of significant overload or underload will tend to themselves be approximately normal. This heuristic refinement should give a rough idea about the stochastic fluctuations, adequate for many engineering applications; e.g., it shows that the stochastic fluctuations in $X(t)$ should be roughly of order $\sqrt{X(t)}$. However, we caution that this is very rough; the standard deviation might well be off by a factor of 2 or more. Refined stochastic approximations are still needed.

9. Conclusions

In §2 we specified the single $G_t/M_t/s_t + GI_t$ fluid queue; it differs from Liu and Whitt (2010) by having M_t service and GI_t abandonment instead of both being GI. The M_t service eliminates the need to solve a fixed-point equation to find the service content density b. In §§3 and 4 we showed that a single fluid queue can be analyzed by assuming that the arrival rate function λ, the staffing function s, and the service rate function μ are all piecewise polynomials. However, that did not permit an extension to networks because the departure rate function does not inherit that property. In §5 we used asymptotic methods to show how to analyze the single fluid queue without having to assume either (i) that the arrival rate function is piecewise polynomial or (ii) that there are only finitely many switches between overloaded and underloaded intervals in each finite interval. In §6 and §7 we showed how to treat the $(G_t/M_t/s_t + GI_t)^m/M_t$ and $(G/GI/s + GI)^m/M$ networks with proportional routing. Theorem 10 established the existence of unique vector of arrival rate functions, allowing for feedback, and thus a corresponding unique performance description for the entire network. The performance functions at each queue are given in §4.

As discussed in §9 of Liu and Whitt (2010), we have conducted simulation experiments showing that the fluid model provides very accurate approximations for very large-scale many-server queueing systems; we show the results of one such experiment in §EC.6. The approximations are also excellent for the *mean values* of the corresponding queueing random variables when the scale is

quite small, e.g., when there are 20 servers or fewer; e.g., see Figure 7 of Liu and Whitt (2010). We have provided a heuristic stochastic refinement in §8; it approximates the number of customers in the queueing system, first by a Poisson distribution having the fluid value as its mean, and then by a normal distribution.

There are many directions for future research. It remains to establish supporting many-server heavy-traffic limits, including stochastic refinements. It remains to examine the algorithms provided by Theorem 10 and Remark 2; it remains to extend Theorem 10 to GI and GI_t service. It remains to develop alternative approximations for time-varying many-server queueing systems, where the staffing adjusts dynamically (appropriately) to the time-varying demand, so that the system tends to be critically loaded at all times, as opposed to switching between overloaded intervals and underloaded intervals.

10. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgment

This research was supported by NSF grant CMMI 0948190.

References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Buzacott, J. A., J. G. Shanthikumar. 1992. *Stochastic Models of Manufacturing Systems*. Prentice Hall, Englewood Cliffs, NJ.

Chen, H., A. Mandelbaum. 1991. Discrete flow networks: Bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16**(2) 408–446.

Chen, H., D. D. Yao. 2001. *Fundamentals of Queueing Networks*. Springer, New York.

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

Goodman, J. B., W. A. Massey. 1984. The non-ergodic Jackson network. *J. Appl. Probab.* **21**(4) 860–869.

Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* **16**(1) 13–39.

Liu, Y., W. Whitt. 2010. The $G_t/GI/s_t + GI$ many-server fluid queue. Working paper, Columbia University, New York, http://www.columbia.edu/~ww2040/allpapers.html.

Mandelbaum, A., W. A. Massey, M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* **30**(1–2) 149–201.

Massey, W. A., W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* **13**(1–3) 183–250.

Nelson, B., M. Taaffe. 2004. The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II–the multiclass network. *INFORMS J. Comput.* **16**(3) 275–283.

Newell, G. F. 1982. *Applications of Queueing Theory*, 2nd ed. Chapman and Hall, London.

Pang, G., W. Whitt. 2010. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65**(4) 325–364.

Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* **4** 193–267.

Teschl, G. 2000. *Ordinary Differential Equations and Dynamical Systems*. Lecture Notes, University of Vienna, Vienna. Accessed July 2011, http://www.mat.univie.ac.at/~gerald/ftp/book-ode/.

Whitt, W. 1983. The queueing network analyzer. *Bell System Tech. J.* **62**(9) 2779–2815.

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

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Yom-Tov, G., A. Mandelbaum. 2010. The Erlang-R queue: Time-varying QED queues with reentrant customers in support of healthcare staffing. Working paper, The Technion, Haifa, Israel.