Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Scheduling to Differentiate Service in a Multiclass Service System

Yunan Liu, Xu Sun, Kyle Hovey

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–18
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Scheduling to Differentiate Service in a Multiclass Service System

Yunan Liu,[a] Xu Sun,[b] Kyle Hovey[c]

[a] Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695; [b] Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida 32603; [c] U.S. Department of Defense, Washington, District of Columbia 20301

Contact: yliu48@ncsu.edu, https://orcid.org/0000-0001-9961-2610 (YL); xusun@ufl.edu, https://orcid.org/0000-0003-2560-7370 (XS); kahovey@ncsu.edu (KH)

Received: January 3, 2019
Revised: April 29, 2020
Accepted: August 5, 2020
Published Online in Articles in Advance: March 8, 2021

Subject Classifications: queues: nonstationary, transient results; queues: approximations, balking and reneging, limit theorems
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2020.2075

Abstract. Motivated by large-scale service systems, we study a multiclass queueing system having class-dependent service rates and heterogeneous abandonment distributions. Our objective is to devise proper staffing and scheduling schemes to achieve differentiated services for each class. Formally, for a class-specific delay target $w_i > 0$ and threshold $\alpha_i \in (0,1)$, we concurrently determine an appropriate staffing level (number of servers) and a server-assignment rule (assigning newly idle servers to a waiting customer from one of the classes), under which the percentage of class-i customers waiting more than w_i does not exceed α_i. We tackle the problem under the efficiency-driven many-server heavy-traffic limiting regime, where both the demand volume and the number of servers grow proportionally to infinity. Our main findings are as follows: (a) class-level service differentiation is obtained by using a delay-based dynamic prioritization scheme; (b) the proposed scheduling rule achieves an important state-space collapse, in which all waiting time processes evolve as fixed proportions of a one-dimensional state-descriptor called the *frontier process*; (c) the frontier process solves a stochastic Volterra equation and is thus a non-Markovian process; (d) the proposed staffing-and-scheduling solution can be readily extended to time-varying settings. In this paper, we establish heavy-traffic limit theorems to show that our solution is asymptotically correct for large systems, and we numerically demonstrate that it performs reasonably well even for relatively small systems.

Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2020.2075.

Keywords: dynamic scheduling • dynamic prioritization • time-varying staffing • efficiency-driven • heavy-traffic approximations • service differentiation • tail probability of delay

1. Introduction

In this paper, we study the problem of achieving differentiated service for a service system in which customers of K different classes (each having its own dedicated queue) are served by a pool of statistically identical servers. The problem is traditionally formulated as a stochastic optimization problem, where the objective is to minimize the staffing level subject to a set of prescribed performance targets. In this work, we are especially interested in satisfying the following quality-of-service (QoS) constraints:

$$\mathbb{P}(V_i > w_i) \le \alpha_i, \quad 1 \le i \le K, \tag{1}$$

for class-specific delay target w_i and tail-probability target $\alpha_i \in (0,1)$, $1 \le i \le K$, where V_i denotes the waiting time of an arbitrary class-i customer. In words, the set of constraints requires that a class-i customer waits longer than w_i time units with a probability no greater than α_i. We refer to the left-hand side of (1) as the *tail probability of delay* (TPoD). The input of the optimization problem comprises the staffing (salary) costs, the QoS metrics, and various system parameters such as arrival rates, service times, and customers' patience-time distribution. The output of the problem specifies a proper staffing level and a server-assignment rule that dictates how to pair a newly available server with a customer when there are customers from more than one class waiting.

Ideally, one 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, due to the complexity of the model, this optimal staffing problem can be extremely difficult to solve; the solution also 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 across all customer classes. Loosely speaking, we seek an appropriate staffing level and a scheduling policy under which

$$\mathbb{P}(V_i > w_i) \approx \alpha_i, \quad 1 \le i \le K.$$

Hereinafter, we refer to this problem as the *service differentiation problem*.

Applied Relevance of TPoD-Based Metrics. TPoD-based QoS metrics have been widely used in service systems where the manager 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 today's multimedia (or omni-channel) contact centers, wherein one looks at the service level not just for phone calls, but for email, live chat, and social media channels. Whereas there is a longstanding tradition for contact centers to target the service level of answering 80% of calls within 20 seconds, different QoS targets might be available for other channels. Indeed, a collection of insights shared by call-center management experts seems to suggest the following rule of thumb: answer 80% of live chats in 40 seconds, 95% of emails within four hours, and 80% of social media posts within 20 minutes (see Preece et al. 2018). TPoD is also widely used in health care. For instance, inpatient wards of Singapore hospitals strive to keep delays below six hours; the probability that delay is below six hours, referred to as the *six-hour service level*, is closely monitored and found to vary between 4% and 37% over time (see Shi et al. 2016). Another relevant 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. 2019, p. 724), "CTAS level i patients need to be seen by a physician *for the first time* within w_i minutes $100\alpha_i$% of the time," with $(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120)$ and $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8)$. Whereas the model considered here (see Figure 2) might not fully capture complicated dynamics of and procedures in a hospital ED, we feel that the methodological framework developed in this paper could inform decision making on effective allocation of medical resources needed for the initial treatment of ED patients.

In addition to call-center and hospital settings, our modeling framework and proposed solutions may be applied to other service systems that share similar features, such as immigration offices in which the employees have to select cases to expedite in the face of a large backlog of applications of different preference levels. In summary, we believe that our framework provides a useful tool to understand how scarce service resources should be allocated in systems whose service strategies are driven either by revenue or less tangible aspects such as social welfare.

The multiclass, multiserver queueing system considered in this paper captures *three* salient features of real-world service systems. First, service requirements are class-specific. This assumption makes our model especially appealing in practical settings. For instance, banking call centers receive requests as simple as balance inquiries and as complex as dealing with fraudulent activities. Whereas the former can be handled relatively quickly, the latter tends to be more difficult to handle and thus requires longer services. Class-dependent services also arise in health-care settings wherein the service time of patients with more severe symptoms can be twice as much as that of those showing milder symptoms (see, e.g., table 2 of Ding et al. 2019). Second, in addition to the incorporation of customer abandonment, we allow reneging behavior of customers to be class-dependent as well. In the call-center context, this feature reflects the reality that callers may hang up due to prolonged waiting times and that callers who are calling about fraud may be more patient than those who are calling about balance inquiries. Third, different customer classes could have different QoS requirements. In particular, we allow both the delay target and the associated tail-probability target to be class-specific. It is worth mentioning that any one of these features would prevent treating the customers as one superclass.

Despite its theoretical and practical importance, very few studies dealing with *many-server* service systems have chosen to work with class-dependent service rates. As noted by Gurvich et al. (2008), "If one assumes class-dependent service rates, the resulting problem is much more difficult"; indeed, the critical-loading condition gives rise to a multidimensional piecewise-linear diffusion that is not amenable to analysis. Similar remarks were made by, for example, Kim et al. (2018), who argued that "the class-dependent service rates case involves a complicated partial differential equation, which is not amenable to solving directly".

1.1. Our Contributions

1.1.1. A Different Scaling to Analytically Treat Class-Dependent Services with TPoD Targets.

To our best knowledge, our paper is the first to provide an analytic treatment of *class-dependent services* with TPoD targets. We first develop tractable asymptotic limits for the queueing model with class-dependent service rates (Theorem 1) and then apply our new results to obtain analytic formulas for our control parameters (Theorems 2 and 4). Such analytic limits require a different asymptotic scaling: more precisely speaking, we choose *not* to scale down the delay targets, as was commonly done in the literature. Under this new scaling condition, our proposed solution gives rise to novel limits and exhibits nice properties inherited from the staffing problem for a single-class queue, albeit for services that are class-dependent. Indeed, under our control policy, the system dynamics can be fully described (asymptotically) by a multidimensional

Ornstein-Uhlenbeck-type process that further translates to a one-dimensional state-descriptor (see Theorem 1 and its proof). Such a dimensional reduction seems impossible for critical loading systems due to the limit being a multidimensional piecewise-linear diffusion.

Besides facilitating the analysis of class-dependent services, this scaling may be more suitable to systems where customer waiting times are comparable to (or longer than) service times. For example, many call centers are found to be severely understaffed (see, e.g., figures 4–5 of Huang et al. 2017). Also, service times in hospital EDs tend to be (somewhat surprisingly) small—approximately six minutes according to Yom-Tov and Mandelbaum (2014)—whereas the median waiting time to receive the initial treatment is estimated to about 30 minutes.

1.1.2. A New Scheduling Rule. Our point of departure in addressing the service differentiation problem is to propose a novel server-assignment rule based on customers' elapsed delays. For models with constant arrival rates, the service scheduling policy reads as follows: let $H_i(t)$ denote the real-time delay of the head-of-line customer in queue i at time t; then our scheduling rule always assigns the next available server to the head-of-line customer in queue i^*, with i^* satisfying

$$i^* \in \arg\max_{1\le i\le K}\left\{\frac{H_i(t)}{w_i} + \frac{1}{\sqrt{\lambda}}\ \kappa_i\right\}, \tag{2}$$

where λ denotes the aggregated arrival rate and κ_i are K control parameters (yet to be determined). Note that the allocation scheme specified by (2) exhibits a *separation of scales*, which involves two prioritization regulators: "first-order" term $H_i(t)/w_i$ and "second-order" term $\kappa_i/\sqrt{\lambda}$. The advantage of our new prioritization rule (2) is that not only can it hold the class-i delay around its delay target w_i (controlled by the first-order term $H_i(t)/w_i$), it can also differentiate the probability of class-i delay exceeding w_i (i.e., class-i TPoD) according to the desired class-specific probability target α_i (controlled by the second-order term $\kappa_i/\sqrt{\lambda}$). This stands in stark contrast to Gurvich and Whitt (2010) and Sun and Whitt (2018), wherein the identical probability target α is considered across all classes.

1.1.3. A New Scaling Limit. At the heart of our solution approach is a functional central limit theorem (FCLT) for various quantities of interest and an important state-space-collapse (SSC) result showing that, under the new scheduling policy (paired with a version of the square-root staffing rule), all waiting-time processes reduce to a simple functional of a one-dimensional process, henceforth referred to as the *frontier process*. In contrast to the existing literature wherein standard diffusion processes, such as Ornstein-Uhlenbeck (OU) and piecewise-linear diffusion processes, often arise as the scaling limit, our frontier process $\hat{H}$ is a *non-Markovian* Gaussian process. Specifically, $\hat{H}$ uniquely solves a *stochastic Volterra equation* (SVE):

$$\begin{aligned}\hat{H}(t) = &\int_0^t L(t,s)\hat{H}(s)\mathrm{d}s + \int_0^t J(t,s)\mathrm{d}\mathcal{W}(s)\\ &+ K(t),\end{aligned} \tag{3}$$

where $\mathcal{W}$ is standard Brownian motion and $K(\cdot)$, $L(\cdot,\cdot)$, and $J(\cdot,\cdot)$ are all simple analytic functions of the model inputs (see Theorem 1 for detailed formulas). We remark that the non-Markovian probability structure of the process limit $\hat{H}$ is attributed to our practical assumption of the class-dependent service rates; indeed, we show that such non-Markovian limit will degenerate to an OU process when service rates are equal across all customer classes. Although an SVE admits no analytic solutions in general, we are able to provide some useful analysis: (i) we establish conditions that guarantee the existence and uniqueness of the solution to SVE (3) and show that it is Gaussian; (ii) we develop efficient iterative algorithms to compute the first two statistical moments of $\hat{H}$, which are essential to obtaining the desired control parameters for our staffing and scheduling scheme; and (iii) we give closed-form expressions for $\hat{H}$ in some special cases.

To our knowledge, this seems to be the first appearance of an SVE-type limit in the heavy-traffic queueing theory literature, and our analysis may serve as useful building blocks for future research on multiclass queueing systems.

1.1.4. Extension to Time-Varying Arrivals. In various settings, it is reasonable to assume that the input stream of customers to the system during a given time horizon follows a nonhomogeneous Poisson process. Of course, if the arrival rates are time-varying, then the associated staffing and scheduling solution ought to be time-dependent as well. Indeed, to handle systems with strongly time-varying arrival rates, we advance a time-varying version of the proposed staffing and scheduling scheme that involves $K+1$ control functions. By establishing heavy-traffic stochastic-process limits for the waiting-time processes, we identify a set of "asymptotically correct" control functions with which the intended service differentiation can be achieved over any finite time horizon (see Theorem 4). The computation of these control functions relies on the first and second moments of the frontier process for which we develop an efficient iterative algorithm (see Remark 6). Figure 1 illustrates the entire procedure.

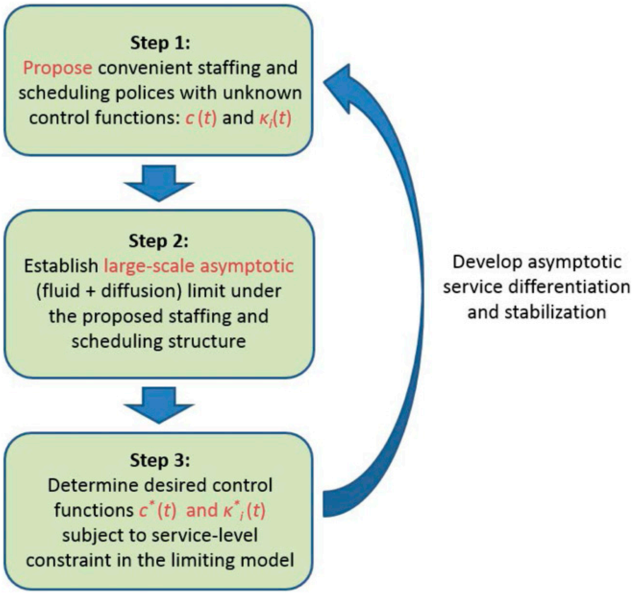

Step 1:
Propose convenient staffing and scheduling polices with unknown control functions: $c(t)$ and $\kappa_i(t)$
Step 2:
Establish large-scale asymptotic (fluid + diffusion) limit under the proposed staffing and scheduling structure
Step 3:
Determine desired control functions $c^*(t)$ and $\kappa^*_i(t)$ subject to service-level constraint in the limiting model
Develop asymptotic service differentiation and stabilization

Figure 1. (Color online) A Road Map

1.2. Paper Organization

In Section 2, we review the related literature. In Section 3, we describe the queueing model and introduce a class of new staffing and scheduling policies. In Section 4, we introduce our asymptotic framework and present the limit theorems that help us pin down various control parameters to achieve intended service differentiation. In Section 5, we extend our framework substantially to deal with systems with strongly time-varying arrival rates by plugging a set of control (design) functions into the staffing and scheduling policy. We provide a semiclosed form solution for the desired control functions by establishing heavy-traffic limits for the waiting-time processes. In Section 6, we report numerical examples. Finally, we make concluding remarks in Section 7. The e-companion presents all the technical proofs omitted from the main paper. Additional supplementary results are given in a longer online appendix (Liu et al. 2018).

2. Related Literature

Here we primarily focus on two lines of research that are most relevant for the present study. The first stream focuses on the asymptotically optimal policies under various objective functions. The other stream is more recent and deals with constrained staffing problems with service-level requirements.

The problem of staffing and scheduling multiclass queueing systems is notoriously difficult, even for simple Markov models. Using the conventional heavy-traffic scaling, Van Mieghem (1995) showed a $c\mu$-type index rule to be asymptotically optimal. Mandelbaum and Stolyar (2004) advanced a generalized $c\mu$-type policy for parallel-server processing networks and established asymptotic optimality of the proposed policy. Extensions of these index rules to allow customer abandonment were considered by Ata and Tongarlak (2013) and Kim and Ward (2013). Moving from conventional heavy-traffic to many-server heavy-traffic approximations, Atar et al. (2010, 2011) extended the $c\mu$-type rule to a system with customer abandonment modeled as a multiclass overloaded $M/M/s+M$ queue. Also under the overloaded regime, Puha and Ward (2019) considered static priority scheduling for a multiclass $G/GI/N+GI$ fluid queue. However, the optimality of these greedy policies does not extend in general to critically-loaded systems. Instead, dynamic prioritization schemes were proposed by Harrison and Zeevi (2004) and Atar et al. (2004), who applied a diffusion control framework and obtained asymptotically policies via the solution to a partial differential equation. More recently, Kim et al. (2018) incorporated the customer patience-time distribution into an optimal scheduling problem and developed a near-optimal policy that can be implemented by customer contact centers to further improve performance metrics. However, because the specification of the aforementioned policies requires numerically solving a Bellman equation, it is difficult to see how sensitive the control policies are to the changes in the system parameters. As noted by Gans et al. (2003), this asymptotic analysis “has so far shed little qualitative light” on the structure of optimal controls in the many-server critically-loaded regime. This in part motivates a complementary stream of research that we will survey.

The formulation of our problem follows more closely the constraint-satisfaction framework as adopted by Gurvich et al. (2008) and Gurvich and Whitt (2010) (see also Soh and Gurvich 2016). Based on a simple idle-server-based priority control, Gurvich et al. (2008) developed a joint staffing and scheduling control and established the asymptotic optimality of the proposed policy. By focusing on ratio scheduling and routing policies, Gurvich and Whitt (2010) sought “good and simple” policies and established the SSC associated with the heavy-traffic limit showing that the ratio rules are asymptotically optimal. It is important to emphasize that, in contrast to Gurvich et al. (2008) and Gurvich and Whitt (2010), we allow service times to be class-dependent. (The “pool-dependent service rates” condition in Gurvich and Whitt 2010 essentially assumes service rates to be class-independent under their fast-pool-first routing rule.) More recently, 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 the authors were not able to address class-dependent services either—explicit staffing and scheduling formulas only exist for the special case where the service rate equals the abandonment rate. In addition, they restricted attention to an identical TPoD target α, whereas we allow the targets α_i to be *class-specific*.

Finally, the extension of our approach to the time-varying setting is related to the vast literature on the staffing problems of single-class queuing systems with time-varying arrivals. We hereby only review time-varying staffing methods that are closely related to the present paper (namely, those based on many-server limits and offered-load analysis). Jennings et al. (1996) developed a modified-offered-load approximation approach for stabilizing the probability of delays. For more general models with abandonments, Feldman et al. (2008) devised a simulation-based iterative staffing algorithm to achieve the intended performance target. Focusing on overloaded regimes, Liu and Whitt (2012) showed that controlling probability of abandonment is equivalent to controlling mean queueing delay. More recently, Liu (2018) developed an analytic staffing method to stabilize the TPoD and proved asymptotic correctness of the staffing algorithm; for this reason, the present study may be seen as a multiclass extension to the model of Liu (2018) by allowing each class to have its own TPoD target. Such an extension, as previously alluded to, is not straightforward and introduces major hurdles that lead to nontrivial technical contributions.

3. Problem Formulation and Proposed Solutions

We describe the multiclass queueing model in Section 3.1. We introduce our staffing and dynamic scheduling rules in Sections 3.2 and 3.3.

3.1. A Multiclass V Model

Consider a V-model having $K \geq 2$ customer queues served by one common service pool. Customers arrive to the ith queue according to a homogeneous Poisson process A_i with rate λ_i. For mathematical convenience, we assume that initially there are customers in queue—thus all servers are busy at time zero. In addition, the elapsed waiting time of the "oldest" customer does not exceed θ, where θ is some positive constant. This is easily satisfied as long as all customers arrived in the finite past.

We assume that class-i service times are *independent and identically distributed* (i.i.d.) random variables following an exponential distribution with class-dependent service rate μ_i. Class-i customers may choose to abandon from the ith queue according to i.i.d. abandonment times following a general distribution, with *cumulative distribution function* (CDF) $F_i(x)$, complementary CDF $F_i^c(x) \equiv 1 - F_i(x)$, *probability density function* (PDF) $f_i(x)$, and hazard rate $h_{F_i}(x) \equiv f_i(x)/F^c(x)$. We assume that service times and patience times are mutually independent, independent of the arrival processes. In addition, we will assume that $F_i^c(x) > 0$ on any compact interval.

The system adopts a work-conserving policy; that is, no customers wait in queue if there is an available server. In addition, a *first-come first-served* (FCFS) queueing discipline is implemented within each customer class. Figure 2 gives a graphical illustration of a three-class system. Let $Q_i(t)$ represent the number of customers waiting in the ith 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 ith queue, respectively, up to time t. By flow conservation,

$$Q_i(t) = Q_i(0) + A_i(t) - E_i(t) - R_i(t). \tag{4}$$

Let $B_i(t)$ be the number of busy servers currently serving class-i customers at time t, and let $D_i(t)$ 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

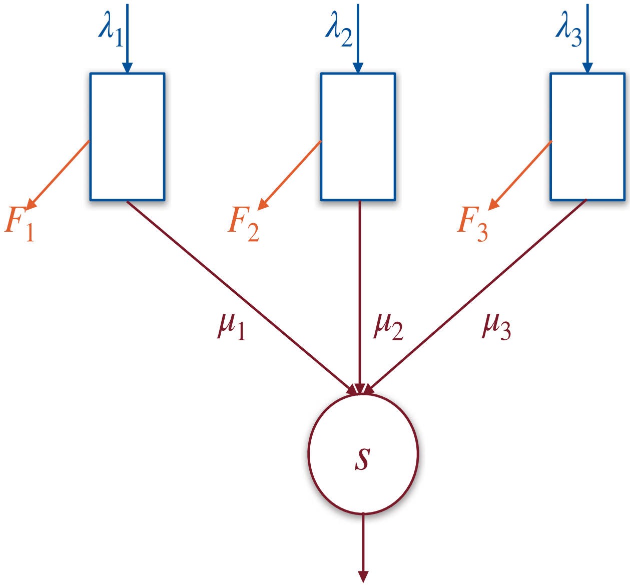
$$B_i(t) = B_i(0) + E_i(t) - D_i(t). \tag{5}$$

Finally, let $X_i(t)$ denote the total number of class-i customers in the system at time t. Adding up (4) and (5) yields

$$\begin{aligned} X_i(t) &= Q_i(t) + B_i(t) \\ &= X_i(0) + A_i(t) - D_i(t) - R_i(t). \end{aligned} \tag{6}$$

Alternatively, one can derive (6) directly from flow conservation.

Figure 2. (Color online) A Multiclass V Model with Class-Dependent Services and Abandonments



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 ith queue, that is, 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 ith queue. Let $V_i(t)$ represent the class-i *potential waiting time* (PWT) at time t, that is, the waiting time of a potential class-i customer arriving at time t who has infinite patience. Following Aras et al. (2018), Liu and Whitt (2014b), we can conveniently express the enter-service process and the queue-length process for each customer class in the following way:

$$E_i(t) = \sum_{k=1}^{A_i(t-H_i(t))} \mathbf{1}_{\{\gamma_{i,k} > V_i(\xi_{i,k})\}}, \tag{7}$$

$$Q_i(t) = \sum_{k=A_i(t-H_i(t))}^{A_i(t)} \mathbf{1}_{\{\xi_{i,k}+\gamma_{i,k} > t\}}, \tag{8}$$

where $\mathbf{1}_A$ denotes the indicator function of event (set) A, the random variables $-w_i \le \xi_{i,1} < \xi_{i,2} < \cdots$ 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 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).

3.2. The Staffing Rule

We start by introducing a version of the square-root staffing rule, which consists of two terms: (i) the offered-load (first-order term) and (ii) the safety staffing level (second-order term).

The Offered-Load Staffing Term. Here we adopt the *offered-load analysis*, which estimates the required service capacity by estimating how much capacity would be used if there were not a limit on its availability. For example, consider a single-class $M/GI/s + GI$ model having Poisson arrivals with rate λ, 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 $M/GI/s + GI$ model is complicated, the corresponding $M/GI/\infty$ infinite-server model remains remarkably tractable, where the number of customers or busy servers in steady state follows a Poisson distribution with mean

$$m_\infty \equiv \mathbb{E}[X_\infty] = \lambda \int_{-\infty}^{t} G^c(t-u)\mathrm{d}u = \lambda \mathbb{E}[S], \tag{9}$$

where S denotes the service time in the $M/GI/\infty$ queue. If the objective is to stabilize the expected delay at a target w, then one will need to set the staffing levels to a modified version of (9), namely,

$$\begin{aligned} m_{\mathrm{DIS}} &\equiv \underbrace{F^c(w)\lambda}_{\text{effective arrival rate}} \times \int_{-\infty}^{t} G^c(t-u)\mathrm{d}u \\ &= F^c(w)\lambda\mathbb{E}[S], \end{aligned} \tag{10}$$

where we have used DIS to denote the "delayed-infinite-server approximation," as in Liu and Whitt (2012) (see also Liu and Whitt 2014a, 2017 for additional analysis on DIS). 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 $m_{\mathrm{DIS}}(t)$ 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 aforementioned offered-load analysis by choosing the first-order staffing level as

$$m \equiv \sum_{i=1}^{K} m_i \text{ for } m_i \equiv F_i^c(w_i) \cdot (\lambda_i/\mu_i), \tag{11}$$

where each term in the sum of (11) is obtained by replacing (F, w, G, λ) in (10) with the class-dependent primitives $(F_i, w_i, \exp(\mu_i), \lambda_i)$.

The Safety-Staffing Term. Unfortunately, m is not effective for stabilizing class-dependent TPoDs, because m does not include the class-dependent probability targets α_i. Our strategy is to refine the staffing level by adding a second-order safety staffing term driven by the class-dependent probability targets α_i. Let $\lambda \equiv \sum_{i=1}^{K} \lambda_i$ be the aggregate demand rate. We envision a staffing formula consisting of two pieces, namely,

$$s = \left\lceil m + \sqrt{\lambda} c \right\rceil, \tag{12}$$

where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x, and $c \equiv c(\alpha_1, \ldots, \alpha_K)$ is a control parameter that depends on $(\alpha_1, \ldots, \alpha_K)$ and will be determined later.

Remark 1 (Role of the Safety-Staffing Coefficient c)**.** Note that the first-order staffing term m in (12) is on the order of λ, whereas the second-order term is on the order of $\sqrt{\lambda}$. Given that the offered-load m depends on delay target w_i, arrival rate λ_i, service rate μ_i, and patience-time distribution F_i, the remaining flexibility in the staffing formula depends entirely on the single parameter 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.

3.3. A Dynamic Prioritization Scheme

Following step 1 in Figure 1, we next introduce a delay-based scheduling rule that involves additional K control parameters. To implement such a scheduling rule, we track the elapsed waiting time of all waiting customers. Because customers of the same class are served in the FCFS scheme, 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 (2), where the first term $H_i(t)/w_i$ is the HWT scaled by the delay target and $\kappa_i \equiv \kappa(\alpha_i)$, referred to as the second-order class-i *prioritization regulator*, is an α_i-dependent control parameter to be specified later. Furthermore, we define what we call the *frontier process* as

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

Remark 2 (Understanding the Proposed Dynamic Scheduling Policy)**.** Our proposed scheduling rule exhibits an important separation of scales. To the best of our knowledge, this is a feature unique to the present study and absent from previous research. 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 nonlinear extension). The second-order term $(1/\sqrt{\lambda})\kappa_i$ helps accomplish the class-dependent probability target α_i. Intuitively, such a control parameter κ_i should satisfy the following properties:

(i) *Monotonicity.* For fixed time t, κ_i should be a decreasing function of α_i, because a bigger value of α_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 κ_i should satisfy $\kappa_i < 0$ $(\kappa_i \geq 0)$ (Benchmarking with the case $\alpha_i = 0.5$, κ_i should base on the value of α_i to adjust the priority levels by adding a positive or negative weight to $H_i(t)/w_i$). See numerical examples in Section 6 for more discussions of the structure of κ_i.

Moreover, the proposed scheduling policy is in alignment with the practice of real-world services such as 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. (2019). This makes this rule especially appealing, as the intrinsic fairness of the policy helps achieve ethical expectations set forth by the CTAS guideline. Furthermore, when $w_i = w$ and $\alpha_i = \alpha$ for all $1 \leq i \leq K$, rule (2) degenerates to the *global FCFS* scheduling policy.

In the next section, we will first establish an FCLT result under our proposed staffing-and-scheduling rule with control parameters c and κ_i (step 2 in Figure 1); using the FCLT result, we will next obtain the exact formulas of c and κ_i so that the resulting solution is asymptotically feasible with respect to the TPoD-based service-level constraints (step 3 in Figure 1).

4. Asymptotic Analysis

In this section, we present our main results. Section 4.1 gives the asymptotic framework and states the FCLT and functional weak law of large numbers (FWLLN) results for the multiclass V model operating under our control policy introduced in Sections 3.2–3.3. In Section 4.2, we utilize the FCLT results to obtain the desired control parameters κ_i and c that are expected to achieve TPoD-based service-level differentiation asymptotically. Section 4.3 provides a more detailed discussion of the important case of class-independent service rate. All proofs are given in the e-companion.

4.1. Many-Server FCLT Under the Proposed Control Policy

We consider an asymptotic framework in which the system scale (here the average arrival λ) grows to infinity. Following the convention in the literature, we will use n in place of λ as our scaling parameter. This gives rise to a sequence of K-class V models indexed by n. Let $A_i^n(t)$ be the class-i arrival process in the nth model having a rate $n\lambda_i$, where, by slight abuse of notation, we used λ_i to denote the baseline arrival rate. Our staffing rule would then satisfy

$$s^n = \lceil nm + \sqrt{n}c \rceil, \tag{13}$$

where m and c are the offered-load in (11) and safety staffing term (yet to be determined).

Let H_i^n and V_i^n be the class-i HWT and PWT in the nth model. Our scheduling rule satisfies

$$i^* \in \arg\max_{1 \leq i \leq K} \left\{ \frac{H_i^n(t)}{w_i} + \frac{1}{\sqrt{n}} \, \kappa_i \right\}, \tag{14}$$

where κ_i is a control parameter yet to be determined.

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

$$\Lambda_i(s,t) \equiv \lambda_i(t-s), \quad \bar{A}_i^n(s,t) \equiv n^{-1} A_i^n(s,t) \quad \text{and}$$
$$\hat{A}_i^n(s,t) \equiv n^{-1/2}\big(A_i^n(s,t) - n\Lambda_i(s,t)\big).$$

The sequence of processes $\bar{A}_i^n$ and $\hat{A}_i^n$ satisfy an FWLLN and FCLT, namely,

$$\big(\bar{A}_i^n(s,t), \hat{A}_i^n(s,t)\big) \Rightarrow \big(\Lambda_i(s,t), \hat{A}_i(s,t)\big) \text{ as } n \to \infty,$$

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

Remark 3 (General Arrival Processes)**.** Our main results can be easily extended to general arrival processes (which are not necessarily Poisson), as long as their CLT-scaled versions satisfy the FCLT

$$\hat{A}^n_i(s,t) \Rightarrow c_{\lambda_i} \mathcal{W}_{\lambda_i} \circ \Lambda_i(s,t) \text{ as } n \to \infty,$$

for some $c_{\lambda_i} > 0$. These types of G 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 Liu et al. 2019 and He et al. 2016 for construction and analysis of such G arrival processes). In this case, our FCLT result in Theorem 1 can be easily adjusted by simply multiplying $\mathcal{W}_{\lambda_j}$ by c_{λ_i}. For Poisson or nonhomogenous Poisson processes, $c_{\lambda_i} = 1$.

To proceed, define the CLT-scaled versions

$$\hat{B}^n_i(t) \equiv n^{-1/2}\big(B^n_i(t) - nm_i\big),$$
$$\hat{H}^n_i(t) \equiv n^{1/2}\big(H^n_i(t) - w_i\big) \text{ and}$$
$$\hat{V}^n_i(t) \equiv n^{1/2}\big(V^n_i(t) - w_i\big).$$

In addition, we define the CLT-scaled frontier process to be $\hat{H}^n(t) \equiv n^{1/2}(H^n(t) - 1)$. We are now ready to provide the FCLT for all relevant quantity of interests. In the next theorem, we focus on delay-related performance (see Theorem EC.1 in the e-companion for the FCLT of queue-length processes).

Theorem 1 (FCLT Under the Proposed Control Policy)**.** *Suppose the system operates under the proposed staffing and scheduling rule and there is an initial convergence of* $(\hat{H}^n, \hat{B}^n_1, \ldots, \hat{B}^n_K)$ *to zero at* $t = 0$.

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

$$(\hat{H}^n_1,\ldots,\hat{H}^n_K,\hat{V}^n_1\ldots,\hat{V}^n_K)\Rightarrow(\hat{H}_1,\ldots,\hat{H}_K,\hat{V}_1\ldots,\hat{V}_K) \quad \text{in } \mathcal{D}^{2K} \quad \text{as} \quad n\to\infty, \tag{15}$$

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

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

$$\hat{H}_i(t) \equiv w_i\big(\hat{H}(t) - \kappa_i\big) \quad \text{and}$$
$$\hat{V}_i(t) \equiv w_i\big(\hat{H}(t + w_i) - \kappa_i\big); \tag{16}$$

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

$$K(t) \equiv \eta^{-1}\left(\int_0^t \sum_{i=1}^K \psi_i \kappa_i e^{\mu_i(s-t)} ds - c\right),$$
$$L(t,s) \equiv \eta^{-1}\left(\sum_{i=1}^K e^{\mu_i(s-t)}\big(\eta_i\mu_i - \psi_i\big)\right),$$
$$J(t,s) \equiv \eta^{-1}\left(2\sum_{i=1}^K e^{2\mu_i(s-t)} F^c_i(w_i)\lambda_i\right)^{1/2} \quad \textit{for}$$
$$\eta_i \equiv w_i\lambda_i F^c_i(w_i), \quad \psi_i \equiv w_i\lambda_i f_i(w_i) \tag{17}$$

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

Remark 4 (Abandonment from a Subset of Classes)**.** In practice, it is possible that customer abandonment may only occur in certain classes, not all. In this case, one may consider a more general model, where the K customer classes divide into two categories, $\mathcal{I}_0$ and $\mathcal{I}_1$, such that $\mathcal{I}_0 \cup \mathcal{I}_1 = \{1, \ldots, m\}$, and only classes in the subset $\mathcal{I}_1$ have queue abandonment, while those in $\mathcal{I}_2$ are infinitely patient. This is a more general and practical framework, because it covers commonly considered special cases: if $\mathcal{I}_1 = \phi$, then all classes are infinitely patient (i.e., the no-abandonment model); if $\mathcal{I}_0 = \phi$, then all classes are impatient as treated in Theorem 1 (i.e., the abandonment model).

We refer to the aforementioned more general case as the "partial-abandonment" model; we point out that both our proofs and analytic results can be readily extended to cover the partial-abandonment case. As long as F_i and f_i satisfy the aforementioned regularity conditions for $i \in \mathcal{I}_1$, then all the results in Theorem 1 remain valid, with $F^c_j(w_j)$ and $f_j(w_j)$ in (17) being replaced with 1 and 0, respectively, for all $j \in \mathcal{I}_0$. We omit the proofs because it is similar to the proof of Theorem 1.

Theorem 1 provides the FCLT for waiting times under our proposed staffing and scheduling policy with the second-order terms c and κ_i yet to be determined. Such FCLT results will be used later to achieve the intended TPoD-based service differentiation. Part (b) of Theorem 1 gives a nice SSC result—both limiting HWT and PWT processes are deterministic functionals of the one-dimensional frontier process $\hat{H}$. The intuition behind the SSC is that all these normalized HWTs (plus the second-order prioritization regulator) in (2) do not differ much from each other under the delay-based scheduling rule. It is significant that the one-dimensional state-descriptor $\hat{H}$ under the SSC is the solution to an 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. We next provide a proof sketch of the theorem. The full proof of Theorem 1 is given in Section EC.1 of the e-companion.

Proof Sketch of Theorem 1.

Step 1: We first show that each component within the curly brackets in (14) can be made arbitrarily close to the frontier process by choosing n large enough. This is essentially an SSC result and follows from a key observation that, at any given point in time, the number of total departures required for an HoL customer to enter service under the proposed scheduling policy is of order $O(1)$.

Step 2: We then use (7) to obtain a simple relation between $\hat{H}_i^n$ and $\hat{B}_i^n$. Based on the fact that the difference between $\hat{H}_i^n(t)$ and $w_i(\hat{H}^n(t) - \kappa_i)$ 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 $(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}^n)$. This allows us to apply Gronwall's inequality to establish the stochastic boundedness of the sequence $\{(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}^n); n \in \mathbb{N}\}$, which in turn enables us to deduce the desired FWLLN results.

Step 3: By appealing to the continuous mapping with the established FWLLN, we establish the Brownian limits for the sequence $\{(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}^n)\}$. Specifically, the limiting processes $(\hat{B}_1, \ldots, \hat{B}_K, \hat{H})$ collectively satisfy the following set of OU-type stochastic integral equations:

$$\hat{B}_i(t) + \eta_i \hat{H}(t) = -\int_0^t \mu_i \hat{B}_i(u)\mathrm{d}u - \int_0^t \psi_i \hat{H}(u)\mathrm{d}u + \int_0^t \psi_i \kappa_i \mathrm{d}u + G_i(t) \tag{18}$$

$$\text{for } i = 1, \ldots, K, \text{ and } \sum_{i=1}^K \hat{B}_i(t) = c,$$

where η_i and ψ_i are specified by (17) and

$$\begin{aligned} G_i(t) &\equiv \hat{E}_{i,1}(t) + \hat{E}_{i,2}(t) - \hat{D}_i(t), \\ \hat{E}_{i,1}(t) &\equiv F_i^c(w_i)\int_{-w_i}^{t-w_i} \sqrt{\lambda_i}\mathrm{d}\mathcal{W}_{\lambda_i}(u), \\ \hat{E}_{i,2}(t) &\equiv \sqrt{F_i^c(w_i)F_i(w_i)}\int_{-w_i}^{t-w_i} \sqrt{\lambda_i}\mathrm{d}\mathcal{W}_{\theta_i}(u), \\ \hat{D}_i(t) &\equiv \int_0^t \sqrt{\mu_i m_i}\mathrm{d}\mathcal{W}_{\mu_i}(u), \end{aligned} \tag{19}$$

for $\mathcal{W}_{\lambda_i}, \mathcal{W}_{\theta_i}, \mathcal{W}_{\mu_i}$ being independent standard Brownian motions. Next, the FCLT for the HWT and PWT processes follows by the converging-together lemma with the established FCLT for the frontier process. Step 4: Note that each equation in (18) allows us to write $\hat{B}_i$ as a function of $\hat{H}$. Plugging them into the equation $\sum_{i=1}^K \hat{B}_i(t) = c$ with some algebraic simplifications yields the desired result. The sum $\sum_{i=1}^K \hat{B}_i(t)$ is a constant because the overall staffing level is a deterministic function of which its "diffusion" term is c (see (EC.14) for details). Our proof draws heavily from Aras et al. (2017, 2018), Liu and Whitt (2014b). □

Remark 5 (Separation of Variability)**.** The diffusion limits $(\hat{B}_1, \ldots, \hat{B}_K, \hat{H})$ satisfy a $(K+1)$-dimensional stochastic differential equation (SDE), and, according to (16), there are $3K$ independent Brownian motions $\mathcal{W}_{\lambda_i}, \mathcal{W}_{\theta_i}, \mathcal{W}_{\mu_i}$ stemming from the independent random sources (arrival, abandonment, and service) of all K customer classes.

The following result is a direct consequence of Theorem 1.

Corollary 1 (FWLLN)**.** *Suppose that all conditions of Theorem 1 are satisfied. Then there is a joint convergence for the LLN-scaled processes, namely,*

$$(H_1^n, \ldots, H_K^n, V_1^n, \ldots, V_K^n) \Rightarrow (w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}) \quad in \quad \mathcal{D}^{2K} \quad as \quad n \to \infty,$$

where $\mathfrak{e}$ represents the constant function of one.

As a consequence of Corollary 1, the abandonment probabilities are also stabilized, namely,

$$\mathbb{P}(\text{a class-}i\text{ customer will abandon}) \approx F_i(w_i)$$

for n large enough. We conclude this section by making several important observations about the dynamics of the limit frontier process $\hat{H}$.

Remark 6 (A Closer Look at the SVE (3))**.**

(i) *Existence and Uniqueness of Solutions.* The process defined by (3) belongs to the class of *linear* Volterra integral equations for which existence/uniqueness of solutions is a classical result (see, e.g., Ito 1979). Also, it is a Gaussian process, because the dynamics are driven by a Brownian motion—thus its mean function and covariance function completely determine all of the finite-dimensional distributions.

(ii) *Analytic solutions in special cases.* The SVE (3) in general has no analytic solution, except for some special cases. For example, if $\mu_i = h_{F_i}(w_i)$ for all $1 \le i \le K$ so that the drift term $L(t,s) = 0$, then the SVE (3) is a simple Brownian integral that admits an analytic solution. Another special case arises when L and J are separable functions in t and s, which is the case when service rates are class-independent (see Section 4.3 for discussions of this important special case).

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

(iv) *Existence of limiting distribution.* From (18), we see that $\{(\hat{B}_1, \ldots, \hat{B}_K, \hat{H})\}$ is a multidimensional OU-type process. Thus, we would expect that the limiting distribution of $\hat{H}$ exists as $t \to \infty$. As we shall demonstrate in the next section, the frontier process $\hat{H}$ does have a limiting distribution under suitable regularity conditions.

4.2. Finding the Right Control Parameters

Recall that our goal is to derive appropriate staffing and scheduling rules under which

$$\alpha_i \approx \mathbb{P}(V_i^n > w_i) = \mathbb{P}(\hat{V}_i^n > 0).$$

On the other hand, part (b) of Theorem 1 implies that

$$\mathbb{P}(\hat{V}_i(\infty) > 0) = \mathbb{P}(\hat{H}(\infty) > \kappa_i) = \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\kappa_i - \mathbb{E}[\hat{H}(\infty)]}{\sigma_{\hat{H}}(\infty)}\right),$$

where we have used $\mathcal{N}(\mu, \sigma^2)$ to denote a normal random variable with mean μ and variance σ^2, and assumed that the limiting distribution $\hat{H}(\infty)$ exists, having mean $\mathbb{E}[\hat{H}(\infty)]$ and standard deviation $\sigma_{\hat{H}}(\infty) = \sqrt{\mathrm{Var}(\hat{H}(\infty))}$.

One should expect that the limiting distribution of the waiting time process $\hat{V}_i$ in Theorem 1 would coincide with the limit of the steady-state sequence $\{\hat{V}_i^n; n \in \mathbb{N}\}$, namely,

$$\mathbb{P}(\hat{V}_i(\infty) \le x) = \lim_{n \to \infty} \mathbb{P}(\hat{V}_i^n \le x). \tag{20}$$

But this is not automatic, because an interchange of iterated limits is involved. Despite proving to be a valid approximation for many previous models studied, a formal justification for such an interchange of limits tends to be difficult and would often qualify for a separate paper. Thus, we shall have to contend ourselves with the conjecture (20) and seek control parameters κ_i and c such that $\mathbb{P}(\hat{V}_i(\infty) > 0) = \alpha_i$, in which case κ_i and c ought to satisfy

$$\kappa_i - \mathbb{E}[\hat{H}(\infty)] = z_{1-\alpha_i} \sigma_{\hat{H}}(\infty), \tag{21}$$

where z_α is the α-quantile of a standard normal random variable, that is, $z_\alpha = \mathbb{P}(\mathcal{N}(0,1) \le \alpha)$.

One obvious solution to (21) is to make $\mathbb{E}[\hat{H}(\infty)] = 0$ and then set $\kappa_i = z_{1-\alpha_i} \sigma_{\hat{H}}(\infty)$. From (3), it is straightforward to see that a necessary condition for $\hat{H}(\infty)$ to have mean zero is that $K(\infty) = 0$, which in turn requires our control parameters to satisfy the following relation: $c = \sum_i (\psi_i \kappa_i / \mu_i)$. In this case, the task of finding the right control parameters boils down to finding the stationary variance of the process $\hat{H}$. To this end, let us define

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

$$\mathcal{J}(t) \equiv \eta^{-1}\left(2\sum_{i=1}^{K} e^{-2\mu_i t} F_i^c(w_i)\lambda_i\right)^{1/2},$$

so that $L(t,s) = \mathcal{L}(t-s)$ and $J(t,s) = \mathcal{J}(t-s)$. Furthermore, let $\mathcal{L}^{(k)}$ denote the kth convolution of function $\mathcal{L}$. One can then introduce the resolvent kernel of function $\mathcal{L}$ as

$$\mathcal{R}(t) \equiv \sum_{k=1}^{\infty} \mathcal{L}^{(k)}(t). \tag{22}$$

We next give an analytic expression for the frontier process.

Theorem 2 (Solution to the SVE)**.** *For the resolvent kernel $\mathcal{R}$ specified by (22), the following results hold:*

(i) The resolvent kernel $\mathcal{R}$ is integrable over the positive real line if and only if $\psi_j > 0$ for some $j \in \{1, \ldots, K\}$. In particular, $\mathcal{R}$ is uniformly bounded under the stated condition. In addition, the resolvent kernel $\mathcal{R}$ uniquely solves the following fixed-point equation:

$$\mathcal{R}(t) = \mathcal{L}(t) + \int_0^t \mathcal{L}(t-s)\mathcal{R}(s)\mathrm{d}s \quad \text{for} \quad t \ge 0. \tag{23}$$

(ii) The solution to (3) is given by the Brownian integral

$$\hat{H}(t) = K(t) + \int_0^t \mathcal{J}(t-s)\mathrm{d}\mathcal{W}(s) + \int_0^t \mathcal{R}(t-s)K(s)\mathrm{d}s + \int_0^t \mathcal{R}(t-s) \times \int_0^s \mathcal{J}(s-u)\mathrm{d}\mathcal{W}(u)\mathrm{d}s. \tag{24}$$

(iii) The expectation and variance of $\hat{H}(t)$ are given by

$$\mathbb{E}[\hat{H}(t)] = K(t) + \int_0^t \mathcal{R}(t-s)K(s)\mathrm{d}s = K(t) + \int_0^t \mathcal{R}(s)K(t-s)\mathrm{d}s, \tag{25}$$

$$\mathrm{Var}(\hat{H}(t)) = \int_0^t \left(\mathcal{J}(u) + \int_0^u \mathcal{R}(s)\mathcal{J}(u-s)\mathrm{d}s\right)^2 \mathrm{d}u. \tag{26}$$

(iv) If the condition specified in part (i) is satisfied, then the limiting distribution of $\hat{H}$ exists as $t \to \infty$. In particular, the limiting variance can be written as

$$\mathrm{Var}(\hat{H}(\infty)) = \lim_{t \to \infty} \mathrm{Var}(\hat{H}(t)) = \int_0^\infty \left(\mathcal{J}(u) + \int_0^u \mathcal{R}(u-s)\mathcal{J}(s)\mathrm{d}s\right)^2 \mathrm{d}u. \tag{27}$$

(*v*) *If, in addition, the control parameters satisfy the relation* $c = \sum_i(\psi_i\kappa_i/\mu_i)$*, then the mean of the limiting distribution of* $\hat{H}$ *is zero, namely,* $\mathbb{E}[\hat{H}(\infty)] = 0$.

The fixed-point equation in part (i) of Theorem 2 leads to a natural iterative algorithm for computing $\mathcal{R}$ over any compact interval $[0,T]$. For practical implementation, one might choose to run the algorithm on a (sufficiently) long time horizon $[0,T]$, and use the value of $\mathrm{Var}(\hat{H}(T))$ as a reasonable approximation for $\mathrm{Var}(\hat{H}(\infty))$. From Theorem 2, we see that the limiting variance of the $\hat{H}$ can be calculated via the formula provided in (27). By setting $\sigma_{\hat{H}}(\infty) = \sqrt{\mathrm{Var}(\hat{H}(\infty))}$ and $\kappa_i = z_{1-\alpha_i}\sigma_{\hat{H}}(\infty)$, we achieve what we set out to do.

4.3. The Case of Class-Independent Service Rate

It is well known that the case of class-dependent service rate can be more complex (see, e.g., Kim et al. 2018). In this subsection, we assume that service rates are class-independent, that is, $\mu_i = \mu$ for all $1 \le i \le K$. Under this assumption, the results greatly simplify, yielding functions L and K that are separable in t and s, so that the SVE in (3) degenerates to a much more tractable Ornstein-Uhlenbeck (OU) process. The following result formalizes this observation.

Corollary 2 (Class-Independent Services)**.** *When* $\mu_i = \mu$*, the limiting frontier process* $\hat{H}$ *is a one-dimensional OU process, namely,*

$$\eta \mathrm{d}\hat{H}(t) = -\sum_i \psi_i \hat{H}(t)\mathrm{d}t + \left(2\sum_i F_i^c(w_i)\lambda_i\right)^{1/2}\mathrm{d}\mathcal{W} + \left(\sum_i \psi_i\kappa_i - c\mu\right)\mathrm{d}t.$$

Remark 7 (SVE vs. OU)**.** The process $\widehat{H}$ that solves the SVE (3) differs from a regular OU process by having a two-parameter drift $L(t,s)$ and volatility $J(t,s)$. Essentially, it is such a characteristic that deprives the solution of the Markov property. The non-Markovian property stems from the practical assumption of class-specific service rate μ_i. We provide some intuitive explanation: if the service rates are class-independent, then the heavy-traffic limit, as noted in the literature, can be fully characterized by a one-dimensional process, because the times all customers spend in service are statistically equal. On the other hand, when service rates are class-dependent, one must resort to a K-dimensional process to track the service content in order to maintain a Markov property, but the Markov property is never completely preserved by a one-dimensional process (e.g., the frontier process).

This special case allows us to gain a richer insight into how various model parameters and QoS constraints affect our staffing and scheduling solution, as the desired control parameters now admit explicit expression, namely,

$$c^* \equiv \sum_{i=1}^{K}\frac{\psi_i\kappa_i}{\mu} \quad \text{and}$$
$$\kappa_i^* \equiv z_{1-\alpha_i}\sqrt{\frac{\sum_{j=1}^{K}\lambda_j F_j^c(w_i)}{\left(\sum_{j=1}^{K}\eta_j\right)\left(\sum_{j=1}^{K}\psi_j\right)}}. \tag{28}$$

The constants in (28) can be used to compute the required average number of servers and scheduling threshold. When $K = 1$, our staffing Equation (28) degenerates to the ED+QED staffing equations (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 scheduling and staffing, to improve the service to the next level (e.g., reducing w_i by Δw_i, or reducing α_i by $\Delta\alpha_i$) and to determine how many extra servers are needed and how much the scheduling threshold κ_i should be adjusted.

5. Extension to Time-Varying Models

So far, we have focused our attention on time-stationary arrivals. In the real world, many service systems (e.g., customer contact centers and hospital EDs) tend to face arrival rates that are significantly time-varying. Hence, it is practically relevant and theoretically interesting to develop staffing and scheduling solutions that are helpful for time-varying systems. Fortunately, as will be demonstrated in a moment, our modeling framework can conveniently accommodate time-varying customer flow with slight modifications.

To fix ideas, we assume that arrivals of each class follow a nonhomogeneous Poisson process (NHPP) with a known arrive-rate function. We find it convenient to let the class-i arrival process start 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 zero. In the face of time-varying arrivals, it makes sense to consider a transient version of the service differentiation problem, namely,

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

where T (e.g., $T = 8$ hours) denotes a finite time horizon and $V_i(t)$ 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 α_i.

Paralleling the asymptotic framework advanced in Section 4.1, we consider a sequence of models with an associated sequence of arrival processes having time-varying arrival-rate functions. More formally, let $A_i^n(t)$ be the class-i NHPP arrival process in the nth 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. Accordingly, we propose using a time-dependent staffing function

$$s^n(t) = \lceil nm(t) + \sqrt{n}c(t) \rceil, \tag{29}$$

where

$$\begin{aligned} m(t) &\equiv \sum_{i=1}^{K} m_i(t), \qquad \text{where} \\ m_i(t) &\equiv \int_0^t \underbrace{F_i^c(w_i)\lambda_i(u - w_i)}_{\text{effective class-}i\text{ arrival rate}} \; e^{-\mu_i(t-u)} \mathrm{d}u \end{aligned} \tag{30}$$

and $c(\cdot)$ is a safety staffing function yet to be determined. In a similar fashion, we consider a time-varying dynamic prioritization scheme that always assigns the next available server to the head-of-line customer in queue i^*, with i^* satisfying

$$i^* \in \arg\max_{1 \le i \le K} \left\{ \frac{H_i^n(t)}{w_i} + \frac{1}{\sqrt{n}} \; \kappa_i(t) \right\}, \tag{31}$$

where $\kappa_i(\cdot)$ are K control functions yet to be determined.

Remark 8 (Releasing Busy Servers). With possibly time-varying staffing levels, one needs to specify how to manage the system when all servers are busy yet the staffing is scheduled to decrease. There is a well-established procedure, called *server switching*, to handle this. More precisely, when a server is scheduled to depart, the customer in service is not required to stay with the server until service is complete. Instead, one allows the service in progress to be handed off to another available server. Moreover, one does not force a customer out of service if the staffing is scheduled to decrease when all are busy. Instead, one releases the first server that becomes free after the time of scheduled staffing decrease. Of course, with this assumption, the (actual) number of servers itself forms a random process, yet the impact is inconsequential (on the diffusion scale).

Our next result is a direct extension of Theorem 1 to the time-varying queueing system.

Theorem 3 (FCLT Under the Proposed Control Policy). *Suppose that the system uses the staffing rule (29) and the scheduling policy given by (31). If there is an initial convergence of $(\hat{H}^n, \hat{B}_1^n, \ldots, \hat{B}_K^n)$ to zero at $t = 0$, then we have the following:*

(a) There is a joint convergence for the CLT-scaled waiting time processes,

$$\begin{aligned} (\hat{H}_1^n, \ldots, \hat{H}_K^n, \hat{V}_1^n \ldots, \hat{V}_K^n) &\Rightarrow (\hat{H}_1, \ldots, \hat{H}_K, \hat{V}_1 \ldots, \hat{V}_K) \\ &\quad \text{in} \quad \mathcal{D}^{2K} \quad \text{as} \quad n \to \infty, \end{aligned} \tag{32}$$

where all process limits are deterministic functionals of a one-dimensional process $\hat{H}$, namely,

$$\begin{aligned} \hat{H}_i(t) &\equiv w_i\big(\hat{H}(t) - \kappa_i(t)\big) \quad \text{and} \\ \hat{V}_i(t) &= w_i\big(\hat{H}(t + w_i) - \kappa_i(t + w_i)\big); \end{aligned} \tag{33}$$

the process $\hat{H}$ uniquely solves an SVE in the form of (3) for

$$\begin{aligned} L(t,s) &\equiv \frac{\sum_{i=1}^K \eta_i(s) e^{\mu_i(s-t)} \big(\mu_i - h_{F_i}(w_i)\big)}{\eta(t)}, \\ J(t,s) &\equiv \frac{\big(\sum_{i=1}^K e^{2\mu_i(s-t)} \big(F_i^c(w_i)\lambda_i(s - w_i) \; + \; \mu_i m_i(s)\big)\big)^{1/2}}{\eta(t)}, \\ K(t) &\equiv \frac{\sum_{i=1}^K \left(\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} \big(\mu_i - h_{F_i}(w_i)\big)\kappa_i(s)\mathrm{d}s \right) - c(t)}{\eta(t)}, \end{aligned} \tag{34}$$

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

(b) In addition, $\hat{H}$ is a Gaussian process with

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

$$\begin{aligned} M_{\hat{H}} &= \Gamma(M_{\hat{H}}), \qquad \text{where} \\ \Gamma(M_{\hat{H}})(t) &\equiv \int_0^t L(t,s) M_{\hat{H}}(s)\mathrm{d}s + K(t), \end{aligned} \tag{35}$$

(ii) and covariance $C_{\hat{H}}(t,s) \equiv Cov(\hat{H}(t), \hat{H}(s))$, $0 \le s,t$, uniquely solving the FPE

$$C_{\hat{H}} = \Theta(C_{\hat{H}}),$$

where the operator Θ is defined as

$$\begin{aligned} \Theta(C_{\hat{H}})(t,s) \equiv &- \int_0^t \int_0^s L(t,u)L(s,v)C_{\hat{H}}(u,v)\mathrm{d}v\mathrm{d}u \\ &+ \int_0^t L(t,u)C_{\hat{H}}(u,s)\mathrm{d}u + \int_0^s L(s,v) \\ &\times C_{\hat{H}}(t,v)\mathrm{d}v + \int_0^{s\wedge t} J(t,u)J(s,u)\mathrm{d}u. \end{aligned} \tag{36}$$

Remark 9 (Algorithms). The operators Γ and Θ are shown to be contractions in appropriate functional spaces (see Section EC.1 in the e-companion). 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.1 in the e-companion for detailed discussions.

Given the SSC achieved by the time-varying version of the staffing and schedule rule, we now focus on investigating the one-dimensional process $\hat{H}$. When n is large, we hope to satisfy

$$\begin{aligned}\alpha_i &\equiv \mathbb{P}\big(V_i^n(t) > w_i\big) = \mathbb{P}\big(\hat{V}_i^n(t) > 0\big) \approx \mathbb{P}\big(\hat{V}_i(t) > 0\big)\\ &= \mathbb{P}\big(\hat{H}(t+w_i) - \kappa_i(t+w_i) > 0\big)\\ &= \mathbb{P}\Big(\mathcal{N}\Big(M_{\hat{H}}(t+w_i), \sigma^2_{\hat{H}}(t+w_i)\Big) > \kappa_i(t+w_i)\Big)\\ &= \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\kappa_i(t+w_i) - M_{\hat{H}}(t+w_i)}{\sigma_{\hat{H}}(t+w_i)}\right) \qquad (37)\end{aligned}$$

for all $t \geq -w_i$, where $\sigma_{\hat{H}}(t) = \sqrt{\mathrm{Var}(\hat{H}(t))} = \sqrt{C_{\hat{H}}(t,t)}$ is the standard deviation of $\hat{H}(t)$ at t. Equation (37) further simplifies to

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

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

$$\kappa_i(t) - M_{\hat{H}}(t) = z_{1-\alpha_i}\sigma_{\hat{H}}(t). \qquad (38)$$

One obvious solution to (38) is to choose $c(\cdot)$ appropriately (based on any given $\kappa_i(\cdot)$) in such a way that $K(t)$ in (34) vanishes so that $M_{\hat{H}}(t) = 0$ for all t (note that FPE (35) now has a unique solution $M_{\hat{H}}(t) = 0$ when $K(t) = 0$). This leads up to a set of desired control functions to be stated in our next theorem. The result also guarantees uniqueness of solutions to (38) among the proposed class of control rules. Let $P_{a,t}^{(n)}(i)$ denote the probability that a class-i customer will abandon in the nth model at time t.

Theorem 4 (Asymptotic Performance Stabilization)**.** *Consider a V-system operating under the staffing and scheduling control as specified by* (29) *and* (31), *respectively. Then we have the following:*

(*i*) *Condition* (38) *is satisfied if*

$$\begin{aligned}c(t) &= \sum_{i=1}^{K}\Big\{\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}\\ &\quad \times \big(\mu_i - h_{F_i}(w_i)\big)\kappa_i(s)\mathrm{d}s\Big\}, \qquad (39)\\ \kappa_i(t) &= z_{1-\alpha_i}\sigma_{\hat{H}}(t), \qquad 1 \leq i \leq K. \qquad (40)\end{aligned}$$

(*ii*) *The use of the aforementioned formulas of c and κ_i leads to the desired performance stabilization, that is,*

$$\begin{aligned}&\mathbb{P}\big(V_i^n(t) > w_i\big) \to \alpha_i \quad \textit{and}\\ &\mathbb{P}\big(H_i^n(t) > w_i\big) \to \alpha_i \quad \textit{as} \quad n \to \infty\end{aligned}$$

for $1 \leq i \leq K$. In addition,

$$(V_i^n, H_i^n) \Rightarrow (w_i\mathfrak{e}, w_i\mathfrak{e}) \quad \textit{in} \quad \mathcal{D}^2 \quad \textit{as} \quad n \to \infty \quad \textit{for} \quad i = 1, \ldots, K. \qquad (41)$$

In particular, the abandonment probabilities are also stabilized, namely,

$$P_{a,t}^{(n)}(i) \to F_i(w_i) \quad \textit{as} \quad n \to \infty.$$

(*iii*) *The safety-staffing term $c(\cdot)$ that achieves the aforementioned performance stabilization is unique; in contrast, the set of terms $(\kappa_1(\cdot), \ldots, \kappa_K(\cdot))$ are unique up to adding any common function $\Delta(\cdot)$.*

A few comments on Theorem 4 are in order. The main idea behind part (i) is to choose appropriate control functions $c(\cdot)$ and $\kappa_i(\cdot)$ to tilt the mean of the error term $\hat{V}_i^n(t)$ (rather than the mean of $V_i^n(t)$), so that asymptotically the probability mass of $\{\hat{V}_i^n(t) > 0\}$ (or $\{V_i^n(t) > w_i\}$) can be set to the desired α_i at all time t. Indeed, part (ii) of the theorem states that all constraints will be asymptotically binding with the selected control functions. Moreover, from part (iii) of Theorem 4 it follows that the staffing component of solutions to (38) is unique; that is, one cannot find another solution making all constraints binding (asymptotically) by employing fewer servers. In this respect, the resulting solution can be considered asymptotically optimal *among the specific class of control rules as considered in the present paper*. Lastly, the asserted "uniqueness" of prioritization regulators $(\kappa_1, \ldots, \kappa_K)$ is also intuitive in that applying any $\tilde{\kappa}_i(t) = \kappa_i(t) + \Delta(t)$ for $1 \leq i \leq K$ will not make a difference in our proposed scheduling rule.

Remark 10 (Structure of the Control Functions)**.** From (39), we see that the second-order safety staffing term $c(\cdot)$ depends on the second-order prioritization regulator $\kappa_i(\cdot)$, which in turn relies on α_i through z_{α_i}. Paralleling Remark 2, $\kappa_i(t)$ is decreasing in α_i, and its sign depends on how α_i compares with 0.5, that is, $\kappa_i(t) > 0$ ($\kappa_i(t) < 0$) if $\alpha_i < 0.5$ ($\alpha_i > 0.5$). 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 $\kappa_1(t) - \kappa_2(t) > 0$ is increasing in $\sigma_{\hat{H}}(t)$, which characterizes the system's overall stochastic variability (recall from Remark 6 that the variability of $\hat{H}$ captures the randomness of all events, including arrivals, service times, and abandonment times). This suggests that not only staffing levels will increase but also the prioritization scheme (scheduling rule) becomes more discriminative as the system exhibits higher levels of volatility. Finally, we emphasize that w_i (α_i) is the first-order (second-order) QoS target, because a slight change in w_i (α_i) affects the first-order (second-order) term in both (29) and (31).

Corollary 3 (Frontier Process $\hat{H}$ When Service Rates Are Class-Independent)**.** *When $\mu_i = \mu$, we have the following:*

(*i*) *The limiting frontier process* $\hat{H}$ *satisfies the one-dimensional OU process*

$$\eta(t)\hat{H}(t) = -\int_0^t \eta(u)\hat{H}(u)\mathrm{d}u + \mathcal{S}(t) + G(t), \tag{42}$$

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

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

(*ii*) *The SDE* (42) *has a unique solution*

$$\hat{H}(t) = \frac{1}{R(t)}\bigg(\int_0^t e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}\tilde{J}(u)\mathrm{d}\mathcal{W}(u) \\ + \int_0^t e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}R(u)\mathrm{d}K(u) \\ + \int_0^t e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}K(u)\mathrm{d}R(u)\bigg), \tag{43}$$

where $\mathcal{W}$ *is a standard Brownian motion, and*

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

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

$$\sigma_{\hat{H}}^2(t) \equiv \mathrm{Var}\big(\hat{H}(t)\big) = \frac{1}{R^2(t)}\int_0^t e^{2\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}\tilde{J}^2(u)\mathrm{d}u.$$

If $K = 1$, then our multiclass 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*

$$c(t) = z_{1-\alpha}e^{-\mu t}\bigg(Z(t) - \big(\mu - h_F(w)\big)\int_0^t Z(s)\mathrm{d}s\bigg), \tag{44}$$

$$\text{with} \quad Z(t) \equiv e^{(\mu - h_F(w))t} \\ \times\sqrt{\int_0^t e^{2h_F(w)}\big(F^c(w)\lambda(u-w) + \mu m(u)\big)\mathrm{d}u}. \tag{45}$$

It is easy to check that (44) and (45) coincide with the staffing equations (7) and (8) in Liu (2018), except for a time shift by w. This is due to the slightly different initial condition here.

6. Numerical Studies

In this section, we conduct extensive numerical experiments to test the effectiveness of our proposed staffing and scheduling solution. In Section 6.1, we describe in detail the architecture of the simulation, which includes the generation of virtual customers and statistical estimation methods for the tail probabilities. In Section 6.2, we consider time-stationary models of different system scales. In Section 6.3, we consider a base model having time-varying arrival rates and class-independent service rates. More numerical instances, including class-dependent service rates, mixed arrival rates, higher QoS targets, and a five-class example, are provided in the e-companion (Liu et al. 2018).

6.1. Implementation Details

All Monte Carlo simulations were conducted using MATLAB. We sample the values of the performance functions at fixed time points $\Delta T, 2\Delta T, \ldots, N\Delta T = T$, where $T = 24$ is the length of the time interval, the step size (sampling resolution) is $\Delta T = 0.01$, and $N = T/\Delta T = 2{,}400$ is the total number of samples in $[0, T]$. To collect simulated data of PWT, on each simulation run, we create frequent *virtual arrivals* at all queues with interarrival time ΔT. These virtual customers behave like real customers while in the queue and capture what the system experience would be like for customers 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 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 jth $(1 \le j \le N)$ class-i virtual customer arrives at queue i at time $j\Delta T$. If this customer is removed (from the head of the line) at time t, then the system collects a sample for the class-i PWT at time $j\Delta T$ on the lth run: $V_i^l(j\Delta T) = t - j\Delta T$. The class-$i$ mean PWT and TPoD at time $t_j \equiv j\Delta T$ are estimated by averaging n (here $n = 5{,}000$) independent copies of $V_i(j\Delta T)$ and indicators $\mathbf{1}_{\{V_i(j\Delta T) > w_i\}}$; namely, we use the unbiased Monte Carlo estimators

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

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

6.2. Time-Stationary Models

We start by looking at time-stationary settings where there are two customer classes, each having an exponential abandonment-time distribution with PDF $f_i(x) = \theta_i e^{-\theta_i x}, i = 1, 2$. System parameters include

$\lambda_1 = 1, \lambda_2 = 1.5$, $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 50$, $\theta_1 = 0.6$, and $\theta_2 = 0.3$; QoS parameters are given as $w_1 = 0.5, w_2 = 1$, $\alpha_1 = 0.2$, and $\alpha_2 = 0.8$. We calculate the desired control parameters c and κ_2 based on the formulas provided in Section 4.2. In doing so, we numerically compute the variance of $\hat{H}(T)$ for T large enough using Equation (26), and use it as an approximation for the stationary variance $\sigma\hat{H}(\infty)$ (see Equation (27)).

Because our method is based on asymptotic analysis as $n \to \infty$, an important question is how effective our proposed solution is when applied to systems of different sizes. To seek an answer to this question, we let the scale parameter n vary from 50 to 5 with everything else being fixed. Figure 3 shows the class-dependent TPoD over a finite time horizon estimated from Monto Carlo simulations for three different values of n, namely, 50, 10, and 5, as displayed in panels (a), (b), and (c), respectively. For each case, we provide the $100(1-\beta)\% = 95\%$ confidence intervals obtained from 1,000 simulation runs. From the plots, we see that, for large or moderately sized systems, our proposed staffing and scheduling scheme with properly chosen control parameters tends to perform exceptionally well in terms of its ability to achieve the intended performance stabilization. When the system size becomes even smaller, that is, $n = 5$, the approximation error tends to be more evident, which is not very surprising, given that all of our results were achieved under a many-server heavy-traffic environment.

6.3. A Two-Class Base Model with Time-Varying Arrivals

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

$$\lambda_i(t) = \bar{\lambda}_i\big(1 + r_i \sin(\gamma_i t + \phi_i)\big), \qquad 1 \le i \le K, \tag{46}$$

with average rate $\bar{\lambda}_i$, relative amplitude $|r_i| < 1$, frequency γ_i, and phase ϕ_i. We first consider a two-class V model, where class 1 and class 2 represent high- and low-priority customers, respectively. We let $\bar{\lambda}_1(t) = 1, \bar{\lambda}_2(t) = 1.5, r_1 = 0.2, r_2 = 0.3$, $\gamma_1 = \gamma_2 = 1$, $\phi_1 = 0$, $\phi_2 = -1$. (See Table EC.1 in Section EC.2.3.4 of the e-companion for the case of class-dependent γ_i.) Abandonment times follow class-dependent exponential distributions with PDF $f_i(x) = \theta_i e^{-\theta_i x}$. We let $\theta_1 = 0.6$ and $\theta_2 = 0.3$. Service rates are class-independent and standardized so that $\mu_1 = \mu_2 = 1$, with mean service time $1/\mu_i = 1$. (See Section EC.2.3.1 in the e-companion for an example of class-dependent μ_i.) 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 $w_1 = 0.5, w_2 = 1$, $\alpha_1 = 0.2, \alpha_2 = 0.8$.

In Figure 4, we calculate and plot the required control functions in a finite time interval $[0, T]$, with $T = 24$, including the offered-load function $m(t)$ in (30),

Figure 3. (Color online) Simulation Estimates of Class-Dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ for Three Systems of Different Sizes with 95% Confidence Intervals Simulated and 1,000 Independent Runs

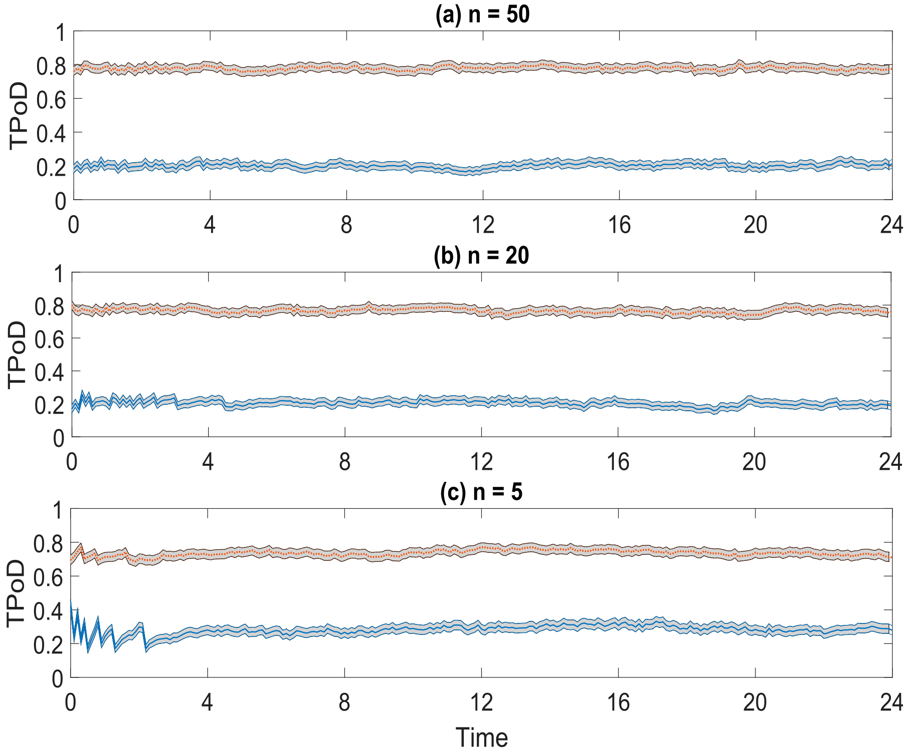
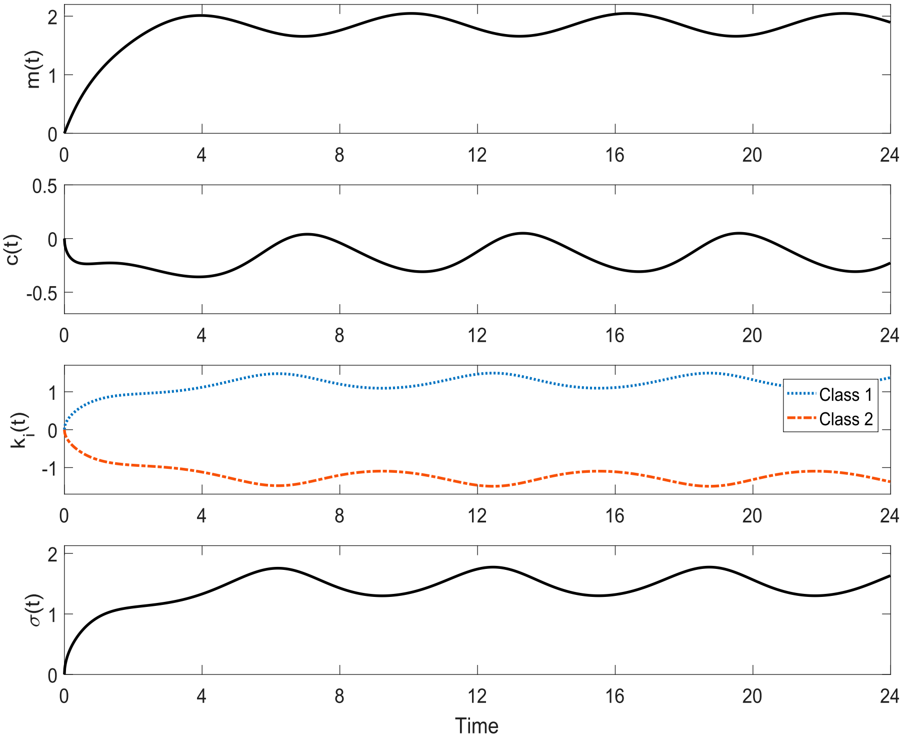


Figure 4. (Color online) Computed Control Functions for a Two-Class Base-Case Example: $m(t)$, $c(t)$, $\kappa_i(t)$, and $\sigma(t)$, $i = 1, 2$



the second-order staffing term $c(t)$ in (39), the second-order prioritization regulators (40), and the standard deviation process of $\hat{H}$ in (43). Consistent with discussions in Remarks 2 and 10, 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 4, we conduct Monte Carlo simulation experiments to test the effectiveness of our proposed solution. For our

Figure 5. (Color online) Plots of (i) Arrival Rates (Top Panel), (ii) Simulation Estimates of Class-Dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (Middle Panel), and (iii) Time-Varying Staffing Level (Bottom Panel) for the Two-Class Base-Case Example with 5,000 Independent Runs

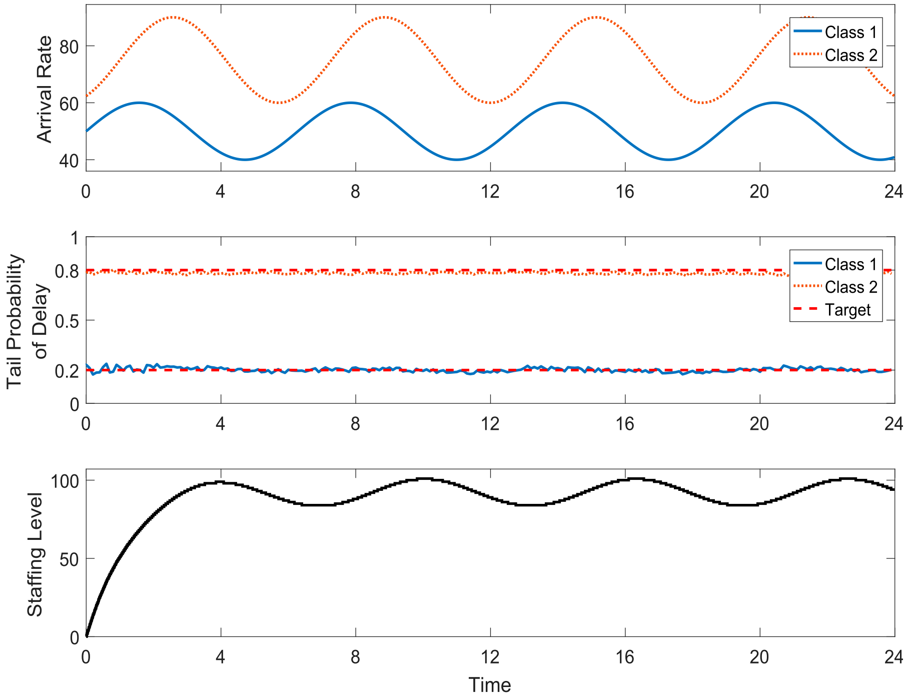


Table 1. A Two-Class Base-Case Example: Average, Max, and Min of (Simulated) TPoDs and Relative Differences to Their Target Levels, Using Floored, Rounded, and Ceiled Versions of the Time-Varying Staffing Formula

		Average			Max			Min		
Class		Floor	Round	Ceiling	Floor	Round	Ceiling	Floor	Round	Ceiling
1	α_1	0.2312	0.2091	0.1819	0.2515	0.2324	0.2053	0.2065	0.1862	0.1611
	%	(+15.59)	(+4.53)	(−9.06)	(+25.76)	(+16.21)	(+2.65)	(+3.27)	(−6.92)	(−19.47)
2	α_2	0.8085	0.7855	0.7612	0.8284	0.7995	0.7753	0.7950	0.7673	0.7473
	%	(+1.06)	(−1.81)	(−4.85)	(+3.55)	(−0.06)	(−3.09)	(−0.62)	(−4.09)	(−6.58)

base case, we let $n = 50$ and generate 5,000 independent runs. Specifically, at each time $0 \le t \le T$ on an arbitrary run, we schedule the next customer into service according to (31), using the control function κ_i given in Figure 4. We plot (i) arrival rates, (ii) simulations of TPoD, and (iii) staffing functions in Figure 5, using a sampling resolution (i.e., step size) $\Delta t = 0.01$. From a visual inspection of the middle panel of Figure 5, 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 Section EC.2.1.

Staffing Discretization. In practice (and in our simulation experiments), our time-varying staffing formula needs to be discretized to integer values. Table 1 gives the time-averaged, maximum, and minimum simulation results for the two TPoDs 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 (α 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.

7. Concluding Remarks

In this paper, we studied a service differentiation problem for a multiclass queueing system with class-dependent services and abandonment distributions. Motivated by call-center and health-care applications, we measure class-dependent service levels using the so-called TPoD, that is, the probability that the waiting time exceeds a delay target. Under a many-server asymptotic framework, we proposed a joint staffing and scheduling policy that can achieve intended service differentiation across all customer classes expressed in terms of TPoD constraints. We showed that there is a natural time-varying extension of the proposed staffing and scheduling solution to cope with time-varying arrivals. This modified solution is both state-dependent (based on real-time-elapsed customer delays) and time-dependent (capturing a time variability from the arrival processes), and can achieve TPoD-based performance stabilization at all times. Supplementing our limit theorems on asymptotic differentiation and stabilization, we also conducted 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.

Acknowledgments

The authors thank the area editor, associate editor, and two referees for their helpful comments. Special thanks to Ping Cao and Ward Whitt for providing constructive comments.

References

Aras AK, Chen X, Liu Y (2018) Many-server Gaussian limits for non-Markovian queues with customer abandonment. *Queueing Systems* 89(1):81–125.

Aras AK, Liu Y, Whitt W (2017) Heavy-traffic limit for the initial content process. *Stochastic Systems* 7(1):95–142.

Ata B, Tongarlak MH (2013) On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems* 74(1):65–104.

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. *Oper. Res.* 58(5):1427–1439.

Atar R, Giat C, Shimkin N (2011) On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems* 67(2):127–144.

Atar R, Mandelbaum A, Reiman M (2004) Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 14(3):1084–1134.

Ding Y, Park E, Nagarajan M, Grafstein E (2019) Patient prioritization in emergency department triage systems: An empirical study of canadian triage and acuity scale (CTAS). *Manufacturing Service Oper. Management* 21(4):723–741.

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Gurvich I, Whitt W (2010) Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* 58(2):316–328.

Gurvich I, Armony M, Mandelbaum A (2008) Service-level differentiation in call centers with fully flexible servers. *Management Sci.* 54(2):279–294.

Harrison JM, Zeevi A (2004) Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* 52(2): 243–257.

He B, Liu Y, Whitt W (2016) Staffing a service system with non-Poisson non-stationary arrivals. *Probab. Engrg. Inform. Sci.* 30(4):593–621.

Huang J, Mandelbaum A, Zhang H, Zhang J (2017) Refined models for efficiency-driven queues with applications to delay announcements and staffing. *Oper. Res.* 65(5):1380–1397.

Ito I (1979) On the existence and uniqueness of solutions of stochastic integral equations of the Volterra type. *Kodai Math. J.* 2(2):158–170.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10):1383–1394.

Kim J, Ward AR (2013) Dynamic scheduling of a *GI/GI/1 + GI* queue with multiple customer classes. *Queueing Systems* 75(2–4):339–384.

Kim J, Randhawa RS, Ward AR (2018) Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing Service Oper. Management* 20(2):285–301.

Kleinrock L (1964) A delay dependent queue discipline. *Naval Res. Logist.* 11(3–4):329–341.

Li N, Stanford DA, Taylor P, Ziedins I (2017) Non-linear accumulating priority queues with equivalent linear proxies. *Oper. Res.* 65(6):1712–1726.

Liu Y (2018) Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Oper. Res.* 66(2):514–534.

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6):1551–1564.

Liu Y, Whitt W (2014a) Stabilizing performance in networks of queues with time-varying arrival rates. *Probab. Engrg. Inform. Sci.* 28(4): 419–449.

Liu Y, Whitt W (2014b) Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1):378–421.

Liu Y, Whitt W (2017) Stabilizing performance in a service system with time-varying arrivals and customer feedback. *Eur. J. Oper. Res.* 256(2):473–486.

Liu Y, Sun X, Hovey K (2018) Online appendix to "Staffing and scheduling to differentiate service in time-varying multiclass service systems." Accessed February 22, 2021, https://yunanliu.wordpress.ncsu.edu/files/2020/08/SLEDapp120218.pdf.

Liu R, Kuhl M, Liu Y, Wilson J (2019) Modeling and simulation of nonstationary non-Poisson processes. *INFORMS J. Comput.* 31(2): 347–366.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Oper. Res.* 52(6):836–855.

Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* 57(5):1189–1205.

Preece D, Sherlock F, Bischoff B (2018) What are the industry standards for call centre metrics? Accessed February 22, 2021, https://www.callcentrehelper.com/industry-standards-metrics-125584.htm.

Puha AL, Ward AR (2019) Fluid limits for multiclass many server queues with general reneging distributions and head-of-line scheduling. Working paper, California State University San Marcos, San Marcos.

Shi P, Chou MC, Dai J, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* 62(1):1–28.

Soh SB, Gurvich I (2016) Call center staffing: Service-level constraints and index priorities. *Oper. Res.* 65(2):537–555.

Sun X, Whitt W (2018) Delay-based service differentiation in a many-server queue with time-varying arrival rates. *Stochastic Systems* 8(3):230–263.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized c/μ rule. *Ann. Appl. Probab.* 5(3): 809–833.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer, New York).

Yom-Tov GB, Mandelbaum A (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing Service Oper. Management* 16(2):283–299.

Yunan Liu is an associate professor in the Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests include queueing theory, stochastic modeling, applied probability, simulations, optimal control, online learning, and their applications in call-center, health-care, transportation, blockchain, and manufacturing systems. His work was awarded first place in the INFORMS Junior Faculty Interest Group Paper Competition in 2016.

Xu Sun is an assistant professor of data analytics and applied operations research in the Department of Industrial and Systems Engineering at the University of Florida. His research includes the analysis and optimization of stochastic service systems, the theory of stochastic-process limits, and the design and control of sustainable urban transport systems. He was a finalist of the 2017 Best Student Paper Competition in the INFORMS Finance Section and a recent recipient of the Travel Award from Institute for Mathematics and its Applications.

Kyle Hovey obtained his PhD degree from the graduate program of operations research at North Carolina State University, and he is currently an operations researcher for the Department of Defense. His research interests include stochastic systems, simulation, queueing theory, and data visualizations.