Submitted to *Operations Research*
manuscript OPRE-2009-06-259; OPRE-2011-12-597.R1

Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals

Yunan Liu
Department of Industrial Engineering, Room 446, 400 Daniels Hall, North Carolina State University, Raleigh, NC 27695, yliu48@ncsu.edu

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, ww2040@columbia.edu

An algorithm is developed to determine time-dependent staffing levels to stabilize the time-dependent abandonment probabilities and expected delays at positive target values in the $M_t/GI/s_t + GI$ many-server queueing model, which has a nonhomogeneous Poisson arrival process (the M_t), general service times (the first GI) and allows customer abandonment according to a general patience distribution (the $+GI$). New offered-load and modified-offered-load approximations involving infinite-server models are developed for that purpose. Simulations show that the approximations are effective. A many-server heavy-traffic limit in the efficiency-driven regime shows that: (i) the proposed approximations achieve the goal asymptotically as the scale increases, and (ii) it is not possible to simultaneously stabilize the mean queue length in the same asymptotic regime.

Key words: staffing; capacity planning; many-server queues; queues with time-varying arrivals; queues with abandonment; infinite-server queues, offered-load approximations; service systems;

History: Submitted on June 3, 2009; revisions (regarded as new submission) on December 1, 2011, and March 28, 2012; accepted July 12, 2012

1. Introduction

In this paper we develop new staffing algorithms for many-server queueing systems with time-varying arrivals, focusing on the challenging case in which service times are relatively long and there is customer abandonment from queue; see Green et al. (2007) for background. Specifically, we develop formula-based algorithms to stabilize the time-dependent abandonment probability and the expected (potential) delay (the delay before starting service for a customer arriving at time t with infinite patience) at any (necessarily related) fixed targets, across a wide range of possible targets, in the $M_t/GI/s_t + GI$ queueing model. This model has a nonhomogeneous Poisson arrival

process (the M_t), a large number s_t of homogeneous servers working in parallel (a function of time t, which is to be determined), independent and identically distributed (i.i.d.) service times with a general distribution (the first GI), unlimited waiting space, the first-come first-served (FCFS) service discipline, and customer abandonment from the queue of customers waiting to start service, with i.i.d. times to abandon having a general distribution (the second GI).

Our results extend Feldman et al. (2008), which introduced a simulation-based *iterative staffing algorithm* (ISA) to stabilize the time-dependent delay probability (the probability that an arrival must wait in queue before starting service). As illustrated by Figure 3 in that paper, the ISA was shown to be remarkably effective at stabilizing the delay probability in the fully Markovian $M_t/M/s_t + M$ model, for all possible quality-of-service (QoS) levels, provided that the arrival rate is suitably large, so that a large number of servers (e.g., 100) is actually required. The target delay probability was allowed to range from 0.1 (high QoS) to 0.9 (low QoS). Indeed, as illustrated by Figures 5 and 6 of that paper, with ISA staffing the delay probability becomes essentially the same as in a corresponding stationary model with constant arrival rate, and thus also showing that the analytically-based modified-offered-load (MOL) approximation (reviewed here in §2) succeeds in stabilizing delays for this $M_t/M/s_t + M$ model. Thus, from the perspective of the delay probability, the effect of the time-varying arrival rate can be eliminated by applying the ISA (or the MOL) to choose the staffing level appropriately.

Since the ISA is based on simulation, it can easily be applied to associated non-Markovian models and even to much more general models. Indeed, experiments showed that the ISA is also effective for stabilizing the delay probability in the more general $M_t/GI/s_t + GI$ model considered here. The ISA also has the advantage of providing automatic verification: Since ISA is based on simulation, we can confirm that ISA achieves its goal in the final simulation results.

Even though the ISA can stabilize the delay probability, it is unable to eliminate the effect of the time-varying arrival rate entirely. When ISA is applied to stabilize the delay probability, the ISA also stabilizes other performance measures to some extent, but as illustrated by Figure 4 of Feldman et al. (2008) and Figure 6 of the accompanying e-companion, significant fluctuations are

seen in the time-varying abandonment probability and average delay with a sinusoidal arrival-rate function when the target QoS is relatively low (the target delay probability is high). An open problem posed in §8 of Feldman et al. (2008) was to develop a way to stabilize the abandonment probability across the full range of target abandonment probabilities.

We address that open problem in this paper. Moreover, we do so with a formula-based algorithm, instead of simulation. Our formula-based algorithm applies to the general $M_t/GI/s_t + GI$ model. Since all performance measures tend to be stabilized together at customary targets with higher QoS, we succeed in obtaining an effective formula-based algorithm for staffing to meet all the standard performance measures at customary targets with higher QoS.

There are two important steps here. The first is to introduce an entirely new *offered-load* (OL) framework, involving two *infinite-server* (IS) queues in series, which we call the *delayed-infinite-server* (DIS) approximating model. (We give background on OL approximations in §2.) The first IS queue represents the waiting room (queue), while the second represents the service facility. The mean number of busy servers in the second IS queue represents the new OL. When the targeted QoS is relatively low (the abandonment probability is high), the OL itself directly yields a good staffing function, called DIS staffing. Moreover, when we use DIS staffing, the entire DIS model serves as a useful approximation for the original $M_t/GI/s_t + GI$ model, so that we obtain useful formulas approximating other performance measures, such as the time-varying mean queue length; see Theorem 1 below. We thus see that the mean queue length cannot be stabilized at the same time as the abandonment probability under higher abandonment-probability targets, and we can quantify the fluctuations in the mean queue length.

We substantiate the good performance of the DIS approximation for heavily loaded systems by establishing a heavy-traffic limit theorem. In particular, we apply our recent heavy-traffic fluid limits for many-server models with time-varying arrival rate and staffing in Liu and Whitt (2012a,b,c) to show that the DIS approximation is asymptotically correct in the overloaded or efficiency-driven (ED) many-server heavy-traffic regime; see Garnett et al. (2002) for background. (Related limits appear in Kang and Ramanan (2010), Kaspi and Ramanan (2011), Mandelbaum et al. (1998),

Whitt (2006), but Mandelbaum et al. (1998) is restricted to the Markovian special case, while the others assume fixed staffing.) As a corollary, we deduce that it is not possible for any algorithm to simultaneously stabilize all performance measures across the full range of target values in the many-server heavy-traffic regime.

Unfortunately, however, the simple DIS approximation does not perform well in the common case in which the target QoS is high (the abandonment probability target is low), which tends to take the system out of the ED regime. In the second step we treat that case by introducing a new *modified-offered-load* (MOL) approximation, which uses the new DIS offered load; we call this the DIS-MOL approximation. We will show that the DIS-MOL approximation is not too dismal!

Paralleling previous MOL approximations in Jennings et al. (1996) and Feldman et al. (2008) (also see Jagerman (1975) and Massey and Whitt (1994, 1997)), our MOL approximation uses steady-state performance formulas from the associated stationary $M/GI/s+GI$ model in a time-varying manner, using an arrival rate determined by the DIS OL. This second step is not immediate either, because the $M/GI/s+GI$ model tends to be intractable. In this step, we apply the approximation for all steady-state performance measures in this model by an associated $M/M/s+M(n)$ model from Whitt (2005), which uses an exponential service time with the same mean and a state-dependent Markovian abandonment process. We conduct simulations to show that this new formula-based DIS-MOL staffing algorithm stabilizes the abandonment probability and the expected waiting time effectively across a wide range of targets. As indicated above, we actually achieve more: Since all performance measures tend to be stabilized together when the target QoS is high, in that case we actually achieve a formula-based staffing algorithm for both the delay probability and the abandonment probability, plus several other performance measures as well.

Here is how the rest of this paper is organized: We start in §2 by reviewing the analytical OL and MOL approaches to the staffing problem. Next in §3 we develop the DIS model and give explicit expressions for all the key performance measures. In §4 we show that the DIS approximation achieves its goal of stabilizing abandonment probabilities and expected delays as the scale increases.

There we also prove that it is impossible to simultaneously asymptotically stabilize all performance functions. In §5 we develop the new MOL approximation for normally loaded systems.

In §6 we perform simulation experiments to validate the approximations, considering the Markovian $M_t/M/s_t + M$ examples with sinusoidal arrival-rate function. We also consider corresponding many-server service systems with non-exponential service times and abandonment times in the e-companion and a longer version available on the authors' web pages. Our simulations add real system constraints including the discretization issues and specified staffing intervals.

In §7 we present extra details for the asymptotic results in §4. Finally, in §8 we draw conclusions. We present additional material in the e-companion and a longer version available on the authors' web pages; the contents are specified in §EC.1.

2. Background on Offered Load Approximations

The general idea of an OL approximation is to initially assume that there are as many resources (here servers) as needed and then see how many are actually used. After we have determined how many servers would be needed if there were no constraint on their availability, we staff to provide the number of servers needed. Since the model is stochastic, the time-dependent number of servers used itself is random, so we must use some deterministic time-dependent partial characterization of this time-dependent distribution, such as the time-dependent mean or that time-dependent mean plus some multiple of the associated time-dependent standard deviation, as in the staffing algorithm proposed by Jennings et al. (1996).

As a consequence, the basic OL approximation is an IS approximation: The time-dependent number in the $M_t/GI/s_t + GI$ system is approximated by the time-dependent number of busy servers in the corresponding $M_t/GI/\infty$ model, having the same arrival processes and service times. This step is effective because the IS model is remarkably easy to analyze. The number of busy servers in the $M_t/GI/\infty$ model at time t has a Poisson distribution with the mean

$$m_0(t) \equiv \int_{-\infty}^{t} \lambda(s) P(S > t - s)\, ds, \tag{1}$$

where $\equiv$ denotes "equality by definition" and S is a generic service time; see Eick et al. (1993a,b). (The subscript 0 will be explained later.)

The IS approximation leads to the classical *square-root-staffing* (SRS) formula

$$s_\gamma(t) = m_0(t) + \beta_\gamma \sqrt{m_0(t)}, \tag{2}$$

where γ is the target level of performance, $s_\gamma(t)$ is the required staffing level at time t for that target, β_γ is an associated QoS parameter, and $m_0(t)$ is the mean number of busy servers in the IS model. This mean $m_0(t)$ serves as the appropriate notion of *offered load* (OL) at time t, which is independent of both γ and the abandonment-time cdf. The SRS in (2) is based on a normal approximation for the exact Poisson distribution. If γ is the target delay probability, then β_γ is chosen to satisfy $P(N(0,1) > \beta_\gamma) = \gamma$.

The MOL approximation is a refinement of the OL approximation above, which is needed because in reality there are not infinitely many servers, so that the number in system is not actually so well approximated by a normal distribution. For the $M_t/GI/s_t + GI$ model, the MOL approximation for the performance at time t is the steady state performance in the associated stationary $M/GI/s+GI$ model with constant arrival rate

$$\lambda_0^{MOL}(t) \equiv \frac{m_0(t)}{E[S]}, \tag{3}$$

where $m_0(t)$ is the offered load and S is a random service time. The MOL staffing at time t is the smallest staffing level $s(t)$ such that the performance target is met. Since performance is difficult to analyze in the general $M/GI/s+GI$ model, here we propose using the approximation in Whitt (2005). For further discussion, see §5.

An alternative to using the exact $M/GI/s+GI$ steady-state formula for the delay probability in the MOL approximation is to use a heavy-traffic approximation for it. That leads to a direct application of the SRS, but with a new formula for the QoS parameter β_γ. This approach was used with the delay probability target in §4 of Jennings et al. (1996)) for the $M/M/s$ model and in

Feldman et al. (2008) for the more general $M/M/s+M$ model. For the $M/M/s+M$ model, the Garnett function approximating the delay probability obtained from Garnett et al. (2002) is

$$P(Delay) \approx \omega(-\beta, \sqrt{\mu/\theta}) \equiv G_1(\beta), \tag{4}$$

where $\omega(x,y) \equiv [1 + h(-xy)/(yh(x))]^{-1}$, $h(x) \equiv \phi(x)/\Phi(x)$, $\phi(x) \equiv (1/\sqrt{2\pi})e^{-x^2/2}$, $\Phi(x) \equiv \int_{-\infty}^{x} \phi(y)dy$. To stabilize the delay probability at target γ, it suffices to obtain a β_γ by inverting the Garnett function, letting $P(Delay) \equiv \gamma$.

3. The Delayed-Infinite-Server (DIS) Approximation

Here we use IS models in a new way. Instead of directly replacing the $M_t/GI/s_t+GI$ model by its $M_t/GI/\infty$ counterpart, which ignores the customer abandonment, we represent our model as two IS facilities in series: first the waiting room (or the queue), and then the service facility, as depicted in Figure 1.

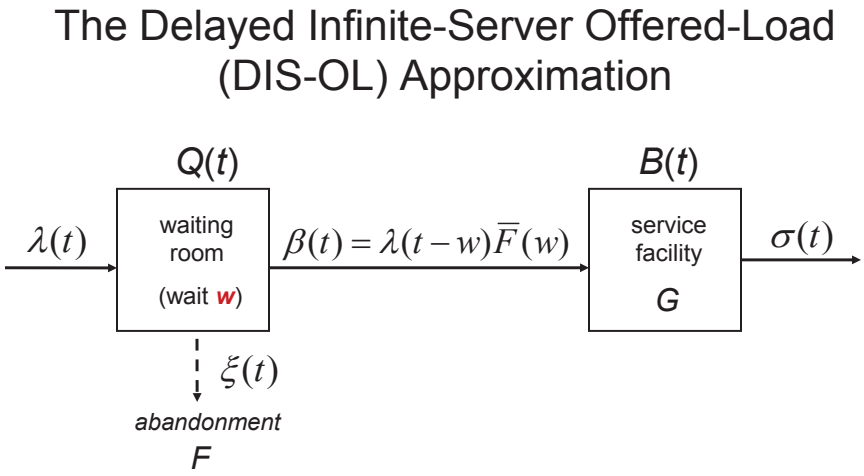

Figure 1 The delayed infinite-server (DIS) approximation for the $M_t/GI/s_t+GI$ queueing model. The contents $Q(t)$ and $B(t)$ are independent Poisson random variables for each t; the three flows are Poisson processes.

We start by assuming that our goal is to have every arrival that does not elect to abandon wait exactly time w before entering service. To achieve that goal in our approximation, we require that all external arrivals enter the waiting room that has infinite capacity and spend the fixed time w there before they move on to the service facility. (That is done in the approximation, not in the actual system.) While in the waiting room, each customer may abandon instead of entering service, after which the customer is lost. As in the original model, the abandonment times of successive arrivals to the queue are i.i.d. random variables with cumulative distribution function (cdf) F. The resulting model is the approximating DIS model.

In the DIS model with parameter w, the customer always enters service after spending time w in the queue, if the customer has not yet abandoned. That rule is possible because the service facility (the second IS queue) has infinitely many servers. We assume the system starts empty at time 0 and we let the first customer enter service after time w. Thus, for $t \geq w$, customers enter the service facility at rate $\lambda(t-w)\bar{F}(w)$, where $\lambda(\cdot)$ is the arrival-rate function and $\bar{F} \equiv 1 - F$.

Since all arrivals wait precisely the target duration w before entering service, if they do not elect to abandon, the approximate abandonment probability is always $F(w)$. Hence we can initially specify either the target abandonment probability $\alpha \equiv P(Ab)$ or the target delay w. If F is continuous, then there always is a w such that $F(w) = \alpha$ for any given α. If F is also strictly increasing, then $w = F^{-1}(\alpha)$. We assume that F is continuous and strictly increasing. Hence we can work with either α or w in the DIS model.

We now describe the performance in the approximating DIS model depicted in Figure 1. We start with a targeted waiting time w and the original $M_t/GI/s_t + GI$ model, specified by the arrival-rate function λ, the service-time cdf G and the abandonment-time cdf F. Let S and A be generic service-time and abandonment-time random variables; i.e., $G(x) \equiv P(S \leq x)$ and $F(x) \equiv P(A \leq x)$ for $x \geq 0$. Assume that $E[S] < \infty$. (We do not need to assume that $E[A] < \infty$ because in our approximation abandonments can only occur before time w.) Since F is continuous, F has no point mass at w, i.e., $P(A = w) = 0$, so there is no ambiguity about customer action after waiting in queue for time w.

The approximating model thus becomes a network of two $M_t/GI/\infty$ queues in series. The waiting room has arrival-rate function λ and service times distributed as $T \equiv A \wedge w \equiv \min\{A, w\}$, while the service facility has arrival rate $\lambda(t-w)\bar{F}(w)$, where $\bar{F}(x) \equiv 1 - F(x)$, and the given service-time cdf G. Let F_T be the associated cdf of the truncated random variable T, i.e.,

$$F_T(x) \equiv P(T \le x) = F(x), \quad 0 \le x < w, \quad F_T(x) = 1, \quad x \ge w. \tag{5}$$

We see that T has a point probability mass at w, since $P(T = w) = P(A \ge w) = \bar{F}(w)$.

As in Eick et al. (1993a), we assume that the system starts in the infinite past (at $t = -\infty$), with the policy above (customers entering service after waiting w if they have not yet abandoned). With this convention, all processes are defined on the entire real line. That is convenient both for some formulas and for representing the dynamic steady state associated with periodic arrival-rate functions, as in Eick et al. (1993b). If we want the system to start empty at time 0, then we can simply let $\lambda(t) = 0$ for all $t < 0$.

We can split the arrival process into two independent Poisson processes, one for the customers who will eventually be served and the other for the customers who will eventually abandon. Each customer is eventually served with probability $\bar{F}(w)$. We can further modify these two Poisson processes to obtain independent Poisson processes for the process counting customers entering service (at the times they enter service) and the process counting customers abandoning (at the times they abandon). Each can be represented as the departure process from an $M_t/GI/\infty$ queue, which corresponds to the original Poisson arrival process modified by having its points translated to the right by i.i.d. random variables; that is well known to preserve the Poisson property. For the counting process counting customers entering service, the service time is the constant value w; for the counting process counting customers abandoning, the service time is by the random value $(T|T < w)$. By this construction, we have proved that the arrival process to the service facility is indeed a nonhomogeneous Poisson process with rate $\lambda(t-w)\bar{F}(w)$ at time t. We can thus apply established results for the $M_t/GI/\infty$ model, in particular Theorem 1 of Eick et al. (1993a).

Let $Q(t,y)$ denote the number of customers in queue at time t that have remaining time before abandonment greater than y and let $Q(t) \equiv Q(t,0)$ denote the total number of customers in queue at time t. The random variable $Q(t,y)$ is depicted in Figure 2, which parallels Figure 1 of Eick et al. (1993a). We put a point at (x,y) in the plane if there is an arrival at x with abandonment time y. Thus $Q(t,y)$ will be the shaded region above the interval $[t-w,t]$ as shown. All arrivals with abandonment times greater than w will be served. All arrivals before time $t-w$ with abandonment times greater than w will have entered service before time t.

Let $B(t,y)$ denote the number of customers in the service facility at time t that have remaining service time greater than y and let $B(t) \equiv B(t,0)$ denote the total number of busy servers (number of customers in the service facility) at time t. Let $X(t) \equiv Q(t) + B(t)$ be the number of customers in the system at time t. Let $W(t)$ be the potential waiting time at time t, i.e., the virtual waiting time before entering service of a customer with infinite patience arriving at time t.

We now summarize the main structural results for the approximation. For a nonnegative random

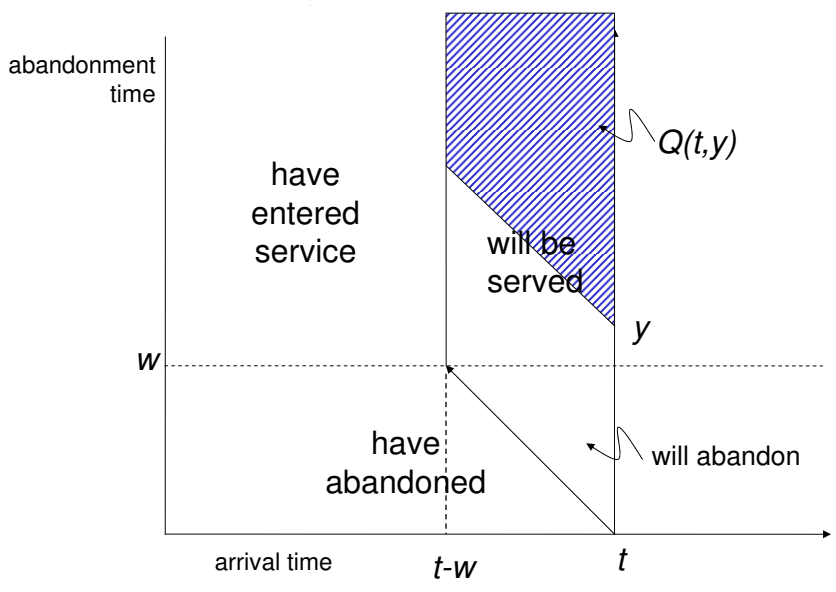

Figure 2 The random variable $Q(t,y)$ representing the queue content at time t that has remaining time before abandonment greater than y, in the DIS approximation.

variable Z with finite mean $E[Z]$ and cdf H, let Z_e denote a random variable with the associated *stationary-excess cdf* (or residual-lifetime cdf) H_e, defined by

$$H_e(y) \equiv P(Z_e \le y) \equiv \frac{1}{E[Z]} \int_0^y \bar{H}(x)\, dx, \quad y \ge 0. \tag{6}$$

The moments of Z_e can be easily expressed in terms of the moments of Z via

$$E[Z_e^k] = \frac{E[Z^{k+1}]}{(k+1)E[Z]}, \quad k \ge 1. \tag{7}$$

Both $B(t,y)$ and $Q(t,y)$ are functions of the model parameter α (or $w = F^{-1}(\alpha)$), but in this section we drop the subscript for simplicity.

THEOREM 1. *Consider the DIS approximation for the $M_t/GI/s_t+GI$ model specified above, starting in the distant past with specified delay target (parameter) $w \ge 0$ or abandonment probability target $\alpha = F(w)$. The approximation makes $W(t) = w$ with probability 1 and the probability of abandonment $F(w)$ for all arrivals. Moreover, $Q(t,y_1)$ and $B(t,y_2)$ are independent Poisson random variables for each t and each pair (y_1,y_2) with $y_1 > 0$ and $y_2 > 0$, having means*

$$\begin{aligned} E[Q(t,y)] &= \int_{t-w}^{t} \lambda(x)\bar{F}(t+y-x)\, dx = \int_0^w \lambda(t-x)\bar{F}(y+x)\, dx\ , \\ E[B(t,y)] &= \bar{F}(w) \int_{-\infty}^{t} \lambda(x-w)\bar{G}(t+y-x)\, dx = \bar{F}(w) \int_0^{\infty} \lambda(t-w-x)\bar{G}(y+x)\, dx. \end{aligned} \tag{8}$$

The total numbers of customers in queue and in service at time t, $Q(t)$ and $B(t)$ respectively, are independent Poisson random variables with means

$$\begin{aligned} E[Q(t)] = E[Q(t,0)] &= \int_{-\infty}^{t} \lambda(x)\bar{F}_T(t-x)\, dx = \int_{t-w}^{t} \lambda(x)\bar{F}_T(t-x)\, dx \\ &= E\left[\int_{t-T}^{t} \lambda(x)\, dx\right] = E[\lambda(t-T_e)]E[T], \end{aligned} \tag{9}$$

$$\begin{aligned} E[B(t)] = E[B(t,0)] &= \bar{F}(w) \int_{-\infty}^{t} \lambda(x-w)\bar{G}(t-x)\, dx \\ &= \bar{F}(w) E\left[\int_{t-w-S}^{t-w} \lambda(x)\, dx\right] = \bar{F}(w)E[\lambda(t-w-S_e)]E[S]. \end{aligned} \tag{10}$$

Thus, $X(t)$, the total number of customers in the system at time t is a Poisson random variable

with a mean $E[Q(t)] + E[B(t)]$. *The processes counting the numbers of customers abandoning and entering service are independent Poisson processes with rate functions* ξ *and* β, *respectively, where*

$$\begin{aligned} \xi(t) &= \int_0^w \lambda(t-x)\, dF(x) = E[\lambda(t-T)|T<w], \\ \beta(t) &= \lambda(t-w)\bar{F}(w). \end{aligned} \tag{11}$$

The departure process (*counting the number of customers completing service*) *is a Poisson process with rate*

$$\sigma(t) = \bar{F}(w) \int_0^\infty \lambda(t-w-x)\, dG(x) = \bar{F}(w) E[\lambda(t-w-S)]. \tag{12}$$

Proof. For the most part, these results are a direct application of Theorem 1 of Eick et al. (1993a). Understanding is facilitated by drawing pictures of Poisson random measure. Here we elaborate on only one point: To establish the claim that $Q(t, y_1)$ is independent of $B(t, y_2)$ for each (y_1, y_2), first observe that the departure process from the queue prior to time t is independent of $\{Q(t, y) : y \geq 0\}$. That implies that the process of customers entering service prior to time t is also independent of $\{Q(t, y) : y \geq 0\}$. But $B(t, y_2)$ depends only on the history of Q up to time t through its own arrival process up to time t. ∎

As discussed in Eick et al. (1993a), the last two representations for $E[Q(t)]$ and $E[B(t)]$ in (10) are appealing because they show random time lags, but these random time lags appear inside the expectation in a nonlinear fashion. We get convenient explicit formulas when the arrival rate function λ has special structure, e.g., when λ is sinusoidal or a polynomial, as we show in §EC.2. As in Eick et al. (1993a) and Massey and Whitt (1997), the polynomial arrival rate functions yield useful approximations. We can see directly that the targeted wait of w before starting service increases the random time lags in $E[B(t)]$ by w. This will be negligible if w is small, but not otherwise. As noted in Eick et al. (1993a), the time lag in $E[B(t)]$ involving $w + S_e$ is different from the time lag $w + S$ appearing in the departure rate $\sigma(t)$.

Since the departure process from the service facility is a Poisson process, we see that the approximation extends immediately to yield corresponding approximations for any number of $M_t/GI/s_t + GI$ models in series, in the spirit of Massey and Whitt (1993).

We now indicate how the DIS approximation can be used to specify the staffing function $s_\alpha(t)$ in the original $M_t/GI/s_t + GI$ model in order to achieve target $\alpha \equiv P(Ab)$. For any given target abandonment probability α with the direct DIS approximation, the number of busy servers at time t would be the random variable $B_\alpha(t)$. With the DIS approximation, $B_\alpha(t)$ has a Poisson distribution with mean $m_\alpha(t) \equiv E[B_\alpha(t)]$, for which we give an explicit formula. Hence $B_\alpha(t)$ is approximately normally distributed with both mean and variance equal to $m_\alpha(t)$.

The *simple DIS staffing approximation* is to simply set $s_\alpha(t) = m_\alpha(t)$. We will show that the simple DIS staffing policy is effective when the QoS is low, but not when the QoS is high.

4. Asymptotic Stability

In this section we prove that simple DIS staffing $s_\alpha(t) = m_\alpha(t)$ is effective in stabilizing the abandonment probability and the expected delay at any positive target values α and w related by $\alpha = F(w)$ asymptotically as the scale increases. For that purpose, we apply the many-server heavy-traffic FWLLN established in Liu and Whitt (2012a,b,c). That FWLLN involves a sequence of $G_t/GI/s_t + GI$ queueing models indexed by n. Model n has a general arrival process with time-varying arrival rate $\lambda_n(t) \equiv n\lambda(t)$ (which covers the M_t assumption of the current paper), i.i.d. service times with cumulative distribution function (cdf) G, a time-varying number of servers $s_{n,\alpha}(t) \equiv \lceil n\, s_\alpha(t) \rceil$ (the least integer above $ns_\alpha(t)$) and customer abandonment from queue, where the patience times of successive customers to enter queue are i.i.d. with general cdf F. The two cdf's G and F are fixed independent of n, and differentiable, with positive probability density functions (pdf's) g and f.

Our scaling of the fixed functions λ and s induces the familiar many-server heavy-traffic scaling. Under that scaling, and under regularity conditions, Liu and Whitt (2012b,c) establish a FWLLN with convergence of the appropriately scaled stochastic processes to deterministic performance functions associated with the fluid model analyzed in Liu and Whitt (2012a). Under this scaling and these regularity conditions, we now show that the DIS staffing achieves stability asymptotically; i.e., the time-dependent mean delay and abandonment probability are stabilized as $n \to \infty$.

To state the result, let $Q_n(t)$ be the number of customers waiting in queue at time t in the n^{th} queueing model. Let $W_n(t)$ be the corresponding potential waiting time, i.e., the virtual waiting time at time t if there were an arrival at time t, assuming that arrival had unlimited patience. Let $A_n(t)$ be the number of customers that have abandoned in the interval $[0,t]$. Let $A_n(t,u)$ be the number of customers among arrivals in $[0,t]$ that have abandoned in the interval $[0,t+u]$. Let $\bar{Q}_n(t) \equiv n^{-1}Q_n(t)$, $\bar{A}_n(t) \equiv n^{-1}A_n(t)$ and $\bar{A}_n(t,u) \equiv n^{-1}A_n(t,u)$ be the associated FWLLN-scaled processes. (The process $W_n(t)$ is not scaled.) Let $\Lambda(t) \equiv \int_0^t \lambda(s)\,ds$. Let 1_C be the indicator variable, which is equal to 1 if event C occurs and is equal to 0 otherwise.

Since the arrival processes are nonhomogeneous Poisson processes here, both a functional central limit theorem (FCLT) and a FWLLN hold for the arrival processes, as required in Liu and Whitt (2012b,c). To establish convergence of expected potential waiting times, we assume the the regularity conditions in Liu and Whitt (2012a,b,c) are satisfied, namely, all model parameters (λ, F and G) are piecewisely continuous and differentiable. We assume in addition that the service times have finite second moments.

THEOREM 2. (*asymptotic stability*) *Consider a sequence of $M_t/GI/s_t+GI$ models with the many-server heavy-traffic scaling specified above. Suppose that these systems start empty at time* 0*, the regularity conditions in* Liu and Whitt (2012a,b,c) *are satisfied and $E[S^2] < \infty$. Then, with the abandonment-probability target α, under the DIS staffing $s_{n,\alpha}(t) \equiv \lceil n\, s_\alpha(t) \rceil$, where*

$$s_\alpha(t) = m_\alpha(t) \equiv E[B_\alpha(t)] = \bar{F}(w) \int_0^{t-w} \bar{G}(x)\lambda(t-w-x)dx \cdot 1_{\{t>w\}}, \tag{13}$$

the expected delays and abandonment probabilities are stabilized at their targets w and α with $\alpha = F(w)$ asymptotically as $n \to \infty$; i.e., for any time b with $w < b < \infty$,

$$\begin{aligned} &\sup_{0\le t\le b}\{|\bar{Q}_n(t) - E[Q(t)]|\} \;\Rightarrow\; 0, \quad \sup_{0\le t\le b}\{|W_n(t)-w|\} \Rightarrow 0, \quad E[W_n(t)] \to w, \quad t \ge 0, \\ &\sup_{0\le t\le b}\{|\bar{A}_n(t) - A(t)|\} \Rightarrow 0 \quad \textit{and} \quad \sup_{0\le t\le b,\, w<u<b}\{|\bar{A}_n(t,t+u) - A(t,u)|\} \Rightarrow 0 \end{aligned} \tag{14}$$

as $n \to \infty$, where

$$E[Q(t)] = E[Q(t,0)] \equiv \int_0^w \lambda(t-x)\bar{F}(x)\,dx,$$

$$A(t) \equiv \int_0^t \xi(s)\,ds, \quad \xi(t) \equiv \int_0^w \lambda(t-x)f(x)\,dx \quad \textit{and} \quad A(t,u) \equiv \Lambda(t)\alpha, \quad u > w. \tag{15}$$

REMARK 1. (stabilizing abandonment probabilities, not rates) From (15), we see that the abandonment rate $\xi(t)$ is not stabilized. Instead, the proportion of arrivals arriving in any time interval $[t_1, t_2]$ that eventually abandon before starting service approaches α. That is consistent with our goal to stabilize the abandonment probability, as opposed to the abandonment rate. That is achieved starting empty by not staffing until time w.

Proof. Under the regularity conditions specified in Liu and Whitt (2012a,b,c), the appropriately scaled versions of the stochastic processes describing the performance in the $G_t/GI/s_t + GI$ queueing models indexed by n converge to corresponding deterministic functions describing the performance of an associated deterministic fluid model having capacity $s(t)$ and fluid input arriving at rate $\lambda(t)$ at time t. The performance of this limiting fluid model is characterized in Liu and Whitt (2012a).

A key assumption in Liu and Whitt (2012a,b,c) is that the fluid model alternates between underloaded intervals and overloaded intervals, with critical loading only holding at isolated points of switching from one regime to another. When we use the DIS staffing in order to stabilize abandonments at a positive target α, we are forcing the system to always operate in an overloaded regime after an initial transient required for the capacity to be filled. Thus, by staffing in that way, we consider the special case in which the system remains overloaded after an initial underloaded interval; i.e., there is a single switching point in the fluid model at $t = w$ (the delay target). In the terminology of Garnett et al. (2002), the limit is for the quality-driven (underloaded) many-server heavy-traffic regime before time w and the efficiency-driven (overloaded) many-server heavy-traffic regime after time w.

In §10 of Liu and Whitt (2012a) it was shown how to staff the fluid model in order to stabilize delays and abandonments. (Abandonment probabilities in the queueing model correspond to proportions of entering fluid that abandon before entering service in the fluid model.) For simplicity, in §10 of Liu and Whitt (2012a) it was assumed that the fluid model starts empty at time 0, and

so we made that assumption, but it was shown how to treat more general initial conditions in §H of the appendix of Liu and Whitt (2012a). When starting empty, the staffing policy to stabilize delays of all fluid to enter service at a target w provides no staffing at all, and thus allows no fluid to enter service, until time w. The stabilizing staffing function after time w is given by (13) above. Given that the delays are stabilized at w, the proportion of arriving fluid at each time to abandon before entering service is $\alpha = F(w)$. Since F is continuous and strictly increasing, it has a unique inverse F^{-1}, so that we could start with α instead of w, and then let $w = F^{-1}(\alpha)$.

Moreover, as discussed in §4 of Liu and Whitt (2012a), there is an intimate connection between the fluid content at time t and the mean of the number of busy servers in an associated IS model. As a consequence, this staffing function for the fluid model coincides with the simple DIS staffing, adjusted appropriately for the scaling factor n; i.e., in the fluid model, $s_\alpha(t) = E[B_{\alpha,n}(t)]/n$, if we also assume that the fluid model starts empty and we first provide staffing at time w, where w is chosen so that $w = F^{-1}(\alpha)$. Thus we can combine the results above to deduce that DIS staffing achieves its goal asymptotically.

As a consequence, if we use the simple DIS staffing in the fluid model, we succeed in stabilizing the delays in the fluid model. We then apply the FWLLN with that particular staffing function. Then the FWLLN in Liu and Whitt (2012b,c) directly applies the stated limits for $\bar{Q}_n(t)$, $W_n(t)$ and $\bar{A}_n(t)$ in (14), Since all fluid waits exactly w before entering service, if it does not abandon, the same will be true asymptotically for the stochastic model. Thus we get the stated limit for $\bar{A}_n(t,u)$ with $u > w$.

Finally, we apply uniform integrability to get the convergence of means $E[W_n(t)] \to w$ for each t from the established convergence $W_n(t) \Rightarrow w$, using p. 31 of Billingsley (1999). To obtain the uniform integrability, we show that $\sup_{n\geq 1} E[W_n(t)^2] < \infty$. That is proved in §7.1. ∎

From the representation of the DIS approximating mean queue length $E[Q_\alpha(t)]$ in Theorem 1, we can show that it cannot be a constant function with a time-varying arrival rate function. Hence, Theorem 2 together with two additional lemmas below implies the following corollary.

COROLLARY 1. (*the mean queue length is not stabilized asymptotically*) *Suppose that the conditions of Theorem* 2 *hold with the arrival rate function* λ *not being a constant function. Then, under DIS staffing, the scaled number in queue* $n^{-1}Q_n(t)$ *converges to the mean DIS queue length, which cannot be a constant function of* t *after time* w*. Thus, the DIS staffing function that asymptotically stabilizes the abandonment probability does not asymptotically stabilize the mean queue length.*

We use the following lemma to prove Corollary 1. We prove this lemma and the following one in §7.2.

LEMMA 1. (*uniqueness of time-shifted integral*) *If*

$$m(t) = \int_0^w \lambda(t-x)\bar{F}(x)\,dx, \quad t \geq w, \tag{16}$$

is a positive constant for all $t \geq w$*, then* λ *is a constant function for* $t \geq 0$*.*

We now observe that the DIS staffing is essentially the only staffing function that can stabilize abandonments and delays.

COROLLARY 2. (*uniqueness*) *The DIS staffing in* (13) *is the unique staffing function that stabilizes abandonment and delays at positive values in the fluid model. Consequently, any other sequence of staffing functions* $\{s_n : n \geq 1\}$ *that asymptotically stabilizes abandonment and delays must satisfy* $n^{-1}s_n(t) \to s_\alpha(t)$ *as* $n \to \infty$*.*

However, the MOL staffing function for the stochastic system is unique only in the order of $o(n)$, according to the fluid scaling. We use the following lemma to prove Corollary 2. The following lemma shows that there are not multiple staffing functions that produce identical potential waiting time functions or identical abandonment rate functions in the limiting fluid model.

LEMMA 2. (*unique fluid model staffing functions yielding given targeted performance*) *For the* $G_t/GI/s_t + GI$ *deterministic fluid model specified by the model data* $(\lambda, s, G, F, b(0,\cdot), q(0,\cdot))$ *satisfying the assumptions of* Liu and Whitt (2012a) *starting empty at time* 0*, the DIS staffing in* (13) *is the unique staffing yielding the positive constant target delay* w *and abandonment* $\alpha = F(w)$*.*

We can combine Corollaries 1 and 2 to deduce the following corollary

COROLLARY 3. (*impossibility of simultaneous stabilization*) *There does not exist a staffing function that can simultaneously stabilize the abandonment probability and the mean queue length at positive targets asymptotically in the many-server heavy-traffic regime.*

5. The DIS-MOL Approximation

Consistent with Theorem 2, in simulation experiments we see that the simple DIS staffing is remarkably effective in stabilizing the abandonment probability and the expected delays in large-scale queueing models with relatively low QoS (high abandonment probabilities and expected delays); see Figure 4 for an example. Unfortunately, however, the DIS approximation does not perform so well for higher QoS (lower abandonment probabilities and expected delays). Such lighter loads tend to move the system from the ED asymptotic regime to the QED asymptotic regime. (With higher QoS, the scale must be extremely large, such as $n = 1000$ or more, before the DIS staffing is effective. Experience indicates that the required scale increases as α decreases.)

Fortunately, for such higher QoS, a new MOL approximation performs remarkably well; see Figure 5 for an example. We now develop it. Let $P_t(Ab)$ be the time-dependent probability that an arrival at time t will eventually abandon. For a stationary model, the offered load can be defined as $\lambda(1 - P(Ab))E[S]$, the rate customers enter service, $\lambda(1 - P(Ab))$, multiplied by the mean service time, $E[S]$. A candidate time-dependent generalization would be the pointwise-stationary approximation for the offered load, $\lambda(t)(1-P_t(Ab))E[S]$, but $\lambda(t)(1-P_t(Ab))$ is the rate of arrivals at time t that will eventually enter service; it is not the rate customers actually enter service at time t. By Little's law applied to the service facility in the stationary model, $\lambda(1 - P(Ab))E[S]$ is also the expected number of busy servers in steady state. Experience indicates that the mean number of busy servers in the IS system tends to be a far better analog of the offered load in a nonstationary model.

Thus, to obtain the DIS-MOL approximation for the staffing $s(t)$ at time t to achieve the target

α, we let $w = F^{-1}(\alpha)$ and we look at the steady-state distribution of the stationary $M/GI/s+GI$ model applied with $s = s(t)$ and arrival rate

$$\lambda_\alpha^{MOL}(t) \equiv \frac{m_\alpha(t)}{E[S](1-\alpha)}. \tag{17}$$

For each fixed time t, we let the DIS-MOL staffing level $s_\alpha^{MOL}(t)$ be the least integer staffing level such that the stationary abandonment probability $P(Ab)$ in the $M/GI/s+GI$ model with arrival rate $\lambda = \lambda_\alpha^{MOL}(t)$ in (17) and $s = s_\alpha^{MOL}(t)$ satisfies $P(Ab) \le \alpha$.

Since the stationary $M/GI/s+GI$ model itself is quite complicated, we apply the approximation from Whitt (2005), which is based on an associated Markovian $M/M/s+M(n)$ model with state-dependent abandonment rates. Alternatively, since that approach approximates the general service-time distribution by an exponential distribution with the same mean, one can use the exact solution or asymptotic approximations for the associated $M/M/s+GI$ model from Zeltyn and Mandelbaum (2005). Either way, we are exploiting the property that the service-time distribution beyond its mean tends not to matter to much in the stationary $M/GI/s+GI$ model; see Whitt (2005, 2006). In an application this property can be checked with simulation.

Remark 2. (sensitivity to the service-time distribution beyond its mean) Since the stationary $M/GI/s/0$ loss model and the stationary $M/GI/\infty$ IS model have the celebrated insensitivity property, i.e., since the steady-state performance is independent of the service-time distribution beyond its mean, it is not too surprising that the stationary $M/GI/s+GI$ model should exhibit *approximate* insensitivity to the service-time distribution beyond its mean. However, unlike the performance in these stationary models, the insensitivity is lost when we abandon the stationarity, as was shown in Davis et al. (1995) for the loss model. Similarly, the transient performance in the $M_t/GI/s+GI$ model with time-varying arrival rate is more sensitive to the service-time distribution beyond its mean, as shown by the example in §2 of Liu and Whitt (2012a). That sensitivity is captured in our approximation through the impact of the service-time distribution beyond its mean on the transient performance of the time-varying IS and DIS models.

6. An $M_t/M/s_t+M$ Example with a Sinusoidal Arrival-Rate Function

In this section, we use simulation experiments to show that both the abandonment probability $P_t(Ab)$ and the expected delay $E[W(t)]$ are indeed stabilized (independent of time) in Markovian $M_t/M/s_t+M$ examples for low QoS targets with the DIS algorithm and for all QoS targets with the DIS-MOL algorithm. We show that the new methods also work for non-exponential service and patience distributions in §EC.4 and the longer version available on the authors' web pages.

Our staffing procedure applies to arbitrary arrival-rate functions, because we can apply Theorem 1 to calculate the DIS OL function $m(t)$. Following common practice in the study of time-varying arrival rates, we consider a sinusoidal arrival-rate function

$$\lambda(t) = a + b \cdot \sin(ct), \quad 0 \le t \le T. \tag{18}$$

This sinusoidal example is convenient because we can apply Theorem 1 to obtain explicit analytical formulas for the offered load $m_\alpha(t)$, paralleling Eick et al. (1993b). This sinusoidal example also roughly captures the spirit of real systems, as seen in call center data, as in Figure 7 of Feldman et al. (2008). We obtain a concrete many-server example by letting $a=100$, $b=20$, $c=1$ and $T=20$ in (18). Here we let the individual service rate μ be 1 and the individual abandonment rate θ be 0.5. (Choosing $\mu=1$ corresponds to measuring time in units of mean service times.)

An important issue for applications is the rate of fluctuation in the arrival-rate function compared to the expected service time. Since a cycle of the sinusoidal arrival-rate function in (18) is $2\pi/c$ and we have set $c=1$, the length of a cycle is $2\pi \approx 6.3$ Thus there will be slightly more than three cycles during the interval $[0, 20]$; e.g., see Figure 4. If we measure time in hours, then the mean service time is one hour and the full time period is slightly less than one day. Then the three peaks in the arrival rate in Figure 4 occur approximately at at 2am, 8am and 2pm. Thus in this example the fluctuation in the arrival rate function are relatively fast compared to the expected service time. From Jennings et al. (1996) and Feldman et al. (2008), we know that the pointwise-stationary approximation does not perform well for this example. That is demonstrated for the abandonment probability and the expected delay in the appendix.

We now turn to DIS staffing. We apply Theorem 1 and Eick et al. (1993b) to calculate the DIS performance functions. Letting $\bar{F}(x) = e^{-\theta x}$, $\bar{G}(x) = e^{-\mu x}$, and $\lambda(t) \equiv 0$ for $t < 0$, we obtain

$$E[B(t)] = \begin{cases} e^{-\theta w}\frac{a}{\mu} - (\frac{a}{\mu} - \frac{bc}{\mu^2+c^2})e^{-\theta w-\mu(t-w)} + \frac{be^{-\theta w}}{\mu\sqrt{1+c^2/\mu^2}}\sin[c(t-w)-\phi], & t \geq w \\ 0, & 0 \leq t < w, \end{cases} \tag{19}$$

$$E[Q(t)] = \begin{cases} \frac{a}{\theta}(1-e^{-\theta w}) + b\sqrt{\frac{x^2+y^2}{\theta^2+c^2}}\sin(ct+\beta-\gamma), & t \geq w \\ \frac{a}{\theta} - (\frac{bc}{\theta^2+c^2} - \frac{a}{\theta})e^{-\theta t} + \frac{b}{\sqrt{\theta^2+c^2}}\sin(ct-\gamma), & 0 \leq t < w, \end{cases} \tag{20}$$

$$\textit{where} \quad \phi = \arctan(c/\mu), \quad \gamma = \arctan(a/\theta), \quad \beta = \arctan(y/x)$$

$$x = 1 - e^{-\theta w}\cos(cw), \quad y = e^{-\theta w}\sin(cw)$$

From (19), we see that $E[B(t)]$ is asymptotically sinusoidal as $t \to \infty$, eventually coinciding with the formula for the periodic steady state, given for the more general $M_t/GI/s_t + GI$ model in Theorem **??**. From (20), we see that $E[Q(t)]$ is sinusoidal when $t \geq w$.

We first compare the DIS-MOL staffing function $s_\alpha^{MOL}(t)$ to the DIS staffing function, which

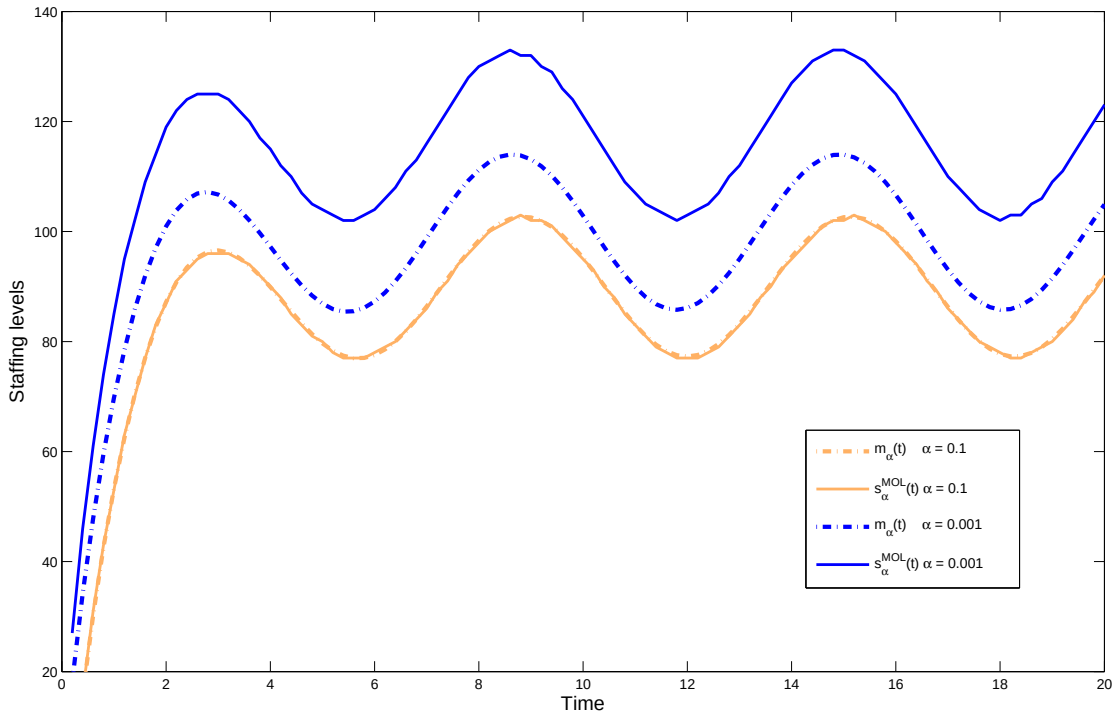

Figure 3 A comparison of the OL function $m_\alpha(t)$ and the DIS-MOL staffing function $s_\alpha^{MOL}(t)$ for the $M_t/M/s_t + M$ example with sinusoidal arrival rate for abandonment probability targets $\alpha = 0.1$ and 0.001.

is the OL $m_\alpha(t)$. In Figure 3, we plot $\{s_\alpha^{MOL}(t) : 0 \le t \le T\}$ and $\{m_\alpha(t) : 0 \le t \le T\}$ $(T = 20)$ with two abandonment probability targets: $\alpha = 0.1$ and $\alpha = 0.001$. These two targets are extreme cases. The first one ($\alpha = 10\%$) represents a low QoS; the second one ($\alpha = 0.1\%$) represents a high QoS. Figure 3 shows that $\{s_\alpha^{MOL}(t) : 0 \le t \le T\}$ and $\{m_\alpha(t) : 0 \le t \le T\}$ coincide when the QoS is low. Therefore, in order to stabilize abandonment under low QoS, we can just staff the system with the OL function $m_\alpha(t)$, given in Theorem 1. However, when the QoS is high, Figure 3 shows that the DIS-MOL staffing function is significantly above the OL function, which makes the MOL refinement necessary. Our experiments indicate that the DIS-MOL staffing is never less than the OL (DIS staffing).

We now use simulation experiments to show that DIS-MOL approximation achieves the desired time-stable performances under all QoS targets. First we evaluate the simple DIS approximation at low QoS targets (where it is nearly identical to DIS-MOL). Figure 4 shows simulation estimates of key performance measures with target abandonment probability α for $0.05 \le \alpha \le 0.20$. (Additional details about the simulation estimates are given in §EC.3.) Figure 4 shows that both the abandonment probability and the expected delay are stabilized at the targets α and $w = F^{-1}(\alpha) = -1/\theta \log(1-\alpha)$. Moreover, as predicted, the queue-length processes are not stabilized; they agree closely with the formulas in (20). All these quantities were estimated by performing multiple (5000) independent replications under the staffing function $s_\alpha(t) = m_\alpha(t) = E[B(t)]$ in (19).

Figure 4 shows that the simple DIS approximation works quite well for α between 0.05 and 0.20, and associated expected delays ranging from 0.1 to 0.45. However, at least 5% abandonment may not be acceptable. For higher QoS targets, the DIS-MOL approximation is needed.

In Figure 5, we again plot the expected queue lengths, abandonment probabilities, delay probabilities and expected delays, using the DIS-MOL approximations with relatively low abandonment probability targets $0.005 \le \alpha \le 0.02$. The DIS-MOL approximation works remarkably well after the initial transient period.

In the DIS model, since all customers are required to wait in queue for w before entering service, unlike the expected queue length, we are not able to produce an approximate delay probability,

because the DIS approximation predicts that every customer should be delayed. Of course, in the original stochastic model, the delay probability is always less than 1. The delay probability increases as the abandonment probability α increases. When α is big enough, e.g., $\alpha = 0.2$, the delay probability is nearly 1, as shown in Figure 4; when α is as small as 0.005, the delay probability goes down to 0.2, as shown in Figure 5. Figure 4 and 5 show that the delay probabilities are stabilized when α is small and are sinusoidal otherwise. This phenomenon is consistent with Figure 4 in Feldman et al. (2008), which shows the abandonment probabilities when the delay probability is the target. There, both were stabilized in with high QoS, but the abandonment probabilities were not stabilized under low QoS (high delay probability targets).

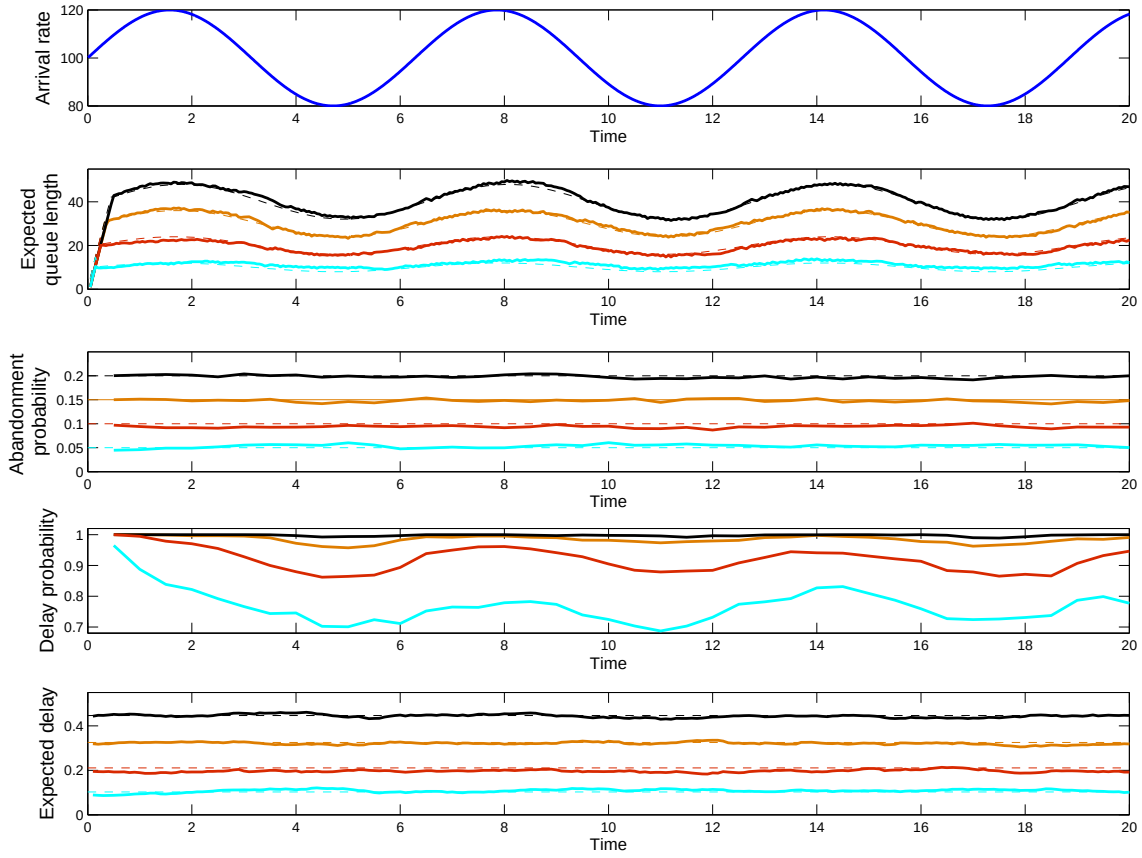

Figure 4 A comparison of the simple DIS approximation with simulation: Estimated time-dependent expected queue lengths, abandonment probabilities, delay probabilities and expected delays for the $M_t/M/s_t + M$ example with sinusoidal arrival under the OL Staffing, with low QoS ($\alpha = 0.05, 0.10, 0.15, 0.20$).

We conclude this section by remarking that the DIS-MOL approximation is consistent with a square-root-staffing (SRS) formula. Paralleling (2), the candidate new SRS formula for the $M_t/GI/s_t+GI$ model based on the DIS OL with abandonment target α would be

$$s_\alpha(t) = m_\alpha(t) + \beta_\alpha \sqrt{m_\alpha(t)}. \tag{21}$$

Formula (21) differs from (2) by using the DIS OL $m_\alpha(t)$ instead of the standard OL $m_0(t)$ in (1). In §EC.5 we verify that the DIS-MOL approximation is indeed consistent with the SRS staffing in (21); i.e., we show that the DIS-MOL staffing function $s_\alpha^{MOL}(t)$ has the form of the SRS formula (21). Following Feldman et al. (2008), for an abandonment probability target α, we let $D_\alpha(t) \equiv s_t^{MOL} - m_\alpha(t)$ be the difference between the DIS-MOL staffing and the OL functions, and let $\beta_\alpha(t) \equiv D_\alpha(t)/\sqrt{m_\alpha(t)}$ be the *implicit QoS function* for DIS-MOL. Figure EC.4 shows that $\beta_\alpha(t) \approx \beta_\alpha$, independent of t, where β_α is a nonnegative decreasing function of α. However, it remains to find

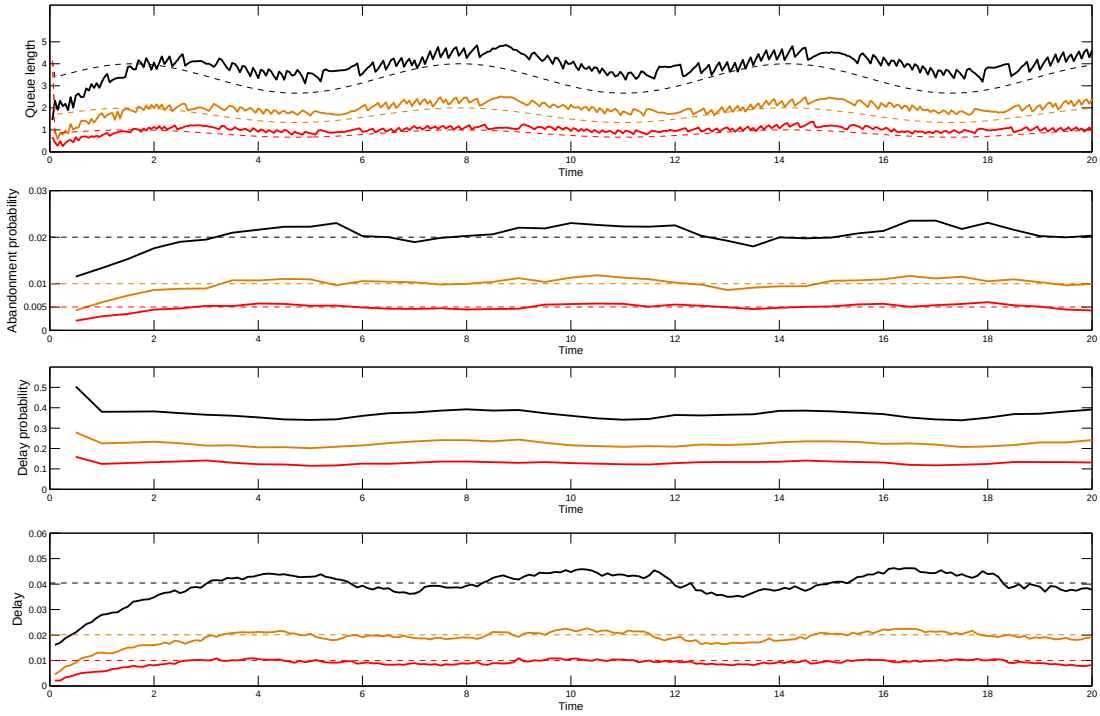

Figure 5 Estimated time-dependent expected queue lengths, abandonment probabilities, delay probabilities and expected delays for the $M_t/M/s_t+M$ example with sinusoidal arrival under the DIS-MOL staffing with relatively high QoS ($\alpha = 0.005$, 0.01 and 0.02).

a simple formula for the new QoS constant β_α. In §EC.5 we show results of fitting it to simulation data.

7. Additional Proofs for §4

7.1. A Bound to Justify Uniform Integrability in Theorem 2

To justify the uniform integrability used to imply convergence in moments $E[W_n(t)] \to w$ given the established convergence in distribution $W_n(t) \Rightarrow w$, we show that $\sup_{n\geq 1} E[W_n(t)^2] < \infty$, where $W_n(t)$ is the potential waiting time at time t in model n; see p. 31 of Billingsley (1999).

LEMMA 3. *Under the assumptions of Theorem* 2, $\sup_{n\geq 1} E[W_n(t)^2] < \infty$.

Proof. It suffices to use a crude upper bound. Thus we simplify. First, we bound the arrival process in model n, $N_n(t)$, above by a stationary Poisson process over the interval $[0,T]$ by letting the fluid arrival rate be $\lambda_{bd} \geq \lambda(t)$ for all t in a bounded interval $[0,T]$. Similarly, we bound the staffing function in the fluid model below by a positive capacity s_{bd} with $0 < s_{bd} \leq s(t)$, $0 \leq t \leq T$, which is possible by Assumption 11 of Liu and Whitt (2012a). We also assume that there is no abandonment. These modifications can be shown to increase the process $W_n(t)$ in sample path stochastic order, as in Whitt (1981).

With those simplifications, we have a sequence of classical $M/GI/s/\infty$ models, with arrival rate $n\lambda_{bd}$ and staffing level $\lceil ns_{bd} \rceil$, $n \geq 1$. We now focus on these models, without changing the notation. Let $W_n^a(k)$ be the actual delay of the k^{th} arrival and observe that $W_n(t) \leq W_n(N_n(t)+) \leq W_n^a(N_n(t)) + S$, where $N_n(t)$ is the homogeneous Poisson arrival process with rate $n\lambda_{bd}$ and S is a generic service time independent of $W_n^a(N_n(t))$. Next we bound the given $M/GI/s$ models with the customary FCFS service discipline above by the associated $M/G/s$ model assigning the customers to servers in a cyclic or round robin order. In particular, we next apply the stochastic bounds in Wolff (1977, 1987) for the moments to deduce that

$$E[(W_n^a(N_n(t)))^2] \leq E[W_n^{a,c}(N_n(t)))^2],$$

where the additional superscript c on the right side denotes the cyclic service discipline.

With the cyclic service discipline, the model is equivalent to separate $GI/GI/1$ models. However, the arrival process at each server in model n has Erlang $E_{\lceil s_{bd}n \rceil}$ interarrival times, which change with n. This will not be a serious difficulty, because we can relate these arrival processes back to the original Poisson arrival process. Next, the upper bound can be further bounded above by the sum of all service times assigned to that single server up to time T, i.e.,

$$E[W_n^{a,c}(N_n(t)))^2] \le E\left[\left(\sum_{i=1}^{N_n^c(T)} S_i\right)^2\right], \quad 0 \le t \le T, \tag{22}$$

where $N_n^c(T)$ is the number of arrivals assigned to that individual server, which is a renewal process with Erlang interarrival times. Since every $\lceil ns_{bd} \rceil^{\text{th}}$ arrival in the original system is assigned to this server, $N_n^c(T) \le (N_n(T)/s_{bd}n) + 1$. Combining these results, we get

$$E[(W_n^a(N_n(t)))^2] \le E\left[\left(\sum_{i=1}^{(N_n(T)/s_{bd}n)+1} S_i\right)^2\right], \quad 0 \le t \le T, \tag{23}$$

where $N_n(t)$ is a Poisson process with rate $\lambda_{bd}n$. Hence,

$$E[N_n(T)/s_{bd}n] = \frac{\lambda_{bd}n}{s_{bd}n} = \frac{\lambda_{bd}}{s_{bd}} < \infty \quad \text{and} \quad Var\left(N_n(T)/s_{bd}n\right) = \frac{\lambda_{bd}n}{(s_{bd}n)^2} \le \frac{\lambda_{bd}}{s_{bd}^2} < \infty,$$

from which the desired uniform bound follows, using formulas for compound Poisson random variables and the condition that $E[S^2] < \infty$. ∎

7.2. Proofs of Two Lemmas in §4

We now prove the two lemmas used with Theorem 2 to prove Corollary 2.

Proof of Lemma 1. Let $m_w \equiv \int_0^w \bar{F}(x)\,dx$ and $p_w(x) \equiv \bar{F}(x)/m_w$, $0 \le x < w$. Then the function m in (16) can be expressed as an integral weighted average on the interval $[w, \infty)$ with respect to the positive pdf p_w, namely,

$$m(t) = m_w \int_0^w \lambda(t-x)p_w(x)\,dx, \quad t \ge w.$$

Now extend the pdf p and the arrival rate function λ to the entire real line by letting $p_w(x) = 0$ if $x \le 0$ or if $x \ge w$ and $\lambda(t) = 0$ for $t < 0$. Then $m(t) = c$ for all $t \ge w$ if and only if the convolution integral

$$m_{c,w}(t) \equiv m_w \int_{-\infty}^{\infty} \lambda_{c,w}(t-x)p_w(x)\,dx = 0 \quad \text{for all} \quad t \ge 0, \tag{24}$$

where $\lambda_{c,w}(t) \equiv \lambda(t+w) - c$ for all $t \geq 0$. In particular, $m_{c,w}$ is the convolution of the function $\lambda_{c,w}$ with respect to the density p, which is decreasing and positive on its interval of support, $[0, w]$; i.e., $m_{c,w} = p_w \star \lambda_{c,w}$, as on p. 143 of Feller (1971). Now let

$$\hat{p}_w(s) \equiv \int_{-\infty}^{\infty} e^{-st} p_w(t)\, dt$$

for positive real s, and similarly for the other functions. Exploiting basic properties of transforms of convolutions, we get $\hat{m}_{c,w}(s) = \hat{\lambda}_{c,w}(s)\hat{p}_w(s)$ for all positive real s. If $m_{c,w}(t) = 0$ for all t, then necessarily $\hat{m}_{c,w}(s) = 0$ for all positive real s. Since $\hat{p}_w(s) > 0$ for all positive real s, we deduce that necessarily $\hat{\lambda}_{c,w}(s) = 0 \quad \text{for all} \quad s > 0$. Since $\hat{\lambda}_{c,0}(s) = e^{sw}\hat{\lambda}_{c,w}(s)$, where $\hat{\lambda}_{c,0}(s)$ coincides with the ordinary Laplace transform of $\lambda_{c,w}(t)$, we see that $\tilde{\lambda}_{c,0}(s) = 0$ for all positive real s. By §VII.6 of Feller (1971), that implies that $\lambda_{c,w}(t) = 0$ for all t. ∎

Proof of Lemma 2. Since the fluid model starts empty, in order to have all fluid experience delay of w, no staffing at all can be provided until time w. After time w the staffing must be as in (13) in order to achieve maintain the target delay w. In turn, that fixed delay must be obtained to provide the fixed abandonment proportion $\alpha = F(w)$.

To elaborate, there must be a first time that the staffing deviates from the DIS staffing. We can trace the implications of a change in the staffing function in the fluid model, referring to results in Liu and Whitt (2012a). First, a change in the staffing function s necessarily changes the rate fluid enters service $b(\cdot, 0)$. To see that, first observe that a change in s necessarily changes $\hat{a}$ in (19) (of Liu and Whitt (2012a), like all references in this proof) to an associated $\bar{a}$. In particular, (19) implies that the function $\hat{a}$ is a monotone function of the derivative s'. If we increase s' over some interval $[0, u]$, then, $\hat{a}$ necessarily increases over $[0, u]$. From the monotonicity of the fixed point equation for the rate fluid enters service in (18), we can apply Theorem 2 to deduce that the $b(\cdot, 0)$ increases as well, and similarly for a decrease. Hence, a change in the staffing function s forces a change in the rate fluid enters service, $b(\cdot, 0)$, which is monotone in s'. That change in the function $b(\cdot, 0)$ in turn forces a change in the BWT w by virtue of the ODE in Theorem 3. However, here an increase in $b(t, 0)$ forces a decrease in $w(t)$. Thus the first change of staffing changes the BWT

$w(t)$. The change in the BWT w forces corresponding changes in the fluid density in queue $q(t,x)$ by Corollary 2, the PWT $v(t)$ by Theorem 5, and in the abandonment rate function α by (7). The change in $b(\cdot,0)$ produces changes in the fluid density in service $b(t,x)$ via (15) and the service completion rate σ via (9). Since we are concerned with stabilizing the PWT $v(t)$ at w, the change in $v(t)$ implies that the DIS staffing is the unique staffing that achieves the stabilizing goals. ∎

8. Conclusions

We have developed a systematic formula-based procedure to stabilize the abandonment probability and the expected delay in an $M_t/GI/s_t+GI$ model with a time-varying arrival rate, across a wide range of performance targets, by providing an appropriate staffing function $s(t)$. The first step in §3 involves the delayed infinite-server (DIS) model with two $M_t/GI/\infty$ IS models in series. In this model (but not in the actual system) each customer arrives at the first IS system (representing the queue) and stays there for a fixed time $w = F^{-1}(\alpha)$ (the target). Customers may abandon from this first queue. If they do not, then they proceed to the second IS system (representing the service facility). Since the number of busy servers in the service facility, $B(t)$, is a Poisson random variable for each t, it is easy to analyze. Its mean $m_\alpha(t) \equiv E[B(t)]$ as a function of the abandonment probability target α is the offered-load function used in this paper. We gave explicit formulas for $m_\alpha(t)$ in §3 and §EC.2

We found that the DIS mean $m_\alpha(t)$ itself provides an excellent staffing function for low Quality-of-Service (QoS) targets. Indeed, in §4 we proved that it achieves the stabilizing goal asymptotically as the scale increases. However, to obtain a staffing function that works for all QoS levels, we developed a new modified-offered-load (MOL) approximation in §5, obtaining our overall DIS-MOL approximation. As in previous MOL approximations, our MOL approximation exploits the associated stationary model with a constant arrival rate depending on the appropriate offered load. To treat the steady-state of the stationary $M/GI/s+GI$ model, we use an approximation developed in Whitt (2005), which is based on an associated Markovian $M/M/s+M(n)$ model with state-dependent abandonment rates.

Our simulation experiments have shown that the DIS-MOL approximation not only stabilizes expected delays and abandonment probabilities, but it also describes other performance measures, e.g., the expected queue length $E[Q(t)]$. As in Feldman et al. (2008), we find that other performance measures are stabilized to a great extent, but not fully (across a wide range of performance targets). That was illustrated in Figures 4 and 5. Indeed, in Corollary 3 we showed that it is not possible to simultaneously stabilize all performance measures asymptotically in the efficiency-driven many-server heavy-traffic regime. However, just as in Feldman et al. (2008), we find that DIS-MOL simultaneously stabilizes all standard performance measures with higher QoS targets, when the system is in the quality-and-efficiency driven regime.

In §6 we showed that a modification of the classical square-root-staffing (SRS) formula in (2) can be applied for staffing to meet abandonment-probability targets provided that we use the DIS offered-load function $m_\alpha(t)$. However, in general, it remains to determine an appropriate QoS parameter β_α. In EC.5 we developed an explicit approximation formula for the QoS parameter β_α for the Markovian $M_t/M/s_t + M$ special case; see equation (EC.1). Its special linear separable structure reveals how performance depends on the model parameters.

We have demonstrated that the DIS-MOL approximation is remarkably effective by performing simulation experiments for both Markovian and non-Markovian models with sinusoidal arrival-rate functions, when the arrival rates are not too small (around 100 with mean service $ES = 1$). In general, the performance of DIS-MOL tends to improve as the scale (arrival rate and number of servers) increases. We have also considered both smaller and larger systems, in particular, for average arrival rates ranging from $s = 20$ to $s = 1000$. The performance of DIS-MOL is spectacular for $s = 1000$ and still reasonable for $s = 20$.

While conducting the simulation experiments, we considered several discretization issues: how to convert a continuous staffing function into an integer-valued staffing function; what is the consequence of agents being required to finish their current services when called to leave; how does the size of a fixed-staffing period affect this approach (discussed in the e-companion).

Much work remains to be done in the future. For example, it remains to establish supporting theory for the DIS-MOL approximation, paralleling Massey and Whitt (1994). So far, we only can conjecture that the DIS-MOL staffing is never less than the DIS staffing. We also need asymptotic results supporting the excellent performance of the DIS-MOL approximation under a wide range of targets. We conjecture that it is asymptotically correct in the QED many-server heavy-traffic regime (in a meaningful way, e.g., that $\sqrt{n}P_t^n(Ab) \to \alpha$ as $n \to \infty$, independent of t, where α_n is the target in model n, which is required to satisfy $\sqrt{n}\alpha_n \to \alpha$ as $n \to \infty$). It also remains to stabilize performance measures in multi-class multi-pool systems and in systems with different service disciplines.

Acknowledgments.

This research was begun while the first author was a doctoral student at Columbia University. The research was supported by NSF grants DMI-0457095, CMMI-0948190 and CMMI-1066372.

References

Billingsley, 1999. *Convergence of Probability Measures*, second ed., Wiley, New York.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Davis, J. L., W. A. Massey, W. Whitt. 1995. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Sci.* **41** 1107–1116.

Eick, S. G., W. A. Massey, W. Whitt. 1993a. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

Eick, S. G., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.

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

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*, second ed., Wiley, New York.

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

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

Jagerman, D. L. 1975. Nonstationary blocking in telephone traffic. *Bell System Tech. J.* **54** 625–661.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42** 1383–1394.

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

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

Liu, Y., W. Whitt. 2012a. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* **71** 405–444.

Liu, Y., W. Whitt. 2012b. A many-server fluid limit for the $G_t/GI/s_t + GI_t$ queueing model experiencing periods of overloading. *Operations Research Letters*, doi:10.1016/j.orl.2012.05.010

Liu, Y., W. Whitt. 2012c. Many-server heavy-traffic limit for queues with time-varying parameters. Columbia University, NY. 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** 149–201.

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

Massey, W. A., W. Whitt. 1994. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Ann. Appl. Probabil.* **4** 1145–1160.

Massey, W. A., W. Whitt. 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems* **25** 157–172.

Whitt, W. 1981. Comparing counting processes processes and queues. *Adv. Appl. Prob.* **14** 207–220.

Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Sci.* **51** 221–235.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* **54** 37–54.

Wolff, R. W. 1977. An upper bound for multi-channel queues. *J. Appl. Prob.* **14** 884–888.

Wolff, R. W. 1987. An upper bound for multi-channel queues. *J. Appl. Prob.* **24** 547–551.

Zeltyn S., A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems* 51, 361–402.