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 service-level (SL) differentiation in a multi-class many-server queueing system with time-varying (TV) arrival rates. We address the following question: How many servers are required (staffing) over time and how does the system manager make dynamic scheduling decisions to minimize staffing cost, subject to class-level quality-of-service constraints? With that goal in mind, we propose two types of ratio-control rules, namely (i) the head-of-line-delay-ratio (HLDR) rule, and (ii) the time-varying queue-ratio (TVQR) rule. The former is a blind scheduling rule that extends the accumulating-priority control proposed by Kleinrock (1964) while the latter is a TV version of the FQR rule considered by Gurvich and Whitt (2010). Both rules achieve an important state-space collapse (SSC) in the many-server heavy-traffic (MSHT) limit. We consider a variety of SL types and exploit SSC to construct asymptotically feasible solutions for the joint-staffing-and-scheduling problem. We establish the asymptotic optimality of the proposed solutions under some special choice of the system parameters. A by-product of our study is the sample-path version of the TV MSHT Little’s law that is a consequence of the MSHT limits.

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. Our study is largely motivated by hospital emergency departments and modern call centers, where (i) service demands typically vary significantly over time, and (ii) multiple job classes are present competing for resources. For instance, Canadian emergency departments classify patients into five acuity levels, with the highest acuity level 1 resuscitation, and the lowest acuity level 5 non-urgent. According to the Canadian triage and acuity scale (CTAS) guideline, “CTAS level $i$ patients need to be treated within $w_i$ minutes” with $(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120)$. In common call centers, there is the well-known x-y rule, meaning that $x\%$ of calls have to be answered within $y$ seconds.

In these contexts, the problem of achieving differentiated service with the minimum possible staffing level has two components: staffing and scheduling. In the staffing phase, one must decide how many servers should be placed in the system over time to fulfill the TV service demands. In the scheduling phase, one must decide how the

*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
next-available server should be assigned to customers in real time so that the performance metrics are not violated. A direct-analysis approach for the joint-staffing-and-scheduling problem is very difficult, often intractable. This prompts consideration of good and simple solution that results in near-optimal performance in a suitable many-server heavy-traffic (MSHT). Indeed, our proposed solutions are simple yet near-optimal for specified formulations.

Our work aims to extend, as much as possible, the result of [16] to the time-varying setting. Gurvich and Whitt [16] 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 MSHT limit; also see [14, 15]. However, it remains unclear how to integrate staffing control with scheduling decisions so as to minimize the staffing cost without the performance metrics being violated. Indeed, our propose solutions utilize the TV class-dependent arrival-rate functions.

It is significant that TV 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 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.

The staffing control we propose of matching demand for services with capacity to supply them is based on the square-root-staffing (SRS) algorithm which, in turn, uses the offered-load process of an infinite server queue. The approach follows the early papers [17, 12, 25]. The algorithm is also used in Theorem 5.1 in the electronic companion of [11], Theorem 2 in [30] and Section 2.6 in [41].

For the scheduling component, we propose the head-of-line-delay-ratio (HLDR) control, aiming to keep the head-of-line delays at desired ratios. The HLDR rule is appealing because it is a blind scheduling policy, i.e., it does not depend on any model parameters. Our HLDR controls are generalizations of the dynamic-priority control proposed by [23], which has recently been proposed for delay-based service differentiation in emergency departments based on acuity by [32] and [33]; 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 HLDR scheduling control, we also consider the time-varying queue-ratio (TVQR) rule designed to stabilize delay ratios at desired targets. This alternative is not blind, because it requires knowledge
of the arrival rates.

We find an asymptotic equivalence between the HLDR rule and the TVQR rule. This 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 limits in Theorem 4.1 and Theorem 4.2, which generalize the SP-MSHT-LL for the stationary model that is a consequence of Theorem 4.3 in [14] and is discussed after equation (13) in §3 of [16]. In particular, the SP-MSHT-LL states, for large scale systems that are approximately in the quality-and-efficiency-diven (QED) MSHT regime, that

\[ Q_i(t) \approx \lambda_i(t)V_i(t) \quad \text{for all} \quad t, \]

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 \( i \).

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 [18].

This paper is related to the problem of optimally scheduling a many-server queueing system with several classes of customers. For queues in which each class-\( i \) customer has a marginal delay cost \( c_i \) and a mean service time \( 1/\mu_i \), a static priority policy, referred to as \( c\mu \) rule, has proven asymptotically optimal; see [26, 36]. For overloaded queues with customer abandonments, [3, 4] showed that a modified priority rule, referred to as \( c\mu/\theta \), is asymptotically optimal for the long-run average cost. With a finite-horizon cost function, [8] established the asymptotic optimality of a static priority rule under the assumed condition on the ordering of the abandonment rates and running costs. The corresponding discounted-cost control problem was studied by [2, 5, 19]. Most of these papers assume class-dependent service and convert the original scheduling problem into a more tractable diffusion control problem. In contrast, our goal is to achieve desired delay-based differentiated service in the presence of time-varying arrivals with minimum possible staffing. We do not consider delay or abandonment costs.

We hasten to admit that we are by no means the first to study scheduling problems with time-varying queueing effects. Many researchers in the IEEE community have conducted research on scheduling queues with time-varying arrivals or services and our review here is by no means exhaustive. For wireless channels, [31] showed that the proposed exponential scheduling rule is near-optimal. The paper [27] considered dynamic routing and power allocation for a wireless network with time-varying channels and proposed a joint routing and power allocation policy that stabilizes the system and provides bounded average delay guarantees. For queues with asynchronously varying service rates, [1] showed that the proposed online scheduling rule, referred to as the Modified Largest Weighted Delay First, is throughput optimal. Ingolfsson et al. [20] adopted a joint approach to staffing and scheduling that allows for inclusion of time-varying arrival rates in scheduling workforce. While these papers make an excellent
contribution to scheduling in the presence of time-varying effects, none considered SL constraints.

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, [24] 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. 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 [28]. Important precedents for TV MSHT limits in the QED regime are [25] and [30].

1.2. Main Contributions.

1. We conduct what we think is the first study of SL differentiation in the presence of TV arrival rates. We consider a variety of SL types and construct asymptotically feasible solutions for the joint-staffing-and-scheduling problem. We establish the asymptotic optimality of the proposed solutions under some special choice of the system parameters.

2. We establish the first MSHT limits for ratio controls for TV multi-class many-server queues. In particular, we analyze the proposed HLDR rule and the TVQR 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 proposed control rules achieve the desired delay ratios in every sample path.

3. We show that insight can be gained into the proposed controls 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 value of of HLDR and TVQR scheduling rules with TV arrival rates. We define the model and introduce the staffing minimization problem in §3. We describe the proposed solutions to the joint staffing and scheduling problems and state the main analytical results in §4. We provide the proof of the MSHT limits in §6. The proofs for asymptotic feasibility and optimality are given in §7. We provide background on the simulation methodology and more numerical results in the supplement.
2. Initial Simulation Experiments. We illustrate the FQR, HLDR and TVQR scheduling rules with a two-class $M_t/M/s_t + M$ model having sinusoidal arrival-rate functions and staffing chosen to stabilize the aggregate performance. 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 nonhomogenous Poisson processes (NHPP’s) with arrival-rate functions

\begin{equation}
\lambda_i(t) = a_i + b_i \sin(d_i t) \quad \text{for} \quad 0 \leq t \leq T, \quad i = 1, 2.
\end{equation}

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 $\mu_i$ and abandonment rates $\theta_i$.

Let the time-varying 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 $m_i$ is given in (3.3), and the SRS staffing formula is given by (3.4).

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 [14], we infer that the queue ratio should be approximately equal to $(1/2)(60/90) = 1/3$. Hence one would want to use the FQR rule with target queue ratio $r = 1/3$. With this value, we understand that the ratio $Q_1/Q_2$ is expected to be around the target 1/3, while the delay ratio should be about 1/2. We set the fixed staffing level using the SRS staffing rule with $c = 0.25$, 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 [14].

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 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.

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, Figure 3 shows the corresponding results when we simply change the sign of $b_1$ from $-$ to $+$, which makes $AR(t) = 2/3$ for all $t$.

2.4. TV Arrivals with Abandonment. We now consider these same scheduling rules in the two-class model when there is customer abandonment. For simplicity, assume 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 safety staffing function $c \equiv 0$, which is consistent with the heuristic of “simply staffing to the offered load,” as discussed in paragraph 3 of §6 of [11].

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. The Queueing Model and the Problem Formulation. 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 potential delay in §3.4 and introduce problem formulations with different SL types in §3.5. We define the HLDR and TVQR rules in §3.6 and §3.7, respectively.

3.1. Notation and Conventions. We denote by $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$, respectively, the sets of all real numbers, 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 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 $\sigma$-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 [21] 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) $A^n_i(t)$ with rate $n\lambda_i(t)$. These NHPPs are mutually independent. For $i \in \mathcal{I}$, let

$$\Lambda_i(t) \equiv \int_0^t \lambda_i(u)du, \quad \hat{A}^n_i(t) \equiv n^{-1/2} \left(A^n_i(t) - n\Lambda_i(t)\right).$$

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

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

where $W_i$ represents a standard Brownian motion for each $i \in \mathcal{I}$. Denote by $A^n \equiv \sum_{i \in \mathcal{I}} A^n_i$ the aggregate arrival process. By the assumed independence, $A^n$ is a NHPP satisfying a FCLT; that is

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

where $\lambda_\Sigma(t) \equiv \sum_{i \in \mathcal{I}} \lambda_i(t)$ and $\Lambda(t) \equiv \int_0^t \lambda_\Sigma(u)du$. 
Service times and patience times are mutually independent and exponentially distributed, but these can be class-dependent. Let $\mu_i$ and $\theta_i$ 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 [28].

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

$$m^n_i(t) = \int_{-\infty}^t G^n_i(t-u)\lambda^n_i(u)du = \mathbb{E} \left[ \int_{t-S_i}^t \lambda^n_i(u)du \right] = \mathbb{E} [\lambda^n_i(t-S_i,e)] \mathbb{E} [S_i],$$

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

$$G_{i,e}(t) \equiv P(S_{i,e} \leq t) \equiv \frac{1}{\mathbb{E}[S_i]} \int_0^t G^n_i(u)du, \quad \text{for} \quad t \geq 0;$$

see [10]. When $S_i$ has an exponential distribution with $\mathbb{E}[S_i] = 1/\mu_i$, as we have assumed, then $m^n_i$ satisfies the ordinary differential equation (ODE)

$$m^n_i(t) = \lambda^n_i(t) - \mu_i m^n_i(t).$$

Equation (3.5) 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 (3.4) follows the early papers [17, 12, 25]. That scaling also is used in Theorem 5.1 in the electronic companion of [11], Theorem 2 in [30] and Section 2.6 in [41].
With time-varying staffing \( s^a(t) \), we need to specify how we manage the system when all servers are busy when the staffing is scheduled to decrease. What we do is to immediately enforce that staffing change, so that we need to force a customer out of service. In the single-class case we can let one customer to return to the head of the queue, as in [30]. 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 \( Q^a_{0,i}(t) \) to denote the number of class-\( i \) customers in the HPQ. We will show that the high-priority queue has no impact on the asymptotic behavior, regardless of the class identities of pushed-back customers; i.e., the content of this high-priority queue is asymptotically negligible in the MSHT scaling, and thus does not affect the limit.

**Remark 3.3.** 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.

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

\[
Q^a_i(t) = Q^a_i(0) + A^a_i(t) - \Psi^a_i(t) - R^a_i(t) 
\]

\[
(3.6)
\]

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

\[
Q^a_{0,i}(t) + B^a_i(t) = Q^a_{0,i}(0) + B^a_i(0) + \Psi^a_i(t) - D^a_i(t) 
\]

\[
(3.7)
\]

where \( \Pi^d_i \) are unit-rate Poisson processes independent of \( \Pi^a_i \) and \( \Pi^{ab}_i \) given in (3.6). Let \( X^a_i(t) \) denote the total number of class-\( i \) customers in system at time \( t \). Adding
\(X_i^n(t) = Q_i^n(t) + Q_{0,i}^n(t) + B_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - R_i^n(t).\)

Alternatively, one can derive (3.8) directly from flow conservation.

Finally, let \(Q_{0,i}^n(t) \equiv \sum_{i \in \mathcal{I}} Q_{0,i}^n(t), \) \(Q^n(t) \equiv \sum_{i \in \mathcal{I}} Q_i^n(t)\) 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 (3.8) over \(i \in \mathcal{I}\) yields

\[(3.9) \quad X^n(t) = Q^n(t) + Q_{0}^n(t) + B^n(t) = X^n(0) + A^n(t) - D^n(t) - \sum_{i \in \mathcal{I}} R_i^n(t)\]

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

\[(3.10) \quad A_0^n(t) \equiv \sum_{u \in S^n_a(t)} 1(B^n(u-) = s^n(u-)).\]

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

\[(3.11) \quad D_0^n(t) \equiv \sum_{u \in S^n_d(t)} 1(Q_0^n(u-) > 0) + \int_0^t 1(Q_0^n(u-) > 0)dD^n(u).\]

From (3.10) and (3.11), it follows that

\[Q_0^n(t) = A_0^n(t) - D_0^n(t)\]

\[= \sum_{u \in S^n_a(t)} 1(B^n(u-) = s^n(u-)) - \sum_{u \in S^n_d(t)} 1(Q_0^n(u-) > 0)\]

\[- \int_0^t 1(Q_0^n(u-) > 0)dD^n(u).\]

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

\[(3.13) \quad Z^n(t) \equiv s^n(0) - s^n(t) - D^n(t), \quad t \geq 0.\]
and apply the one-dimensional reflection mapping $\psi$ to $Z^n$ to get

$$\Upsilon^n_0(t) \equiv \psi(Z^n)(t) \equiv Z^n(t) - \inf_{0 \leq u \leq t} \{Z^n(u)\};$$

(3.14)

e.g., see §13.5 in [37]. The following lemma shows that $\Upsilon^n_0$ serves as an upper bound for $Q^n_0$.

**Lemma 3.1.** Let $Q^n_0$ and $\Upsilon^n_0$ be as given in (3.12) and (3.14) respectively. Then

$$Q^n_0(t) \leq \Upsilon^n_0(t) \quad \text{for all } t \geq 0 \quad \text{w.p.1.}$$

**Proof of Lemma 3.1.** By (3.14) and (3.13), it is not hard to see that

$$\Upsilon^n_0(t) = \sum_{u \in S^n(t)} 1 - \sum_{u \in S^n(u)} 1(\Upsilon^n_0(u) > 0) - \int_0^t 1(\Upsilon^n_0(u) > 0) dD^n(u).$$

(3.15)

Combining (3.12) and (3.15) 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 §6 we will show that $\Upsilon^n_0(t)$ is asymptotically negligible in the MSHT scaling, and so $Q^n_0(t)$ has no impact on the MSHT limit.

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

$$V^n_i(t) \equiv \inf\{s \geq 0 : \Psi^n_i(t + s) \geq Q^n_i(0) + A^n_i(t)\}.$$

One may attempt to incorporate the abandonment process $R^n_i$ into the expression and write

$$V'^n_i(t) \equiv \inf\{s \geq 0 : \Psi'^n_i(t + s) + R^n_i(t + s) \geq Q^n_i(0) + A^n_i(t)\},$$

(3.16)

but the representation (3.16) is incorrect, because the term $R^n_i(t + s)$ may include class-$i$ customers that arrived after time $t$ and then abandoned; see §1 in [35].

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 [35]. Following the notation of that paper, we define $R^n_i(t, s)$ to be the number of class-$i$ customers who arrived before time $t$ but have abandoned over the time interval $[t, s)$. Then the potential delay in queue $i$ at time $t$ can be represented as the following first-passage time

$$V^n_i(t) \equiv \inf\{s \geq 0 : \Psi^n_i(t + s) + R^n_i(t + s) > Q^n_i(0) + A^n_i(t)\}.$$

(3.17)
3.5. The Problem Formulation. We are now ready to formulate the staffing minimization problem subject to the SL constraints.

We start with mean-waiting-time formulation. In particular, we solve at any given point $t$ in time the following optimization problem:

\begin{equation}
\begin{aligned}
\text{minimize } & s^n(t) \\
\text{subject to: } & \mathbb{E}[V^n_i(u)] \leq T^n_i(u) \quad \text{for } u \leq t, \ i \in \mathcal{I}.
\end{aligned}
\end{equation}

Recall that $V^n_i(t)$, as in (3.17), represents the waiting time of a virtual customer of class $i$ that arrives at time $t$. These SL constraints stipulate that the expected delay in queue $i$ at time $t$ shall not exceed the target $T^n_i(t)$. Here we allow the SL targets $T^n_i(\cdot)$ be functions in time. We will let $T^n_i$ scale with $n$ so as to put our system into the QED MSHT regime. Particularly we make the following assumption.

**Assumption 1.** (QED scaling for SL targets) The SL target functions $T^n_i(\cdot)$ are scaled so that $T^n_i(\cdot) = T_i(\cdot)/\sqrt{n}$ for some pre-specified functions $T_i$, $i \in \mathcal{I}$.

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

**Definition 3.1.** We say that a joint-staffing-and-scheduling policy $(s, \pi)$ is admissible if (i) the staffing component $s$ follows the SRS rule (3.4), and (ii) the scheduling component $\pi$ is nonanticipative. We let $\Pi$ be the set of all admissible policies.

**Definition 3.2.** (asymptotic feasibility for the mean-waiting-time formulation) A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically feasible for (3.18) if $(s^n, \pi^n) \in \Pi$ and

\begin{equation}
\limsup_{n \to \infty} \mathbb{E}[V^n_i(t)/T^n_i(t)] \leq 1 \quad \text{for all } t, \ i \in \mathcal{I}.
\end{equation}

**Definition 3.3.** (asymptotic optimality for the mean-waiting-time formulation) A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically optimal for (3.18), if it is asymptotically feasible and for any other sequence $\{(s^m, \pi^m)\}$ that is asymptotically feasible,

\begin{equation}
[s^n(t) - s^m(t)]^+ = o(n^{1/2}) \quad \text{as } n \to \infty, \quad \text{for all } t.
\end{equation}

We next consider an alternative formulation representing the goal of common callcenters. This formulation aims to control the tail probability of the waiting time of each class. At any given point $t$ in time, we solve the optimization problem

\begin{equation}
\begin{aligned}
\text{minimize } & s^n(t) \\
\text{subject to: } & \mathbb{P}(V^n_i(u) > T^n_i(u)) \leq \alpha \quad \text{for } u \leq t, \ i \in \mathcal{I}.
\end{aligned}
\end{equation}
Again, $V_i^n$ and $T_i^n$ represent the real-time virtual waiting time of class $i$ customers and time-dependent SL target respectively. Then the set of constraints requires that the probability that a class $i$ customer who arrives at time $t$ waits longer than $T_i^n(t)$ time units is no greater than $\alpha$.

As was carefully discussed by [16], this seemingly reasonable formulation can be problematic; see also [13]. Problem may arise because one can simply choose not to serve any class-$i$ customer who has waited longer than the performance target, without violating any of the SL constraints. The difficulty can be circumvented by adding a global SL constraint as was done in §2.2.1 of [16]. Such a formulation and its corresponding solution will be considered shortly. At the moment, we will discuss the asymptotical feasibility for problem (3.21) despite the fact that this formulation is somewhat problematic.

**Definition 3.4.** (asymptotic feasibility for the tail-probability formulation) A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically feasible for (3.21) if, $(s^n, \pi^n) \in \Pi$, and for every $\epsilon > 0$,

$$\limsup_{n \to \infty} P \left( \frac{V_i^n(t)}{T_i^n(t)} \geq 1 + \epsilon \right) \leq \alpha \quad \text{for all } t, \quad i \in \mathcal{I}. \quad (3.22)$$

As was alluded to earlier, a global SL constraint is sometimes required for the tail-probability formulation to be well-posed, which naturally leads to our third formulation which we call the mixed formulation. At each time $t$, we solve the following optimization problem:

$$\begin{align*}
\text{minimize} & \quad s^n(t) \\
\text{subject to:} & \quad \mathbb{E}[Q^n(u)] \leq q^n(u) \quad \text{for } u \leq t, \\
& \quad P \left( \frac{V_i^n(u)}{T_i^n(t)} \leq \frac{T_i^n(t)}{T_i^n(u)} \right) \leq \alpha \quad \text{for } u \leq t, \quad i = 1, \ldots, K - 1.
\end{align*} \quad (3.23)$$

We recall that $Q^n(t)$ represents the total number of waiting customers in system at time $t$. Again, we let the target function $q^n$ scale with $n$ so as to force the system to operate in the QED regime. In particular, we make the following assumption by which the underlying staffing rule has to be in the form of (3.4).

**Assumption 2.** (QED scaling for SL targets) the SL target function $q^n(\cdot)$ is scaled so that $q^n(\cdot) = \sqrt{n}q(\cdot)$ for some pre-specified functions $q$.

**Definition 3.5.** (asymptotic feasibility for the mixed formulation) A sequence of staffing functions and scheduling policies $\{(s^n, \pi^n)\}$ is said to be asymptotically feasible for (3.23) if, $(s^n, \pi^n) \in \Pi$, and for every $\epsilon > 0$,

$$\begin{align*}
\limsup_{n \to \infty} \mathbb{E}[Q^n(t)/q^n(t)] & \leq 1 \quad \text{for all } t, \quad \text{and} \\
\limsup_{n \to \infty} P \left( \frac{V_i^n(t)}{T_i^n(t)} \geq 1 + \epsilon \right) & \leq \alpha \quad \text{for all } t, \quad i = 1, \ldots, K - 1. \quad (3.24)
\end{align*}$$
Definition 3.6. (asymptotic optimality for the mixed formulation) A sequence of staffing functions and scheduling policies \( \{(s^n, \pi^n)\} \) is said to be asymptotically optimal for (3.23), if it is asymptotically feasible and for any other sequence \( \{(s'^n, \pi^n)\} \) that is asymptotically feasible,

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

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

\[
\tilde{w}^n_i(t) \equiv w^n_i(t)/v_i(t) \quad \text{for each} \quad i \in \mathcal{I}.
\]

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

\[
\tilde{w}^n(t) \equiv \max_{i \in \mathcal{I}} \{\tilde{w}^n_1(t), \ldots, \tilde{w}^n_K(t)\} = \max_{i \in \mathcal{I}} \{w^n_1(t)/v_i(t), \ldots, w^n_K(t)/v_K(t)\}.
\]

Let \( \tau(t) \) denote the customer class that has the maximum weighted HoL delay; i.e.,

\[
\tau(t) \equiv \{i \in \mathcal{I} : \tilde{w}^n_i(t) = \tilde{w}^n(t)\}.
\]

We can then spell out the assignment processes \( \Psi^n_i(\cdot) \):

\[
\Psi^n_i(t) = \sum_{u \in \mathcal{T}^n(t)} 1(\tau(u) = i),
\]

where \( \mathcal{T}^n(t) \) is the collection of time instances up to time \( t \) at which a scheduling decision is to be made and \( \tau(\cdot) \) is given by (3.28). Here ties are broken arbitrarily. For instance, if \( \tilde{w}^n_i(t) = \tilde{w}^n_i(t') = \tilde{w}^n(t) \) for \( i \neq i' \), then the next-available server chooses to serve either queue \( i \) or queue \( i' \) with equal probabilities.

Remark 3.4 (reduction to AP and global FCFS). 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 at 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, HLDR selects the waiting customer with the highest accumulating priority for service.

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 [34].
3.7. The TVQR Control. As indicated earlier, our HLDR control is intimately related to TV version of the QR rule studied in [14]. We briefly review the FQR control, which is a special case of the more general QR control introduced by [14], 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 $v$ to be the diffusion-scaled queue-length processes and $\hat{v}$ control uses a vector function $v$ itself can be extended to a large class of TV arrival-rate functions.

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 to serve a class $i$ customer where

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

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

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 [14] 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

$$ X_i^n(\cdot) \equiv n^{-1/2} (X_i^n(\cdot) - n \cdot m_i(\cdot)) \quad \text{and} \quad \hat{X}_i^n(\cdot) \equiv n^{-1/2} (X_i^n(\cdot) - n \cdot m(\cdot)), $$

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

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

be the diffusion-scaled queue-length processes and $\hat{Q}_n \equiv n^{-1/2} Q^n$ and $\hat{Q}_0 \equiv n^{-1/2} Q^n_0$ be the aggregate quantities. The same scaling was used by [11, 30, 41]. As usual, we scale the delay processes by multiplying by $\sqrt{n}$ instead of dividing by $\sqrt{n}$ as in (4.2):

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

We impose the following regularity conditions:

**Assumption 3.** (A1) For each $i \in \mathcal{I}$, the arrival-rate function $\lambda_i(\cdot)$ is differentiable with bounded first derivative; 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 functions $\lambda_i(\cdot)$ are bounded away from zero; i.e., there exists $\lambda_* > 0$ such that $\lambda_* \equiv \min_{i \in \mathcal{I}} \inf_{t \geq 0} \lambda_i(t) > 0$ for all $t$. 


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

(A3) All control functions \( v_i(\cdot) \) are continuous and bounded from above and away from zero; 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 is a set of interacting diffusion processes.

**Theorem 4.1 (QED MSHT FCLT for HLDR).** Suppose that the system is staffed according to (3.4), operates under the HLDR scheduling rule and Assumptions A1 - A3 hold. If, in addition, there is convergence of the initial distribution at time 0, i.e., if

\[
(\hat{X}^n_1(0), \ldots, \hat{X}^n_K(0), \hat{Q}^n_1(0), \ldots, \hat{Q}^n_K(0)) \Rightarrow (X^{(d)}_1(0), \ldots, X^{(d)}_K(0), Q^{(d)}_1(0), \ldots, Q^{(d)}_K(0))
\]

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

\[
\begin{aligned}
(\hat{X}^n_1, \ldots, \hat{X}^n_K, \hat{Q}^n_1, \ldots, \hat{Q}^n_K, \hat{V}^n_1, \ldots, \hat{V}^n_K, \hat{w}^n_1, \ldots, \hat{w}^n_K)
\Rightarrow
(X^{(d)}_1, \ldots, X^{(d)}_K, Q^{(d)}_1, \ldots, Q^{(d)}_K, V^{(d)}_1, \ldots, V^{(d)}_K, w^{(d)}_1, \ldots, w^{(d)}_K)
\end{aligned}
\]

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

\[
X^{(d)}_i(t) = X^{(d)}_i(0) - \mu_i \int_0^t X^{(d)}_i(u) du - \left( \theta_i - \mu_i \right) \int_0^t \gamma(u)^{-1} \lambda_i(u) \lambda_i(u) du + \int_0^t \gamma(u)^{-1} \lambda_i(u) \lambda_i(u) \mu_i(u) dW_i(u)
\]

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 \( i \in \mathcal{I} \),

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

### 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 (3.4), the behavior on the fluid scale is determined by the offered load \( m(t) \equiv m_1(t) + \cdots + m_K(t) \), where the individual per-class offered loads \( m_i(\cdot) \) depend on the specified \( \lambda_i(\cdot) \) and \( \mu_i \) 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 (3.4) is specified by the SRS safety
function \( c \), which appears explicitly in the diffusion limit. Hence, in the limit, the
remaining flexibility in the staffing depends entirely on the single function \( c(\cdot) \), 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 \((\hat{X}_1^{n}(\cdot), \ldots, \hat{X}_K^{n}(\cdot))\)
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 \([X^{(d)}(u) - c(u)]^+\), 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 role of customer abandonment. While customer abandonment does in-
fluence 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 determin-
ing these limiting ratios. It is wiped out in the heavy-traffic diffusion limit. For the
\( n \)-th model, both arrivals and departures occur at a time scale of \( n^{-1} \). But because
the queue-lengths live on the order of \( n^{1/2} \) in the QED, abandonments occur at a time
scale of \( n^{-1/2} \) indicating a much slower rate. This observation is consistent with [38]
for the basic \( M/M/s + M \) Erlang-A model.

4.2.4. the sample-path MSHT Little’s law. We obtain the SP MSHT LL directly
from the conclusion of Theorem 4.1. In particular, for each \( i \), we see that, almost
surely,
\[
Q_i^{(d)}(t) = \lambda_i(t) V_i^{(d)}(t) \quad \text{for all} \quad t \geq 0.
\]
For the \( n \)-th system, we have
\[
\hat{Q}_i^n(t) = \lambda_i(t) \hat{V}_i^n(t) + o(1) \quad \text{as} \quad n \to \infty
\]
or
\[
Q_i^n(t) = \lambda_i^n(t) V_i^n(t) + o(\sqrt{n}) \quad \text{as} \quad n \to \infty.
\]
That is, the limit tells us that \( Q_i^n(t) \) 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.
Fig 5: Sample paths of the queue-length process $Q_i(\cdot)$ and the scaled delay process $v_i(\cdot)\cdot w_i(\cdot)$ for $i = 1, 2$ with the HLDR scheduling policy.

4.2.5. Impact of the arrival-rate and the weight functions. Given the limit for the queue-length processes in (4.6), 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,

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

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

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

$$\Rightarrow \left(X^{(d)}_1, \ldots, X^{(d)}_K, Q^{(d)}_1, \ldots, Q^{(d)}_K, V^{(d)}_1, \ldots, V^{(d)}_K, w^{(d)}_1, \ldots, w^{(d)}_K\right) \text{ in } D^{4K}$$

as $n \to \infty$ where the diffusion limits $X^{(d)}_i$ satisfy

$$X^{(d)}_i(t) = X^{(d)}_i(0) - \mu_i \int_0^t X^{(d)}_i(u)du$$

$$- (\theta_i - \mu_i) \int_0^t \gamma^{-1} v_i \lambda_i \left[ X^{(d)}(u) - c \right]^+ du + \sqrt{2\lambda_i} W_i(t).$$
in which $\gamma = \sum_{i \in \mathcal{I}} v_i \lambda_i$ and $X^{(d)}(t) = \sum_{i \in \mathcal{I}} X^{(d)}_i(t)$; for each $i \in \mathcal{I}$

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

Corollary 4.1 is in agreement with Theorem 4.3 in [14] 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 [14] has $[\bar{X}]^+$ and $[\bar{X}]^-$ 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 [14] 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 (4.10) is in alignment with previous work on AP by [23] and [33], where the objective is to achieve desired ratios of stationary mean waiting times experienced by customers from the different classes. By focusing on the QED MSHT regime, we are able to obtain a much stronger sample-path result.

If $\mu_i = \mu$ and $\theta_i = \theta$, $u \in \mathcal{I}$, then the limit of the aggregate content process $X^{(d)}$ is a one-dimensional diffusion. Hence, the limit is essentially the same as that for the single-class $M_t/M/s_t + M$ model as considered by [41] where the analysis draws upon [30].

**Corollary 4.2 (class-independent services and abandonments).** Suppose that the conditions in Theorem 4.1 are satisfied and $\mu_i = \mu$, and $\theta_i = \theta$, $i \in \mathcal{I}$. Then

$$\left( \hat{X}^n, \hat{Q}^n_1, \ldots, \hat{Q}^n_K, \hat{V}^n_1, \ldots, \hat{V}^n_K, \hat{w}^n_1, \ldots, \hat{w}^n_K \right) \Rightarrow \left( X^{(d)}, Q^{(d)}_1, \ldots, Q^{(d)}_K, V^{(d)}_1, \ldots, V^{(d)}_K, w^{(d)}_1, \ldots, w^{(d)}_K \right)$$

where

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

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

$$Q_i^{(d)}(\cdot) \equiv \gamma(\cdot)^{-1} v_i(\cdot) \lambda_i(\cdot) \left[ X^{(d)}(\cdot) - c(\cdot) \right]^+,$$
$$V_i^{(d)}(\cdot) = w_i^{(d)}(\cdot) \equiv v_i(\cdot) \cdot \gamma(\cdot)^{-1} \left[ X^{(d)}(\cdot) - c(\cdot) \right]^+.$$
If we assume further that $\theta = \mu$ in Corollary 4.2, then the aggregate model is known to behave like an $M_t/M/\infty$ model. Let $\theta = \mu = 1$ in (4.11). From 4.11 it holds that

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

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

4.4. The MSHT FCLT for TVQR in the QED Regime.

We now turn to the TVQR control as described by §3.7. Mimicking the analysis of [14], one can establish the MSHT limits, regarding the TVQR rule, via hydrodynamic limits. However, the proof in [14] is quite involved and in turn relies on additional general state space collapse (SSC) results from [9]. 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 §6, 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.

\textbf{Theorem 4.2 (QED MSHT FCLT for TVQR).} Suppose that the system is staffed according to (3.4), operates under the TVQR scheduling rule and Assumptions A1 - A2 hold. If, in addition,

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

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

$$(4.13) \implies \left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n, \hat{w}_1^n, \ldots, \hat{w}_K^n\right)$$

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

$$(4.14) X_i^{(d)}(t) = X_i^{(d)}(0) - \mu_i \int_0^t X_i^{(d)}(u)du$$

$$- (\theta_i - \mu_i) \int_0^t r_i(u) [X^{(d)}(u) - c(u)]^+ du + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}dW_i(u)$$

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

$$(4.15) Q_i^{(d)}(\cdot) = r_i(\cdot) \left[ X^{(d)}(\cdot) - c(\cdot) \right]^+, \quad \text{and} \quad V_i^{(d)}(\cdot) = w_i^{(d)}(\cdot) \equiv \frac{r_i(\cdot)}{\lambda_i(\cdot)} \cdot \left[ X^{(d)}(\cdot) - c(\cdot) \right]^+. $$
We gain several insights from the theorem above: (a) with the TVQR, the desired queue-ratio is achieved in the limit despite the fact that arrival rates are changing; (b) 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

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

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

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

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

### 4.6. Extending the QIR Limits to TV Arrivals.

Even though [14] 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 [14] 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 \( D([0, t], \mathbb{R}) \) 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.

### 4.7. The Proposed Solution.

For each formulation introduced above, we propose a solution that consists of a staffing component and a scheduling component. Recall that \((v_1, \ldots, v_K)\) and \((r_1, \ldots, r_K)\) are the ratio functions in the HLDR and TVQR rule respectively and \(c\) is the TV safety staffing function.
4.7.1. Mean-Waiting-Time Formulation. We start with the mean-waiting-time formulation as given by (3.18).

- **staffing:** Choose $c^*$ that satisfies $\mathbb{E} \left[ X^{(d)}(t) - c^*(t) \right]^+ = \vartheta(t)$ with
  \begin{equation}
  (4.16) \quad \vartheta(t) = \sum_{i \in \mathcal{I}} \lambda_i(t) T_i(t).
  \end{equation}

- **scheduling:** (a) Apply HLDR with ratio functions
  \begin{equation}
  (4.17) \quad v^* \equiv (v_1^*(t), \ldots, v_K^*(t)) = (T_1(t), \ldots, T_K(t)),
  \end{equation}
  or (b) use TVQR with ratio functions
  \begin{equation}
  (4.18) \quad r^* \equiv (r_1^*(t), \ldots, r_K^*(t)) = (\lambda_1(t) T_1(t), \ldots, \lambda_K(t) T_K(t))/\vartheta(t).
  \end{equation}

Informally, our MSHT FCLT in Theorem 4.1 justifies the following approximation:

\[
\mathbb{E}[V_{ni}(t)]/T_{ni}(t) \approx \mathbb{E} \left[ \frac{V_i^{(d)}(t)}{T_i(t)} \right] = \mathbb{E} \left[ \frac{X^{(d)}(t) - c^*(t)}{\vartheta(t)} \right] = 1.
\]

**Theorem 4.3.** (asymptotic feasibility and optimality of the mean-waiting-time formulation) Let $s^n$ be determined through the square-root staffing in (3.4) with $c^*$ as specified above. Set $\pi^n$ to HLDR with ratio functions $v^*$. Then, the sequence $\{(s^n, \pi^n)\}$ is asymptotically feasible for (3.18). If, in addition, we have $\mu_i = \mu$ and $\theta_i = \theta$ for $i \in \mathcal{I}$, then the sequence $\{(s^n, \pi^n)\}$ is also asymptotically optimal.

4.7.2. Tail-Probability Formulation. For the tail-probability formulation given in (3.21), we propose the following solution.

- **staffing:** Choose $c^*$ that satisfies $\mathbb{P} \left( X^{(d)}(t) > \vartheta(t) + c^*(t) \right) = \alpha$, for $t \geq 0$.
- **scheduling:** (a) apply HLDR with ratio functions given in (4.17), or (b) use TVQR with ratio functions given in (4.18).

Informally, our MSHT FCLT in Theorem 4.1 supports the use of the following approximation:

\[
\mathbb{P} \left( V_{ni}(t) > T_{ni}(t) \right) \approx \mathbb{P} \left( \frac{V_i^{(d)}(t)}{T_i(t)} > \frac{X^{(d)}(t) - c^*(t)}{\vartheta(t)} \right).
\]

**Theorem 4.4.** (asymptotic feasibility of the tail-probability formulation) Let $s^n$ be determined through the square-root staffing in (3.4) with $c^*$ as specified above. Set $\pi^n$ to HLDR with ratio functions $v^*$. Then, the sequence $\{(s^n, \pi^n)\}$ is asymptotically feasible for (3.21).
4.7.3. **Mixed Formulation.** For the mixed formulation given in (3.23), our proposed solution is stated as follows.

**staffing:** Choose \( c^*(\cdot) \) that satisfies \( E \left[ X^{(d)}(t) - c^*(t) \right] = q(t) \), for each \( t \geq 0 \).

**scheduling:** For the function \( c^* \) as determined above, choose \( x(t) \) satisfying \( P \left( X^{(d)}(t) > x(t) + c^*(t) \right) = \alpha \), for \( t \geq 0 \). For each \( t \leq T \), set \( T_K(t) = \left[ x(t) - \sum_{i=1}^{K-1} \lambda_i(t)T_i(t) \right] / \lambda_K(t) \). Then apply HLDR with ratio functions given in (4.17), or (b) use TVQR with ratio functions given in (4.18).

**Theorem 4.5.** (asymptotic feasibility and optimality of the mixed formulation) Let \( s^n \) be determined through the square-root staffing in (3.4) with \( c^* \) as specified above. Set \( \pi^n \) to HLDR with ratio functions \( v^* \). Then, the sequence \( \{(s^n, \pi^n)\} \) is asymptotically feasible for (3.23). If, in addition, we have \( \mu_i = \mu, \theta_i = \theta \) for \( i \in I \) and \( \theta \leq \mu \), then the sequence \( \{(s^n, \pi^n)\} \) is also asymptotically optimal.

5. **Computing the Optimal Safety Staffing Function.** Successful application of these staffing-and-scheduling solutions requires effective computation of the minimum safety staffing function \( c^* \). In this section we illustrate how the function \( c^* \) can be calculated explicitly under some special choice of the system parameters.

5.1. **The Experimental Setting.** We use the same numerical example as in §2 but choosing \( (a_1, b_1, d_1) = (60, -20, 2/5) \) and \( (a_2, b_2, d_2) = (90, 30, 2/5) \) in (2.1). Assume that \( \mu_i = \theta_i = 1, i = 1, 2 \), from Corollary 4.2 and the discussion below, we know that the aggregate model behaves like an infinite-server queue. As a consequence, its diffusion limit \( X^{(d)} \) is an OU process with time-varying diffusion parameter. One can easily show that, for each \( t \geq 0 \), the random variable \( X^{(d)}(t) \) is normally distributed with mean 0 and variance \( m(t) \). In addition, suppose that the SL targets for class-1 and class-2 are \( T_{n1} \equiv 1/6 \) and \( T_{n2} \equiv 1/3 \) respectively.

In the context of a hospital ED, where an average treatment time is about 90 minutes, a cycle would be about 5\( \pi \) time longer which is about 24 hours or a day, and the SL targets are 15 minutes and 30 minutes respectively for high-acuity and low-acuity patients. Thus, our parameter choice can provide insight for EDs.

We consider the mean-waiting-time formulation and the tail-probability formulation introduced in (3.18) and (3.21), respectively. For each formulation, we provide below an explicit expression for the corresponding minimum safety staffing function \( c^* \). We then apply the solutions in §4.7.1 and §4.7.2 to conduct the simulation studies.

5.2. **Calculating the Minimum Safety Staffing Level.** To calculate the minimum safety staffing function \( c^* \) for the mean-waiting-time formulation, let

\[
\vartheta(t) = E \left[ X^{(d)}(t) - c^*(t) \right] + .
\]
Each $X^{(d)}(t)$ is a normal random variable with mean 0 and variance $m(t)$. Thus,

\begin{equation}
(5.1) \quad c^*(t) = \sqrt{m(t)} \cdot \tilde{c}(t)
\end{equation}

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

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

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

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

Because $X^{(d)}(t)$ is normally distributed with mean 0 and variance $m(t)$, we have

\begin{equation}
(5.2) \quad c^*(t) = \Phi^{-1}(1 - \alpha) \sqrt{m(t)} - \vartheta(t).
\end{equation}

5.3. Insights. Figure 6 depicts the potential delays over the time interval $[0, 50]$ for the HLDR rule (left) and the TVQR rule (right) with $c^*$ derived from (5.1). We plot the potential delays for both classes. All estimates were obtained by averaging over 2000 independent replications. Figure 6 shows that both HLDR and TVQR stabilize the potential delay of each class at the associated SL target.

Fig 6: Potential delays 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 $\mu = 1$ and abandonment rate $\theta = 1$ and minimum staffing function $c^*$ derived from (5.1).

Figure 7 plots the tail probabilities over the time interval $[0, 50]$ for the HLDR rule (plots at the top) and the TVQR rule (plots at the bottom) with $c^*$ derived from (5.2).
Here we tested three different tail-probability targets, $\alpha = 0.25, 0.5, 0.75$. We plot the tail probabilities for both classes. All estimates were obtained by averaging over 2000 independent replications. Figure 7 shows that, for all three cases, both HLDR and TVQR stabilize the tail probabilities of each class at the desired level.

Fig 7: Tail probabilities 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 $\mu = 1$ and abandonment rate $\theta = 1$ and minimum staffing function $c^+$ derived from (5.2).

6. Proof of MSHT FCLT for HLDR and TVQR.

Proof of Theorem 4.1. For any $x \in \mathcal{D}$, let $x(t_1, t_2) \equiv x(t_2) - x(t_1)$. In addition, let $L_{n,t}^i(s)$ denote the number of class-$i$ customers who arrived after time $t$ but have abandoned in the interval $[t, s)$. With the HLDR control, the queue-length processes satisfy

\begin{equation}
Q_i^n(t-) = A^n_i[H_i(t), t] - L_{i,H}^n(t)H_i(t), t) = A^n_i[t - w^n_i(t), t] - L_{i,t}^{n,t-w^n_i(t)}[t - w^n_i(t), t].
\end{equation}

Let

\begin{equation}
\hat{R}_i^n(\cdot) \equiv n^{-1/2}R_i^n(\cdot), \quad \hat{R}_i^{n,t}(t+) \equiv n^{-1/2}R_i^{n,t}(t+) \quad \text{and} \quad \hat{L}_i^{n,t}(t+) \equiv n^{-1/2}L_i^{n,t}(t+).
\end{equation}
By the definition of $R_i^n$, $R_i^{n,t}$ and $L_i^{n,t}$, we have

\begin{equation}
\hat{R}_i^n[t,s] = \hat{R}_i^{n,t}(s) + \hat{L}_i^{n,t}(s).
\end{equation}

Combining (3.2), (4.2), (6.1) and (6.2) yields

\begin{equation}
\hat{Q}_i^n(t) = \hat{A}_i^n[t - w_i^n(t), t] + n^{1/2} \int_{t-w_i^n(t)}^t \lambda_i(u)du - \hat{L}_i^{n,t-w_i^n(t)}[t - w_i^n(t), t]
\end{equation}

\begin{equation}
= \hat{A}_i^n[t - w_i^n(t), t] + n^{1/2}\lambda_i(t)w_i^n(t) - \hat{L}_i^{n,t-w_i^n(t)}[t - w_i^n(t), t] + e_i^n(t)
\end{equation}

where

\begin{equation}
e_i^n(t) \equiv n^{1/2} \int_{t-w_i^n(t)}^t \lambda_i(u)du - n^{1/2}\lambda_i(t)w_i^n(t).
\end{equation}

Introduce the auxiliary process

\begin{equation}
\hat{K}_i^n(t) \equiv \hat{A}_i^n[t - w_i^n(t), t] - \hat{L}_i^{n,t-w_i^n(t)}[t - w_i^n(t), t] + e_i^n(t) \quad \text{for } i \in I.
\end{equation}

Then inserting (6.6) into (6.4) yields

\begin{equation}
\hat{Q}_i^n(t) = \lambda_i(t)\hat{w}_i^n(t) + \hat{K}_i^n(t), \quad i \in I.
\end{equation}

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

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

\begin{equation}
(Q_1^n(t), \ldots, Q_K^n(t)) \propto (v_1(t)\lambda_1(t), \ldots, v_K(t)\lambda_K(t)) \quad \text{for all } t \leq T.
\end{equation}

This is essentially a state-space-collapse (SSC) result in the many-server diffusion limit. Finally, by a similar argument as in [14] (first SSC and then diffusion limits), we obtain the diffusion limits for $\hat{X}_i^n(\cdot)$. The limits for the queue-length processes and delay processes follow immediately.

1. Stochastic Boundedness of $\{\hat{X}_i^n(\cdot); n \in \mathbb{N}\}$ and $\{\hat{Q}_i^n(\cdot); n \in \mathbb{N}\}$. Here we exploit a martingale decomposition, as in [28] and [30]. Specifically the processes

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

\begin{equation}
= n^{-1/2} \left[ \Pi_i^d \left( \mu_i \int_0^t B_i^n(u)du \right) - \mu_i \int_0^t B_i^n(u)du \right]
\end{equation}
and
\[
\hat{Y}_t^n(t) = n^{-1/2} \left[ R_t^n(t) - \theta_t \int_0^t Q_t^n(u) du \right]
\]
\[
= n^{-1/2} \left[ \Pi_{ab}^n \left( \theta_t \int_0^t Q_t^n(u) du \right) - \theta_t \int_0^t Q_t^n(u) du \right]
\]
are square-integrable martingales with respect to a proper filtration. The associated quadratic variation processes are
\[
\langle \hat{D}_t^n \rangle(t) = \frac{\mu_t}{n} \int_0^t B_t^n(u) du \quad \text{and} \quad \langle \hat{Y}_t^n \rangle(t) = \frac{\theta_t}{n} \int_0^t Q_t^n(u) du.
\]
Both \(\{\hat{D}_t^n(\cdot); n \in \mathbb{N}\}\) and \(\{\hat{Y}_t^n(\cdot); n \in \mathbb{N}\}\) are stochastically bounded due to Lemma 5.8 of [28], which is based on the Lenglart-Rebolledo inequality, stated as Lemma 5.7 there.

From (3.5), it follows
\[
m_t^n(0) + \int_0^t \lambda_t(u) du - \frac{\mu_t}{n} \int_0^t m_t^n(u) du.
\]
Scaling both sides of (6.11) by \(n\) and subtracting it from (3.8) gives us
\[
X_t^n(t) - n m_t^n(0) + \int_0^t \lambda_t(u) du - \mu_t \int_0^t m_t^n(u) du.
\]
Dividing both sides by \(n^{1/2}\) yields
\[
\hat{X}_t^n(t) = \hat{X}_t^n(0) - \mu_t \int_0^t \hat{X}_t^n(u) du + \mu_t \int_0^t \hat{Q}_t^n(u) du
\]
\[
- (\theta_t - \mu_t) \int_0^t \hat{Q}_t^n(u) du + \hat{A}_t^n(t) - \hat{D}_t^n(t) - \hat{Y}_t^n(t).
\]
Let \(\bar{a} \equiv \max_i \mu_i \vee \max_i \theta_t\) and
\[
\hat{M}_t^n(t) \equiv \hat{A}_t^n(t) - \hat{D}_t^n(t) - \hat{Y}_t^n(t).
\]
Note that \(\{\hat{M}_t^n; n \in \mathbb{N}\}\) is stochastically bounded. Using (6.12) - (6.13), we have
\[
\left| \hat{X}_t^n(t) \right| \leq \left| \hat{X}_t^n(0) \right| + \bar{a} \int_0^t \left[ \left| \hat{X}_t^n(u) \right| + \hat{Q}_t^n(u) + \hat{Q}_{0,i}^n(u) \right] du + \left| \hat{M}_t^n(t) \right|.
\]
Adding up (6.14) over \( i \in \mathcal{I} \), we obtain

\[
\sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(t) \right| \leq \sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(0) \right| + \bar{a} \int_0^t \left\{ \sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(u) \right| + \hat{Q}_n(u) + \hat{Q}_0^n(u) \right\} du + \sum_{i \in \mathcal{I}} \left| \hat{M}_i^n(t) \right|.
\]

In addition,

\[
\hat{Q}_n(t) + \hat{Q}_0^n(t) = \left[ \hat{X}_n(t) - c(t) \right]^+ \leq \sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(t) \right| + \left| c(t) \right|.
\]

Plugging (6.16) into (6.15) yields

\[
\sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(t) \right| \leq \sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(0) \right| + \bar{a} \int_0^t |c(u)| du + 2\bar{a} \int_0^t \sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(u) \right| du + \sum_{i \in \mathcal{I}} \left| \hat{M}_i^n(t) \right|.
\]

An application of the Gronwall’s inequality with (6.17) establishes the stochastic boundedness of \( \left\{ \sum_{i \in \mathcal{I}} \left| \hat{X}_i^n(\cdot) \right| ; n \in \mathbb{N} \right\} \). Thus for \( i \in \mathcal{I} \) the sequence \( \left\{ \hat{X}_i^n(\cdot) ; n \in \mathbb{N} \right\} \) is stochastically bounded. Then the stochastic boundedness of \( \left\{ \hat{Q}_n(\cdot) ; n \in \mathbb{N} \right\} \) and \( \left\{ \hat{Q}_0^n(\cdot) ; n \in \mathbb{N} \right\} \) follows easily by (6.16).

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

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

and

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

Applying the continuous mapping theorem (CMT) with integration in (6.19), we have

\[
\bar{D}_i^n(\cdot) \equiv \mu_i \int_0^t \bar{B}_i^n(u) du \Rightarrow \mu_i \int_0^t m_i(u) du \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.
\]

Then apply the CMT with composition in (6.20) to obtain

\[
\bar{D}_i^n(\cdot) = n^{-1/2} \left[ \Pi_i^d (n\mu_i \int_0^t \bar{B}_i^n(u) du) - n\mu_i \int_0^t \bar{B}_i^n(u) du \right]
\]

\[
= n^{-1/2} \left( \Pi_i^d \circ n\bar{D}_i^n(\cdot) - n\bar{D}_i^n(\cdot) \right) \Rightarrow W_i \left( \mu_i \int_0^t m_i(u) du \right) \quad \text{in} \quad \mathcal{D}
\]
as $n \to \infty$ where we have used $W_i$ to denote a standard Brownian motion. It is a simple exercise to show via (6.21) that

$$
\dot{D}^n(t) \equiv n^{-1/2} \left[ \bar{D}^n(t) - n \sum_{i \in \mathcal{I}} \mu_i \int_0^t \bar{b}_i^n(u) du \right] \Rightarrow W \left( \sum_{i \in \mathcal{I}} \mu_i \int_0^\cdot m_i(u) \right) \quad \text{in } \mathcal{D}
$$

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

2. Asymptotic Negligibility of $\{\dot{Q}_n^0(\cdot); n \in \mathbb{N}\}$. 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 $\{\dot{Q}_n^0(\cdot); n \in \mathbb{N}\}$, we first argue that $\dot{\Upsilon}_n^0(\cdot)$ vanishes as $n \to \infty$. For that purpose, define $\dot{Z}_n^0(\cdot) \equiv n^{-1/2} \Upsilon_n^0(\cdot).$ By (3.12),

$$
\dot{\Upsilon}_n^0(t) = \dot{Z}_n^0(t) - \sup_{u \leq t} \left\{ -\dot{Z}_n^0(u) \right\}.
$$

Combining (3.4), (3.13), (6.11) and (6.22) and some algebraic manipulation leads easily to

$$
\dot{Z}_n^0(t) = -n^{1/2} \int_0^t \lambda(u) du - \chi^n(t)
$$

where

$$
\chi^n(t) \equiv \dot{D}^n(t) + \sum_{i \in \mathcal{I}} \mu_i \int_0^t \left[ \dot{X}_i^n(u) - \dot{Q}_i^n(u) - \dot{Q}_0^n(u) \right] du + c(t).
$$

By the C-tightness of $\dot{D}^n$, and the stochastic boundedness of $\dot{X}_i^n(u), \dot{Q}_i^n$ and $\dot{Q}_0^n$, we deduce that the sequence of $\{\chi^n(\cdot); n \in \mathbb{N}\}$ is stochastically bounded and C-tight. Define

$$
u^n(t) \equiv \arg \max_{u \leq t} \left\{ -\dot{Z}_n^0(u) \right\} = \arg \max_{u \leq t} \left\{ n^{1/2} \int_0^t \lambda(u) du + \chi^n(t) \right\}.
$$

From (6.23) - (6.24), it follows

$$
\dot{\Upsilon}_0^n(t) = -n^{1/2} \int_{\nu^n(t)}^t \lambda(u) du - \chi^n(t) + \chi^n(u^n(t)) \geq 0
$$

Combining the inequality in (6.25) and the stochastic boundedness of $\chi^n(\cdot)$ allows us to conclude

$$
\sup_{t \leq T} \{ t - u^n(t) \} = O_p(n^{-1/2}).
$$
For a cadlag (right continuous with left limits) function $x(\cdot)$, define $|x|_T^* = \sup_{t \leq T} |x(t)|$. Using (6.25), we can easily deduce

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

In virtue of the established C-tightness of $X^n$,

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

Since $\epsilon$ is arbitrarily chosen, we have proven

(6.27) \quad $\hat{Y}_0^n(\cdot) \equiv n^{-1/2} Y_0^n(\cdot) \Rightarrow 0 \quad \text{in} \quad D \quad \text{as} \quad n \rightarrow \infty.$

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

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

3. State Space Collapse. By (6.4)

(6.29) \quad $n^{1/2} \int_{t-w_i^n(t)}^t \lambda_i(u)du = \hat{Q}_i^n(t) - \hat{A}_i^n(t-w_i^n(t),t) + \hat{L}_i^n(t-w_i^n(t)) + \hat{Y}_i^n(t-w_i^n(t),t)$.

Note that the right hand side is stochastically bounded owing to the stochastic boundedness of $\hat{Q}_n, \hat{A}_n$ and $\hat{R}_n$, along with the relation (6.3). By Assumption A1, the integrator $\lambda_i$ is strictly positive. Hence $\{n^{1/2}w_i^n(\cdot); n \in \mathbb{N}\}$ is stochastically bounded, for $i \in I$.

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

(6.30) \quad $\langle \hat{Y}_i^n(\cdot) \rangle = \frac{\theta_i}{n} \int_0^\infty Q_i^n(u)du \Rightarrow 0 \quad \text{in} \quad D \quad \text{as} \quad n \rightarrow \infty$

drawing upon Section 7.1 of [28]. The convergence in (6.30) implies

(6.31) \quad $\hat{R}_i^n(\cdot) - \theta_i \int_0^\cdot \hat{Q}_i^n(u)du \Rightarrow 0 \quad \text{in} \quad D \quad \text{as} \quad n \rightarrow \infty$.
by applying the Lenglart-Rebolledo inequality; see p. 30 of [22]. In view of
\[
\int_{t-w_i^n(t)}^{t} \hat{Q}_i^n(u)du \leq \left| \hat{Q}_i^n \right|_{T} |w_i^n|_T
\]
and that the random variable \( \left| \hat{Q}_i^n \right|_{T} |w_i^n|_T \) is independent of \( t \) and converges to 0 in \( \mathbb{R} \) as \( n \to \infty \), we conclude that \( \hat{R}_i^n[t - w_i^n(t), t] \) vanishes uniformly over \([0, T]\) as desired.

Next consider the term \( e_i^n \) given in (6.5). By Taylor expansion
\begin{align*}
|e_i^n(t)| &\equiv n^{1/2} \int_{t-w_i^n(t)}^{t} \lambda_i(u)du - n^{1/2} \lambda_i(t)w_i^n(t) \\
&= n^{1/2} \lambda_i(t)w_i^n(t) + n^{1/2}(w_i^n(t))^2 \lambda_i(t) + o_p\left(n^{1/2}(w_i^n(t))^2\right) - n^{1/2} \lambda_i(t)w_i^n(t) \\
&= n^{1/2}(w_i^n(t))^2 \lambda_i(t) + o_p\left(n^{1/2}(w_i^n(t))^2\right) \\
&= O_p\left(n^{1/2}(w_i^n T)^2\right)
\end{align*}
where the last equality is due to Assumption A1 which guarantees the boundedness of \( |\lambda'_i(\cdot)| \) over any compact intervals. The random variable \( n^{1/2}(w_i^n T)^2 \) is independent of time \( t \) and converges to zero as \( n \to \infty \) because \( n^{1/2}(w_i^n T)^2 \) is stochastically bounded and \( |w_i^n T| \) goes to zero as \( n \) approaches infinity. We thus establish the asymptotic negligibility of \( \tilde{K}_i^n(\cdot) \), for \( i \in \mathcal{I} \).

From here the proof follows closely that of Theorem 4.2 with simple modifications. Define the imbalance process
\begin{equation}
\Xi_i^n(\cdot) \equiv \frac{\hat{w}_i^n(\cdot)}{v_i(\cdot)} - \frac{\sum_{i \in \mathcal{I}} \lambda_i(\cdot) \hat{w}_i^n(\cdot)}{\sum_{i \in \mathcal{I}} \lambda_i(\cdot) v_i(\cdot)} \quad \text{for} \quad i \in \mathcal{I}.
\end{equation}
At each decision epoch, the HLDR rule choose a class with maximum positive imbalance and assign the head-of-line customer from that queue to the next available server.

Suppose that \( \Xi_i^n(0) \neq 0 \). Our analysis below indicates that it takes infinitesimally small time for the imbalance process \( \Xi_i^n \) to hit zero. Hence, assume without loss of generality that \( \Xi_i^n(0) = 0 \). We aim to show that, for each \( i \in \mathcal{I} \), the imbalance process \( \Xi_i^n(\cdot) \) is infinitely close to the zero function; i.e., for an arbitrary \( \epsilon > 0 \),
\begin{equation}
\mathbb{P}(|\Xi_i^n|_T > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.
\end{equation}
Define the stopping time (depending on \( \epsilon \))
\[
\bar{\tau}_i^n = \inf \{ t > 0 : |\Xi_i^n(t)| > \epsilon \}.
\]
Then to establish (6.34), it suffices to show $\mathbb{P}(\tau_i^n \leq T) \to 0$ as $n \to \infty$. Note that a positive imbalance guarantees the existence of a negative imbalance. Thus the problem further boils down to showing, for each $i \in I$,

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

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

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

With the initial condition $\Xi_i^n(0) = 0$, such a random time $\sigma_i^n$ is guaranteed to exist on the event $C$. Using the definition of $\tau_i^n$ and $\sigma_i^n$ allows us to conclude that $\Xi_i^n(t) \leq -\epsilon/4$ and $\tilde{Q}_i^n(t) > 0$ for all $t \in (\sigma_i^n, \tau_i^n]$. It is easily verifiable that the two conditions $(i) - (ii)$ described in the proof of Theorem 4.2 are satisfied for $k = i, \eta_1 = \sigma_i^n$ and $\eta_2 = \tau_i^n$.

Let $\Delta_i^n$ be the queue-imbalance process given in (6.56) with $r_i(\cdot)$ there being replaced by $\gamma^{-1}(\cdot)\lambda_i(\cdot)v_i(\cdot)$. Using (6.7) and applying union bound, we have

$$
\mathbb{P}(\tau_i^n \leq T) \leq \mathbb{P}(\Xi_i^n(\sigma_i^n) \geq -\epsilon/4, \Xi_i^n(\tau_i^n) < -\epsilon)
$$

$$
\leq \mathbb{P}\left(\Delta_i^n(\sigma_i^n) > -\lambda_i(\sigma_i^n)v_i(\sigma_i^n)\epsilon/2, \Delta_i^n(\tau_i^n) < -3\lambda_i(\tau_i^n)v_i(\tau_i^n)\epsilon/4\right)
$$

$$
+ \mathbb{P}\left(\sup_{t \leq T} | -\tilde{K}_i^n(t) + r_i(t)\Sigma_i \tilde{K}_i^n(t) | > \lambda_i v_i \epsilon/4\right).
$$

(6.36) Repeating the argument in the proof of Theorem 4.2, one can easily argue that $\tau_i^n - \sigma_i^n = O_p(n^{-1/2})$, and so

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

(6.37) By the established asymptotic negligibility of $\tilde{K}_i^n$, we have

$$
\mathbb{P}\left(\sup_{t \leq T} | -\tilde{K}_i^n(t) + r_i(t)\Sigma_i \tilde{K}_i^n(t) | > \lambda_i v_i \epsilon/4\right) \to 0 \quad \text{as} \quad n \to \infty.
$$

(6.38) Combining (6.36), (6.37) and (6.38) yields (6.35) and hence (6.34).

In view of (6.34), we conclude

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

with $\Xi_i^n$ given in (6.33). Combining the above with (6.7) yields

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

(6.39) Using the convergence-together lemma, we have

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

(6.40)
4. Diffusion Limits. Using the CMT with integration in (6.39), we obtain (6.41)

\[ \Upsilon_i^n(\cdot) \equiv \int_0^\infty \hat{Q}_i^n(u)du - \int_0^\infty \gamma(u)^{-1}v_i(u)\lambda_i(u)\hat{Q}_i^n(u)du \to 0 \quad \text{in } D \quad \text{as } n \to \infty. \]

Combining (6.12), (6.16) and (6.41) gives (6.42)

\[ \hat{X}_i^n(t) = \hat{X}_i^n(0) - \mu_i \int_0^t \hat{X}_i^n(u)du - (\theta_i - \mu_i)\Upsilon_i^n(\cdot) + \int_0^t \hat{Q}_i^n(u)du + \hat{A}_i^n(t) - \hat{D}_i^n(t) \]

\[ - \hat{Y}_i^n(t) - (\theta_i - \mu_i) \int_0^t \gamma(u)^{-1}v_i(u)\lambda_i(u) \left\{ \left[ \hat{X}_i^n(u) - c(u) \right]^+ - \hat{Q}_i^n(u) \right\} du. \]

An application of Theorem 4.1 of [28] together with (3.2), (6.21), (6.28), (6.30) and (6.41) allows us to establish the many-server heavy-traffic limit for \( \{\hat{X}_i^n(\cdot); n \in \mathbb{N}\} \):

\[ \left( \hat{X}_i^n, \dot{\hat{X}}_i^n, \hat{Q}_i^n \right) \Rightarrow \left( X_i^{(d)}, \dot{X}_i^{(d)}, Q_i^{(d)} \right) \quad \text{in } D^K \quad \text{as } n \to \infty, \]

where \( X_i^{(d)} \) satisfies the differential equation (4.5). Then apply the convergence-together lemma with (6.40) we conclude (6.43)

\[ \left( \hat{X}_i^n, \dot{\hat{X}}_i^n, \hat{Q}_i^n, \hat{Q}_i^n \right) \Rightarrow \left( X_i^{(d)}, \dot{X}_i^{(d)}, Q_i^{(d)}, Q_i^{(d)} \right) \quad \text{in } D^{2K} \]

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

5. Potential Delay Asymptotics. To establish heavy-traffic stochastic-process limits for potential delays, we follow the solution approach as in Section 3 of [35]. 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 (3.17) have proper fluid and diffusion limits. For each \( i \in \mathcal{I} \), introduce the fluid-scaled processes

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

Clearly we have (6.44)

\[ \left( \tilde{A}_i^n, \tilde{\Psi}_i^n, \tilde{R}_i^n, \tilde{Q}_i^n \right) \Rightarrow (\Lambda_i, \Lambda_i, 0, 0) \quad \text{in } D^4 \quad \text{as } n \to \infty. \]

Now define (6.45)

\[ \hat{\Psi}_i^n(\cdot) \equiv n^{-1/2} (\tilde{\Psi}_i^n(\cdot) - n\Lambda_i(\cdot)) \]

Then (6.46)

\[ \left( \hat{A}_i^n, \hat{\Psi}_i^n, \hat{R}_i^n, \hat{Q}_i^n \right) \Rightarrow \left( A_i^{(d)}, \Psi_i^{(d)}, R_i^{(d)}, Q_i^{(d)} \right) \quad \text{in } D^4 \]
as \( n \to \infty \) where \( \bar{X}_i^n, \hat{\Psi}_i^n, \hat{Q}_i^n \) and \( \bar{R}_i^n \) are given in (3.2), (6.45), (4.2) and (6.2) respectively, and

\[
R_i^{(d)}(\cdot) \equiv \theta_i \int_0^\infty Q_i^d(u)du, \quad \Psi_i^{(d)}(\cdot) \equiv Q_i^d(0) + A_i^{(d)}(\cdot) - R_i^{(d)}(\cdot)
\]

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 \( V_i^n(t) \):

\[
V_i^{n,l}(t) \leq V_i^n(t) \leq V_i^{n,u}(t)
\]

where

\[
V_i^{n,l}(t) \equiv \inf\{s \geq 0 : \Psi_i^n(t + s) + R_i^n(t + s) \geq Q_i^n(0) + A_i^n(t)\} = \inf\{s \geq 0 : \bar{\Psi}_i^n(t + s) + \bar{R}_i^n(t + s) \geq \bar{Q}_i^n(0) + \bar{A}_i^n(t)\}
\]

and

\[
V_i^{n,u}(t) \equiv \inf\{s \geq 0 : \Psi_i^n(t + s) + R_i^n(t) \geq Q_i^n(0) + A_i^n(t)\} = \inf\{s \geq 0 : \bar{\Psi}_i^n(t + s) \geq \bar{Q}_i^n(0) + \bar{A}_i^n(t) - \bar{R}_i^n(t)\}
\]

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

\[
\bar{U}^{n,l}_i(t) \equiv \inf\{s \geq 0 : \Psi_i^n(s) + \bar{R}_i^n(s) \geq \bar{Q}_i^n(0) + \bar{A}_i^n(t)\}
\]

\[
\bar{U}^{n,u}_i(t) \equiv \inf\{s \geq 0 : \Psi_i^n(s) \geq \bar{Q}_i^n(0) + \bar{A}_i^n(t) - \bar{R}_i^n(t)\}
\]

One may attempt to apply the corollary of [29] together with (6.44), (6.46) to get

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

in \( \mathcal{D} \) as \( n \to \infty \), and then use (6.47) - (6.49) to conclude the desired results. However the right-hand side of (6.49) does not satisfy the conditions of the corollary. In particular, \( \bar{Q}_i^n(0) + \bar{A}_i^n - \bar{R}_i^n \) is not necessarily nondecreasing. To resolve the problem, we use the same linear-interpolation technique as illustrated in Fig. 1 of [35]. The key is to construct a process \( \hat{V}_i^{n,u} \) such that \( \hat{V}_i^{n,u}(t) \geq V_i^{n,u}(t) \) for all \( t \geq 0 \) and

\[
n^{1/2} \hat{V}_i^{n,u} \Rightarrow \frac{Q_i^{(d)}}{\Lambda_i}.
\]

A standard sandwiching argument allows us to conclude

\[
\hat{V}_i^n(\cdot) \equiv n^{1/2} V_i^n(\cdot) \Rightarrow \frac{Q_i^{(d)}(\cdot)}{\Lambda_i(\cdot)} = v_k(t) \cdot \gamma(\cdot)^{-1} \left[ X(\cdot) - c(\cdot) \right]^+ \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty
\]
juxtapose (6.43).

Condition (6.50) holds if the error caused by these linear interpolations is asymptotically negligible. The proof of Lemma 7.1 in [35] applies here if we replace the departure process $D_n$ there with our assignment process $\Psi^n_i$.

To sum up, we have shown that

\[
(6.51) \quad \left( \hat{X}^n_1, \ldots, \hat{X}^n_K, \hat{Q}^n_1, \ldots, \hat{Q}^n_K, \hat{V}^n_1(\cdot), \ldots, \hat{V}^n_K \right) \Rightarrow \left( X^{(d)}_1, \ldots, X^{(d)}_K, Q^{(d)}_1, \ldots, Q^{(d)}_K, V^{(d)}_1, \ldots, V^{(d)}_K \right) \text{ in } D^{3K} \text{ as } n \to \infty.
\]

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 $\eta_1$ and $\eta_2$ 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 (3.9) and the non-idling condition (i), one can easily derive

\[
(6.52) \quad \sum_{i \in \mathcal{I}} A^n_i(\eta_1, \eta_2) - D^n(\eta_1, \eta_2) - \sum_{i \in \mathcal{I}} R^n_i(\eta_1, \eta_2) = s^n(\eta_1, \eta_2) + Q^n_0(\eta_1, \eta_2) + \sum_{i \in \mathcal{I}} Q^n_i(\eta_1, \eta_2)
\]

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

\[
(6.53) \quad Q^n_k(\eta_1, \eta_2) = A^n_k(\eta_1, \eta_2) - R^n_k(\eta_1, \eta_2).
\]

Combining (6.52) and (6.52) yields

\[
(6.54) \quad \sum_{i \neq k} A^n_i(\eta_1, \eta_2) - D^n(\eta_1, \eta_2) - \sum_{i \neq k} R^n_i(\eta_1, \eta_2) = s^n(\eta_1, \eta_2) + Q^n_0(\eta_1, \eta_2) + \sum_{i \neq k} Q^n_i(\eta_1, \eta_2).
\]
Now using (6.11) and (6.54) and following similar derivation used for (6.12), we have (6.55)

\[ n^{1/2} \int_{\eta_1}^{\eta_2} \lambda_k(u)du = \sum_{i \neq k} \tilde{A}_i^n(\eta_1, \eta_2) - \tilde{D}_i^n(\eta_1, \eta_2) - \sum_{i \neq k} \tilde{R}_i^n(\eta_1, \eta_2) - c(\eta_1, \eta_2) - \tilde{Q}_i^n(\eta_1, \eta_2) \]

\[ - \sum_{i \neq k} \tilde{Q}_i^n(\eta_1, \eta_2) - \sum_{i \in I} \mu_i \left( \int_{\eta_1}^{\eta_2} \tilde{X}_i^n(u)du - \int_{\eta_1}^{\eta_2} \tilde{Q}_i^n(u)du \right) \]

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) \( \sum_{i \in I} r_i(\cdot) = 1 \). Next define for each \( i \in I \) the imbalance process

\[ \Delta_i^n(\cdot) \equiv \tilde{Q}_i^n(\cdot) - r_i(\cdot)\tilde{Q}_i^n(\cdot). \]

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

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

\[ \mathbb{P} (|\Delta_i^n|_T > \epsilon) \to 0 \quad \text{as} \quad n \to \infty. \]

Define a stopping time (depending on \( \epsilon \))

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

Then to establish (6.57), it suffices to show \( \mathbb{P}(\tilde{\tau}_i^n \leq T) \to 0 \) as \( n \to \infty \). Note that \( \sum_{i \in I} \Delta_i^n(\cdot) = 0 \). Thus the problem further boils down to showing

\[ \mathbb{P}(\tau_i^n \leq T) \to 0 \quad \text{as} \quad n \to \infty \]

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

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

With the initial condition \( \Delta_i^n(0) = 0 \), such a random time \( \sigma_i^n \) 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 \( \tau_i^n \) and \( \sigma_i^n \) allows us to conclude that \( \Delta_i^n(t) \leq -\epsilon/2 \) and \( \tilde{Q}_i^n(t) > 0 \) 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 (6.55) it follows (6.58)

\[ n^{1/2} \int_{\sigma_i^n}^{\tau_i^n} \lambda_i(u)du \leq \sum_{j \neq i} \tilde{A}_j^n(\sigma_i^n, \tau_i^n) - \tilde{D}_j^n(\sigma_i^n, \tau_i^n) - \sum_{j \neq k} \tilde{R}_j^n(\sigma_i^n, \tau_i^n) - c(\sigma_i^n, \tau_i^n) - \tilde{Q}_j^n(\sigma_i^n, \tau_i^n) \]

\[ - \sum_{j \neq i} \tilde{Q}_j^n(\sigma_i^n, \tau_i^n) - \sum_{i \in I} \mu_i \left( \int_{\sigma_i^n}^{\tau_i^n} \tilde{X}_i^n(u)du - \int_{\sigma_i^n}^{\tau_i^n} \tilde{Q}_i^n(u)du \right). \]
That all terms on the right side are stochastically bounded implies the stochastic boundedness of the sequence \( \{ n^{1/2} (\tau^n_i - \sigma^n_i); n \in \mathbb{N} \} \).

Define \( \Gamma^n_i(t_1, t_2) = \tau_i(t_2) Q^n(t_2) - \tau_i(t_1) Q^n(t_1) \) and let \( \epsilon' = \epsilon/4 \), using union bound, we obtain

\[
\mathbb{P}(\tau^n_i \leq T) \leq \mathbb{P}(\Delta^n_i(\tau^n_i) < -\epsilon \Delta^n_i(\sigma^n_i) \geq -\epsilon/2),
\]

\[
\leq \mathbb{P}\left( \hat{Q}^n_i(\tau^n_i) - \hat{Q}^n_i(\sigma^n_i) - \Gamma^n_i(\sigma^n_i, \tau^n_i) < -\epsilon/2 \right),
\]

\[
\leq \mathbb{P}\left( \hat{Q}^n_i(\tau^n_i) - \hat{Q}^n_i(\sigma^n_i) - \Gamma^n_i(\sigma^n_i, \tau^n_i) < -\epsilon/2, \Gamma^n_i[\sigma^n_i, \tau^n_i] \leq \epsilon' \right) + \mathbb{P}\left( \hat{Q}^n_i(\tau^n_i) - \hat{Q}^n_i(\sigma^n_i) < -\epsilon/4 \right) + \mathbb{P}\left( \Gamma^n_i(\sigma^n_i, \tau^n_i) > \epsilon/4 \right)
\]

(6.59)

Recall that our goal is to show \( \mathbb{P}(\tau^n_i \leq T) \) goes to zero as \( n \to \infty \). To that end, we argue that both terms at the right end of (6.59) 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 \( [\sigma^n_i, \tau^n_i] \). Thus, if no customer abandoned the queue, then we must have

\[
\mathbb{P}\left( \hat{Q}^n_i(\tau^n_i) - \hat{Q}^n_i(\sigma^n_i) < -\epsilon/4 \right) = 0
\]

by the fact that \( Q^n_i \) is nondecreasing over \( [\sigma^n_i, \tau^n_i] \). With customer abandonments, we have

\[
\mathbb{P}\left( \hat{Q}^n_i(\tau^n_i) - \hat{Q}^n_i(\sigma^n_i) < -\epsilon/4 \right) \leq \mathbb{P}\left( \hat{R}^n_i(\tau^n_i) - \hat{R}^n_i(\sigma^n_i) < -\epsilon/4 \right)
\]

(6.60)

because only abandonments can cause \( Q^n_i \) to decrease over \( [\sigma^n_i, \tau^n_i] \). The following lemma plays a crucial role in the rest of proof. Its proof is deferred to the end of the section.

**Lemma 6.1.** Both \( \{ \hat{Q}^n(\cdot); n \in \mathbb{N} \} \) and \( \{ \hat{R}^n_i(\cdot); n \in \mathbb{N} \} \) are C-tight under the assumptions of Theorem 4.2.

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

\[
\mathbb{P}\left( \hat{R}^n_i(\tau^n_i) - \hat{R}^n_i(\sigma^n_i) < -\epsilon/4 \right) \to 0 \quad \text{as} \quad n \to \infty.
\]

Combining the above with (6.60) allows us to conclude that

\[
\mathbb{P}\left( \hat{Q}^n_i(\tau^n_i) - \hat{Q}^n_i(\sigma^n_i) < -\epsilon/4 \right) \to 0 \quad \text{as} \quad n \to \infty.
\]

(6.61)

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

\[
\mathbb{P}\left( \Gamma^n_i(\sigma^n_i, \tau^n_i) > \epsilon/4 \right) \to 0 \quad \text{as} \quad n \to \infty.
\]

(6.62)
Combining (6.59), (6.61) and (6.62) yields
\[ \mathbb{P}(\tau^n_i \leq T) \to 0 \quad \text{as} \quad n \to \infty \]
which in turn implies
\[ \Delta^n_i(\cdot) \equiv \hat{Q}^n_i(\cdot) - r_i(\cdot)\hat{Q}^n(\cdot) \Rightarrow 0 \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \]
for all \( i \in \mathcal{I} \). The convergence can be strengthened to joint convergence by the fact that all the limits are deterministic process. This is again a SSC result. Repeating step 4 - 5 in the proof the HLDR rule as in \S 6 leads us to the conclusion of Theorem 4.2.

**Proof of Lemma 6.1.** By (6.16), \( \{\hat{Q}^n(\cdot); n \in \mathbb{N}\} \) is C-tight if \( \{\hat{X}^n_i; n \in \mathbb{N}\} \) is C-tight for \( i \in \mathcal{I} \). The latter holds true if the martingales \( \hat{A}^n_i, \hat{B}^n_i \) and \( \hat{Y}^n_i \) are C-tight, owing to (6.12) and the established stochastic boundedness of \( \hat{X}^n_i \) and \( \hat{Q}^n_i \). But \( \hat{A}^n_i, \hat{B}^n_i \) and \( \hat{Y}^n_i \) are C-tight due to (3.2), (6.21) and (6.30). Hence \( \{\hat{Q}^n(\cdot); n \in \mathbb{N}\} \) is C-tight. The C-tightness of \( \{\hat{R}^n_i(\cdot); n \in \mathbb{N}\} \) follows from (6.31) and the stochastic boundedness of \( \{\hat{Q}^n_i(\cdot); n \in \mathbb{N}\} \) drawing upon the stochastic boundedness of \( \{\hat{Q}^n(\cdot); n \in \mathbb{N}\} \).

7. Proofs for Asymptotic Feasibility and Optimality. The following lemma is the crucial ingredient in the proof of asymptotic optimality.

**Lemma 7.1.** (comparison principle for piecewise-linear diffusions) Consider the following two stochastic integral equations:
\[ \begin{align*}
X(t) &= X(0) - \beta_1 \int_0^t X(u)du + \beta_2 \int_0^t [X(u) - c(u)]^+ du + \int_0^t \sigma(u)dW(u), \\
X'(t) &= X'(0) - \beta_1 \int_0^t X'(u)du + \beta_2 \int_0^t [X'(u) - c'(u)]^+ du + \int_0^t \sigma(u)dW'(u),
\end{align*} \]
where \( X(0) \overset{a.s.}{=} X'(0) \) and \( \beta_1 \geq \beta_2 \geq 0 \). Let \( Q(t) \equiv [X(t) - c(t)]^+ \) and \( Q'(t) \equiv [X'(t) - c'(t)]^+ \). If \( c' \leq c \) and \( c'(t) < c(t) \) for some \( t \geq 0 \), then \( \mathbb{E}[Q'(t)] - \mathbb{E}[Q(t)] > 0 \).

**Proof of Lemma 7.1.** Note that the expectation of a random variable depends only on its probability distribution. We can thus define \( X \) and \( X' \) on the same probability space where \( W'(t) \equiv W(t) \). The idea is to couple two diffusion processes in such a way that they agree as often as possible. A direct application of Theorem 1.3 of [42] allows us to conclude that \( X(t) \leq X'(t) \) almost surely. Let \( A_t \) be the event \( \{X(t) \geq c(t)\} \). Then \( Q'(t) - Q(t) = [X'(t) - c'(t)]^+ \geq 0 \) on \( A^c_t \), and thus \( \mathbb{E}[Q'(t)1_{A^c_t}] \geq \mathbb{E}[Q(t)1_{A^c_t}] \).
Similarly,
\[ Q'(t) - Q(t) = X'(t) - X(t) + c'(t) - c(t) \quad \text{on} \quad A_t. \]

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

Combining the above yields
\[ \mathbb{E}[Q'(t)] - \mathbb{E}[Q(t)] = \mathbb{E}[Q'(t)1_{A_t}] - \mathbb{E}[Q(t)1_{A_t}] + \mathbb{E}[Q'(t)1_{A_t}] - \mathbb{E}[Q(t)1_{A_t}] > 0. \]

This completes the proof of Lemma 7.1.

**Proof of Theorem 4.3.** Towards proving asymptotic feasibility, we apply Fatou’s lemma and Theorem 4.1 to conclude
\[ \limsup_{n \to \infty} \mathbb{E}[V^n_i(t)/T^n_i(t)] \leq \mathbb{E}[\lambda(t)/\vartheta(t)] = \mathbb{E}[Q^{(d)}(t)/\vartheta(t)] = 1, \quad \text{for} \quad i \in \mathcal{I}. \]

To prove asymptotic optimality, suppose by way of contradiction that condition (3.20) is violated for \((s^n, \pi^n)\) at time \(t\). Using Fatou’s lemma, we get
\[ \liminf_{n \to \infty} \mathbb{E}[V^n_i(t)/T^n_i(t)] \geq \mathbb{E}[Q^{(d)}(t)/\vartheta(t)] > \mathbb{E}[Q^{(d)}(t)/\vartheta(t)] = 1, \]

where the second inequality follows by applying Lemma 7.1 with \(\beta_1 = \mu, \beta_2 = \mu - \theta\) and \(\sigma(t) = \sqrt{\lambda(t) + \mu m(t)}\).

**Proof of Theorem 4.4.** To establish asymptotic feasibility, apply Portmanteau theorem and Theorem 4.1 to get
\[ \limsup_{n \to \infty} \mathbb{P}(V^n_i(t)/T^n_i(t) \geq 1 + \epsilon) \leq \mathbb{P}(V^{(d)}_i(t)/T_i(t) \geq 1 + \epsilon) \]
\[ = \mathbb{P}(Q^{(d)}(t)/\vartheta(t) \geq 1 + \epsilon) \leq \alpha, \quad \text{for} \quad i \in \mathcal{I}. \]

**Proof of Theorem 4.5.** The proof follows closely the steps in the proof of Theorem 4.3 and Theorem 4.4. By Fatou’s lemma,
\[ \limsup_{n \to \infty} \mathbb{E}[Q^n(t)/\varphi^n(t)] \leq \mathbb{E}[Q^{(d)}(t)/\varphi(t)] = 1. \]

By Portmanteau theorem, we have, for \(i = 1, \ldots, K - 1,\)
\[ \limsup_{n \to \infty} \mathbb{P}(V^n_i(t)/T^n_i(t) \geq 1 + \epsilon) \leq \mathbb{P}(V^{(d)}_i(t)/T_i(t) \geq 1 + \epsilon) \]
\[ = \mathbb{P}(X^{(d)}(t) - c^*(t) \geq (1 + \epsilon) \cdot x(t)) \leq \alpha. \]

Towards proving asymptotic optimality, suppose by way of contradiction that condition (3.25) is violated for \((s^n, \pi^n)\) at time \(t\). Using Fatou’s lemma again, we obtain
\[ \liminf_{n \to \infty} \mathbb{E}[Q^n(t)/\varphi^n(t)] \geq \mathbb{E}[Q^{(d)}(t)/\varphi(t)] > \mathbb{E}[Q^{(d)}(t)/\varphi(t)] = 1, \]

where the second inequality follows by applying Lemma 7.1.
Acknowledgment. The authors thank the anonymous reviewers and the editors whose careful reviews have led to an essentially rewritten and much improved paper. The authors are grateful to Yunan Liu and Kyle Hovey for a careful reading and helpful constructive comments. This research was supported by NSF grant CMMI 1634133.

REFERENCES


X. SUN AND W. WHITT

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