Stabilizing Performance in a Service System with Time-Varying Arrivals and Customer Feedback

Yunan Liu and Ward Whitt

Department of Industrial Engineering, North Carolina State University, Raleigh, NC 27695
IEOR Department, Columbia University, New York, NY 10027-6699

June 23, 2017

Abstract

Analytical offered-load and modified-offered-load (MOL) approximations are developed to determine staffing levels that stabilize performance at designated targets in a non-Markovian many-server queueing model with time-varying arrival rates, customer abandonment from queue and random feedback with additional feedback delay in an infinite-server or finite-server queue. To provide a flexible model that can be readily fit to system data, the model has Bernoulli routing, where the feedback probabilities, service-time, patience-time and feedback-delay distributions all are general and may depend on the visit number. Simulation experiments confirm that the new MOL approximations are effective. A many-server heavy-traffic FWLLN shows that the performance targets are achieved asymptotically as the scale increases.

(The figures have been removed from this version to create a PDF/A complient version of the paper file. Please see the journal or the authors' web pages for a full version of the paper.)

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 service systems with time-varying arrival rates in order to stabilize performance at designated targets; see Green et al. (2007) for a review and Stolletz (2008), Defraeye and van Nieuwenhuyse (2013), Liu and Whitt (2014), Yom-Tov and Mandelbaum (2014) and He et al. (2016) for recent related work. We continue to focus on service systems that can be modeled as many-server queues with customer abandonment from queue and non-exponential distributions, but here in addition we consider Bernoulli feedback with additional delay after completing service.

A queue with delayed feedback after completing service is a special queueing network, about which there is an enormous literature, but our concern is with the time-dependent performance of a non-stationary non-Markov model, which is well beyond exact analysis. We do assume that the arrival process is a *nonhomogeneous Poisson process* (NHPP), but the service-time, patience-time and delay-before-return distributions all can be non-exponential and can change upon successive feedbacks. At first glance, it would seem that previous methods do not apply to the generalization with multiple delayed feedbacks having changing parameters. Our main innovation is to propose an approximation involving a series of infinite-server models. Instead of the natural two-queue model for a single delayed feedback shown on the left in Figure **??**, with an orbit queue for the customers experiencing extra delay in addition to the usual queue, we propose the five-queue series model on the right, which also has separate queues for the customers waiting and in service upon first visit and upon second visit, as well as for those customers being delayed in between the two visits; we elaborate on the model below.

Previous work has shown that a time-varying arrival-rate function and a non-exponential service-time distribution can have a significant impact on performance; see Eick et al. (1993) for discussion of the basic $M_t/GI/\infty$ infinite-server special case. Figures 1 and 2 of Jennings et al. (1996) dramatically show the poor performance that can occur if we use stationary methods to set staffing levels, either using the overall average arrival rate or using the pointwise-stationary approximation (PSA), which uses a stationary model in a nonstationary way, letting the arrival rate in the stationary model at time t be the actual arrival rate $\lambda(t)$ at time t. As reviewed in Green et al. (2007), the PSA can be effective with relative short service times, but tends to fail badly with longer service times. The additional delayed feedback adds to the challenge because it can significantly alter the time-varying demand, not only in magnitude but also in timing. For example, the delayed feedback can amplify or damp the peak demand and shift it in time.

The literature exposes two common reasons for feedback after completing service: First, de Vericourt and Zhou (2005) focus on call center customers that may return later because the initial service was unsatisfactory. Second, Yom-Tov and Mandelbaum (2014) focus on the treatment of patients by a doctor in a hospital that may naturally occur in stages, starting with an initial screening and continuing later after tests have been ordered and completed. Our paper is closely related to Yom-Tov and Mandelbaum (2014), 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. They showed that the MOL approximation has great potential for improved performance analysis in healthcare, where the service times tend to be relatively long, so that PSA does not apply.

Motivated by these applications, we consider a feedback model that has appealing flexibility. In particular, instead of the Markovian routing with fixed feedback probability p and one fixed service-time distribution considered in Yom-Tov and Mandelbaum (2014), we consider history-dependent Bernoulli routing, where there may be any number of visits and the feedback probability p and the service-time distribution and the subsequent delay distribution (before returning for a new service) all may vary with the visit number. We focus on the common important case of at most one feedback, which seems to be a more realistic model than Markovian routing, which produces a geometric random number of feedbacks. It is significant that the approach here also extends directly to any finite number of feedbacks; we demonstrate by also considering examples with two feedback opportunities. Our methods also extend directly to time-dependent feedback probabilities, but we do not examine that here. (The justification is that a time-dependent independent thinning of an NHPP is again an NHPP; see §2.3 and §2.4 of Ross (1996).)

We also allow customer abandonment, which often tends to be more realistic for many service systems, as observed by Garnett et al. (2002). The patience-time distributions are also allowed to be non-exponential and depend on the visit number. Just as in Yom-Tov and Mandelbaum (2014), we use the general offered-load (OL) method with the MOL refinement, as reviewed in Jennings et al. (1996), Green et al. (2007), Liu and Whitt (2012c) and Whitt (2013). There are difference between the MOL methods designed to stabilize the delay probability and the abandonment probability, as discussed in Liu and Whitt (2012c), but the main contribution here beyond Yom-Tov and Mandelbaum (2014) is the new method for computing the time-varying offered load. Because the offered load is the primary determinant of performance, the performance impact from more faithfully representing the service and feedback process in a time-varying setting can be great.

To analyze this new feedback model with customer abandonment, we draw on Liu and Whitt (2012c) in which we developed a *delayed-infinite-server* (DIS) offered-load approximation and a new DIS-MOL (*DIS-modified-offered-load*) algorithm to determine time-dependent staffing levels in order to stabilize expected delays and abandonment probabilities at specified *quality of service* (QoS) targets in a many-server queue with time-varying arrival rates. The model in Liu and Whitt (2012c) was $M_t/GI/s_t + GI$ model, having arrivals according to an NHPP 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 (FCFS) 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. (2015) and Brown et al. (2005).

We refer to the base model with a single feedback considered here 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, depending on the visit number, while the orbit queue has a single service-time cdf H, with all waiting customers entering service in a FCFS order. We develop approximations for the number of customers waiting before service and in service upon each visit and the number of customers in orbit. When we refer to the number of customers in the system or the waiting time, we do not include the orbit queue.

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 before starting service on any visit 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, as shown in Figure **??**. (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 in Liu and Whitt (2012c), the DIS algorithm is ineffective for high-QoS (low waiting-time) targets. We develop a new DIS-MOL approximation for that case and conduct simulation experiments to show that it is effective. (The DIS-MOL approximation is also asymptotically correct as the scale increases.) 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 for the steady-state performance 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. In the online appendix 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. We defer the proof to the online appendix. 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. Finally, in §7 we draw conclusions. Additional supporting material appears in the online appendix Liu and Whitt (2015).

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 with FCFS service, 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 **??**.

2.1 The Approximating Five-Queue DIS Model

The approximating DIS model, depicted on the right in Figure **??**, 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. (1993). Theorem 1 of Eick et al. (1993) 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

$$m(t) \equiv m_1(t) + m_2(t) \equiv E[B_1(t)] + E[B_2(t)]. \tag{2.1}$$

As before, all flows are Poisson processes, with rate functions as depicted in Figure **??**. 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. (1993). (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

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

Let 1_C be the indicator variable, which is equal to 1 if event C occurs and is equal to 0 otherwise.

Theorem 2.1 (*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

$$\beta_1(t) = \lambda(t-w)\bar{F}_1(w) \quad \textit{and} \quad \beta_2(t) = \lambda_F(t-w)\bar{F}_2(w).$$

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 online appendix Liu and Whitt (2015). As discussed in Eick et al. (1993), 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 online 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 2.1 and obtain the following concrete formulas for the performance measures. We let $x^+ \equiv \max(x, 0)$.

Corollary 2.1 (*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* $w>0$. *All results in Theorem 2.1 hold with rate functions*

$$\begin{aligned}
\xi_1(t) &= \int_0^{t\wedge w}\lambda(t-x)dF_1(x), \qquad \beta_1(t)=\lambda(t-w)\cdot\bar{F}_1(w)\cdot 1_{\{t\geq w\}},\\
\sigma_1(t) &= \bar{F}_1(w)\int_0^{(t-w)^+}\lambda(t-w-x)dG_1(x),\\
\lambda_F(t) &= \int_0^{(t-w)^+}p\sigma_1(t-y)dH(y)\\
&= p\bar{F}_1(w)\int_0^{(t-w)^+}\int_0^{t-w-y}\lambda(t-w-x-y)dG_1(x)dH(y),\\
\xi_2(t) &= \int_0^{(t-w)^+\wedge w}\lambda_F(t-z)dF_2(z)\\
&= (1-p)\bar{F}_1(w)\int_0^{(t-w)^+\wedge w}\int_0^{t-w-z}\int_0^{t-w-y-z}\lambda(t-w-x-y-w)dG_1(x)dH(y)dF_2(z),\\
\beta_2(t) &= \lambda_{F_1}(t-w)\bar{F}_2(w)\cdot 1_{\{t\geq 2w\}}\\
&= p\bar{F}_1(w)\bar{F}_2(w)\int_0^{(t-2w)^+}\int_0^{t-2w-y}\lambda(t-2w-x-y)dG_1(x)dH(y),\\
\sigma_2(t) &= \int_0^{(t-2w)^+}\beta_2(t-z)dG_2(z)\\
&= p\bar{F}_1(w)\bar{F}_2(w)\int_0^{(t-2w)^+}\int_0^{t-2w-z}\int_0^{t-2w-y-z}\lambda(t-2w-x-y-z)dG_1(x)dH(y)dG_2(z),
\end{aligned}$$

and mean number of customers in these five IS queues

$$\begin{aligned}
E[Q_1(t)] &= \int_0^{t\wedge w}\lambda(t-x)\bar{F}_1(x)dx, \quad E[B_1(t)]=\bar{F}_1(w)\int_0^{(t-w)^+}\lambda(t-w-x)\bar{G}_1(x)dx,\\
E[O(t)] &= \int_0^{(t-w)^+}p\sigma_1(t-x)\bar{H}(x)dx\\
&= p\bar{F}_1(w)\int_0^{(t-w)^+}\int_0^{t-w-y}\lambda(t-w-x-y)dG_1(x)\bar{H}(y)dy,\\
E[Q_2(t)] &= \int_0^{(t-w)^+\wedge w}\lambda_F(t-z)\bar{F}_2(z)dz,\\
&= p\bar{F}_1(w)\int_0^{(t-w)^+\wedge w}\int_0^{t-w-z}\int_0^{t-w-y-z}\lambda(t-w-x-y-z)dG_1(x)dH(y)\bar{F}_2(z)dz,\\
E[B_2(t)] &= \bar{F}_2(w)\int_0^{(t-2w)^+}\lambda_F(t-w-z)\bar{G}_2(z)dz\\
&= p\bar{F}_1(w)\bar{F}_2(w)\int_0^{(t-2w)^+}\int_0^{t-2w-z}\int_0^{t-2w-y-z}\lambda(t-2w-x-y-z)dG_1(x)dH(y)\bar{G}_2(z)dz.
\end{aligned}$$

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 give explicit formulas for the case of sinusoidal arrival rate functions, which are often used to create stylized models. In the longer online 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 $0 \leq w_1 \leq w_2 \leq \cdots \leq w_n$. The goal is to design an appropriate staffing function $s(t)$ for $0 \leq t \leq T$ 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.

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. (In the rest of this paper we restrict attention to M_t arrivals. The greater generality provides a basis for extensions. See He et al. (2016) for a discussion of G_t arrivals.) 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 and the limit corresponds to the associated fluid model studied directly in Liu and Whitt (2012a). 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

$$N^{(e)}(t) = N^{(b)}(\Lambda(t)), \quad t \geq 0, \tag{3.1}$$

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

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds \tag{3.2}$$

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.,

$$n^{-1} N^{(b)}(nt) \Rightarrow t \quad \text{in} \quad D \quad \text{as} \quad n \to \infty, \tag{3.3}$$

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

$$N_n^{(e)}(t) \equiv N^{(b)}(n\Lambda(t)), \quad t \geq 0, \tag{3.4}$$

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 (3.3), we deduce that $N_n^{(e)}$ also obeys the FWLLN

$$\bar{N}_n^{(e)}(t) \equiv n^{-1}N^{(e)}(nt) \Rightarrow \Lambda(t) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty. \tag{3.5}$$

We remark that the limit is the cumulative external arrival rate function of the fluid model in Liu and Whitt (2012a).

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 2.1, in particular,

$$s(t) = m(t) = m_1(t) = m_2(t) = E[B_1(t)] + E[B_2(t)]. \tag{3.6}$$

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.

Theorem 3.1 (*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* (3.4) *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*

$$s(t) = m(t) = m_1(t) + m_2(t) = E[B_1(t)] + E[B_2(t)], \tag{3.7}$$

as given in Theorem 2.1. *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* $w < b < \infty$,

$$\begin{aligned} \sup_{0 \leq t \leq b} \{|\bar{Q}_n^{(i)}(t) - E[Q^{(i)}(t)]|\} &\Rightarrow 0, \quad \sup_{0 \leq t \leq b} \{|W_n^{(i)}(t) - w|\} \Rightarrow 0, \\ \sup_{0 \leq t \leq b} \{|\bar{A}_n^{(i)}(t) - A^{(i)}(t)|\} &\Rightarrow 0, \quad E[W_n^{(i)}(t)] \to w, \quad \textit{and} \\ \sup_{0 \leq t \leq b_i,\, w_i < u < b_i} \{|\bar{A}_n^{(i)}(t, t+u) - A^{(i)}(t,u)|\} &\Rightarrow 0, \quad t \geq 0, \end{aligned} \tag{3.8}$$

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

$$E[Q^{(i)}(t)] = E[Q^{(i)}(t,0)] \equiv \int_0^{w_i} \lambda_i(t-x)\bar{F}_i(x)\,dx, \quad A^{(i)}(t) \equiv \int_0^t \xi_i(s)\,ds$$

$$\xi_i(t) \equiv \int_0^{w_i} \lambda_i(t-x) f_i(x)\,dx \quad \textit{and} \quad A^{(i)}(t,u) \equiv \Lambda_i(t)\alpha_i, \quad u > w_i. \tag{3.9}$$

We give the proof in the appendix. 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.

We remark that the DIS-MOL algorithm also can be shown to be asymptotically correct as the scale increases in the setting of Theorem 3.1, but that result is misleading, because the FWLLN in Theorem 3.1 expresses a MSHT limit in the ED regime, where the waiting-time target w and abandonment probabilities α_i are held fixed while the scale increases. It remains to establish a MSHT limit in the QED regime where instead the probability of delay is held fixed as the scale increases.

4 The Refined DIS-MOL Approximation

Just as in Liu and Whitt (2012c), simulation experiments to be discussed in §5 show that the DIS approximation is effective under low-QoS (high-waiting-time) targets, but is ineffective under the common high-QoS (low-waiting-time) targets. Thus, we develop a refined DIS-MOL staffing algorithm here. 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

$$\lambda_{mol}(t) \equiv \sum_{i=1}^{2} \lambda_{mol,i}(t) \quad \text{and} \quad \lambda_{mol,i}(t) \equiv \frac{m_i(t)}{(1-\alpha_i)E[S_i]}, \tag{4.1}$$

where $m_i(t) = E[B_i(t)]$ for each i. We enforce the additivity in (4.1) 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 (4.1) is 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 (4.1) to use in the stationary $M/GI/s+GI$ model.

The MOL arrival rate in (4.1) 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 (4.1) 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 (4.1), 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 (4.1). 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

$$F_{mol}(t) = \frac{\lambda_{mol,1}(t)F_1 + \lambda_{mol,2}(t)F_2}{\lambda_{mol}(t)} \tag{4.2}$$

so that

$$\begin{aligned} 1 - \alpha_{mol}(t) &= \frac{\lambda_{mol,1}(t)(1-\alpha_1) + \lambda_{mol,2}(t)(1-\alpha_2)}{\lambda_{mol}(t)} && (4.3)\\ \text{and} \qquad G_{mol}(t) &= \frac{(1-\alpha_1)\lambda_{mol,1}(t)G_1 + (1-\alpha_2)\lambda_{mol,2}(t)G_2}{(1-\alpha_{mol}(t))\lambda_{mol}(t)}. && (4.4) \end{aligned}$$

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 (4.4), we have

$$E[S_{mol}(t)] = \frac{(1-\alpha_1)\lambda_{mol,1}(t)E[S_1] + (1-\alpha_2)\lambda_{mol,2}(t)E[S_2]}{(1-\alpha_{mol}(t))\lambda_{mol}(t)} \tag{4.5}$$

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 4.1 (*additivity*) *With these definitions, we maintain the important MOL additivity assuming that*

$$m_{mol}(t) \equiv (1-\alpha_{mol}(t))\lambda_{mol}(t)E[S_{mol}(t)]. \tag{4.6}$$

$$\begin{aligned} \text{Then} \qquad m_{mol}(t) &\equiv (1-\alpha_{mol}(t))\lambda_{mol}(t)E[S_{mol}(t)] \\ &= (1-\alpha_1)\lambda_{mol,1}(t)E[S_1] + (1-\alpha_2)\lambda_{mol,2}(t)E[S_2] = m(t). && (4.7) \end{aligned}$$

Proof We start with (4.6) and then apply the definition of $E[S_{mol}(t)]$ in (4.5) to get the second line. We then apply (4.1). ■

4.2 Computing the DIS-MOL Staffing Function

For each time t, we apply the constant arrival rate in (4.1), abandonment cdf in (4.2) and service-time cdf in (4.4) 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

$$E[W_s(\infty)] = \sum_{i=s}^{\infty} E[W_s(\infty)|Q(\infty) = i] \cdot \pi_i = \sum_{i=s}^{\infty}\sum_{k=0}^{s-i} \frac{1}{s\mu + \delta_k} \cdot \pi_i, \tag{4.8}$$

where μ is the reciprocal of the mean service time in (4.5) 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 (4.3), 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 use 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

$$\lambda(t) = \bar{\lambda}(1 + r\sin(t)) = 100(1 + r\sin(t)), \quad t \geq 0, \tag{5.1}$$

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 and larger values of r in the appendix. The same stable performance is seen for $r = 0.5$, but some degradation in performance is seen where the staffing decreases for $r = 0.8$.

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 online appendix.)

Figures 1 and 2 show the results of the simulation experiment for high and low waiting-time targets. In Figure 1 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 2 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 1 we see very smooth and accurate plots of the expected waiting times and abandonment

Figure 1: 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 (5.1) 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.

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 (5.1) 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.

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 2 with the lower waiting-time targets. There is greater variability with the lower waiting-time targets.

Figure 2 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 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 by applying the simple *square root staffing* (SRS) formula

$$s(t) \equiv m(t) + \beta\sqrt{m(t)}, \tag{5.2}$$

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,

$$\beta_{DIS-MOL}(t) = \frac{s_{DISMOL}(t) - m(t)}{\sqrt{m(t)}} \tag{5.3}$$

for the example considered in Figure 2. Figure 3 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 online appendix.

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.

Figure 3: 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 2 as a function of the waiting-time target w.

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 online 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,

$$\lambda_1(t) = 100(1 + 0.2\sin(t)), \quad \text{and} \quad \lambda_2(t) = 60(1 + 0.2\sin(0.8t + 2)). \tag{6.1}$$

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 4 and 5. The results are good, just as in §5.

Figure 4: 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 (6.1), 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 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 (6.1), 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 6 and 7 show the results of the simulation experiment for high and low waiting-time targets, respectively, again based on 2000 independent replications, each starting empty.

Figure 6: 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 (5.1) 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.

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 (5.1) 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.

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 cdf G_3 and patience cdf F_3. The new DIS model has *eight* IS queues in series, as depicted in Figure **??**.

Since there are now three customer classes, characterized by their class-dependent service-time and patience-time distributions, we easily generalize results in Theorem 2.1 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 (4.1)–(4.4) to

$$\begin{aligned}
\lambda_{MOL}(t) &\equiv \sum_{i=1}^{3} \lambda_{mol,i}(t), \quad \text{where} \quad \lambda_{mol,i}(t) \equiv \frac{m_i(t)}{(1-\alpha_i)E[S_i]}, \quad i = 1,2,3, \\
F_{mol}(t) &= \frac{\sum_{k=1}^{3} \lambda_{mol,k}(t) F_k}{\lambda_{mol}(t)}, \quad (1 - \alpha_{mol}(t)) = \frac{\sum_{k=1}^{3} \lambda_{mol,k}(t)(1-\alpha_k)}{\lambda_{mol}(t)}, \\
G_{mol}(t) &= \frac{\sum_{k=1}^{3} (1-\alpha_2)\lambda_{mol,2}(t) G_2}{(1-\alpha_{mol}(t))\lambda_{mol}(t)}.
\end{aligned}$$

Figures of simulation experiments in the online appendix verify the effectiveness of our DIS and DIS-MOL approaches just as in Figures 1 and 2. We remark that this analysis can generalize to the case of any finite number of feedbacks.

7 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 **??** 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 (2014) 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 online 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 (2014).

Theorem 2.1 here and Theorem 1 of the online appendix 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 3.1 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, including models with retrials from blocked arrivals as in the large literature reported in Artalejo (2010), but such more general models remain to be examined carefully.

Acknowledgement

We thank the National Science Foundation for support: NSF grants CMMI 1362310 (first author) and CMMI 1066372 and 1265070 (second author). This research began as part of the first author's doctoral dissertation at Columbia University.

References

Armony M, Israelit S, Mandelbaum A, Marmor Y, Tseytlin Y, Yom-Tov G (2015) Patient flow in hospitals: a data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Artalejo JR (2010) Accessible bibliography on retrial queues: progress in 2000-2009. *Mathematics of Computer Modeling* 51:1071–1081.

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

de Vericourt F, Zhou YP (2005) Managing response time in a call-routing problem with service failure. *Oper. Res.* 53(6):968–981.

Defraeye M, van Nieuwenhuyse I (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 SG, Massey WA, Whitt W (1993) The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41:731–742.

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

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

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16:13–29.

He B, Liu Y, Whitt W (2016) Staffing a service system with non-Poisson nonstationary arrrivals, probability in the Engoneering and Informational Sciences, to appear.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42:1383–1394.

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

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

Liu Y, Whitt W (2014) Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 28:419–449.

Liu Y, Whitt W (2015) Online appendix: Stabilizing performance in many-server queues with time-varying arrivals and customer feedback, columbia University, http://www.columbia.edu/∼ww2040/StabFeedbackapp.pdf.

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

Ross SM (1996) *Stochastic Processes* (New York: Wiley), second edition.

Stolletz R (2008) Approximation of the nonstationary $M(t)/M(t)/c(t)$ queue using stationary models: the stationary backlog-carryover approach. *European Journal of Operations Research* 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* (New York: Springer).

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

Whitt W (2013) Offered load analysis for staffing. *Manufacturing and Service Operations Management* 15(2):166–169.

Yom-Tov G, Mandelbaum A (2014) Erlang R: a time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Oper. Management* 16(2):283–299.