ONLINE APPENDIX
to
Staffing to Stabilize the Tail Probability of Delay in Service Systems with Time-Varying Demand

Yunan Liu

December 16, 2016

Abstract

Analytic formulas are developed to set the time-dependent number of servers in order to stabilize the tail probability of customer waiting times for the $G_t/GI/s_t + GI$ queueing model, which has a non-stationary non-Poisson arrival process (the G_t), non-exponential service times (the first GI), and allows customer abandonment according to a non-exponential patience dis-tribution (the $+GI$). Specifically, for any delay target $w > 0$ and probability target $\alpha \in (0, 1)$, we determine appropriate staffing levels (the s_t) so that the time-varying probability that the waiting time exceeds a maximum acceptable value w is stabilized at α at all times. In addition, effective approximating formulas are provided for other important performance functions such as the probabilities of delay and abandonment, and the means of delay and queue length. Many-server heavy-traffic limit theorems in the efficiency-driven regime are developed to show that

(i) the proposed staffing function achieves the goal asymptotically as the scale increases, and (ii) the proposed approximating formulas for other performance measures are asymptotically accurate as the scale increases. Extensive simulations show that both the staffing functions and the performance approximations are effective, even for smaller systems having around 3 servers.

Acronym	Meaning
ccdf	complementary cumulative distribution function
cdf	cumulative distribution function
DIS	delayed infinite server
DIS-MOL	delayed infinite-server modified-offered-load approximation
FCLT	functional central limit theorem
ED	efficiency driven
ERP	equilibrium renewal process
FWLLN	functional weak law of large numbers
TTGA	Two-Term Gaussian approximation
i.i.d.	independent and identically distributed
MPS	marginal price of staffing
MSHT	many-server heavy-traffic
NHPP	non-homogeneous Poisson process
NNPP	nonstationary non-Poisson process
OL	overloaded
pdf	probability density function
PoA	probability of abandonment
PoD	probability of delay
PWT	potential waiting time
QED	quality-and-efficiency driven
QoS	quality of service
SCV	squared coefficient of variation
TPoD	tail probability of delay

Table 1: Summary of useful acronyms used in the main paper.

1 Overview

This appendix provides additional supplementary material to the main paper. In §2 we present a full Markovian model. In §3 we present an example with long service times with mean service time $\mathrm{E}[S] = 4$. In §§4–5 we present additional results of the examples with large and small arrival rates supplementing §5.3 of the main paper. In §§6–7, we present additional details of the lightly-loaded and heavily-loaded systems considered in §5.4 of the main paper. In §8, we confirm that the TTGA staffing works well for real-world examples, including realistic arrival rates estimated from real hospital data and call-center data in the SEEStat database (SEE Center [2014]). In §9 we supplement §5.6 of the main paper by presenting additional examples with other arrival-rate functions (e.g., constant, piecewise linear and on-and-off arrival rates). In §10 we provide addition examples with non-exponential service to supplement §5.7 in the main paper. In §11 we provide proofs that are omitted in the main paper. In §12, we provide the final explicit expressions for the approximating formulas in §4 of the main paper. In §13, we present the explicit form of the TTGA formula for the example in §2 of the main paper. In §14 we provide the implementation details of simulations.

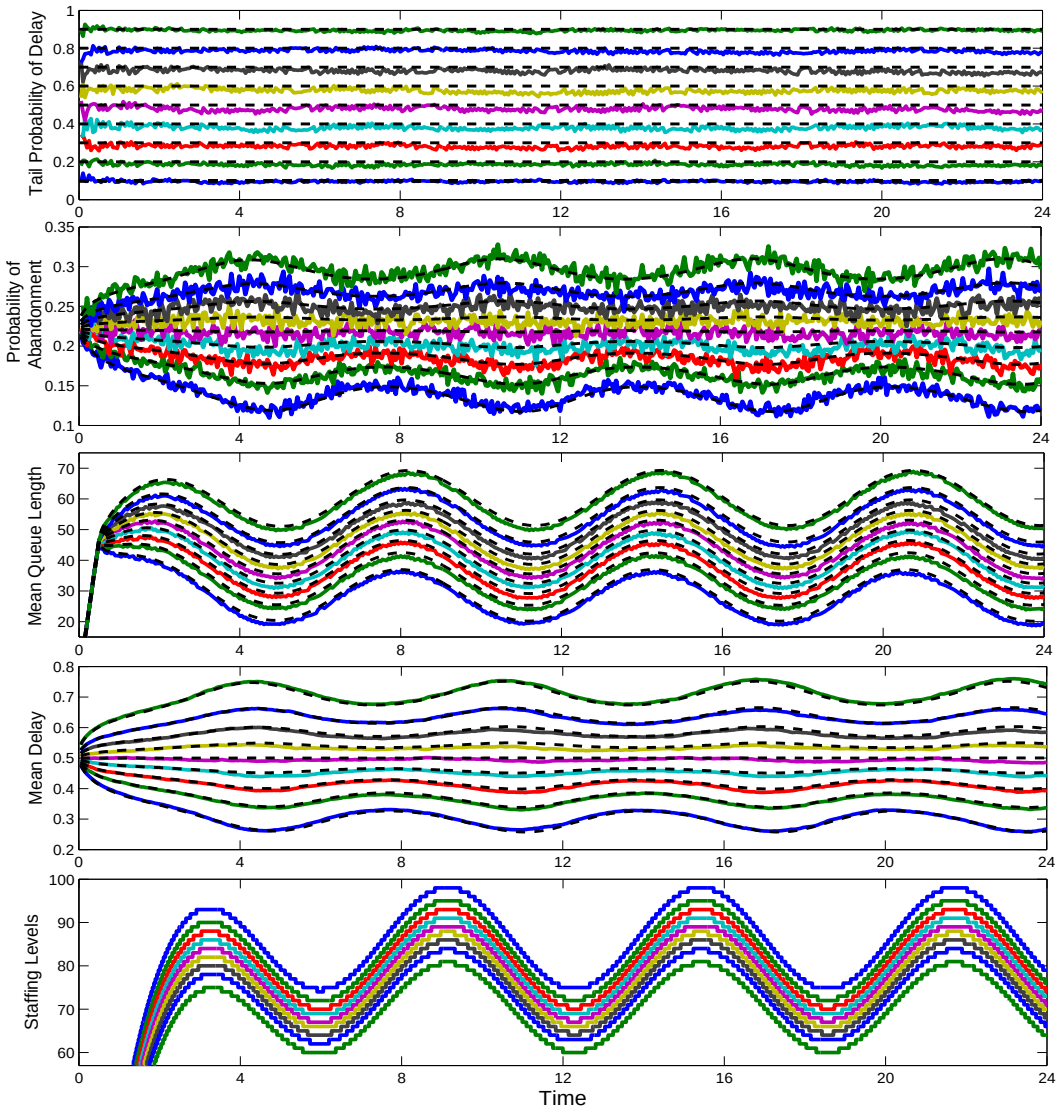

Figure 1: Performance measures of the $M_t/M/s_t + M$ example with $\lambda(t) = 100 + 20\sin t$, mean service time $\mathrm{E}[S] = 1/\mu = 1$, abandonment rate $\theta = 0.5$ and QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$

2 A Full Markovian Example

In this section we consider an $M_t/M/s_t + M$ full Markovian example, having an NHPP arrival process, exponential service times with rate μ and exponential patience times with rate θ. In this case, the staffing formulas (2), (9) and (10) in the main paper simplifies to

$$s_w^{(1)}(t) = e^{w(\mu-\theta)-\mu t}\int_0^{t-w}\lambda(u)e^{\mu u}du \quad \text{and} \quad s_{w,\alpha}^{(2)}(t) \quad = z_\alpha e^{-\mu t}\left(Z(t) - (\mu-\theta)\int_w^{t\vee w} Z(u)du\right)$$

where

$$Z(t) = e^{(\mu-\theta)t}\sqrt{\int_w^{t\vee w} e^{2\theta x}(2\mu s_w^{(1)}(x) + \dot{s}_w^{(1)}(x))dx}.$$

We have the model parameters: $\lambda(t) = 100 + 20\sin t$, $\mu = 1$, and $\theta = 0.5$. Figure 1 reports shows that TPoDs are stabilized at desired levels for $w = 0.5$, $\alpha = 0.1, 0.2, \ldots, 0.9$, and other performance measures are well approximated.

3 Long Service Times

Service systems such as hospitals have long service times, which is usually more difficult to treat. We now confirm that TTGA works well for systems with long service times. We consider the $H_2(t)/M/s_t + H_2$ model having all parameters given in §2 except for a modified mean service time $\mathrm{E}[S] = 1/\mu = 4$. Figure 2 shows that TTGA continue to perform well. The TPoDs are even smoother because the case of longer service times requires more servers, thus larger system size, which reduces discretization errors. We recognize that the staffing levels for $\mathrm{E}[S] = 4$ (between 200 and 350, as shown in the last subplot of Figure 2) are much higher comparing with the case $\mathrm{E}[S] = 1$ (between 60 and 100, as shown in the last plot of Figure 1 in the main paper). Therefore, this result also supplement our discussion on large-scale systems in §4.

Targets	Avg (diff. to target)	Max (diff. to target)	Min (diff. to target)
0.9	0.9040 (+0.0040)	0.9184 (+0.0184)	0.8878 (-0.0122)
0.8	0.7991 (-0.0009)	0.8154 (+0.0154)	0.7804 (-0.0196)
0.7	0.6989 (-0.0011)	0.7122 (+0.0122)	0.6832 (-0.0168)
0.6	0.5968 (-0.0032)	0.6146 (+0.0146)	0.5750 (-0.0250)
0.5	0.5033 (+0.0033)	0.5250 (+0.0250)	0.4852 (-0.0148)
0.4	0.4042 (+0.0042)	0.4254 (+0.0254)	0.3808 (-0.0192)
0.3	0.3034 (+0.0034)	0.3174 (+0.0174)	0.2812 (-0.0188)
0.2	0.2023 (+0.0023)	0.2170 (+0.0170)	0.1862 (-0.0138)
0.1	0.1026 (+0.0026)	0.1146 (+0.0146)	0.0900 (-0.0100)

Table 2: Comparison with TPoD targets (average, min and max), with $w = 0.5$ and $\bar{\lambda} = 1000$.

4 Large-Scale Systems

We now consider large arrival rates to supplement the Theorems 2-3 and the arguments in §5.3 of the main paper. We repeat the $H_2(t)/M/s_t + H_2$ example in §2 with $\bar{\lambda} = 1000$. Figure 3 shows that TPoD is stabilized at all desired targets and PoA, mean queue length and mean delay all closely match with their approximating formulas. The PoD and utilization are both close to 1 (thus omitted here). Table 2 summarizes the minimum, maximum and the average values of the TPoD and compare to the associated TPoD targets. This example helps visualize the asymptotic

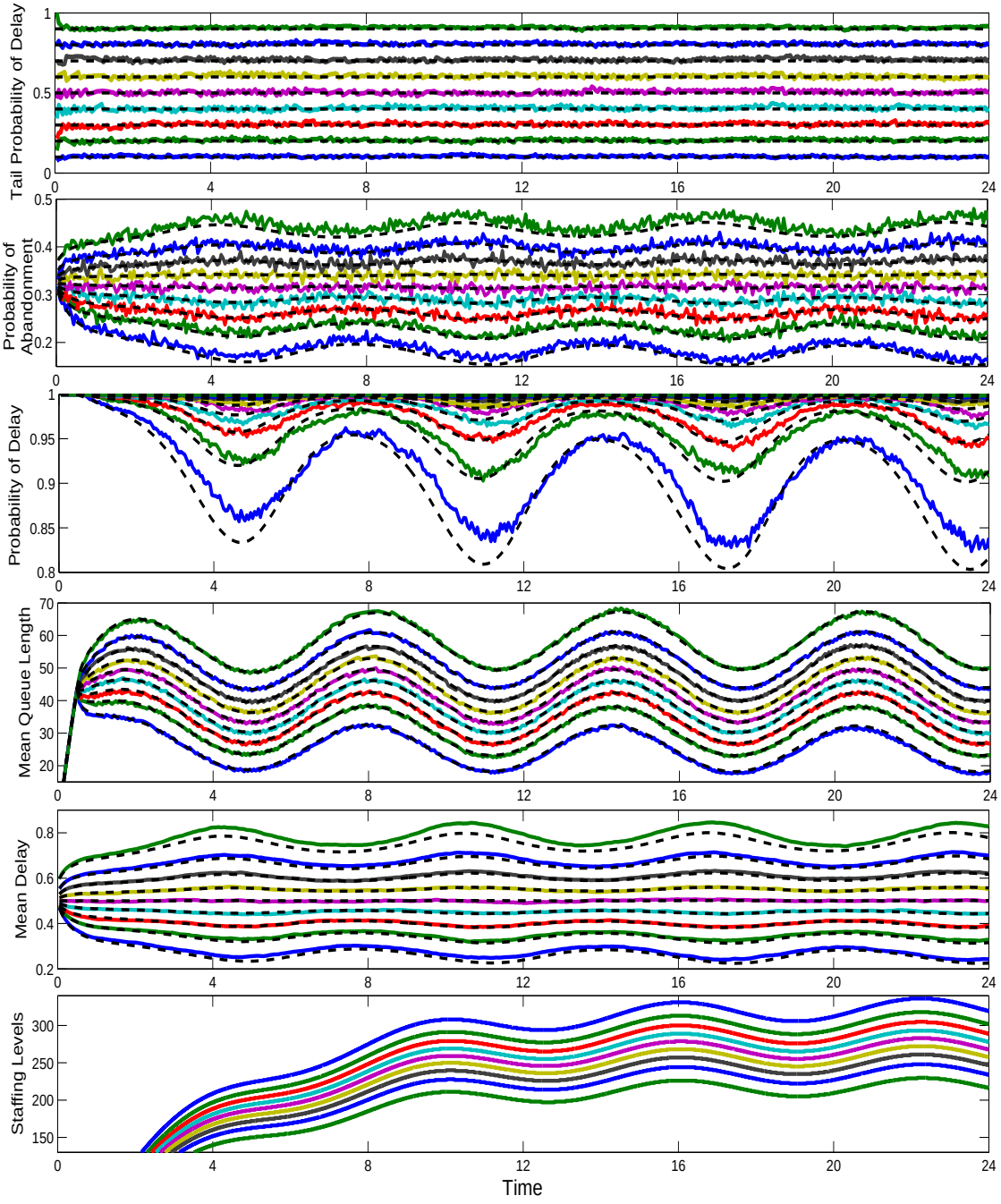

Figure 2: Performance measures of the $H_2(t)/M/s_t + H_2$ example with $\lambda(t) = 100 + 20\sin t$, large mean service time $\mathrm{E}[S] = 1/\mu = 4$, and QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$

stability of the TTGA staffing and the asymptotic accuracy of the performance functions as the scale increases.

5 Small-Scale Systems

Supplementing §5.3 in the main paper, here we present more details of the example with $\bar{\lambda} = 10$. We have seen in Figure 4 of the main paper that TPoD of this example has much bigger fluctuation due to higher sensitivity to staffing discretization. Mimicking Table 1 in the main paper, Table 3 shows the average, maximum, and minimum of the TPoD using ceiling and flooring, and compares them with the associated TPoD targets.

	Avg (diff. to target)		Max (diff. to target)		Min (diff. to target)	
Targets	ceiling	flooring	ceiling	flooring	ceiling	flooring
0.9	0.9204 (+0.0204)	0.9381 (+0.0381)	0.9438 (+0.0438)	0.9658 (+0.0658)	0.8910 (-0.0090)	0.9124 (+0.0124)
0.8	0.7940 (-0.0060)	0.8381 (+0.0381)	0.8410 (+0.0410)	0.8782 (+0.0782)	0.7538 (-0.0462)	0.7968 (-0.0032)
0.7	0.6858 (-0.0142)	0.7383 (+0.0383)	0.7414 (+0.0414)	0.7894 (+0.0894)	0.6384 (-0.0616)	0.6894 (-0.0106)
0.6	0.5875 (-0.0125)	0.6351 (+0.0351)	0.6406 (+0.0406)	0.6846 (+0.0846)	0.5344 (-0.0656)	0.5862 (-0.0138)
0.5	0.4875 (-0.0125)	0.5312 (+0.0312)	0.5460 (+0.0460)	0.5876 (+0.0876)	0.4238 (-0.0762)	0.4762 (-0.0238)
0.4	0.3908 (-0.0092)	0.4328 (+0.0328)	0.4418 (+0.0418)	0.4880 (+0.0880)	0.3392 (-0.0608)	0.3786 (-0.0214)
0.3	0.2910 (-0.0090)	0.3312 (+0.0312)	0.3354 (+0.0354)	0.3912 (+0.0912)	0.2470 (-0.0530)	0.2874 (-0.0126)
0.2	0.1968 (-0.0032)	0.2342 (+0.0342)	0.2498 (+0.0498)	0.2824 (+0.0824)	0.1540 (-0.0460)	0.1946 (-0.0054)
0.1	0.1020 (+0.0020)	0.1310 (+0.0310)	0.1324 (+0.0324)	0.1584 (+0.0584)	0.0780 (-0.0220)	0.0968 (-0.0032)

Table 3: Comparison with TPoD targets (average, min and max), with $w = 0.5$ and $\lambda(t) = 10 + 2\sin t$.

Simulation estimates of the performance measures are displayed in Figure 4. Notice that for small systems, the utilization can be as low as 0.7 and the PoD can reach 0.4, which indicate that the system is closer to the critically loaded state or the QED regime. Performance approximations degrade with smaller arrival rates, but are still acceptable.

The spikes (jumps) of TPoD are synchronized with the changes of the staffing level at future times. Figure 5 plots the TPoD of target $\alpha = 0.9$ and the corresponding staffing function. There is a time lag with length 0.5 between jumps of the TPoD and jumps in opposite directions of the staffing function, because the delay of an arrival at time t will be realized approximately 0.5 time unit later when this arrival enters service, so it is linked to future changes of staffing levels. The TPoD will drop straight down immediately at t if there an extra server is added at time $t + 0.5$.

6 Lightly Loaded Systems

In the main paper, we have already shown examples with delay target $w = 0.05$. In this section, we present two examples: (i) delay target $w = 0$ and $\alpha = 0.1, \ldots, 0.9$, and (ii) moderate delay target $w = 0.3$ and very small TPoD target $\alpha = 0.02, 0.04, 0.06, 0.08, 0.1$.

We remark that the case $w = 0$ is not supported by the asymptotic stability theorem in the main paper (Theorem 2) which clearly requires $w > 0$. Nevertheless, Figure 6 confirms that TTGA

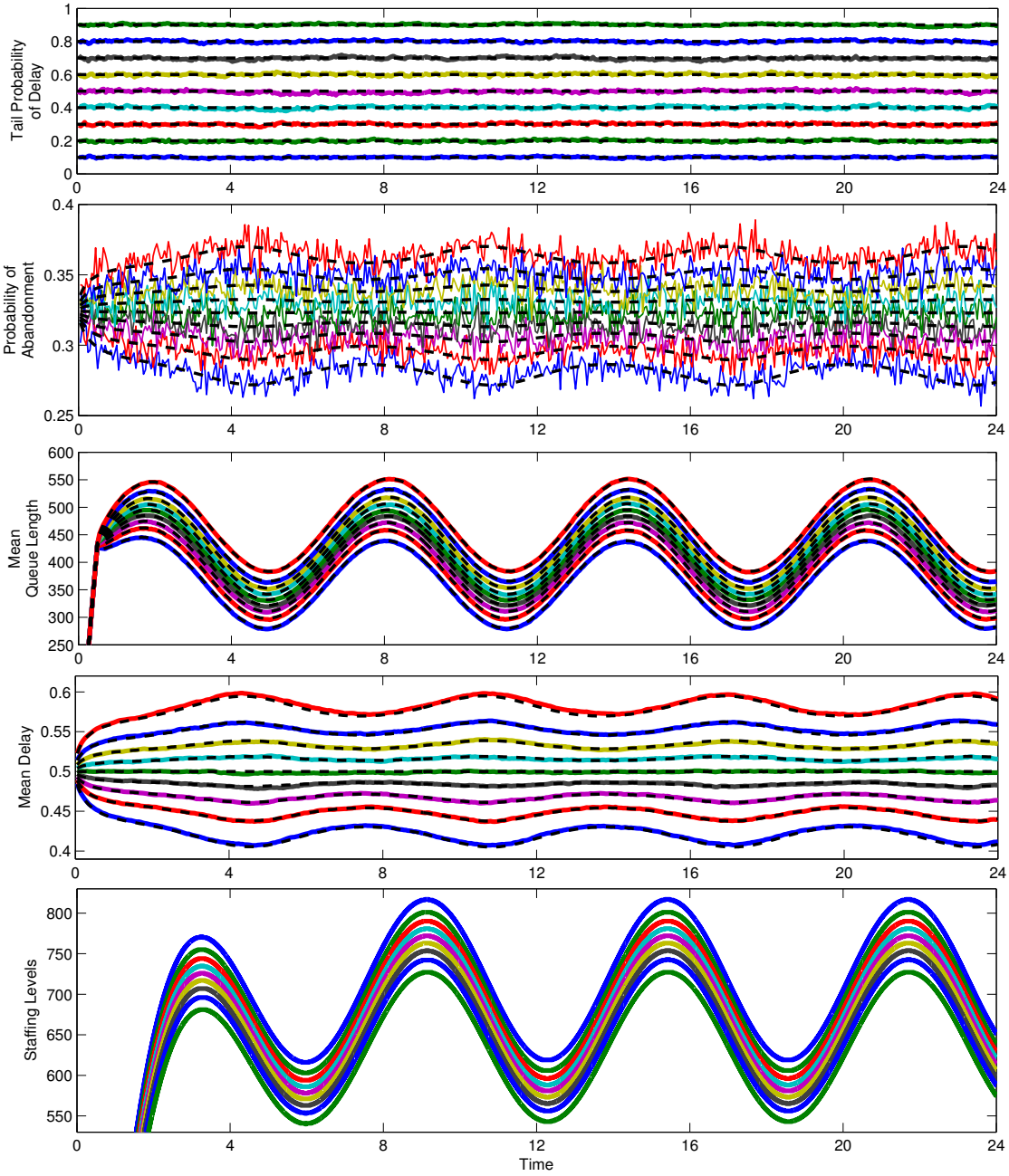

Figure 3: Performance measures of the $H_2(t)/M/s_t + H_2$ model with $\lambda(t) = 1000 + 200 \sin t$, and QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$

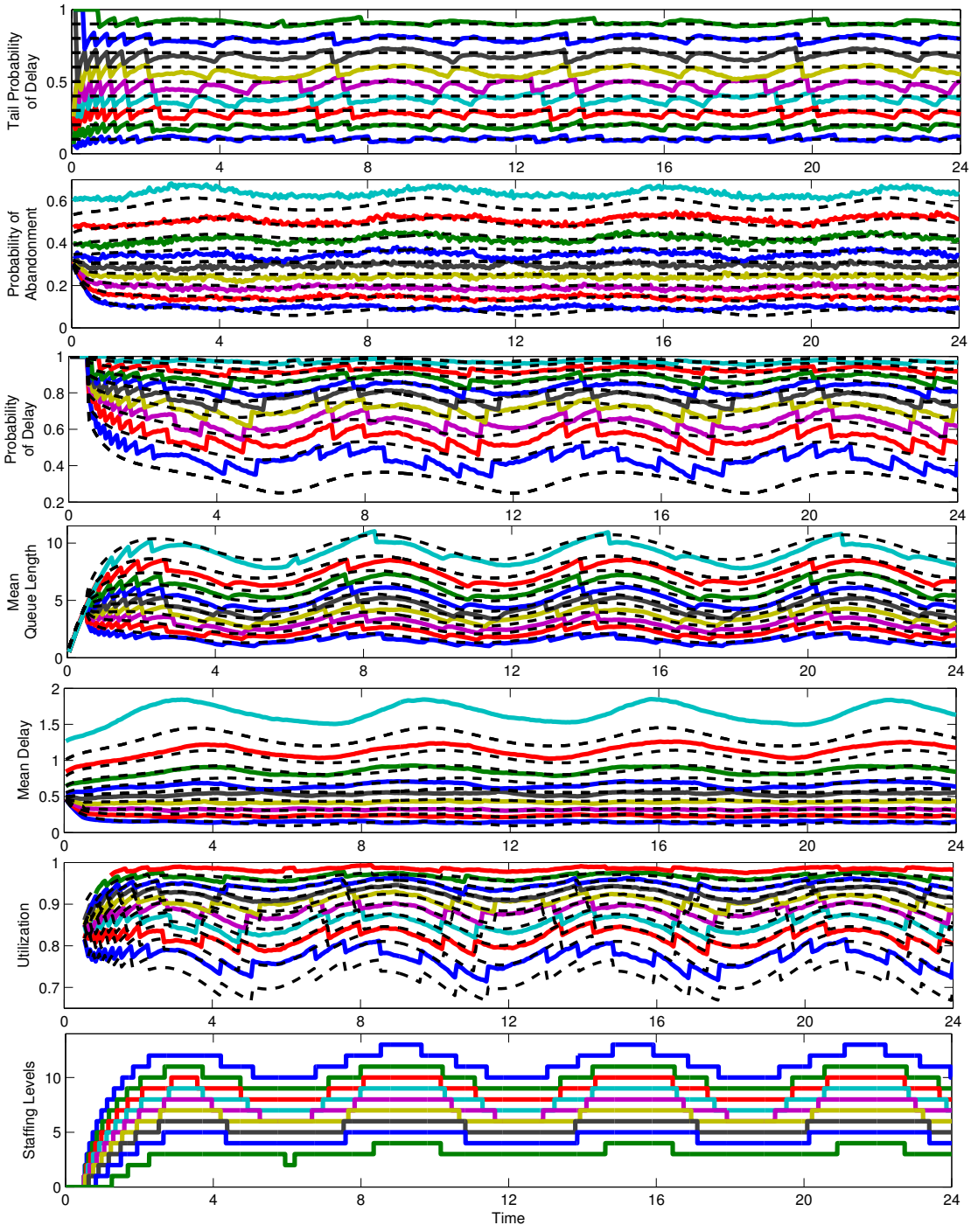

Figure 4: Performance measures of the $H_2(t)/M/s_t + H_2$ model with $\lambda(t) = 10 + 2\sin t$, and QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$

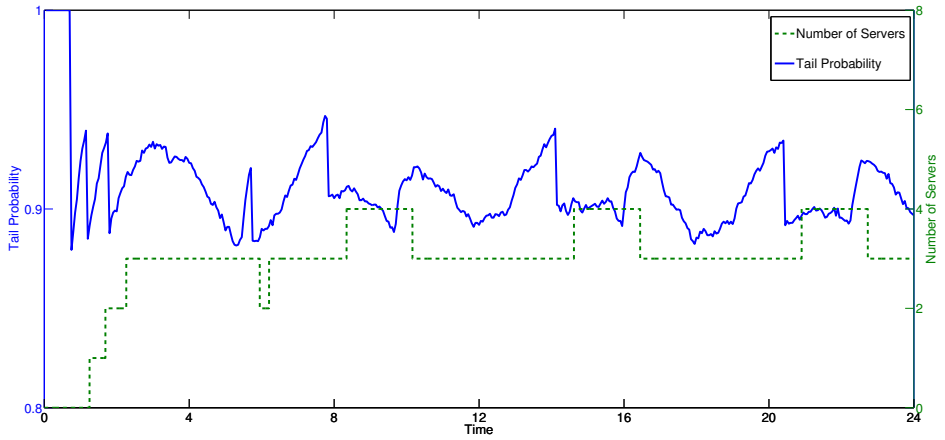

Figure 5: The jumps of TPoD caused by the future jumps of staffing levels, $\alpha = 0.9$, $w = 0.5$, $\bar{\lambda} = 10$.

performs well for $w = 0$.

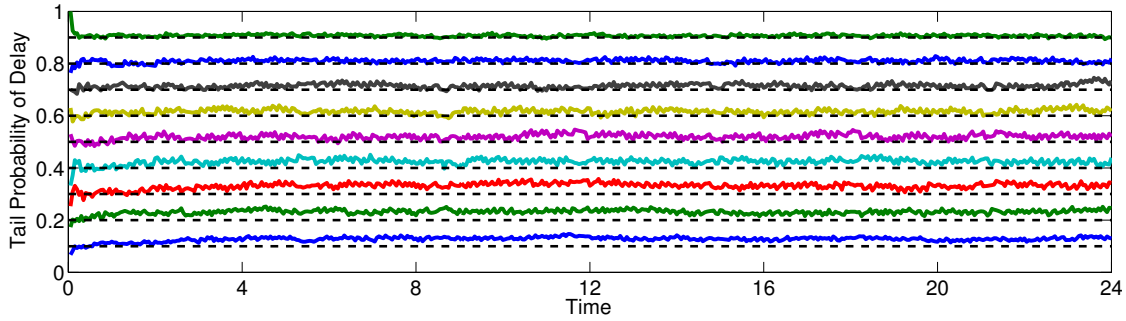

Figure 6: Stabilized TPoDs of the $H_2(t)/M/s_t + H_2$ model with $\lambda(t) = 100 + 20 \sin t$, QoS targets $w = 0$, $\alpha = 0.1, 0.2, \ldots, 0.9$

When w and α are small, it is expected that the PoD will not be close to 1 and the system will be in QED regime (instead of ED regime which is required for TTGA perform well). Nevertheless, Figure 7 confirms the good performance of TTGA. We note that the PoD is between 0.4 and 0.8 so the system is indeed in the QED regime.

7 Heavily Loaded Systems

Supplementing §5.4 of the main paper, we provide additional results for large dealys $w = 3$ and 6. Figure 8 provides the performance approximations for $w = 3$, which is omitted in §5.4 of the main paper. We observe that large delay causes (i) high PoAs, (ii) long waiting lines, and (iii) small number number of servers.

Keeping all other parameters unchanged, we now consider an even bigger delay target $w = 6$. See Figure 9 for the TPoDs under the TTGA staffing function by ceiling and flooring. A big w reduces the fluctuation of the staffing levels (thus resulting in a flatter and slowly changing TTGA formula), this can be understood from the staffing function $s_w^{(1)}(t)$ in Corollary 3. As a result, the TPoD becomes more sensitive to the staffing discretization. For example, if it takes a long time

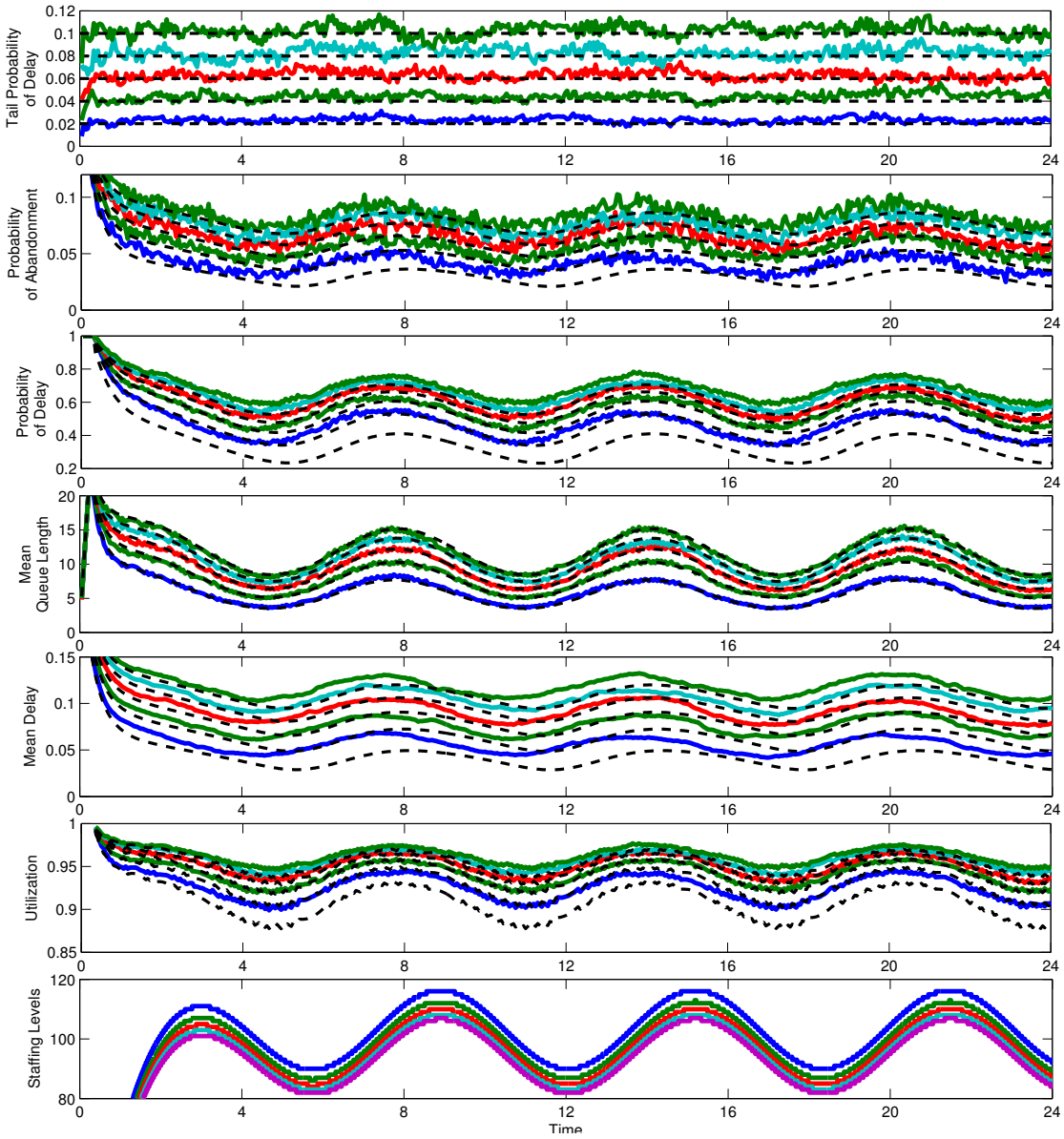

Figure 7: Performance measures of the $H_2(t)/M/s_t + H_2$ model with $\lambda(t) = 100 + 20\sin t$, QoS targets $w = 0.3$, $\alpha = 0.02, 0.04, \ldots, 0.1$

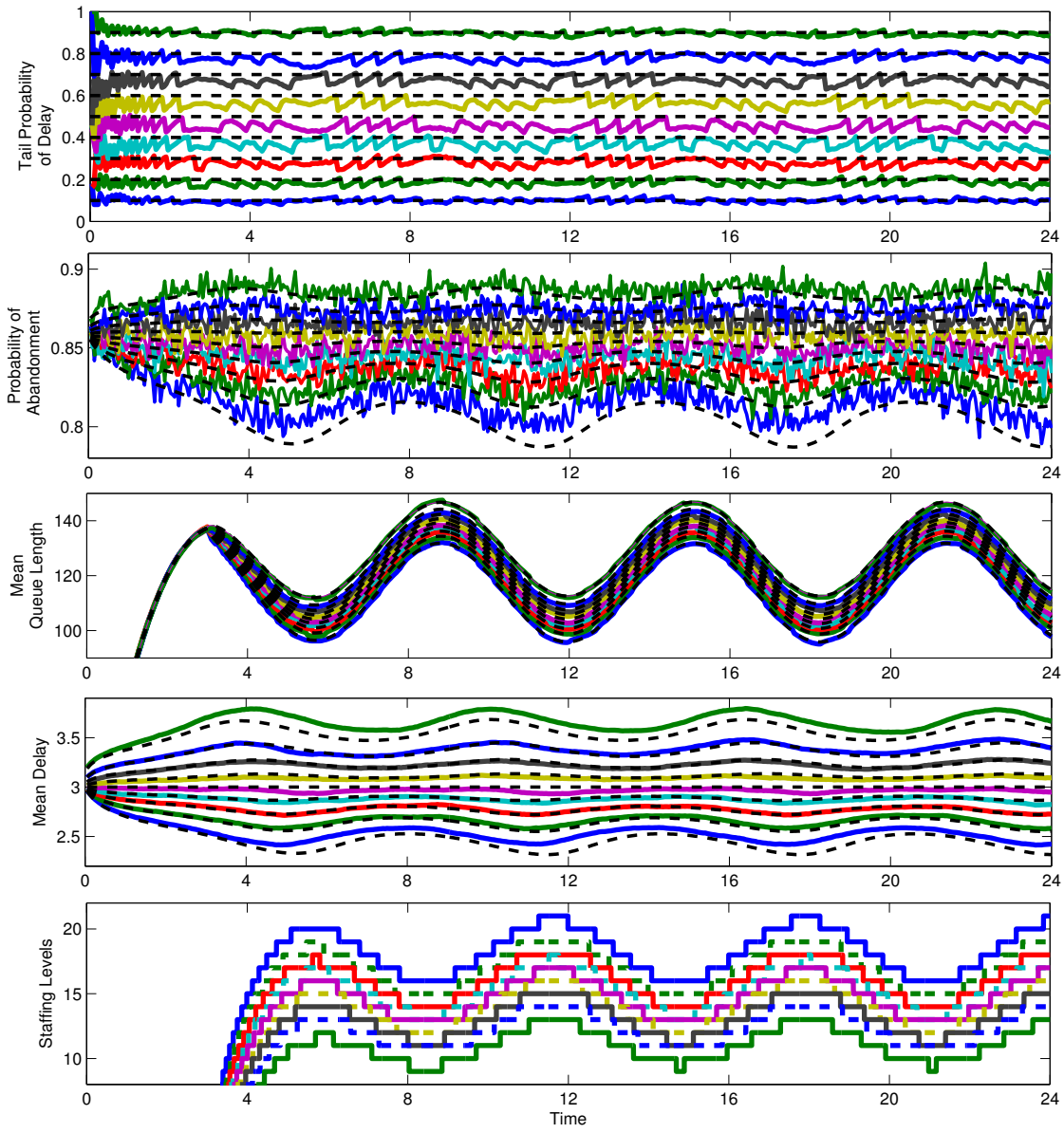

Figure 8: Performance measures of the $H_2(t)/M/s_t + H_2$ model with $\lambda(t) = 100 + 20\sin t$, QoS targets $w = 3$, $\alpha = 0.1, \ldots, 0.9$

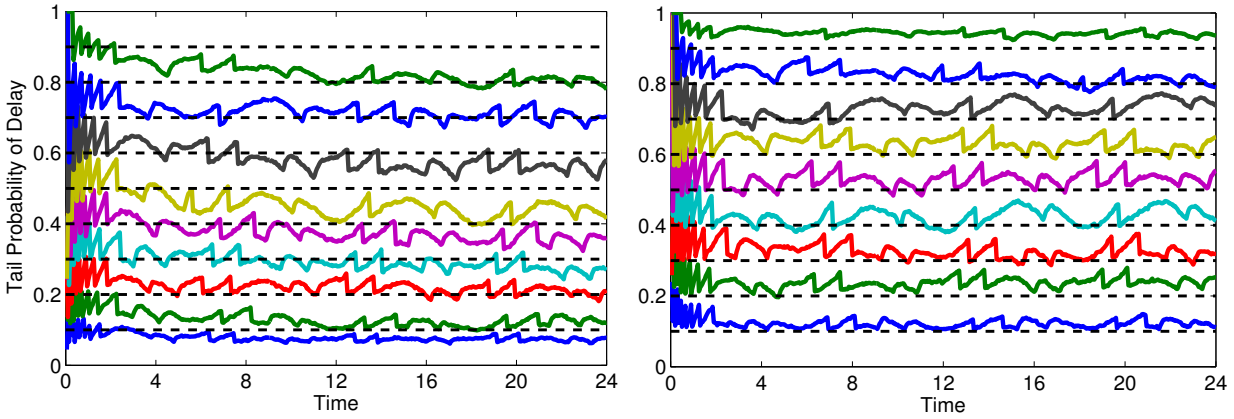

Figure 9: TPoDs of the $H_2(t)/M/s_t + H_2$ model with $w = 6$, $\lambda(t) = 100 + 20\sin t$ using ceiling (left) and flooring (right) for discretization

for the staffing level to increase from 6 to 7 before discretization, and we choose ceiling, then the system will be over staffed for a long time. This naturally makes stabilizing TPoD for large w very difficult.

8 Additional Examples with Real Hospital and Call-Center Data

We consider some additional realistic examples: The first one has the arrival rate estimated from the emergency room records in the SEEStat database (SEE Center [2014]). The second example has arrival rates estimated from a call center of a U.S. bank, obtained from SEEStat center (SEE Center [2014]).

8.1 Another Example with Real-Hospital Arrival Rates

We now consider an $M_t/M/s_t + M$ model with an arrival rate, obtained from the emergency room records in the SEEStat database (SEE Center [2014]), see Figure 10. This arrival rate is computed by averaging hourly arrival rates during weekdays from January 2004 to October 2007. Because the waiting times are long and abandonment is low in hospitals, we set the delay target $w = 2$ hours, mean service time $1/\mu = 2$ hours, mean patience time $1/\theta = 4$ hours.

Figure 10 reports the the TTGA staffing levels and the associated time-dependent TPoDs, with $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$. Despite the drastically changing arrival rate and low staffing levels (e.g., the average staffing level between time 2 and 10 is 3 for $\alpha = 0.9$), we conclude that the TTGA staffing method successfully achieves time-stable performance for TPoD at desired targets for arrival rates estimated from real-hospital data.

8.2 Additional Examples with Real Call-Center Arrival Rates

We now consider another realistic example having arrival rates estimated from a real call center, obtained from SEEStat center (SEE Center [2014]). Comparing to the health care systems, the delay target, mean service and mean patience times are muchh smaller in call centers. Suggested by Feldman et at. [2008], we let both the mean service time and mean patience time be 6 minutes and delay target w be 3 minutes, i.e., $\mu = \theta = 10, w = 0.05$ as we measure the time by hours. Figure 11 reports the arrival rate, TPoD, and staffing functions with $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$. Figure 11

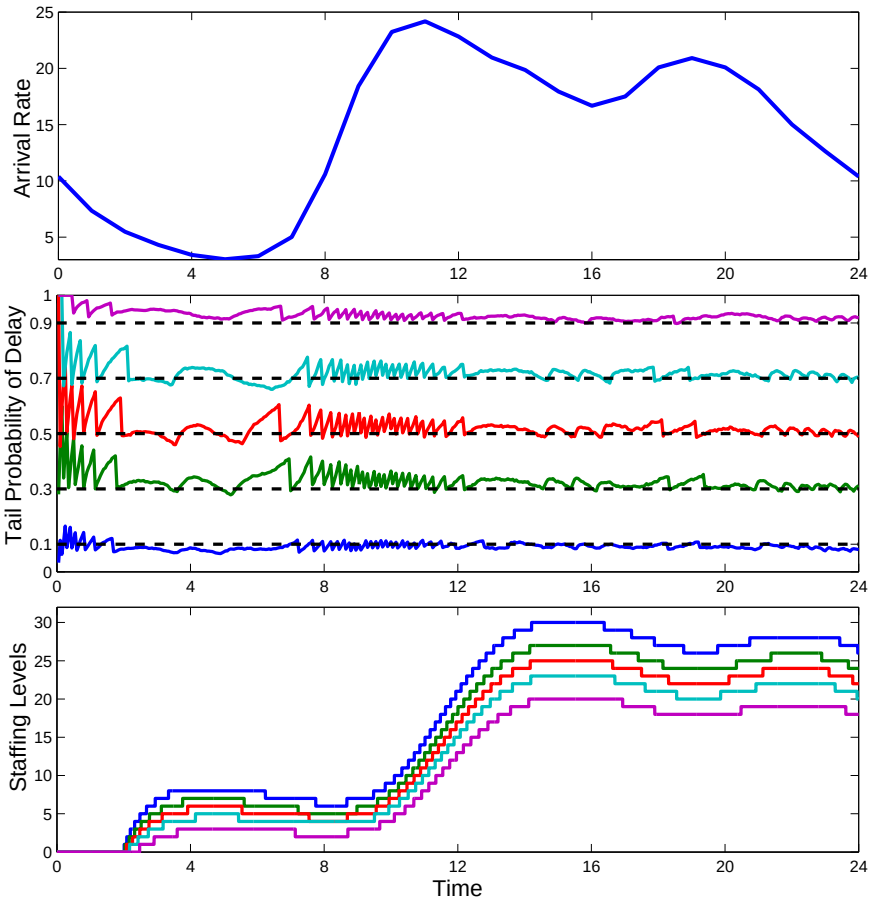

Figure 10: Arrival rate, TPoD, and staffing functions of the EW model, $w = 2$, $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$

once again confirms that the TTGA staffing method can indeed be applied to real service systems, where arrival rates vary significantly (here from 0 to 100).

During the hours 0 to 6, the TPoD cannot be well stabilized because the arrival rate is too small (close to 0) so the call center can only staff either 1 or 0 agent. As the arrival rate rapidly increases after hour 6 (from 0 to 100), the TPoD can be well stabilized. Again, the big fluctuations in TPoD are caused by the smaller system size (note the average staffing levels for the 5 targets are from 2 to 7).

9 Other Arrival Rates

We next the $H_2(t)/M/s_t + H_2$ example in §2 with other arrival rate functions, including

(i) Quadratic: $\lambda(t) = 90 + 5t - 0.15t^2$ (topleft in Figure 12);

(ii) Piecewise constant: $\lambda(t)$ alternates between 80 and 120 in every 5 time units (topright in Figure 12);

(iii) Constant: $\lambda(t) = 100$ (topleft in Figure 13);

(iv) Piecewise linear: $\lambda(t)$ varies linearly between 80 and 120 in every 5 time units (topright in Figure 13);

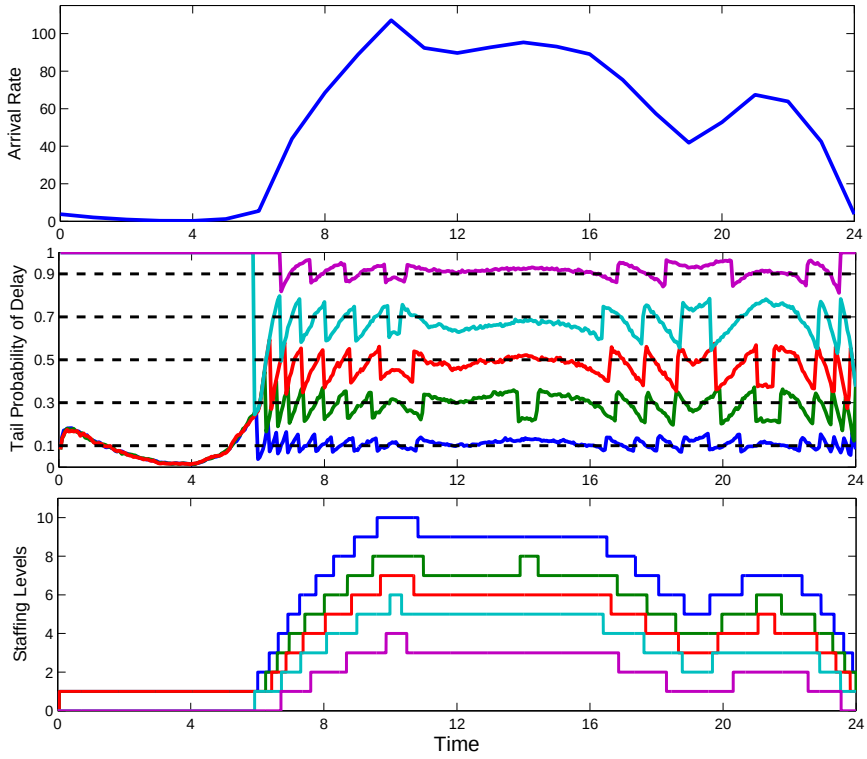

Figure 11: Arrival rate, TPoDs, and TTGA staffing functions for the model with real call-center arrival rate, with $w = 3$ minutes, $\alpha = 0.1, 0.3, \ldots, 0.9$.

(iv) On-and-off: $\lambda(t)$ alternates between 100 and 0 in every 2 time units (Figure 15).

See Figure 12 for Cases (i) and (ii), and Figure 13 for Cases (iii) and (iv). These results show that TTGA continues to perform well. In right-hand plot of Figure 12, we observe that when the arrival rate is low with $\lambda(t) = 80$ (high with $\lambda(t) = 120$), the mean queue length is low (high), but countering to intuition, both the PoA and mean delay are high (low). Similarly, In right-hand plot of Figure 13, when the arrival rate increases (decreases), the mean queue length increases (decreases), but both the PoA and mean delay decrease (increase).

We remark that all performance measures quickly achieve time-stable performances for the case of constant arrival rate (Case (iii)), which is consistent with Corollary 1 in the main paper. Supplementing Corollary 1, the next Corollary provides the long run performance approximation formulas for models with constant arrival rates. Its proof directly follows from Theorem 3.

Corollary 9.1 (*Long-run performance approximation formulas for the $G/M/s_t+GI$ model*) *If the arrival rate is a constant $\bar{\lambda}$, then as $t \to \infty$, the performance approximation formulas in Theorem 3 simplifies to*

$$\tilde{V}(t) \to \mathrm{E}[(w + (\mathcal{Z} - z_\alpha)\sigma_{V^*})^+], \quad \tilde{Q}(t) \to \mathrm{E}[(X^* - s_{w,\alpha} + \sigma_{X^*}\mathcal{Z})^+],$$
$$\tilde{p}_{de}(t) \to \Phi\left(\frac{w}{\sigma_{V^*}} - z_\alpha\right), \quad \tilde{p}_{ab}(t) \to \int_0^\infty \Phi\left(\frac{w-x}{\sigma_{V^*}} - z_\alpha\right) f(x)dx, \quad \textit{and}$$
$$\tilde{u}(t) \to \frac{\mathrm{E}[(X^* + \sigma_{X^*}\mathcal{Z})^+ \wedge s_{w,\alpha}]}{s_{w,\alpha}}, \quad \textit{as} \quad t \to \infty,$$

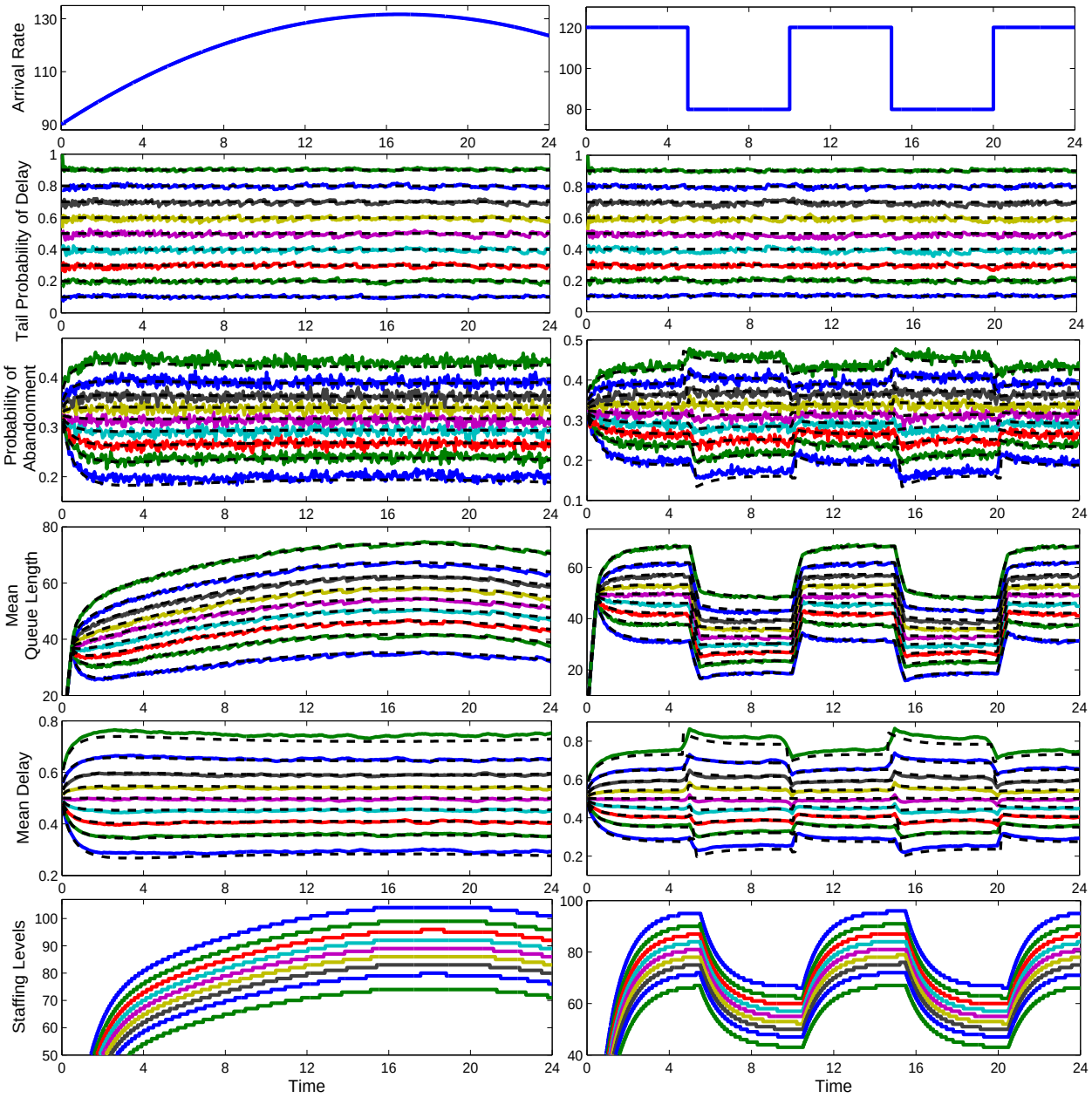

Figure 12: Performance measures of the $H_2(t)/M/s_t + H_2$ model having (i) quadratic arrival rate (left) and (ii) piecewise constant arrival rate, with QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$.

where

$$\sigma_{V^*} = \frac{C}{2\lambda f(w)}, \quad X^* = \lambda \int_0^w \bar{G}(x)dx - z_\alpha \sqrt{\frac{\lambda C \bar{F}(w)}{2h_F(w)}} + s_{w,\alpha}, \tag{1}$$

$$\sigma_{X^*} = \lambda \int_0^w \bar{F}(x)((c_\lambda^2 - 1)\bar{F}(x) + 1)dx + \frac{\lambda C \bar{F}(w)}{2h_F(w)}, \quad s_{w,\alpha} = s_w^{(1)} + \beta_{w,\alpha}\sqrt{s_w^{(1)}} \tag{2}$$

is given in Corolloary 1 and C is defined in Corollary 3.

9.1 Marginal Price of Staffing

We now demonstrate how our analytic TTGA staffing formulas can help estimate the *marginal price of staffing* (MPS), that is, in order to improve the service to a next level (e.g., reducing w or α for Δw by $\Delta\alpha$), how much additional staffing (extra servers) is needed. We address this question by considering the case of constant arrival rate, which represents the average level of the staffing functions. Assuming the density f is differentiable, taking partial derivatives of the staffing formula in (2) with respect to w and α yields

$$-\frac{\partial s_{w,\alpha}}{\partial w} = f(w)\frac{\lambda}{\mu} + \frac{\sqrt{\lambda} z_\alpha [(c_\lambda^2 - 1)(-f^2(w) + \bar{F}(w)\dot{f}(w)) + 2\dot{f}(w)]}{2\mu\sqrt{2f(w)[(c_\lambda^2 - 1)\bar{F}(w) + 2]}}, \tag{3}$$

$$-\frac{\partial s_{w,\alpha}}{\partial \alpha} = \frac{\sqrt{\lambda f(w)[(c_\lambda^2 - 1)\bar{F}(w) + 2]}}{\sqrt{2}\mu\phi(\Phi^{-1}(1-\alpha))}. \tag{4}$$

The above two equations can help estimate the MPS of TTGA. For instance, in order to reduce the delay (probability) target from w to $w - \Delta w$ (from α to $\alpha - \Delta\alpha$), we have to increase the staffing level by adding approximately $-\partial s_{w,\alpha}/\partial w \cdot \Delta w$ ($-\partial s_{w,\alpha}/\partial \alpha \cdot \Delta\alpha$) servers. Using the example considered in the left-hand plot of Figure 13, we plot the partial derivatives in (3) and (4) for $n = 100$, 50 and 10. In the left-hand plot of Figure 14, we fix $\alpha = 0.5$ and let w increase from 0 to 6 with a step size 0.1. It shows that the MPS is monotonically decreasing in w. In the right-hand plot of Figure 14, we fix $w = 0.5$ and let α increase from 0.02 to 0.98 with step size 0.02. We observe that the MPS is high when α is close to 0 or 1 but low when $\alpha \approx 0.5$. For instance, for $\Delta\alpha = 0.1$ and $n = 100$, we need to add to the staffing function $(-\partial s_{0.5,\alpha}/\partial\alpha|_{\alpha=0.5}) \times \Delta\alpha \approx 30 \times 0.1 = 3$ servers if we hope to reduce α from 0.5 to $0.5 - \Delta\alpha = 0.4$. For $n = 10$, we need to only add around $9 \times \Delta\alpha \approx 1$ server in order to reduce α from 0.5 to 0.4. This example also demonstrate the impact of adding one server, which is consistent with results in §5 here and §5.2 of the main paper.

9.2 On-and-Off Arrivals

Unlike the perfectly stabilized TPoDs in Figures 12-13, the example with on-and-off arrivals (with rates alternating between 100 and 0) exhibits some performance degradations. Because the arrival rate jumps drastically by adding or subtracting 100 servers at a time, the required TTGA staffing functions will accordingly increase or decrease extremely fast. Given full staffing flexibility, we can make sure the staffing level increases at desired speed. However, since we do not kick customers out of service before they finish service, our real staffing level cannot decrease as fast as desired. As shown in the last plot of Figure 15, the actual number of servers can be higher than the planned TTGA staffing function (shown in Subplot 3 of Figure 15) by at most 2 servers when the staffing function decreases. As a result, the system becomes inevitably overstaffed as the staffing function decreases. This explains the periodic drops of the TPoDs as shown in Figure 15. Nevertheless, our TTGA method can successfully control the TPoDs at or below the desired targets.

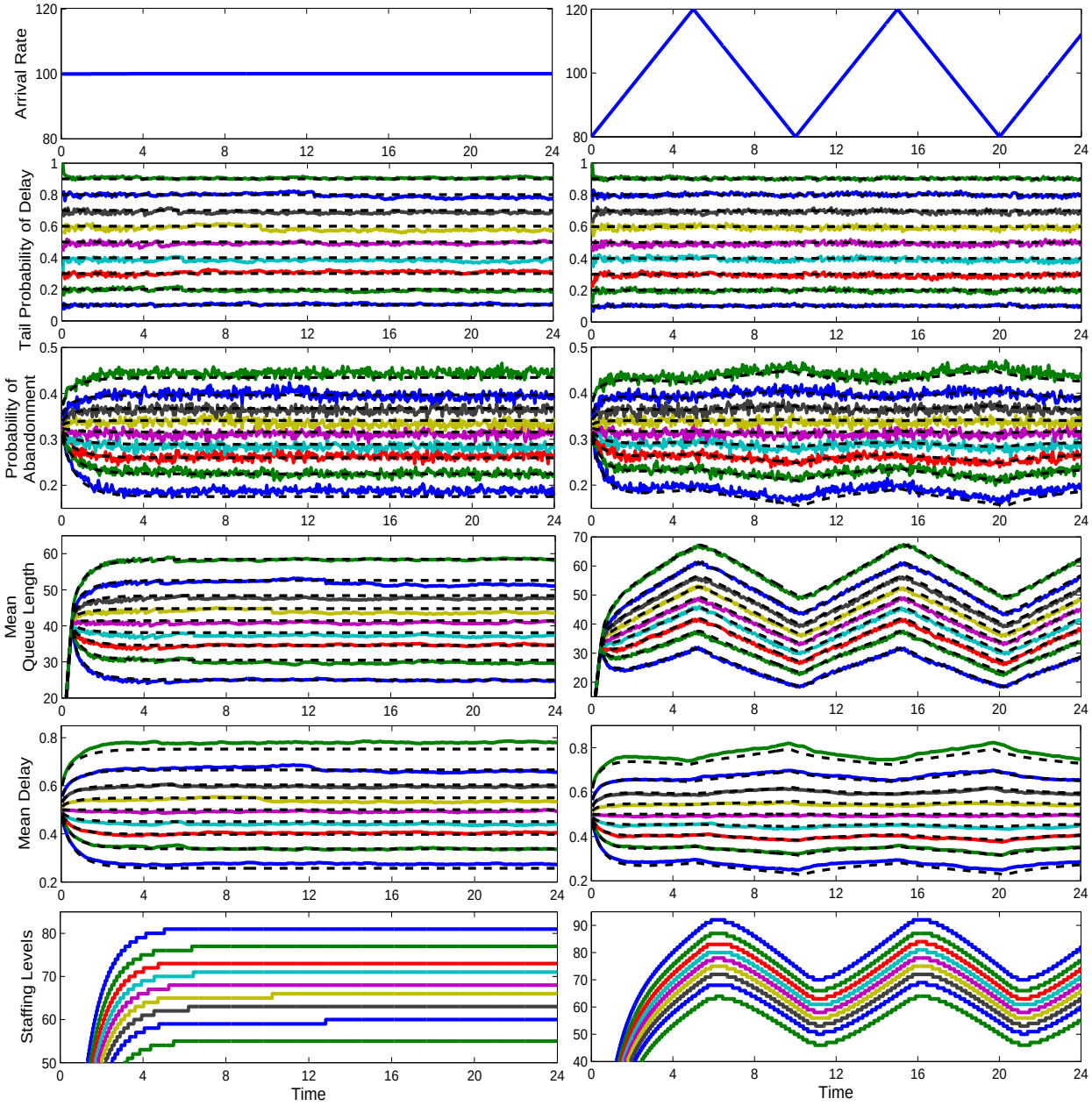

Figure 13: Performance measures of the $H_2(t)/M/s_t + H_2$ model having (i) constant arrival rate (left) and (ii) piecewise linear arrival rate, with QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$.

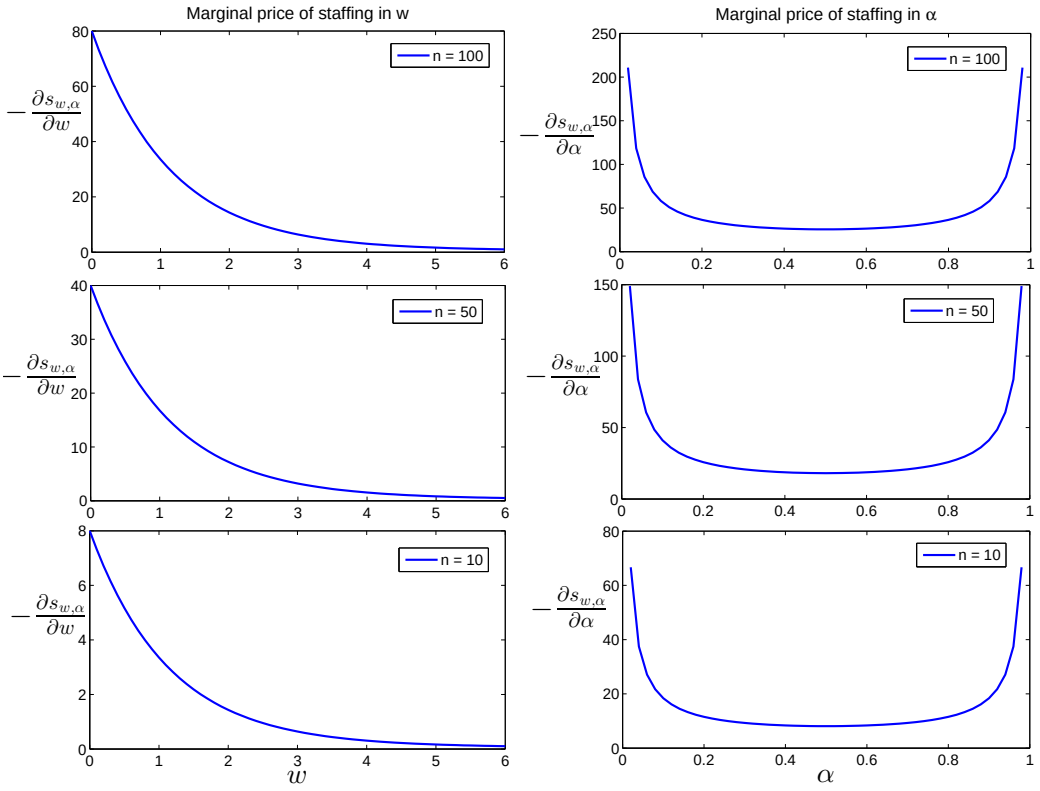

Figure 14: MPS with respect to w (left) and α (right) when the arrival rate is 100 and t is large

10 Non-exponential Service Distribution

We supplement §5.6 of the main paper by considering additional examples with (i) lognormal and (ii) hyperexponential service distributions.

10.1 Lognormal Serivce

First, we provide Figure 16 (an analog of Figure 13 of the main paper) to verify the effectiveness of TTGA for lognormal service times with $c_s^2 = 1$ (i.e., $LN(1,1)$) and $c_s^2 = 0.25$ (i.e., $LN(1, 0.25)$).

Next, we substantiate the important role of c_s^2 for the heuristic TTGA formula for GI service. In particular, we compare the performance of TPoDs for the $H_2(t)/LN(1,4)/s_t + H_2$ example using two staffing levels: (i) the TTGA formula with $c_s^2 = 4$ and (ii) the TTGA formula for M service (i.e., with $c_s^2 = 1$), see Figure 17 for this comparison. We observe that the heuristic staffing formula which incorporate the service SCV c_s^2 indeed achieves time-stable TPoDs, thus outperforming the case with $c_s^2 = 1$. It is not surprising to see that even the case (ii) staffing level achieves somewhat acceptable TPoD performance, because this refinement with c_s^2 only affects the secondary staffing term $s_{w,\alpha}^{(2)}$ so the two versions of TTGA formulas are only slightly different.

10.2 Explaining Where the c_s^2 Comes From When Service Is GI

We now provide more insights into the TTGA staffing formula ((9) in the main paper) and explain where the term c_s^2 comes from. First, the term $I^2(t)$ in (20) of the main paper is obtained by summing three terms $I_\lambda^2(t)$, $I_s^2(t)$ and $I_a^2(t)$. These terms characterize the fluctuations of the

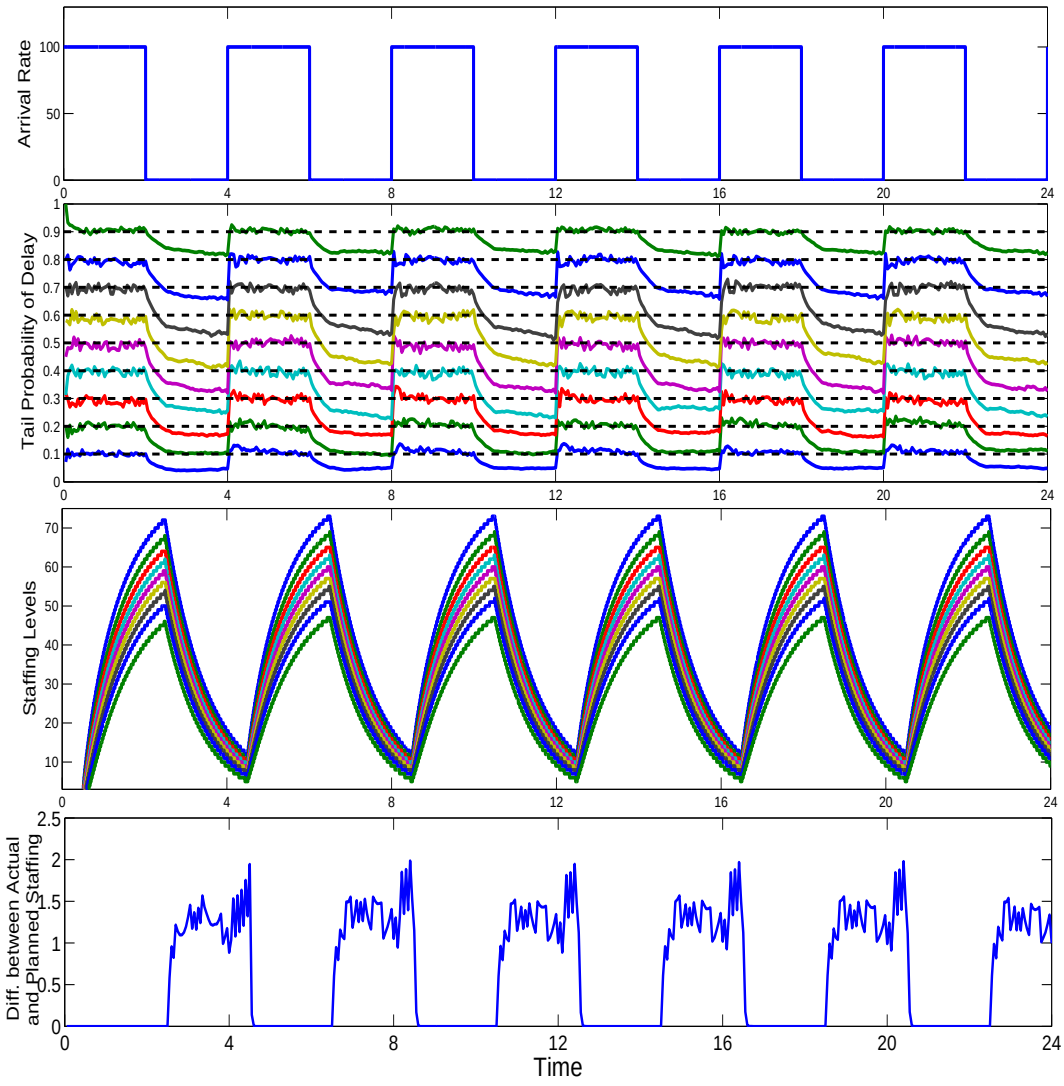

Figure 15: Performance measures of the $H_2(t)/M/s_t + H_2$ model with QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$, and on-and-off arrival rate function

limiting diffusion processes for the $G_t/M/s_t + GI$ model, brought by three independent random processes: arrival, service completion and abandonment. See Theorem 4.2 in Liu and Whitt [2014a] for details. In particular,

$$I_\lambda^2(t) = \frac{c_\lambda^2 F^c(w(t))b(t,0)}{(\lambda(t-w(t))F^c(w(t)))^2}, \quad I_s^2(t) = \frac{b(t,0)-\dot{s}(t)}{(\lambda(t-w(t))F^c(w(t)))^2}, \quad I_a^2(t) = \frac{F^c(w(t))F(w(t))}{(\lambda(t-w(t))F^c(w(t)))^2}.$$

Here the term $I_s^2(t)$ is obtained from the diffusion limit of the NHPP departure process having rate $\mu s(t)$ (because of the assumption of M service). When the service times are nonexponential, we approximate the departure process by a renewal process with time changes, of which the limiting diffusion process is a Brownian motion multiplied by the service SCV c_s. In other words, we use a revised version

$$I_s^2(t) = \frac{c_s^2(b(t,0)-\dot{s}(t))}{(\lambda(t-w(t))F^c(w(t)))^2},$$

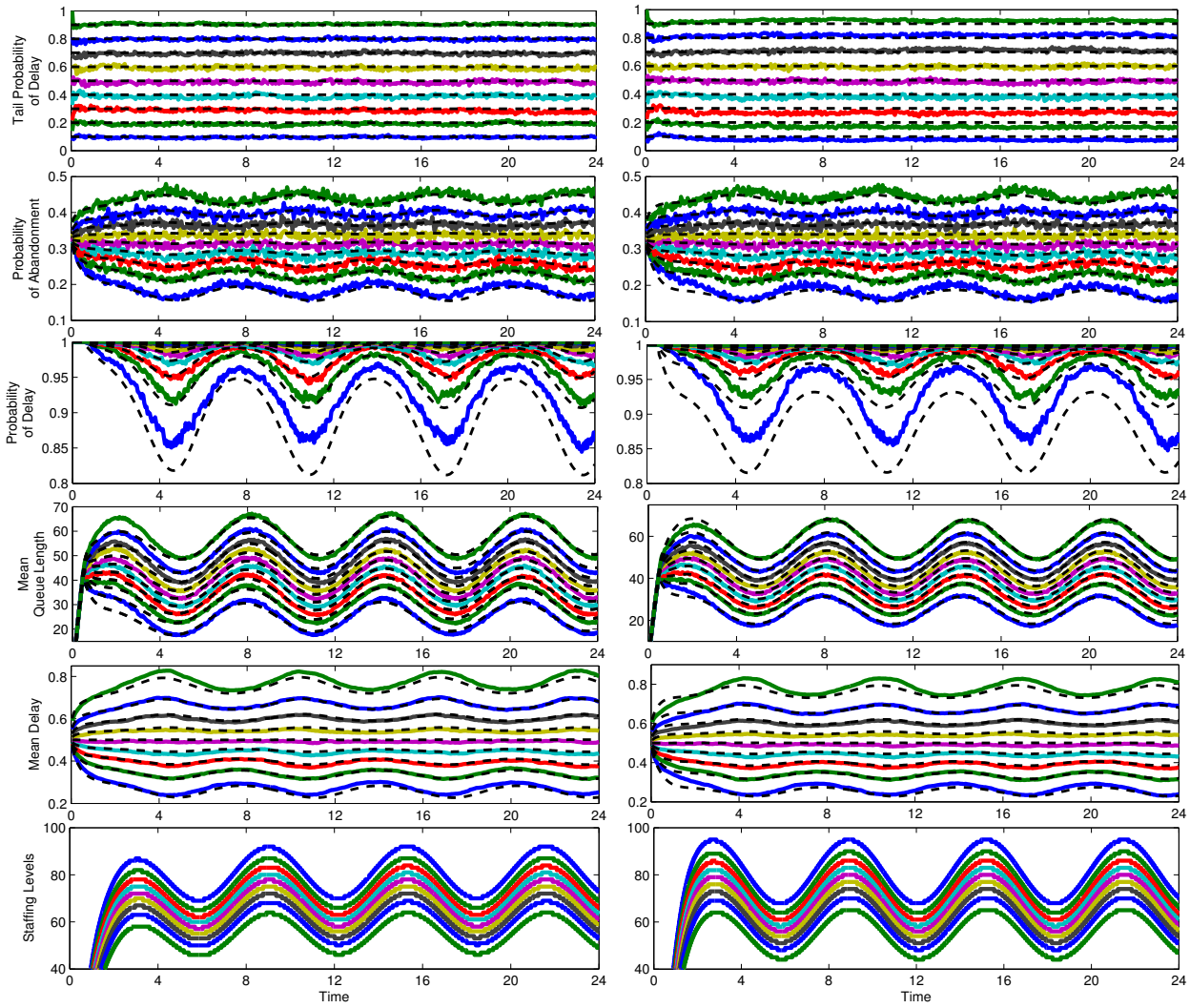

Figure 16: Performance measures of the $H_2(t)/LN(1, c_s^2)/s_t + H_2$ model with arrival rate $\lambda(t) = 100 + 20\sin t$, SCV $c_s^2 = 1$(left) and $c_s^2 = 0.25$(right), QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$,

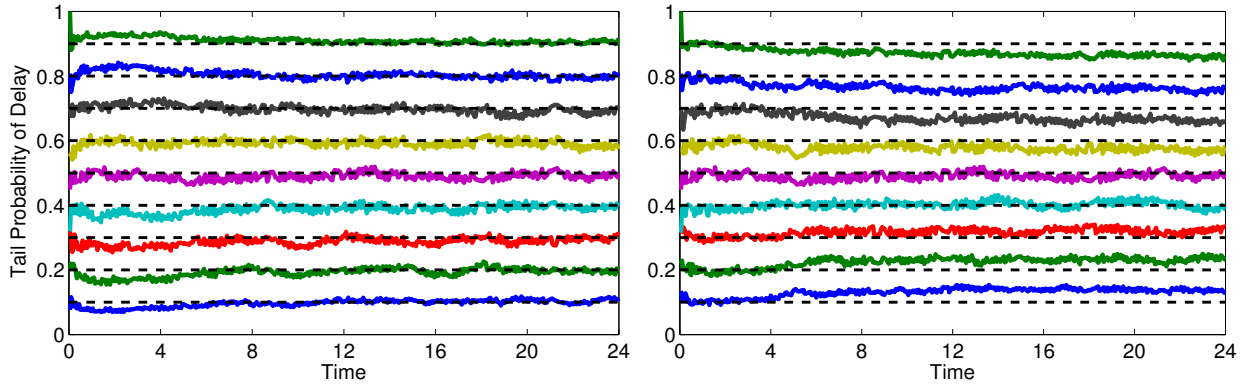

Figure 17: Comparison of TPoD of the $H_2(t)/LN(1,4)/s_t + H_2$ model with arrival rate $\lambda(t) = 100 + 20\sin t$, SCV $c_s^2 = 4$, QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$, when using (i) TTGA staffing function with $c_s^2 = 4$ (left), and (ii) TTGA staffing function for exponential service with $c_s^2 = 1$ (right)

which leads to the general TTGA formula in (10) of the main paper. This heuristic refinement is based on our understanding of what the more general FCLT with GI service will be. Of course, it remains to provide rigorous MSHT limit theorems for the $G_t/GI/s_t + GI$ model, see Liu and Whitt [2014b] for recent developments.

10.3 Hyperexponential Service

To end this section, we repeat the $H_2(t)/GI/s_t + H_2$ example by letting the service distribution be H_2 with cdf

$$G(x) = 1 - p_s e^{-\mu_1 x} - (1 - p_s)e^{-\mu_2 x},$$

where $\mu_1 = 2p_s\mu, \mu_2 = 2(1-p_s)\mu, \mu = 1, p_s = (5+\sqrt{15})/10$, so that $c_s^2 = 4$ and $\mathrm{E}[S] = 1$. Figure 18 shows that the TPoD is stabilized and other performance measures are well approximated, except for a warm-up period.

11 Additional Proofs

11.1 Proof of Theorem 1

The proof of Theorem 1 follows from the proof of Theorem 2 in Liu and Whitt [2012a]. Theorem 2 in Liu and Whitt [2012a] establishes the asymptotic stability of the DIS staffing function for achieving the constant mean delay, by considering the $M_t/GI/s_t + GI$ model. In particular, it states that for the n^{th} $M_t/GI/s_t + GI$ model having arrival rate $\lambda_n(t) \equiv n\lambda(t)$, if the staffing level $s_n(t) = \lceil n\, s_w^{(1)}(t) \rceil$ with $s_w^{(1)}(t)$ given in (7) of the main paper, then

$$\sup_{0<t\leq T} |E[W_n(t)] - w| \to 0, \quad \text{as} \quad n \to \infty.$$

The proof of Theorem 2 in Liu and Whitt [2012a] applies (i) the FWLLN result for the $G_t/GI/s_t + GI$ developed in Liu and Whitt [2012c], and (ii) the staffing formula to stabilize the fluid waiting time for the $G_t/GI/s_t + GI$ fluid model developed in Theorem 8 in Liu and Whitt [2012a]. Since both results (i) and (ii) allow G_t arrival process, we can quickly generate Theorem 2 to the case of G_t arrival. Thus the proof of Theorem 1 in the main paper is completed.

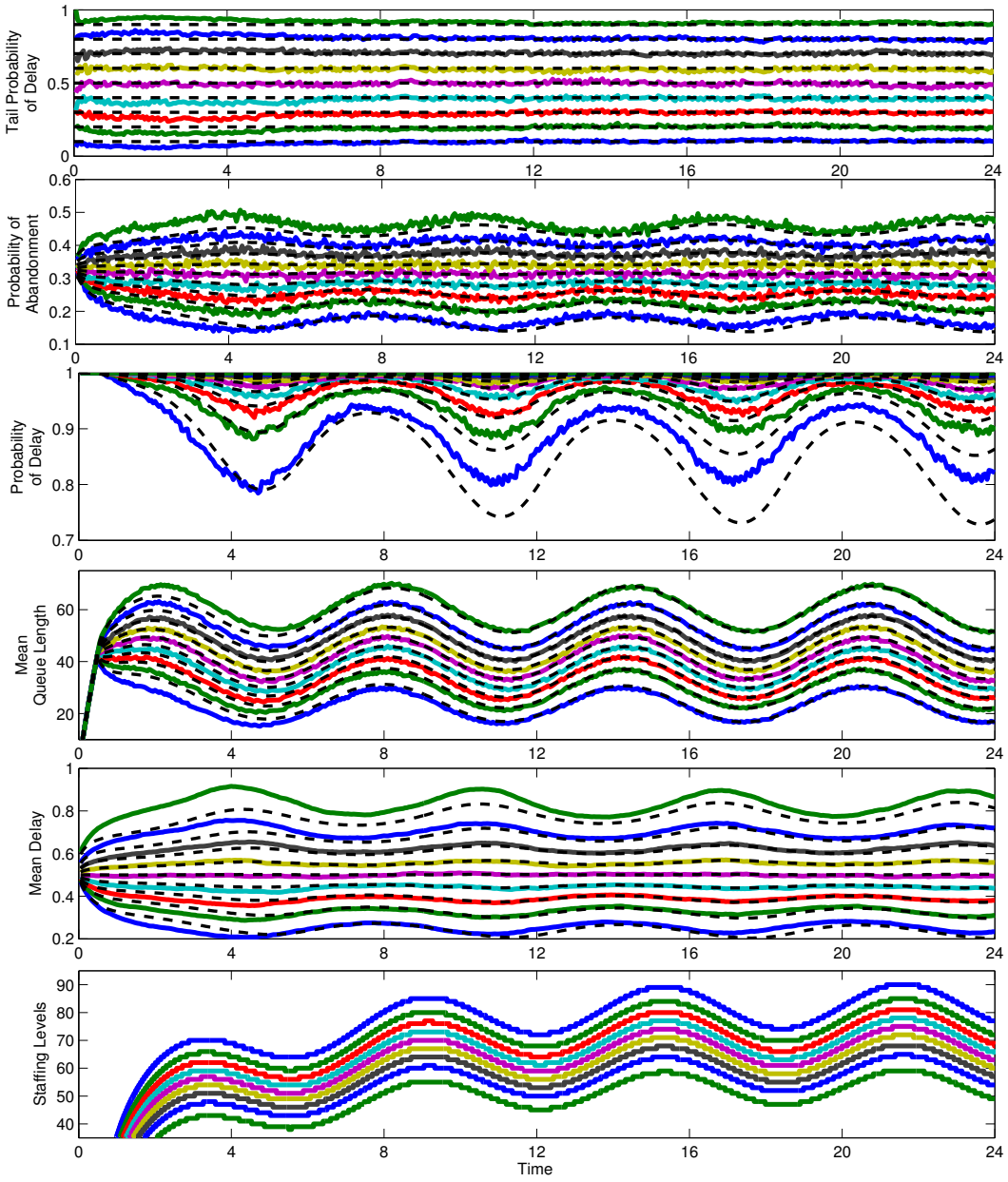

Figure 18: Performance measures of the $H_2(t)/H_2/s_t + H_2$ model with arrival rate $\lambda(t) = 100 + 20\sin t$, SCV $c_s^2 = 4$, QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$,

11.2 Proofs of corollaries

Proof of Corollary 1. When $\lambda(t) = \lambda$, (2) in the main paper implies that

$$s_w^{(1)}(t) = \bar{F}(w) \int_0^{t-w} \lambda \bar{G}(x) dx = \frac{\bar{F}(w)\lambda}{\mu}(1 - e^{-\mu(t-w)}) \tag{5}$$

Letting $t \to \infty, s_w^{(1)}(t) \sim \bar{F}(w)\lambda/\mu$. Define

$$C \equiv (c_\lambda^2 - 1)\bar{F}(w) + 2. \tag{6}$$

Then from (5) and (10) in the main paper, we have

$$\begin{aligned} e^{-\mu t} Z(t) &= e^{-h_F(w)t} \sqrt{\int_w^t e^{2h_F(v)x} \bar{F}(w)\lambda \left(C - e^{-\mu(x-v)}\right) dx} \\ &= \sqrt{\bar{F}(w)\lambda} \sqrt{\frac{C}{2h_F(w)}(1 - e^{2h_F(w)(w-t)}) - \frac{e^{\mu w}}{2h_F(w) - \mu}(e^{-\mu t} - e^{(2h_F(w)-\mu)w - 2h_F(w)t})} \\ &\sim \sqrt{\frac{C\bar{F}(w)\lambda}{2h_F(w)}} \end{aligned} \tag{7}$$

Similarly, we have

$$\lim_{t\to\infty} e^{-\mu t} \int_w^t Z(u) du = \lim_{t\to\infty} \frac{Z(t)}{\mu e^{\mu t}} = \frac{1}{\mu} \sqrt{\frac{C\bar{F}(w)\lambda}{2h_F(w)}} \tag{8}$$

Combining (7) and (8) yields that

$$\begin{aligned} s_{w,\alpha}^{(2)}(t) &= z_\alpha \left(e^{-\mu t} Z(t) - (\mu - h_F(w)) e^{-\mu t} \int_w^t Z(u) du \right) \cdot \mathbf{1}_{\{t \geq w\}} \\ &\sim \frac{z_\alpha h_F(w)}{\mu} \sqrt{\frac{C\bar{F}(w)\lambda}{2h_F(w)}} = z_\alpha \sqrt{[(c_\lambda^2 - 1)\bar{F}(w) + 2] h_F(w) s_w^{(1)} / 2\mu}. \end{aligned}$$

Proof of Corollary 2. NHPP arrival implies that $c_\lambda = 1$. Hence,

$$(s_{w,\alpha}^{(2)}(t))^2 = z_\alpha^2 \, e^{-2\mu t} \int_w^t e^{2h_F(w*)x} (2\mu s_w^{(1)}(x) + \dot{s}_w^{(1)}(x)) dx = z_\alpha^2 \, e^{-2\mu t} \int_w^t d(e^{2\mu x} s_v^{(1)}(x)) = z_\alpha^2 \, s_w^{(1)}(t),$$

where the forth equality holds because $s_w^{(1)}(w) = 0$.

Proof of Corollary 3. If arrival rate is $\lambda(t) = \bar{\lambda}(1 + r\sin(\gamma t + \phi)$, and let $C \equiv (c_\lambda^2 - 1)\bar{F}(w) + 1$, $\varphi \equiv \arctan(\gamma/\mu)$, and $\eta \equiv \varphi + \arctan(\gamma/(2h_F(w))$. We have

$$\begin{aligned}
s_w^{(1)}(t) &= \bar{F}(w)\int_0^{t-w} e^{-\mu x}(\bar{\lambda} + r\bar{\lambda}\sin\gamma(t+\phi/\gamma - w - x))dx \\
&= \bar{\lambda}\bar{F}(w)\left(\int_0^{t-w} e^{-\mu x}dx + r\int_0^{t-w} e^{-\mu x}\sin\gamma(t+\phi/\gamma - w - x)dx\right) \\
&= \bar{\lambda}\bar{F}(w)\left[\frac{1-e^{-\mu(t-w)}}{\mu} - r\left(\frac{\gamma e^{-\mu(t-w)}}{\mu^2+\gamma^2}(\mu\sin\phi - \gamma\cos\varphi)\right.\right. \\
&\qquad \left.\left. -\frac{1}{\mu^2+\gamma^2}(\mu\sin\gamma(t+\phi/\gamma - w) - \gamma\cos\gamma(t+\phi/\gamma - w))\right)\right] \\
&= \bar{\lambda}\bar{F}(w)\left[\frac{1-e^{-\mu(t-w)}}{\mu} - \frac{r}{\sqrt{\mu^2+\gamma^2}}\left(e^{-\mu(t-w)}\sin(\phi-\varphi) - \sin(\gamma(t+\phi/\gamma - w) - \varphi)\right)\right] \qquad (9) \\
&\sim \bar{\lambda}\bar{F}(w)\left(\frac{1}{\mu} + \frac{r}{\sqrt{\mu^2+\gamma^2}}\sin(\gamma(t+\phi/\gamma - w) - \varphi)\right)
\end{aligned}$$

Next, we compute

$$\begin{aligned}
&\int_w^t e^{2h_F(w)x}\left(\left[(c_\lambda^2-1)\bar{F}(w)+2\right](\mu s_w^{(1)}(x) + \dot{s}_w^{(1)}(x)) - \dot{s}_w^{(1)}(x)\right)dx \\
&= \int_w^t e^{2h_F(w)x}(\mu C s_w^{(1)}(x) + (C-1)\dot{s}_w^{(1)}(x))dx \\
&= \mu C\int_w^t e^{2h_F(w)x}s_w^{(1)}(x)dx + (C-1)\int_w^t e^{2h_F(w)x}\dot{s}_w^{(1)}(x)dx \qquad (10)
\end{aligned}$$

We compute the two integrands in (10) respectively. For the first integrand,

$$\begin{aligned}
&\int_w^t e^{2h_F(w)x}s_w^{(1)}(x)dx \\
&= \int_w^t e^{2h_F(w)x}\bar{\lambda}\bar{F}(w)\left[\frac{1-e^{-\mu(x-w)}}{\mu} - \frac{r}{\sqrt{\mu^2+\gamma^2}}\left(e^{-\mu(x-w)}\sin(\phi-\varphi) - \sin(\gamma(x+\phi/\gamma - w) - \varphi)\right)\right]dx \\
&= \bar{\lambda}\bar{F}(w)\left(\frac{1}{\mu}\int_w^t e^{2h_F(w)x}dx + \frac{r}{\sqrt{\mu^2+\gamma^2}}\int_w^t e^{2h_F(w)x}\sin(\gamma(x+\phi/\gamma - w) - \varphi)dx -\right. \\
&\qquad \left. e^{\mu w}(\frac{1}{\mu} - \frac{r\sin(\phi-\varphi)}{\sqrt{\mu^2+\gamma^2}})\int_w^t e^{(2h_F(w)-\mu)x}dx\right) \\
&= \bar{\lambda}\bar{F}(w)\left(\frac{e^{2h_F(w)t} - e^{2h_F(w)w}}{2\mu h_F(w)} + \frac{r(e^{2h_F(w)t}\sin(\gamma(t+\phi/\gamma - w) - \eta) - e^{2h_F(w)w}\sin(\phi-\eta))}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}} -\right. \\
&\qquad \left.\left(\frac{1}{\mu} - \frac{r\sin(\phi-\varphi)}{\sqrt{\mu^2+\gamma^2}}\right)\frac{e^{\mu w}\left(e^{(2h_F(w)-\mu)t} - e^{(2h_F(w)-\mu)w}\right)}{2h_F(w)-\mu}\right) \qquad (11)
\end{aligned}$$

For the second integrand, we have

$$\int_w^t e^{2h_F(w)x}\dot{s}_w^{(1)}(x)dx = \int_w^t e^{2h_F(w)x}ds_w^{(1)}(x) = e^{2h_F(w)t}s_w^{(1)}(t) - 2h_F(w)\int_w^t e^{2h_F(w)(x)}s_w^{(1)}(x)dx. \tag{12}$$

Note that the integrand in the second term of (12) coincide with the first integrand in (10), which is calculated in (11). Next, we establish the convergence of $e^{-\mu t}Z(t)$ as $t \to \infty$.

$$\begin{aligned}
e^{-\mu t}Z(t) &= e^{-h_F(w)t}\sqrt{\mu C\int_w^t e^{2h_F(w)x}s_w^{(1)}(x)dx + (C-1)\int_w^t e^{2h_F(w)x}\dot{s}_w^{(1)}(x)dx} \\
&= \sqrt{\begin{array}{l}\bar{\lambda}\bar{F}(w)(\mu C-(C-1)2h_F(w))\left(\dfrac{(1-e^{2h_F(w)(w-t)})}{2\mu h_F(w)}\right. \\ +\dfrac{r(\sin(\gamma(t+\phi/\gamma-w)-\eta)-e^{2h_F(w)(w-t)}\sin(\phi-\eta))}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}} \\ \left.-e^{-2h_F(w)t}(\dfrac{1}{\mu}-\dfrac{r\sin(\phi-\varphi)}{\sqrt{\mu^2+\gamma^2}})\dfrac{e^{\mu w}(e^{(2h_F(w)-\mu)t}-e^{(2h_F(w)-\mu)w})}{2h_F(w)-\mu}\right)+(C-1)s_w^{(1)}(t)\end{array}} \\
&\sim \sqrt{\begin{array}{l}\bar{\lambda}\bar{F}(w)(\mu C-(C-1)2h_F(w))\left(\dfrac{1}{2\mu h_F(w)}+\dfrac{r\sin(\gamma(t+\phi/\gamma-w)-\eta)}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}}\right) \\ +(C-1)\bar{\lambda}\bar{F}(w)\left(\dfrac{1}{\mu}+\dfrac{r}{\sqrt{\mu^2+\gamma^2}}\sin(\gamma(t+\phi/\gamma-w)-\varphi)\right)\end{array}}
\end{aligned} \tag{13}$$

Similarly to (8), $e^{-\mu t}\int_w^t Z(u)du \sim (1/\mu)\lim_{t\to\infty} e^{-\mu t}Z(t)$. Therefore,

$$\begin{aligned}
s_{w,\alpha}^{(2)}(t) &\sim \frac{z_\alpha h_F(w)}{\mu}\sqrt{\begin{array}{l}\bar{\lambda}\bar{F}(w)(\mu C-(C-1)2h_F(w))\left(\dfrac{1}{2\mu h_F(w)}+\dfrac{r\sin(\gamma(t+\phi/\gamma-w)-\eta)}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}}\right) \\ +(C-1)\bar{\lambda}\bar{F}(w)\left(\dfrac{1}{\mu}+\dfrac{r}{\sqrt{\mu^2+\gamma^2}}\sin(\gamma(t+\phi/\gamma-w)-\varphi)\right)\end{array}} \\
&= \frac{z_\alpha f(w)\sqrt{\bar{\lambda}}}{\mu\sqrt{\bar{F}(w)}}\sqrt{\begin{array}{l}(\mu C-(C-1)2h_F(w))\left(\dfrac{1}{2\mu h_F(w)}+\dfrac{r\sin(\gamma(t+\phi/\gamma-w)-\eta)}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}}\right) \\ +(C-1)\left(\dfrac{1}{\mu}+\dfrac{r}{\sqrt{\mu^2+\gamma^2}}\sin(\gamma(t+\phi/\gamma-w)-\varphi)\right)\end{array}} \\
&= \frac{z_\alpha f(w)\sqrt{\bar{\lambda}}}{\mu\sqrt{\bar{F}(w)}}\sqrt{\begin{array}{l}\dfrac{C}{2h_F(w)}+\dfrac{r(\mu C-2(C-1)h_F(w))}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}}\sin(\gamma(t+\phi/\gamma-w)-\eta) \\ +\dfrac{r(C-1)}{\sqrt{\mu^2+\gamma^2}}\sin(\gamma(t+\phi/\gamma-w)-\varphi)\end{array}}
\end{aligned} \tag{14}$$

12 Explicit Expressions for Approximating Formulas in §4

To facilitate computations of the approximating performance functions in Theorem 3, we simplify these formulas and provide explicit expressions in terms of the Gaussian pdf ϕ and cdf Φ, especially for $\mathrm{E}[V(t)], \mathrm{E}[Q(t)]$, and $u(t)$. To calculate the expectation and variance of a Gaussian random variable X truncated below at 0 and above $a > 0$, we have the following formula

$$\mathrm{E}[X^+ \wedge a] = a\Phi\left(\frac{\mu - a}{\sigma}\right) + \mu\left(\Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(-\frac{\mu}{\sigma}\right)\right) + \sigma\left(\phi\left(\frac{\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)\right)$$

Specifically, if $a = +\infty, \mathrm{E}[X^+] = \mu\Phi\left(\frac{\mu}{\sigma}\right) + \sigma\phi\left(\frac{\mu}{\sigma}\right)$. Therefore, we have

$$\begin{aligned}
\mathrm{E}[V(t)] &= (w - z_\alpha\sigma_{V^*}(t))\Phi\left(\frac{w}{\sigma_{V^*}(t)} - z_\alpha\right) + \sigma_{V^*}(t)\phi\left(\frac{w}{\sigma_{V^*}(t)} - z_\alpha\right) \\
\mathrm{E}[Q(t)] &= (X^*(t) - s_{w,\alpha}(t))\Phi\left(\frac{X^*(t) - s_{w,\alpha}(t)}{\sigma_{X^*}(t)}\right) + \sigma_{X^*}(t)\phi\left(\frac{X^*(t) - s_{w,\alpha}(t)}{\sigma_{X^*}(t)}\right) \\
u(t) &= s_{w,\alpha}(t)\Phi\left(\frac{X^*(t) - s_{w,\alpha}(t)}{\sigma_{X^*}(t)}\right) + X^*(t)\left(\Phi\left(\frac{s_{w,\alpha}(t) - X^*(t)}{\sigma_{X^*}(t)}\right) - \Phi\left(-\frac{X^*(t)}{\sigma_{X^*}(t)}\right)\right) \\
&\quad + \sigma_{X^*}(t)\left(\phi\left(\frac{X^*(t)}{\sigma_{X^*}(t)}\right) - \phi\left(\frac{s_{w,\alpha}(t) - X^*(t)}{\sigma_{X^*}(t)}\right)\right)
\end{aligned}$$

13 Staffing Formula of the Main Example

$s_w^{(1)}(t)$ is the same as (9), $s_{w,\alpha}^{(2)}(t)$ is the same as (9) in the main paper where

$$Z(t) = e^{\mu t}\sqrt{\begin{aligned}
&\bar{\lambda}\bar{F}(w)(\mu C - (C-1)2h_F(w))\left(\frac{(1 - e^{2h_F(w)(w-t)})}{2\mu h_F(w)}\right. \\
&+ \frac{r(\sin(\gamma(t + \phi/\gamma - w) - \eta) - e^{2h_F(w)(w-t)}\sin(\phi - \eta))}{\sqrt{(\mu^2 + \gamma^2)(4h_F^2(w) + \gamma^2)}} \\
&\left. - e^{-2h_F(w)t}\left(\frac{1}{\mu} - \frac{r\sin(\phi - \varphi)}{\sqrt{\mu^2 + \gamma^2}}\right)\frac{e^{\mu w}(e^{(2h_F(w)-\mu)t} - e^{(2h_F(w)-\mu)w})}{2h_F(w) - \mu}\right) + (C-1)s_w^{(1)}(t)
\end{aligned}}$$

and φ, η are defined as in Corollary 3.

14 Implementation Details

All numerical calculations and simulations are implemented in MATLAB. We sample the values of the performance functions at fixed time points $\Delta T, 2\Delta T, \ldots, N\Delta T = T$ where $T = 24$ is the length of the time interval, $\Delta T = 0.05$, and $N = T/\Delta T = 480$ is the total number of samples in $[0, T]$.

In each simulation replication r, if a customer arrives at time τ and enters service at time t, the potential waiting time at τ is $V^r(\tau) = t - \tau$. Let $B^r(\tau)$ and $Q^r(\tau)$ be the number of customers waiting in queue and in service at time τ. The mean delay and mean queue length at each time τ are estimated by the averages of $V^r(\tau)$ and $Q^r(\tau)$ over all 5000 replications. We estimate the TPoD and PoD at time τ using the average of the indicator variable $\mathbf{1}_{\{V^r(\tau)>w\}}$ and $\mathbf{1}_{\{V^r(\tau)>0\}}$. The service utilization is estimated by the average of the ratio $B^r(\tau)/s(\tau)$.

References

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

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*. **54**(2) 324–338.

Liu, Y., W. Whitt. 2012a. Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research*. **60**(6) 1551–1564.

Liu, Y., W. Whitt. 2012b. The $G_t/GI/s_t+GI$ Many-Server Fluid Queue. *Queueing Systems*. **71**(4) 405–444.

Liu, Y., W. Whitt. 2012c. A Many-Server Fluid Limit for the $G_t/GI/s_t+GI$ Queueing Model experiencing Periods of Overloading. *Operations Research Letters*. **40** 307–312.

Liu, Y., W. Whitt. 2014a. Many-Server Heavy-Traffic Limit for Queues with Time-Varying Parameters. *Annals of Applied Probability*. **24**(1) 378–421.

Liu, Y., W. Whitt. 2014b. Many-Server FCLT Limits for the $G_t/GI/s_t+GI$ Queue. Working paper.

SEE Center, Technion. 2014 SEEStat database. URL `http://seeserver.iem.technion.ac.il/see-terminal/.`