# ONLINE APPENDIX

# to

# Staffing and Scheduling to Differentiate Service in Time-Varying Multiclass Service Systems

Yunan Liu[1], Xu Sun[2], and Kyle Hovey[1]

[1]Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695

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

December 3, 2018

**Abstract**

Motivated by large-scale service systems, we study an overloaded multi-class queueing system having time-varying arrivals and customer abandonments. Our objective is to devise appropriate staffing and scheduling policies to achieve differentiated service levels for each customer class. Formally, for a class-specific delay target $w_i > 0$ and probability target $\alpha_i \in (0, 1)$, we concurrently determine a proper staffing level (number of servers) and a scheduling rule (assigning newly idle servers to a waiting customer from one of the classes), under which the probability that a class-$i$ customer waits more than $w_i$ does not exceed $\alpha_i$ at all times. For this purpose, we propose a joint staffing and scheduling policy that is both time dependent (coping with the time variability in arrival pattern) and state dependent (capturing the stochastic variability in service times and arrival times). The proposed framework enables us to treat class-dependent service rate. Effectiveness of our proposed staffing and scheduling plocies is substantiated by heavy traffic limit theorems (as the system scale increases). We also conduct computer simulation experiments to provide engineering confirmations and practical insights.

**Overview.** This appendix provides additional supplementary material to the main paper. In §1 we give a road map of our approach. In §2, we provide all the technical proofs omitted from the main paper. In §3, we give additional numerical studies. First, in Table 1 we give all acronyms used here and in the main paper.

# 1 Road Map of Our Approach

We first explain the main steps of the our approach to achieve the asymptotic service-level differentiation and stabilization, see Figure 1. First, we propose our TV-SRS and TV-DPS policies having some unknown control functions, namely, $c$ and $\kappa_i$, $1 \leq i \leq K$. See §§2.2–2.3 of the main paper for detailed structure of TV-SRS and TV-DPS. In both formulas, we have determined the first order terms (nominal staffing term of TV-SRS, and normalized HWT term of TV-DPS), and we leave the second order terms undetermined ($c$ for TV-SRS and $\kappa_i$ for TV-DPS). What is complicated here is how to treat these second-order terms, and they are driven by the probability targets $\alpha_i$.

Next, we develop convenient many-server FWLLN and F-CLT limits under our proposed TV-SRS and TV-DPS, which is the main focus of our analysis. In Theorem 1 and Corollary 1, we give the FWLLN and FCLT limits for various system functions, including class-dependent queue length, number of busy servers, PWT and HWT. Theorem 1 also establishes an SSC meaning that all $K$ class-dependent PWT and HWT degenerate to a one-dimensional *frontier* process $\widehat{H}$. In Corollary 2, we further analyze this frontier process: we show that (i) $\widehat{H}$ uniquely solves an SVE of which the drift and volatility terms are explicit functions of the model input parameters (thus independent with the unknown control functions); (ii) the solution $\widehat{H}$ is Gaussian; and (iii) the transient mean and variance process of $\widehat{H}$ can be numerically computed using contraction-based algorithms (which converges geometrically fast).

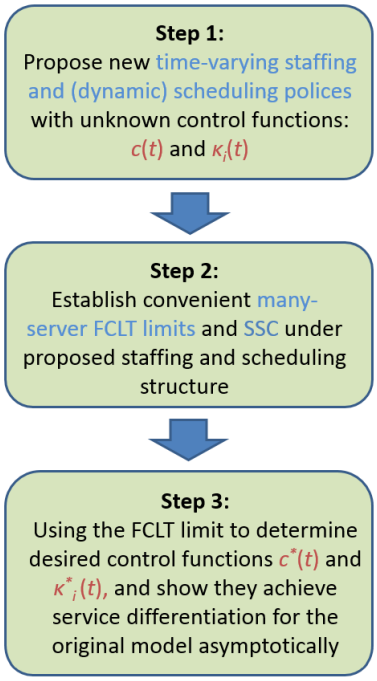


Figure 1: Road map of our approach

Finally, we utilize this frontier process $\widehat{H}$ to determine the "optimal" control functions $c^*$ and $\kappa_i^*$. The idea here is to use our control functions to shift the mean value of $\widehat{V}_i$ so that the probability mass $\{\widehat{V}_i > 0\}$ is controlled at $\alpha_i$. We give the resulting (unique) formulas in Proposition 1; here the variance $\mathrm{Var}(\widehat{H})(t)$ plays an important

| Acronym | Meaning |
|---|---|
| ASL | average staffing level |
| CCDF | complementary cumulative distribution function |
| CDF | cumulative distribution function |
| CTAS | Canadian triage and acuity scale |
| DIS | delayed infinite server |
| DIS-MOL | delayed infinite-server modified-offered-load approximation |
| ED | emergency department |
| ERP | equilibrium renewal process |
| FCLT | functional central limit theorem |
| FPE | fixed-point equation |
| FWLLN | functional weak law of large numbers |
| HWT | head-of-line waiting time |
| HT | heavy traffic |
| i.i.d. | independent and identically distributed |
| KPI | key performance indicator |
| MOL | modified offered load |
| MPSS | marginal price of staffing and scheduling |
| MSHT | many-server heavy-traffic |
| MSL | maximum staffing level |
| 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 |
| SL | service level |
| SSC | state-space collapse |
| SVE | stochastic Volterra equation |
| TPoD | tail probability of delay |
| TTGA | two-term Gaussian approximation |
| TV | time varying |
| TV-SRS | time-varying square-root staffing |
| TV-DPS | time-varying dynamic prioritization scheduling |

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

role in these formulas. In Theorem 2, we show that the *complete* TV-SRS and TV-DPS policies (with structure given in §§2.2–2.3 and control functions given in Proposition 1) successively achieve asymptotic service-level differentiation and stabilization.

Corollaries 2–4 aim to provide additional insights by studying important special cases. For example, Corollary 2 shows that when service rates are class independent, the SVE degenerates to an OU process with time-varying parameters and admits an analytic solution.

# 2 Proofs

We hereby provide all proofs that are omitted in the main paper, including Theorems 1–2, Propositions 1-2, Corollaries 2–4.

## 2.1 Proof of Theorem 1

**Main steps of the proof.** Step 1: We first show that each component within the curly bracket in (15) of the main paper is at most $O(1/\sqrt{n})$ away from the frontier process, that is, $H_i^n(t)/w_i + n^{-1/2}\kappa_i(t) = H^n(t) + O(1/n)$ (or $\widehat{H}_i^n(t) = w_i(\widehat{H}^n(t) - \kappa_i(t)) + O(1/\sqrt{n})$). This is essentially a SSC result and follows from a key observation that, at any given point in time, the number of total departures required for a HoL customer to enter service under the TV-DPS policy is of order $O(1)$. Step 2: We then use (6) of the main paper to obtain a simple relation between $\widehat{H}_i^n$ and $\widehat{B}_i^n$. Based on the fact that the difference between $\widehat{H}_i^n(t)$ and $w_i(\widehat{H}^n(t) - \kappa_i(t))$ can be made arbitrarily small for $n$ large enough, we are able to establish a set of $K$ differential equations and one linear equation jointly satisfied by $(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{H}^n)$. This allows us to apply the Gronwall's inequality to establish the stochastic boundedness of the sequence $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{H}^n); n \in \mathbb{N}\}$, which in turn enables us to deduce the desired FWLLN results. Step 3: An application of the continuous mapping theorem with the established FWLLN allows us to establish the Brownian limits given in (21) of the main paper for the corresponding CLT-scaled processes. Applying the continuous mapping theorem again with these Brownian limits yields the joint convergence of $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{H}^n)\}$. Next, the FCLT for the HWT and PWT processes follows by converging-together lemma with the established FCLT for the frontier process. Step 4: Finally, the FCLT for the queue-length processes follows by first exploiting the relation between $Q_i^n$ and $H_i^n$ and then applying the continuous mapping theorem. □

**Step 1: SSC for the pre-limit HWT and PWT processes.** We start by observing the relation between $H_i^n(t)$ and $V_i^n(t)$

$$V_i^n(t - H_i^n(t)) = H_i^n(t) + O(1/n) \tag{1}$$

*under the TV-DPS rule*, where the error term $O(1/n)$ will follow if the number of customers (from other queues) who have a higher service priority over the HoL customer in the $i$th queue is of order $O(1)$; i.e., it only requires $O(1)$ number of service completions before the HoL customer of the $i$th queue enters service. To see that relation (1) is true, suppose customer $A$ enters service from the $i$th queue at time $t$ and customer $B$ becomes the new HoL customer in queue $i$. Then customer $B$ must have arrived at the system at time $t - H_i^n(t-)$. Further we use $a_i^n$ to denote the inter-arrival time between $A$ and $B$. It is immediate that customer $A$ arrived at the system at time $t - H_i^n(t-) - a_i^n$. Suppose $\kappa_i \equiv 0,\ i \in \mathcal{I} \equiv \{1, \ldots, K\}$ (the case where $\kappa_i$ are not zero functions can be analyzed in a similar fashion). Then under the TV-DPS policy, only those class-$j$ customers who arrived during the interval

$$\left( t - \frac{w_j \left( H_i^n(t) + a_i^n \right)}{w_i}, t - \frac{w_j H_i^n(t)}{w_i} \right) \tag{2}$$

could enter service prior to the time at which customer $B$ enters service. To proceed, we make the following observation: If $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are two independent Poisson processes with rate $\lambda^{(1)}$ and $\lambda^{(2)}$, respectively, then the number of arrivals from $\mathcal{P}^{(2)}$ between two successive arrivals of $\mathcal{P}^{(1)}$ follows a geometric distribution with parameter with parameter $\frac{\lambda^{(1)}}{\lambda^{(1)}+\lambda^{(2)}}$. Now because the interval (2) has a length of $(w_j a_i^n / w_i)$, the number class-$j$ customer who can enter service before $B$ is stochastically bounded by a geometric distributed random variable with mean $\frac{w_j \lambda_j^{\uparrow}}{w_i \lambda_i^{\downarrow}}$ and variance $\left( \frac{w_j \lambda_j^{\uparrow}}{w_i \lambda_i^{\downarrow}} \right)^2 + \frac{w_j \lambda_j^{\uparrow}}{w_i \lambda_i^{\downarrow}}$. By the same token, we can argue that the total number of customers who will enter service before $B$ enters service is bounded by the sum of $K-1$ geometric random variables with mean $\ell_i^{(1)} = \frac{\sum_{j \neq i} w_j \lambda_j^{\uparrow}}{w_i \lambda_i^{\downarrow}}$ and variance $\ell_i^{(2)} = \sum_{j \neq i} \left( \frac{w_j \lambda_j^{\uparrow}}{w_i \lambda_i^{\downarrow}} \right)^2 + \ell_i^{(1)}$.

On the other hand, the inter-arrival times of class $i$ live on the order of $O(1/n)$. Using the same reasoning for (1), we have

$$H_i^n(t)/w_i + n^{-1/2} \kappa_i(t) = H^n(t) - O(1/n),$$

or equivalently,

$$\widehat{H}_i^n(t) = w_i(\widehat{H}^n(t) - \kappa_i(t)) - O(1/\sqrt{n}), \tag{3}$$

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

**Step 2: The FWLLN.** Here we prove the desired FWLLN results by showing the stochastic boundedness of the corresponding CLT-scaled processes; see §5.2 of [5] for a precise definition of stochastic boundedness.

In what follows, 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. To that end, introduce the LLN- and CLT-scaled empirical process

$$\begin{aligned} \bar{U}^n(t,x) &\equiv \frac{1}{n}\sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{X_i \le x\}} \quad \text{for} \quad t \ge 0, \quad 0 \le x \le 1, \quad \text{and} \\ \widehat{U}^n(t,x) &\equiv \sqrt{n}\left(\bar{U}^n(t,x) - \mathbb{E}\left[\bar{U}^n(t,x)\right]\right) = \frac{1}{\sqrt{n}}\left(\sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{X_i \le x\}} - x\right), \end{aligned} \tag{4}$$

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0,1]$. [4] 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. Paralleling (3.3) - (3.6) in [1], we break the enter-service process $E_i^n(t)$ in (6) of the main paper into three pieces, namely,

$$E_i^n(t) = E_{i,1}^n(t) + E_{i,2}^n(t) + E_{i,3}^n(t), \tag{5}$$

where

$$E_{i,1}^n(t) \equiv \sqrt{n}\int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\mathrm{d}\widehat{A}_i^n(u), \quad t \ge 0, \tag{6}$$

$$E_{i,2}^n(t) \equiv \sqrt{n}\int_{-H_i^n(0)}^{t-H_i^n(t)} \int_0^1 \mathbf{1}_{\{y > F_i^c(V_i^n(u))\}}\mathrm{d}\widehat{U}_i^n(\bar{A}_i^n(u), y) \quad t \ge 0, \tag{7}$$

$$E_{i,3}^n(t) \equiv n\int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\lambda_i(u)\mathrm{d}u \quad t \ge 0, \tag{8}$$

for $\bar{A}_i^n, \widehat{A}_i^n$ given by (16) of the main paper and $\widehat{U}_i^n$ is a CLT-scaled empirical process specified by (4).

Define the fluid version and CLT-scaled version of the enter-service process as

$$\varepsilon_i(t) \equiv \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)\mathrm{d}u, \tag{9}$$

$$\widehat{E}_i^n(t) \equiv n^{-1/2}\left(E_i^n(t) - n\varepsilon_i(t)\right) = n^{-1/2}\left(E_i^n(t) - n\int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)\mathrm{d}u\right). \tag{10}$$

Following the decomposition given in (5) - (8), we can write

$$\widehat{E}_i^n(t) = \widehat{E}_{i,1}^n(t) + \widehat{E}_{i,2}^n(t) + \widehat{E}_{i,3}^n(t), \tag{11}$$

where

$$\widehat{E}_{i,1}^n(t) \equiv n^{-1/2}E_{i,1}^n(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\mathrm{d}\widehat{A}_i^n(u) \quad t \ge 0 \tag{12}$$

$$\widehat{E}_{i,2}^n(t) \equiv n^{-1/2}E_{i,2}^n(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} \int_0^1 \mathbf{1}_{\{y > F_i^c(V_i^n(u))\}}\mathrm{d}\widehat{U}_i^n(\bar{A}_i^n(u), y) \quad t \ge 0, \tag{13}$$

$$\widehat{E}_{i,3}^n(t) \equiv n^{-1/2}\left(E_{i,3}^n(t) - n\int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)\mathrm{d}u\right) \quad t \ge 0. \tag{14}$$

For the term $\widehat{E}^n_{i,3}$, we further deduce

$$\begin{aligned}
\widehat{E}^n_{i,3}(t) &= \sqrt{n}\left[\int_{-H^n_i(0)}^{t-H^n_i(t)} F^c_i(V^n_i(u))\lambda_i(u)\mathrm{d}u - \int_{-w_i}^{t-w_i} F^c_i(w_i)\lambda_i(u)\mathrm{d}u\right] \\
&= \sqrt{n}\int_0^t F^c_i(H^n_i(u))\lambda_i(u-H^n_i(u))\mathrm{d}u - \sqrt{n}\int_0^t F^c_i(w_i)\lambda_i(u-w_i)\mathrm{d}u \\
&\quad - \int_0^t F^c_i(H^n_i(u))\lambda_i(u-H^n_i(u))\mathrm{d}\widehat{H}^n_i(u) + O(n^{-1/2}) \\
&= -\int_0^t \left\{f_i(\zeta^n_i(u))\lambda_i(u-\zeta^n_i(u)) + F^c_i(\zeta^n_i(u))\lambda'_i(u-\zeta^n_i(u))\right\} w_i(\widehat{H}^n(u)-\kappa_i(u))\mathrm{d}u \\
&\quad - \int_0^t w_i F^c_i(H^n_i(u))\lambda_i(u-H^n_i(u))\mathrm{d}(\widehat{H}^n(u)-\kappa_i(u)) + O(n^{-1/2}),
\end{aligned} \tag{15}$$

where the second equality follows by a change of variables, namely $t \to t - H^n_i(t)$, plus the relation (1), while the third equality follows from (3) and applying the mean-value theorem with $\zeta^n_i(t)$ satisfying

$$H^n_i(t) \wedge w_i \le \zeta^n_i(t) \le H^n_i(t) \vee w_i. \tag{16}$$

On the other hand, the conservation of flow implies

$$E^n_i(t) = B^n_i(t) + D^n_i(t), \tag{17}$$

where we have used $D^n_i(t)$ to denote the number of class-$i$ customers that have completed service by time $t$. From (9) it follows

$$\varepsilon_i(t) = \int_{-w_i}^{t-w_i} F^c_i(w_i)\lambda_i(u)\mathrm{d}u = \int_0^t F^c_i(w_i)\lambda_i(u-w_i)\mathrm{d}u = m_i(t) + \int_0^t \mu_i m_i(u)\mathrm{d}u, \tag{18}$$

where the last equality follows from (10) of the main paper. Multiplying both sides of (18) by $n$ and subtracting it from (17) yields

$$\begin{aligned}
E^n_i(t) - n\varepsilon_i(t) &= B^n_i(t) - nm_i(t) + D^n_i(t) - n\int_0^t \mu_i m_i(u)\mathrm{d}u \\
&= B^n_i(t) - nm_i(t) + \left(D^n_i(t) - \mu_i\int_0^t B^n_i(u)\mathrm{d}u\right) + \mu_i\left(\int_0^t B^n_i(u)\mathrm{d}u - \int_0^t nm_i(u)\mathrm{d}u\right).
\end{aligned}$$

Next divide both sides by $n^{1/2}$ to get

$$\widehat{E}^n_i(t) = \widehat{B}^n_i(t) + \mu_i\int_0^t \widehat{B}^n_i(u)\mathrm{d}u + \widehat{D}^n_i(t) \quad \text{or} \quad \mathrm{d}\widehat{B}^n_i(t) + \mu_i\widehat{B}^n_i(t)\mathrm{d}t = \mathrm{d}\widehat{E}^n_i(t) - \mathrm{d}\widehat{D}^n_i(t), \tag{19}$$

where we have defined

$$\widehat{D}^n_i(t) \equiv n^{-1/2}\left(D^n_i(t) - \mu_i\int_0^t B^n_i(u)\mathrm{d}u\right).$$

Because the baseline input exceeds the output at all times, it holds that

$$B^n(t) = s^n(t) \tag{20}$$

with arbitrarily high probability for $n$ sufficiently large. Hence, it suffices to focus on the sample paths for which relation (20) holds. In this case we can easily deduce

$$\sum_{i=1}^{K} \widehat{B}_i^n(t) = n^{-1/2}\left(B^n(t) - nm(t)\right) = n^{-1/2}\left(s^n(t) - nm(t)\right) = c(t). \tag{21}$$

Upon substituting (11) - (13) and (15) into(19), we obtain, for $i = 1, \ldots, K$,

$$\begin{aligned}
&\mathrm{d}\widehat{B}_i^n(t) + w_i F_i^c(H_i^n(t))\lambda_i(t - H_i^n(t))\mathrm{d}\widehat{H}^n(t) \\
=& -\mu_i \widehat{B}_i^n(t)\mathrm{d}t - \left[f_i(\zeta_i^n(t))\lambda_i(t-\zeta_i^n(t)) + F_i^c(\zeta_i^n(t))\lambda_i'(t-\zeta_i^n(t))\right] w_i \widehat{H}^n(t)\mathrm{d}t \\
&+ \left[f_i(\zeta_i^n(t))\lambda_i(t-\zeta_i^n(t)) + F_i^c(\zeta_i^n(t))\lambda_i'(t-\zeta_i^n(t))\right] w_i \kappa_i(t)\mathrm{d}t \\
&+ w_i F_i^c(H_i^n(t))\lambda_i(t - H_i^n(t))\mathrm{d}\kappa_i(t) + \mathrm{d}\widehat{E}_{i,1}^n(u) + \mathrm{d}\widehat{E}_{i,2}^n(u) - \mathrm{d}\widehat{D}_i^n(u) + O(n^{-1/2}).
\end{aligned} \tag{22}$$

We can then use (21) to write $\widehat{B}_K^n = c(t) - \sum_{i=1}^{K-1} \widehat{B}_i^n$. Plugging it into (22) for $i = K$, we obtain a set of $K$ linear differential equations with respect to the $K$-dimensional process $(\widehat{B}_1^n, \ldots, \widehat{B}_{K-1}^n, \widehat{H}^n)$. Similar to what was done to (5.14) in [1], we apply the Gronwall's inequality together with the stochastic boundedness of $\widehat{E}_{i,1}^n$, $\widehat{E}_{i,2}^n$, and $\widehat{D}_i^n$ plus the assumed properties of $\lambda_i$, $f_i$ and $F_i^c$ to conclude the stochastic boundedness of the sequence $\{(\widehat{B}_1^n, \ldots, \widehat{B}_{K-1}^n, \widehat{H}^n); n \in \mathbb{N}\}$. In particular, the sequence $\{\widehat{H}^n; n \in \mathbb{N}\}$ is stochastically bounded. In view of (3) and (1), we have that $\{\widehat{H}_n; n \in \mathbb{N}\}$ and $\{\widehat{V}_n; n \in \mathbb{N}\}$ are stochastically bounded, for $i = 1, \ldots, K$. This implies the FWLLN for the HWT and PWT processes, that is, as $n \to \infty$,

$$(H^n, H_1^n, \ldots, H_K^n, V_1^n, \ldots, V_K^n) \Rightarrow (\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}) \quad \text{in} \quad \mathcal{D}^{2K+1}, \tag{23}$$

where the joint convergence holds due to converging-together lemma (Theorem 11.4.5. in [6]).

**Step 3: The FCLT for the waiting time processes.** Similar to the proof of Lemma 5.1 in [1], we invoke the continuous mapping theorem with (12) and (23) to get

$$\widehat{E}_{i,1}^n(t) \Rightarrow \widehat{E}_{i,1}(t) \equiv F_i^c(w_i) \int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)}\mathrm{d}\mathcal{W}_{\lambda_i}(u), \tag{24}$$

where $\mathcal{W}_{\lambda_i}$ is a standard Brownian motion.

To proceed, we argue that, as $n \to \infty$,

$$\widehat{E}_{i,2}^n(t) \Rightarrow \widehat{E}_{i,2}(t) \equiv \sqrt{F_i^c(w_i)F_i(w_i)} \int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)}\mathrm{d}\mathcal{W}_{\theta_i}(u), \tag{25}$$

for $\mathcal{W}_{\theta_i}$ being a standard Brownian independent of $\mathcal{W}_{\lambda_i}$. The essential structure of the proof for (25) is exactly the same as that of A.7.2 in [1], which in turn draws on Theorem 7.1.4 in [2]. 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

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

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

$$\widehat{D}_i^n(\cdot) = n^{-1/2}\left[\Pi_i^d\left(n\mu_i \int_0^{\cdot} \bar{B}_i^n(u)\mathrm{d}u\right) - n\mu_i \int_0^{\cdot} \bar{B}_i^n(u)\mathrm{d}u\right] \Rightarrow \mathcal{W}_{\mu_i}\left(\mu_i \int_0^{\cdot} m_i(u)\mathrm{d}u\right) \quad \text{as} \quad n \to \infty, \tag{26}$$

where we have defined $\Pi_i^d$ 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 (19) of the main paper, we will need to strengthen (24), (25) and (26) 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 (17). 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 [**?** ] for a reference, which is a variant of Theorem 7.6 of [5].

To derive a set of SDEs satisfied by the CLT-scaled processes $(\widehat{H}^n, \widehat{B}_1^n, \ldots, \widehat{B}_K^n)$, we seek to simplify the right-hand side of (15). First we note that the inequality (16) and the convergence in (25) imply

$$\zeta_i^n(t) = w_i + O(n^{-1/2}) = H_i^n(t) + O(n^{-1/2}). \tag{27}$$

We then use integration by parts to deduce

$$\begin{aligned}
&- \int_0^t w_i F_i^c(\zeta_i^n(u))\lambda_i'(u - \zeta_i^n(u))(\widehat{H}^n(u) - \kappa_i(u))\mathrm{d}u \\
&\quad - \int_0^t w_i F_i^c(H_i^n(u))\lambda_i(u - H_i^n(u))\mathrm{d}(\widehat{H}^n(u) - \kappa_i(u)) \\
&= - w_i F_i^c(\zeta_i^n(t))\lambda_i(t - \zeta_i^n(t))(\widehat{H}^n(t) - \kappa_i(t)) \\
&\quad + \int_0^t w_i \left\{F_i^c(\zeta_i^n(u))\lambda_i(u - \zeta_i^n(u)) - F_i^c(H_i^n(u))\lambda_i(u - H_i^n(u))\right\}\mathrm{d}(\widehat{H}^n(u) - \kappa_i(u)) \\
&\quad + \int_0^t w_i \lambda_i(u - \zeta_i^n(u))(\widehat{H}^n(u) - \kappa_i(u))\mathrm{d}F_i^c(\zeta_i^n(u)) \\
&= - w_i F_i^c(w_i)\lambda_i(t - w_i)(\widehat{H}^n(t) - \kappa_i(t)) + O(n^{-1/2}),
\end{aligned} \tag{28}$$

where the last equality holds due to (27). Upon plugging (28) into (15), we obtain

$$\widehat{E}^n_{i,3}(t) = -\int_0^t w_i f_i(w_i)\lambda_i(u-w_i)(\widehat{H}^n(u)-\kappa_i(u))\mathrm{d}u - w_i F^c_i(w_i)\lambda_i(t-w_i)(\widehat{H}^n(t)-\kappa_i(t)) + O(n^{-1/2}).$$

Now plugging (11) and the equation above into (19), we get

$$\begin{aligned}
&\widehat{B}^n_i(t) + w_i F^c_i(w_i)\lambda_i(t-w_i)\widehat{H}^n(t) \\
=& -\mu_i\int_0^t \widehat{B}^n_i(u)\mathrm{d}u - \int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\widehat{H}^n(u)\mathrm{d}u + \int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\kappa_i(u)\mathrm{d}u \\
&+ w_i F^c_i(w_i)\lambda_i(t-w_i)\kappa_i(t) + \widehat{E}^n_{i,1}(t) + \widehat{E}^n_{i,2}(t) - \widehat{D}^n_i(t) + O(n^{-1/2}) \quad \text{for} \quad i=1,\ldots,K.
\end{aligned} \tag{29}$$

The joint convergence $(\widehat{H}^n, \widehat{B}^n_i, \ldots, \widehat{B}^n_K) \Rightarrow (\widehat{H}, \widehat{B}_i, \ldots, \widehat{B}_K)$ then follows by applying the continuous mapping theorem (see Theorem 4.1 of [5]) to (20) and (29), with the *joint* convergence of $\widehat{E}^n_{i,1}, \widehat{E}^n_{i,2}$ and $\widehat{D}^n_i$, as specified by (24), (25) and (26), respectively. Alternatively, one can subtract (29) by (20) in the main paper and invoke the Gronwall's inequality to show that the difference between the pre-limit and the limit is bounded by a negligible term as $n \to \infty$, as was done in the proof of (4.7) in [1]. The convergence of $\{\widehat{H}^n_i; n \in \mathbb{N}\}$ and $\{\widehat{V}^n_i; n \in \mathbb{N}\}$ follow easily from and (3) and (1), respectively.

**Step 4: The FCLT for the queue-length processes.** To show that $\{\widehat{Q}^n_i; n \in \mathbb{N}\}$ converges to the corresponding limit, we decompose the right-hand side of (7) in the main paper into three terms, namely,

$$Q^n_i(t) = Q^n_{i,1}(t) + Q^n_{i,2}(t) + Q^n_{i,3}(t), \tag{30}$$

where

$$Q^n_{i,1}(t) \equiv \sqrt{n}\int_{t-H^n_i(t)}^t F^c_i(t-u)\mathrm{d}\widehat{A}^n_i(u), \quad t \geq 0, \tag{31}$$

$$Q^n_{i,2}(t) \equiv \sqrt{n}\int_{t-H^n_i(t)}^t \int_0^1 \mathbf{1}_{\{x > F^c_i(t-u)\}}\mathrm{d}\widehat{U}^n_i(\bar{A}^n_i(u), x) \quad t \geq 0, \tag{32}$$

$$Q^n_{i,3}(t) \equiv n\int_{t-H^n_i(t)}^t F^c_i(t-u)\lambda_i(u)\mathrm{d}u \quad t \geq 0, \tag{33}$$

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

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

$$\text{where} \quad \widehat{Q}_{i,1}^n(t) \equiv \int_{t-H_i^n(t)}^t F_i^c(t-u)\mathrm{d}\widehat{A}_i^n(u) \;\Rightarrow\; \int_{t-w_i}^t F_i^c(t-u)\mathrm{d}\widehat{A}_i(u), \tag{34}$$

$$\widehat{Q}_{i,2}^n(t) \equiv \int_{t-H_i^n(t)}^t \int_0^1 \mathbf{1}_{\{x>F_i^c(t-u)\}}\mathrm{d}\widehat{U}_i^n(\bar{A}_i^n(u), x)$$

$$\Rightarrow \int_{t-w_i}^t \sqrt{F_i^c(t-u)F_i(t-u)\lambda_i(u)}\mathrm{d}\mathcal{W}_{\theta_i}(u), \tag{35}$$

$$\widehat{Q}_{i,3}^n(t) \equiv \sqrt{n}\int_{t-H_i^n(t)}^{t-w_i} F_i^c(t-u)\lambda_i(u)\mathrm{d}u \;\Rightarrow\; F_i^c(w_i)\lambda_i(t-w_i)\widehat{H}_i(t). \tag{36}$$

Here the proof for (34) and (35) is very similar to that of (24) and (25), and the proof for (36) is also straightforward. □

## 2.2 Proof of Proposition 1

The multi-dimensional SDE in (20) of the main paper is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\mu_i t}\tilde{B}_i(t)\right) = e^{\mu_i t}\left(-w_i F_i^c(w_i)\lambda_i(t-w_i)\widehat{H}(t) - \int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\widehat{H}(u)\mathrm{d}u + y_i(t) + G_i(t)\right), \tag{37}$$

where

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

Integrating (37) from 0 to $t$ yields

$$\begin{aligned}
\tilde{B}_i(t) &= e^{-\mu_i t}\int_0^t e^{\mu_i s}\left(-w_i F_i^c(w_i)\lambda_i(s-w_i)\widehat{H}(s) - \int_0^s w_i f_i(w_i)\lambda_i(u-w_i)\widehat{H}(u)du + y_i(s) + G_i(s)\right)\mathrm{d}s \\
&= e^{-\mu_i t}\left(-\int_0^t e^{\mu_i s} w_i F_i^c(w_i)\lambda_i(s-w_i)\widehat{H}(s)\mathrm{d}s - \int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\widehat{H}(u)\int_u^t e^{\mu_i s}\mathrm{d}s\mathrm{d}u\right. \\
&\quad \left. + \int_0^t e^{\mu_i s} y_i(s)ds + \int_0^t e^{\mu_i s} G_i(s)\mathrm{d}s\right) \\
&= \int_0^t w_i\lambda_i(s-w_i)\left(-F_i^c(w_i)e^{\mu_i(s-t)} - f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\widehat{H}(s)\mathrm{d}s \\
&\quad + \int_0^t e^{\mu_i(s-t)} y_i(s)ds + \int_0^t e^{\mu_i(s-t)} G_i(s)\mathrm{d}s.
\end{aligned}$$

Summing up over $i$ from 1 to $K$, we have

$$\begin{aligned}
\int_0^t c(s)\mathrm{d}s = \sum_{i=1}^K \tilde{B}_i(t) &= \int_0^t \sum_{i=1}^K w_i\lambda_i(s-w_i)\left(-F_i^c(w_i)e^{\mu_i(s-t)} - f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\widehat{H}(s)\mathrm{d}s \\
&+ \sum_{i=1}^K \int_0^t e^{\mu_i(s-t)}\left(w_i F_i^c(w_i)\lambda_i(s-w_i)\kappa_i(s) + \int_0^s w_i f_i(w_i)\lambda_i(u-w_i)\kappa_i(u)du\right)\mathrm{d}s \\
&+ \sum_{i=1}^K \int_0^t e^{\mu_i(s-t)}\int_0^s \sqrt{F_i^c(w_i)\lambda_i(u-w_i)+\mu_i m_i(u)}\mathrm{d}\mathcal{W}_i(u)\mathrm{d}s \\
&= \sum_{i=1}^K \int_0^t w_i\lambda_i(s-w_i)\left(-F_i^c(w_i)e^{\mu_i(s-t)} - f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\widehat{H}(s)\mathrm{d}s \\
&+ \sum_{i=1}^K \int_0^t w_i\lambda_i(s-w_i)\kappa_i(u)\left(F_i^c(w_i)e^{\mu_i(s-t)} + f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\mathrm{d}u \\
&+ \sum_{i=1}^K \int_0^t \frac{1-e^{\mu_i(u-t)}}{\mu_i}\sqrt{F_i^c(w_i)\lambda_i(u-w_i)+\mu_i m_i(u)}\mathrm{d}\mathcal{W}_i(u). \qquad (38)
\end{aligned}$$

Differentiating (38) yields

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

And further aggregating the independent Brownian motions $\mathcal{W}_1, \ldots, \mathcal{W}_K$ into $\mathcal{W}$ yields the SVE in (26) of the main paper.

**Uniqueness and existence of solution to the SVE (25) of the main paper.** Consider two functions $x, y \in \mathbb{C}$ satisfying an equation

$$x(t) = \int_0^t L(t,s)x(s)\mathrm{d}s + y(t). \qquad (39)$$

we show that (39) 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

$$\psi(x) = \int_0^t L(t,s)x(s)\mathrm{d}s + y(t). \qquad (40)$$

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

$$x = \psi(x). \qquad (41)$$

We first prove that $\psi$ 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

$$\begin{aligned} |\psi(x_1)(t) - \psi(x_2)(t)| &\le \int_0^t |L(t,s)| \mathrm{d}s \cdot \|x_1 - x_2\|_T \\ &\le \|x_1 - x_2\|_T \left( \frac{\sum_{i=1}^K w_i \lambda_i^{\uparrow} \left( \mu_i F_i^c(w_i) + f_i(w_i) \right)}{\sum_{i=1}^K w_i \lambda_i^{\downarrow} F_i^c(w_i)} \right) t. \end{aligned} \tag{42}$$

Hence, we have $\|\psi(x_1) - \psi(x_2)\|_T \le L^{\uparrow} T \|x_1 - x_2\|_T$, where the constant

$$L^{\uparrow} = \frac{\sum_{i=1}^K w_i \lambda_i^{\uparrow} \left( \mu_i F_i^c(w_i) + f_i(w_i) \right)}{\sum_{i=1}^K w_i \lambda_i^{\downarrow} F_i^c(w_i)} < \infty, \tag{43}$$

which is guaranteed by the strict positivity assumptions on $w_i$, $\lambda_i$ and $F_i^c$ for all $1 \le i \le K$. In case $L^{\uparrow} T > 1$, we can partition the interval $[0, T]$ to successive smaller intervals with length $\Delta T$ 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 (41) has a unique solution over the entire interval $[0, T]$.

Consequently, the function $\phi$ specified by (39) is well-defined because $\phi(y)$ has one and only one image for any $y$. So we conclude that SDE (25) in the main paper has a unique solution $\widehat{H}$. If fact, we can write solution as

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

**Remark 2.1** *The strict positivity assumptions on $\lambda_i$ and $F_i^c$ for all classes $1 \le i \le K$ can be relaxed. Note that the contraction property (42) continues to hold as long as there exists a class $i$ such that $\lambda_i^{\downarrow}$ and $F_i^c(w_i)$ are both positive.*

To show that $\widehat{H}$ is Gaussian, we again use the contraction $\psi$ defined in (40). We follow the steps that establish strong solutions in [3]. 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 $\psi$ 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 $\mu_k$ and $\sigma_k^2$ 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 $\mu_k$ and $\sigma_k^2$, 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 (26) of the main paper yields

$$m_{\widehat{H}}(t) = \int_0^t L(t,s) m_{\widehat{H}}(s) \mathrm{d}s + K(t), \qquad \text{where} \quad m_{\widehat{H}}(t) = \mathbb{E}[\widehat{H}(t)]. \tag{44}$$

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 $x_1, x_2 \in \mathbb{C}$,

$$|\Gamma(x_1)(t) - \Gamma(x_2)(t)| \leq \int_0^t |L(t,s)||x_1(s) - x_2(s)| ds \leq L^{\uparrow} t \|x_1 - x_2\|_t,$$

where the finite upperbound $L^{\uparrow}$ is given by (43). The rest of the proof is similar.

To treat the variance of $\widehat{H}$, consider the SVE (25) of the main paper at $0 \leq s, t \leq T$

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

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

$$C = \Theta(C), \tag{45}$$

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

$$\begin{aligned} \Theta(C)(t,s) = &- \int_0^t \int_0^s L(t,u) h(s,v) C(u,v) \mathrm{d}v \mathrm{d}u + \int_0^t L(t,u) C(u,s) \mathrm{d}u \\ &+ \int_0^s L(s,v) C(t,v) \mathrm{d}v + \int_0^{s \wedge t} J(t,u) J(s,u) \mathrm{d}u. \end{aligned} \tag{46}$$

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

$$\begin{aligned} |\Theta(x_1)(t,s) - \Theta(x_2)(t,s)| \leq &\int_0^t \int_0^s |L(t,u) L(s,v)| \cdot |x_1(u,v) - x_2(u,v)| \mathrm{d}v \mathrm{d}u \\ &+ \int_0^t |L(t,u)| \cdot |x_1(u,s) - x_2(u,s)| \mathrm{d}u + \int_0^s |L(s,v)| \cdot |x_1(t,v) - x_2(t,v)| \mathrm{d}v \\ \leq &\left( (L^{\uparrow})^2 ts + L^{\uparrow} t + L^{\uparrow} s \right) \|x_1 - x_2\|_T. \end{aligned}$$

The contraction property is guaranteed if we pick some small enough $\Delta T > 0$ such that $\left((L^{\uparrow})^2\Delta T^2 + 2L^{\uparrow}\Delta T\right) < 1$. 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 2.2 (Numerical Algorithm for $\sigma^2_{\widehat{H}}(t)$)** *The above proof of the existence and uniqueness of the FPE (45) 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 $k^{\text{th}}$ iteration, let $C^{(k+1)} = \Theta\left(C^{(i)}\right)$ with $\Theta$ given in (46).*

$(iii)$ *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 $0 \le t \le T$, which will be used later to devise required control functions $c$ and $\kappa_i$.* □

## 2.3 Proof of Proposition 2

First note that the FPE (28) of the main paper specifies a well-defined function $\phi : \mathbb{C} \to \mathbb{C}$ such that

$$M_{\widehat{H}} = \phi(K). \tag{47}$$

See the proof of the uniqueness and existence of the SVE (specifically, see (39)–(43)) for details. (In fact, it is not hard to see that the function $\phi$ in (47) is Liptchitz continuous and linear.)

Let $(\boldsymbol{\kappa}^*, c^*) \equiv (\kappa_1^*, \ldots, \kappa_K^*, c^*)$, with $\kappa_i^*$ and $c^*$ given in (34) and (33) of the main paper. Let $K^*$ and $M^*_{\widehat{H}}$ be the corresponding version of (27) of the main paper and the mean of $\widehat{H}$. (We know that $K^*(t) = M^*_{\widehat{H}}(t) = 0$.) So we have

$$\kappa_i^*(t) = \kappa_i^*(t) - M^*_{\widehat{H}}(t) = z_{1-\alpha_i}\sigma_{\widehat{H}}(t), \qquad 1 \le i \le K. \tag{48}$$

Now consider another solution to $(\tilde{\boldsymbol{\kappa}}, \tilde{c})$ to (32) of the main paper, 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 (27) of the main paper and mean of $\widehat{H}$. By (32) of the main paper, we have

$$\kappa_i^*(t) + \Delta\kappa_i(t) - \tilde{M}_{\widehat{H}}(t) = z_{1-\alpha_i}\sigma_{\widehat{H}}(t),, \qquad 1 \le i \le K. \tag{49}$$

Comparing (48) with (49), we must have

$$\Delta\kappa_i(t) = \tilde{M}_{\widehat{H}}(t) - M^*_{\widehat{H}}(t) \equiv \Delta\kappa(t) \qquad \text{for all } 1 \le i \le K. \tag{50}$$

Hence, any alternative solution to (32) of the main paper (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

$$\Delta\kappa(t) = \tilde{M}_{\widehat{H}}(t) - M^*_{\widehat{H}}(t) = \int_0^t L(t,s)\left(\tilde{M}_{\widehat{H}}(s) - M^*_{\widehat{H}}(s)\right)\mathrm{d}s + \left(\tilde{K}(t) - K^*(t)\right),$$
$$\text{or equivalently} \quad \Delta\kappa = \tilde{M}_{\widehat{H}} - M^*_{\widehat{H}} = \phi\left(\tilde{K} - K^*\right),$$

meaning

$$\Delta\kappa(t) = \int_0^t L(t,s)\Delta\kappa(s)\mathrm{d}s + \left(\tilde{K}(t) - K^*(t)\right). \tag{51}$$

By (50) and (27) of the main paper, we have

$$\tilde{K}(t) - K^*(t) = \frac{\Delta\kappa(t)\sum_{i=1}^K \left(\eta_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))\mathrm{d}s\right) - \Delta c(t)}{\eta(t)}. \tag{52}$$

Finally, combining (51) with (52), 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 above holds by (27) of the main paper. Therefore, we can see that $c$ is indeed unique, but $\kappa_i$ is only unique up to adding a arbitrary common function $\Delta$, which is consistent with our intuition. □

## 2.4 Proof of Theorem 2

The FCLT limits in Theorem 2 implies the FWLLN, that is, we have

$$(H_i^n, V_i^n) \Rightarrow (w_i\mathfrak{e}, w_i\mathfrak{e}) \quad \text{in} \quad \mathcal{D}^2, \quad \text{for} \quad 1 \le i \le K, \quad \text{as} \quad n \to \infty,$$

where $\mathfrak{e}(t) = 1$. To prove part (i) of Theorem 2, it is sufficient to show that $\{V_i^n, n \ge 1\}$ and $\{H_i^n, n \ge 1\}$ are *uniformly integrable* (u.i.).

We first prove that the queue length $Q_i^n$ is u.i. To do so, note that $Q_i^n$, which is further bounded by the queue length of an $M_t/GI/\infty$ infinite-server model, having arrival rate $\lambda_i^n(t)$ and service hazard rate

$\tilde{h}_i(x) = \min\{h_i(x), \mu_i\}$. Denote its queue length by $X^n_\infty(t)$. We have $\bar{Q}^n(t) \leq_{st} \bar{X}^n_\infty(t)$. Because $X^n_\infty(t)$ is a Poisson r.v., the u.i. of $\bar{X}^n_\infty(t)$ is straightforward. Specifically, we have

$$\sup_n \mathbb{E}\left[(\bar{X}^n_\infty(t))^2\right] = \sup_n \left[\frac{\int_0^t \lambda_i^n(t-x)G_i(x)\mathrm{d}x}{n} + \left(\frac{\int_0^t \lambda_i^n(t-x)G_i(x)\mathrm{d}x}{n}\right)^2\right] < \infty, \tag{53}$$

where $G_i$ is the cdf having hazard rate $\tilde{h}_i$. See Proposition A.2.2 in [2].

Next, we write the PWT

$$V_i^n(t) = \sum_{j=0}^{Q_i^n(t)} U_j,$$

where $U_j$ is the time between the $j^{\text{th}}$ and $(j+1)^{\text{th}}$ departure times of existing waiting customers at queue $i$. Here a departure includes abandonment and entrance to service. Then

$$\begin{aligned}\mathbb{E}\left[V_i^n(t)^2\right] &= \mathbb{E}\left[\sum_{j=0}^{Q_i^n(t)} (U_j)^2 + \sum_{j\neq k} U_j U_k\right] \\ &\leq (\mathbb{E}[Q_i^n(t)]+1)\frac{2\ell_i^2+\ell_i}{(nm^{\downarrow}\tilde{\mu})^2} + \mathbb{E}[Q_i^n(t)^2 + Q_i^n(t)]\frac{\ell_i^2}{(nm^{\downarrow}\tilde{\mu})^2}\end{aligned}$$

where $\tilde{\mu} \equiv \min_{1\leq i\leq K} \mu_i$.

Using the bound in (53), we have $\sup_n \mathbb{E}\left[V_i^n(t)^2\right] < \infty$, which implies u.i. of $V_i^n(t)$. The u.i. of $H_i^n$ is straightforward because $0 \leq H_i^n(t) \leq T + w_i$.

We next prove part (ii) of Theorem 2. 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 (22) of the main paper. □

## 2.5 Proof of Corollaries 2–4

**Proof of Corollary 2.** Because the functions $L(t,s)$ and $J(t,s)$ are now separable in $t$ and $s$, SDE of $\widehat{H}$ becomes

$$\widehat{H}(t) = \frac{1}{R(t)}\int_0^t \tilde{L}(s)\widehat{H}(s)\mathrm{d}s + \frac{1}{R(t)}\int_0^t \tilde{J}(s)\mathrm{d}\mathcal{W}(s) + K(t), \tag{54}$$

where $R(t)$, $\tilde{L}(t)$ and $\tilde{J}(t)$ are specified in Proposition 1. Multiplying $R(t)$ on both sides and differentiating (54) yield

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

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,

$$\begin{aligned}\widehat{H}(t)=&\int_0^t e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\frac{\tilde{J}(u)}{R(u)}\mathrm{d}\mathcal{W}(u)\\&+\int_0^t e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\mathrm{d}K(u)+\int_0^t e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}\frac{K(u)R'(u)}{R(u)}\mathrm{d}u.\end{aligned}\tag{55}$$

Note that

$$e^{-\int_u^t \frac{R'(v)-\tilde{L}(v)}{R(v)}\mathrm{d}v}=e^{\log R(u)-\log R(t)}e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}\mathrm{d}v}=\frac{R(u)}{R(t)}e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}\mathrm{d}v}.\tag{56}$$

Combining (55) and (56) yields the desired solution in Corollary 2. The variance formula in Corollary 2 easily follows from the isometry of the Brownian integral. □

**Proof of Corollary 3.** When $\lambda_i(t)=\lambda_i$, and $\mu_i=\mu$ we automatically achieve

$$m_i(t)=m_i\equiv\frac{\lambda_i F_i^c(w_i)}{\mu},\qquad \eta(t)=\eta\equiv\sum_{i=1}^K\eta_i=\sum_{i=1}^K w_i\lambda_i F_i^c(w_i)\tag{57}$$

and the variance formula simplifies to

$$\begin{aligned}\sigma^2_{\widehat{H}}(t)&=\frac{\sum_{i=1}^K\left(F_i^c(w_i)\lambda_i+\mu m_i\right)}{e^{2\mu t}\eta^2}\int_0^t e^{\frac{2}{\eta}\left(\eta\mu u+\int_u^t\sum_{i=1}^K\eta_i(\mu-h_{F_i}(w_i))dv\right)}du\\&=\frac{2\sum_{i=1}^K F_i^c(w_i)\lambda_i}{\eta^2}e^{-\frac{2t}{\eta}\sum_{i=1}^K\eta_i h_{F_i}(w_i)}\int_0^t e^{\frac{2u}{\eta}\left(\sum_{i=1}^K\eta_i h_{F_i}(w_i)\right)}du.\end{aligned}$$

Now, the integral is of an elementary function and after simplification and letting $t$ go to infinity, we have the following expression for the asymptotic variance:

$$\sigma^2_{\widehat{H}}(t)=\frac{\sum_{i=1}^K F_i^c(w_i)\lambda_i}{\eta\sum_{i=1}^K\eta_i h_{F_i}(w_i)\sum_{i=1}^K\eta_i h_{F_i}(w_i)}\left(1-e^{-\frac{2t}{\eta}\sum_{i=1}^K\eta_i h_{F_i}(w_i)}\right)\to\frac{\sum_{i=1}^K\lambda_i F_i^c(w_i)}{\sum_{i=1}^K w_i\lambda_i F_i^c(w_i)\sum_{i=1}^K w_i\lambda_i f_i(w_i)},$$

Therefore, at $t\to\infty$, we have

$$\kappa_i(t)\to\kappa_i=z_{1-\alpha_i}\sqrt{\frac{\sum_{i=1}^K\lambda_i F_i^c(w_i)}{\sum_{i=1}^K w_i\lambda_i F_i^c(w_i)\sum_{i=1}^K w_i\lambda_i f_i(w_i)}},\qquad c_i(t)\to c_i\equiv\frac{w_i\lambda_i f_i(w_i)}{\mu}\kappa_i.$$

And the convergence occurs exponentially fast. □

**Proof of Corollary 4.** When $K = 1$, the variance formula (39) 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 Corollary 5. □

# 3 Additional Numerical Studies

## 3.1 Implementation Details

All Monte Carlo simulations were conducted using 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, 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 frequent *virtual arrivals* at all queues with interarrival time $\Delta T$. 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 $j^{\text{th}}$ ($1 \le j \le N$) 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 $l^{\text{th}}$ 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 $\mathbf{1}_{\{V_i(j\Delta T) > w_i\}}$, namely, we use the unbiased Monte-Carlo estimators

$$\widehat{\mathbb{E}[V_i(t_j)]} \equiv \frac{1}{m} \sum_{l=1}^{m} V_i^l(j\Delta T) \qquad \text{and} \qquad \widehat{\mathbb{P}(V_i(t_j) > w_i)} \equiv \frac{1}{m} \sum_{l=1}^{m} \mathbf{1}_{\{V_i^l(j\Delta T) > w_i\}}.$$

The numerical integrations (for the variance formulas and control functions) were done using the trapezoidal method in MATLAB.

### 3.2 Additional simulations for TPoD using the two-class base model

#### 3.2.1 Confidence intervals

Supplementing Figure 4 in the main paper, we provide a the $100(1-\beta)\% = 99\%$ confidence intervals (CIs) for the two-class based model. See Figure 2 for the variance bands (shaded in grey). Recall at every $t \in [0,T]$, we use the unbiased estimator $Y_i(t) \equiv \mathbf{1}_{\{V_i(t)>w_i\}}$ to estimator the class-$i$ TPoD, which has mean $\mathbb{E}[Y_i(t)] \approx \alpha_i$ and variance $\text{Var}(Y_i(t)) \approx \alpha_i(1-\alpha_i)$. Based on this, the half width of the CI is $z_{1-\beta/2}\sqrt{\text{Var}(Y_i(t))}/\sqrt{m} \approx 0.0146$, when the number of samples is $m = 5000$. Figure 2 shows that our Monte-Carlo simulations indeed yield accurate estimations of TPoDs, which meet desired targets at all times.

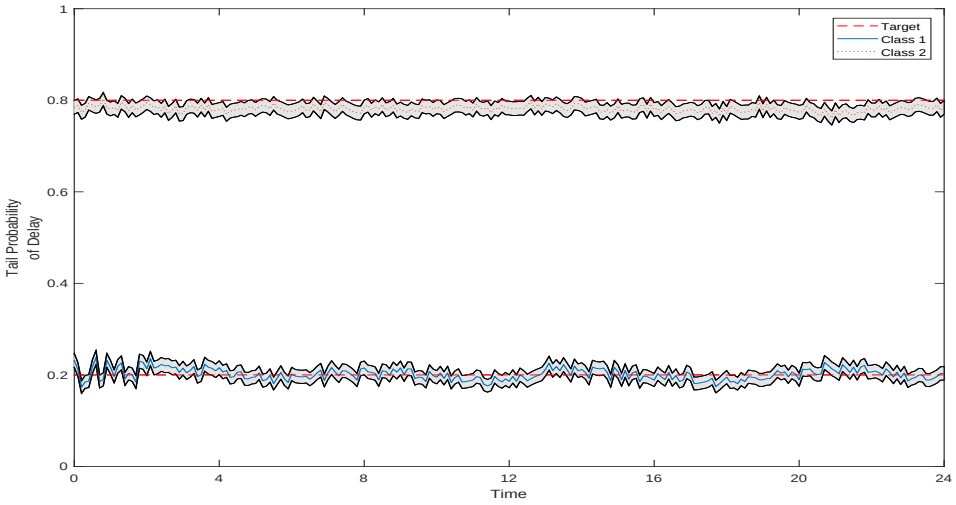


Figure 2: Simulation comparison for a two-class model with 99% confidence intervals simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$, with $\mu_1 = \mu_2 = 1$, $n = 50$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 5000 independent runs.

#### 3.2.2 Class-dependent service rates

We extend the two-class base model in §4.1 of the main paper to the case of class-dependent service rates, $\mu_1 = 0.5$ and $\mu_2 = 1$. In this case we numerically compute the variance process of $\widehat{H}(t)$ and required control functions using our contraction based algorithm given in Remark 2.2. We pick the error tolerance $\epsilon = 10^{-6}$; and our contraction-based algorithm converges in 42 iterations. Similar to the case of class-independent service rates, TV-SRS and TV-DPS continue to achieve good TPoD performance. See Figure 3 for the simulation results.

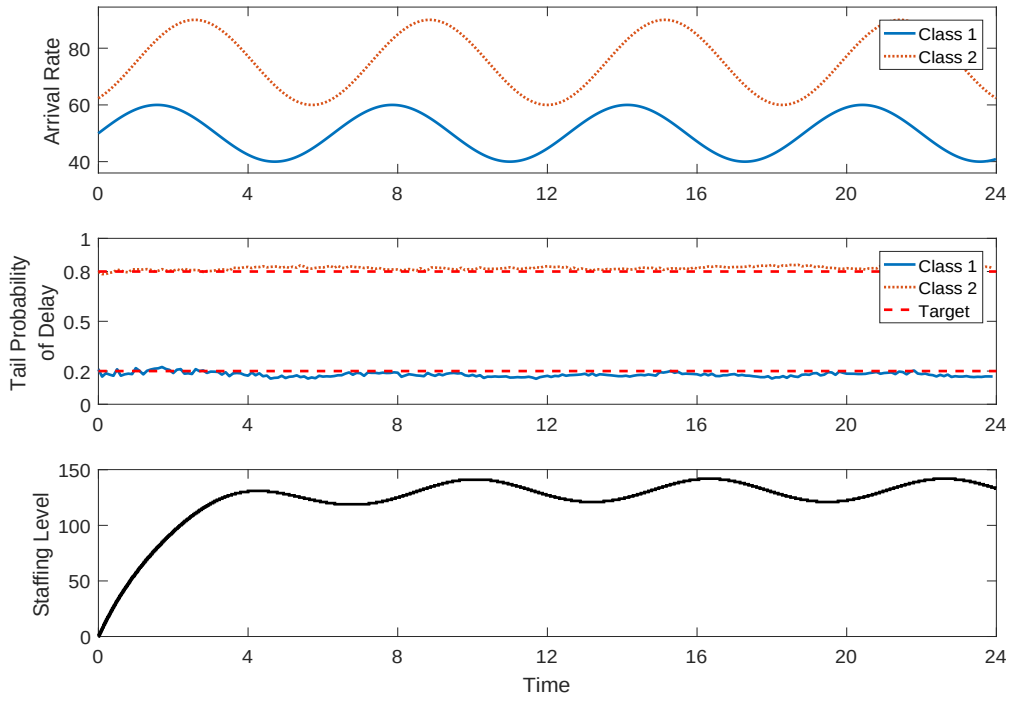


Figure 3: Simulation comparison for a two-class model with class-dependent rates: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 50$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 5000 independent runs.

### 3.2.3 Bigger staffing intervals

We extend our discussion in §4.1.2 of the main paper to further study the impact of inflexible staffing functions. We do so by increasing $\Delta_s$. In Figure 4, we compare the TPoD performance for a system staffed according to the MSL method with $\Delta_s = 0.5$ and $\Delta_s = 2$. As expected, when increasing the interval length, the staffing level becomes too rough to cope with the time-varying demand, thus unable to achieve desired performance stabilization.

### 3.2.4 Smaller delay targets

In §4.2.1 of the main paper, we have considered the case of small delay targets $w_1 = 0.1$, $w_2 = 0.2$, Figure 6 there shows that our methods continue to achieve stable performance for both classes. We now set $w_1$ and $w_2$ to values that are closer to 0. In Figure 5, we give simulations for two cases of smaller $w_i$: (a) $w_1 = 0.05$, $w_2 = 0.1$; and (b) $w_1 = 0.01$, $w_2 = 0.02$. The other parameters are the same as in Figure 6 of the main paper. According to Figure 5, we see that stable TPoD performance is achieved for case (a); however, our

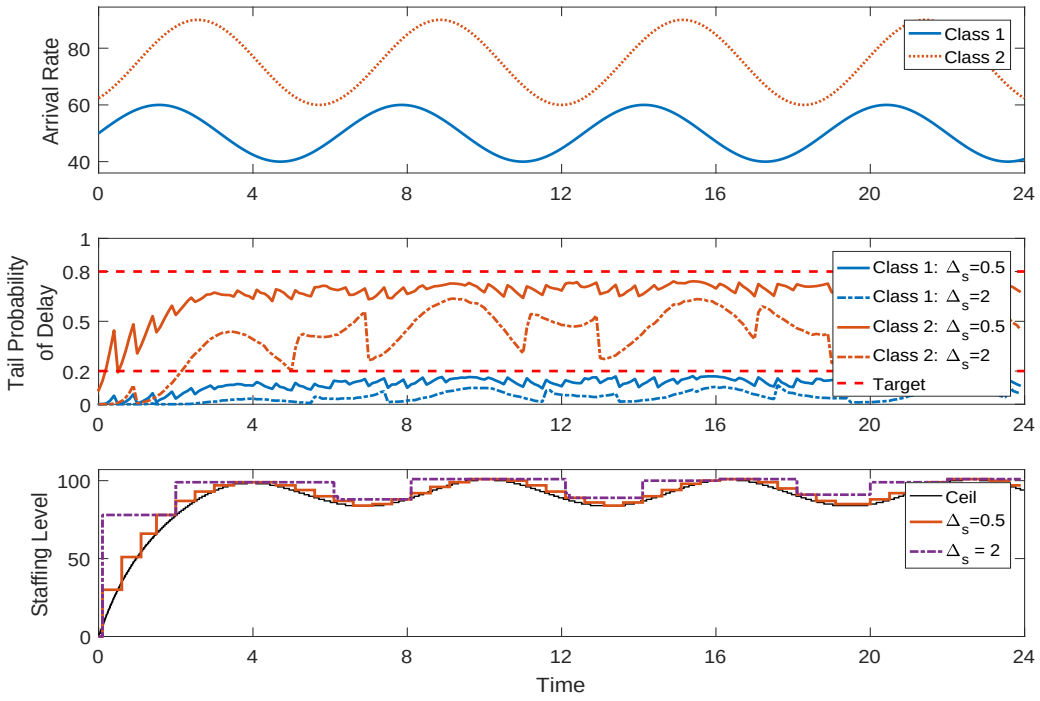


Figure 4: Simulation comparison for a two-class model with fixed-staffing intervals: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying MSL staffing $\Delta_s = 0.5, 2$ (bottom panel), with $\mu_1 = \mu_2 = 1$, $n = 50$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 5000 independent runs.

method is no longer effective for case (b). This is consistent with our intuitions. First, all heavy-traffic results on asymptotic service differentiation have been established under the assumption that $w_i > 0$, which puts the model in the efficiency-driven regime (namely, the system is asymptotically overloaded). However, for any finite $n$ (hereby $n = 50$), a small $w_i \approx 0$ will place the system in the quality-and-efficiency driven (QED) regime. Therefore, effective TPoD control with extremely small $w_i$ will require the knowledge of the corresponding FCLT limits in the QED regime. In the case $w_i = 0$ so that TPoD degenerates to PoD, our method should fail completely. After all, $w_i$ appears in the denominator of the TV-DPS control formula. We remark that the problem of achieving differentiated and stabilized PoD performance remains an open problem.

### 3.2.5 Relaxation of the assumption on class-dependent arrival times

Recall that at the beginning of §2 of the main paper, we made the following assumptions on the arrival processes.

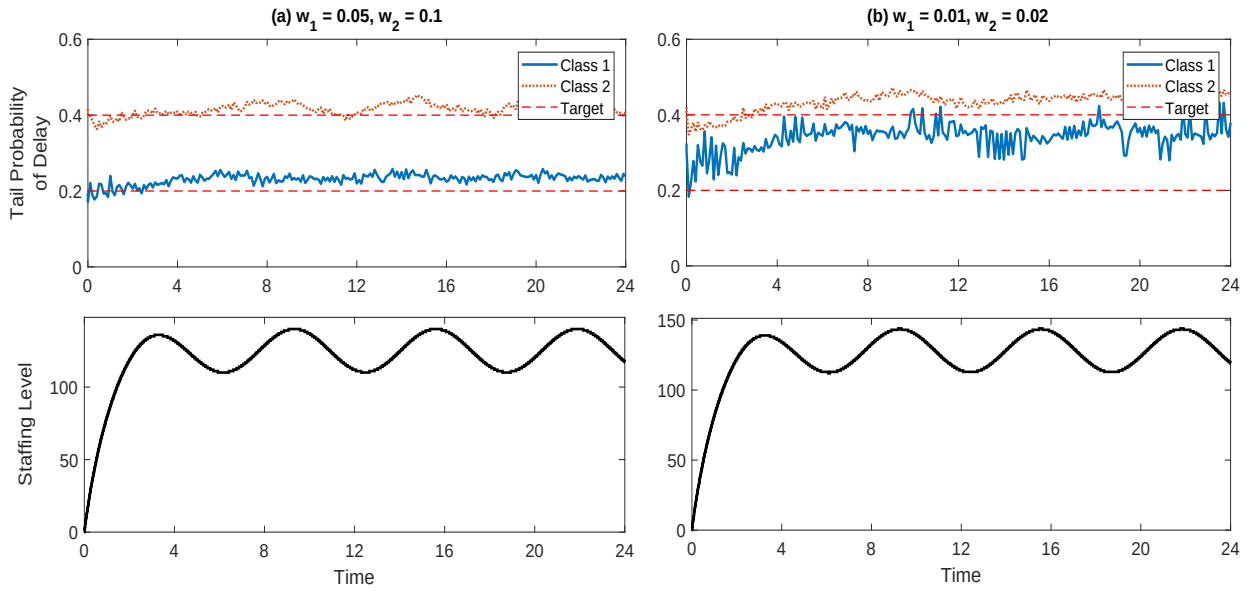


Figure 5: Simulation comparison for a two-class model with smaller delay targets: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing (bottom panel), with (a) $w_1 = 0.05$, $w_2 = 0.1$ (left) and (b) $w_1 = 0.01$, $w_2 = 0.02$, $\mu_1 = \mu_2 = 1$, $n = 50$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, and 5000 independent runs.

**Assumption 3.1 (Arrivals begin at different times)** *Each class-$i$ arrival process $A_i(t)$ begins at time $-w_i < 0$, $1 \le i \le K$.*

Such an assumption is imposed to facilitate the mathematical treatment. We now elaborate: Given class-dependent delay targets $0 < w_1 \le w_2 \le \cdots \le w_K$ (without loss of generality we order them in the increasing order), each class-$i$ arrival process begins at a different (negative) time $-w_i$, and class-$i$ arrivals begin to occur earlier than class-$j$ arrivals for $i > j$. By time 0 at which we begin to serve all customers following TV-SRS and TV-DPS, we already have enough candidate customers, and more important, each class-$i$ HoL customer is "old" enough (reaching the specific class-$i$ delay target $w_i$). This provide a clean condition for us to implement TV-DPS. (Note that we do not serve any customer until time 0 because our TV-SRS $s(t) = 0$ for $t < 0$.)

We now relax Assumption 3.1 by allowing customers of all classes to arrive at the same time $t = 0$. Without loss of generality, we can simply let all customer classes to begin their arrival processes at time $-w_K = \max_{1 \le i \le K} w_i < 0$. (It suffices to later shift time by $w_K$ units.)

**Assumption 3.2 (Arrivals begin at the same times)** *All arrival processes $\{A_i(t), 1 \le i \le K\}$ begin at the same time $-w^* \equiv -\max_{1 \le i \le K} w_i < 0$.*

Comparing Assumption 3.2 to Assumption 3.1, we now allow additional arrivals to occur before time 0 for classes $1, 2, \ldots, K-1$. Specifically, extra class-$i$ arrivals occur in the augmented interval $[-w_K, -w_i)$, $1 \le i \le K-1$. In the $n^{\text{th}}$ system, Assumption 2 add approximately $O(n(w_K - w_i))$ initial customers to the $i^{\text{th}}$ queue, $1 \le i \le K-1$.

We first give a numerical example to illustrate the effect of adding additional customers at time 0 under Assumption 3.2 with the system operating under TV-SRS and TV-DPS based on Assumption 3.1 (the formulas currently used in the main paper). In Figure 6, we reuse our two-class base example. The second plot shows that that the additional initial customers can have significant impact on the TPoD performance; they cause an initial rampdown (understaffing), see the dotted-and-dashed lines. We believe this performance is still acceptable, because stable TPoD is achieved after some initial "warmup" time.

Nevertheless, we next discuss how to adjust our staffing and scheduling formulas to eliminate the initial TPoD bumps. To treat the new (more complex) assumption 3.2, we follow the discussions in Remark 3 of the main paper. We partition the negative interval $[-w_K, 0)$ into $K$ consecutive subintervals: $\mathcal{I}_1 \equiv [-w_K, -w_K + w_1), \mathcal{I}_2 \equiv [-w_K + w_2, -w_K + w_3), \ldots, \mathcal{I}_K \equiv [-w_K + w_{K-1}, 0)$. In the interval $\mathcal{I}_1$, we do not serve any customers. In $\mathcal{I}_2$, we act as if there is one customer class, namely, class 1 (we compute the TV-SRS formula using arrival rate $\lambda_1(t)\mathbf{1}_{\{t\in\mathcal{I}_1\}}$). In $\mathcal{I}_3$, we pretend that there are only two classes, namely class 1 and 2 (we only serve the first two classes using 2-class TV-SRS and TV-DPS formulas calculated based on arrival rates $\lambda_1(t)\mathbf{1}_{\{t\in\mathcal{I}_2\}}$ and $\lambda_2(t)\mathbf{1}_{\{t\in\mathcal{I}_1\cup\mathcal{I}_2\}}$). In the interval $\mathcal{I}_i$, $3 \le i \le K$, we serve the first $i-1$ classes and compute the TV-SRS and TV-DPS formulas using arrival rates $\lambda_1(t)\mathbf{1}_{\{t\in\mathcal{I}_{i-1}\}}$, $\lambda_2(t)\mathbf{1}_{\{t\in\mathcal{I}_{i-2}\cup\mathcal{I}_{i-1}\}}$, …, $\lambda_{i-1}(t)\mathbf{1}_{\{t\in\mathcal{I}_1\cup\cdots\cup\mathcal{I}_{i-1}\}}$. At time 0 and beyond, the TV-SRS and TV-DPS rules are implemented in the usual way for all classes (using arrival rates $\lambda_1(t)\mathbf{1}_{\{-w_1\le t\le T\}}$, …, $\lambda_K(t)\mathbf{1}_{\{-w_K\le t\le T\}}$).

Using the same two-class example, we plot the simulated TPoD curves in Figure 6 under the refined policy; see the two solid lines in plot 2. This shows that our refinement on our "piecewise" TV-SRS and TV-DPS in consecutive intervals indeed achieve stable TPoD performance starting from the beginning (without needing a warmup period). See plot 3 of Figure 6 for the two staffing functions.

### 3.3 Staffing and scheduling to differentiate the mean PWT

We have shown in Theorem 2 of the main paper that TV-SRS and TV-DPS achieves asymptotic stabilization for both mean delay $\mathbb{E}[V_i(t)]$ and TPoD $\mathbb{P}(H_i(t) > w_i)$ at desired targets $w_i$ and $\alpha_i$.

We will next provide simulations to confirm the effectiveness of our methods using $\mathbb{E}[V_i(t)]$ as the performance metric. First, we note that this is a performance metric that depends only on $w_i$, not $\alpha_i$; and our fine-tuning second-order control functions (second-order safety staffing term $c$ and second-order scheduling

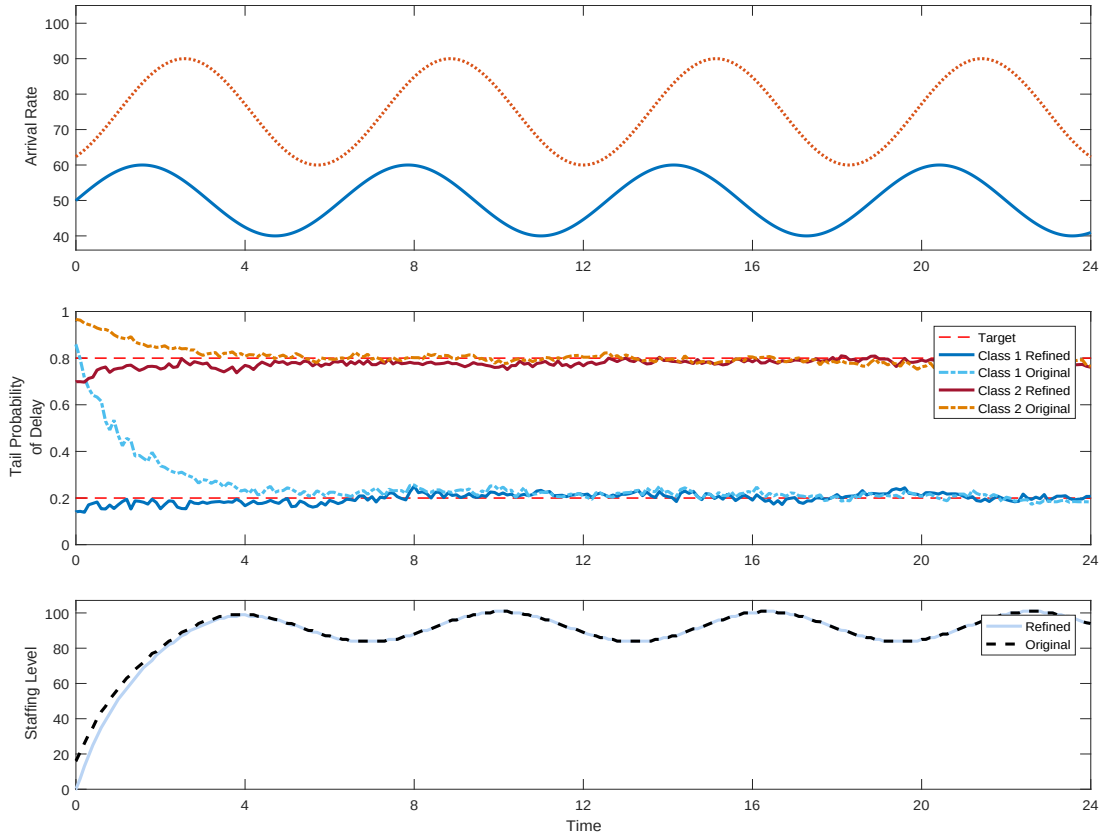


Figure 6: Simulation comparison for a two-class model with both classes begin to arrival at the same time $-w_2$: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ under both staffing functions (middle panel); and (iii) two staffing functions (bottom panel), with $w_1 = 0.05$, $w_2 = 0.1$, $\mu_1 = \mu_2 = 1$, $n = 50$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 5000 independent runs.

threshold $\kappa_i$) are developed to account for the probability target $\alpha_i$. Indeed, it is clear from (14) and (15) of the main paper that these second-order terms will play negligible roles as the scale $n$ increases, which is consistent with our asymptotic stability results in Theorem 2. For simplicity, in this section we will set the probability target $\alpha_i = 0$ for all $1 \le i \le K$, so that $c(t) = \kappa_i(t) = 0$ and TV-DPS and TV-SRS degenerate to the simpler *HWT-based dynamic prioritization* and *offered-load staffing*

$$i^* \in \arg \max_{1 \le i \le K} \left\{ \frac{H_i(t)}{w_i} \right\} \qquad \text{and} \qquad s(t) = m(t). \tag{58}$$

This simplification is intuitive: The basic idea of using the second-order terms to set the tail probability $\mathbb{P}(V_i^n(t) > w_i)$ to $\alpha_i$ is to asymptotically adjust the mean of the nearly Gaussian distributed $\widehat{V}_i^n(t)$ so that the probability mass above 0 is equal to $\alpha_i$, see (31) of the main paper. According to the symmetric structure of the limiting Gaussian distribution, we should set $\alpha_i = 0$ in order to have the mean of $V_i^n(t)$ balanced at

$w_i$.

We next give simulation results. consider sinusoidal arrival rates

$$\lambda_i(t) = \bar{\lambda}_i \left(1 + r_i \sin(\gamma_i t + \phi_i)\right), \qquad 1 \le i \le K, \tag{59}$$

with average rate $\bar{\lambda}_i$, relative amplitude $|r_i| < 1$, frequency $\gamma_i$, and phase $\phi_i$. We first consider a two-class base model, where Class 1 and Class 2 represent high and low priority customers respectively. We let $\bar{\lambda}_1(t) = 1, \bar{\lambda}_2(t) = 1.5, r_1 = 0.2, r_2 = 0.3, \gamma_1 = \gamma_2 = 1, \phi_1 = 0, \phi_2 = -1$. Abandonment times follow class-dependent exponential distributions with PDF $f_i(x) = \theta_i e^{-\theta_i x}$. We let $\theta_1 = 0.6$, $\theta_2 = 0.3$. Service rates are class-independent and standardized so that $\mu_1 = 0.5$, $\mu_2 = 1$ (with mean service time $1/\mu_i = 1$). To prioritize Class 1, we set higher QoS levels (i.e., lower target wait time). We set our target QoS parameters as $w_1 = 0.5$ and $w_2 = 1$ ($\alpha_1 = \alpha_2 = 0.5$).

Paralleling §4 in the main paper, we consider the following cases:

(i) **Base case.** The two-class model described above with sinusoidal arrival rates in (59), and scale $n = 50$. See Figure 7 and Table 9. The relative performance difference is controlled under 3%.

(ii) **High QoS.** The example in (i) with smaller delay targets $w_1 = 0.1, w_2 = 0.2$. See Figure 8.

(iii) **Smaller arrival rate.** The example in (i) having the sinusoidal arrival rates in (59) with $n = 5$. See Figure 9.

(iv) **Mixed arrivals.** The example in (i) with class-1 arrival rate reduced by an order of magnitude (i.e., $\bar{\lambda}_1 = 0.1$. See Figure 10.

(v) **Five-class example.** The five-class example in §4.2.5 of the main paper, with $\alpha_i = 0.5$, $1 \le i \le 5$. See Figure 11.

Table 2: Five Class Model: Class specific parameters and QoS target levels

| | Class parameters | | | | | | Delay targets |
|---|---|---|---|---|---|---|---|
| Class | $\bar{\lambda}$ | $r$ | $\gamma$ | $\phi$ | $\theta$ | $\mu$ | $w$ |
| 1 | 1.0 | 0.20 | 0.5 | 0 | 0.6 | 1 | 0.2 |
| 2 | 1.5 | 0.30 | 1.0 | -1 | 0.3 | 1 | 0.4 |
| 3 | 1.2 | 0.05 | 1.3 | 1 | 0.5 | 1 | 0.6 |
| 4 | 1.1 | 0.15 | 1.6 | -2 | 1.0 | 1 | 0.8 |
| 5 | 1.6 | 0.40 | 2.0 | 2 | 1.2 | 1 | 1.0 |

The Monte-Carlo simulations are conducted by generating $m$ independent runs, with $m = 5000$ when the scale $n = 50$ and $m = 20000$ when $n = 5$. From Figures 7–11, we conclude that our staffing and scheduling rules in (58) successfully achieve stabilized performance across all classes.

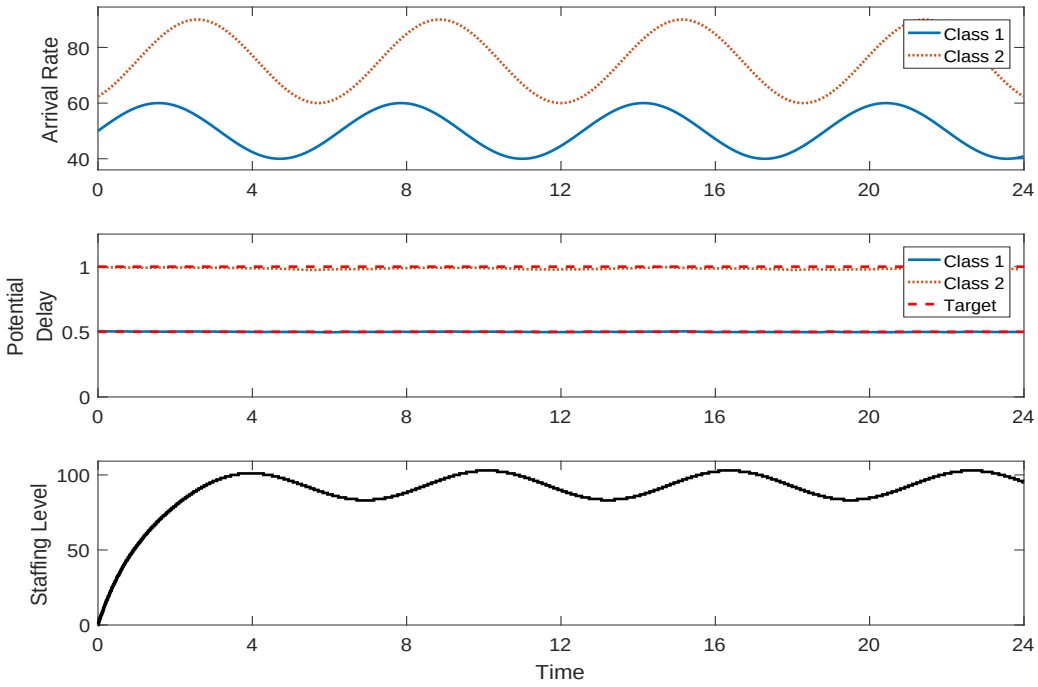


Figure 7: Simulation comparison for a two-class model: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $\mathbb{E}[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 50$, $w_1 = 0.5$, $w_2 = 1$, and 5000 independent runs.

Table 3: A two-class base case: Average, max, and min of (simulated) delay and relative differences to their target levels, using floored, rounded, and ceiled versions of the TV-SRS formula.

| | | Avg. | | | Max | | | Min | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Class | | Floor | Round | Ceiling | Floor | Round | Ceiling | Floor | Round | Ceiling |
| 1 | $w_1$ | 0.4980 | 0.4983 | 0.4991 | 0.5019 | 0.5013 | 0.5035 | 0.4940 | 0.4945 | 0.4967 |
| | % | (-0.39) | (-0.34) | (-0.18) | (+0.38) | (+0.25) | (+0.69) | (-1.20) | (-1.09) | (-0.66) |
| 2 | $w_2$ | 0.9823 | 0.9828 | 0.9845 | 0.9928 | 0.9902 | 0.9965 | 0.9720 | 0.9717 | 0.9755 |
| | % | (-1.77) | (-1.72) | (-1.55) | (-0.72) | (-0.98) | (-0.35) | (-2.80) | (-2.83) | (-2.45) |

## References


[1] A. Korhan Aras, Xinyun Chen, and Yunan Liu. Many-server Gaussian limits for non-Markovian queues with customer abandonment. *Queueing Systems*, 89(1):81–125, 2018.

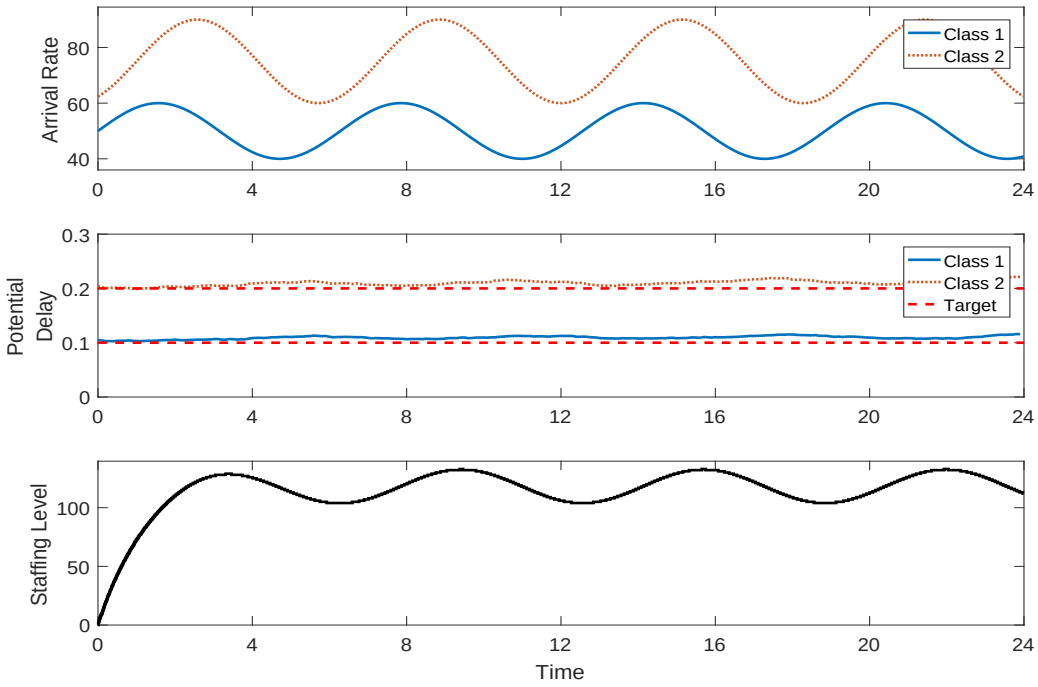


Figure 8: Simulation comparison for a two-class model with smaller delay targets: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $\mathbb{E}[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 50$, $w_1 = 0.1$, $w_2 = 0.2$, and 5000 independent runs.

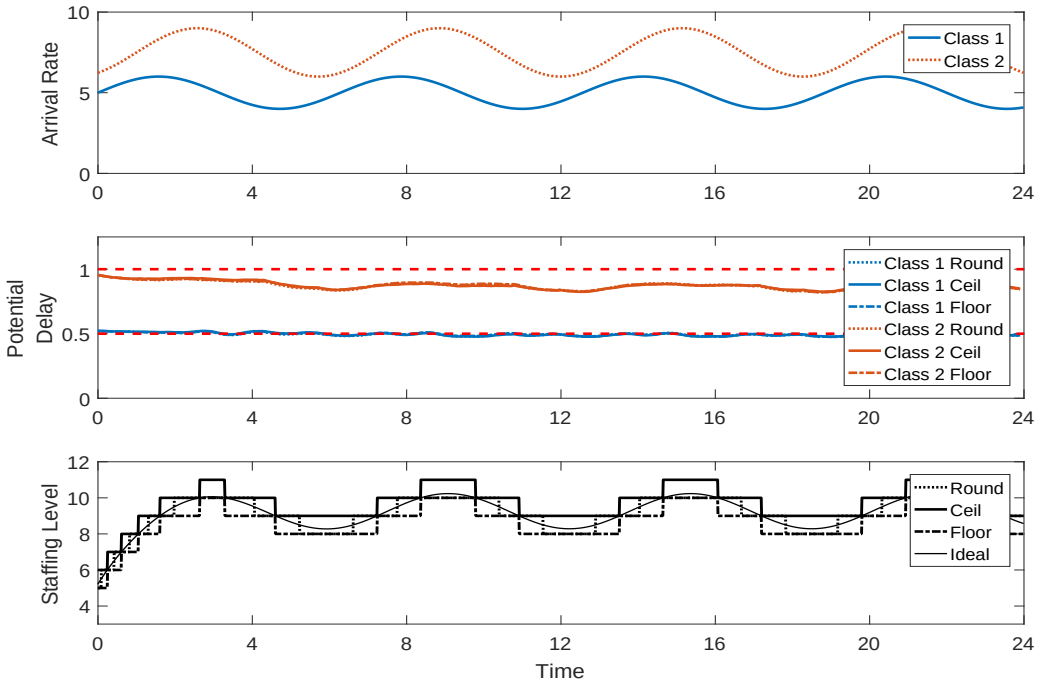


Figure 9: Simulation comparison for a two-class model with smaller arrival rates: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $\mathbb{E}[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 5$, $w_1 = 0.5$, $w_2 = 1$, and 20000 independent runs.

[2] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

[3] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 1991.

[4] Elena V Krichagina and Anatolii A Puhalskii. A heavy-traffic analysis of a closed queueing system with a $GI/\infty$ service center. *Queueing Systems*, 25(1-4):235–280, 1997.

[5] Guodong Pang, Rishi Talreja, Ward Whitt, et al. Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probability Surveys*, 4:193–267, 2007.

[6] Ward Whitt. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media, 2002.

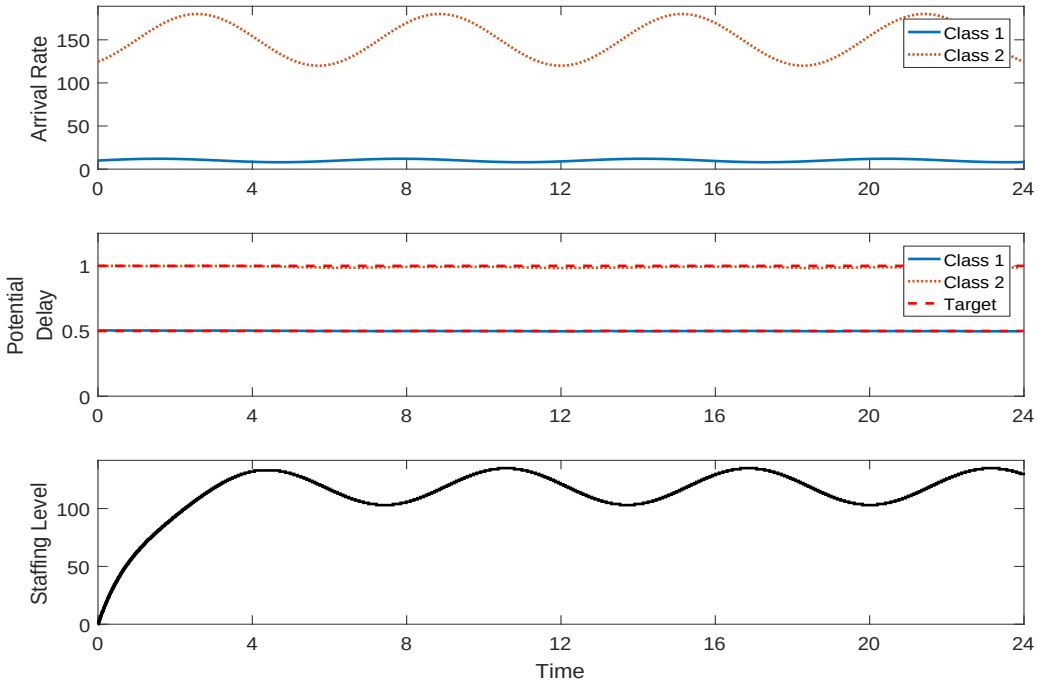


Figure 10: Simulation comparison for a two-class model with mixed magnitudes of arrivals: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $\mathbb{E}[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 100$, $\bar{\lambda}_1 = 0.1$, $w_1 = 0.1$, $w_2 = 0.2$, and 5000 independent runs.

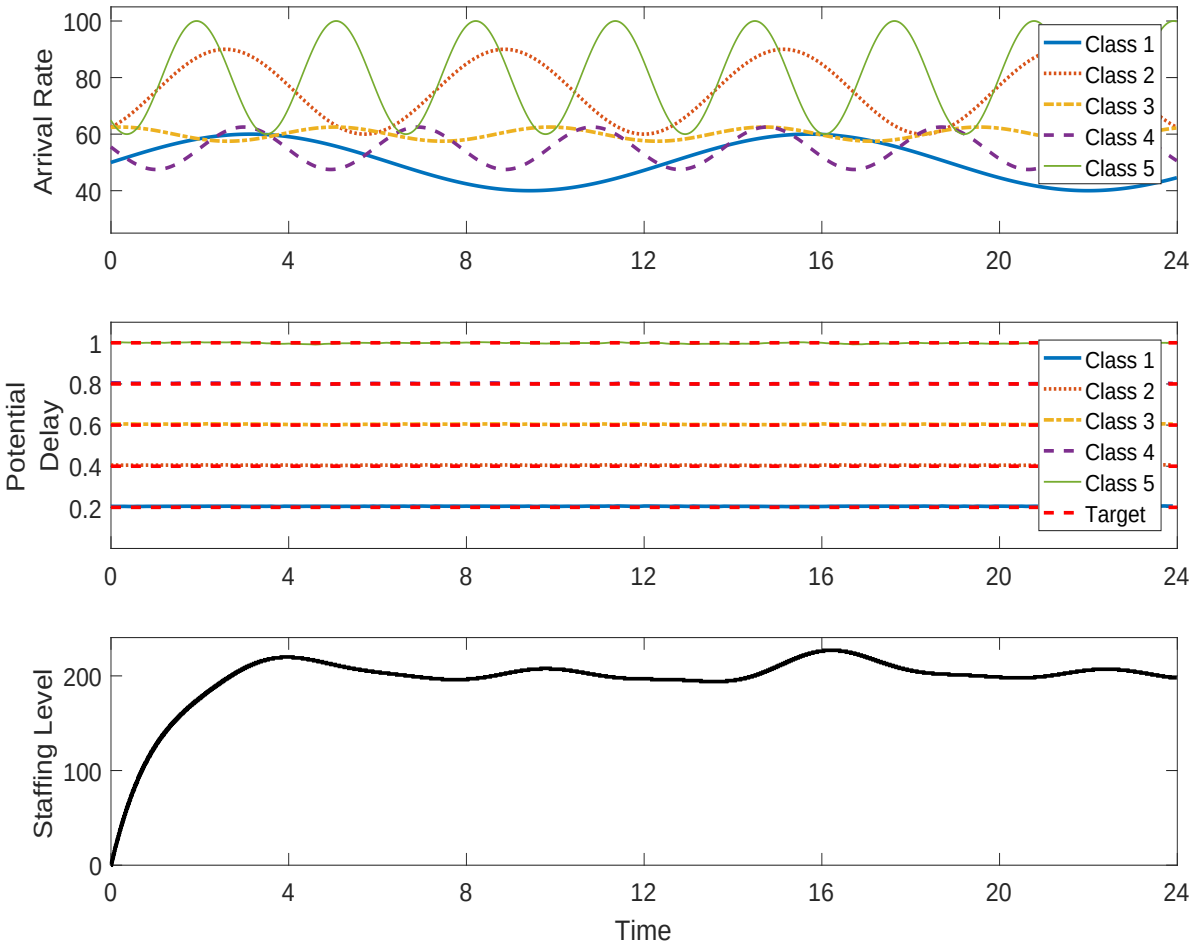


Figure 11: Simulation comparison for a fixe-class model: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $\mathbb{E}[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with 5000 independent runs, and model parameters given in Table2