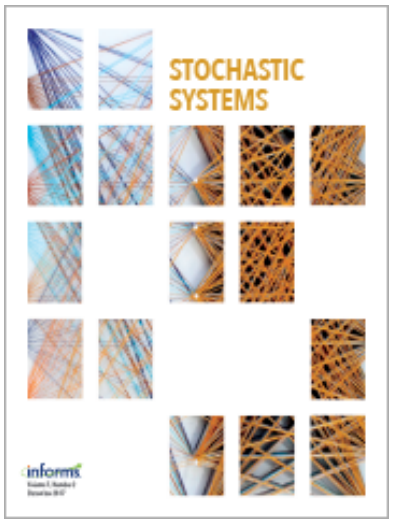


Stochastic Systems

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

Delay-Based Service Differentiation with Many Servers and Time-Varying Arrival Rates

Xu Sun, Ward Whitt

To cite this article:
Xu Sun, Ward Whitt (2018) Delay-Based Service Differentiation with Many Servers and Time-Varying Arrival Rates. Stochastic Systems

Published online in Articles in Advance 24 Sep 2018

. https://doi.org/10.1287/stsy.2018.0015

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

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–34
ISSN 1946-5238 (print), ISSN 1946-5238 (online)

Delay-Based Service Differentiation with Many Servers and Time-Varying Arrival Rates

Xu Sun,[a] **Ward Whitt**[a]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
Contact: xs2235@columbia.edu, http://orcid.org/0000-0003-2560-7370 (XS); ww2040@columbia.edu, http://orcid.org/0000-0003-4298-9964 (WW)

Received: July 8, 2017
Revised: January 22, 2018
Accepted: April 9, 2018
Published Online in Articles in Advance: September 24, 2018

https://doi.org/10.1287/stsy.2018.0015

Abstract. We study the problem of *staffing* (specifying a time-varying number of servers) and *scheduling* (assigning newly idle servers to a waiting customer from one of K classes) in the many-server V model with class-dependent time-varying arrival rates. In order to stabilize performance at class-dependent delay targets, we propose the blind (model-free) head-of-line delay-ratio (HLDR) scheduling rule, which extends an earlier dynamic-priority rule that exploits the head-of-line delay information. We study the HLDR rule in the quality-and-efficiency-driven many-server heavy-traffic (MSHT) regime. We staff to the MSHT fluid limit plus a control function in the diffusion scale. We establish a MSHT limit for the Markov model, which has dramatic state-space collapse, showing that the targeted ratios are attained asymptotically. In the MSHT limit, meeting staffing goals reduces to a one-dimensional control problem for the aggregate queue content, which may be approximated by recently developed staffing algorithms for time-varying single-class models. Simulation experiments confirm that the overall procedure can be effective, even for non-Markov models.

Funding: This research was supported by the National Science Foundation [Grant CMMI 1634133].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsy.2018.0015.

Keywords: service differentiation • many-server heavy-traffic limit • time-varying arrivals • ratio control • scheduling of customers to enter service • sample-path Little's law

1. Introduction and Summary

In this paper, we study delay-based service differentiation via ratio controls in a multiclass many-server service system with time-varying arrival rates. We aim to keep the ratios of the delays of different classes nearly constant over time at specified targets.

1.1. A Time-Varying V Model in the Quality-and-Efficiency-Driven Many-Server Heavy-Traffic Regime

In particular, we study the *time-varying* (TV) V model—that is, the multiclass extension of the $M_t/M/s_t + M$ many-server Markovian queueing model with unlimited waiting space and abandonment from queue. There is a TV number $s(t)$ of homogeneous servers working in parallel. Arrivals from K classes come according to independent nonhomogeneous Poisson processes (NHPPs), with arrivals of class-i occurring at a TV rate $\lambda_i(t)$. If possible, class-i customers enter service immediately upon arrival; otherwise, they join the end of a class-i queue, thereafter to be served in order of arrival. The customer service times and patience times (time to abandon from queue after arrival) are mutually independent exponential random variables, independent of the arrival process. The mean service time and patience time of each class i customer are $1/\mu_i$ and $1/\theta_i$, respectively.

For this model, we study the combined problem of staffing (choosing the function $s(t)$) and scheduling (assigning a newly idle server to the *head-of-line* (HoL) customer in one of the K queues). We do not allow a server to be idle when there is a waiting customer. We propose a variant of the *square-root-staffing* (SRS) rule for staffing and a *head-of-line delay-ratio* (HLDR) scheduling rule and establish supporting results. This approach is attractive because it is transparent and flexible; for example, it can be applied to non-Markov models; see Section 1.3.6.

1.1.1. Staffing. In particular, our SRS staffing function is

$$s(t) = m(t) + \tilde{c}(t)\sqrt{m(t)}, \tag{1}$$

where $m(t)$ is the *offered load*—that is, the expected number of busy servers in the associated infinite-server model (obtained by acting as if $s(t) = \infty$), and $\tilde{c}(t)$ 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 $m_i(t)$, each of which can be represented as the integral

$$m_i(t) \equiv \int_0^t e^{-\mu_i s}\lambda_i(t-s)\,ds, \tag{2}$$

or as the solution of the ordinary differential equation

$$\dot{m}_i(t) = \lambda_i(t) - \mu_i m_i(t). \tag{3}$$

The SRS approach to TV staffing in (1) follows Jennings et al. (1996) and Feldman et al. (2008) for the single-class case, with (2) coming from theorems 1 and 6 of Eick et al. (1993); see Green et al. (2007) and Whitt (2017) for reviews.

1.1.2. Releasing Busy Servers. With TV staffing, we need to specify what happens at the times when staffing is scheduled to decrease but all servers are busy. Even for constant staffing levels, variants of this issue commonly arise in service systems when the servers are people, because human servers work on shifts and may be busy at the end of the shift. In applications, we may assume that the server completes the service in progress after completing the shift, but then the staffing is actually higher than stipulated at those times. When the service times are relatively long, we may want to allow server switching upon departure, which we assume is being used here; see Ingolfsson et al. (2007) and Liu and Whitt (2012a).

For simplicity in the mathematical analysis, we try to avoid this issue as much as possible. Thus, we assume that server switching is being used, so that the server that has completed service most recently is released. Then, to maintain work conservation, we assume that the customer that was being served (the most recent customer to enter service) is pushed back into a queue. Because the service-time distribution is exponential, the remaining service time has the same distribution as a new service time. This push-back scheme is the standard approach for the $M_t/M/s_t + M$ model; for example, see Puhalskii (2013).

However, there is an additional complication for multiclass queues, because we want to maintain work conservation and class identity. Hence, we assume that the most recent arrival is pushed out of service and placed in a special high-priority queue, so that the order of entering service is not altered by this feature; see Section 3.3. As part of our proof of the many-server heavy-traffic (MSHT) functional central limit theorem (FCLT), we show that the impact of this high-priority queue is asymptotically negligible; see Step 2 of the proof of Theorem 1.

1.1.3. The HLDR Scheduling Rule. HLDR exploits the HoL waiting time $U_i(t)$ of class-i at time t. HLDR uses a prespecified TV vector function $v(t) \equiv (v_1(t), \ldots, v_K(t))$. The HLDR scheduling rule assigns the newly available server to the 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—that is, it does not depend on any model parameters.

1.1.4. A New Quality-and-Efficiency-Driven MSHT FCLT. We establish a new *many-server heavy-traffic functional central limit theorem* that support the combined SRS staffing and HLDR scheduling for the TV model. As usual, we consider a sequence of models indexed by the number of servers, n, and let $n \to \infty$. 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

$$s^n(t) = m^n(t) + \tilde{c}(t)\sqrt{m^n(t)} = nm(t) + \sqrt{n}c(t) \quad \text{for} \quad c(t) \equiv \tilde{c}(t)\sqrt{m(t)}\,, \tag{4}$$

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 MSHT fluid limit coincides (with the appropriate scaling by n) with the offered load in for the infinite-server model, as given in Section 1.1.1; for example, see section 9 in Massey and Whitt (1993), Mandelbaum et al. (1998), and section 4 of Liu and Whitt (2012a). The second expression in (4) is appealing for the simple direct way that n appears.

We show that the scaling in (4) puts the model into the *quality-and-efficiency-driven* (QED) MSHT regime; that is, we establish a nondegenerate joint MSHT FCLT for the (appropriately scaled) number of class-i customers in the

system at time t for all i, together with associated delay processes, where the target HoL delay ratios hold almost surely for all t in the limit process; see Theorem 1. Our MSHT FCLT is consistent with previous QED MSHT limits for both stationary models in Halfin and Whitt (1981) and Garnett et al. (2002) and for nonstationary models in Mandelbaum et al. (1998), theorem 2 in Puhalskii (2013), and section 2.6 in Whitt and Zhao (2017). Just like $\tilde{c}(t)$ in (1), the function $c(t)$ in (4) is a control that we use to achieve performance objectives, for example, stabilize performance of the K classes over time at designated targets.

1.2. The Accumulating-Priority Discipline for Healthcare Applications

The stationary version of HLDR, where the vector function $v(t)$ above is independent of t coincides with the *accumulating-priority* (AP) scheduling rule studied by Stanford et al. (2014), Sharif et al. (2014), Li and Stanford (2016), and Li et al. (2017), which in turn coincides with a dynamic-priority rule proposed by Kleinrock (1964). If $v_i(t) = 1$ for all $i \in \mathcal{I}$ and t; that is, 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).

As discussed in Sharif et al. (2014), there is strong motivation for this scheduling policy in healthcare. In particular, Canadian emergency departments (EDs) classify patients into five acuity levels. According to the Canadian Triage and Acuity Scale (CTAS) guideline (Bullard et al. 2014, p. 20), "CTAS level i patients need to be treated within w_i minutes" with $(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120)$. In this context, we establish additional insight for AP by (1) studying staffing as well as scheduling; and (2) establishing MSHT limit and extending to a TV setting.

There is also motivation for the TV extension here from healthcare, because the arrival rates in EDs are strongly time-dependent and the service times are relatively long, as can be seen from Armony et al. (2015) and Whitt and Zhang (2017). One of the great appeals of the AP and HLDR scheduling rules is that they also apply without change in a TV environment, but we contribute by exposing how these rules *perform* in a TV environment. Our framework also allows TV targets.

1.3. Extending Previous Ratio Controls to a Time-Varying Setting

The HLDR scheduling rule is also closely related to ratio rules considered by Gurvich and Whitt (2009a, b; 2010); also see Dai and Tezcan (2008, 2011). These papers considered more general (stationary) models with multiple pools of servers, and the associated routing as well as scheduling, but we only consider a single service pool in this paper. The papers Gurvich and Whitt (2009a, b; 2010) establish MSHT limits for these ratio controls, showing that they induce a simplifying state-space collapse, that permit achieving performance goals asymptotically. Here, we extend those results (for a single service pool) to a TV setting. The technical complexity is significantly less here, because by restricting attention to the single-pool case, we do not need to consider the hydrodynamic limits in Gurvich and Whitt (2009a).

1.3.1. Fixed-Queue-Ratio Scheduling in a TV Setting. The paper by Gurvich and Whitt (2010) showed that analogs of HLDR based on queue lengths instead of HoL delays, called *fixed-queue-ratio* (FQR) *controls*, are effective for achieving delay-based service-differentiation. (Gurvich and Whitt 2009b also considers variants of HLDR in sections 3.3 and 3.4 and its internet supplement.)

Just as with HLDR and AP, FQR extends directly to a TV environment. We started this study by conducting simulation experiments to investigate how FQR and HLDR perform in a TV environment. We present some of the results here in Section 2.

These two scheduling rules often both work well in a TV environment, but not always: If the ratios of the arrival rates of different classes are time-varying, then FQR can seriously fail to stabilize delays. However, we also introduce a modified TVQR that achieves the same performance as HLDR asymptotically; see Theorem 2. In contrast to HLDR, TVQR is not a blind control, because it requires the arrival rates, although those can be estimated, as suggested in definition 3.4 of Gurvich and Whitt (2009b).

1.3.2. State Space Collapse and the Sample-Path TV MSHT Little's Law. The successes and failures of FQR in the TV setting can be explained by a *sample-path* (SP) MSHT *Little's law* (LL) that is a consequence of the TV MSHT limits in Theorems 1 and 2, which generalize the SP-MSHT-LL for the stationary model that is a consequence of theorem 4.3 in Gurvich and Whitt (2009a) and is discussed after equation 13 in section 3 of Gurvich and Whitt (2010). In particular, for large-scale systems that are approximately in the QED MSHT regime,

$$Q_i(t) \approx \lambda_i(t) V_i(t), \quad 0 \le t \le T < \infty \quad \text{for all } i, \tag{5}$$

where $Q_i(t)$ is the queue length, $\lambda_i(t)$ is the arrival rate and $V_i(t)$ is the class-i potential delay at time t (the delay of a hypothetical class-i customer if it were to arrive at time t and had infinite patience). Formally, this can be expressed via the following corollary to our Theorem 1:

Corollary 1 (SP TV MSHT Little's Law). *Under the conditions of Theorem 1, the relations in (5) are valid asymptotically as $n \to \infty$.*

The SP-MSHT-LL in (5) holds because of the dramatic state-space collapse (SSC) associated with the QED MSHT limits under the ratio rules. Because of the SSC, the queue and delay ratios are related by

$$\frac{Q_i(t)}{Q_j(t)} \approx \frac{\lambda_i(t)}{\lambda_j(t)} \frac{V_i(t)}{V_j(t)}, \quad 0 \le t \le T < \infty \quad \text{for all } i. \tag{6}$$

As a consequence, we also obtain the following corollary to our Theorem 1:

Corollary 2 (Queue Ratios and Delay Ratios). *Under the conditions of Theorem 1, asymptotically as $n \to \infty$, a queue ratio can be stable over time together with a delay ratio if and only if the TV ratio of the arrival rates is stable.*

It is significant that the ratio of class-dependent arrival rates may be TV in applications. For instance, section 3.5 of Whitt and Zhang (2017) shows that the proportion of arrivals to the Israeli emergency department that are admitted to an internal ward of the hospital varied strongly over time. Because the admitted patients tend to be among the more critical patients, we infer that there is likely to be a difference in the TV arrival rates of patients classified by acuity.

1.3.3. Scaling the Tail-Probability Delay Targets in the QED MSHT Limit. One-way to achieve delay-based service differentiation is to have class-dependent targets for the delay tail probabilities. In particular, for the sequence of models indexed by n, the goal may be expressed as

$$P(V_i^n(t) \ge w_i^n) \approx \alpha, \quad 0 \le t \le T \quad \text{for all } i, \tag{7}$$

where the class-i targets w_i^n are chosen to produce the desired service-level differentiation. The targets w_i^n could be TV as well, but we leave that out because we are usually interested in stable performance over time in the TV setting.

A key component of the QED MSHT FCLT supporting (7) is the QED scaling of the delay probability targets, which follows assumption 2.1 of Gurvich and Whitt (2010). Because the QED MSHT scaling makes queue lengths be of order $O(\sqrt{n})$, while waiting times are of order $O(1/\sqrt{n})$, waiting times and queue lengths are scaled very differently in the QED MSHT scaling. In order to get a nondegenerate QED MSHT limit for $P(V_i^n(t) \ge w_i^n)$, we assume that

$$\sqrt{n} w_i^n \to w_i \quad \text{as} \quad n \to \infty \quad \text{for} \quad 0 < w_i < \infty, \quad \text{for all } i. \tag{8}$$

As a consequence, we get

$$P(V_i^n(t) \ge w_i^n) = P(\sqrt{n}\, V_i^n(t) \ge \sqrt{n}\, w_i^n) \to P(\hat{V}_i(t) \ge w_i), \quad \text{for all } i, \quad 0 \le t \le T, \tag{9}$$

where $\hat{V}(t) \equiv (\hat{V}_1(t), \ldots, \hat{V}_K(t))$ is the limit process we shall establish with the selected scheduling policy.

The scaling of the delay targets here and in Gurvich and Whitt (2010) makes the tail probabilities similar to the delay probabilities $P(V_i^n(t) \ge 0)$ that are known to be well stabilized in the QED MSHT regime for the single-class model; for example, see Feldman et al. (2008). In contrast, if we do not scale the delay probability targets, then we are forced into the *efficiency-driven* (ED) MSHT regime. Liu (2018) has shown that this ED scaling can also be effective for stabilizing tail probabilities. The approach in Liu (2018) evidently should become relatively more effective as the delay targets increase. Such large targets often occur in healthcare; for example, as in the six-hour boarding time limit in the Singapore hospital discussed in section 1.1.1 of Shi et al. (2016).

1.3.4. Reduction to a One-Dimensional Stochastic Control Problem. Given the one-dimensional MSHT limit associated with HLDR, it suffices to stabilize the tail probability of the limit of the total (aggregate) queue length,

$\hat{Q}(t)$ at a specified target. Paralleling the big display between equations 13 and 14 of Gurvich and Whitt (2010), we have

$$\begin{aligned}\lim_{n\to\infty} P(V_i^n(t) \ge w_i^n) &= P(\hat{V}_i(t) \ge w_i) = P(\hat{Q}_i(t) \ge \lambda_i(t)w_i) \\ &= P\left(\sum_{i=1}^K \hat{Q}_i(t) \ge \sum_{i=1}^K \lambda_i(t)w_i\right) = P\left(\hat{Q}(t) \ge \sum_{i=1}^K \lambda_i(t)w_i\right).\end{aligned} \tag{10}$$

Hence, given $\lambda_i(t)$ and w_i for all i, we can stabilize all processes at the target levels—that is, we can achieve $P(V_i^n(t) \ge w_i^n) \approx \alpha$ for all i, if we can find a control function $c(t)$ that achieves

$$P\left(\hat{Q}(t) \ge \sum_{i=1}^K \lambda_i(t)w_i\right) = \alpha. \tag{11}$$

1.3.5. The Benefits of Additional Structure: Four Cases. Given that we are taking advantage of the SSC provided by 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 paper, we identify four cases. Case 1 is the general model with parameters μ_i and θ_i depending on the class i, for which Theorem 1 shows that the limit in reduction above is $\hat{Q}(t) = [\hat{X}(t) - c(t)]^+$, where $\hat{X}(t)$ is a sum of the components of a K-dimensional diffusion process (and so not itself a diffusion process). We obtain the other three 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 TV K-dimensional Ornstein–Uhlenbeck 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 cases 3 and 4. Case 3 has $\theta_i = \theta$ and $\mu_i = \mu$ for all i; then, the limit process is a 1-dimensional diffusion process. In Case 3, we can establish asymptotic optimality for the proposed solutions to the combined staffing and scheduling problem and 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 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 (2017) and Whitt and Zhao (2017).

Case 4 combines cases 2 and 3, having $\theta_i = \mu_i = \mu$ for all i. Case 4 is the ideal situation where we can provide an explicit solution for the staffing function. The simplification provided by having the abandonment rate equal to the service rate can be explained by the connection to infinite-server models; see section 6 of Feldman et al. (2008). We verify the effectiveness of our HLDR policy with a simulation example in Section 5.3.

1.3.6. Staffing for the Aggregate Queueing Model. In Section 1.3.4, we observed that we can apply the limit process from the MSHT limit to obtain a stochastic control problem for the staffing. An alternative is to use a staffing algorithm for the aggregated queueing model associated with the given model. Within the QED MSHT framework, we can obtain an appropriate model by constructing an associated sequence of single-class models for which the aggregate queue length process has the same QED MSHT limit as obtained for the TV multiclass model.

For example, in Case 3 in Section 1.3.5, the aggregate model is directly an $M_t/M/s_t + M$ model, which has been studied in Feldman et al. (2008) and Liu and Whitt (2012b) and subsequently. Indeed, as long as the service and patience distributions are the same for all classes, the aggregate model is a $G_t/GI/s_t + GI$ model, for which staffing algorithms have been developed in He et al. (2016), Liu and Whitt (2012b), and Whitt and Zhao (2017). We illustrate for the case of a multiclass $M_t/GI/s_t/M$ model with a lognormal service distribution in Section 5.3.

However, there are significant difficulties in the general case 1, because the service times and patience times lose the independence property. Analogous difficulties in conventional heavy-traffic limits for multiclass single-server queues were exposed and studied in Fendick et al. (1989, 1991) and Fendick and Whitt (1989).

1.4. Optimizing and Satisficing by Focusing on Ratio Rules

The standard approach to the staffing-and-scheduling problem for the Markovian queueing model is to formulate a Markov decision process, as in Puterman (1994), starting by specifying relevant costs (for example, for

waiting and for abandonment) and rewards (for completed service, for example, 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. Following great success for queueing models with conventional high-traffic (HT) scaling, for example, as for the $c\mu$ rule (Van Mieghem 1995, Mandelbaum and Stolyar 2004), this approach was applied to scheduling in many-server queues by Atar et al. (2004), Harrison and Zeevi (2004), and Atar (2005) and continues to be a major direction of research, as can be seen from Arapostathis et al. (2015) and Arapostathis and Pang (2016). (The substantial body of related work can be traced from these references.) The MSHT limits are used to produce a limiting diffusion control problem. Unfortunately, the resulting Hamilton–Jacobi–Bellman equations for the limiting diffusion control problems tend to be difficult to solve, so that it is hard to extract useful applied results.

That impasse led Gurvich and Whitt (2009a, b; 2010) to focus on ratio scheduling and routing policies. Instead of optimal policies, they sought "good" policies. In the language of Herbert Simon (Simon 1947, 1979), they suggested satisficing instead of optimizing. In part, this was because the implications of a seemingly natural optimization framework are not so evident; see Milner and Olsen (2008) and section 2 of Gurvich et al. (2008). For example, a tail-probability constraint can permit the scheduler simply to not serve any class-i customer who has waited longer than the performance target.

In contrast, if fixed ratios over time are maintained, then we directly understand the implications of the scheduling rule. To put this in a formal optimization framework, we would say that obtaining fixed or nearly fixed ratios is not a means to another end, but is in fact part of *the goal* (the objective). From that perspective, the SSC associated with the MSHT limit shows that the ratio rules are asymptotically optimal.

Nevertheless, Gurvich and Whitt (2009a, b; 2010) devoted considerable effort to establishing asymptotic optimality of ratio rules for conventional cost models, where it exists, for example, for the generalized $c\mu$ rule in section 3.2 of Gurvich and Whitt (2009b). Where the ratio rules fell short, they focused on the weaker notion of asymptotic feasibility. We will do the same here.

1.5. Organization

In Section 2, we present results of initial simulation experiments to show the value of HLDR and TVQR scheduling rules with TV arrival rates. In Section 3, we define the model and introduce the staffing minimization problem. In Section 4, we state our main analytical results and describe the proposed solutions to the joint staffing and scheduling problems. In Section 5, we present the simulation results implementing the full algorithm for examples from Case 4 and showing that it performs well. We include an example with a lognormal service-time distribution. In Sections 6 and 7, we provide the proofs of the MSHT limits and for asymptotic feasibility and optimality. In Section 8, we conclude with discussing directions for future research. We provide background on the simulation methodology and more numerical results in the online appendix.

2. Initial Simulation Experiments

We illustrate the FQR, HLDR, and TVQR scheduling rules with a two-class $M_t/M/s_t + M$ model having sinusoidal arrival-rate functions and staffing chosen to stabilize the aggregate performance. (TVQR is defined in Section 3.7.)

2.1. The Experimental Setting

Let the two arrival-rate functions be

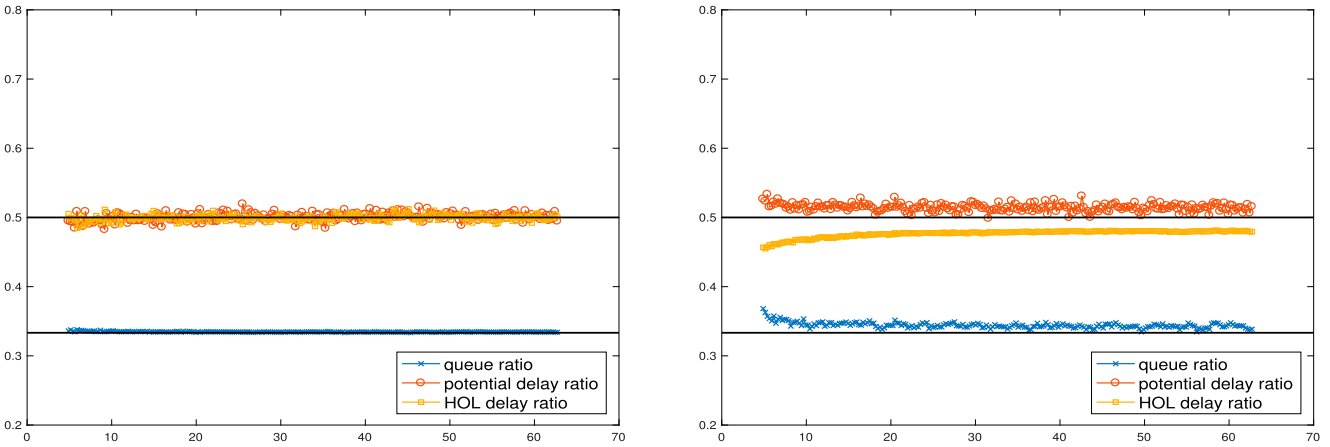
$$\lambda_i(t) = a_i + b_i \sin(d_i t) \quad \text{for } 0 \le t \le T, \quad i = 1, 2. \tag{12}$$

Let the TV staffing functions be as in (4).

2.2. Stationary Arrivals

We start with the stationary case without customer abandonment from queue, letting $(a_1, b_1) = (60, 0)$ and $(a_2, b_2) = (90, 0)$ in (12) (so that the time-scaling factors d_i play no role) with $\mu_1 = \mu_2 = \mu = 1$ and $\theta_1 = \theta_2 = 0$. Suppose that the objective is to achieve a delay ratio $v = 1/2$. From the SP MSHT Little's law in (5), we infer that the queue ratio should be approximately equal to $(1/2)(60/90) = 1/3$. Hence, one would want to use the FQR rule with target queue ratio $r = 1/3$. With this value, we understand that the ratio Q_1/Q_2 is expected to be around the target 1/3, while the delay ratio should be about 1/2. We set the fixed staffing level using the SRS staffing rule with $c \equiv 0.25$, yielding the constant staffing level $s = 170$ to meet the constant offered load of 150. We obtain

Figure 1. Queue and Delay Ratios for a Two-Class Stationary $M/M/s$ Queue with Arrival Rate Functions $\lambda_1 = 60$, $\lambda_2 = 90$, Common Service Rate $\mu = 1$, Without Abandonment ($\theta_1 = \theta_2 = 0$), and $\tilde{c} = 0.25$



FQR HLDR

our simulation estimates by performing 2,000 independent replications; see the online appendix for further explanation.

Figure 1 shows the queue ratio and two delay ratios over the time interval $[5, 70]$ for the FQR rule (left) and the HLDR rule (right). We plot both the potential delay and the HoL delay. Because the HoL delay at time t is the elapsed delay of the customer in queue that is next to enter service, the HoL customer will experience additional delay before entering service, we expect it to be somewhat less than the HoL potential delay. Figure 1 shows that both FQR and HLDR stabilize the queue ratio at the target $r = 1/3$ and the delay ratio at the associated level $v = 1/2$. For FQR, this is as predicted by theorem 4.3 of Gurvich and Whitt (2009a).

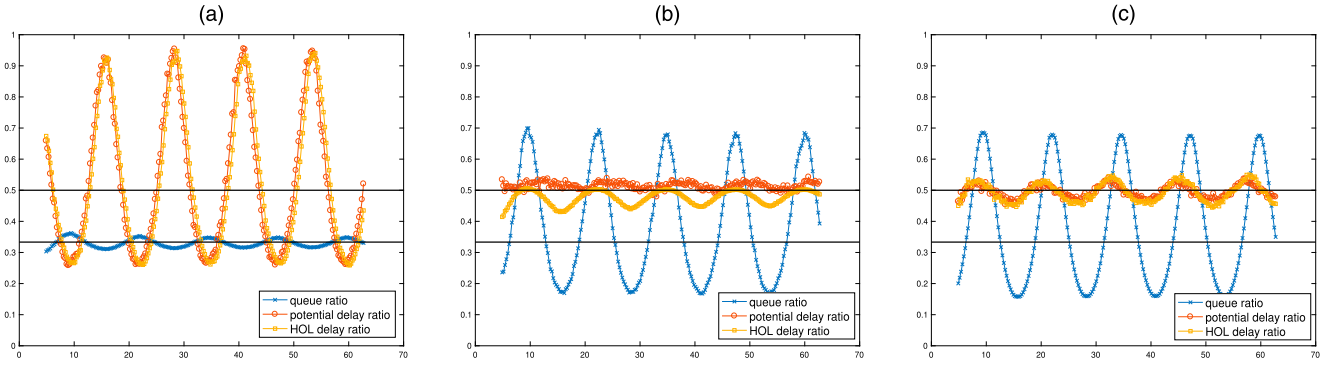
2.3. TV Arrivals Without Abandonment

Now consider TV arrival-rate functions by choosing $(a_1, b_1, d_1) = (60, -20, 1/2)$ and $(a_2, b_2, d_2) = (90, 30, 1/2)$ in (12), so that the overall arrival-rate function is

$$\lambda(t) = \lambda_1(t) + \lambda_2(t) = 150 + 10\sin(t/2).$$

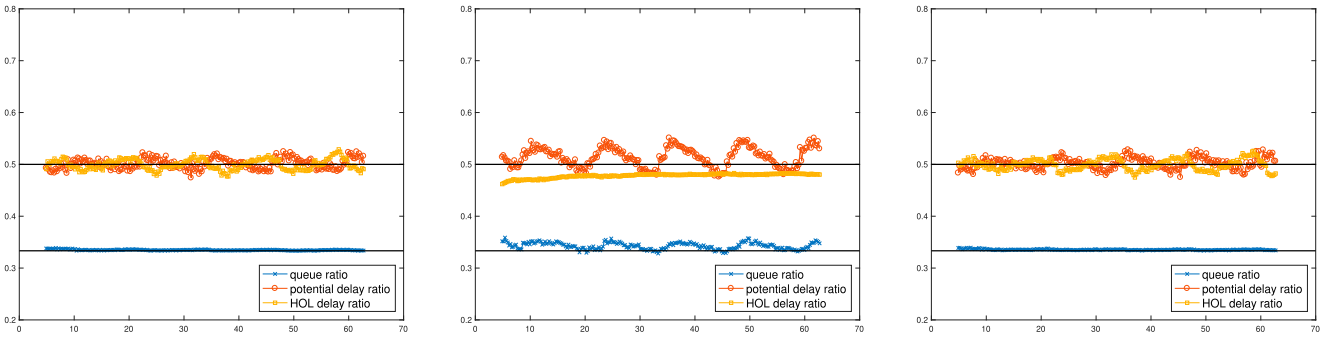
Again, let $\mu_1 = \mu_2 = \mu = 1$ and $\theta_1 = \theta_2 = 0$. With $d_1 = d_2 = 1/2$, the cycle length is $4\pi \approx 12.57$, which is about one half day if we measure time in hours. Figure 2 shows the results. Figure 2, (a) and (b), plots the same set of performance measures for FQR and HLDR shown in Figure 1. Figure 2(a) shows that FQR is again effective at

Figure 2. Queue and Delay Ratios for a Two-Class $M_t/M/s_t$ Queue with Arrival Rate Functions $\lambda_1(t) = 60 - 20\sin(t/2)$, $\lambda_2 = 90 + 30\sin(t/2)$, Common Service Rate $\mu = 1$, Without Abandonment ($\theta_1 = \theta_2 = 0$), and $\tilde{c} = 0.25$



FQR HLDR TVQR

Figure 3. Queue and Delay Ratios for a Two-Class $M_t/M/s_t$ Queue with Arrival-Rate Functions $\lambda_1(t) = 60 + 20\sin(t/2)$, $\lambda_2 = 90 + 30\sin(t/2)$, Common Service Rate $\mu = 1$, Without Abandonment ($\theta_1 = \theta_2 = 0$), and $\tilde{c} = 0.25$



stabilizing the queue lengths, but is now highly ineffective at indirectly stabilizing delays. Similarly, Figure 2(b) shows that HLDR is remarkably effective at directly stabilizing the ratio of the delays, but it does not indirectly stabilize the queue lengths. Figure 2(c) shows that the specially designed TV modification of FQR performs much like HLDR.

What we see in Figure 2 can be explained by (5): The ratio of the arrival rates varies from $(60 - 20)/(90 + 30) = 1/3$ to $(60 + 20)/(90 - 30) = 4/3$, a factor of 4. To see that, we encounter no such difficulty if the aggregate arrival rate is highly TV, while the ratio $AR(t)$ is constant. To illustrate, Figure 3 shows the corresponding results when we simply change the sign of b_1 from $-$ to $+$, which makes $AR(t) = 2/3$ for all t.

2.4. TV Arrivals with Abandonment

We now consider these same scheduling rules in the two-class model when there is customer abandonment. For simplicity, let the abandonment rates be class-invariant with rate $\theta = 0.5$. (The mean time to abandon is twice the mean service time.) From our experiments, we see that abandonment affects our ability to stabilize the ratios, but that it has less and less impact as the scale increases (and has none at all in the MSHT limit). To demonstrate the impact of scale, we plot the queue and delay ratios as a function of system size for the two-class example in Figure 4. Here, we use safety staffing function $c \equiv 0$, which is consistent with the heuristic of "simply staffing to the offered load," as discussed in paragraph 3 of section 6 of Feldman et al. (2008, p. 338).

Figure 4 shows the queue and delay ratios as a function of system size for the same two-class $M_t/M/s_t + M$ queue, but with abandonment rates $\theta_1 = \theta_2 = 0.5$. Figure 4 shows that these scheduling controls become more effective as the scale increases, consistent with out later MSHT limit.

Remark 1 (Class-Dependent Service). *The online appendix shows the corresponding results for the two-class $M_t/M/s_t + M$ queue with class-dependent service times.*

3. Formulation

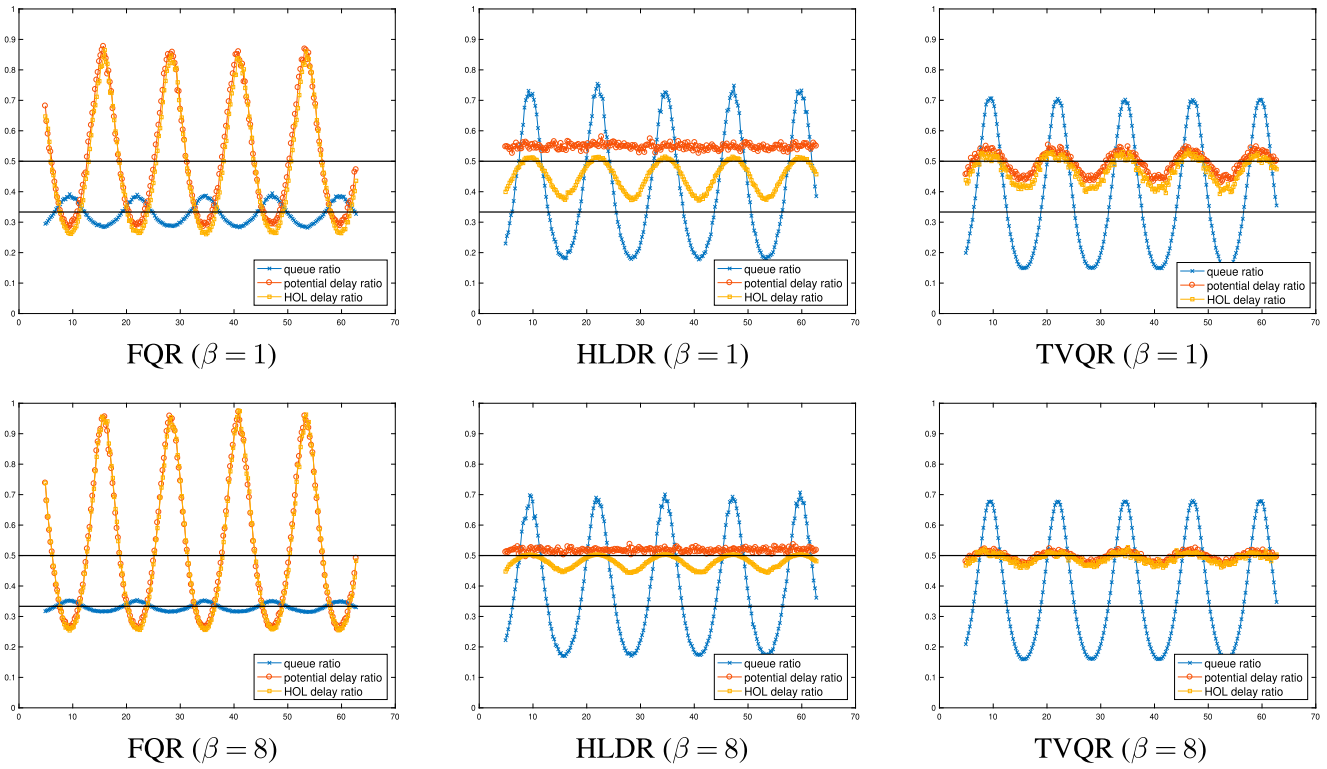
We specify our notation and conventions in Section 3.1 and lay out the preliminaries of the time-varying multiclass queueing model in Section 3.2. We formalize the high-priority queue for customers pushed out of service because of staffing decrease in Section 3.3. We then define the potential delay in Section 3.4 and introduce problem formulations with different SL types in Section 3.5. We define the HLDR and TVQR rules in Sections 3.6 and 3.7, respectively.

3.1. Notation and Conventions

We denote by $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{N}$, respectively, the sets of all real numbers, nonnegative reals and nonnegative integers. For real numbers a and b, $a \wedge b \equiv \min(a, b)$, $a \vee b \equiv \max(a, b)$ and $[a]^+ \equiv a \vee 0$. We use $\lceil a \rceil$ to denote the least integer that is greater than or equal to a. $1(A)$ denotes the indicator function of event (set) A.

The space of right-continuous $\mathbb{R}$-valued functions on $\mathbb{R}_+$ with left-hand limit is denoted by $\mathcal{D} \equiv \mathcal{D}(\mathbb{R}_+, \mathbb{R})$ and is endowed with Skorokhod's J_1-topology and the Borel σ-algebra. For a function $\{x(t); t \in \mathbb{R}_+\}$ in $\mathcal{D}$, let $x(t-)$ represent the left-hand limit at t for $t > 0$ and $\Delta x(t) \equiv x(t) - x(t-)$. All stochastic processes are assumed to be random elements of $\mathcal{D}$. Convergence in distribution (weak convergence) in $\mathcal{D}$ has the standard meaning and is denoted by $\Rightarrow$. The quadratic variation process of a locally square integrable martingale $\{M(t); t \in \mathbb{R}_+\}$ is denoted by $\{\langle M \rangle(t); t \in \mathbb{R}_+\}$.

Figure 4. Queue and Delay Ratios as a Function of System Size for a Two-Class $M_t/M/s_t + M$ Queue with Arrival Rate Functions $\lambda_1(t) = \beta \cdot (60 - 20\sin(t/2))$, $\lambda_2 = \beta \cdot (90 + 30\sin(t/2))$, Service Rate $\mu = 1$, Abandonment Rates $\theta_1 = \theta_2 = 0.5$, and Safety Staffing Function $c \equiv 0$: The Cases $\beta = 1$ and $\beta = 8$



FQR ($\beta = 1$) HLDR ($\beta = 1$) TVQR ($\beta = 1$)

FQR ($\beta = 8$) HLDR ($\beta = 8$) TVQR ($\beta = 8$)

We refer the reader to Jacod and Shiryaev (2013), Pang et al. (2007), and Whitt (2002) for background in weak-convergence and martingale theory. All random entities introduced in this paper are supported by a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

3.2. Preliminaries

There is a set $\mathcal{I} \equiv \{1, \ldots, K\}$ of customer classes. As indicated in Section 1.1.4, for the MSHT FCLT, we consider a sequence of models indexed by the number of servers. In model n, the arrival processes $A_i^n(t)$ are independent NHPPs with rates $n\lambda_i(t)$. For $i \in \mathcal{I}$, let

$$\Lambda_i(t) \equiv \int_0^t \lambda_i(u)\mathrm{d}u, \qquad \hat{A}_i^n(t) \equiv n^{-1/2}(A_i^n(t) - n\Lambda_i(t)). \tag{13}$$

The sequence of processes $\{\hat{A}_i^n\}$ satisfies a FCLT; that is,

$$\hat{A}_i^n(\,\cdot\,) \Rightarrow W_i \circ \Lambda_i(\,\cdot\,) \equiv \hat{A}_i(\,\cdot\,) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{14}$$

where W_i represents a standard Brownian motion for each $i \in \mathcal{I}$. Denote by $A^n \equiv \sum_{i\in\mathcal{I}} A_i^n$ the aggregate arrival process. By the assumed independence, A^n is an NHPP satisfying a FCLT as well with arrival rate function $\lambda(t) \equiv \sum_{i\in\mathcal{I}} \lambda_i(t)$ and associated cumulative rate function $\Lambda(t) \equiv \int_0^t \lambda(u)\mathrm{d}u$. As in Section 1, the service times and patience times are mutually independent, independent of the arrival processes, and exponentially distributed, but these can be class-dependent. Let μ_i and θ_i denote the service rate and abandonment rate of class-i customers, respectively.

Remark 2 (More General Arrival Processes). *We could generalize the arrival processes from M_t to G_t, and the analysis would still go through, provided that we follow the composition construction as by equation* 2.2 *in Whitt* (2015) *and assume a FCLT for the base process; see section* 7.3 *of Pang et al.* (2007).

As in Section 1.1.4, we staff according to (4), which matches the inflow and outflow on the fluid scale; that is, both the queue and the idleness are zero on the fluid scale. As indicated in Section 1.1.2, with time-varying staffing $s^n(t)$,

we need to specify how we manage the system when all servers are busy and the staffing is scheduled to decrease. What we do is to immediately enforce that staffing change, so that we force a customer out of service. In the single-class case, we can let one customer return to the head of the queue, as in Puhalskii (2013). In the multiple-class case, the identity of the class that is moved out of service has an effect on the system state. Our remedy is to create a *high-priority queue* (HPQ) and let any customer that was forced out of service join the back of the HPQ.

To be specific, we assume that the most recent customer to enter service is forced back into the HPQ, so that entering service in order of arrival is maintained. We stipulate that customers in the HPQ have the highest service priority; that is, the next available server always chooses to serve the HoL customer in the HPQ first. In addition, we require that *no customers abandon the HPQ*. Henceforth, we use $Q^n_{0,i}(t)$ to denote the number of class-i customers in the HPQ. We will show that the high-priority queue has no impact on the asymptotic behavior, regardless of the class identities of pushed-back customers; that is, the content of this high-priority queue is asymptotically negligible in the MSHT scaling, and thus does not affect the limit.

We assume a work-conserving policy—that is, no customers wait in queue if there is an available server. Let $Q^n_i(t)$ represent the number of customers in the ith queue, let $\Psi^n_i(t)$ represent the number of customers that have entered service (including any pushed back into the high-priority queue, if any), and let $R^n_i(t)$ represent the number of abandonments of class-i customers, respectively, all up to time t. By flow conservation

$$\begin{aligned} Q^n_i(t) &= Q^n_i(0) + A^n_i(t) - \Psi^n_i(t) - R^n_i(t) \\ &= Q^n_i(0) + \Pi^a_i\left(n\Lambda_i(t)\right) - \Psi^n_i(t) - \Pi^{qb}_i\left(\theta_i \int_0^t Q^n_i(u)\mathrm{d}u\right), \end{aligned} \tag{15}$$

where Π^a_i and Π^{qb}_i are independent unit-rate Poisson processes. Let $B^n_i(t)$ be the number of busy servers serving a class-i customer at time t and $D^n_i(t)$ the cumulative number of class-i customer that have departed *due to service completion* up to time t. Again by flow conservation, we get

$$\begin{aligned} Q^n_{0,i}(t) + B^n_i(t) &= Q^n_{0,i}(0) + B^n_i(0) + \Psi^n_i(t) - D^n_i(t) \\ &= B^n_i(0) + \Psi^n_i(t) - \Pi^d_i\left(\mu_i \int_0^t B^n_i(u)\mathrm{d}u\right), \end{aligned} \tag{16}$$

where Π^d_i are unit-rate Poisson processes independent of Π^a_i and Π^{qb}_i given in (15). Let $X^n_i(t)$ denote the total number of class-i customers in system at time t. Adding up (15) and (16) yields

$$X^n_i(t) = Q^n_i(t) + Q^n_{0,i}(t) + B^n_i(t) = X^n_i(0) + A^n_i(t) - D^n_i(t) - R^n_i(t). \tag{17}$$

Alternatively, one can derive (17) directly from flow conservation.

Finally, let $Q^n_0(t) \equiv \sum_{i\in\mathcal{I}} Q^n_{0,i}(t)$, $Q^n(t) \equiv \sum_{i\in\mathcal{I}} Q^n_i(t)$, and $X^n(t) \equiv \sum_{i\in\mathcal{I}} X^n_i(t)$ be the total number of high- and low-priority customers in queue(s) and the aggregate number of customers in system respectively. Adding up (17) over $i \in \mathcal{I}$ yields

$$X^n(t) = Q^n(t) + Q^n_0(t) + B^n(t) = X^n(0) + A^n(t) - D^n(t) - \sum_{i\in\mathcal{I}} R^n_i(t), \tag{18}$$

where we have defined $B^n(t) \equiv \sum_{i\in\mathcal{I}} B^n_i(t)$ and $D^n(t) \equiv \sum_{i\in\mathcal{I}} D^n_i(t)$.

3.3. The High-Priority Queue

To formally describe the dynamics of the HPQ, we use $\mathcal{S}^n_a(t) \equiv \{u \in [0,t] : \Delta s^n(u) = -1\}$ ($\mathcal{S}^n_d(t) \equiv \{u \in [0,t] : \Delta s^n(u) = 1\}$) to represent the collection of time instances at which the staffing decreases (increases). Then customers enter the HPQ according to the process

$$A^n_0(t) \equiv \sum_{u\in\mathcal{S}^n_a(t)} 1(B^n(u-) = s^n(u-)). \tag{19}$$

Let $D^n_0(t)$ denote the number of departures from the HPQ (number of customers that reenter the service facility from the HPQ) up to time t. Then, it holds that

$$D^n_0(t) \equiv \sum_{u\in\mathcal{S}^n_d(t)} 1(Q^n_0(u-) > 0) + \int_0^t 1(Q^n_0(u-) > 0)\mathrm{d}D^n(u). \tag{20}$$

From (19) and (20), it follows that

$$\begin{aligned} Q_0^n(t) &= A_0^n(t) - D_0^n(t) \\ &= \sum_{u\in\mathcal{G}_a^n(t)} 1(B^n(u-) = s^n(u-)) - \sum_{u\in\mathcal{G}_d^n(t)} 1(Q_0^n(u-) > 0) \\ &\quad - \int_0^t 1(Q_0^n(u-) > 0)\mathrm{d}D^n(u). \end{aligned} \tag{21}$$

We now develop a more tractable upper-bound process for the contents of the HPQ. For that purpose, we consider a net-input process that allows additional arrivals, but has the same departure rules. For that purpose, let the new net-input process be defined by

$$Z^n(t) \equiv s^n(0) - s^n(t) - D^n(t), \quad t \ge 0, \tag{22}$$

and apply the one-dimensional reflection mapping ψ to Z^n to get

$$\Upsilon_0^n(t) \equiv \psi(Z^n)(t) \equiv Z^n(t) - \inf_{0\le u\le t}\{Z^n(u)\}; \tag{23}$$

for example, see section 13.5 in Whitt (2002). The following lemma shows that Υ_0^n serves as an upper bound for Q_0^n.

Lemma 1. *Let Q_0^n and Υ_0^n be as given in (21) and (23), respectively. Then*

$$Q_0^n(t) \le \Upsilon_0^n(t) \quad \textit{for all} \quad t \ge 0 \qquad w.p.1.$$

Proof of Lemma 1. By (23) and (22), it is not hard to see that

$$\Upsilon_0^n(t) = \sum_{u\in\mathcal{G}_a^n(t)} 1 - \sum_{u\in\mathcal{G}_d^n(t)} 1(\Upsilon_0^n(u-) > 0) - \int_0^t 1(\Upsilon_0^n(u-) > 0)\mathrm{d}D^n(u). \tag{24}$$

Combining (21) and (24) gives the desired result. We can apply mathematical induction over successive event times. We see that the upper bound system can have extra arrivals, but must have the same departures whenever the two processes are equal. □

In Section 6, we will show that $\Upsilon_0^n(t)$ is asymptotically negligible in the MSHT scaling, and so $Q_0^n(t)$ has no impact on the MSHT limit.

3.4. Potential Delays

Without customer abandonment, the potential delay in queue i at time t can be represented as the following first-passage time:

$$V_i^n(t) \equiv \inf\{s \ge 0 : \Psi_i^n(t+s) \ge Q_i^n(0) + A_i^n(t)\}.$$

One may attempt to incorporate the abandonment process R_i^n into the expression and write

$$V_i^n(t) \equiv \inf\{s \ge 0 : \Psi_i^n(t+s) + R_i^n(t+s) \ge Q_i^n(0) + A_i^n(t)\}, \tag{25}$$

but the representation (25) is *incorrect*, because the term $R_i^n(t+s)$ may include class-i customers that arrived after time t and then abandoned; see section 1 in Talreja and Whitt (2009).

To formally define the potential delay of class i at some time $t \ge 0$, we exclude the abandonment of customers who arrived after time t; see section 4 of Talreja and Whitt (2009). Following the notation of that paper, we define $R_i^{n,t}(s)$ to be the number of class-i customers who arrived before time t but have abandoned over the time interval $[t, s)$. Then, the potential delay in queue i at time t can be represented as the following first-passage time

$$V_i^n(t) \equiv \inf\{s \ge 0 : \Psi_i^n(t+s) + R_i^{n,t}(t+s) > Q_i^n(0) + A_i^n(t)\}. \tag{26}$$

3.5. The Optimization Formulation

We now introduce several formulations, each aiming to minimize the total cost over a finite interval $[0, T]$, subject to the service-level constraints.

3.5.1. A Mean-Waiting-Time Formulation. We start with mean-waiting-time formulation

$$\begin{aligned} &\text{minimize } \int_0^T s^n(u)\mathrm{d}u\,, \\ &\text{subject to: } \mathbb{E}[V_i^n(u)] \le w_i^n(u) \quad \text{for } u \le T, \quad i \in \mathcal{I}, \end{aligned} \tag{27}$$

where $V_i^n(t)$, as in (26), represents the waiting time of *a virtual customer* of class i that arrives at time t. These SL constraints stipulate that the expected delay in queue i at time t shall not exceed the target $w_i^n(t)$. Here, we allow the SL targets $w_i^n(\cdot)$ be functions in time.

As in Section 1.3.3, we scale w_i^n with n to put our system into the QED MSHT regime.

Assumption 1 (QED Scaling for SL Targets). *The SL target functions $w_i^n(\cdot)$ are scaled so that $w_i^n(\cdot) = w_i(\cdot)/\sqrt{n}$ for some prespecified functions w_i, $i \in \mathcal{I}$.*

We now define the set of admissible policies. To this end, we say that a scheduling policy is *nonanticipative* if a decision at any time is based on the history up to that time and not upon future events.

Definition 1 (Admissible Policies). We say that a joint-staffing-and-scheduling policy (s, π) is admissible if (1) the staffing component s follows the SRS rule (4); and (2) the scheduling component π is nonanticipative. We let Π be the set of all admissible policies.

Definition 2 (Asymptotic Feasibility for the Mean-Waiting-Time Formulation). A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically feasible for (27) if $(s^n, \pi^n) \in \Pi$ and

$$\limsup_{n\to\infty} \mathbb{E}[V_i^n(t)/w_i^n(t)] \le 1 \quad \text{for all } t, \quad i \in \mathcal{I}. \tag{28}$$

Definition 3 (Asymptotic Optimality for the Mean-Waiting-Time Formulation). A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically optimal for (27), if it is asymptotically feasible and for any other sequence $\{(s'^n, \pi^n)\}$ that is asymptotically feasible.

$$[s^n(t) - s'^n(t)]^+ = o(n^{1/2}) \quad \text{as} \quad n \to \infty, \quad \text{for all } t. \tag{29}$$

3.5.2. A Tail-Probability Formulation. We next consider an alternative formulation representing the goal of common call centers. This formulation aims to control the tail probability of the waiting time of each class. The optimization problem is

$$\begin{aligned} &\text{minimize } \int_0^T s^n(u)\mathrm{d}u \\ &\text{subject to: } \mathbb{P}\,(V_i^n(u) > w_i^n(u)) \le \alpha \quad \text{for } u \le T, \quad i \in \mathcal{I}. \end{aligned} \tag{30}$$

The set of constraints requires that the probability that a class i customer who arrives at time t waits longer than $w_i^n(t)$ time units is no greater than α.

As mentioned in Section 1.4, this seemingly reasonable formulation can be problematic; for example, because one can simply choose not to serve any class-i customer who has waited longer than the performance target, without violating any of the SL constraints. The difficulty can be circumvented by adding a global SL constraint as was done in section 2.2.1 of Gurvich and Whitt (2010). Such a formulation and its corresponding solution will be considered shortly. At the moment, we will discuss the asymptotical feasibility for problem (30) despite the fact that this formulation is somewhat problematic.

Definition 4 (Asymptotic Feasibility for the Tail-Probability Formulation). A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically feasible for (30) if, $(s^n, \pi^n) \in \Pi$, and for every $\epsilon > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\,(V_i^n(t)/w_i^n(t) \ge 1 + \epsilon) \le \alpha \quad \text{for all } t, \quad i \in \mathcal{I}. \tag{31}$$

3.5.3. A Mixed Formulation. As indicated above, a global SL constraint is sometimes required for the tail-probability formulation to be well-posed, which naturally leads to our third formulation which we call the mixed formulation:

$$\begin{aligned} &\text{minimize} \int_0^T s^n(u)Du \\ &\text{subject to: } \mathbb{E}[Q^n(u)] \le q^n(u) \quad \text{for } u \le T, \\ &\qquad\qquad \mathbb{P}(V_i^n(u) \le w_i^n(t)) \le \alpha \quad \text{for } u \le T, \quad i = 1, \ldots, K-1. \end{aligned} \tag{32}$$

We recall that $Q^n(t)$ represents the total number of waiting customers in system at time t. Again, we let the target function q^n scale with n so as to force the system to operate in the QED regime. In particular, we make the following assumption by which the underlying staffing rule has to be in the form of (4).

Assumption 2 (QED Scaling for SL Targets). *The SL target function $q^n(\cdot)$ is scaled so that $q^n(\cdot) = \sqrt{n}q(\cdot)$ for some prespecified function q.*

Definition 5 (Asymptotic Feasibility for the Mixed Formulation). A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically feasible for (32), if $(s^n, \pi^n) \in \Pi$, and for every $\epsilon > 0$,

$$\begin{aligned} &\limsup_{n\to\infty} \mathbb{E}[Q^n(t)/q^n(t)] \le 1 \quad \text{for all } t, \quad \text{and} \\ &\limsup_{n\to\infty} \mathbb{P}(V_i^n(t)/w_i^n(t) \ge 1 + \epsilon) \le \alpha \quad \text{for all } t, \quad i = 1, \ldots, K-1. \end{aligned} \tag{33}$$

Definition 6 (Asymptotic Optimality for the Mixed Formulation). A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically optimal for (32), if it is asymptotically feasible and for any other sequence $\{(s'^n, \pi^n)\}$ that is asymptotically feasible,

$$[s^n(t) - s'^n(t)]^+ = o(n^{1/2}) \quad \text{as} \quad n \to \infty, \quad \text{for all } t. \tag{34}$$

3.6. The HLDR Control

We now formalize the HLDR scheduling rule that uniquely determines the assignment processes $\Psi_i(\cdot)$. Let $U_i^n(t)$ be the HoL delay of customer i. Then, the HoL customer in queue i arrived at time $H_i^n(t) \equiv t - U_i^n(t)$. Now, introduce a set of weight/control functions $v(\cdot) \equiv (v_1(\cdot), \ldots, v_K(\cdot))$ and define a weighted HoL delay

$$\tilde{U}_i^n(t) \equiv U_i^n(t)/v_i(t) \quad \text{for each } i \in \mathcal{I}. \tag{35}$$

In addition, use $\tilde{U}^n(t)$ to represent the maximum of those weighted HoL delays—that is,

$$\tilde{U}^n(t) \equiv \max_{i\in\mathcal{I}} \left\{\tilde{U}_1^n(t), \ldots, \tilde{U}_K^n(t)\right\} = \max_{i\in\mathcal{I}} \{U_1^n(t)/v_i(t), \ldots, U_K^n(t)/v_K(t)\}. \tag{36}$$

Let $\tau(t)$ denote the customer class that has the maximum weighted HoL delay; that is,

$$\tau(t) \equiv \left\{i \in \mathcal{I} : \tilde{U}_i^n(t) = \tilde{U}^n(t)\right\}. \tag{37}$$

We can then spell out the assignment processes $\Psi_i^n(\cdot)$:

$$\Psi_i^n(t) = \sum_{u\in\mathcal{T}^n(t)} 1(\tau(u) = i), \tag{38}$$

where $\mathcal{T}^n(t)$ is the collection of time instances up to time t at which a scheduling decision is to be made and $\tau(\cdot)$ is given by (37). Here, ties are broken arbitrarily. For instance, if $\tilde{U}_i^n(t) = \tilde{U}_{i'}^n(t) = \tilde{U}^n(t)$ for $i \neq i'$, then the next-available server chooses to serve either queue i or queue i' with equal probabilities.

3.7. The TVQR Control

As indicated earlier, our HLDR control is intimately related to TV version of the QR rule studied in Gurvich and Whitt (2009a). We briefly review the FQR control, which is a special case of the more general QR control introduced by Gurvich and Whitt (2009a), in the context of multiclass queue with a single pool of independent and identically distributed (i.i.d.) servers. Again, let $Q_i^n(t)$ be the queue length of class i and Q^n 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

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

that is, 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

$$i^* \equiv i^*(t) \in \arg\max_{i \in \mathcal{I}} \{Q_i^n(t) - r_i(t) Q^n(t)\}.$$

4. Main Results

In Section 4.1, we state our main result and then discuss important insights that it provides in Section 4.2. We establish corollaries for important special cases in Section 4.3. In Section 4.4, we establish the associated result for the TVQR rule, and in Section 4.5, we discuss the asymptotic equivalence. In Section 4.6, we observe that the results in Gurvich and Whitt (2009a) themselves can be extended to a large class of TV arrival-rate functions. Finally, in Section 4.7, we propose solutions to the joint-staffing-and scheduling problems formulated in Section 3.5.

4.1. The MSHT FCLT for HLDR in the QED Regime

We first introduce the diffusion-scaled processes

$$\hat{X}_i^n(\cdot) \equiv n^{-1/2}\,(X_i^n(\cdot) - nm_i(\cdot)) \quad \text{and} \quad \hat{X}^n(\cdot) \equiv n^{-1/2}\,(X^n(\cdot) - nm(\cdot)), \tag{39}$$

where $X_i^n(t)$ represents the number of class-i customers in system at time t. Let

$$\hat{Q}_i^n(\cdot) \equiv n^{-1/2} Q_i^n(\cdot) \quad \text{and} \quad \hat{Q}_{0,i}^n(\cdot) \equiv n^{-1/2} Q_{0,i}^n(\cdot), \tag{40}$$

be the diffusion-scaled queue-length processes and $\hat{Q}^n \equiv n^{-1/2} Q^n$ and $\hat{Q}_0^n \equiv n^{-1/2} Q_0^n$ be the aggregate quantities. The same scaling was used by Feldman et al. (2008), Puhalskii (2013), and Whitt and Zhao (2017). As usual, we scale the delay processes by multiplying by $\sqrt{n}$ instead of dividing by $\sqrt{n}$ as in (40):

$$\hat{V}_i^n(t) \equiv n^{1/2} V_i^n(t) \quad \text{and} \quad \hat{U}_i^n(t) \equiv n^{1/2} U_i^n(t) \quad \text{for } i \in \mathcal{I}. \tag{41}$$

We impose the following regularity conditions:

Assumption 3. (A1) *For each $i \in \mathcal{I}$, the arrival-rate function $\lambda_i(\cdot)$ is differentiable with bounded first derivative; that is, there exists a constant $M_1 > 0$ such that $|\lambda_i'(t)| < M_1$ for all $i \in \mathcal{I}$ and $t \geq 0$. The functions $\lambda_i(\cdot)$ are bounded away from zero; that is, there exists $\lambda_* > 0$ such that $\lambda_* \equiv \min_{i \in \mathcal{I}} \inf_{t \geq 0} \lambda_i(t) > 0$ for all t.*

(A2) *The safety-staffing function $c(\cdot)$ is continuous.*

(A3) *All control functions $v_i(\cdot)$ are continuous and bounded from above and away from zero; that is, $v_* \equiv \min_{i \in \mathcal{I}} \inf_{t \geq 0} v_i(t) > 0$ and $v^* \equiv \max_{i \in \mathcal{I}} \sup_{t \geq 0} v_i(t) < \infty$.*

Our main results establishes a MSHT FCLT for HLDR in the QED regime. The limit is a set of interacting diffusion processes.

Theorem 1 (QED MSHT FCLT for HLDR). *Suppose that the system is staffed according to (4), operates under the HLDR scheduling rule, and Assumptions A1–A3 hold. If, in addition, there is convergence of the initial distribution at time 0, that is, if*

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

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

$$\begin{aligned} &\left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n, \hat{U}_1^n, \ldots, \hat{U}_K^n\right) \\ &\Rightarrow \left(\hat{X}_1, \ldots, \hat{X}_K, \hat{Q}_1, \ldots, \hat{Q}_K, \hat{V}_1, \ldots, \hat{V}_K, \hat{U}_1, \ldots, \hat{U}_K\right) \quad \text{in} \quad \mathcal{D}^{4K}, \end{aligned} \tag{42}$$

as $n \to \infty$, *where the diffusion limits* $\hat{X}_i(\cdot)$ *satisfy*

$$\begin{aligned}\hat{X}_i(t) = \hat{X}_i(0) - \mu_i \int_0^t \hat{X}_i(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t \gamma(u)^{-1} v_i(u)\lambda_i(u) \\ \times \left[\hat{X}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}\mathrm{d}W_i(u),\end{aligned} \tag{43}$$

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

$$\begin{aligned}\hat{Q}_i(\cdot) &\equiv \gamma(\cdot)^{-1} v_i(\cdot)\lambda_i(\cdot)\left[\hat{X}(\cdot) - c(\cdot)\right]^+, \\ \hat{V}_i(\cdot) &= \hat{U}_i(\cdot) \equiv v_i(\cdot)\cdot\gamma(\cdot)^{-1}\left[\hat{X}(\cdot) - c(\cdot)\right]^+.\end{aligned} \tag{44}$$

4.2. Important Insights

We can draw several important insights from Theorem 1.

4.2.1. The Role of the SRS Safety Functions *c*. Given that the staffing is done by (4), 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 (4) is specified by the SRS safety function c, which appears explicitly in the diffusion limit. 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 limit.

4.2.2. State-Space Collapse. While the stochastic limit process $(\hat{X}_1, \ldots, \hat{X}_K)$ for the K-dimensional scaled number-in-system process $(\hat{X}_1^n, \ldots, \hat{X}_K^n)$ is a K-dimensional diffusion, depending on the K i.i.d. standard Brownian motions W_i, the limits for the other processes are all a functional of the one-dimensional limit process $\hat{X}$, in particular of $\left[\hat{X} - c\right]^+$, so that there is great state-space collapse. In particular, the limit processes $\hat{Q}_i$, $\hat{V}_i$ and $\hat{U}_i$ are deterministic functionals of each other, as shown by (44). Although the potential and HoL delays are not the same, their limits are the same.

4.2.3. The Role of Customer Abandonment. Although customer abandonment does influence the queue-length and waiting-time limit processes of interest through the one-dimensional limit process $\hat{X}$, customer abandonment plays no roles in determining these limiting ratios. It is wiped out in the heavy-traffic diffusion limit. For the n-th model, both arrivals and departures occur at a time scale of n^{-1}. But because the queue-lengths live on the order of $n^{1/2}$ in the QED, abandonments occur at a time scale of $n^{-1/2}$, indicating a much slower rate. This observation is consistent with Whitt (2006) for the basic $M/M/s + M$ Erlang-A model.

4.2.4. The Sample-Path MSHT Little's Law. We obtain the SP MSHT LL directly from the conclusion of Theorem 1. In particular, for each i, we see that, almost surely,

$$\hat{Q}_i(t) = \lambda_i(t)\hat{V}_i(t) \quad \text{for all } t \geq 0. \tag{45}$$

For the n-th system, we have

$$\hat{Q}_i^n(t) = \lambda_i(t)\hat{V}_i^n(t) + o(1) \quad \text{as} \quad n \to \infty \tag{46}$$

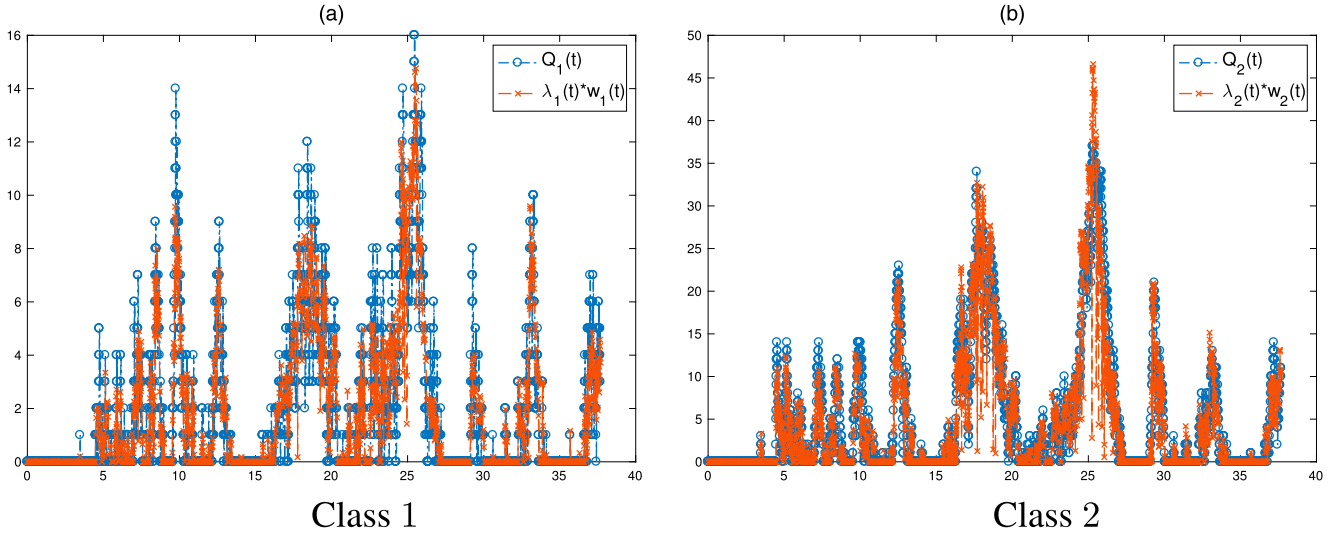
or

$$Q_i^n(t) = \lambda_i^n(t)V_i^n(t) + o(\sqrt{n}) \quad \text{as} \quad n \to \infty. \tag{47}$$

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

Figure 5 depicts the individual sample paths of $Q_i(\cdot)$ and $\lambda_i(\cdot)V_i(\cdot)$ on the same plot for $i = 1, 2$ with the HLDR policy for the base case. Figure 5, (a) and (b), shows that, with the HLDR rule, the sample paths change over time, but the two curves agree closely with error of small order, which strongly supports the SP-MSHT-LL.

Figure 5. Sample Paths of the Queue-Length Process $Q_i(\cdot)$ and the Scaled Delay Process $v_i(\cdot)V_i(\cdot)$ for $i = 1, 2$ with the HLDR Scheduling Policy



4.2.5. Impact of the Arrival Rate and the Weight Functions. Given the limit for the queue-length processes in (44), we see that the proportion of class k queue length of the total queue length is *increasing* in its instantaneous arrival rate $\lambda_k(t)$ but decreasing in the instantaneous rate $1/v_k(t)$.

4.3. Important Special Cases

Theorem 1 applies to the stationary model as an important special case.

Corollary 3 (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,*

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

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

$$\begin{aligned}&\left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n, \hat{U}_1^n, \ldots, \hat{U}_K^n\right)\\ &\Rightarrow \left(\hat{X}_1, \ldots, \hat{X}_K, \hat{Q}_1, \ldots, \hat{Q}_K, \hat{V}_1, \ldots, \hat{V}_K, \hat{U}_1, \ldots, \hat{U}_K\right) \quad \text{in} \quad \mathcal{D}^{4K},\end{aligned}$$

as $n \to \infty$ where the diffusion limits $\hat{X}_i$ satisfy

$$\begin{aligned}\hat{X}_i(t) = \hat{X}_i(0) &- \mu_i \int_0^t \hat{X}_i(u)\mathrm{d}u\\ &- (\theta_i - \mu_i) \int_0^t \gamma^{-1} v_i \lambda_i \left[\hat{X}(u) - c\right]^+ \mathrm{d}u + \sqrt{2\lambda_i} W_i(t),\end{aligned}$$

in which $\gamma = \sum_{i \in \mathcal{J}} v_i \lambda_i$ and $\hat{X} \equiv \sum_{i \in \mathcal{J}} \hat{X}_i$; for each $i \in \mathcal{J}$

$$\hat{Q}_i(\cdot) \equiv v_i \lambda_i \gamma^{-1} \left[\hat{X}(\cdot) - c\right]^+ \quad \text{and} \quad \hat{V}_i(\cdot) = \hat{U}_i(\cdot) \equiv v_i \cdot \gamma^{-1} \left[\hat{X}(\cdot) - c\right]^+. \tag{48}$$

Corollary 3 is in agreement with theorem 4.3 in Gurvich and Whitt (2009a) if one replaces the (state-dependent) ratio function $\tilde{p}_i$ there by a fixed ratio parameter $\gamma^{-1} v_i \lambda_i$. This suggests some form of asymptotic equivalence between the HLDR control and the TVQR control. In fact, we will show in Section 4.5 that an asymptotic equivalence exists not only for time-stationary models but also in time-varying settings. Theorem 4.3 in Gurvich and Whitt (2009a) has $[\hat{X}]^+$ and $[\hat{X}]^-$ in equation 6, whereas (43) in the present paper uses $[\hat{X} - c]^+$ and $[\hat{X} - c]^-$. The discrepancies are due to different centering component being used. In Gurvich and Whitt (2009a), the number of customers in system is centered by the number of servers, whereas we use $nm(t)$ to be the centering term.

Remark 3 (Consistent with Previous AP Results). *The result in* (48) *is in alignment with previous work on AP by Kleinrock* (1964) *and Stanford et al.* (2014), *where the objective is to achieve desired ratios of stationary mean waiting times experienced by customers from the different classes. By focusing on the QED MSHT regime, we are able to obtain a much stronger sample-path result.*

If $\mu_i = \mu$ and $\theta_i = \theta$, $u \in \mathcal{I}$, then the limit of the aggregate content process $\hat{X}$ 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 4 (Class-Independent Services and Abandonments). *Suppose that the conditions in Theorem 1 are satisfied and* $\mu_i = \mu$, *and* $\theta_i = \theta$, $i \in \mathcal{I}$. *Then*

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

where

$$\hat{X}(t) = \hat{X}(0) - \mu \int_0^t \left(\hat{X}(u) \wedge c(u)\right) \mathrm{d}u - (\theta - \mu) \int_0^t \left[\hat{X}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda(u) + \mu m(u)}\mathrm{d}W(u); \tag{49}$$

for each $i \in \mathcal{I}$,

$$\begin{aligned} \hat{Q}_i(\cdot) &\equiv \gamma(\cdot)^{-1} v_i(\cdot)\lambda_i(\cdot)\left[\hat{X}(\cdot) - c(\cdot)\right]^+, \\ \hat{V}_i(\cdot) = \hat{U}_i(\cdot) &\equiv v_i(\cdot)\cdot\gamma(\cdot)^{-1}\left[\hat{X}(\cdot) - c(\cdot)\right]^+. \end{aligned} \tag{50}$$

If we assume further that $\theta = \mu$ in Corollary 4, then the aggregate model is known to behave like an $M_t/M/\infty$ model. Let $\theta = \mu = 1$ in (49). From (49) it holds that

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

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

4.4. The MSHT FCLT for TVQR in the QED Regime

We now turn to the TVQR control as described by Section 3.7. Mimicking the analysis of Gurvich and Whitt (2009a), one can establish the MSHT limits, regarding the TVQR rule, via hydrodynamic limits. However, the proof in Gurvich and Whitt (2009a) is quite involved and in turn relies on additional general state-space collapse 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 Section 6, 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 2 (QED MSHT FCLT for TVQR). *Suppose that the system is staffed according to* (4), *operates under the TVQR scheduling rule and Assumptions A1 and A2 hold. If, in addition,*

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

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

$$\left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n, \hat{U}_1^n, \ldots, \hat{U}_K^n\right) \Rightarrow \left(\hat{X}_1, \ldots, \hat{X}_K, \hat{Q}_1, \ldots, \hat{Q}_K, \hat{V}_1, \ldots, \hat{V}_K, \hat{U}_1, \ldots, \hat{U}_K\right) \tag{51}$$

in $\mathcal{D}^{4K}$ *where the diffusion limits* $\hat{X}_i(\cdot)$ *satisfy*

$$\hat{X}_i(t) = \hat{X}_i(0) - \mu_i \int_0^t \hat{X}_i(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t r_i(u)\left[\hat{X}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}\mathrm{d}W_i(u), \tag{52}$$

where $W_i(\cdot)$ are standard Brownian motions. For each $i \in \mathcal{I}$

$$\hat{Q}_i(\cdot) \equiv r_i(\cdot)\left[\hat{X}(\cdot) - c(\cdot)\right]^+, \quad \textit{and} \quad \hat{V}_i(\cdot) = \hat{U}_i(\cdot) \equiv \frac{r_i(\cdot)}{\lambda_i(\cdot)} \cdot \left[\hat{X}(\cdot) - c(\cdot)\right]^+. \tag{53}$$

We gain several insights from the theorem above: (1) With the TVQR, the desired queue ratio is achieved in the limit despite the fact that arrival rates are changing; (2) from (53), it follows that both the potential and the HoL delays are *inversely* proportional to the arrival rate and proportional to the time-varying queue ratio.

4.5. 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

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

One can easily verify that the stochastic Equation (43) becomes Equation (52).

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

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

Direct calculation allows us to translate Equation (52) into (43).

4.6. Extending the QIR Limits to TV Arrivals

Even though Gurvich and Whitt (2009a) establishes MSHT results for stationary models, we now observe that these results extend immediately to a large class of models with TV arrival rates. In particular, we now observe that the theorems 3.1, 4.1, and 4.3 in Gurvich and Whitt (2009a) directly extend to TV arrival-rate functions that are piecewise-constant, with all changes in the arrival rates occurring on a finite subset of the given bounded interval $[0, T]$. The given proof then applies recursively over the successive subintervals, using the convergence of the terminal values on each interval as the convergence of the initial values required for the next interval. Because any function in $\mathcal{D}([0, t], \mathbb{R})$ on a bounded interval can be approximated by a piecewise-constant function over $[0, T]$, this result is quite general. However, to treat the case of smooth arrival rate functions, as considered here, a further limit-interchange argument is required. Although the remaining argument may be complex, there should be little doubt that the extension holds.

4.7. The Proposed Solution

For each formulation introduced above, we propose a solution that consists of a staffing component 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.

4.7.1. Mean-Waiting-Time Formulation.

We start with the mean-waiting-time formulation as given by (27).

▷ **Staffing:** Choose c^* that satisfies $\mathbb{E}\,[\hat{X}(t) - c^*(t)]^+ = \vartheta(t)$ with

$$\vartheta(t) \equiv \sum_{i \in \mathcal{I}} \lambda_i(t) w_i(t). \tag{54}$$

▷ **Scheduling:** (1) Apply HLDR with ratio functions

$$v^* \equiv (v_1^*(t), \ldots, v_K^*(t)) = (w_1(t), \ldots, w_K(t)), \tag{55}$$

or (2) use TVQR with ratio functions

$$r^* \equiv (r_1^*(t), \ldots, r_K^*(t)) = (\lambda_1(t)w_1(t), \ldots, \lambda_K(t)w_K(t))/\vartheta(t). \tag{56}$$

Informally, our MSHT FCLT in Theorem 1 justifies the following approximation:

$$\mathbb{E}[V_i^n(t)]/w_i^n(t) \approx \mathbb{E}\left[\hat{V}_i(t)\right]/w_i(t) = \mathbb{E}\left[\hat{X}(t) - c^*(t)\right]^+/\vartheta(t) = 1.$$

Theorem 3 (Asymptotic Feasibility and Optimality of the Mean-Waiting-Time Formulation). *Let s^n be determined through the square-root staffing in (4) with c^* as specified above. Set π^n to HLDR with ratio functions v^*. Then, the sequence $\{(s^n, \pi^n)\}$ is asymptotically feasible for (27). If, in addition, we have $\mu_i = \mu$ and $\theta_i = \theta$ for $i \in \mathcal{I}$, then the sequence $\{(s^n, \pi^n)\}$ is also asymptotically optimal.*

4.7.2. Tail-Probability Formulation. For the tail-probability formulation given in (30), we propose the following solution.

▷ **Staffing:** Choose c^* that satisfies $\mathbb{P}(\hat{X}(t) > \vartheta(t) + c^*(t)) = \alpha$, for $t \geq 0$.

▷ **Scheduling:** (1) apply HLDR with ratio functions given in (55), or (2) use TVQR with ratio functions given in (56).

Informally, our MSHT FCLT in Theorem 1 supports the use of the following approximation:

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

Theorem 4 (Asymptotic Feasibility of the Tail-Probability Formulation). *Let s^n be determined through the square-root staffing in (4) with c^* as specified above. Set π^n to HLDR with ratio functions v^*. Then, the sequence $\{(s^n, \pi^n)\}$ is asymptotically feasible for (30).*

4.7.3. Mixed Formulation. For the mixed formulation given in (32), our proposed solution is stated as follows.

▷ **Staffing:** Choose $c^*(\cdot)$ that satisfies $\mathbb{E}[\hat{X}(t) - c^*(t)]^+ = q(t)$, for each $t \geq 0$.

▷ **Scheduling:** For the function c^* as determined above, choose $x(t)$ satisfying $\mathbb{P}(\hat{X}(t) > x(t) + c^*(t)) = \alpha$, for $t \geq 0$.

For each $t \leq T$, set $w_K(t) = [x(t) - \sum_{i=1}^{K-1} \lambda_i(t) w_i(t)]/\lambda_K(t)$. Then apply HLDR with ratio functions given in (55), or (2) use TVQR with ratio functions given in (56).

Theorem 5. (Asymptotic Feasibility and Optimality of the Mixed Formulation). *Let s^n be determined through the square-root staffing in (4) with c^* as specified above. Set π^n to HLDR with ratio functions v^*. Then, the sequence $\{(s^n, \pi^n)\}$ is asymptotically feasible for (32). If, in addition, we have $\mu_i = \mu, \theta_i = \theta$ for $i \in \mathcal{I}$ and $\theta \leq \mu$, then the sequence $\{(s^n, \pi^n)\}$ is also asymptotically optimal.*

5. Simulation Confirmation

Successful application of the proposed solutions to the joint-staffing-and-scheduling problem in Section 4.7 requires effective computation of the *minimum* safety staffing function c^*. In this section, we illustrate how the function c^* can be calculated explicitly in Case 4 in Section 1.3.5, where $\theta_i = \mu_i = \mu$ for all i. Then we present results of simulation experiments to show how HLDR and TVQR perform.

5.1. Calculating the Minimum Safety Staffing Level with $\mu = \theta$

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

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

We apply Corollary 4 and the following remark, which identifies $\hat{X}(t)$ as an OU process. Because $\hat{X}(t)$ is normally distributed with mean 0 and variance $m(t)$, it holds that

$$c^*(t) = \Phi^{-1}(1-\alpha)\sqrt{m(t)} - \vartheta(t). \tag{57}$$

To calculate the minimum safety staffing function c^* for the *mean-waiting-time* formulation, let

$$\vartheta(t) = \mathbb{E}\left[\hat{X}(t) - c^*(t)\right]^+.$$

It is readily verifiable that

$$c^*(t) = \sqrt{m(t)} \cdot \tilde{c}(t), \tag{58}$$

where $\tilde{c}(t)$ is the unique root of the equation

$$\frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} - x\Phi^c(x) = \vartheta(t)/\sqrt{m(t)}. \tag{59}$$

Remark 4 (Avoiding the Scale Parameter n in Applications). *In applications, the original targets w_i^n will be used in calculating the safety staffing. We now explain how to apply* (58). (*The discussion for the tail-probability formulation is similar.*) *By* (58), *the safety staffing is*

$$n^{1/2}c^*(t) = \sqrt{nm(t)}\tilde{c}(t) = \sqrt{m^n}\tilde{c}(t),$$

where the offered load $m^n(t)$ is calculated according to (2) *using the original arrival-rate functions λ_i^n. Thus, the key is to compute $\tilde{c}(t)$ by solving* (59). *The left side of* (59) *is independent of the scaling parameter n while the right side becomes*

$$\frac{\vartheta(t)}{\sqrt{m(t)}} = \frac{\sum_i n\lambda_i(t)\cdot n^{-1/2}w_i(t)}{\sqrt{nm(t)}} = \frac{\sum_i \lambda_i^n(t)w_i^n(t)}{\sqrt{m^n(t)}}.$$

Thus, there will be no use of the scaling parameter n. The scaling is only used for the proof of asymptotic feasibility and optimality of the proposed solutions.

5.2. The Experimental Setting

For our simulation experiments, we start by considering the same two-class Markov V example in Section 2 but choosing $(a_1, b_1, d_1) = (60, -20, 2/5)$ and $(a_2, b_2, d_2) = (90, 30, 2/5)$ in (12). We assume that $\mu_i = \theta_i = 1, i = 1, 2$. In addition, we stipulate that the SL targets for class 1 and 2 are $w_1^n \equiv 1/6$ and $w_2^n \equiv 1/3$, respectively.

We have chosen the parameters to relate to a hospital emergency room (ER) where patients are classified into two categories, namely, high-acuity and low-acuity patients. In the context of an ER where the average treatment time is about 90 minutes, a cycle would be about 5π times longer, which is about 24 hours, and the SL targets are 15 minutes and 30 minutes for high-acuity and low-acuity patients, respectively. Abandonments from the queue can be interpreted as patients who left without being seen or patients who were diverted to other facilities before receiving treatment. Thus, our parameter choice may provide insight for hospital ERs.

Remark 5 (Supporting Healthcare Data). *According to the National Hospital Ambulatory Medical Care Survey, United States,* 2010–2011, "*The median wait time to be treated in the ED was about* 30 *minutes, and the median treatment time was slightly more than* 90 *minutes in* 2010–2011." *The Centers for Disease Control and Prevention reported in May* 2014 *that average emergency department wait times* (*about* 30 *minutes*) *and treatment times* (*about* 90 *minutes*), *which add up to roughly two hours in the ER.*

Customer abandonment is less prominent in hospitals than in modern call centers, but it is a factor. Nevertheless, it would have been more reasonable to assume $\theta < \mu$, but that takes us out of the tractable Case 4 *in Section* 1.3.5. *With $\mu = \theta$, the equation in* (4) *simplifies greatly, yielding an OU process with TV variance. Indeed, corollary* 5.1 *in the e-companion of Feldman et al.* (2008) *has shown that $\hat{X}(t)$ is normally distributed with zero mean and variance $m(t)$.*

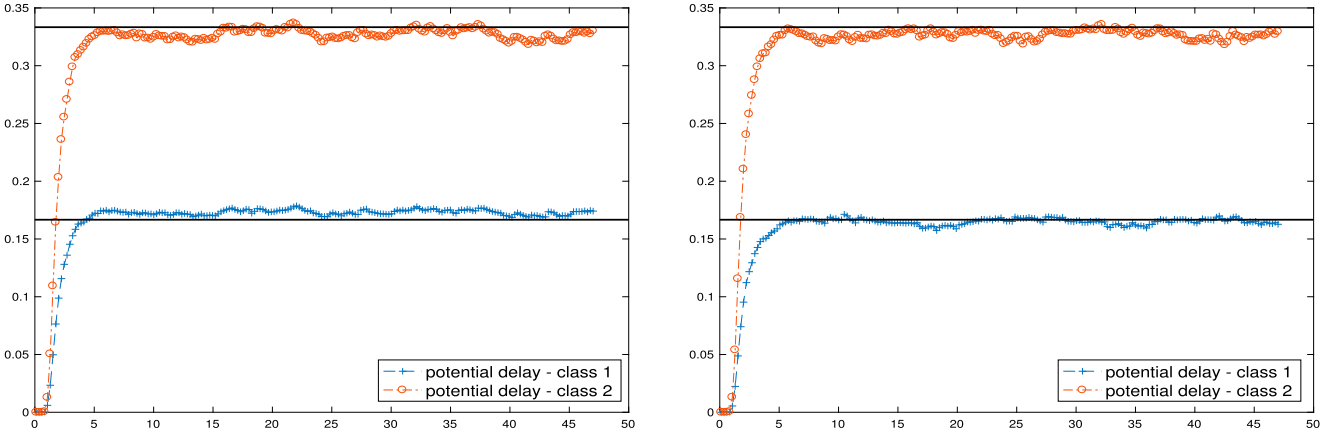
5.3. The Simulation Results

In Section 5.3.1, we report simulation results for the example described in Section 5.2. We consider both the mean-waiting-time formulation and the tail-probability formulation introduced in (27) and (30). For each formulation, we use the explicit expression for the corresponding minimum safety staffing function c^* from Section 5.1. We then apply the solutions in Sections 4.7.1 and 4.7.2 to conduct the simulation studies. We extend our method to lognormal service times in Section 5.3.2. With nonexponential service times, we use the staffing method introduced in section 3 of He et al. (2016), which also applies to non-Poisson arrival processes.

In both cases, we use periodic steady-state formulas for the offered load, so we do not try to staff to meet an unrealistic initial startup period, but we could do so by applying (2) or (3) with $\lambda(t) = 0$ for $t \le 0$; for example, to treat the sinusoidal case, we could apply equation 19 of Liu and Whitt (2012b).

5.3.1. Exponential Service Times. Figure 6 depicts the estimated expected potential delays over the time interval $[0, 50]$ for the HLDR rule (left) and the TVQR rule (right) with c^* derived from (58). We plot these estimated expected potential delays for both classes. All estimates were obtained by averaging over 2,000 independent replications. Figure 6 shows that both HLDR and TVQR stabilize the expected potential delay of each class at the associated SL target.

Figure 6. Estimated Expected Potential Delays 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 (58)

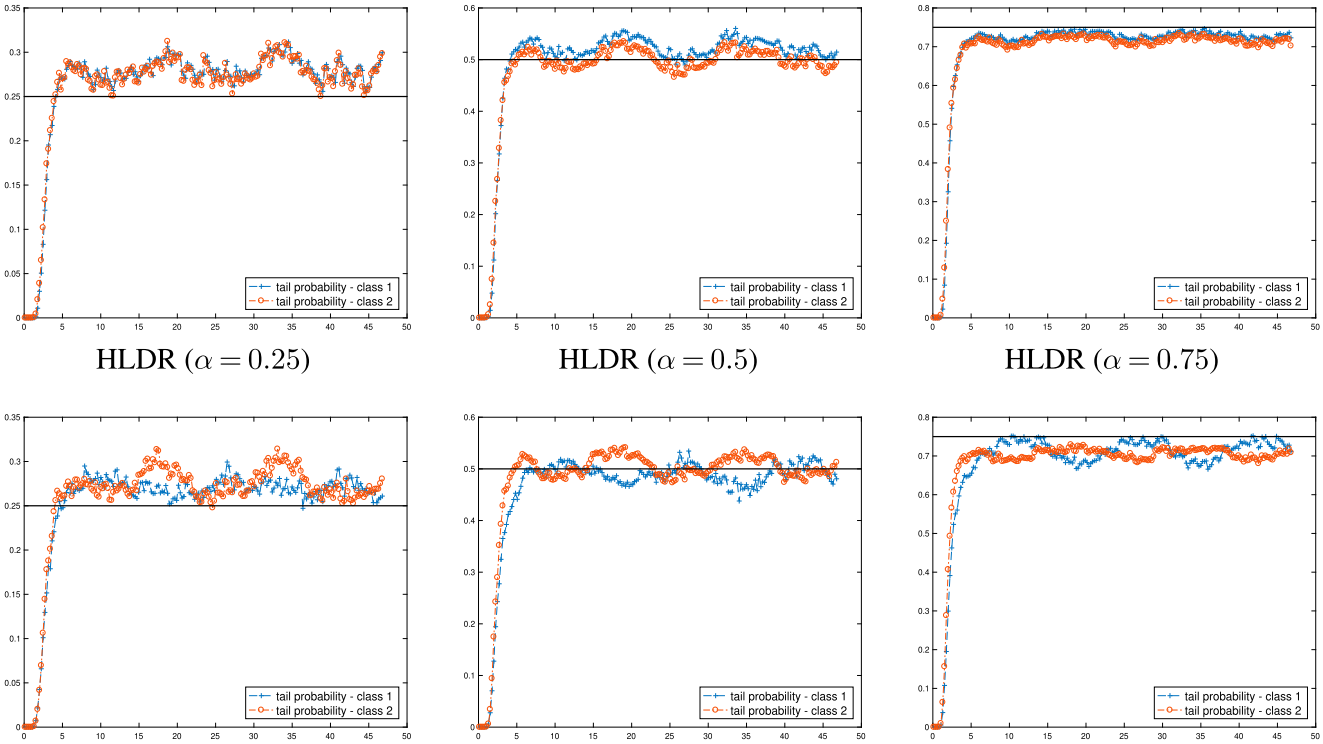


HLDR TVQR

Figure 7 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 (57). 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 2,000 independent replications. Figure 7 shows that, for all three cases, both HLDR and TVQR stabilize the tail probabilities of each class at the desired level.

5.3.2. Lognormal Service Times. For the last experiment, we consider nonexponential service-time distributions. In particular, we examine cases with lognormal service times. Let μ and σ^2 denote the parameters of the normal

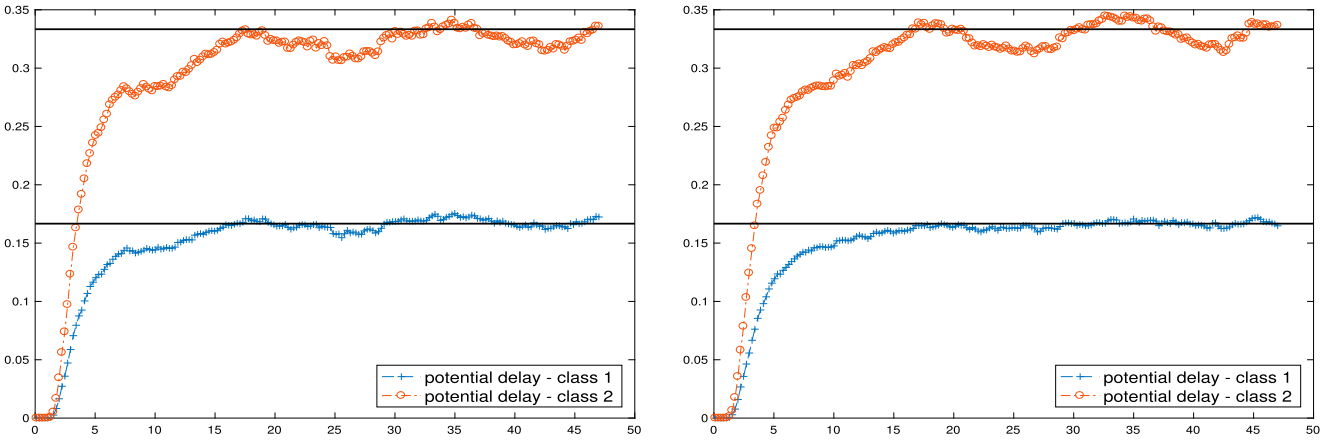
Figure 7. Tail Probabilities for a Two-Class $M_t/M/s_t + M$ Queue with Arrival-Rate Functions $\lambda_1(t) = 60 - 20\sin(2t/5)$, $\lambda_2 = 90 + 30\sin(2t/5)$, Common Service Rate $\mu = 1$, Abandonment Rate $\theta = 1$, and Minimum Staffing Function c^* Derived from (57)



HLDR ($\alpha = 0.25$) HLDR ($\alpha = 0.5$) HLDR ($\alpha = 0.75$)

TVQR ($\alpha = 0.25$) TVQR ($\alpha = 0.5$) TVQR ($\alpha = 0.75$)

Figure 8. Estimated Expected Potential Delays for a Two-Class $M_t/G/s_t + M$ Queue with Arrival-Rate Functions $\lambda_1(t) = 60 - 20\sin(2t/5)$, $\lambda_2 = 90 + 30\sin(2t/5)$, and Abandonment Rate $\theta = 1$



Note. Service times follow a lognormal distribution with mean 1 and variance 4.

distribution, so that, if S has a lognormal distribution, then $\ln(S)$ is distributed normally with mean μ and variance σ^2.

The associated mean and variance of a lognormal random variable are

$$\mathbb{E}[S] = \exp(\mu + \sigma^2/2) \quad \text{and} \quad \text{Var}[S] = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1).$$

Hence,

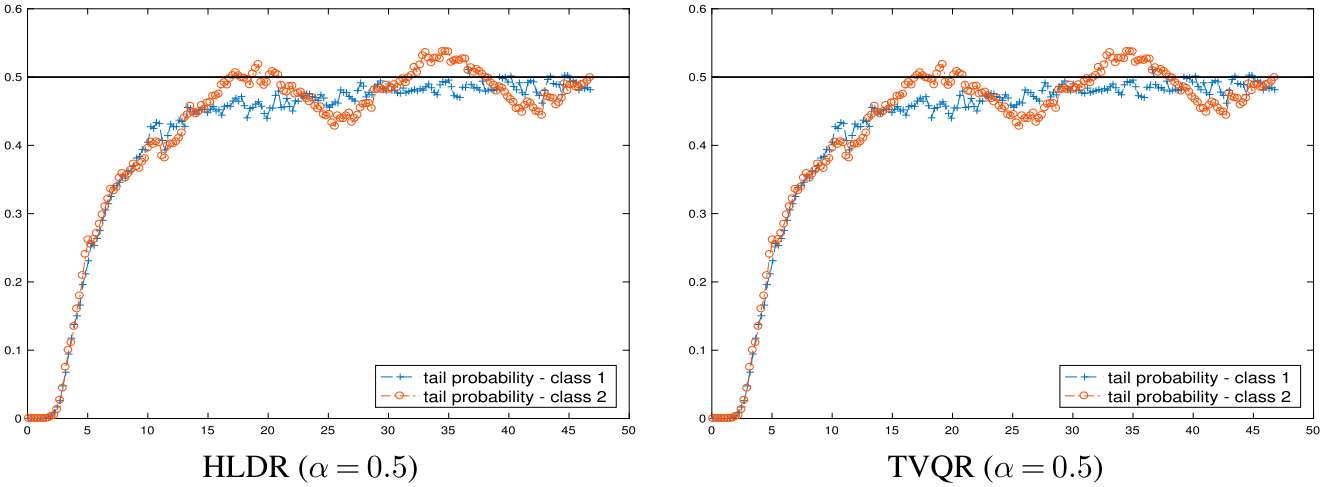
$$scv[S] \equiv \frac{\text{Var}[S]}{(\mathbb{E}[S])^2} = \exp(\sigma^2) - 1;$$

that is, the squared coefficient of variation (scv) is uniquely determined by the parameter σ^2.

We would like to construct a lognormal random variable (r.v.) with scv equal to 2. We therefore choose σ^2 satisfying $\exp(\sigma^2) - 1 = 4$. Direct calculation gives $\sigma^2 = \ln 5$. In addition, we require the r.v. to be mean $-$ 1. Then, the parameter μ has to satisfy $\mu + \sigma^2/2 = 0$ which yields $\mu = -(\ln 5)/2$. More generally, if we require that $scv[S] = c$ and $\mathbb{E}[S] = 1$, then $\sigma^2 = \ln(c + 1)$ and $\mu = -\ln(c + 1)/2$.

Figure 8 depicts the estimated expected potential delays over the time interval $[0, 50]$ for the HLDR rule (left) and the TVQR rule (right). We show the potential delays for both classes. All estimates were obtained by averaging over

Figure 9. Tail Probabilities for a Two-Class $M_t/G/s_t + M$ Queue with Arrival-Rate Functions $\lambda_1(t) = 60 - 20\sin(2t/5)$, $\lambda_2 = 90 + 30\sin(2t/5)$, and Abandonment Rate $\theta = 1$



Note. Service times follow a lognormal distribution with mean 1 and variance 4.

2,000 independent replications. Figure 8 shows that both HLDR and TVQR stabilize performance at the appropriate target, after the initial warmup period.

Figure 9 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). Here, we assume that the target tail probability $\alpha = 0.5$. We plot the tail probabilities for both classes. All estimates were obtained by averaging over 2,000 independent replications. Figure 9 shows that both HLDR and TVQR perform reasonably well.

We see that the warmup period due to starting empty before performance is stabilized is longer with lognormal service times. An explanation and quantitative approximation are given in formula 20 of Eick et al. (1993).

6. Proofs of MSHT FCLTs for HLDR and TVQR

Proof of Theorem 1. For any $x \in \mathcal{D}$, let $x[t_1, t_2) \equiv x(t_2-) - x(t_1-)$. In addition, let $L_i^{n,t}(s)$ 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

$$Q_i^n(t-) = A_i^n[H_i^n(t), t) - L_i^{n,H_i^n(t)}[H_i^n(t), t) = A_i^n[t - U_i^n(t), t) - L_i^{n,t-U_i^n(t)}[t - U_i^n(t), t). \tag{60}$$

Let

$$\hat{R}_i^n(\cdot) \equiv n^{-1/2} R_i^n(\cdot), \quad \hat{R}_i^{n,t}(t+\cdot) \equiv n^{-1/2} R_i^{n,t}(t+\cdot) \quad \text{and} \quad \hat{L}_i^{n,t}(t+\cdot) \equiv n^{-1/2} L_i^{n,t}(t+\cdot). \tag{61}$$

By the definition of R_i^n, $R_i^{n,t}$ and $L_i^{n,t}$, we have

$$\hat{R}_i^n[t, s] = \hat{R}_i^{n,t}(s) + \hat{L}_i^{n,t}(s). \tag{62}$$

Combining (14), (40), (60), and (61) yields

$$\begin{aligned} \hat{Q}_i^n(t-) &= \hat{A}_i^n[t - U_i^n(t), t) + n^{1/2} \int_{t-U_i^n(t)}^{t} \lambda_i(u)\mathrm{d}u - \hat{L}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t) \\ &= \hat{A}_i^n[t - U_i^n(t)), t) + n^{1/2}\lambda_i(t)U_i^n(t) - \hat{L}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t) + e_i^n(t), \end{aligned} \tag{63}$$

where

$$e_i^n(t) \equiv n^{1/2} \int_{t-U_i^n(t)}^{t} \lambda_i(u)\mathrm{d}u - n^{1/2}\lambda_i(t)U_i^n(t). \tag{64}$$

Introduce the auxiliary process

$$\hat{K}_i^n(t) \equiv \hat{A}_i^n[t - U_i^n(t)), t) - \hat{L}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t) + e_i^n(t) \quad \text{for} \quad i \in \mathcal{I}. \tag{65}$$

Then, inserting (65) into (63) yields

$$\hat{Q}_i^n(t-) = \lambda_i(t)\hat{U}_i^n(t) + \hat{K}_i^n(t), \quad i \in \mathcal{I}. \tag{66}$$

We will later show that the auxiliary processes $\hat{K}_i^n(\cdot)$ vanish uniformly over compact intervals as n grows to infinity.

We lay out the path ahead. We start off by showing that both $\{\hat{X}_i^n(\cdot); n \in \mathbb{N}\}$ and $\{\hat{Q}^n(\cdot); n \in \mathbb{N}\}$ are stochastically bounded. We then argue that the sequence of HoL delay processes $\{n^{1/2}U_i^n(\cdot); n \in \mathbb{N}\}$ are stochastically bounded, which shows that $U_i^n(\cdot)$ lives on the order of $O(n^{-1/2})$. We then prove that the queue-length processes are asymptotically proportional to the weights; that is,

$$(\hat{Q}_1^n(t), \ldots, \hat{Q}_K^n(t)) \propto (v_1(t)\lambda_1(t), \ldots, v_K(t)\lambda_K(t)) \quad \text{for all} \quad t \le T.$$

This is essentially a state-space-collapse result in the many-server diffusion limit. Finally, by a similar argument as in Gurvich and Whitt (2009a) (first SSC and then diffusion limits), we obtain the diffusion limits for $\hat{X}_i^n(\cdot)$. The limits for the queue-length processes and delay processes follow immediately.

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

$$\begin{aligned}\hat{D}_i^n(t) &\equiv n^{-1/2}\left[D_i^n(t) - \mu_i\int_0^t B_i^n(u)\mathrm{d}u\right]\\ &= n^{-1/2}\left[\Pi_i^d\left(\mu_i\int_0^t B^n(u)\mathrm{d}u\right) - \mu_i\int_0^t B_i^n(u)\mathrm{d}u\right]\end{aligned} \tag{67}$$

and

$$\begin{aligned}\hat{Y}_i^n(t) &\equiv n^{-1/2}\left[R_i^n(t) - \theta_i\int_0^t Q_i^n(u)\mathrm{d}u\right]\\ &= n^{-1/2}\left[\Pi_i^{ab}\left(\theta_i\int_0^t Q_i^n(u)\mathrm{d}u\right) - \theta_i\int_0^t Q_i^n(u)\mathrm{d}u\right]\end{aligned} \tag{68}$$

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

$$\langle\hat{D}_i^n\rangle(t) = \frac{\mu_i}{n}\int_0^t B_i^n(u)\mathrm{d}u \quad\text{and}\quad \langle\hat{Y}_i^n\rangle(t) = \frac{\theta_i}{n}\int_0^t Q_i^n(u)\mathrm{d}u. \tag{69}$$

Both $\{\hat{D}_i^n(\cdot); n\in\mathbb{N}\}$ and $\{\hat{Y}_i^n(\cdot); n\in\mathbb{N}\}$ 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), it follows

$$m_i(t) = m_i(0) + \int_0^t \lambda_i(u)\mathrm{d}u - \mu_i\int_0^t m_i(u)\mathrm{d}u. \tag{70}$$

Scaling both sides of (70) by n and subtracting it from (17) gives us

$$\begin{aligned}X_i^n(t) - nm_i(t) &= X_i^n(0) - nm_i(0)\\ &\quad + A_i^n(t) - n\int_0^t \lambda_i(u)\mathrm{d}u - D^n(t) + n\mu_i\int_0^t m_i(u)\mathrm{d}u - R_i^n(t).\end{aligned}$$

Dividing both sides by $n^{1/2}$ yields

$$\begin{aligned}\hat{X}_i^n(t) &= \hat{X}_i^n(0) - \mu_i\int_0^t \hat{X}_i^n(u)\mathrm{d}u\\ &\quad + \mu_i\int_0^t \hat{Q}_{0,i}^n(u)\mathrm{d}u - (\theta_i-\mu_i)\int_0^t \hat{Q}_i^n(u)\mathrm{d}u + \hat{A}_i^n(t) - \hat{D}_i^n(t) - \hat{Y}_i^n(t).\end{aligned} \tag{71}$$

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

$$\hat{\mathcal{M}}_i^n(t) \equiv \hat{A}_i^n(t) - \hat{D}_i^n(t) - \hat{Y}_i^n(t). \tag{72}$$

Note that $\{\mathcal{M}_i^n; n\in\mathbb{N}\}$ is stochastically bounded. Using (71) and (72), we have

$$\left|\hat{X}_i^n(t)\right| \le \left|\hat{X}_i^n(0)\right| + \bar{a}\int_0^t\left[\left|\hat{X}_i^n(u)\right| + \hat{Q}_i^n(u) + \hat{Q}_{0,i}^n(u)\right]\mathrm{d}u + \left|\hat{\mathcal{M}}_i^n(t)\right|. \tag{73}$$

Adding up (73) over $i\in\mathcal{I}$ and letting $\hat{\mathbb{X}}^n \equiv \sum_{i\in\mathcal{I}}\left|\hat{X}_i^n\right|$, we obtain

$$\hat{\mathbb{X}}^n(t) \le \hat{\mathbb{X}}^n(0) + \bar{a}\int_0^t\left[\hat{\mathbb{X}}^n(u) + \hat{Q}^n(u) + \hat{Q}_0^n(u)\right]\mathrm{d}u + \sum_{i\in\mathcal{I}}\left|\hat{\mathcal{M}}_i^n(t)\right|. \tag{74}$$

In addition,

$$\hat{Q}^n(t) + \hat{Q}_0^n(t) = \left[\hat{X}^n(t) - c(t)\right]^+ \le \hat{\mathbb{X}}^n(t) + \left|c(t)\right|. \tag{75}$$

Plugging (75) into (74) yields

$$\hat{\mathbb{X}}^n(t) \le \hat{\mathbb{X}}^n(0) + \bar{a}\int_0^t |c(u)|\mathrm{d}u + 2\bar{a}\int_0^t \hat{\mathbb{X}}^n(u)\mathrm{d}u + \sum_{i\in\mathcal{I}}\left|\hat{\mathcal{M}}_i^n(t)\right|. \tag{76}$$

An application of the Gronwall's inequality with (76) establishes the stochastic boundedness of $\{\hat{\mathbb{X}}^n; n\in\mathbb{N}\}$. Thus, for $i\in\mathcal{I}$ the sequence $\{\hat{X}_i^n(\cdot); n\in\mathbb{N}\}$ is stochastically bounded. Then, the stochastic boundedness of $\{\hat{Q}^n(\cdot); n\in\mathbb{N}\}$ and $\{\hat{Q}_0^n(\cdot); n\in\mathbb{N}\}$ follows easily by (75).

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 (39) and (40), we must have

$$\bar{X}_i^n(\cdot) \equiv \frac{X_i^n(\cdot)}{n} \Rightarrow m_i(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty, \tag{77}$$

and

$$\bar{B}_i^n(\cdot) \equiv \frac{B_i^n(\cdot)}{n} = \frac{X_i^n(\cdot) - Q_i^n(\cdot) - Q_{0,i}^n(\cdot)}{n} \Rightarrow m_i(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{78}$$

Applying the continuous mapping theorem (CMT) with integration in (78), we have

$$\bar{D}_i^n(\cdot) \equiv \mu_i\int_0^{\cdot} \bar{B}_i^n(u)\mathrm{d}u \Rightarrow \mu_i\int_0^{\cdot} m_i(u)\mathrm{d}u \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{79}$$

Then, apply the CMT with composition in (79) to obtain

$$\begin{aligned}\hat{D}_i^n(\cdot) &= n^{-1/2}\left[\Pi_i^d\left(n\mu_i\int_0^{\cdot}\bar{B}_i^n(u)\mathrm{d}u\right) - n\mu_i\int_0^{\cdot}\bar{B}_i^n(u)\mathrm{d}u\right]\\ &= n^{-1/2}\left(\Pi_i^d\circ n\bar{D}_i^n(\cdot) - n\bar{D}_i^n(\cdot)\right) \Rightarrow W_i\left(\mu_i\int_0^{\cdot} m_i(u)\mathrm{d}u\right) \quad \text{in} \quad \mathcal{D},\end{aligned} \tag{80}$$

as $n\to\infty$ where we have used W_i to denote a standard Brownian motion. It is a simple exercise to show via (80) that

$$\hat{D}^n(\cdot) \equiv n^{-1/2}\left[D^n(\cdot) - n\sum_{i\in\mathcal{I}}\mu_i\int_0^{\cdot}\bar{B}_i^n(u)\mathrm{d}u\right] \Rightarrow W\left(\sum_{i\in\mathcal{I}}\mu_i\int_0^{\cdot} m_i(u)\right) \quad \text{in} \quad \mathcal{D}, \tag{81}$$

as $n\to\infty$ where W represents a reference Brownian motion.

2. *Asymptotic Negligibility of* $\{\hat{Q}_0^n(\cdot); n\in\mathbb{N}\}$. 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 $\{\hat{Q}_0^n(\cdot); n\in\mathbb{N}\}$, we first argue that $\hat{\Upsilon}_0^n(\cdot) \equiv n^{-1/2}\Upsilon_0^n(\cdot)$ vanishes as $n\to\infty$. For that purpose, define $\hat{Z}^n(\cdot) \equiv n^{-1/2}Z^n(\cdot)$. By (21),

$$\hat{\Upsilon}_0^n(t) = \hat{Z}^n(t) - \sup_{u\le t}\left\{-\hat{Z}^n(u)\right\}. \tag{82}$$

Combining (4), (22), (70), and (81) and some algebraic manipulation leads easily to

$$\hat{Z}^n(t) = -n^{1/2}\int_0^t \lambda(u)\mathrm{d}u - \mathcal{X}^n(t), \tag{83}$$

where

$$\mathcal{X}^n(t) \equiv \hat{D}^n(t) + \sum\mu_i\int_0^t\left[\hat{X}_i^n(u) - \hat{Q}_{0,i}^n(u) - \hat{Q}_i^n(u)\right]\mathrm{d}u + c(t).$$

By the C-tightness of $\hat{D}^n$ and the stochastic boundedness of $\hat{X}_i^n(u)$, $\hat{Q}_i^n$, and $\hat{Q}_{0,i}^n$, we deduce that the sequence of $\{\mathcal{X}^n(\cdot); n\in\mathbb{N}\}$ is stochastically bounded and C-tight. Define

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

From (82) and (83), it follows

$$\hat{\Upsilon}_0^n(t) = -n^{1/2}\int_{u^n(t)}^t \lambda(u)\mathrm{d}u - \mathscr{X}^n(t) + \mathscr{X}^n(u^n(t)) \geq 0. \tag{84}$$

Combining the inequality in (84) and the stochastic boundedness of $\mathscr{X}^n(\cdot)$ allows us to conclude

$$\sup_{t\leq T}\{t - u^n(t)\} = O_p(n^{-1/2}). \tag{85}$$

For a cadlag (right continuous with left limits) function $x(\cdot)$, define $|x|_T^* \equiv \sup_{t\leq T}|x(t)|$. Using (84), we can easily deduce

$$\mathbb{P}\left(\left|\hat{\Upsilon}_0^n\right|_T^* > \epsilon\right) \leq \mathbb{P}\left(\sup_{t\leq T}\{-\mathscr{X}^n(t) + \mathscr{X}^n(u^n(t))\} \geq \epsilon\right).$$

In virtue of the established C-tightness of $\mathscr{X}^n$,

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

Because ϵ is arbitrarily chosen, we have proven

$$\hat{\Upsilon}_0^n(\cdot) \equiv n^{-1/2}\Upsilon_0^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathscr{D} \quad \text{as} \quad n\to\infty. \tag{86}$$

It is immediate by Lemma 1 and the definition of $\hat{Q}_0^n$ and $\hat{\Upsilon}_0^n$ that $\hat{Q}_0^n(t) \leq \hat{\Upsilon}_0^n(t)$ for all $t \leq T$. Hence, we must have

$$\left(\hat{Q}_0^n, \hat{Q}_{0,1}^n, \ldots, \hat{Q}_{0,K}^n\right) \Rightarrow 0 \quad \text{in} \quad \mathscr{D}^{K+1} \quad \text{as} \quad n\to\infty. \tag{87}$$

3. *State-Space Collapse* by (63)

$$n^{1/2}\int_{t-U_i^n(t)}^t \lambda_i(u)\mathrm{d}u = \hat{Q}_i^n(t-) - \hat{A}_i^n[t - U_i^n(t), t) + \hat{L}_i^{n,t-U_i^n(t)}[t - U_i^n(t), t). \tag{88}$$

Note that the right-hand side is stochastically bounded owing to the stochastic boundedness of $\hat{Q}^n$, $\hat{A}_i^n$ and $\hat{R}_i^n$, along with the relation (62). By Assumption A1, the integrant λ_i is strictly positive. Hence, $\{n^{1/2}U_i^n(\cdot); n\in\mathbb{N}\}$ is stochastically bounded, for $i\in\mathcal{I}$.

Toward proving the asymptotic negligibility of $\hat{K}_i^n(\cdot)$, we show that $\hat{A}_i^n[t - U_i^n(t), t)$, $\hat{L}_i^n[t - U_i^n(t), t)$ and $e_i^n(t)$ vanish as $n\to\infty$. That $\hat{A}_i^n[t - U_i^n(t), t)$ converge to zero uniformly over $[0, T]$ is straightforward because $\hat{A}_i^n(\cdot)$ converges weakly to a Brownian motion (with a time shift) and the maximum time increment $|U_i^n|_T^*$ converges to zero in $\mathbb{R}$ as $n\to\infty$ due to the stochastic boundedness of $\{n^{1/2}U_i^n; n\in\mathbb{N}\}$. To see that $\hat{R}_i^n[t - U_i^n(t), t)$ vanishes as n grows to infinity, note that the quadratic variation

$$\langle\hat{Y}_i^n\rangle(\cdot) = \frac{\theta_i}{n}\int_0^{\cdot} Q_i^n(u)\mathrm{d}u \Rightarrow 0 \quad \text{in} \quad \mathscr{D} \quad \text{as} \quad n\to\infty, \tag{89}$$

drawing upon section 7.1 of Pang et al. (2007). The convergence in (89) implies

$$\hat{R}_i^n(\cdot) - \theta_i\int_0^{\cdot} \hat{Q}_i^n(u)\mathrm{d}u \Rightarrow 0 \quad \text{in} \quad \mathscr{D} \quad \text{as} \quad n\to\infty, \tag{90}$$

by applying the Lenglart–Rebolledo inequality; see Karatzas and Shreve (2012). In view of

$$\int_{t-U_i^n(t)}^t \hat{Q}_i^n(u)\mathrm{d}u \leq \left|\hat{Q}^n\right|_T^* |U_i^n|_T^*,$$

and that the random variable $\left|\hat{Q}^n\right|_T^* |U_i^n|_T^*$ is independent of t and converges to 0 in $\mathbb{R}$ as $n\to\infty$, we conclude that $\hat{R}_i^n[t - U_i^n(t), t]$ vanishes uniformly over $[0, T]$ as desired.

Next consider the term e_i^n given in (64). By Taylor expansion

$$\begin{aligned}|e_i^n(t)| &\equiv \left| n^{1/2} \int_{t-U_i^n(t)}^{t} \lambda_i(u)\mathrm{d}u - n^{1/2}\lambda_i(t)U_i^n(t) \right| \\ &= \left| n^{1/2}\lambda_i(t)U_i^n(t) + n^{1/2}(U_i^n(t))^2\lambda_i'(t) + o_p\left(n^{1/2}(U_i^n(t))^2\right) - n^{1/2}\lambda_i(t)U_i^n(t) \right|, \quad (91) \\ &= \left| n^{1/2}(U_i^n(t))^2\lambda_i'(t) + o_p\left(n^{1/2}(U_i^n(t))^2\right) \right| \\ &= O_p(n^{1/2}(|U_i^n|_T^*)^2),\end{aligned}$$

where the last equality is due to Assumption A1, which guarantees the boundedness of $|\lambda_i'(\cdot)|$ over any compact intervals. The random variable $n^{1/2}(|U_i^n|_T^*)^2$ is independent of time t and converges to zero as $n \to \infty$ because $n^{1/2}|U_i^n|_T^*$ is stochastically bounded and $|U_i^n|_T^*$ goes to zero as n approaches infinity. We thus establish the asymptotic negligibility of $\hat{K}_i^n(\cdot)$, for $i \in \mathcal{I}$.

From here, the proof follows closely that of Theorem 2 with simple modifications. Define the imbalance process

$$\Xi_i^n(\cdot) \equiv \frac{\hat{U}_i^n(\cdot)}{v_i(\cdot)} - \frac{\Sigma_{i\in\mathcal{I}}\lambda_i(\cdot)\hat{U}_i^n(\cdot)}{\Sigma_{i\in\mathcal{I}}\lambda_i(\cdot)v_i(\cdot)} \quad \text{for } i \in \mathcal{I}. \tag{92}$$

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

Suppose that $\Xi_i^n(0) \neq 0$. Our analysis below indicates that it takes infinitesimally small time for the imbalance process Ξ_i^n to hit zero. Hence, assume without loss of generality that $\Xi_i^n(0) = 0$. We aim to show that, for each $i \in \mathcal{I}$, the imbalance process $\Xi_i^n(\cdot)$ is infinitely close to the zero function; that is, for an arbitrary $\epsilon > 0$,

$$\mathbb{P}(|\Xi_i^n|_T^* > \epsilon) \to 0 \quad \text{as} \quad n \to \infty. \tag{93}$$

Define the stopping time (depending on ϵ)

$$\tilde{\tau}_i^n \equiv \inf\{t > 0 : |\Xi_i^n(t)| > \epsilon\}.$$

Then, to establish (93), it suffices to show $\mathbb{P}(\tilde{\tau}_i^n \le T) \to 0$ as $n \to \infty$. Note that a positive imbalance guarantees the existence of a negative imbalance. Thus, the problem further boils down to showing, for each $i \in \mathcal{I}$,

$$\mathbb{P}(\tau_i^n \le T) \to 0 \quad \text{as} \quad n \to \infty, \tag{94}$$

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

$$\sigma_i^n \equiv \sup\{t \ge 0 | t < \tau_i^n, \Xi_i^n(t) \ge -\epsilon/4\}.$$

With the initial condition $\Xi_i^n(0) = 0$, such a random time σ_i^n is guaranteed to exist on the event C. Using the definition of τ_i^n and σ_i^n allows us to conclude that $\Xi_i^n(t) \le -\epsilon/4$ and $\hat{Q}^n(t) > 0$ for all $t \in (\sigma_i^n, \tau_i^n]$. It is easily verifiable that the two conditions $(i) - (ii)$ described in the proof of Theorem 2 are satisfied for $k = i, \eta_1 = \sigma_i^n$ and $\eta_2 = \tau_i^n$.

Let Δ_i^n be the queue-imbalance process given in (114) with $r_i(\cdot)$ there being replaced by $\gamma^{-1}(\cdot)\lambda_i(\cdot)v_i(\cdot)$. Using (66) and applying union bound, we have

$$\begin{aligned}\mathbb{P}(\tau_i^n \le T) &\le \mathbb{P}\left(\Xi_i^n(\sigma_i^n) \ge -\epsilon/4,\ \Xi_i^n(\tau_i^n) < -\epsilon\right) \\ &\le \mathbb{P}\left(\Delta_i^n(\sigma_i^n-) > -\lambda_i(\sigma_i^n-)v_i(\sigma_i^n-)\,\epsilon/2,\ \Delta_i^n(\tau_i^n-) < -3\lambda_i(\tau_i^n-)v_i(\tau_i^n-)\,\epsilon/4\right) \quad (95) \\ &\quad + \mathbb{P}\left(\sup_{t\le T}\left| -\hat{K}_i^n(t) + r_i(t)\Sigma_{i\in\mathcal{I}}\hat{K}_i^n(t)\right| > \lambda_* v_*\,\epsilon/4\right).\end{aligned}$$

Repeating the argument in the proof of Theorem 2, one can easily argue that $\tau_i^n - \sigma_i^n = O_p(n^{-1/2})$, and so

$$\mathbb{P}\left(\Delta_i^n(\sigma_i^n) > -\lambda_i(\sigma_i^n)v_i(\sigma_i^n)\,\epsilon/2,\ \Delta_i^n(\tau_i^n) < -3\lambda_i(\tau_i^n)v_i(\tau_i^n)\,\epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{96}$$

By the established asymptotic negligibility of $\hat{K}_i^n$, we have

$$\mathbb{P}\left(\sup_{t\le T}\left|-\hat{K}_i^n(t)+r_i(t)\Sigma_{i\in\mathcal{I}}\hat{K}_i^n(t)\right|>\lambda_*v_*\epsilon/4\right)\to 0 \quad \text{as} \quad n\to\infty. \tag{97}$$

Combining (95), (96), and (97) yields (94) and hence (93).

In view of (93), we conclude

$$\Xi_i^n \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty,$$

with Ξ_i^n given in (92). Combining the above with (66) yields

$$\Theta_i^n(\cdot)\equiv\hat{Q}_i^n(\cdot)-\gamma(\cdot)^{-1}v_i(\cdot)\lambda_i(\cdot)\hat{Q}^n(\cdot)\Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty, \quad \text{for } i\in\mathcal{I}. \tag{98}$$

Using the convergence-together lemma, we have

$$(\Theta_1^n,\ldots,\Theta_K^n)\Rightarrow(0,\ldots,0) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n\to\infty. \tag{99}$$

4. *Diffusion Limits.* Using the CMT with integration in (98), we obtain

$$\Upsilon_i^n(\cdot)\equiv\int_0^{\cdot}\hat{Q}_i^n(u)\mathrm{d}u-\int_0^{\cdot}\gamma(u)^{-1}v_i(u)\lambda_i(u)\hat{Q}^n(u)\mathrm{d}u\Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{100}$$

Combining (71), (75), and (100) gives

$$\begin{aligned}\hat{X}_1^n(t)=\hat{X}_i^n(0)-\mu_i\int_0^t\hat{X}_i^n(u)\mathrm{d}u-(\theta_i-\mu_i)\Upsilon_i^n(t)+\mu_i\int_0^t\hat{Q}_{0,i}^n(u)\mathrm{d}u+\hat{A}_i^n(t)-\hat{D}_i^n(t)\\-\hat{Y}_i^n(t)-(\theta_i-\mu_i)\int_0^t\gamma(u)^{-1}v_i(u)\lambda_i(u)\left\{\left[\hat{X}^n(u)-c(u)\right]^+-\hat{Q}_0^n(u)\right\}\mathrm{d}u.\end{aligned} \tag{101}$$

An application of theorem 4.1 of Pang et al. (2007) together with (14), (80), (87), (89), and (100) allows us to establish the many-server heavy-traffic limit for $\{\hat{X}_i^n(\cdot);n\in\mathbb{N}\}$:

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

where $\hat{X}_i$ satisfies the differential Equation (43). Then apply the convergence-together lemma with (99) we conclude

$$\left(\hat{X}_1^n,\ldots,\hat{X}_K^n,\hat{Q}_1^n,\ldots,\hat{Q}_K^n\right)\Rightarrow\left(\hat{X}_1,\ldots,\hat{X}_K,\hat{Q}_1,\ldots,\hat{Q}_K\right) \quad \text{in} \quad \mathcal{D}^{2K}, \tag{102}$$

as $n\to\infty$ where the limiting processes $\hat{Q}_i$ are given in (44).

5. *Potential Delay Asymptotics.* To establish heavy-traffic stochastic-process limits for potential delays, we follow the solution approach as in section 3 of Talreja and Whitt (2009). Paralleling the proof of theorem 3.1 in that paper, we decompose the proof into two steps. The first step is to show that all processes in (26) have proper fluid and diffusion limits. For each $i\in\mathcal{I}$, introduce the fluid-scaled processes

$$\bar{A}_i^n(\cdot)\equiv A_i^n(\cdot)/n, \qquad \bar{\Psi}_i^n(\cdot)\equiv\Psi_i^n(\cdot)/n, \qquad \bar{Q}_i^n(\cdot)\equiv Q_i^n(\cdot)/n \quad \text{and} \quad \bar{R}_i^n(\cdot)\equiv R_i^n(\cdot)/n.$$

Clearly, we have

$$(\bar{A}_i^n,\bar{\Psi}_i^n,\bar{R}_i^n,\bar{Q}_i^n)\Rightarrow(\Lambda_i,\Lambda_i,0,0) \quad \text{in} \quad \mathcal{D}^4 \quad \text{as} \quad n\to\infty. \tag{103}$$

Now define

$$\hat{\Psi}_i^n(\cdot)\equiv n^{-1/2}(\Psi_i^n(\cdot)-n\Lambda_i(\cdot)). \tag{104}$$

Then,

$$\left(\hat{A}_i^n,\hat{\Psi}_i^n,\hat{R}_i^n,\hat{Q}_i^n\right)\Rightarrow\left(\hat{A}_i,\hat{\Psi}_i,\hat{R}_i,\hat{Q}_i\right) \quad \text{in} \quad \mathcal{D}^4 \tag{105}$$

as $n\to\infty$ where $\hat{A}_i^n,\hat{\Psi}_i^n,\hat{Q}_i^n$ and $\hat{R}_i^n$ are given in (14), (104), (40), and (61), respectively, and

$$\hat{R}_i(\cdot)\equiv\theta_i\int_0^{\cdot}\hat{Q}_i(u)\mathrm{d}u, \qquad \hat{\Psi}_i(\cdot)\equiv\hat{Q}_i(0)+\hat{A}_i(\cdot)-\hat{Q}_i(\cdot)-\hat{R}_i(\cdot),$$

where $\hat{A}_i$ and $\hat{Q}_i$ are given in (14) and (40), respectively.

The second step is to construct a lower and an upper bound for the process V_i^n:

$$V_i^{n,l}(t) \le V_i^n(t) \le V_i^{n,u}(t), \tag{106}$$

where

$$\begin{aligned} V_i^{n,l}(t) &= \inf\{s \ge 0 : \bar{\Psi}_i^n(t+s) + \bar{R}_i^n(t+s) \ge \bar{Q}_i^n(0) + \bar{A}_i^n(t)\} \\ V_i^{n,u}(t) &= \inf\{s \ge 0 : \bar{\Psi}_i^n(t+s) \ge \bar{Q}_i^n(0) + \bar{A}_i^n(t) - \bar{R}_i^n(t)\}. \end{aligned} \tag{107}$$

One may attempt to apply the corollary of Puhalskii (1994) together with (103) and (105) to get

$$n^{1/2} V_i^{n,l}(\cdot) \Rightarrow \hat{Q}_i(\cdot)/\Lambda_i'(\cdot) \quad \text{and} \quad n^{1/2} V_i^{n,u}(\cdot) \Rightarrow \hat{Q}_i(\cdot)/\Lambda_i'(\cdot),$$

in $\mathcal{D}$ as $n \to \infty$, and then use (106) and (107) to conclude the desired results. However the right-hand side of the second line in (107) does not satisfy the conditions of the corollary. In particular, $\bar{Q}_i^n(0) + \bar{A}_i^n - \bar{R}_i^n$ is not necessarily nondecreasing. To resolve the problem, we use the same linear-interpolation technique as illustrated in figure 1 of Talreja and Whitt (2009). The key is to construct a process $\tilde{V}_i^{n,u}$ such that $\tilde{V}_i^{n,u}(t) \ge V_i^{n,u}(t)$ for all $t \ge 0$ and

$$n^{1/2} \tilde{V}_i^{n,u}(\cdot) \Rightarrow \hat{Q}_i(\cdot)/\Lambda_i'(\cdot). \tag{108}$$

A standard sandwiching argument allows us to conclude

$$\hat{V}_i^n(\cdot) \equiv n^{1/2} V_i^n(\cdot) \Rightarrow \frac{\hat{Q}_i(\cdot)}{\Lambda_i'(\cdot)} = v_k(t) \cdot \gamma(\cdot)^{-1} \left[\hat{X}(\cdot) - c(\cdot)\right]^+ \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty,$$

jointly with (102).

Condition (108) holds if the error caused by these linear interpolations is asymptotically negligible. The proof of lemma 7.1 in Talreja and Whitt (2009) applies here if we replace the departure process D_n there with our assignment process Ψ_i^n.

To sum up, we have shown that

$$\begin{aligned} &\left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n(\cdot), \ldots, \hat{V}_K^n\right) \\ &\Rightarrow \left(\hat{X}_1, \ldots, \hat{X}_K, \hat{Q}_1, \ldots, \hat{Q}_K, \hat{V}_1, \ldots, \hat{V}_K\right) \quad \text{in} \quad \mathcal{D}^{3K} \quad \text{as} \quad n \to \infty. \quad \square \end{aligned} \tag{109}$$

Proof of Theorem 2. The key is to observe that, whenever the queue ratio moves away from the target, it always takes the scheduler $O(n^{-1/2})$ time to correct the digression. To give an idea on why the system behaves asymptotically as stated in Theorem 2, consider a many-server queue with two customer classes. Suppose that the system is to maintain a fixed queue ratio r_1/r_2. Then, if ever $Q_1/Q_2 < r_1/r_2$, the next available server always chooses to serve a class-2 customer until after the inequality changes direction; that is, $Q_1/Q_2 \ge r_1/r_2$. Notice that departures occur at the rate of order $O(n)$ whereas the queue lengths live on the scale of $O(n^{1/2})$. Thus it always takes $O(n^{-1/2})$ amount of time before the inequality changes direction. The proof below formalizes this intuition.

We start by analyzing a scenario in which no customer of certain class enters service over a time interval. More precisely, let η_1 and η_2 be $[0, T]$-valued random variable satisfying $\eta_1 \le \eta_2$. Fix $k \in \mathcal{J}$ and let H denote any event under which

1. No server has ever been idle over the period $[\eta_1, \eta_2]$;
2. No class-k customer enters service over $[\eta_1, \eta_2]$.

Working with the same notation $x(t_1, t_2] \equiv x(t_2) - x(t_1)$ for a function $x(\cdot)$ in t and exploiting (18) and the nonidling condition (*i*), one can easily derive

$$\sum_{i \in \mathcal{J}} A_i^n(\eta_1, \eta_2] - D^n(\eta_1, \eta_2] - \sum_{i \in \mathcal{J}} R_i^n(\eta_1, \eta_2] = s^n(\eta_1, \eta_2] + Q_0^n(\eta_1, \eta_2] + \sum_{i \in \mathcal{J}} Q_i^n(\eta_1, \eta_2]. \tag{110}$$

Moreover, by condition (*ii*), no customer enters service from the k-th queue and so

$$Q_k^n(\eta_1, \eta_2] = A_k^n(\eta_1, \eta_2] - R_k^n(\eta_1, \eta_2]. \tag{111}$$

Combining (110) and (110) yields

$$\sum_{i\neq k} A_i^n(\eta_1,\eta_2] - D^n(\eta_1,\eta_2] - \sum_{i\neq k} R_i^n(\eta_1,\eta_2] = s^n(\eta_1,\eta_2] + Q_0^n(\eta_1,\eta_2] + \sum_{i\neq k} Q_i^n(\eta_1,\eta_2]. \tag{112}$$

Now, using (70) and (112) and following similar derivation used for (71), we have

$$\begin{aligned} n^{1/2}\int_{\eta_1}^{\eta_2} \lambda_k(u)\mathrm{d}u \quad &= \sum_{i\neq k} \hat{A}_i^n(\eta_1,\eta_2] - \hat{D}^n(\eta_1,\eta_2] - \sum_{i\neq k} \hat{R}_i^n(\eta_1,\eta_2] - c(\eta_1,\eta_2] - \hat{Q}_0^n(\eta_1,\eta_2] \\ &\quad - \sum_{i\neq k} \hat{Q}_i^n(\eta_1,\eta_2] - \sum_{i\in\mathcal{I}} \mu_i\left(\int_{\eta_1}^{\eta_2} \hat{X}_i^n(u)\mathrm{d}u - \int_{\eta_1}^{\eta_2} \hat{Q}_i^n(u)\mathrm{d}u\right). \end{aligned} \tag{113}$$

Recall the set of ratio functions $r(\cdot) \equiv (r_1(\cdot),\ldots,r_K(\cdot))$ with the constraints: (1) each component $r_i(\cdot)$ is continuous in t; and (2) $\sum_{i\in\mathcal{I}} r_i(\cdot) = 1$. Next, define for each $i\in\mathcal{I}$ the imbalance process

$$\Delta_i^n(\cdot) \equiv \hat{Q}_i^n(\cdot) - r_i(\cdot)\hat{Q}^n(\cdot). \tag{114}$$

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 $\hat{Q}_i^n(\cdot)$ is infinitely close to $\hat{Q}_i^n(\cdot)$ as n grows. More precisely, we aim to show that, for each $i\in\mathcal{I}$ and $\epsilon>0$,

$$\mathbb{P}\left(|\Delta_i^n|_T^* > \epsilon\right) \to 0 \quad \text{as} \quad n\to\infty. \tag{115}$$

Define a stopping time (depending on ϵ)

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

Then, to establish (115), it suffices to show $\mathbb{P}(\tilde{\tau}_i^n \le T) \to 0$ as $n\to\infty$. Note that $\sum_{i\in\mathcal{I}} \Delta_i^n(\cdot) = 0$. Thus the problem further boils down to showing

$$\mathbb{P}(\tau_i^n \le T) \to 0 \quad \text{as} \quad n\to\infty,$$

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

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

With the initial condition $\Delta_i^n(0) = 0$, such a random time σ_i^n is guaranteed to exist on the event C. Taking $k=i$, $\eta_1 = \sigma_i^n$ and $\eta_2 = \tau_i^n$ and using the definition of τ_i^n and σ_i^n allows us to conclude that $\Delta_i^n(t) \le -\epsilon/2$ and $\hat{Q}^n(t) > 0$ for all $t\in(\sigma_i^n,\tau_i^n]$. Therefore both condition (i) and (ii) hold for $\eta_1 = \sigma_i^n$ and $\eta_2 = \tau_i^n$. From (113), it follows

$$\begin{aligned} n^{1/2}\int_{\sigma_i^n}^{\tau_i^n} \lambda_i(u)\mathrm{d}u \le \sum_{j\neq i} \hat{A}_j^n(\sigma_i^n,\tau_i^n] - \hat{D}^n(\sigma_i^n,\tau_i^n] - \sum_{i\neq k} \hat{R}_i^n(\sigma_i^n,\tau_i^n] - c(\sigma_i^n,\tau_i^n] - \hat{Q}_0^n(\sigma_i^n,\tau_i^n] \\ - \sum_{j\neq i} \hat{Q}_i^n(\sigma_i^n,\tau_i^n] - \sum_{i\in\mathcal{I}} \mu_i\left(\int_{\sigma_i^n}^{\tau_i^n} \hat{X}_i^n(u)\mathrm{d}u - \int_{\sigma_i^n}^{\tau_i^n} \hat{Q}_i^n(u)\mathrm{d}u\right). \end{aligned} \tag{116}$$

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

$$\begin{aligned} \mathbb{P}(\tau_i^n \le T) &\le \mathbb{P}(\Delta_i^n(\tau_i^n) < -\epsilon, \Delta_i^n(\sigma_i^n) \ge -\epsilon/2), \\ &\le \mathbb{P}\left(\hat{Q}_i^n(\tau_i^n) - \hat{Q}_i^n(\sigma_i^n) - \Gamma_i^n(\sigma_i^n,\tau_i^n] < -\epsilon/2\right) \\ &\le \mathbb{P}\left(\hat{Q}_i^n(\tau_i^n) - \hat{Q}_i^n(\sigma_i^n) - \Gamma_i^n(\sigma_i^n,\tau_i^n] < -\epsilon/2, \Gamma_i^n(\sigma_i^n,\tau_i^n] \le \epsilon'\right) \\ &\quad + \mathbb{P}\left(\hat{Q}_i^n(\tau_i^n) - \hat{Q}_i^n(\sigma_i^n) - \Gamma_i^n(\sigma_i^n,\tau_i^n] < -\epsilon/2, \Gamma_i^n(\sigma_i^n,\tau_i^n] > \epsilon'\right) \\ &\le \mathbb{P}\left(\hat{Q}_i^n(\tau_i^n) - \hat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) + \mathbb{P}(\Gamma_i^n(\sigma_i^n,\tau_i^n] > \epsilon/4). \end{aligned} \tag{117}$$

Recall that our goal is to show $\mathbb{P}(\tau_i^n \le T)$ goes to zero as $n \to \infty$. To that end, we argue that both terms at the right end of (117) converge to zero as n grows to infinity.

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

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

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

$$\mathbb{P}\left(\hat{Q}_i^n(\tau_i^n) - \hat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) \le \mathbb{P}\left(\hat{R}_i^n(\tau_i^n) - \hat{R}_i^n(\sigma_i^n) < -\epsilon/4\right), \tag{118}$$

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. Its proof is deferred to the end of the section.

Lemma 2. *Both* $\{\hat{Q}^n(\cdot); n \in \mathbb{N}\}$ *and* $\{\hat{R}_i^n(\cdot); n \in \mathbb{N}\}$ *are C-tight under the assumptions of Theorem 2.*

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

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

Combining the above with (118) allows us to conclude that

$$\mathbb{P}\left(\hat{Q}_i^n(\tau_i^n) - \hat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{119}$$

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

$$\mathbb{P}(\Gamma_i^n(\sigma_i^n, \tau_i^n] > \epsilon/4) \to 0 \quad \text{as} \quad n \to \infty. \tag{120}$$

Combining (117), (119), and (120) yields

$$\mathbb{P}(\tau_i^n \le T) \to 0 \quad \text{as} \quad n \to \infty,$$

which in turn implies

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

for all $i \in \mathcal{I}$. The convergence can be strengthened to joint convergence by the fact that all the limits are deterministic process. This is again a SSC result. Repeating steps 4 and 5 in the Proof of Theorem 1 leads us to the conclusion of Theorem 2. □

Proof of Lemma 2. By (75), $\{\hat{Q}^n(\cdot); n \in \mathbb{N}\}$ is C-tight if $\{\hat{X}_i^n; n \in \mathbb{N}\}$ is C-tight for $i \in \mathcal{I}$. The latter holds true if the martingales $\hat{A}_i^n$, $\hat{D}_i^n$, and $\hat{Y}_i^n$ are C-tight, owing to (71) and the established stochastic boundedness of $\hat{X}_i^n$ and $\hat{Q}^n$. But $\hat{A}_i^n$, $\hat{D}_i^n$, and $\hat{Y}_i^n$ are C-tight, due to (14), (80), and (89). Hence $\{\hat{Q}^n(\cdot); n \in \mathbb{N}\}$ is C-tight. The C-tightness of $\{\hat{R}_i^n(\cdot); n \in \mathbb{N}\}$ follows from (90) and the stochastic boundedness of $\{\hat{Q}_i^n(\cdot); n \in \mathbb{N}\}$ drawing upon the stochastic boundedness of $\{\hat{Q}^n(\cdot); n \in \mathbb{N}\}$. □

7. Proofs for Asymptotic Feasibility and Optimality

The following lemma is the crucial ingredient in the proof of asymptotic optimality. We remark that it would also be possible to apply theorem 7.1 of Feldman et al. (2008), which applies to birth–death processes.

Lemma 3 (Comparison Principle for Piecewise-Linear Diffusions). *Consider the following two stochastic integral equations:*

$$\begin{aligned} X(t) &= X(0) - \beta_1 \int_0^t X(u)\mathrm{d}u + \beta_2 \int_0^t [X(u) - c(u)]^+ \mathrm{d}u + \int_0^t \sigma(u)\mathrm{d}W(u), \\ X'(t) &= X'(0) - \beta_1 \int_0^t X'(u)\mathrm{d}u + \beta_2 \int_0^t [X'(u) - c'(u)]^+ \mathrm{d}u + \int_0^t \sigma(u)\mathrm{d}W'(u), \end{aligned} \tag{121}$$

where $X(0) \overset{\text{a.s.}}{=} X'(0)$ *and* $\beta_1 \ge \beta_2 \ge 0$. *Let* $Q(t) \equiv [X(t) - c(t)]^+$ *and* $Q'(t) \equiv [X'(t) - c'(t)]^+$. *If* $c' \le c$ *and* $c'(t) < c(t)$ *for some* $t \ge 0$, *then* $\mathbb{E}[Q'(t)] - \mathbb{E}[Q(t)] > 0$.

Proof of Lemma 3. Note that the expectation of a random variable depends only on its probability distribution. We can thus define X and X' on the same probability space where $W'(t) \equiv W(t)$. The idea is to couple two diffusion processes in such a way that they agree as often as possible. A direct application of theorem 1.3 of Yamada (1973) allows us to conclude that $X(t) \le X'(t)$ almost surely. Let A_t be the event $\{X(t) \ge c(t)\}$. Then,

$$Q'(t) - Q(t) = [X'(t) - c'(t)]^+ \ge 0 \quad \text{on} \quad A_t^c,$$

and thus

$$\mathbb{E}[Q'(t)1_{A_t^c}] \ge \mathbb{E}[Q(t)1_{A_t^c}].$$

Similarly,

$$Q'(t) - Q(t) = X'(t) - X(t) + c'(t) - c(t) \quad \text{on} \quad A_t.$$

Hence,

$$\mathbb{E}[Q'(t)1_{A_t}] - \mathbb{E}[Q(t)1_{A_t}] \ge \mathbb{P}(A_t)(c(t) - c'(t)) > 0.$$

Combining the above yields

$$\mathbb{E}[Q'(t)] - \mathbb{E}[Q(t)] = \mathbb{E}[Q'(t)1_{A_t}] - \mathbb{E}[Q(t)1_{A_t}] + \mathbb{E}[Q'(t)1_{A_t^c}] - \mathbb{E}[Q(t)1_{A_t^c}] > 0.$$

This completes the proof of Lemma 3. □

Proof of Theorem 3. Toward proving asymptotic feasibility, we apply Fatou's lemma and Theorem 1 to conclude

$$\limsup_{n\to\infty} \mathbb{E}[V_i^n(t)/w_i^n(t)] \le \mathbb{E}[\hat{V}_i(t)/w_i(t)] = \mathbb{E}[\hat{Q}(t)/\vartheta(t)] = 1, \quad \text{for } i \in \mathcal{J}.$$

To prove asymptotic optimality, suppose by way of contradiction that condition (29) is violated for (s'^n, π^n) at time t. Using Fatou's lemma, we get

$$\liminf_{n\to\infty} \mathbb{E}[V_i'^n(t)/w_i^n(t)] \ge \mathbb{E}[\hat{Q}'(t)/\vartheta(t)] > \mathbb{E}[\hat{Q}(t)/\vartheta(t)] = 1,$$

where the second inequality follows by applying Lemma 3 with $\beta_1 = \mu, \beta_2 = \mu - \theta$ and $\sigma(t) = \sqrt{\lambda(t) + \mu m(t)}$. □

Proof of Theorem 4. To establish asymptotic feasibility, apply Portmanteau theorem and Theorem 1 to get

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(V_i^n(t)/w_i^n(t) \ge 1+\epsilon) &\le \mathbb{P}(\hat{V}(t)/w_i(t) \ge 1+\epsilon) \\ &= \mathbb{P}(\hat{Q}(t)/\vartheta(t) \ge 1+\epsilon) \le \alpha, \quad \text{for } i \in \mathcal{J}. \quad \square\end{aligned}$$

Proof of Theorem 5. The proof follows closely the steps in the proof of Theorems 3 and 4. By Fatou's lemma,

$$\limsup_{n\to\infty} \mathbb{E}[Q^n(t)/q^n(t)] \le \mathbb{E}[\hat{Q}(t)/q(t)] = 1.$$

By Portmanteau theorem, we have, for $i = 1, \ldots, K-1$,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(V_i^n(t)/w_i^n(t) \ge 1+\epsilon) &\le \mathbb{P}(\hat{V}(t)/w_i(t) \ge 1+\epsilon) \\ &= \mathbb{P}\left(\hat{X}(t) - c^*(t) \ge (1+\epsilon)\cdot x(t)\right) \le \alpha.\end{aligned}$$

Toward proving asymptotic optimality, suppose by way of contradiction that condition (34) is violated for (s'^n, π^n) at time t. Using Fatou's lemma again, we obtain

$$\liminf_{n\to\infty} \mathbb{E}[Q'^n(t)/q^n(t)] \ge \mathbb{E}[\hat{Q}'(t)/q(t)] > \mathbb{E}[\hat{Q}(t)/q(t)] = 1,$$

where the second inequality follows by applying Lemma 3. □

8. Directions for Future Research

There are many important directions for future research.

1. The major item left undone in the present study is identifying the control functions $c^*(t)$ for the one-dimensional limit process $\hat{X}(t)$ in the staffing solutions in Section 4.7. We would like to obtain explicit solutions as illustrated in Section 5.3 for Case 3 in Section 1.3.5.

2. However, the paper can be of great value without identifying the control functions $c^*(t)$ by applying previous staffing algorithms for single-class models for that purpose (for which there are not yet supporting MSHT theory). The combination of HLDR with previously developed single-class staffing algorithms applied to the aggregate model, as suggested in Section 1.3.6, remains to be carefully explored.

3. A great appeal of the present approach is that it extends naturally to non-Markov models, as illustrated in Section 5.3.2. In particular, the ratio scheduling rules directly extend outside the Markovian domain. Hence, it remains to also treat the staffing. For that purpose, it is significant that there has already been success staffing for single-class non-Markovian many-server models, for example, as in He et al. (2016), Liu and Whitt (2012b), and Whitt and Zhao (2017). However, we do remark that many of these approaches involve the overloaded ED and underloaded QD MSHT limiting regimes instead of the QED regime used here. It remains to be investigated how these developed algorithms apply in this multiclass settings.

4. It remains to establish supporting QED MSHT limits for TV non-Markovian models.

Acknowledgments.

The authors thank Yunan Liu, Kyle Hovey, the anonymous reviewers, and the editors for helpful constructive comments.

References

Arapostathis A, Pang G (2016) Ergodic diffusion control of multiclass multi-pool networks in the Halfin-Whitt regime. *Ann. Appl. Probab.* 26(5):3110–3153.

Arapostathis A, Biswas A, Pang G (2015) Ergodic control of multi-class $M/M/N + M$ queues in the Halfin-Whitt regime. *Ann. Appl. Probab.* 25(6): 3511–3570.

Armony M, Israelit S, Mandelbaum A, Marmor Y, Tseytlin Y, Yom-Tov G (2015) Patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Atar R (2005) Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15(4):2606–2650.

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

Atar R, Shaki YY, Shwartz A (2011) A blind policy for equalizing cumulative idleness. *Queueing Systems* 67(4):275–293.

Bullard MJ, Chan T, Brayman C, Warren D, Musgrave E, Unger B (2014) Revisions to the Canadian emergency department triage and acuity scale (CTAS) guidelines. *Can. J. Emerg. Medicine* 10(2):136–142.

Dai J, Tezcan T (2008) Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* 59(2):95–134.

Dai J, Tezcan T (2011) State space collapse in many-server diffusion limits of parallel server systems. *Math. Oper. Res.* 36(2):271–320.

Eick SG, Massey WA, Whitt W (1993) The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41(4):731–742.

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

Fendick KW, Whitt W (1989) Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proc. IEEE* 71(1):171–194.

Fendick KW, Saksena V, Whitt W (1989) Dependence in packet queues. *IEEE Trans Commun.* 37(11):1173–1183.

Fendick KW, Saksena V, Whitt W (1991) Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans Commun.* 39(8):1231–1244.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

Gurvich I, Whitt W (2009a) Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* 34(2):363–396.

Gurvich I, Whitt W (2009b) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.

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

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

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

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

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

Ingolfsson A, Akhmetshina E, Budge S, Li Y, Wu X (2007) A survey and experimental comparison of service-level approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS J. Comput.* 19(2):201–214.

Jacod J, Shiryaev AN (2013) *Limit Theorems for Stochastic Processes*, Vol. 288 (Springer Science & Business Media, Berlin).

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

Karatzas I, Shreve S (2012) *Brownian Motion and Stochastic Calculus*, Vol. 113 (Springer Science & Business Media, Berlin).

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

Li N, Stanford DA (2016) Multi-server accumulating priority queues with heterogeneous servers. *Eur. J. Oper. Res.* 252(3):866–878.

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

Liu Y (2018) Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Oper. Res.* 66(6):1000–1000.
Liu Y, Whitt W (2012a) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71(4):405–444.
Liu Y, Whitt W (2012b) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6):1551–1564.
Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* 52(6):836–855.
Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1-2):149–201.
Massey WA, Whitt W (1993) Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1):183–250.
Milner JM, Olsen T (2008) Service-level agreements in call centers: Perils and prescriptions. *Management Sci.* 54(2):238–252.
Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surv.* 4:193–267.
Puhalskii A (1994) On the invariance principle for the first passage time. *Math. Oper. Res.* 19(4):946–954.
Puhalskii AA (2013) On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Math. Methods Oper. Res.* 78(1):119–148.
Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley, New York).
Sharif AB, Stanford DA, Taylor P, Ziedins I (2014) A multi-class multi-server accumulating priority queue with application to health care. *Oper. Res. Health Care* 3(2):73–79.
Shi P, Chou MC, Dai JG, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Sci.* 62(1):1–28.
Simon HA (1947) *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization* (Macmillan, New York).
Simon HA (1979) Rational decision making in business organizations. *Amer. Econom. Rev.* 69(4):493–513.
Stanford DA, Taylor P, Ziedins I (2014) Waiting time distributions in the accumulating priority queue. *Queueing Systems* 77(3):297–330.
Talreja R, Whitt W (2008) Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing. *Management Sci.* 54(8):1513–1527.
Talreja R, Whitt W (2009) Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab.* 19(6):2137–2175.
Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c - \mu$ rule. *Ann. Appl. Probab.* 5(3) 809–833.
Whitt W (2002) *Stochastic-Process Limits* (Springer Science & Business Media, Berlin).
Whitt W (2006) Sensitivity of performance in the erlang-a queueing model to changes in the model parameters. *Oper. Res.* 54(2):247–260.
Whitt W (2015) Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* 81(4):341–378.
Whitt W (2017) Time-varying queues, Working paper, Columbia University, New York, http://www.columbia.edu/~ww2040/allpapers.html.
Whitt W, Zhang X (2017) A data-driven model of an emergency department. *Oper. Res. Health Care* 12:1–15.
Whitt W, Zhao J (2017) Staffing to stabilizing blocking in loss models with non-Markovian arrivals. *Naval Res. Logist.* 64(3):177–202.
Yamada T (1973) On a comparison theorem for solutions of stochastic differential equations and its applications. *J. Math. Kyoto Univ.* 13(3):497–512.