Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Stabilizing Performance in Many-Server Queues with Time-Varying Arrivals and Customer Feedback

Yunan Liu
Department of Industrial Engineering, North Carolina State University, Raleigh, NC 27695, yliu48@ncsu.edu, http://www.ncsu.edu/∼yliu48

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

Analytical approximations are developed to determine the time-dependent offered load (effective demand) and appropriate staffing levels that stabilize performance at designated targets in a many-server queueing model with time-varying arrival rates, customer feedback and abandonment. To provide a flexible model that can be readily fit to system data, Markovian routing is not assumed. Instead, the model has history-dependent Bernoulli routing, with at most finitely many feedbacks, where the feedback probabilities, service-time and patience distributions all may depend on the visit number. Before returning to receive a new service, the fed-back customers experience delays in an infinite-server or finite-capacity queue, where the parameters may again depend on the visit number. A many-server heavy-traffic FWLLN shows that the performance targets are achieved asymptotically as the scale increases. A new refined modified-offered-load approximation is developed to obtain good results with low waiting-time targets. Simulation experiments confirm that the approximations are effective.

Key words: staffing algorithms for service systems; time-varying arrival rates; many-server queues; queues with feedback; retrials; stabilizing performance.

History: submitted August 11, 2013

1. Introduction

This paper is part of an ongoing effort to develop effective methods to set staffing levels (the time-dependent number of servers) in queueing systems with time-varying arrival rates in order to stabilize performance at designated targets; see Green et al. (2007) for a review and Defraeye and van Nieuwenhuyse (2013) and Stolletz (2008) for recent related work. Here we focus on many-server queues with Bernoulli feedback after completing service, so that this paper is closely related to Yom-Tov and Mandelbaum (2013), where a *modified-offered-load* (MOL) approximation was proposed

to help set staffing levels at a queue with time-varying arrival rates and Markovian feedback after a delay in an *infinite-server* (IS) queue. Yom-Tov and Mandelbaum (2013) showed that the MOL approximation has great potential for applications in healthcare, where there are longer service times; see Armony et al. (2011) for further background.

Motivated by these same applications, we consider an alternative feedback model, which we think has appealing flexibility and often may be more realistic. In particular, instead of Markovian routing with fixed feedback probability p and one fixed service-time distribution, we consider history-dependent Bernoulli routing, where the feedback probability p and the service-time distribution may vary with the visit number. We focus on the case of at most one feedback, but the methods extend directly to any finite number of feedbacks. (We also consider examples with two feedback opportunities.) We also allow customer abandonment, which tends to be more realistic for multi-server queues, as suggested by Garnett et al. (2002). The associated patience-time distributions are also allowed to depend on the visit number. These history-dependent parameters significantly complicate the analysis, producing a multi-class model, but our approach addresses it in an interesting and effective way; e.g., see the new multi-queue offered load models in Figures 1 and 9.

To analyze this new feedback model, we draw on Liu and Whitt (2012c) in which we developed a new offered-load approximation and a new algorithm to determine time-dependent staffing levels in order to stabilize expected delays and abandonment probabilities at specified QoS targets in a many-server queue with time-varying arrival rates. We considered the $M_t/GI/s_t+GI$ model, having arrivals according to a *nonhomogeneous Poisson process* (NHPP, the M_t) with arrival rate function $\lambda(t)$, independent and identically distributed (i.i.d.) service times with a general distribution (the first GI), a time-varying number of servers (the s_t, to be determined), i.i.d. patience times with a general distribution (times to abandon from queue, the final $+GI$), unlimited waiting space and the first-come first-served service discipline. We included non-exponential service and patience distributions as well as time-varying arrivals because they commonly occur; e.g. see Armony et al. (2011) and Brown et al. (2005).

As in Feldman et al. (2008), Jennings et al. (1996), Yom-Tov and Mandelbaum (2013) and references therein, the new *Delayed-Infinite-Server Modified-Offered-Load* (DIS-MOL) algorithm in Liu and Whitt (2012c) exploits IS queues and is an MOL approximation. The DIS offered load is obtained by considering two IS queues in series, the first representing the waiting room and the second representing the service facility. In this artificial construction for generating an appropriate offered load, each arrival is required to stay the targeted waiting time, say w, in the waiting room if that customer does not elect to abandon. If the patience-time cumulative distribution function (cdf) is F, each arrival abandons with probability $\alpha = F(w)$, so that the abandonment target α is

linked to the expected waiting time target w. The expected number of busy servers in the second IS queue, $m_\alpha(t)$, serves as the new offered load to be used in the new MOL approximation.

The simple DIS algorithm staffs directly according to the offered load, letting $s_\alpha(t) = \lceil m_\alpha(t) \rceil$, where $m_\alpha(t)$ is the DIS OL given in §2 and $\lceil x \rceil$ is the ceiling function, giving the least integer greater than or equal to x. The DIS algorithm itself is effective for low *Quality of Service* (QoS) targets, but it is ineffective for high QoS targets. The new MOL approximation, dubbed DIS-MOL, uses an approximation for the performance in the corresponding stationary $M/GI/s+GI$ model from Whitt (2005) to set staffing at each time t (the usual minimum number of servers such that the QoS target is met), where the arrival rate is set equal to

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

where S is a generic service time. As $\alpha \downarrow 0$, the OL $m_\alpha(t)$ converges to the conventional OL, which we refer to as $m_0(t)$, and the DIS-MOL arrival-rate function $\lambda_\alpha^{mol}(t)$ in (1) converges to the conventional MOL arrival-rate function, which we refer to as $\lambda_0^{mol}(t)$, so that this new method is closely related to previous OL and MOL approximations. If abandonment is not deemed appropriate in the model, we can use the conventional MOL approximation with a delay-probability target, using $m_0(t)$ and $\lambda_0^{mol}(t)$ in the model developed here.

Since IS models and networks of IS models are easy to analyze, as shown by Eick et al. (1993b,a) and Massey and Whitt (1993), explicit formulas exist for the OL $m_\alpha(t)$, making DIS-MOL remarkably easy to implement; see Theorem 1, (19) and (20) of Liu and Whitt (2012c). Simulation experiments confirmed that the DIS-MOL algorithm is effective in stabilizing expected delays and abandonment probabilities over a wide range of QoS targets, ranging from low QoS (heavy loads) to high QoS (light loads). For example, for about 100 servers, a high QoS abandonment probability target might be 0.01, 0.02 or 0.03, while a low QoS abandonment probability target might be 0.10, 0.20 or 0.30 (an order of magnitude larger, i.e., ten times greater). As shown by Yom-Tov and Mandelbaum (2013) and Liu and Whitt (2012c), these MOL approximations also can be useful for much smaller systems.

In this paper, we primarily focus on the special case of at most one feedback, but it will be evident that the methods extend directly to any finite number of feedbacks. We refer to this base model as $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$. The main queue has the two service-time cdf's G_i and patience cdf's F_i, while the orbit queue has a service-time cdf H. We also consider the associated $(M_t/GI, GI/s_t+GI, GI)+(GI/s_t+GI)$ model in which the orbit queue has finite capacity; in that case, it also has a staffing function and a patience distribution. The goal is to stabilize expected potential waiting times (the virtual waiting time of an arrival with infinite patience) at a fixed

value w for all time and $i = 1, 2$. Since these models are special kinds of two-class queueing models, we also consider the more elementary $\sum_{i=1}^{2}(M_t/GI + GI)/s_t$ two-class queue, in which the two classes arrive according to two independent NHPP's with arrival rate functions $\lambda^{(i)}(t)$ and their own service-time cdf's G_i and patience cdf's F_i, $i = 1, 2$, but there is a single service facility with a time-varying number of servers $s(t)$, again to be determined.

The approximating DIS model for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ feedback queue has five IS queues in series; see §2. (If there are k possible feedbacks, then the DIS model has $2 + 3k$ IS queues in series; see §6.1 for the case $k = 2$.) We show that the simple DIS algorithm (staffing directly to the DIS offered load) is effective for all three models with low QoS targets. To provide theoretical support, we prove a new functional weak law of large numbers (FWLLN) showing that any positive waiting-time target w is achieved asymptotically as the scale (arrival rate and number of servers) increases.

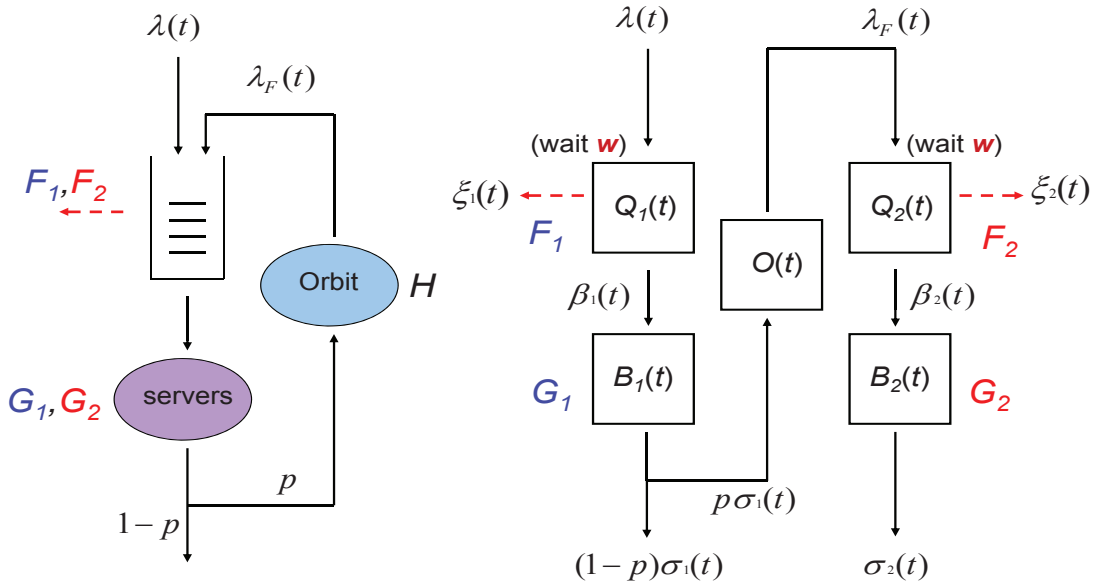
However, as before, the DIS algorithm is ineffective for low waiting-time targets. We develop a new DIS-MOL approximation for that case and conduct simulation experiments to show that it is effective. Given previous MOL approximations, ideally the MOL approximation for the main case would involve a stationary $(M/GI, GI/s + GI, GI) + (GI/\infty)$ feedback queue to apply at each time t. Since no steady-state performance results exist for such a complex stationary model, we develop an *aggregate* single-class stationary $M/GI/s + GI$ model. With this new aggregate approximating stationary model, we are able to apply the algorithm from Whitt (2005) just as in Liu and Whitt (2012c). Fortunately, simulation experiments confirm that this aggregation approach is effective.

Here is how the rest of this paper is organized: We start in §2 by giving explicit expressions for all the key performance functions of the new $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ queue with Bernoulli feedback and an IS orbit queue, with fixed delay target w. We also give explicit formulas in structured special cases when the arrival-rate function is sinusoidal. In §3 we state the supporting many-server heavy-traffic FWLLN showing that the DIS approximation asymptotically stabilizes the expected delay as the scale increases. In §4 we develop the new DIS-MOL approximation. In §5 we show the results of simulation experiments to support the approximations. In §6 we show that the good results also hold for (i) the more elementary $\sum_{i=1}^{2}(M_t/GI + GI)/s_t$ two-class queue, (ii) the more complicated $(M_t/GI, GI/s_t + GI, GI) + (GI/s_t + GI)$ queue with Bernoulli feedback and a $(GI/s_t + GI)$ finite-capacity orbit queue and (iii) the generalization of the base model allowing two feedback opportunities. In §7 we prove the FWLLN stated in §3. Finally, in §8 we draw conclusions. Additional supporting material appears in an online appendix.

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

We now develop the DIS approximation for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model, which has Bernoulli feedback with probability p for each new customer completing service; otherwise the

customer departs. Customers arrive according to an external NHPP arrival process with arrival rate function λ. The original (feedback) arrivals have i.i.d. service times and patience times distributed as generic random variables S_1 with cdf G_1 and A_1 with cdf F_1 (S_2 with cdf G_2 and A_2 with cdf F_2), respectively. Customers that are fed back encounter i.i.d. delays distributed as the generic random variable U with cdf H. The arrival-rate function of the fed-back customers is λ_F. This feedback model is depicted on the left in Figure 1.



(a) The $(M_t/GI,GI/s_t + GI,GI) + (GI/\infty)$ Model with Delayed Customer Feedbacks

(b) The Delayed Infinite-Server (DIS) Approximation

Figure 1 The $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model with delayed customer feedback and its Delayed Infinite-Server (DIS) approximation. The approximating offered load is $m(t) = m_1(t) + m_2(t) \equiv E[B_1(t)] + E[B_2(t)]$.

2.1. The Approximating Five-Queue DIS Model

The approximating DIS model, depicted on the right in Figure 1, has *five* IS queues in series, the first two for the external arrivals, in queue and in service, the third for the IS orbit queue (which is directly an IS queue) and the last two for the fed-back customers, in queue and in service. Since all arrivals to a queue are forced to remain in the waiting room a constant time w unless they abandon in this approximating model, the service times in the first and fourth IS queues (representing the waiting room) are distributed as $T_1 \equiv A_1 \wedge w$ and $T_2 \equiv A_2 \wedge w$, respectively. The service times in the second and fifth IS queues (representing the service facility) are distributed as S_1 and S_2, and the service times in the third IS queue (representing the orbit queue) are distributed as U. The performance functions for the five IS queues are then calculated recursively using Eick et al. (1993b). Theorem 1 of Eick et al. (1993b) implies that the departure process from the $M_t/GI/\infty$

IS queue is itself an NHPP with an explicitly specified rate function. It is also well know that an independent thinning of an NHPP is again an NHPP. Thus all five IS queues are $M_t/GI/\infty$ models.

In the DIS approximation for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model, we let $Q_i(t)$ and $B_i(t)$ be the number of customers in waiting room i and in service facility i at time t, $i = 1, 2$. We let $O(t)$ be the number of customers in the orbit room at time t. The approximating offered load (OL) function, which of course is a function of the waiting time target w, is

$$m(t) \equiv m_1(t) + m_2(t) \equiv E[B_1(t)] + E[B_2(t)]. \tag{2}$$

As before, all flows are Poisson processes, with rate functions as depicted in Figure 1. The abandonment rates from the two waiting rooms (IS queues 1 and 4) are $\xi_i(t)$; The rates into service from the waiting rooms (IS queues 2 and 5) are $\beta_i(t)$; the departure rate of original customers from the service facility (both fed-back and not) is $\sigma_1(t)$; the departure rates from the system of original customers and fed-back customers are $(1-p)\sigma_1(t)$ and $\sigma_2(t)$; and the feedback rate (leaving the service facility and entering the orbit IS queue) is $p\sigma_1(t)$.

2.2. The DIS Performance Functions

In this section we display the performance functions for the DIS approximation of the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model. All these performance functions are crucial in providing time-varying staffing functions and predicting system performance under these staffing policies. The next theorem generalizes Theorem 1 in Liu and Whitt (2012c) and follows directly from Eick et al. (1993b). (Also see Massey and Whitt (1993).)

For a non-negative random variable X with finite mean $E[X]$ and cdf F_X, let X_e denote a random variable with the associated *stationary-excess* cdf (or residual-lifetime cdf) F_X^e, defined by

$$F_X^e(x) \equiv P(X_e \le x) \equiv \frac{1}{E[X]} \int_0^x \bar{F}_X(y)dy, \quad x \ge 0,$$

where $\bar{F}_X(y) \equiv 1 - F_X(y)$. The moments of X_e can be easily expressed in terms of the moments of X via

$$E[X_e^k] = \frac{E[X^{k+1}]}{(k+1)E[X]}, \quad k \ge 1.$$

THEOREM 1. (*performance functions starting from the infinite past*) *Consider the DIS approximation for the* $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ *model specified in §2, starting empty in the distant past with specified delay target (parameter)* $w \ge 0$. *The total numbers of customers in the waiting*

rooms, service facilities, and in the orbit at time t, $Q_i(t)$, $B_i(t)$ *and* $O(t)$ *are independent Poisson random variables with means*

$$\begin{aligned}
E[Q_1(t)] &= E\left[\int_{t-T_1}^{t} \lambda(x)\,dx\right] = E[\lambda(t-T_{1,e})]E[T_1],\\
E[B_1(t)] &= \bar{F}_1(w)E\left[\int_{t-w-S_1}^{t-w} \lambda(x)\,dx\right] = \bar{F}_1(w)E[\lambda(t-w-S_{1,e})]E[S_1],\\
E[O(t)] &= p\,E\left[\int_{t-U}^{t} \sigma_1(x)\,dx\right] = p\,E[\sigma_1(t-U_e)]E[U],\\
E[Q_2(t)] &= E\left[\int_{t-T_2}^{t} \lambda_F(x)\,dx\right] = E[\lambda_F(t-T_{2,e})]E[T_2],\\
E[B_2(t)] &= \bar{F}_2(w)E\left[\int_{t-w-S_2}^{t-w} \lambda_F(x)\,dx\right] = \bar{F}_2(w)E[\lambda_F(t-w-S_{2,e})]E[S_2],
\end{aligned}$$

where $T_i \equiv A_i \wedge w$. *Thus,* $X(t)$, *the total number of customers in the system at time t is a Poisson random variable with a mean* $E[Q_1(t)]+E[Q_2(t)]+E[B_1(t)]+E[B_2(t)]$. *The processes counting the numbers of customers abandoning from waiting room 1 and 2 are independent Poisson processes with rate functions* $\xi_i(t)$, *where*

$$\begin{aligned}
\xi_1(t) &= \int_0^w \lambda(t-x)\,dF_1(x) = E[\lambda(t-T_1)1_{\{T_1<w\}}],\\
\xi_2(t) &= \int_0^w \lambda_F(t-x)\,dF_2(x) = E[\lambda_F(t-T_2)1_{\{T_2<w\}}].
\end{aligned}$$

The processes counting the numbers of customers entering service facility 1 and 2 are independent Poisson processes with rate functions $\beta_1(t)$ *and* $\beta_2(t)$, *where*

$$\begin{aligned}
\beta_1(t) &= \lambda(t-w)\bar{F}_1(w),\\
\beta_2(t) &= \lambda_F(t-w)\bar{F}_2(w).
\end{aligned}$$

The departure processes (counting the number of customers completing service) from service facility 1 and 2 are independent Poisson processes with rate $(1-p)\,\sigma_1(t)$ *and* $\sigma_2(t)$, *where*

$$\begin{aligned}
\sigma_1(t) &= \bar{F}_1(w)\int_0^\infty \lambda(t-w-x)\,dG_1(x) = \bar{F}_1(w)E[\lambda(t-w-S_1)],\\
\sigma_2(t) &= \bar{F}_2(w)\int_0^\infty \lambda_F(t-w-x)\,dG_2(x) = \bar{F}_2(w)E[\lambda_F(t-w-S_2)].
\end{aligned}$$

The process counting the numbers of customers entering the second waiting room is a Poisson process with rate function λ_F, *where*

$$\lambda_F(t) = p\int_0^\infty \sigma_1(t-x)\,dH(x) = (1-p)E[\sigma_1(t-U)].$$

When the arrival rate is constant, i.e., $\lambda(t) = \lambda$, the steady-state performance functions can be easily obtained using simple calculations for a five-queue IS network, which in particular simplifies to five IS queues in series; see the appendix. As discussed in Eick et al. (1993b), Massey and Whitt (1993), Liu and Whitt (2012c), simple linear and quadratic approximations derived from Taylor series for general arrival-rate functions can be convenient. These approximations show simple time lags and space shifts; see the appendix.

In applications, a typical objective is to design a staffing function for a specified planning period $[0,T]$ (e.g., $T = 24$ for a day). To treat that case, we let $\lambda(t) = 0$ for $t < 0$ into Theorem 1 and obtain the following concrete formulas for the performance measures.

COROLLARY 1. (*performance functions of the initially empty DIS model*) *Consider the initially empty DIS approximation for the* $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ *model with delay target* $w > 0$, *starting at time 0. All results in Theorem 1 hold with rate functions*

$$\begin{aligned}
\alpha_1(t) &= \int_0^{t\wedge w} \lambda(t-x)dF_1(x),\\
\beta_1(t) &= \lambda(t-w)\cdot \bar{F}_1(w)\cdot 1_{\{t\ge w\}},\\
\sigma_1(t) &= \bar{F}_1(w)\int_0^{t-w} \lambda(t-w-x)dG_1(x)\cdot 1_{\{t\ge w\}},\\
\lambda_F(t) &= \int_0^{t-w} p\sigma_1(t-y)dH(y)\cdot 1_{\{t\ge w\}}\\
&= p\bar{F}_1(w)\int_0^{t-w}\int_0^{t-w-y} \lambda(t-w-x-y)dG_1(x)dH(y)\cdot 1_{\{t\ge w\}},\\
\alpha_2(t) &= \int_0^{(t-w)\wedge w} \lambda_F(t-z)dF_2(z)\cdot 1_{\{t\ge w\}}\\
&= (1-p)\bar{F}_1(w)\int_0^{(t-w)\wedge w}\int_0^{t-w-z}\int_0^{t-w-y-z} \lambda(t-w-x-y-w)dG_1(x)dH(y)dF_2(z)\cdot 1_{\{t\ge w\}},\\
\beta_2(t) &= \lambda_{F_1}(t-w)\bar{F}_2(w)\cdot 1_{\{t\ge 2w\}}\\
&= p\bar{F}_1(w)\bar{F}_2(w)\int_0^{t-2w}\int_0^{t-2w-y} \lambda(t-2w-x-y)dG_1(x)dH(y)\cdot 1_{\{t\ge 2w\}},\\
\sigma_2(t) &= \int_0^{t-2w} \beta_2(t-z)dG_2(z)\cdot 1_{\{t\ge 2w\}}\\
&= p\bar{F}_1(w)\bar{F}_2(w)\int_0^{t-2w}\int_0^{t-2w-z}\int_0^{t-2w-y-z} \lambda(t-2w-x-y-z)dG_1(x)dH(y)dG_2(z)\cdot 1_{\{t\ge 2w\}},
\end{aligned}$$

and mean number of customers in these five IS queues

$$\begin{aligned}
E[Q_1(t)] &= \int_0^{t\wedge w} \lambda(t-x)\bar{F}_1(x)dx,\\
E[B_1(t)] &= \bar{F}_1(w)\int_0^{t-w} \lambda(t-w-x)\bar{G}_1(x)dx\cdot 1_{\{t\ge w\}},\\
E[O(t)] &= \int_0^{t-w} p\sigma_1(t-x)\bar{H}(x)dx\cdot 1_{\{t\ge w\}}
\end{aligned}$$

$$= p\bar{F}_1(w)\int_0^{t-w}\int_0^{t-w-y}\lambda(t-w-x-y)dG_1(x)\bar{H}(y)dy\cdot 1_{\{t\ge w\}},$$

$$E[Q_2(t)] = \int_0^{(t-w)\wedge w}\lambda_F(t-z)\bar{F}_2(z)dz\cdot 1_{\{t\ge w\}},$$

$$= p\bar{F}_1(w)\int_0^{(t-w)\wedge w}\int_0^{t-w-z}\int_0^{t-w-y-z}\lambda(t-w-x-y-z)dG_1(x)dH(y)\bar{F}_2(z)dz\cdot 1_{\{t\ge w\}},$$

$$E[B_2(t)] = \bar{F}_2(w)\int_0^{t-2w}\lambda_F(t-w-z)\bar{G}_2(z)dz\cdot 1_{\{t\ge 2w\}}$$

$$= p\bar{F}_1(w)\bar{F}_2(w)\int_0^{t-2w}\int_0^{t-2w-z}\int_0^{t-2w-y-z}\lambda(t-2w-x-y-z)dG_1(x)dH(y)\bar{G}_2(z)dz\cdot 1_{\{t\ge 2w\}}.$$

The total number of busy servers (or number of customers in service) at time t is $B(t)\equiv B_1(t)+B_2(t)$. As in Liu and Whitt (2012c), we let $m(t)\equiv E[B(t)]=E[B_1(t)]+E[B_2(t)]$ be the DIS OL function.

In the appendix we also consider a slightly generalized scheme. Suppose the system is not empty at the beginning of the day (at time 0) and the initial number of waiting customers in the system along with their elapsed waiting times are observed (not random). For instance, there are n customers waiting in a single line at time 0 and their elapsed waiting times are $0\le w_1\le w_2\le\cdots\le w_n$. The goal is to design an appropriate staffing function $s(t)$ for $0\le t\le T$ such that the average customer waiting times can be stabilized during $[0,T]$ (e.g., $T=8$ or $T=24$). A typical example is the Manhattan DMV office. On a regular morning, by the opening of the office (8:00 am), which may have a line of waiting customers outside the door. This variant is also analyzed in the appendix.

2.3. Sinusoidal Arrival Rate

Since many service systems have daily cycles, it is natural to consider sinusoidal and other periodic arrival rates, as was done in Jennings et al. (1996), Feldman et al. (2008), Liu and Whitt (2012c). For periodic arrival processes, we can simply focus on the dynamic steady state if we start the system at the infinite past.

THEOREM 2. *Consider the DIS approximation for the $(M_t/GI,GI/s_t+GI,GI)+(GI/\infty)$ model specified above, starting in the distant past with specified delay target $w>0$ and with sinusoidal arrival-rate function $\lambda(t)=a+b\cdot\sin(ct)$. Then $Q_1(t)$, $B_1(t)$, $O(t)$, $Q_2(t)$ and $B_2(t)$ are independent Poisson random variables having sinusoidal means*

$$E[Q_1(t)] = E[T_1](a+\gamma(T_{1,e})b\cdot\sin(ct-\theta(T_{1,e}))),$$

$$E[B_1(t)] = \bar{F}_1(w)E[S_1]\,(a+\gamma(S_{1,e})b\cdot\sin[c(t-w)-\theta(S_{1,e})])\,,$$

$$E[O(t)] = p\,\bar{F}_1(w)E[U]\,(a+\gamma(S_1)\gamma(U_e)b\cdot\sin[c(t-w)-\theta(S-1)-\theta(U_e)])\,,$$

$$E[Q_2(t)] = p\,\bar{F}_2(w)E[T_2]\,(a+\gamma(S_1)\gamma(U)\gamma(T_{2,e})b\cdot\sin[c(t-w)-\theta(S_1)-\theta(U)-\theta(T_{2,e})])\,,$$

$$E[B_2(t)] = p\,\bar{F}_1(w)\bar{F}_2(w)E[S_2]\,(a+\gamma(S_1)\gamma(U)\gamma(S_{2,e})b\cdot\sin[c(t-2w)-\theta(S_1)-\theta(U)-\theta(S_{2,e})])\,,$$

where $\theta(X) \equiv \arctan(\phi_1(X)/\phi_2(X))$, $\gamma(X) \equiv \sqrt{\phi_1(X)^2 + \phi_2(X)^2}$, $\phi_1(X) \equiv E[\sin(cX)]$, $\phi_2(X) \equiv E[\cos(cX)]$. *The abandonment rates from the two waiting rooms are sinusoidal*

$$\begin{aligned}
\xi_1(t) &= aF_1(w) + \tilde{\gamma}(A)b \cdot \sin[ct - \tilde{\theta}(A)], \\
\xi_2(t) &= a\,p\,F_2(w)\bar{F}_1(w) + p\bar{F}_1(w)\gamma(S_1)\gamma(U)\tilde{\gamma}(A)b \cdot \sin[c(t-w) - \theta(S_2) - \theta(U) - \tilde{\theta}(A)],
\end{aligned}$$

where $\tilde{\theta}(X) \equiv \tilde{\phi}_1(X)/\tilde{\phi}_2(X)$, $\tilde{\gamma}(X) \equiv \sqrt{\tilde{\phi}_1(X)^2 + \tilde{\phi}_2(X)^2}$, $\tilde{\phi}_1(X) \equiv E[\sin(cX)1_{\{X<w\}}]$, $\tilde{\phi}_2(X) \equiv E[\cos(cX)1_{\{X<w\}}]$. *The rates of entering the two service facilities are sinusoidal*

$$\begin{aligned}
\beta_1(t) &= \lambda(t-w)\bar{F}_1(w), \\
\beta_2(t) &= p\bar{F}_1(w)\bar{F}_2(w)\left(a + \gamma(S_2)\gamma(U)b \cdot \sin[c(t-2w) - \theta(S_2) - \theta(U)]\right),
\end{aligned}$$

The departure rates from the two service facilities are sinusoidal

$$\begin{aligned}
\sigma_1(t) &= \bar{F}_1(w)\left(a + \gamma(S_1)b \cdot \sin[c(t-w) - \theta(S_1)]\right), \\
\sigma_2(t) &= p\bar{F}_1(w)\bar{F}_2(w)\left(a + \gamma(S_2)^2\gamma(U)b \cdot \sin[c(t-2w) - 2\theta(S_2) - \theta(U)]\right).
\end{aligned}$$

The arrival rate to the second waiting room is sinusoidal

$$\lambda_F(t) = p\bar{F}_1(w)\left(a + \gamma(S_1)\gamma(U)b \cdot \sin[c(t-w) - \theta(S_1) - \theta(U)]\right).$$

Remark 1. (extreme values of the sinusoidal performance functions) Note the extreme values of $E[Q_1(t)]$, $E[B_1(t)]$, $E[O(t)]$, $E[Q_2(t)]$ and $E[B_2(t)]$ occur at

$$\begin{aligned}
t_{Q_1} &= t_\lambda + \theta(T_{1,e})/c, \\
t_{B_1} &= t_\lambda + w + \theta(S_{1,e})/c, \\
t_O &= t_\lambda + w + (\theta(S_1) + \theta(U_e))/c, \\
t_{Q_2} &= t_\lambda + w + (\theta(S_1) + \theta(U) + \theta(T_{2,e}))/c, \\
t_{B_2} &= t_\lambda + 2w + (\theta(S_1) + \theta(U) + \theta(S_{2,e}))/c,
\end{aligned}$$

respectively, where $t_\lambda = \pi/2c + n\pi/c$ for n integer are times at which the extreme values of $\lambda(t)$ occurs. Their extreme values are

$$\begin{aligned}
E[Q_1(t_{Q_1})] &= E[T_1](a + \gamma(T_{1,e})b), \\
E[B_1(t_{B_1})] &= \bar{F}_1(w)E[S_1]\left(a + \gamma(S_{1,e})b\right), \\
E[O(t_O)] &= p\,\bar{F}_1(w)E[U]\left(a + \gamma(S_1)\gamma(U_e)b\right), \\
E[Q_2(t_{Q_2})] &= p\,\bar{F}_1(w)E[T_2]\left(a + \gamma(S_1)\gamma(U)\gamma(T_{2,e})b\right), \\
E[B_2(t_{B_2})] &= p\,\bar{F}(w)^2E[S]\left(a + \gamma(S_1)\gamma(U)\gamma(S_{2,e})b\right),
\end{aligned}$$

respectively.

It is interesting to investigate how the new feature of delayed feedback influence the variation of the OL function. In particular, we want to see if the relative amplitude of the new OL function is flattened or exaggerated compared to the old one. However, the general scheme is complicated because the OL function strongly depends not only on the basic model parameters F_i, G_i, H and λ, it also depends on the target service level w. For the rest of this section, we assume that $F_1 = F_2 = F$ and $G_1 = G_2 = G$. Under that condition, we consider two special cases: (i) exponential service (S) and orbit (U) times and (ii) deterministic service and orbit times. Let $RA(m)$ and $RA(m^*)$ be the relative amplitude (relative variation around the average) of the new and old OL functions, respectively. We also want to investigate the time lag incurred by the feedback structure. Let the phase difference of the two OL functions be $\Delta PH(m, m^*) \equiv Phase(m^*) - Phase(m)$. The following result is proved in the appendix.

THEOREM 3. *Consider the DIS approximation for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model specified above with $F_1 = F_2 = F$ and $G_1 = G_2 = G$. Let the system start empty in the distant past with specified delay target $w > 0$ and with sinusoidal arrival-rate function $\lambda(t) = a + b \cdot \sin(ct)$. Then the OL function $m(t) \equiv E[B_1(t)] + E[B_2(t)]$ is sinusoidal*

$$m(t) = \bar{F}(w)E[S]\left(a(1 + p\,\bar{F}(w)) + b\,\gamma(S_e)\sqrt{u^2 + v^2}\,\sin[c(t-w) - \bar{\theta}]\right), \tag{3}$$

where $\bar{\theta} \equiv \arctan(u/v)$, $u \equiv \sin[\theta(S_e)] + p\,\bar{F}(w)\gamma(S)\gamma(U)\sin(\tilde{\theta})$, $v \equiv \cos[\theta(S_e)] + p\,\bar{F}(w)\gamma(S)\gamma(U)\cos(\tilde{\theta})$, $\tilde{\theta} \equiv cw + \theta(S) + \theta(U) + \theta(S_e)$, $\theta(X) \equiv \phi_1(X)/\phi_2(X)$, $\gamma(X) \equiv \sqrt{\phi_1(X)^2 + \phi_2(X)^2}$, $\phi_1(X) \equiv E[\sin(cX)]$, $\phi_2(X) \equiv E[\cos(cX)]$.

(i) If both service (S) and orbit (U) times are exponential, then

$$RA(m) < RA(m^*) \quad \textit{if} \quad \left(1 + \frac{c^2}{\mu^2}\right)\left(1 + \frac{c^2}{\delta^2}\right) > 1.$$

(ii) If both service and orbit times are deterministic, then $RA(m) \le RA(m^*)$.

Furthermore, in both cases

$$\lim_{c\to 0} \frac{RA(m)}{RA(m^*)} = 1,$$
$$\lim_{c\to 0} \Delta PH(m, m^*) = 0.$$

3. Asymptotic Effectiveness as the Scale Increases

In this section we state the many-server heavy-traffic FWLLN for the $(G_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model with Bernoulli feedback after a random delay in an IS orbit queue, implying that the DIS staffing algorithm is effective in stabilizing the expected waiting times for all customers at a fixed positive value w asymptotically as the scale increases. The associated abandonment

probability targets $\alpha_i = F_i(w)$ for $i = 1, 2$, where $i = 1$ corresponds to external arrivals and $i = 2$ corresponds to feedback after completing service, are then achieved asymptotically as well.

Paralleling Liu and Whitt (2012b,c), the FWLLN involves a sequence of $(G_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ models indexed by n. As before, we let the service and patience distributions G_i, F_i, H be independent of n. The cdf's G_i, F_i and H are differentiable, with positive finite probability density functions (pdf's) g_i, f_i and h.

In Liu and Whitt (2012c) we assumed that the arrival process $N_n(t)$ was NHPP, but greater generality is allowed by Liu and Whitt (2012b,a). In order to simplify the proof, we make the DIS staffing simply be proportional to the scale parameter n. We achieve that by letting the arrival rate in model n be a scaled version of a fixed arrival rate function. As in Liu and Whitt (2012c), that works directly if we assume that the external arrival process is an NHPP, but to allow greater generality we assume a specific process representation.

We now assume that the queue has a base external arrival counting process that can be expressed as

$$N^{(e)}(t) = N^{(b)}(\Lambda(t)), \quad t \geq 0, \tag{4}$$

where $\Lambda(t)$ is a differentiable cumulative rate function with

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds \tag{5}$$

where $\lambda(t)$ is specified as part of the model data. and $N^{(b)} \equiv \{N^{(b)}(t) : t \geq 0\}$ is a rate-1 stationary point process satisfying a FWLLN, i.e.,

$$n^{-1} N^{(b)}(nt) \Rightarrow t \quad \text{in} \quad D \quad \text{as} \quad n \to \infty, \tag{6}$$

where $\Rightarrow$ denotes convergence in distribution in the function space D with the topology of uniform convergence over bounded subintervals of the domain $[0, \infty)$ as in Whitt (2002).

In that framework, we then define the external arrival process in model n by letting

$$N_n^{(e)}(t) \equiv N^{(b)}(n\Lambda(t)), \quad t \geq 0, \tag{7}$$

which gives it cumulative arrival rate function $\Lambda_n(t) = n\Lambda(t)$, a simple multiple of the base arrival rate function. On account of this construction and assumption (6), we deduce that $N_n^{(e)}$ also obeys the FWLLN

$$\bar{N}_n^{(e)}(t) \equiv n^{-1} N^{(e)}(nt) \Rightarrow \Lambda_1(t) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty, \tag{8}$$

where the limit is the cumulative external arrival rate function of the fluid model.

Since the external arrival rate has been constructed by simple scaling, the associated DIS staffing can be constructed by simple scaling as well; see §4 of Liu and Whitt (2012a). Hence, in model n,

we can use a time-varying number of servers $s_n(t) \equiv \lceil n\,s(t) \rceil$ (the least integer above $ns(t)$), which we assume is set by the DIS staffing algorithm, which is a scaled version of the staffing in the associated fluid model with cumulative arrival rate Λ, already specified in Theorem 1, in particular,

$$s(t) = m(t) = m_1(t) = m_2(t) = E[B_1(t)] + E[B_2(t)]. \tag{9}$$

We define the following performance functions for the n^{th} model: Let $N_n(t)$ be the total number of (external plus internal) arrivals in the interval $[0,t]$; let $Q_n^{(i)}(t)$ be the number of customers of type i waiting in queue at time t; let $W_n^{(i)}(t)$ be the corresponding potential waiting time, i.e., the virtual waiting time at time t if there were an arrival at time t of type i, assuming that arrival had unlimited patience; let $A_n^{(i)}(t)$ be the number of type i customers that have abandoned from queue in the interval $[0,t]$; let $A_n^{(i)}(t,u)$ be the number of type-i customers among arrivals to the queue in $[0,t]$ that have abandoned in the interval $[0,t+u]$; let $D_n^{(i)}(t)$ be the number of type-i customers to complete service in the interval $[0,t]$; let $D_n^{(1,2)}(t)$ be the number of type-1 customers to complete service that have been fed back in the interval $[0,t]$; let $D_n^{(2)}(t)$ be the number of type-2 customers to arrive back at the queue in the interval $[0,t]$. Define associated FWLLN-scaled processes: by letting $\bar{N}_n(t) \equiv n^{-1}N_n(t)$, and similarly for the other processes except the process $W_n^{(i)}(t)$ is not scaled. Let 1_C be the indicator variable, which is equal to 1 if event C occurs and is equal to 0 otherwise.

THEOREM 4. (*asymptotic effectiveness*) *Consider a sequence of* $(G_t/GI, GI/s_t+GI, GI)+(GI/\infty)$ *models indexed by* n *with the external arrival processes in* (7) *and the many-server heavy-traffic scaling specified above. Suppose that these systems start empty at time* 0*, the regularity conditions in* Liu and Whitt (2012b,a) *are satisfied* (*including the finite positive densities*) *and* $E[S_i^2] < \infty$ *for all* i*. Then, with any expected waiting time target* $w > 0$ *and associated abandonment-probability targets* $\alpha_i = F_i(w) > 0$*,* $i = 1,2$*, use the DIS staffing* $s_n(t) \equiv \lceil n\,s(t) \rceil$*, where*

$$s(t) = m(t) = m_1(t) + m_2(t) = E[B_1(t)] + E[B_2(t)], \tag{10}$$

as given in Theorem 1*. Then the expected delays and abandonment probabilities are stabilized at their targets* w *and* α_i *for* $i = 1,2$ *asymptotically as* $n \to \infty$*. Moreover, for any time* b *with* $w < b < \infty$*,*

$$\begin{aligned} &\sup_{0\le t\le b}\{|\bar{Q}_n^{(i)}(t) - E[Q^{(i)}(t)]|\} \;\Rightarrow\; 0, \quad \sup_{0\le t\le b}\{|W_n^{(i)}(t) - w|\} \Rightarrow 0, \quad E[W_n^{(i)}(t)] \to w, \quad t \ge 0, \\ &\sup_{0\le t\le b}\{|\bar{A}_n^{(i)}(t) - A^{(i)}(t)|\} \Rightarrow 0 \quad \textit{and} \quad \sup_{0\le t\le b_i,\, w_i<u<b_i}\{|\bar{A}_n^{(i)}(t,t+u) - A^{(i)}(t,u)|\} \Rightarrow 0 \end{aligned} \tag{11}$$

as $n \to \infty$*, where (with* $\lambda_1 = \lambda$ *and* $\lambda_2 = \lambda_F$*)*

$$\begin{aligned} E[Q^{(i)}(t)] = E[Q^{(i)}(t,0)] &\equiv \int_0^{w_i} \lambda_i(t-x)\bar{F}_i(x)\,dx, \quad A^{(i)}(t) \equiv \int_0^t \xi_i(s)\,ds \\ \xi_i(t) &\equiv \int_0^{w_i} \lambda_i(t-x) f_i(x)\,dx \quad \textit{and} \quad A^{(i)}(t,u) \equiv \Lambda_i(t)\alpha_i, \quad u > w_i. \end{aligned} \tag{12}$$

We give the proof in §7. Essentially the same argument yields corresponding FWLLN's for the $\sum_{i=1}^{2}(M_t/GI+GI)/s_t$ two-class queue and the $(M_t/GI, GI/s_t+GI, GI)+(GI/s_t+GI)$ model when the orbit queue has finite capacity.

4. The Refined DIS-MOL Approximation

Paralleling the DIS-MOL approximation in Liu and Whitt (2012c), we let the DIS-MOL staffing be the time-varying number of servers needed in the stationary $M/GI/s+GI$ model with time-varying total arrival rate $\lambda_{mol}(t)$, regarded as constant at each time t, depending on the offered loads $m_i(t)$, and associated parameters according to

$$\lambda_{MOL}(t) \equiv \sum_{i=1}^{2} \lambda_{mol,i}(t), \tag{13}$$

where

$$\lambda_{mol,i}(t) \equiv \frac{m_i(t)}{(1-\alpha_i)E[S_i]} \tag{14}$$

where $m_i(t) = E[B_i(t)]$ for each i. We enforce the additivity in (13) and the additivity $m(t) = m_1(t)+m_2(t)$.

We now elaborate on our reasoning. As in Liu and Whitt (2012c), the idea behind (1) and (14) is that we want to exploit the basic offered load relation for the stationary model, which corresponds to Little's law applied to the service facility, i.e., $m = \lambda E[S]$. However, the arrival rate should be adjusted for abandonment. Hence, if λ is the external arrival rate, not adjusted for abandonment, then $m = \lambda(1-\alpha)E[S]$ and $\lambda = m/(1-\alpha)E[S]$. However, now we have two classes of customers with different parameters, so we have $m_i = \lambda_i(1-\alpha_i)E[S_i]$ for each i, which leads to $\lambda_i = m_i/(1-\alpha_i)E[S_i]$ for each i. The total arrival rate is the sum of these two arrival rates. When we substitute $m_i(t)$ for m_i, we obtain our DIS-MOL arrival rates (14) to use in the stationary $M/GI/s+GI$ model.

The MOL arrival rate in (13) generalizes the relatively simple formula $\lambda_{MOL}(t) = m_\alpha(t)/(1-\alpha)E[S]$ for a single queue in Liu and Whitt (2012c). Formula (13) reduces to that when $F_i = F$ for all i, so that $\alpha_i = \alpha$, and $G_i = G$ for all i, so that $E[S_i] = E[S]$ for all i. Given the MOL arrival rate function in (13), we apply the approximations for the performance in the stationary $M/GI/s+GI$ model from Whitt (2005), just as in Liu and Whitt (2012c), except we use w as the target for the expected waiting time.

4.1. Constructing the Aggregate Stationary Model

We have just constructed the aggregate DIS-MOL arrival rate in (13). In order to produce a stationary $M/GI/s+GI$ model for each time t, it now remains to define appropriate aggregate

service-time and patience cdf's G_{mol} and F_{mol} to be used in the stationary model at time t. We let these be defined as appropriate averages. In particular, we let

$$F_{mol}(t) = \frac{\lambda_{mol,1}(t)F_1 + \lambda_{mol,2}(t)F_2}{\lambda_{mol}(t)} \tag{15}$$

so that

$$1 - \alpha_{mol}(t) = \frac{\lambda_{mol,1}(t)(1-\alpha_1) + \lambda_{mol,2}(t)(1-\alpha_2)}{\lambda_{mol}(t)} \tag{16}$$

and

$$G_{mol}(t) = \frac{(1-\alpha_1)\lambda_{mol,1}(t)G_1 + (1-\alpha_2)\lambda_{mol,2}(t)G_2}{(1-\alpha_{mol}(t))\lambda_{mol}(t)}. \tag{17}$$

Let $S_{mol}(t)$ and $A_{mol}(t)$ be generic random variables with the cdf's G_{mol} and F_{mol} at time t. From (17), we have

$$E[S_{mol}(t)] = \frac{(1-\alpha_1)\lambda_{mol,1}(t)E[S_1] + (1-\alpha_2)\lambda_{mol,2}(t)E[S_2]}{(1-\alpha_{mol}(t))\lambda_{mol}(t)} \tag{18}$$

Since these definitions are averages, we meet the obvious consistency condition that $G_{mol}(t) = G$ if $G_1 = G_2 = G$ and $F_{mol}(t) = F$ if $F_1 = F_2 = F$.

PROPOSITION 1. (*additivity*) *With these definitions, we maintain the important MOL additivity assuming that*

$$m_{mol}(t) \equiv (1-\alpha_{mol}(t))\lambda_{mol}(t)E[S_{mol}(t). \tag{19}$$

Then

$$\begin{aligned} m_{mol}(t) &\equiv (1-\alpha_{mol}(t))\lambda_{mol}(t)E[S_{mol}(t) \\ &= (1-\alpha_1)\lambda_{mol,1}(t)E[S_1] + (1-\alpha_2)\lambda_{mol,2}(t)E[S_2] = m_1(t) + m_2(t) = m(t), \end{aligned} \tag{20}$$

as it should.

Proof. We start with (19) and then apply the definition of $E[S_{mol}(t)]$ in (18) to get the second line. We then apply (14). ∎

4.2. Computing the DIS-MOL Staffing Function

For each time t, we apply the constant arrival rate in (13), abandonment cdf in (15) and service-time cdf in (17) in order to obtain a stationary $M/GI/s+GI$ model, which of course depends on t. We numerically select the staffing level $s_{mol}(t)$ to be the smallest value for which the expected steady-state potential waiting time (virtual waiting time for a customer, if that customer had unlimited patience) is less than the target w.

To do so, we exploit the approximating state-dependent Markovian $M/M/s+M(n)$ model for the stationary $M/GI/s+GI$ queue, developed in Whitt (2005). With that model, we first compute

the steady-state distribution $\pi_i \equiv P(Q(\infty) = i)$, $i \geq 0$, for the $M/M/s + M(n)$ queue, as indicated in §7 of Whitt (2005). We next compute the expected steady-state potential waiting time by conditioning on the total number of customers in the queue. As a function of the number of servers s, we write

$$E[W_s(\infty)] = \sum_{i=s}^{\infty} E[W_s(\infty)|Q(\infty) = i] \cdot \pi_i = \sum_{i=s}^{\infty} \sum_{k=0}^{s-i} \frac{1}{s\mu + \delta_k} \cdot \pi_i, \tag{21}$$

where μ is the reciprocal of the mean service time in (18) and δ_k is the state-dependent abandonment rate in (3.4) of Whitt (2005). The goal here is to find an $s_{mol}(t)$ such that $s_{mol}(t) = \min\{s > 0, E[W_s(\infty)] < w\}$ for each stationary $(M/GI/s + GI)_t$ model.

In closing this section, we also remark that we could also be staffing at time t to satisfy the new abandonment target $\alpha_{mol}(t)$ given in (16), i.e., we could choose the minimum number of servers so that the steady-state probability of abandonment is below $\alpha_{mol}(t)$. This is so because if the potential waiting time is indeed w for an arrival, then the probability that this arrival will abandon is approximately $F_{mol}(t, w) = \alpha_{mol}(t)$.

5. Comparison with Simulations

We now describe results of simulation experiments to show the effectiveness of the approximations.

5.1. The Base Model

Our base model is the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model with Bernoulli feedback after a random delay in an IS orbit queue. (We consider other models in §6 and the appendix.) Just as in Feldman et al. (2008), Liu and Whitt (2012c), for our base case we let the system start empty and we use a sinusoidal arrival rate function with average offered load for new arrivals of approximately 100, so that the staffing would fluctuate around 100 for the external arrivals alone. (We also consider cases with smaller arrival rates in the appendix.) In particular, we use the arrival rate function

$$\lambda(t) = \bar{\lambda}(1 + r\sin(t)) = 100(1 + r\sin(t)), \quad t \geq 0, \tag{22}$$

for relative amplitudes r, denoted by $M_t(r)$; here we let $r = 0.2$. We let the feedback probability be $p = 0.2$, but we let the mean service times for the original and fed-back customers be $\mu_1^{-1} \equiv E[S_1] = 1$ and $\mu_2^{-1} \equiv E[S_2] = 5$, respectively, so that the offered loads of the two kinds of customers are roughly equal. In the appendix we obtain similar results for the corresponding model with $p = 0.5$ and $\mu_2^{-1} \equiv E[S_2] = 2$, which has more similar mean service times.

We let the three service-time distributions be hyperexponential (H_2) with *squared coefficient of variation* (scv, variance divided by the square of the mean) $c^2 = 4$, with balanced means, as on p.

137 of Whitt (1982); we thus write $H_2(m,4)$ with specified mean m. We let the patience times of the original and fed-back customers be exponential, but with different means, denoted by $M(m)$. In particular, we consider the $(M_t(r)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (p, H_2(1,4)/\infty)$ model with $r = p = 0.2$. All service-time distributions are H_2, while all patience distributions are M, but the means vary, so that the complex refined DIS-MOL formulas in §4 associated with the aggregate model are needed, and are tested in these experiments. We also consider corresponding models with non-exponential patience cdf's in the appendix.

5.2. Results from the Simulation Experiment

We simulated the model above starting empty over the time interval $[0,20]$. We estimated the performance functions by taking averages from 2000 independent replications. (Additional details are given in the appendix.)

Figures 2 and 3 show the results of the simulation experiment for high and low waiting-time targets, respectively, In Figure 2 the waiting-time targets are $w = 0.10, 0.20, 0.30, 0.40$, so that the simple DIS staffing is used, while in Figure 3 the waiting-time targets are $w = 0.01, 0.02, 0.03, 0.04$, ten times smaller, so that the refined DIS-MOL staffing is used. The performance functions are averages based on 2000 independent replications.

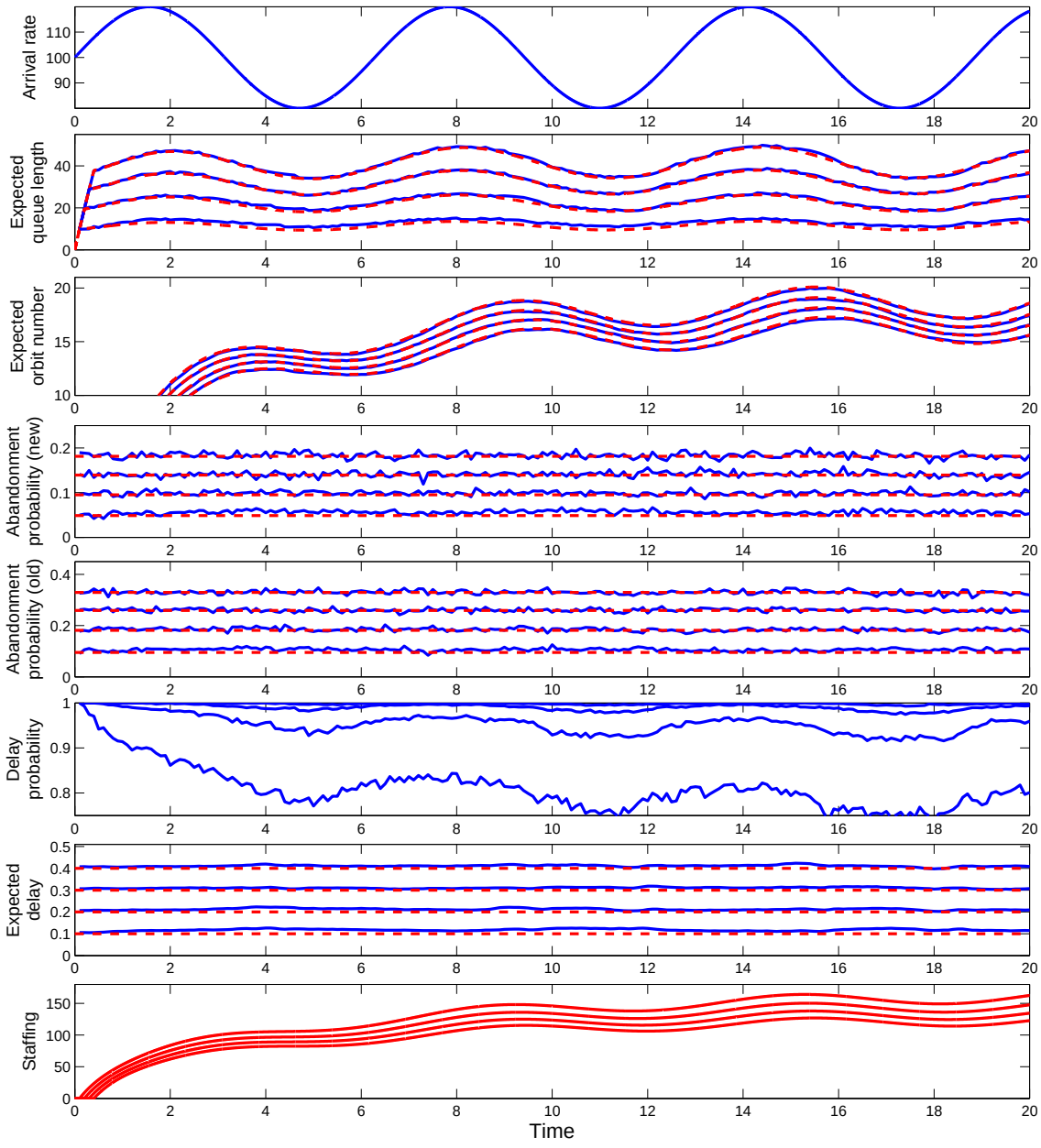
Consistent with Liu and Whitt (2012c) and the FWLLN in §3, with the higher waiting-time targets in Figure 2 we see very smooth and accurate plots of the expected waiting times and abandonment probabilities, which are the performance functions to be stabilized, but strongly fluctuating expected queue lengths and delay probabilities, which agree closely with the formulas in §2. With the higher waiting-time targets, there is higher abandonment probability, so that the maximum staffing is about 160 instead of about $100 + 100 = 200$ in Figure 3 with the lower waiting-time targets. There is greater variability with the lower waiting-time targets.

Figure 3 shows that, consistent with experience in Feldman et al. (2008) and Liu and Whitt (2012c), all performance functions tend to be stabilized simultaneously with the lower waiting-time targets, after an initial startup effect due to starting empty. The delay probability starts at 1 because the stabilizing staffing algorithm does not start staffing until time $w > 0$. That feature ensures that all arrivals wait exactly w in the limiting fluid model (see §10 of Liu and Whitt (2012a)), but it would probably not be used in applications.

5.3. Square Root Staffing

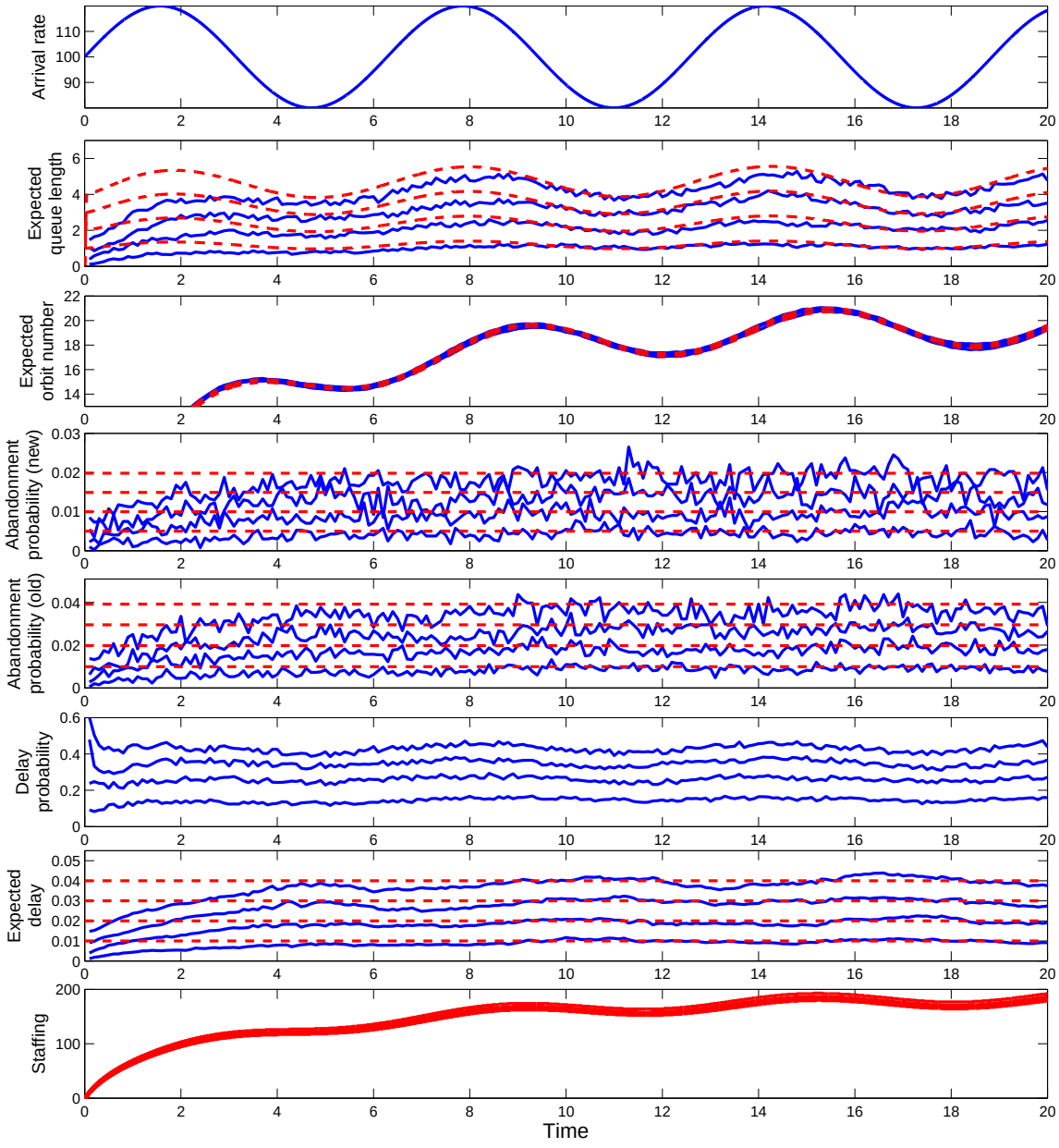
We emphasize that the DIS OL $m(t)$ given explicitly in §2 is the key quantity being computed. The DIS OL quantifies the essential demand, combining the impact of the random service times with the time-varying arrival rate, both of which are complicated by the feedback. The relatively

Figure 2 Performance functions in the $(M_t(0.2)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (22) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (low QoS) targets ($w = 0.10$, 0.20, 0.30 and 0.40) and simple DIS staffing.



complicated DIS-MOL staffing, which requires an algorithm for computing an approximation for the steady-state performance in the stationary $M/GI/s + GI$ model, is of course also important in identifying the exact staffing level required to stabilize the expected potential waiting times at the target w. However, except for the specific QoS parameter β, the same goal could be achieved

Figure 3 Performance functions in the $(M_t(0.2)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (22) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing.



by applying the simple *square root staffing* (SRS) formula

$$s(t) \equiv m(t) + \beta\sqrt{m(t)}, \tag{23}$$

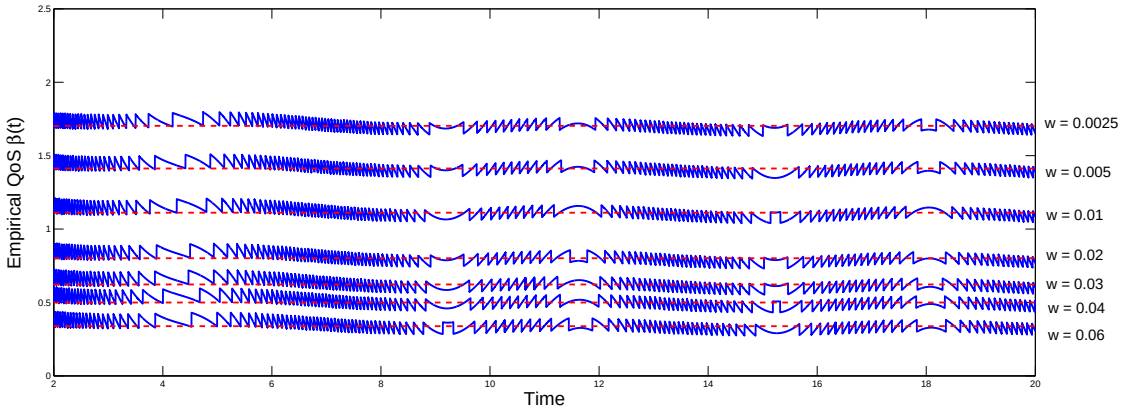
with this DIS OL $m(t)$. Without the DIS-MOL step, we could just search for the appropriate constant β to use in the SRS formula. The DIS OL already succeeds in eliminating the dependence on time.

As in Feldman et al. (2008), we demonstrate the importance of the DIS OL in the present context by plotting the implied empirical QoS,

$$\beta_{DIS-MOL}(t) = \frac{s_{DISMOL}(t) - m(t)}{\sqrt{m(t)}} \tag{24}$$

for the example considered in 3. Figure 4 shows that the DIS-MOL staffing is indeed equivalent to SRS staffing for an appropriate QoS parameter β, which is given on the y axis on the left, as a function of the target w on the right. We present similar empirical QoS plots for other examples in the appendix.

Figure 4 The empirical Quality of Service (QoS) provided by the DIS-MOL staffing in the $(M_t(0.2)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (0.2, H_2(1,4)/\infty)$ example of Figure 3 as a function of the waiting-time target w.



The DIS OL is appropriate for smaller models as well, but then the actual staffing and the resulting performance are complicated because the discretization becomes very important. However, the DIS OL remains an important first step to identify the effective time-dependent demand.

6. Other Models

In this section we discuss the other two models mentioned in the introduction. We first discuss the $\sum_{i=1}^{2}(M_t/GI+GI)/s_t$ two-class queue, in which the two classes arrive according to two independent NHPP's. We then discuss the $(M_t/GI, GI/s_t+GI, GI)+(GI/s_t+GI)$ feedback model in which the orbit queue has finite capacity. Afterwards, we discuss the model with two feedback opportunities. More examples are discussed in the appendix.

6.1. Two-Class Queue

In this section we consider the associated $\sum_{i=1}^{2}(M_t/GI+GI)/s_t$ two-class queue, in particular, the $\sum_{i=1}^{2}(M_t/H_2(m_i,4)+M(m_i)/s_t$ model with $H_2(m,4)$ service-time cdf's for both classes with $m_1=1.0$ and $m_2=0.6$ and $M(m)$ patience cdf's for both classes with $m_1=2.0$ and $m_2=1.0$. We let the arrival processes be independent NHPP's, but with different sinusoidal arrival-rate functions, in particular,

$$\lambda_1(t)=100(1+0.2\sin(t)), \quad \text{and} \quad \lambda_2(t)=60(1+0.2\sin(0.8t+2)). \tag{25}$$

The analysis of this model is more elementary. First, there is no orbit queue. We get the DIS OL by simply applying the DIS approximation to the two classes separately. That yields the per-class OL's $m_i(t)=E[B_i(t)]$ for $i=1,2$ and then we add to get the total OL: $m(t)=m_1(t)+m_2(t)$. Given this overall DIS OL, we apply the same refined DIS-MOL approximation in §4. The results of simulation experiments for high and low waiting-time targets,based on 2000 independent replications, are shown in Figures 5 and 6. The results are good, just as in §5.

Figure 5 Performance functions in the $\sum_{i=1}^{2}(M_t/H_2(m_i,4)+M(m_i)/s_t$ two-class model with the two sinusoidal arrival-rate functions in (25), service-time means $m_1=1.0$ and $m_2=0.6$ and patience means $m_1=2.0$ and $m_2=1.0$: four cases of identical high waiting-time (low QoS) targets ($w=0.10$, 0.20, 0.30 and 0.40) and simple DIS staffing at both queues.

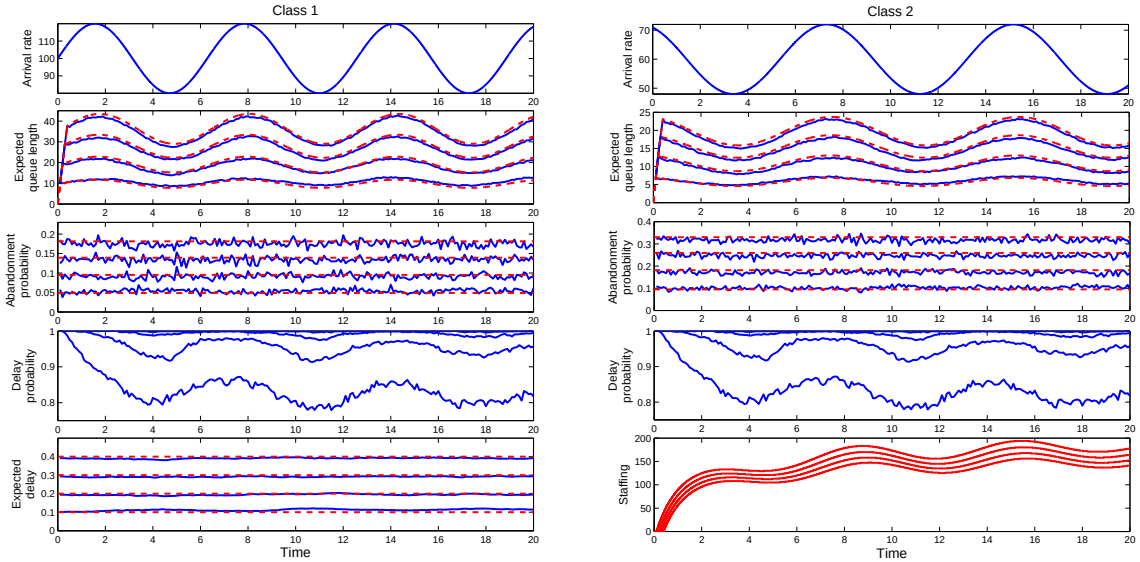
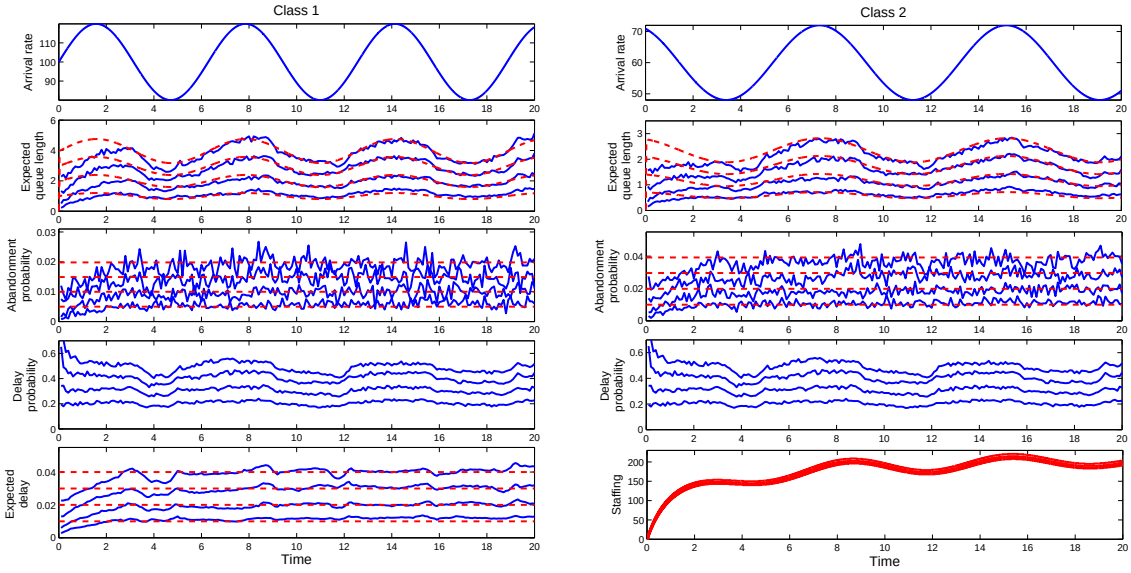


Figure 6 Performance functions in the $\sum_{i=1}^{2}(M_t/H_2(m_i,4)+M(m_i)/s_t$ two-class model with the two sinusoidal arrival-rate functions in (25), service-time means $m_1=1.0$ and $m_2=0.6$ and patience means $m_1=2.0$ and $m_2=1.0$: four cases of identical low waiting-time (high QoS) targets ($w=0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing at both queues.



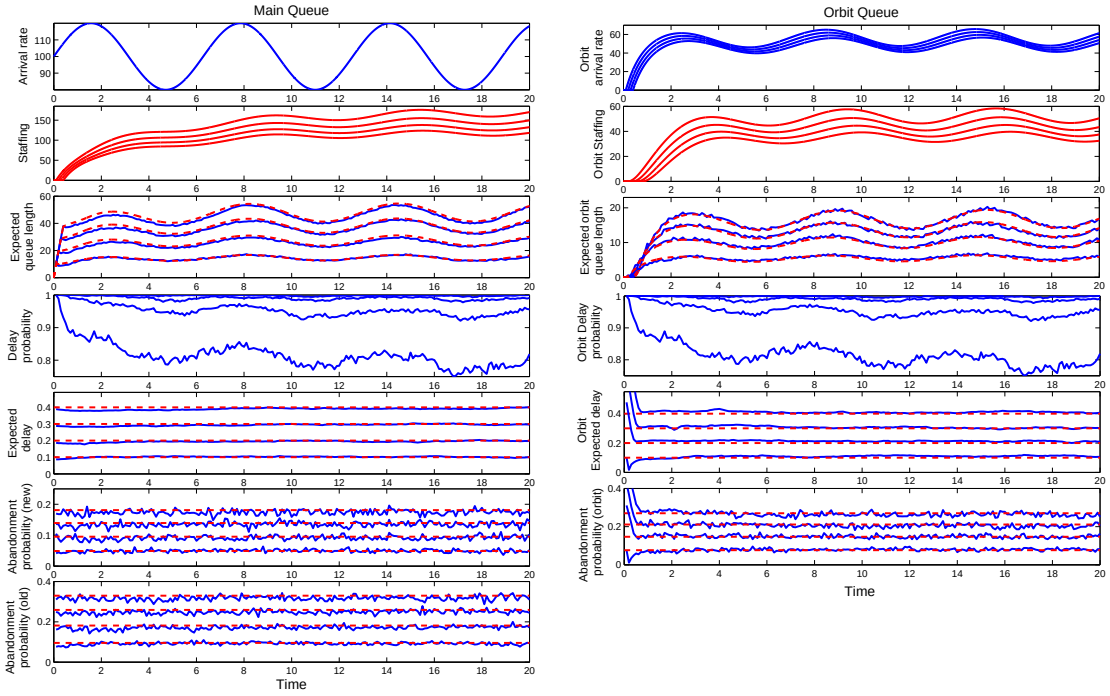
6.2. A Finite-Capacity Orbit Queue

In this section we consider the associated $(M_t/GI,GI/s_t+GI,GI)+(GI/s_t+GI)$ model with Bernoulli feedback after a random delay in a *finite-capacity* orbit queue. We use the same waiting-time targets to set the staffing levels in the orbit queue and the main queue. In particular, we consider the $(M_t(r)/H_2(1,4),H_2(10/6,4)/s_t+M(2),M(1))+(p,H_2(1,4)/s_t+M(1))$ model with $r=0.2$ and $p=0.6$. Just as in §5, all service-time distributions are H_2, while all patience distributions are M, but the means vary, so that the complex refined DIS-MOL formulas in §4 associated with the aggregate model are needed. Figures 7 and 8 show the results of the simulation experiment for high and low waiting-time targets, respectively, again based on 2000 independent replications, each starting empty.

6.3. Two Feedback Opportunities

In this section we consider a modification of the base model in which there are two feedback opportunities. Each customer that has been fed back once returns again with probability p_2 after another delay in an IS orbit queue with cdf H_2. Upon return, these customers have service-time cdf G_3 and patience cdf F_3. The new DIS model has *eight* IS queues in series, as depicted in Figure 9.

Figure 7 Performance functions in the $(M_t(0.2)/H_2(1,4), H_2(10/6,4)/s_t + M(2), M(1)) + (0.6, H_2(1,4)/s_t + M(1))$ model with the sinusoidal arrival rate in (22) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.6$ and a finite-capacity orbit queue: four cases of identical high waiting-time (low QoS) targets ($w = 0.10$, 0.20, 0.30 and 0.40) and simple DIS staffing at both queues.



Since there are now three customer classes, characterized by their class-dependent service-time and patience-time distributions, we easily generalize results in Theorem 1 to include the formulas for class 3. We have

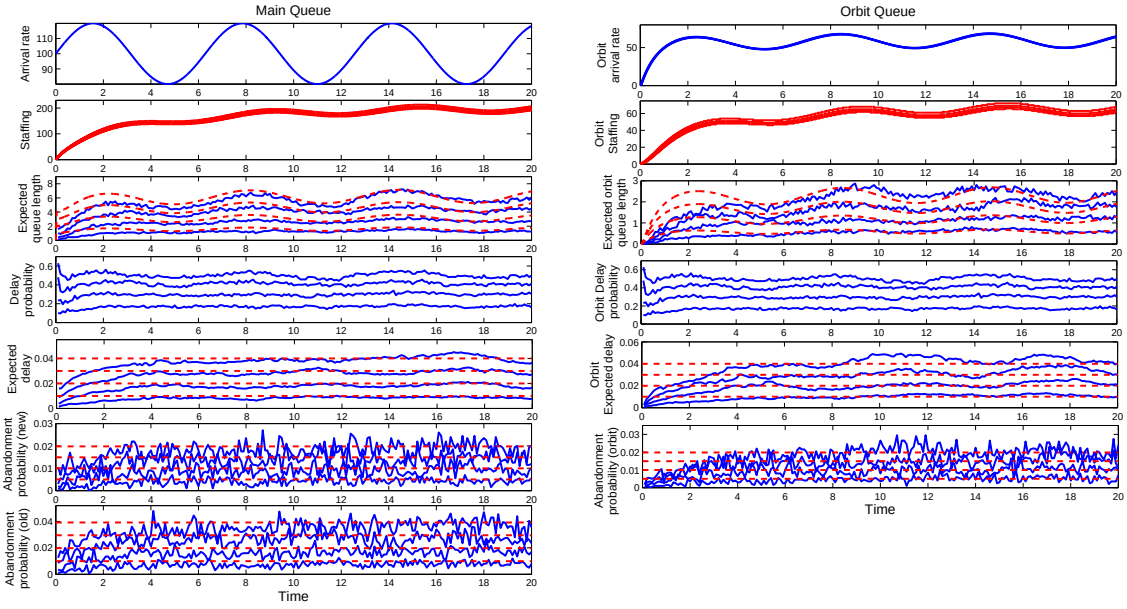
$$\begin{aligned}
E[O_2(t)] &= p_2\, E\left[\int_{t-U_2}^{t} \sigma_2(x)\, dx\right] = p_2\, E[\sigma_2(t-U_{2,e})]E[U_2], \\
E[Q_3(t)] &= E\left[\int_{t-T_3}^{t} \lambda_{F,2}(x)\, dx\right] = E[\lambda_{F,2}(t-T_{3,e})]E[T_3], \\
m_3(t) \equiv E[B_3(t)] &= \bar{F}_3(w) E\left[\int_{t-w-S_3}^{t-w} \lambda_{F,2}(x)\, dx\right] = \bar{F}_3(w) E[\lambda_{F,2}(t-w-S_{3,e})]E[S_3], \\
\lambda_{F,2}(t) &= p \int_0^\infty \sigma_2(t-x)\, dH_2(x) = (1-p_2)E[\sigma_2(t-U_2)],
\end{aligned}$$

where $T_3 \equiv A_3 \wedge w$, and A_3, S_3 and U_2 follow cdfs F_3, G_3 and H_3.

Regarding the DIS-MOL approximation, we generalize (13)–(17) to

$$\begin{aligned}
\lambda_{MOL}(t) &\equiv \sum_{i=1}^{3} \lambda_{mol,i}(t), \quad \text{where} \quad \lambda_{mol,i}(t) \equiv \frac{m_i(t)}{(1-\alpha_i)E[S_i]}, \quad i = 1,2,3, \\
F_{mol}(t) &= \frac{\sum_{k=1}^{3} \lambda_{mol,k}(t) F_k}{\lambda_{mol}(t)}, \quad (1-\alpha_{mol}(t)) = \frac{\sum_{k=1}^{3} \lambda_{mol,k}(t)(1-\alpha_k)}{\lambda_{mol}(t)},
\end{aligned}$$

Figure 8 Performance functions in the $(M_t(0.2)/H_2(1,4), H_2(10/6,4)/s_t + M(2), M(1)) + (0.6, H_2(1,4)/s_t + M(1))$ model with the sinusoidal arrival rate in (22) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.6$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing.



$$G_{mol}(t) = \frac{\sum_{k=1}^{3}(1-\alpha_2)\lambda_{mol,2}(t)G_2}{(1-\alpha_{mol}(t))\lambda_{mol}(t)}.$$

Figures of simulation experiments in the appendix verify the effectiveness of our DIS and DIS-MOL approaches just as in Figures 2 and 3. We remark this analysis can generalize to the case of any finite feedbacks.

7. Proof of Theorem 4

We first act as if the service facility can be partitioned into two parts, one dedicated to the new arrivals, with the other dedicated to the fed-back customers. In model n, the capacities of these two parts are $s_{i,n}(t) \equiv \lceil ns_i(t) \rceil$ for $i = 1, 2$. For the fluid model, the corresponding capacities are $s_i(t) = m_i(t) \equiv E[B_i(t)]$ for $i = 1, 2$. We first discuss the fluid limit and then establish the FWLLN for the partitioned system. Afterwards, we show that the performance in the original system is asymptotically equivalent to the performance in the partitioned system.

7.1. The Partitioned Fluid Model

It is significant that the limit in the FWLLN for the each component of the partitioned system is a deterministic fluid model. The fluid model for the first component also has parameter vectors $(\lambda, s_1, F_1, G_1, w, \alpha_1)$, but they have a different interpretation: Now $\lambda(t)$ is the arrival rate of the

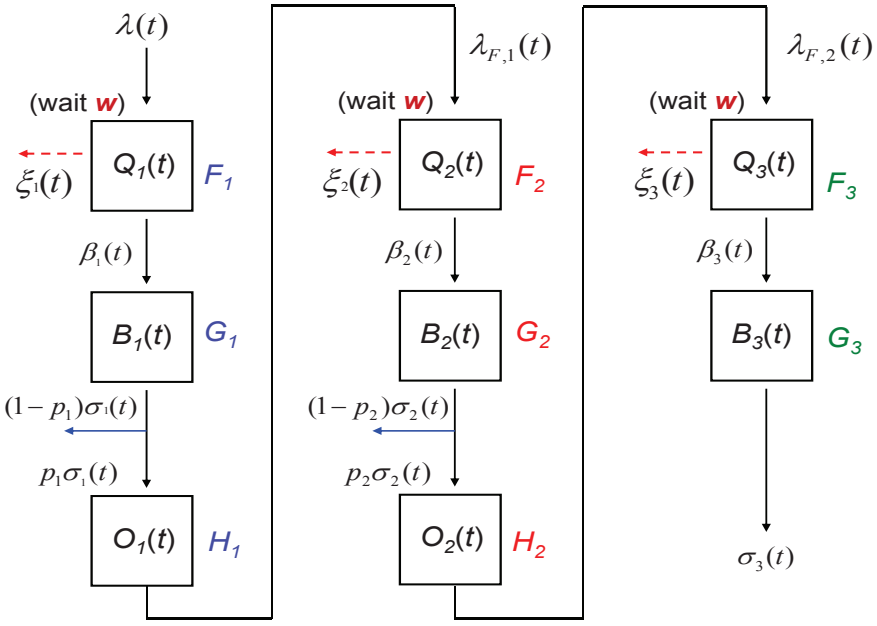


Figure 9 The DIS approximation for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty) + (GI/\infty)$ model with two delayed customer feedback opportunities. Here there are two IS orbit queues. The approximating offered load is $m(t) = m_1(t) + m_2(t) + m_3(t) \equiv E[B_1(t)] + E[B_2(t)] + E[B_3(t)]$.

divisible deterministic fluid at time t. A proportion $F_1(x)$ of the fluid to directly enter the queue from the external input abandons by time x of entering the queue if it has not yet entered service; a proportion $G_1(x)$ of the fluid to directly enter service from the external input completes service by time x after it has begun service. The staffing function $s_1(t)$ stabilizes the waiting time in the fluid model at w. We refer to §4 of Liu and Whitt (2012a) for a discussion of the connection between the DIS model and the fluid model and §10 of Liu and Whitt (2012a) for the explicit performance functions achieving the waiting-time target w.

Just as in Liu and Whitt (2012a), the content of the two types of fluid in service and queue are described by two-parameter deterministic functions $B_i(t,y)$ and $Q_i(t,y)$; $B_i(t,y)$ is the quantity of type-i fluid in service at time t that has been so for time at most y, while $Q_i(t,y)$ is the quantity of type-i fluid in queue at time t that has been so for time at most y. The total content of type-i fluid in service and in queue at time t are thus $B_i(t) = B_i(t,\infty)$ and $Q_i(t) = Q_i(t,\infty)$, respectively. The overall totals are the sums over the two types.

Given the staffing function that we have used, we can verify that the type-i fluid content in service is $B_i(t) = m_i(t)$ and the overall content is $B(t) = m(t)$ for all $t > w$, and that all fluid waits exactly time w before entering service if it does not first abandon. We summarize these observations

in the following theorem. (We first establish this result for the partitioned model and then the original model.)

THEOREM 5. (*DIS staffing stabilizes the waiting time in the fluid model with feedback*) *The DIS staffing in* §2 *is the unique staffing that stabilizes the waiting time at* w *and the abandonment probabilities at* $\alpha_i = F_i(w)$ *for* $i = 1, 2$ *in the* $(G_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ *fluid queue with Bernoulli feedback. All fluid waits in queue exactly time* w *before entering service if it has not abandoned. Just as in Theorem 1, the abandonment rates of the two kinds of fluid are* $\xi_i(t)$*, the rates that the two kinds of fluid enter service are* $\beta_i(t)$*, the service-completion rates of the two kinds of fluid are* $\sigma_i(t)$ *and the feedback arrival rate function is* $\lambda_F(t)$*.*

7.2. The FWLLN for the Partitioned System

For the partitioned system, we can establish the FWLLN recursively, just as we analyze the DIS model in §2. We first consider the model with staffing functions $s_{1,n}(t)$ containing only the external arrivals. For this model, just as in Liu and Whitt (2012c), we can apply the established FWLLN in Liu and Whitt (2012b) to obtain the desired FWLLN. Since the waiting time target is w, we can use §10 of Liu and Whitt (2012a) to uniquely characterize the limiting fluid model, which has staffing function $s_i(t)$.

We now proceed forward to the next queue. From this initial FWLLN for the first partition of the system, we obtain the limit for the sequence of scaled departure counting processes of these customers, denoted by $\{\bar{D}_n^{(1)} : n \geq 1\}$. Given that $\bar{D}_n^{(1)} \Rightarrow \bar{D}^{(1)}$ in D, we can next obtain the corresponding limit for the sequence of customers fed back after service completion, denoted by $\{\bar{D}_n^{(1,2)} : n \geq 1\}$. For that purpose, let $\{X_{n,1,k} : k \geq 1\}$ be a sequence of i.i.d. routing random variables with $X_{n,i,k} = 2$ if the j^{th} departure in $D_n^{(1)}$ is fed back. Then we can represent $D_n^{(1,2)}(t)$ explicitly as

$$D_n^{(1,2)}(t) = \sum_{k=1}^{D_n^{(1)}(t)} 1_{\{X_{n,1,k}=2\}}, \quad t \geq 0, \tag{26}$$

and the associated scaled version as

$$\bar{D}_n^{(1,2)}(t) = \sum_i \bar{Z}_n(t) \circ \bar{D}_n^{(1)}(t), \quad t \geq 0, \tag{27}$$

where $\circ$ is the composition function and

$$\bar{Z}_n(t) \equiv \frac{1}{n} \sum_{k=1}^{\lfloor nt \rfloor} 1_{\{X_{n,1,k}=2\}} \Rightarrow pt \quad \text{in} \quad D \tag{28}$$

We now apply the continuous mapping theorem in §3.4 of Whitt (2002) for the continuous composition functions appearing in (27), see Theorem 13.2.1 of Whitt (2002), with the established

limit for $D_n^{(1)}$ and the FWLLN for partial sums of i.i.d. random variables $\bar{Z}_{n,i,j}$. to obtain the limit $\bar{D}_n^{(1,2)} \Rightarrow \bar{D}^{(1,2)}$.

Given the limit for $\bar{D}_n^{(1,2)}$ just established, we can apply the FWLLN for the IS orbit queue in Pang and Whitt (2010) to obtain the FWLLN for all the processes associated with the orbit queue, including its departure process, which serves as the arrival process to the second part of the partitioned system, serving the fed-back customers.

Finally, we obtain a corresponding FWLLN for the second partition of the partitioned system, serving the fed-back customers, using the same reasoning as above. Since the waiting-time target is w for both classes the fluid models are uniquely determined by Theorem 8 in §10 of Liu and Whitt (2012a). Hence all the performance functions are as described. It only remains to show that the partitioned system is asymptotically equivalent to the original system. We first discuss the relation between the corresponding fluid models in the partitioned system.

7.3. Additivity of Fluid Models

We now observe that the limiting fluid model in the theorem is actually equivalent to the fluid limit for the partitioned system, because both systems have the common constant waiting time w. This equivalence is a consequence of the following more general theorem about fluid models, which we state without proof.

THEOREM 6. (*additivity of fluid models*) *Two fluid models with the FCFS discipline indexed by i that are combined into a two-class FCFS fluid queue by having total arrival-rate function $\lambda = \lambda_1 + \lambda_2$ and staffing $s(t) = s_1(t) + s_2(t)$ have additive performance with*

$$B(t,x) = B_1(t,x) + B_2(t,x) \quad \textit{and} \quad Q(t,x) = Q_1(t,x) + Q_2(t,x) \quad \textit{for all} \quad t, x \tag{29}$$

if and only if the two boundary waiting functions $w_i(t)$ coincide, in which case $w(t) = w_1(t) = w_2(t)$ for all t.

7.4. Asymptotic Equivalence

Even though the limiting fluid models of the partitioned system and the original system are the same, it remains to show that the established FWLLN for the partitioned system implies a corresponding FWLLN for the original system, with identical limits. The problem is that the two kinds of customers interact in the original system, so that the partitioning is not actually valid for each n. However, we can show that the customers from the different components of the partition interact over an asymptotically small part of the total capacity. Thus, the difference can be shown to be asymptotically negligible. To visually think of the separation, we can think of the servers being

numbered, with arrivals from one class taking the smallest numbered free server, while arrivals from the other class taking the largest numbered free server. Then the two classes contend only in the middle, when the system becomes full (which will be the case here after an initial transient period).

We will sketch the argument to show the asymptotic equivalence. To do so, we observe from §10 of Liu and Whitt (2012a) that a small perturbation of the waiting-time target w in the fluid model yields a controlled uniformly small perturbation of the staffing over any bounded time interval $[a,b]$, where $a > w$. Let $s_i(t,w)$ be the staffing function for the two classes ($i = 1$ for external input and $i = 2$ for the fedback fluid) at time t as a function of the constant waiting-time target w. It follows that, for any $\epsilon > 0$, there exists $\delta \equiv \delta(\epsilon) > 0$ so that

$$s_i(t, w+\epsilon) - \delta < s_i(t,w) < s_i(t, w-\epsilon) + \delta \quad \text{for} \quad a \le t \le b \quad \text{and} \quad i = 1,2. \tag{30}$$

Moreover, by the FWLLN for the partitioned system just established, the scaled content $\bar{B}_i^n(t,w)$ can be made arbitrarily close to the staffing $s(t,w)$, i.e, for any $a > w > 0$,

$$\sup_{a \le t \le b} \{|\bar{B}_i^n(t,w) - s_i(t,w)|\} \Rightarrow 0 \quad \text{as} \quad n \to \infty. \tag{31}$$

Hence, given $w > \epsilon > 0$, suppose that the waiting-time target is required to fall in the interval $[w-\epsilon, w+\epsilon]$. Then, there exists $\delta \equiv \delta(\epsilon) > 0$ and n_0 such that for $n \ge n_0$

$$s_i(t, w+\epsilon) - 2\delta < \bar{B}_i^n(t,w) < s_i(t, w-\epsilon) + 2\delta \quad \text{for} \quad a \le t \le b \quad \text{and} \quad i = 1,2. \tag{32}$$

Of course, in our combined system we also have $s(t,w) = s_1(t,w) + s_2(t,w)$, but we now have slack so that the content of one class can be too large, while the content of the other class is too small. Since $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ and we can let ϵ be arbitrarily small, we achieve the fluid limit of the partitioned model for the original model. Hence, the proof of Theorem 4 is complete.

In closing, we remark that an alternative proof can be done by the compactness argument, where we show that the sequence of scaled queueing processes are tight and then uniquely characterize the limit in terms of the fluid model. Tightness for the sequence of class-i scaled departure counting processes holds because the increments, conditional on any history, are stochastically bounded over any bounded interval by a constant rate Poisson process, with rate equal to the supremum of the staffing function multiplied by the supremum of the service-time hazard-rate function, which is bounded because the system starts empty and the service-time distributions have positive finite densities. ∎

8. Conclusions

In this paper we have extended the two-queue approximating *Delayed-Infinite-Server* (DIS) model for the $M_t/GI/s_t + GI$ model in Liu and Whitt (2012c) to the corresponding five-queue approximating DIS model depicted in Figure 1 for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model with Bernoulli feedback after a random delay in an infinite-server orbit queue and a corresponding six-queue approximating DIS model for the corresponding model with a $(GI/s_t + GI)$ finite-capacity orbit queue. These models present attractive alternatives to the Erlang-R model in Yom-Tov and Mandelbaum (2013) because the fed-back customers can have different service-time and patience cdf's. The same approach extends to any finite number of feedbacks; the case of two feedbacks is discussed in §6.3 and the appendix. The approach applies to systems with or without customer abandonment. Without customer abandonment, the offered load is $m_\alpha(t)$ for $\alpha = 0$; then we would use a delay-probability target, as in Feldman et al. (2008), Jennings et al. (1996) and Yom-Tov and Mandelbaum (2013).

Theorems 1 and 2 give explicit expressions for all DIS performance functions in general and with sinusoidal arrival rate functions. Moreover, we have presented results of simulation experiments showing that the DIS offered load (OL) itself provides staffing that successfully stabilizes abandonment probabilities and expected waiting times with low QoS targets. Theorem 4 establishes a FWLLN showing that the DIS staffing achieves its performance goals asymptotically as the scale increases.

In §4 we have also developed a new aggregate approximating single-class *Delayed-Infinite-Server Modified-Offered-Load* (DIS-MOL) approximation to set staffing levels with low waiting-time (high QoS) targets. We showed that we can use either the aggregate abandonment probability target or the waiting-time target, but the waiting-time target tends to produce a faster algorithm, in part because the abandonment probability target $F_{mol}(w;t)$ is a time-dependent function. We have presented results of simulation experiments in §5 and §6 showing that the new DIS and DIS-MOL staffing algorithms are effective across a wide range of QoS targets.

The queue with Bernoulli feedback after an additional delay in a finite-capacity orbit queue is a special case of a network of many-server queues with feedback. Our excellent results in this case indicate that the methods should apply to more general networks of queues, including multiple queues and customer classes, with various forms of routing, but such more general models remain to be examined carefully.

Acknowledgement

This research began as part of the first author's doctoral dissertation at Columbia University. The second author received support from NSF grants CMMI 1066372 and 1265070.

References

Armony, M., S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov. 2011. Patient flow in hospitals: a data-based queueing-science perspective. New York University, http://www.stern.nyu.edu/om/faculty/armony/.

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. Stat. Assoc.* **100** 36–50.

Defraeye, M., I. van Nieuwenhuyse. 2013. Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems* **54**(4) 1558–1567.

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

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

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. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 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 Oper. Management* **16** 13–29.

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

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 $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Oper. Res. Letters* **40** 307–312.

Liu, Y., W. Whitt. 2012c. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* **60** 1551–1564.

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

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

Stolletz, R. 2008. Approximation of the nonstationary $M(t)/M(t)/c(t)$ queue using stationary models: the stationary backlog-carryover approach. *Eur. J. Oper. Res.* **190**(2) 478–493.

Whitt, W. 1982. Approximating a point process by a renewal process: two basic methods. *Oper. Res.* **30** 125–147.

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

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

Yom-Tov, G., A. Mandelbaum. 2013. The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. *Manufacturing and Service Oper. Management* Working paper, the Technion.