Stochastic Systems

arXiv: math.PR/0000000

DELAY-BASED SERVICE DIFFERENTIATION IN A MANY-SERVER QUEUE WITH TIME-VARYING ARRIVAL RATES

By Xu Sun* and Ward Whitt*

We study the problem of delay-based service differentiation in a multi-class many-server queueing system with time-varying (TV) arrival rates. Gurvich and Whitt (2010) showed that fixed-queue-ratio (FQR) controls for scheduling (selecting customers from queue to enter service) can achieve desired class-dependent delay targets in a stationary setting. We show that this good property can break down with with class-dependent TV arrival rates. By simulation and many-server heavy-traffic (MSHT) limits, we show that FQR can fail to stabilize delays ratios when the ratios of the arrival-rate functions for different classes are not nearly constant. To remedy this deficiency, we propose an alternative family of ratio controls that exploits head-of-line delay information. The head-of-line-delay-ratio (HLDR) control is a blind scheduling rule that extends the accumulating-priority control proposed by Kleinrock (1964), which has recently been reconsidered by Stanford et al. (2014) and Sharif et al. (2014). By simulation and MSHT limits, we show that HLDR stabilizes delay ratios at desired targets. We find that these TV results can be explained by the sample-path version of the TV MSHT Little's law that is a consequence of the MSHT limit.

1. Introduction. In this paper, we study delay-based service differentiation in a many-server queue via ratio controls in the presence of diverse customer needs (multiple customer classes) and time-varying (TV) arrival rates. Motivation for our study is provided by hospital emergency departments, where the arrival rate is typically strongly TV and the customer (patient) lengths-of-stay and delays are long enough to interact with that time variation; e.g., see §3 of [2] and §3 of [40].

Gurvich and Whitt [15] showed that *fixed-queue-ratio* (FQR) *controls* that schedule (select the next customer to enter service from queue when a server becomes free) aiming to keep the queue lengths at fixed ratios also are effective for achieving delay-based service-differentiation in stationary large-scale service systems modeled as many-server queues, delicately balancing the service levels of the different classes. (Routing new arrivals to alternative service pools was considered there too, but we consider only scheduling for a single service pool.) Indeed, the goals are achieved asymptotically in the *many-server heavy-traffic* (MSHT) *limit*; also see [13, 14].

We wanted to see how the FQR control performs with time-varying arrival rates,

*Department of Industrial Engineering and Operations Research, Columbia University

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

so we conducted simulation experiments to investigate. We found that FQR controls remain quite effective for balancing the queue lengths over time, keeping them near desired ratios, but that the FQR controls can be highly ineffective at the indirect goal of stabilizing delays at fixed ratios; see Figure 2 in §2.

The property that causes difficulties for FQR is class-dependent arrival rates, i.e., where the ratios of the arrival rates of two different classes varies strongly over time. It is thus significant that class-dependent arrival rates may indeed occur in applications. For example, §3.5 of [40] shows that the proportion of arrivals to the Israeli emergency department (ED) that are admitted to an internal ward of the hospital varied strongly over time. Since the admitted patients tend to be among the more critical patients, we infer that there is likely to be a difference in the arrival rates of patients classified by acuity.

Thus, we investigated alternative ratio controls designed to stabilize delays at target ratios. For that purpose, we propose the the *head-of-line-delay-ratio* (HLDR) *control*, aiming to keep the head-of-line delays at fixed ratios. The HLDR rule is appealing because it is a *blind* scheduling policy, i.e., it does not depend on any model parameters. We establish many-server heavy-traffic limits and conduct simulation experiments showing that the HLDR control is consistently effective at stabilizing the ratios of the delays experienced by the different classes at desired targets, while the FQR rule is not; see §2.

Our HLDR controls are generalizations of the dynamic-priority control proposed by Kleinrock (1964), which has recently been proposed for delay-based service differentiation in emergency departments based on acuity by Stanford et al. (2014) and Sharif et al. (2014); they call their proposed scheduling rule the accumulating-priority control, because they let the customers of different classes accumulate priority over the time they spend waiting in queue at different (constant) rates. Our analysis provides new results and insights for this accumulating-priority control.

In addition to the FQR and HLDR scheduling controls, we also consider a TV variation of QR (TVQR) designed to stabilize delay ratios at desired targets and a TV version of HLDR (TV-HLDR) designed to stabilize queue ratios at desired targets. We find that these too are effective, being also asymptotically correct in the MSHT limit, although the direct HLDR tends to outperform TVQR in simulation experiments. These alternative controls are not blind, because they also require knowledge of the arrival rates.

We find that our results can be explained to a large extent by a *sample-path (SP) MSHT Little's law (LL)* that is a consequence of the TV MSHT limit in Theorem 4.1, which is a generalization of the the SP-MSHT-LL for the stationary model that is a consequence of Theorem 4.3 in [13] and is discussed after equation (13) in §3 of [15]. In particular, the SP-MSHT-LL states, for large scale systems that are approximately

in the quality-and-efficiency-diven (QED) MSHT regime, that

$$Q_i(t) \approx \lambda_i(t)V_i(t) \quad \text{for all} \quad t, \tag{1.1}$$

where $Q_i(t)$ is the queue length, $\lambda_i(t)$ is the arrival rate and $V_i(t)$ is the potential delay at time t for class i.

If we consider ratios $QR(t) \equiv Q_1(t)/Q_2(t)$, $AR(t) \equiv \lambda_1(t)/\lambda_2(t)$ and $DR(t) \equiv V_1(t)/V_2(t)$, then as a consequence of (1.1) we have

$$QR(t) \approx AR(t) * DR(t) \quad \text{for all} \quad t. \tag{1.2}$$

Given (4.8), $QR(t)$ and $DR(t)$ can both be nearly constant over time only if $AR(t)$ is nearly constant over time. The new SP-MSHT-LL implies that it is impossible to stabilize queue ratios and delay ratios simultaneously with these ratio controls in the MSHT limit when the ratio of the asymptotic arrival rates is time-varying. Otherwise, all four ratio controls stabilize both queue ratios and delay ratios; e.g., see Figure 3 in §2.

1.1. *Related Literature.* There is a large literature on scheduling customers or jobs in an optimal or near optimal way. A classical textbook on scheduling is [7]; a more recent textbook with a strong scheduling focus is [17].

The problem of optimally scheduling a stationary many-server queueing system with several classes of impatient customers was extensively studied in [5, 18]; see also [4] and [1]. These papers assume class-dependent service and convert the original scheduling problem into a a more tractable diffusion control problem. In contrast, our goal is not to develop an optimal or asymptotically optimal scheduling policy, but rather to achieve desired delay-based differentiated service in the presence of multiple customer classes and time-varying arrivals. We do not study optimal control.

This paper is related to the literature on making delay announcements, because delay-history-based announcements have been considered. Armony et al. [3] considered announcing the delay of the last customer to enter service (LES) and analyzed the impact on system performance. Studies of announcements using the LES delay and the closely related head-of-line (HoL) delay as candidates to use in delay announcements were studied by [19, 20, 21, 22]. In contrast to our findings, for predicting the delay to be experienced of a new arrival in a stationary model, they showed that queue-length-based estimators tend to be more accurate than delay-history-based estimators provided that they are unbiased, because the queue-length estimators better reflect the state at the time of the new arrival, but delay-history-based estimators are robust to model assumptions. For TV arrival rates, Figure 2 of [21] shows that HoL delay predictors can have a significant bias due to the different state seen by the HoL customer upon arrival. In [22], fluid models are shown to be effective to improve the accuracy of these delay predictors.

The present work also relates to the literature on stabilizing performance of queueing systems with a time-varying arrival-rate function. It has been shown that in face of time-varying arrival rates, it is impossible to stabilize certain performance measures at the same time in heavy-traffic limits. For example, [27] showed that it is not possible to stabilize the abandonment probability and mean queue length at the same time in an $M_t/GI/s_t + GI$ many-server queueing model, while [39] showed for a $G_t/G_t/1$ queue that it is not possible to stabilize the queue-length and waiting time simultaneously; also see [28] for extensive simulation studies. These examples are single-class models. We take a step further by exposing the impact of TV arrival rates on stabilizing delay and queue ratios in multi-class models.

Our proofs draw on the martingale theory of weak convergence. An overview of martingale proofs of heavy-traffic limits of the time-stationary many-server queue with abandonment in critical loading can be found in [30]. Important precedents for TV MSHT limits in the QED regime are [29] and [32].

1.2. *Main Contributions.*

1. We conduct what we think is the first study of scheduling rules for assigning customers waiting in queue to newly available servers in the presence of TV arrival rates.
2. We show that, with TV arrival rates, the FQR control may fail badly in stabilizing delay ratios, even though it stabilizes queue ratios well. To provide a remedy, we propose a new HLDR control, which can be regarded as a generalization of the accumulating-priority rule rule studied by [34] and [33], which in turn is a variant of the dynamic-priority rule of [26]. Our HLDR control is more general because, first, we consider TV arrival rates, and associate with each class a TV control function rather than a single control parameter and, second, we go beyond the steady-state analysis and examine the time-varying behavior via many-server heavy-traffic analysis.
3. We establish the first MSHT limits for ratio controls for TV multi-class many-server queues. In particular, we we analyze the proposed HLDR rule for a TV multi-class queue with a single pool of exponential servers and multiple customer classes. With class-dependent service, we show that the queueing system can be uniquely characterized by a set of interacting diffusions in the MSHT limit. These MSHT limits show that the HLDR control achieves the desired delay ratios in every sample path.
4. We show that insight can be gained into the four candidate controls – FQR, HLDR, TVQR and TV-HLDR – by focusing on the SP TV MSHT Little's law. It shows that the queue ratios and delay ratios for two classes can both be stabilized together if and only if the ratio of the arrival rates for the classes is not TV.

1.3. *Organization.* In §2 we present results of initial simulation experiments to show the deficiencies of FQR and the advantages of HLDR with TV arrival rates. We define the model and the controls in §3. We state the main analytical results in §4 and provide the proof of our main theorem in §5. We present a short proof of the MSHT limits for the TVQR control in the appendix. We provide background on the simulation methodology and more numerical results in the supplement.

2. Initial Simulation Experiments. We illustrate the FQR and HLDR 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. Our analysis methods are more general, not being limited to two classes or sinusoidal arrival rate functions.

2.1. *The Experimental Setting.* Let the arrival processes for the two classes be independent nonhomogeneous Poisson processes (NHPP's) with arrival-rate functions

$$\lambda_i(t) = a_i + b_i \sin(d_i t) \quad \text{for} \quad 0 \le t \le T, \quad i = 1, 2. \tag{2.1}$$

Let the service times and patience times (before abandonment from queue) be mutually independent exponential random variables (and independent of the arrival processes), with constant class-dependent service rates μ_i and abandonment rates θ_i.

Let the time-staffing staffing level, the number $s(t)$ of servers working at time t, be based on the *square-root-safety* (SRS) *staffing rule*, which in turn is based on the *time-varying offered-load* $m(t)$, i.e., the time-varying mean number of busy servers in the associated infinite-server model, which is the sum of the offered loads $m_i(t)$ for the two classes: where

$$m_i(t) = \int_{-\infty}^{t} G_i^c(t-u)\lambda(u)\mathrm{d}u = \mathbb{E}\left[\int_{t-S_i}^{t} \lambda(u)\mathrm{d}u\right] = \mathbb{E}\left[\lambda(t-S_{i,e})\right]\mathbb{E}[S_i], \tag{2.2}$$

with S_i representing a generic class-i service-time random variable with cumulative distribution function (cdf) $G_i(t)$, $G_i^c(t) \equiv 1 - G_i(t) \equiv \mathbb{P}(S_i > t)$ and $S_{i,e}$ denotes a random variable with the associated stationary-excess cdf, defined by

$$G_{i,e}(t) \equiv \mathbb{P}(S_{i,e} \le t) \equiv \frac{1}{\mathbb{E}[S_i]} \int_0^t G_i^c(u)\mathrm{d}u, \quad \text{for} \quad t \ge 0;$$

see [9] . When S_i has an exponential distribution with $E[S_i] = 1/\mu_i$, as we have assumed, then m_i satisfies the ordinary differential equation

$$\dot{m}_i(t) = \lambda_i(t) - \mu_i m_i(t). \tag{2.3}$$

Given the offered-load $m(t) = m_1(t) + m_2(t)$, we can apply the SRS staffing formula

$$s(t) = \lceil m(t) + c\sqrt{m(t)} \rceil, \quad t \ge 0, \tag{2.4}$$

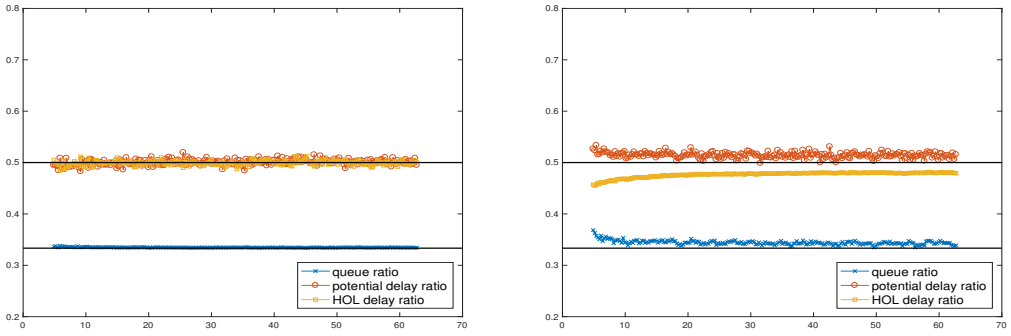
where c is the *quality-of-service* (QoS) parameter, which can be chosen to stabilize the delay probability at a desired target. The SRS staffing in (2.4) was supported first by a direct infinite-server (IS) approximation and then by associated modified-offered-load (MOL) and MSHT limits in [24]; see [12, 10, 41] for reviews and elaborations.

With time-varying staffing $s(t)$, we need to specify how we manage the system when all servers are busy when the staffing is scheduled to decrease. For greater reality, we may let the first server to complete their current service leave after that service is complete, which assumes that service switching is allowed when designated servers are scheduled to leave. In the model for our MSHT limits, we immediately push one server back into a high-priority queue and let that customer receive a new service, with rate depending on the class of that customer. We then show that the content of this high-priority queue is asymptotically negligible in the MSHT scaling, and thus does not affect the limit.

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 (2.1) (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 [13], 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 QoS coefficient $c = 0.25$, yielding the constant staffing level $s = 170$ to meet the constant offered load of 150. We obtain our simulation estimates by performing 2000 independent replications; see the 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 head-of-line (HoL) delay. In general (with abandonments), the potential delay at time t is the virtual delay, i.e., the delay that would be experienced by a hypothetical arrival at time t that is infinitely patient. Here it is measured by the actual delay experienced by arrivals. In contrast, the HoL delay at time t is the elapsed delay of the customer in queue that is next to enter service. Because the HoL customer will experience additional delay before entering service, we expect it to be somewhat less than the HoL potential delay. All estimates were obtained by averaging over 2000 independent replications. 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 [13].

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 (2.1), so that



(a) FQR (b) HLDR

Fig 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 the QoS parameter $c = 0.25$.

the overall arrival-rate function is

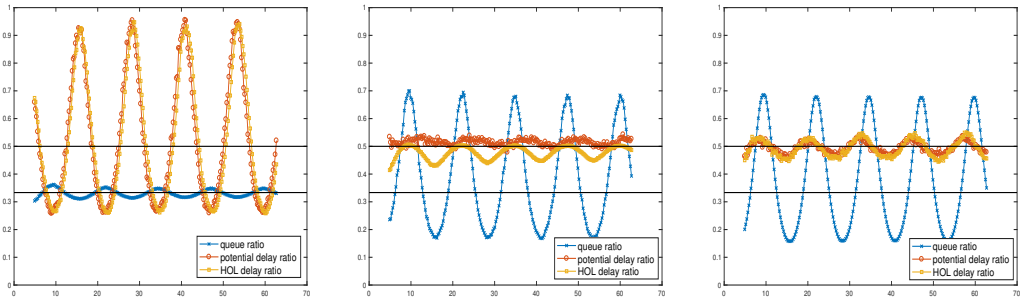
$$\lambda(t) = \lambda_1(t) + \lambda_2(t) = 150 + 10\sin(t/2).$$

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. In the context of a hospital ED, where a mean length of stay is about 4 hours, a cycle would be about 4 time longer, so that a day corresponds to about half a cycle. Thus, our parameter choice can provide insight for ED's.

Panels 2a and 2b of Figure 2 plot the same set of performance measures for FQR and HLDR shown in Figure 1. Panel 2a shows that FQR is again effective at stabilizing the queue lengths, but is now highly ineffective at indirectly stabilizing delays. Similarly, Panel 2b shows that HLDR is remarkably effective at directly stabilizing the ratio of the delays, but it does not indirectly stabilize the queue lengths. Panel 2c shows that the specially designed TV modification of FQR performs much like HLDR.

What we see in Figure 2 can be explained by (1.1) and (4.8): the ratio of the arrival rates $AR(t)$ 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, 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 customers abandonment. For simplicity, assume



(a) FQR (b) HLDR (c) TVQR

Fig 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 the QoS parameter $c = 0.25$.

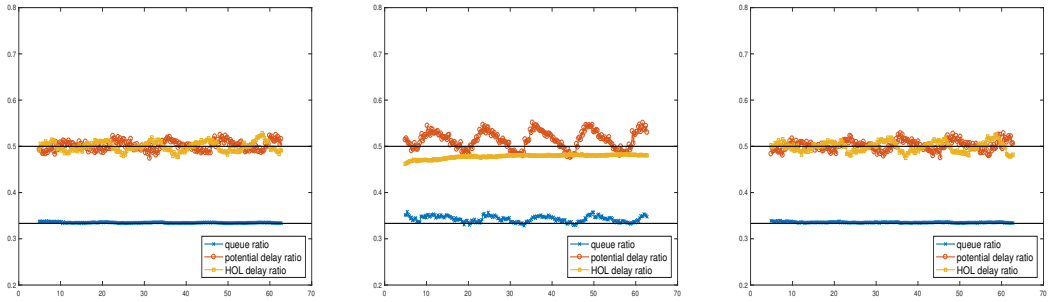
that impatience times are class-invariant following an exponential distribution with rate $\theta = 0.5$. This implies that the impatience time is two times longer than the service time on average. From our experiments, we see that abandonment affects our ability to stabilize the ratios, but that it 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 QoS parameter $c = 0$, which is consistent with the heuristic of "simply staffing to the offered load," as discussed in paragraph 3 of §6 of [10].

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 $\theta_1 = \theta_2 = 0.5$ and the QoS coefficient $c = 0$. Figure 4 shows that these scheduling controls become more effective as the scale increases, consistent with out later MSHT limit.

REMARK 2.1 (class-dependent service). The appendix shows the corresponding results for the two-class $M_t/M/s_t + M$ queue with class-dependent service times.

3. A Family of Time-Varying Multi-Class Queueing Models. We specify our notation and conventions in §3.1 and lay out the preliminaries of the time-varying multi-class queueing model in §3.2. We formalize the high-priority queue for customers pushed out of service because of staffing decrease in §3.3. We then define the HLDR and TVQR rules in §3.4 and §3.5, respectively.

3.1. *Notation and Conventions.* We denote by $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$, respectively, the sets of all real numbers, non-negative 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



(a) FQR (b) HLDR (c) TVQR

Fig 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 the QoS parameter $c = 0.25$.

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 lefthand 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 lefthand limit at t with the convention that $x(0-) = 0$ and $\Delta x(t) \equiv x(t) - x(t-)$. All stochastic processes are assumed to have trajectories from and are considered as 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 $\{\langle M\rangle(t); t \in \mathbb{R}_+\}$. We refer the reader to [23] for background in weak-convergence and martingale theory. All random entities introduced in this paper are supported by a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

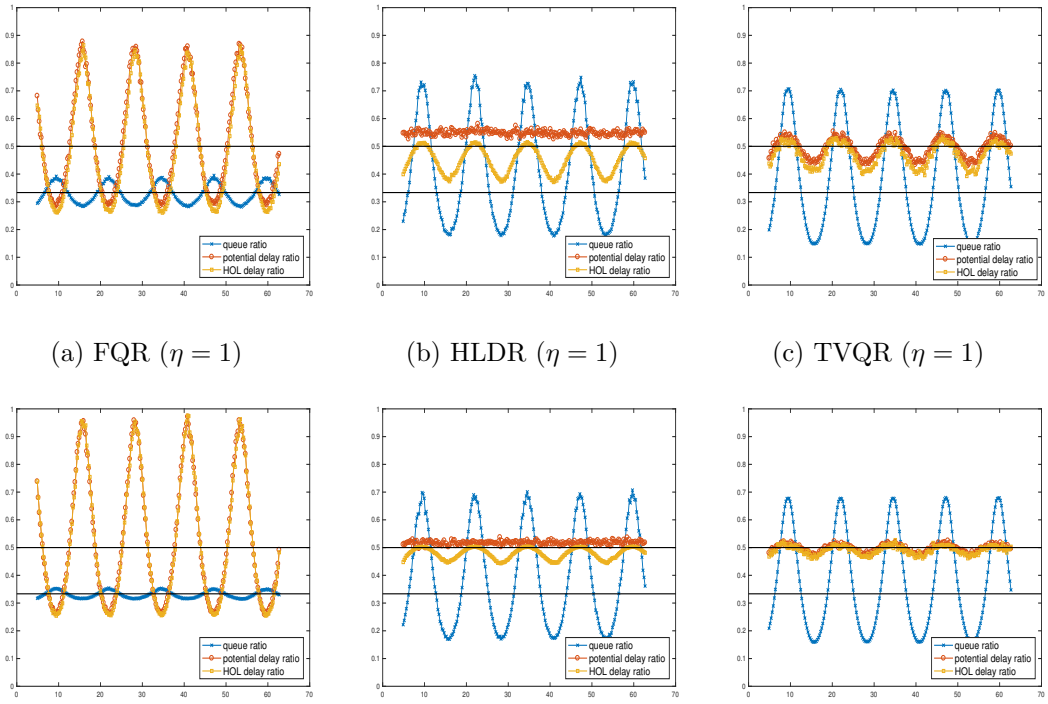
3.2. *Preliminaries.* There is a set $\mathcal{I} \equiv \{1, \ldots, K\}$ of customer classes. In the n-th model, the arrivals of class i follow a non-homogeneous Poisson process (NHPP) $A_i^n(t)$ with rate $n\lambda_i(t)$. These NHPPs are mutually independently. For $i \in \mathcal{I}$, let

$$\Lambda_i(t) \equiv \int_0^t \lambda_i(u)\mathrm{d}u, \qquad \widehat{A}_i^n(t) \equiv n^{-1/2}\left(A_i^n(t) - n\Lambda_i(t)\right) \tag{3.1}$$

The sequence of processes $\{\widehat{A}_i^n\}$ satisfies a functional central limit theorem (FCLT); i.e.,

$$\widehat{A}_i^n(\cdot) \Rightarrow W_i \circ \Lambda_i(\cdot) \equiv A_i^{(d)}(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{3.2}$$

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 a NHPP



(a) FQR ($\eta = 1$) (b) HLDR ($\eta = 1$) (c) TVQR ($\eta = 1$)

(d) FQR ($\eta = 8$) (e) HLDR ($\eta = 8$) (f) TVQR ($\eta = 8$)

Fig 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 $\lambda_1(t) = \eta\cdot(60-20\sin(t/2))$, $\lambda_2 = \eta\cdot(90+30\sin(t/2))$, service rate $\mu = 1$, abandonment rates $\theta_1 = \theta_2 = 0.5$ and the QoS coefficient $c = 0.0$: the cases $\eta = 1$ and $\eta = 8$.

satisfying a FCLT; that is

$$\widehat{A}^n(\cdot) \equiv n^{-1/2}\left(A^n - n\int_0^{\cdot}\lambda_\Sigma(u)\mathrm{d}u\right) \Rightarrow W\circ\Lambda(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty$$

where $\lambda_\Sigma(t) \equiv \sum_{i\in\mathcal{I}}\lambda_i(t)$ and $\Lambda(t) \equiv \int_0^t \lambda_\Sigma(u)\mathrm{d}u$.

As in §2, the service times and patience times are mutually independent 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 3.1. (*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 (2.2) in [39] and assume a FCLT for the base process; see §7.3 of [30].

We allow the staffing level (number of servers) to be time-varying as well, but we go beyond the the TV SRS staffing function in (2.4) as a function of n. First, the offered load as a function of n is $m^n(t) = nm(t)$, where $m(t) \equiv \sum_{i\in\mathcal{I}} m_i(t)$ and m_i satisfies the ODE for the offered load in (2.3). Now we let

$$s^n(t) = n\, m(t) + n^{1/2}c(t), \quad t \geq 0, \tag{3.3}$$

where $c(t)$ is a general function to be specified. Thus, we obtain the n-dependent version of (2.4) if we let $c(t) = c\sqrt{m(t)}$, with $m(t) \equiv \sum_{i\in\mathcal{I}} m_i(t)$ and m_i satisfying the ODE for the offered load in (2.3), but we allow other possibilities.

Equation (2.3) stipulates that the inflow and outflow have to be matched on the fluid scale. This is in line with numerous studies of stochastic processing networks which follow a hierarchy where one first considers a static planning problem given demand information and then invokes the standard Brownian motion machinery (second order refinement). From a technical point of view, such special type of growth behavior for $s^n(\cdot)$ forces the system to reside in the Quality-and-Efficiency-Driven (QED) MSHT limiting regime. The hypothesis (3.3) follows the early papers [16, 11, 29]. That scaling also is used in Theorem 5.1 in the electronic companion of [10], Theorem 2 in [32] and Section 2.6 in [41].

If the staffing is scheduled to decrease when the servers are all busy, we immediately enforce that staffing change, so that we need to force a customer out of service. In the single-class case it is possible to let one customer to return to the head of the queue, as in [32]. 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 HPQ have the highest service priority; i.e., 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 $Q^n_{0,i}(t)$ 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.

We assume a work-conserving policy, i.e., no customers wait in queue if there are servers available. 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

$$\begin{aligned} Q^n_i(t) &= Q^n_i(0) + A^n_i(t) - \Psi^n_i(t) - R^n_i(t) \\ &= Q^n_i(0) + \Pi^a_i\left(n\Lambda_i(t)\right) - \Psi_i(t) - \Pi^{ab}_i\left(\theta_i \int_0^t Q^n_i(u)\mathrm{d}u\right), \end{aligned} \tag{3.4}$$

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

$$\begin{aligned} Q_{0,i}^n(t) + B_i^n(t) &= Q_{0,i}^n(0) + B_i^n(0) + \Psi_i^n(t) - D_i^n(t) \\ &= B_i^n(0) + \Psi_i^n(t) - \Pi_i^d\left(\mu_i \int_0^t B_i^n(u)\mathrm{d}u\right), \end{aligned} \tag{3.5}$$

where Π_i^d are unit-rate Poisson processes independent of Π_i^a and Π_i^{ab} given in (3.4). Let $X_i^n(t)$ denote the total number of class-i customers in system at time t. Adding up (3.4) and (3.5) yields

$$X_i^n(t) = Q_i^n(t) + Q_{0,i}^n(t) + B_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - R_i^n(t). \tag{3.6}$$

Alternatively, one can derive (3.6) directly from flow conservation.

Finally, let $Q_0^n(t) \equiv \sum_{i\in\mathcal{I}} Q_{0,i}^n(t)$, $Q^n(t) \equiv \sum_{i\in\mathcal{I}} Q_i^n(t)$ and $X^n(t) \equiv \sum_{i\in\mathcal{I}} X_i^n(t)$ be the total number of high- and low- priority customers in queue(s) and the aggregate number of customers in system respectively. Adding up (3.6) over $i \in \mathcal{I}$ yields

$$X^n(t) = Q^n(t) + Q_0^n(t) + B^n(t) = X^n(0) + A^n(t) - D^n(t) - \sum_{i\in\mathcal{I}} R_i^n(t) \tag{3.7}$$

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

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

$$A_0^n(t) \equiv \sum_{u\in\mathcal{S}_a^n(t)} 1(B^n(u-) = s^n(u-)). \tag{3.8}$$

Let $D_0^n(t)$ 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

$$D_0^n(t) \equiv \sum_{u\in\mathcal{S}_d^n(t)} 1(Q_0^n(u-) > 0) + \int_0^t 1(Q_0^n(u-) > 0)\mathrm{d}D^n(u). \tag{3.9}$$

From (3.8) and (3.9), it follows that

$$\begin{aligned} Q_0^n(t) &= A_0^n(t) - D_0^n(t) \\ &= \sum_{u\in\mathcal{S}_a^n(t)} 1(B^n(u-) = s^n(u-)) - \sum_{u\in\mathcal{S}_d^n(t)} 1(Q_0^n(u-) > 0) \\ &\quad - \int_0^t 1(Q_0^n(u-) > 0)\mathrm{d}D^n(u). \end{aligned} \tag{3.10}$$

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

$$Z^n(t) \equiv s^n(0) - s^n(t) - D^n(t), \quad t \geq 0. \tag{3.11}$$

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

$$\Upsilon_0^n(t) \equiv \psi(Z^n)(t) \equiv Z^n(t) - \inf_{0 \leq u \leq t} \{Z^n(u)\}; \tag{3.12}$$

e.g., see §13.5 in [37]. The following lemma shows that Υ_0^n serves as an upper bound for Q_0^n.

LEMMA 3.1. *Let Q_0^n and Υ_0^n be as given in (3.10) and (3.12) respectively. Then*

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

Proof of Lemma 3.1. By (3.12) and (3.11), it is not hard to see that

$$\Upsilon_0^n(t) = \sum_{u \in \mathcal{S}_a^n(t)} 1 - \sum_{u \in \mathcal{S}_d^n(t)} 1(\Upsilon_0^n(u-) > 0) - \int_0^t 1(\Upsilon_0^n(u-) > 0) \mathrm{d}D^n(u). \tag{3.13}$$

Combining (3.10) and (3.13) 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 §5.2 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. *The HLDR Control.* We now formalize the HLDR scheduling rule that uniquely determines the assignment processes $\Psi_i(\cdot)$. Let $w_i^n(t)$ be the head-of-line (HoL) delay of customer i. Then the HoL customer in queue i arrived at time $T_i^n(t) \equiv t - w_i^n(t)$. Now introduce a set of weight/control functions $v(\cdot) \equiv (v_1(\cdot), \ldots, v_K(\cdot))$ and define a weighted HoL delay

$$\tilde{w}_i^n(t) \equiv w_i^n(t)/v_i(t) \quad \text{for each} \quad i \in \mathcal{I}. \tag{3.14}$$

In addition, use $\tilde{w}^n(t)$ to represent the maximum of those weighted HoL delays, i.e.,

$$\begin{aligned} \tilde{w}^n(t) &\equiv \max_{i \in \mathcal{I}} \{\tilde{w}_1^n(t), \ldots, \tilde{w}_K^n(t)\} \\ &= \max_{i \in \mathcal{I}} \{w_1^n(t)/v_i(t), \ldots, w_K^n(t)/v_K(t)\} \quad \text{for} \quad t \geq 0. \end{aligned} \tag{3.15}$$

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

$$\tau(t) \equiv \{i \in \mathcal{I} : \tilde{w}_i^n(t) = \tilde{w}^n(t)\}\,. \tag{3.16}$$

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

$$\Psi_i^n(t) = \sum_{u \in \mathcal{T}^n(t)} 1(\tau(u) = i), \tag{3.17}$$

where

$$\begin{aligned} \mathcal{T}^n(t) \equiv \{u \in [0,t] : \Delta A^n(u) = 1, B^n(u-) < s^n(u-)\} \\ \cup \{u \in [0,t] : \Delta D^n(u) = 1, Q^n(u-) > 0\} \end{aligned} \tag{3.18}$$

is the collection of time instances at which an assignment decision is to be made and $\tau(\cdot)$ is given by (3.16). Here ties are broken arbitrarily. For instance, if $\tilde{w}_i^n(t) = \tilde{w}_{i'}^n(t) = \tilde{w}^n(t)$ for $i \neq i'$, then the next-available server chooses to serve either queue i or queue i' with equal probabilities.

REMARK 3.2 (reduction to AP). The HLDR rule reduces to the accumulating-priority (AP) rule if all $v_i(t) = v_i$; i.e., if the weight functions $v_i(\cdot)$ are constant functions. In that case, we can think that waiting customers accumulate priority ar a constant rate while in the queue, with customer from class i accumulating priority at a rate $1/v_i$. When a server becomes free, HDLR selects the waiting customer with the highest accumulating priority for service.

REMARK 3.3 (reduction to the global FCFS). If $v_i(t) = 1$ for all $i \in \mathcal{I}$ and t; i.e., all classes accumulate priority at an equal constant rate, then the HLDR reduces to *global first-come-first-serve (FCFS)*, as in [35].

3.5. *The TVQR Control.* As indicated earlier, our HLDR control is intimately related to the more general QR controls studied in [13]. We briefly review the FQR control, which is a special case of the more general QR control introduced by [13], in the context of multi-class queue with a single pool of 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 $v \equiv (v_1, \ldots, v_K)$. Upon service completion, the available server admits to service the customer from the head of the queue i^* where

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

i.e., the next-available-server always chooses to serve the queue with the greatest queue imbalance.

Here instead of using fixed ratios we introduce a time-varying vector function $v(\cdot) \equiv (v_1(\cdot), \ldots, v_K(\cdot))$ and the next-available-server choose serve a class i customer where

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

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

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

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

$$V_i^n(t) \equiv \inf\{s \geq 0 : \Psi_i^n(t+s) + R_i^n(t+s) \geq Q_i^n(0) + A_i^n(t)\}, \tag{3.19}$$

but the representation (3.19) is *incorrect*, because the term $R_i^n(t+s)$ may include class-i customers that arrived after time t and then abandoned; see §1 in [36].

To formally define the potential delay of class i at some time $t \geq 0$, we exclude the abandonment of customers who arrived after time t; see §4 of [36]. Following the notation of that paper, we add another superscript t to the abandonment process, which indicates that only the abandonment of customers arriving before time t are included. Then the potential delay in queue i at time t can be represented as the following first-passage time

$$V_i^n(t) \equiv \inf\{s \geq 0 : \Psi_i^n(t+s) + R_i^{n,t}(t+s) > Q_i^n(0) + A_i^n(t)\}. \tag{3.20}$$

Finally, note that the potential delay and the HoL delay have a simple relation:

$$w_i^n(t_{i,k}^n) = V_i^n\left(t_{i,k}^n - w_i^n(t_{i,k}^n)\right), \tag{3.21}$$

where $\{t_{i,k}^n; k \in \mathbb{N}\}$ are time instances at which a customer enters service from queue i.

4. Main Results. In §4.1 we state our main result and then discuss important insights that it provides in §4.2. We establish corollaries for important special cases in §4.3. In §4.4 we establish the associated result for the TVQR rule and in §4.5 we discuss the asymptotic equivalence. Finally, in §4.6 we observe that the results in [13] themselves can be extended to a large class of TV arrival-rate functions.

4.1. *The MSHT FCLT for HLDR in the QED Regime.* We first introduce the diffusion-scaled processes

$$\widehat{X}_i^n(\cdot) \equiv n^{-1/2}\left(X_i^n(\cdot) - n \cdot m_i(\cdot)\right) \quad \text{and} \quad \widehat{X}^n(\cdot) \equiv n^{-1/2}\left(X^n(\cdot) - n \cdot m(\cdot)\right), \tag{4.1}$$

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

$$\widehat{Q}_i^n(\cdot) \equiv n^{-1/2} Q_i^n(\cdot) \quad \text{and} \quad \widehat{Q}_{0,i}^n(\cdot) \equiv n^{-1/2} Q_{0,i}^n(\cdot) \tag{4.2}$$

be the diffusion-scaled queue-length processes and $\widehat{Q}^n \equiv n^{-1/2}Q^n$ and $\widehat{Q}^n_0 \equiv n^{-1/2}Q^n_0$ be the aggregate quantities. The same scaling was used by [10, 32, 41]. As usual, we scale the delay processes by multiplying by $\sqrt{n}$ instead of dividing by $\sqrt{n}$ as in (4.2):

$$\widehat{V}^n_i(t) \equiv n^{1/2}V^n_i(t) \quad \text{and} \quad \widehat{w}^n_i(t) \equiv n^{1/2}w^n_i(t) \quad \text{for} \quad i \in \mathcal{I}. \tag{4.3}$$

We impose the following regularity conditions:

(A1) For each $i \in \mathcal{I}$, the arrival-rate function $\lambda_i(\cdot)$ is differentiable with bounded first derivative; i.e., 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 function $\lambda(\cdot)$ is bounded away from zero; i.e., there exists $\lambda_* > 0$ such that $\lambda(t) \geq \lambda_*$ for all t.

(A2) The safety-staffing function $c(\cdot)$ is continuous and hence integrable.

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

Our main results establishes a MSHT FCLT for HLDR in the QED regime. The limit process is a diffusion process.

THEOREM 4.1 (QED MSHT FCLT for HLDR). *Suppose that the system is staffed according to (3.3), operates under the HLDR scheduling rule and Conditions A1 - A3 hold. If, in addition, there is convergence of the initial distribution at time* 0*, i.e., if*

$$(\widehat{X}^n_1(0), \ldots, \widehat{X}^n_K(0), \widehat{Q}^n_1(0), \ldots, \widehat{Q}^n_K(0)) \Rightarrow (X^{(d)}_1(0), \ldots, X^{(d)}_K(0), Q^{(d)}_1(0), \ldots, Q^{(d)}_K(0))$$

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

$$\begin{aligned} &\left(\widehat{X}^n_1(\cdot), \ldots, \widehat{X}^n_K(\cdot), \widehat{Q}^n_1(\cdot), \ldots, \widehat{Q}^n_K(\cdot), \widehat{V}^n_1(\cdot), \ldots, \widehat{V}^n_K(\cdot), \widehat{w}^n_1(\cdot), \ldots, \widehat{w}^n_K(\cdot)\right) \\ &\Rightarrow \left(X^{(d)}_1(\cdot), \ldots, X^{(d)}_K(\cdot), Q^{(d)}_1(\cdot), \ldots, Q^{(d)}_K(\cdot), V^{(d)}_1(\cdot), \ldots, V^{(d)}_K(\cdot), w^{(d)}_1(\cdot), \ldots, w^{(d)}_K(\cdot)\right) \end{aligned} \tag{4.4}$$

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

$$\begin{aligned} X^{(d)}_i(t) = X^{(d)}_i(0) - \mu_i \int_0^t X^{(d)}_i(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t \gamma(u)^{-1} v_i(u)\lambda_i(u) \\ \times \left[X^{(d)}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}\mathrm{d}W_i(u) \end{aligned} \tag{4.5}$$

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

$$\begin{aligned} Q^{(d)}_i(\cdot) &\equiv \gamma(\cdot)^{-1} v_i(\cdot)\lambda_i(\cdot)\left[X^{(d)}(\cdot) - c(\cdot)\right]^+ \\ V^{(d)}_i(\cdot) &= w^{(d)}_i(\cdot) \equiv v_i(\cdot)\cdot\gamma(\cdot)^{-1}\left[X^{(d)}(\cdot) - c(\cdot)\right]^+. \end{aligned} \tag{4.6}$$

4.2. *Important Insights.* We can draw several important insights from Theorem 4.1.

4.2.1. *the role of the SRS safety functions $c(\cdot)$.* Given that the staffing is done by (3.3), 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 $m_i(\cdot)$ depend on the specified $\lambda_i(\cdot)$ and μ_i for $i \in \mathcal{I}$. (The functions $\lambda_i(\cdot)$ and $m_i(\cdot)$ are scaled up by n in the limit.) The remaining component of the staffing in (3.3) 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 $c(\cdot)$, which remains to be specified. The limiting performance impact of the staffing function $c(\cdot)$ can be seen directly in the limit.

4.2.2. *state-space collapse.* While the stochastic limit process $(X_1^{(d)}(\cdot), \ldots, X_K^{(d)}(\cdot))$ for the K-dimensional scaled number-in-system process $(\widehat{X}_1^n(\cdot), \ldots, \widehat{X}_K^n(\cdot))$ 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 $X^{(d)}(\cdot) \equiv X_1^{(d)}(\cdot) + \cdots + X_K^{(d)}(\cdot)$, in particular of $\left[X^{(d)}(u) - c(u)\right]^+$, so that there is great state-space collapse. In particular, the limit processes $Q_i^{(d)}(\cdot)$, $V_i^{(d)}(\cdot)$ and $w_i^{(d)}(\cdot)$ are deterministic functionals of each other, as shown by (4.6). While the potential and HoL delays are not the same, their limits are the same.

4.2.3. *the sample-path MSHT Little's law.* We obtain the SP MSHT LL directly from the conclusion of Theorem 4.1. In particular, for each i, we see that, almost surely,

$$Q_i^{(d)}(t) = \lambda_i(t) V_i^{(d)}(t) \quad \text{for all} \quad t \geq 0. \tag{4.7}$$

For the n-th system, we have

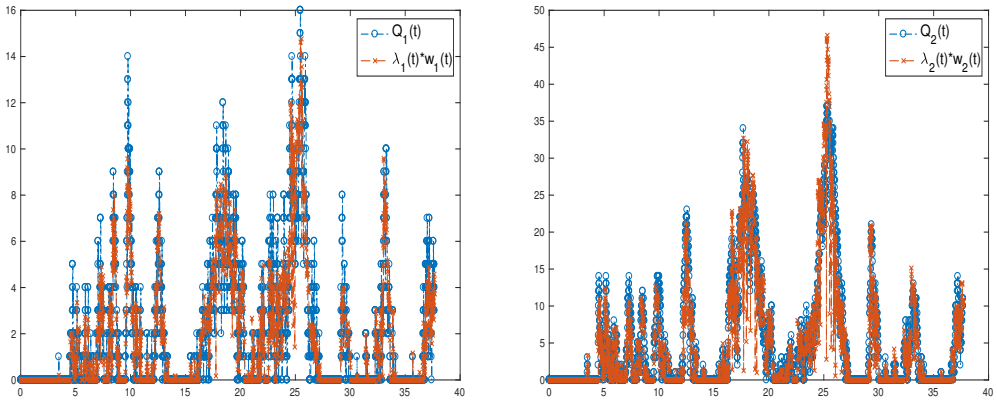
$$\widehat{Q}_i^n(t) = \lambda_i(t) \widehat{V}_i^n(t) + o(1) \quad \text{as} \quad n \to \infty \tag{4.8}$$

or

$$Q_i^n(t) = n\lambda_i(t) V_i^n(t) + o(\sqrt{n}) \quad \text{as} \quad n \to \infty. \tag{4.9}$$

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

Figure 5 depicts the individual sample paths of $Q_i(\cdot)$ and $\lambda_i(\cdot)w_i(\cdot)$ on the same plot for $i = 1, 2$ with the HLDR policy for the base case. Panel (a) and Panel (b) show 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.



(a) Class 1 (b) Class 2

Fig 5: Sample paths of the queue-length process $Q_i(\cdot)$ and the scaled delay process $v_i(\cdot)w_i(\cdot)$ for $i = 1, 2$ with the HLDR scheduling policy.

4.2.4. *the limiting ratios are deterministic.* While the limit in Theorem 4.1 is of course random, the limits for all the ratios of interest are deterministic. That is, as $n \to \infty$, the ratios converge in probability to deterministic limits uniformly in t over bounded intervals:

$$\begin{aligned} V_i^n(t)/V_j^n(t) &\Rightarrow v_i(t)/v_j(t), \\ w_i^n(t)/w_j^n(t) &\Rightarrow v_i(t)/v_j(t) \quad \text{and} \\ Q_i^n(t)/Q_j^n(t) &\Rightarrow v_i(t)\lambda_i(t)/v_j(t)\lambda_j(t). \end{aligned} \tag{4.10}$$

4.2.5. *the role of customer abandonment.* While customer abandonment does influence the queue-length and waiting-time limit processes of interest through the one-dimensional limit process $X^{(d)}(u)$, customer abandonment plays no roles in determining these limiting ratios in (4.10). 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 [38] for the basic $M/M/s + M$ Erlang-A model.

4.2.6. *impact of the arrival-rate and the weight functions.* Given the limit for the queue ratios in (4.10), 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 4.1 applies to the stationary model as an important special case.

COROLLARY 4.1 (the stationary case). *Let* $\lambda_i(t) = \lambda_i, v_i(t) = v_i$ *and* $c(t) = c$ *for* $t \geq 0$. *If, in addition,*

$$(\widehat{X}_1^n(0), \ldots, \widehat{X}_K^n(0), \widehat{Q}_1^n(0), \ldots, \widehat{Q}_K^n(0)) \Rightarrow (X_1^{(d)}(0), \ldots, X_K^{(d)}(0), Q_1^{(d)}(0), \ldots, Q_K^{(d)}(0))$$

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

$$\left(\widehat{X}_1^n(\cdot), \ldots, \widehat{X}_K^n(\cdot), \widehat{Q}_1^n(\cdot), \ldots, \widehat{Q}_K^n(\cdot), \widehat{V}_1^n(\cdot), \ldots, \widehat{V}_K^n(\cdot), \widehat{w}_1^n(\cdot), \ldots, \widehat{w}_K^n(\cdot)\right)$$
$$\Rightarrow \left(X_1^{(d)}(\cdot), \ldots, X_K^{(d)}(\cdot), Q_1^{(d)}(\cdot), \ldots, Q_K^{(d)}(\cdot), V_1^{(d)}(\cdot), \ldots, V_K^{(d)}(\cdot), w_1^{(d)}(\cdot), \ldots, w_K^{(d)}(\cdot)\right)$$

in $\mathcal{D}^{4K}$ *where the diffusion limits* $X_i^{(d)}$ *satisfy*

$$X_i^{(d)}(t) = X_i^{(d)}(0) - \mu_i \int_0^t X_i^{(d)}(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t \gamma^{-1} v_i \lambda_i \left[X^{(d)}(u) - c\right]^+ \mathrm{d}u + \sqrt{2\lambda_i} W_i(t).$$

in which $\gamma = \sum_{i \in \mathcal{I}} v_i \lambda_i$ *and* $X^{(d)} \equiv \sum_{i \in \mathcal{I}} X_i^{(d)}$; *for each* $i \in \mathcal{I}$

(4.11)
$$Q_i^{(d)}(\cdot) \equiv v_i \lambda_i \gamma^{-1} \left[X^{(d)}(\cdot) - c\right]^+ \quad \text{and} \quad V_i^{(d)}(\cdot) = w_i^{(d)}(\cdot) \equiv v_i \cdot \gamma^{-1} \left[X^{(d)}(\cdot) - c\right]^+.$$

Corollary 4.1 is in agreement with Theorem 4.3 in [13] 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 §4.5 that an asymptotic equivalence exists not only for time-stationary models but also in time-varying settings. Theorem 4.3 in [13] has $[\hat{X}]^+$ and $[\hat{X}]^-$ in the equation (6) whereas (4.5) in the present paper uses $[X^{(d)} - c]^+$ and $[X^{(d)} - c]^-$. The discrepancies are due to different centering component being used. In [13] the number of customers in system is centered by the number of servers whereas we use $nm(t)$ to be the centering term.

REMARK 4.1 (consistent with previous AP results). The result in (4.11) is in alignment with previous work on AP by [26] and [34], 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.

For $K = 1$, Theorem 4.1 reduces to Theorem 2.5 of [41], which in turn draws upon [32].

COROLLARY 4.2 (the single-class case). *Suppose that the conditions in Theorem 4.1 are satisfied and $K = 1$. Let $\lambda_1 = \lambda, \theta_1 = \theta$ and $\mu_1 = 1$. Then*

$$\widehat{X}^n(\cdot) \Rightarrow X^{(d)}(\cdot) \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad n \to \infty$$

where

$$\begin{aligned} X^{(d)}(t) = X^{(d)}(0) &- \int_0^t \left(X^{(d)}(u) \wedge c(u)\right) \mathrm{d}u \\ &- \theta \int_0^t \left[X^{(d)}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda(u) + m(u)} \mathrm{d}W(u). \end{aligned} \tag{4.12}$$

For the special case $\theta = \mu$, it is well known that the resulting $M_t/M/s_t + M$ model is equivalent to an $M_t/M/\infty$ model. Let $\theta = \mu = 1$ in (4.12). Then it holds that

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

Here the diffusion limit $X^{(d)}$ is an Ornstein-Uhlenbeck (OU) process with time-varying volatility.

4.4. *The MSHT FCLT for TVQR in the QED Regime.* We now turn to the TVQR control as described by §3.5. Mimicking the analysis of [13], one can establish the MSHT limits, regarding the TVQR rule, via hydrodynamic limits. However, the proof in [13] is quite involved and in turn relies on additional general state space collapse (SSC) results from [8]. 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 the appendix, adopts a similar stopping-time argument as used by [6] in the analysis of an inverted-V system under the Longest-Idle-Pool-First routing rule.

THEOREM 4.2 (Diffusion Limit with the TVQR Rule). *Suppose that the system operates under the TVQR rule. If, in addition,*

$$(\widehat{X}_1^n(0), \ldots, \widehat{X}_K^n(0), \widehat{Q}_1^n(0), \ldots, \widehat{Q}_K^n(0)) \Rightarrow (X_1^{(d)}(0), \ldots, X_K^{(d)}(0), Q_1^{(d)}(0), \ldots, Q_K^{(d)}(0))$$

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

(4.13)

$$\begin{aligned} &\left(\widehat{X}_1^n(\cdot), \ldots, \widehat{X}_K^n(\cdot), \widehat{Q}_1^n(\cdot), \ldots, \widehat{Q}_K^n(\cdot), \widehat{V}_1^n(\cdot), \ldots, \widehat{V}_K^n(\cdot), \widehat{w}_1^n(\cdot), \ldots, \widehat{w}_K^n(\cdot)\right) \\ &\Rightarrow \left(X_1^{(d)}(\cdot), \ldots, X_K^{(d)}(\cdot), Q_1^{(d)}(\cdot), \ldots, Q_K^{(d)}(\cdot), V_1^{(d)}(\cdot), \ldots, V_K^{(d)}(\cdot), w_1^{(d)}(\cdot), \ldots, w_K^{(d)}(\cdot)\right) \end{aligned}$$

in $\mathcal{D}^{4K}$ *where the diffusion limits* $X_i^{(d)}(\cdot)$ *satisfy*

(4.14)
$$X_i^{(d)}(t) = X_i^{(d)}(0) - \mu_i \int_0^t X_i^{(d)}(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t r_i(u) \left[X^{(d)}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}\mathrm{d}W_i(u)$$

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

(4.15)
$$Q_i^{(d)}(\cdot) \equiv r_i(\cdot)\left[X^{(d)}(\cdot) - c(\cdot)\right]^+, \quad \textit{and} \quad V_i^{(d)}(\cdot) = w_i^{(d)}(\cdot) \equiv \frac{r_i(\cdot)}{\lambda_i(\cdot)} \cdot \left[X^{(d)}(\cdot) - c(\cdot)\right]^+.$$

We gain several insights from the theorem above: (i) with the TVQR, the desired queue-ratio is achieved in the limit despite the fact that arrival rates are changing; (ii) from (4.15) 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

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

One can easily verify that the stochastic equation (4.5) becomes the equation (4.14).

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

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

Direct calculation allows us to translate equation (4.14) into (4.5).

4.6. *Extending the QIR Limits to TV Arrivals.* Even though [13] 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 [13] 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. Since 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. While the remaining argument may be complex, there should be little doubt that the extension holds.

5. Proof of Theorem 4.1. For any $x \in \mathcal{D}$, let $x[t_1,t_2] \equiv (t_2-) - x(t_1-)$. With the HLDR control, the queue-length processes satisfy

$$\begin{aligned} Q_i^n(t-) &= A_i^n[T_i(t),t] - R_i^{n,T_i(t)}[T_i(t),t] \\ &= A_i^n[t - v_i(t)\tilde{w}^n(t),t] - R_i^{n,t-v_i(t)\tilde{w}^n(t)}[t - v_i(t)\tilde{w}^n(t),t]. \end{aligned} \tag{5.1}$$

The first equality trivially holds due to the definition of $T_i(t)$. For the second equality, note that $t-v_i(t)\tilde{w}^n(t) \le t-w_i^n(t) = T_t(t)$. If $t-v_i(t)\tilde{w}^n(t) = T_i(t)$, the second equality trivially holds. If $t - v_i(t)\tilde{w}^n(t) \neq T_i(t)$, it suffices to argue that all customers that arrive over the period $\big(t-v_i(t)\tilde{w}^n(t), T_i(t)\big)$ abandon the queue before time t. Suppose for the sake of contradiction that a class-i customer arrived during this period but never abandoned before time t. Then it must be the case that the customer hasn't yet been assigned at time t and hence is still in the queue at time t. This contradicts the definition of $T_i(t)$ because a customer who arrived before time $T_i(t)$ can no longer be in the queue.

Let

$$\widehat{R}_i^n(\cdot) \equiv n^{-1/2}R_i^n(\cdot) \quad \text{and} \quad \widehat{R}_i^{n,t}(t+\cdot) \equiv n^{-1/2}R_i^{n,t}(t+\cdot). \tag{5.2}$$

It follows from (3.2), (4.2), (5.1) and (5.2) that

$$\begin{aligned} \widehat{Q}_i^n(t-) &= \widehat{A}_i^n[v_i(t)\tilde{w}^n(t)),t] + n^{1/2}\int_{t-v_i(t)\tilde{w}^n(t)}^t \lambda_i(u)\mathrm{d}u \\ &\qquad - \widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t - v_i(t)\tilde{w}^n(t),t] \\ &= \widehat{A}_i^n[v_i(t)\tilde{w}^n(t)),t] + n^{1/2}v_i(t)\lambda_i(t)\tilde{w}^n(t) \\ &\qquad - \widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t - v_i(t)\tilde{w}^n(t),t] + e_i^n(t) \end{aligned} \tag{5.3}$$

with

$$e_i^n(t) \equiv n^{1/2}\int_{t-v_i(t)\tilde{w}^n(t)}^t \lambda_i(u)\mathrm{d}u - n^{1/2}v_i(t)\lambda_i(t)\tilde{w}^n(t). \tag{5.4}$$

Adding up (5.3) over $i \in \mathcal{I}$ and rearranging, we obtain

$$n^{1/2}\gamma(t)\tilde{w}^n(t) = \widehat{Q}^n(t-) - \sum_{i\in\mathcal{I}} \widehat{A}_i^n[t - v_i(t)\tilde{w}^n(t), t] + \sum_{i\in\mathcal{I}} \widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t - v_i(t)\tilde{w}^n(t), t] - \sum_{i\in\mathcal{I}} e_i^n(t).$$

Inserting the above the expression into (5.3) yields

$$\widehat{Q}_i^n(t-) = \gamma(t)^{-1}v_i(t)\lambda_i(t)\widehat{Q}^n(t-) + \widehat{A}_i^n[t - v_i(t)\tilde{w}^n(t), t] - \widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t - v_i(t)\tilde{w}^n(t), t] + e_i^n(t) - \gamma(t)^{-1}K^n(t), \tag{5.5}$$

with

$$K^n(t) \equiv \sum_{i\in\mathcal{I}} \widehat{A}_i^n[t - v_i(t)\tilde{w}^n(t), t] - \sum_{i\in\mathcal{I}} \widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t - v_i(t)\tilde{w}^n(t), t] + \sum_{i\in\mathcal{I}} e_i^n(t).$$

We will show that, with the exception of term $\gamma(t)^{-1}v_i(t)\lambda_i(t)\widehat{Q}^n$, all components on the right hand side vanish as n grows to infinity.

We lay out the path ahead. (i) We start off by showing that both $\{\widehat{X}_i^n(\cdot); n \in \mathbb{N}\}$ and $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ are stochastically bounded. We then argue that the sequence of HoL delay processes $\{n^{1/2}\tilde{w}^n(\cdot); n \in \mathbb{N}\}$ are stochastically bounded, which shows that $\tilde{w}^n(\cdot)$ (defined by (3.15)) lives on the order of $O(n^{-1/2})$. (ii) The result of (i) allows us to prove that the queue-length processes are asymptotically proportional to the weights; i.e.,

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

This is essentially a state-space-collapse (SSC) result in the many-server diffusion limit. (iii) By a similar argument as in [13] (first SSC and then diffusion limits), we obtain the diffusion limits for $\widehat{X}_i^n(\cdot)$. (iv) The limits for the queue-length processes and delay processes follow immediately.

5.1. *Stochastic Boundedness of* $\{\widehat{X}_i^n(\cdot); n \in \mathbb{N}\}$ *and* $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$. Here we exploit a martingale decomposition, as in [30] and [32]. Specifically the processes

$$\begin{aligned}\widehat{D}_i^n(t) &\equiv n^{-1/2}\left[D_i^n(t) - \mu_i\int_0^t B_i^n(u)\mathrm{d}u\right] \\ &= n^{-1/2}\left[\Pi_i^d\left(\mu_i\int_0^t B^n(u)\mathrm{d}u\right) - \mu_i\int_0^t B_i^n(u)\mathrm{d}u\right]\end{aligned} \tag{5.6}$$

and

$$\begin{aligned}\widehat{Y}_i^n(t) &\equiv n^{-1/2}\left[R_i^n(t) - \theta_i \int_0^t Q_i^n(u)\mathrm{d}u\right] \\ &= n^{-1/2}\left[\Pi_i^{ab}\left(\theta_i \int_0^t Q_i^n(u)\mathrm{d}u\right) - \theta_i \int_0^t Q_i^n(u)\mathrm{d}u\right]\end{aligned} \tag{5.7}$$

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

$$\langle \widehat{D}_i^n\rangle(t) = \frac{\mu_i}{n}\int_0^t B_i^n(u)\mathrm{d}u \quad \text{and} \quad \langle \widehat{Y}_i^n\rangle(t) = \frac{\theta_i}{n}\int_0^t Q_i^n(u)\mathrm{d}u. \tag{5.8}$$

Both $\{\widehat{D}_i^n(\cdot); n \in \mathbb{N}\}$ and $\{\widehat{Y}_i^n(\cdot); n \in \mathbb{N}\}$ are stochastic bounded due to Lemma 5.8 of [30], which is based on the Lenglart-Rebolledo inequality, stated as Lemma 5.7 there.

Now use (3.7) to write

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - R_i^n(t). \tag{5.9}$$

From (2.3), it follows

$$m_i(t) = m_i(0) + \int_0^t \lambda_i(u)\mathrm{d}u - \mu_i \int_0^t m_i(u)\mathrm{d}u. \tag{5.10}$$

Scaling both sides of (5.10) by n and subtracting it from (5.9) gives us

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

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

$$\begin{aligned}\widehat{X}_i^n(t) = \widehat{X}_i^n(0) - \mu_i \int_0^t \widehat{X}_i^n(u)\mathrm{d}u - (\theta_i - \mu_i)\int_0^t \widehat{Q}_i^n(u)\mathrm{d}u \\ + \mu_i \int_0^t \widehat{Q}_{0,i}^n(u)\mathrm{d}u + \widehat{A}_i^n(t) - \widehat{D}_i^n(t) - \widehat{Y}_i^n(t).\end{aligned} \tag{5.11}$$

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

$$\mathcal{M}_i^n(t) \equiv \widehat{A}_i^n(t) - \widehat{D}_i^n(t) - \widehat{Y}_i^n(t). \tag{5.12}$$

Clearly $\{\mathcal{M}_i^n; n \in \mathbb{N}\}$ is stochastically bounded for $i \in \mathcal{I}$. From (5.11) - (5.12), it follows

$$\left|\widehat{X}_i^n(t)\right| \le \left|\widehat{X}_i^n(0)\right| + \bar{a}\int_0^t \left|\widehat{X}_i^n(u)\right|\mathrm{d}u + \bar{a}\int_0^t \left(\widehat{Q}_i^n(u) + \widehat{Q}_{0,i}^n(u)\right)\mathrm{d}u + \left|\mathcal{M}_i^n(t)\right|. \tag{5.13}$$

Adding up (5.13) over $i \in \mathcal{I}$, we obtain

$$\begin{aligned}\sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n(t)\right| \le \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n(0)\right| + \bar{a}\int_0^t \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n(u)\right| \mathrm{d}u \\ + \bar{a}\int_0^t \left(\widehat{Q}^n(u) + \widehat{Q}_0^n(u)\right) \mathrm{d}u + \sum_{i\in\mathcal{I}} |\mathcal{M}_i^n(t)|\,.\end{aligned} \tag{5.14}$$

Also note that

$$\widehat{Q}^n(\cdot) + \widehat{Q}_0^n(\cdot) = n^{-1/2}\left[X^n(\cdot) - s^n(\cdot)\right]^+ = \left[\widehat{X}^n(\cdot) - c(\cdot)\right]^+ \le \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n\right| + \left|c(\cdot)\right| \tag{5.15}$$

where the last inequality is owing to the basic inequality $\left|\widehat{X}^n\right| \le \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n\right|$. Plugging (5.15) into (5.14) yields

$$\sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n(t)\right| \le \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n(0)\right| + \bar{a}\int_0^t |c(u)|\,\mathrm{d}u + 2\bar{a}\int_0^t \sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n(u)\right| \mathrm{d}u + \sum_{i\in\mathcal{I}} |\mathcal{M}_i^n(t)|\,. \tag{5.16}$$

An application of the Gronwall's inequality with (5.16) establishes the stochastic boundedness of $\left\{\sum_{i\in\mathcal{I}}\left|\widehat{X}_i^n\right|; n \in \mathbb{N}\right\}$. Thus for $i \in \mathcal{I}$ the sequence $\{\widehat{X}_i^n(\cdot); n \in \mathbb{N}\}$ is stochastically bounded. Then the stochastic boundedness of $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ and $\{\widehat{Q}_0^n(\cdot); n \in \mathbb{N}\}$ follows easily by (5.15).

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 [30]. Indeed, by (4.1) and (4.2), we must have

$$\overline{X}_i^n(\cdot) \equiv \frac{X_i^n(\cdot)}{n} \Rightarrow m_i(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{5.17}$$

and

$$\overline{B}_i^n(\cdot) \equiv \frac{B_i^n(\cdot)}{n} = \frac{X_i^n(\cdot) - Q_i^n(\cdot) - Q_{0,i}^n(\cdot)}{n} - \Rightarrow m_i(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{5.18}$$

Applying the continuous mapping theorem (CMT) with integration in (5.18), we have

$$\overline{D}_i^n(\cdot) \equiv \mu_i \int_0^{\cdot} \overline{B}_i^n(u)\mathrm{d}u \Rightarrow \mu_i \int_0^{\cdot} m_i(u)\mathrm{d}u \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{5.19}$$

Then apply the CMT with composition in (5.19) to obtain

$$\begin{aligned}\widehat{D}_i^n(\cdot) &= n^{-1/2}\left[\Pi_i^d\left(n\mu_i \int_0^{\cdot} \overline{B}_i^n(u)\mathrm{d}u\right) - n\mu_i \int_0^{\cdot} \overline{B}_i^n(u)\mathrm{d}u\right] \\ &= n^{-1/2}\left(\Pi_i^d \circ n\overline{D}_i^n(\cdot) - n\overline{D}_i^n(\cdot)\right) \Rightarrow W_i\left(\mu_i \int_0^{\cdot} m_i(u)\mathrm{d}u\right) \quad \text{in} \quad \mathcal{D}\end{aligned} \tag{5.20}$$

as $n \to \infty$ where we have used W_i to denote a standard Brownian motion. It is a simple exercise to show via (5.20) that

(5.21)
$$\widehat{D}^n(\cdot) \equiv n^{-1/2}\left[D^n(\cdot) - n\sum_{i\in\mathcal{I}}\mu_i\int_0^{\cdot}\overline{B}_i^n(u)\mathrm{d}u\right] \Rightarrow W\left(\sum_{i\in\mathcal{I}}\mu_i\int_0^{\cdot}m_i(u)\right) \quad \text{in} \quad \mathcal{D}$$

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

5.2. *Asymptotic Negligibility of* $\{\widehat{Q}_0^n(\cdot); n \in \mathbb{N}\}$. The argument required here is a variant of Theorem 13.5.2 (b) in [37], 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 $\{\widehat{Q}_0^n(\cdot); n \in \mathbb{N}\}$, we first argue that $\widehat{\Upsilon}_0^n(\cdot) \equiv n^{-1/2}\Upsilon_0^n(\cdot)$ vanishes as $n \to \infty$. For that purpose, define $\widehat{Z}^n(\cdot) \equiv n^{-1/2}Z^n(\cdot)$. By (3.10),

$$\widehat{\Upsilon}_0^n(t) = \widehat{Z}^n(t) - \sup_{u\le t}\left\{-\widehat{Z}^n(u)\right\}. \tag{5.22}$$

Combining (3.3), (3.11), (5.10) and (5.21) and some algebraic manipulation leads easily to

$$\widehat{Z}^n(t) = -n^{1/2}\int_0^t \lambda(u)\mathrm{d}u - \mathcal{X}^n(t) \tag{5.23}$$

where

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

In view of condition A2, the C-tightness of $\widehat{D}^n$, and the stochastic boundedness of $\widehat{X}_i^n(u)$, $\widehat{Q}_i^n$ and $\widehat{Q}_{0,i}^n$, we deduce that the sequence of $\{\mathcal{X}^n(\cdot); n \in \mathbb{N}\}$ is stochastically bounded and C-tight. Define

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

From (5.22) - (5.23), it follows

$$\widehat{\Upsilon}_0^n(t) = -n^{1/2}\int_{u^n(t)}^t \lambda(u)\mathrm{d}u - \mathcal{X}^n(t) + \mathcal{X}^n(u^n(t)) \ge 0 \tag{5.24}$$

Combining the inequality in (5.24) and the stochastic boundedness of $\mathcal{X}^n(\cdot)$ allows us to conclude

$$\sup_{t\le T}\{t - u^n(t)\} = O_p(n^{-1/2}). \tag{5.25}$$

For a cadlag (right continuous with left limits) function $x(\cdot)$, define $|x|_T^* \equiv \sup_{t\le T}|x(t)|$. Using (5.24), we can easily deduce

$$\mathbb{P}\left(\left|\widehat{\Upsilon}_0^n\right|_T^* > \epsilon\right) \le \mathbb{P}\left(\sup_{t\le T}\{-\mathcal{X}^n(t) + \mathcal{X}^n(u^n(t))\} \ge \epsilon\right).$$

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

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

Since ϵ is arbitrarily chosen, we have proven

$$\widehat{\Upsilon}_0^n(\cdot) \equiv n^{-1/2}\Upsilon_0^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n\to\infty. \tag{5.26}$$

It is immediate by Lemma 3.1 and the definition of $\widehat{Q}_0^n$ and $\widehat{\Upsilon}_0^n$ that $\widehat{Q}_0^n(t) \le \widehat{\Upsilon}_0^n(t)$ for all $t \le T$. Hence, we must have

$$\left(\widehat{Q}_0^n(\cdot), \widehat{Q}_{0,1}^n(\cdot), \ldots, \widehat{Q}_{0,K}^n(\cdot)\right) \Rightarrow 0 \quad \text{in} \quad \mathcal{D}^{K+1} \quad \text{as} \quad n\to\infty. \tag{5.27}$$

5.3. *State Space Collapse.* By Condition A3,

$$\int_{t-v_*\tilde{w}^n(t)}^t \lambda_i(u)\mathrm{d}u \le \int_{t-v_i(t)\tilde{w}^n(t)}^t \lambda_i(u)\mathrm{d}u. \tag{5.28}$$

Inserting (5.28) into (5.3) yields

$$n^{1/2}\int_{t-v_*\tilde{w}^n(t)}^t \lambda_i(u)\mathrm{d}u \le \widehat{Q}_i^n(t-) - \widehat{A}_i^n[v_i(t)\tilde{w}^n(t)),t] + \widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t-v_i(t)\tilde{w}^n(t),t]. \tag{5.29}$$

Adding up (5.29) over $i\in\mathcal{I}$, we obtain

$$n^{1/2}\int_{t-v_*\tilde{w}^n(t)}^t \lambda(u)\mathrm{d}u \le \widehat{Q}^n(t-) - \sum_{i\in\mathcal{I}}\widehat{A}_i^n[v_i(t)\tilde{w}^n(t)),t] + \sum_{i\in\mathcal{I}}\widehat{R}_i^{n,t-v_i(t)\tilde{w}^n(t)}[t-v_i(t)\tilde{w}^n(t),t].$$

Note that the right hand side is stochastically bounded, due to the stochastic boundedness of $\widehat{Q}^n$, $\widehat{A}_i^n$ and $\widehat{R}_i^n$. In view of Condition A1, we have shown that $\{n^{1/2}\tilde{w}^n(\cdot); n\in\mathbb{N}\}$ is stochastically bounded.

We now argue that $\widehat{A}_i^n[t-v_i\tilde{w}^n(t),t]$, $\widehat{R}_i^n[t-v_i\tilde{w}^n(t),t]$ and $e_i^n(t)$ vanish uniformly over $[0,T]$ as $n\to\infty$. That $\widehat{A}_i^n[t-v_i\tilde{w}^n(t),t]$ converge to zero uniformly over $[0,T]$ is straightforward since $\widehat{A}_i^n(\cdot)$ converges weakly to a Brownian motion (with a time shift)

and the maximum time increment $|\tilde{w}^n|_T^*$ converges to zero in $\mathbb{R}$ as $n \to \infty$. To see that $\widehat{R}_i^n[t - v_i\tilde{w}^n(t), t]$ vanishes as n grows to infinity, note that the quadratic variation

$$\langle \widehat{Y}_i^n \rangle(\cdot) = \frac{\theta_i}{n} \int_0^{\cdot} Q_i^n(u)\mathrm{d}u \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{5.30}$$

drawing upon Section 7.1 of [30]. The convergence in (5.30) implies

$$\widehat{R}_i^n(\cdot) - \theta_i \int_0^{\cdot} \widehat{Q}_i^n(u)\mathrm{d}u \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{5.31}$$

by applying the Lenglart-Rebolledo inequality; see p. 30 of [25]. In view of

$$\int_{t-v_i\tilde{w}^n(t)}^t \widehat{Q}_i^n(u)\mathrm{d}u \le v^* \left|\widehat{Q}^n\right|_T^* |\tilde{w}^n|_T^*$$

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

Next consider the term e_i^n given in (5.4). By Taylor expansion

$$\begin{aligned}
|e_i^n(t)| &\equiv \left| n^{1/2} \int_{t-v_i(t)\tilde{w}^n(t)}^t \lambda_i(u)\mathrm{d}u - n^{1/2} v_i(t)\lambda_i(t)\tilde{w}^n(t) \right| \\
&= \Big| n^{1/2} v_i(t)\tilde{w}^n(t)\lambda_i(t) + n^{1/2}\big(v_i(t)\tilde{w}^n(t)\big)^2 \lambda_i'(t) \\
&\qquad + o_p\Big(\big(v_i(t)\tilde{w}^n(t)\big)^2\Big) - n^{1/2} v_i(t)\lambda_i(t)\tilde{w}^n(t) \Big| \\
&= \left| n^{1/2}\big(v_i(t)\tilde{w}^n(t)\big)^2 \lambda_i'(t) + o_p\Big(\big(v_i(t)\tilde{w}^n(t)\big)^2\Big) \right| \\
&= O_p\left(n^{1/2}(|\tilde{w}^n|_T^*)^2\right)
\end{aligned} \tag{5.32}$$

where the last equality is due to Conditions A1 and A3 which guarantee the boundedness of $|\lambda_i'(\cdot)|$ and $v_i(\cdot)$ over any compact intervals.

The random variable $n^{1/2}(|\tilde{w}^n|_T^*)^2$ is independent of time t and converges to zero as $n \to \infty$ because $n^{1/2}|\tilde{w}^n|_T^*$ is stochastically bounded and $|\tilde{w}^n|_T^*$ goes to zero as n approaches infinity.

Now we have shown that $\widehat{A}_i^n[t - v_i(t)\tilde{w}^n(t), t]$, $\widehat{R}_i^n[t - v_i(t)\tilde{w}^n(t), t]$ and $e_i^n(t)$ vanish uniformly over $[0, T]$ as n grows to infinity. In view of (5.5), we conclude

$$\Theta_i^n(\cdot) \equiv \widehat{Q}_i^n(\cdot) - \gamma(\cdot)^{-1} v_i(\cdot)\lambda_i(\cdot)\widehat{Q}^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \quad \text{for} \quad i \in \mathcal{I}, \tag{5.33}$$

By the convergence-together lemma,

$$(\Theta_1^n(\cdot), \ldots, \Theta_K^n(\cdot)) \Rightarrow (0, \ldots, 0) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n \to \infty. \tag{5.34}$$

5.4. *Diffusion Limits.* Using the CMT with integration in (5.33), we obtain

(5.35)
$$\Upsilon_i^n(\cdot) \equiv \int_0^{\cdot} \widehat{Q}_i^n(u)\mathrm{d}u - \int_0^{\cdot} \gamma(u)^{-1} v_i(u)\lambda_i(u)\widehat{Q}^n(u)\mathrm{d}u \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.$$

Note that

$$\widehat{Q}^n(\cdot) + \widehat{Q}_0^n(\cdot) = \left[\widehat{X}^n(\cdot) - c(\cdot)\right]^+. \tag{5.36}$$

Combining (5.11), (5.35) and (5.36) gives

(5.37)
$$\begin{aligned}\widehat{X}_i^n(t) = \widehat{X}_i^n(0) - \mu_i \int_0^t \widehat{X}_i^n(u)\mathrm{d}u - (\theta_i - \mu_i)\Upsilon_i^n(t) + \mu_i \int_0^t \widehat{Q}_{0,i}^n(u)\mathrm{d}u + \widehat{A}_i^n(t) - \widehat{D}_i^n(t) \\ - \widehat{Y}_i^n(t) - (\theta_i - \mu_i)\int_0^t \gamma(u)^{-1} v_i(u)\lambda_i(u)\left(\left[\widehat{X}^n(u) - c(u)\right]^+ - \widehat{Q}_0^n(u)\right)\mathrm{d}u.\end{aligned}$$

An application of Theorem 4.1 of [30] together with (3.2), (5.20), (5.27), (5.30) and (5.35) allows us to establish the many-server heavy-traffic limit for $\{\widehat{X}_i^n(\cdot); n \in \mathbb{N}\}$:

$$\left(\widehat{X}_1^n(\cdot), \ldots, \widehat{X}_K^n(\cdot)\right) \Rightarrow \left(X_1^{(d)}(\cdot), \ldots, X_K^{(d)}(\cdot)\right) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n \to \infty,$$

where $X_i^{(d)}$ satisfies the differential equation (4.5). Then apply the convergence-together lemma with (5.34) we conclude

$$\begin{aligned}&\left(\widehat{X}_1^n(\cdot), \ldots, \widehat{X}_K^n(\cdot), \widehat{Q}_1^n(\cdot), \ldots, \widehat{Q}_K^n(\cdot)\right) \\ &\quad \Rightarrow \left(X_1^{(d)}(\cdot), \ldots, X_K^{(d)}(\cdot), Q_1^{(d)}(\cdot), \ldots, Q_K^{(d)}(\cdot)\right) \quad \text{in} \quad \mathcal{D}^{2K}\end{aligned} \tag{5.38}$$

as $n \to \infty$ where the limiting processes $Q_i^{(d)}$ are given in (4.6).

5.5. *Potential Delay Asymptotics.* To establish heavy-traffic stochastic-process limits for potential delays, we follow the solution approach as in Section 3 of [36]. 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 (3.20) have proper fluid and diffusion limits. For each $i \in \mathcal{I}$, introduce the fluid-scaled processes

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

Clearly we have

$$\left(\overline{A}_i^n(\cdot), \overline{\Psi}_i^n(\cdot), \overline{R}_i^n(\cdot), \overline{Q}_i^n(\cdot)\right) \Rightarrow (\Lambda_i(\cdot), \Lambda_i(\cdot), 0, 0) \quad \text{in} \quad \mathcal{D}^4 \quad \text{as} \quad n \to \infty. \tag{5.39}$$

Now define

$$\widehat{\Psi}_i^n(\cdot) \equiv n^{-1/2}\left(\Psi_i^n(\cdot) - n\Lambda_i(\cdot)\right) \tag{5.40}$$

Then

$$\left(\widehat{A}_i^n(\cdot), \widehat{\Psi}_i^n(\cdot), \widehat{R}_i^n(\cdot), \widehat{Q}_i^n(\cdot)\right) \Rightarrow \left(A_i^{(d)}(\cdot), \Psi_i^{(d)}(\cdot), R_i^{(d)}(\cdot), Q_i^{(d)}(\cdot)\right) \quad \text{in} \quad \mathcal{D}^4 \tag{5.41}$$

as $n \to \infty$ where $\widehat{A}_i^n, \widehat{\Psi}_i^n, \widehat{Q}_i^n$ and $\widehat{R}_i^n$ are given in (3.2), (5.40), (4.2) and (5.2) respectively, and

$$R_i^{(d)}(\cdot) \equiv \theta_i \int_0^{\cdot} Q_i^d(u)\mathrm{d}u, \qquad \Psi_i^{(d)}(\cdot) \equiv Q_i^d(0) + A_i^{(d)}(\cdot) - Q_i^{(d)}(\cdot) - R_i^{(d)}(\cdot)$$

where $A_i^{(d)}$ and $Q_i^{(d)}$ are given in (3.2) and (4.2) respectively.

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

$$V_i^{n,l}(t) \le V_i^n(t) \le V_i^{n,u}(t) \tag{5.42}$$

where

$$\begin{aligned} V_i^{n,l}(t) &\equiv \inf\{s \ge 0 : \Psi_i^n(t+s) + R_i^n(t+s) \ge Q_i^n(0) + A_i^n(t)\} \\ &= \inf\{s \ge 0 : \overline{\Psi}_i^n(t+s) + \overline{R}_i^n(t+s) \ge \overline{Q}_i^n(0) + \overline{A}_i^n(t)\} \end{aligned} \tag{5.43}$$

and

$$\begin{aligned} V_i^{n,u}(t) &\equiv \inf\{s \ge 0 : \Psi_i^n(t+s) + R_i^n(t) \ge Q_i^n(0) + A_i^n(t)\} \\ &= \inf\{s \ge 0 : \overline{\Psi}_i^n(t+s) \ge \overline{Q}_i^n(0) + \overline{A}_i^n(t) - \overline{R}_i^n(t)\} \end{aligned} \tag{5.44}$$

For all $n \ge 1$, define the first-passage-time processes $\overline{U}_i^{n,l} \equiv (\overline{U}_i^{n,l}(t), t \ge 0)$ and $\overline{U}_i^{n,u} \equiv (\overline{U}_i^{n,u}(t), t \ge 0)$ where

$$\overline{U}_i^{n,l}(t) \equiv \inf\{s \ge 0 : \overline{\Psi}_i^n(s) + \overline{R}_i^n(s) \ge \overline{Q}_i^n(0) + \overline{A}_i^n(t)\} \tag{5.45}$$

$$\overline{U}_i^{n,u}(t) \equiv \inf\{s \ge 0 : \overline{\Psi}_i^n(s) \ge \overline{Q}_i^n(0) + \overline{A}_i^n(t) - \overline{R}_i^n(t)\} \tag{5.46}$$

One may attempt to apply the corollary of [31] together with (5.39), (5.41) to get

$$n^{1/2}V_i^{n,l} = n^{1/2}(\overline{U}_i^{n,l} - e)^+ \Rightarrow \frac{Q_i^{(d)}}{\Lambda_i'} \quad \text{and} \quad n^{1/2}V_i^{n,u} = n^{1/2}(\overline{U}_i^{n,l} - e)^+ \Rightarrow \frac{Q_i^{(d)}}{\Lambda_i'}$$

in $\mathcal{D}$ as $n \to \infty$, and then use (5.42) - (5.44) to conclude the desired results. However the right-hand side of the condition in (5.44) does not satisfy the conditions of the corollary. In particular, $\overline{Q}_i^n(0) + \overline{A}_i^n - \overline{R}_i^n$ is not necessarily nondecreasing. To resolve the problem, we use the same linear-interpolation technique as illustrated in Fig. 1 of that paper. 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

$$n^{1/2}\tilde{V}_i^{n,u} \Rightarrow \frac{Q_i^{(d)}}{\Lambda_i'}. \tag{5.47}$$

A standard sandwiching argument allows us to conclude

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

jointly with (5.38).

Condition (5.47) holds if the error caused by these linear interpolations is asymptotically negligible. The proof of Lemma 7.1 in [36] 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

$$\begin{aligned} &\left(\widehat{X}_1^n(\cdot), \ldots, \widehat{X}_K^n(\cdot), \widehat{Q}_1^n(\cdot), \ldots, \widehat{Q}_K^n(\cdot), \widehat{V}_1^n(\cdot), \ldots, \widehat{V}_K^n(\cdot) \right) \\ &\Rightarrow \left(X_1^{(d)}(\cdot), \ldots, X_K^{(d)}(\cdot), Q_1^{(d)}(\cdot), \ldots, Q_K^{(d)}(\cdot), V_1^{(d)}(\cdot), \ldots, V_K^{(d)}(\cdot) \right) \quad \text{in} \quad \mathcal{D}^{3K} \end{aligned} \tag{5.48}$$

as $n \to \infty$.

5.6. *HoL Delay Asymptotics.* Scaling both sides of (3.21) by $n^{1/2}$, we have

$$\widehat{w}_i^n(t_{i,k}^n) = \widehat{V}_i^n \left(t_{i,k}^n - w_i^n(t_{i,k}^n) \right)$$

for all $t_{i,k}^n$. Note that the set $\{t_{i,k}^n\}$ becomes dense in $[0,T]$ as $n \to \infty$. From the proof of Theorem 4.1, it follows that the process $w_i^n(\cdot)$ converges to zero uniformly over any compact interval. We thus obtain, by taking $n \to \infty$, that

$$\widehat{w}_i^n(\cdot) - \widehat{V}_i^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{5.49}$$

The convergence in (5.49) can be strengthened to joint convergence by the convergence-together lemma, that is

$$\left(\widehat{w}_1^n(\cdot) - \widehat{V}_1^n(\cdot), \ldots, \widehat{w}_K^n(\cdot) - \widehat{V}_K^n(\cdot) \right) \Rightarrow (0, \ldots, 0) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n \to \infty.$$

Then use the convergence-together lemma to conclude that

$$(\widehat{w}_1^n(\cdot), \ldots, \widehat{w}_K^n(\cdot)) \Rightarrow \left(w_1^{(d)}(\cdot), \ldots, w_K^{(d)}(\cdot) \right) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n \to \infty$$

jointly with (5.48) where $w_i^{(d)}(\cdot)$ is given by (4.6).

Acknowledgment. This research was supported by NSF grant CMMI 1634133.

REFERENCES

[1] Arapostathis, A., Biswas, A. and Pang, G. (2015). Ergodic control of multi-class $M/M/N+M$ queues in the Halfin–Whitt regime. *The Annals of Applied Probability* **25** 3511–3570.

[2] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y. and Yom-Tov, G. (2015). Patient Flow in Hospitals: a Data-based Queueing-science Perspective. *Stochastic Systems* **5** 146–194.

[3] Armony, M., Shimkin, N. and Whitt, W. (2009). The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57** 66–81.

[4] Atar, R., Giat, C. and Shimkin, N. (2010). The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* **58** 1427–1439.

[5] Atar, R., Mandelbaum, A., Reiman, M. I. et al. (2004). Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **14** 1084–1134.

[6] Atar, R., Shaki, Y. Y. and Shwartz, A. (2011). A blind policy for equalizing cumulative idleness. *Queueing Systems* **67** 275–293.

[7] Conway, R. W., Maxwell, W. L. and Miller, L. W. (2003). *Theory of scheduling*. Courier Corporation.

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

[9] Eick, S. G., Massey, W. A. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research* **41** 731–742.

[10] Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54** 324–338.

[11] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4** 208–227.

[12] Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management* **16** 13–39.

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

[14] Gurvich, I. and Whitt, W. (2009). Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems. *Manufacturing and Service Operations Management* **11** 237–253.

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

[16] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations research* **29** 567–588.

[17] Harchol-Balter, M. (2013). *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press.

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

[19] Ibrahim, R. and Whitt, W. (2009). Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11** 397–415.

[20] Ibrahim, R. and Whitt, W. (2009). Real-Time Delay Estimation in Overloaded Multiserver Queues with Abandonment. *Management Science* **55** 1729–1742.

[21] Ibrahim, R. and Whitt, W. (2011). Real-Time Delay Estimation Based on Delay History in Many-Server Service Systems with Time-Varying Arrivals. *Production and Operations Management* **20** 654–667.

[22] Ibrahim, R. and Whitt, W. (2011). Wait-Time Predictors for Customer Service Systems with Time-Varying Demand and Capacity. *Operations Research* **59** 1106–1118.

[23] Jacod, J. and Shiryaev, A. N. (2013). *Limit theorems for stochastic processes* **288**. Springer Science & Business Media.

[24] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science* **42** 1383–1394.

[25] Karatzas, I. and Shreve, S. (2012). *Brownian motion and stochastic calculus* **113**. Springer Science & Business Media.

[26] Kleinrock, L. (1964). A delay dependent queue discipline. *Naval Research Logistics (NRL)* **11** 329–341.

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

[28] Ma, N. and Whitt, W. (2015). Using simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate. In *Proceedings of the 2015 Winter Simulation Conference* 2598–2609. IEEE Press.

[29] Mandelbaum, A., Massey, W. A. and Reiman, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* **30** 149–201.

[30] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 7.

[31] Puhalskii, A. (1994). On the invariance principle for the first passage time. *Mathematics of Operations Research* **19** 946–954.

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

[33] Sharif, A. B., Stanford, D. A., Taylor, P. and Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care* **3** 73–79.

[34] Stanford, D. A., Taylor, P. and Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems* **77** 297–330.

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

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

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

[38] Whitt, W. (2006). Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operations Research* **54** 247–260.

[39] Whitt, W. (2015). Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* **81** 341–378.

[40] Whitt, W. and Zhang, X. (2017). A data-driven model of an emergency department. *Operations Research for Health Care* **12** 1–15.

[41] Whitt, W. and Zhao, J. (2017). Staffing to stabilizing blocking in loss models with non-Markovian arrivals. *Naval Research Logistics*.

APPENDIX

to

Delay-Based Service Differentiation in a Many-Server Queue with Time-Varying Arrival Rates

by

Xu Sun and Ward Whitt

APPENDIX A: A SHORT PROOF OF THEOREM 4.2

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 4.2, 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; i.e., $Q_1/Q_2 \geq r_1/r_2$. 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 $\eta_1 \leq \eta_2$. Fix $k \in \mathcal{I}$ and let H denote any event under which

(i) no server has ever been idle over the period $[\eta_1, \eta_2]$;
(ii) 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 (3.7), one can easily derive

(A.1)
$$\sum_{i\in\mathcal{I}} A_i^n[\eta_1,\eta_2] - D^n[\eta_1,\eta_2] - \sum_{i\in\mathcal{I}} R_i^n[\eta_1,\eta_2] = X^n[\eta_1,\eta_2] = s^n[\eta_1,\eta_2] + \sum_{i\in\mathcal{I}} Q_i^n[\eta_1,\eta_2]$$

where the second equality follows from the non-idling condition (i). Moreover, by condition (ii) we have

$$A_k^n[\eta_1,\eta_2] = Q_k^n[\eta_1,\eta_2] \tag{A.2}$$

because there is no departure from the k-th queue. Combining (A.1) and (A.1) yields

(A.3)
$$\sum_{i\neq k} A_i^n[\eta_1,\eta_2] - D^n[\eta_1,\eta_2] - \sum_{i\in\mathcal{I}} R_i^n[\eta_1,\eta_2] = X^n[\eta_1,\eta_2] = s^n[\eta_1,\eta_2] + \sum_{i\neq k} Q_i^n[\eta_1,\eta_2].$$

From (5.10) it follows that

$$\int_{\eta_1}^{\eta_2} \lambda_\Sigma(u)\mathrm{d}u - \int_{\eta_1}^{\eta_2} m(u)\mathrm{d}u = m(\eta_2) - m(\eta_1) \tag{A.4}$$

Scaling both sides of (A.4) by n and subtracting it from (A.3) gives

$$\sum_{i\neq k}\left(A_i^n[\eta_1,\eta_2]-n\int_{\eta_1}^{\eta_2}\lambda_i(u)\mathrm{d}u\right)-\left(D^n[\eta_1,\eta_2]-n\int_{\eta_1}^{\eta_2}m(u)\mathrm{d}u\right)-\sum_{i\in\mathcal{I}}R_i^n[\eta_1,\eta_2]$$
$$=s^n[\eta_1,\eta_2]-n\cdot m[\eta_1,\eta_2]+\sum_{i\neq k}Q_i^n[\eta_1,\eta_2]+n\int_{\eta_1}^{\eta_2}\lambda_k(u)\mathrm{d}u$$

which in turn yields (after scaling both sides by $n^{-1/2}$ and rearranging terms)

$$\begin{aligned}&\sum_{i\neq k}\widehat{A}_i^n[\eta_1,\eta_2]-\widehat{D}^n[\eta_1,\eta_2]-\int_{\eta_1}^{\eta_2}c(u)\mathrm{d}u-c[\eta_1,\eta_2]-\sum_{i\neq k}\widehat{Q}_i^n[\eta_1,\eta_2]\\&=\sum_{i\in\mathcal{I}}\widehat{R}_i^n[\eta_1,\eta_2]+n^{1/2}\int_{\eta_1}^{\eta_2}\lambda_k(u)\mathrm{d}u.\end{aligned}\tag{A.5}$$

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

$$\Delta_i^n(\cdot)\equiv\widehat{Q}_i^n(\cdot)-r_i(\cdot)\widehat{Q}^n(\cdot).\tag{A.6}$$

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.

Assume, without loss, that all queues start empty at time zero, i.e., $\widehat{Q}_i^n(0)=0$ for $i\in\mathcal{I}$. Hence $\widehat{Q}^n(0)=0$ and $\Delta_i^n(0)=0$ for all $i\in\mathcal{I}$. We aim to show that, for each $i\in\mathcal{I}$, the process $\widehat{Q}_i^n(\cdot)$ is infinitely close to $\widehat{Q}_i^n(\cdot)$ as n grows. More precisely, We aim to show that, for each $i\in\mathcal{I}$ and $\epsilon>0$,

$$\mathbb{P}\left(|\Delta_i^n|_T^*>\epsilon\right)\to 0\quad\text{as}\quad n\to\infty.\tag{A.7}$$

Define for $i\in\mathcal{I}$ a stopping time (depending on ϵ)

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

Then to establish (A.7), it suffices to show $\mathbb{P}\left(\tilde{\tau}_i^n\leq T\right)\to 0$ as $n\to\infty$. Note that

$$\sum_{i\in\mathcal{I}}\Delta_i^n(\cdot)=\sum_{i\in\mathcal{I}}\widehat{Q}_i^n(\cdot)-\sum_{i\in\mathcal{I}}r_i(\cdot)\widehat{Q}^n(\cdot)=0.$$

The problem further boils down to showing, for each $i\in\mathcal{I}$,

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

where $\tau_i^n \equiv \inf\{t > 0 : \Delta_i^n(t) < -\epsilon\}$. On the event $C \equiv \{\tau_i \le T\}$, 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 $\widehat{Q}^n(t) > 0$ 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 (A.5), it follows immediately

(A.8)
$$n^{1/2}\int_{\sigma_i^n}^{\tau_i^n} \lambda_k(u)\mathrm{d}u \le \sum_{j\ne i} \widehat{A}_j^n[\sigma_i^n, \tau_i^n] - \widehat{D}^n[\sigma_i^n, \tau_i^n] - \int_{\sigma_i^n}^{\tau_i^n} c(u)\mathrm{d}u - \sum_{j\ne i} \widehat{Q}_i^n[\sigma_i^n, \tau_i^n] - c[\sigma_i^n, \tau_i^n]$$

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

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

$$\begin{aligned}
\mathbb{P}(\tau_i^n \le T) &\le \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) - \Gamma_i^n[\sigma_i^n, \tau_i^n] < -\epsilon/2\right) \\
&\le \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) - \Gamma_i^n[\sigma_i^n, \tau_i^n] < -\epsilon/2, \Gamma_i^n[\sigma_i^n, \tau_i^n] \le \epsilon'\right) \\
&\quad + \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) - \Gamma_i^n[\sigma_i^n, \tau_i^n] < -\epsilon/2, \Gamma_i^n[\sigma_i^n, \tau_i^n] > \epsilon'\right) \\
&\le \mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) + \mathbb{P}\left(\Gamma_i^n[\sigma_i^n, \tau_i^n] > \epsilon/4\right)
\end{aligned} \tag{A.9}$$

Recall that our goal is to show $\mathbb{P}(\tau_i^n \le T)$ goes to zero as $n \to \infty$. To that end, it suffices to show that both terms at the right end of (A.9) 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

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

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

$$\mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) \le \mathbb{P}\left(\widehat{R}_i^n(\tau_i^n) - \widehat{R}_i^n(\sigma_i^n) < -\epsilon/4\right), \tag{A.10}$$

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 A.1. *Both* $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ *and* $\{\widehat{R}_i^n(\cdot); n \in \mathbb{N}\}$ *are C-tight under the assumption of Theorem 4.2.*

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

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

Combining the above with (A.10) allows us to conclude that

$$\mathbb{P}\left(\widehat{Q}_i^n(\tau_i^n) - \widehat{Q}_i^n(\sigma_i^n) < -\epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{A.11}$$

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

$$\mathbb{P}\left(\Gamma_i^n[\sigma_i^n, \tau_i^n] > \epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty. \tag{A.12}$$

Combining (A.9), (A.11) and (A.12) yields

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

which in turn implies

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

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 step 4 - 6 in the proof the HLDR rule as in §5.4 - §5.6 leads us to the conclusion of Theorem 4.2.

Proof of Lemma A.1. By (5.15), $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ is C-tight if $\left\{\widehat{X}_i^n; n \in \mathbb{N}\right\}$ is C-tight for $i \in \mathcal{I}$. The latter holds true if the martingales $\widehat{A}_i^n$, $\widehat{D}_i^n$ and $\widehat{Y}_i^n$ are C-tight, owing to (5.11) and the established stochastic boundedness of $\widehat{X}_i^n$ and $\widehat{Q}^n$. But $\widehat{A}_i^n$, $\widehat{D}_i^n$ and $\widehat{Y}_i^n$ are C-tight ,due to (3.2), (5.20) and (5.30). Hence $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$ is C-tight. The C-tightness of $\{\widehat{R}_i^n(\cdot); n \in \mathbb{N}\}$ follows from (5.31) and the stochastic boundedness of $\{\widehat{Q}_i^n(\cdot); n \in \mathbb{N}\}$ drawing upon the stochastic boundedness of $\{\widehat{Q}^n(\cdot); n \in \mathbb{N}\}$.

DEPARTMENT OF IEOR, COLUMBIA UNIVERSITY,
NEW YORK, NY, 10027
E-MAIL: xs2235@columbia.edu