E-Companion

This e-companion provides supplementary materials to the main paper. In §EC.1, we provide all the technical proofs omitted from the main paper. In §EC.2, we give additional numerical studies.

EC.1. Proofs

EC.1.1. A More Comprehensive Version of Theorem 1 and Its Proof.

We hereby state and prove a more comprehensive version of Theorem 1, where the FCLT for the queue lengths is also established which have their own rights.

THEOREM EC.1. *Suppose the system operates under the proposed staffing and scheduling rule and there is an initial convergence of* $(\widehat{H}^n, \widehat{B}_1^n, \ldots, \widehat{B}_K^n)$ *to zero at* $t=0$.

(a) Then there is a joint convergence for the CLT-scaled waiting time processes:

$$\left(\widehat{H}_1^n, \ldots, \widehat{H}_K^n, \widehat{W}_1^n \ldots, \widehat{W}_K^n\right) \Rightarrow \left(\widehat{H}_1, \ldots, \widehat{H}_K, \widehat{W}_1 \ldots, \widehat{W}_K\right) \quad in \quad \mathcal{D}^{2K} \quad as \quad n \to \infty,$$

where the limits on the right-hand side are well-defined stochastic processes.

(b) The limits for all HWT and PWT processes are deterministic functionals of a one-dimensional process $\widehat{H}$*, namely,*

$$\widehat{H}_i(t) \equiv w_i(\widehat{H}(t) - \kappa_i) \quad and \quad \widehat{V}_i(t) \equiv w_i(\widehat{H}(t+w_i) - \kappa_i);$$

the process $\widehat{H}$ *uniquely solves the SVE* (3) *where*

$$K(t) \equiv \eta^{-1}\left(\int_0^t \sum_{i=1}^K \psi_i \kappa_i e^{\mu_i(s-t)} \mathrm{d}s - c\right), \; L(t,s) \equiv \eta^{-1}\left(\sum_{i=1}^K e^{\mu_i(s-t)}\left(\eta_i \mu_i - \psi_i\right)\right),$$
$$J(t,s) \equiv \eta^{-1}\left(2\sum_{i=1}^K e^{2\mu_i(s-t)} F_i^c(w_i)\lambda_i\right)^{1/2} \quad for \quad \eta_i \equiv w_i\lambda_i F_i^c(w_i), \quad \psi_i \equiv w_i\lambda_i f_i(w_i)$$

and $\eta \equiv \sum_{i=1}^K \eta_i$.

(c) The FCLT for each queue-length process is the sum of two terms, namely,

$$\widehat{Q}_i(t) \equiv \int_{t-w_i}^t \sqrt{\lambda_i F_i^c(t-u)}\mathrm{d}\mathcal{W}_i + \lambda_i F_i^c(w_i)\widehat{H}_i(t),$$

where $\{\mathcal{W}_i; i=1,\ldots,K\}$ *are* K *independent standard Brownian motions.*

As alluded to in the main paper, the proof of FCLT for the waiting-time processes proceeds in four major steps.

Step 1: SSC for the pre-limit HWT and PWT processes. Let $a_i^n(t)$ denote the inter-arrival time between the HoL customer in queue i and the most recent class-i customer who entered service. By the way the scheduling rule operates,

$$H^n(t) - a_i^n(t)/w_i < H_i^n(t)/w_i + n^{-1/2}\kappa_i \le H^n(t). \tag{EC.1}$$

For a fixed time t, it is not difficult to see that $a_i^n(t)$ is first-order stochastically dominated by an exponential random variable with rate $n\lambda_i F_i^c(\bar{T})$ for $\bar{T} \equiv T + \theta$. To proceed, we would like to establish a uniform bound for $a_i^n(t)$ over all $t \le T$. To this end, we make the following observation: (i) For each class, the number of arrivals over any compact time interval is $O(n)$; and (ii) the maximum of n i.i.d. exponential random variables is $O(\log n)$. As an immediate consequence, we have $\sup_{t\le T}\{a_i^n(t)\} = O(n^{-1}\log n)$. Combining with (EC.1) yields $H_i^n(t)/w_i + n^{-1/2}\kappa_i = H^n(t) - O(n^{-1}\log n)$, or, equivalently,

$$\widehat{H}_i^n(t) = w_i(\widehat{H}^n(t) - \kappa_i) - O(n^{-1/2}\log n), \tag{EC.2}$$

where we recall that $\widehat{H}^n$ is the CLT-scaled frontier process, i.e., $\widehat{H}^n(t) \equiv n^{1/2}\left(H^n(t) - 1\right)$.

We next argue that under the proposed scheduling policy the PWT and the HWT satisfy

$$V_i^n(t - H_i^n(t)) = H_i^n(t) + O(n^{-1}\log n). \tag{EC.3}$$

The above relation evidently holds true for $K = 1$, because the PWT at the time of arrival of the HoL customer is the HoL customer's elapsed waiting time (i.e., the HWT) at time t plus the additional time until the next departure. For $K \ge 2$, we aims to establish (EC.3) by showing that the number of service completions needed for the HoL customer of queue i to enter service is no greater than the sum of $K-1$ geometric random variables. To see this is the case, suppose at time t customer A enters service from queue i and customer B becomes the new HoL customer in queue i. Then customer B must have arrived at the system at time $t - H_i^n(t)$. By the definition of $a_i^n(t)$, customer A arrived at the system at time $t - H_i^n(t) - a_i^n(t)$. Suppose $\kappa_i \equiv 0,\ i \in \mathcal{I} \equiv \{1, \ldots, K\}$ (the case where κ_i are not zeros can be analyzed in a similar fashion). Then under the proposed scheduling policy, only those class-j customers who arrived during the interval

$$\left(t - \frac{w_j\left(H_i^n(t) + a_i^n(t)\right)}{w_i}, t - \frac{w_j H_i^n(t)}{w_i}\right) \tag{EC.4}$$

could enter service prior to the time at which customer B enters service. To proceed, we make the following observation: The number of arrivals from a Poisson process with arrival rate $\lambda^{(2)}$ over an exponentially distributed time with rate $\lambda^{(1)}$ follows a geometric distribution with parameter $\frac{\lambda^{(1)}}{\lambda^{(1)}+\lambda^{(2)}}$. Now because the interval (EC.4) has a length of $(w_j a_i^n/w_i)$, the number class-j customers who have a higher service priority over B is stochastically dominated by a geometric random variable with mean $\frac{w_j\lambda_j}{w_i\lambda_i F_i^c(\bar{T})}$. This shows that the total number of customers who will enter service before B is first-order stochastically dominated by the sum of $K-1$ geometric random variables. This gives (EC.3) for $K \ge 2$.

Step 2: The FWLLN. We will first prove that the sequence $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{H}^n); n \in \mathbb{N}\}$ is stochastically bounded; see §5.2 of Pang et al. (2007) for a precise definition of stochastic boundedness. To that end, introduce the LLN- and CLT-scaled empirical process

$$\begin{aligned}
&\bar{U}^n(t,x) \equiv \frac{1}{n}\sum_{k=1}^{\lfloor nt\rfloor} \mathbf{1}_{\{X_i \le x\}} \quad \text{for} \quad t \ge 0, \quad 0 \le x \le 1, \quad \text{and} \\
&\widehat{U}^n(t,x) \equiv \sqrt{n}\left(\bar{U}^n(t,x) - \mathbb{E}\left[\bar{U}^n(t,x)\right]\right) = \frac{1}{\sqrt{n}}\left(\sum_{k=1}^{\lfloor nt\rfloor} \mathbf{1}_{\{X_i \le x\}} - x\right),
\end{aligned} \tag{EC.5}$$

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0,1]$. Krichagina and Puhalskii (1997) have shown that $\widehat{U}^n \Rightarrow \widehat{U}$ in $\mathcal{D}_{\mathcal{D}}$ as $n \to \infty$, where $\widehat{U}$ is the standard Kiefer process (see for example Aras et al. (2018) for a review of Kiefer process). We now break the enter-service process $E_i^n(t)$ in (7) into three pieces, namely,

$$E_i^n(t) = E_{i,1}^n(t) + E_{i,2}^n(t) + E_{i,3}^n(t), \tag{EC.6}$$

where we defined

$$E_{i,1}^n(t) \equiv \sqrt{n}\int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\mathrm{d}\widehat{A}_i^n(u), \quad E_{i,2}^n(t) \equiv \sqrt{n}\int_{-H_i^n(0)}^{t-H_i^n(t)}\int_0^1 \mathbf{1}_{\{y>F_i^c(V_i^n(u))\}}\mathrm{d}\widehat{U}_i^n(\bar{A}_i^n(u), y),$$

and

$$E_{i,3}^n(t) \equiv n\int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\lambda_i \mathrm{d}u,$$

where $\widehat{U}_i^n$ is a CLT-scaled empirical process specified by (EC.5). The decomposition in (EC.6) exhibits a separation of randomness. Specifically, the three terms capture the variabilities from three separate random sources: the arrival process (by the term $\widehat{A}_i^n$ in $E_{i,1}^n$), the abandonment times (by the Kiefer term $\widehat{U}_i^n$ in $E_{i,2}^n$, and the waiting time V_i^n (by $E_{i,3}^n$) which further depends on the service times.

To proceed, define the CLT-scaled enter-service process as

$$\widehat{E}_i^n(t) \equiv n^{-1/2}\left(E_i^n(t) - n\varepsilon_i(t)\right) \quad \text{for} \quad \varepsilon_i(t) \equiv F_i^c(w_i)\lambda_i t. \tag{EC.7}$$

Following the decomposition given in (EC.6), we can write

$$\widehat{E}_i^n(t) = \widehat{E}_{i,1}^n(t) + \widehat{E}_{i,2}^n(t) + \widehat{E}_{i,3}^n(t), \tag{EC.8}$$

where

$$\widehat{E}_{i,1}^n(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\mathrm{d}\widehat{A}_i^n(u), \ \widehat{E}_{i,2}^n(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)}\int_0^1 \mathbf{1}_{\{y>F_i^c(V_i^n(u))\}}\mathrm{d}\widehat{U}_i^n(\bar{A}_i^n(u), y), \tag{EC.9}$$

and

$$\begin{aligned}
\widehat{E}_{i,3}^n(t) &\equiv n^{-1/2}\left(E_{i,3}^n(t) - n\varepsilon_i(t)\right) = \sqrt{n}\left(\int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\lambda_i \mathrm{d}u - \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i \mathrm{d}u\right) \\
&= \sqrt{n}\lambda_i\int_0^t \left(F_i^c(H_i^n(u)) - F_i^c(w_i)\right)\mathrm{d}u - \lambda_i\int_0^t F_i^c(H_i^n(u))\mathrm{d}\widehat{H}_i^n(u) + O(n^{-1/2}\log n) \\
&= -\lambda_i w_i\left(\int_0^t f_i(\zeta_i^n(u))(\widehat{H}^n(u) - \kappa_i)\mathrm{d}u - \int_0^t F_i^c(H_i^n(u))\mathrm{d}\widehat{H}^n(u)\right) + O(n^{-1/2}\log n),
\end{aligned} \tag{EC.10}$$

where the second equality follows by a change of variables, namely $t \to t - H_i^n(t)$, plus the relation (EC.3), while the third equality follows from (EC.2) and applying the mean-value theorem with

$$\min\{U_i^n(t), w_i\} \le \zeta_i^n(t) \le \max\{U_i^n(t), w_i\}. \tag{EC.11}$$

On the other hand, the conservation of flow implies

$$E_i^n(t) = B_i^n(t) - B_i^n(0) + D_i^n(t), \tag{EC.12}$$

Clearly, from (EC.7) it follows that $\varepsilon_i(t) = \mu_i m_i t$. Multiplying its both sides by n, subtracting it from (EC.12), and dividing both sides by $n^{1/2}$ yields

$$\widehat{E}_i^n(t) = \widehat{B}_i^n(t) - \widehat{B}_i^n(0) + \mu_i \int_0^t \widehat{B}_i^n(u)\mathrm{d}u + \widehat{D}_i^n(t)$$

or

$$\mathrm{d}\widehat{B}_i^n(t) + \mu_i \widehat{B}_i^n(t)\mathrm{d}t = \mathrm{d}\widehat{E}_i^n(t) - \mathrm{d}\widehat{D}_i^n(t) \quad \text{for} \quad \widehat{D}_i^n(t) \equiv n^{-1/2}\left(D_i^n(t) - \mu_i \int_0^t B_i^n(u)\mathrm{d}u\right) \tag{EC.13}$$

Let $B^n(t) \equiv B_1^n(t) + \cdots + B_K^n(t)$. It is routine to show, with the overloading assumption (11), that the event $\mathcal{E}^n \equiv \{B^n(t) = s^n; 0 \le t \le T\}$ holds with arbitrarily high probability by choosing n large enough. Thus, it suffices to focus on the sample paths for which event $\mathcal{E}^n$ holds. In this case we get

$$\sum_{i=1}^K \widehat{B}_i^n(t) = n^{-1/2}\left(B^n(t) - nm\right) = n^{-1/2}\left(s^n - nm(t)\right) = c. \tag{EC.14}$$

Upon substituting (EC.8) - (EC.10) into (EC.13), we obtain, for i = 1, …, K,

$$\begin{aligned}\widehat{B}_i^n(t) + \lambda_i w_i \int_0^t F_i^c(H_i^n(u))\mathrm{d}\widehat{H}^n(u) = &-\mu_i \int_0^t \widehat{B}_i^n(u)\mathrm{d}u - \lambda_i w_i \int_0^t f_i(\zeta_i^n(u))\widehat{H}^n(u)\mathrm{d}u \\ &+ \lambda_i w_i \kappa_i \int_0^t f_i(\zeta_i^n(u))\mathrm{d}u + \widehat{E}_{i,1}^n(u) + \widehat{E}_{i,2}^n(u) - \widehat{D}_i^n(u) + O(n^{-1/2}\log n).\end{aligned} \tag{EC.15}$$

Together with (EC.14), we end up getting $K+1$ linear differential equations with respect to the $(K+1)$-dimensional process $(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{H}^n)$. Paralleling (5.14) in Aras et al. (2018), we apply the Gronwall's inequality together with the stochastic boundedness of $\widehat{E}_{i,1}^n$, $\widehat{E}_{i,2}^n, \widehat{D}_i^n$ plus the assumed properties of f_i, F_i^c to conclude the stochastic boundedness of the sequence $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{H}^n); n \in \mathbb{N}\}$; in particular, the sequences $\{\widehat{H}^n; n \in \mathbb{N}\}$ and $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n); n \in \mathbb{N}\}$ are stochastically bounded.

On the other hand, by the established stochastic boundedness of $\{\widehat{H}^n; n \in \mathbb{N}\}$ together with the relations (EC.2) and (EC.3), we conclude that $\{\widehat{H}_n; n \in \mathbb{N}\}$ and $\{\widehat{V}_n; n \in \mathbb{N}\}$ are stochastically bounded. This implies the FWLLN for the HWT and PWT processes, that is, as $n \to \infty$,

$$(H^n, H_1^n, \ldots, H_K^n, V_1^n, \ldots, V_K^n) \Rightarrow (\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}) \quad \text{in} \quad \mathcal{D}^{2K+1}, \tag{EC.16}$$

where the joint convergence is due to converging-together lemma (Theorem 11.4.5. in Whitt (2002)).

Step 3: The FCLT for the waiting time processes. Similar to the proof of Lemma 5.1 in Aras et al. (2018), we invoke the continuous mapping theorem with (EC.9) and (EC.16) to get

$$\widehat{E}_{i,1}^n(t) \Rightarrow \widehat{E}_{i,1}(t) \equiv F_i^c(w_i) \int_{-w_i}^{t-w_i} \sqrt{\lambda_i}\mathrm{d}\mathcal{W}_{\lambda_i}(u), \tag{EC.17}$$

where $\mathcal{W}_{\lambda_i}$ is a standard Brownian motion. To proceed, we argue that, as $n \to \infty$,

$$\widehat{E}_{i,2}^n(t) \Rightarrow \widehat{E}_{i,2}(t) \equiv \sqrt{F_i^c(w_i)F_i(w_i)} \int_{-w_i}^{t-w_i} \sqrt{\lambda_i}\mathrm{d}\mathcal{W}_{\theta_i}(u), \tag{EC.18}$$

where $\mathcal{W}_{\theta_i}$ is a standard Brownian independent of $\mathcal{W}_{\lambda_i}$. The essential structure of the proof for (EC.18) is exactly the same as that of A.7.2 in Aras et al. (2018), which in turn draws on Theorem 7.1.4 in Ethier and Kurtz (1986). Because the proof can be fully adapted from theirs, we

omit the details. Moreover, as a direct consequence of the established stochastic boundedness of $\{(\widehat{B}_1^n,\ldots,\widehat{B}_K^n);n\in\mathbb{N}\}$, we have the FWLLN for the busy-server processes

$$\left(\bar{B}_1^n,\ldots,\bar{B}_K^n\right)\Rightarrow(m_1\mathfrak{e},\ldots,m_K\mathfrak{e})\quad\text{in}\quad\mathcal{D}^K\quad\text{as}\quad n\to\infty.$$

Next a standard random-time-change argument allows us to derive

$$\widehat{D}_i^n(\cdot)=n^{-1/2}\left[\Pi_i^d\left(n\mu_i\int_0^{\cdot}\bar{B}_i^n(u)\mathrm{d}u\right)-n\mu_i\int_0^{\cdot}\bar{B}_i^n(u)\mathrm{d}u\right]\Rightarrow\mathcal{W}_{\mu_i}\left(\mu_i\int_0^{\cdot}m_i(u)\mathrm{d}u\right)\tag{EC.19}$$

as $n\to\infty$, where we have defined Π_i^d to be a unit-rate Poisson process and $\mathcal{W}_{\mu_i}$ to be a standard Brownian motion independent of $\mathcal{W}_{\lambda_i}$ and $\mathcal{W}_{\theta_i}$. To establish the convergence of (15), we will need to strengthen (EC.17), (EC.18) and (EC.19) to joint convergence. The joint convergence of multiple random elements is equivalent to individual convergence if they are independent. Here $\widehat{E}_{i,1}^n$, $\widehat{E}_{i,2}^n$ and $\widehat{D}_i^n$ are not independent because both $\widehat{E}_{i,1}^n$ and $\widehat{E}_{i,2}^n$ involve the arrival-time sequence, and $\widehat{D}_i^n$ depends on B_i^n which in turn correlates with E_i^n through (EC.12). But they are conditionally independent given A_i^n,H_i^n,V_i^n and B_i^n. Hence, we can establish the joint convergence by first conditioning and then unconditioning. See Lemma 4.1 of Aras et al. (2017) for a reference, which is a variant of Theorem 7.6 of Pang et al. (2007).

To derive a set of SDEs satisfied by the CLT-scaled processes $(\widehat{B}_1^n,\ldots,\widehat{B}_K^n,\widehat{H}^n)$, we seek to simplify the right-hand side of (EC.10). First we note that the inequality (EC.11) and the convergence in (EC.18) imply

$$\zeta_i^n(t)=w_i+O(n^{-1/2}\log n)=H_i^n(t)+O(n^{-1/2}\log n).\tag{EC.20}$$

Using integration by parts, we get that $-\lambda_i w_i\int_0^t F_i^c(H_i^n(u))\mathrm{d}\widehat{H}^n(u)$ is equal to

$$-\lambda_i w_i F_i^c(\zeta_i^n(t))\widehat{H}^n(t)+\lambda_i w_i F_i^c(\zeta_i^n(0))\widehat{H}^n(0)+\lambda_i w_i\int_0^t\widehat{H}^n(u)\mathrm{d}F_i^c(\zeta_i^n(u)).\tag{EC.21}$$

Upon plugging (EC.21) into (EC.10) and making use of (EC.20), we arrive at

$$\widehat{E}_{i,3}^n(t)=-\lambda_i w_i f_i(w_i)\int_0^t(\widehat{H}^n(u)-\kappa_i)\mathrm{d}u+\lambda_i w_i F_i^c(w_i)\widehat{H}^n(0)-\lambda_i w_i F_i^c(w_i)\widehat{H}^n(t)+O(n^{-1/2}\log n).$$

Now plugging (EC.8) and the equation above into (EC.13), we have, for $i=1,\ldots,K$,

$$\begin{aligned}\widehat{B}_i^n(t)+\lambda_i w_i F_i^c(w_i)\widehat{H}^n(t)&=\widehat{B}_i^n(0)+\lambda_i w_i F_i^c(w_i)\widehat{H}^n(0)-\mu_i\int_0^t\widehat{B}_i^n(u)\mathrm{d}u\\&\quad-\lambda_i w_i f_i(w_i)\int_0^t\widehat{H}^n(u)\mathrm{d}u+\lambda_i w_i f_i(w_i)\kappa_i t+\widehat{E}_{i,1}^n(t)+\widehat{E}_{i,2}^n(t)-\widehat{D}_i^n(t)+O(n^{-1/2}\log n).\end{aligned}\tag{EC.22}$$

The joint convergence $(\widehat{B}_i^n,\ldots,\widehat{B}_K^n,\widehat{H}^n)\Rightarrow(\widehat{B}_i,\ldots,\widehat{B}_K,\widehat{H})$ then follows by applying the continuous mapping theorem (see Theorem 4.1 of Pang et al. (2007)) to (EC.14) and (EC.22), with the *joint* convergence of $\widehat{E}_{i,1}^n,\widehat{E}_{i,2}^n$, and $\widehat{D}_i^n$ as specified by (EC.17), (EC.18), and (EC.19), respectively. The convergence of $\{\widehat{H}_i^n;n\in\mathbb{N}\}$ and $\{\widehat{V}_i^n;n\in\mathbb{N}\}$ follow easily from and (EC.2) and (EC.3), respectively.

Step 4: Deriving the SVE for the frontier process. The multi-dimensional SDE (18) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\mu_i t}\tilde{B}_i(t)\right) = e^{\mu_i t}\left(-\eta_i \widehat{H}(t) - \int_0^t \psi_i \widehat{H}(u)\mathrm{d}u + y_i(t) + G_i(t)\right), \tag{EC.23}$$

where

$$\tilde{B}_i(t) \equiv \int_0^t \widehat{B}_i(u)\mathrm{d}u \quad \text{and} \quad y_i(t) \equiv \int_0^t w_i f_i(w_i)\lambda_i\kappa_i \mathrm{d}u.$$

Integrating (EC.23) from 0 to t yields

$$\begin{aligned}\tilde{B}_i(t) &= e^{-\mu_i t}\int_0^t e^{\mu_i s}\left(-\eta_i\widehat{H}(s) - \int_0^s \psi_i \widehat{H}(u)du + y_i(s) + G_i(s)\right)\mathrm{d}s \\ &= -\int_0^t \eta_i e^{\mu_i(s-t)}\widehat{H}(s)\mathrm{d}s - \int_0^t \psi_i \frac{1-e^{\mu_i(s-t)}}{\mu_i}\widehat{H}(s)\mathrm{d}s \\ &\quad + \int_0^t \psi_i\kappa_i\frac{1-e^{\mu_i(s-t)}}{\mu_i}\mathrm{d}s + \int_0^t e^{\mu_i(s-t)}G_i(s)\mathrm{d}s.\end{aligned}$$

Summing up over i from 1 to K, we have

$$\begin{aligned}\int_0^t c\mathrm{d}s = \sum_{i=1}^K \tilde{B}_i(t) = -\sum_{i=1}^K\Bigg(&\int_0^t\left(\eta_i e^{\mu_i(s-t)}\widehat{H}(s) - \psi_i\frac{1-e^{\mu_i(s-t)}}{\mu_i}\widehat{H}(s) + \psi_i\kappa_i\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\mathrm{d}s \\ &+ \int_0^t \frac{1-e^{\mu_i(u-t)}}{\mu_i}\sqrt{F_i^c(w_i)\lambda_i + \mu_i m_i}\mathrm{d}\mathcal{W}_i(u)\Bigg),\end{aligned}$$

where the second equality holds by aggregating three independent Brownian motions $\mathcal{W}_{\mu_i}$, $\mathcal{W}_{\theta_i}$ and $\mathcal{W}_{\lambda_i}$ in (19) into one independent standard Brownian motion $\mathcal{W}_i$. Differentiating the above and further aggregating the independent Brownian motions $\mathcal{W}_1,\ldots,\mathcal{W}_K$ into $\mathcal{W}$ yields the SVE in (3).

FCLT for the queue-length processes. To derive the FCLT for the queue-length processes, we decompose the right-hand side of (8) into three terms, namely,

$$Q_i^n(t) = Q_{i,1}^n(t) + Q_{i,2}^n(t) + Q_{i,3}^n(t), \tag{EC.24}$$

where

$$Q_{i,1}^n(t) \equiv \sqrt{n}\int_{t-H_i^n(t)}^t F_i^c(t-u)\mathrm{d}\widehat{A}_i^n(u), \quad t\geq 0, \tag{EC.25}$$

$$Q_{i,2}^n(t) \equiv \sqrt{n}\int_{t-H_i^n(t)}^t\int_0^1 \mathbf{1}_{\{x>F_i^c(t-u)\}}\mathrm{d}\widehat{U}_i^n(\bar{A}_i^n(u),x) \quad t\geq 0, \tag{EC.26}$$

$$Q_{i,3}^n(t) \equiv n\lambda_i\int_{t-H_i^n(t)}^t F_i^c(t-u)\mathrm{d}u \quad t\geq 0, \tag{EC.27}$$

Similar to (EC.6), the decomposition of the queue length above can also be explained by the separation of variabilities in the arrival process, service times, and abandonment times.

Accordingly, the centered and normalized queue-length process can be decomposed into three terms

$$\widehat{Q}_i^n(t) \equiv n^{-1/2}\left(Q_i^n(t) - nq_i(t)\right) = \widehat{Q}_{i,1}^n(t) + \widehat{Q}_{i,2}^n(t) + \widehat{Q}_{i,3}^n(t),$$

where

$$\widehat{Q}^n_{i,1}(t) \equiv \int_{t-H^n_i(t)}^{t} F^c_i(t-u)\mathrm{d}\widehat{A}^n_i(u) \;\Rightarrow\; \int_{t-w_i}^{t} F^c_i(t-u)\mathrm{d}\widehat{A}_i(u), \tag{EC.28}$$

$$\widehat{Q}^n_{i,2}(t) \equiv \int_{t-H^n_i(t)}^{t}\int_0^1 \mathbf{1}_{\{x>F^c_i(t-u)\}}\mathrm{d}\widehat{U}^n_i(\bar{A}^n_i(u),x) \Rightarrow \int_{t-w_i}^{t}\sqrt{\lambda_i F^c_i(t-u)F_i(t-u)}\mathrm{d}\mathcal{W}_{\theta_i}(u), \tag{EC.29}$$

$$\widehat{Q}^n_{i,3}(t) \equiv \sqrt{n}\lambda_i \int_{t-H^n_i(t)}^{t-w_i} F^c_i(t-u)\mathrm{d}u \;\Rightarrow\; \lambda_i F^c_i(w_i)\widehat{H}_i(t). \tag{EC.30}$$

Here the proof for (EC.28) and (EC.29) is very similar to that of (EC.17) and (EC.18), and the proof for (EC.30) is also straightforward. □

EC.1.2. Proof of Theorem 2.

To establish part (i), we find that there exist multiple integrability criteria that we can apply to the resolvent of linear Volterra equations. Here we choose to use the results laid out by Levin (1977). For this purpose, write $\mathcal{L} = \mathcal{L}_1 - \mathcal{L}_2$ where

$$\mathcal{K}_1(t) \equiv \eta^{-1}\left(\sum_{i=1}^{K} e^{-\mu_i t}\eta_i\mu_i\right) \quad \text{and} \quad \mathcal{K}_2(t) \equiv \eta^{-1}\left(\sum_{i=1}^{K} e^{-\mu_i t}\psi_i\right).$$

For an arbitrary number $\delta \geq 0$, we see that

$$\int_0^\infty \mathcal{L}_1(t)e^{-\delta t}\mathrm{d}t = \sum_{i=1}^{K}\eta_i\mu_i/\eta(\mu_i+\delta) \leq 1, \;\; \int_0^\infty \mathcal{L}_2(t)e^{-\delta t}\mathrm{d}t = \sum_{i=1}^{K}\psi_i/\eta(\mu_i+\delta) \geq 0. \tag{EC.31}$$

It is not difficult to see from (EC.31) that

$$\int_0^\infty \mathcal{L}(t)e^{-\delta t}\mathrm{d}t \neq 1 \quad \text{for all} \quad \delta \geq 0 \quad \text{if and only if} \quad \psi_j > 0 \quad \text{for some} \quad j.$$

Part (i) of the proposition then follows by applying Theorem 1.1. of Levin (1977). Part (ii) is a well-known result, see, e.g., Levin (1977). For part (iii), the expression for the first moment is immediate. To derive the variance formula, we apply Itô Isometry to get

$$\begin{aligned}\mathrm{Var}(\widehat{H}(t)) &= \int_0^t \left(\mathcal{J}(t-u) + \int_u^t \mathcal{R}(t-s)\mathcal{J}(s-u)\mathrm{d}s\right)^2 \mathrm{d}u \\ &= \int_0^t \left(\mathcal{J}(t-u) + \int_0^{t-u} \mathcal{R}(s)\mathcal{J}(t-u-s)\mathrm{d}s\right)^2 \mathrm{d}u,\end{aligned}$$

where the last expression, upon using a change of variables, leads to (26). To show that the result of part (iv) is true, notice that function $\mathcal{J}$ has exponential decay. Thus, the limit of the right-hand side of (26), as $t \to \infty$, is finite, if function $\mathcal{R}$ is uniformly bounded over the positive real line. But the required uniform boundedness follows as an immediate consequence of the conclusion of part (i) under the specified condition. Convergence (finiteness) of $\mathbb{E}[\widehat{H}(t)]$ as $t \to \infty$ is immediate due to (25), the integrability of $\mathcal{R}$, and the uniform boundedness of function K. To establish part (v), it suffices to show that (a) K is vanishing as $t \to \infty$ and (b) $\mathcal{R}$ is integrable. Condition (a) is automatically guaranteed by our specific choice of control parameters whereas condition (b) follows directly from part (i). This completes the proof of the proposition. □

EC.1.3. Proof of Theorem 3.

The proof of part (a) follows closely the steps of that for Theorem 1. Thus, we only show the proof of part (b).

Uniqueness and existence of solution to the SVE (3). Consider two functions $x, y \in \mathbb{C}$ (space of continuous functions) satisfying an equation

$$x(t) = \int_0^t L(t,s)x(s)\mathrm{d}s + y(t). \tag{EC.32}$$

we show that (EC.32) specifies a well-defined function $\phi : \mathbb{C} \to \mathbb{C}$ such that $x = \psi(y)$. To do so, for a given y, we define the operator

$$\psi(x)(t) \equiv \int_0^t L(t,s)x(s)\mathrm{d}s + y(t). \tag{EC.33}$$

Therefore, x solves the *fixed-point equation* (FPE)

$$x = \psi(x). \tag{EC.34}$$

We first prove that ψ is a contraction over a finite interval $[0,T]$. Specifically, let $x_1, x_2 \in \mathbb{C}$, and use the uniform norm $\|x\|_T = \sup_{\{0\le t\le T\}} |x(t)|$. We have

$$|\psi(x_1)(t) - \psi(x_2)(t)| \le \int_0^t |L(t,s)|\mathrm{d}s \cdot \|x_1 - x_2\|_T \le L^{\uparrow} T \|x_1 - x_2\|_T, \tag{EC.35}$$

where the constant

$$L^{\uparrow} = \frac{\sum_{i=1}^K w_i \lambda_i^{\uparrow} \left(\mu_i F_i^c(w_i) + f_i(w_i)\right)}{\sum_{i=1}^K w_i \lambda_i^{\downarrow} F_i^c(w_i)} < \infty. \tag{EC.36}$$

In case $L^{\uparrow}T > 1$, we can partition the interval $[0,T]$ to successive smaller intervals with length ΔT satisfying $\Delta T < 1/L^{\uparrow}$. This will recursively guarantee the contraction property over all smaller intervals. Hence, the Banach fixed point theorem implies that the FPE (EC.34) has a unique solution over the entire interval $[0,T]$.

Consequently, the function ϕ specified by (EC.32) is well-defined because $\phi(y)$ has one and only one image for any y. So we conclude that (3) has a unique solution $\widehat{H}$. If fact, we can write (3) as

$$\widehat{H}(t) = \phi\left(\int_0^t J(t,s)\mathrm{d}\mathcal{W}(s) + K(t)\right).$$

To show that $\widehat{H}$ is Gaussian, we again use the contraction ψ defined in (EC.33). We follow the steps that establish strong solutions in Karatzas and Shreve (1991). Define a sequence of processes $\{\widehat{H}^{(k)}, k = 0,1,2,\ldots\}$ such that $\widehat{H}^{(0)}(t) = 0$, and $\widehat{H}^{(k+1)} = \psi(\widehat{H}^{(k)})$ with $y(t) = \int_0^t J(t,s)\mathrm{d}\mathcal{W}(s,\omega)$ for $k \ge 0$. (For each Brownian path and associated Brownian integral, we recursively define the sequence.) We can show that $\widehat{H}^{(k)}$ is Gaussian using an inductive argument. Specifically, $\widehat{H}^{(k+1)}$ is Gaussian because both $\int_0^t L(t,s)\widehat{H}^{(k)}(s)\mathrm{d}s$ and $\int_0^t J(t,s)\mathrm{d}\mathcal{W}(s,\omega)$ are Gaussian. Because ψ is a contraction, we know that $\widehat{H}$ is the almost sure limit of $\widehat{H}^{(k)}$, which implies weak convergence. Hence, $\widehat{H}$ is again Gaussian (because the limit of convergent Gaussian processes is again Gaussian). To elaborate, we

may consider the characteristic function of $\widehat{H}^{(k)}(t)$: $\Phi_k(s) = e^{is\mu_k - s^2\sigma_k^2/2}$ (with μ_k and σ_k^2 being the mean and variance of $\widehat{H}^{(k)}$), which must converge to the characteristic function of $\widehat{H}$. Convergence of $\Phi_k(s)$ at all s implies the convergence of μ_k and σ_k^2, which implies that the characteristic function of $\widehat{H}$ has the form $e^{is\mu_\infty - s^2\sigma_\infty^2/2}$, which concludes the Gaussian distribution.

Treating the mean and variance of $\widehat{H}$. Taking expectation in (3) yields

$$m_{\widehat{H}}(t) = \int_0^t L(t,s) m_{\widehat{H}}(s)\mathrm{d}s + K(t), \qquad \text{where} \quad m_{\widehat{H}}(t) = \mathbb{E}[\widehat{H}(t)]. \tag{EC.37}$$

It remains to show that the FPE $x = \Gamma(x)$ has a unique solution, where $x \in \mathbb{C}$ and the operator

$$\Gamma(x)(t) = \int_0^t L(t,s)x(s)\mathrm{d}s + K(t).$$

We can do so by showing that $\Gamma : \mathbb{C} \to \mathbb{C}$ is another contraction. Specifically, for $x_1, x_2 \in \mathbb{C}$,

$$|\Gamma(x_1)(t) - \Gamma(x_2)(t)| \le \int_0^t |L(t,s)||x_1(s) - x_2(s)|\mathrm{d}s \le L^{\uparrow} t \|x_1 - x_2\|_t,$$

where the finite upperbound $L^{\uparrow}$ is given by (EC.36). The rest of the proof is similar.

To treat the variance of $\widehat{H}$, consider the SVE (3) at $0 \le s, t \le T$

$$\begin{aligned} H(t) - \int_0^t L(t,u)H(u)\mathrm{d}u &= \int_0^t J(t,u)\mathrm{d}\mathcal{W}(u), \\ H(s) - \int_0^s L(s,v)H(v)\mathrm{d}v &= \int_0^s J(s,v)\mathrm{d}\mathcal{W}(v). \end{aligned}$$

Multiplying the two equations and taking expectation yield that

$$\begin{aligned} C(t,s) = &- \int_0^t \int_0^s L(t,u)h(s,v)C(u,v)\mathrm{d}v\mathrm{d}u + \int_0^{s\wedge t} J(t,u)J(s,u)\mathrm{d}u \\ &+ \int_0^t L(t,u)C(u,s)\mathrm{d}u + \int_0^s h(s,v)C(t,v)\mathrm{d}v, \end{aligned}$$

where $C(t,s) = \mathrm{Cov}(\widehat{H}(t), \widehat{H}(s))$, or equivalently, an FPE

$$C = \Theta(C), \tag{EC.38}$$

where $C(\cdot,\cdot) \in \mathbb{C}([0,T]^2)$, and the operator

$$\begin{aligned} \Theta(C)(t,s) = &- \int_0^t \int_0^s L(t,u)h(s,v)C(u,v)\mathrm{d}v\mathrm{d}u + \int_0^t L(t,u)C(u,s)\mathrm{d}u \\ &+ \int_0^s L(s,v)C(t,v)\mathrm{d}v + \int_0^{s\wedge t} J(t,u)J(s,u)\mathrm{d}u. \end{aligned} \tag{EC.39}$$

Using the norm $\|x\|_T = \sup_{0\le s,t\le T} |x(t,s)|$, we next prove that Θ is a contraction. Specifically, for $x_1, x_2 \in \mathbb{C}([0,T]^2)$, we have

$$\begin{aligned} |\Theta(x_1)(t,s) - \Theta(x_2)(t,s)| \le &\int_0^t \int_0^s |L(t,u)L(s,v)| \cdot |x_1(u,v) - x_2(u,v)|\mathrm{d}v\mathrm{d}u \\ &+ \int_0^t |L(t,u)| \cdot |x_1(u,s) - x_2(u,s)|\mathrm{d}u + \int_0^s |L(s,v)| \cdot |x_1(t,v) - x_2(t,v)|\mathrm{d}v \end{aligned}$$

$$\leq \left((L^{\uparrow})^2 ts + L^{\uparrow} t + L^{\uparrow} s\right) \|x_1 - x_2\|_T.$$

The contraction property is guaranteed if we pick a small $\Delta T > 0$ such that $((L^{\uparrow})^2 \Delta T^2 + 2L^{\uparrow}\Delta T) < 1$. According to the Banach contraction theorem, we have the uniqueness and existence in the small block $[0, \Delta T]^2$. The uniqueness and existence of $C(\cdot,\cdot)$ over the entire region $[0,T]\times[0,T]$ can be proved by recursively dealing with small blocks of the form $[i\Delta T,(i+1)\Delta T]\times[j\Delta T,(j+1)\Delta T]$. □

Remark EC.1 (Numerical Algorithm for $\sigma^2_{\widehat{H}}(t)$). The above proof of the existence and uniqueness of the FPE (EC.38) automatically suggests the following recursive algorithm to compute the covariance $C(t,s)$ and variance $\sigma^2_{\widehat{H}}(t)$. To begin with, we pick an acceptable error target $\epsilon > 0$.

Algorithm:

(i) Pick an initial candidate $C^{(0)}(\cdot,\cdot)$;

(ii) In the k^{th} iteration, let $C^{(k+1)} = \Theta\left(C^{(k)}\right)$ with Θ given in (EC.39).

(iii) If $\|C^{(k+1)} - C^{(k)}\|_T < \epsilon$, stop; otherwise, $k = k+1$ and go back to step (ii).

According to the Banach contraction theorem, this algorithm should converge geometrically fast. When it finally terminates, we set $\sigma^2_{\widehat{H}}(t) = C(t,t)$, for $0 \leq t \leq T$, which will be used later to devise required control functions c and κ_i. The algorithm to compute the mean $M_{\widehat{H}}$ is similar. □

EC.1.4. Proof of Theorem 4

The conclusion of part (i) is immediate. The conclusion of part (ii) is also straightforward because the TPoD for class-i customers

$$\begin{aligned}
\mathbb{P}(V_i^n(t) > w_i) &= \mathbb{P}(\sqrt{n}(V_i^n(t) - w_i) > 0) = \mathbb{P}(\widehat{V}_i^n(t) > 0) \\
&\to \mathbb{P}(\widehat{V}_i(t) > 0) = \mathbb{P}\left(w_i\left(\widehat{H}(t+w_i) - \kappa_i(t+w_i)\right) > 0\right) \\
&= \mathbb{P}\left(\widehat{H}(t+w_i) > \kappa_i(t+w_i)\right) = \mathbb{P}\left(\mathcal{Z} > \frac{\kappa_i(t+w_i)}{\sigma_{\widehat{H}}(t+w_i)}\right) = \mathbb{P}\left(\mathcal{Z} > z_{\alpha_i}\right) = \alpha_i,
\end{aligned}$$

where the third equality holds by (33).

To prove part (iii), note that the FPE (35) specifies a well-defined function $\phi : \mathbb{C} \to \mathbb{C}$ such that

$$M_{\widehat{H}} = \phi(K).$$

See the proof of the uniqueness and existence of the SVE (specifically, see (EC.32)–(EC.36)) for details. To proceed, let $(\boldsymbol{\kappa}^*, c^*) \equiv (\kappa_1^*, \ldots, \kappa_K^*, c^*)$, with κ_i^* and c^* given in (40) and (39). Let K^* and $M^*_{\widehat{H}}$ be the corresponding version of (34) and the mean of $\widehat{H}$. (We know that $K^*(t) = M^*_{\widehat{H}}(t) = 0$.) So we have

$$\kappa_i^*(t) = \kappa_i^*(t) - M^*_{\widehat{H}}(t) = z_{1-\alpha_i}\sigma_{\widehat{H}}(t), \qquad 1 \leq i \leq K. \tag{EC.40}$$

Now consider another solution to $(\tilde{\boldsymbol{\kappa}}, \tilde{c})$ to (38), with $(\tilde{\boldsymbol{\kappa}}, \tilde{c}) \equiv (\kappa_1^* + \Delta\kappa_1, \ldots, \kappa_K^* + \Delta\kappa_K, c^* + \Delta c)$. Let $\tilde{K}$ and $\tilde{M}_{\widehat{H}}$ be the corresponding version of (34) and mean of $\widehat{H}$. By (38), we have

$$\kappa_i^*(t) + \Delta\kappa_i(t) - \tilde{M}_{\widehat{H}}(t) = z_{1-\alpha_i}\sigma_{\widehat{H}}(t),, \qquad 1 \leq i \leq K. \tag{EC.41}$$

Comparing (EC.40) with (EC.41), we must have

$$\Delta\kappa_i(t) = \tilde{M}_{\widehat{H}}(t) - M^*_{\widehat{H}}(t) \equiv \Delta\kappa(t) \qquad \text{for all } 1 \le i \le K. \tag{EC.42}$$

Hence, any alternative solution to (38) (if any) has the form $(\kappa_1^* + \Delta\kappa, \ldots, \kappa_K^* + \Delta\kappa, c^* + \Delta c)$. Next, $M^*_{\widehat{H}} = \phi(K^*)$ and $\tilde{M}_{\widehat{H}} = \phi(\tilde{K})$ imply that

$$M^*_{\widehat{H}}(t) = \int_0^t L(t,s) M^*_{\widehat{H}}(s)\mathrm{d}s + K^*(t) \quad \text{and} \quad \tilde{M}_{\widehat{H}}(t) = \int_0^t L(t,s)\tilde{M}_{\widehat{H}}(s)\mathrm{d}s + \tilde{K}(t),$$

which leads to

$$\begin{aligned}\Delta\kappa(t) = \tilde{M}_{\widehat{H}}(t) - M^*_{\widehat{H}}(t) &= \int_0^t L(t,s)\left(\tilde{M}_{\widehat{H}}(s) - M^*_{\widehat{H}}(s)\right)\mathrm{d}s + \left(\tilde{K}(t) - K^*(t)\right), \\ &= \int_0^t L(t,s)\Delta\kappa(s)\mathrm{d}s + \left(\tilde{K}(t) - K^*(t)\right), \end{aligned} \tag{EC.43}$$

where the last equality holds by the first equality. By (EC.42) and (34), we have

$$\tilde{K}(t) - K^*(t) = \frac{\Delta\kappa(t)\sum_{i=1}^K \left(\eta_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))\mathrm{d}s\right) - \Delta c(t)}{\eta(t)}. \tag{EC.44}$$

Finally, combining (EC.43) with (EC.44), we must have, for any $\Delta\kappa$,

$$\Delta c(t) = \Delta\kappa(t)\sum_{i=1}^K \left(\eta_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))\mathrm{d}s\right) - \eta(t)\left(\Delta\kappa(t) - \int_0^t L(t,s)\Delta\kappa(s)\mathrm{d}s\right) = 0,$$

where the last equality holds by (34). This establishes part (iii). □

EC.1.5. Proof of Corollary 3.

Because the functions $L(t,s)$ and $J(t,s)$ are now separable in t and s, SDE (3) becomes

$$\widehat{H}(t) = \frac{1}{R(t)}\int_0^t \tilde{L}(s)\widehat{H}(s)\mathrm{d}s + \frac{1}{R(t)}\int_0^t \tilde{J}(s)\mathrm{d}\mathcal{W}(s) + K(t), \tag{EC.45}$$

where $R(t)$, $\tilde{L}(t)$ and $\tilde{J}(t)$ are specified in Corollary 3. Multiplying $R(t)$ on both sides and differentiating (EC.45) yields

$$\frac{R'(t) - \tilde{L}(t)}{R(t)}\widehat{H}(t)\mathrm{d}t + \mathrm{d}\widehat{H}(t) = \frac{\tilde{J}(t)}{R(t)}\mathrm{d}\mathcal{W}(t) + K'(t)\mathrm{d}t + \frac{K(t)R'(t)}{R(t)}\mathrm{d}t.$$

Multiplying $e^{\int_0^t \frac{R'(t)-\tilde{L}(t)}{R(t)}dv}$ on both sides and integrating from 0 to t yields

$$\begin{aligned} e^{\int_0^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\widehat{H}(t) = &\int_0^t e^{\int_0^u \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\frac{\tilde{J}(u)}{R(u)}\mathrm{d}\mathcal{W}(u) \\ &+ \int_0^t e^{\int_0^u \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\mathrm{d}K(u) + \int_0^t e^{\int_0^u \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\frac{K(u)R'(u)}{R(u)}\mathrm{d}u. \end{aligned}$$

or equivalently $\quad \widehat{H}(t) = \int_0^t e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\frac{\tilde{J}(u)}{R(u)}\mathrm{d}\mathcal{W}(u)$

$$+\int_0^t e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\mathrm{d}K(u)+\int_0^t e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\frac{K(u)R'(u)}{R(u)}\mathrm{d}u. \quad \text{(EC.46)}$$

Note that

$$e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}=e^{\log R(u)-\log R(t)}e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}\mathrm{d}v}=\frac{R(u)}{R(t)}e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}\mathrm{d}v}. \quad \text{(EC.47)}$$

Combining (EC.46) and (EC.47) yields the solution in (43). The variance formula in Corollary 3 easily follows from the isometry of the Brownian integral. □

EC.1.6. Proof of Corollary 4

When $K=1$, the variance formula simplifies to

$$\sigma(t)=\frac{e^{-h_F(w)t}}{\eta(t)}\sqrt{\int_0^t e^{2h_F(w)u}\left(F^c(w)\lambda(u-w)+\mu m(u)\right)\mathrm{d}u}.$$

Therefore, the second-order staffing term

$$\begin{aligned}c(t)&=z_{1-\alpha}e^{-\mu t}\left(e^{-h_F(w)t}e^{\mu t}\sqrt{\int_0^t e^{2h_F(w)u}\left(F^c(w)\lambda(u-w)+\mu m(u)\right)\mathrm{d}u}\right.\\&\qquad\left.-(\mu-h_F(w))\int_0^t e^{-h_F(w)s}e^{\mu s}\sqrt{\int_0^s e^{2h_F(w)u}\left(F^c(w)\lambda(u-w)+\mu m(u)\right)\mathrm{d}u\mathrm{d}s}\right)\\&=z_{1-\alpha}e^{-\mu t}\left(Z(t)-(\mu-h_F(w))\int_0^t Z(s)\mathrm{d}s\right)\end{aligned}$$

for $Z(t)$ given in statement of the corollary. □

EC.2. Additional Numerical Studies

EC.2.1. Implementation Details

All Monte Carlo simulations were conducted using MATLAB. We sample the values of the performance functions at fixed time points $t_1,\ldots,t_N$, with $t_i\equiv i\Delta T$, $1\le i\le N$, $T=24$, the step size (sampling resolution) is $\Delta T=0.01$, and $N=T/\Delta T=2400$ is the total number of samples in $[0,T]$. To collect simulated data of PWT, on each simulation run, we create *virtual arrivals* to all queues at $t_1,\ldots,t_N$. These virtual customers behave like real customers while in the queue and capture what the system experience would be like for a customer had they arrived at the given sampling time points. However, these virtual customers, when they are eventually moved to the head of the queue and assigned with a server, will not enter service; instead, they are removed immediately from the system after their elapsed waiting times have been recorded. For instance, the j^{th} $(1\le j\le N)$ class-i virtual customer arrives at queue i at time $j\Delta T$. If this customer is removed (from the head of the line) at time t, then the system collects a sample for the class-i PWT at time $j\Delta T$ on the l^{th} run: $V_i^l(j\Delta T)=t-j\Delta T$. The class-$i$ mean PWT and TPoD at time $t_j\equiv j\Delta T$ are estimated by averaging m (e.g., $m=5000$) independent copies of $V_i(j\Delta T)$ and indicators $\mathbf{1}_{\{V_i(j\Delta T)>w_i\}}$.

EC.2.2. Fixed Staffing Intervals

In practice, system managers are often unable to add and remove servers in a nearly continuous manner; they must staff at a certain level for a fixed period of time (i.e. shifts in a hospital). We further expand upon the discretization of the continuous staffing function, by letting staffing decisions be limited to fixed intervals, in which the staffing levels must remain constants. We explore the impact of modifying our prescribed staffing formula to mimic this practical constraint. For a given staffing interval Δ_s (e.g., 30 minutes) and a continuous staffing formula $s(t)$, we consider two Δ_s-based discretization methods (i) *average staffing level* (ASL) and (ii) *maximum staffing level* (MSL), which are given by

$$s^{\mathrm{ASL}}(t) \equiv \sum_{i=1}^{\lceil T/\Delta_s \rceil} \bar{s}_i \mathbf{1}_{\{t\in[(i-1)\Delta_s, i\Delta_s\}}, \qquad \bar{s}_i \equiv \frac{1}{\Delta_s}\int_{(i-1)\Delta_s}^{i\Delta_s \wedge T} s(u)\mathrm{d}u,$$
$$s^{\mathrm{MSL}}(t) \equiv \sum_{i=1}^{\lceil T/\Delta_s \rceil} s_i^{\uparrow} \mathbf{1}_{\{t\in[(i-1)\Delta_s, i\Delta_s\}}, \qquad s_i^{\uparrow} \equiv \sup_{(i-1)\Delta_s \le u \le i\Delta_s \wedge T} s(u),$$

where $x \wedge y \equiv \min(x, y)$. MSL sets the staffing level in each interval as the maximum of staffing

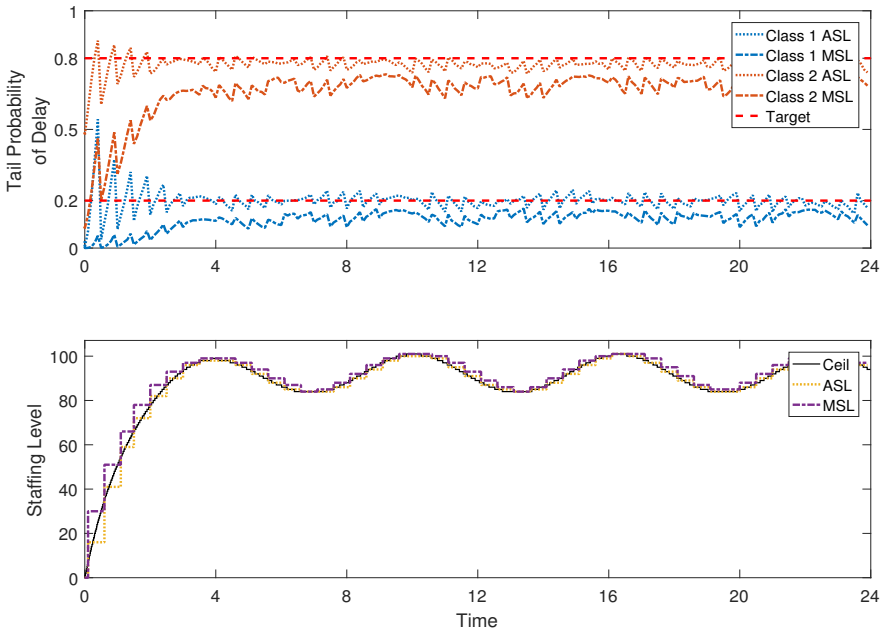

Figure EC.1 Plots of (i) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (top panel) and (ii) time-varying ASL and MSL staffing levels having $\Delta_s = 0.5$ (bottom panel) for the two-class base-case example with 5000 independent runs.

function, ensuring target QoS to be met as we will be slightly overstaffing the system; while ASL uses the average staffing level in each interval to ensure a smaller absolute deviation from the TPoD target. We again simulate our two-class base-case example, but with staffing formulas calculated according to the ASL and MSL methods. We give our simulation results with $\Delta_s = 0.5$ (30 minutes) in Figure EC.1. We observe that both ASL and MSL achieve relative performance stabilization after an initial

warm-up period (approximately the interval $[0,4]$). During the warm-up period, the rate of change in the required staffing is high and an inflexible staffing interval is not able to respond dynamically enough to meet demand. Indeed, ASL achieves better stabilization around the targets while MSL ensures meeting service levels at all times, leading to higher QoS than required. We consider other values for the staffing interval Δ_s in the e-companion.

EC.2.3. Additional numerical examples

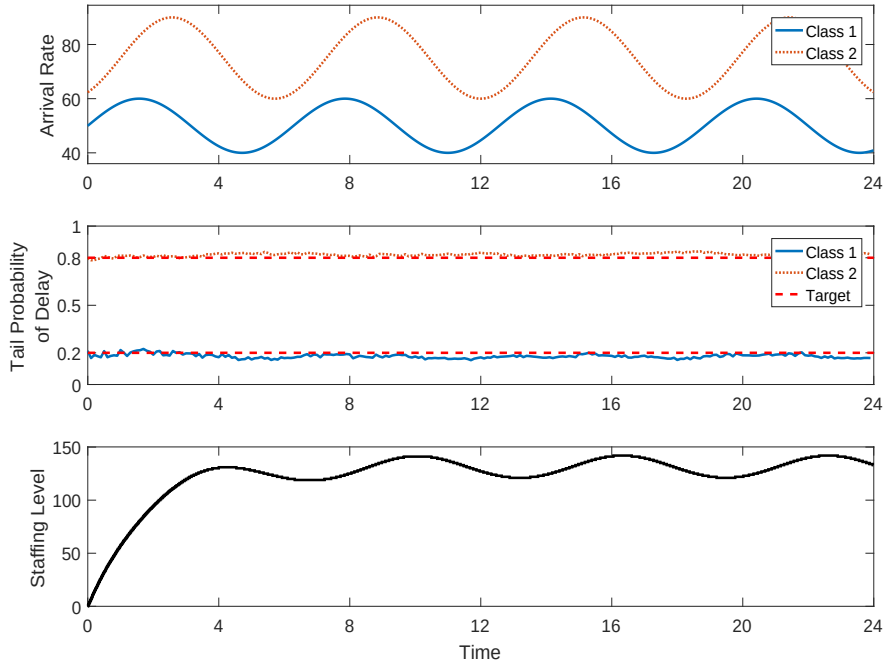

Figure EC.2 Plots of (i) arrival rates (top panel); (ii) simulation estimates of class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel), and (iii) time-varying staffing level (bottom panel) for a two-class model with class-dependent rates wherein $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 50$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$.

EC.2.3.1. Class-dependent service rates Results in §4 enables us to treat the case of class-dependent service rates, which has strong practical relevance. Consider our two-class base-model example with modified service rates $(\mu_1, \mu_2) = (0.5, 1)$ (so that a high priority class requires significantly more time in service). In this case we numerically compute the variance of $\widehat{H}(t)$ and the required control functions using our contraction based algorithm given in Remark EC.1. In the numerical experiment, it takes 42 iterations for the algorithm to converge with an error tolerance $\epsilon = 10^{-6}$. Figure EC.2 shows that our methods continue to achieve desired service-level differentiation and performance stabilization.

EC.2.3.2. Mixed arrival rates We look at the case where arrival rates are of different orders of magnitude. This is relevant because, in practice, certain customer classes may have infrequent arrivals as compared to other classes, see Ding et al. (2019). We modify the arrival rates in our two-class base-case example so that $\bar{\lambda}_1 = 0.1$, but set $n = 100$ so that the overall system size remains

comparable to the base case. We see from Figure EC.3 that even though the majority of arrivals to the system are from Class 1, we have effective TPoD stabilization for both classes.

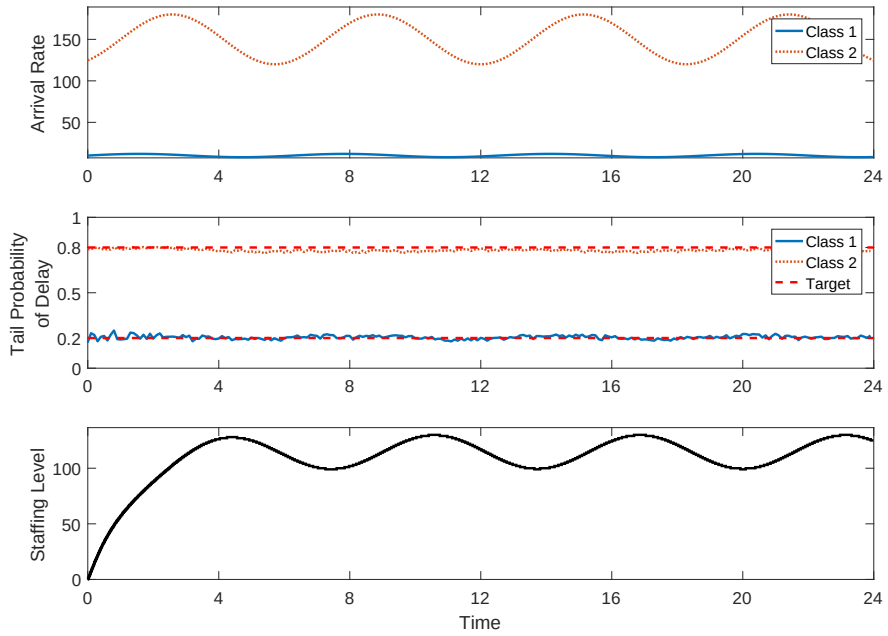

Figure EC.3 Plots of (i) arrival rates (top panel); (ii) simulation estimates of class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel), and (iii) time-varying staffing level (bottom panel) for a two-class model with mixed arrival rates wherein $\bar{\lambda}_1 = 0.1$, $\bar{\lambda}_2 = 1.5$, $n = 100$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$.

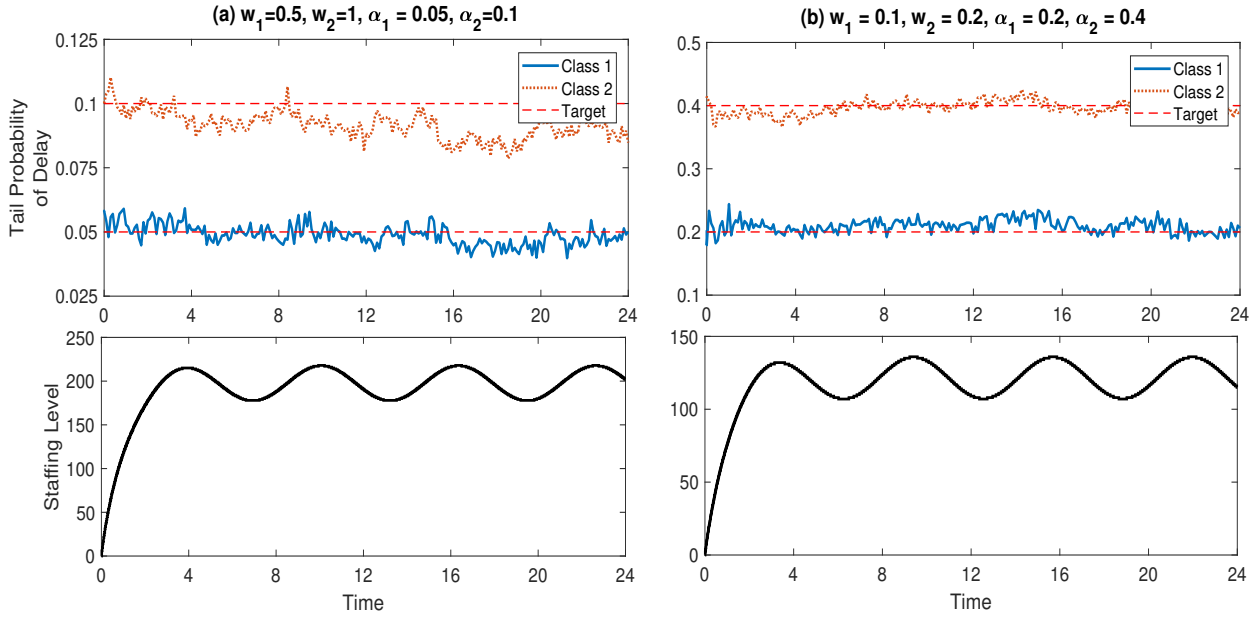

Figure EC.4 The two-class model with high QoS targets: (a) $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.05$, $\alpha_2 = 0.1$ (left), (b) $w_1 = 0.1$, $w_2 = 0.2$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$ (right).

EC.2.3.3. Higher QoS targets In our base model, we set $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$ to test if TPoDs can be indeed significantly differentiated. We now validate the effectiveness of rules (29) and (31) when both classes have higher QoS targets. Figure EC.4 gives the simulation results with (i) smaller probability targets $\alpha_1 = 0.05$ and $\alpha_2 = 0.1$ ($w_1 = 0.5$, $w_2 = 1$); and (ii) smaller delay targets $w_1 = 0.1$ and $w_2 = 0.2$ ($\alpha_1 = 0.2$, $\alpha_2 = 0.4$). Figure EC.4 shows that TPoD's remain relatively stable in both cases.

EC.2.3.4. A Five-Class Example We now consider a five-class V model, having class-dependent sinusoidal arrival rates as in (46), exponential abandonment and service times. All model input parameters and QoS parameters are given in Table EC.1.

Table EC.1 Five Class Model: Class specific parameters and QoS target levels

	Class parameters						Service levels	
Class	λ	r	γ	ϕ	θ	μ	w	α
1	1.0	0.20	0.5	0	0.6	1	0.2	0.1
2	1.5	0.30	1.0	-1	0.3	1	0.4	0.3
3	1.2	0.05	1.3	1	0.5	1	0.6	0.5
4	1.1	0.15	1.6	-2	1.0	1	0.8	0.7
5	1.6	0.40	2.0	2	1.2	1	1.0	0.9

The control functions are given in the left-hand panel of Figure EC.5. In this example, we intentionally let the sinusoidal arrival rates have class-dependent periods, frequencies, and relative amplitudes (see right-hand panel of Figure EC.5). Nevertheless, our method continues to successfully achieves stable TPoD-based service levels across all 5 classes.

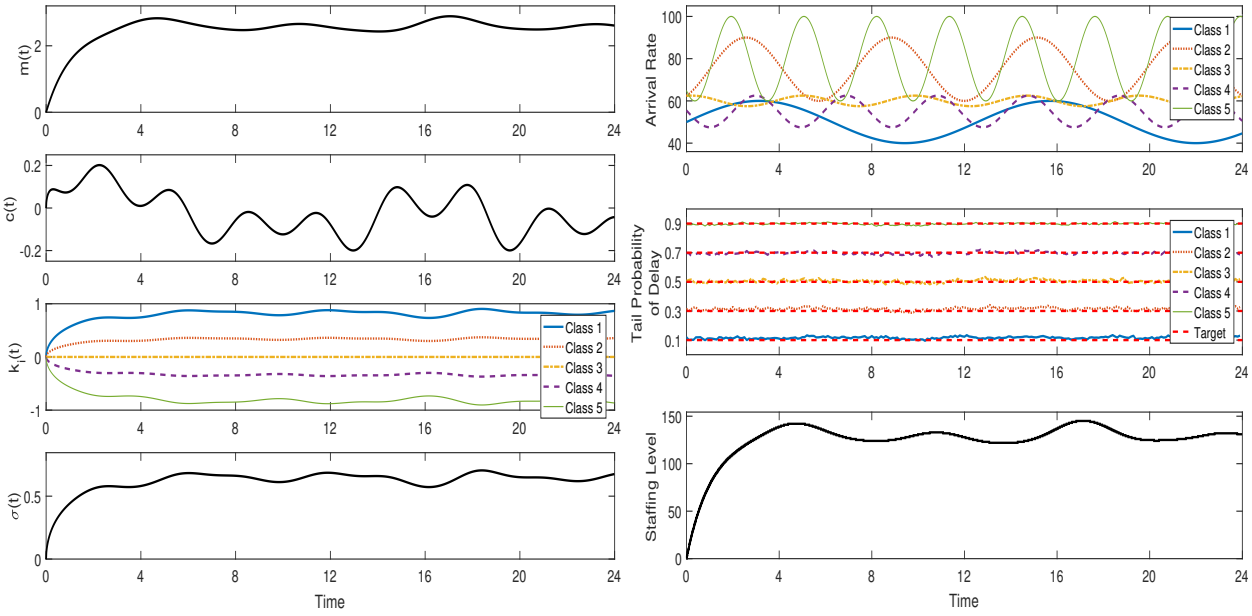

Figure EC.5 A five-class example: (i) Computed control functions $m(t)$, $c(t)$, and $\kappa_i(t)$ for $i = 1, \ldots, 5$ (left), (ii) Simulation comparisons for TPoD $\mathbb{P}(V_i(t) > w_i)$, $i = 1, \ldots, 5$ (right), with $n = 50$, input and QoS parameters given in Table EC.1, and 5000 samples.