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

Yunan Liu
Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695

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

Kyle Hovey
Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695

Motivated by large-scale service systems, we study a 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 new staffing and scheduling policies that are 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 rates. Effectiveness of our proposed staffing and scheduling policies 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.

Key words: service differentiation; time-varying queues; time-varying staffing; dynamic prioritization; stabilizing performance; efficiency-driven

1. Introduction

In this paper, we study a service-level differentiation problem for a many-server service system with (a finite number) $K$ customer classes each having its own dedicated queue and time-varying arrival rate. The problem of achieving differentiated service can be framed as concurrent determination of staffing (i.e., number of servers) and scheduling (i.e., pairing a newly available server with a customer when there are customers from more than one class waiting) rule to satisfy a set of prescribed performance targets. In the
present study, we are especially interested in satisfying the following service-level constraints:

\[ \mathbb{P}(V_i(t) > w_i) \leq \alpha_i, \quad 1 \leq i \leq K, \quad 0 < t < T, \]

for class-specific delay target \( w_i \) and tail-probability target \( \alpha_i \in (0, 1) \), \( 1 \leq i \leq K \), finite time horizon \( T \) (e.g., \( T = 24 \)), where \( V_i(t) \) is the delay of a class-\( i \) customer arriving at time \( t \). In words, the set of constraints requires that a class \( i \) customer who arrives at time \( t \) waits longer than \( w_i \) time units with a probability no greater than \( \alpha_i \). We refer to the left-hand side of (1) as the tail probability of delay (TPoD). Such TPoD-based quality-of-service (QoS) metrics have been widely used in service systems, such as the 80/20 rule in call centers (Aksin et al. 2007, Gans et al. 2003), the 6-hour service level in Singapore hospitals (Shi et al. 2016).

Ideally, we would like to use the minimum possible staffing to meet those targets, in which case one expects that all the constraints in (1) are binding or nearly binding. Note that the minimum staffing level depends critically on the space of scheduling policies. Here, instead of solving an optimal staffing problem subject to constraints, we seek simple and effective scheduling rules that can achieve performance stabilization in a finite time period across all customer classes. Loosely speaking, we look for a staffing function and a scheduling policy under which

\[ \mathbb{P}(V_i(t) > w_i) \approx \alpha_i, \quad 1 \leq i \leq K, \quad 0 < t < T. \]

From now on we refer to the above problem as the service differentiation and performance stabilization problem.

Motivations for the present study largely arise from human-operated service systems where the system operator needs to determine how to economically plan and fairly allocate scarce service resources (e.g., number of servers) to meet the diverse needs of its customers. One notable example is the Canadian triage and acuity scale (CTAS) guideline that classifies patients in the emergency department (ED) into five acuity levels, where each acuity level is associated with a prescribed performance target, consisting of a threshold time and the proportion of patients whose waiting time should not exceed that threshold. According to
the CTAS guideline Ding et al. (2018), Bullard et al. (2014), “CTAS level $i$ patients need to be seen by a physician within $w_i$ minutes $100\alpha_i\%$ of the time”, with

$$(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120) \quad \text{and} \quad (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8).$$

In this setting, healthcare personnel represents the service resource which can be effectively staffed and scheduled to meet the CTAS targets. Similar multi-level triage policies have been widely adopted in many other EDs (not limited to those in Canada), see Fernandes et al. (2005).

Other common examples of service differentiation include large customer contact centers where callers are often segmented into different classes each having a specified service deadline. Service differentiation is also important in today’s multi-media (or omni-channel) contact centers where one looks at the service level not just for the voice transactions alone, but for emails or webchat interactions. Each of these channels requires that we define what our service level is. There may have been 80% of the voice calls answered within 20 seconds, but in email that may equate to 80% of the emails responded within four hours, or 80% of the chat requests answered within 90 seconds; see Taylor (2011). In addition to customer contact centers, our modeling framework and proposed solutions may be applied to other service systems that share similar features. Examples include immigration offices in which the employees have to select cases to expedite in the face of a large backlog of immigration/permit applications as well as amusement parks where service providers have to tradeoff serving the fastpass and regular customers; see Kostami and Ward (2009). To summarize, our framework provides a useful tool to understand how scarce service resources should be allocated in the aforementioned systems where service strategies are not driven by revenue but rather less tangible aspects such as social welfare.

The multi-class multi-server queueing system considered in this paper captures salient features of real-world service systems. First, we assume the demand function to be time varying for each class. This assumption is primarily motivated by empirical studies showing that demand arrivals in real-world service systems typically vary strongly over time; see Green et al. (2007). Second, we incorporate customer abandonments to reflect the fact that patients waiting in the ED may leave the system without being seen and callers may hang up due to prolonged waiting times. Third, we allow service times to have class-dependent service
rates; this makes our model especially useful in practical settings. One example is the CTAS example where treatment times are evidently different for patients of various acuity levels.

1.1. Literature

Our paper relates to two different streams of research, and we will review each in turn.

First the staffing component of our proposed solution is related to works on development of time-varying staffing functions to stabilize performance of relevant queueing systems having time-varying arrival-rate functions. The point-wise stationary approximation has been proven useful in staffing systems with shorter service times and slowly varying arrival rates; see Green et al. (2007) for a review. The modified-offered-load (MOL) approximation has been used to design staffing functions to control performance functions including the probability of delay (PoD), mean waiting time, and probability of abandonment (PoA). A key step of MOL is to staff according to the offered-load function of the corresponding infinite-server queue (which estimates the total service resource needed if there were no constraint on the capacity), see He et al. (2016), Jennings et al. (1996), Liu and Whitt (2012), Yom-Tov and A. (2014), Whitt (2013). Feldman et al. (2008) developed a simulation-based iterative staffing algorithm (ISA) to stabilize performance the PoD; the idea of ISA has been extended by Defraeye and van Nieuwenhuyse (2013) to treat TPoD. Recently, Liu (2018) developed an analytic staffing function to stabilize the TPoD and proved the a corresponding asymptotic stability result. Staffing functions based on Gaussian variance approximations are studied in Pender and Massey (2017). To the best of our knowledge, prior to our work there exists no result on joint staffing and scheduling decisions in overloaded time-varying queues.

Second, the scheduling component of our proposed solution to the service differentiation problem relates to a vast body of research on optimal scheduling control for queueing systems. Using the conventional heavy-traffic (HT) scaling, Van Mieghem (1995) showed the celebrated $c\mu$ rule to be asymptotically optimal; see also Mandelbaum and Stolyar (2004). Similar approaches were adopted by Atar et al. (2004), Harrison and Zeevi (2004), Atar (2005) for critically loaded systems and by Atar et al. (2010) for overloaded systems in the many-server setting. More recently, Kim et al. (2018) incorporated the customer patience-time distribution into an optimal scheduling problem. Using heavy-traffic analysis, they proposed
a near-optimal scheduling policies that can be implemented by customer contact centers to further improve performance metrics. In this paper, we too use generally distributed patience time and devise control solutions that account for temporal changes in customer patience. Finally, we point out the empirical work of Ding et al. (2018) that used patient-level data to analyze patient routing behaviors; their empirical findings suggest that the Canadian EDs apply a delay-dependent prioritization across different triage levels. The formulation of our problem is mostly related to the constraint-satisfaction approach as adopted by Gurvich et al. (2008) and Gurvich and Whitt (2010); see also Soh and Gurvich (2016). By focusing on ratio scheduling and routing policies, Gurvich and Whitt (2010) sought “good and simple” policies and established the state-space-collapse (SSC) associated with the HT limit showing that the ratio rules are asymptotically optimal. More recent, Sun and Whitt (2018) applied ratio rules in a time-varying environment to achieve service differentiation in a critically-loaded system. It is worth noting that all of these papers assume a critical loading system (hence the delay becomes negligible as the system size increases). These results may not be applicable to systems that are operating in the overloaded regime so that customer waiting times are comparable to their service times, thus not negligible (e.g., healthcare systems).

In this paper, we focus on the treatment of overloaded systems where the queue length and customer waiting times are no longer negligible compared to service times (In this sense, overloaded systems can be more difficult to analyze than critically loaded systems). For this reason, we do not scale down the delay targets, as was done in Gurvich and Whitt (2010), Soh and Gurvich (2016), Sun and Whitt (2018). Moreover, motivated by CTAS-type service levels, we allow the probability targets $\alpha_i$ to be different across classes as opposed to identical targets as used by Gurvich and Whitt (2010), Sun and Whitt (2018). Finally, we consider class-dependent service rates $\mu_i$ rather than a common service rate $\mu$ for all classes as was imposed by Gurvich et al. (2008), Gurvich and Whitt (2010), Soh and Gurvich (2016), Sun and Whitt (2018).

1.2. Contributions

First, we propose a new time-varying staffing rule and time-varying dynamic scheduling policy for the multi-class model, having time-varying arrivals, customer abandonment, and class-dependent service, abandonment, and arrival rates. Our dynamic scheduling rule prioritizes customers based on their elapsed delays
with a time-varying class-dependent prioritization regulator. Second, under our new staffing and scheduling policies, we establish many-server heavy-traffic (MSHT) functional central limit theorem (FCLT) for various quantities of interest. By proving an SSC result for the waiting time processes, we show that the multi-class delay functions reduce to a simple one-dimensional process (called the frontier process). Because we allow service rates to be class dependent, our frontier process uniquely solves a stochastic Volterra equation, which is in sharp contrast with the existing literature wherein Ornstein-Uhlenbeck (or piecewise linear diffusion) processes often arise as the scaling limit. Third, based on the SSC and further analysis of this frontier process, we obtain the desired analytic control functions for our time-varying staffing and scheduling policies; The computation of these control functions replies on the first and second moment of the limiting frontier process for which we develop efficient algorithms. We prove that they asymptotically achieve TPoD-based service-level differentiation for all classes at customized service targets. See Figure 1 for an illustration of our main steps. Last, we consider important special cases to gain useful insights of our staffing and scheduling policies. We also conduct extensive simulation experiments to substantiate the effectiveness and robustness of our results.

Organization of the paper. In §2, we describe the multi-class time-varying queueing model and introduce our staffing and scheduling policies. In §3, we present the limit theorems which establish asymptotic service differentiation and stabilization. In §4 we report numerical examples. Finally, we make concluding remarks in §5. All the technical proofs are given in the e-companion. We provide additional results in a longer online appendix Liu et al. (2018b).

2. Problem Formulation and Proposed Solutions

We describe the time-varying multi-class queueing model in §2.1. We introduce our time-varying staffing and dynamic scheduling rules in §2.2 and §2.3.
2.1. A Multiclass V model

Consider a V-model having $K \geq 2$ customer queues served by one common service pool. Customers arrive to the $i^{th}$ queue according to a non-homogeneous Poisson process (NHPP) $A_i$ with rate function $\lambda_i(\cdot)$. In what follows, we will be using $\Lambda_i(\cdot)$ to denote the corresponding cumulative arrival function, i.e., $\Lambda_i(t) \equiv \int_{t_0}^{t} \lambda_i(u)du$, if the process starts at time $t_0$. For mathematical convenience, we assume that the class-$i$ arrival process $A_i$ starts at time $-w_i$. This assumption facilitates the mathematical treatment, because the proposed scheduling policy (to be specified later) can be simply implemented at time 0. We discuss how this assumption can be relaxed in Remark 3.

We assume class-$i$ service times are independent and identically distributed (i.i.d.) random variables following an exponential distribution with class-dependent service rate $\mu_i$. Class-$i$ customers may choose to abandon from the $i^{th}$ queue according to i.i.d. abandonment times following a general distribution, with cumulative distribution function (CDF) $F_i(x)$, complementary CDF (CCDF) $F_i^c(x) \equiv 1 - F_i(x)$, probability density function (PDF) $f_i(x)$, and hazard rate $h_i(x) \equiv f_i(x)/F_i^c(x)$. We assume that service times and patience times are mutually independent, independent of the arrival processes. Throughout this paper, we will assume $\lambda_i(\cdot)$ to be bounded away from zero and infinity, having piecewise bounded first-order derivative. In addition, we assume the PDF $f_i(x) > 0$ for $x \geq 0$ so that the CCDF $F_i^c(x) > 0$ on any compact interval.
The system adopts a work-conserving policy, i.e., no customers wait in queue if there is an available server. Figure 2 gives a graphical illustration of a three-class system. Let \( Q_i(t) \) represent the number of customers waiting in the \( i \)th queue. We use \( E_i(t) \) and \( R_i(t) \) to denote the number of customers that have entered service and that have abandoned from the \( i \)th queue, respectively, up to time \( t \). By flow conservation,

\[
Q_i(t) = Q_i(0) + A_i(t) - E_i(t) - R_i(t).
\]  

Let \( B_i(t) \) be the number of busy servers currently serving class-\( i \) customers at time \( t \) and \( D_i(t) \) be the cumulative number of class-\( i \) customers that have departed due to service completion up to time \( t \). Again by flow conservation, we get

\[
B_i(t) = B_i(0) + E_i(t) - D_i(t).
\]  

Finally, let \( X_i(t) \) denote the total number of class-\( i \) customers in the system at time \( t \). Adding up (3) and (4) yields

\[
X_i(t) = Q_i(t) + B_i(t) = X_i(0) + A_i(t) - D_i(t) - R_i(t).
\]  

**Two waiting times.** We now introduce two types of waiting-time processes that we will exploit heavily in the subsequent analysis. Let \( H_i(t) \) denote the head-of-line waiting time (HWT) of the \( i \)th queue, i.e., the waiting time of the class-\( i \) customer who has been waiting the longest (if there is any); \( H_i(t) = 0 \) if there is no customer waiting in the \( i \)th queue. Let \( V_i(t) \) represent the class-\( i \) potential waiting time (PWT) at time \( t \), i.e., the waiting time of a potential class-\( i \) customer arriving at time \( t \) who has infinite patience. Based on these two waiting times, we can conveniently express the enter-service process and the queue-length process for each customer class in the following way:

\[
E_i(t) = \sum_{k=1}^{\text{A}_i(t-H_i(t))} 1_{\{\gamma_{i,k} > V_i(\xi_{i,k})\}},
\]

\[
Q_i(t) = \sum_{k=\text{A}_i(t-H_i(t))}^{\text{A}_i(t)} 1_{\{\xi_{i,k} + \gamma_{i,k} > t\}},
\]

where \( 1_A \) denotes the indicator function of event (set) \( A \), the random variables \(-w_i \leq \xi_{i,1} < \xi_{i,2} < \cdots\) denote the successive arrival times of class-\( i \) customers, and \( \gamma_{i,1}, \gamma_{i,2}, \ldots \) denote the i. i. d. patience times with CDF \( F_i \). As will become clear in the subsequent analysis, these representations are useful in deriving
the limiting FCLT results. To complete the model, it remains to specify (i) the staffing level for the service pool (which plans the overall service capacity for all customer classes), and (ii) the scheduling policy used to pair a newly available server with a waiting customer from one of $K$ classes (which determines how to dynamically allocate the overall service capacity to serve each customer class).

### 2.2. A Time-Varying Square-Root Staffing Rule

Following Step 1 in Figure 1, we now introduce a time-varying square-root staffing (TV-SRS) rule, which consists of two terms: (i) the nominal staffing level (first-order term) and (ii) the safety staffing level (second-order term).

**First-order nominal staffing term.** To set the nominal staffing level, we adopt the offered load analysis which estimates the required service capacity by estimating how much capacity would be used if there were no limit on its availability. For example, consider a single-class $M_t/GI/s_t + GI$ model having Poisson arrivals rate $\lambda(t)$, independent and i.i.d. service times with a general distribution $G$ (the first $GI$), and i.i.d. customer abandonment following a general distribution $F$ (the $+GI$). Although the $M_t/GI/s_t + GI$ model is complicated, the corresponding $M_t/GI/\infty$ infinite-server model remains remarkably tractable, where the number of customers (or busy servers) follows a Poisson distribution with mean

$$m_\infty(t) \equiv \mathbb{E}[X_\infty(t)] = \int_0^t \lambda(u) G^c(t - u) \, du.$$  \hspace{1cm} (8)

If the objective is to stabilize the expected delay at any given point in time at a target $w$, one will need to set the staffing levels to a modified version of (8), namely,

$$m_{DIS}(t) \equiv \int_0^t \underbrace{F^c(w) \lambda(u - w)}_{\text{effective arrival rate}} \underbrace{G^c(t - u) \, du},$$  \hspace{1cm} (9)

where we have used DIS to denote the “delayed-infinite-server approximation”, as in Liu and Whitt (2012). The effective arrival rate can be justified by the fact that, if every arrival who does not elect to abandon waits $w$ time units, then a fraction $F(w)$ of arrivals will abandon the queue before entering service. In other words, one can think of $m_{DIS}(t)$ as the mean number of busy servers needed to serve all customers who are willing to wait for $w$ time units.
For our multiclass V model with class-dependent delay \( w_i \), we follow the above offered-load analysis by setting the nominal staffing level as

\[
m(t) \equiv \sum_{i=1}^{K} m_i(t), \quad \text{where} \quad m_i(t) \equiv \int_{0}^{t} F_c^i(w_i) \lambda_i(u - w_i) e^{-\mu_i(t-u)} du,
\]

where each term in the sum of (10) is obtained by replacing \((F, w, G, \lambda)\) in (9) with the class-dependent primitives \((F_i, w_i, \exp(\mu_i), \lambda_i)\).

**Second-order safety staffing term.** Unfortunately, \( m(t) \) is not effective for stabilizing class-dependent TPoDs, because \( m(t) \) does not include the class-dependent probability targets \( \alpha_i \). Our strategy is to refine the staffing level by adding a second-order safety staffing term that is a function driven by the class-dependent probability targets \( \alpha_i \). Let \( \lambda(t) \equiv \sum_{i=1}^{K} \lambda_i(t) \) be the aggregate demand function, and \( \bar{\lambda} \equiv T^{-1} \int_{0}^{T} \lambda(t) dt \) be the average arrival rate over \([0, T]\). We envision a staffing function consisting of two pieces, namely,

\[
s(t) = \left[ m(t) + \sqrt{\bar{\lambda} c(t)} \right],
\]

where \([x]\) is smallest integer that is greater than or equal to \( x \), and \( c(t) \equiv c(t, \alpha_1, \ldots, \alpha_K) \) is a time-varying and \((\alpha_1, \ldots, \alpha_K)\)-dependent piecewise continuous control function, which will be determined later. We refer to such a staffing formula (11) as TV-SRS.

**Remark 1 (Role of the Safety Staffing Functions \( c \)).** Note that the first-order nominal term \( m(t) \) in (11) lives on the order of \( \bar{\lambda} \), while the second-term term lives on the order \( \sqrt{\bar{\lambda}} \). Given that the offered load \( m(t) \) depends on delay target \( w_i \), arrival rate \( \lambda_i(t) \), service rate \( \mu_i \) and patience-time distribution \( F_i \), the remaining flexibility in the staffing formula depends entirely on the single control function \( c \), which will be determined to satisfy the performance targets, as specified by (1). Hence, the overall staffing level \( s \) depends on probability targets \((\alpha_1, \ldots, \alpha_K)\) only through \( c \).

With time-varying staffing \( s(t) \), we need to specify how we manage the system when all servers are busy and the staffing is scheduled to decrease. What we do is to allow server switching: When that server is scheduled to depart, we do not require that the customer in service stay in service with the same server until service is complete. Instead, we allow the service in progress to be handed off to another available server.
Moreover, we do not force a customer out of service if the staffing is scheduled to decrease when all are busy. Instead, we release the first server that becomes free after the time of scheduled staffing decrease. With this assumption, the actual number of busy servers is a random process. Henceforth, we use $s_d(t)$ to denote the number of busy servers that are due to depart. Let $D(t) \equiv \sum_{i \in I} D_i(t)$ and $Z(t) \equiv s(0) - s(t) - D(t)$. One can easily argue

$$s_d(t) \leq \psi(Z)(t) \equiv Z(t) - \inf_{0 \leq u \leq t} Z(u),$$

where we have defined $\psi$ to be the one-dimensional reflection mapping; e.g., see §13.5 in Whitt (2002).

### 2.3. A Time-Varying Dynamic Prioritization Scheduling Rule

Following Step 1 in Figure 1, we next introduce a delay-based dynamic scheduling rule which is both time dependent and state dependent. To implement such a scheduling policy, we track the elapsed waiting time of all waiting customers. Because customers are served under FCFS within each class, it suffices to track the HWTs, namely, $(H_1(t), \ldots, H_K(t))$.

We route the next class-$i^*$ head-of-line (HoL) customer (if any) into service, with $i^*$ satisfying

$$i^* \in \arg \max_{1 \leq i \leq K} \left\{ \frac{H_i(t)}{w_i} + \frac{1}{\sqrt{\lambda}} \kappa_i(t) \right\},$$

where the first term $H_i(t)/w_i$ is the HWT scaled by the delay target, and $\kappa_i(t) \equiv \kappa_i(t, \alpha_i)$, referred to as the second-order class-$i$ prioritization regulator, is a time-varying and $\alpha_i$-dependent piecewise continuous control function to be specified later. We refer to such a scheduling rule as the time-varying dynamic prioritization scheduling (TV-DPS) policy. Furthermore, we define what we call the frontier process as

$$H(t) \equiv \frac{H_{i^*}(t)}{w_{i^*}} + \frac{1}{\sqrt{\lambda}} \kappa_{i^*}(t).$$

**Remark 2 (Understanding TV-DPS).** The first-order term $H_i(t)/w_i$ is designed to guarantee that the class-$i$ delay is close to its target $w_i$ (it is controlling the relative delay imbalance $(H_i(t) - w_i)/w_i$, rather than the absolute delay imbalance). The idea of exploiting the head-of-line delay information dates back to Kleinrock (1964); see also Li et al. (2017) for a non-linear extension. The second-order term $(1/\sqrt{\lambda})\kappa_i(\cdot)$ helps accomplish the class-dependent probability target $\alpha_i$. Intuitively, such a control function $\kappa_i$ should satisfy the following properties:
(i) **Monotonicity.** For fixed time $t$, $\kappa_i(t)$ should be a decreasing function of $\alpha_i$, because a bigger value of $\alpha_i$ means a lower service quality, which yields a lower prioritization level for class $i$;

(ii) **Sign.** For a class $i$ with probability target $\alpha_i > 0.5$ ($\alpha_i \leq 0.5$), the fine-tuning prioritization regulator $\kappa_i$ should satisfy $\kappa_i(t) < 0$ ($\kappa_i(t) \geq 0$) (Benchmarking with the case $\alpha_i = 0.5$, $\kappa_i$ should base on the value of $\alpha_i$ to adjust the priority levels by adding a positive or negative weight to $H_i(t)/w_i$). See numerical examples in §4 for more discussions of the structure of $\kappa_i$.

Our TV-DPS rule is both time dependent (accounting for time variability in the arrival processes) and state dependent (dynamically capturing the system’s stochasticity). To the best of our knowledge, this is a feature unique to the present study and absent from previous research. Moreover, our proposed scheduling policy is in alignment with the current practice of Canadian EDs where patients are routed not only by triage level (static) priorities but also by their actual (dynamic) wait time, as documented by Ding et al. (2018). This makes this rule especially appealing as the intrinsic fairness of the TV-DPS policy helps achieve ethical expectations set forth by the CTAS guideline. Furthermore, when $w_i = w$ and $\alpha_i = \alpha$ for all $1 \leq i \leq K$, TV-DPS degenerates to the global FCFS scheduling policy.

**Remark 3 (Relaxation of the Assumption of Different Arrival Times).** We assumed that each class-$i$ arrival process begins at a different (negative) time $-w_i$, so that by time 0 (at which we begin to serve all customers following TV-SRS and TV-DPS) we already have enough candidate customers. More important, each class-$i$ HoL customer is “old” enough (meaning they have reached the specific class-$i$ delay target $w_i$). This provide a clean condition for our mathematical treatment.

We now briefly discuss the situation where customers of each class start to arrive at time zero. Suppose there are three classes with delay targets $w_1$, $w_2$ and $w_3$, respectively. Without loss of generality, we assume $w_1 < w_2 < w_3$. Then a modified version of the TV-DPS rule proceeds as follows. Over the period $[0, w_1)$, we do not serve any customers. During $[w_1, w_2)$, we act as if there is one customer class, namely, class 1. During $[w_2, w_3)$, we pretend that there are only two classes, namely class 1 and 2 (i.e., choose to serve the first two classes only), and apply the rule (14) for $K = 2$. At time $w_3$ and beyond, the TV-DPS rule is implemented in the usual way for all classes. See discussions and examples in §3.2.5 of the online appendix.
In the next section, we will first establish an MSHT FCLT result under our TV-SRS and TV-DPS rules with unknown control parameters $c$ and $\kappa_i$ (Step 2 in Figure 1); using the FCLT limit, we will next obtain the exact formulas of $c$ and $\kappa_i$ so that the TPoD-based service-level constraints are asymptotically satisfied as the scale increases (Step 3 in Figure 1).

3. Asymptotic Service-Level Differentiation

In this section, we present our main results. §3.1 gives the asymptotic framework and states the many-server FCLT and FWLLN results for the multi-class V model operating under the TV-SRS and TV-DPS policies introduced in §§2.2–2.3. In §3.2 we utilize the FCLT results to obtain the desired control factors $\kappa_i$ and $c$ and show that they asymptotically achieve TPoD-based service-level differentiation and performance stabilization. §3.3 provides a more detailed discussion of the important case of class-independent service rate. All proofs are given in the e-companion.

3.1. Many-Server FCLT Limits under TV-SRS and TV-DPS

To establish an FCLT limit, we consider an asymptotic framework in which the system scale (here the average arrival $\bar{\lambda}$) grows to infinity. Following the convention in the literature, we will use $n$ in place of $\bar{\lambda}$ as our scaling parameter. This gives rise to a sequence of $K$-class V models indexed by $n$. Let $A^n_i(t)$ be the class-$i$ NHPP arrival process in the $n^{th}$ model, having a rate function $n\lambda_i(\cdot)$ where, by slight abuse of notation, we used $\lambda_i(t)$ to denote the baseline arrival rate at time $t$. Our TV-SRS function satisfies

$$s^n(t) = \lceil nm(t) + \sqrt{n}c(t) \rceil,$$

where $m$ and $c$ are the offered-load function in (10) and safety staffing term (yet to be determined).

Let $H^n_i$ and $V^n_i$ be the class-$i$ HWT and PWT in the $n^{th}$ model. Our TV-DPS satisfies

$$i^* \in \arg \max_{1 \leq i \leq K} \left\{ \frac{H^n_i(t)}{w_i} + \frac{1}{\sqrt{n}} \kappa_i(t) \right\},$$

where $\kappa_i$ is a control function yet to be determined.

For $1 \leq i \leq K$, let

$$\Lambda_i(t) \equiv \int_{-w_i}^{t} \lambda_i(u)du, \quad \bar{A}_i^n(t) \equiv n^{-1}A_i^n(t) \quad \text{and} \quad \tilde{A}_i^n(t) \equiv n^{-1/2} \left( A_i^n(t) - n\Lambda_i(t) \right).$$
The sequence of processes $\hat{A}_n^i$ and $\hat{A}_n^i$ satisfy a FWLLN and FCLT, namely,

$$(\hat{A}_n^i(\cdot), \hat{A}_n^i(\cdot)) \Rightarrow (\Lambda_i(\cdot), \hat{A}_i(\cdot)) \text{ in } D^2 \text{ as } n \to \infty,$$

for $\hat{A}_i(\cdot) \equiv W_{\lambda i} \circ \Lambda_i(\cdot)$, where $x \circ y(\cdot) \equiv x(y(t))$, $W_{\lambda_i}$ being a standard Brownian motion, and $D \equiv D([0, \infty), \mathbb{R})$ is the space of right-continuous $\mathbb{R}$-valued functions on $\mathbb{R}_+$ with lefthand limit, which is endowed with the Skorokhod $J_1$-topology, and $\Rightarrow$ means convergence in distribution (weak convergence).

**Remark 4 (More general $G_t$ arrivals).** Our main results below can be easily extended to more general $G_t$ arrival processes (which are not necessarily NHPPs), as long as their CLT-scaled versions satisfy the FCLT

$$\hat{A}_n^i(\cdot) \Rightarrow c_{\lambda_i} W_{\lambda_i} \circ \Lambda_i(\cdot) \text{ in } D \text{ as } n \to \infty,$$

for some $c_{\lambda_i} > 0$. These types of $G_t$ arrival processes can be used to model over-dispersed and under-dispersed arrival processes (i.e., when the variance-to-mean ratio of the number of arrivals is not close to 1), see Gerhardt and Nelson (2009), He et al. (2016), Liu et al. (2018a), Liu (2018) for construction and analysis of such $G_t$ arrival processes. In this case, our FCLT limits in Theorem 1 can be easily adjusted by simply multiplying $W_{\lambda_j}$ by the constant $c_{\lambda_i}$. For NHPPs, $c_{\lambda_i} = 1$.

Following the notations in §2.1, we use $Q^n_i(t)$ and $B^n_i(t)$ to denote the number of class-$i$ customers in queue and in service at time $t$, respectively in the $n^{th}$ V model. Their sum, denoted by $X^n_i(t)$, represents the total number of class-$i$ customers in system at time $t$. We now define their corresponding CLT-scaled versions

$$\hat{B}^n_i(t) \equiv n^{-1/2} (B^n_i(t) - nm_i(t)), \quad \hat{Q}^n_i(t) \equiv n^{-1/2} (Q^n_i(t) - nq_i(t)), \quad \hat{X}^n_i(t) \equiv n^{-1/2} (X^n_i(t) - nx_i(t))$$

where $m_i$ is given by (10), $q_i(t) \equiv \int_{t-w_i}^t F^c_i(t-u)\lambda_i(u)du$, and $x_i \equiv m_i + q_i$. In addition, let

$$\hat{H}^n_i(t) \equiv n^{1/2} (H^n_i(t) - w_i) \quad \text{and} \quad \hat{V}^n_i(t) \equiv n^{1/2} (V^n_i(t) - w_i)$$

be the CLT-scaled HWT and PWT processes, respectively. Finally, we define the CLT-scaled frontier process

$$\hat{H}^n(t) \equiv n^{1/2} (H^n(t) - 1).$$
Suppose the system operates under TV-SRS in (15) and TV-DPS in (16). Then there is a joint convergence for the CLT-scaled processes:

\[ \left( \hat{H}^n, \hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}_1^n, \ldots, \hat{H}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n, \hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n \right) \Rightarrow \left( \bar{H}, \bar{B}_1, \ldots, \bar{B}_K, \bar{H}_1, \ldots, \bar{H}_K, \bar{V}_1, \ldots, \bar{V}_K, \bar{X}_1, \ldots, \bar{X}_K, \bar{Q}_1, \ldots, \bar{Q}_K \right) \quad \text{in} \quad \mathcal{D}^{5K+1}, \]  

(20)
as \( n \to \infty \), where the FCLT limits on the right-hand side are well-defined stochastic processes.

(i) The limiting processes \( \left( \hat{H}, \hat{B}_1, \ldots, \hat{B}_K \right) \) jointly satisfy the following set of \( K \) Ornstein-Uhlenbeck (OU) type stochastic integral equations

\[
\hat{B}_i(t) + \eta_i(t)\hat{H}(t) = -\int_0^t \mu_i \hat{B}_i(u)du - \int_0^t \psi_i(u)\hat{H}(u)du + \int_0^t \psi_i(u)\kappa_i(u)du
+ \eta_i(t)\kappa_i(t) + G_i(t) \quad \text{for} \quad i = 1, \ldots, K, \quad \text{and} \quad \sum_{i=1}^K \hat{B}_i(t) = c(t),
\]

(21)
where \( \eta_i(t) \equiv w_i \lambda_i (t - w_i) F_i^c(w_i) \), \( \psi_i(t) \equiv w_i \lambda_i (t - w_i) f_i(w_i) \),

\[
G_i(t) \equiv \hat{E}_{i,1}(t) + \hat{E}_{i,2}(t) - \hat{D}_i(t), \quad \hat{E}_{i,1}(t) \equiv F_i^c(w_i) \int_{-w_i}^{1-w_i} \sqrt{\lambda_i(u)} dW_{\lambda_i}(u),
\]

\[
\hat{E}_{i,2}(t) \equiv \sqrt{F_i^c(w_i) F_i(w_i)} \int_{-w_i}^{1-w_i} \sqrt{\lambda_i(u)} dW_{\bar{\theta}_i}(u), \quad \hat{D}_i(t) \equiv \int_0^t \sqrt{\mu_i m_i(u)} dW_{\mu_i}(u),
\]

(22)
and \( W_{\lambda_i}, W_{\bar{\theta}_i}, W_{\mu_i} \) are independent standard Brownian motions.

(ii) The FCLT limits for all HWT and PWT processes are deterministic functionals of a one-dimensional process \( \hat{H} \), namely,

\[
\hat{H}_i(t) = w_i (\hat{H}(t) - \kappa_i(t)), \quad \text{and} \quad \hat{V}_i(t) = w_i (\hat{H}(t + w_i) - \kappa_i(t + w_i));
\]

(23)

(iii) The FCLT limit for each queue-length process is the sum of three terms, namely, \( \hat{Q}_i(t) \equiv \hat{Q}_{i,1}(t) + \hat{Q}_{i,2}(t) + \hat{Q}_{i,3}(t) \), where

\[
\hat{Q}_{i,1}(t) \equiv \int_{-w_i}^t F_i^c(t - u) \sqrt{\lambda_i(u)} dW_{\lambda_i}(u),
\]

\[
\hat{Q}_{i,2}(t) \equiv \int_{-w_i}^t \sqrt{F_i^c(t - u) F_i(t - u) \lambda_i(u)} dW_{\bar{\theta}_i}(s),
\]

\[
\hat{Q}_{i,3}(t) \equiv \lambda_i (t - w_i) F_i^c(w_i) \hat{H}_i(t).
\]

(iv) Finally, the FCLT limits for number in system is given by \( \hat{X}_i(t) = \hat{B}_i(t) + \hat{Q}_i(t) \).
**Remark 5 (SSC and Separation of Variability).** Theorem 1 provides the FCLT limits for waiting times and queue lengths under TV-SRS and TV-DPS with the second-order terms \( c \) and \( \kappa_i \) yet to be determined. Such FCLT results will be used later to achieve asymptotic performance differentiation and stabilization. Part (ii) of Theorem 1 gives a nice SSC result: The diffusion limits \((\widehat{H}, \widehat{B}_1, \ldots, \widehat{B}_K)\) satisfy the \((K + 1)\)-dimensional stochastic differential equation (SDE), and according to (23), both limiting HWT and PWT processes are deterministic functionals of the one-dimensional limiting frontier process \( \widehat{H} \). The intuition behind the SSC is that all these normalized HWTs (plus the second-order prioritization regulator) in (14) do not differ much from each other under the TV-DPS policy. In fact, the difference between any two classes is of order \( O(1/n) \). In addition, there are \( 3K \) independent Brownian motions \( W_{\lambda_i}, W_{\theta_i}, W_{\mu_i} \), stemming from the independent random sources (arrival, abandonment and service) of all \( K \) customer classes. We will see later in Proposition 1, these sources of randomness jointly contribute to the variability of the one-dimensional process \( \widehat{H} \).

We next provide an FWLLN result for the V model operating under the TV-SRS and TV-DPS rule. For that purpose, we define the LLN-scaled processes as follows

\[
\bar{B}_i^n(t) \equiv n^{-1} B_i^n(t), \quad \bar{Q}_i^n(t) \equiv n^{-1} Q_i^n(t), \quad \bar{X}_i^n(t) \equiv n^{-1} X_i^n(t) \quad \text{for} \quad 1 \leq i \leq K. \tag{24}
\]

The next result is a direct consequence of Theorem 1.

**Corollary 1 (FWLLN).** Suppose that the system operates under TV-SRS in (15) and TV-DPS in (16).

Then we have the joint convergence for the LLN-scaled processes

\[
(\bar{B}_1^n, \ldots, \bar{B}_K^n, \bar{Q}_1^n, \ldots, \bar{Q}_K^n, \bar{X}_1^n, \ldots, \bar{X}_K^n, H_1^n, \ldots, H_K^n, V_1^n, \ldots, V_K^n) \Rightarrow (m_1, \ldots, m_K, q_1, \ldots, q_K, x_1, \ldots, x_K, w_1 \varepsilon, \ldots, w_K \varepsilon, w_1 \varepsilon, \ldots, w_K \varepsilon) \quad \text{in} \quad \mathcal{D}^{5K} \tag{25}
\]

as \( n \to \infty \), where the function \( \varepsilon(t) = 1 \).

Below we provide a proof sketch of the theorem. The details are given in §EC.1.
Proof sketch of Theorem 1. Step 1: We first show that each component within the curly bracket in (16) 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 $\tilde{H}_i^n(t) = w_i(\tilde{H}_i^n(t) - \kappa_i(t)) + O(1/\sqrt{n})$). 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) to obtain a simple relation between $\tilde{H}_i^n$ and $\tilde{B}_i^n$. Based on the fact that the difference between $\tilde{H}_i^n(t)$ and $w_i(\tilde{H}_i^n(t) - \kappa_i(t))$ 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 $(\tilde{B}_1^n, ..., \tilde{B}_K^n, \tilde{H}^n)$. This allows us to apply the Gronwall’s inequality to establish the stochastic boundedness of the sequence $\{(\tilde{B}_1^n, ..., \tilde{B}_K^n, \tilde{H}^n); n \in \mathbb{N}\}$, 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 (22) for the corresponding CLT-scaled processes. Applying the continuous mapping theorem again with these Brownian limits yields the joint convergence of $\{(\tilde{B}_1^n, ..., \tilde{B}_K^n, \tilde{H}^n)\}$. 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.

We next take a closer look at the dynamics of the limit frontier process $\tilde{H}$. Define

$$\eta(t) \equiv \sum_{i=1}^{K} \eta_i(t) = \sum_{i=1}^{K} w_i \lambda_i(t - w_i) F_i^c(w_i).$$  \hspace{1cm} (26)

Asymptotically, a customer enters service at $t$ only when he arrived at $t - w_i$, and that the fraction of customers who do not abandonment during $w_i$ is $F_i^c(w_i)$. Hence, according to Little’s law, $\eta(t)$ can be interpreted as the time-varying number of customers (of all types) waiting to be processed at $t$, excluding those who will later abandon.

Note that each equation in (21) allows us to write $\tilde{B}_i$ as a function of $\tilde{H}$. Plugging them into the equation $\sum_{i=1}^{K} \tilde{B}_i(t) = c(t)$ plus some algebraic simplifications yields the result below.
Proposition 1 (Distribution of the frontier process \( \hat{H} \)). The process \( \hat{H} \) uniquely solves the following stochastic Volterra equation (SVE)

\[
\hat{H}(t) = \int_0^t L(t, s) \hat{H}(s) \, ds + \int_0^t J(t, s) \, dW(s) + K(t),
\]

where

\[
L(t, s) = \sum_{i=1}^K \frac{\eta_i(s)e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))}{\eta(t)}
\]

\[
J(t, s) = \frac{\sqrt{\sum_{i=1}^K e^{2\mu_i(s-t)}(F_i^c(w_i)\lambda_i(s-w_i) + \mu_i m_i(s))}}{\eta(t)}
\]

\[
K(t) = \sum_{i=1}^K \left( \eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))\kappa_i(s) \, ds \right) - c(t)
\]

\( W \) is a standard Brownian motion. In addition, \( \hat{H} \) is a Gaussian process with

(i) mean \( M_{\hat{H}}(t) \equiv \mathbb{E}[\hat{H}(t)], \, 0 \leq t \leq T \), uniquely solving the fixed-point equation (FPE)

\[
M_{\hat{H}} = \Gamma(M_{\hat{H}}), \quad \text{where} \quad \Gamma(M_{\hat{H}})(t) = \int_0^t L(t, s) M_{\hat{H}}(s) \, ds + K(t),
\]

(ii) covariance \( C_{\hat{H}}(t, s) \equiv \text{Cov}(\hat{H}(t), \hat{H}(s)), \, 0 \leq s, t \leq T \), uniquely solving the FPE

\[
C_{\hat{H}} = \Theta(C_{\hat{H}}),
\]

where the operator \( \Theta \) is defined as

\[
\Theta(C_{\hat{H}})(t, s) = -\int_0^t \int_0^s L(t, u) L(s, v) C_{\hat{H}}(u, v) \, dv \, du + \int_0^t L(t, u) C_{\hat{H}}(u, s) \, du + \int_s^t L(s, v) C_{\hat{H}}(t, v) \, dv + \int_0^s J(t, u) J(s, u) \, du.
\]

The FCLT for \( \hat{H} \) satisfies a SVE rather than an ordinary SDE which is more commonly seen in the literature. This is solely because the service rates are assumed to be class-dependent. We summarize our key findings regarding the SVE in Remark 6.

Remark 6 (A closer look at the SVE (27)).

(i) Analytic solutions in special cases. Such an SVE (27) in general has no analytic solution, except for some special cases. For example, if \( \mu_i = h_{F_i}(w_i) \) for all \( 1 \leq i \leq K \) so that the drift term \( L(t, s) = 0 \), then the SVE (27) is a simple Brownian integral which admits an analytic solution. Another important case is when \( L(t, s) \) and \( J(t, s) \) are separable functions in \( t \) and \( s \), which is the case when service rates are class independent (see §3.3 for discussions of this important special case).
(ii) **Variability.** The SVE is driven by the Brownian motion $W$, which rises from aggregating all $3K$ independent Brownian motions $W_{\lambda_i}, W_{\theta_i}, W_{\mu_i}, 1 \leq i \leq K$ in (22); see the proof of Proposition 1 in §EC.1 for details. Indeed, the stochastic variability of the frontier waiting time process is collectively determined by the randomness in the arrivals, service times and abandonment times.

(iii) **Dependence on control functions.** The terms $L$ and $J$ are functions of model inputs $(\lambda_i, F_i, \mu_i, w_i)$ only, thus independent of the control functions $\kappa_i$ and $c$, which only appears in $K$. Hence, varying $\kappa_i$ and $c$ will affect the mean of $\hat{H}$, but not its variance. This is a crucial observation, because as will become clear later in §3.2, (i) computing the variance of $\hat{H}$ (which is uncontrollable via $c$ and $\kappa_i$) and (ii) appropriately shifting the mean of $\hat{H}$ (by adjusting our control functions) are critical in achieving desired class-dependent service levels.

(iv) **Algorithms.** We prove Proposition 1 by showing that the operators $\Gamma$ and $\Theta$ are both contractions in appropriate functional spaces, see §EC.1. In addition, our proof naturally leads to effective numerical algorithms for computing $M_{\hat{H}}$ and $C_{\hat{H}}$ (in fact, our algorithms converge geometrically fast). See Remark EC.2 in the e-companion for detailed discussions. It is obvious that $M_{\hat{H}}(t) = 0$ if $K(t) = 0$ (because a zero function now solves the FPE (29)). This will indeed be the case considered later in §3.2.

### 3.2. Asymptotic Service Differentiation and Stabilization

Given the SSC achieved by TV-SRS and TV-DPS, we now focus on investigating the one-dimensional process $\hat{H}$. When $n$ is large we hope to satisfy

\[
\alpha_i \equiv \mathbb{P}(V_i^n(t) > w_i) = \mathbb{P}(\hat{V}_i(t) > 0) \approx \mathbb{P}(\hat{H}(t + w_i) - \kappa_i(t + w_i) > 0)
\]

\[
= \mathbb{P}\left(\mathcal{N}\left(M_{\hat{H}}(t + w_i), \sigma_{\hat{H}}^2(t + w_i)\right) > \kappa_i(t + w_i)\right) = \mathbb{P}\left(\mathcal{N}(0, 1) > \frac{\kappa_i(t + w_i) - M_{\hat{H}}(t + w_i)}{\sigma_{\hat{H}}(t + w_i)}\right)
\]

for all $t \geq -w_i$, where $\mathcal{N}(\mu, \sigma^2)$ is a normal random variable with mean $\mu$ and variance $\sigma^2$, $\sigma_{\hat{H}}(t) = \sqrt{\text{Var}(\hat{H}(t))} = \sqrt{C_{\hat{H}}(t, t)}$ is the standard deviation of $\hat{H}(t)$ at $t$. Equation (31) further simplifies to

\[
\mathbb{P}\left(\mathcal{N}(0, 1) > \frac{\kappa_i(t) - M_{\hat{H}}(t)}{\sigma_{\hat{H}}(t)}\right) \approx \alpha_i, \quad t \geq 0,
\]

in which case we should choose appropriate control functions $\kappa_i$ and $c$ so that

\[
\kappa_i(t) - M_{\hat{H}}(t) = z_{1 - \alpha_i} \sigma_{\hat{H}}(t),
\]

(33)
where \( z_\alpha \) is the \( \alpha \)-quantile of a standard Gaussian random variable, that is, \( z_\alpha \) satisfies \( \mathbb{P}(\mathcal{N}(0, 1) \leq \alpha) = z_\alpha \).

One obvious solution to (33) is that, for any \( \kappa_i \), we can choose \( c \) appropriately so that \( K(t) \) in (28) is set to 0, so that \( M_{\tilde{H}}(t) = 0 \) for all \( t \) (note that FPE (29) now has a unique solution \( M_{\tilde{H}}(t) = 0 \) when \( K(t) = 0 \)), and this leads to the control formulas below in (34)–(35). We next show that these control functions are indeed the unique solutions to (33).

**Proposition 2 (Asymptotically unique control functions).** The condition (33) is satisfied if and only if

\[
c(t) = \sum_{i=1}^{K} \left\{ \eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s)e^{\eta_i(s-t)}(\mu_i - h_{F_i}(w_i))\kappa_i(s)ds \right\},
\]

\[
\kappa_i(t) = z_{1-\alpha_i}\sigma_{\tilde{H}}(t), \quad 1 \leq i \leq K.
\]

See \( \S \)EC.1 for the proof for Proposition 2. The safety staffing term \( c \) is indeed uniquely given by (34). However, to be rigorous, we should say that the prioritization regulators \( \kappa_1, \ldots, \kappa_K \) are unique up to adding any common function \( \Delta \), that is, applying any \( \bar{\kappa}_i(t) = \kappa_i(t) + \Delta(t) \) for \( 1 \leq i \leq K \) which will not make a difference in our TV-DPS rule.

**Remark 7 (Structure of the control functions).** The main idea here is that we choose appropriate control functions \( c \) and \( \kappa_i \) to tilt the mean of the error term \( \tilde{V}_i^n(t) \) (rather than the mean of \( V_i^n(t) \)), so that asymptotically the probability mass of \( \{ \tilde{V}_i^n(t) > 0 \} \) (or \( \{ V_i^n(t) > w_i \} \)) can be set to desired \( \alpha_i \) at all time \( t \). We observe from (34) that the second-order safety staffing term \( c \) depends on \( \alpha_i \) through the second-order prioritization regulator \( \kappa_i \), and \( \kappa_i \) depends on \( \alpha_i \) through \( z_{\alpha_i} \). Consistent with Remark 2, \( \kappa_i \) is decreasing in \( \alpha_i \), and its sign depends on how \( \alpha_i \) compares with 0.5, that is, \( \kappa_i(t) > 0 \) (\( \kappa_i(t) < 0 \)) if \( \alpha_i < 0.5 \) (\( \alpha_i > 0.5 \)). When the probability target \( \alpha_i = 0.5 \) for all \( 1 \leq i \leq K \), we have \( c(t) = \kappa_i(t) = 0 \) so that we lose the second-order terms in both TV-SRS and TV-DPS formulas. Another interesting observation is that a bigger system variability leads to more contrasting prioritization standards. To elaborate, consider the case \( \alpha_1 < 0.5 < \alpha_2 \) so that \( z_{1-\alpha_1} > 0 > z_{1-\alpha_2} \) and \( \kappa_1(t) > 0 > \kappa_2(t) \), the difference of the two prioritization regulators \( \kappa_1(t) - \kappa_2(t) \) is increasing in \( \sigma_{\tilde{H}}(t) \), which characterizes the system's overall stochastic variability (recall from Remark 6 that the variability of \( \tilde{H} \) captures the randomness of all
events, including arrivals, service times and abandonment times). This implies that in a more random environment, we rely less (more) on the state-dependent portion (deterministic control regulator) of TV-DPS to inform the scheduling decision; as the the system environment becomes more volatile, information of the system state becomes less useful. Finally, we emphasize that \( w_i (\alpha_i) \) is the first-order (second-order) QoS target, because a slight change in \( w_i (\alpha_i) \) affects the first-order (second-order) term in both the TV-SRS and TV-DPS formulas.

The next theorem establishes the asymptotic effectiveness of our methods.

**Theorem 2 (Asymptotic service differentiation and performance stabilization).** Under TV-SRS (15) and TV-DPS (16) with \( c_i(\cdot) \) and \( \kappa_i(\cdot) \) specified in (34) and (35), we have the following asymptotic stability results:

(i) Mean PWT and mean HWT are both asymptotically stabilized for all classes:

\[
\mathbb{E}[V^n_i(t)] \to w_i \quad \text{and} \quad \mathbb{E}[H^n_i(t)] \to w_i \quad \text{as} \quad n \to \infty, \quad 1 \leq i \leq K, \quad 0 < t \leq T.
\]  

(ii) TPoDs for PWT and HWT are both asymptotically stabilized for all classes:

\[
\mathbb{P}(V^n_i(t) > w_i) \to \alpha_i \quad \text{and} \quad \mathbb{P}(H^n_i(t) > w_i) \to \alpha_i \quad \text{as} \quad n \to \infty, \quad 1 \leq i \leq K, \quad 0 < t \leq T.
\]

**Remark 8 (Differentiation of the mean PWT).** If the goal is to asymptotically differentiate and stabilize the mean PWT \( \mathbb{E}[V^n_i(t)] \) rather than the TPoD, then we are free to choose any arbitrary \( \alpha_i \) in (34) and (35), because the second-order terms of TV-SRS and TV-DPS play a negligible role as the scale \( n \) increases. However, for a finite \( n \), numerical examples show that it is helpful to set \( \alpha_i = 0 \) for all \( 1 \leq i \leq K \), so that the probability mass \( \{V^n_i(t) > w_i\} \approx 0 \) and the symmetric structure of the nearly Gaussian distribution “guarantees” a balanced mean at \( w_i \). See §3.3 of the online appendix for more discussions and numerical examples.

### 3.3. The Case of Class-Independent Service Rate

It is well known that the case of class-dependent service rate can be more complex, see Kim et al. (2018) for example. In this subsection, we assume that service rates are class independent, that is, \( \mu_i = \mu \) for
all \(1 \leq i \leq K\). Under this assumption, we show that the results are simplified significantly; indeed, the functions \(L\) and \(K\) are now separable in \(t\) and \(s\) so that the SVE in (27) degenerates to a much more tractable Ornstein-Uhlenbeck (OU) process with time-varying drift and volatility. We summarize our results below.

**Corollary 2 (Frontier process \(\hat{\mathcal{H}}\) when service rates are class independent).** Suppose \(\mu_i = \mu, 1 \leq i \leq K\), then

(i) the limiting frontier process \(\hat{\mathcal{H}}\) satisfies the one-dimensional OU type SDE

\[
\eta(t)\hat{\mathcal{H}}(t) = -\int_0^t \eta(u)\hat{\mathcal{H}}(u)du + \mathcal{S}(t) + \mathcal{G}(t),
\]

where \(G(t) \equiv \sum_{i=1}^{K} G_i(t)\), for \(G_i(t)\) being the Brownian-driven terms given in Theorem 1, and

\[
\mathcal{S}(t) \equiv \sum_{i=1}^{K} \eta_i(t)\kappa_i(t) + \int_0^t \sum_{i=1}^{K} \eta_i(u) h_{F_i}(w_i)\kappa_i(u)du - c(t) - \mu \int_0^t c(u)du.
\]

(ii) The SDE (38) has a unique solution

\[
\hat{\mathcal{H}}(t) = \frac{1}{R(t)} \left( \int_0^t e^{\int_u^t \frac{L(v)}{\mu} dv} J(u)dW(u) + \int_0^t e^{\int_u^t \frac{L(v)}{\mu} dv} R(u)dL(u) + \int_0^t e^{\int_u^t \frac{L(v)}{\mu} dv} K(u)dR(u) \right),
\]

where \(W\) is a standard Brownian motion,

\[
R(t) = e^{\mu t}\eta(t), \quad \tilde{L}(t) = e^{\mu t} \sum_{i=1}^{K} \eta_i(t)(\mu - h_{F_i}(w_i)), \quad \tilde{J}(t) = e^{\mu t} \sqrt{\sum_{i=1}^{K} (F_i^c(w_i)\lambda_i(t - w_i) + \mu m_i(t)).}
\]

(iii) The variance of \(\hat{\mathcal{H}}(t)\) is

\[
\sigma_{\hat{\mathcal{H}}}(t) \equiv \text{Var} \left( \hat{\mathcal{H}}(t) \right) = \frac{1}{R^2(t)} \int_0^t e^{2\int_u^t \frac{L(v)}{\mu} dv} \tilde{J}^2(u)du.
\]

We next consider some special cases to obtain insights.

**Corollary 3 (Constant arrival rates).** When \(\lambda_i(t) = \lambda_i\), we have

\[
m_i(t) \sim m_i = \frac{\lambda_i F_i^c(w_i)}{\mu}, \quad c(t) \sim c = \sum_{i=1}^{K} \frac{w_i \lambda_i f_i(w_i)}{\mu} \kappa_i,
\]

\[
\kappa_i(t) \sim \kappa_i \equiv z_{1-\alpha_i} \left[ \frac{\sum_{j=1}^{K} \lambda_j F_j^c(w_i)}{\left( \sum_{j=1}^{K} \lambda_j f_j(w_j)w_j \right)^{1/2}} \right].
\]

where we say \(f(t) \sim g(t)\) if \(f(t)/g(t) \to 1\) as \(t \to \infty\).
Remark 9 (Average Staffing and Prioritization Levels). The constants in (41) and (42) can be used to compute the required average number of servers and scheduling threshold. When $K = 1$, our staffing formula (41) degenerates to the ED+QED staffing formulas (30) and (31) in Mandelbaum and Zeltyn (2009) which asymptotically controls the TPoD for the stationary $M/M/n + G$ model.

In addition, these analytic formulas can provide an estimate of the marginal prices of staffing and scheduling (MPSS), that is, to improve the service to the next level (e.g., reducing $w_i$ by $\Delta w_i$, or reducing $\alpha_i$ by $\Delta \alpha_i$), how many extra servers are need and how much should the scheduling threshold $\kappa_i$ be adjusted? See the online appendix for numerical examples on MPSS.

If $K = 1$, then our multi-class V model degenerates to a single-class $M_t/M/s_t + GI$ model.

Corollary 4 (The single-class case). When $K = 1$, the second-order staffing term $c(t)$ simplifies to

$$c(t) = z_1 \alpha e^{-\mu t} \left( Z(t) - (\mu - h_F(w)) \int_0^t Z(s) ds \right),$$

with

$$Z(t) \equiv e^{(\mu - h_F(w))t} \sqrt{\int_0^t e^{2h_F(w)} (F^c(w) \lambda(u - w) + \mu m(u))},$$

It is easy to check that (43) and (44) coincide with the staffing formulas (7) and (8) in Liu (2018), except for a time shift by $w$. This is due to the slightly different initial condition here.

4. Numerical Studies

In this section, we provide numerical examples and simulation comparisons to test the effectiveness of our TV-SRS and TV-DPS formulas. In §4.1 we first consider a base model having two customer classes and state-independent service rates. We next give additional simulation experiments in §4.2, including cases with smaller arrival rates and number of servers, higher quality of service, mixed scales of arrival rates, state-dependent service rates, and a five-class example.

4.1. A Two-Class Base Model

Because sinusoidal functions capture the periodic structure in realistic arrival patterns (see Chen et al. (2018), Feldman et al. (2008), Liu and Whitt (2012)), we consider sinusoidal arrival rates

$$\lambda_i(t) = \bar{\lambda}_i \left( 1 + r_i \sin(\gamma_i t + \phi_i) \right), \quad 1 \leq i \leq K,$$

(45)
with average rate $\bar{\lambda}_i$, relative amplitude $|r_i| < 1$, frequency $\gamma_i$, and phase $\phi_i$. We first consider a two-class V 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 $\mu_1 = \mu_2 = 1$ (with mean service time $1/\mu_i = 1$). To prioritize Class 1, we set higher QoS levels (i.e., lower target wait time and tail probability of delay). We set our target model parameters as $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$.

![Figure 3](image-url)  
*Figure 3*  
Computed control functions for a two-class base case: $m(t)$, $c(t)$, $\kappa_i(t)$ and $\sigma(t)$, $i = 1, 2$.

In Figure 3, we calculate and plot the required control functions for TV-SRS and TV-DPS in a finite time interval $[0, T]$, with $T = 24$, including the offered-load function $m(t)$ in (10), the second-order staffing term $c(t)$ in (34), the second-order prioritization regulators (35), and the standard deviation process of $\hat{H}$ in (39). Consistent with discussions in Remarks 2 and 7, we observe that $\kappa_1(t) > 0$ and $\kappa_2(t) < 0$ because $\alpha_1 = 0.2 < 0.5 < 0.8 = \alpha_2$. In addition, the second-order safety staffing term, $c(t)$, can be alternating between positive and negative.
Using these control functions in Figure 3, we conduct Monte-Carlo simulation experiments to test the effectiveness of TV-SRS and TV-DPS. For our base case, we let \( n = 50 \) and generate 5000 independent runs. Specifically, at each time \( 0 \leq t \leq T \) on an arbitrary run, we schedule the next customer into service according to our TV-DPS in (14) using the control function \( \kappa_i \) given in Figure 3. We plot (i) arrival rates, (ii) simulations of TPoD, and (iii) staffing functions, in Figure 4, using a sampling resolution (i.e., step size) \( \Delta t = 0.01 \). From a visual inspection of the middle panel of Figure 4, we see that our method effectively achieves stabilization of TPoD \( \mathbb{P}(V_i(t) > w_i) \) for both classes at their (differentiated) targets (dashed lines). Implementation details of the simulations are discussed in §EC.2.1.

### 4.1.1. Staffing Discretization

In practice (and in our simulation experiments), our TV-SRS formula needs to be discretized to integer values. Table 1 gives the time-averaged, maximum, and minimum simulation results for the two TPoD’s and their relative differences from targets \( (\mathbb{P}(V_i(t) > w_i) - \alpha_i)/\alpha_i \), using three staffing discretization methods (flooring, rounding, ceiling).
Table 1 exhibits the impact of adding and removing a server on the TPoD performance. As shown in the table, the discretization method seems to play a bigger role when the target QoS is high ($\alpha$ is small), as in the case of Class 1. In contrast, Class 2 with low QoS is relatively insensitive to the discretization method. We conduct the remainder of the simulations with the ceiling discretization method. As the scale $n$ increases, the discretization becomes insignificant and all methods will provide nearly equivalent TPoD performance.

### Table 1

<table>
<thead>
<tr>
<th>Class</th>
<th>Avg.</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Floor</td>
<td>Round</td>
<td>Ceiling</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.2312</td>
<td>0.2091</td>
<td>0.1819</td>
</tr>
<tr>
<td>%</td>
<td>(+15.59)</td>
<td>(+4.53)</td>
<td>(-9.06)</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.8085</td>
<td>0.7855</td>
<td>0.7612</td>
</tr>
<tr>
<td>%</td>
<td>(+1.06)</td>
<td>(-1.81)</td>
<td>(-4.85)</td>
</tr>
</tbody>
</table>

#### 4.1.2. 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 $\Delta_s$ (e.g., 30 minutes) and a continuous TV-SRS formula $s(t)$, we consider two $\Delta_s$-based discretization methods (i) average staffing level (ASL) and (ii) maximum staffing level (MSL), which are given by

$$s^{\text{ASL}}(t) = \sum_{i=1}^{\lceil T/\Delta_s \rceil} \bar{s}_i 1_{\{i \leq u \leq (i+1)\Delta_s\}},$$

$$s^{\text{MSL}}(t) = \sum_{i=1}^{\lceil T/\Delta_s \rceil} s_i^\uparrow 1_{\{i \leq u \leq (i+1)\Delta_s\}},$$

where $x \land y \equiv \min(x, y)$. MSL sets the staffing level in each interval as the the maximum of TV-SRS, ensur-
Figure 5 Simulation comparison for a two-class base case: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $P(V_i(t) > w_i)$ (middle panel); and (iii) time-varying ASL and MSL staffing levels having $\Delta_s = 0.5$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 5000 independent runs.

We again simulate our two-class base case model, 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 5. 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 $\Delta_s$ in the e-companion.

4.2. Other Cases

We next test the robustness of TV-SRS and TV-DPS by considering variates of the base model, including (i) higher QoS targets (§4.2.1), (ii) smaller system scale (§4.2.2), (iii) mixed class scales (§EC.2.2.2), (iv) class-dependent service rates (§4.2.4), and (v) a five-class example (§4.2.5).
4.2.1. 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 TV-SRS and TV-DPS when both classes have higher QoS targets. Figure 6 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 6 shows that TPoD’s remain relatively stable in both cases. See §3.2.4 in the online appendix for examples with higher QoS targets.

4.2.2. Smaller arrival rates  Our method is based on asymptotic analysis of the V model when $n \rightarrow \infty$, so it is evident that we should be able to achieve desired TPoD performance when $n$ is relatively large. An important question is how effective TV-SRS and TV-DPS are for a small-scale system. We now consider the two-class base model having a smaller scale $n = 5$. Due to the increased stochastic variability in small-scale models, we increase our sample size to 20000 independent runs in our Monte-Carlo simulations. Figure 7 shows that: (i) Due to the small arrival rate, the required staffing level is small, so that addition and removal of a single server from time to time lead to bigger TPoD bumps; (ii) Different staffing discritization methods now play bigger roles, that is, adding a server to the staffing level at all times can cause a much larger performance change; and nevertheless, (iii) our TV-SRS and TV-DPS yield relatively stable TPoD performance. This example shows that results derived from the large-scale (many-server) limits may have strong practical relevance, even for small-scale systems.
4.2.3. 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. (2018). We modify the arrival rates in our two-class base case so that $\tilde{\lambda}_1 = 0.1$, but set $n = 100$ so that the overall system size remains comparable to the base case. We see from Figure EC.1 in the e-companion, that even though the majority of arrivals to the system are from Class 1, we have effective TPoD stabilization for both classes.

4.2.4. Class-dependent service rates  Results in §3 enables us to treat the case of class-dependent service rates, which has strong practical relevance. Taking the CTAS example, a patient of higher acuity may require a much longer treatment, resulting in a smaller service rate. We now consider our two-class base model with $\mu_1 = 0.5$ and $\mu_2 = 1$ (so that a high priority class requires significantly more time in service). To obtain the control parameters, we calculated $\text{Var}(\hat{H}(t))$ according to the contraction-based algorithm given in Remark EC.2 of the e-companion. Simulations show that our methods continue to achieve desired service-level differentiation and performance stabilization; see Figure EC.2 in the e-companion. We discuss the implementation and convergence behavior of the our algorithm for $\text{Var}(\hat{H}(t))$ in Remark EC.2.
4.2.5. A Five-Class Example  Finally, motivated by the CTAS example, we now consider a five-class V model, having class-dependent sinusoidal arrival rates as in (45), exponential abandonment and service times. All model input parameters and QoS parameters are given in Table 2. The control functions are given in the left-hand panel of Figure 8. In this example, we intentionally let the sinusoidal arrival rates have class-dependent periods, frequencies, and relative amplitudes (see right-hand panel of Figure 8). Nevertheless, our method continues to successfully achieves stable TPoD-based service levels across all 5 classes.

5. Concluding Remarks

In this paper, we studied a service differentiation problem for a time-varying queueing system with multiple customer classes. Motivated by call center and health care applications, we measure class-dependent service levels using the so-called TPoD, that is, the probability the waiting time exceeds a delay target. Under a many-server asymptotic framework, we proposed a TV-SRS rule and a TV-DPS policy that can asymptotically achieve TPoD-based performance stabilization, across all customer classes, at any time $t$. Our new TV-DPS rule is both state dependent (based the real-time elapsed customer delays) and time dependent (capturing a time variability from the arrival processes). Supplementing our limit theorems on asymptotic differentiation and stabilization (Theorems 1 and 2), we also conduct extensive simulation experiments to provide engineering confirmation and practical insights. Numerical results show that our proposed solution works effectively in a wide range of model settings.

There are several avenues for future research in this area. One natural extension would be to consider a more general network with heterogeneous pools of servers under the setting of skill-based routing; this will
make the model more practical for service systems such as call centers. Another interesting direction is to consider scheduling policies that exploit other system state information, such as queue lengths.

References


Figure 8 A five-class based model: (i) Computed control functions $m(t)$, $c(t)$, and $\kappa_i(t)$ for $i = 1, \ldots, 5$ (left), (ii) Simulation comparisons for TPoD $P(V_i(t) > w_i)$, $i = 1, \ldots, 5$ (right), with $n = 50$, input and QoS parameters given in Table 2, and 5000 samples.


**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. Additional numerical examples appear in the longer online appendix Liu et al. (2018b).

**EC.1. Proofs**

Here we prove Theorems 1–2, Propositions 1–2, and Corollary 2. The proofs of Corollaries 3 and 4 are more straightforward, which are given in the online appendix.

Because we assume each baseline arrival-rate function $\lambda_i$ is bounded away from zero and infinity, we define $\lambda_i^\downarrow \equiv \inf_{0 \leq t \leq T} \lambda(t) > 0$ and $\lambda_i^\uparrow \equiv \sup_{0 \leq t \leq T} \lambda(t) < \infty$.

**EC.1.1. Proof of Theorem 1.**

The proof proceeds in four major steps, as indicated by the proof sketch presented on p.16 of the main paper.

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

$$V^n_i(t - H^n_i(t)) = H^n_i(t) + O(n^{-1} \log n) \quad (EC.1)$$

under the TV-DPS rule, where the error term will follow, as we will explain momentarily, if at any time the number of service completions (each taking $O(1/n)$ amount of time) needed before the HoL customer of the $i$th queue enters service is of order $O(1)$; i.e., 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 $O(1)$. To elaborate, 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 $t - H^n_i(t)$. Further we use $\xi^n_i$ to denote the inter-arrival time between $A$ and $B$. It is immediate that customer $A$ arrived at the system at time $t - H^n_i(t) - \xi^n_i$. 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

$$\left( t - \frac{w_j (H^n_i(t) + \xi^n_i)}{w_i}, t - \frac{w_j H^n_i(t)}{w_i} \right) \quad (EC.2)$$

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 $\mathcal{D}^{(2)}$ between two successive arrivals of $\mathcal{D}^{(1)}$ follows a geometric distribution with parameter $\frac{\lambda^{(1)}}{\lambda^{(1)}+\lambda^{(2)}}$. Now because the interval (EC.2) has a length of $(w_j x^n_i/w_i)$, the number class-$j$ customer who can enter service before $B$ is stochastically bounded by a geometric distributed random variable with mean $\frac{w_j \lambda_j}{w_i \lambda_i}$ and variance $\left( \frac{w_j \lambda_j}{w_i \lambda_i} \right)^2 + \frac{w_j \lambda_j}{w_i \lambda_i}$. 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 $\ell^{(1)}_i = \sum_{j \neq i} \frac{w_j \lambda_j}{w_i \lambda_i}$ and variance $\ell^{(2)}_i = \sum_{j \neq i} \left( \frac{w_j \lambda_j}{w_i \lambda_i} \right)^2 + \ell^{(1)}_i$. The logarithmic error term arises due to the fact that the total number of such $\xi^n_i$'s over any compact time interval is at most $O(n)$ and that the maximum of $O(n)$ i.i.d. random variables with a finite moment generating function is of order $O(\log n)$.

On the other hand, each inter-arrival time of class $i$ has a length of $O(1/n)$ under our MSHT scaling. Using the same reasoning for (EC.1), we have

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

or equivalently,

$$\hat{H}^n_i(t) = w_i(\hat{H}^n(t) - \kappa_i(t)) - O(n^{-1/2} \log n),$$

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

**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 Pang et al. (2007) for a precise definition of stochastic boundedness. In what follows, we will first prove that the sequence $\{(\hat{B}^n_i, \ldots, \hat{B}^n_K; \hat{H}^n); n \in \mathbb{N}\}$ is stochastically bounded. To that end, introduce the LLN- and CLT-scaled empirical process

$$\bar{U}^n(t, x) \equiv \frac{1}{n} \sum_{k=1}^{[nt]} 1_{\{X_k \leq x\}} \quad \text{for} \quad t \geq 0, \quad 0 \leq x \leq 1, \quad \text{and}$$

$$\hat{U}^n(t, x) \equiv \sqrt{n} \left( \bar{U}^n(t, x) - \mathbb{E} \left[ \bar{U}^n(t, x) \right] \right) = \frac{1}{\sqrt{n}} \left( \sum_{k=1}^{[nt]} 1_{\{X_k \leq x\}} - x \right),$$

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0, 1]$. Krichagina and Puhalskii (1997) have shown that $\hat{U}^n \Rightarrow \hat{U}$ in $\mathcal{D}_P$ as $n \to \infty$, where $\hat{U}$ is the standard Kiefer process. Paralleling (3.3) - (3.6) in Aras et al. (2018), we break the enter-service process $E^n_i(t)$ in (6) into three pieces, namely,

$$E^n_i(t) = E^n_{i,1}(t) + E^n_{i,2}(t) + E^n_{i,3}(t),$$

where

$$E^n_{i,1}(t) \equiv \sqrt{n} \int_{-H^n_i(0)}^{t-H^n_i(t)} F^n_i(V^n_i(u))d\hat{A}^n_i(u), \quad t \geq 0,$$
Following the decomposition given in (EC.5) - (EC.8), we can write
\[
E_{i,2}^n(t) \equiv \sqrt{n} \int_{-H_i^n(0)}^{t-H_i^n(t)} \int_0^1 1_{(y > F_i^n(V_i^n(u)))} \, d\hat{U}_i^n(A_i^n(u), y) \quad t \geq 0, \tag{EC.7}
\]
\[
E_{i,3}^n(t) \equiv n \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u)) \lambda_i(u) \, du \quad t \geq 0, \tag{EC.8}
\]
for \( \bar{A}_i^n, \hat{A}_i^n \) given by (17) and \( \hat{U}_i^n \) is a CLT-scaled empirical process specified by (EC.4).

Define the fluid version and CLT-scaled version of the enter-service process as
\[
\varepsilon_i(t) \equiv \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \, du, \tag{EC.9}
\]
\[
\hat{E}_i^n(t) \equiv n^{-1/2} (E_i^n(t) - n \varepsilon_i(t)) = n^{-1/2} \left( E_i^n(t) - n \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \, du \right). \tag{EC.10}
\]

Following the decomposition given in (EC.5) - (EC.8), we can write
\[
\hat{E}_i^n(t) = \hat{E}_{i,1}^n(t) + \hat{E}_{i,2}^n(t) + \hat{E}_{i,3}^n(t), \tag{EC.11}
\]
where
\[
\hat{E}_{i,1}^n(t) \equiv n^{-1/2} E_{i,1}^n(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u)) \, d\hat{A}_i^n(u) \quad t \geq 0, \tag{EC.12}
\]
\[
\hat{E}_{i,2}^n(t) \equiv n^{-1/2} E_{i,2}^n(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} \int_0^1 1_{(y > F_i^n(V_i^n(u)))} \, d\hat{U}_i^n(\bar{A}_i^n(u), y) \quad t \geq 0, \tag{EC.13}
\]
\[
\hat{E}_{i,3}^n(t) \equiv n^{-1/2} \left( E_{i,3}^n(t) - n \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \, du \right) \quad t \geq 0. \tag{EC.14}
\]

For the term \( \hat{E}_{i,3}^n \), we further deduce
\[
\hat{E}_{i,3}^n(t) = \sqrt{n} \left( \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u)) \lambda_i(u) \, du - \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \, du \right)
\]
\[
= \sqrt{n} \int_0^t F_i^c(H_i^n(u)) \lambda_i(u - H_i^n(u)) \, du - \sqrt{n} \int_0^t F_i^c(w_i) \lambda_i(u - w_i) \, du
\]
\[
- \int_0^t F_i^c(\bar{H}_i^n(u)) \lambda_i(u - H_i^n(u)) \, d\hat{H}_i^n(u) + O(n^{-1/2} \log n) \tag{EC.15}
\]
\[
= - \int_0^t \{ f_i(c_i^n(u)) \lambda_i(u - c_i^n(u)) + F_i^c(c_i^n(u)) \lambda'_i(u - c_i^n(u)) \} \, w_i(\bar{H}_i^n(u) - \kappa_i(u)) \, du
\]
\[
- \int_0^t w_i F_i^c(\bar{H}_i^n(u)) \lambda_i(u - H_i^n(u)) \, d(\hat{H}_i^n(u) - \kappa_i(u)) + O(n^{-1/2} \log n),
\]
where the second equality follows by a change of variables, namely \( t \rightarrow t - H_i^n(t) \), plus the relation (EC.1), while the third equality follows from (EC.3) and applying the mean-value theorem with \( \zeta_i^n(t) \) satisfying
\[
H_i^n(t) \land w_i \leq \zeta_i^n(t) \leq H_i^n(t) \lor w_i. \tag{EC.16}
\]
On the other hand, the conservation of flow implies

\[ E_i^n(t) = B_i^n(t) + D_i^n(t), \]  

(EC.17)

where we have used \( D_i^n(t) \) to denote the number of class-\( i \) customers that have completed service by time \( t \). From (EC.9) it follows

\[ \varepsilon_i(t) = \int_0^t F_i^n(w_i) \lambda_i(u - w_i) \, du = m_i(t) + \int_0^t \mu_i m_i(u) \, du, \]  

(EC.18)

where the last equality follows from (10). Multiplying both sides of (EC.18) by \( n \), subtracting it from (EC.17), and dividing both sides by \( n^{1/2} \) yields,

\[ \hat{E}_i^n(t) = \hat{B}_i^n(t) + \mu_i \int_0^t \hat{B}_i^n(u) \, du + \hat{D}_i^n(t) \quad \text{or} \quad d \hat{B}_i^n(t) + \mu_i \hat{B}_i^n(t) \, dt = d \hat{E}_i^n(t) - d \hat{D}_i^n(t), \]  

(EC.19)

where we have defined

\[ \hat{D}_i^n(t) \equiv n^{-1/2} \left( D_i^n(t) - \mu_i \int_0^t B_i^n(u) \, du \right). \]

Let \( B^n(t) \equiv B_1^n(t) + \cdots + B_K^n(t) \). It is standard routine to show, with the overloading assumption (10),

\[ B^n(t) = s^n(t) + s_0^n(t) \]  

(EC.20)

holds with arbitrarily high probability by choosing \( n \) large enough. Hence, it suffices to focus on the sample paths for which (EC.20) holds. In this case one can easily derive

\[ \sum_{i=1}^K \hat{B}_i^n(t) = n^{-1/2} (B^n(t) - nm(t)) = n^{-1/2} (s^n(t) + s_0^n(t) - nm(t)) = c(t) + \hat{s}_0^n(t), \]  

(EC.21)

for \( \hat{s}_0^n(t) \equiv n^{-1/2} s_0^n(t) \). Appealing to (12) and (EC.18), we obtain

\[ \hat{s}_0^n(t) \leq \psi \left( \sum_{i \in \mathcal{I}} \hat{D}_i^n(t) + \sum_{i \in \mathcal{I}} \mu_i \int_0^t \hat{B}_i^n(u) \, du - n^{-1/2} \sum_{i \in \mathcal{I}} \varepsilon_i(t) \right) \leq \sum_{i \in \mathcal{I}} \left| \hat{D}_i^n(t) \right| + \sum_{i \in \mathcal{I}} \mu_i \int_0^t \left| \hat{B}_i^n(u) \right| \, du. \]  

(EC.22)

Upon substituting (EC.11) - (EC.13) and (EC.15) into (EC.19), we obtain

\[ d \hat{B}_i^n(t) + w_i F_i^n(H_i^n(t)) \lambda_i(t - H_i^n(t)) \, d \hat{H}_i^n(t) \]

\[ = - \mu_i \hat{B}_i^n(t) \, dt - [f_i(\zeta_i^n(t)) \lambda_i(t - \zeta_i^n(t)) + F_i^n(\zeta_i^n(t)) \lambda_i'(t - \zeta_i^n(t))] w_i \hat{H}_i^n(t) \, dt \]

\[ + [f_i(\zeta_i^n(t)) \lambda_i(t - \zeta_i^n(t)) + F_i^n(\zeta_i^n(t)) \lambda_i'(t - \zeta_i^n(t))] w_i \kappa_i(t) \, dt \]

\[ + w_i F_i^n(H_i^n(t)) \lambda_i(t - H_i^n(t)) \, d \kappa_i(t) + d \hat{E}_{i,1}^n(u) + d \hat{E}_{i,2}^n(u) \]

\[ - d \hat{D}_i^n(u) + O(n^{-1/2} \log n) \quad \text{for} \quad i = 1, \ldots, K. \]  

(EC.23)
For an arbitrarily given $k \in \mathcal{I}$, from (EC.21) it follows $\hat{B}_k^n(t) = c(t) + \hat{s}_d^n(t) - \sum_{i \neq k} \hat{B}_i^n(t)$. Plugging it into (EC.23) for $i = k$, we obtain a set of $K$ linear differential equations with respect to the $K$-dimensional process $(\hat{B}_1^n, \ldots, \hat{B}_{k-1}^n, \hat{B}_{k+1}^n, \ldots, \hat{B}_K^n, \hat{H}^n)$. Similar to what was done to (5.14) in Aras et al. (2018), we apply the Gronwall’s inequality together with the stochastic boundedness of $\hat{E}_{i,1}$, $\hat{E}_{i,2}$, $\hat{B}_i^n$, the assumed properties of $\lambda_i$, $f_i$, $F_i^e$ plus the technical estimate (EC.22) to conclude the stochastic boundedness of the sequence $\{(\hat{B}_1^n, \ldots, \hat{B}_{k-1}^n, \hat{B}_{k+1}^n, \ldots, \hat{B}_K^n, \hat{H}^n) ; n \in \mathbb{N}\}$; as an immediate consequence, $\{\hat{B}_k^n ; n \in \mathbb{N}\}$ is stochastically bounded. Similar to the proof of Lemma 1 in Sun and Whitt (2018), we use (EC.22) and the established stochastic boundedness of $\{((\hat{B}_1^n, \ldots, \hat{B}_K^n) ; n \in \mathbb{N}\}$ to deduce

$$\hat{s}_d^n \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \quad \text{(EC.24)}$$

On the other hand, by the established stochastic boundedness of $\{\hat{H}^n ; n \in \mathbb{N}\}$ plus the relations (EC.3) and (EC.1), it holds that $\{\hat{H}_n ; n \in \mathbb{N}\}$ and $\{\hat{V}_n ; n \in \mathbb{N}\}$ are stochastically bounded, for $i = 1, \ldots, K$. This implies the FWLLN for the HWT and PWT processes, that is, as $n \to \infty$,

$$(H^n, H_1^n, \ldots, H_K^n, V_1^n, \ldots, V_K^n) \Rightarrow (\epsilon, w_1 \epsilon, \ldots, w_K \epsilon, w_1 \epsilon, \ldots, w_K \epsilon) \quad \text{in} \quad \mathcal{D}^{2K+1}, \quad \text{(EC.25)}$$

where the joint convergence holds 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.12) and (EC.25) to get

$$\hat{E}_{i,1}^n(t) \Rightarrow \hat{E}_{i,1}(t) \equiv F_i^c(w_i) \int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)} \, dW_{\lambda_i}(u), \quad \text{(EC.26)}$$

where $W_{\lambda_i}$ is a standard Brownian motion.

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

$$\hat{E}_{i,2}^n(t) \Rightarrow \hat{E}_{i,2}(t) \equiv \sqrt{F_i^c(w_i) F_i(w_i)} \int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)} \, dW_{\theta_i}(u), \quad \text{(EC.27)}$$

for $W_{\theta_i}$ being a standard Brownian independent of $W_{\lambda_i}$. The essential structure of the proof for (EC.27) 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 $\{((\hat{B}_1^n, \ldots, \hat{B}_K^n) ; n \in \mathbb{N}\}$, we have the FWLLN for the busy-server processes

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

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

$$\hat{D}_i^n(\cdot) = n^{-1/2} \left[ \Pi_i^n \left( n \mu_i \int_0^{\hat{D}_i^n(u)} du \right) - n \mu_i \int_0^{\hat{D}_i^n(u)} du \right] \Rightarrow W_{\mu_i} \left( \mu_i \int_0^{m_i(u)} du \right) \quad \text{as} \quad n \to \infty, \quad \text{(EC.28)}$$
We then use integration by parts to deduce both (2017) for a reference, which is a variant of Theorem 7.6 of Pang et al. (2007). To 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 \( \hat{H}_i^n \), \( \hat{B}_1^n, \ldots, \hat{B}_K^n \), we seek to simplify the right-hand side of (EC.15). First we note that the inequality (EC.16) and the convergence in (EC.27) imply

\[
\zeta^n_i(t) = w_i + O(n^{-1/2}) = H^n_i(t) + O(n^{-1/2} \log n).
\]

To establish the convergence of (20), we will need to strengthen (EC.26), (EC.27) and (EC.28) to joint convergence. The joint convergence of multiple random elements is equivalent to individual convergence if they are independent. Here \( \hat{E}_{i,1}^n \), \( \hat{E}_{i,2}^n \) and \( \hat{D}_i^n \) are not independent because both \( \hat{E}_{i,1}^n \) and \( \hat{E}_{i,2}^n \) involve the arrival-time sequence, and \( \hat{D}_i^n \) depends on \( B^n_i \) which in turn correlates with \( E^n_i \) through (EC.17). But they are conditionally independent given \( A^n_i, H^n_i, V^n_i, B^n_i \). 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).

We then use integration by parts to deduce

\[
- \int_0^t w_i F_i^n(\zeta^n_i(u)) \lambda_i(u - \zeta^n_i(u))(\hat{H}^n_i(u) - \kappa_i(u))du - \int_0^t w_i F_i^n(H^n_i(u)) \lambda_i(u - H^n_i(u))d(\hat{H}^n_i(u) - \kappa_i(u))
\]

\[
= - w_i F_i^n(\zeta^n_i(t)) \lambda_i(t - \zeta^n_i(t))(\hat{H}^n_i(t) - \kappa_i(t))
+ \int_0^t w_i \{ F_i^n(\zeta^n_i(u)) \lambda_i(u - \zeta^n_i(u)) - F_i^n(H^n_i(u)) \lambda_i(u - H^n_i(u)) \} d(\hat{H}^n_i(u) - \kappa_i(u))
+ \int_0^t w_i \lambda_i(u - \zeta^n_i(u))(\hat{H}^n_i(u) - \kappa_i(u))dF_i^n(\zeta^n_i(u))
\]

\[
= - w_i F_i^n(w_i) \lambda_i(t - w_i)(\hat{H}^n_i(t) - \kappa_i(t)) + O(n^{-1/2} \log n),
\]

where the last equality holds due to (EC.29). Upon plugging (EC.30) into (EC.15), we obtain

\[
\hat{E}_i^n(t) = - \int_0^t w_i f_i(w_i) \lambda_i(u - w_i)(\hat{H}^n_i(u) - \kappa_i(u))du - w_i F_i^n(w_i) \lambda_i(t - w_i)(\hat{H}^n_i(t) - \kappa_i(t)) + O(n^{-1/2} \log n).
\]

Now plugging (EC.11) and the equation above into (EC.19), we get, for \( i = 1, \ldots, K \),

\[
\hat{B}_i^n(t) + w_i F_i^n(w_i) \lambda_i(t - w_i)\hat{H}^n_i(t)
= - \mu_i \int_0^t \hat{B}_i^n(u)du - \int_0^t w_i f_i(w_i) \lambda_i(u - w_i)\hat{H}^n_i(u)du + \int_0^t w_i f_i(w_i) \lambda_i(u - w_i)\kappa_i(u)du
+ w_i F_i^n(w_i) \lambda_i(t - w_i)\kappa_i(t) + \hat{E}_{i,1}^n(t) + \hat{E}_{i,2}^n(t) - \hat{D}_i^n(t) + O(n^{-1/2} \log n).
\]

The joint convergence \( (\hat{H}^n, \hat{B}_1^n, \ldots, \hat{B}_K^n) \Rightarrow (\hat{H}, \hat{B}_1, \ldots, \hat{B}_K) \) then follows by applying the continuous mapping theorem (see Theorem 4.1 of Pang et al. (2007)) to (EC.20) and (EC.31), with the joint convergence
of \(\tilde{s}_d, \tilde{E}^{n}_{i,1}, \tilde{E}^{n}_{i,2}, \) and \(\tilde{D}^{n}_i\) as specified by (EC.24), (EC.26), (EC.27), and (EC.28), respectively. Alternatively, one can subtract (EC.31) by (21) 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 Aras et al. (2018). The convergence of \(\{\tilde{H}^n_i; n \in \mathbb{N}\}\) and \(\{\tilde{V}^n_i; n \in \mathbb{N}\}\) follow easily from and (EC.3) and (EC.1), respectively.

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

\[
Q^n_i(t) = Q^n_{i,1}(t) + Q^n_{i,2}(t) + Q^n_{i,3}(t),
\]

where

\[
Q^n_{i,1}(t) = \sqrt{n} \int_{t-H^n_i(t)}^t F^c_i(t-u) d\tilde{A}^n_i(u), \quad t \geq 0,
\]

\[
Q^n_{i,2}(t) = \sqrt{n} \int_{t-H^n_i(t)}^t \int_0^1 1_{\{x>F^c_i(t-u)\}} d\tilde{U}^n_i(\tilde{A}^n_i(u), x) \quad t \geq 0,
\]

\[
Q^n_{i,3}(t) = n \int_{t-H^n_i(t)}^t F^c_i(t-u) \lambda_i(u) du \quad t \geq 0.
\]

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

\[
\tilde{Q}^n_i(t) = n^{-1/2} (Q^n_i(t) - nq_i(t)) = \tilde{Q}^n_{i,1}(t) + \tilde{Q}^n_{i,2}(t) + \tilde{Q}^n_{i,3}(t),
\]

where

\[
\tilde{Q}^n_{i,1}(t) = \int_{t-H^n_i(t)}^t F^c_i(t-u) d\tilde{A}^n_i(u) \Rightarrow \int_{t-w_i}^t F^c_i(t-u) d\tilde{A}_i(u),
\]

\[
\tilde{Q}^n_{i,2}(t) = \int_{t-H^n_i(t)}^t \int_0^1 1_{\{x>F^c_i(t-u)\}} d\tilde{U}^n_i(\tilde{A}^n_i(u), x) \Rightarrow \int_{t-w_i}^t \sqrt{F^c_i(t-u)F_i(t-u)} \lambda_i(u) dW_{\tilde{q}_i}(u),
\]

\[
\tilde{Q}^n_{i,3}(t) = \sqrt{n} \int_{t-H^n_i(t)}^{t-w_i} F^c_i(t-u) \lambda_i(u) du \Rightarrow F^c_i(w_i) \lambda_i(t-w_i) \tilde{H}_i(t).
\]

Here the proof for (EC.36) and (EC.37) is very similar to that of (EC.26) and (EC.27), and the proof for (EC.38) is also straightforward.
EC.1.2. Proof of Proposition 1.

The multi-dimensional SDE (21) is equivalent to

\[
d\left(e^{\mu_t \tilde{B}(t)}\right) = e^{\mu_t} \left(-w_i F_i^c(w_i) \lambda_i(t - w_i) \hat{H}(t) - \int_0^t w_i f_i(w_i) \lambda_i(u - w_i) \hat{H}(u) du + y_i(t) + G_i(t)\right),
\]

(EC.39)

where

\[
\tilde{B}_i(t) = \int_0^t \tilde{B}_i(u) du \quad \text{and} \quad y_i(t) = w_i F_i^c(w_i) \lambda_i(t - w_i) \kappa_i(t) + \int_0^t w_i f_i(w_i) \lambda_i(u - w_i) \kappa_i(u) du.
\]

Integrating (EC.39) from 0 to \(t\) yields

\[
\tilde{B}_i(t) = e^{-\mu_t} \int_0^t e^{\mu_s} \left(-w_i F_i^c(w_i) \lambda_i(s - w_i) \hat{H}(s) - \int_0^s w_i f_i(w_i) \lambda_i(u - w_i) \hat{H}(u) du + y_i(s) + G_i(s)\right) ds
\]

\[
= e^{-\mu_t} \left(-\int_0^t e^{\mu_s} w_i F_i^c(w_i) \lambda_i(s - w_i) \hat{H}(s) ds - \int_0^t w_i f_i(w_i) \lambda_i(u - w_i) \hat{H}(u) \int_u^t e^{\mu_s} ds du + \int_0^t e^{\mu_s} y_i(s) ds + \int_0^t e^{\mu_s} G_i(s) ds\right)
\]

\[
= \int_0^t w_i \lambda_i(s - w_i) \left(-F_i^c(w_i) e^{\mu_i(s - t)} - f_i(w_i) \frac{1 - e^{\mu_i(s - t)}}{\mu_i}\right) \hat{H}(s) ds + \int_0^t e^{\mu_i(s - t)} y_i(s) ds + \int_0^t e^{\mu_i(s - t)} G_i(s) ds.
\]

Summing up over \(i\) from 1 to \(K\), we have

\[
\int_0^t c(s) ds = \sum_{i=1}^K \int_0^t \tilde{B}_i(t) = \sum_{i=1}^K \int_0^t \sum_{i=1}^K w_i \lambda_i(s - w_i) \left(-F_i^c(w_i) e^{\mu_i(s - t)} - f_i(w_i) \frac{1 - e^{\mu_i(s - t)}}{\mu_i}\right) \hat{H}(s) ds
\]

\[
+ \sum_{i=1}^K \int_0^t e^{\mu_i(s - t)} \left(w_i F_i^c(w_i) \lambda_i(s - w_i) \kappa_i(s) + \int_0^s w_i f_i(w_i) \lambda_i(u - w_i) \kappa_i(u) du\right) ds
\]

\[
+ \sum_{i=1}^K \int_0^t e^{\mu_i(s - t)} \int_0^s \sqrt{F_i^c(w_i)} \lambda_i(u - w_i) + \mu_i m_i(u) dW_i(u) ds
\]

\[
= \sum_{i=1}^K \int_0^t w_i \lambda_i(s - w_i) \left(-F_i^c(w_i) e^{\mu_i(s - t)} - f_i(w_i) \frac{1 - e^{\mu_i(s - t)}}{\mu_i}\right) \hat{H}(s) ds + \sum_{i=1}^K \int_0^t w_i \lambda_i(s - w_i) \kappa_i(u) \left(F_i^c(w_i) e^{\mu_i(s - t)} + f_i(w_i) \frac{1 - e^{\mu_i(s - t)}}{\mu_i}\right) du
\]

\[
+ \sum_{i=1}^K \int_0^t \frac{1 - e^{\mu_i(u - t)}}{\mu_i} \sqrt{F_i^c(w_i)} \lambda_i(u - w_i) + \mu_i m_i(u) dW_i(u), \quad \text{(EC.40)}
\]

where the second equality holds by aggregating three independent Brownian motions \(W_{\mu_i}, W_{\theta_i}\) and \(W_{\lambda_i}\) in (22) into one independent standard Brownian motion \(W_i\) for each \(1 \leq i \leq K\). Differentiating (EC.40) yields

\[
c(t) = -\sum_{i=1}^K w_i \lambda_i(t - w_i) F_i^c(w_i) \hat{H}(t) + \int_0^t \sum_{i=1}^K w_i \lambda_i(s - w_i) e^{\mu_i(s - t)} (\mu_i F_i^c(w_i) - f_i(w_i)) \hat{H}(s) ds
\]
Hence, we have
\[
\|\text{image for any } y \text{ point theorem implies that the FPE (EC.43) has a unique solution over the entire interval } [0, T].
\]
This will recursively guarantee the contraction property over all smaller intervals. Hence, the Banach fixed
point theorem implies that the FPE (EC.43) has a unique solution over the entire interval [0, T].

And further aggregating the independent Brownian motions \( W_1, \ldots, W_K \) into \( \mathcal{W} \) yields the SVE in (27).

**Uniqueness and existence of solution to the SVE** (27). Consider two functions \( x, y \in \mathbb{C} \) (space of continuous functions) satisfying an equation
\[
x(t) = \int_0^t L(t, s)x(s)ds + y(t).
\]
we show that (EC.41) 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
\[
\psi(x)(t) \equiv \int_0^t L(t, s)x(s)ds + y(t).
\]
Therefore, \( x \) solves the fixed-point equation (FPE)
\[
x = \psi(x).
\]
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 \( \|x\|_T = \sup_{0 \leq t \leq T} |x(t)| \). We have
\[
|\psi(x_1)(t) - \psi(x_2)(t)| \leq \int_0^t |L(t, s)|ds \cdot \|x_1 - x_2\|_T
\leq \|x_1 - x_2\|_T \left( \sum_{i=1}^K \frac{w_i \lambda_i}{\|F_i^c(w_i)\|} \right) \frac{\sum_{i=1}^K w_i \lambda_i}{\sum_{i=1}^K w_i \lambda_i} F_i^c(w_i) t.
\]
Hence, we have \( \|\psi(x_1) - \psi(x_2)\|_T \leq L^T \|x_1 - x_2\|_T \), where the constant
\[
L^T = \sum_{i=1}^K \frac{w_i \lambda_i}{\|F_i^c(w_i)\|} (\mu_i F_i^c(w_i) + f_i(w_i)) < \infty,
\]
which is guaranteed by the strict positivity assumptions on \( w_i, \lambda_i \) and \( F_i^c \) for all \( 1 \leq i \leq K \). In case \( L^T > 1 \), we can partition the interval \([0, T]\) to successive smaller intervals with length \( \Delta T \) satisfying \( \Delta T < 1/L^T \). This will recursively guarantee the contraction property over all smaller intervals. Hence, the Banach fixed point theorem implies that the FPE (EC.43) has a unique solution over the entire interval \([0, T]\).

Consequently, the function \( \phi \) specified by (EC.41) is well-defined because \( \phi(y) \) has one and only one image for any \( y \). So we conclude that (27) has a unique solution \( \hat{H} \). If fact, we can write (27) as
\[
\hat{H} = \phi \left( \int_0^t J(\cdot, s)dW(s) + K(\cdot) \right).
\]
Remark EC.1. The strict positivity assumptions on \( \lambda_i \) and \( F_i^c \) for all classes \( 1 \leq i \leq K \) can be relaxed. Note that the contraction property (EC.44) continues to hold as long as there exists some (not all) class \( i \) such that \( \lambda_i^+ \) and \( F_i^c(w_i) \) are both positive.

To show that \( \hat{H} \) is Gaussian, we again use the contraction \( \psi \) defined in (EC.42). We follow the steps that establish strong solutions in Karatzas and Shreve (1991). Define a sequence of processes \( \{\hat{H}^{(k)}, k = 0, 1, 2, \ldots\} \) such that \( \hat{H}^{(0)}(t) = 0 \), and \( \hat{H}^{(k+1)} = \psi(\hat{H}^{(k)}) \) with \( y(t) = \int_0^t J(t, s) dW(s, \omega) \) for \( k \geq 0 \). (For each Brownian path and associated Brownian integral, we recursively define the sequence.) We can show that \( \hat{H}^{(k)} \) is Gaussian using an inductive argument. Specifically, \( \hat{H}^{(k+1)} \) is Gaussian because both \( \int_0^t L(t, s) \hat{H}^{(k)}(s) ds \) and \( \int_0^t J(t, s) dW(s, \omega) \) are Gaussian. Because \( \psi \) is a contraction, we know that \( \hat{H} \) is the almost sure limit of \( \hat{H}^{(k)} \), which implies weak convergence. Hence, \( \hat{H} \) is again Gaussian (because the limit of convergent Gaussian processes is again Gaussian). To elaborate, we may consider the characteristic function of \( \hat{H}^{(k)}(t) \): \( \Phi_k(s) = e^{is\mu_k - s^2\sigma_k^2/2} \) (with \( \mu_k \) and \( \sigma_k^2 \) being the mean and variance of \( \hat{H}^{(k)}(t) \)), which must converge to the characteristic function of \( \hat{H} \). 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 \( \hat{H} \) has the form \( e^{is\mu_\infty - s^2\sigma_\infty^2/2} \), which concludes the Gaussian distribution.

_Treating the mean and variance of \( \hat{H} \)._ Taking expectation in (27) yields

\[
\hat{m}_\hat{H}(t) = \int_0^t L(t, s) \hat{m}_\hat{H}(s) ds + K(t), \quad \text{where} \quad \hat{m}_\hat{H}(t) = \mathbb{E}[\hat{H}(t)]. \tag{EC.46}
\]

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

\[
\Gamma(x)(t) = \int_0^t L(t, s)x(s)ds + K(t).
\]

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

\[
|\Gamma(x_1)(t) - \Gamma(x_2)(t)| \leq \int_0^t |L(t, s)||x_1(s) - x_2(s)| ds \leq L^+t||x_1 - x_2||_1,
\]

where the finite upperbound \( L^+ \) is given by (EC.45). The rest of the proof is similar.

To treat the variance of \( \hat{H} \), consider the SVE (27) at \( 0 \leq s, t \leq T \)

\[
H(t) - \int_0^t L(t, u)H(u)du = \int_0^t J(t, u)dW(u),
\]

\[
H(s) - \int_0^s L(s, v)H(v)dv = \int_0^s J(s, v)dW(v).
\]

Multiplying the two equations and taking expectation yield that

\[
C(t, s) = -\int_0^t \int_0^s L(t, u)h(s, v)C(u, v)dvdudu + \int_0^s \int_0^t J(t, u)J(s, u)du
\]
\[
+ \int_0^t L(t, u)C(u, s)du + \int_0^s h(s, v)C(t, v)dv,
\]
where \( C(t, s) = \text{Cov}(\hat{H}(t), \hat{H}(s)) \), or equivalently, an FPE
\[
C = \Theta(C), \tag{EC.47}
\]
where \( C(\cdot, \cdot) \in \mathbb{C}([0, T]^2) \), and the operator
\[
\Theta(C)(t, s) = -\int_0^t \int_0^s L(t, u)h(s, v)C(u, v)dvdu + \int_0^t L(t, u)C(u, s)du \\
+ \int_0^s L(s, v)C(t, v)dv + \int_0^{s+t} J(t, u)J(s, u)du. \tag{EC.48}
\]
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
\[
|\Theta(x_1)(t, s) - \Theta(x_2)(t, s)| \leq \int_0^t \int_0^s |L(t, u)L(s, v)| \cdot |x_1(u, v) - x_2(u, v)|dvdu \\
+ \int_0^t |L(t, u)| \cdot |x_1(u, s) - x_2(u, s)|du + \int_0^s |L(s, v)| \cdot |x_1(t, v) - x_2(t, v)|dv \\
\leq ((L^\uparrow)^2 ts + L^\uparrow t + L^\uparrow s) \|x_1 - x_2\|_T.
\]
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.2 (Numerical Algorithm for \( \sigma^2_{\hat{H}}(t) \)).** The above proof of the existence and uniqueness of the FPE (EC.47) automatically suggests the following recursive algorithm to compute the covariance \( C(t, s) \) and variance \( \sigma^2_{\hat{H}}(t) \). To begin with, we pick an acceptable error target \( \epsilon > 0 \).

**Algorithm:**
1. Pick an initial candidate \( C^{(0)}(\cdot, \cdot) \);
2. In the \( k\)th iteration, let \( C^{(k+1)} = \Theta \left( C^{(k)} \right) \) with \( \Theta \) given in (EC.48).
3. 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 \( \sigma^2_{\hat{H}}(t) = C(t, t) \), for \( 0 \leq t \leq T \), which will be used later to devise required control functions \( c \) and \( \kappa_i \). The algorithm to compute the mean \( M_{\hat{H}} \) is similar.

**EC.1.3. Proof of Proposition 2**

First note that the FPE (29) specifies a well-defined function \( \phi: \mathbb{C} \to \mathbb{C} \) such that
\[
M_{\hat{H}} = \phi(K). \tag{EC.49}
\]
See the proof of the uniqueness and existence of the SVE (specifically, see (EC.41)–(EC.45)) for details.

Let \((\kappa^*, c^*) \equiv (\kappa_1^*, \ldots, \kappa_K^*, c^*)\), with \(\kappa_i^*\) and \(c^*\) given in (35) and (34). Let \(K^*\) and \(M^*_H\) be the corresponding version of (28) and the mean of \(\tilde{H}\). (We know that \(K^*(t) = M^*_H(t) = 0\).) So we have

\[
\kappa_i^*(t) = \kappa_i^*(t) - M^*_H(t) = z_{1-i}\sigma_H(t), \quad 1 \leq i \leq K. \tag{EC.50}
\]

Now consider another solution to \((\check{H}, \check{c})\) to (33), with \((\check{H}, \check{c}) \equiv (\kappa_1^* + \Delta \kappa_1, \ldots, \kappa_K^* + \Delta \kappa_K, c^* + \Delta c)\). Let \(\tilde{K}\) and \(\tilde{M}_H\) be the corresponding version of (28) and mean of \(\tilde{H}\). By (33), we have

\[
\kappa_i^*(t) + \Delta \kappa_i(t) = \tilde{M}_H(t) = \tilde{M}_H(t) = \tilde{H}(t), \quad 1 \leq i \leq K. \tag{EC.51}
\]

Comparing (EC.50) with (EC.51), we must have

\[
\Delta \kappa_i(t) = \tilde{M}_H(t) - M^*_H(t) = \Delta \kappa(t) \quad \text{for all } 1 \leq i \leq K. \tag{EC.52}
\]

Hence, any alternative solution to (33) (if any) has the form \((\kappa_1^* + \Delta \kappa_1, \ldots, \kappa_K^* + \Delta \kappa, c^* + \Delta c)\). Next, \(M^*_H = \phi(K^*)\) and \(\tilde{M}_H = \phi(\tilde{K})\) imply that

\[
M^*_H(t) = \int_0^t L(t, s)M^*_H(s)ds + K^*(t) \quad \text{and} \quad \tilde{M}_H(t) = \int_0^t L(t, s)\tilde{M}_H(s)ds + \tilde{K}(t),
\]

which leads to

\[
\Delta \kappa(t) = \tilde{M}_H(t) - M^*_H(t) = \int_0^t L(t, s)\left(\tilde{M}_H(s) - M^*_H(s)\right)ds + \left(\tilde{K}(t) - K^*(t)\right),
\]

\[
\equiv \int_0^t L(t, s)\Delta \kappa(s)ds + \left(\tilde{K}(t) - K^*(t)\right), \tag{EC.53}
\]

where the last equality holds by the first equality. By (EC.52) and (28), we have

\[
\tilde{K}(t) - K^*(t) = \frac{\Delta \kappa(t) \sum_{i=1}^K \left(\eta_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))ds\right) - \Delta c(t)}{\eta(t)}. \tag{EC.54}
\]

Finally, combining (EC.53) with (EC.54), we must have, for any \(\Delta \kappa\),

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

where the last equality above holds by (28). 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. \(\square\)
EC.1.4. Proof of Theorem 2

The FCLT limits in Theorem 1 implies the FWLLN, that is, we have

\[(H^n_i, V^n_i) \Rightarrow (w_i e, w_i e) \text{ in } D^2, \text{ for } 1 \leq i \leq K, \text{ as } n \to \infty,\]

where \(e(t) = 1\). To prove part (i) of Theorem 2, it is sufficient to show that \(\{V^n_i, n \geq 1\}\) and \(\{H^n_i, n \geq 1\}\) are uniformly integrable (u.i.).

We first prove that the queue length \(Q^n_i\) is u.i. To do so, note that \(Q^n_i\), which is further bounded by the queue length of an \(M_i/GI/\infty\) infinite-server model, having arrival rate \(\lambda^n_i(t)\) and service hazard rate \(\hat{h}_i(x) = \min\{h_i(x), \mu_i\}\). Denote its queue length by \(X^n_\infty(t)\). We have \(\hat{Q}^n(t) \leq_{st} \hat{X}^n_\infty(t)\). Because \(X^n_\infty(t)\) is a Poisson r.v., the u.i. of \(\hat{X}^n_\infty(t)\) is straightforward. Specifically, we have

\[
\sup_n \mathbb{E}[(\hat{X}^n_\infty(t))^2] = \sup_n \left[ \frac{\int_0^t \lambda^n_i(t-x)G_i(x)dx}{n} + \left( \frac{\int_0^t \lambda^n_i(t-x)G_i(x)dx}{n} \right)^2 \right] < \infty, \tag{EC.55}
\]

where \(G_i\) is the CDF having hazard rate \(\hat{h}_i\). See Proposition A.2.2 in Ethier and Kurtz (1986).

Next, we write the PWT

\[
V^n_i(t) = \sum_{j=0}^{Q^n_i(t)} U_j,
\]

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

\[
\mathbb{E}[V^n_i(t)^2] = \left[ \sum_{j=0}^{Q^n_i(t)} (U_j)^2 + \sum_{j \neq k} U_j U_k \right] \leq (\mathbb{E}[Q^n_i(t)] + 1) \left( \frac{(\ell^{(1)} + 1)^2 + \ell^{(2)}}{(nm\bar{\mu})^2} + \mathbb{E}[Q^n_i(t)^2 + Q^n_i(t)]\right) \left( \frac{\ell^{(1)} + 1}{(nm\bar{\mu})^2} \right)
\]

where \(\bar{\mu} \equiv \min_{1 \leq i \leq K} \mu_i\). Using the bound in (EC.55), we have \(\sup_n \mathbb{E}[V^n_i(t)^2] < \infty\), which implies u.i. of \(V^n_i\).

The u.i. of \(H^n_i\) is straightforward because \(0 \leq H^n_i(t) \leq T + w_i\).

The TPoD for class-\(i\) customers

\[
P(V^n_i(t) > w_i) = P(\sqrt{n}(V^n_i(t) - w_i) > 0) = P(\hat{V}^n_i(t) > 0)
\]

\[
\rightarrow P(\hat{V}_i(t) > 0) = P(\hat{V}_i(t) - \bar{\kappa}(t + w_i) > 0)
\]

\[
= P\left( \frac{\hat{H}(t + w_i) - \kappa_i(t + w_i)}{\hat{\sigma}(t + w_i)} \right) = P(Z > \frac{\kappa_i(t + w_i)}{\hat{\sigma}(t + w_i)}) = P(Z > z_{\alpha_i}) = \alpha_i,
\]

where the third equality holds by (23). \(\Box\)
EC.1.5. Proof of Corollary 2.

Because the functions \( L(t, s) \) and \( J(t, s) \) are now separable in \( t \) and \( s \), SDE (27) becomes

\[
\hat{H}(t) = \frac{1}{R(t)} \int_0^t \hat{L}(s) \hat{H}(s) \mathrm{d}s + \frac{1}{R(t)} \int_0^t \hat{J}(s) \mathrm{d}W(s) + K(t),
\]

(EC.56)

where \( R(t), \hat{L}(t) \) and \( \hat{J}(t) \) are specified in Proposition 1. Multiplying \( R(t) \) on both sides and differentiating (EC.56) yields

\[
\frac{R'(t) - \hat{L}(t)}{R(t)} \hat{H}(t) \mathrm{d}t + \frac{R'(t) - \hat{J}(t)}{R(t)} \mathrm{d}W(t) = \frac{\hat{J}(t)}{R(t)} \mathrm{d}W(t) + K'(t) \mathrm{d}t + \frac{K(t)R'(t)}{R(t)} \mathrm{d}t.
\]

Multiplying \( e^{\int_0^t \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} \) on both sides and integrating from 0 to \( t \) yields

\[
e^{\int_0^t \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} \hat{H}(t) = \int_0^t e^{\int_0^u \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} \frac{\hat{J}(u)}{R(u)} \mathrm{d}W(u) + \int_0^t e^{\int_0^u \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} K(u) \mathrm{d}u.
\]

or equivalently

\[
\hat{H}(t) = \int_0^t e^{-\int_u^t \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} \frac{\hat{J}(u)}{R(u)} \mathrm{d}W(u) + \int_0^t e^{-\int_u^t \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} K(u) \frac{R'(u)}{R(u)} \mathrm{d}u.
\]

(EC.57)

Note that

\[
e^{-\int_u^t \frac{R'(r) - \hat{L}(r)}{R(r)} \mathrm{d}v} = e^{\log R(u) - \log R(t)} e^{\int_u^t \frac{\hat{L}(r)}{R(r)} \mathrm{d}v} = \frac{R(u)}{R(t)} e^{\int_u^t \frac{\hat{L}(r)}{R(r)} \mathrm{d}v}.
\]

(EC.58)

Combining (EC.57) and (EC.58) yields the solution in (39). The variance formula in Corollary 2 easily follows from the isometry of the Brownian integral.

\[\square\]

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 \( t_i = i \Delta T, 1 \leq i \leq N, T = 24 \), 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 ($1 \leq j \leq N$) 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 $t^{th}$ run: $V_i^j(j\Delta T) = t - j\Delta T$. 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 $V_i^j(j\Delta T)$ and indicators $1_{\{V_i^j(j\Delta T) > w_i\}}$.

EC.2.2. Additional numerical examples

EC.2.2.1. Class-dependent service rates  We modify the two-class base model in §4.1 by setting $\mu_1 = 0.5$ and $\mu_2 = 1$. In this case we numerically compute the variance of $\hat{H}(t)$ and required control functions using our contraction based algorithm given in Remark EC.2. It our algorithm 42 iterations to converge with an error tolerance $\epsilon = 10^{-6}$. Similar to the case of class-independent service rates, TV-SRS and TV-DPS continue to achieve good TPoD performance. See Figure EC.2 for the simulation results.

EC.2.2.2. Mixed arrival rates  We test the case where arrival rates are of different orders of magnitude. We modify the arrival rates in our two-class base case 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.1 that, even though the majority of arrivals to the system are from Class 1, we have effective TPoD stabilization for both classes.

![Figure EC.1](image-url) Simulation comparison for a two-class model with mixed arrival rates: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $P(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $\bar{\lambda}_1 = 0.1, \bar{\lambda}_2 = 1.5, n = 100, w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8$. 
EC.2.2.3. Other examples  In the longer online appendix Liu et al. (2018b), we conduct additional simulation experiments, including inflexible staffing with a bigger $\Delta_s$, higher QoS targets, and relaxed initial condition (with customers starting to arrive at the same time). We also consider the case of staffing and scheduling to achieve class-differentiated mean PWT (rather than the TPoD).

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