Staffing and Scheduling to Differentiate Service in Many-Server Service Systems

Xu Sun

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Staffing and Scheduling to Differentiate Service in Many-Server Service Systems

Xu Sun

This dissertation contributes to the study of a queueing system with a single pool of multiple homogeneous servers to which multiple classes of customers arrive in independent streams. The objective is to devise appropriate staffing and scheduling policies to achieve specified class-dependent service levels expressed in terms of tail probability of delays. Here staffing and scheduling are concerned with specifying a time-varying number of servers and assigning newly idle servers to a waiting customer from one of K classes, respectively. More 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 and a scheduling rule, under which the probability that a class-i customer waits more than w_i does not exceed α_i at all times. For this purpose, we propose new staffing-and-scheduling solutions under the critically-loaded and overloaded regimes. In both cases, the proposed solutions are both time dependent (coping with the time variability in the arrival pattern) and state dependent (capturing the stochastic variability in service and arrival times). We prove heavy-traffic limit theorems to substantiate the effectiveness of our proposed staffing and scheduling policies. We also conduct computer simulation experiments to provide engineering confirmation and practical insight.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to express my deepest gratitude to my advisor Prof. Ward Whitt for all his guidance and support through my doctoral studies. Prof. Whitt is a thought leader and an authority in the field of queueing theory. He is widely acknowledged as having good taste in research problems. I have learned from him that "good taste in research emerges gradually from studying broadly and deeply, listening to good speakers, and reading good papers". His ample knowledge and keen intuition have never ceased to amaze me. The knowledge and skills that he has instilled in me, both mathematical and engineering, have been invaluable to me. I could not have asked for a better mentor through my academic journey.

I would like to thank the other members of my dissertation committee: Agostino Capponi, Yunan Liu, Karl Sigman and David Yao. Under the direction of Prof. Capponi and Prof. Yao, I worked on another research project in which we applied weak convergence analysis techniques to investigate how the topology of a banking system network affects systemic risk. Although I finally decided not include that work in the thesis because of the topic difference, I do appreciate the rewarding learning experience and I am grateful to them for being wonderful mentors. I am indebted to Yunan for generously offering his time and guidance on our collaborative research and for offering his support during my job market year. I also thank Prof. Sigman for introducing me to the world of random processes and probabilistic simulation algorithms through his Queueing Theory and Simulation classes.

I'd like to extend my sincere thanks to all my PhD friends for their company and to all the IEOR staff members for creating such a warm and supportive environment. I'd also like to thank my friend and collaborator Bo Sun from Hong Kong University of Science and Technology for introducing me to another promising area for future study.

Finally, I owe a debt of gratitude to my Mom and Dad for their unconditional love and

affection all the time. My special thanks go to my beloved wife Lisha Zhou who has been a constant and never-ending source of strength, courage and inspiration. Without her love and support throughout the years, I would not have made it through the darkest days. I dedicate this work to her.

To my beloved wife Lisha Zhou

Chapter 1

Introduction

Queueing models provide powerful tools for describing the phenomenon of congestion, and find many applications in everyday life. Examples of such applications range from systems with visible queues (e.g., convenient stores, amusement parks, roads and bridges) to systems with invisible queues (e.g., call centers, ticketing systems, computer and communication networks). Queueing models consist of workstations with one or more shared servers, a finite or infinite buffer, and customers or jobs that arrive at those stations and require some amount of service. A typical queueing model describes the arrival process of customers at each node, the service-time distributions, a service discipline (describing in what order the customers in queue will be served) and a routing rule which directs customers to the available servers.

In many application areas, there are multiple customer or job classes. In general, anything that separates the customers into groups will lead to different customer types. One common example of multiclass service system is customer contact center where callers are often segmented into different classes based on request types. For instance, a banking call center may receive requests as simple as balance enquiries and as complex as dealing with fraudulent activities. While the former can be handled relatively quickly, fraudulent activities tend to be more difficult to handle and more urgent than obtaining balance information. Furthermore, callers who are calling about fraud may be more patient than those who request for balance enquiry services while waiting. This suggests that callers of different types may differ significantly in their service requirements, degree of impatience,

and urgency level. How to allocate the limited service resources to reasonably accommodate diverse customer needs has been an area of persistent pursuit.

In addition to multiple customer types, real-world applications tend to have time-varying customer arrival rates. This is in contrast to most analytical queueing models which assume a constant customer arrival rate. The arrival rate may depend upon time but be independent of the system state but we do not treat that feature. For instance, arrival rates change due to the time of day, the day of the week, or the season of the year. Of course, the arrival rate may depend upon the state of the system as well. Figure 1.1 taken from [Ye *et al.*, 2019] plots the mean arrival count volume for each time interval of each day-of-week for two call types in a US call center. The two plots demonstrate strong time-varying patterns and display obvious time-of-day effects. Such temporal variability is also typical for healthcare systems. Indeed, patient arrival rates in a hospital emergency department can very significantly over the course of the day; see e.g., Figure 1.2 taken from [Armony *et al.*, 2015], where the arrival rate of emergency department visits varies by a factor of 5. [Kim and Whitt, 2014] identified supporting evidence for the daily emergency department arrivals to fit a nonhomogenous Poisson process; see also the findings in [Maman, 2009]. We thus assume in our model that the arrival process of each follows a nonhomogenous Poisson process.

In this research, we study a service-level differentiation problem for a many-server service system with 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 a 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, expressed in terms of *tail probability of delay*.

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 satisfy diverse customer needs, and our choice of performance measures is primarily due to the fact that tail probability of delay is one of the prevalent performance measures in service industry. One notable example is the *Canadian triage and acuity scale* (CTAS) guideline that classifies patients in the emer-

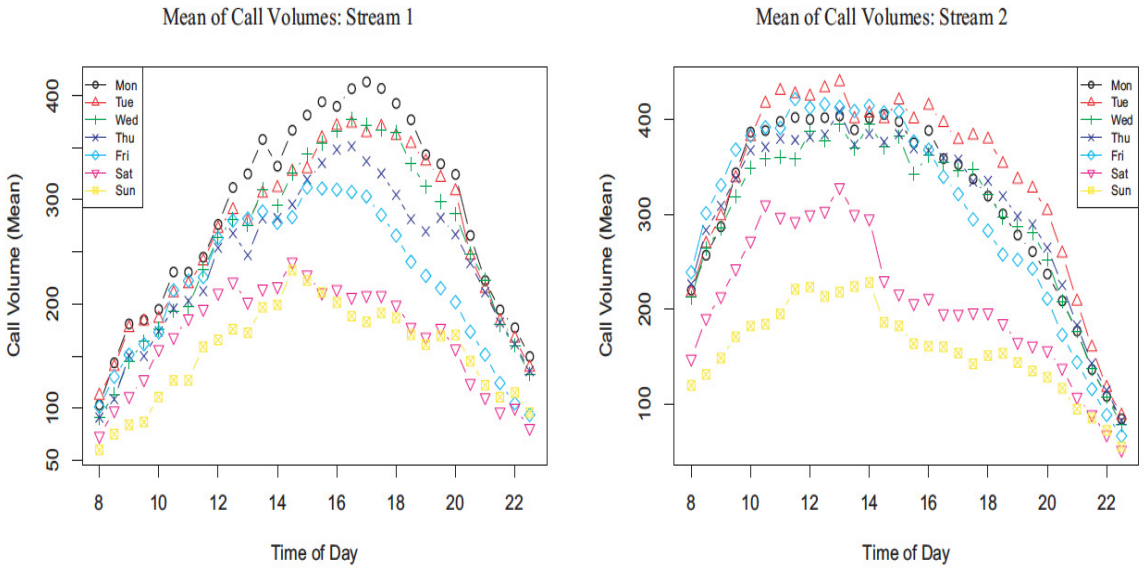

Figure 1.1: US call center: mean call volumes of different time intervals and different day-of-week

gency department into five acuity levels. Each acuity level is associated with a prescribed performance target, expressed in terms of 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], "CTAS level i patients need to be seen by a physician within w_i minutes $100\alpha_i\%$ of the time", with

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

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 emergency departments (not limited to those in Canada), see [Fernandes *et al.*, 2005].

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 web chat 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

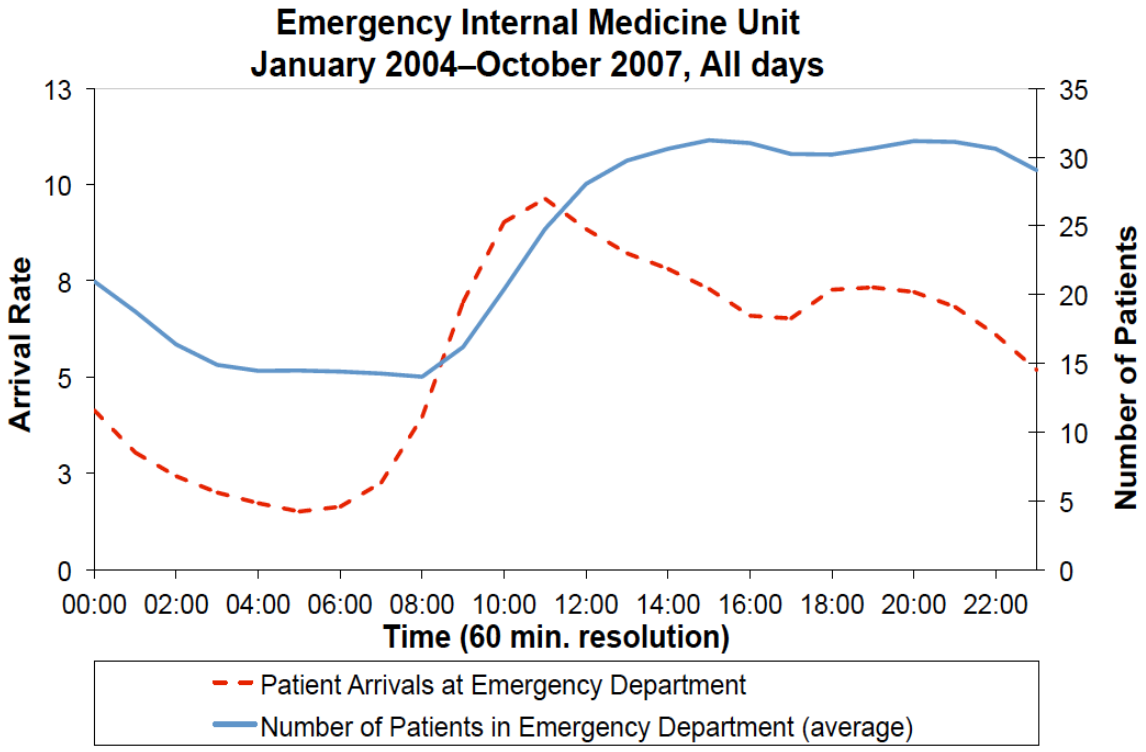

Figure 1.2: Average number of patients and arrival rate by hour of the day.

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, such as immigration offices in which the employees have to select cases to expedite in the face of a large backlog of immigration/permit applications and amusement parks where service providers have to tradeoff serving the fast-pass and regular customers; see [Kostami and Ward, 2009]. Our framework provides a useful tool to understand how scarce service resources should be allocated in the aforementioned systems where service strategies are either driven by revenue (e.g., banking call centers and amusement parts) or less tangible aspects such as social welfare (e.g., hospital emergency departments and immigration offices).

To summarize, the multiclass many-server queueing system considered here 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 emergency department 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. For instance, in a hospital emergency department, treatment times are evidently different for patients of various acuity levels with high acuity patients tending to stay longer before being discharged or admitted into the hospital.

The class of control policies proposed in this work consist of two components: (i) a staffing formula and (ii) a scheduling rule, relating to two different streams of research respectively, namely *staffing time-varying queues* and *scheduling in service systems*. We will discuss each in turn.

1.1 Staffing Time-Varying Queues

The staffing component of our proposed solution is related to works on the development of time-varying staffing functions to stabilize performance of relevant queueing systems having time-varying arrival-rate functions. The main idea is to adopt the *offered load analysis* which estimates the required service capacity by caculating how much capacity would be used if there were not limit on its availability.

Here a key underlying assumption is that demand is exogenous in a sense that the demand for service is not altered by the amount of service capacity being provided. More precisely, the congestion level should have no impact on customer arrival behaviors. Moreover, a successful application of the offered load analysis requires complete knowledge of the arrival rate. However, in many real-world applications this rate is not directly visible, so it must be estimated. For our purpose we will assume (throughout this work) that the arrival rates (or arrival-rate functions) are known. It is important to note that, for the targeted applications, the demand rates need to be easily and reliably estimated using appropriate operational data.

To illustrate the main idea, consider a single-class $M_t/GI/s_t+GI$ model having Poisson

arrivals rate $\lambda(t)$, independent and Independent and identically distributed (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 sufficiently complicated, the corresponding $M_t/GI/\infty$ infinite-server model remains highly tractable, where the number of customers (or busy servers) $X_\infty(t)$ at time t follows a Poisson distribution with mean $m_\infty(t)$ which is expressed in terms of the arrival-rate function λ and the service-time distribution G as

$$m_\infty(t) \equiv \mathbb{E}[X_\infty(t)] = \int_0^t \lambda(u) G^c(t-u) \mathrm{d}u. \tag{1.1}$$

This leads to the Gaussian approximation $X_\infty(t) \approx N(m_\infty(t), m_\infty(t))$, where $N(\mu, \sigma^2)$ denotes a random variable with a normal distribution having mean μ and variance σ^2. (We have variance equal to the mean because of the Poisson distribution.) If we choose $s(t)$ so that $\mathbb{P}(N(m_\infty(t), m_\infty(t)) > s(t)) = \alpha$, then we obtain the classical square-root staffing formula

$$s(t) = \lceil m_\infty(t) + \beta\sqrt{m_\infty(t)} \rceil,$$

where $\lceil x \rceil$ is smallest integer that is greater than or equal to x and $\beta = \Phi^{-1}(1-\alpha)$ is a *quality-of-service* parameter. It is easily to see from (1.1) that

$$m_\infty(t) = \int_0^t \lambda(u) G^c(t-u) \mathrm{d}u \to \mathbb{E}[\lambda(t - S_e)\mathbb{E}[S]] \quad \text{as} \quad t \to \infty, \tag{1.2}$$

where S and S_e are random variables with the service-time distribution G and the associated stationary-excess distribution, i.e., $\mathbb{P}(S_e \le x) \equiv (1/\mathbb{E}[S]) \int_0^x P(S > u) \mathrm{d}u$. The final expression in (1.2) supports the notion of a time lag, showing that the extent of the lag is related to S_e instead of the service time S. This random time lag can be explained by renewal theory. For a stationary $M/G/\infty$ model, the remaining service times of the customers in service conditioned on that number in service are distributed according to S_e. If the mean service time is relatively short, then the random time lag S_e can be ignored, which leads to the point-wise stationary approximation (PSA), namely

$$m_{PSA}(t) = \lambda(t)\mathbb{E}[S].$$

Indeed, the PSA algorithm has been proven useful in staffing systems with shorter service times and slowly varying arrival rates; see [Green *et al.*, 2007] for a review.

Another approach, often referred to as the *modified offered load* approach, has been adopted to design staffing functions that make it possible to stabilize various performance metrics including the probability of delay, mean waiting time, and fraction of abandonment, see [Jennings *et al.*, 1996; Liu and Whitt, 2012]. The key 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 service resources. An excellent survey on offered load analysis was provided by [Whitt, 2013]. [Feldman *et al.*, 2008] developed a simulation-based iterative staffing algorithm to stabilize *probability of delay*; the idea of this simulation-based iterative algorithm has been extended by [Defraeye and van Nieuwenhuyse, 2013] to treat tail probability of delay. Recently, [Liu, 2018] developed an analytic staffing function to stabilize the tail probability of delay and proved the a corresponding asymptotic stability result. To the best of our knowledge, prior to the current work there exists no result on joint staffing and scheduling decisions in time-varying queues to satisfy class-dependent performance metrics.

1.2 Scheduling in Service Systems

The scheduling component of our proposed solution relates to a vast body of research on scheduling. Scheduling deals with the problem of deciding which of the outstanding requests is to be allocated resources. There are many different scheduling algorithms. The simplest scheduling algorithm is perhaps the first-come first-served discipline that processes jobs in the order that they arrive. Various priority schemes can be implemented even without automatic customer classification. For example, the shortest-remaining-time-first policy, which is a preemptive version of the shortest-job-first scheduling, always selects the job with the smallest amount of time remaining until completion. By giving priority to customers whose service times are shorter, the shortest-remaining-time-first rule and the shortest-job-first policy can minimize the mean waiting time of the system, see e.g., [Schrage and Miller, 1966]. Yet these scheduling rules are found infrequently in practice due to the perceived unfairness (unless that class of customers is given a dedicated server, as in supermarket check-out systems) and/or due to the difficulty of estimating service times

accurately. Other commonly used scheduling policies include the *earliest-deadline-first* rule that keeps searching for the job closest to its deadline, which will be the next to be scheduled for processing, see e.g., [Doytchinov *et al.*, 2001] and the references therein.

In this work we are particularly interested in systems in which arriving customers are segmented into different classes. The standard approach to the optimal scheduling problem for the multiclass Markovian model is to formulate a Markov decision process, as in [Puterman, 1994], starting by specifying relevant costs (e.g., for waiting and for abandonment) and rewards (for completed service, e.g., throughput). For queueing problems such as these, a direct application is difficult, so that it is natural to seek asymptotic optimality in the presence of heavy-traffic scaling. Using the conventional heavy-traffic 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. Similar kind of asymptotic optimality result was established by [Stolyar, 2004] for a "max-weight" resource-pooling scheme in a general switch. [Ye and Yao, 2008] considered a broad class of utility-maximizing resource allocation schemes in the context of stochastic processing networks with concurrent occupancy of resources; they established heavy-traffic optimality of the utility-maximizing allocation; see also [Ye and Yao, 2012]. 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 work, we too allow for 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 emergency departments 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] in the context of time-stationary systems; 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 associated with the many-server heavy-traffic limit showing that the ratio rules are asymptotically optimal. However, 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).

1.3 Organization and Contribution

In Chapter 2, we introduce the queueing model with a single pool of multiple homogeneous servers to which K classes of customers arrive in independent streams. Following the convention (see, e.g., §5.1. of [Gans *et al.*, 2003a]), we refer to this model as the V system. Depending on how a service system is organized, different network topologies may arise. These include "V", "N", "X", "W" and "M" systems as displayed in Figure 16 of [Gans *et al.*, 2003a]. We formally describe what we mean by *service-level differentiation* in the setting of a V system. Following the literature, we distinguish between two operational regimes, namely the critically-loaded and overloaded regimes, characterized through the scaling conditions for the delay targets.

In Chapter 3, we solve the problem under the critically-loaded regime. We propose two scheduling rules, namely, the *time-varying queue-ratio* rule and the *head-of-line delay-ratio* rule, each having K ratio-control functions that can be used to achieve prescribed performance targets. Under the proposed control policies, we establish the functional central limit theorem for various quantities of interest. In particular, we establish state-space collapse by showing that all queue-length and waiting-time processes reduce to a simple function of a one-dimensional process under the proposed scheduling policies. Because of the state-space collapse, the queue-length and waiting-time processes of each class are related through a sample-path version of the heavy-traffic Little's Law. Based on the heavy-trafficlimits, we identify the desired control functions that allow us to asymptotically achieve service-level differentiation for all classes at customized performance targets.

Chapter 4 focuses on the overloaded regime. First we introduce a *time-varying square-root-staffing* rule and a *time-varying dynamic prioritization scheduling* having K control

functions. We then show that all waiting-time processes reduce to a simple function of a one-dimensional process called the *frontier process* under the proposed policy. 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. Based on the state-space collapse and further analysis of this frontier process, we identify the desired control functions for our staffing-and-scheduling policies. The computation of these control functions relies on the first and second moment of the limiting frontier process for which we develop efficient algorithms. We prove that the proposed policies asymptotically achieve service-level differentiation for all classes at customized service targets.

For both critically-loaded and overloaded systems, 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.

Chapter 2

Model Description and Problem Formulation

2.1 A Multiclass V Model

Consider a V system having K customer queues served by one common service pool. Let $A_i(t)$ denote the stochastic process counting the number of arrivals to the i^{th} queue in the interval $[t_i^0, 0]$, given the process starts at time t_i^0. We assume that $A_i(t)$ follows a non-homogeneous Poisson process (NHPP) with rate function λ_i. In what follows, we will be using $\Lambda_i(t)$ to denote the corresponding cumulative arrival function, i.e., $\Lambda_i(t) \equiv \int_{t_i^0}^{t} \lambda_i(u)\mathrm{d}u$.

We assume class-i service times are 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 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 $h_{F_i}(x) \equiv f_i(x)/F^c(x)$. We assume that service times and patience times are mutually independent, independent of the arrival processes. 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. 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

$$Q_i(t) = Q_i(0) + A_i(t) - E_i(t) - R_i(t). \tag{2.1}$$

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

$$B_i(t) = B_i(0) + E_i(t) - D_i(t). \tag{2.2}$$

Finally, let $X_i(t)$ denote the total number of class-i customers in the system at time t. Adding up (2.1) and (2.2) yields

$$X_i(t) = Q_i(t) + B_i(t) = X_i(0) + A_i(t) - D_i(t) - R_i(t). \tag{2.3}$$

Alternatively, one can derive (2.3) directly from flow conservation.

Two waiting times. We now introduce two types of waiting-time processes that we will exploit heavily in the subsequent analysis. Let $U_i(t)$ 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); $U_i(t) = 0$ 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:

$$E_i(t) = \sum_{k=1}^{A_i(t-U_i(t))} \mathbf{1}_{\{\gamma_{i,k} > V_i(\xi_{i,k})\}}, \tag{2.4}$$

$$Q_i(t) = \sum_{k=A_i(t-U_i(t))}^{A_i(t)} \mathbf{1}_{\{\xi_{i,k}+\gamma_{i,k} > t\}}, \tag{2.5}$$

where $\mathbf{1}_A$ denotes the indicator function of event (set) A, the random variables $t_i^0 \le \xi_{i,1} < \xi_{i,2} < \cdots$ 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 functional central limit theorem (FCLT). To complete the model, it remains to specify (i) a proper staffing level $s(t)$ (number of servers in the system) at time t 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). The choice of proper staffing levels and appropriate scheduling rules is guided by some prescribed performance targets as we will discuss momentarily.

2.2 Releasing Busy Servers

With possibly time-varying staffing levels, 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) total number of servers itself forms a random process. Henceforth, we use $s_d(t)$ to denote the number of busy servers that are due to depart at time t. Note that $s_d(t)$ can only increase at the time of scheduled staffing decrease and that it can only decrease when a server finishes a service *and* the value of $s_d(t)$ remains positive immediately prior to this job completion. This makes the process $s_d(\cdot)$ behave essentially like a queue, which allows us to bound s_d by a more tractable upper-bound process expressed through the one-dimensional reflection mapping ψ (see, e.g., §13.5 in [Whitt, 2002]):

$$s_d(t) \le Y(t) \equiv \psi(Z)(t) = Z(t) - \inf_{0 \le u \le t} Z(u), \tag{2.6}$$

where

$$Z(t) \equiv s(0) - s(t) - D(t) \tag{2.7}$$

with $D(t) \equiv \sum_{i \in \mathcal{I}} D_i(t)$ being the aggregate departure process. A rigorous proof of (2.6) is elementary and thus omitted.

2.3 Operational Regimes

Before we formalize the problem statement, we briefly review the constraint satisfaction problem for an $M/M/s+G$ queue, where one chooses the minimum staffing s that adheres to a given chance constraint. The constraint may be expressed in terms of tail of delay or tail probability of delay (TPoD). A straightforward approach is to apply exact formulae for performance measures of the $M/M/s+G$ queue. However, these formulae for performance measures are relatively complicated, involving double integration of the patience-time distribution. In addition, they provide no intuition and give rise to numerical problems for large s. For these reasons, an asymptotic approach is often pursued. Depending on the application context, two asymptotic operational regimes often arise: the critically-loaded (quality-and-efficiency-driven) and the overloaded (efficiency-driven) regime, each of which corresponds to a different approximate solution of the constraint satisfaction problem.

More formally, the critically-loaded regime corresponds to the least staffing level that adhere to the constraint $\mathbb{P}(V > 0) \leq \alpha$ or $\mathbb{P}(V > w_r/\sqrt{\lambda}) \leq \alpha$, where λ is the arrival rate and w_r is some constant; that is, the delay target is of order $O(1/\sqrt{\lambda})$. It has long been observed that the critically-loaded regime enables one to achieve high levels of efficiency and service quality for λ large enough; see, e.g., [Zeltyn and Mandelbaum, 2005]. In contrast, the overloaded regime corresponds to the least staffing that adhere to the constraint $\mathbb{P}(V > w_o) \leq \alpha$ given that the delay target w_o is in the order of a mean service time; see, e.g., [Mandelbaum and Zeltyn, 2009].

For the multiclass V model introduced in §2.1, we need a proper way to measure the overall load of the system. To this purpose, let $\bar{\lambda} \equiv T^{-1} \int_0^T \lambda(t)\mathrm{d}t$ for $\lambda(t) \equiv \sum_{i=1}^K \lambda_i(t)$; that is, $\lambda(t)$ is the aggregate demand function and $\bar{\lambda}$ is the corresponding time average over a planning horizon $[0, T]$. We are especially interested in satisfying the following service-level constraints:

$$\mathbb{P}\left(V_i(t) > w_i^{\bar{\lambda}}\right) \leq \alpha_i, \quad 1 \leq i \leq K, \quad 0 < t < T, \tag{2.8}$$

for class-specific delay target $w_i^{\bar{\lambda}}$ (which may or may not depend on $\bar{\lambda}$) and tail-probability target $\alpha_i \in (0, 1)$, $1 \leq i \leq K$, a finite time horizon T (e.g., $T = 24$), where $V_i(t)$ is the PWT of class i at time t, defined as the time that a class-i customer arriving at time t would have

to wait given that his/her patience is infinite. In words, the set of constraints requires that a class i customer who arrives at time t waits longer than $w_i^{\bar{\lambda}}$ time units with a probability no greater than α_i. We refer to the left-hand side of (2.8) as the TPoD. Such TPoD-based *quality-of-service* metrics have been widely used in service systems, such as the 80/20 rule in call centers [Aksin *et al.*, 2007; Gans *et al.*, 2003b], 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 (2.8) 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

$$\mathbb{P}\left(V_i(t) > w_i^{\bar{\lambda}}\right) \approx \alpha_i, \quad 1 \le i \le K, \quad 0 < t < T. \tag{2.9}$$

From now on we refer to the above problem as the *service-level differentiation* problem or joint-staffing-and-scheduling problem.

To put the V model into the *critically-loaded* regime, we need to scale the delay targets so that

$$\bar{\lambda}^{1/2} w_i^{\bar{\lambda}} \to w_i \quad \text{as} \quad \bar{\lambda} \to \infty \quad \text{for} \quad i \in \mathcal{I}. \tag{2.10}$$

This scaling follows Assumption 2.1 of [Gurvich and Whitt, 2010] that makes queue lengths be of order $O(\sqrt{\bar{\lambda}})$, while waiting times are of order $O(1/\sqrt{\bar{\lambda}})$.

In contrast, if we do not scale the delay targets, then we are forced into the overloaded regime, in which case

$$w_i^{\bar{\lambda}} \equiv w_i \quad \text{for} \quad i \in \mathcal{I}. \tag{2.11}$$

[Liu, 2018] has shown that this scaling can also be effective for stabilizing tail probabilities for the single-class model. The approach in [Liu, 2018] evidently should become relatively more effective as the delay targets increase. Such large targets often occur in healthcare; e.g., as in the 6-hour boarding time limit in the Singapore hospital.

Similar to the constraint-satisfaction problem for single-class models, the scalings (2.10) and (2.11) give rise to different (approximate) solutions to the staffing-and-scheduling problem given by (2.9) for our multiclass V system. We thus treat them separately in subsequent chapters (Chapter 3 and Chapter 4).

Chapter 3

The Critically-Loaded Regime

This chapter deals with the service-level differentiation problem in the critically-loaded regime. For this purpose, we propose a variant of the *square-root-staffing* (SRS) rule for staffing in §3.1 and two scheduling rules, namely, the time-varying queue-ratio (TVQR) and head-of-line delay-ratio (HLDR) rule in §3.2 and §3.3, respectively. We establish supporting results in §3.4 and propose staffing-and-scheduling solutions in §3.5.

Throughout this chapter, we impose two additional assumptions: (i) The system starts at time zero, i.e., $t_i^0 = 0$ ($i \in \mathcal{I}$); (ii) patience times are mutually independent and exponentially distributed. The mean patience time of each class i customer is $1/\theta_i$; that is, $F_i(x) = 1 - e^{-\theta_i x}$ for $i \in \mathcal{I}$. In addition, to further simplify the analysis, we assume the tail-probability targets α_i in (2.8) to be class-independent throughout this chapter; that is, $\alpha_i = \alpha$ for some $\alpha \in (0, 1)$.

3.1 Staffing

For the time-varying V model introduced earlier, our SRS staffing function is

$$\lceil s(t) = m(t) + \sqrt{\bar{\lambda}} c(t) \rceil, \tag{3.1}$$

where $m(t)$ is the *offered load* process, i.e., the expected number of busy servers in the associated infinite-server model (obtained by acting as if $s(t) = \infty$) and $c(t)$ is a control function to meet desired performance targets. Because the classes can be considered separately in

an infinite-server model, the offered load $m(t)$ is the sum of the corresponding single-class offered loads $m_i(t)$, i.e.,

$$m(t) = m_1(t) + \cdots + m_K(t) \quad \text{for} \quad t \geq 0,$$

where each of these offered load processes can be represented as the integral

$$m_i(t) \equiv \int_0^t e^{-\mu_i s} \lambda_i(t-s)\, ds \tag{3.2}$$

or as the solution of the ordinary differential equation

$$\dot{m}_i(t) = \lambda_i(t) - \mu_i m_i(t). \tag{3.3}$$

The SRS approach to time-varying staffing in (3.1) follows [Jennings *et al.*, 1996] and [Feldman *et al.*, 2008] for the single-class case, with (3.2) coming from Theorems 1 and 6 of [Eick *et al.*, 1993].

3.2 The TVQR Control

The TVQR rule is a time-varying version of the fixed-queue-ratio (FQR) rule studied in [Gurvich and Whitt, 2010]. We briefly review the FQR control, which is a special case of the more general queue-ratio control introduced by [Gurvich and Whitt, 2009], in the context of multiclass queue with a single pool of i.i.d. servers. Again, let $Q_i(t)$ be the queue length of class i, and let $Q(t)$ be the corresponding aggregate quantity. The FQR control uses a vector function $r \equiv (r_1, \ldots, r_K)$. Upon service completion, the available server admits to service the customer from the head of the queue i^* where

$$i^\star \equiv i^\star(t) \in \arg\max_{i \in \mathcal{I}} \{Q_i(t) - r_i Q(t)\};$$

i.e., the next-available-server always chooses to serve the queue with the greatest queue imbalance. Here instead of using fixed ratios we introduce a time-varying vector function $r(\cdot) \equiv (r_1(\cdot), \ldots, r_K(\cdot))$ and the next-available-server choose to serve a class i customer where

$$i^\star \equiv i^\star(t) \in \arg\max_{i \in \mathcal{I}} \{Q_i(t) - r_i(t) Q(t)\},$$

with ties broken evenly. Intuitively, the TVQR control makes

$$\frac{Q_i(t)}{r_i(t)} \approx \frac{Q_j(t)}{r_j(t)} \qquad 0 \le t \le T \quad \text{for} \quad i,j \in \mathcal{I}, \tag{3.4}$$

This is essentially a state-space collapse (SSC) result. In §3.4 we justify this form of SSC by establishing many-server heavy-traffic (MSHT) limit theorems for the TVQR control policy.

Another important result stemming from Theorem 3.4.1 and 3.4.2 is the time-varying Sample-Path heavy-traffic Little's law. In particular, for large-scale V systems that are approximately in the critically-loaded regime, we have

$$Q_i(t) \approx \lambda_i(t) V_i(t), \qquad 0 \le t \le T \quad \text{for} \quad i \in \mathcal{I}. \tag{3.5}$$

To illustrate the usefulness of the TVQR rule, set

$$r_i(t) = \frac{\lambda_i(t) w_i^{\bar{\lambda}}}{\sum_i \lambda_i(t) w_i^{\bar{\lambda}}}$$

in (3.4). Paralleling the big display between (13) and (14) on p. 322 of [Gurvich and Whitt, 2010], we have

$$\mathbb{P}\left(V_i(t) \ge w_i^{\bar{\lambda}}\right) \approx \mathbb{P}\left(Q_i(t) \ge \lambda_i(t) w_i^{\bar{\lambda}}\right) \approx \mathbb{P}\left(Q(t) \ge \sum_i \lambda_i(t) w_i^{\bar{\lambda}}\right),$$

where the first and second approximations follow from (3.5) and (3.4) ,respectively. Hence, given $\lambda_i(t)$ and $w_i^{\bar{\lambda}}$ for all i, we can stabilize all PWT processes at the target levels, i.e., we can achieve $P(V_i(t) \ge w_i^{\bar{\lambda}}) \approx \alpha$ for all i, if we can find a control function $c(t)$ that achieves

$$\mathbb{P}\left(Q(t) \ge \sum_i \lambda_i(t) w_i^{\bar{\lambda}}\right) \approx \alpha. \tag{3.6}$$

3.3 The HLDR Control

We next describe the HLDR scheduling rule that uniquely determines the enter-service processes $E_i(t)$. For that purpose, we introduce a set of control functions $v(t) \equiv (v_1(t), \ldots, v_K(t))$. Recall that $U_i(t)$ is the elapsed waiting time of the HoL customer in queue i. Define a weighted HoL delay

$$\tilde{U}_i(t) \equiv U_i(t)/v_i(t) \quad \text{for} \quad i \in \mathcal{I}.$$

Then the HLDR rule routes the next class-$i^\star$ HoL customer into service, with $i^\star$ satisfying

$$i^\star \equiv i^\star(t) \in \arg\max_{1\le i\le K} \left\{\tilde{U}_i(t)\right\},$$

with ties broken evenly. In words, the HLDR scheduling rule assigns the newly available server to the head-of-line (HoL) class-i customer that has the maximum value of $U_i(t)/v_i(t)$. The HLDR rule is appealing because it is a *blind* scheduling policy, i.e., it does not depend on any model parameters. In addition, we introduce $U(t)$ to represent the maximum of those weighted HoL delays, i.e.,

$$U(t) \equiv \max\left\{\tilde{U}_1(t),\ldots,\tilde{U}_K(t)\right\} = \max\left\{U_1(t)/v_1(t),\ldots,U_K(t)/v_K(t)\right\}. \tag{3.7}$$

The stationary version of HLDR, where the vector function $v(t)$ above is independent of t coincides with the *accumulating-priority* scheduling rule studied by [Stanford *et al.*, 2014; Sharif *et al.*, 2014]. 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. If $v_i(t) = 1$ for all $i \in \mathcal{I}$ and t; i.e., all classes accumulate priority at an equal constant rate, then the HLDR reduces to *global first-come first-serve*, as in [Talreja and Whitt, 2008].

3.4 MSHT FCLT Limits

To establish MSHT FCLT limits, 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, so that (2.10) becomes

$$n^{1/2} w_i^n \to w_i \quad \text{as} \quad n \to \infty \quad \text{for} \quad i \in \mathcal{I}.$$

This gives rise to a sequence of K-class V models indexed by n. As usual, we keep the service and abandonment rates unchanged, but let the arrival-rate and staffing functions in model n be $\lambda_i^n(t) \equiv n\lambda_i(t)$, so that the offered load is $m^n(t) = nm(t)$, and

$$\lceil s^n(t) = nm(t) + \sqrt{n}c(t) \rceil, \tag{3.8}$$

where $m(t)$ corresponds to the MSHT fluid limit, obtained from the associated *functional weak law of large numbers* (FWLLN). It is significant that the fluid limit coincides (with the

appropriate scaling by n) with the offered load in for the infinite-server model, as given in §3.1; e.g., see [Mandelbaum *et al.*, 1998]. The second expression in (3.8) is appealing for the simple direct way that n appears. In model n, the arrival processes $A_i^n(t)$ are independent NHPP's with rates $n\lambda_i(t)$. For $i \in \mathcal{I}$, let

$$\Lambda_i(t) \equiv \int_0^t \lambda_i(u)\mathrm{d}u, \qquad \widehat{A}_i^n(t) \equiv n^{-1/2}\left(A_i^n(t) - n\Lambda_i(t)\right).$$

The sequence of processes $\{\widehat{A}_i^n\}$ satisfies a FCLT; i.e.,

$$\widehat{A}_i^n(\cdot) \Rightarrow \mathcal{W}_i^a \circ \Lambda_i(\cdot) \equiv \widehat{A}_i(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{3.9}$$

where $\mathcal{W}_i^a$ represents a standard Brownian motion for each $i \in \mathcal{I}$. $\mathcal{D} \equiv \mathcal{D}(\mathbb{R}_+, \mathbb{R})$ 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).

We next introduce the diffusion-scaled processes

$$\widehat{X}_i^n(t) \equiv n^{-1/2}\left(X_i^n(t) - nm_i(t)\right) \quad \text{and} \quad \widehat{Q}_i^n(t) \equiv n^{-1/2}Q_i^n(t), \tag{3.10}$$

where $X_i^n(t)$ and $Q_i^n(t)$ represent the number of class-i customers in system and in queue at time t, respectively. The same scaling was used by [Feldman *et al.*, 2008; Whitt and Zhao, 2017]. As usual, we scale the delay processes by multiplying by $\sqrt{n}$ instead of dividing by $\sqrt{n}$:

$$\widehat{V}_i^n(t) \equiv n^{1/2}V_i^n(t) \quad \text{and} \quad \widehat{U}_i^n(t) \equiv n^{1/2}U_i^n(t) \quad \text{for} \quad i \in \mathcal{I}.$$

Mimicking the analysis of [Gurvich and Whitt, 2009], one can establish the MSHT limits, regarding the TVQR rule, via hydrodynamic limits. However, the proof in [Gurvich and Whitt, 2009] is quite involved and in turn relies on additional general SSC results from [Dai and Tezcan, 2011]. Owing to the simpler structure of the V system, we are able to avoid using the hydrodynamic functions and develop a much shorter and elementary proof. The proof, which is deferred to §3.7, adopts a similar stopping-time argument as used by [Atar *et al.*, 2011] in the analysis of an inverted-V system under the Longest-Idle-Pool-First routing rule.

Theorem 3.4.1 (MSHT FCLT for TVQR) *Suppose that the system is staffed according to* (3.8)*, operates under the TVQR scheduling rule. All control functions* $r_i(\cdot)$ *and* $c(\cdot)$

are continuous. If, in addition,

$$(\widehat{X}_1^n(0),\ldots,\widehat{X}_K^n(0),\widehat{Q}_1^n(0),\ldots,\widehat{Q}_K^n(0)) \Rightarrow (\widehat{X}_1(0),\ldots,\widehat{X}_K(0),\widehat{Q}_1(0),\ldots,\widehat{Q}_K(0))$$

in $\mathbb{R}^{2K}$ *as* $n\to\infty$*, then we have the joint convergence*

$$\left(\widehat{X}_1^n,\ldots,\widehat{X}_K^n,\widehat{Q}_1^n,\ldots,\widehat{Q}_K^n,\widehat{V}_1^n,\ldots,\widehat{V}_K^n\right) \Rightarrow \left(\widehat{X}_1,\ldots,\widehat{X}_K,\widehat{Q}_1,\ldots,\widehat{Q}_K,\widehat{V}_1,\ldots,\widehat{V}_K\right) \tag{3.11}$$

in $\mathcal{D}^{3K}$ *as* $n\to\infty$*, where the diffusion limits* $\widehat{X}_i$ *satisfy*

$$\begin{aligned}\widehat{X}_i(t) = \widehat{X}_i(0) - \mu_i\int_0^t \widehat{X}_i(u)\mathrm{d}u - (\theta_i-\mu_i)\int_0^t r_i(u)\left[\widehat{X}(u)-c(u)\right]^+\mathrm{d}u \\ + \int_0^t \sqrt{\lambda_i(u)+\mu_i m_i(u)}\mathrm{d}\mathcal{W}_i(u),\end{aligned} \tag{3.12}$$

where $\widehat{X}\equiv\sum_{i\in\mathcal{I}}\widehat{X}_i$ *and* $\mathcal{W}_i(\cdot)$ *are standard Brownian motions. For each* $i\in\mathcal{I}$*,*

$$\widehat{Q}_i(\cdot)\equiv r_i(\cdot)\left[\widehat{X}(\cdot)-c(\cdot)\right]^+ \quad \textit{and} \quad \widehat{V}_i(\cdot)=\frac{r_i(\cdot)}{\lambda_i(\cdot)}\cdot\left[\widehat{X}(\cdot)-c(\cdot)\right]^+, \tag{3.13}$$

Our next main result establishes a MSHT FCLT for HLDR in the critically-loaded regime. The limit is a set of interacting diffusion processess.

Theorem 3.4.2 (MSHT FCLT for HLDR) *Suppose that the system is staffed according to* (3.8) *with* $c(\cdot)$ *being continuous, operates under the HLDR scheduling rule. All control functions* $v_i(\cdot)$ *are continuous and bounded from above and away from zero; i.e.,* $v_*\equiv\min_{i\in\mathcal{I}}\inf_{t\geq 0}v_i(t)>0$ *and* $v^*\equiv\max_{i\in\mathcal{I}}\sup_{t\geq 0}v_i(t)<\infty$*. If, in addition, there is convergence of the initial distribution at time* 0*, i.e., if*

$$(\widehat{X}_1^n(0),\ldots,\widehat{X}_K^n(0),\widehat{Q}_1^n(0),\ldots,\widehat{Q}_K^n(0)) \Rightarrow (\widehat{X}_1(0),\ldots,\widehat{X}_K(0),\widehat{Q}_1(0),\ldots,\widehat{Q}_K(0))$$

in $\mathbb{R}^{2K}$ *as* $n\to\infty$*, then we have the joint convergence*

$$\begin{aligned}&\left(\widehat{X}_1^n,\ldots,\widehat{X}_K^n,\widehat{Q}_1^n,\ldots,\widehat{Q}_K^n,\widehat{V}_1^n,\ldots,\widehat{V}_K^n,\widehat{U}_1^n,\ldots,\widehat{U}_K^n\right)\\ &\Rightarrow\left(\widehat{X}_1,\ldots,\widehat{X}_K,\widehat{Q}_1,\ldots,\widehat{Q}_K,\widehat{V}_1,\ldots,\widehat{V}_K,\widehat{U}_1,\ldots,\widehat{U}_K\right)\end{aligned} \tag{3.14}$$

in $\mathcal{D}^{4K}$ *as* $n\to\infty$*, where the diffusion limits* $\widehat{X}_i$ *satisfy*

$$\begin{aligned}\widehat{X}_i(t) = \widehat{X}_i(0) - \mu_i\int_0^t \widehat{X}_i(u)\mathrm{d}u - (\theta_i-\mu_i)\int_0^t \gamma(u)^{-1}v_i(u)\lambda_i(u)\\ \times\left[\widehat{X}(u)-c(u)\right]^+\mathrm{d}u + \int_0^t\sqrt{\lambda_i(u)+\mu_i m_i(u)}\mathrm{d}\mathcal{W}_i(u)\end{aligned} \tag{3.15}$$

with $\gamma(\cdot) \equiv \sum_{i\in\mathcal{I}} v_i(\cdot)\lambda_i(\cdot)$, $\widehat{X} \equiv \sum_{i\in\mathcal{I}} \widehat{X}_i$ *and* $\mathcal{W}_i(\cdot)$ *i.i.d. standard Brownian motions. For each* $i \in \mathcal{I}$,

$$\widehat{Q}_i(\cdot) \equiv \gamma(\cdot)^{-1} v_i(\cdot)\lambda_i(\cdot)\left[\widehat{X}(\cdot) - c(\cdot)\right]^+, \quad \widehat{V}_i(\cdot) = \widehat{U}_i(\cdot) \equiv v_i(\cdot)\cdot\gamma(\cdot)^{-1}\left[\widehat{X}(\cdot) - c(\cdot)\right]^+. \tag{3.16}$$

Remark 3.4.1 (State-Space Collapse) *In both Theorem 3.4.1 and 3.4.2, we see that while the stochastic limit process* $(\widehat{X}_1, \ldots, \widehat{X}_K)$ *for the* K*-dimensional scaled number-in-system process* $(\widehat{X}_1^n, \ldots, \widehat{X}_K^n)$ *is a* K*-dimensional diffusion, depending on the* K *i.i.d. standard Brownian motions* $\mathcal{W}_i$*, the limits for the other processes are all a functional of the one-dimensional limit process* $\widehat{X}$*, in particular of* $\left[\widehat{X} - c\right]^+$*, so that there is great SSC. In particular, the limit processes* $\widehat{Q}_i$*,* $\widehat{V}_i$ *and* $\widehat{U}_i$ *are functions of each other, as shown by (3.13) for TVQR and by (3.16) for HLDR.*

Remark 3.4.2 (Asymptotic Equivalence of HLDR and TVQR) *We first observe that for a specific set of control functions* $v(\cdot) \equiv (v_1(\cdot), \ldots, v_K(\cdot))$ *used in the HLDR rule, one can always construct a set of time-varying queue-ratio functions* $r(\cdot) \equiv (r_1(\cdot), \ldots, r_K(\cdot))$ *such that the resulting TVQR control and the HLDR control are asymptotically equivalent.*

Fix the set of control functions $v(\cdot) \equiv (v_1(\cdot), \ldots, v_K(\cdot))$*. Let*

$$r_k(\cdot) = \frac{v_k(\cdot)\lambda_k(\cdot)}{\sum_{i\in\mathcal{I}} v_i(\cdot)\lambda_i(\cdot)} \quad \textit{for each} \quad k \in \mathcal{I}.$$

One can easily verify that the stochastic equation (3.15) becomes the equation (3.12).

We then observe that for a specific set of queue-ratio functions $r(\cdot) \equiv (r_1(\cdot), \ldots, r_K(\cdot))$*, one can always find a set of control functions* $v(\cdot) \equiv (v_1(\cdot), \ldots, v_K(\cdot))$ *used in the HLDR rule such that the resulting HLDR control and the TVQR control are asymptotically equivalent. In fact, the construction is also straightforward. Let*

$$v_k(\cdot) = \frac{r_k(\cdot)}{\lambda_k(\cdot)} \quad \textit{for each} \quad k \in \mathcal{I}.$$

Direct calculation allows us to translate equation (3.12) into (3.15).

Several important insights can be glimpsed from Theorem 3.4.1 and 3.4.2.

The Role of the SRS Safety Functions c Given that the staffing is done by (3.8), the behavior on the fluid scale is determined by the offered load $m(t) \equiv m_1(t) + \cdots + m_K(t)$, where the individual per-class offered loads m_i depend on the specified λ_i and μ_i for $i \in \mathcal{I}$. The remaining component of the staffing in (3.8) is specified by the SRS safety function c, which appears explicitly in the diffusion limits. Hence, in the limit, the remaining flexibility in the staffing depends entirely on the single function c, which remains to be specified. The limiting performance impact of the staffing function c can be seen directly in the limits, namely, (3.12) and (3.15).

The Sample-Path Heavy-Traffic Little's Law As an immediate consequence of Theorem 3.4.1 and Theorem 3.4.2, we obtain the sample-path heavy-traffic Little's Law for both scheduling control policies. In particular, for each i, we see that,

$$\widehat{Q}_i(t) = \lambda_i(t)\widehat{V}_i(t) \quad \text{for all} \quad t \geq 0.$$

For the n^{th} system, we have

$$\widehat{Q}_i^n(t) = \lambda_i(t)\widehat{V}_i^n(t) + o(1) \quad \text{or} \quad Q_i^n(t) = \lambda_i^n(t)V_i^n(t).$$

That is, the limit tells us that $Q_1^n(t)$ is $O(\sqrt{n})$, while the error is of a smaller order.

In what follows, we discuss several important special cases. We will primarily focus on the HLDR rule. The discuss for TVQR is similar. First, Theorem 3.4.2 applies to the stationary model as an important special case.

Corollary 3.4.1 (the stationary case) *Let $\lambda_i(t) = \lambda_i, v_i(t) = v_i$ and $c(t) = c$ for $t \geq 0$. If, in addition,*

$$(\widehat{X}_1^n(0), \ldots, \widehat{X}_K^n(0), \widehat{Q}_1^n(0), \ldots, \widehat{Q}_K^n(0)) \Rightarrow (\widehat{X}_1(0), \ldots, \widehat{X}_K(0), \widehat{Q}_1(0), \ldots, \widehat{Q}_K(0))$$

in $\mathbb{R}^{2K}$ as $n \to \infty$, then we have the joint convergence

$$\begin{aligned} &\left(\widehat{X}_1^n, \ldots, \widehat{X}_K^n, \widehat{Q}_1^n, \ldots, \widehat{Q}_K^n, \widehat{V}_1^n, \ldots, \widehat{V}_K^n, \widehat{U}_1^n, \ldots, \widehat{U}_K^n\right) \\ \Rightarrow &\left(\widehat{X}_1, \ldots, \widehat{X}_K, \widehat{Q}_1, \ldots, \widehat{Q}_K, \widehat{V}_1, \ldots, \widehat{V}_K, \widehat{U}_1, \ldots, \widehat{U}_K\right) \end{aligned}$$

in $\mathcal{D}^{4K}$ as $n \to \infty$, where the diffusion limits $\widehat{X}_i$ satisfy

$$\widehat{X}_i(t) = \widehat{X}_i(0) - \mu_i \int_0^t \widehat{X}_i(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t \gamma^{-1} v_i \lambda_i \left[\widehat{X}(u) - c\right]^+ \mathrm{d}u + \sqrt{2\lambda_i}\mathcal{W}_i(t),$$

in which $\gamma = \sum_{i\in\mathcal{I}} v_i\lambda_i$ *and* $\widehat{X} \equiv \sum_{i\in\mathcal{I}} \widehat{X}_i$*; for each* $i \in \mathcal{I}$*,*

$$\widehat{Q}_i(\cdot) \equiv v_i\lambda_i\gamma^{-1}\left[\widehat{X}(\cdot) - c\right]^+ \quad \textit{and} \quad \widehat{V}_i(\cdot) = \widehat{U}_i(\cdot) \equiv v_i \cdot \gamma^{-1}\left[\widehat{X}(\cdot) - c\right]^+. \tag{3.17}$$

Corollary 3.4.1 is in agreement with Theorem 4.3 in [Gurvich and Whitt, 2009] if one replaces the (state-dependent) ratio function $\tilde{p}_i$ there by a fixed ratio parameter $\gamma^{-1}v_i\lambda_i$. Theorem 4.3 in [Gurvich and Whitt, 2009] has $[\hat{X}]^+$ and $[\hat{X}]^-$ in the equation (6) whereas (3.15) in the present paper uses $[\widehat{X}-c]^+$ and $[\widehat{X}-c]^-$. The discrepancies are due to different centering component being used. In [Gurvich and Whitt, 2009] the number of customers in system is centered by the number of servers whereas we use $nm(t)$ to be the centering term.

If $\mu_i = \mu$ and $\theta_i = \theta$, $u \in \mathcal{I}$,then the limit of the aggregate content process $\widehat{X}$ is a one-dimensional diffusion. Hence, the limit is essentially the same as that for the single-class $M_t/M/s_t + M$ model as considered by [Whitt and Zhao, 2017] where the analysis draws upon [Puhalskii, 2013].

Corollary 3.4.2 (class-independent services) *Suppose that the conditions in Theorem 3.4.2 are satisfied and* $\mu_i = \mu$*,* $i \in \mathcal{I}$*. Then*

$$\left(\widehat{X}^n, \widehat{Q}_1^n, \ldots, \widehat{Q}_K^n, \widehat{V}_1^n, \ldots, \widehat{V}_K^n, \widehat{U}_1^n, \ldots, \widehat{U}_K^n\right) \Rightarrow \left(\widehat{X}, \widehat{Q}_1, \ldots, \widehat{Q}_K, \widehat{V}_1, \ldots, \widehat{V}_K, \widehat{U}_1, \ldots, \widehat{U}_K\right)$$

where

$$\begin{aligned}\widehat{X}(t) = \widehat{X}(0) - \mu\int_0^t \widehat{X}(u)\mathrm{d}u - \sum_{i\in\mathcal{I}}(\theta_i - \mu)\int_0^t \gamma(u)^{-1}v_i(u)\lambda_i(u) \\ \times\left[\widehat{X}(u) - c(u)\right]^+\mathrm{d}u + \int_0^t \sqrt{\lambda(u) + \mu m(u)}\mathrm{d}\mathcal{W}(u).\end{aligned} \tag{3.18}$$

For each $i \in \mathcal{I}$*,*

$$\widehat{Q}_i(\cdot) \equiv \gamma(\cdot)^{-1}v_i(\cdot)\lambda_i(\cdot)\left[\widehat{X}(\cdot) - c(\cdot)\right]^+, \quad \widehat{V}_i(\cdot) = \widehat{U}_i(\cdot) \equiv v_i(\cdot)\cdot\gamma(\cdot)^{-1}\left[\widehat{X}(\cdot) - c(\cdot)\right]^+. \tag{3.19}$$

If we assume further that $\theta_i = \mu$ in Corollary 3.4.2, then the aggregate model is known to behave like an $M_t/M/\infty$ model. Let $\theta = \mu = 1$ in (3.18). From 3.18 it holds that

$$\widehat{X}(t) = \widehat{X}(0) - \mu\int_0^t \widehat{X}(u)\mathrm{d}u + \int_0^t \sqrt{\lambda(u) + \mu m(u)}\mathrm{d}\mathcal{W}(u).$$

Hence the diffusion limit of the aggregate content process $\widehat{X}$ is an Ornstein-Uhlenbeck (OU) process with time-varying variance.

3.5 The Proposed Solution

We propose a solution that consists of a staffing component and and a scheduling component. Recall that v and r are the ratio functions in the HLDR and TVQR rule respectively and c is the TV safety staffing function.

▷ **staffing:** Choose c^* that satisfies $\mathbb{P}\left(\widehat{X}(t) > \vartheta(t) + c^*(t)\right) = \alpha$ with

$$\vartheta(t) \equiv \sum_{i\in\mathcal{I}} \lambda_i(t) w_i. \tag{3.20}$$

▷ **scheduling:** (a) Apply HLDR with ratio functions

$$v^* \equiv (v_1^*(t), \ldots, v_K^*(t)) = (w_1, \ldots, w_K), \tag{3.21}$$

or (b) use TVQR with ratio functions

$$r^* \equiv (r_1^*(t), \ldots, r_K^*(t)) = (\lambda_1(t)w_1, \ldots, \lambda_K(t)w_K)/\vartheta(t). \tag{3.22}$$

Informally, our FCLT supports the use of the following approximation:

$$\mathbb{P}\left(V_i^n(t) > w_i^n\right) \approx \mathbb{P}\left(\widehat{V}_i(t) > w_i\right) = \mathbb{P}\left(\left[\widehat{X}(t) - c^*(t)\right]^+ > \vartheta(t)\right).$$

Given that we are taking advantage of the SSC provided by TVQR and HLDR, the form of the limit reveals how difficult is the overall control problem. The difficulty depends critically upon the model parameters μ_i and θ_i.

In this work, we identify three cases. **Case 1** is the general model with parameters μ_i and θ_i depending on the class i, for which Theorems 3.4.1 and 3.4.1 show that the limit in reduction above is $\widehat{Q}(t) = [\widehat{X}(t) - c(t)]^+$, where $\widehat{X}(t)$ is a sum of the components of a K-dimensional diffusion process. We obtain the other two cases by imposing additional conditions on the service and abandonment rates. **Case 2** has $\theta_i = \mu_i$ for all i; then the limit process has the structure of a time-varying K-dimensional OU diffusion process, complicated by a time-varying variance. The K-dimensional structure of the limit process in cases 1 and 2 reveals inherent challenges in analyzing the multiclass model.

The strongest positive conclusions are for case 3. **Case 3** has $\mu_i = \mu$ for all i; then the limit process is a 1-dimensional diffusion process. In Case 3, we can effectively reduce

the staffing component to the staffing problem for the associated single-class $M_t/M/s_t + M$ model. It remains to solve the 1-dimensional diffusion control problem to find the staffing function. For practical applications, this result strongly supports applying TVQR or HLDR together with heuristic staffing algorithms for the single-class $M_t/M/s_t + M$ model, such as the modified-offered-load approximation or the iterative-staffing-algorithm in [Feldman *et al.*, 2008]; these are surveyed in [Whitt and Zhao, 2017].

3.6 Numerical Studies

Successful application of the proposed solutions to the joint-staffing-and-scheduling problem in §3.5 requires effective computation of the *minimum* safety staffing function c^*. In this section we illustrate how the function c^* can be calculated explicitly for a special case where $\theta_i = \mu_i = \mu$ for all i. Then we present results of simulation experiments to show how HLDR and TVQR perform.

To calculate the minimum safety staffing function c^* for the *tail-probability* formulation, let

$$\alpha = \mathbb{P}\left(\widehat{X}(t) > c(t) + \vartheta(t)\right).$$

We apply Corollary 3.4.2 and the following remark, which identifies $\widehat{X}(t)$ as an OU process. Because $\widehat{X}(t)$ is normally distributed with mean 0 and variance $m(t)$, it holds that

$$c^*(t) = \Phi^{-1}(1-\alpha)\sqrt{m(t)} - \vartheta(t). \tag{3.23}$$

For our simulation experiments, we start by considering a two-class Markov V model. The arrival-rate functions are given by

$$\lambda_i(t) = a_i + b_i \sin(d_i t) \quad \text{for} \quad 0 \le t \le T, \quad i = 1, 2,$$

where $(a_1, b_1, d_1) = (60, -20, 2/5)$ and $(a_2, b_2, d_2) = (90, 30, 2/5)$. We assume that $\mu_i = \theta_i = 1, i = 1, 2$. In addition, we stipulate that the delay targets for class-1 and class-2 are $w_1^n \equiv 1/6$ and $w_2^n \equiv 1/3$ respectively.

Figure 3.1 plots the tail probabilities over the time interval $[0, 50]$ for the HLDR rule (plots at the top) and the TVQR rule (plots at the bottom) with c^* derived from (3.23).

Here we tested three different tail-probability targets, $\alpha = 0.25, 0.5, 0.75$. We plot the tail probabilities for both classes. All estimates were obtained by averaging over 2000 independent replications. Figure 3.1 shows that, for all three cases, both HLDR and TVQR stabilize the tail probabilities of each class at the desired level.

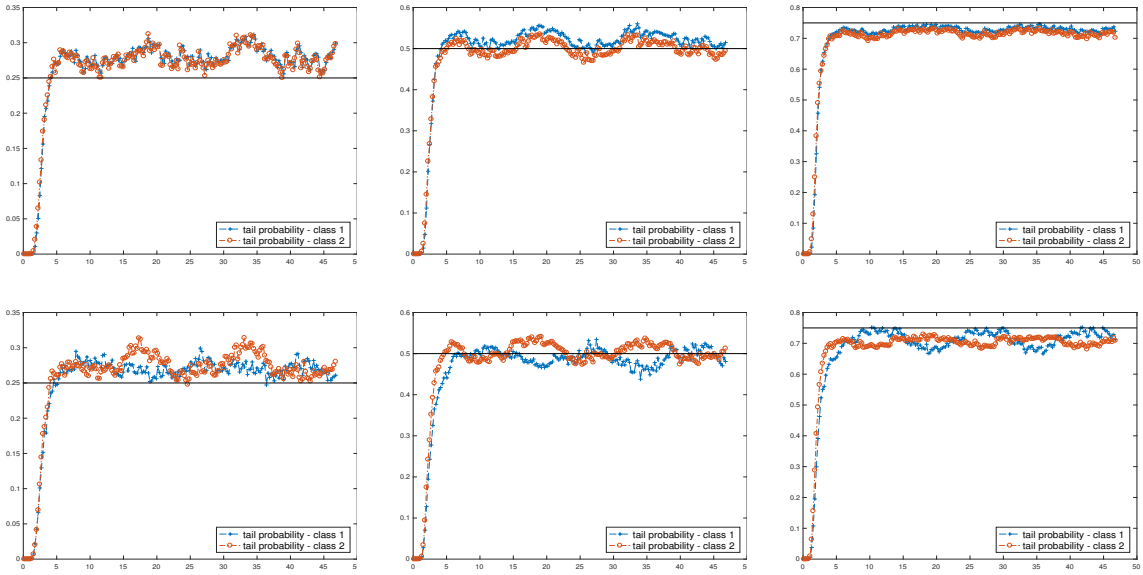

Figure 3.1: Tail probabilities for a two-class $M_t/M/s_t+M$ queue with arrival-rate functions $\lambda_1(t) = 60 - 20\sin(2t/5)$, $\lambda_2 = 90 + 30\sin(2t/5)$, common service rate $\mu = 1$, abandonment rate $\theta = 1$ and minimum staffing function c^* derived from (3.23).

3.7 Proofs of Theorem 3.4.1 - 3.4.2

We now provide the proofs for Theorem 3.4.1 and Theorem 3.4.2.

Proof of Theorem 3.4.1.

The proof proceeds in four steps.

1. Stochastic Boundedness of $\{\widehat{X}_i^n(\cdot); n \in \mathbb{N}\}$ and $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ Here we exploit a martingale decomposition, as in [Pang *et al.*, 2007] and [Puhalskii, 2013]. Specifically the

processes

$$\widehat{D}_i^n(t) \equiv n^{-1/2}\left[D_i^n(t) - \mu_i \int_0^t B_i^n(u)\mathrm{d}u\right] \quad \text{and} \quad \widehat{R}_i^n(t) \equiv n^{-1/2}\left[R_i^n(t) - \theta_i \int_0^t Q_i^n(u)\mathrm{d}u\right]$$

are square-integrable martingales with respect to a proper filtration. The associated quadratic variation processes are

$$\langle \widehat{D}_i^n \rangle(t) = \frac{\mu_i}{n}\int_0^t B_i^n(u)\mathrm{d}u \quad \text{and} \quad \langle \widehat{R}_i^n \rangle(t) = \frac{\theta_i}{n}\int_0^t Q_i^n(u)\mathrm{d}u.$$

Both $\{\widehat{D}_i^n(\cdot); n \in \mathbb{N}\}$ and $\{\widehat{R}_i^n(\cdot); n \in \mathbb{N}\}$ are stochastically bounded due to Lemma 5.8 of [Pang *et al.*, 2007], which is based on the Lenglart-Rebolledo inequality, stated as Lemma 5.7 there.

From (3.3), it follows

$$m_i(t) = m_i(0) + \int_0^t \lambda_i(u)\mathrm{d}u - \mu_i \int_0^t m_i(u)\mathrm{d}u. \tag{3.24}$$

Scaling both sides of (3.24) by n, subtracting it from (2.3), and dividing both sides by $n^{1/2}$ yields

$$\widehat{X}_i^n(t) = \widehat{X}_i^n(0) - \mu_i \int_0^t \widehat{X}_i^n(u)\mathrm{d}u - (\theta_i - \mu_i)\int_0^t \widehat{Q}_i^n(u)\mathrm{d}u + \widehat{A}_i^n(t) - \widehat{D}_i^n(t) - \widehat{R}_i^n(t). \tag{3.25}$$

Let $\bar{a} \equiv \max_i \mu_i \vee \max_i \theta_i$ and

$$\hat{\mathcal{M}}_i^n(t) \equiv \widehat{A}_i^n(t) - \widehat{D}_i^n(t) - \widehat{R}_i^n(t). \tag{3.26}$$

Note that $\{\mathcal{M}_i^n; n \in \mathbb{N}\}$ is stochastically bounded. Using (3.25) - (3.26), we have

$$\left|\widehat{X}_i^n(t)\right| \le \left|\widehat{X}_i^n(0)\right| + \bar{a}\int_0^t \left[\left|\widehat{X}_i^n(u)\right| + \widehat{Q}_i^n(u)\right]\mathrm{d}u + \left|\hat{\mathcal{M}}_i^n(t)\right|. \tag{3.27}$$

Adding up (3.27) over $i \in \mathcal{I}$ and letting $\hat{\mathbb{X}}^n \equiv \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n\right|$, we obtain

$$\hat{\mathbb{X}}^n(t) \le \hat{\mathbb{X}}^n(0) + \bar{a}\int_0^t \left[\hat{\mathbb{X}}^n(u) + \widehat{Q}^n(u)\right]\mathrm{d}u + \sum_{i\in\mathcal{I}}\left|\hat{\mathcal{M}}_i^n(t)\right|. \tag{3.28}$$

To proceed, define $\hat{s}_d^n(t) \equiv n^{-1/2}s_d^n(t)$ and note that

$$\widehat{Q}^n(t) = \left[\widehat{X}^n(t) - c(t) - \hat{s}_d^n(t)\right]^+ \le \hat{\mathbb{X}}^n(t) + \left|c(t)\right|, \tag{3.29}$$

where the inequality follows from the nonnegativity of $s_d^n(t)$. Plugging (3.29) into (3.28) gives us

$$\hat{\mathbb{X}}^n(t) \leq \hat{\mathbb{X}}^n(0) + \bar{a}\int_0^t |c(u)|\,\mathrm{d}u + 2\bar{a}\int_0^t \hat{\mathbb{X}}^n(u)\mathrm{d}u + \sum_{i\in\mathcal{I}}\left|\hat{\mathcal{M}}_i^n(t)\right|. \tag{3.30}$$

An application of the Gronwall's inequality with (3.30) establishes the stochastic boundedness of $\left\{\hat{\mathbb{X}}^n; n\in\mathbb{N}\right\}$. Thus for $i\in\mathcal{I}$ the sequence $\{\widehat{X}_i^n; n\in\mathbb{N}\}$ is stochastically bounded. Then the stochastic boundedness of $\{\widehat{Q}^n; n\in\mathbb{N}\}$ follows easily by (3.29). In addition, $\{\widehat{B}^n; n\in\mathbb{N}\}$ is also stochastically bounded due to the relation $\widehat{X}_i^n = \widehat{B}_i^n + \widehat{Q}_i^n$.

We next use the established stochastic boundedness to derive the fluid limit for the number of customers in system and the number of busy servers, as in [Pang *et al.*, 2007]. Indeed, by (3.10) we must have

$$\left(\bar{X}_i^n, \bar{Q}_i^n\right) \equiv n^{-1}\left(X_i^n, Q_i^n\right) \Rightarrow (m_i, 0) \quad \text{in} \quad \mathcal{D}^{2K} \quad \text{as} \quad n\to\infty \tag{3.31}$$

and thus

$$\bar{B}_i^n \equiv \frac{B_i^n}{n} = \frac{X_i^n - Q_i^n}{n} \Rightarrow m_i \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{3.32}$$

Applying the continuous mapping theorem (CMT) with integration in (3.32), we have

$$\bar{D}_i^n(t) \equiv \mu_i\int_0^t \bar{B}_i^n(u)\mathrm{d}u \Rightarrow \mu_i\int_0^t m_i(u)\mathrm{d}u \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{3.33}$$

Then apply the CMT with composition in (3.33) to obtain

$$\widehat{D}_i^n(t) \equiv n^{-1/2}\left[D_i^n(t) - \mu_i\int_0^t B_i^n(u)\mathrm{d}u\right] \Rightarrow \mathcal{W}_i^{\mu}\left(\mu_i\int_0^t m_i(u)\mathrm{d}u\right) \quad \text{in} \quad \mathcal{D} \tag{3.34}$$

as $n\to\infty$ where we have used $\mathcal{W}_i^{\mu}$ to denote a standard Brownian motion. Similarly, we have

$$\widehat{R}_i^n(t) \equiv n^{-1/2}\left[R_i^n(t) - \theta_i\int_0^t Q_i^n(u)\mathrm{d}u\right] \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \tag{3.35}$$

as $n\to\infty$.

2. Asymptotic Negligibility of $\{\hat{s}_d^n; n\in\mathbb{N}\}$ The argument required here is a variant of Theorem 13.5.2 (b) in [Whitt, 2002], but the extra term needed to get convergence is nonlinear instead of $c_n e$ there and we exploit stochastic boundedness rather than convergence, so we give the direct argument

To establish the uniform asymptotic negligibility of $\{\hat{s}_d^n; n \in \mathbb{N}\}$, we first argue that $\widehat{Y}^n \equiv n^{-1/2}Y^n$, where Y^n is defined by (2.6), vanishes as $n \to \infty$. For that purpose, define $\widehat{Z}^n \equiv n^{-1/2}Z^n$ for Z^n given in (2.7). Some algebraic manipulation leads easily to

$$\widehat{Z}^n(t) = -n^{1/2}\int_0^t \lambda(u)\mathrm{d}u - \mathcal{X}^n(t) \tag{3.36}$$

where

$$\mathcal{X}^n(t) \equiv \widehat{D}^n(t) + \sum \mu_i \int_0^t \widehat{B}_i^n(u)\mathrm{d}u,$$

for $\widehat{D}^n(t) \equiv \sum_{i\in\mathcal{I}} \widehat{D}_i^n(t)$. By the C-tightness of $\widehat{D}^n$ and the stochastic boundedness of $\widehat{B}^n$, we deduce that $\{\mathcal{X}^n(\cdot); n \in \mathbb{N}\}$ is stochastically bounded and C-tight. By (2.6),

$$\widehat{Y}_0^n(t) = \widehat{Z}^n(t) + \sup_{u\le t}\left\{-\widehat{Z}^n(u)\right\}. \tag{3.37}$$

Define

$$u^n(t) \equiv \arg\max_{u\le t}\left\{-\widehat{Z}^n(u)\right\} = \arg\max_{u\le t}\left\{n^{1/2}\int_0^t \lambda(u)\mathrm{d}u + \mathcal{X}^n(t)\right\}.$$

From (3.36) - (3.37) it follows

$$\widehat{Y}_0^n(t) = -n^{1/2}\int_{u^n(t)}^t \lambda(u)\mathrm{d}u - \mathcal{X}^n(t) + \mathcal{X}^n(u^n(t)) \ge 0 \tag{3.38}$$

Combining the inequality in (3.38) and the stochastic boundedness of $\mathcal{X}^n(\cdot)$ allows us to conclude

$$\sup_{t\le T}\{t - u^n(t)\} = O_p(n^{-1/2}). \tag{3.39}$$

For a cadlag (right continuous with left limits) function $x(\cdot)$, define $|x|_T^* \equiv \sup_{t\le T}|x(t)|$. Using (3.38), we can easily deduce

$$\mathbb{P}\left(\left|\widehat{Y}^n\right|_T^* > \epsilon\right) \le \mathbb{P}\left(\sup_{t\le T}\{-\mathcal{X}^n(t) + \mathcal{X}^n(u^n(t))\} \ge \epsilon\right).$$

In virtue of the established C-tightness of $\mathcal{X}^n$ and (3.39),

$$\mathbb{P}\left(\sup_{t\le T}\{-\mathcal{X}^n(t) + \mathcal{X}^n(u^n(t))\} \ge \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Since ϵ is arbitrarily chosen, we have proven

$$\widehat{Y}^n \equiv n^{-1/2}Y^n \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.$$

Hence, we must have

$$\hat{s}_d^n \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{3.40}$$

3. State Space Collapse Define for each $i \in \mathcal{I}$ the imbalance process

$$\Delta_i^n(\cdot) \equiv \widehat{Q}_i^n(\cdot) - r_i(\cdot)\widehat{Q}^n(\cdot). \tag{3.41}$$

At each decision epoch, the QR rule chooses a class with maximum positive imbalance and assign the head-of-line customer from that queue to the next available server.

Suppose that $\Delta_i^n(0) \neq 0$. Our analysis below indicates that it takes infinitesimally small time for the imbalance process Δ_i^n to hit zero. Hence, assume without loss of generality that $\Delta_i^n(0) = 0$. We aim to show that, for each $i \in \mathcal{I}$, the process $\widehat{Q}_i^n(\cdot)$ is infinitely close to $\widehat{Q}_i^n(\cdot)$ as n grows. More precisely, we aim to show that, for each $i \in \mathcal{I}$ and $\epsilon > 0$,

$$\mathbb{P}\left(|\Delta_i^n|_T^* > \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.42}$$

Define a stopping time (depending on ϵ)

$$\tilde{\tau}_i^n \equiv \inf\left\{t > 0 : |\Delta_i^n(t)| > \epsilon\right\}$$

Then to establish (3.42), it suffices to show $\mathbb{P}\left(\tilde{\tau}_i^n \leq T\right) \to 0$ as $n \to \infty$. Note that $\sum_{i\in\mathcal{I}} \Delta_i^n(\cdot) = 0$. Thus the problem further boils down to showing

$$\mathbb{P}\left(\tau_i^n \leq T\right) \to 0 \quad \text{as} \quad n \to \infty$$

where $\tau_i^n \equiv \inf\left\{t > 0 : \Delta_i^n(t) < -\epsilon\right\}$. On the event $C \equiv \{\tau_i^n \leq T\}$, let us define another random time σ_i^n

$$\sigma_i^n \equiv \sup\left\{t \geq 0 | t < \tau_i^n, \Delta_i^n(t) \geq -\epsilon/2\right\}.$$

With the initial condition $\Delta_i^n(0) = 0$, such a random time σ_i^n is guaranteed to exist on the event C.

Working with the notation $x(t_1, t_2] \equiv x(t_2) - x(t_1)$ for a function $x(\cdot)$ in t and exploiting (2.3), one can easily derive

$$\sum_{i\in\mathcal{I}} A_i^n(\sigma_i^n, \tau_i^n] - D^n(\sigma_i^n, \tau_i^n] - \sum_{i\in\mathcal{I}} R_i^n(\sigma_i^n, \tau_i^n] = s^n(\sigma_i^n, \tau_i^n] + s_d^n(\sigma_i^n, \tau_i^n] + \sum_{i\in\mathcal{I}} Q_i^n(\sigma_i^n, \tau_i^n] \tag{3.43}$$

Moreover, no customer enters service from the k^{th} queue over $[\sigma_i^n, \tau_i^n]$ and so

$$Q_k^n(\sigma_i^n, \tau_i^n] = A_k^n(\sigma_i^n, \tau_i^n] - R_k^n(\sigma_i^n, \tau_i^n]. \tag{3.44}$$

Combining (3.43) and (3.44) yields

$$\sum_{i\neq k} A_i^n(\sigma_i^n,\tau_i^n] - D^n(\sigma_i^n,\tau_i^n] - \sum_{i\neq k} R_i^n(\sigma_i^n,\tau_i^n] = s^n(\sigma_i^n,\tau_i^n] + s_d^n(\sigma_i^n,\tau_i^n] + \sum_{i\neq k} Q_i^n(\sigma_i^n,\tau_i^n]. \tag{3.45}$$

From (3.45) it follows easily

$$\begin{aligned} n^{1/2}\int_{\sigma_i^n}^{\tau_i^n} &= \sum_{i\neq k}\widehat{A}_i^n(\sigma_i^n,\tau_i^n] - \widehat{D}^n(\sigma_i^n,\tau_i^n] - \sum_{i\neq k}\widehat{R}_i^n(\sigma_i^n,\tau_i^n] \\ &\quad - \sum_{i\neq k}\theta_i\int_{\sigma_i^n}^{\tau_i^n}\widehat{Q}_i^n(u)\mathrm{d}u - c(\sigma_i^n,\tau_i^n] - \hat{s}_d^n(\sigma_i^n,\tau_i^n] - \sum_{i\neq k}\widehat{Q}_i^n(\sigma_i^n,\tau_i^n]. \end{aligned}$$

That all terms on the right side are stochastically bounded implies the stochastic boundedness of the sequence $\{n^{1/2}(\tau_i^n-\sigma_i^n); n\in\mathbb{N}\}$.

Define $\Gamma_i^n(t_1,t_2] \equiv r_i(t_2)\widehat{Q}^n(t_2) - r_i(t_1)\widehat{Q}^n(t_1)$ and let $\epsilon' = \epsilon/4$, using union bound, we obtain

$$\begin{aligned} \mathbb{P}\left(\tau_i^n \le T\right) &\le \mathbb{P}\left(\Delta_i^n(\tau_i^n) < -\epsilon, \Delta_i^n(\sigma_i^n) \ge -\epsilon/2\right) \\ &\le \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) - \Gamma_i^n(\sigma_i^n,\tau_i^n] < -\epsilon/2\right) \\ &\le \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) - \Gamma_i^n(\sigma_i^n,\tau_i^n] < -\epsilon/2, \Gamma_i^n(\sigma_i^n,\tau_i^n] \le \epsilon'\right) \\ &\quad + \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) - \Gamma_i^n(\sigma_i^n,\tau_i^n] < -\epsilon/2, \Gamma_i^n(\sigma_i^n,\tau_i^n] > \epsilon'\right) \\ &\le \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) + \mathbb{P}\left(\Gamma_i^n(\sigma_i^n,\tau_i^n] > \epsilon/4\right) \end{aligned} \tag{3.46}$$

Recall that our goal is to show $\mathbb{P}\left(\tau_i^n \le T\right)$ goes to zero as $n\to\infty$. To that end, we argue that both terms at the right end of (3.46) converge to zero as n grows to infinity.

For the first term, notice that no customer entered service from queue i under the TVQR rule over $[\sigma_i^n,\tau_i^n]$. Thus, if no customer abandoned the queue, then we must have

$$\mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) = 0$$

by the fact that Q_i^n is nondecreasing over $[\sigma_i^n,\tau_i^n]$. With customer abandonments, we have

$$\mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) \le \mathbb{P}\left(n^{-1/2}R_i^n(\tau_i^n) - n^{-1/2}R_i^n(\sigma_i^n) < -\epsilon/4\right), \tag{3.47}$$

because only abandonments can cause Q_i^n to decrease over $[\sigma_i^n,\tau_i^n]$. The following lemma plays a crucial role in the rest of proof.

Lemma 3.7.1 *Both* $\{\widehat{Q}^n; n \in \mathbb{N}\}$ *and* $\{n^{-1/2}R_i^n; n \in \mathbb{N}\}$ *are C-tight under the assumptions of Theorem 3.4.1.*

Because $\{n^{-1/2}R_i^n(\cdot); n \in \mathbb{N}\}$ is C-tight and $\tau_i^n - \sigma_i^n = O_p(n^{-1/2})$,

$$\mathbb{P}\left(n^{-1/2}R_i^n(\tau_i^n) - n^{-1/2}R_i^n(\sigma_i^n) < -\epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Combining the above with (3.47) allows us to conclude that

$$\mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.48}$$

Similarly, by the C-tightness of $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ and that $\tau_i^n - \sigma_i^n = O_p(n^{-1/2})$, we have

$$\mathbb{P}\left(\Gamma_i^n(\sigma_i^n, \tau_i^n] > \epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.49}$$

Combining (3.46), (3.48) and (3.49) yields

$$\mathbb{P}\left(\tau_i^n \leq T\right) \to 0 \quad \text{as} \quad n \to \infty$$

which in turn implies

$$\Delta_i^n(\cdot) \equiv \widehat{Q}_i^n(\cdot) - r_i(\cdot)\widehat{Q}^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{3.50}$$

for all $i \in \mathcal{I}$.

4. Diffusion Limits An application of Theorem 4.1 of [Pang *et al.*, 2007] together with (3.9), (3.34), (3.35), (3.40) and (3.50) allows us to establish the many-server heavy-traffic limit for $\{\widehat{X}_i^n; n \in \mathbb{N}\}$:

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

where $\widehat{X}_i$ satisfies the differential equation (3.12). Then apply the convergence-together lemma with (3.50) we conclude

$$\left(\widehat{X}_1^n, \ldots, \widehat{X}_K^n, \widehat{Q}_1^n, \ldots, \widehat{Q}_K^n\right) \Rightarrow \left(\widehat{X}_1, \ldots, \widehat{X}_K, \widehat{Q}_1, \ldots, \widehat{Q}_K\right) \quad \text{in} \quad \mathcal{D}^{2K} \quad \text{as} \quad n \to \infty, \tag{3.51}$$

where the limiting processes $\widehat{Q}_i$ are given in (3.13). The FCLT for the PWT processes follows by applying the two-parameter version of Puhalskii's invariance principle for first passage time; see Theorem 2.9 in [Talreja and Whitt, 2009]. □

Proof of Theorem 3.4.2.

It suffices to show the SSC associated with the HLDR rule. The rest of proof resembles that of Theorem 3.4.1. Let $a_i^n(t)$ denote the inter-arrival time between the HoL customer in queue i and the most recent class-i customer who entered service. By the way the HLDR control operates,

$$U^n(t) - a_i^n(t)/v_i(t) < U_i^n(t)/v_i(t) \leq U^n(t). \tag{3.52}$$

Without customer abandonments, it is clear that $a_i^n(t)$ is first-order stochastically dominated by exponential random variable with rate $n\lambda_i^{\downarrow}$. In the presence of impatient customers, $a_i^n(t)$ is stochastically dominated by an exponential random variable with rate $n\lambda_i^{\downarrow}F_i^c(T)$. To proceed, we would like to establish a uniform bound for $a_i^n(t)$ over all $t \leq T$. For this purpose, we make the following observation: (i) For each class, the number of arrivals over any compact time interval is $O_p(n)$; and (ii) the maximum of n i.i.d. exponential random variables is $O_p(\log n)$. As an immediate consequence, we have $\sup_{t\leq T}\{a_i^n(t)\} = O(n^{-1}\log n)$. Combining with (3.52) yields

$$U_i^n(t)/v_i(t) = U^n(t) - O_p(n^{-1}\log n),$$

or, equivalently,

$$\widehat{U}_i^n(t) = v_i(t)\widehat{U}^n(t) - O_p(n^{-1/2}\log n), \tag{3.53}$$

where we defined $\widehat{U}^n(t) \equiv n^{1/2}U^n(t)$.

For any $x \in \mathcal{D}$, let $x[t_1, t_2] \equiv x(t_2) - x(t_1-)$. In addition, let $R_i^{n,t}(s)$ denote the number of class-i customers who arrived after time t but have abandoned in the interval $[t, s]$. With the HLDR control, the queue-length processes satisfy

$$Q_i^n(t) = A_i^n[t - U_i^n(t), t] - R_i^{n,t-U_i^n(t)}[t - U_i^n(t), t]. \tag{3.54}$$

Define $\widehat{R}_i^{n,t}(s) \equiv n^{-1/2}R_i^{n,t}(s)$. By (3.54) we have

$$\begin{aligned}\widehat{Q}_i^n(t) &= \widehat{A}_i^n[t - U_i^n(t), t] + n^{1/2}\int_{t-U_i^n(t)}^{t}\lambda_i(u)\mathrm{d}u - \widehat{R}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t] \\ &= \widehat{A}_i^n[t - U_i^n(t), t] + \lambda_i(t)\widehat{U}_i^n(t) - \widehat{R}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t] + e_i^n(t)\end{aligned} \tag{3.55}$$

where

$$e_i^n(t) \equiv n^{1/2}\int_{t-U_i^n(t)}^{t}\lambda_i(u)\mathrm{d}u - n^{1/2}\lambda_i(t)U_i^n(t).$$

By using the first equality in (3.55), we conclude the stochastic boundedness of $\{n^{1/2}U_i^n; n \in \mathbb{N}\}$; in particular, we have $U_i^n \Rightarrow 0$ as $n \to \infty$, which in turn implies the asymptotic negligibility of $\widehat{A}_i^n[t - U_i^n(t), t]$, $\widehat{R}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t]$ and $e_i^n(t)$ over $[0, T]$. In view of the second equality in (3.55), we get

$$\widehat{Q}_i^n(\cdot) - \lambda_i(\cdot)\widehat{U}_i^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{3.56}$$

Next, combining (3.53) and (3.56) yields

$$\widehat{Q}_i^n(\cdot) - \lambda_i(\cdot)v_i(\cdot)\widehat{U}^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.$$

This is essentially the desired SSC result; that is,

$$\widehat{Q}_i^n(\cdot) - \gamma(\cdot)^{-1}v_i(\cdot)\lambda_i(\cdot)\widehat{Q}^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \quad \text{for} \quad i \in \mathcal{I}.$$

□

Chapter 4

The Overloaded Regime

In this chapter, we study the service-level differentiation problem in the overloaded regime. For mathematical convenience, we assume that arrival process for the class-i customers A_i starts at time $-w_i$; i.e., $t_i^0 = -w_i$. 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 4.2.2.

4.1 A Time-Varying Square-Root Staffing Rule

We now introduce the 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. Because the objective is to stabilize the expected delay at any given point in time at a target w, we will need to set the staffing levels to a modified version of (3.2), namely,

$$m_{\rm DIS}(t) \equiv \int_0^t \underbrace{F^c(w)\lambda(u-w)}_{\text{effective arrival rate}} G^c(t-u)\mathrm{d}u, \tag{4.1}$$

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 $m_{\rm DIS}(t)$ 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

$$m(t) \equiv \sum_{i=1}^{K} m_i(t), \qquad \text{where} \quad m_i(t) \equiv \int_0^t \underbrace{F_i^c(w_i)\lambda_i(u-w_i)}_{\text{effective class-}i\text{ arrival rate}} \; e^{-\mu_i(t-u)}\mathrm{d}u, \tag{4.2}$$

where each term in the sum of (4.2) is obtained by replacing (F, w, G, λ) in (4.1) 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 α_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 α_i. We envision a staffing function consisting of two pieces, namely,

$$s(t) = \left\lceil m(t) + \sqrt{\bar{\lambda}}c(t) \right\rceil. \tag{4.3}$$

where $c(t) \equiv c(t, \alpha_1, \ldots, \alpha_K)$ 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 (4.3) as time-varying square-root-staffing (TV-SRS).

Remark 4.1.1 (Role of the Safety Staffing Functions c) *Note that the first-order nominal term $m(t)$ in (4.3) 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 μ_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 (2.8). Hence, the overall staffing level s depends on probability targets $(\alpha_1, \ldots, \alpha_K)$ only through c.*

4.2 A Time-Varying Dynamic Prioritization Scheduling Rule

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 first-come first-serve within each class, it suffices to track the HWTs, namely, $(U_1(t), \ldots, U_K(t))$.

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

$$i^\star \in \arg\max_{1\le i\le K} \left\{ \underbrace{\frac{U_i(t)}{w_i}}_{\text{normalized HWT}} + \frac{1}{\sqrt{\lambda}}\,\kappa_i(t) \right\}, \tag{4.4}$$

where the first term $U_i(t)/w_i$ 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 α_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

$$U(t) \equiv \frac{U_{i^\star}(t)}{w_{i^\star}} + \frac{1}{\sqrt{\lambda}}\kappa_{i^\star}(t). \tag{4.5}$$

Remark 4.2.1 (Understanding TV-DPS) *The first-order term $U_i(t)/w_i$ is designed to guarantee that the class-i delay is close to its target w_i (it is controlling the relative delay imbalance $(U_i(t)-w_i)/w_i$, rather than the absolute delay imbalance). The second-order term $(1/\sqrt{\lambda})\kappa_i(\cdot)$ helps accomplish the class-dependent probability target α_i. Intuitively, such a control function κ_i should satisfy the following properties:*

(i) Monotonicity. For fixed time t, $\kappa_i(t)$ 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 $\alpha_i > 0.5$ ($\alpha_i \le 0.5$), the fine-tuning prioritization regulator κ_i should satisfy $\kappa_i(t) < 0$ ($\kappa_i(t) \ge 0$) (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 $U_i(t)/w_i$).

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 first-come first-serve scheduling policy.*

Remark 4.2.2 (Relaxation of the assumption on 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 (4.5) for* $K = 2$*. At time* w_3 *and beyond, the TV-DPS rule is implemented in the usual way for all classes.*

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 κ_i; using the FCLT limit, we will next obtain the exact formulas of c and κ_i so that the TPoD-based service-level constraints are asymptotically satisfied as the scale increases.

4.3 Achieving Service-Level Differentiation

In this section, we present our main results. §4.3.1 gives the asymptotic framework and states the MSHT FCLT and FWLLN results for the multiclass V model operating under

the TV-SRS and TV-DPS policies introduced in §§4.1–4.2. In §4.3.2 we utilize the FCLT results to obtain the desired control factors κ_i and c and show that they asymptotically achieve TPoD-based service-level differentiation and performance stabilization. All proofs are given in the appendix.

4.3.1 Many-Server FCLT Limits

Again, 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 $A_i^n(t)$ 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

$$s^n(t) = \lceil nm(t) + \sqrt{n}c(t) \rceil, \tag{4.6}$$

where m and c are the offered-load function in (4.2) and safety staffing term (yet to be determined).

Let U_i^n and V_i^n be the class-i HWT and PWT in the n^{th} model. Our TV-DPS satisfies

$$i^* \in \arg\max_{1\le i\le K} \left\{ \frac{U_i^n(t)}{w_i} + \frac{1}{\sqrt{n}}\, \kappa_i(t) \right\}, \tag{4.7}$$

where κ_i is a control function yet to be determined.

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

$$\Lambda_i(t) \equiv \int_{-w_i}^{t} \lambda_i(u)\mathrm{d}u, \quad \bar{A}_i^n(t) \equiv n^{-1}A_i^n(t) \quad \text{and} \quad \widehat{A}_i^n(t) \equiv n^{-1/2}\left(A_i^n(t) - n\Lambda_i(t)\right). \tag{4.8}$$

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

$$(\bar{A}_i^n(\cdot), \widehat{A}_i^n(\cdot)) \Rightarrow (\Lambda_i(\cdot), \widehat{A}_i(\cdot)) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty, \tag{4.9}$$

for $\widehat{A}_i(\cdot) \equiv \mathcal{W}_{\lambda_i} \circ \Lambda_i(\cdot)$, where $x \circ y(t) \equiv x(y(t))$, $\mathcal{W}_{\lambda_i}$ being a standard Brownian motion.

Remark 4.3.1 (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

$$\widehat{A}_i^n(\cdot) \Rightarrow c_{\lambda_i}\mathcal{W}_{\lambda_i} \circ \Lambda_i(\cdot) \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad n \to \infty,$$

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). In this case, our FCLT limits in Theorem 4.3.1 can be easily adjusted by simply multiplying $\mathcal{W}_{\lambda_j}$ by the constant c_{λ_i}. For NHPPs, $c_{\lambda_i} = 1$.

Following the notations in §2.1, we use $Q_i^n(t)$ and $B_i^n(t)$ 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 $X_i^n(t)$, represents the total number of class-i customers in system at time t. We now define their corresponding CLT-scaled versions

$$\widehat{B}_i^n(t) \equiv n^{-1/2}\left(B_i^n(t) - nm_i(t)\right), \ \widehat{Q}_i^n(t) \equiv n^{-1/2}\left(Q_i^n(t) - nq_i(t)\right)$$
$$\text{and} \quad \widehat{X}_i^n(t) \equiv n^{-1/2}\left(X_i^n(t) - nx_i(t)\right)$$

where m_i is given by (4.2), $q_i(t) \equiv \int_{t-w_i}^t F_i^c(t-u)\lambda_i(u)\mathrm{d}u$, and $x_i \equiv m_i + q_i$. In addition, let

$$\widehat{U}_i^n(t) \equiv n^{1/2}\left(U_i^n(t) - w_i\right) \quad \text{and} \quad \widehat{V}_i^n(t) \equiv n^{1/2}\left(V_i^n(t) - w_i\right) \tag{4.10}$$

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

$$\widehat{U}^n(t) \equiv n^{1/2}\left(U^n(t) - 1\right).$$

Theorem 4.3.1 (MSHT FCLT limits under TV-SRS and TV-DPS) *Suppose the system operates under TV-SRS in* (4.6) *and TV-DPS in* (4.7)*. Then there is a joint convergence for the CLT-scaled processes:*

$$\begin{aligned}&\left(\widehat{U}^n, \widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{U}_1^n, \ldots, \widehat{U}_K^n, \widehat{V}_1^n \ldots, \widehat{V}_K^n, \widehat{X}_1^n, \ldots, \widehat{X}_K^n, \widehat{Q}_1^n, \ldots, \widehat{Q}_K^n\right)\\ \Rightarrow &\left(\widehat{U}, \widehat{B}_1, \ldots, \widehat{B}_K, \widehat{U}_1, \ldots, \widehat{U}_K, \widehat{V}_1 \ldots, \widehat{V}_K, \widehat{X}_1, \ldots, \widehat{X}_K, \widehat{Q}_1, \ldots, \widehat{Q}_K\right)\end{aligned} \tag{4.11}$$

in $\mathcal{D}^{5K+1}$ as $n \to \infty$, where the FCLT limits on the right-hand side are well-defined stochastic processes.

(i) *The limiting processes* $(\widehat{U}, \widehat{B}_1, \ldots, \widehat{B}_K)$ *jointly satisfy the following set of* K *OU type stochastic integral equations and one linear equation, namely,*

$$\widehat{B}_i(t) + \eta_i(t)\widehat{U}(t) = -\int_0^t \mu_i \widehat{B}_i(u)\mathrm{d}u - \int_0^t \psi_i(u)\widehat{U}(u)\mathrm{d}u + \int_0^t \psi_i(u)\kappa_i(u)\mathrm{d}u + \eta_i(t)\kappa_i(t) + G_i(t) \quad \textit{for} \quad i = 1, \ldots, K, \quad \textit{and} \quad \sum_{i=1}^{K} \widehat{B}_i(t) = c(t), \tag{4.12}$$

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

$$G_i(t) \equiv \widehat{E}_{i,1}(t) + \widehat{E}_{i,2}(t) - \widehat{D}_i(t), \quad \widehat{E}_{i,1}(t) \equiv F_i^c(w_i)\int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)}\mathrm{d}\mathcal{W}_{\lambda_i}(u),$$
$$\widehat{E}_{i,2}(t) \equiv \sqrt{F_i^c(w_i)F_i(w_i)}\int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)}\mathrm{d}\mathcal{W}_{\theta_i}(u), \quad \widehat{D}_i(t) \equiv \int_0^t \sqrt{\mu_i m_i(u)}\mathrm{d}\mathcal{W}_{\mu_i}(u), \tag{4.13}$$

and $\mathcal{W}_{\lambda_i}, \mathcal{W}_{\theta_i}, \mathcal{W}_{\mu_i}$ *are independent standard Brownian motions.*

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

$$\widehat{U}_i(t) \equiv w_i(\widehat{U}(t) - \kappa_i(t)), \quad \textit{and} \quad \widehat{V}_i(t) = w_i(\widehat{U}(t + w_i) - \kappa_i(t + w_i)). \tag{4.14}$$

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

$$\widehat{Q}_{i,1}(t) \equiv \int_{t-w_i}^{t} F_i^c(t - u)\sqrt{\lambda_i(u)}\mathrm{d}\mathcal{W}_{\lambda_i}(u),$$
$$\widehat{Q}_{i,2}(t) \equiv \int_{t-w_i}^{t} \sqrt{F_i^c(t - u)F_i(t - u)\lambda_i(u)}\mathrm{d}\mathcal{W}_{\theta_i}(s),$$
$$\widehat{Q}_{i,3}(t) \equiv \lambda_i(t - w_i)F_i^c(w_i)\widehat{U}_i(t).$$

(iv) *Finally, the FCLT limit for number in system is given by* $\widehat{X}_i(t) = \widehat{B}_i(t) + \widehat{Q}_i(t)$.

Remark 4.3.2 (SSC and Separation of Variability) *Theorem 4.3.1 provides the FCLT limits for waiting times and queue lengths under TV-SRS and TV-DPS with the second-order terms* c *and* κ_i *yet to be determined. Such FCLT results will be used later to achieve asymptotic performance differentiation and stabilization. Part (ii) of Theorem 4.3.1 gives a nice SSC result: The diffusion limits* $(\widehat{U}, \widehat{B}_1, \ldots, \widehat{B}_K)$ *satisfy the* $(K + 1)$*-dimensional*

stochastic differential equation (SDE), and according to (4.14), 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 (4.5) do not differ much from each other under the TV-DPS policy. In addition, 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. We will see later in Proposition 4.3.1, these sources of randomness jointly contribute to the variability of the one-dimensional process* $\widehat{U}$*.*

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

$$\bar{B}_i^n(t) \equiv n^{-1}B_i^n(t), \quad \bar{Q}_i^n(t) \equiv n^{-1}Q_i^n(t), \quad \bar{X}_i^n(t) \equiv n^{-1}X_i^n(t) \quad \text{for} \quad 1 \le i \le K. \tag{4.15}$$

The next result is a direct consequence of Theorem 4.3.1.

Corollary 4.3.1 (FWLLN) *Suppose that the system operates under TV-SRS in (4.6) and TV-DPS in* (4.7)*. Then we have the joint convergence for the LLN-scaled processes*

$$\begin{aligned} &\left(\bar{B}_1^n, \ldots, \bar{B}_K^n, \bar{Q}_1^n, \ldots, \bar{Q}_K^n, \bar{X}_1^n, \ldots, \bar{X}_K^n, U_1^n, \ldots, U_K^n, V_1^n, \ldots, V_K^n\right) \\ &\Rightarrow (m_1, \ldots, m_K, q_1, \ldots, q_K, x_1, \ldots, x_K, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}) \quad \textit{in} \quad \mathcal{D}^{5K} \end{aligned} \tag{4.16}$$

as $n \to \infty$*, where the function* $\mathfrak{e}(t) = 1$*.*

Below we provide a proof sketch of the theorem. The details are given in §4.6.

Proof sketch of Theorem 4.3.1. Step 1: We first show that each component within the curly bracket in (4.7) is at most $O(n^{-1}\log n)$ away from the frontier process, that is, $U_i^n(t)/w_i + n^{-1/2}\kappa_i(t) = U^n(t) + O(n^{-1}\log n)$. 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 (2.4) to obtain a simple relation between $\widehat{U}_i^n$ and $\widehat{B}_i^n$. Based on the fact that the difference between $\widehat{U}_i^n(t)$ and $w_i(\widehat{U}^n(t) - \kappa_i(t))$ can be made arbitrarily small for n large enough, we are able to establish a set of K differential equations and one linear equation jointly satisfied by $(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{U}^n)$. This allows us to apply the Gronwall's

inequality to establish the stochastic boundedness of the sequence $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{U}^n); n \in \mathbb{N}\}$, 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 (4.13) for the corresponding CLT-scaled processes. Applying the continuous mapping theorem again with these Brownian limits yields the joint convergence of $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{U}^n)\}$. 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 U_i^n and then applying the continuous mapping theorem. □

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

$$\eta(t) \equiv \sum_{i=1}^{K} \eta_i(t) = \sum_{i=1}^{K} w_i \lambda_i (t - w_i) F_i^c(w_i). \tag{4.17}$$

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 (4.12) allows us to write $\widehat{B}_i$ as a function of $\widehat{U}$. Plugging them into the equation $\sum_{i=1}^{K} \widehat{B}_i(t) = c(t)$ plus some algebraic simplifications yields the result below.

Proposition 4.3.1 (Distribution of the frontier process $\widehat{U}$) *The process $\widehat{U}$ uniquely solves the following stochastic Volterra equation (SVE)*

$$\widehat{U}(t) = \int_0^t L(t,s)\widehat{U}(s)\mathrm{d}s + \int_0^t J(t,s)\mathrm{d}\mathcal{W}(s) + K(t), \tag{4.18}$$

where

$$L(t,s) \equiv \frac{\sum_{i=1}^{K} \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right)}{\eta(t)}, \tag{4.19}$$

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

$$K(t) \equiv \frac{\sum_{i=1}^{K} \left(\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right) \kappa_i(s) ds\right) - c(t)}{\eta(t)}, \tag{4.20}$$

$\mathcal{W}$ is a standard Brownian motion. In addition, $\widehat{U}$ is a Gaussian process with

(i) *mean* $M_{\widehat{U}}(t) \equiv \mathbb{E}[\widehat{U}(t)]$, $0 \le t \le T$, *uniquely solving the fixed-point equation (FPE)*

$$M_{\widehat{U}} = \Gamma(M_{\widehat{U}}), \qquad \text{where} \quad \Gamma(M_{\widehat{U}})(t) \equiv \int_0^t L(t,s) M_{\widehat{U}}(s) ds + K(t), \tag{4.21}$$

(ii) *covariance* $C_{\widehat{U}}(t,s) \equiv Cov(\widehat{U}(t), \widehat{U}(s))$, $0 \le s,t \le T$, *uniquely solving the FPE*

$$C_{\widehat{U}} = \Theta(C_{\widehat{U}}),$$

where the operator Θ *is defined as*

$$\begin{aligned} \Theta(C_{\widehat{U}})(t,s) \equiv & - \int_0^t \int_0^s L(t,u) L(s,v) C_{\widehat{U}}(u,v) dv du \\ & + \int_0^t L(t,u) C_{\widehat{U}}(u,s) du + \int_0^s L(s,v) C_{\widehat{U}}(t,v) dv \qquad (4.22) \\ & + \int_0^{s \wedge t} J(t,u) J(s,u) du. \qquad (4.23) \end{aligned}$$

The FCLT for $\widehat{U}$ 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 4.3.3.

Remark 4.3.3 (A closer look at the SVE (4.18)**)**

(i) ***Analytic solutions in special cases.*** *Such an SVE* (4.18) *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* (4.18) *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 §4.4 for discussions of this important special case).*

(ii) ***Variability.*** *The SVE is driven by the Brownian motion* $\mathcal{W}$*, which rises from aggregating all* $3K$ *independent Brownian motions* $\mathcal{W}_{\lambda_i}, \mathcal{W}_{\theta_i}, \mathcal{W}_{\mu_i}$, $1 \le i \le K$ *in* (4.13)*; see the proof of Proposition 4.3.1 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 κ_i and c, which only appears in K. Hence, varying κ_i and c will affect the mean of $\widehat{U}$, but not its variance. This is a crucial observation, because as will become clear later in §4.3.2, (i) computing the variance of $\widehat{U}$ (which is uncontrollable via c and κ_i) and (ii) appropriately shifting the mean of $\widehat{U}$ (by adjusting our control functions) are critical in achieving desired class-dependent service levels.*

(iv) ***Algorithms.*** *We prove Proposition 4.3.1 by showing that the operators Γ and Θ are both contractions in appropriate functional spaces. In addition, our proof naturally leads to effective numerical algorithms for computing $M_{\widehat{U}}$ and $C_{\widehat{U}}$ (in fact, our algorithms converge geometrically fast). It is obvious that $M_{\widehat{U}}(t) = 0$ if $K(t) = 0$ (because a zero function now solves the FPE (4.21)). This will indeed be the case considered later in §4.3.2.*

4.3.2 The Proposed Solution

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

$$\begin{aligned}
\alpha_i &\equiv \mathbb{P}(V_i^n(t) > w_i) = \mathbb{P}(\widehat{V}_i^n(t) > 0) \\
&\approx \mathbb{P}(\widehat{V}_i(t) > 0) = \mathbb{P}(\widehat{U}(t + w_i) - \kappa_i(t + w_i) > 0) \\
&= \mathbb{P}\left(\mathcal{N}\left(M_{\widehat{U}}(t + w_i), \sigma^2_{\widehat{U}}(t + w_i)\right) > \kappa_i(t + w_i)\right) \\
&= \mathbb{P}\left(\mathcal{N}(0, 1) > \frac{\kappa_i(t + w_i) - M_{\widehat{U}}(t + w_i)}{\sigma_{\widehat{U}}(t + w_i)}\right)
\end{aligned} \tag{4.24}$$

for all $t \geq -w_i$, where $\mathcal{N}(\mu, \sigma^2)$ is a normal random variable with mean μ and variance σ^2, $\sigma_{\widehat{U}}(t) = \sqrt{\mathrm{Var}(\widehat{U}(t))} = \sqrt{C_{\widehat{U}}(t, t)}$ is the standard deviation of $\widehat{U}(t)$ at t. Equation (4.24) further simplifies to

$$\mathbb{P}\left(\mathcal{N}(0, 1) > \frac{\kappa_i(t) - M_{\widehat{U}}(t)}{\sigma_{\widehat{U}}(t)}\right) \approx \alpha_i, \qquad t \geq 0, \tag{4.25}$$

in which case we should choose appropriate control functions κ_i and c so that

$$\kappa_i(t) - M_{\widehat{U}}(t) = z_{1-\alpha_i}\sigma_{\widehat{U}}(t), \tag{4.26}$$

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

One obvious solution to (4.26) is that, for any κ_i, we can choose c appropriately so that $K(t)$ in (4.20) is set to 0, so that $M_{\widehat{U}}(t) = 0$ for all t (note that FPE (4.21) now has a unique solution $M_{\widehat{U}}(t) = 0$ when $K(t) = 0$), and this leads to the control formulas below in (4.27)–(4.28). We next show that these control functions are indeed the unique solutions to (4.26).

Proposition 4.3.2 (Asymptotically unique control functions) *The condition* (4.26) *is satisfied if and only if*

$$c(t) = \sum_{i=1}^{K} \left\{ \eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right) \kappa_i(s) ds \right\}, \tag{4.27}$$

$$\kappa_i(t) = z_{1-\alpha_i} \sigma_{\widehat{U}}(t), \qquad 1 \leq i \leq K. \tag{4.28}$$

See the appendix for the proof for Proposition 4.3.2. The safety staffing term c is indeed uniquely given by (4.27). However, to be rigorous, we should say that the prioritization regulators $\kappa_1, \ldots, \kappa_K$ are unique up to adding any common function Δ, that is, applying any $\tilde{\kappa}_i(t) = \kappa_i(t) + \Delta(t)$ for $1 \leq i \leq K$ which will not make a difference in our TV-DPS rule.

Remark 4.3.4 (Structure of the control functions) *The main idea here is that we choose appropriate control functions c and κ_i to tilt the mean of the error term $\widehat{V}_i^n(t)$ (rather than the mean of $V_i^n(t)$), so that asymptotically the probability mass of $\{\widehat{V}_i^n(t) > 0\}$ (or $\{V_i^n(t) > w_i\}$) can be set to desired α_i at all time t. We observe from* (4.27) *that the second-order safety staffing term c depends on α_i through the second-order prioritization regulator κ_i, and κ_i depends on α_i through z_{α_i}. Consistent with Remark 4.2.1, κ_i 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 $\alpha_i < 0.5$ ($\alpha_i > 0.5$). 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*

$\kappa_1(t) - \kappa_2(t) > 0$ is increasing in $\sigma_{\widehat{U}}(t)$, which characterizes the system's overall stochastic variability (recall from Remark 4.3.3 that the variability of $\widehat{U}$ 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 (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 the TV-SRS and TV-DPS formulas.

The next theorem establishes the asymptotic effectiveness of our methods.

Theorem 4.3.2 (Asymptotic service differentiation and performance stabilization) *Under TV-SRS* (4.6) *and TV-DPS* (4.7) *with $c_i(\cdot)$ and $\kappa_i(\cdot)$ specified in* (4.27) *and* (4.28)*, we have the following asymptotic stability results: TPoDs for PWT and HWT are both asymptotically stabilized for all classes:*

$$\mathbb{P}(V_i^n(t) > w_i) \to \alpha_i \quad \textit{and} \quad \mathbb{P}(U_i^n(t) > w_i) \to \alpha_i \qquad \textit{as} \quad n \to \infty \tag{4.29}$$

for $1 \le i \le K, 0 < t \le T$.

4.4 The Case of Class-Independent Service Rate

This section provides a more detailed discussion of the important 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 $1 \le i \le K$. 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 (4.18) degenerates to a much more tractable OU process with time-varying drift and volatility. We summarize our results below.

Corollary 4.4.1 (Frontier process $\widehat{U}$ when service rates are class independent) *Suppose $\mu_i = \mu, 1 \le i \le K$, then*

(*i*) *the limiting frontier process* $\widehat{U}$ *satisfies the one-dimensional OU type SDE*

$$\eta(t)\widehat{U}(t) = -\int_0^t \eta(u)\widehat{U}(u)\mathrm{d}u + \mathcal{S}(t) + G(t), \tag{4.30}$$

where $G(t) \equiv \sum_{i=1}^K G_i(t)$*, for* $G_i(t)$ *being the Brownian-driven terms given in Theorem 4.3.1, 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* (4.30) *has a unique solution*

$$\widehat{U}(t) = \frac{1}{R(t)}\left(\int_0^t e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}\tilde{J}(u)\mathrm{d}\mathcal{W}(u) + \int_0^t e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}R(u)\mathrm{d}K(u) + \int_0^t e^{\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}K(u)\mathrm{d}R(u)\right), \tag{4.31}$$

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

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

(*iii*) *The variance of* $\widehat{U}(t)$ *is*

$$\sigma^2_{\widehat{U}}(t) \equiv \mathrm{Var}\left(\widehat{H}(t)\right) = \frac{1}{R^2(t)}\int_0^t e^{2\int_u^t \frac{\tilde{L}(v)}{R(v)}dv}\tilde{J}^2(u)\mathrm{d}u. \tag{4.32}$$

We next consider some special cases to obtain insights.

Corollary 4.4.2 (Constant arrival rates) *When* $\lambda_i(t) = \lambda_i$*, we have*

$$m_i(t) \sim m_i \equiv \frac{\lambda_i F_i^c(w_i)}{\mu}, \qquad c(t) \sim c \equiv \sum_{i=1}^K \frac{w_i\lambda_i f_i(w_i)}{\mu}\kappa_i, \tag{4.33}$$

$$\kappa_i(t) \sim \kappa_i \equiv z_{1-\alpha_i}\sqrt{\frac{\sum_{j=1}^K \lambda_j F_j^c(w_i)}{\left(\sum_{j=1}^K \lambda_j f_j(w_j)w_j\right)\left(\sum_{j=1}^K \lambda_j F_j^c(w_j)w_j\right)}}. \tag{4.34}$$

where we say $f(t) \sim g(t)$ *if* $f(t)/g(t) \to 1$ *as* $t \to \infty$*.*

Remark 4.4.1 (Average staffing and prioritization levels) *The constants in* (4.33) *and* (4.34) *can be used to compute the required average number of servers and scheduling threshold. When* $K = 1$*, our staffing formula* (4.33) *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* Δw_i*, or reducing* α_i *by* $\Delta\alpha_i$*), how many extra servers are need and how much should the scheduling threshold* κ_i *be adjusted?*

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

Corollary 4.4.3 (The single-class case) *When* $K = 1$*, the second-order staffing term* $c(t)$ *simplifies to*

$$c(t) = z_{1-\alpha} e^{-\mu t} \left(Z(t) - (\mu - h_F(w)) \int_0^t Z(s) ds \right), \tag{4.35}$$

$$\text{with} \quad Z(t) \equiv e^{(\mu - h_F(w))t} \sqrt{\int_0^t e^{2h_F(w)} \left(F^c(w)\lambda(u-w) + \mu m(u) \right)}. \tag{4.36}$$

It is easy to check that (4.35) and (4.36) 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.

4.5 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 §4.5.1 we first consider a base model having two customer classes and state-independent service rates. We next give additional simulation experiments in §4.5.2, 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.

4.5.1 A Two-Class Base Model

Because sinusoidal functions capture the periodic structure in realistic arrival patterns, we consider sinusoidal arrival rates

$$\lambda_i(t) = \bar{\lambda}_i \left(1 + r_i \sin(\gamma_i t + \phi_i)\right), \qquad 1 \le i \le K, \tag{4.37}$$

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 $\bar{\lambda}_1(t) = 1, \bar{\lambda}_2(t) = 1.5, r_1 = 0.2, r_2 = 0.3, \gamma_1 = \gamma_2 = 1, \phi_1 = 0, \phi_2 = -1$. 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 $w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8$.

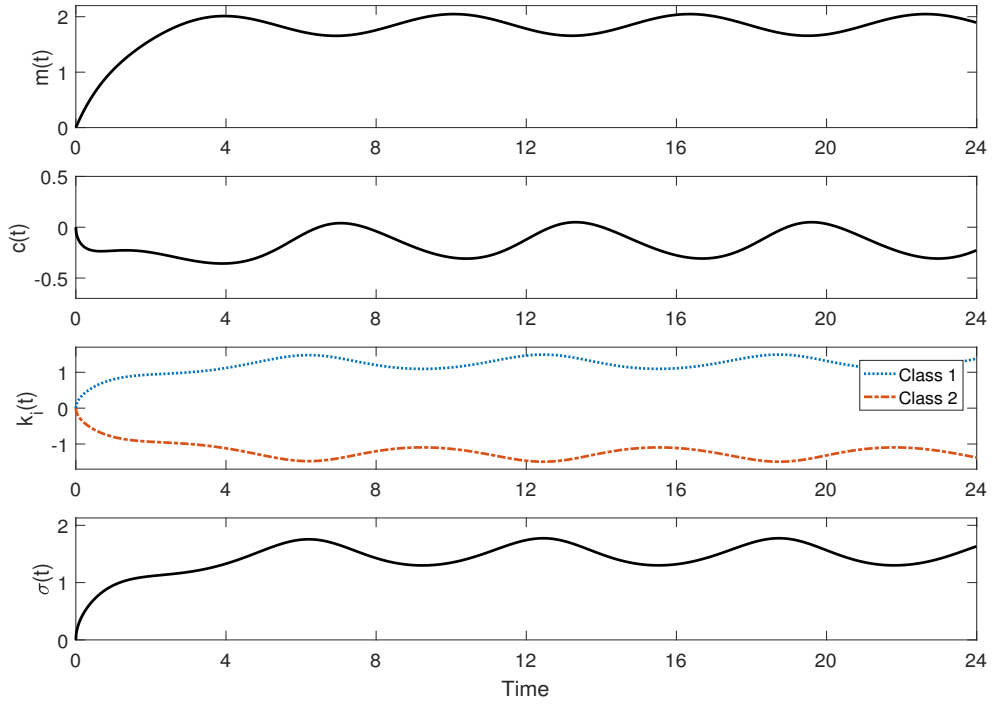

Figure 4.1: Computed control functions for a two-class base case: $m(t)$, $c(t)$, $\kappa_i(t)$ and $\sigma(t)$, $i = 1, 2$.

In Figure 4.1, 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 (4.2), the second-order staffing term $c(t)$ in (4.27), the second-order prioritization regulators (4.28), and the standard deviation process of $\widehat{U}$ in (4.31). Consistent with discussions in Remarks 4.2.1 and 4.3.4, 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.

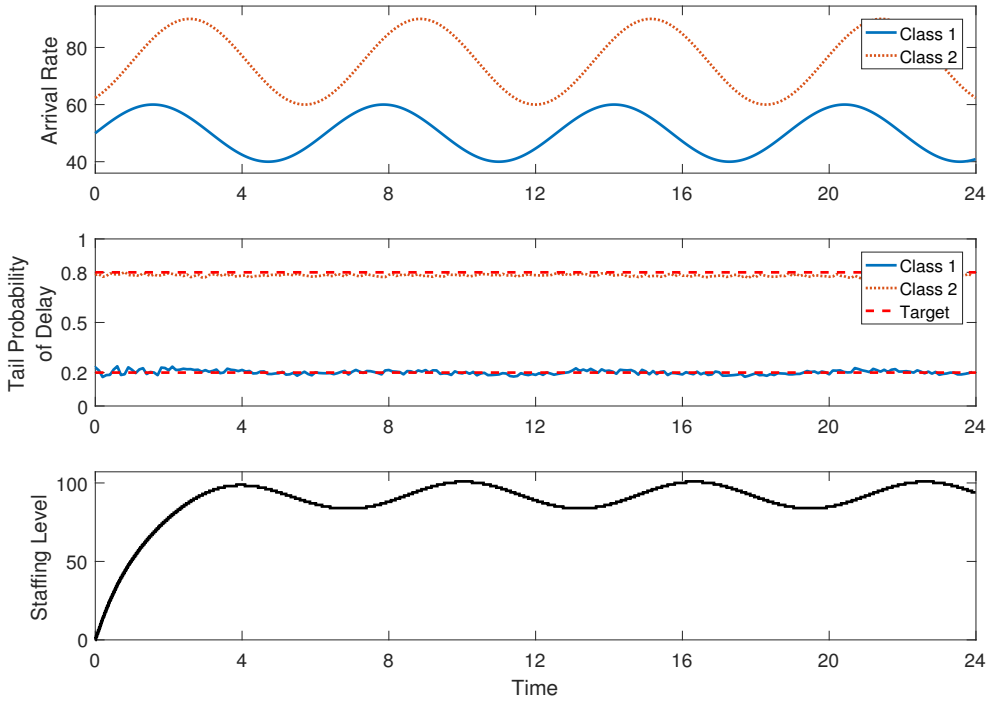

Figure 4.2: Simulation comparison for a two-class base case: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 5000 independent runs.

Using these control functions in Figure 4.1, 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 (4.5) using the control function κ_i given in Figure 4.1. We plot (i) arrival rates, (ii) simulations of TPoD, and (iii) staffing functions, in Figure 4.2, using a sampling resolution (i.e., step size)

$\Delta t = 0.01$. From a visual inspection of the middle panel of Figure 4.2, 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).

4.5.2 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 (§4.5.2.1), (ii) smaller system scale (§4.5.2.2), (iii) class-dependent service rates (§4.5.2.3), and (iv) a five-class example (§4.5.2.4).

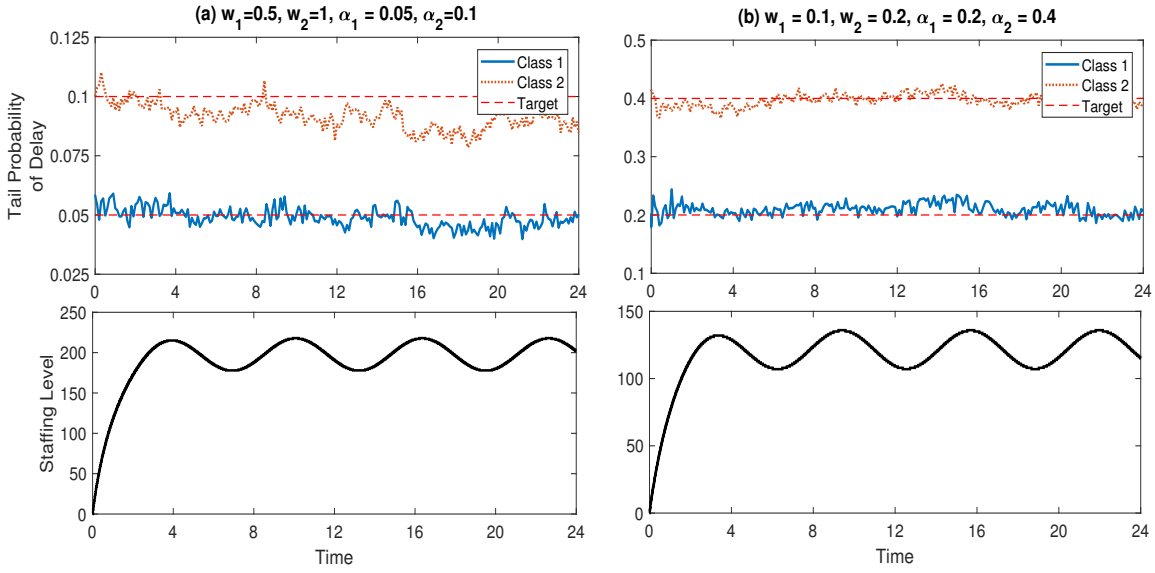

Figure 4.3: The two-class based model with high QoS targets: (a) $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.05$, $\alpha_2 = 0.1$ (left), (b) $w_1 = 0.1$, $w_2 = 0.2$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$ (right).

4.5.2.1 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 4.3 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 4.3 shows that TPoD's remain relatively stable in both cases.

4.5.2.2 Smaller arrival rates

Our method is based on asymptotic analysis of the V model when $n \to \infty$, 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.

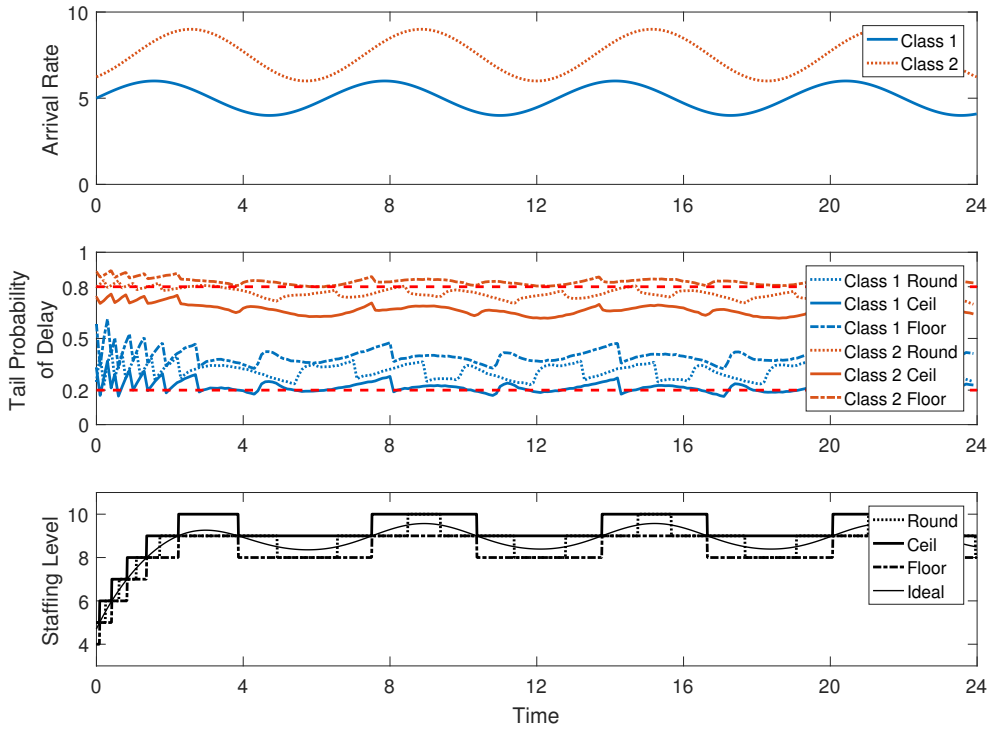

Figure 4.4: Simulation comparison for a small-scale two-class model: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $n = 5$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 20000 independent runs, under three staffing discretizations.

Figure 4.4 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.

4.5.2.3 Class-dependent service rates

Results in §4.3 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 $\mathrm{Var}(\widehat{U}(t))$ according to the contraction-based algorithm given in the appendix. Simulations show that our methods continue to achieve desired service-level differentiation and performance stabilization; see Figure 4.5.

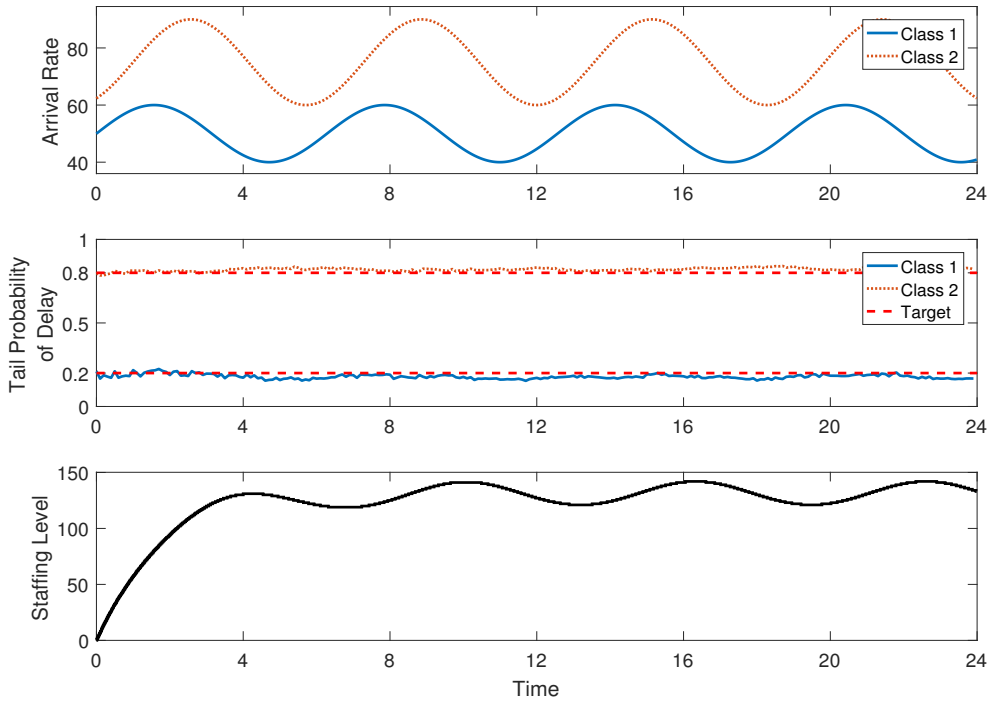

Figure 4.5: Simulation comparison for a two-class model with class-dependent rates: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 50$, $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$.

4.5.2.4 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 (4.37), exponential abandonment and service times. All model input parameters and QoS parameters are given in Table 4.1.

Table 4.1: Five Class Model: Class specific parameters and QoS target levels

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

The control functions are given in the left-hand panel of Figure 4.6. In this example, we intentionally let the sinusoidal arrival rates have class-dependent periods, frequencies, and relative amplitudes (see right-hand panel of Figure 4.6). Nevertheless, our method continues to successfully achieves stable TPoD-based service levels across all 5 classes.

4.6 Proof of Theorem 4.3.1

Because we assume each baseline arrival-rate function λ_i is bounded away from zero and infinity, we define $\lambda_i^{\downarrow} \equiv \inf_{0\le t\le T} \lambda(t) > 0$ and $\lambda_i^{\uparrow} \equiv \sup_{0\le t\le T} \lambda(t) < \infty$. The proof proceeds in four major steps, as indicated by the proof sketch presented in the main paper.

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

$$U^n(t) - a_i^n(t)/w_i < U_i^n(t)/w_i + n^{-1/2}\kappa_i(t) \le U^n(t), \tag{4.38}$$

where $a_i^n(t)$ is first-order stochastically dominated by an exponential random variable with rate $n\lambda_i^{\downarrow}F_i^c(T_i)$ for $T_i \equiv T + w_i$. By using the same argument as in the proof of Theorem

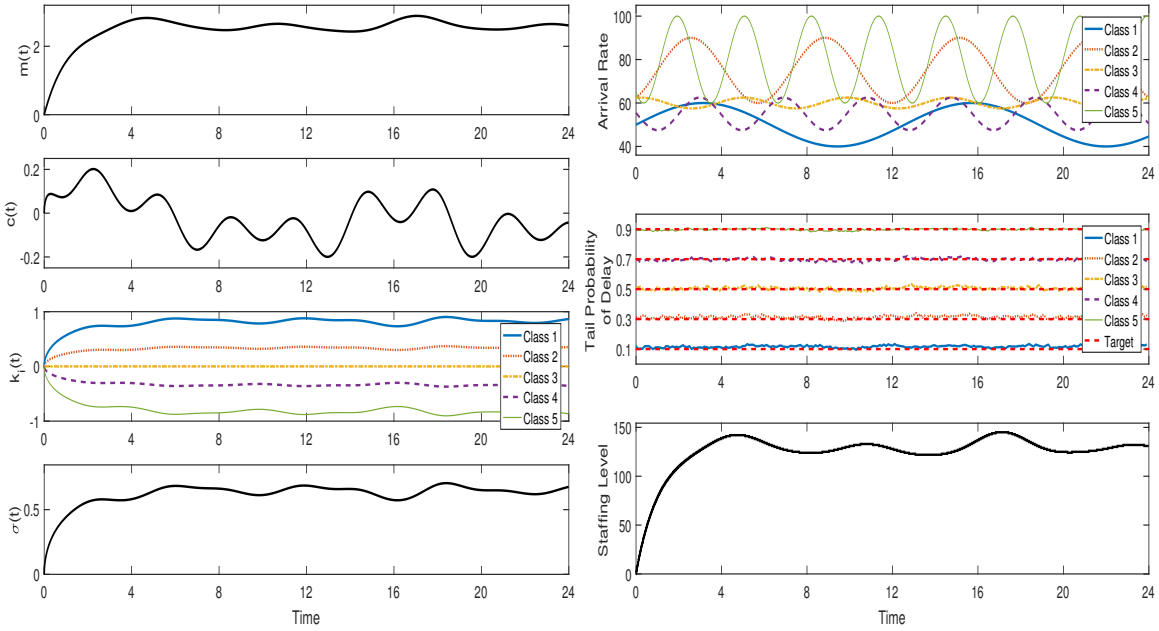

Figure 4.6: A five-class based model: (i) Computed control functions $m(t)$, $c(t)$, and $\kappa_i(t)$ for $i = 1, \ldots, 5$ (left), (ii) Simulation comparisons for TPoD $\mathbb{P}(V_i(t) > w_i)$, $i = 1, \ldots, 5$ (right), with $n = 50$, input and QoS parameters given in Table 4.1, and 5000 samples.

3.4.2, we conclude $\sup_{t \le T}\{a_i^n(t)\} = O(n^{-1} \log n)$. Combining with (4.38) yields

$$U_i^n(t)/w_i + n^{-1/2}\kappa_i(t) = U^n(t) - O_p(n^{-1} \log n),$$

or, equivalently,

$$\widehat{U}_i^n(t) = w_i(\widehat{U}^n(t) - \kappa_i(t)) - O_p(n^{-1/2} \log n), \tag{4.39}$$

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

We next argue that *under the TV-DPS rule*, the PWT and the HWT satisfy

$$V_i^n(t - U_i^n(t)) = U_i^n(t) + O_p(n^{-1} \log n). \tag{4.40}$$

The above relation evidently holds true for $K = 1$, because the PWT at the time of arrival of the HoL customer is the HoL customer's elapsed waiting time (i.e., the HWT) at time t plus the additional time until the next departure. For $K \ge 2$, we aims to establish (4.40) by showing that the number of service completions needed for the HoL customer of queue i to enter service is no greater than the sum of $K - 1$ geometric random variables. To see

this is the case, suppose at time t customer A enters service from queue i and customer B becomes the new HoL customer in queue i. Then customer B must have arrived at the system at time $t - U_i^n(t)$. By the definition of $a_i^n(t)$ it follows that customer A arrived at the system at time $t - U_i^n(t) - a_i^n(t)$. Suppose $\kappa_i \equiv 0,\ i \in \mathcal{I} \equiv \{1, \ldots, K\}$ (the case where κ_i are not zero functions can be analyzed in a similar fashion). Then under the TV-DPS policy, only those class-j customers who arrived during the interval

$$\left(t - \frac{w_j \left(U_i^n(t) + a_i^n(t)\right)}{w_i}, t - \frac{w_j U_i^n(t)}{w_i}\right) \tag{4.41}$$

could enter service prior to the time at which customer B enters service. To proceed, we make the following observation: The number of arrivals from a Poisson process with arrival rate $\lambda^{(2)}$ over an exponentially distributed time with rate $\lambda^{(1)}$ follows a geometric distribution with parameter $\frac{\lambda^{(1)}}{\lambda^{(1)}+\lambda^{(2)}}$. Now because the interval (4.41) has a length of $(w_j a_i^n / w_i)$, the number class-j customers who have a higher service priority over B is stochastically dominated by a geometric random variable with mean $\frac{w_j \lambda_j^{\uparrow}}{w_i \lambda_i^{\downarrow} F_i^c(T_i)}$. This shows that the number of customers (from the other classes) who have higher service priority over B can be bounded by the sum of $K-1$ geometric random variables. This gives (4.40) for $K \geq 2$.

Step 2: The FWLLN. Here we prove the desired FWLLN results by showing the stochastic boundedness of the corresponding CLT-scaled processes. In what follows, we will first prove that the sequence $\{(\widehat{B}_1^n, \ldots, \widehat{B}_K^n, \widehat{U}^n); n \in \mathbb{N}\}$ is stochastically bounded. To that end, introduce the LLN- and CLT-scaled empirical process

$$\begin{aligned} \bar{K}^n(t,x) &\equiv \frac{1}{n} \sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{X_i \leq x\}} \quad \text{for} \quad t \geq 0, \quad 0 \leq x \leq 1, \quad \text{and} \\ \widehat{K}^n(t,x) &\equiv \sqrt{n}\left(\bar{K}^n(t,x) - \mathbb{E}\left[\bar{K}^n(t,x)\right]\right) = \frac{1}{\sqrt{n}} \left(\sum_{k=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{X_i \leq x\}} - x\right), \end{aligned} \tag{4.42}$$

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0,1]$. [Krichagina and Puhalskii, 1997] have shown that $\widehat{K}^n \Rightarrow \widehat{K}$ in $\mathcal{D}_{\mathcal{D}}$ as $n \to \infty$, where $\widehat{K}$ is the standard Kiefer process. Paralleling (3.3) - (3.6) in [Aras *et al.*, 2018], we break the enter-service process $E_i^n(t)$ in (2.4) into three pieces, namely,

$$E_i^n(t) = E_{i,1}^n(t) + E_{i,2}^n(t) + E_{i,3}^n(t), \tag{4.43}$$

where

$$E_{i,1}^n(t) \equiv \sqrt{n} \int_{-U_i^n(0)}^{t-U_i^n(t)} F_i^c(V_i^n(u)) \mathrm{d}\widehat{A}_i^n(u), \quad t \geq 0, \tag{4.44}$$

$$E_{i,2}^n(t) \equiv \sqrt{n} \int_{-U_i^n(0)}^{t-U_i^n(t)} \int_0^1 \mathbf{1}_{\{y > F_i^c(V_i^n(u))\}} \mathrm{d}\widehat{K}_i^n(\bar{A}_i^n(u), y) \quad t \geq 0, \tag{4.45}$$

$$E_{i,3}^n(t) \equiv n \int_{-U_i^n(0)}^{t-U_i^n(t)} F_i^c(V_i^n(u)) \lambda_i(u) \mathrm{d}u \quad t \geq 0, \tag{4.46}$$

for $\bar{A}_i^n, \widehat{A}_i^n$ given by (4.8) and $\widehat{K}_i^n$ is a CLT-scaled empirical process specified by (4.42).

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

$$\varepsilon_i(t) \equiv \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \mathrm{d}u, \tag{4.47}$$

$$\widehat{E}_i^n(t) \equiv n^{-1/2} \left(E_i^n(t) - n\varepsilon_i(t) \right) = n^{-1/2} \left(E_i^n(t) - n \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \mathrm{d}u \right). \tag{4.48}$$

Following the decomposition given in (4.43) - (4.46), we can write

$$\widehat{E}_i^n(t) = \widehat{E}_{i,1}^n(t) + \widehat{E}_{i,2}^n(t) + \widehat{E}_{i,3}^n(t), \tag{4.49}$$

where

$$\widehat{E}_{i,1}^n(t) \equiv n^{-1/2} E_{i,1}^n(t) = \int_{-U_i^n(0)}^{t-U_i^n(t)} F_i^c(V_i^n(u)) \mathrm{d}\widehat{A}_i^n(u) \quad t \geq 0, \tag{4.50}$$

$$\widehat{E}_{i,2}^n(t) \equiv n^{-1/2} E_{i,2}^n(t) = \int_{-U_i^n(0)}^{t-U_i^n(t)} \int_0^1 \mathbf{1}_{\{y > F_i^c(V_i^n(u))\}} \mathrm{d}\widehat{K}_i^n(\bar{A}_i^n(u), y) \quad t \geq 0, \tag{4.51}$$

$$\widehat{E}_{i,3}^n(t) \equiv n^{-1/2} \left(E_{i,3}^n(t) - n \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \mathrm{d}u \right) \quad t \geq 0. \tag{4.52}$$

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

$$\begin{aligned}
\widehat{E}_{i,3}^n(t) &= \sqrt{n} \left(\int_{-U_i^n(0)}^{t-U_i^n(t)} F_i^c(V_i^n(u)) \lambda_i(u) \mathrm{d}u - \int_{-w_i}^{t-w_i} F_i^c(w_i) \lambda_i(u) \mathrm{d}u \right) \\
&= \sqrt{n} \int_0^t F_i^c(U_i^n(u)) \lambda_i(u - U_i^n(u)) \mathrm{d}u - \sqrt{n} \int_0^t F_i^c(w_i) \lambda_i(u - w_i) \mathrm{d}u \\
&\quad - \int_0^t F_i^c(U_i^n(u)) \lambda_i(u - U_i^n(u)) \mathrm{d}\widehat{U}_i^n(u) + O_p(n^{-1/2} \log n) \\
&= - \int_0^t \left\{ f_i(\zeta_i^n(u)) \lambda_i(u - \zeta_i^n(u)) + F_i^c(\zeta_i^n(u)) \lambda_i'(u - \zeta_i^n(u)) \right\} w_i(\widehat{U}^n(u) - \kappa_i(u)) \mathrm{d}u \\
&\quad - \int_0^t w_i F_i^c(U_i^n(u)) \lambda_i(u - U_i^n(u)) \mathrm{d}(\widehat{U}^n(u) - \kappa_i(u)) + O_p(n^{-1/2} \log n),
\end{aligned} \tag{4.53}$$

where the second equality follows by a change of variables (namely, $t \to t - U_i^n(t)$) plus the relation (4.40), while the third equality follows from (4.39) and applying the mean-value theorem with $\zeta_i^n(t)$ satisfying

$$\min\{U_i^n(t), w_i\} \leq \zeta_i^n(t) \leq \max\{U_i^n(t), w_i\}. \tag{4.54}$$

On the other hand, by using conservation of flow, we get

$$E_i^n(t) = B_i^n(t) + D_i^n(t), \tag{4.55}$$

From (4.47) it follows

$$\varepsilon_i(t) = m_i(t) + \int_0^t \mu_i m_i(u)\mathrm{d}u, \tag{4.56}$$

where the equality follows from (4.2). Multiplying both sides of (4.56) by n, subtracting it from (4.55), and dividing both sides by $n^{1/2}$ yields

$$\widehat{E}_i^n(t) = \widehat{B}_i^n(t) + \mu_i \int_0^t \widehat{B}_i^n(u)\mathrm{d}u + \widehat{D}_i^n(t) \quad \text{or} \quad \mathrm{d}\widehat{B}_i^n(t) + \mu_i \widehat{B}_i^n(t)\mathrm{d}t = \mathrm{d}\widehat{E}_i^n(t) - \mathrm{d}\widehat{D}_i^n(t), \tag{4.57}$$

where

$$\widehat{D}_i^n(t) \equiv n^{-1/2}\left[D_i^n(t) - \mu_i \int_0^t B_i^n(u)\mathrm{d}u\right].$$

Let $B^n(t) \equiv \sum_{i\in\mathcal{I}} B_i^n(t)$. It is routine to show, with the overloading assumption (4.2), that

$$B^n(t) = s^n(t) + s_d^n(t) \tag{4.58}$$

holds with arbitrarily high probability by choosing n large enough. Thus, it suffices to focus on the sample paths for which (4.58) holds. In this case we get

$$\sum_{i=1}^K \widehat{B}_i^n(t) = n^{-1/2}\left(B^n(t) - nm(t)\right) = n^{-1/2}\left(s^n(t) + s_d^n(t) - nm(t)\right) = c(t) + \hat{s}_d^n(t), \tag{4.59}$$

where $\hat{s}_d^n(t) \equiv n^{-1/2} s_d^n(t)$. Moreover, by using (2.6) and (4.56) we deduce

$$\begin{aligned}\hat{s}_d^n(t) &\leq \psi\left(-\sum_{i\in\mathcal{I}} \widehat{D}_i^n(t) - \sum_{i\in\mathcal{I}} \mu_i \int_0^t \widehat{B}_i^n(u)\mathrm{d}u - n^{1/2}\sum_{i\in\mathcal{I}} \varepsilon_i(t)\right) \\ &\leq 2\left|\sum_{i\in\mathcal{I}} \widehat{D}_i^n(t) + \sum_{i\in\mathcal{I}} \mu_i \int_0^t \widehat{B}_i^n(u)\mathrm{d}u\right| \leq 2\left|\sum_{i\in\mathcal{I}} \widehat{D}_i^n(t)\right| + 2\sum_{i\in\mathcal{I}} \mu_i \int_0^t \left|\widehat{B}_i^n(u)\right| \mathrm{d}u,\end{aligned} \tag{4.60}$$

where the second equality follows from the continuity and Lipschitz properties of the reflection mapping. Upon substituting (4.49) - (4.51) and (4.53) into (4.57), we obtain

$$\begin{aligned}
&\mathrm{d}\widehat{B}_i^n(t) + w_i F_i^c(U_i^n(t))\lambda_i(t-U_i^n(t))\mathrm{d}\widehat{U}^n(t)\\
=&-\mu_i\widehat{B}_i^n(t)\mathrm{d}t - \left[f_i(\zeta_i^n(t))\lambda_i(t-\zeta_i^n(t)) + F_i^c(\zeta_i^n(t))\lambda_i'(t-\zeta_i^n(t))\right] w_i\widehat{U}^n(t)\mathrm{d}t\\
&+\left[f_i(\zeta_i^n(t))\lambda_i(t-\zeta_i^n(t)) + F_i^c(\zeta_i^n(t))\lambda_i'(t-\zeta_i^n(t))\right] w_i\kappa_i(t)\mathrm{d}t\\
&+ w_i F_i^c(U_i^n(t))\lambda_i(t-U_i^n(t))\mathrm{d}\kappa_i(t) + \mathrm{d}\widehat{E}_{i,1}^n(u) + \mathrm{d}\widehat{E}_{i,2}^n(u)\\
&-\mathrm{d}\widehat{D}_i^n(u) + O_p(n^{-1/2}\log n) \quad \text{for} \quad i=1,\dots,K.
\end{aligned} \tag{4.61}$$

Together with (4.59), we end up getting $K+1$ linear differential equations with respect to the $(K+1)$-dimensional process $(\widehat{B}_1^n,\dots,\widehat{B}_K^n,\widehat{H}^n)$. 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 $\widehat{E}_{i,1}^n$, $\widehat{E}_{i,2}^n$, $\widehat{D}_i^n$, the third bound in (4.60) plus the assumed properties of λ_i, f_i, F_i^c to conclude the stochastic boundedness of the sequence $\{(\widehat{B}_1^n,\dots,\widehat{B}_K^n,\widehat{H}^n); n\in\mathbb{N}\}$. In particular, the sequences $\{\widehat{U}^n; n\in\mathbb{N}\}$ and $\{(\widehat{B}_1^n,\dots,\widehat{B}_K^n); n\in\mathbb{N}\}$ are stochastically bounded. Having established the stochastic boundedness of $\{(\widehat{B}_1^n,\dots,\widehat{B}_K^n); n\in\mathbb{N}\}$, we can replicate the proof of (3.40) to get

$$\hat{s}_d^n \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{4.62}$$

On the other hand, from the established stochastic boundedness of $\{\widehat{U}^n; n\in\mathbb{N}\}$ together with the relations (4.39) and (4.40), we deduce that $\{\widehat{U}_n; n\in\mathbb{N}\}$ and $\{\widehat{V}_n; n\in\mathbb{N}\}$ are stochastically bounded, for $i=1,\dots,K$. This implies the FWLLN for the HWT and PWT processes, that is, as $n\to\infty$,

$$(U^n, U_1^n,\dots,U_K^n, V_1^n,\dots,V_K^n) \Rightarrow (\mathfrak{e}, w_1\mathfrak{e},\dots,w_K\mathfrak{e}, w_1\mathfrak{e},\dots,w_K\mathfrak{e}) \quad \text{in} \quad \mathcal{D}^{2K+1}, \tag{4.63}$$

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 (4.50) and (4.63) to get

$$\widehat{E}_{i,1}^n(t) \Rightarrow \widehat{E}_{i,1}(t) \equiv F_i^c(w_i)\int_{-w_i}^{t-w_i}\sqrt{\lambda_i(u)}\mathrm{d}\mathcal{W}_{\lambda_i}(u), \tag{4.64}$$

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

$$\widehat{E}^n_{i,2}(t) \Rightarrow \widehat{E}_{i,2}(t) \equiv \sqrt{F^c_i(w_i)F_i(w_i)} \int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)} \mathrm{d}\mathcal{W}_{\theta_i}(u), \tag{4.65}$$

where $\mathcal{W}_{\theta_i}$ is a standard Brownian independent of $\mathcal{W}_{\lambda_i}$. The essential structure of the proof for (4.65) 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, from the established stochastic boundedness of $\{(\widehat{B}^n_1, \ldots, \widehat{B}^n_K); n \in \mathbb{N}\}$, it follows the FWLLN for the busy-server processes

$$\left(\bar{B}^n_1, \ldots, \bar{B}^n_K\right) \Rightarrow (m_1, \ldots, m_K) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n \to \infty.$$

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

$$\widehat{D}^n_i(\cdot) = n^{-1/2}\left[\Pi^d_i\left(n\mu_i \int_0^{\cdot} \bar{B}^n_i(u)\mathrm{d}u\right) - n\mu_i \int_0^{\cdot} \bar{B}^n_i(u)\mathrm{d}u\right] \Rightarrow \mathcal{W}_{\mu_i}\left(\mu_i \int_0^{\cdot} m_i(u)\mathrm{d}u\right), \tag{4.66}$$

$n \to \infty$, where we have defined Π^d_i to be a unit-rate Poisson process and $\mathcal{W}_{\mu_i}$ to be a standard Brownian motion independent of $\mathcal{W}_{\lambda_i}$ and $\mathcal{W}_{\theta_i}$. To establish the convergence of (4.11), we will need to strengthen (4.64), (4.65) and (4.66) to joint convergence. The joint convergence of multiple random elements is equivalent to individual convergence if they are independent. Here $\widehat{E}^n_{i,1}$, $\widehat{E}^n_{i,2}$ and $\widehat{D}^n_i$ are not independent because both $\widehat{E}^n_{i,1}$ and $\widehat{E}^n_{i,2}$ involve the arrival-time sequence, and $\widehat{D}^n_i$ depends on B^n_i which in turn correlates with E^n_i through (4.55). But they are conditionally independent given A^n_i, U^n_i, V^n_i and B^n_i. Hence, we can establish the joint convergence by first conditioning and then unconditioning. See Theorem 7.6 of [Pang *et al.*, 2007] for a reference.

To derive a set of SDEs satisfied by the CLT-scaled processes $(\widehat{B}^n_1, \ldots, \widehat{B}^n_K, \widehat{U}^n)$, we seek to simplify the right-hand side of (4.53). First we note that the inequality (4.54) and the convergence in (4.65) imply

$$\zeta^n_i(t) = w_i + O(n^{-1/2}) = U^n_i(t) + O(n^{-1/2}\log n). \tag{4.67}$$

We then use integration by parts to deduce

$$\begin{aligned}
&-\int_0^t w_i F_i^c(\zeta_i^n(u))\lambda_i'(u-\zeta_i^n(u))(\widehat{U}^n(u)-\kappa_i(u))\mathrm{d}u \\
&\quad -\int_0^t w_i F_i^c(U_i^n(u))\lambda_i(u-U_i^n(u))\mathrm{d}(\widehat{U}^n(u)-\kappa_i(u)) \\
=&-w_i F_i^c(\zeta_i^n(t))\lambda_i(t-\zeta_i^n(t))(\widehat{U}^n(t)-\kappa_i(t)) \\
&\quad +\int_0^t w_i\left\{F_i^c(\zeta_i^n(u))\lambda_i(u-\zeta_i^n(u))-F_i^c(U_i^n(u))\lambda_i(u-U_i^n(u))\right\}\mathrm{d}(\widehat{U}^n(u)-\kappa_i(u)) \\
&\quad +\int_0^t w_i\lambda_i(u-\zeta_i^n(u))(\widehat{U}^n(u)-\kappa_i(u))\mathrm{d}F_i^c(\zeta_i^n(u)) \\
=&-w_i F_i^c(w_i)\lambda_i(t-w_i)(\widehat{U}^n(t)-\kappa_i(t))+O(n^{-1/2}\log n),
\end{aligned} \tag{4.68}$$

where the last equality holds due to (4.67). Upon plugging (4.68) into (4.53), we obtain

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

Now plugging (4.49) and the equation above into (4.57), we get, for $i=1,\ldots,K$,

$$\begin{aligned}
&\widehat{B}_i^n(t)+w_i F_i^c(w_i)\lambda_i(t-w_i)\widehat{U}^n(t) \\
=&-\mu_i\int_0^t \widehat{B}_i^n(u)\mathrm{d}u-\int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\widehat{U}^n(u)\mathrm{d}u+\int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\kappa_i(u)\mathrm{d}u \\
&+w_i F_i^c(w_i)\lambda_i(t-w_i)\kappa_i(t)+\widehat{E}_{i,1}^n(t)+\widehat{E}_{i,2}^n(t)-\widehat{D}_i^n(t)+O(n^{-1/2}\log n).
\end{aligned} \tag{4.69}$$

The joint convergence $(\widehat{B}_i^n,\ldots,\widehat{B}_K^n,\widehat{U}^n)\Rightarrow(\widehat{B}_i,\ldots,\widehat{B}_K,\widehat{U})$ then follows by applying the continuous mapping theorem (see Theorem 4.1 of [Pang *et al.*, 2007]) to (4.58) and (4.69), with the *joint* convergence of $\hat{s}_d^n$, $\widehat{E}_{i,1}^n,\widehat{E}_{i,2}^n$, and $\widehat{D}_i^n$ as specified by (4.62), (4.64), (4.65), and (4.66), respectively. Alternatively, one can subtract (4.69) by (4.12) 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 $\{\widehat{U}_i^n;n\in\mathbb{N}\}$ and $\{\widehat{V}_i^n;n\in\mathbb{N}\}$ follow easily from and (4.39) and (4.40), respectively.

Step 4: The FCLT for the queue-length processes. To show that $\{\widehat{Q}_i^n;n\in\mathbb{N}\}$ converges to the corresponding limit, we decompose the right-hand side of (2.5) into three

terms, namely,

$$Q_i^n(t) = Q_{i,1}^n(t) + Q_{i,2}^n(t) + Q_{i,3}^n(t), \tag{4.70}$$

where

$$\begin{aligned} Q_{i,1}^n(t) &\equiv \sqrt{n}\int_{t-U_i^n(t)}^t F_i^c(t-u)\mathrm{d}\widehat{A}_i^n(u), \quad t \geq 0, && (4.71)\\ Q_{i,2}^n(t) &\equiv \sqrt{n}\int_{t-U_i^n(t)}^t \int_0^1 \mathbf{1}_{\{x>F_i^c(t-u)\}}\mathrm{d}\widehat{K}_i^n(\bar{A}_i^n(u),x) \quad t \geq 0, && (4.72)\\ Q_{i,3}^n(t) &\equiv n\int_{t-U_i^n(t)}^t F_i^c(t-u)\lambda_i(u)\mathrm{d}u \quad t \geq 0, && (4.73) \end{aligned}$$

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

$$\begin{aligned} &\widehat{Q}_i^n(t) \equiv n^{-1/2}\left(Q_i^n(t) - nq_i(t)\right) = \widehat{Q}_{i,1}^n(t) + \widehat{Q}_{i,2}^n(t) + \widehat{Q}_{i,3}^n(t),\\ &\text{where} \quad \widehat{Q}_{i,1}^n(t) \equiv \int_{t-U_i^n(t)}^t F_i^c(t-u)\mathrm{d}\widehat{A}_i^n(u) \;\Rightarrow\; \int_{t-w_i}^t F_i^c(t-u)\mathrm{d}\widehat{A}_i(u), && (4.74)\\ &\qquad \widehat{Q}_{i,2}^n(t) \equiv \int_{t-U_i^n(t)}^t \int_0^1 \mathbf{1}_{\{x>F_i^c(t-u)\}}\mathrm{d}\widehat{K}_i^n(\bar{A}_i^n(u),x)\\ &\qquad\qquad \Rightarrow \int_{t-w_i}^t \sqrt{F_i^c(t-u)F_i(t-u)\lambda_i(u)}\mathrm{d}\mathcal{W}_{\theta_i}(u), && (4.75)\\ &\qquad \widehat{Q}_{i,3}^n(t) \equiv \sqrt{n}\int_{t-U_i^n(t)}^{t-w_i} F_i^c(t-u)\lambda_i(u)\mathrm{d}u \;\Rightarrow\; F_i^c(w_i)\lambda_i(t-w_i)\widehat{U}_i(t). && (4.76) \end{aligned}$$

Here the proof for (4.74) and (4.75) is very similar to that of (4.64) and (4.65), and the proof for (4.76) is also straightforward. □

Chapter 5

Conclusions

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. We invested this problem for both critically-loaded and overloaded systems with class-independent service rate and impatient customers. For critically-loaded systems, we proposed a SRS rule and two scheduling policies that can asymptotically achieve TPoD-based performance stabilization for all classes over a finite time horizon. For overloaded systems, we proposed a novel joint-staffing-and scheduling solution that can asymptotically stabilize the TPoD across all customer classes. We established heavy-traffic limit theorems to substantiate the effectiveness of the proposed solution. In addition, we conducted extensive simulation experiments to provide engineering confirmation and practical insight. 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. (i) For the overloaded setting, our proposed scheduling policy will fail when $w_i = 0$ for some $i \in I$. Developing scheduling rules that can handle zero delay targets should be an interesting future research direction. (ii) Instead of exploiting the HoL delay information to differentiate service in an overloaded V system, one may consider queue-ratio based scheduling policies. From the implementation perspective, this would be advantageous to the HoL delay based policy as considered here because a queue-ratio based scheduling rule does not require to track each customer's

elapsed waiting time in queue. (iii) Another natural extension would be to consider a more general network with heterogeneous pools of servers under the setting of skill-based routing; this would make the model more practical for service systems such as call centers. (iv) This research assumes exponential services. However, in many real-world applications service times are not exponentially distributed. Thus, it would be beneficial to devise effective controls in the presence of non-exponential services and establish heavy-traffic limit theorems with general service-time distributions. (v) Finally, customers in queues may switch classes. This is especially true for healthcare settings where patients' health condition may either deteriorate or improve while waiting for treatment. How to staff and schedule to achieve differentiated service with class switching can be a meaningful direction for future research.

Appendix

Proofs of Chapter 4

In this part of the appendix, we provide the proofs for Proposition 4.3.1 and Proposition 4.3.2.

Proof of Proposition 4.3.1.

The multi-dimensional SDE (4.12) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\mu_i t}\tilde{B}_i(t)\right) = e^{\mu_i t}\left(-w_i F_i^c(w_i)\lambda_i(t-w_i)\widehat{H}(t) - \int_0^t w_i f_i(w_i)\lambda_i(u-w_i)\widehat{H}(u)\mathrm{d}u + y_i(t) + G_i(t)\right), \tag{1}$$

where

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

Integrating (1) from 0 to t yields

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

Summing up over i from 1 to K, we have

$$\begin{aligned}
\int_0^t c(s)\mathrm{d}s = \sum_{i=1}^K \tilde{B}_i(t) &= \int_0^t \sum_{i=1}^K w_i\lambda_i(s-w_i)\left(-F_i^c(w_i)e^{\mu_i(s-t)} - f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\widehat{H}(s)\mathrm{d}s \\
&+ \sum_{i=1}^K \int_0^t e^{\mu_i(s-t)}\left(w_iF_i^c(w_i)\lambda_i(s-w_i)\kappa_i(s) + \int_0^s w_if_i(w_i)\lambda_i(u-w_i)\kappa_i(u)du\right)\mathrm{d}s \\
&+ \sum_{i=1}^K \int_0^t e^{\mu_i(s-t)}\int_0^s \sqrt{F_i^c(w_i)\lambda_i(u-w_i)+\mu_im_i(u)}\mathrm{d}\mathcal{W}_i(u)\mathrm{d}s \\
&= \sum_{i=1}^K \int_0^t w_i\lambda_i(s-w_i)\left(-F_i^c(w_i)e^{\mu_i(s-t)} - f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\widehat{H}(s)\mathrm{d}s \\
&+ \sum_{i=1}^K \int_0^t w_i\lambda_i(s-w_i)\kappa_i(u)\left(F_i^c(w_i)e^{\mu_i(s-t)} + f_i(w_i)\frac{1-e^{\mu_i(s-t)}}{\mu_i}\right)\mathrm{d}u \\
&+ \sum_{i=1}^K \int_0^t \frac{1-e^{\mu_i(u-t)}}{\mu_i}\sqrt{F_i^c(w_i)\lambda_i(u-w_i)+\mu_im_i(u)}\mathrm{d}\mathcal{W}_i(u), \qquad (2)
\end{aligned}$$

where the second equality holds by aggregating three independent Brownian motions $\mathcal{W}_{\mu_i}$, $\mathcal{W}_{\theta_i}$ and $\mathcal{W}_{\lambda_i}$ in (4.13) into one independent standard Brownian motion $\mathcal{W}_i$ for each $1 \leq i \leq K$. Differentiating (2) yields

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

And further aggregating the independent Brownian motions $\mathcal{W}_1, \ldots, \mathcal{W}_K$ into $\mathcal{W}$ yields the SVE in (4.18).

Uniqueness and existence of solution to the SVE (4.18)**.** Consider two functions $x, y \in \mathbb{C}$ (space of continuous functions) satisfying an equation

$$x(t) = \int_0^t L(t,s)x(s)\mathrm{d}s + y(t). \qquad (3)$$

we show that (3) specifies a well-defined function $\phi : \mathbb{C} \to \mathbb{C}$ such that $x = \psi(y)$. To do so, for a given y, we define the operator

$$\psi(x)(t) \equiv \int_0^t L(t,s)x(s)\mathrm{d}s + y(t). \tag{4}$$

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

$$x = \psi(x). \tag{5}$$

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

$$\begin{aligned} |\psi(x_1)(t) - \psi(x_2)(t)| &\le \int_0^t |L(t,s)|\mathrm{d}s \cdot \|x_1 - x_2\|_T \\ &\le \|x_1 - x_2\|_T \left(\frac{\sum_{i=1}^K w_i \lambda_i^{\uparrow} \left(\mu_i F_i^c(w_i) + f_i(w_i)\right)}{\sum_{i=1}^K w_i \lambda_i^{\downarrow} F_i^c(w_i)} \right) t. \end{aligned} \tag{6}$$

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

$$L^{\uparrow} = \frac{\sum_{i=1}^K w_i \lambda_i^{\uparrow} \left(\mu_i F_i^c(w_i) + f_i(w_i)\right)}{\sum_{i=1}^K w_i \lambda_i^{\downarrow} F_i^c(w_i)} < \infty, \tag{7}$$

which is guaranteed by the strict positivity assumptions on w_i, λ_i and F_i^c for all $1 \le i \le K$. In case $L^{\uparrow} T > 1$, we can partition the interval $[0, T]$ to successive smaller intervals with length ΔT satisfying $\Delta T < 1/L^{\uparrow}$. This will recursively guarantee the contraction property over all smaller intervals. Hence, the Banach fixed point theorem implies that the FPE (5) has a unique solution over the entire interval $[0, T]$.

Consequently, the function ϕ specified by (3) is well-defined because $\phi(y)$ has one and only one image for any y. So we conclude that (4.18) has a unique solution $\widehat{H}$. If fact, we can write (4.18) as

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

Treating the mean and variance of $\widehat{H}$. Taking expectation in (4.18) yields

$$m_{\widehat{H}}(t) = \int_0^t L(t,s) m_{\widehat{H}}(s)\mathrm{d}s + K(t), \qquad \text{where} \quad m_{\widehat{H}}(t) = \mathbb{E}[\widehat{H}(t)]. \tag{8}$$

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

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

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

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

where the finite upperbound $L^{\uparrow}$ is given by (7). The rest of the proof is similar.

To treat the variance of $\widehat{H}$, consider the SVE (4.18) at $0 \le s, t \le T$

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

Multiplying the two equations and taking expectation yield that

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

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

$$C = \Theta(C), \tag{9}$$

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

$$\begin{aligned} \Theta(C)(t,s) = &- \int_0^t \int_0^s L(t,u)h(s,v)C(u,v)\mathrm{d}v\mathrm{d}u + \int_0^t L(t,u)C(u,s)\mathrm{d}u \\ &+ \int_0^s L(s,v)C(t,v)\mathrm{d}v + \int_0^{s\wedge t} J(t,u)J(s,u)\mathrm{d}u. \end{aligned} \tag{10}$$

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

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

The contraction property is guaranteed if we pick a small $\Delta T > 0$ such that

$$\left((L^{\uparrow})^2 \Delta T^2 + 2L^{\uparrow}\Delta T\right) < 1.$$

According to the Banach contraction theorem, we have the uniqueness and existence in the small block $[0, \Delta T]^2$. The uniqueness and existence of $C(\cdot, \cdot)$ over the entire region $[0, T] \times [0, T]$ can be proved by recursively dealing with small blocks of the form $[i\Delta T, (i + 1)\Delta T] \times [j\Delta T, (j + 1)\Delta T]$.

Remark .0.1 (Numerical Algorithm for $\sigma^2_{\widehat{H}}(t)$) *The above proof of the existence and uniqueness of the FPE* (9) *automatically suggests the following recursive algorithm to compute the covariance $C(t, s)$ and variance $\sigma^2_{\widehat{H}}(t)$. To begin with, we pick an acceptable error target $\epsilon > 0$.*

Algorithm:

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

(ii) In the k^{th} iteration, let $C^{(k+1)} = \Theta\left(C^{(k)}\right)$ with Θ given in (10)*.*

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

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

Proof of Proposition 4.3.2

First note that the FPE (4.21) specifies a well-defined function $\phi : \mathbb{C} \to \mathbb{C}$ such that

$$M_{\widehat{H}} = \phi(K). \tag{11}$$

See the proof of the uniqueness and existence of the SVE (specifically, see (3)–(7)) for details.

Let $(\boldsymbol{\kappa}^*, c^*) \equiv (\kappa_1^*, \ldots, \kappa_K^*, c^*)$, with κ_i^* and c^* given in (4.28) and (4.27). Let K^* and $M^*_{\widehat{H}}$ be the corresponding version of (4.20) and the mean of $\widehat{H}$. (We know that $K^*(t) =$

$M^*_{\widehat{H}}(t) = 0.$) So we have

$$\kappa_i^*(t) = \kappa_i^*(t) - M^*_{\widehat{H}}(t) = z_{1-\alpha_i}\sigma_{\widehat{H}}(t), \qquad 1 \le i \le K. \tag{12}$$

Now consider another solution to $(\tilde{\boldsymbol{\kappa}}, \tilde{c})$ to (4.26), with $(\tilde{\boldsymbol{\kappa}}, \tilde{c}) \equiv (\kappa_1^* + \Delta\kappa_1, \ldots, \kappa_K^* + \Delta\kappa_K, c^* + \Delta c)$. Let $\tilde{K}$ and $\tilde{M}_{\widehat{H}}$ be the corresponding version of (4.20) and mean of $\widehat{H}$. By (4.26), we have

$$\kappa_i^*(t) + \Delta\kappa_i(t) - \tilde{M}_{\widehat{H}}(t) = z_{1-\alpha_i}\sigma_{\widehat{H}}(t),, \qquad 1 \le i \le K. \tag{13}$$

Comparing (12) with (13), we must have

$$\Delta\kappa_i(t) = \tilde{M}_{\widehat{H}}(t) - M^*_{\widehat{H}}(t) \equiv \Delta\kappa(t) \qquad \text{for all } 1 \le i \le K. \tag{14}$$

Hence, any alternative solution to (4.26) (if any) has the form $(\kappa_1^* + \Delta\kappa, \ldots, \kappa_K^* + \Delta\kappa, c^* + \Delta c)$. Next, $M^*_{\widehat{H}} = \phi(K^*)$ and $\tilde{M}_{\widehat{H}} = \phi(\tilde{K})$ imply that

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

which leads to

$$\begin{aligned}\Delta\kappa(t) = \tilde{M}_{\widehat{H}}(t) - M^*_{\widehat{H}}(t) &= \int_0^t L(t,s)\left(\tilde{M}_{\widehat{H}}(s) - M^*_{\widehat{H}}(s)\right)\mathrm{d}s + \left(\tilde{K}(t) - K^*(t)\right), \\ &= \int_0^t L(t,s)\Delta\kappa(s)\mathrm{d}s + \left(\tilde{K}(t) - K^*(t)\right), \end{aligned} \tag{15}$$

where the last equality holds by the first equality. By (14) and (4.20), we have

$$\tilde{K}(t) - K^*(t) = \frac{\Delta\kappa(t)\sum_{i=1}^K \left(\eta_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))\mathrm{d}s\right) - \Delta c(t)}{\eta(t)}. \tag{16}$$

Finally, combining (15) with (16), we must have, for any $\Delta\kappa$,

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

where the last equality above holds by (4.20). Therefore, we can see that c is indeed unique, but κ_i is only unique up to adding a arbitrary common function Δ, which is consistent with our intuition. □

Bibliography

[Aksin *et al.*, 2007] Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688, 2007.

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

[Armony *et al.*, 2015] Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, and Galit B Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.

[Atar *et al.*, 2004] Rami Atar, Avi Mandelbaum, and Martin Reiman. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 14(3):1084–1134, 2004.

[Atar *et al.*, 2010] Rami Atar, Chanit Giat, and Nahum Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

[Atar *et al.*, 2011] Rami Atar, Yair Y. Shaki, and Adam Shwartz. A blind policy for equalizing cumulative idleness. *Queueing Systems*, 67(4):275–293, 2011.

[Atar, 2005] Rami Atar. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 15(4):2606–2650, 2005.

[Dai and Tezcan, 2011] J.G. Dai and Tolga Tezcan. State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research*, 36(2):271–320, 2011.

[Defraeye and van Nieuwenhuyse, 2013] M. Defraeye and I. van Nieuwenhuyse. Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems*, 54(4):1558–1567, 2013.

[Ding *et al.*, 2018] Yichuan Ding, Eric Park, Mahesh Nagarajan, and Eric Grafstein. Patient prioritization in emergency department triage systems: An empirical study of canadian triage and acuity scale (ctas). *to appear in Manufacturing and Service Operations Management*, 2018.

[Doytchinov *et al.*, 2001] Bogdan Doytchinov, John Lehoczky, and Steven Shreve. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Annals of Applied Probability*, pages 332–378, 2001.

[Eick *et al.*, 1993] Stephen G Eick, William A Massey, and Ward Whitt. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742, 1993.

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

[Feldman *et al.*, 2008] Zohar Feldman, Avishai Mandelbaum, William A. Massey, and Ward Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, 2008.

[Fernandes *et al.*, 2005] C. Fernandes, P. Tanabe, N. Gilboy, L. Johnson, R. McNair, A. Rosenau, P. Sawchuk, D.A. Thompson, D.A. Travers, N. Bonalumi, and R.E. Suter. Five level triage: A report from the acep/ena five level triage task force. *Journal of Emergency Nursing*, 31(1):39–50, 2005.

[Gans *et al.*, 2003a] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[Gans *et al.*, 2003b] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[Green *et al.*, 2007] Linda V Green, Peter J Kolesar, and Ward Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

[Gurvich and Whitt, 2009] Itay Gurvich and Ward Whitt. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research*, 34(2):363–396, 2009.

[Gurvich and Whitt, 2010] Itai Gurvich and Ward Whitt. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research*, 58(2):316–328, 2010.

[Gurvich *et al.*, 2008] Itay Gurvich, Mor Armony, and Avishai Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

[Harrison and Zeevi, 2004] J. Michael Harrison and Assaf Zeevi. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research*, 52(2):243–257, 2004.

[Jennings *et al.*, 1996] Otis B. Jennings, Avishai Mandelbaum, William A. Massey, and Ward Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[Kim and Whitt, 2014] Song-Hee Kim and Ward Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

[Kim *et al.*, 2018] Jeunghyun Kim, Ramandeep S Randhawa, and Amy R. Ward. Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing & Service Operations Management*, 20(2):285–301, 2018.

[Kleinrock, 1964] Leonard Kleinrock. A delay dependent queue discipline. *Naval Research Logistics (NRL)*, 11(3-4):329–341, 1964.

[Kostami and Ward, 2009] Vasiliki Kostami and Amy R Ward. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management*, 11(4):644–656, 2009.

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

[Li *et al.*, 2017] Na Li, David A Stanford, Peter Taylor, and Ilze Ziedins. Non-linear accumulating priority queues with equivalent linear proxies. *Operations Research*, 65(6):1712–1726, 2017.

[Liu and Whitt, 2012] Yunan Liu and Ward Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations research*, 60(6):1551–1564, 2012.

[Liu, 2018] Yunan Liu. Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research*, 66(2):514–534, 2018.

[Maman, 2009] Shimrit Maman. *Uncertainty in the demand for service: The case of call centers and emergency departments*. PhD thesis, 2009.

[Mandelbaum and Stolyar, 2004] Avishai Mandelbaum and Alexander L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52(6):836–855, 2004.

[Mandelbaum and Zeltyn, 2009] Avishai Mandelbaum and Sergey Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

[Mandelbaum *et al.*, 1998] Avishai Mandelbaum, William A Massey, and Martin I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1-2):149–201, 1998.

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

[Puhalskii, 2013] Anatolii A. Puhalskii. On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Mathematical Methods of Operations Research*, 78(1):119–148, 2013.

[Puterman, 1994] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[Schrage and Miller, 1966] Linus E Schrage and Louis W Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14(4):670–684, 1966.

[Sharif *et al.*, 2014] Azaz Bin Sharif, David A Stanford, Peter Taylor, and Ilze Ziedins. A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2):73–79, 2014.

[Shi *et al.*, 2016] Pengyi Shi, Mabel C. Chou, J. G. Dai, Ding Ding, and Joe Sim. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science*, 62(1):1–28, 2016.

[Soh and Gurvich, 2016] Seung Bum Soh and Itai Gurvich. Call center staffing: Service-level constraints and index priorities. *Operations Research*, 65(2):537–555, 2016.

[Stanford *et al.*, 2014] David A Stanford, Peter Taylor, and Ilze Ziedins. Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3):297–330, 2014.

[Stolyar, 2004] Alexander L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability*, 14(1):1–53, 2004.

[Talreja and Whitt, 2008] Rishi Talreja and Ward Whitt. Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing. *Management Science*, 54(8):1513–1527, 2008.

[Talreja and Whitt, 2009] Rishi Talreja and Ward Whitt. Heavy-traffic limits for waiting times in many-server queues with abandonment. *The Annals of Applied Probability*, 19(6):2137–2175, 2009.

[Taylor, 2011] Colin Taylor. Metrics that matter – service level, 2011.

[Van Mieghem, 1995] Jan A. Van Mieghem. Dynamic scheduling with convex delay costs: The generalized $c - \mu$ rule. *The Annals of Applied Probability*, pages 809–833, 1995.

[Whitt and Zhao, 2017] Ward Whitt and Jingtong Zhao. Staffing to stabilizing blocking in loss models with non-Markovian arrivals. *Naval Research Logistics*, 64(3):177–202, 2017.

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

[Whitt, 2013] Ward Whitt. Offered load analysis for staffing. *Manufacturing and Operations Management*, 15(2):166–169, 2013.

[Ye and Yao, 2008] Heng-Qing Ye and David D Yao. Heavy-traffic optimality of a stochastic network under utility-maximizing resource allocation. *Operations Research*, 56(2):453–470, 2008.

[Ye and Yao, 2012] Heng-Qing Ye and David D Yao. A stochastic network under proportional fair resource control-diffusion limit with multiple bottlenecks. *Operations Research*, 60(3):716–738, 2012.

[Ye *et al.*, 2019] Han Ye, James Luedtke, and Haipeng Shen. Call center arrivals: When to jointly forecast multiple streams? *Production and Operations Management*, 28(1):27–42, 2019.

[Zeltyn and Mandelbaum, 2005] Sergey Zeltyn and Avishai Mandelbaum. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems*, 51(3-4):361–402, 2005.