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

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

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 abandonment from queue and random feedback with additional delay after completing service. To provide a flexible model that can be readily fit to system data, the model has history-dependent Bernoulli routing, 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 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. A many-server heavy-traffic FWLLN shows that the performance targets are achieved asymptotically as the scale increases.

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; Revision submitted January 9, 2015

1. Overview

This e-companion to the main paper has three more sections. We start in §2 by displaying the explicit DIS performance functions with a sinusoidal arrival rate function. Then in §3 we prove Theorem 2 in the main paper. Finally, in §3.5 we give examples with a lower arrival rate, and thus lower staffing levels.

2. Performance Functions with a 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 1. *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*

$$\begin{aligned}
E[Q_1(t)] &= E[T_1](a + \gamma(T_{1,e})b \cdot \sin(ct - \theta(T_{1,e}))),\\
E[B_1(t)] &= \bar{F}_1(w)E[S_1]\left(a + \gamma(S_{1,e})b \cdot \sin[c(t-w) - \theta(S_{1,e})]\right),\\
E[O(t)] &= p\,\bar{F}_1(w)E[U]\left(a + \gamma(S_1)\gamma(U_e)b \cdot \sin[c(t-w) - \theta(S-1) - \theta(U_e)]\right),\\
E[Q_2(t)] &= p\,\bar{F}_2(w)E[T_2]\left(a + \gamma(S_1)\gamma(U)\gamma(T_{2,e})b \cdot \sin[c(t-w) - \theta(S_1) - \theta(U) - \theta(T_{2,e})]\right),\\
E[B_2(t)] &= p\,\bar{F}_1(w)\bar{F}_2(w)E[S_2]\left(a + \gamma(S_1)\gamma(U)\gamma(S_{2,e})b \cdot \sin[c(t-2w) - \theta(S_1) - \theta(U) - \theta(S_{2,e})]\right),
\end{aligned}$$

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

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

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 2. *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*

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

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* $RA(m) \leq RA(m^*)$.

Furthermore, in both cases

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

3. Proof of the Main Limit Theorem

In this section we prove Theorem 2 in the main paper. 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.

3.1. 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 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 3. (*DIS staffing stabilizes the waiting time in the fluid model with feedback*) *The DIS staffing in §2 of the main paper 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 in the main paper, 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 $\lambda_F(t)$.*

3.2. 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

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

and the associated scaled version as

$$\bar{D}_n^{(1,2)}(t) = \sum_i \bar{Z}_n(t) \circ \bar{D}_n^{(1)}(t), \quad t \geq 0, \tag{3}$$

where $\circ$ is the composition function and

$$\bar{Z}_n(t) \equiv \frac{1}{n}\sum_{k=1}^{\lfloor nt \rfloor} 1_{\{X_{n,1,k}=2\}} \Rightarrow pt \quad \text{in} \quad D \tag{4}$$

We now apply the continuous mapping theorem in §3.4 of Whitt (2002) for the continuous composition functions appearing in (3), 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.

3.3. 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 4. (*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*

$$B(t,x) = B_1(t,x) + B_2(t,x) \quad \textit{and} \quad Q(t,x) = Q_1(t,x) + Q_2(t,x) \quad \textit{for all} \quad t, x \tag{5}$$

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.

3.4. 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

$$s_i(t, w+\epsilon) - \delta < s_i(t,w) < s_i(t, w-\epsilon) + \delta \quad \text{for} \quad a \le t \le b \quad \text{and} \quad i = 1,2. \tag{6}$$

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 $a > w > 0$,

$$\sup_{a \le t \le b} \{|\bar{B}_i^n(t,w) - s_i(t,w)|\} \Rightarrow 0 \quad \text{as} \quad n \to \infty. \tag{7}$$

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 $n \ge n_0$

$$s_i(t, w+\epsilon) - 2\delta < \bar{B}_i^n(t,w) < s_i(t, w-\epsilon) + 2\delta \quad \text{for} \quad a \le t \le b \quad \text{and} \quad i = 1,2. \tag{8}$$

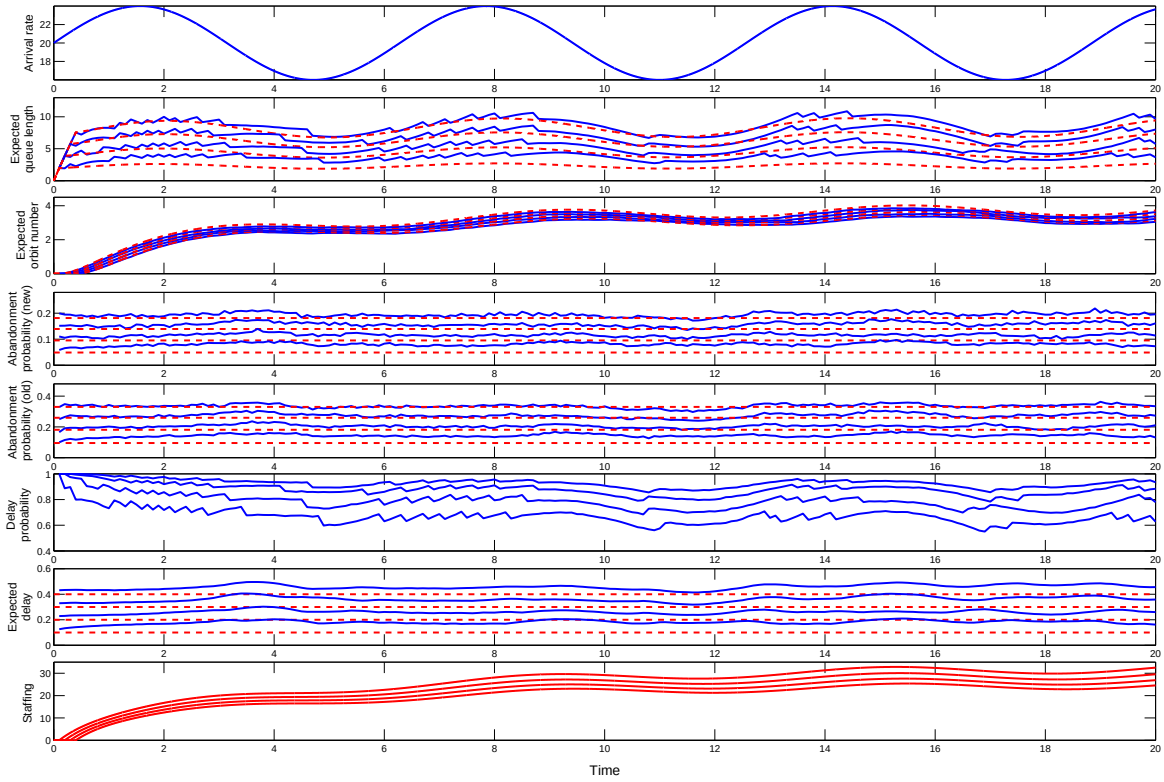
Of course, in our combined system we also have $s(t,w) = s_1(t,w) + s_2(t,w)$, but we now have slack so that the content of one class can be too large, while the content of the other class is too small. Since $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ and we can let ϵ be arbitrarily small, we achieve the fluid limit of the partitioned model for the original model. Hence, the proof of Theorem 2 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. ∎

3.5. Lower Arrival Rates and Staffing

We now supplement §5 by showing in Figures 1 and 2 of the performance functions in the same $(M_t(0.2)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (0.2, H_2(1,4)/\infty)$ model except that $\bar{\lambda}$ is reduced from 100 to 20. As the scale decreases, the discretization becomes a more and more serious issue. Thus there is a limit to the stabilization that can be achieved with very small scale. Here we increase the number of replications to 5000.

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 formula (20) of the main paper for $\bar{\lambda} = 20$ 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.



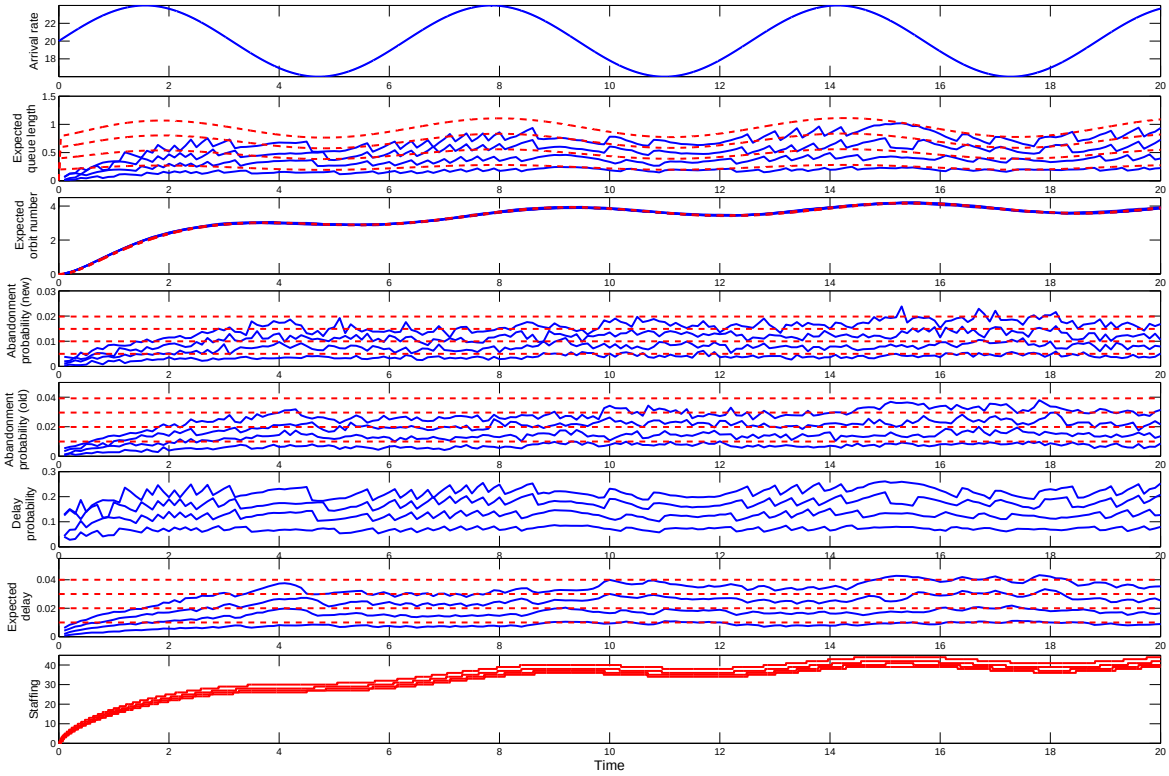
To show the discretization effect, we display the DIS staffing for this model with $\bar{\lambda} = 20$ with high and low waiting-time targets, respectively, in Figures 3 and 4. Figure 3 shows that a difference of 0.1 in the waiting-time target is approximately worth a single server in this context.

For comparison with Figure 3, we also show the DIS staffing in the corresponding model with $\bar{\lambda}$ further reduced to 5 in Figure 5.

Acknowledgement

We thank the National Science Foundation for support: NSF grants CMMI 1362310 (first author)

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 in formula (20) of the main paper for $\bar{\lambda} = 20$ 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.



and CMMI 1066372 and 1265070 (second author). This research began as part of the first author's doctoral dissertation at Columbia University.

References

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.

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.

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

Figure 3 DIS staffing 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 in formula (20) of the main paper for $\bar{\lambda} = 20$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (high QoS) targets ($w = 0.1$, 0.2, 0.3 and 0.4) and DIS-MOL staffing.

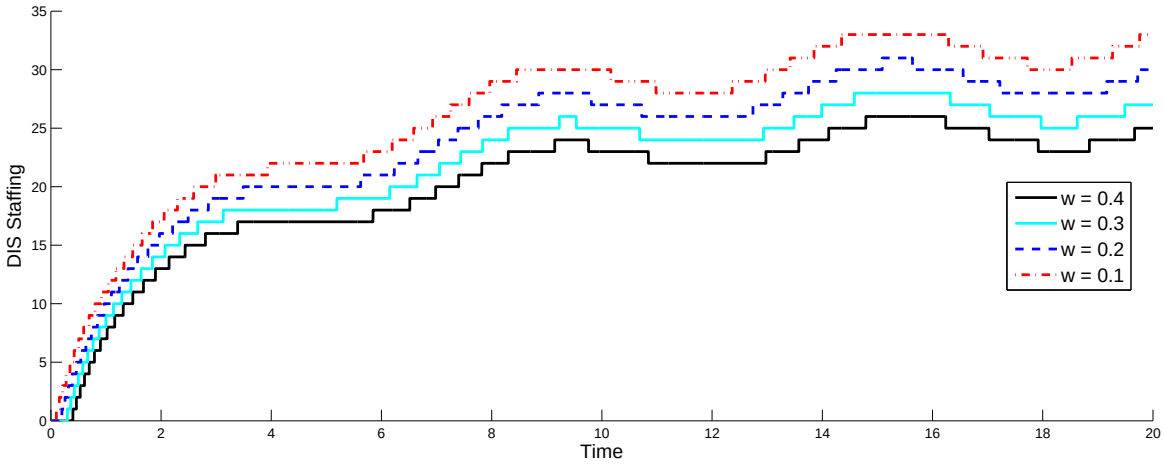
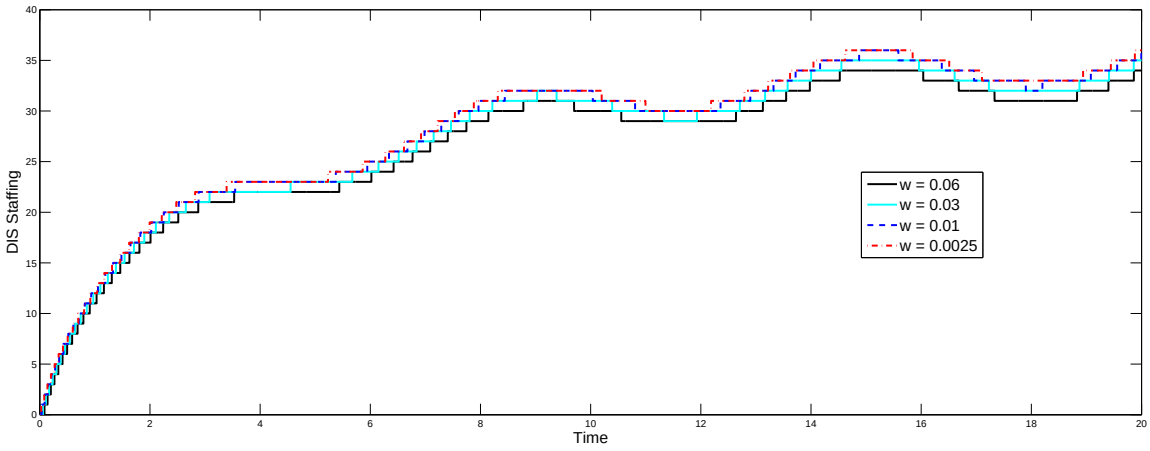


Figure 4 DIS staffing functions in the $(M_t(0.2)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in in formula (20) of the main paper for $\bar{\lambda} = 20$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.0025$, 0.01, 0.03 and 0.06) and DIS-MOL staffing.



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

Figure 5 DIS staffing 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 in formula (20) of the main paper for $\bar{\lambda} = 5$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (high QoS) targets ($w = 0.1$, 0.2, 0.3 and 0.4) and DIS-MOL staffing.