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

Yunan Liu and Ward Whitt

Department of Industrial Engineering
North Carolina State University
Raleigh, NC 27695

Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027

yliu48@ncsu.edu, ww2040@columbia.edu

April 21, 2016

Abstract

Analytical approximations are developed to determine 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 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

1 Overview

This is an appendix to the main paper [9], providing additional supplementary material. In §2 we present additional results supplementing the DIS approximation in §2 of [9]. In §3 we give the proof of the asymptotic effectiveness (Theorem 2 in the main paper). In §4 we elaborate on details of our estimation procedures for the performance functions. In §5 we present additional results from simulation experiments, supplementing §5 and §6 of the main paper. In §6 we present additional estimates of the implied empirical Quality of Service (QoS) function, supplementing §5.3 of [9]. Finally, in §7 we examine the index of dispersion for counts of the various arrival processes in the feedback model to see if we see significant deviation from a nonhomogeneous Poisson process, which might degrade the DIS-MOL approximation.

2 Supplement to the DIS Model in §2 of [9]

In this section we present additional material supplementing the DIS approximation in §2 of the main paper [9]. We first give additional results in special cases.

2.1 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 [2, 4, 7]. For periodic arrival processes, we can simply focus on the dynamic steady state if we start the system at the infinite past.

Theorem 2.1 *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*

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

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*

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

The departure rates from the two service facilities are sinusoidal

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

The arrival rate to the second waiting room is sinusoidal

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

Remark 2.1 (*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 2.2 *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*

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

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

Furthermore, in both cases

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

Proof. First, the offered load formula in (3) within Theorem 3 of [9] can be easily verified by adding up $E[B_1(t)]$ and $E[B_2(t)]$ in Theorem 2 of [9]. So the relative amplitude of $m^*(t)$ and $m(t) \equiv E[B_1(t)]$ are

$$RA(m^*) \quad = \quad \frac{b\,\gamma(S_e)}{a} \quad \text{and} \quad RA(m) = \frac{b\,\gamma(S_e)\,\sqrt{u^2+v^2}}{a(1+p\,\bar{F}(w))}.$$

Therefore, it remains to determine the ratio

$$\begin{aligned} \frac{RA(m)}{RA(m^*)} &= \frac{\sqrt{u^2+v^2}}{1+p\,\bar{F}(w)} \\ &= \frac{\sqrt{1+p^2\bar{F}(w)^2\gamma(S)^2\gamma(U)^2+2p\bar{F}(w)\gamma(S)\gamma(U)\cos[cw+\theta(S)+\theta(U)]}}{1+p\,\bar{F}(w)}. \end{aligned} \tag{2}$$

If S and U are exponentially distributed with rate μ and δ respectively, it is easy to see that

$$\phi_1(S) = \frac{c/\mu}{1+c^2/\mu^2} \quad \text{and} \quad \phi_2(S) = \frac{1}{1+c^2/\mu^2},$$

so that

$$\gamma(S) = \frac{1}{\sqrt{1+c^2/\mu^2}} \quad \text{and} \quad \theta(S) = \arctan\left(\frac{c}{\mu}\right). \tag{3}$$

Similarly, we have

$$\gamma(U) = \frac{1}{\sqrt{1+c^2/\delta^2}} \quad \text{and} \quad \theta(U) = \arctan\left(\frac{c}{\delta}\right). \tag{4}$$

Plugging (3) and (4) into (2) yields

$$\begin{aligned} \frac{RA(m)}{RA(m^*)} &= \frac{\sqrt{1+\frac{p^2\bar{F}(w)^2}{(1+c^2/\mu^2)(1+c^2/\delta^2)}+\frac{2p\bar{F}(w)\cos[cw+\theta(S)+\theta(U)]}{\sqrt{(1+c^2/\mu^2)(1+c^2/\delta^2)}}}}{1+p\,\bar{F}(w)} \\ &< \frac{\sqrt{1+p^2\bar{F}(w)^2+2p\bar{F}(w)}}{1+p\,\bar{F}(w)} = 1, \end{aligned}$$

when the condition in (i) holds. The time lag

$$\begin{aligned}
\Delta PH(m,m^*) &= \arctan\left(\frac{u}{v}\right) - \theta(S_e) \\
&= \arctan\left(\frac{\sin\left[\arctan\left(\frac{c}{\mu}\right)\right] + \frac{p\bar{F}(w)\sin\left[cw+2\arctan\left(\frac{c}{\mu}\right)+\arctan\left(\frac{c}{\delta}\right)\right]}{\sqrt{(1+c^2/\mu^2)(1+c^2/\delta^2)}}}{\cos\left[\arctan\left(\frac{c}{\mu}\right)\right] + \frac{p\bar{F}(w)\cos\left[cw+2\arctan\left(\frac{c}{\mu}\right)+\arctan\left(\frac{c}{\delta}\right)\right]}{\sqrt{(1+c^2/\mu^2)(1+c^2/\delta^2)}}}\right) - \arctan\left(\frac{c}{\mu}\right) \\
&= \arctan\left(\frac{c + \frac{p\bar{F}(w)\sin\left[cw+2\arctan\left(\frac{c}{\mu}\right)+\arctan\left(\frac{c}{\delta}\right)\right]}{\sqrt{1+c^2/\delta^2}}}{\mu + \frac{p\bar{F}(w)\cos\left[cw+2\arctan\left(\frac{c}{\mu}\right)+\arctan\left(\frac{c}{\delta}\right)\right]}{\sqrt{1+c^2/\delta^2}}}\right) - \arctan\left(\frac{c}{\mu}\right) \\
&\to 0 \quad \text{as} \quad c \to 0.
\end{aligned}$$

If S and U are deterministic, then

$$\gamma(S) = \gamma(U) = 1, \quad \theta(S) = c\,S = \frac{c}{\mu} \quad \text{and} \quad \theta(U) = c\,U = \frac{c}{\delta}.$$

Therefore, from (2) we have

$$\begin{aligned}
\frac{RA(m)}{RA(m^*)} &= \frac{\sqrt{1 + p^2\bar{F}(w)^2 + 2p\bar{F}(w)\cos[c(w+S+U)]}}{1 + p\,\bar{F}(w)} \\
&\le \frac{\sqrt{1 + p^2\bar{F}(w)^2 + 2p\bar{F}(w)}}{1 + p\,\bar{F}(w)} = 1. \quad \blacksquare
\end{aligned}$$

Since $S = 1/\mu$, $S_e \sim Unif(0, 1/\mu)$, which implies that $\theta(S_e) = c/2\mu$. Therefore, the time lag

$$\begin{aligned}
\Delta PH(m,m^*) &= \frac{\sin\left(\frac{c}{2\mu}\right) + p\bar{F}(w)\sin\left(cw + \frac{c}{\mu} + \frac{c}{\delta} + \frac{c}{2\mu}\right)}{\cos\left(\frac{c}{2\mu}\right) + p\bar{F}(w)\cos\left(cw + \frac{c}{\mu} + \frac{c}{\delta} + \frac{c}{2\mu}\right)} - \frac{c}{2\mu} \\
&\to 0 \quad \text{as} \quad c \to 0. \quad \blacksquare
\end{aligned}$$

2.2 Additional Results

We first consider how the $(M_t/GI, GI/s_t+GI, GI)+(GI/\infty)$ model simplifies when the arrival-rate function is constant. When the arrival rate is constant, i.e., $\lambda(t) = \lambda$, the steady-state performance functions can be easily obtained. This analysis entails simple calculations for a five-queue IS network, which in particular simplifies to five IS queues in series. Here we display the results for the simple case in which $G_1 = G_2 = G$ and $F_1 = F_2 = F$.

Corollary 2.1 (*steady state performance of DIS-OL*) *Consider the DIS approximation for the* $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ *model in Theorem 1 of [9] with constant arrival rate* λ, $G_1 = G_2 = G$, $F_1 = F_2 = F$ *and delay target* w. *The steady-state (as* $t \to \infty$*) numbers of customers in the waiting rooms, in the service facilities, and in the orbit room,* $Q_1(\infty)$, $B_1(\infty)$, $O(\infty)$, $Q_2(\infty)$, $B_2(\infty)$, *are independent Poisson random variables with means*

$$\begin{aligned}
E[Q_1(\infty)] &= \lambda\,E[T], \\
E[B_1(\infty)] &= \bar{F}(w)\,\lambda\,E[S], \\
E[O(\infty)] &= p\,\bar{F}(w)\,\lambda\,E[U], \\
E[Q_2(\infty)] &= p\,\bar{F}(w)\,\lambda\,E[T], \\
E[B_2(\infty)] &= p\,\bar{F}^2(w)\lambda\,E[S],
\end{aligned}$$

where $T \equiv A \wedge w$. Thus, $X(\infty)$, the steady-state total number of customers in the system is a Poisson random variable with a mean

$$\begin{aligned} E[X(\infty)] &= E[Q_1(\infty)] + E[Q_2(\infty)] + E[B_1(\infty)] + E[B_2(\infty)] \\ &= \lambda \left(1 + p\bar{F}(w)\right) \left(E[T] + \bar{F}(w)E[S]\right). \end{aligned}$$

If the system is initially in steady state, the processes counting the numbers of customers abandoning from waiting room 1 and 2 are Poisson processes with rates

$$\alpha_1 = \lambda\, F(w) \quad \textit{and} \quad \alpha_2 = \lambda\, p\, \bar{F}(w)\, F(w).$$

The processes counting the numbers of customers entering service facility 1 and 2 are Poisson processes with rates

$$\beta_1 = \lambda\, \bar{F}(w) \quad \textit{and} \quad \beta_2 = \lambda\, p\, \bar{F}^2(w).$$

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

$$\sigma_1 = \lambda\, \bar{F}(w) \quad \textit{and} \quad \sigma_2 = p\, \lambda\, \bar{F}^2(w).$$

The process counting the numbers of customers entering the second waiting room is a Poisson process with rate function $\lambda_F = p\, \lambda\, \bar{F}(w)$.

As discussed in [1] and [3], 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. As before, we ignore the fact that these arrival rate functions are necessarily negative on part of the domain. Assuming that the approximations are used with proper care, they can still be very useful.

Theorem 2.3 *Consider the DIS approximation for the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model specified above with $G_1 = G_2 = G$, $F_1 = F_2 = F$, starting in the distant past with specified delay target $w \geq 0$ and with quadratic arrival-rate function $\lambda(t) = a + bt + ct^2$. Then $Q_1(t)$, $B_1(t)$, $O(t)$, $Q_2(t)$ and $B_2(t)$ are independent Poisson random variables having quadratic means*

$$\begin{aligned} E[Q_1(t)] &= E[T](\lambda(t - E[T_e]) + c\,Var(T_e)), \\ E[B_1(t)] &= \bar{F}(w)\, E[S]\ (\lambda(t - w - E[S_e]) + c\,Var(S_e)), \\ E[O(t)] &= p\,\bar{F}(w)\, E[U]\ (\lambda(t - w - E[S] - E(U_e)) + c(Var(S) + Var(U_e))), \\ E[Q_2(t)] &= p\,\bar{F}(w)\, E[T]\ (\lambda(t - w - E[S] - E[U] - E[T_e]) + c(Var(U) + Var(S) + Var(T_e))), \\ E[B_2(t)] &= p\,\bar{F}(w)^2\, E[S]\, (\lambda(t - 2w - E[S] - E[U] - E[S_e]) + c(Var(U) + Var(S) + Var(S_3))). \end{aligned}$$

We next 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 in the next period $[0, T]$ (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), a long line of waiting customers may have already formed outside the door. Since the system is initially

non-empty, analogously the DIS approximation has an initially non-empty waiting room 1, i.e., there are n customers in the first waiting room at time 0. The next Theorem characterizes the OL function of this case.

Let $S_k^{(i)}, A_k^{(i)}$ be the service and patience time of the kth customer in his ith visit, let U_k be the orbit time of the kth customer, $1 \le k \le n$, $i = 1, 2$. Let $R_k \equiv 1$ if the kth customer revisits the system and let $R_k \equiv 0$ otherwise.

Theorem 2.4 (*the initially non-empty DIS model*) *Consider the DIS approximation for the* $(M_t/GI, GI/s_t+GI, GI) + (GI/\infty)$ *model model with customer feedback,* $G_1 = G_2 = G$, $F_1 = F_2 = F$, *and delay target* $w \ge 0$ *specified above, starting at time* 0. *Suppose initially there are* n *customers waiting in the queue with elapsed waiting times* $w_1 \le w_2 \le \cdots \le w_n$. *The approximation makes* $W(t) = w$ *with probability* 1 *and the probability of abandonment* $F(w)$ *for all arrivals by letting* $n(w)$ *customers entering service at* $t = 0$, *where* $n(w) \equiv \inf\{k \ge 1 : w_k \ge w\}$. *Moreover, the total numbers of customers in the waiting rooms, in the service facilities, and in the orbit room at time* t, $\tilde{Q}_1(t)$, $\tilde{B}_1(t)$, $\tilde{O}(t)$, $\tilde{Q}_2(t)$, $\tilde{B}_2(t)$, *can be written as sums of independent random variables:*

$$\begin{aligned}
\tilde{Q}_1(t) &= Q_1(t) + Q_1^1(t), \quad \tilde{B}_1(t) = B_1(t) + B_1^0(t) + B_1^1(t), \quad \tilde{O}(t) = O(t) + O^0(t) + O^1(t), \\
\tilde{Q}_2(t) &= Q_2(t) + Q_2^0(t) + Q_2^1(t), \quad \tilde{B}_2(t) = B_2(t) + B_2^0(t) + B_2^1(t),
\end{aligned}$$

where $Q_1(t)$, $B_1(t)$, $O(t)$, $Q_2(t)$ *and* $B_2(t)$ *are Poisson random variables with mean in Theorem 1 of [9],* $B_1^0(t)$, $O(t)^0$, $Q_2^0(t)$ *and* $B_2^0(t)$ *are Binomial random variables with parameters*

$$\begin{aligned}
n^{B_1} &= n^O = n^{Q_2} = n^{B_2} = n(w), \\
p^{B_1} &= \bar{G}(t), \quad p^O = pP(S^{(1)} < t, S^{(1)} + U > t) = p \int_0^t \bar{H}(t-x) dG(x), \\
p^{Q_2} &= p\, P(S^{(1)} + U < t, S^{(1)} + U + A^{(2)} > t) = p \int_0^t \bar{F}(t-x) dG * H(x), \\
p^{B_2} &= p\, P(S^{(1)} + U + A^{(2)} < t, S^{(1)} + U + A^{(2)} + S^{(2)} > t) = p \int_0^t \bar{G}(t-x) dG * H * F(x),
\end{aligned}$$

and $Q_1^1(t), B_1^1(t)$, $O(t)^1$, $Q_2^1(t)$ *and* $B_2^1(t)$ *can be expressed as sums of independent indicator random variables, in particular,*

$$\begin{aligned}
Q_1^1(t) &= \sum_{k=n(w)+1}^{n} 1_{\{w_k + t \le w,\, A_k^{(1)} > t + w_k\}}, \\
B_1^1(t) &= \sum_{k=n(w)+1}^{n} 1_{\{w_k + t > w,\, A_k^{(1)} > w,\, S_k^{(1)} > t - (w - w_k)\}}, \\
O^1(t) &= \sum_{k=n(w)+1}^{n} 1_{\{w_k + t > w,\, A_k^{(1)} > w,\, S_k^{(1)} < t - (w - w_k),\, U_k + S_k^{(1)} > t - (w - w_k),\, R_k = 1\}}, \\
Q_2^1(t) &= \sum_{k=n(w)+1}^{n} 1_{\{w_k + t > w,\, A_k^{(1)} > w,\, S_k^{(1)} < t - (w - w_k),\, t - (2w - w_k) < U_k + S_k^{(1)} < t - (w - w_k),\, U_k + S_k^{(1)} + A_k^{(2)} > t - (w - w_k),\, R_k = 1\}}, \\
B_2^1(t) &= \sum_{k=n(w)+1}^{n} 1_{\{w_k + t > w,\, A_k^{(1)} > w,\, S_k^{(1)} < t - (w - w_k),\, U_k + S_k^{(1)} < t - (w - w_k),\, A_k^{(2)} > w,\, U_k + S_k^{(1)} + S_k^{(2)} > t - (2w - w_k),\, R_k = 1\}}.
\end{aligned}$$

Remark 2.2 (*interpretations of random variables in Theorem* **??**) *In order to stabilize the potential waiting time for $t > 0$, we let $n(w)$ customers (who have waited longer than target w by $t = 0$) enter service immediately at time 0. According to the same policy, the rest $n - n(w)$ initial customers have to wait for extra time because their elapsed waiting time $w_k < w$ for $n(w) + 1 \le k \le n$. They will enter service after time 0 at different moments $t_k \equiv w - w_k$. It is easy to see that $B_1^0(t)$ ($O^0(t)$, $Q_2^0(t)$ and $B_2^0(t)$) denotes the number of initial $n(w)$ customers who entered service at time 0 and are in the first service facility (orbit, second waiting room and the second service facility) at time t. Similarly, $Q_1^1(t)$ ($B_1^1(t)$, $O(t)^1$, $Q_2^1(t)$ and $B_2^1(t)$) denotes the number of initial $n - n(w)$ customers who entered service after time 0 and are in the first waiting room (first service facility, orbit, second waiting room and the second service facility) at time t.*

3 Proof of Theorem 2 in [9]

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 $s_{i,n}(t) \equiv \lceil n s_i(t) \rceil$ 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 [5] for a discussion of the connection between the DIS model and the fluid model and §10 of [5] for the explicit performance functions achieving the waiting-time target w.

Just as in [5], 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 3.1 (*DIS staffing stabilizes the waiting time in the fluid model with feedback*) *The DIS staffing in §2 of [9] 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 of [9], 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 [7], we can apply the established FWLLN in [6] to obtain the desired FWLLN. Since the waiting time target is w, we can use §10 of [5] 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

$$D_n^{(1,2)}(t) = \sum_{k=1}^{D_n^{(1)}(t)} 1_{\{X_{n,1,k}=2\}}, \quad t \geq 0, \tag{5}$$

and the associated scaled version as

$$\bar{D}_n^{(1,2)}(t) = \sum_i \bar{Z}_n(t) \circ \bar{D}_n^{(1)}(t), \quad t \geq 0, \tag{6}$$

where $\circ$ is the composition function and

$$\bar{Z}_n(t) \equiv \frac{1}{n} \sum_{k=1}^{\lfloor nt \rfloor} 1_{\{X_{n,1,k}=2\}} \Rightarrow pt \quad \text{in} \quad D \tag{7}$$

We now apply the continuous mapping theorem in §3.4 of [11] for the continuous composition functions appearing in (6), see Theorem 13.2.1 of [11], 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 [10] 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 [5]. 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 3.2 (*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*

$$B(t,x) = B_1(t,x) + B_2(t,x) \quad \textit{and} \quad Q(t,x) = Q_1(t,x) + Q_2(t,x) \quad \textit{for all} \quad t, x \tag{8}$$

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 [5] 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

$$s_i(t, w+\epsilon) - \delta < s_i(t,w) < s_i(t, w-\epsilon) + \delta \quad \text{for} \quad a \le t \le b \quad \text{and} \quad i = 1, 2. \tag{9}$$

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 $a > w > 0$,

$$\sup_{a \le t \le b} \{|\bar{B}_i^n(t,w) - s_i(t,w)|\} \Rightarrow 0 \quad \text{as} \quad n \to \infty. \tag{10}$$

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 $n \ge n_0$

$$s_i(t, w+\epsilon) - 2\delta < \bar{B}_i^n(t,w) < s_i(t, w-\epsilon) + 2\delta \quad \text{for} \quad a \le t \le b \quad \text{and} \quad i = 1, 2. \tag{11}$$

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 of [9] 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. ∎

4 Estimation Procedures for the Performance Functions

We now provide extra details about our estimation procedure for the following time-dependent performance functions: (i) the mean potential waiting time, $E[W(t)]$, (ii) the class$-i$ abandonment probability, $P_t^{(i)}(Ab)$, $i = 1, 2$, (iii) the delay probability, $P_t(Delay)$, and (iv) the mean queue length for both the main queue and the orbit queue, $E[Q(t)]$ and $E[O(t)]$. We estimate these performance measures in a time interval $[0, T]$ with $T = 20$. At the j^{th} simulation replication, we periodically generate virtual arrivals (of both classes) at deterministic times $\Delta t, 2\Delta t, 3\Delta t, \ldots, T$, with $\Delta t = 0.1$. These virtual customers have the same patience-time distribution; they abandon as if they were real customers, but they will not be removed from the queue if they abandon. They still wait for their turn to enter service so that we can record their virtual waiting times (although we won't let them enter so they don't affect the dynamics in the service facility). We use indicator variables $\xi_{d,j}(k)$ ($\xi_{a,j}^{(i)}(k)$) to indicate if the k^{th} (type$-i$) virtual customer of the j^{th} replication is delayed (abandons), $i = 1, 2$, $k = 1, \ldots, T/\Delta t$. We use $W_j(k)$ to record the virtual waiting time of the k^{th} virtual arrival, that is $E_j(k) - A_j(k)$, where $E_j(k)$ is the time this virtual customer "enters" service and $A_j(k) = k \cdot \Delta t$ is its arrival time. For the queue-length processes, we sample the continuous-time queue-length process at discrete time points $\Delta t, 2\Delta t, \ldots, T$, $Q_j(k)$ and $O_j(k)$. Here we make sure to exclude (not counting) those virtual arrivals in queue.

We generate $N = 2000$ independent replications to estimate the delay probability, abandonment probability, mean potential waiting time and mean queue lengths. Specifically, for at $t_k \equiv k\Delta t$, $k = 1, \ldots, T/\Delta T$, we approximate $E[W(t_k)]$, $P_{t_k}^{(i)}(Ab)$, $P_{t_k}(Delay)$, $E[Q(t_k)]$ and $E[O(t_k)]$ with

$$\frac{1}{N}\sum_{j=1}^{N} W_j(k), \quad \frac{1}{N}\sum_{j=1}^{N} \xi_{a,j}^{(i)}(k), \quad \frac{1}{N}\sum_{j=1}^{N} \xi_{d,j}(k), \quad \frac{1}{N}\sum_{j=1}^{N} Q_j(k) \quad \text{and} \quad \frac{1}{N}\sum_{j=1}^{N} O_j(k).$$

5 Additional Experiments

In this section we supplement the main paper by presenting additional results from simulation experiments. We start in §5.1 by considering models with higher time variability in arrival rates (represented by bigger relative amplitude r). In §5.2 by considering a variant of the main example in §5 of [9] with much smaller external arrival rate, with $\bar{\lambda}$ reduced from 100 to 20. In §5.4 we consider a variant of the main example in §5 of [9] with more balanced mean service times, now having $E[S_2] = 2.0$ instead of $E[S_2] = 5.0$ (with $E[S_1] = 1$). In §5.5 we consider an example with two feedback opportunities, as shown in Figure 8 of [9]. In all examples we use the arrival rate function

$$\lambda(t) = \bar{\lambda}(1 + r\sin(t)) = 100(1 + r\sin(t)), \quad t \geq 0, \tag{12}$$

for average arrival rate $\bar{\lambda}$ and relative amplitudes r, denoted by $M_t(r)$. We usually let $\bar{\lambda} = 100$ and $r = 0.2$.

5.1 Larger Relative Amplitude

We now supplement §5 by showing in Figures 1, 2, 3 and 4 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 the relative amplitude r is increased from 0.2 to 0.5 and 0.8. As r increases, both the arrival rate and the staffing function (DIS and DIS-MOL) become more variable in time. Figures 1 and 2 show that DIS and DIS-MOL staffing functions continue to work well for $r = 0.5$. However, Figures 3 and 4 show significant performance degradations (with unstable delays) when $r = 0.8$.

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 (12) for $\bar{\lambda} = 100$ and $r = 0.5$, 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.

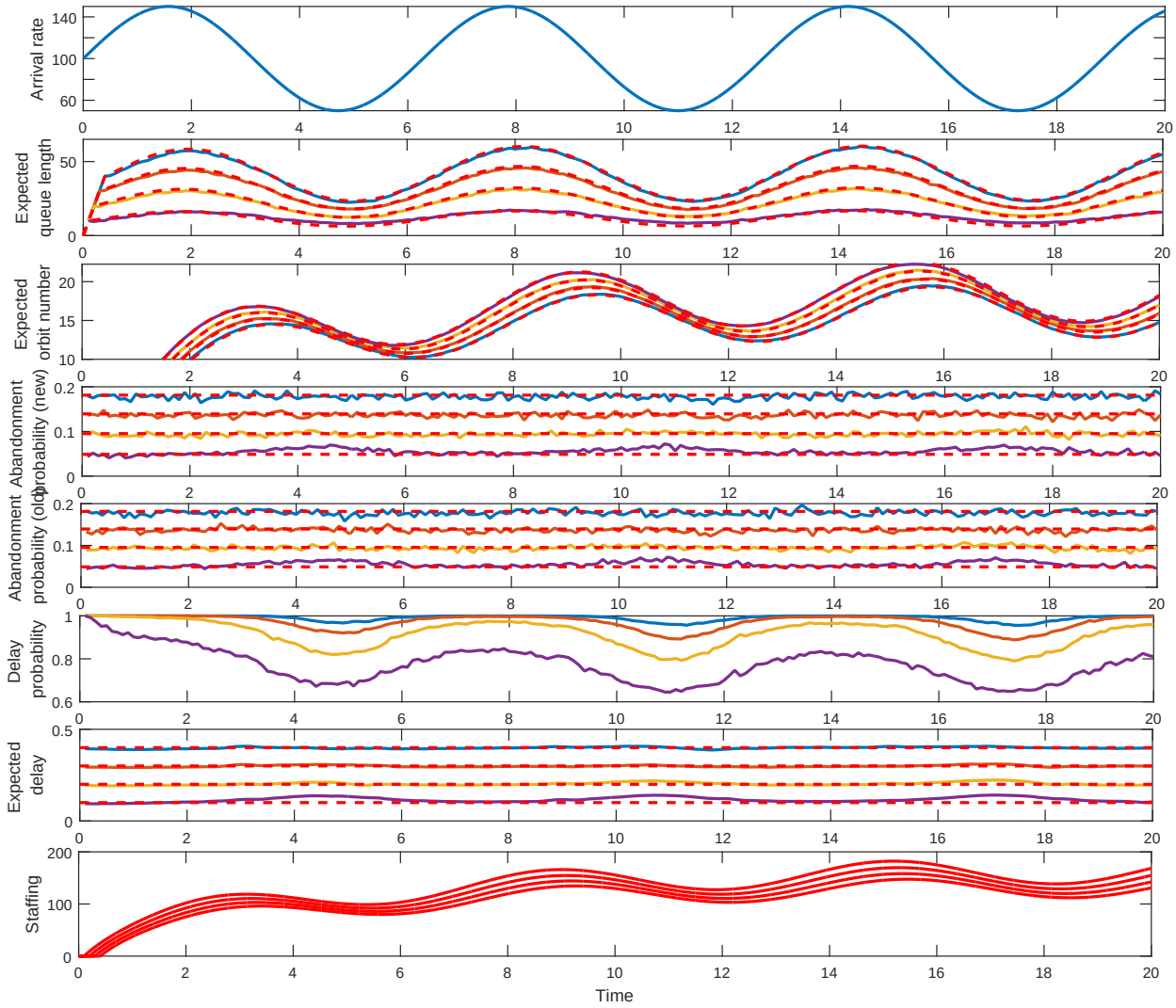
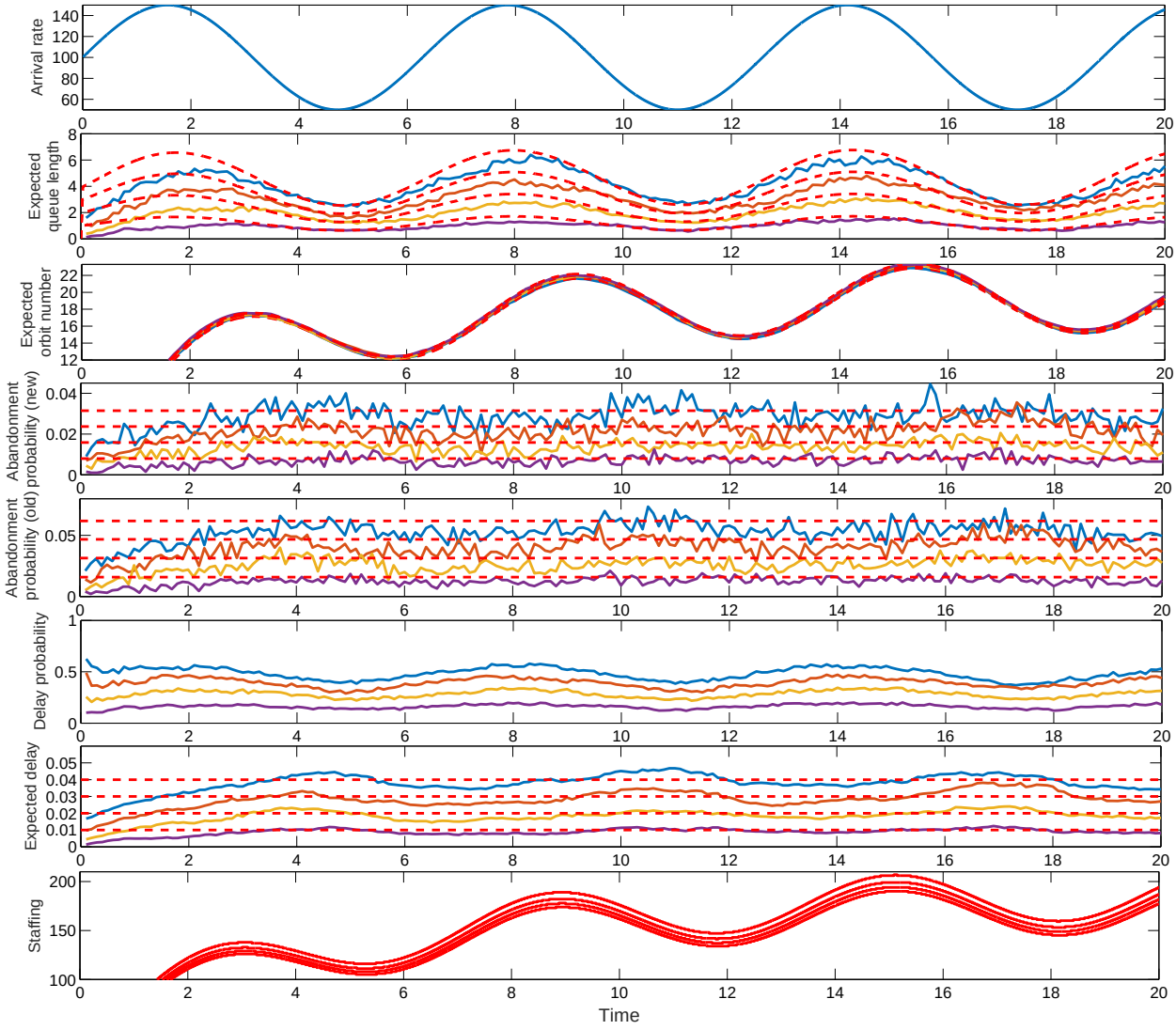


Figure 2: Performance functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.5$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing.

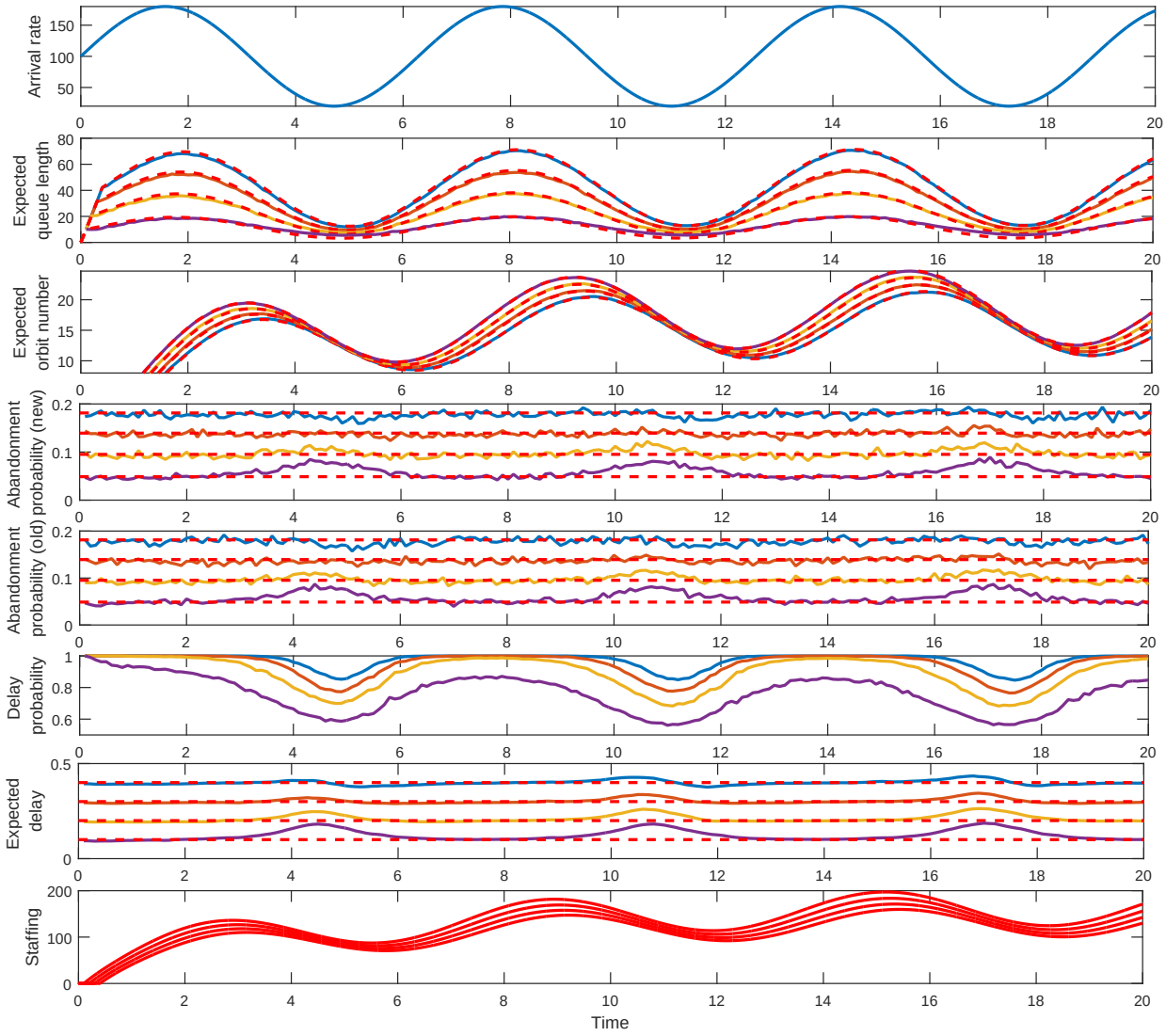


5.2 Lower Arrival Rates and Staffing

We now supplement §5 by showing in Figures 5 and 6 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.

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 7 and 8. Figure 7 shows that a difference of 0.1 in the waiting-time target is approximately worth a single server in this context.

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 (12) for $\bar{\lambda} = 100$ and $r = 0.8$, 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.



For comparison with Figure 7, we also show the DIS staffing in the corresponding model with $\bar{\lambda}$ further reduced to 5 in Figure 9.

5.3 Performance in Model with Non-Exponential Patience Distributions

We now supplement §5 of [9] by showing in Figures 10 and 11 the analog of Figures 2 and 3 for the same $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty)$ model with Bernoulli feedback after a random delay in an IS orbit queue, but with non-exponential patience distributions. Here all distributions in the model are $H_2(m, 4)$. Otherwise the parameters are the same as before. Figures 10 and 11 show excellent performance just as in §5 of [9].

Figure 4: 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 (12) for $\bar{\lambda} = 100$ and $r = 0.8$, 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.

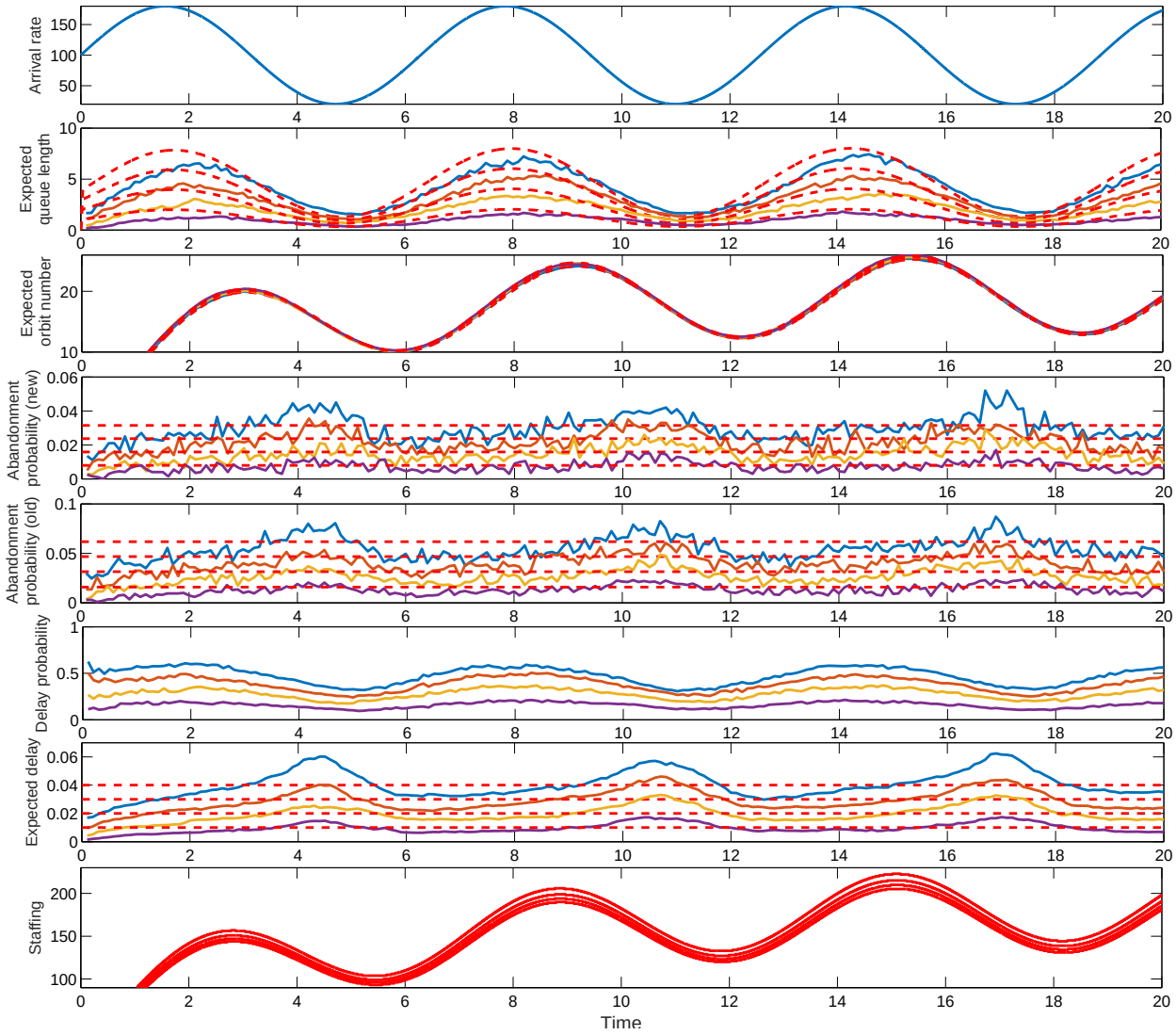


Figure 5: Performance functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 20$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (low QoS) targets ($w = 0.10$, 0.20, 0.30 and 0.40) and simple DIS staffing.

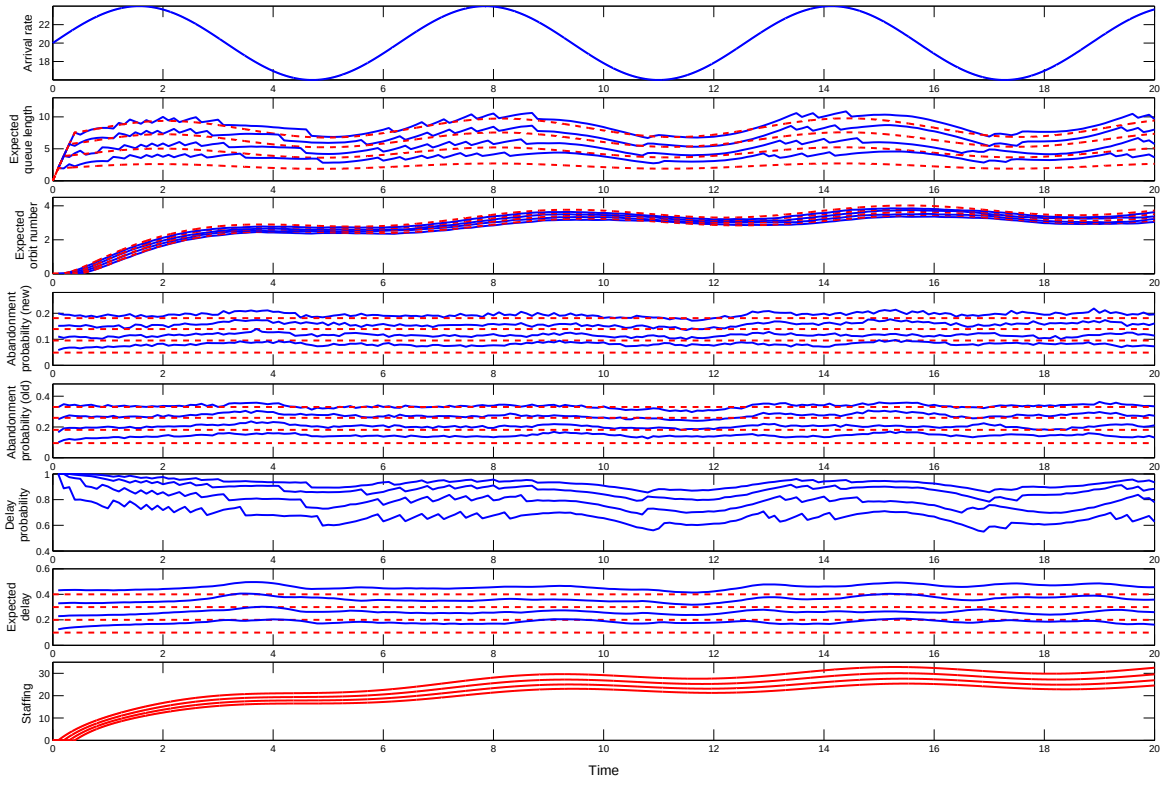


Figure 6: Performance functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 20$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing.

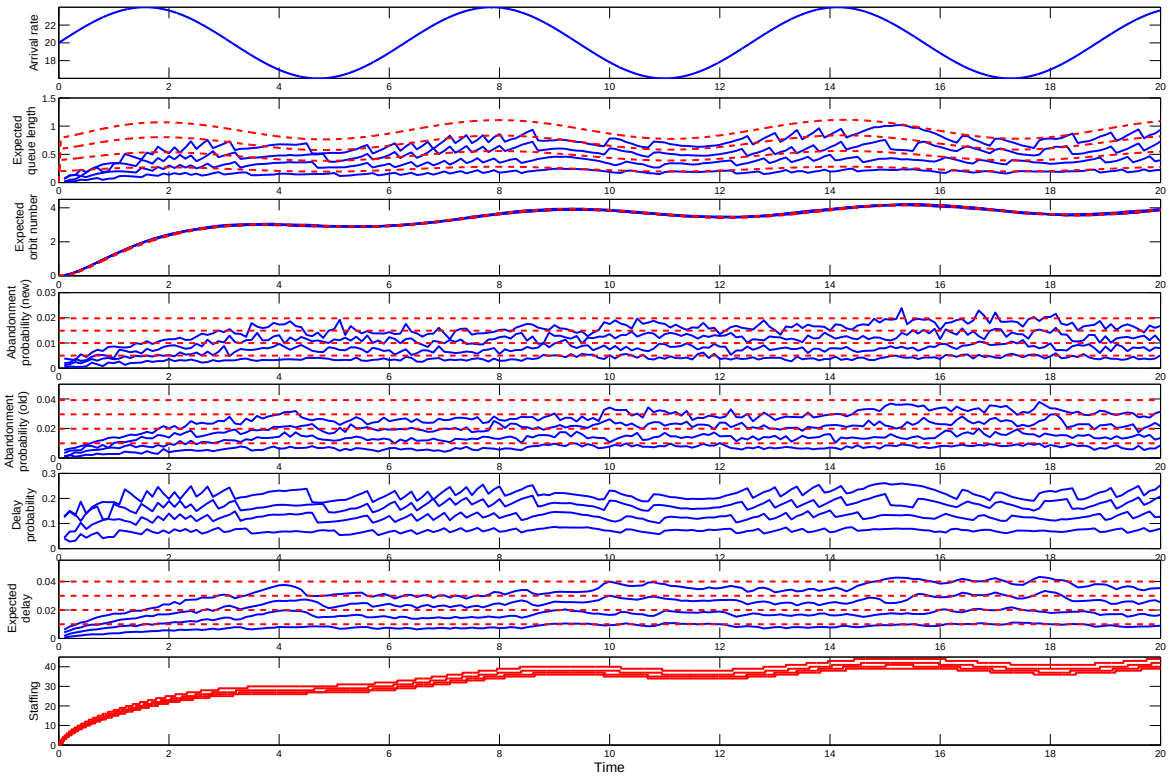


Figure 7: DIS staffing functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 20$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (high QoS) targets ($w = 0.1$, 0.2, 0.3 and 0.4) and DIS-MOL staffing.

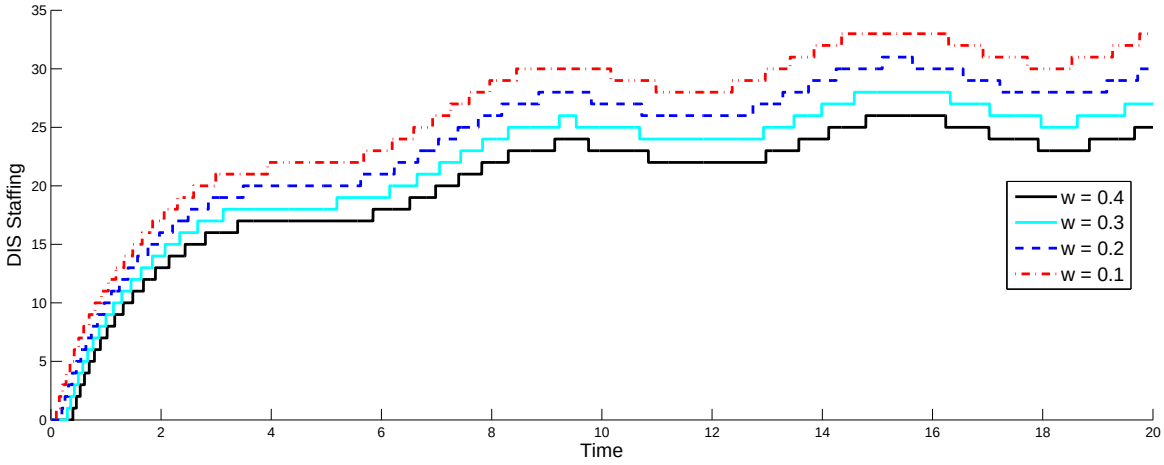


Figure 8: DIS staffing functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 20$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.0025$, 0.01, 0.03 and 0.06) and DIS-MOL staffing.

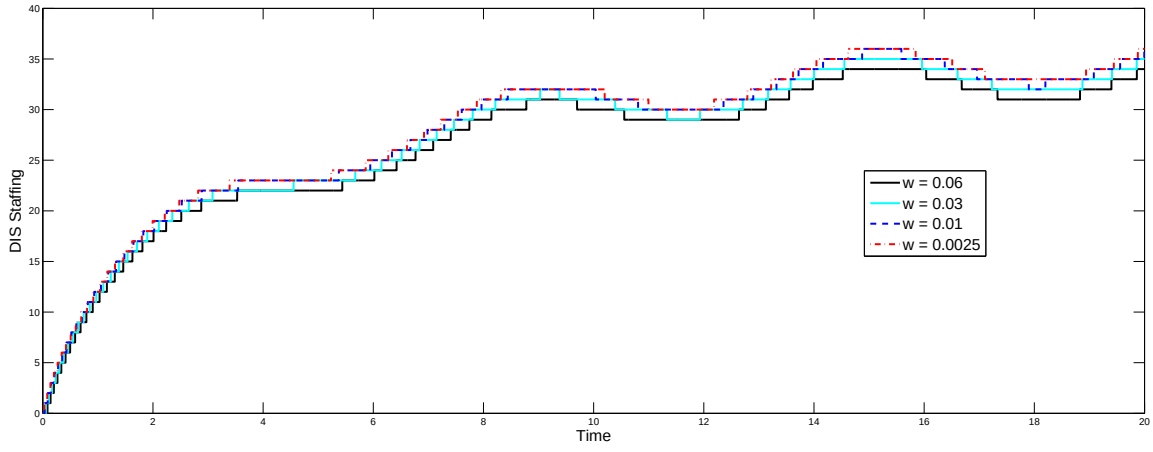


Figure 9: DIS staffing functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 5$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (high QoS) targets ($w = 0.1$, 0.2, 0.3 and 0.4) and DIS-MOL staffing.

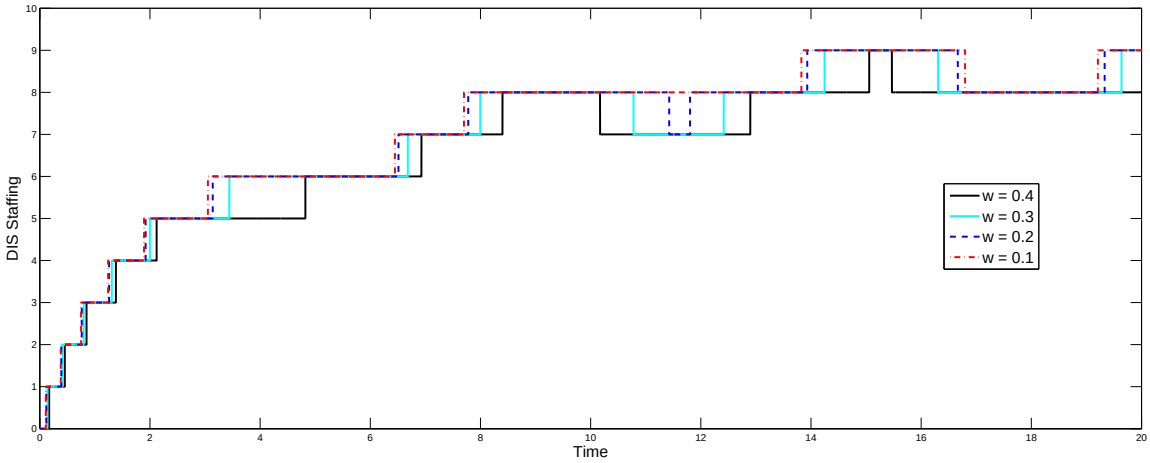


Figure 10: Performance functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{H_2(2,4), H_2(1,4)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of high waiting-time (low QoS) targets ($w = 0.10$, 0.20, 0.30 and 0.40) and simple DIS staffing.

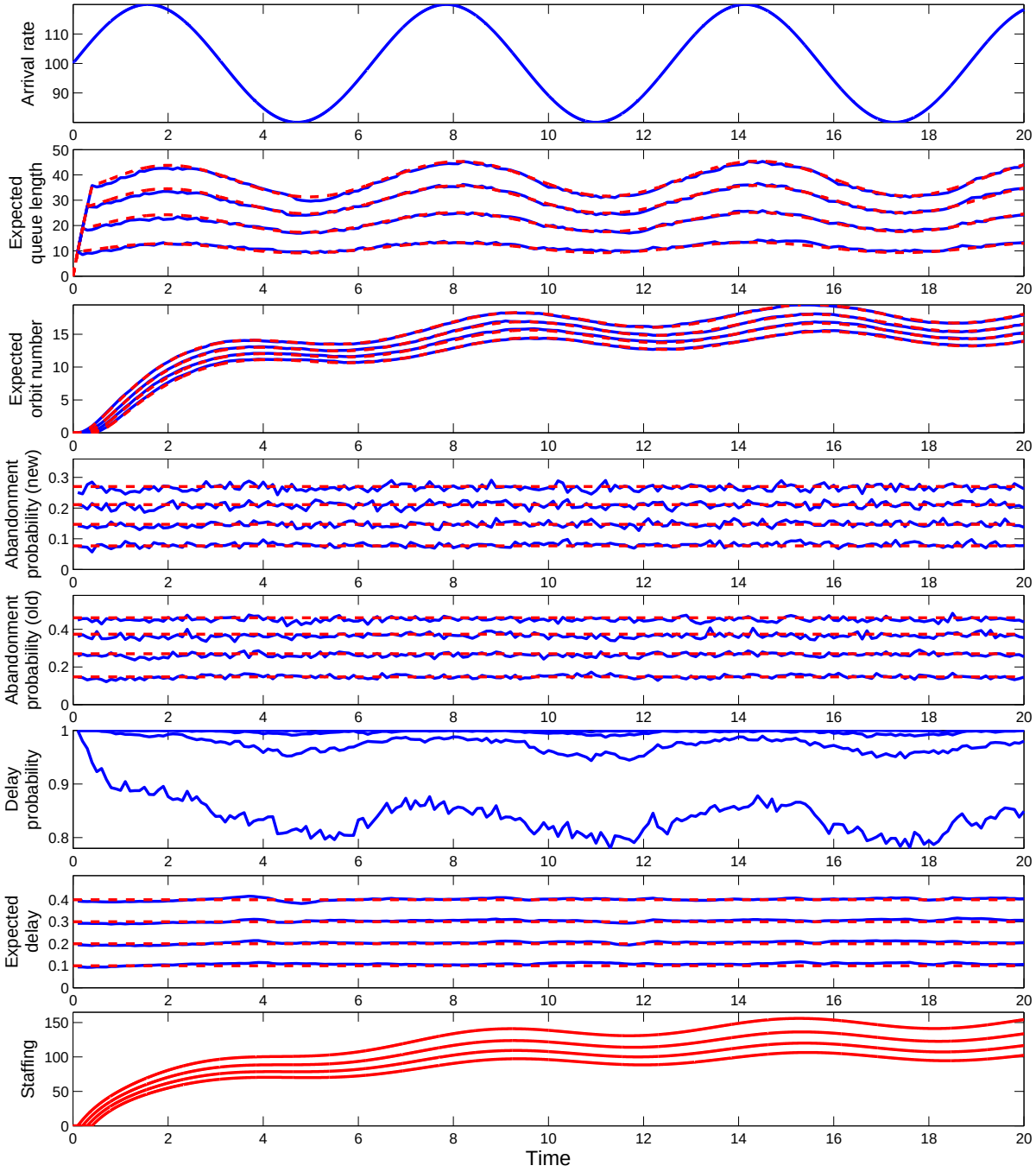
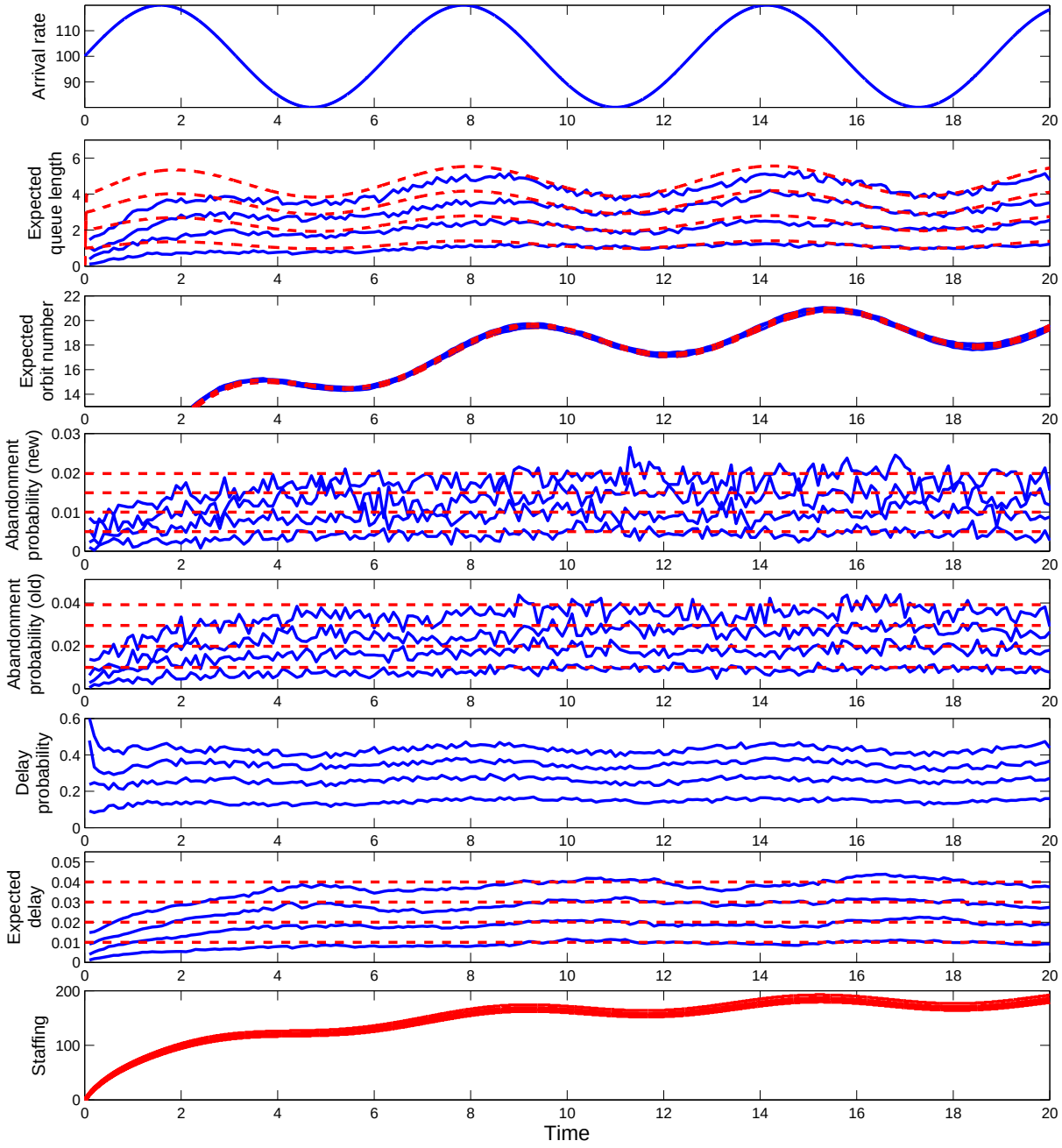


Figure 11: Performance functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{H_2(2,4), H_2(1,4)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing.



5.4 Performance in Model with Similar Mean Service Times

We now consider a variant of 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 in Figures 2 and 3 except the mean service time for the fedback customers is $E[S_2] = 2$. The feedback probability is then increased to 0.5, so that the two contributions to the overall offered load, by new arrivals and fedback customers, are nearly equal. Figures 12 and 13 show the performance functions for the cases of high and low waiting-time targets, respectively.

Figure 12: 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 (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.5$, $E[S_2] = 0.5$ 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.

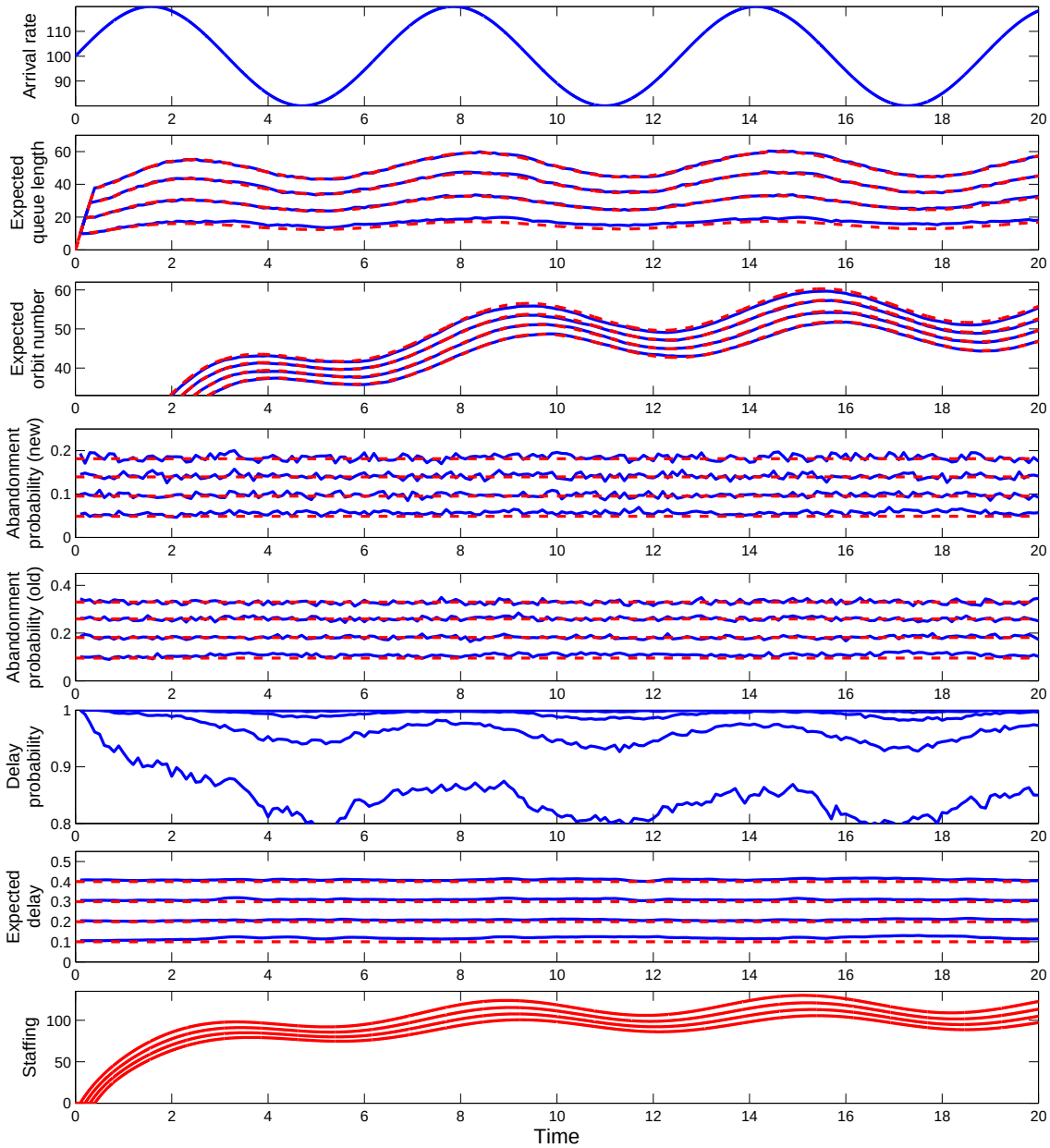
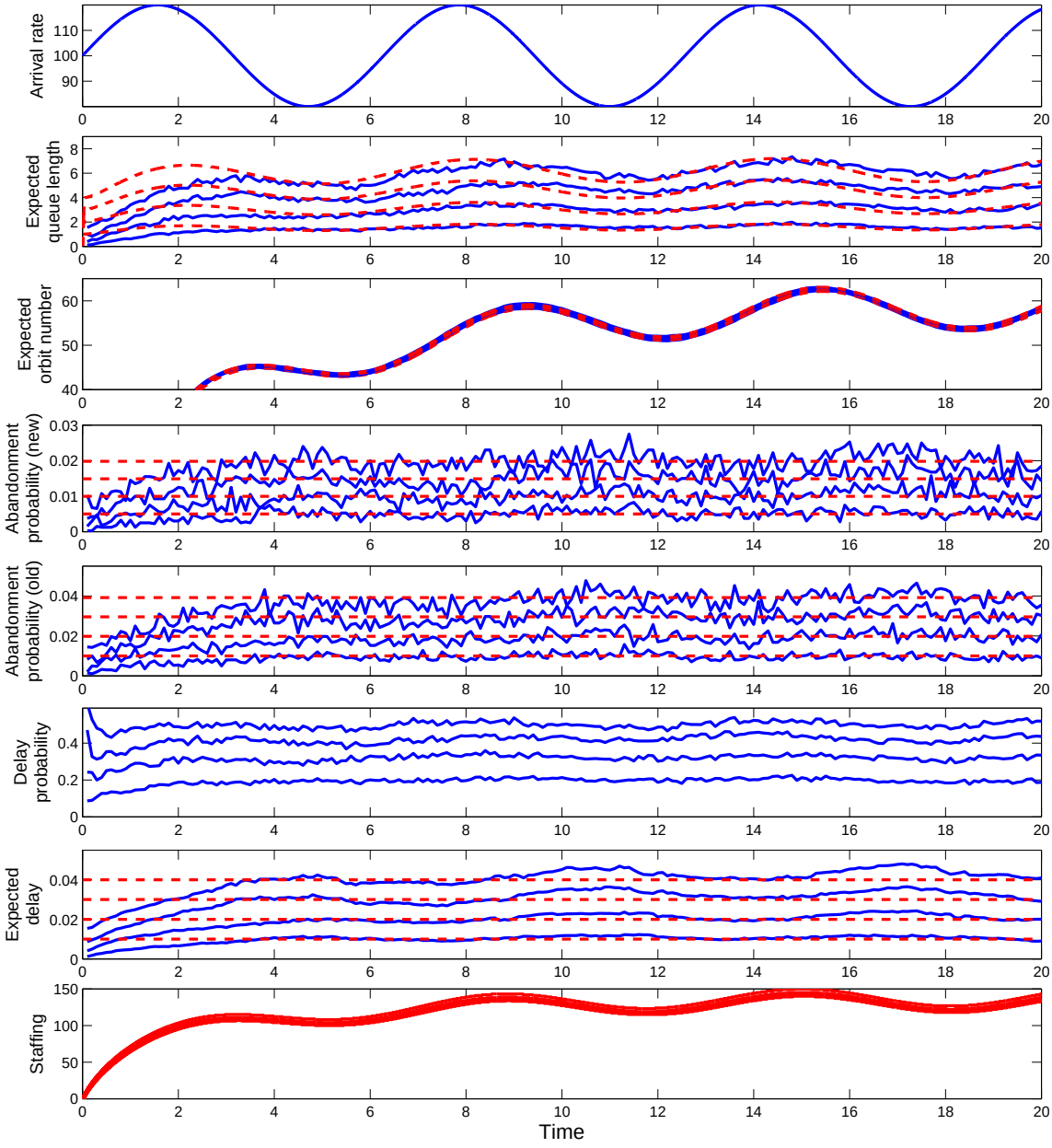


Figure 13: Performance functions in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.5$, $E[S_2] = 0.5$ and an IS orbit queue: four cases of low waiting-time (high QoS) targets ($w = 0.01$, 0.02, 0.03 and 0.04) and DIS-MOL staffing.



5.5 Performance with Two Feedback Opportunities

We now consider an example in which there are two feedback opportunities, as in Figure 8 in §6.3. As before, we let all service-time distributions be hyperexponential $H_2(m, 4)$, but with different means, and all patience distributions be exponential, $M(m)$. Specifically, the three service-time means are $m_1 = 1.0, m_2 = 10/6, m_3 = 2.0$ and the three patience means are $m_1 = 2.0, m_2 = 1.0, m_3 = 10/8$. Just as in §5, the arrival process is an NHPP with sinusoidal arrival-rate function as in (5) of the main paper for $\bar{\lambda} = 100$ and $r = 0.2$. The two feedback probabilities are $p_1 = 0.6, p_2 = 0.5$. Figures 14 and 15 show the main performance functions for the cases of high waiting time targets $(0.10, 0.20, 0.30, 0.40)$ with DIS staffing and low waiting time targets $(0.01, 0.02, 0.03, 0.04)$ with DIS-MOL staffing, respectively.

Figure 14: Performance functions in the $(M_t/\{H_2(m,4), H_2(m,4)\}/s_t + \{M(m), M(m)\}) + (H_2(m,4)/\infty) + (H_2(m,4)/\infty)$ model having two delayed customer feedback opportunities, with sinusoidal arrival rate in (5) for $r = 0.2$, mean service times $m_1 = 1.0, m_2 = 10/6, m_3 = 2.0$, mean patience times $m_1 = 2.0, m_2 = 1.0, m_3 = 10/8$ and feedback probabilities $p_1 = 0.6, p_2 = 0.5$: the cases of high waiting-time (low QoS targets) ($w = 0.10$, 0.20, 0.30 and 0.40) and simple DIS staffing.

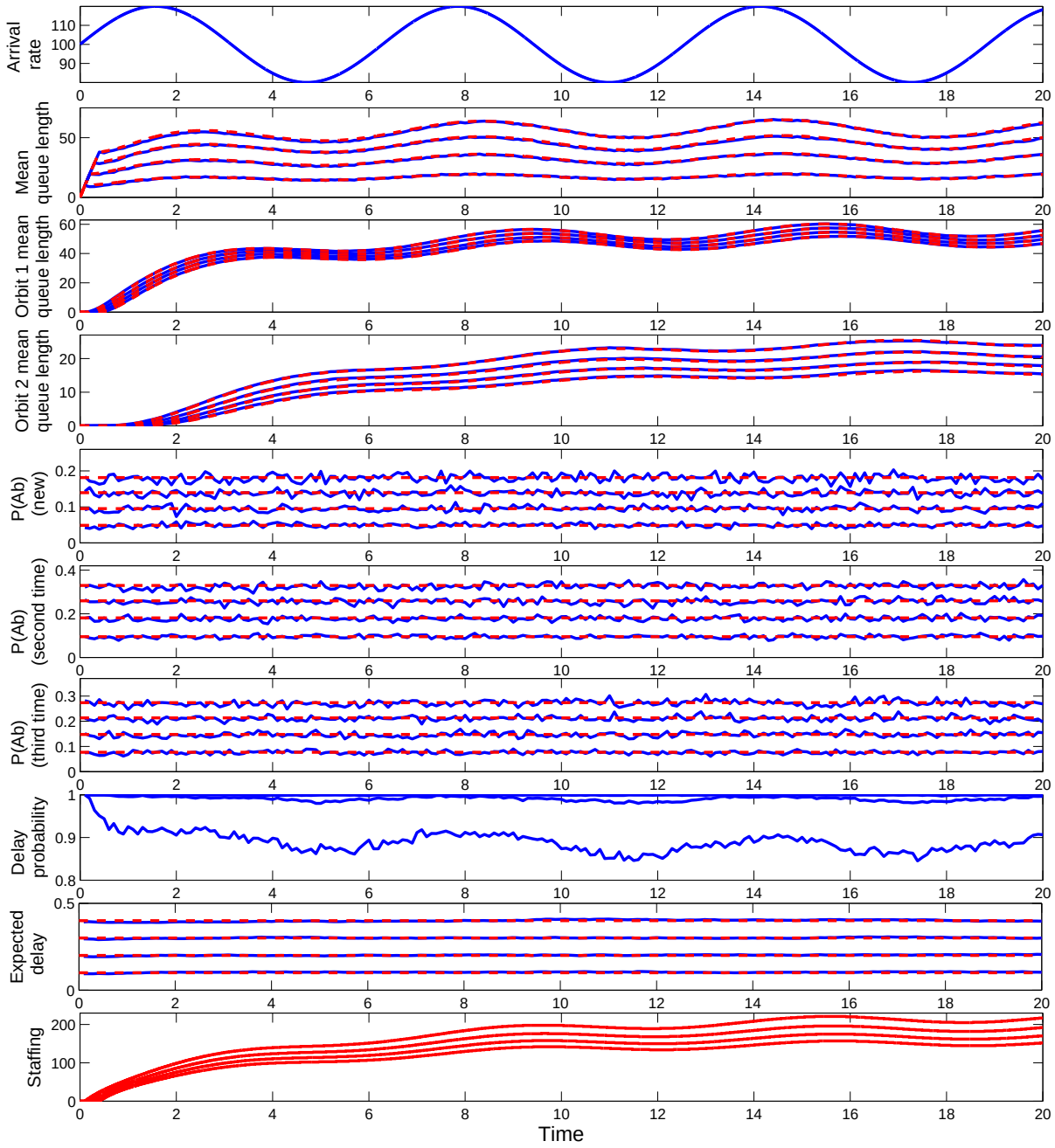
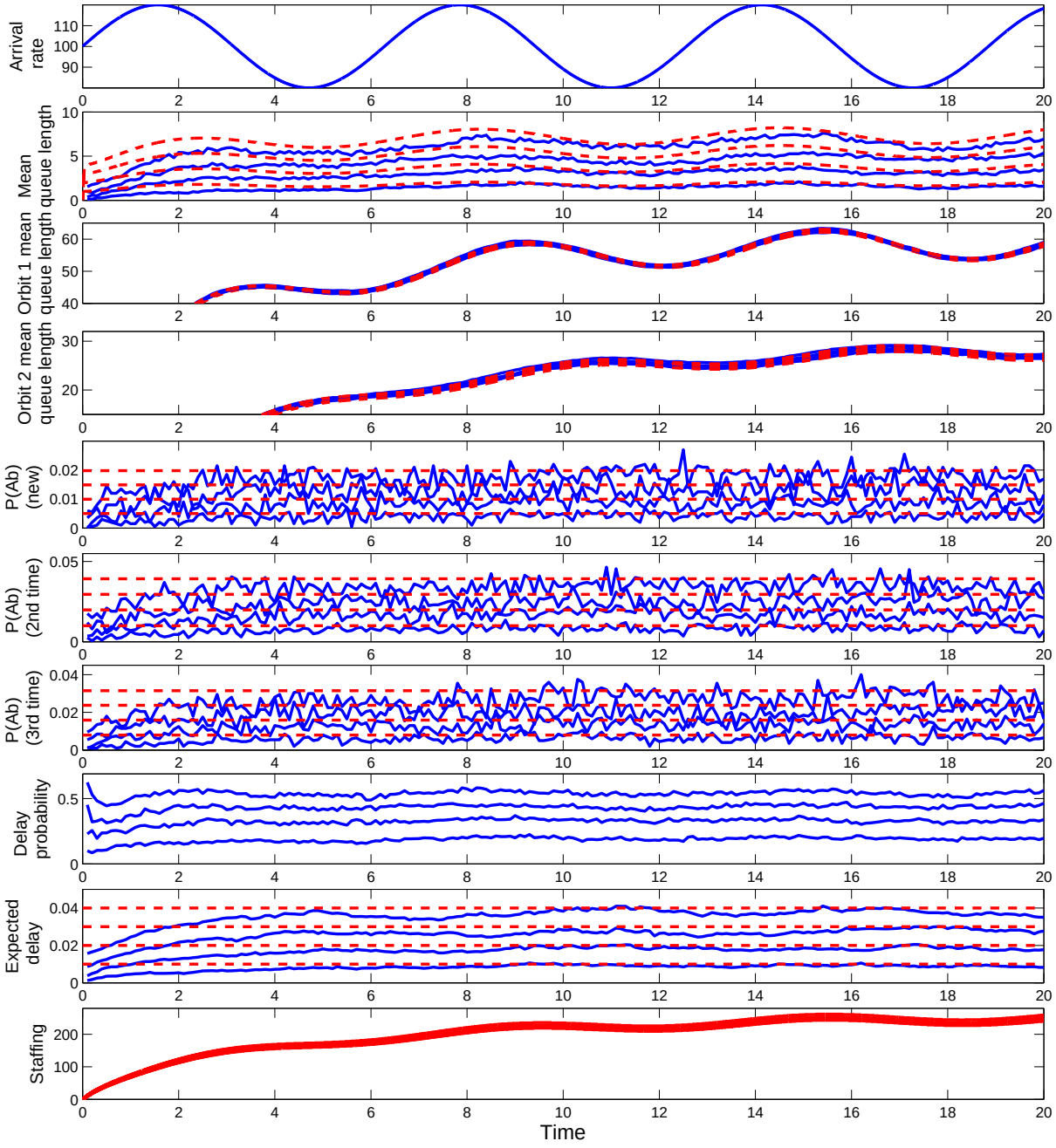


Figure 15: Performance functions in the $(M_t/\{H_2(m,4), H_2(m,4)\}/s_t + \{M(m), M(m)\}) + (H_2(m,4)/\infty) + (H_2(m,4)/\infty)$ model having two delayed customer feedback opportunities, with sinusoidal arrival rate in (5) for $r = 0.2$, mean service times $m_1 = 1.0, m_2 = 10/6, m_3 = 2.0$, mean patience times $m_1 = 2.0, m_2 = 1.0, m_3 = 10/8$ and feedback probabilities $p_1 = 0.6, p_2 = 0.5$: the cases of low waiting-time (high QoS targets) ($w = 0.01$, 0.02, 0.03 and 0.04) and simple DIS staffing.



6 Implied Empirical Quality of Service Functions

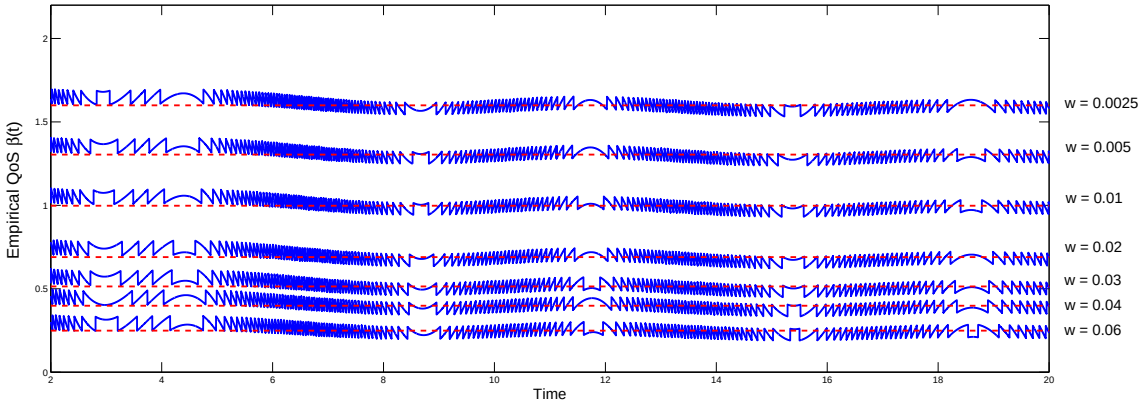
In this section we elaborate on §5.3 of the main paper, where we presented the implied empirical QoS function,

$$\beta_{DIS-MOL}(t) = \frac{s_{DIS-MOL}(t) - m(t)}{\sqrt{m(t)}} \tag{13}$$

associated with DIS-MOL staffing for the base model, denoted by $s_{DIS-MOL}(t)$. We now present corresponding plots for the other models.

We start with the $\sum_{i=1}^{2}(M_t/H_2(m_i, 4) + M(m_i)/s_t$ two-class model. The results are plotted in Figure 16.

Figure 16: The empirical Quality of Service (QoS) provided by the DIS-MOL staffing as a function of the waiting-time target w in the $\sum_{i=1}^{2}(M_t/H_2(m_i, 4) + M(m_i)/s_t$ two-class model example of §6.1 in the main paper.



We next consider the example from §6.2 involving 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 feedback after a delay in a finite-capacity orbit queue. The results are displayed in Figure 17.

We next consider the example with two feedback opportunities from §6.3. Specifically, we consider the $(M_t/GI, GI/s_t + GI, GI) + (GI/\infty) + (GI/\infty)$ model with two delayed customer feedback opportunities. The results are displayed in Figure 18.

Finally, we show the empirical Quality of Service (QoS) provided by the DIS-MOL staffing in the $(M_t(0.2)/H_2(1, 4), H_2(5, 4)/s_t + M(2), M(1)) + (0.2, H_2(1, 4)/\infty)$ model from §5.2 in this appendix with $\bar{\lambda}$ is reduced from 100, first to 20, and then to 5. Figures 19 and 20 show the results for 20 and 5, respectively. It should be compared to the corresponding plots for the model with $\bar{\lambda} = 100$ in Figure 4 of [9]. From the greater thickness of the plots here, we see that the discretization now has a bigger impact. But we see that the OL can be useful.

Figure 17: The empirical Quality of Service (QoS) provided by the DIS-MOL staffing as a function of the waiting-time target w in the $(M_t(0.2)/\{H_2(1,4), H_2(10/6,4)\}/s_t + \{M(2), M(1)\}) + (0.6, H_2(1,4)/s_t + M(1))$ example from §6.2.

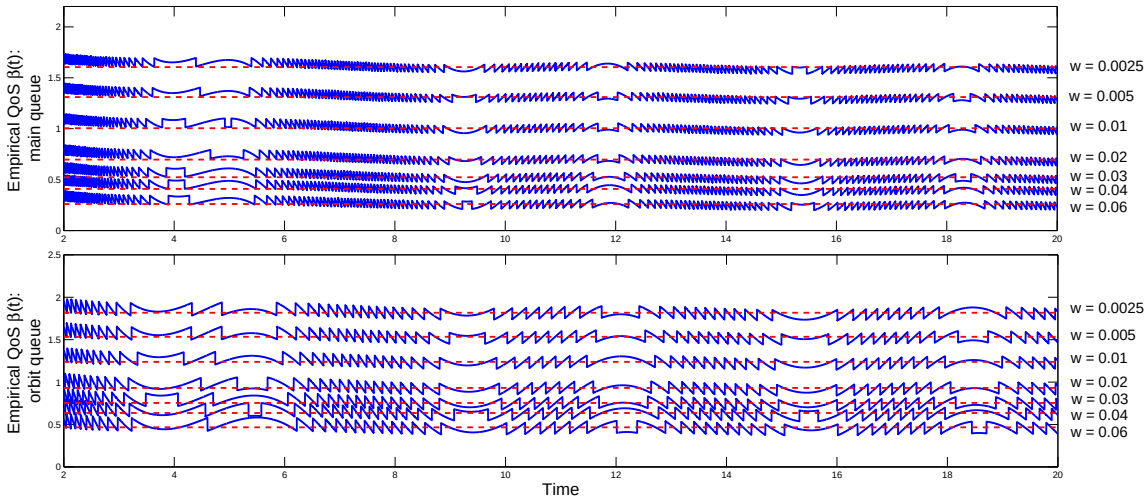


Figure 18: The empirical Quality of Service (QoS) provided by the DIS-MOL staffing as a function of the waiting-time target w in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ example from §6.3.

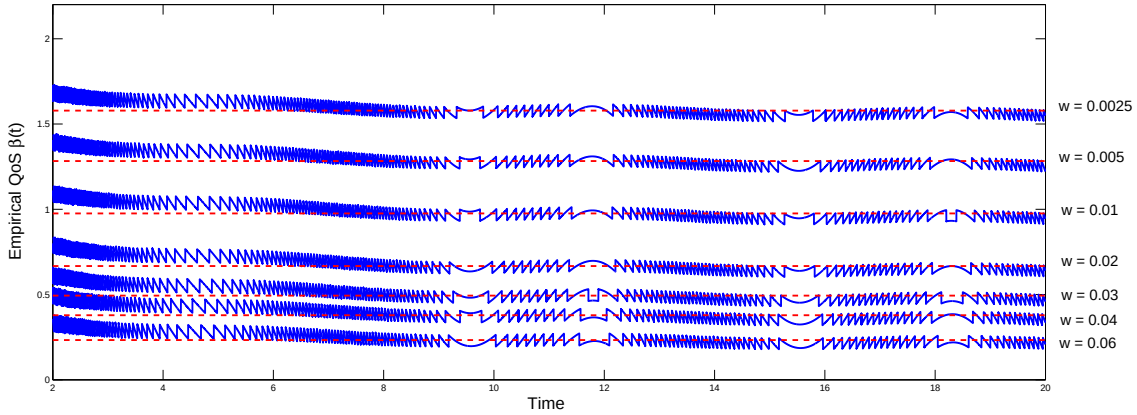


Figure 19: The empirical Quality of Service (QoS) provided by the DIS-MOL staffing as a function of the waiting-time target w in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model from §5.2 in this appendix with $\bar{\lambda}$ is reduced from 100 to 20.

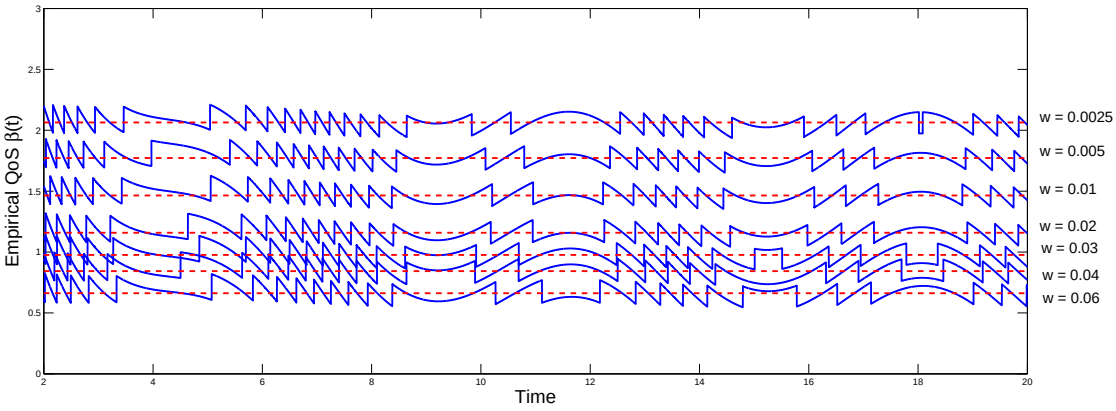
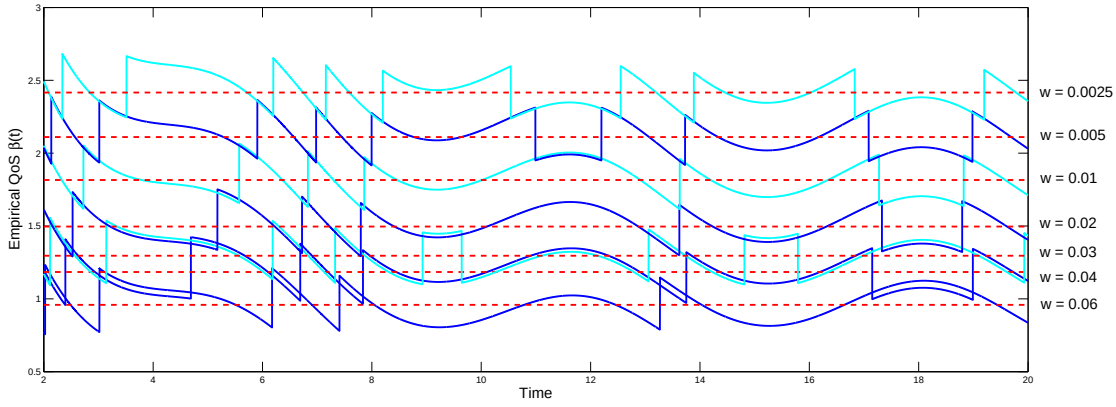


Figure 20: The empirical Quality of Service (QoS) provided by the DIS-MOL staffing as a function of the waiting-time target w in the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.2, H_2(1,4)/\infty)$ model from §5.2 in this appendix with $\bar{\lambda}$ is reduced from 100 to 5.



7 The Index of Disperson for Counts (IDC) of the Flows

A statistical analysis of the departure processes was a major component of our recent paper on feed-forward networks of many-server queues [8].

The DIS approximation for high waiting-time targets is valid for general G_t arrival processes, but the DIS-MOL approximation for low waiting-time targets depends on the M_t NHPP property, in order for the stationary $M/GI/s + GI$ model to yield a reasonable approximation for each t. As discussed in [8], it is possible that there can be significant degradation of the DIS-MOL approximation at a queue if the arrival process is not nearly NHPP. That can be caused by an upstream queue with a high waiting-time (low QoS) target and a service-time distribution that is not nearly exponential, such as the $H_2(m, 4)$ distributions we have been considering. We did not find such performance degradation to be a serious problem in the context of the present paper, so we did not discuss it in the main paper. However, we did investigate the question for the present feedback model. We discuss the results here.

First, we give some background. We assume that the external arrival process is NHPP. Under low waiting-time (high QoS) targets, that will make all queues similar to $M_t/GI/\infty$ IS queues, which have NHPP departure processes. Moreover, independent thinning of an NHPP is again an NHPP. So the NHPP property should propogate forward to all arrival processes in a lightly loaded feedback model.

To understand complications in more heavily loaded models, we first observe that the servers are all likely to be busy simultaneously in a more heavily loaded system. Thus, the departure process should behave like the superposition of a random number of i.i.d. renewal processes, where the inter-renewal times are the service times. Such superposition processes are well studied for stationary models. For a stationary model, the departure process from the upstream queue will tend to be similar to the superposition of a fixed number i.i.d. renewal processes, with inter-renewal times distributed as the service times in that upstream queue. Such superposition processes approach a Poisson process locally as the number of component processes increases, but they also have the same central-limit-theorem behavior as a single renewal process as time increases for any fixed number; see [8] and §9.8 of [11] for more discussion.

In [8] we found that the *index of dispersion for counts* (IDC) revealed when the flow could be regarded as approximately an NHPP as far as performance with the DIS-MOL approximation is concerned. The IDC is the ratio of the variance to the mean of the counting process. That is, if $A(t)$ counts the number of arrivals in $[0, t]$, then the IDC is the function

$$I(t) \equiv \frac{Var(A(t)}{E[A(t)]}, \quad t \geq 0. \tag{14}$$

For an NHPP, $I(t) = 1$ for all $t \geq 0$.

Based on [8], we are motivated to examine the IDC's of the various counting processes arising in our feedback models. We report the results in this section of the appendix. For the most part, we conclude that the flows should be approximately NHPP when it matters. Hence, we did not discuss this feature in the main paper. Our most important finding is that the IDC of the overall departure process from the system could well have an IDC significantly greater than 1, so that the departure process would have significantly more stochastic variability than an NHPP. That could cause performance degradation at a subsequent queue fed by the departure process from the feedback model.

7.1 The Base Example: an IS Orbit Queue

we start by examining the IDC of various counting processes in the base $(M_t(r)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (p, H_2(1,4)/\infty)$ model with an IS orbit queue and $r = p = 0.2$ from §5 of [9]. We consider the same models here as before. All service-time distributions are H_2, while all patience distributions are M, but the means vary, as indicated. Since the departure process from an $M_t/GI/\infty$ IS model is always Poisson, by Theorem 1 of [1], we expect no problem with an IS orbit queue.

Paralleling Figures 2 and 3 of the main paper, Figure 21 (22) shows the results when both queues have high waiting-time,low QoS, targets and DIS staffing (low waiting-time, high QoS, targets and DIS-MOL staffing), respectively. The entrant arrivals are from the process with rate function $\lambda_F(t)$ coming out of the orbit queue; the total arrival process combined the reentrant arrivals with the external arrivals, having rate function $\lambda(t) + \lambda_F(t)$; The total departure process is the aggregate departure process from the system, entering the outside world, corresponding to the rate functions $(1-p)\sigma_1(t) + \sigma_2(t)$; and the total orbiting process is the process entering the orbit queue, with rate function $p\sigma_1(t)$. From the perspective of the performance of DIS-MOL at the main queue, the most important process is the total arrival process.

Figure 22 shows that all the IDC's are consistently near 1 with low waiting-time (high QoS) targets and DIS-MOL staffing, so that all flows are nearly NHPP's. Since the queues are quite lightly loaded with DIS-MOL staffing, they are not too different from $M_t/GI/\infty$ queues. These IDC's are consistent with the theoretical properties that (i) the departure process from an $M_t/GI/\infty$ queue is NHPP and (ii) the independent thinning of an NHPP is again an NHPP. Hence, the NHPP property propagates through the network, approximately.

However, when both queues have high waiting-time (low QoS) targets, we see that the total departure process exhibits an IDC approaching 2, but none of the arrival processes have IDC's that differ significantly from 1. Evidently the thinning of the initial departure process makes the arrival process at the orbit queue nearly NHPP. These IDC plots are consistent with the excellent performance we saw in §6.2 of the main paper. However, we emphasize that DIS staffing with high waiting-time (low QoS) targets is effective without requiring the NHPP property.

However, it is important to note that the overall departure process *does* have an IDC that is significantly greater than 1. Hence, it is possible that this feedback queue, with these high waiting-time (low QoS) targets, could have adverse effect on the performance of a subsequent queue fed by the overall departure process, because the overall departure process is significantly more variable than an NHPP. Experience indicates that the IDC values in the range between 1 and 2 do not affect the performance too much, so this phenomenon appears to be not too serious.

Figure 21: Estimates of the mean $ED(t)$, variance $Var(D(t))$ and IDC $I(t) \equiv Var(D(t))/ED(t)$ for several counting processes from the $(M_t(r)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (p, H_2(1,4)/\infty)$ base model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue, as in §5 of [9]: the cases of high waiting-time (low QoS) targets $w = 0.40$ and 0.10 and DIS staffing.

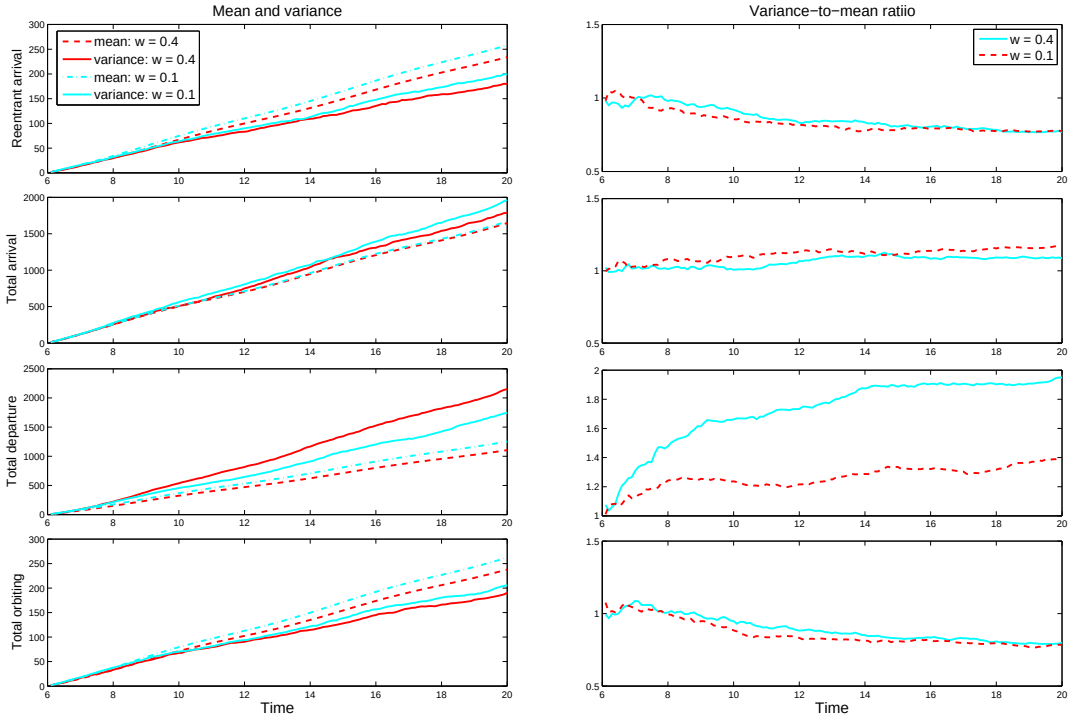
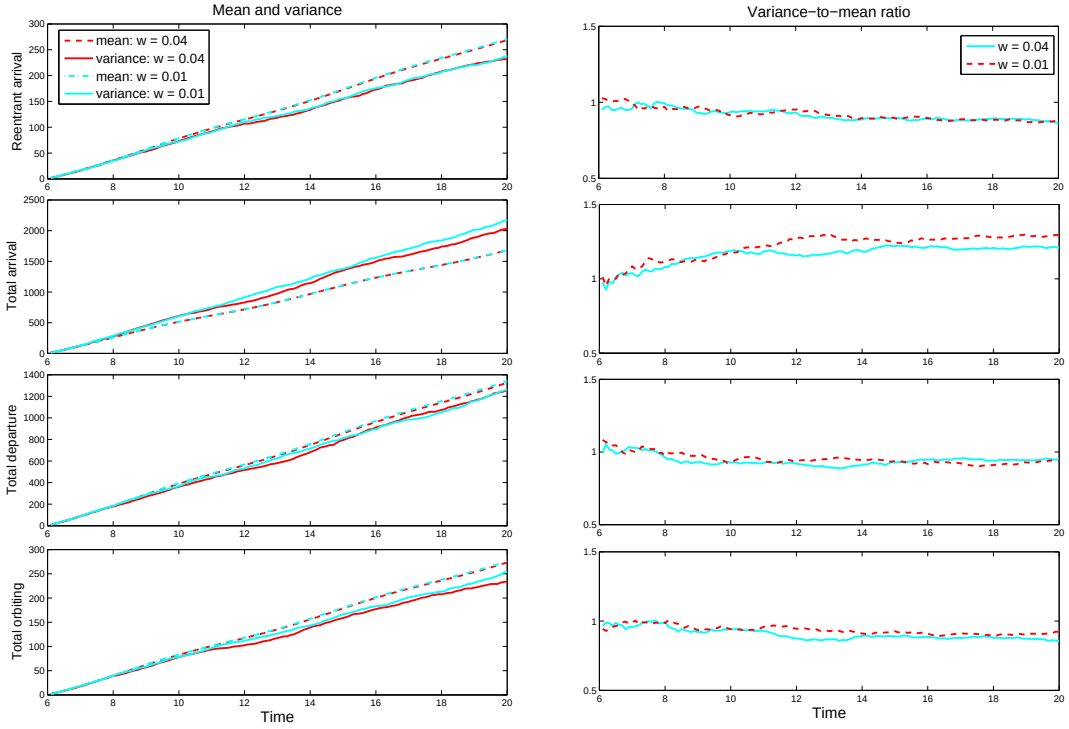


Figure 22: Estimates of the mean $ED(t)$, variance $Var(D(t))$ and IDC $I(t) \equiv Var(D(t))/ED(t)$ for several counting processes from the $(M_t(r)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (p, H_2(1,4)/\infty)$ base model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.2$ and an IS orbit queue, as in §5 of [9]: the cases of low waiting-time (low QoS) targets $w = 0.04$ and 0.01 and DIS-MOL staffing.

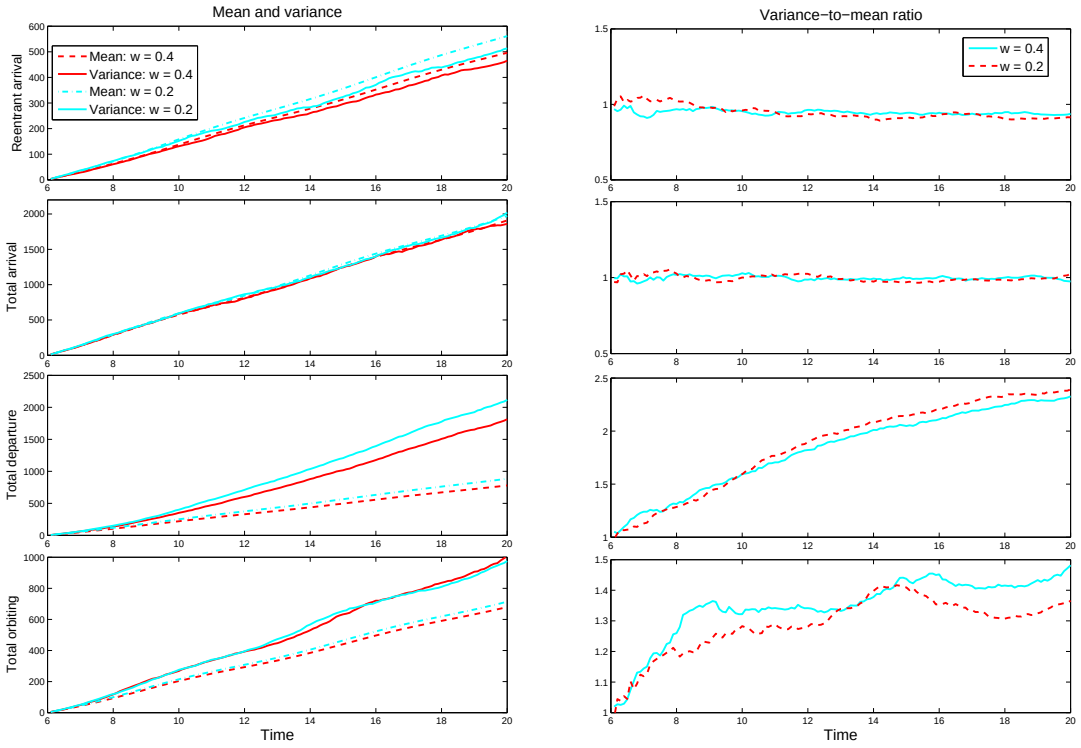


7.2 A Finite-Capacity Orbit Queue

As indicated above, there is a potential problem with a finite-capacity orbit queue. Hence, we now estimate the IDC's of the main flows in the $(M_t(r)/GI, GI/s_t + GI, GI) + (p, GI/s_t + GI)$ model with a finite-capacity orbit queue, as in §6.2 of [9]. Again we use the same model parameters as in the main paper. We will be presenting IDC's in the two cases corresponding to Figures 7 and 8 in the main paper. As is done there, we let the waiting-time targets be the same in the two queues, but we consider both high and low waiting-time targets. In particular, we consider the $(M_t(r)/H_2(1,4), H_2(5,4)/s_t + M(2), M(1)) + (p, H_2(1,4)/s_t + M(1))$ model with $r = 0.2$ and $p = 0.6$.

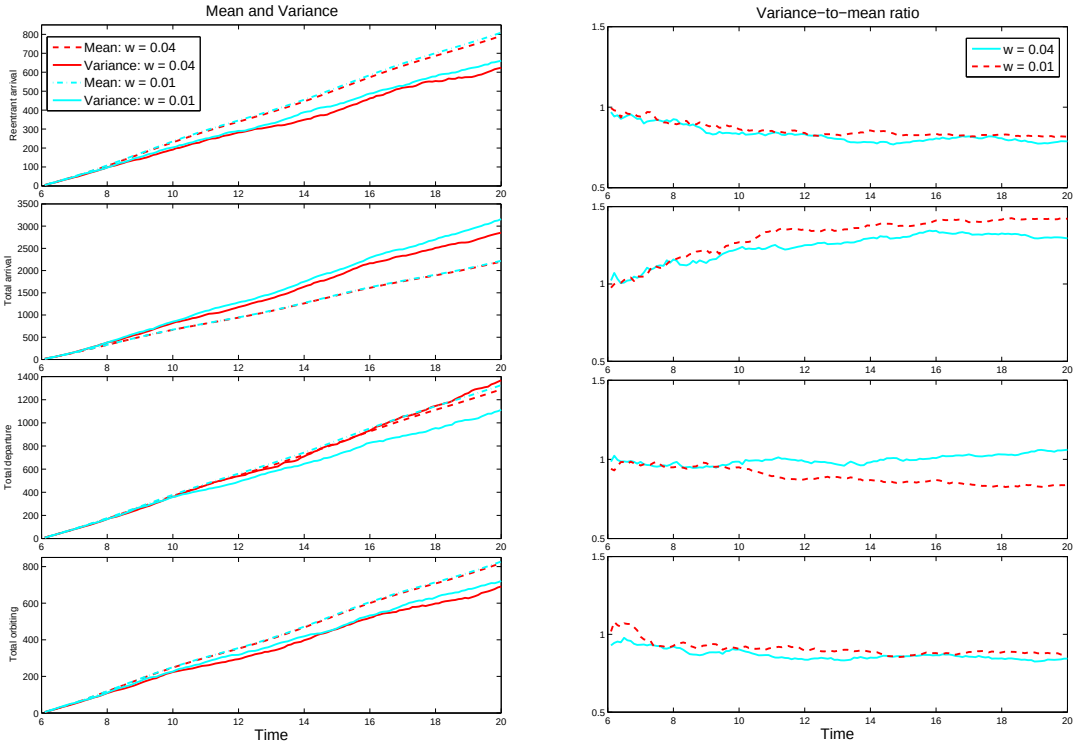
Figures 23 and 24 show the results with high and low waiting-time targets. The same target is used at the finite-capacity orbit queue as at the main queue. Here the feedback probability is $p = 0.6$ instead of $p = 0.2$, as for the IS orbit queue.

Figure 23: Estimates of the mean $ED(t)$, variance $Var(D(t))$ and IDC $I(t) \equiv Var(D(t))/ED(t)$ for several counting processes from the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + (0.6, H_2(1,4)/s_t + M(1))$ model with finite-capacity orbit queue, the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$, Bernoulli feedback with probability $p = 0.6$, as in §6.2 of [9]: the cases of identical high waiting-time (low QoS) targets $W = 0.10$ and $w = 0.40$ and DIS staffing at both queues.



The results are similar to those for the IS orbit queue before. In both cases the total arrival process appears to be approximately an NHPP. However, as before, in the case of high waiting-time (low QoS) targets $W = 0.10$ and $w = 0.40$ and DIS staffing at both queues, the total departure

Figure 24: Estimates of the mean $ED(t)$, variance $Var(D(t))$ and IDC $I(t) \equiv Var(D(t))/ED(t)$ for several counting processes from the $(M_t(0.2)/\{H_2(1,4), H_2(5,4)\}/s_t + \{M(2), M(1)\}) + +(0.6, H_2(1,4)/s_t + M(1))$ model with a finite-capacity orbit queue, the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.2$ and Bernoulli feedback with probability $p = 0.6$, as in §6.2 of [9]: the cases of identical low waiting-time (high QoS) targets $W = 0.01$ and $w = 0.04$ and DIS-MOL staffing at both queues.



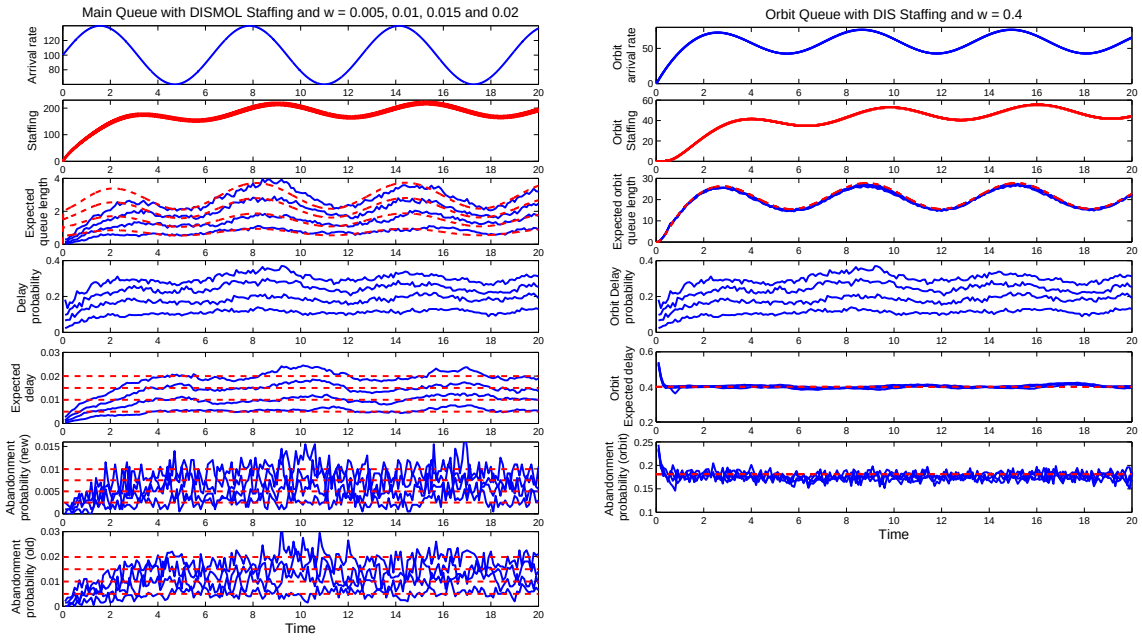
process has an IDC significantly above 1. Hence, there could well be performance degradation at a subsequent queue fed by the overall departure process.

7.3 Identifying a Problem Case

Evidently we are less likely to encounter difficulties with DIS-MOL in this feedback setting than in the feed-forward setting of [8]. Experience in [8] suggests that problems are likely to arise in a mixed case, where the main queue has a high QoS target, while the orbit queue has a low QoS target and a non-exponential service-time cdf. In order to seek out a bad case, we now consider a mixed case, with a high waiting-time target of $w = 0.4$ at the orbit queue, which has an $H_2(1, 4)$ service-time cdf, but then two cases of low waiting-time (high QoS) targets ($w = 0.01$ and 0.02) and DIS-MOL staffing at the main queue. We also increase the relative amplitude of the sinusoidal arrival rate function from $r = 0.2$ to $r = 0.4$ and we make the service-time distributions M at the main queue instead of H_2. (The performance in the $H_2/M/s$ model is typically worse than in the $H_2/H_2/s$ model.) In particular, the model now is $(M_t(0.4)/M(1), M(5/3)/s_t + M(2), M(1)) + (0.6, H_2(1, 4)/s_t + M(2))$. The waiting-time target at the orbit queue is fixed at the high target $w_2 = 0.4$ and DIS staffing is used. For the performance results, we consider four different low waiting-time targets at the main queue: $w = 0.005$, 0.010, 0.015 and 0.020, and DIS-MOL staffing is used there.

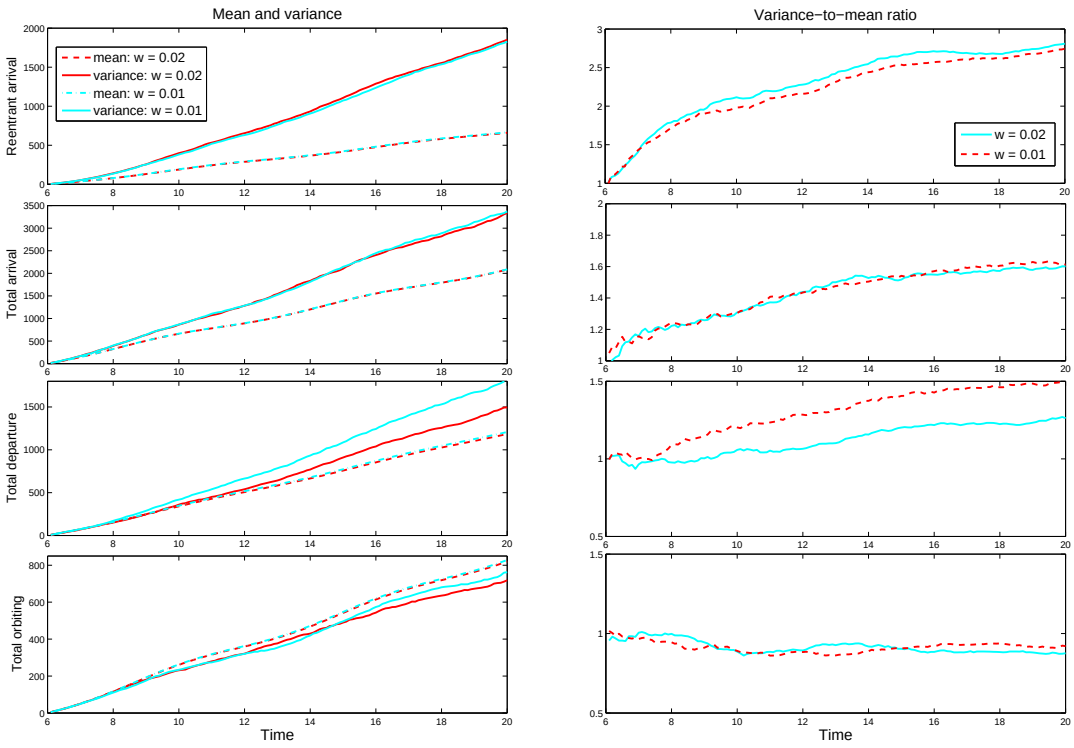
Figure 25 shows the performance results. Now we do see performance degradation, similar to what we saw in [8].

Figure 25: Performance functions in the $(M_t(0.4)/\{M(m), M(m)\}/s_t + \{M(m), M(m)\}) + (0.6, H_2(1, 4)/s_t + M(1))$ model with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.4$, Bernoulli feedback with probability $p = 0.6$ and a finite-capacity orbit queue: mixed performance targets, a high waiting-time target of 0.4 at the orbit queue, which has an $H_2(1, 4)$ service-time cdf: four cases of low waiting-time (high QoS) targets ($w = 0.005$, 0.010, 0.015 and 0.020) and DIS-MOL staffing at the main queue.



The corresponding IDC's of the flows are shown in Figure 26. As in [8], the departure from $I(t) \approx 1$ helps explain the performance degradation seen in Figure 25.

Figure 26: Estimates of the mean $ED(t)$, variance $Var(D(t))$ and IDC $I(t) \equiv Var(D(t))/ED(t)$ for several counting processes from the same $(M_t(0.4)/\{M(m), M(m)\}/s_t + \{M(m), M(m)\}) + (0.6, H_2(1, 4)/s_t + M(1))$ model considered in Figure 25, with the sinusoidal arrival rate in (12) for $\bar{\lambda} = 100$ and $r = 0.4$, Bernoulli feedback with probability $p = 0.6$ and a finite-capacity orbit queue with fixed high waiting-time (low QoS) target $w = 0.4$ and DIS staffing, as in §6.2 of [9]: the cases of low waiting-time (high QoS) targets $w = 0.01$ and $w = 0.02$ and DIS-MOL staffing at the main queue.



As expected, Figure 26 shows that the total arrival process has an IDC significantly greater than 1, providing evidence of a total arrival process that is significantly more variable than an NHPP. This IDC is consistent with the performance degradation seen in Figure 25. This performance degradation is not too severe, and might not interfere with useful engineering applications, but it is detectable. The performance degradation in Figure 25 is evident.

Acknowledgement

This research began as part of the first author's doctoral dissertation at Columbia University. The first author receives support from NSF grant CMMI 1362310, the second author received support from NSF grants CMMI 1066372 and 1265070.

References

[1] Eick, S. G., Massey, W. A. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Oper Res* 41:731–742.

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

[3] Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper Management* 16:13–29.

[4] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Sci* 42:1383–1394.

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

[6] Liu, Y. and Whitt, W. (2012). 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.

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

[8] Liu, Y. and Whitt, W. (2013). Stabilizing performance in feed-forward networks of many-server queues with time-varying arrival rates. Columbia University: Available at: www.columbia.edu/ww2040/allpapers.html.

[9] Liu, Y. and Whitt, W. (2013). Stabilizing performance in many-server queues with time-varying arrivals and customer feedback. Columbia University: Available at: www.columbia.edu/ww2040/allpapers.html.

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

[11] Whitt, W. (2002). *Stochastic-Process Limits*. New York: Springer.