E-Companion

This appendix provides additional supplementary materials to the main paper. In §EC.1, we give the performance of the simulation-based staffing algorithm developed in Defraeye and van Nieuwenhuyse (2013) (Figure 5 there). In §EC.2 and §EC.2.4, we confirm that the TTGA staffing works well for other real-world examples, including the arrival rate estimated from the emergency room records and a call center of a U.S. bank, obtained from SEEStat center (SEE Center (2014)), and examples having other arrival-rate functions (e.g., constant, piecewise linear and on-and-off arrival rates). In §EC.3 we provide additional proofs. In §EC.4 we give explicit staffing formulas of the main example and more detailed expressions of the performance approximation formulas (12). Simulation details are reported in §EC.5. Additional materials appear in a longer appendix.

EC.1. Performance of the Simulation-Based Staffing Algorithm in Defraeye and van Nieuwenhuyse (2013)

To give a direct comparison of the performance of TTGA with the performance of the simulation-based staffing algorithm developed in Defraeye and van Nieuwenhuyse (2013), we hereby consider an example with realistic arrival rate estimated from the emergency department of a Belgian hospital (top panel of Figure 8). Figure EC.1 (which is Figure 5 in Defraeye and van Nieuwenhuyse (2013)) shows the good performance of the simulation-based staffing method, which is a self correcting procedure. By comparing the TTGA performance in Figure 8 to Figure EC.1, we conclude that we can achieve the same goal without needing simulations.

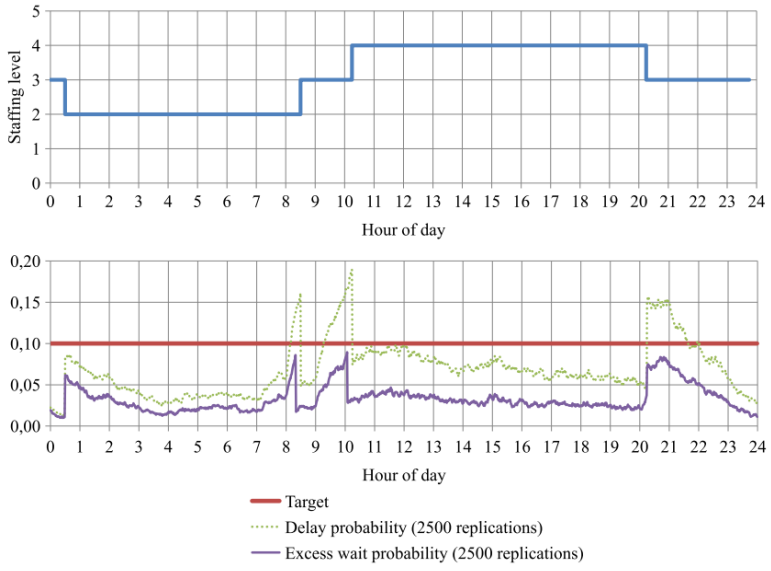

Figure EC.1 Performance of the simulation-based staffing function of Defraeye and van Nieuwenhuyse (2013) (Figure 5 there).

EC.2. Additional Examples

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)).

EC.2.1. Performance Table for the High QoS Example

We quantify the good performance shown in Figure 9 by computing the minimum, maximum and average values of the TPoD performance and compare them to the TPoD targets in Table EC.1.

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 EC.1 Comparison with TPoD targets (average, min and max), with $w = 0.05$.

EC.2.2. 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 EC.2. 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 EC.2 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. In §EC.2 of the online supplement, we apply TTGA to realistic call-center examples having arrival rates estimated from real-call center data with the famous 80-20 rule; there we also show that TTGA works well for the challenging Belgian hospital example considered in Defraeye and van Nieuwenhuyse (2013).

EC.2.3. 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

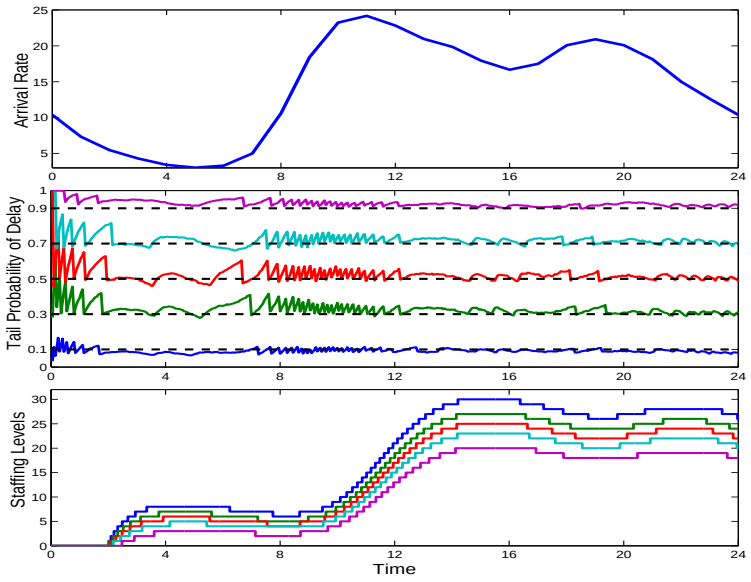

Figure EC.2 Arrival rate, TPoD, and staffing functions of the EW model, $w=2$, $\alpha=0.1,0.3,0.5,0.7,0.9$

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 EC.3 reports the arrival rate, TPoD, and staffing functions with $\alpha=0.1,0.3,0.5,0.7,0.9$. Figure EC.3 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).

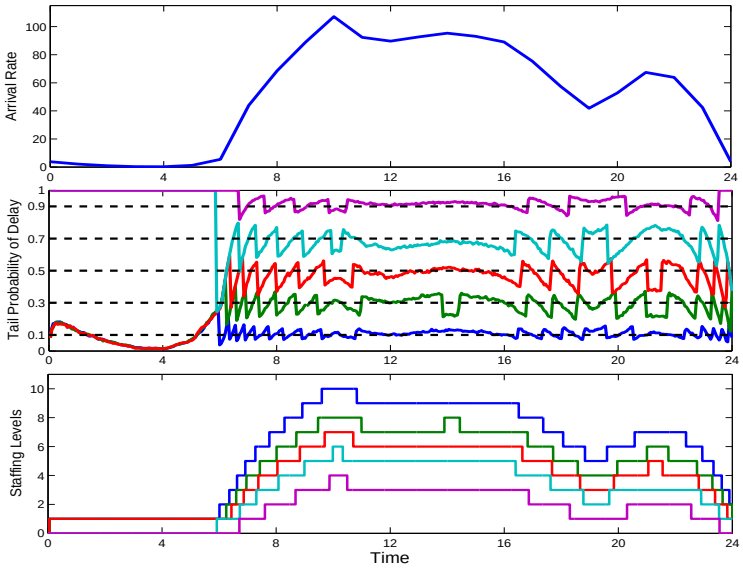

Figure EC.3 Arrival rate, TPoDs, and TTGA staffing functions for the model with real call-center arrival rate, with $w=3$ minutes, $\alpha=0.1,0.3,\ldots,0.9$.

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).

EC.2.4. Other Arrival Rates

We next consider the $H_2(t)/M/s_t+H_2$ example in §4 with other arrival rate functions, including

(i) Quadratic: $\lambda(t)=90+5t-0.15t^2$ (Figure 12 of the longer appendix);

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

(iii) Constant: $\lambda(t)=100$ (see Figure 13 of the longer appendix);

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

(iv) On-and-off: $\lambda(t)$ alternates between 100 and 0 in every 2 time units (Figure EC.5).

See Figures 12 and 13 of the longer appendix for Cases (i), (ii), (iii) and (iv). These results show that TTGA continues to perform well. In right-hand plot of Figure 12 of the longer appendix, 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, 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), see Figure 12 of the longer appendix), 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 EC.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*

$$\tilde{V}(t)\to \mathrm{E}[(w+(\mathcal{Z}-z_\alpha)\sigma_{V^*})^+],\quad \tilde{Q}(t)\to \mathrm{E}[(X^*-s_{w,\alpha}+\sigma_{X^*}\mathcal{Z})^+],\quad \tilde{p}_{de}(t)\to \Phi\left(\frac{w}{\sigma_{V^*}}-z_\alpha\right),$$

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

where

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

$$\sigma_{X^*}=\lambda\int_0^w \bar{F}(x)((c_\lambda^2-1)\bar{F}(x)+1)dx+\frac{\lambda C\bar{F}(w)}{2h_F(w)},$$

$s_{w,\alpha}=s_w^{(1)}+\beta_{w,\alpha}\sqrt{s_w^{(1)}}$ *is given in Corollary 1 and C is defined in Corollary 3.*

EC.2.4.1. 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 by Δw or α 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 Corollary EC.1 with respect to w and α yields

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

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

The above two equations can help estimate the MPS of TTGA. For instance, in order to reduce the

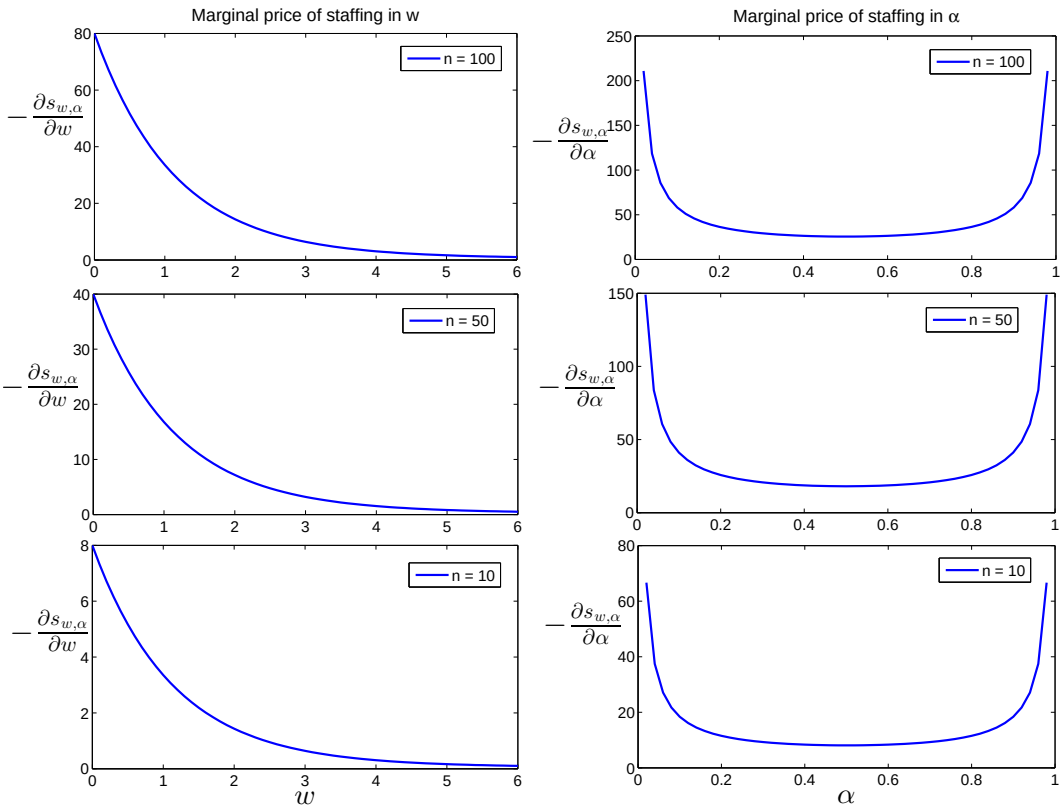

Figure EC.4 MPS with respect to w (left) and α (right) when the arrival rate is 100 and t is large

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 with a constant arrival rate in case (iii), we plot the partial derivatives in (EC.1) and (EC.2) for $n=100$, 50 and 10. In the left-hand plot of Figure EC.4, 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 EC.4, 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 §5.2 of the main paper.

Remark EC.1 (Insights of MPS). We remark that the interesting shapes of $-\partial s_{w,\alpha}/\partial w$ and $-\partial s_{w,\alpha}/\partial\alpha$ demonstrated in Figure EC.4 are quite intuitive. First, when the delay target w is bigger (so the system is more heavily-loaded), the staffing level must be lower. Hence, the system performance is evidently more sensitive to a change of staffing by Δs. This explains why $-\partial s_{w,\alpha}/\partial w$ is a decreasing function in w (e.g., adding 1 server makes a more effective performance improvement when the current staffing is low). Next, when the probability α is close to the boundary of the interval (0, 1), it can be very costly to make further improvement, that is, it requires adding a lot of servers to gain even marginal improvement in α. We now provide insights using plots of normal PDFs in Figure 8 in the main paper: when the staffing function changes by Δs, it makes a more significant impact to the probability $P(W(t)>w)$ for the case $\alpha=0.5$ than cases of $\alpha \approx 0$ and $\alpha \approx 1$, because the normal PDF peaks at $\alpha = 0.5$.

EC.2.4.2. On-and-Off Arrivals Unlike the perfectly stabilized TPoDs in Figure 12 of the online appendix, the example with on-and-off arrivals (with rates alternating between 100 and 0) exhibits some performance degradations. Because the arrival rate jumps drastically between 0 and 100 periodically, 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 EC.5, the actual number of servers can be higher than the planned TTGA staffing function (shown in Subplot 3 of Figure EC.5) 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 EC.5. Nevertheless, our TTGA method can successfully control the TPoDs at or below the desired targets.

EC.3. Additional Proofs

EC.3.1. 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

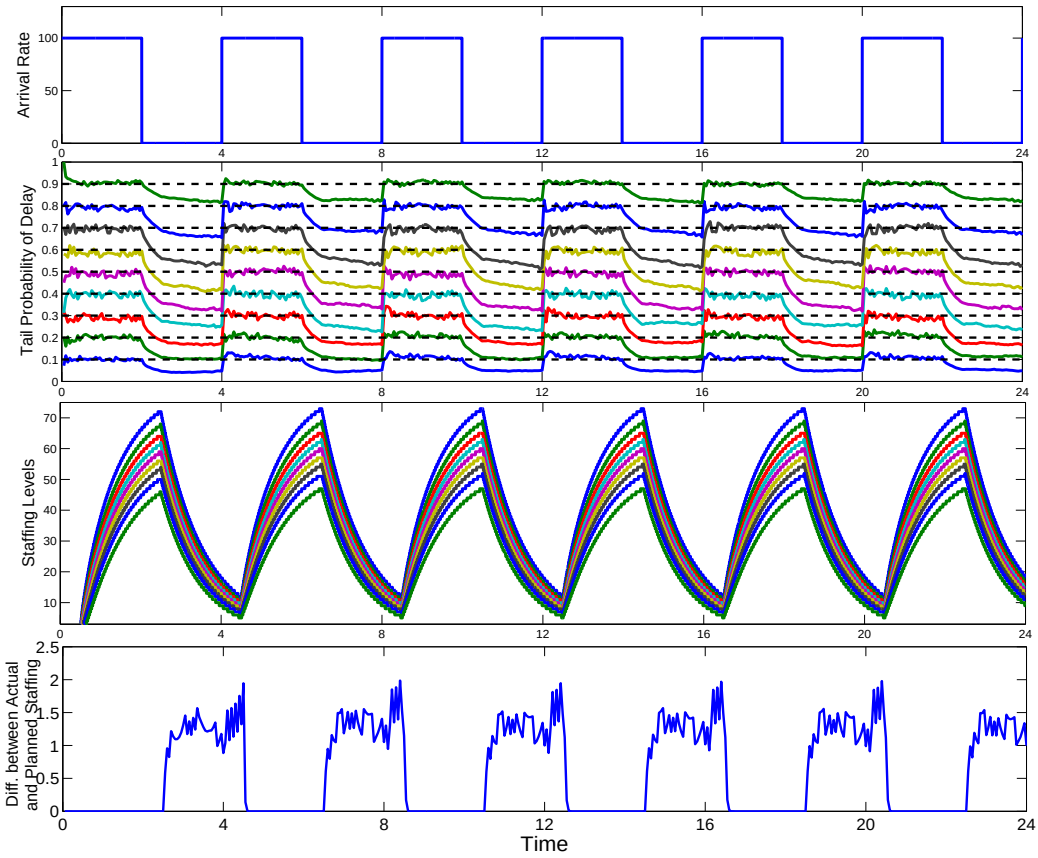

Figure EC.5 Performance measures of the $H_2(t)/M/s_t + H_2$ model with QoS targets $w = 0.5$, $\alpha = 0.1, \ldots, 0.9$, and on-and-off arrival rate function

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 generalize Theorem 2 to the case of G_t arrival. Thus the proof of Theorem 1 in the main paper is completed. □

EC.3.2. Proofs of Corollaries

Proof of Corollary 1. When $\lambda(t) = \lambda$, the OL formula in (5) simplifies to

$$s_w^{(1)}(t) = \bar{F}(w) \int_0^{t-w} \lambda \bar{G}(x) dx = \frac{\bar{F}(w)\lambda}{\mu}(1 - e^{-\mu(t-w)}) \sim \bar{F}(w)\lambda/\mu, \quad \text{as} \quad t \to \infty. \qquad \text{(EC.3)}$$

Define $C \equiv (c_\lambda^2 - 1)\bar{F}(w) + 2$. Next (EC.3) and (8) imply that

$$e^{-\mu t} Z(t) = e^{-h_F(w)t} \sqrt{\int_w^t e^{2h_F(v)x} \bar{F}(w)\lambda \left(C - e^{-\mu(x-v)}\right) dx}$$
$$= \sqrt{\bar{F}(w)\lambda} \sqrt{\frac{C}{2h_F(w)}(1 - e^{2h_F(w)(w-t)}) - \frac{e^{\mu w}}{2h_F(w) - \mu}(e^{-\mu t} - e^{(2h_F(w)-\mu)w - 2h_F(w)t})} \sim \sqrt{\frac{C\bar{F}(w)\lambda}{2h_F(w)}} \tag{EC.4}$$

Similarly, we have

$$\lim_{t\to\infty} e^{-\mu t} \int_w^t Z(u)du = \lim_{t\to\infty} \frac{Z(t)}{\mu e^{\mu t}} = \frac{1}{\mu}\sqrt{\frac{C\bar{F}(w)\lambda}{2h_F(w)}} \tag{EC.5}$$

Combining (EC.4) and (EC.5) yields that

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

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

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

where the last equality holds because $s^{(1)}_w(w) = 0$. $\square$

Proof of Corollary 3. If arrival rate $\lambda(t) = \bar{\lambda}(1 + r\sin(\gamma t + \phi)$, 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

$$s^{(1)}_w(t) = \bar{F}(w) \int_0^{t-w} e^{-\mu x} (\bar{\lambda} + r\bar{\lambda}\sin\gamma(t + \phi/\gamma - w - x))dx$$
$$= \bar{\lambda}\bar{F}(w) \left(\int_0^{t-w} e^{-\mu x} dx + r \int_0^{t-w} e^{-\mu x} \sin\gamma(t + \phi/\gamma - w - x)dx \right)$$
$$= \bar{\lambda}\bar{F}(w) \left[\frac{1 - e^{-\mu(t-w)}}{\mu} - \frac{r}{\sqrt{\mu^2 + \gamma^2}} \left(e^{-\mu(t-w)} \sin(\phi - \varphi) - \sin(\gamma(t + \phi/\gamma - w) - \varphi) \right) \right]$$
$$\sim \bar{\lambda}\bar{F}(w) \left(\frac{1}{\mu} + \frac{r}{\sqrt{\mu^2 + \gamma^2}} \sin(\gamma(t + \phi/\gamma - w) - \varphi) \right) \tag{EC.6}$$

Next, we compute

$$\int_w^t e^{2h_F(w)x} \left(\left[(c_\lambda^2 - 1)\bar{F}(w) + 2 \right] (\mu s^{(1)}_w(x) + \dot{s}^{(1)}_w(x)) - \dot{s}^{(1)}_w(x) \right) dx$$
$$= \int_w^t e^{2h_F(w)x} (\mu C s^{(1)}_w(x) + (C-1)\dot{s}^{(1)}_w(x))dx = \mu C \int_w^t e^{2h_F(w)x} s^{(1)}_w(x)dx + (C-1) \int_w^t e^{2h_F(w)x} \dot{s}^{(1)}_w(x)dx \tag{EC.7}$$

We compute the two integral in (EC.7) respectively. For the first integral,

$$\begin{aligned}
&\int_w^t e^{2h_F(w)x} s_w^{(1)}(x)dx \\
&= \int_w^t e^{2h_F(w)x} \bar{\lambda}\bar{F}(w) \left[\frac{1-e^{-\mu(x-w)}}{\mu} - \frac{r}{\sqrt{\mu^2+\gamma^2}} \left(e^{-\mu(x-w)} \sin(\phi-\varphi) - \sin(\gamma(x+\phi/\gamma-w)-\varphi) \right) \right] dx \\
&= \bar{\lambda}\bar{F}(w) \left(\frac{e^{2h_F(w)t} - e^{2h_F(w)w}}{2\mu h_F(w)} + \frac{r(e^{2h_F(w)t}\sin(\gamma(t+\phi/\gamma-w)-\eta) - e^{2h_F(w)w}\sin(\phi-\eta))}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}} - \right. \\
&\qquad \left. \left(\frac{1}{\mu} - \frac{r\sin(\phi-\varphi)}{\sqrt{\mu^2+\gamma^2}} \right) \frac{e^{\mu w}\left(e^{(2h_F(w)-\mu)t} - e^{(2h_F(w)-\mu)w} \right)}{2h_F(w)-\mu} \right)
\end{aligned} \tag{EC.8}$$

For the second integral, we have

$$\int_w^t e^{2h_F(w)x} \dot{s}_w^{(1)}(x)dx = \int_w^t e^{2h_F(w)x} ds_w^{(1)}(x) = e^{2h_F(w)t} s_w^{(1)}(t) - 2h_F(w) \int_w^t e^{2h_F(w)(x)} s_w^{(1)}(x)dx. \tag{EC.9}$$

Note that the integral in the second term of (EC.9) coincide with the first integral in (EC.7), that is computed in (EC.8). Next, we establish the convergence of $e^{-\mu t}Z(t)$ as $t \to \infty$.

$$\begin{aligned}
e^{-\mu t}Z(t) &= e^{-h_F(w)t} \sqrt{\mu C \int_w^t e^{2h_F(w)x} s_w^{(1)}(x)dx + (C-1)\int_w^t e^{2h_F(w)x}\dot{s}_w^{(1)}(x)dx} \\
&\sim \sqrt{\begin{aligned}
&\bar{\lambda}\bar{F}(w)(\mu C - (C-1)2h_F(w)) \left(\frac{1}{2\mu h_F(w)} + \frac{r\sin(\gamma(t+\phi/\gamma-w)-\eta)}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}} \right) \\
&+ (C-1)\bar{\lambda}\bar{F}(w) \left(\frac{1}{\mu} + \frac{r}{\sqrt{\mu^2+\gamma^2}} \sin(\gamma(t+\phi/\gamma-w)-\varphi) \right)
\end{aligned}}
\end{aligned} \tag{EC.10}$$

Similarly to (EC.5), $e^{-\mu t}\int_w^t Z(u)du \sim (1/\mu)\lim_{t\to\infty} e^{-\mu t}Z(t)$. Therefore,

$$\begin{aligned}
s_{w,\alpha}^{(2)}(t) &\sim \frac{z_\alpha h_F(w)}{\mu} \sqrt{\begin{aligned}
&\bar{\lambda}\bar{F}(w)(\mu C - (C-1)2h_F(w)) \left(\frac{1}{2\mu h_F(w)} + \frac{r\sin(\gamma(t+\phi/\gamma-w)-\eta)}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}} \right) \\
&+ (C-1)\bar{\lambda}\bar{F}(w) \left(\frac{1}{\mu} + \frac{r}{\sqrt{\mu^2+\gamma^2}} \sin(\gamma(t+\phi/\gamma-w)-\varphi) \right)
\end{aligned}} \\
&= \frac{z_\alpha f(w)\sqrt{\bar{\lambda}}}{\mu\sqrt{\bar{F}(w)}} \sqrt{\begin{aligned}
&\frac{C}{2h_F(w)} + \frac{r(\mu C - 2(C-1)h_F(w))}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}} \sin(\gamma(t+\phi/\gamma-w)-\eta) \\
&+ \frac{r(C-1)}{\sqrt{\mu^2+\gamma^2}} \sin(\gamma(t+\phi/\gamma-w)-\varphi)
\end{aligned}}
\end{aligned} \tag{EC.11}$$

□

EC.4. Explicit Staffing Formulas

EC.4.1. Staffing Formula of the Main Example

We hereby give the more detailed staffing formula for the main example in the paper. The first staffing term $s_w^{(1)}(t)$ is the same as (EC.6); the second staffing term $s_{w,\alpha}^{(2)}(t)$ is the same as (7) in

the main paper where

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

and φ,η are defined as in Corollary 3.

EC.4.2. Explicit Expressions for Approximating Formulas in §3

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

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

EC.5. 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 $B^r(\tau)/s(\tau)$.