MATHEMATICS OF OPERATIONS RESEARCH
Vol. 34, No. 2, May 2009, pp. 363–396
ISSN 0364-765X | EISSN 1526-5471 | 09 | 3402 | 0363

DOI 10.1287/moor.1080.0366

Queue-and-Idleness-Ratio Controls in Many-Server Service Systems

Itay Gurvich
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
i-gurvich@kellogg.northwestern.edu

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027,
ww2040@columbia.edu

Motivated by call centers, we study large-scale service systems with multiple customer classes and multiple agent pools, each with many agents. We propose a family of routing rules called *queue-and-idleness-ratio* (QIR) rules. A newly available agent next serves the customer from the head of the queue of the class (from among those he is eligible to serve) whose queue length most exceeds a specified state-dependent proportion of the total queue length. An arriving customer is routed to the agent pool whose idleness most exceeds a specified state-dependent proportion of the total idleness. We identify regularity conditions on the network structure and system parameters under which QIR produces an important *state-space collapse* (SSC) result in the quality-and-efficiency-driven (QED) many-server heavy-traffic limiting regime. The SSC result is applied here to prove stochastic-process limits and in subsequent papers to solve important staffing and control problems for large-scale service systems.

Key words: heavy-traffic; Halfin-Whitt regime; QED regime; state-space collapse; diffusion limits; control of queueing systems
MSC2000 subject classification: Primary: 60K25; secondary: 60K30, 60F17, 90B15, 90B22
OR/MS subject classification: Primary: queues; secondary: diffusion models; limit theorems
History: Received October 21, 2007; revised February 10, 2008, August 27, 2008, and October 26, 2008. Published online in *Articles in Advance* April 3, 2009.

1. Introduction.

1.1. Parallel-server systems.

In this paper we focus on a family of multiclass queueing networks called *parallel-server systems* (PSSs). In a PSS, there are multiple classes of customers (or jobs) and multiple pools of agents (or servers), as depicted in Figure 1. Unlike many queueing networks, in a PSS customers receive at most one service; they depart from the system after a single service completion.

Customers from a common customer class are homogeneous, as are agents within the same service pool; they have common parameters. The agents from each service pool are allowed to serve customers from some designated subset of the customer classes. The allowed routing is depicted by a routing graph, as shown in Figure 1. The nodes of this graph are the queues and the service pools. There is a set $\mathcal{I} = \{1, \ldots, I\}$ of customer classes and a set $\mathcal{J} = \{1, \ldots, J\}$ of service pools. An arc connecting customer-class i to service-pool j indicates that agents from pool j are permitted to serve customers from class i.

We will be considering Markovian PSSs. Customers arrive exogenously according to independent Poisson processes, with one for each class. The class i arrival rate is λ_i. The aggregate arrival rate is $\lambda := \sum_{i \in \mathcal{I}} \lambda_i$. Customers from each class enter service in order of arrival. Pool j contains N_j agents. The service times are mutually independent exponential random variables. When a class i customer is served by a server from pool j, the service rate is $\mu_{i,j}$. Whenever type j agents do not have the required skill to serve class i customers, $\mu_{i,j} = 0$. For some of the results here, we will assume, in addition, that $\mu_{i,j}$ depends only on i, or only on j, or on neither. When customers cannot enter service immediately upon arrival, they go to the end of a queue. Waiting customers from each queue may elect to abandon if they have not yet started service. The times different customers are willing to wait before abandoning are also mutually independent exponential random variables, having mean $1/\theta_i$ for class i. Unlimited patience is obtained by setting $\theta_i = 0$.

PSSs are used to model various manufacturing and service systems, especially call centers; see Gans et al. [14]. Accordingly, we use the terminology *customers* and *agents* instead of *jobs* and *servers*. In the call-center literature, a PSS is often called a call center model with *skill-based routing*. Motivated by that application, we consider many-server PSSs with a large number of agents in each pool.

1.2. The queue-and-idleness-ratio (QIR) rule.

It remains to specify how the customers are assigned to agents. Specifically, we need to specify the *routing rule*, indicating what to do upon customer arrival, and the *scheduling rule*, indicating what to do upon service completion. Our proposed queue-and-idleness-ratio (QIR) rule does both, but in a flexible way that depends on additional parameters that remain to be specified.

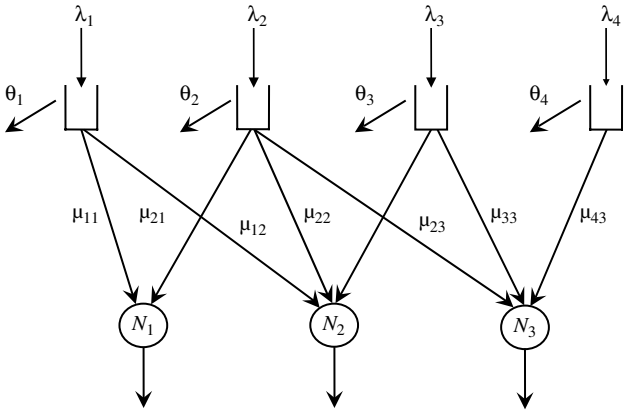

FIGURE 1. A PSS and its corresponding routing graph.

We now explain the QIR control. It uses two vector functions $p(\cdot) := (p_1(\cdot), \ldots, p_I(\cdot))$ and $v(\cdot) := (v_1(\cdot), \ldots, v_J(\cdot))$, which we call *ratio functions*, and which are assumed to satisfy $p_i(x) \geq 0$, $i \in \mathcal{I}$, $v_j(x) \geq 0$, $j \in \mathcal{J}$, $\sum_{i\in\mathcal{I}} p_i(x) = 1$ for all $x \in \mathbb{R}_+$, and $\sum_{j\in\mathcal{J}} v_j(x) = 1$ for all $x \in \mathbb{R}_+$.

Let $\hat{Q}_i^\lambda(t)$ and $\hat{I}_j^\lambda(t)$ be, respectively, the properly scaled class i queue length and type j number of idle agents at time t in the λth system (see Definition 2.1, and the rest of §2.1 for a fuller explanation). Let $\hat{Q}_\Sigma^\lambda(t) := \sum_{i\in\mathcal{I}} \hat{Q}_i^\lambda(t)$ and $\hat{I}_\Sigma^\lambda(t) := \sum_{j\in\mathcal{J}} \hat{I}_j^\lambda(t)$ be the corresponding aggregate quantities. The QIR rule then aims to achieve

$$\hat{Q}_i^\lambda(t) \approx \hat{Q}_\Sigma^\lambda(t) p_i(\hat{Q}_\Sigma^\lambda(t)) \quad \text{and} \quad \hat{I}_j^\lambda(t) \approx \hat{I}_\Sigma^\lambda(t) p_i(\hat{I}_\Sigma^\lambda(t)), \quad i \in \mathcal{I}, \quad j \in \mathcal{J}.$$

That is, it aims to set the scaled queue length of each class and the scaled idleness at each pool so that, when we divide by the corresponding aggregate queue length and aggregate idleness, respectively, they are specified state-dependent ratios. Roughly, QIR will achieve this goal by letting an available agent at time t serve the customer class with the greatest *queue imbalance*, $\hat{Q}_i^\lambda(t) - \hat{Q}_\Sigma^\lambda(t) p_i(\hat{Q}_\Sigma^\lambda(t))$, and by routing an arriving customer at time t to the agent pool with the greatest *idleness imbalance*, $\hat{I}_j^\lambda(t) - \hat{I}_\Sigma^\lambda(t) p_i(\hat{I}_\Sigma^\lambda(t))$. The aim of QIR is to drive these imbalances toward 0. We will show that this goal is achieved asymptotically; i.e., we achieve *asymptotic proportionality*. In fact, the precise definition of the queue and idleness imbalances will be slightly more intricate; see Definition 2.3.

However, even with this added complexity, QIR has important simplicity, because at each decision epoch—customer arrival or service completion—the decision rule uses only *local* and *aggregate* idleness and queue-length information. An available agent will need to know only the queue length of the customer classes that he can serve and the overall number of customers in the system to choose which customer to serve next. In particular, he will neither need to know the queue length of all other customer classes nor will he need to know the detailed occupancy information that specifies the number of type j agents giving service to class i customers. Moreover, the QIR rule does not depend on the model parameters.

1.3. The QED many-server heavy-traffic limiting regime. To establish asymptotic-proportionality results for many-server PSSs, we work in the QED many-server heavy-traffic limiting regime, first formalized by Halfin and Whitt [19] for the $M/M/N$ queue. For the $M/M/N$ model, the QED regime is obtained by letting the aggregate arrival rate λ and the number N of agents grow indefinitely, while holding the service rate fixed, so that the utilization, $\rho := \lambda/N\mu$, approaches its critical value 1 in an appropriate manner. Specifically, considering a sequence of systems indexed by the aggregate arrival rate λ, it is assumed that

$$\sqrt{\lambda}(1 - \rho^\lambda) \to \beta \quad \text{as } \lambda \to \infty, \tag{1}$$

where $-\infty < \beta < \infty$. This QED regime is now widely accepted as the most useful heavy-traffic limiting regime for many-server systems. It is to be contrasted with the *conventional* heavy-traffic regime, in which the number of agents (servers) is held fixed while letting ρ approach one. Halfin and Whitt showed that the limit in (1) holds for the $M/M/N$ model if and only if the steady-state probability that a new arrival must wait before beginning service approaches a limit strictly between 0 and 1. Given the appropriate heavy-traffic condition, as in (1), we then seek to obtain limits for the properly scaled and normalized processes associated with queue-length, waiting time, etc. However, we point out that the situation is not nearly as straightforward for the more general PSSs considered here. In particular, we will use mathematical programs to specify the QED regime in the present context; see §2.

1.4. State-space collapse (SSC). Like many other multiclass queueing networks, PSSs are quite complex and challenging to analyze. To address this complexity for queueing networks, recent research has established special asymptotic techniques to obtain more elementary descriptions of these complex models. The seminal papers of Bramson [6] and Williams [31] provide a magnificent example of such a simplification by showing how complex multiclass queueing networks can be approximated by more elementary diffusion models, known as semimartingale reflected Brownian motions (SRBMs). A key step in this complexity reduction is a *state-space collapse* (*SSC*) result, based on *hydrodynamic limits*, that connects a high-dimensional queue-length process with a lower-dimensional workload process, asymptotically, in the heavy-traffic limit.

These initial papers by Bramson and Williams focus on open queueing networks in the conventional heavy-traffic limiting regime, instead of PSSs in the QED limiting regime, but it is now clear that SSC can serve as a key step in providing simple solutions for many complex stochastic systems. Indeed, SSC results have previously been established in the QED regime. Paralleling the history for the conventional-heavy-traffic regime, the first SSC results for PSSs were based on ad-hoc proofs for specific models and controls. Examples include Armony [1], Gurvich et al. [18], and Armony and Maglaras [2, 3]. Recently, Dai and Tezcan [12] made the important step of extending Bramson's framework to the case of the many-server heavy-traffic regime. They applied their framework to several settings; see Tezcan [28] and Dai and Tezcan [11, 10].

We contribute to this SSC literature by establishing that our proposed QIR controls produce an important SSC for a large class of PSSs in the QED limiting regime. The SSC occurring here is due to the asymptotic proportionality mentioned above. It would be natural to apply Dai and Tezcan [12] for this purpose, but their results cannot be directly applied to the QIR control because of the general structure of the ratio functions. Nevertheless, the key ideas in Dai and Tezcan [12], which in turn build on Bramson [6], lie at the heart of the SSC proofs here.

As a consequence of our SSC results, asymptotically, the multidimensional queue-length and idleness processes will be determined by a process of a lower dimension that contains, in the worst case, only the number of customers in the system of each class. This notion of SSC is somewhat weaker than often seen in the conventional heavy-traffic regime, where the multidimensional queue-length process actually collapses into a one-dimensional process; e.g., see Mandelbaum and Stolyar [22]. However, we too obtain a strong one-dimensional form of SSC in the many-server setting when the service rates are pool-dependent; see Theorem 4.1.

In closing this discussion of SSC, we mention the important paper by Atar [4], which focuses on establishing asymptotic optimality in heavy traffic for PSSs in the QED regime. In §5 we will show that, for special (tree) networks, QIR is equivalent to Atar's control. In these cases the SSC result follows from the analysis in Atar [4].

1.5. The value of QIR and SSC. We illustrate the value of SSC by applying it to establish stochastic-process limits for the basic stochastic process of interest with an arbitrary QIR control, under regularity conditions, in §4. As usual, the limits are diffusion processes. However, the SSC results are important even without subsequent stochastic-process limits. We use SSC and the associated asymptotic proportionality to establish desired properties of QIR controls, even for cases in which a stochastic-process limit remains to be established. In particular, we demonstrate the power of QIR for controlling PSSs in subsequent papers: In Gurvich and Whitt [16], we examine a central question in call-center operations, namely how to jointly determine the *design*, *staffing*, and *routing* (real-time control) of call-centers with multiple customer classes and agent pools. These decisions need to be made so as to minimize labor-related costs while maintaining predetermined quality-of-service (QoS) constraints. The control component of our proposed solution is a special case of QIR called *fixed-queue-ratio* (FQR) routing.

In Gurvich and Whitt [17] we show that, as long as the service-rates are pool dependent, i.e, when $\mu_{i,j} = \mu_j$ for all i and j, QIR with appropriately chosen parameters is asymptotically optimal with respect to convex holding costs in the QED regime. Moreover, we show that in certain cases QIR is partially equivalent to the well-known generalized-$c\mu$ (G$c\mu$) rule. By doing this, we are able to partially extend the important results of Mandelbaum and Stolyar [22] in the conventional heavy-traffic regime to the many-server QED regime.

1.6. Organization of the paper. We elaborate on the model and define the QIR controls in §2. We state our main SSC result (Theorem 3.1) in §3. We apply SSC to establish stochastic-process limits in §4. In §5 we relate QIR to Atar's control in Atar [4] for a class of tree networks, and use that connection to provide an alternative proof of SSC. We provide proofs for the results in §§3, and 4, 5, respectively, in §§6, 7, and 8.

2. The model. We now elaborate on the model description given in §1. The possible routing for a PSS has a natural representation as a *bipartite graph* with vertices $V = \mathcal{J} \cup \mathcal{I}$. The only edges in the graph connect customer classes to agent pools: $E := \{(i, j) \in \mathcal{I} \times \mathcal{J}\colon \mu_{i,j} > 0\}$. An edge (i, j) is present in the routing graph if class i customers can be served by type j agents. Actually, we may choose to use only a subset of the edges of E. The eventual subgraph will depend on the solution of a linear program (LP). Figure 1 depicts a PSS routing graph.

Let $Q_i(t)$ be the queue length of class i customers, and let $I_j(t)$ be the number of idle agents in pool j, at time t. The corresponding aggregate quantities are $Q_\Sigma(t) := \sum_{i\in\mathcal{I}} Q_i(t)$ and $I_\Sigma(t) := \sum_{j\in\mathcal{J}} I_j(t)$. Let $Z_{i,j}(t)$ be the number of type j agents busy giving service to class i customers, and let $Z_j(t) := \sum_{i\in\mathcal{I}} Z_{i,j}(t)$ be the number of busy agents at pool j at time t. Consequently, $I_j(t) = N_j - Z_j(t)$. The overall number of class i customers present in the system at time t is then given by $X_i^\lambda(t) := Q_i(t) + \sum_{j\in\mathcal{J}} Z_{i,j}(t)$. Finally, let $X_\Sigma^\lambda(t)$ be the overall number of customers in the system (in service and in queue), i.e.,

$$X_\Sigma(t) := \sum_{i=1}^{I} X_i(t) = \sum_{i=1}^{I}\left(Q_i(t) + \sum_{j=1}^{J} Z_{i,j}(t)\right).$$

To construct the heavy-traffic framework, we consider a sequence of systems indexed by λ. We add the superscript λ to express the dependence on the index. Thus, $Q_i^\lambda(t)$ stands for the class i queue length at time t in the λth system. The routing graph and service rates $\{\mu_{i,j},\ i \in \mathcal{I},\ j \in \mathcal{J}\}$ do not change with λ.

Notational conventions. For an integer $d > 0$, let $D^d := D^d[0, \infty)$ be the space of all RCLL (right continuous with left limits) functions with values in d-dimensional Euclidean space $\mathbb{R}^d$, equipped with the Skorohod J_1 metric; e.g., see Whitt [30]. We will often use convergence in probability (commonly denoted by $\xrightarrow{P}$) for a sequence of random elements of D^d to the zero function (the function in D^d that is identically 0), here denoted by 0. However, convergence in probability to a deterministic limit is equivalent to convergence in distribution to that limit, denoted by $\Rightarrow$. As a consequence, for a family of stochastic processes, $\{Y^\lambda\colon \lambda > 0\}$ with sample paths in D^d, we will be showing that $Y^\lambda \Rightarrow 0$ in D^d as $\lambda \to \infty$; we will write $Y^\lambda(t) \Rightarrow 0$ in D^d to emphasize that we are considering processes in D^d instead of stationary distributions on $\mathbb{R}$.

Because the limit processes we consider are either the deterministic zero function or diffusion processes, the limit process has continuous sample paths, so the notion of convergence on the underlying function space D^d coincides with uniform convergence on closed bounded intervals. To express that, for a vector-valued process $B(t)$ in D^d, let $\|B\|_{s,T}^* = \sup_{s\le t\le T} \|B(t)\|$, where $\|B(t)\| = \sum_{k=1}^{J} |B_k(t)|$. These are defined similarly for a process $B(t)$ in $D^{d\times m}[0, \infty)$, where $\|B(t)\| = \sum_{k=1}^{d}\sum_{l=1}^{m} |B_{k,l}(t)|$. We omit the subscripts whenever we refer to vectors or to vector processes. For example, N^λ and $X^\lambda(t)$ will stand, respectively, for the vector in $\mathbb{Z}^J$ whose components are N_j^λ and the vector in $\mathbb{R}^I$ whose components are $X_i^\lambda(t)$. Using this notation, we will say that a sequence $\{x^\lambda\colon \lambda > 0\}$ of processes with sample paths in D^d satisfy that $x^\lambda = o_P(1)$ if $\|x^\lambda\|_T^* \Rightarrow 0$ as $\lambda \to \infty$.

We will also consider a weaker notion of convergence, using the space $D_-^d := D^d(0, \infty)$, where the domain is treated as open at the left instead of closed. We again let convergence (to continuous limits) be characterized by uniform convergence over bounded intervals. The restriction to the domain $(0, \infty)$ means that we exclude uniform convergence for intervals of the form $[0, b]$. We have $Y^\lambda(t) \Rightarrow 0$ in $D^d(0, \infty)$ if and only if, for each $0 < s < T < \infty$, $\|Y^\lambda\|_{s,T}^* \Rightarrow 0$.

Finally, we mention conventions for vector products: For $x, y \in \mathbb{R}^d$, xy is the component-wise product ($(xy)_i = x_i y_i$). Whenever $x \in \mathbb{R}^d$ but $y \in \mathbb{R}$, xy should be interpreted so that $(xy)_i = x_i y$. Finally, $x \cdot y$ is the scalar product and e_j is the unit vector with 1 in the jth place and 0 elsewhere.

2.1. Heavy-traffic conditions and the fundamental mathematical program. We need to make two assumptions to put our PSS into the QED many-server heavy-traffic limiting regime. The first is a natural generalization of the condition (1), but that is not enough.

Assumption 2.1 (Heavy-Traffic Conditions). *There are constants $a_i > 0$, $i \in \mathcal{I}$, and $\nu_j > 0$, $j \in \mathcal{J}$, such that, as $\lambda \to \infty$, $\lambda_i/\lambda \to a_i > 0$, $i \in \mathcal{I}$, and $N_j^\lambda/\lambda \to \nu_j > 0$, $j \in \mathcal{J}$. Also, there exist constants $\xi_i \in (-\infty, \infty)$, $i \in \mathcal{I}$ and $\gamma_j \in (-\infty, \infty)$, $j \in \mathcal{J}$, such that, as $\lambda \to \infty$,*

$$\frac{\lambda_i - a_i\lambda}{\sqrt{\lambda}} \to \xi_i \qquad \text{and} \qquad \frac{N_j^\lambda - \nu_j\lambda}{\sqrt{\lambda}} \to \gamma_j.$$

The second QED assumption concerns a mathematical program (in the "fluid scale"). We assume that the constants of Assumption 2.1 are specified. The *fundamental mathematical program* is:

$$\begin{aligned} \text{Minimize} \quad & \rho \\ \text{Subject to:} \quad & \sum_{j\in J} \mu_{i,j}\nu_j x_{i,j} = a_i, \quad i \in \mathcal{I}, \\ & \sum_{i\in\mathcal{I}} x_{i,j} \le \rho, \quad j \in \mathcal{J}, \\ & \rho \ge 0,\ x_{i,j} \ge 0, \quad i \in \mathcal{I},\ \ j \in \mathcal{J}. \end{aligned} \tag{2}$$

Because the constants of Assumption 2.1 are specified, the mathematical program in (2) is an LP. An optimal solution is a vector $(x, \rho) \in \mathbb{R}_+^{I\times J} \times \mathbb{R}_+$ that is feasible for (2) such that no other feasible solution has a lower ρ value. We next impose a critical-loading assumption.

Assumption 2.2 (Critical Loading). *For any optimal solution* $(\bar{x}, \bar{\rho})$ *of* (2), $\sum_{i\in\mathcal{I}} \bar{x}_{i,j} = 1$ *for all* $j \in \mathcal{J}$, *so that* $\bar{\rho} = 1$.

The fact that Assumption 2.2 applies to *any* optimal solution is important. From a practical perspective, this is a natural restriction. If there exists a solution in which some of the agent pools are underloaded, that is, $\sum_{i\in\mathcal{I}} x_{i,j} < 1$ for some $j \in \mathcal{J}$, then it makes sense to decrease the staffing levels. From a mathematical perspective, Assumption 2.2 is imposed to prevent the fluid from drifting into an underloaded state, away from the QED regime.

With Assumption 2.2, it suffices to denote the selected optimal solution by its x coordinate, because the value of ρ will necessarily be 1. Hence, fix an optimal solution $\bar{x}$ for (2) satisfying Assumption 2.2. We now indicate how the chosen optimal solution $\bar{x}$ is used. Because we intend to use QIR for the routing, we do not use $\bar{x}$ for the routing, but $\bar{x}$ plays a critical role in the design. Specifically, *we omit all edges with* $\bar{x}_{i,j} = 0$ *from the network routing graph*; i.e., we do not allow any class i customers to be routed to pool j if $\bar{x}_{i,j} = 0$. *The routing graph includes all edges with* $\bar{x}_{i,j} > 0$*;* we stipulate that the routing graph is $\{(i, j) \in \mathcal{I} \times \mathcal{J}\colon \bar{x}_{i,j} > 0\}$. If, a priori, pool j is unable to serve class i or if we do not want pool j to serve class i, then we enforce that by imposing the constraint $x_{i,j} \le 0$ in the mathematical program.

The routing graph determined by $\bar{x}$ is closely linked to the dynamic control. Indeed, our SSC results for QIR will depend on characteristics of the routing graph. With that in mind, we note that the LP may well have multiple optimal solutions, and different optimal solutions may thus lead to different routing graphs, according to the construction above.

To facilitate the following discussion, we need to be clear about the network graphs under consideration: Our network graphs are simple undirected bipartite graphs, i.e., with at most one edge connecting any two nodes, and with edges only between a customer class and an agent pool. Beyond these basic features, we first require that the graph is connected; i.e., there exists a path between every two nodes in the graph.

Assumption 2.3 (Connected Routing Graph). *The selected optimal solution* $(\bar{x})$ *for* (2) *produces a routing graph determined by the edges* $\mathcal{E}(\bar{x}) := \{(i, j) \in \mathcal{I} \times \mathcal{J}\colon \bar{x}_{i,j} > 0\}$ *that is connected.*

Assumptions 2.1–2.3 are assumed to hold throughout the rest of the paper. The connected-graph assumption is crucial for the ability to instantaneously balance the system asymptotically; see §2.7 of Atar [4] for elaboration. We will actually need additional structure beyond connectedness. Connected graphs can be cyclic or acyclic (a graph is acyclic if there is a unique path between each pair of nodes). This distinction is important, because QIR works well with cyclic networks only when certain parametric conditions hold; see Theorem 3.1 and Remark 3.1.

Given an optimal solution $(\bar{x})$ to (2), $J(i)(\bar{x})$ for $i \in \mathcal{I}$ is defined to be the set of agent pools connected to customer class i; i.e., $J(i)(\bar{x}) = \{j \in \mathcal{J}\colon \bar{x}_{i,j} > 0\}$. Analogously, we let $I(j)(\bar{x})$ for $j \in \mathcal{J}$ be the set of customer classes connected to agent pool j; i.e., $I(j)(\bar{x}) = \{i \in \mathcal{I}\colon \bar{x}_{i,j} > 0\}$. We will often omit the argument $\bar{x}$ when it is clear from the context.

Definition 2.1 (Scaled and Normalized Processes). Fix an optimal solution $\bar{x}$ for (2) for which the edges in $\mathcal{E}(\bar{x})$ induce a connected routing graph. Then, we define the following scaled processes:

$$\hat{X}^\lambda_\Sigma(t) := \frac{X^\lambda_\Sigma(t) - N^\lambda_\Sigma}{\sqrt{\lambda}}; \qquad \hat{I}^\lambda_\Sigma(t) := \frac{I^\lambda_\Sigma(t)}{\sqrt{\lambda}}; \qquad \hat{X}^\lambda_i(t) := \frac{X^\lambda_i(t) - \sum_{j\in\mathcal{J}} \bar{x}_{i,j} N^\lambda_j}{\sqrt{\lambda}}, \quad i \in \mathcal{I};$$

$$\hat{Q}^\lambda_i(t) := \frac{Q^\lambda_i(t)}{\sqrt{\lambda}}, \quad i \in \mathcal{I}; \qquad \hat{I}^\lambda_j(t) := \frac{I^\lambda_j(t)}{\sqrt{\lambda}}, \quad j \in \mathcal{J}; \qquad \hat{Z}^\lambda_{i,j}(t) := \frac{Z^\lambda_{i,j}(t) - \bar{x}_{i,j} N^\lambda_j}{\sqrt{\lambda}}, \quad (i, j) \in \mathcal{I} \times \mathcal{J}.$$

Once we fix a solution $\bar{x}$, which in turn fixes the corresponding routing graph $\mathscr{E}(\bar{x})$, we assume that $Z^{\lambda}_{i,j}(t) \equiv 0$ for all $(i,j) \notin \mathscr{E}(\bar{x})$. This can be relaxed to allow for initial conditions, $Z^{\lambda}_{i,j}(0)$, that are not consistent with the routing graph but here we prohibit this option.

2.2. Definition of QIR. We will impose a smoothness condition on our ratio functions. For that purpose, following convention, we say that an $\mathbb{R}^m$-valued function f on a subset S of $\mathbb{R}^k$ is *locally Hölder continuous* with exponent $\alpha > 0$ if, for every compact subset $K \subset S$, there exists a constant C_K such that $\|f(x) - f(y)\| \le C_K \|x - y\|^{\alpha}$ for all $x, y \in K$, where $\|\cdot\|$ is a chosen norm inducing the usual Euclidean topology, which we take to be the $\mathbb{L}^1$ norm: $\|x\| := \sum_i |x_i|$.

Definition 2.2 (An Admissible State-Dependent Ratio Function). For an integer $d > 0$, a vector-valued function $r \equiv r(\cdot)\colon \mathbb{R}_+ \mapsto \mathbb{R}^d_+$, is an admissible state-dependent ratio function if $\sum_{k=1}^{d} r_k(x) = 1$ for all $x \in \mathbb{R}_+$ and if every component $r_k\colon \mathbb{R}_+ \mapsto \mathbb{R}_+$ is locally Hölder continuous on the open interval $(0, \infty)$ for some exponent $\alpha_k > 0$.

For the following definition, we assume that an optimal solution $\bar{x}$ for (2) is fixed and the routing graph $\mathscr{E}(\bar{x})$ is used. We omit the argument $\bar{x}$ from the notation.

Definition 2.3 (QIR for Admissible State-Dependent Ratio Functions). Given two admissible state-dependent ratio functions v and p, QIR is defined as follows.

- *Upon arrival of a class i customer* at time t, the customer will be routed to an available agent in pool j^*, where

$$j^* := j^*(t) \in \operatorname*{arg\,max}_{j \in J(i),\, \hat{I}^{\lambda}_j(t) > 0} \left\{ \hat{I}^{\lambda}_j(t) - [\hat{X}^{\lambda}_{\Sigma}(t)]^{-} v_j([\hat{X}^{\lambda}_{\Sigma}(t)]^{-}) \right\};$$

i.e., the customer will be routed to an agent pool with the greatest idleness imbalance. If there are no such agents, the customer waits in queue i, to be served in order of arrival.

- *Upon service completion by a type j agent* at time t, the agent will admit to service the customer from the head of queue i^*, where

$$i^* := i^*(t) \in \operatorname*{arg\,max}_{i \in I(j),\, \hat{Q}^{\lambda}_i(t) > 0} \left\{ \hat{Q}^{\lambda}_i(t) - [\hat{X}^{\lambda}_{\Sigma}(t)]^{+} p_i([\hat{X}^{\lambda}_{\Sigma}(t)]^{+}) \right\};$$

i.e., the agent will admit a customer from the queue with the greatest queue imbalance. If there are no such customers, the agent will remain idle.

Ties are broken in an arbitrary but consistent manner, so that the vector-valued stochastic process $(\hat{Q}^{\lambda}, \hat{Z}^{\lambda})$ is a CTMC with stationary transition probabilities.

Remark 2.1 (Degrees of Freedom in Routing and Scheduling). A careful reading of the results and proofs that follow will reveal that the actual way in which j^* or i^* are defined is immaterial as long as they are chosen so that

$$j^* \in \mathscr{V}^{+} := \left\{ j \in J(i)\colon \hat{I}^{\lambda}_j(t) > 0 \text{ and } \hat{I}^{\lambda}_j(t) - [\hat{X}^{\lambda}_{\Sigma}(t)]^{-} v_j([\hat{X}^{\lambda}_{\Sigma}(t)]^{-}) > 0 \right\},$$

$$i^* \in \mathscr{U}^{+} := \left\{ i \in I(j)\colon \hat{Q}^{\lambda}_i(t) > 0 \text{ and } \hat{Q}^{\lambda}_i(t) - [\hat{X}^{\lambda}_{\Sigma}(t)]^{+} p_i([\hat{X}^{\lambda}_{\Sigma}(t)]^{+}) > 0 \right\}.$$

Remark 2.2 (Joint Work Conservation and the Definition of QIR). We note that the QIR routing rule does not work directly with the queue imbalances $\hat{Q}^{\lambda}_i(t) - p_i(\hat{Q}^{\lambda}_{\Sigma}(t))\hat{Q}^{\lambda}_{\Sigma}(t)$. Rather, the queue imbalances are modified so that $[\hat{X}^{\lambda}_{\Sigma}(t)]^{+}$ appears instead of $\hat{Q}^{\lambda}_{\Sigma}(t)$. This modification is important for technical reasons; indeed, the proofs rely on this heavily. We now explain why this is not only a technical assumption. We first observe that, for some simple systems, such are the V and inverted-V models depicted in Figure 3, we have $\hat{Q}^{\lambda}_{\Sigma}(t) = [\hat{X}^{\lambda}_{\Sigma}(t)]^{+}$ under any work-conserving policy. However, in general, that is not the case. Indeed, for the M model in Figure 3, it is possible to have a positive queue for one of the customer classes while there are idle servers in one of the server pools. A reasonable policy will try to prevent such situations, at least approximately, by ensuring that $[\hat{X}^{\lambda}_{\Sigma}(t)]^{+} \approx \hat{Q}^{\lambda}_{\Sigma}(t)$ and $[\hat{X}^{\lambda}_{\Sigma}(t)]^{-} \approx \hat{I}^{\lambda}_{\Sigma}(t)$. Indeed, QIR does exactly that. It guarantees that the queues and idleness have the right proportions, but at the same time the system satisfies

$$[\hat{X}^{\lambda}_{\Sigma}(t)]^{+} \approx \hat{Q}^{\lambda}_{\Sigma}(t) = o_P(1) \qquad \text{and} \qquad [\hat{X}^{\lambda}_{\Sigma}(t)]^{-} \approx \hat{I}^{\lambda}_{\Sigma}(t) = o_P(1).$$

Following Atar [4], we call this *joint work conservation*. We now show that joint work conservation cannot be guaranteed if QIR uses the original queue imbalances, $\hat{Q}^{\lambda}_i(t) - p_i(\hat{Q}^{\lambda}_{\Sigma}(t))\hat{Q}^{\lambda}_{\Sigma}(t)$, rather than the modified ones. To this end, consider the system with four customer classes and three agent pools as depicted in Figure 2. Assume that the ratios are $p_1(\cdot) = p_2(\cdot) = p_3(\cdot) \equiv 0$ and $p_4(\cdot) \equiv 1$. In Figure 2, queues 2 and 4 have some customers

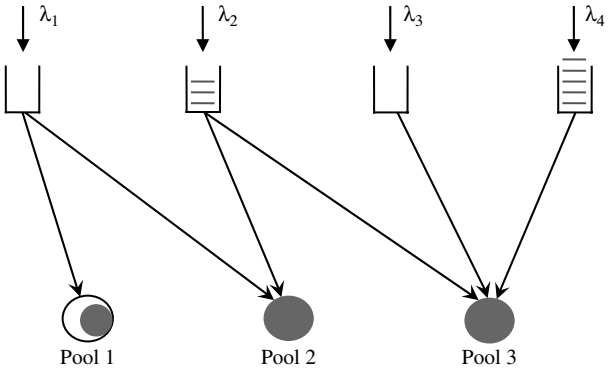

FIGURE 2. An example without joint work conservation.

waiting while the other queues are empty. Also, server pools 2 and 3 are completely busy, while there are some servers available at pool 1. Moreover, assume that at this time t, $\hat{I}_1^\lambda(t) > \hat{Q}_\Sigma^\lambda(t)$ and, in particular, $[\hat{X}_\Sigma^\lambda(t)]^+ = 0$. Then, because $p_2 = 0$ and $p_4 = 1$, a newly available agent in pool 3 will serve a customer from queue 2, which is the queue with the highest value of $\hat{Q}_i^\lambda(t) - p_i(\hat{Q}_\Sigma^\lambda(t))\hat{Q}_\Sigma^\lambda(t)$. However, to move quickly towards joint work conservation, we would prefer this agent from pool 3 to serve class 4. We would prefer to have pool 2 serve class 2, letting pool 1 handle all the class 1 arrivals. That alternative policy will "move" the work toward the nonbusy pool 1. Under QIR, which uses $[\hat{X}_\Sigma^\lambda(t)]^+$, this desired action is exactly what the system would do. In other words, the modified definition of the imbalances is important to force the system to move rapidly towards joint work conservation.

In some cases, QIR does not depend on which form of imbalances are used. Specifically, assume that, for all $i \in \mathcal{I}$, the function $f_i(x) := xp_i(x)$ is strictly increasing. Then, if we change $[\hat{X}_\Sigma^\lambda(t)]^+$ to $\hat{Q}_i^\lambda(t)$ in the set $\{i \in I(j)\colon \hat{Q}_i^\lambda(t) > 0 \text{ and } \hat{Q}_i^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p_i([\hat{X}_\Sigma^\lambda(t)]^+) > 0\}$, then the set is unchanged. Consequently, we can use the original queue imbalances to define QIR. A special case in which this holds is highlighted in Remark 2.3 below. Similar reasoning applies to the idleness ratios. In particular, if the functions $g_i(x) = xv_j(x)$ are strictly increasing, then we can use $\hat{I}_j^\lambda(t) - v_j(\hat{I}_\Sigma^\lambda(t))\hat{I}_\Sigma^\lambda(t)$ in defining the actions upon customer arrival.

REMARK 2.3 (SIMPLIFICATION UNDER FIXED QUEUE RATIOS). If $p(\cdot) \equiv (p_1, \ldots, p_I)$ with $p_i > 0$ for all $i \in \mathcal{I}$, then QIR is achieved by a newly available agent serving the customer from the head of queue i^* where

$$i^* \equiv i^*(t) \in \underset{i \in I(j),\, \hat{Q}_i^\lambda(t) > 0}{\arg\max} \left\{ \frac{\hat{Q}_i^\lambda(t)}{p_i} \right\}.$$

3. State-space collapse under QIR.

THEOREM 3.1 (SSC UNDER QIR). *Fix an optimal solution $\bar{x}$ for* (2) *for which the edges in $\mathcal{E}(\bar{x})$ induce a connected routing graph. Fix the two admissible state-dependent ratio functions p and v. Let QIR be used, following Definition* 2.3. *Suppose that at least one of the following conditions holds with respect to $\bar{x}$.*

CONDITION (C1). *The service rates depend only on the agent type*: *For all $(i,j) \in \mathcal{E}(\bar{x})$, $\mu_{i,j} = \mu_j$.*

CONDITION (C2). *The service rates depend only on the customer class*: *For all $(i,j) \in \mathcal{E}(\bar{x})$, $\mu_{i,j} = \mu_i$.*

CONDITION (C3). *Acyclic routing graph*: *The routing graph determined by $\mathcal{E}(\bar{x})$ is acyclic*; *i.e., it is a tree.*

If, in addition, $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}(0))$ in $\mathbb{R}^{I+I\cdot J}$, then we have SSC:

$$\hat{Q}_i^\lambda(t) - \hat{Q}_\Sigma^\lambda(t)p_i(\hat{Q}_\Sigma^\lambda(t)) \;\Rightarrow\; 0 \qquad \text{and} \qquad \hat{I}_j^\lambda(t) - \hat{I}_\Sigma^\lambda(t)v_j(\hat{I}_\Sigma^\lambda(t)) \;\Rightarrow\; 0 \quad \text{in } D_- \tag{3}$$

as $\lambda \to \infty$ for each $i \in \mathcal{I}$ and $j \in \mathcal{J}$. The convergence in (3) *is strengthened to convergence in D if we assume that*

$$\hat{Q}_i^\lambda(0) - \hat{Q}_\Sigma^\lambda(0)p_i(\hat{Q}_\Sigma^\lambda(0)) \;\Rightarrow\; 0, \quad i \in \mathcal{I}, \qquad \text{and} \qquad \hat{I}_j^\lambda(0) - \hat{I}_\Sigma^\lambda(0)v_j(\hat{I}_\Sigma^\lambda(0)) \;\Rightarrow\; 0, \quad j \in \mathcal{J}. \tag{4}$$

Finally, if Condition (C3) *holds, then,*

$$\frac{1}{\sqrt{\lambda}}\hat{Z}_{ij}^\lambda(t) \;\Rightarrow\; 0 \quad \text{in } D \text{ as } \lambda \to \infty, \;\; i \in \mathcal{I}, \;\; j \in \mathcal{J}. \tag{5}$$

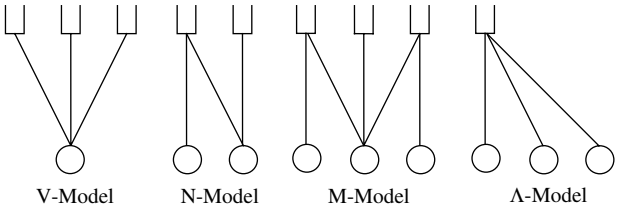

FIGURE 3. The V, N, M, and Λ models.

REMARK 3.1 (NETWORK-GRAPH CHARACTERIZATION). Note that arbitrary connected graphs, including cyclic graphs, are allowed under Conditions (C1) and (C2). However, Condition (C3) rules out cyclic structures. Some simple, yet important, tree structures, are the V, N, M, and Λ (inverted-V) models shown in Figure 3.

At first glance it might seem that the cyclic networks are not required at all, and that one might obtain a cyclic structure (thus satisfying (C3)) by removing some of the edges. However, that is mathematically incorrect. There are examples in which the mathematical program (2) does not have any acyclic solutions that are connected but does have some cyclic solutions. A simple example is constructed as follows: consider a model with two customer classes and two agent groups. Namely, a model with $\mathcal{I} = \mathcal{J} = \{1, 2\}$, with $\lambda_1 = \lambda_2$ (so that $a_1 = a_2 = 0.5$), $\mu_{i,j} = 1$ for all $i, j \in \{1, 2\}$, and finally, with $\nu_1 = \nu_2 = 0.5$. Then, there exists no connected acyclic solution to (2). However, there exist infinitely many cyclic solutions; e.g., $\bar{x}_{i,j} = 0.25$ for all $i, j \in \{1, 2\}$ is an optimal solution that induces a cyclic connected graph.

REMARK 3.2 (CLASS-POOL OCCUPANCY PROCESSES). We now explain why the limit for the class-pool occupancy processes in (5) holds under Condition (C3) but not under any of the other conditions. The reason is the tree structure imposed by Condition (C3). In a tree model, controlling the queues and idleness processes uniquely determines the class-pool occupancy processes $Z^{\lambda}_{i,j}(t)$ through a linear mapping (Equation (11)). That is analogous to what happens in network flows: When the network is a tree, there is a unique way of satisfying all the demand in the network. In the presence of cycles, however, there are many (possibly infinitely many) flows that can satisfy all the demands. In cyclic structures, then, QIR self-selects these agent ratios, not necessarily consistently with the solution $\bar{x}$, and the outcome might depend on the actual ratios p and v. In general, this self-selection might have undesired effects on the system capacity, but not when one of the Conditions (C1) or (C2) holds. Under one of these conditions, the aggregate "fluid" capacity of the system is invariant with respect to the self-selection of the class-pool occupancy processes and is given by $\sum_{j\in\mathcal{J}} \mu_j \bar{\nu}_j$ under Condition (C1), or $\sum_{i\in\mathcal{I}} \mu_i a_i$ under Condition (C2).

EXAMPLE 3.1 (SSC IN A TWO-CLASS MODEL). Even though SSC is an asymptotic property, systems of medium size already exhibit this phenomenon. To illustrate, consider a system with two customer classes, $\mathcal{I} = \{1, 2\}$, and two agents types, $\mathcal{J} = \{1, 2\}$. Let the arrival rates be $\lambda_1 = \lambda_2 = 200$. Assume there is no abandonment. Agents of type 1 can serve both class 1 and class 2 customers. They serve class 1 customers at rate $\mu_{1,1} = 1$ and class 2 customers at rate $\mu_{2,1} = 3$. Agents of type 2 can also give service to both classes, and they do so with rates $\mu_{1,2} = 2$ and $\mu_{2,2} = 3$. Assume that $N_1 = 100$ and $N_2 = 117$ which corresponds roughly to $\bar{\nu}_1 = 1/4$, $\bar{\nu}_2 = 7/24$, and $\gamma_1 = \gamma_2 = 0$, so that an optimal solution for (2) is given by $\bar{x}_{1,1} = 1$, $\bar{x}_{1,2} = 3/7$, $\bar{x}_{2,2} = 4/7$, and $\bar{x}_{i,j} = 0$ otherwise. This solution translates to an N model (see Figure 3).

We simulate this resulting N model to see if SSC holds approximately. Suppose that we use QIR with ratio vector $p(x) \equiv (1/3, 2/3)$ and $v(x) \equiv (0, 1)$. That is, we use a fixed (FQR) ratio function rather than a state-dependent one. With the given ratio vector $p = (1/3, 2/3)$, we should have that $Q_2(t) \approx 2 \cdot Q_1(t)$. Indeed, the simulation results in Figure 4 show that $Q_2(t)$ and $2 \cdot Q_1(t)$ are hardly distinguishable.

4. Stochastic-process limits. We now apply SSC to establish stochastic-process limits under each of the Conditions (C1), (C2), and (C3) of Theorem 3.1. The limits relate the complicated SBR model to much more elementary models, namely, the single-class multitype inverted-V model and the multiclass single-type V model; see Figure 3. Toward this end, we define $\widehat{W}^{\lambda}_i(t) := \sqrt{\lambda} W^{\lambda}_i(t)$ to be the scaled virtual waiting time process of class i customers in the λth system. Also, as before, we set $a_i := \lambda_i/\lambda$. Throughout this section we fix the admissible ratio functions $p(\cdot)$ and $v(\cdot)$. To simplify the notation, we define for all $x \geq 0$:

$$\tilde{p}_i(x) := x p_i(x), \quad i \in \mathcal{I}, \qquad \text{and} \qquad \tilde{v}_i(x) := x v_i(x), \quad j \in \mathcal{J}.$$

The first stochastic-process limit shows the consequence of SSC; the multidimensional limit process is a function of the one-dimensional limiting process $\hat{X}_{\Sigma}(t)$. The joint limits for the queue-length and virtual-waiting-time

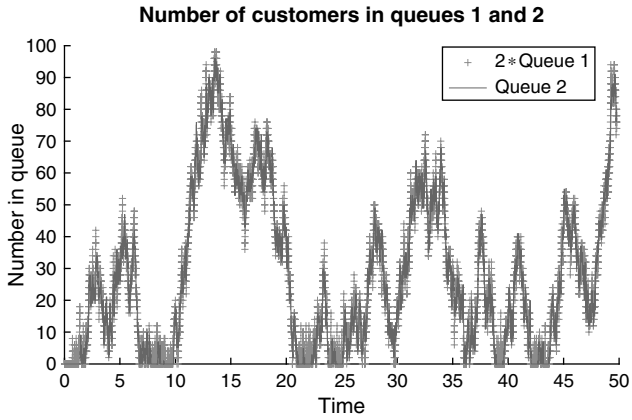

FIGURE 4. The N model for the numerical experiment.

processes imply the heavy-traffic Little's law applied in the heuristic explanation of service-level differentiation in (4) of Gurvich and Whitt [16]. For applications, it is important to realize that we are treating the overall virtual waiting time, including both customers who will be served and customers who will abandon. When the abandonment rate is suitably small, there will be little difference between the overall waiting time and the waiting time conditional on being served.

THEOREM 4.1 (DIFFUSION LIMIT UNDER CONDITION (C1)). *Under the assumptions of Theorem* 3.1 *with Condition* (C1)*, we have the joint convergence*:

$$\begin{aligned}&\big(\hat{X}_\Sigma^\lambda(t), \hat{Q}_1^\lambda(t), \ldots, \hat{Q}_I^\lambda(t), \widehat{W}_1^\lambda(t), \ldots, \widehat{W}_I^\lambda(t), \hat{I}_1^\lambda(t), \ldots, \hat{I}_J^\lambda(t)\big)\\ &\quad\Rightarrow \Big(\hat{X}_\Sigma(t), \tilde{p}_1([\hat{X}_\Sigma(t)]^+), \ldots, \tilde{p}_I([\hat{X}_\Sigma(t)]^+), \frac{1}{a_1}\tilde{p}_1([\hat{X}_\Sigma(t)]^+), \ldots, \frac{1}{a_I}\tilde{p}_I([\hat{X}_\Sigma(t)]^+),\\ &\qquad\qquad \tilde{v}_1([\hat{X}_\Sigma(t)]^-), \ldots, \tilde{v}_J([\hat{X}_\Sigma(t)]^-)\Big)\end{aligned}$$

in D_-^{2I+J+1} *as* $\lambda \to \infty$*, where* $\hat{X}_\Sigma$ *is the unique* (*possibly weak*) *solution of the following one-dimensional SDE*:

$$\hat{X}_\Sigma(t) = \hat{X}_\Sigma(0) - \sum_{j\in\mathcal{J}} \mu_j \gamma_j t + \sum_{j\in\mathcal{J}} \mu_j \int_0^t \tilde{v}_j([\hat{X}_\Sigma(s)]^-)\, ds - \sum_{i\in\mathcal{I}} \theta_i \int_0^t \tilde{p}_i([\hat{X}_\Sigma(s)]^+)\, ds + \sqrt{2}B(t), \tag{6}$$

with $B := \{B(t), t \ge 0\}$ *being a standard Brownian motion. Moreover, the convergence of* $\hat{X}_\Sigma^\lambda(t)$ *can be strengthened to convergence in* D*; i.e,* $\hat{X}_\Sigma^\lambda(t) \Rightarrow \hat{X}_\Sigma(t)$ *in* D *as* $\lambda \to \infty$.

REMARK 4.1 (WHEN (C1) FAILS TO HOLD). The result of Theorem 4.1 is quite strong. While the state of the PSS is characterized by the multidimensional process, $(\hat{Q}_i^\lambda(t), \hat{Z}_{i,j}^\lambda(t);\ i \in \mathcal{I},\ j \in \mathcal{J})$, for each λ, asymptotically as $\lambda \to \infty$ we can characterize the per-class queue-length processes $\hat{Q}_i^\lambda(t)$ and the per-pool idleness processes $\hat{I}_j^\lambda(t)$ in terms of the overall number of customers in the system through the one-dimensional process $\hat{X}_\Sigma(t)$. We claim that Condition (C1) is actually necessary to obtain this result. Under Condition (C1), it suffices to know the number of busy agents $\hat{Z}_j^\lambda(t)$ (or the corresponding number of idle agents $\hat{I}_j^\lambda(t)$) in each pool, to know the departure rate from the system. That makes Theorem 4.1 follow directly from SSC, because it allows us to control the proportions of idle agents. In the absence of Condition (C1), we need more detailed information to know the departure rate from the system. In particular, we need to know the actual values of $(Z_{i,j}^\lambda(t);\ i \in \mathcal{I},\ j \in \mathcal{J})$, over which, in general, we have no control through SSC.

REMARK 4.2 (EQUIVALENCE WITH THE SINGLE-CLASS MODEL). Whenever $\mu_j \equiv \mu$ and $\theta_j \equiv \theta$ for all j, the limit is the same as the one obtained for a sequence of $M/M/N + M$ queues in the Halfin-Whitt regime. With the replacement of the space scaling, $\sqrt{\lambda}$, by the scaling $\sqrt{N_\Sigma^\lambda}$, this limit is given in Theorem 2 of Garnett et al. [15]. If, in addition $\theta = 0$, then the same replacement of scaling leads to the limit process given in Theorem 2 of Halfin and Whitt [19] for a sequence of $M/M/N$ queues with $R + (\sqrt{\mu}\sum_j \gamma_j)\sqrt{R} + o(\sqrt{R})$ agents, where $R = \lambda/\mu$. Specifically, consider a sequence of $M/M/N$ queues with arrival rate λ, service rate μ, and $N^\lambda = R + (\sqrt{\mu}\sum_j \gamma_j)\sqrt{R} + o(\sqrt{R})$. Let $X^\lambda(t)$ be the overall number of customers in the λth $M/M/N$ queue and let $Y^\lambda(t) := [X^\lambda(t) - N^\lambda]/\sqrt{\lambda}$. Then, Theorem 2 of Halfin and Whitt [19] is equivalently stated

as follows: Provided that $Y^\lambda(0) \Rightarrow Y^\lambda(0)$, we have $Y^\lambda(t) \Rightarrow Y(t)$ in $D[0,\infty)$, where Y is a diffusion process satisfying the SDE:

$$Y(t) = Y(0) - \mu \sum_{j\in\mathcal{J}} \gamma_j t + \mu \int_0^t [Y(s)]^- \, ds + \sqrt{2}B(t), \tag{7}$$

with $B := \{B(t),\ t \geq 0\}$ being a standard Brownian motion. As a consequence, then, whenever $\theta_i \equiv 0$ and $\mu_j \equiv \mu$, the PSS and the associated $M/M/N$ queue have asymptotically the same probability law.

Remark 4.3 (Equivalence with the Inverted-V Model). Whenever $\theta_i = \theta$ for all $i \in \mathcal{I}$, the limit we obtain is equal to the limit that we would obtain in the associated inverted-V model, namely, in a model with the same set $\mathcal{J}$ of agent pools, the same service rates $\{\mu_j,\ j \in \mathcal{J}\}$, and the same staffing levels $\{N_j^\lambda,\ j \in \mathcal{J}\}$, but with a single customer class having arrival rate λ. This asymptotic equivalence of the SBR system and the inverted-V model under the assumption of pool-dependent service rates is used extensively in our two subsequent papers (Gurvich and Whitt [16, 17]).

The diffusion limits given in Theorem 4.1 characterize the asymptotic behavior on bounded time periods. The next natural step is to try to understand the asymptotic behavior of the steady-state queue length and waiting time. In some settings one can actually identify the limits of the steady-state variables with the steady-state of the diffusion limit. This, however, requires a limit-interchange argument whose main component is to establish tightness of the scaled steady-state variables. Such an argument was used in simple settings like Halfin and Whitt [19], Garnett et al. [15], Gurvich et al. [18], and also in the more complicated case of Tezcan [28]. Establishing such an interchange is, however, extremely hard in general. Gamarnik and Zeevi [13] and Budhiraja and Lee [8] have developed suitable arguments for generalized Jackson networks in the conventional, single-server, heavy-traffic regime. Their techniques may be adapted, on a case-by-case basis, to certain many-server systems, as was done in Tezcan [28]. A general framework is, however, still missing.

We do not prove the interchange argument for the general PSS case. Assuming that tightness of the scaled steady-state variables holds, however, we can easily link this with the steady-state of the limit diffusion. This link is established in Corollary 4.2 below.

Corollary 4.2 (Steady-State Limits). *Assume that there exists a $\lambda_0 < \infty$ such that a steady state exists for all $\lambda > \lambda_0$. Assume also that the sequence $\{(\hat{Q}_\Sigma^\lambda(\infty), \hat{I}_\Sigma^\lambda(\infty)),\ \lambda > \lambda_0\}$ is tight. Then, as $\lambda \to \infty$,*

$$\begin{aligned} &\hat{X}_\Sigma^\lambda(\infty) \ \Rightarrow\ \hat{X}_\Sigma(\infty); \qquad \hat{I}_j^\lambda(\infty) \ \Rightarrow\ \tilde{v}_j([\hat{X}_\Sigma(\infty)]^-), \quad j \in \mathcal{J}; \\ &\hat{Q}_i^\lambda(\infty) \ \Rightarrow\ \tilde{p}_i([\hat{X}_\Sigma(\infty)]^+); \qquad \textit{and} \qquad \widehat{W}_i^\lambda(\infty) \ \Rightarrow\ \frac{1}{a_i}\tilde{p}_i([\hat{X}_\Sigma(\infty)]^+), \quad i \in \mathcal{I}. \end{aligned} \tag{8}$$

If, in addition, the sequence $\hat{Q}_\Sigma^\lambda(\infty)$ is uniformly integrable, then the convergence in (8) *holds also in expectation.*

The following is a diffusion-limit result under Condition (C2). See §6.3 for the definition of β_i and the construction of the processes $\hat{X}_i^\lambda(t)$.

Theorem 4.3 (Diffusion Limit Under Condition (C2)). *Under the assumptions of Theorem* 3.1 *with Condition* (C2)*, we have the joint convergence*

$$\begin{aligned} &\big(\hat{X}_\Sigma^\lambda(t), \hat{X}_1^\lambda(t), \ldots, \hat{X}_I^\lambda(t), \hat{Q}_1^\lambda(t), \ldots, \hat{Q}_I^\lambda(t), \widehat{W}_1^\lambda(t), \ldots, \widehat{W}_I^\lambda(t), \hat{I}_1^\lambda(t), \ldots, \hat{I}_J^\lambda(t)\big) \\ &\quad \Rightarrow \Big(\hat{X}_\Sigma(t), \hat{X}_1(t), \ldots, \hat{X}_I(t), \tilde{p}_1([\hat{X}_\Sigma(t)]^+), \ldots, \tilde{p}_I([\hat{X}_\Sigma(t)]^+), \frac{1}{a_1}\tilde{p}_1([\hat{X}_\Sigma(t)]^+), \ldots, \frac{1}{a_I}\tilde{p}_I([\hat{X}_\Sigma(t)]^+), \\ &\qquad\qquad \tilde{v}_1([\hat{X}_\Sigma(t)]^-), \ldots, \tilde{v}_J([\hat{X}_\Sigma(t)]^-)\Big) \end{aligned}$$

in D_-^{3I+J+1} as $\lambda \to \infty$, where $(\hat{X}_1(t), \ldots, \hat{X}_I(t))$ is the unique (*possibly weak*) *solution of the following I-dimensional SDE*:

$$\hat{X}_i(t) = \hat{X}_i(0) - \beta_i t - \mu_i \int_0^t \hat{X}_i(s)\, ds - (\theta_i - \mu_i) \int_0^t \tilde{p}_i([\hat{X}_\Sigma(s)]^+)\, ds + \sqrt{2a_i}B_i(t), \tag{9}$$

where the processes $\{\{B_i(t), t \geq 0\},\ i \in \mathcal{I}\}$ are independent standard Brownian motions.

Remark 4.4 (Equivalence with the V Model). The limit we obtain in Theorem 4.3 is equal to the limit that we would obtain in the associated V model, namely, in a model with the same set $\mathcal{I}$ of customer classes and respective arrival and service rates $\{\lambda_i,\ i \in \mathcal{I}\}$ and $\{\mu_i,\ i \in \mathcal{I}\}$, but with a single agent pool having $\sum_{j\in\mathcal{J}} N_j^\lambda$ agents. This stands in contrast to the case of pool-dependent service rates in which the reduction is to a system with multiple agent pools but a single customer class; see Remark 4.3.

The following is a diffusion-limit result under Condition (C3). To abbreviate notation, given x, we let $\tilde{p}(x) := (\tilde{p}_1(x), \ldots, \tilde{p}_I(x))$ and, similarly, $\tilde{v}(x) := (\tilde{v}_1(x), \ldots, \tilde{v}_J(x))$. The mapping $G(\cdot,\cdot)$, that is used in the statement of the theorem, is defined in §5 below.

THEOREM 4.4 (DIFFUSION LIMIT UNDER CONDITION (C3)). *Under the assumptions of Theorem* 3.1 *with Condition* (C3)*, we have the joint convergence*

$$\begin{aligned}&\big(\hat{X}^\lambda_\Sigma(t), \hat{X}^\lambda_1(t), \ldots, \hat{X}^\lambda_I(t), \hat{Q}^\lambda_1(t), \ldots, \hat{Q}^\lambda_I(t), \widehat{W}^\lambda_1(t), \ldots, \widehat{W}^\lambda_I(t), \hat{I}^\lambda_1(t), \ldots, \hat{I}^\lambda_J(t)\big)\\ &\quad\Rightarrow \Big(\hat{X}_\Sigma(t), \hat{X}_1(t), \ldots, \hat{X}_I(t), \tilde{p}_1([\hat{X}_\Sigma(t)]^+), \ldots, \tilde{p}_I([\hat{X}_\Sigma(t)]^+), \frac{1}{a_1}\tilde{p}_1([\hat{X}_\Sigma(t)]^+), \ldots, \frac{1}{a_I}\tilde{p}_I([\hat{X}_\Sigma(t)]^+),\\ &\qquad\qquad \tilde{v}_1([\hat{X}_\Sigma(t)]^-), \ldots, \tilde{v}_J([\hat{X}_\Sigma(t)]^-)\Big)\end{aligned}$$

in D^{3I+J+1}_- *as* $\lambda \to \infty$*, where* $(\hat{X}_1(t), \ldots, \hat{X}_I(t))$ *is the unique* (*possibly weak*) *solution of the following* I*-dimensional SDE*:

$$\begin{aligned}\hat{X}_i(t) = \hat{X}_i(0) - \beta_i t - \sum_{j\in J(i)} \mu_{i,j} \int_0^t G_{ij}\big(\hat{X}(s) - \tilde{p}([\hat{X}_\Sigma(s)]^+), -\tilde{v}([\hat{X}_\Sigma(s)]^-)\big)\, ds\\ - \theta_i \int_0^t \tilde{p}_i([\hat{X}_\Sigma(s)]^+)\, ds + \sqrt{2a_i} B_i(t), \quad i \in \mathcal{I}, \end{aligned} \tag{10}$$

where the processes $\{\{B_i(t),\ t \ge 0\},\ i \in \mathcal{I}\}$ *are independent standard Brownian motions.*

5. QIR and Atar's control. It turns out that for certain tree networks, under an additional restriction on the ratio vector $v(\cdot)$, QIR is equivalent to a special case of the control rule introduced in Atar [4]. Consequently, the SSC for these networks follows from the analysis in Atar [4]. To state the condition, fix an optimal solution $\bar{x}$ for (2) for which the edges in $\mathcal{E}(\bar{x})$ induce a connected routing graph. Then, we will consider networks that satisfy, in addition, the following condition:

CONDITION (C4). *Only one pool with cross-trained agents*: *There exists at most one* $j \in \mathcal{J}$ *with skill set* $I(j)(\bar{x})$ *containing more than one element*; *denote this pool by* j^**. Also, we require that* $v = e_{j^*}$.

Networks that satisfy this structural condition are depicted in Figure 3. Clearly, these networks are trees, so that they also satisfy Condition (C3). We accomplish two objectives in this section: First, we show that under Condition (C4), QIR and an instance of the control constructed in Atar [4] are equivalent; that is proved in Lemma 5.1. Second, we show that, although an SSC result does not appear explicitly in Atar [4], the controls constructed in Atar [4] lead to SSC for tree networks (under Condition (C3)). That appears in Theorem 5.1, whose proof follows from the analysis in Atar [4]; see §8.

Together, these two results provide an alternative proof of SSC for QIR with the networks satisfying Condition (C4). A special case of this equivalence—for the N-model (see Figure 3) with constant ratio functions—was argued and used in Dai and Tezcan [11] in the context of linear-holding-cost minimization.

We start, then, by defining the control used in Atar [4] and stating and proving the SSC result based on his analysis. We refer to the control in Atar [4] as *generalized QIR* (GQIR), even though QIR is not a special case of GQIR. Indeed, GQIR applies only to settings in which the chosen optimal solution $\bar{x}$ is such that the induced graph $\mathcal{E}(\bar{x})$ is a tree. It is not applicable when $\mathcal{E}(\bar{x})$ contains cycles.

Toward the construction of GQIR, define a function $G := G(\alpha, \beta)$: $\mathbb{R}^{I+J} \mapsto \mathbb{R}^{IJ}$, which is defined as the unique solution to the following set of linear equations (see Equation (43) in Atar [4]):

$$\sum_{j\in\mathcal{J}} z_{i,j} = \alpha_i, \qquad \sum_{i\in\mathcal{I}} z_{i,j} = \beta_j, \tag{11}$$

with no additional constraints. Atar [4] shows that the tree structure guarantees a unique solution. This uniqueness can then be used to show that the mapping $G(\cdot,\cdot)$ defines a linear mapping on the domain

$$D_G := \left\{(\alpha,\beta) \in \mathbb{R}^{I+J}\colon \sum_i \alpha_i = \sum_j \beta_j\right\}.$$

Moreover, the mapping G is such that

$$\hat{Z}^\lambda_{ij}(t) := G_{ij}(\hat{X}^\lambda(t) - \hat{Q}^\lambda(t), -\hat{I}^\lambda(t)), \tag{12}$$

that is, given the vector-valued processes $\hat{X}^\lambda(t)$, $\hat{Q}^\lambda(t)$, and $\hat{I}^\lambda(t)$, the values of $\hat{Z}^\lambda(t)$ can be calculated using this mapping. We define a new stochastic process $\check{Z}^\lambda(t)$ in terms of the triple $(\hat{X}^\lambda(t), \hat{Q}^\lambda(t), \hat{I}^\lambda(t))$ through

$$\check{Z}^\lambda_{ij}(t) := G_{ij}(\hat{X}^\lambda(t) - [\hat{X}^\lambda_\Sigma(t)]^+ p([\hat{X}^\lambda_\Sigma(t)]^+), -[\hat{X}^\lambda_\Sigma(t)]^- v([\hat{X}^\lambda_\Sigma(t)]^-)), \quad i \in \mathcal{I}, \quad j \in \mathcal{J}. \tag{13}$$

We are now ready to introduce the GQIR control, as it was constructed in Atar [4]; see §§2.5 and 2.6 of Atar [4].

Definition 5.1 (Generalized QIR: GQIR). Suppose that the routing graph is a tree.

- *Upon an arrival of a class i customer* at time t, if there are any idle agents, route the customer to any agent pool

$$j := j(t) \in \operatorname*{arg\,max}_{k \in J(i),\, I^\lambda_k > 0} (\check{Z}^\lambda_{ik}(t) - \hat{Z}^\lambda_{ik}(t))^+;$$

if all agents in $J(i)$ are busy, then the customer waits in queue, to be served in order of arrival.

- *Upon a service completion by an agent from agent pool j* at time t, if a customer in $I(j)$ is available, then admit to service the customer from the head of any nonempty queue

$$i := i(t) \in \operatorname*{arg\,max}_{k \in I(j),\, \hat{Q}^\lambda_k(t) > 0} (\check{Z}^\lambda_{kj}(t) - \hat{Z}^\lambda_{kj}(t))^+;$$

if all queues in $I(j)$ are empty, then the agent remains idle.

Ties are broken in an arbitrary but consistent manner, so that the vector-valued stochastic process $(\hat{Q}^\lambda, \hat{Z}^\lambda)$ is a CTMC with stationary transition probabilities.

The following lemma holds for all λ; hence we fix λ and omit it from the notation. Also, we add a superscript QIR (or GQIR) to all relevant processes to indicate explicitly the dependence on the control.

Lemma 5.1 (Equivalence of QIR and GQIR Under Condition (C4)). *Fix an optimal solution $\bar{x}$ for* (2) *for which the edges in $\mathcal{E}(\bar{x})$ induce a connected routing graph and assume that Condition* (C4) *holds. Assume that $(X^{GQIR}(0), Z^{GQIR}(0)) = (X^{QIR}(0), Z^{QIR}(0))$ and that the same rule is used to break ties under QIR and GQIR. Then, $\{(X^{GQIR}, Z^{GQIR}),\ t \geq 0\} \stackrel{d}{=} \{(X^{QIR}(t), Z^{QIR}(t)),\ t \geq 0\}$.*

The following is the SSC result for GQIR for general tree networks (going beyond the domain of equivalence). It is a consequence of parts (i) and (iv) in Proposition 1 of Atar [4].

Theorem 5.1 (SSC under GQIR). *Fix an optimal solution $\bar{x}$ for* (2) *for which the edges in $\mathcal{E}(\bar{x})$ induce a connected routing graph. Fix two admissible state-dependent ratio functions p and v and an optimal solution $\bar{x}$ for* (2)*. Suppose that GQIR is used as defined in Definition* 5.1 *and that Condition* (C3) *holds. If, in addition, $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}(0))$ in $\mathbb{R}^{I+I\cdot J}$, then all the conclusions of Theorem* 3.1 *hold, including* (5).

6. Proof of Theorem 3.1. The rest of this paper is devoted to proving the previous results. In this section we prove Theorem 3.1. The proof decomposes into three separate proofs, one for each of the three conditions. Sections 6.2, 6.3, and 6.4 are dedicated to Conditions (C1), (C2), and (C3), respectively. However, we start with preliminaries.

6.1. Preliminaries. Our approach to the construction of the underlying stochastic processes follows a martingale approach that is, by now, quite common; see Pang et al. [23] for an overview.

Sample-path construction. We begin with a sample-path construction that is based on independent unit-rate Poisson processes A_i, $S_{i,j}$, and R_i on $\mathbb{R}_+$ for $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Given these Poisson processes, let

$$X^\lambda_i(t) := X^\lambda_i(0) + A_i(\lambda_i t) - \sum_{j \in \mathcal{J}} S_{i,j}\left(\mu_{i,j} \int_0^t Z^\lambda_{i,j}(s)\, ds\right) - R_i\left(\theta_i \int_0^t Q^\lambda_i(s)\, ds\right), \quad t \geq 0. \tag{14}$$

By direct construction, the stochastic process $\{(X^\lambda_1(t), \ldots, X^\lambda_I(t))\colon t \geq 0\}$ has the correct distribution and is a process in D^I. (A formal argument would follow the proof of Lemma 2.1 in Pang et al. [23].) Equation (14) does not yet fully define the system dynamics because the dynamics of the queue-length processes $Q^\lambda_i(t)$ and the busy-agent processes $Z^\lambda_{i,j}(t)$ are not yet specified. To complete the definition, let $A^\lambda_{ij}(t)$, $(i,j) \in \mathcal{I} \times \mathcal{J}$, be the cumulative number of class i customers routed to an idle type j agent immediately upon arrival by time t; let

$\Phi^\lambda_{i,j}(t)$, $(i,j) \in \mathcal{I} \times \mathcal{J}$, be the cumulative number of class i customers assigned to a type j agent after waiting in queue, by time t. Then, we write

$$\begin{aligned} Z^\lambda_{i,j}(t) &= Z^\lambda_{i,j}(0) + A^\lambda_{i,j}(t) + \Phi^\lambda_{i,j}(t) - S_{i,j}\left(\mu_{i,j}\int_0^t Z^\lambda_{i,j}(s)\,ds\right), \\ Q^\lambda_i(t) &= Q^\lambda_i(0) + A^\lambda_i(t) - \sum_{j\in\mathcal{J}} A^\lambda_{i,j}(t) - \sum_{j\in\mathcal{J}} \Phi^\lambda_{i,j}(t) - R_i\left(\theta_i\int_0^t Q^\lambda_i(s)\,ds\right), \\ I^\lambda_j(t) &= N^\lambda_j - \sum_{i\in\mathcal{I}} Z^\lambda_{i,j}(t). \end{aligned} \tag{15}$$

We let $D^\lambda_j(t) := \sum_{k\in\mathcal{I}} S_{kj}(\mu_{k,j}\int_0^t Z^\lambda_{k,j}(s)\,ds)$ be the cumulative number of service completions by type j agents. The general description of the queueing network above does not yet reflect the specifics of QIR (see Definition 2.3). To include them, let the scaled queue and idleness imbalances be

$$\begin{aligned} \widehat{U}^\lambda_i(t) &:= \widehat{Q}^\lambda_i(t) - [\widehat{X}^\lambda_\Sigma(t)]^+ p_i([\widehat{X}^\lambda_\Sigma(t)]^+), \quad i\in\mathcal{I}, \\ \widehat{V}^\lambda_j(t) &:= \widehat{I}^\lambda_j(t) - [\widehat{X}^\lambda_\Sigma(t)]^- v_j([\widehat{X}^\lambda_\Sigma(t)]^-), \quad j\in\mathcal{J}. \end{aligned} \tag{16}$$

Establishing SSC means showing that all of these processes converge weakly to 0 in the appropriate sense. Now we can write

$$\begin{aligned} A^\lambda_{i,j}(t) &= \int_0^t 1\left\{\hat{I}^\lambda_j(s-) > 0,\ j \in \arg\max_{j\in J(i)}\{\widehat{V}^\lambda_j(s-)\}\right\} dA^\lambda_i(s), \\ \Phi^\lambda_{i,j}(t) &= \int_0^t 1\left\{\widehat{Q}^\lambda_i(s-) > 0,\ i \in \arg\max_{i\in I(j)}\{\widehat{U}^\lambda_i(s-)\}\right\} dD^\lambda_j(s). \end{aligned} \tag{17}$$

These processes are uniquely defined because we have assumed that ties are broken in a consistent manner.

Martingales. We now develop a martingale representation. We start by defining the σ-algebras

$$\begin{aligned} \mathcal{F}^\lambda(t) := \sigma\Bigg\{ & A_i(\lambda_i s), X^\lambda_i(s), Q^\lambda_i(s), Z^\lambda_{i,j}(s), R_i\left(\theta_i\int_0^s Q^\lambda_i(u)\,du\right), \\ & S_{i,j}\left(\mu_{i,j}\int_0^s Z^\lambda_{i,j}(u)\,du\right) : i\in\mathcal{I},\ j\in\mathcal{J},\ s\le t\Bigg\}, \quad t\ge 0, \end{aligned} \tag{18}$$

and make them complete by including all the null sets. The collection of all these σ-algebras $\mathbb{F}^\lambda := \{\mathcal{F}^\lambda(t),\ t\ge 0\}$ is then the *filtration*.

We now exploit a *martingale decomposition*, as in Pang et al. [23]. First, we assume that $E[Q^\lambda_\Sigma(0)] < \infty$ for all λ, but that is without loss of generality, because it can later be relaxed, as in §6.3 of Pang et al. [23]. A straightforward adaptation of Lemmas 3.2 and 3.4 in Pang et al. [23] to our setting establishes that the processes $A_i(\lambda_i t)$, $R_i(\theta_i\int_0^t Q^\lambda_i(s)\,ds)$, and $S_{i,j}(\mu_{i,j}\int_0^t Z^\lambda_{i,j}(s)\,ds)$ admit martingale decompositions with respect to the filtration $\mathbb{F}^\lambda$. Specifically, the stochastic processes

$$\begin{aligned} M^\lambda_{i,j}(t) &:= S_{i,j}\left(\mu_{i,j}\int_0^t Z^\lambda_{i,j}(s)\,ds\right) - \mu_{i,j}\int_0^t Z^\lambda_{i,j}(s)\,ds, \quad i\in\mathcal{I},\ \ j\in\mathcal{J}, \\ M^\lambda_{A_i}(t) &:= A_i(\lambda_i t) - \lambda_i t, \quad i\in\mathcal{I}, \\ M^\lambda_{R_i}(t) &:= R_i\left(\theta_i\int_0^t Q^\lambda_i(s)\,ds\right) - \theta_i\int_0^t Q^\lambda_i(s)\,ds, \quad i\in\mathcal{I}, \end{aligned} \tag{19}$$

are square-integrable martingales with predictable quadratic variations defined by

$$\begin{aligned} \langle M^\lambda_{i,j}\rangle(t) &:= \mu_{i,j}\int_0^t Z^\lambda_{i,j}(s)\,ds, \quad i\in\mathcal{I},\ \ j\in\mathcal{J}, \\ \langle M^\lambda_{A_i}\rangle(t) &:= \lambda_i t, \quad i\in\mathcal{I}, \\ \langle M^\lambda_{R_i}\rangle(t) &:= \theta_i\int_0^t Q^\lambda_i(s)\,ds, \quad i\in\mathcal{I}. \end{aligned} \tag{20}$$

Note that $M^\lambda_{R_i}(t) = 0$, $t\ge 0$ for all i with $\theta_i = 0$.

As integrals of predictable processes with respect to counting processes, the processes $A^\lambda_{i,j}(t)$ and $\Phi^\lambda_{i,j}(t)$ also have corresponding martingale decompositions for all i and j, e.g.,

$$M^\lambda_{A_{i,j}}(t) := \int_0^t 1\left\{\hat{I}^\lambda_j(s-) > 0,\ j \in \arg\max_{j\in J(i)} \hat{V}^\lambda_j(s-)\right\} d(A^\lambda_i(s) - \lambda_i s) \tag{21}$$

is a square-integrable martingale with predictable quadratic variation

$$\langle M^\lambda_{A_{i,j}}\rangle(t) = \int_0^t 1\left\{\hat{I}^\lambda_j(s-) > 0,\ j \in \arg\max_{j\in J(i)} \hat{V}^\lambda_j(s-)\right\} \lambda_i\, ds. \tag{22}$$

This is a consequence of the preservation of the martingale property under stochastic integration. Specifically, fixing any $T > 0$, the martingale $M^\lambda_{A_i}(t \wedge T) = A_i(\lambda_i(t \wedge T)) - \lambda_i(t \wedge T)$ is square integrable and in particular it is in the space $\mathscr{H}^2$ defined on p. 124 of Protter [24]. Because the integrand in (21) is left continuous (because it is defined through the left limits), it is predictable. Because the integrand is also bounded, we can apply Theorem IV.2.11 (p. 129) of Protter [24] to ensure that the stochastic integral in (21) is itself a square integrable martingale. Finally, item (ii) of Lemma 5.77 (p. 85) in Van der Vaart [29] guarantees that the predictable quadratic variation process of (21) is as given in (22). A similar argument is used for the process $\Phi^\lambda_{i,j}(t)$.

Instead of giving detailed expressions for the system dynamics in terms of these martingales here, we will give the required expressions where needed.

Stochastic boundedness and tightness. Our proofs will make extensive use of the concepts of tightness and stochastic boundedness; again see Pang et al. [23] for an overview. In particular, we will be using these notions for stochastic processes in $D^d := D[0,\infty)^d$. We say that a sequence of stochastic processes Y^λ: $\{Y^\lambda(t):\ t \ge 0\}$ in D^d is *stochastically bounded* if, for all $T > 0$,

$$\lim_{A\to\infty} \limsup_{\lambda\to\infty} P\{\|Y^\lambda\|^*_T > A\} = 0,$$

where the norm $\|\cdot\|^*_T$ is as defined in §2. That is a minor modification of the definition for random variables.

We make especial use of C-tightness. A sequence of stochastic processes $\{Y^\lambda\} := \{Y^\lambda(t):\ t \ge 0\}$ in D^d is said to be *C-tight* if, in addition to being tight (as random elements of D^d), every convergent subsequence converges to a limit that is a.s. continuous. Theorem 15.5 from Billingsley [5] is very useful in establishing C-tightness. We restate it here:

Theorem 6.1 (C-Tightness). *Consider a sequence of stochastic processes $\{Y^\lambda\}$ in D^d. Suppose that $\{Y^\lambda(0)\}$ is stochastically bounded in $\mathbb{R}$ and, for each $\epsilon > 0$ and $T > 0$,*

$$\lim_{\delta\to 0} \limsup_{\lambda\to\infty} P\{w_{Y^\lambda}(\delta, T) \ge \epsilon\} = 0 \quad \textit{where } w_x(\delta, T) := \sup_{0\le s<t\le T:\,|t-s|\le\delta} \{|x(t) - x(s)|\}.$$

Then $\{Y^\lambda\}$ is C-tight.

We will want to establish stochastic boundedness and C-tightness for various martingale processes. We use the general notation $M^\lambda(t)$ (or $\widehat{M}^\lambda(t)$ when referring to the scaled version $M^\lambda(t)/\sqrt{\lambda}$ using the scaling in Definition 2.1) for martingale components and refer to specific attributes of the martingale in consideration only where this is needed.

Here are some important general properties:

Lemma 6.1 (Properties of the Martingales). *Let $\widehat{M}^\lambda(t)$ be any of the scaled martingales introduced above (or a finite sum of such), using the scaling in Definition* 2.1*, and assume that $Q^\lambda(0)/\lambda \xrightarrow{P} 0$. Then, the sequence of processes $\{\langle\widehat{M}^\lambda\rangle(t)\}$ is C-tight. Also, the sequence of processes $\{\widehat{M}^\lambda(t)\}$ is stochastically bounded. Finally, for any $\epsilon > 0$,*

$$\limsup_{\lambda\to\infty} P\left\{\sup_{0\le t\le T/\sqrt{\lambda}} |\widehat{M}^\lambda(t)| > \epsilon\right\} = 0. \tag{23}$$

Proof. We begin with C-tightness. Let $\langle\widehat{M}^\lambda\rangle(t)$ be the scaled version of any of the predictable-quadratic-variation processes defined in (20). Then,

$$|\langle\widehat{M}^\lambda\rangle(t) - \langle\widehat{M}^\lambda\rangle(s)| \le c(t-s)\frac{1}{\lambda}\max\left\{\lambda, \sum_{j\in\mathcal{J}} N^\lambda_j, \int_s^t Q^\lambda_\Sigma(u)\,du\right\}, \tag{24}$$

for some positive constant c. By Assumption 2.1, $\sum_{j\in\mathcal{J}} N_j^\lambda \le c_1\lambda$ for all λ large enough and for some constant c_1. Using the trivial inequality $Q_\Sigma^\lambda(u) \le Q_\Sigma^\lambda(0) + \sum_{i\in\mathcal{I}} A_i^\lambda(u)$, the assumed convergence $Q^\lambda(0)/\lambda \xrightarrow{P} 0$, and the renewal strong law of large numbers, we also have that

$$\limsup_{\lambda\to\infty} P\left\{\frac{\int_s^t Q_\Sigma^\lambda(u)\,du}{\lambda} > c_2(t-s)\right\} = 0,$$

for some constant c_2 large enough. Plugging this back into (24), we conclude that

$$\lim_{\delta\to 0}\limsup_{\lambda\to\infty} P\left\{\sup_{0\le s<t\le T:\,|t-s|\le\delta} |\langle\widehat{M}^\lambda\rangle(t) - \langle\widehat{M}^\lambda\rangle(s)| > \epsilon\right\} = 0,$$

and, consequently, that $\langle\widehat{M}^\lambda\rangle(t)$ is C-tight.

The last two conclusions follow from Lemma 5.8 of Pang et al. [23], which is based on the Lenglart-Rebolledo inequality, stated as Lemma 5.7 in that reference. The extension to finite sums follows from basic properties of the quadratic variation processes (see Problem 1.5.7 in Karatzas and Shreve [20] and using the inequality $2\langle M_1, M_2\rangle \le (\langle M_1\rangle + \langle M_1\rangle)$ (see Problem 1.8.9 in Lipster and Shirayev [21]).

6.2. SSC under Condition (C1). For this case, it suffices to consider a less detailed construction of the system dynamics than the one in §6.1. Specifically, we keep Q and I in (15), but instead of Z, we write

$$Z_j^\lambda(t) = Z_j^\lambda(0) + \sum_{i\in\mathcal{I}} A_{i,j}^\lambda(t) + \sum_{i\in\mathcal{I}} \Phi_{i,j}^\lambda(t) - S_j\left(\mu_j\int_0^t Z_j^\lambda(s)\,ds\right), \tag{25}$$

where $Z_j^\lambda(t) := \sum_{i\in\mathcal{I}} Z_{i,j}^\lambda(t)$ is the number of busy agents in pool j and $S_j(\cdot)$ is a unit-rate Poisson process, so that the number of service completions in agent pool j is now given by

$$D_j^\lambda(t) = S_j\left(\mu_j\int_0^t Z_j^\lambda(s)\,ds\right). \tag{26}$$

Also, instead of the martingales $M_{i,j}^\lambda(t)$ in (19), we use the martingales

$$M_j^\lambda(t) := S_j\left(\mu_j\int_0^t Z_j^\lambda(s)\,ds\right) - \mu_j\int_0^t Z_j^\lambda(s)\,ds, \quad j\in\mathcal{J}.$$

Finally, instead of X_i in (14), we use

$$X_\Sigma^\lambda(t) = X_\Sigma^\lambda(0) + \sum_{i\in\mathcal{I}} A_i(\lambda_i t) - \sum_{j\in\mathcal{J}} S_j\left(\mu_j\int_0^t Z_j^\lambda(s)\,ds\right) - \sum_{i\in\mathcal{I}} R_i\left(\theta_i\int_0^t Q_i^\lambda(s)\,ds\right). \tag{27}$$

Following §6.1 with respect to the martingale decomposition and applying some algebraic manipulations, we write

$$X_\Sigma^\lambda(t) = X_\Sigma^\lambda(0) + \left(\lambda - \sum_{j\in\mathcal{J}}\mu_j N_j^\lambda\right)t + \sum_{j\in\mathcal{J}}\mu_j\int_0^t I_j^\lambda(s)\,ds - \sum_{i\in\mathcal{I}}\theta_i\int_0^t Q_i^\lambda(s)\,ds + M_\Sigma^\lambda(t),$$

$$M_\Sigma^\lambda(t) := \sum_{i\in\mathcal{I}} M_{A_i}^\lambda(t) - \sum_{j\in\mathcal{J}} M_j^\lambda(t) - \sum_{i\in\mathcal{I}} M_{R_i}^\lambda(t).$$

By Assumption 2.1,

$$\sum_{j\in\mathcal{J}}\mu_j\gamma_j = \lim_{\lambda\to\infty}\frac{\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda - \lambda}{\sqrt{\lambda}},$$

and we define $\beta := \sum_{j\in\mathcal{J}}\mu_j\gamma_j$. Hence, we may write

$$\widehat{X}_\Sigma^\lambda(t) = \widehat{X}_\Sigma^\lambda(0) - \beta t + \sum_{j\in\mathcal{J}}\mu_j\int_0^t \widehat{I}_j^\lambda(s)\,ds - \sum_{i\in\mathcal{I}}\theta_i\int_0^t \widehat{Q}_i^\lambda(s)\,ds + \widehat{M}_\Sigma^\lambda(t) + o_P(1). \tag{28}$$

The $o_P(1)$ term will play no role in our subsequent analysis and we will omit it.

Having specified the system dynamics, it would be natural to apply Dai and Tezcan [12] to prove the SSC result. Unfortunately, we cannot apply Dai and Tezcan [12] directly. Indeed, a key assumption in Dai and Tezcan [12] is the *homogeneity* condition in their Equation (5.9) or, in its weaker version, in their Equation (7.1). When interpreted to our setting, the homogeneity condition requires that there exist $c_1, c_2 > 0$ and $0 \le \alpha \le 1$ such that

$$\alpha^{c_1} f_i(x) \le f_i(\alpha x) \le \alpha^{c_2} f_i(x),$$

for $f_i(x)$: $\mathbb{R}_+ \to \mathbb{R}$ given by $f_i(x) = x - xp_i(x)$ where $p_i(\cdot)$ is a ratio function. A similar requirement applies to the idleness imbalances. Because our ratio functions need not satisfy those assumptions, we cannot apply Dai and Tezcan [12]. In §7 of that work, it is shown that the homogeneity requirement can be removed if one can, in advance, establish the stochastic boundeness of the scaled dynamics processes. That stochastic boundeness can then be used to prove SSC even when the homogeneity does not hold. In our setting, however, the stochastic boundeness and SSC are interwined and it seems impossible to establish the stochastic boundeness in advance, before proving the SSC.

To overcome this difficulty, we devise a *stopping argument*. Specifically, let

$$\hat{B}^\lambda(t) := \sum_{i \in \mathcal{I}} \widehat{U}_i^\lambda(t), \tag{29}$$

with $\widehat{U}_i^\lambda(t)$ as defined in (16). We first establish SSC assuming that all the processes are stopped at the bounded stopping time

$$T^\lambda := \varsigma^\lambda \wedge T \quad \text{where } \varsigma^\lambda := \inf\{t \ge 0 \mid \hat{B}^\lambda(t) \ge 2\hat{B}^\lambda(0) \vee 1\}. \tag{30}$$

Because all the processes involved are assumed to be right continuous, the stopped processes are well defined (see Propositions 1.1.13 and 1.2.18 of Karatzas and Shreve [20]). Note that the value of $\hat{B}^\lambda(T^\lambda)$ can be greater than $2\hat{B}^\lambda(0) \vee 1$, because there can be a jump at time T^λ. Still, because all arrival and service-completion processes have jumps of size 1, there exists a constant K such that

$$\hat{B}^\lambda(T^\lambda) \le 2\hat{B}^\lambda(0) \vee 1 + K/\sqrt{\lambda}, \quad \text{w.p. 1.} \tag{31}$$

The idea of the stopping argument is that, while it is hard to characterize the limits on the interval $[0, T]$ directly, it is easier to establish these limits for the stopped processes. Once these limits are established, showing that $\varsigma^\lambda \xrightarrow{P} \infty$ will imply that the same limiting behavior holds on $[0, T]$. For the stopped processes, our proof consists of two main modules. First, Lemma 6.2 establishes that the sequence of stopped processes is stochastically bounded and that the sequence of processes $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$ is C-tight. Then, Theorem 6.2 establishes SSC for the stopped processes using the results of Proposition 6.2 and the Bramson [6] and Dai and Tezcan [12] SSC frameworks. Finally, the SSC result is extended to the whole interval $[0, T]$ in Lemma 6.6. Throughout, it is assumed that the conditions of Theorem 3.1 hold in addition to Condition (C1).

Lemma 6.2 (Stochastic Boundedness of Scaled Queueing Processes). *The sequences* $\hat{Q}_\Sigma^\lambda(t \wedge T^\lambda)$, $\hat{I}_\Sigma^\lambda(t \wedge T^\lambda)$, *and* $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$ *are stochastically bounded. In addition, the sequence* $\{\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)\colon \lambda > 0\}$ *is C-tight.*

Proof. By the definition of $\widehat{U}_i^\lambda(t)$ in (16) and Definition 2.2 for state-dependent ratio functions,

$$\hat{B}^\lambda(t) = \sum_{i \in \mathcal{I}} \widehat{U}_i^\lambda(t) = \hat{Q}_\Sigma^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+, \tag{32}$$

where the actual state-dependent ratio functions drop out when we sum over $i \in \mathcal{I}$. In particular,

$$\hat{Q}_\Sigma^\lambda(t) = [\hat{X}_\Sigma^\lambda(t)]^+ + \hat{B}^\lambda(t) \le |\hat{X}_\Sigma^\lambda(t)| + \hat{B}^\lambda(t). \tag{33}$$

By Definition 2.1, $\hat{X}_\Sigma^\lambda(t) = (X_\Sigma^\lambda(t) - N_\Sigma^\lambda)/\sqrt{\lambda} = (Q_\Sigma^\lambda + \sum_{j \in \mathcal{I}} Z_j^\lambda(t) - N_\Sigma^\lambda)/\sqrt{\lambda}$, so that $\hat{I}_\Sigma^\lambda = \hat{Q}_\Sigma^\lambda - \hat{X}_\Sigma^\lambda$. Consequently,

$$\hat{I}_\Sigma^\lambda(t) \le |\hat{X}_\Sigma^\lambda(t)| + \hat{Q}_\Sigma^\lambda(t) \le 2|\hat{X}_\Sigma^\lambda(t)| + \hat{B}^\lambda(t). \tag{34}$$

Plugging these into Equation (28) and applying Gronwall's inequality (see Theorem 4.1 and Lemmas 4.1 and 5.6 in Pang et al. [23] or Problem 5.2.7 in Karatzas and Shreve [20]), we have

$$\|\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)\|_T^* \le c_1\big(|\hat{X}_\Sigma^\lambda(0)| + |\beta|T + \|\hat{B}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\widehat{M}_\Sigma^\lambda(\cdot \wedge T^\lambda)\|_T^*\big)e^{c_2 T},$$

for some positive constants c_1 and c_2. By Lemma 6.1, $\widehat{M}^\lambda_\Sigma(t)$ is stochastically bounded and, by the definition of T^λ and Equation (31), $\widehat{B}^\lambda(t \wedge T^\lambda)$ is stochastically bounded. Here we also use the fact that $\widehat{B}^\lambda(0)$ is itself stochastically bounded by the assumed convergence of $\widehat{X}^\lambda(0)$.

Finally, the sequence $\widehat{X}^\lambda(0)$ is stochastically bounded because it converges; see Corollary 5.2 in Pang et al. [23]. The stochastic boundedness of $\widehat{X}^\lambda_\Sigma(t \wedge T^\lambda)$ now follows from it being bounded by a sum of stochastically bounded sequences; see Lemma 5.5 in Pang et al. [23]. Finally, $\hat{I}^\lambda_\Sigma(t \wedge T^\lambda)$ and $\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)$ are now stochastically bounded by applying (33) and (34).

It remains to establish the claimed C-tightness of $\widehat{X}^\lambda(t \wedge T^\lambda)$. Using (28) once more, we have

$$|\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s)| \le |\beta|(t-s) + \sum_{j\in\mathcal{J}} \mu_j \int_s^t \hat{I}^\lambda_j(s)\,ds + \sum_{i\in\mathcal{I}} \theta_i \int_s^t \hat{Q}^\lambda_i(s)\,ds + |\widehat{M}^\lambda_\Sigma(t) - \widehat{M}^\lambda_\Sigma(s)|,$$

and, as a consequence,

$$|\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s)| \le c_3\big(|\beta|(t-s) + (t-s)\|\hat{I}^\lambda_j(\cdot \wedge T^\lambda)\|^*_T + (t-s)\|\hat{Q}^\lambda_\Sigma(\cdot \wedge T^\lambda)\|^*_T + |\widehat{M}^\lambda_\Sigma(t) - \widehat{M}^\lambda_\Sigma(s)|\big),$$

for all $0 \le s < t \le T^\lambda$. Using (33), (34), and Gronwall's inequality, we obtain

$$|\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s)| \le c_3\big(|\beta|(t-s) + |\widehat{M}^\lambda_\Sigma(t) - \widehat{M}^\lambda_\Sigma(s)| + (t-s)\|\widehat{B}^\lambda(\cdot \wedge T^\lambda)\|^*_T\big)e^{c_4 T},$$

for $0 \le s < t \le T^\lambda$ and for some positive constants c_3 and c_4. With the definitions in Theorem 6.1,

$$w_{\hat{X}^\lambda_\Sigma(\cdot\wedge T^\lambda)}(\delta, T) \le c_5\delta + c_6\delta\|\widehat{B}^\lambda(\cdot \wedge T^\lambda)\|^*_T + c_7 w_{\widehat{M}^\lambda_\Sigma(\cdot\wedge T^\lambda)}(\delta, T), \tag{35}$$

for positive constants c_5, c_6, and c_7. Since $\widehat{B}^\lambda(t \wedge T^\lambda)$ is stochastically bounded by the definition of T^λ, the C-tightness of $\widehat{X}^\lambda(t \wedge T^\lambda)$ is established if we prove the tightness of $\widehat{M}^\lambda_\Sigma(t \wedge T^\lambda)$. This, however, follows immediately from Theorem 5.6 in Pang et al. [23] which allows us to deduce the C-tightness of the sequence of martingales $\widehat{M}^\lambda_\Sigma(t \wedge T^\lambda)$ from the C-tightness of the sequence of predictable-quadratic-variation processes, $\langle \widehat{M}^\lambda_\Sigma \rangle(\cdot \wedge T^\lambda)$, which, in turn, follows from Lemma 6.1. Finally, $\widehat{X}^\lambda_\Sigma(t \wedge T^\lambda)$ is C-tight by (35). □

We are now ready to prove SSC for the stopped processes. With the definition of the processes $\widehat{U}^\lambda_i(t)$ and $\widehat{V}^\lambda_j(t)$ in (16), SSC for the stopped processes is equivalent to the following theorem, as we show in the corollary below:

THEOREM 6.2 (SSC FOR THE STOPPED-PROCESSES). *For any $\epsilon > 0$ and s, $0 < s < T$,*

$$\limsup_{\lambda\to\infty} P\Big\{\sup_{s\le t\le T^\lambda} \sum_{i\in\mathcal{I}} |\widehat{U}^\lambda_i(t)| > \epsilon\Big\} = 0 \qquad \textit{and} \qquad \limsup_{\lambda\to\infty} P\Big\{\sup_{s\le t\le T^\lambda} \sum_{j\in\mathcal{J}} |\widehat{V}^\lambda_j(t)| > \epsilon\Big\} = 0. \tag{36}$$

If, in addition, Equation (4) *holds, then*

$$\limsup_{\lambda\to\infty} P\Big\{\sup_{0\le t\le T^\lambda} \sum_{i\in\mathcal{I}} |\widehat{U}^\lambda_i(t)| > \epsilon\Big\} = 0 \qquad \textit{and} \qquad \limsup_{\lambda\to\infty} P\Big\{\sup_{0\le t\le T^\lambda} \sum_{j\in\mathcal{J}} |\widehat{V}^\lambda_j(t)| > \epsilon\Big\} = 0. \tag{37}$$

The following simple corollary shows that Theorem 6.2 implies the desired form of SSC.

COROLLARY 6.3. *Theorem* 6.2 *implies that*

$$\begin{aligned} \hat{Q}^\lambda_i(t \wedge T^\lambda) - \hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)p_i(\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)) &\Rightarrow 0 \quad \textit{as } \lambda \to \infty \textit{ for all } i \in \mathcal{I}, \\ \hat{I}^\lambda_j(t \wedge T^\lambda) - \hat{I}^\lambda_\Sigma(t \wedge T^\lambda)v_j(\hat{I}^\lambda_\Sigma(t \wedge T^\lambda)) &\Rightarrow 0 \quad \textit{as } \lambda \to \infty \textit{ for all } j \in \mathcal{J}, \end{aligned} \tag{38}$$

where the convergence is in D or D_- depending, as before, on whether Equation (4) *holds.*

PROOF. We only prove the result for the queue processes, because the proof for the idleness processes is similar. Given (16), Theorem 6.2 implies that $\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda) - [\hat{X}^\lambda_\Sigma(t \wedge T^\lambda)]^+ \Rightarrow 0$, where the convergence is in D if Equation (4) holds, and in D_- otherwise. In turn, using the continuity of the state-dependent ratio functions and applying the continuous mapping theorem (see §3.4 of Whitt [30]), we have that

$$p_i(\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)) - p_i([\hat{X}^\lambda_\Sigma(t \wedge T^\lambda)]^+) \Rightarrow 0, \quad \text{as } \lambda \to \infty.$$

Consequently,

$$\widehat{U}^\lambda_i(t \wedge T^\lambda) - \big(\hat{Q}^\lambda_i(t \wedge T^\lambda) - \hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)p_i(\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda))\big) \Rightarrow 0,$$

implying finally that

$$\hat{Q}^\lambda_i(t \wedge T^\lambda) - \hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)p_i(\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)) \Rightarrow 0,$$

where the convergence is in D or D_- depending, as before, on whether Equation (4) holds. □

In general, then, it suffices to consider the processes $\widehat{U}_i^\lambda(t)$, $i \in \mathcal{I}$, and $\widehat{V}_j^\lambda(t)$, $j \in \mathcal{J}$, to establish the result of Theorem 3.1. We will prove all the results only for the processes $\widehat{U}_i^\lambda(t)$, $i \in \mathcal{I}$, as the results for $\widehat{V}_j^\lambda(t)$, $j \in \mathcal{J}$, follow similarly.

In preparation for the proof of Theorem 6.2, we first introduce some of the required framework. Our SSC proof will be based on the general SSC framework of Bramson [6], which was recently extended to the many-server setting by Dai and Tezcan [12]. Because some of the assumptions in Dai and Tezcan [12] do not apply in our setting, we give an independent proof.

The framework of Bramson [6] is based on one key idea: By using an appropriate new scaling of time and space, called the *hydrodynamic scaling*, we can examine the system dynamics over short time intervals. These hydrodynamically scaled (HS) processes are shown to be uniformly approximated by a family of deterministic functions, called the hydrodynamic limit (HL) functions, that satisfy certain equations, known as the hydrodynamic model (HM) equations. This uniform approximation guarantees that, whenever the HL function exhibits the desired SSC, so will the HS processes, as well as the original scaled process, through the appropriate mapping between the original process and its HS version. See Bramson [6] for a more detailed introduction to these concepts.

We begin by defining the HS processes and the HM equations corresponding to our system. To simplify the notation, let

$$R_i^\lambda(t) := R_i\left(\theta_i \int_0^t Q_i^\lambda(s)\,ds\right) \qquad \text{and} \qquad \Theta_i^\lambda(t) := \sqrt{\lambda}[\hat{X}_\Sigma^\lambda(t)]^+ p_i([\hat{X}_\Sigma^\lambda(t)]^+), \quad i \in \mathcal{I}.$$

We will also use $D_j^\lambda(t)$, which was defined in (26). We start with the basic processes indexed by λ:

$$\mathbb{X}^\lambda(t) := \left(A_i^\lambda(t), A_{i,j}^\lambda(t), \Phi_{i,j}^\lambda(t), D_j^\lambda(t), R_i^\lambda(t), \Theta_i^\lambda(t), Q_i^\lambda(t), U_i^\lambda(t), Z_j^\lambda(t);\ i \in \mathcal{I},\ j \in \mathcal{J}\right). \tag{39}$$

Then, for any fixed $L > 0$, and every nonnegative integer m with $m < \sqrt{\lambda}T$, we construct the *hydrodynamically scaled* (*HS*) *processes*

$$\mathbb{X}^{\lambda,m} := \left(A_i^{\lambda,m}(t), A_{i,j}^{\lambda,m}(t), \Phi_{i,j}^{\lambda,m}(t), D_j^{\lambda,m}(t), R_i^{\lambda,m}(t), \Theta_i^{\lambda,m}(t), Q_i^{\lambda,m}(t), U_i^{\lambda,m}(t), Z_j^{\lambda,m}(t);\ i \in \mathcal{I},\ j \in \mathcal{J}\right),$$

for $t \in [0, L]$ as follows:

$$\begin{aligned}
A_i^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(A_i^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - A_i^\lambda\left(\frac{m}{\sqrt{\lambda}}\right)\right),\\
A_{i,j}^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(A_{i,j}^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - A_{i,j}^\lambda\left(\frac{m}{\sqrt{\lambda}}\right)\right),\\
\Phi_{i,j}^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(\Phi_{i<j}^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - \Phi_{i,j}^\lambda\left(\frac{m}{\sqrt{\lambda}}\right)\right),\\
D_j^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(D_j^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - D_j^\lambda\left(\frac{m}{\sqrt{\lambda}}\right)\right),\\
R_i^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(R_i^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - R_i^\lambda\left(\frac{m}{\sqrt{\lambda}}\right)\right),\\
\Theta_i^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(\Theta_i^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - \Theta_i^\lambda\left(\frac{m}{\sqrt{\lambda}}\right)\right),\\
Q_i^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(Q_i^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right)\right),\\
U_i^{\lambda,m}(t) &:= \left(\widehat{U}_i^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right)\right), \qquad \text{and}\\
Z_j^{\lambda,m}(t) &:= \frac{1}{\sqrt{\lambda}}\left(Z^\lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}}\right) - N_j^\lambda\right).
\end{aligned} \tag{40}$$

Remark 6.1 (The Time and Space Scaling). The HS processes in Equation (40) have a more elementary time-and-space scaling than in Bramson [6] and Dai and Tezcan [12]—see §6.1 of Dai and Tezcan [12] and,

in particular, Equation (6.1) there. The more complex scaling is required if we cannot prove a priori that the queue and idleness processes are stochastically bounded. In the absence of stochastic boundedness, the resulting notion of SSC is the multiplicative SSC, defined in Theorem 5.1 and Remark 5.3 of Dai and Tezcan [12]. However, when stochastic boundedness is available, SSC and multiplicative SSC are equivalent. Then both can be proved using the more elementary scaling, as illustrated in the proof of Theorem 7.3 in Dai and Tezcan [12]. Because our SSC argument will focus initially on the stopped processes, the stochastic boundedness established in Lemma 6.2 allows us to use the more elementary time-and-space scaling. □

We will show that the HS processes are uniformly approximated by Lipschitz, and thus absolutely continuous, deterministic functions called the *hydrodynamic limit* (*HL*) *functions*, denoted by

$$\tilde{X} := \big(\tilde{A}_i(t), \tilde{A}_{ij}(t), \tilde{\Phi}_{ij}(t), \tilde{D}_j(t), \tilde{R}_i(t), \tilde{\Theta}_i(t), \widetilde{Q}_i(t), \tilde{U}_i(t), \tilde{Z}_j(t);\ i \in \mathcal{I},\ j \in \mathcal{J}\big),$$

satisfying the following hydrodynamic model (HM) equations. However, the following HM equations, in general, do not determine the HL functions uniquely. Thus we will show that there is some HL function satisfying the HM equations that is suitably close to the HS process above. Because we are not aiming for uniqueness, the hydrodynamic model might contain additional equations, but we specify only those that are relevant for our purposes.

The *hydrodynamic model* (*HM*) *equations* are:

$$\begin{aligned}
\tilde{A}_i(t) &= a_i t, \quad i \in \mathcal{I},\\
\tilde{\Theta}_i(t) = \tilde{R}_i(t) &= 0, \quad \text{for all } t \ge 0,\\
\widetilde{Q}_i(t) &= \widetilde{Q}_i(0) + \tilde{A}_i(t) - \sum_{j\in\mathcal{J}} \tilde{A}_{ij}(t) - \sum_{j\in\mathcal{J}} \tilde{\Phi}_{ij}(t), \quad i \in \mathcal{I},\\
\widetilde{Q}_i(t) &\ge 0, \quad i \in \mathcal{I},\\
\tilde{A}_{ij}(t), \tilde{\Phi}_{ij}(t), &\quad i \in \mathcal{I},\ j \in \mathcal{J} \text{ are nondecreasing},\\
\tilde{Z}_j(t) &= \tilde{Z}_j(0) + \sum_{i\in\mathcal{I}} \tilde{A}_{ij}(t) + \sum_{i\in\mathcal{I}} \tilde{\Phi}_{ij}(t) - \mu_j \bar{\nu}_j t, \quad j \in \mathcal{J}.
\end{aligned} \tag{41}$$

The HM equations in (41) appear in the hydrodynamic model of an arbitrary policy. Letting $\tilde{I}^+(t) := \{i \in \mathcal{I}\colon \tilde{U}_i(t) > 0\}$, and $J(\tilde{I}^+(t)) := \{j \in \mathcal{J}\colon i \in J(i),\ \text{for some } i \in \tilde{I}^+(t)\}$, we also define the following HM equations that are specific to QIR:

$$\sum_{i\in\mathcal{I}} \tilde{U}_i(t) \ge 0 \qquad \text{and} \qquad \tilde{U}_i(t) = \tilde{U}_i(0) + \tilde{A}_i(t) - \sum_{j\in\mathcal{J}} \tilde{A}_{ij}(t) - \sum_{j\in\mathcal{J}} \tilde{\Phi}_{ij}(t), \quad i \in \mathcal{I}, \tag{42}$$

$$\sum_{i\in\tilde{I}^+(t)} \sum_{j\in J(i)} d\tilde{\Phi}_{ij}(t) = \sum_{j\in J(\tilde{I}^+(t))} \mu_j \bar{\nu}_j, \quad \text{and in particular,} \tag{43}$$

$$\sum_{i\in\tilde{I}^+(t)} d\tilde{U}_i(t) \le -c, \quad \text{whenever } \tilde{I}^+(t) \ne \varnothing \text{ for some constant } c > 0. \tag{44}$$

We have stipulated that these HL functions are Lipschitz. The relevant set of HL functions will depend on positive real parameters k, L_k, and N. Specifically, for appropriate parameters, the set of HL functions will be a subset of the family E' of functions in D^d that satisfy $|x(0)| \le k$ and $|x(t_2) - x(t_1)| \le N|t_2 - t_1|$ for all $t_1, t_2 \in [0, L_k]$. By the Arzela-Ascoli theorem in Billingsley [5, p. 221], the set E' is a compact subset of C^d and thus of D^d for appropriate d. Because the HL functions are Lipschitz, they are absolutely continuous; e.g., see §5.4 of Royden [26].

In addition, the HL functions must satisfy the HM equations above. The existence of HL functions satisfying those HM equations will be a consequence of our analysis and, in particular, of Lemmas 6.4 and 6.5 below. Note that the final HM equation, Equation (44), implies that there is a constant upper bound c on the rate at which $\sum_{i\in\mathcal{I}}[\tilde{U}_i(t)]^+$ decreases whenever it is positive. Consequently, it reaches 0 within a finite time, after which it stays at 0, by the first inequality in (42); see Lemma 5.2 of Dai [9] for technical support. We will use this fact in the proof of Theorem 6.2.

First, however, we identify the relations between the HS processes $\mathbb{X}^{\lambda,m}$ and the HL functions $\tilde{X}$. This is done in the following theorem, which shows that, for λ large enough, the HS process $\mathbb{X}^{\lambda,m}$ is close enough to some HL function $\tilde{X}$, satisfying the HM Equations (41)–(44).

THEOREM 6.4 (UNIFORM APPROXIMATION BY HL FUNCTIONS). *For any $T > 0$, $\delta > 0$, and $\epsilon > 0$, there exist $k := k(\delta, \epsilon, T)$, L_k, λ_0, and subsets $\mathcal{K}^{\lambda,k}$ of the underlying probability space Ω, such that for all $\lambda \geq \lambda_0$:*

(i) $\|\widehat{U}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\widehat{Z}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\widehat{Q}^\lambda(\cdot \wedge T^\lambda)\|_T^* \leq k$ *on* $\mathcal{K}^{\lambda,k}$,

(ii) *for each* $\omega \in \mathcal{K}^{\lambda,k}$ *and* m *with* $m < \sqrt{\lambda}T^\lambda$, *there exists an HL function* (*a Lipschitz function satisfying the HM equations*) $\widetilde{X}$, *depending on* λ *and* m, *such that* $\|\mathbb{X}^{\lambda,m} - \widetilde{X}\|_{L_k}^* \leq \epsilon$;

(iii) *finally,* $P\{\mathcal{K}^{\lambda,k}\} \geq 1 - \delta$.

Theorem 6.4 captures the essence of the hydrodynamic limit approach: The idea is to show that, on a suitably large subset of the sample space, and for all λ sufficiently large, the HS process is close enough to some HL function satisfying the HM equations. We postpone the proof of Theorem 6.4 and apply it now to prove Theorem 6.2.

PROOF OF THEOREM 6.2. The proof of the first conclusion consists of two steps. First, we focus on an HL function, fix ϵ and k, and show that there exists some finite time $s^* := s^*(k, \epsilon)$ such that $\widetilde{U}_i(t) = 0$ for $t \geq s^*$ and for all i provided that the HL function $\widetilde{X}$ satisfies the HM Equations (41)–(44) and $\sum_{i\in\mathcal{I}} |\widetilde{U}_i(0)| \leq k + \epsilon$.

Property (44) implies that the HL function $\sum_{i\in\mathcal{I}}[\widetilde{U}_i(t)]^+$ decreases at a rate of at least c until it reaches 0, after which it stays there; see Lemma 5.2 of Dai [9] for technical support. Equation (44) directly controls only the positive part, but Equation (42) implies that the negative part is dominated by the positive part in absolute value. Hence, when $\sum_{i\in\mathcal{I}}[\widetilde{U}_i(t)]^+ = 0$, we also have $\widetilde{U}_i(t) = 0$ for all i by virtue of Equation (42). These functions must remain 0 thereafter, because the positive part cannot increase.

Thus we have the existence of $s^* := s^*(k, \epsilon)$ such that $\widetilde{U}_i(t) = 0$ for $t \geq s^*$ for all i for any HL function $\widetilde{X}$ satisfying the HM Equations (41)–(44) with $\sum_{i\in\mathcal{I}} |\widetilde{U}_i(0)| \leq k + \epsilon$. We now come to the second step of the proof of the first conclusion, in which we find an HL function appropriately related to the HS process: Fix $\delta > 0$, $k > 0$, and λ_0 so that for all $\lambda \geq \lambda_0$, $P\{\mathcal{K}^{\lambda,k}\} \geq 1 - \delta$. Also, choose $L_k \geq 2\lceil s^* \rceil$, and finally, fix $\omega \in \mathcal{K}^{\lambda,k}$. Consider a time $t \leq T^\lambda$ with $t \geq s^*/\sqrt{\lambda}$ (if such time exists), and let

$$m^\lambda(t) := \max\{m\colon\ m < \sqrt{\lambda}T^\lambda,\ (m + s^*)/\sqrt{\lambda} \leq t\}.$$

Then $t \in [m^\lambda(t)/\sqrt{\lambda},\ (m^\lambda(t) + L_k)/\sqrt{\lambda}]$, and, by definition of the HS process, $\widehat{U}_i^\lambda(t) = \widehat{U}_i^{\lambda, m^\lambda(t)}(\sqrt{\lambda}t - m^\lambda(t))$. Now, by Theorem 6.4, for λ large enough, there exists an HL function $\widetilde{X}$ that satisfies the HM Equations (41)–(44) with $|U_i^{\lambda, m^\lambda(t)} - \widetilde{U}_i|_{L_k}^* \leq \epsilon$. In particular, we have $\sum_{i\in\mathcal{I}} |\widetilde{U}_i(0)| \leq k + \epsilon$, which by our previous argument implies that $\sum_{i\in\mathcal{I}} |\widetilde{U}_i(t)| = 0$, $t \geq s^*$. Since $L_k \geq \sqrt{\lambda}t - m^\lambda(t) \geq s^*$,

$$\sum_{i\in\mathcal{I}} \left|\widetilde{U}_i(\sqrt{\lambda}t - m^\lambda(t))\right| = 0 \quad \text{on } \mathcal{K}^{\lambda,k}.$$

Combining these relations, we then have $\sum_{i\in\mathcal{I}} |\widehat{U}_i^\lambda(t)| \leq \epsilon$. Hence,

$$\sup_{s^*/\sqrt{\lambda} \leq t \leq T^\lambda} \sum_{i\in\mathcal{I}} |\widehat{U}_i^\lambda(t)| \leq \epsilon \quad \text{on } \mathcal{K}^{\lambda,k}, \tag{45}$$

where we naturally set the value to be 0 whenever $T^\lambda < s^*/\sqrt{\lambda}$. Because the same holds for any $\omega \in \mathcal{K}^{\lambda,k}$, for all $\lambda \geq \lambda_0$,

$$P\left\{\sup_{s^*/\sqrt{\lambda} \leq t \leq T^\lambda} \sum_{i\in\mathcal{I}} |\widehat{U}_i^\lambda(t)| > \epsilon\right\} \leq \delta,$$

from which (36) readily follows.

For the second conclusion, assume that (4) holds and let

$$\widetilde{\Omega}^\lambda = \left\{w \in \Omega\colon\ \sum_{i\in\mathcal{I}} |\widehat{U}_i^\lambda(0)| \leq \epsilon\right\}.$$

Then, (4) implies that $P\{\widetilde{\Omega}^\lambda\} \to 1$ as $\lambda \to \infty$. Consider the process $U_i^{\lambda,0}(t)$ and its corresponding approximation $\widetilde{U}_i$ from Theorem 6.4. Then, on $\mathcal{K}^{\lambda,k} \cap \widetilde{\Omega}^\lambda$, we must have $\sum_{i\in\mathcal{I}} |\widetilde{U}_i(0)| \leq \epsilon$, and repeating the same argument we used above, we will have $\sum_{i\in\mathcal{I}} |\widetilde{U}_i(t)| \leq \epsilon$ for all $t \geq 0$. In particular, $\widehat{U}_i^\lambda(t) \leq 2\epsilon$ for all $t \leq T^\lambda \wedge s^*/\sqrt{\lambda}$. Adding this to (45), we have on $\mathcal{K}^{\lambda,k} \cap \widetilde{\Omega}^\lambda$,

$$\sup_{0 \leq t \leq T} \sum_{i\in\mathcal{I}} |\widehat{U}_i^\lambda(t)| \leq 2\epsilon.$$

Since $P\{\mathscr{K}^{\lambda,k} \cap \widetilde{\Omega}^{\lambda}\} \geq 1 - 2\delta$ for all λ large enough, we have established that

$$\limsup_{\lambda\to\infty} P\left\{ \sup_{s\leq t\leq T^{\lambda}} \sum_{i\in\mathcal{I}} |\widehat{U}_i^{\lambda}(t)| > 2\epsilon \right\} \leq 2\delta,$$

implying (37).

Proof of Theorem 6.4. The proof is similar to corresponding proofs in Bramson [6] and in Dai and Tezcan [12]. Specifically, it parallels the proofs in §§C.2–C.3 of Dai and Tezcan [12]; however, there are some minor differences. Hence, we write out the proofs, but abbreviate whenever the proof closely follows Bramson [6] or Dai and Tezcan [12]. The proof is divided into three lemmas. Lemma 6.3 shows that, on a large enough subspace of the sample space, the process $\mathbb{X}^{\lambda,m}$ is almost Lipschitz. This is used in Lemma 6.4 to establish the uniform approximation by cluster points. Together with Lemmas 4.1 and 4.2 and Proposition 4.1 of Bramson [6], Lemmas 6.3 and 6.4 here imply the uniform approximation by a Lipschitz function. Bramson [6] elaborates on the compactness and cluster-point structure. Finally, Lemma 6.5 below establishes that each cluster point satisfies Equations (41)–(44). That primarily means the last two equations: (43) and (44).

We start by defining some important sets. Fix k, λ, and L_k and define the following sets:

$$\begin{aligned}
\Omega_1^{\lambda,k} &:= \{\omega\in\Omega\colon \|\widehat{U}^{\lambda}(\cdot\wedge T^{\lambda})\|_T^* + \|\widehat{Z}^{\lambda}(\cdot\wedge T^{\lambda})\|_T^* + \|\widehat{Q}^{\lambda}(\cdot\wedge T^{\lambda})\|_T^* \leq k\},\\
\Omega_2^{\lambda,k} &:= \Omega_2^{\lambda,k}(\epsilon) := \left\{\omega\in\Omega\colon \max_{m<\sqrt{\lambda}T^{\lambda}} \|A^{\lambda,m}(t) - at\|_{L_k}^* \leq \epsilon\right\},\\
\Omega_3^{\lambda,k} &:= \Omega_3^{\lambda,k}(\epsilon) := \left\{\omega\in\Omega\colon \max_{m<\sqrt{\lambda}T^{\lambda}} \|D^{\lambda,m}(t) - \mu\bar{\nu}t\|_{L_k}^* \leq \epsilon\right\},\\
\Omega_4^{\lambda,k} &:= \Omega_4^{\lambda,k}(\epsilon,N) := \left\{\omega\in\Omega\colon \max_{m<\sqrt{\lambda}T^{\lambda}} \sup_{t_1,t_2\leq L_k} \|\mathbb{X}^{\lambda,m}(t_2) - \mathbb{X}^{\lambda,m}(t_1)\| \leq N|t_2-t_1| + \epsilon\right\},
\end{aligned} \tag{46}$$

where $A^{\lambda,m}(t) := (A_1^{\lambda,m}(t),\dots,A_I^{\lambda,m}(t))$, $a := (a_1,\dots,a_I)$, $D^{\lambda,m}(t) := (D_1^{\lambda,m}(t),\dots,D_J^{\lambda,m}(t))$, $\mu\bar{\nu} := (\mu_1\bar{\nu}_1,\dots,\mu_J\bar{\nu}_J)$, and N is some fixed constant that depends only on the vectors a, $\bar{\nu}$ as well as I and J (and whose specific value will be made explicit in the proof of the following lemma). Set $\mathscr{K}^{\lambda,k} := \bigcap_{i=1}^4 \Omega_i^{\lambda,k}$. The following lemma is the analogue of Propositions 6.2 and 6.3 in Dai and Tezcan [12].

Lemma 6.3. *For any $\epsilon,\delta>0$, there exist k and L_k so that $\liminf_{\lambda\to\infty} P\{\mathscr{K}^{\lambda,k}\} \geq 1-\delta$ for $\mathscr{K}^{\lambda,k}$ defined above.*

Proof. By Lemma 6.2, $\hat{X}_{\Sigma}^{\lambda}(t\wedge T^{\lambda})$, $\hat{Q}_{\Sigma}^{\lambda}(t\wedge T^{\lambda})$, and $\hat{I}_{\Sigma}^{\lambda}(t\wedge T^{\lambda})$ are stochastically bounded. Since $\hat{Z}_j(t) = -\hat{I}_j^{\lambda}(t)$ and, by definition, $\sum_{i\in\mathcal{I}}|\widehat{U}_i^{\lambda}(t)| \leq \hat{\bar{Q}}_{\Sigma}^{\lambda}(t) + |\hat{X}_{\Sigma}^{\lambda}(t)|$, we have

$$\lim_{k\to\infty}\limsup_{\lambda\to\infty} P\{\Omega - \Omega_1^{\lambda,k}\} = 0. \tag{47}$$

We now turn to the set $\Omega_3^{\lambda,k}$ (the argument for the set $\Omega_2^{\lambda,k}$ is omitted because it follows similarly and is even easier). By construction, for $t\leq L_k$,

$$\begin{aligned}
D_j^{\lambda,m}(t) &= \frac{1}{\sqrt{\lambda}}\left(S_j\left(\mu_j\int_0^{(m+t)/\sqrt{\lambda}} Z_j^{\lambda}(s)\,ds\right) - S_j\left(\mu_j\int_0^{m/\sqrt{\lambda}} Z_j^{\lambda}(s)\,ds\right)\right)\\
&\stackrel{d}{=} \frac{1}{\sqrt{\lambda}}\left(S_j\left(\mu_j\int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_j^{\lambda}(s)\,ds\right)\right),
\end{aligned}$$

where the equivalence in distribution (as processes) follows from the properties of the Poisson process. As a consequence,

$$\left\| D_j^{\lambda,m}(t) - \frac{1}{\sqrt{\lambda}}\mu_j\int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_j^{\lambda}(s)\,ds\right\|_{L_k}^* \stackrel{d}{=} \frac{1}{\sqrt{\lambda}}\|S_j(\mu_j N_j^{\lambda}t) - \mu_j N_j^{\lambda}t\|_{\psi^{\lambda}}^*,$$

where

$$\psi^{\lambda} = \frac{\int_{m/\sqrt{\lambda}}^{(m+L_k)/\sqrt{\lambda}} Z_j^{\lambda}(s)\,ds}{N_j^{\lambda}}.$$

Using the fact that $Z_j^{\lambda}(t)\leq N_j^{\lambda}$ and carefully applying Proposition 4.3 in Bramson [6], we have

$$P\left\{\left\| D_j^{\lambda,m}(t) - \frac{1}{\sqrt{\lambda}}\mu_j\int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_j^{\lambda}(s)\,ds\right\|_{L_k}^* > \epsilon L_k\right\} \leq \frac{\epsilon\sqrt{\lambda}}{\mu_j N_j^{\lambda} L_k}.$$

Bounding the distribution of the maximum by summation over the individual probability distributions, we then have

$$P\left\{\max_{m<\sqrt{\lambda}T^\lambda}\left\|D_j^{\lambda,m}(t)-\frac{1}{\sqrt{\lambda}}\mu_j\int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}}Z_j^\lambda(s)\,ds\right\|_{L_k}^{*}>\epsilon L_k\right\}\le\frac{\epsilon\lambda T}{\mu_j N_j^\lambda L_k}\le\frac{2\epsilon T}{\mu_j\bar{\nu}_j L_k},\tag{48}$$

for all λ large enough, since $N_j^\lambda/\lambda\to\bar{\nu}_j$ as $\lambda\to\infty$. We can then replace ϵ with ϵ/L_k and choose L_k large enough so that the probability in (48) is bounded by $\delta/8J$. Finally, since $\hat{I}_\Sigma^\lambda(t\wedge T^\lambda)$ is stochastically bounded (by Lemma 6.2), and since $N_j^\lambda=\bar{\nu}_j\lambda+\gamma_j\sqrt{\lambda}+o(\sqrt{\lambda})$, we have

$$P\left\{\max_{m<\sqrt{\lambda}T^\lambda}\left\|\frac{1}{\sqrt{\lambda}}\mu_j\int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}}Z_j^\lambda(s)\,ds-\mu_j\bar{\nu}_j t\right\|_{L_k}^{*}>\epsilon\right\}\le\delta/8J,$$

for all λ large enough. Repeating the same argument for all $j\in\mathcal{J}$ and fixing L_k sufficiently large, we then have, for all λ large enough,

$$P\left\{\max_{m<\sqrt{\lambda}T^\lambda}\left\|D^{\lambda,m}(t)-\mu\bar{\nu}t\right\|_{L_k}^{*}>\epsilon\right\}\le\delta/4,$$

so that $P\{\Omega_3^{\lambda,k}\}\ge 1-\delta/4$. A similar, but easier, argument for $\Omega_2^{\lambda,k}$ shows that $P\{\Omega_2^{\lambda,k}\}\ge 1-\delta/4$ for all λ large enough. The details are omitted.

We turn now to the set $\Omega_4^{\lambda,k}$. We first prove probability bounds for each process separately and finally combine all the bounds to obtain the corresponding bound for the multidimensional process $\mathbb{X}^{\lambda,m}$. We start from the process $D_j^{\lambda,m}(t)$, $j\in\mathcal{J}$. Since $Z_j^\lambda(s)\le N_j^\lambda$ for all $s\ge 0$, we have that

$$\sup_{t_1,t_2\le L_k}\left|D_j^{\lambda,m}(t_2)-D_j^{\lambda,m}(t_1)\right|\le\frac{1}{\sqrt{\lambda}}\sup_{t_1,t_2\le L_k}=\left|S_j(\mu_j N_j^\lambda t_2/\sqrt{\lambda})-S_j(\mu_j N_j^\lambda t_1/\sqrt{\lambda})\right|.$$

For $t_1,t_2\le L_k$,

$$\left|S_j(\mu_j N_j^\lambda t_2/\sqrt{\lambda})-S_j(\mu_j N_j^\lambda t_1/\sqrt{\lambda})\right|\le\mu_j\frac{N_j^\lambda}{\sqrt{\lambda}}|t_2-t_1|+2\sup_{t\le L_k/\sqrt{\lambda}}\left|S_j(\mu_j N_j^\lambda t)-\mu_j N_j^\lambda t\right|.$$

Fix $\epsilon'>0$. We then have

$$\begin{aligned}&P\left\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,t_2\le L_k}|D_j^{\lambda,m}(t_2)-D_j^{\lambda,m}(t_1)|>\mu_j\frac{N_j^\lambda}{\lambda}|t_2-t_1|+\epsilon'\right\}\\&\qquad\le\sqrt{\lambda}T\cdot P\left\{\sup_{t\le L_k/\sqrt{\lambda}}|S_j(\mu_j N_j^\lambda t)-\mu_j N_j^\lambda t|>\frac{\epsilon'}{2}\sqrt{\lambda}\right\}.\end{aligned}\tag{49}$$

Fixing $\delta'>0$, applying Proposition 4.3 of Bramson [6] to the right-hand side of (49), and using the fact that $N_j^\lambda/\lambda\to\bar{\nu}_j$ as $\lambda\to\infty$, we have, for fixed $\delta'>0$,

$$P\left\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,t_2\le L_k}|D_j^{\lambda,m}(t_2)-D_j^{\lambda,m}(t_1)|>\mu_j\bar{\nu}_j|t_2-t_1|+\epsilon'\right\}\le\delta',\tag{50}$$

for all λ large enough. A similar argument is repeated for $A_i^{\lambda,m}(t)$, $i\in\mathcal{I}$, to show that

$$P\left\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,t_2\le L_k}|A_i^{\lambda,m}(t_2)-A_i^{\lambda,m}(t_1)|>a_i|t_2-t_1|+\epsilon'\right\}\le\delta'.\tag{51}$$

We now treat the more complicated routing processes $\Phi_{i,j}^{\lambda,m}(t)$ and $A_{i,j}^{\lambda,m}(t)$. We do so by relating their increments to those of the previously treated processes $D_j^{\lambda,m}(t)$ and $A_i^{\lambda,m}(t)$. For $\Phi_{i,j}^{\lambda,m}(t)$ and $D_j^{\lambda,m}(t)$, it is important not to try to match the routed customers to the service of those same customers; instead we think of departures allowing new customers to be assigned to agents. In particular, we apply (17) to get, for all $i\in\mathcal{I}$, $j\in\mathcal{J}$,

$$\begin{aligned}\left|A_{i,j}^{\lambda,m}(t_2)-A_{i,j}^{\lambda,m}(t_1)\right|&\le\left|A_i^{\lambda,m}(t_2)-A_i^{\lambda,m}(t_1)\right|,\\\left|\Phi_{i,j}^{\lambda,m}(t_2)-\Phi_{i,j}^{\lambda,m}(t_1)\right|&\le\left|D_j^{\lambda,m}(t_2)-D_j^{\lambda,m}(t_1)\right|.\end{aligned}\tag{52}$$

Combining (50)–(52), we obtain

$$\begin{aligned} &P\Big\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,\,t_2\le L_k}|A_{i,j}^{\lambda,m}(t_2)-A_{i,j}^{\lambda,m}(t_1)|>N'|t_2-t_1|+\epsilon'\Big\}\le\delta', \quad i\in\mathcal{I},\ \ j\in\mathcal{J},\\ &P\Big\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,\,t_2\le L_k}|\Phi_{i,j}^{\lambda,m}(t_2)-\Phi_{i,j}^{\lambda,m}(t_1)|>N'|t_2-t_1|+\epsilon'\Big\}\le\delta', \quad i\in\mathcal{I},\ \ j\in\mathcal{J}, \end{aligned} \tag{53}$$

for all λ large enough, where $N':=\max_i\{a_i\}\vee\max_j\{\mu_j\bar{\nu}_j\}$.

Now consider the processes $R_i^{\lambda,m}(t)$, $i\in\mathcal{I}$. By construction,

$$R_i^{\lambda,m}(t)=R_i\Big(\theta_i\int_0^{(t+m)/\sqrt{\lambda}}Q_i^\lambda(s)\,ds\Big)-R_i\Big(\theta_i\int_0^{m/\sqrt{\lambda}}Q_i^\lambda(s)\,ds\Big).$$

Hence,

$$|R_i^{\lambda,m}(t_2)-R_i^{\lambda,m}(t_1)|\le\theta_i\frac{1}{\sqrt{\lambda}}\int_{(m+t_1)/\sqrt{\lambda}}^{(m+t_2)/\sqrt{\lambda}}Q_i^\lambda(s)\,ds+\sup_{t\le L_k}\Big|R_i^{\lambda,m}(t)-\frac{1}{\sqrt{\lambda}}\int_{(m+t_1)/\sqrt{\lambda}}^{(m+t_2)/\sqrt{\lambda}}Q_i^\lambda(s)\,ds\Big|.$$

On $\Omega_1^{\lambda,k}$, however, $\|\widehat{Q}^\lambda\|_{T^\lambda}^*\le k$ and

$$\sup_{t\le L_k}\Big|R_i^{\lambda,m}(t)-\frac{1}{\sqrt{\lambda}}\int_{(m+t_1)/\sqrt{\lambda}}^{(m+t_2)/\sqrt{\lambda}}Q_i^\lambda(s)\,ds\Big|\le_{st}\frac{1}{\sqrt{\lambda}}\sup_{t\le L_k/\sqrt{\lambda}}\big|R_i(\theta_i k\sqrt{\lambda}t)-\theta_i k\sqrt{\lambda}t\big|.$$

Applying Proposition 4.3 of Bramson [6] once again, we have, for all λ large enough,

$$P\Big\{\sup_{t_1,\,t_2\le L_k}|R_i^{\lambda,m}(t_2)-R_i^{\lambda,m}(t_1)|>\theta_i\frac{1}{\sqrt{\lambda}}|t_2-t_1|k+\epsilon'/2\Big\}\le\delta'/2+P\{(\Omega_1^{\lambda,k})^c\}.$$

Since $\theta_i L_k/\sqrt{\lambda}\le\epsilon'/2$ and $P\{(\Omega^{\lambda,k})^c\}\le\delta'/2$ for all λ large enough,

$$P\Big\{\sup_{t_1,\,t_2\le L_k}|R_i^{\lambda,m}(t_2)-R_i^{\lambda,m}(t_1)|>\epsilon'\Big\}\le\delta' \tag{54}$$

for all λ large enough. Now note that

$$Q_i^{\lambda,m}(t)=Q_i^{\lambda,m}(0)+A_i^{\lambda,m}(t)-\sum_{j\in\mathcal{J}}A_{i,j}^{\lambda,m}(t)-\sum_{j\in\mathcal{J}}\Phi_{i,j}^{\lambda,m}(t)-R_i^{\lambda,m}(t).$$

Let $N''=(2I+2J)N'$. Fixing $\delta''>0$ and $\epsilon''>0$, we can then choose new values of δ' and ϵ' and combine (51)–(54) to obtain

$$P\Big\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,\,t_2\le L_k}|Q_i^{\lambda,m}(t_2)-Q_i^{\lambda,m}(t_1)|>N''|t_2-t_1|+\epsilon''\Big\}\le\delta'', \quad i\in\mathcal{I}, \tag{55}$$

for all λ large enough. A similar argument is used for $Z_j^{\lambda,m}(t)$ to show that

$$P\Big\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,\,t_2\le L_k}|Z_j^{\lambda,m}(t_2)-Z_j^{\lambda,m}(t_1)|>N''|t_2-t_1|+\epsilon''\Big\}\le\delta'', \quad j\in\mathcal{J},$$

for all λ large enough. We omit this argument.

We now turn to the processes $\Theta_i^\lambda(t)$, $i\in\mathcal{I}$. Consider $\omega\in\Omega_1^{\lambda,k}$. Then, $\|\hat{X}_\Sigma^\lambda\|_{T^\lambda}^*\le k$, and the Ḧolder condition on the ratio function (see Definition 2.2) implies that there exist constants c_k and α_k, depending on k and ϵ' but not on λ, so that for all $0<t_1\le t_2\le T^\lambda$,

$$[\hat{X}_\Sigma^\lambda(t_2)]^+p_i([\hat{X}_\Sigma^\lambda(t_2)]^+)-[\hat{X}_\Sigma^\lambda(t_1)]^+p_i([\hat{X}_\Sigma^\lambda(t_1)]^+)\le c_k|\hat{X}_\Sigma^\lambda(t_2)-\hat{X}_\Sigma^\lambda(t_1)|^{\alpha_k}+\epsilon'/2;$$

see for example Equation (118) in Atar [4]. Hence,

$$\begin{aligned} &P\Big\{\max_{m<\sqrt{\lambda}T^\lambda}\sup_{t_1,\,t_2\le L_k}|\Theta_i^{\lambda,m}(t_2)-\Theta_i^{\lambda,m}(t_1)|>\epsilon'\Big\}\\ &\quad\le P\{(\Omega_1^{\lambda,k})^c\}+P\Big\{\sup_{0\le t_1<t_2\le T^\lambda:\,|t_2-t_1|\le L_k/\sqrt{\lambda}}c_k|\hat{X}_\Sigma^\lambda(t_2)-\hat{X}_\Sigma^\lambda(t_1)|^{\alpha_k}>\epsilon'/2\Big\}. \end{aligned}$$

Because we can choose k large enough so that $P\{\Omega_1^{\lambda,k}\} \geq 1-\delta'/2$ for any λ large enough, because $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$ is C-tight by Lemma 6.2, and because we can move the exponent α_k, first outside the supremum and then to the other side by raising to the reciprocal power, we can conclude that

$$P\left\{ \max_{m<\sqrt{\lambda}T^\lambda} \sup_{t_1,\, t_2 \leq L_k} |\Theta_i^{\lambda,m}(t_2) - \Theta_i^{\lambda,m}(t_1)| > \epsilon' \right\} \leq \delta'. \tag{56}$$

Combining (55) and (56) and using the fact that

$$\begin{aligned} U_i^{\lambda,m}(t) &= Q_i^{\lambda,m}(t) - \frac{1}{\sqrt{\lambda}}\Theta_i^\lambda((m+t)/\sqrt{\lambda}) \\ &= U_i^{\lambda,m}(0) + A_i^{\lambda,m}(t) - \sum_{j\in\mathcal{J}} A_{i,j}^{\lambda,m}(t) - \sum_{j\in\mathcal{J}} \Phi_{i,j}^{\lambda,m}(t) - R_i^{\lambda,m}(t) - \Theta_i^{\lambda,m}(t), \end{aligned}$$

we can choose new values of ϵ' and δ' so that

$$P\left\{ \max_{m<\sqrt{\lambda}T^\lambda} \sup_{t_1,\, t_2 \leq L_k} |U_i^{\lambda,m}(t_2) - U_i^{\lambda,m}(t_1)| > N''|t_2 - t_1| + \epsilon'' \right\} \leq \delta'' \tag{57}$$

for all λ large enough. Now set $N = 8(I + J + IJ)$. Then, we can combine (50), (51), and (53)–(57) and choose new values δ', δ'', ϵ', and ϵ'' appropriately to conclude that $P\{\Omega_4^{\lambda,k}\} \geq 1-\delta/4$ for all λ large enough. Finally, we can combine the four inequalities for $P\{\Omega_i^{\lambda,k}\}$ above to conclude that $P\{\mathcal{K}^{\lambda,k}\} \geq 1-\delta$ for some k, L_k, λ_0 and for all $\lambda \geq \lambda_0$. □

Having proved that the family of HS processes can be approximated by Lipschitz functions, we now want to establish the uniform approximation by HL functions satisfying the HM equations. For this important step we have the following lemma, which is the analog of Proposition 6.1 in Bramson [6]. The proof is exactly as in Bramson [6] and is hence omitted.

Lemma 6.4. *For all $\epsilon > 0$, there exists $k > 0$, $L_k > 0$, and λ_0, so that, for all $\lambda > \lambda_0$, $\omega \in \mathcal{K}^{\lambda,k}$, and $m < \sqrt{\lambda}T^\lambda$, there exists an HL function $\widetilde{\mathbb{X}}$ (a Lipschitz function satisfying the HM equations) such that $\|\mathbb{X}^{\lambda,m} - \widetilde{\mathbb{X}}\|_{L_k}^* \leq \epsilon$.*

Lemmas 6.3 and 6.4 combined show that, given $\epsilon > 0$ and $\delta > 0$, we can choose a set $\mathcal{K}^{\lambda,k}$ (which also depends on ϵ) with $P\{\mathcal{K}^{\lambda,k}\} > 1-\delta$, on which all the process are stochastically bounded and any HS process can be approximated by an HL function. Lemmas 4.1 and 4.2 and Proposition 4.1 of Bramson [6] imply that the approximating HL function is Lipschitz.

To establish Theorem 6.4, it remains only to show that all the hydrodynamic limits satisfy Equations (41)–(44). That is done in the following lemma. The lemma below is mostly an analogue of Proposition 6.6 in Dai and Tezcan [12]. The major difference is to show that the QIR-specific Equations (42)–(44) hold for any hydrodynamic limit $\widetilde{X}$.

Lemma 6.5. *Fix $k > 0$, $L_k > 0$, and let $\widetilde{X}$ be a hydrodynamic limit of the family of HS processes $\mathbb{X}^{\lambda,m}$ over $[0, L_k]$. Then $\widetilde{X}$ satisfies Equations* (41)–(44).

Proof. Using the definitions of the HL function and the set $\mathcal{K}^{\lambda,k}$, it is immediate that any HL function satisfies Equations (41)–(42); see the proof of Proposition 6.6 in Dai and Tezcan [12]. We turn, then, to prove that any hydrodynamic limit $\widetilde{X}$ satisfies (43).

Toward that end, recall the definition $\tilde{I}^+(t) = \{i \in \mathcal{I}\colon \widetilde{U}_i(t) > 0\}$ and consider $t \geq 0$ with $\tilde{I}^+(t) \neq \varnothing$. Let $\tilde{\epsilon}(t) := \min_{i\in\tilde{I}^+(t)} \widetilde{U}_i(t)$. Because every HL function x is absolutely continuous (and thus continuous), we must have an interval $[t-\tau, t+\tau]$ such that for all $u \in [t-\tau, t+\tau]$, $\min_{i\in\tilde{I}^+(u)} \widetilde{U}_i(u) \geq \tilde{\epsilon}(t)/2$. Moreover, by the continuity of $\widetilde{U}_i(t)$, the interval can be chosen so that $\widetilde{U}_i(u) \leq \tilde{\epsilon}(t)/8$ for all $i \notin \tilde{I}^+(t)$ and $u \in [t-\tau, t+\tau]$. Next, since $\widetilde{X}$ is an HL function, we can fix k and argue that when there exists λ large enough, $\omega \in \mathcal{K}^{\lambda,k}$ and $m < \sqrt{\lambda}T^\lambda$, so that $\|U_i^{\lambda,m} - \widetilde{U}_i\|_{L_k}^* \leq \epsilon$ for $\epsilon \leq \tilde{\epsilon}(t)/8$. In particular, we can choose L_k (and appropriately choose a new value of λ) so that there exists a neighborhood $[t-\tau', t+\tau']$ of t such that for all $u \in [t-\tau', t+\tau']$, $U_i^{\lambda,m}(u) \geq (3/8)\tilde{\epsilon}(t)$, $i \in \tilde{I}^+(t)$, and $U_i^{\lambda,m}(u) \leq (2/8)\tilde{\epsilon}(t)$, $i \notin \tilde{I}^+(t)$. Then, for any $j \in J(\tilde{I}^+(t)) = \{j \in \mathcal{J}\colon i \in I(j)$ for some $i \in \tilde{I}^+(t)\}$,

$$\arg\max_{i\in I(j)} U_i^{\lambda,m}(u) \in \tilde{I}^+(t) \tag{58}$$

for all $u \in [t-\tau', t+\tau']$. In turn, by the definition of $\Phi_{i,j}^\lambda(t)$ (see Equation (17)) we have that

$$\sum_{i\in\tilde{I}^+(t)} \sum_{j\in J(i)} (\Phi_{i,j}^{\lambda,m}(t+\tau') - \Phi_{i,j}^{\lambda,m}(t-\tau')) = \sum_{j\in J(\tilde{I}^+(t))} (D_j^{\lambda,m}(t+\tau') - D_j^{\lambda,m}(t-\tau')), \tag{59}$$

which corresponds to the fact that every service completion is followed by an admission of a customer from a class $i \in \tilde{I}^+(t)$. By definition, $\|D_j^{\lambda,m}(t) - \mu_j\bar{\nu}_j t\|^*_{L_k} \le \epsilon$ on $\mathcal{K}^{\lambda,k}$. Hence,

$$\sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)}(\Phi_{i,j}^{\lambda,m}(t+\tau') - \Phi_{i,j}^{\lambda,m}(t-\tau')) \ge \sum_{j\in J(\tilde{I}^+(t))}\mu_j\bar{\nu}_j 2\tau' - \epsilon,$$

and, since $\|\mathbb{X}^{\lambda,m} - \tilde{X}\| \le \epsilon$,

$$\sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)}(\tilde{\Phi}_{ij}(t+\tau') - \tilde{\Phi}_{ij}(t-\tau')) \ge \sum_{j\in J(\tilde{I}^+(t))}\mu_j\bar{\nu}_j 2\tau' - 2J\epsilon.$$

Because ϵ was chosen arbitrarily and the bound holds for any $\tilde{\tau} \le \tau'$,

$$\sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)} d\tilde{\Phi}_{ij}(t) \ge \sum_{j\in J(\tilde{I}^+(t))}\mu_j\bar{\nu}_j.$$

Equation (43) now follows. Finally, for Equation (44), note that

$$d\sum_{i\in\tilde{I}^+(t)}\tilde{U}_i^\lambda(t) = \sum_{i\in\tilde{I}^+(t)} d\tilde{A}_i(t) - \sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)} d\tilde{A}_{ij}(t) - \sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)} d\tilde{\Phi}_{ij}(t) \le \sum_{i\in\tilde{I}^+(t)} a_i - \sum_{j\in J(\tilde{I}^+(t))}\mu_j\bar{\nu}_j,$$

where the last inequality follows from (43). We now observe that, because of Assumptions 2.2 and 2.3, there exists $c > 0$ such that for any strict subset $B \subset \mathcal{I}$,

$$\sum_{i\in B} a_i - \sum_{i\in B}\sum_{j\in J(i)}\mu_j\bar{\nu}_j \le -c.$$

Equation (44) now follows since $\tilde{I}^+(t)$ is necessarily a strict subset of $\mathcal{I}$. We have thus established that any HL function $\tilde{X}$ satisfies (41)–(44) and the proof of the lemma is complete. □

With Lemma 6.5 we have completed the proofs of Theorems 6.4 and 6.2 for the stopped processes. Theorem 6.2, under Condition (C1), is then established by showing in the following lemma that ς^λ approaches ∞ weakly. This implies that the SSC extends to the whole interval $[0, T]$.

LEMMA 6.6. *For each $T > 0$, $P\{\varsigma^\lambda \le T\} \to 0$ as $\lambda \to \infty$.*

PROOF. By definition, $\hat{B}^\lambda(t) = \sum_{i\in\mathcal{I}}\hat{U}_i^\lambda(t) \ge 0$. Using (30), we can write

$$P\{\varsigma^\lambda \le T\} = P\{|\hat{B}^\lambda|^*_{\varsigma^\lambda\wedge T} \ge 2\hat{B}^\lambda(0)\vee 1\} = P\Big\{\sup_{0\le t\le T^\lambda}\sum_{i\in\mathcal{I}}\hat{U}_i^\lambda(t) > 2\hat{B}^\lambda(0)\vee 1\Big\}, \tag{60}$$

where we use the equivalence of the events $\{\varsigma^\lambda \le T\}$ and $\{\hat{B}^\lambda(T^\lambda) \ge 2\hat{B}^\lambda(0)\vee 1\}$ as follows from the definition of ς^λ in Equation (30). To establish that $P\{\varsigma^\lambda \le T\} \to 0$ as $\lambda \to \infty$, it suffices to prove that the right-hand side of (60) converges to 0. Toward that end, fix $\delta > 0$. Then, there exists k such that $P\{\mathcal{K}^{\lambda,k}\} \ge 1-\delta$ for all λ large enough. Fix $0 < \epsilon < 1/2$. By Theorem 6.4, for all λ large enough and $\omega \in \mathcal{K}^{\lambda,k}$, there exists $\tilde{U}_i(t)$ that satisfies (42)–(44) such that $\|U^{\lambda,0}(t) - \tilde{U}(t)\|^*_{L_k} \le \epsilon$.

By Equation (44),

$$d\sum_{i\in\tilde{I}^+(t)}\tilde{U}_i(t) \le -c$$

for some positive constant c. Consequently, for all $t \le L_k$,

$$\sum_{i\in\mathcal{I}} U_i^{\lambda,0}(t) \le \hat{B}^\lambda(0)\vee\left(\tfrac{1}{2}+\epsilon\right).$$

Since $\hat{U}_i(t/\sqrt{\lambda}) = U_i^{\lambda,0}(t)$ for all $t \le L_k$, we have $\sum_{i\in\mathcal{I}}\hat{U}_i^\lambda(t) \le 2\hat{B}^\lambda(0)\vee 1$ for all $t \le L_k/\sqrt{\lambda}$ and

$$P\left\{\sup_{0\le t\le L_k/\sqrt{\lambda}}\sum_{i\in\mathcal{I}}\hat{U}_i^\lambda(t) > 2\hat{B}^\lambda(0)\vee 1\right\} \le \delta.$$

Choosing $L_k \ge 2\lceil s^*\rceil$ with s^* as defined in the proof of Theorem 6.2 and repeating the arguments in that proof, we have

$$P\left\{\sup_{L_k/\sqrt{\lambda}\le t\le T^\lambda}\sum_{i\in\mathcal{I}}|\hat{U}_i^\lambda(t)| > \epsilon\right\} \le \delta,$$

so that

$$P\Big\{\sup_{L_k/\sqrt{\lambda}\le t\le T^\lambda}\sum_{i\in\mathcal{J}}|\widehat{U}_i^\lambda(t)|>2\hat{B}^\lambda(0)\vee 1\Big\}\le\delta.$$

We conclude by noting that

$$P\Big\{\sup_{0\le t\le T^\lambda}\sum_{i\in\mathcal{J}}\widehat{U}_i^\lambda(t)>2\hat{B}^\lambda(0)\vee 1\Big\}\le P\Big\{\sup_{0\le t\le L_k/\sqrt{\lambda}}\sum_{i\in\mathcal{J}}\widehat{U}_i^\lambda(t)>2\hat{B}^\lambda(0)\vee 1\Big\}$$
$$+P\Big\{\sup_{L_k/\sqrt{\lambda}\le t\le T}\sum_{i\in\mathcal{J}}\widehat{U}_i(t)>2\hat{B}^\lambda(0)\vee 1\Big\}\le 2\delta.$$

Because δ was arbitrary, the proof is complete. □

With Lemma 6.6, we have completed the proof of Theorem 6.2 and in turn the proof of Theorem 3.1 under Condition (C1). Specifically, we have shown that, for all $0<s<T$,

$$\limsup_{\lambda\to\infty}P\Big\{\sup_{s\le t\le T}\sum_{i\in\mathcal{J}}|\widehat{U}_i^\lambda(t)|>\epsilon\Big\}=0. \tag{61}$$

If, in addition, Equation (4) holds, then

$$\limsup_{\lambda\to\infty}P\Big\{\sup_{0\le t\le T}\sum_{i\in\mathcal{J}}|\widehat{U}_i^\lambda(t)|>\epsilon\Big\}=0. \tag{62}$$

The proof of SSC for the processes $\hat{V}_j^\lambda(t)$ follows similarly. Before turning to the proof of Theorem 3.1 under Condition (C2), we state the following corollary that will be of use in §7.

COROLLARY 6.5. *The sequences $\hat{Q}_\Sigma^\lambda(t)$, $\hat{I}_\Sigma^\lambda(t)$, and $\hat{X}_\Sigma^\lambda(t)$ are stochastically bounded; i.e.,*

$$\lim_{A\to\infty}\limsup_{\lambda\to\infty}P\{\|\hat{Q}_\Sigma^\lambda\|_T^*+\|\hat{I}_\Sigma^\lambda\|_T^*+\|\hat{X}_\Sigma^\lambda\|_T^*>A\}=0. \tag{63}$$

Also, the process $\hat{X}_\Sigma^\lambda(t)$ is C-tight.

PROOF. These results have already been proved for the stopped processes in Lemma 6.2. This additional result then follows from Lemma 6.6.

6.3. SSC under Condition (C2). In §6.2, because of the pool-dependent service rates, we could use a somewhat less detailed description of the system dynamics than the general description given in Equations (14)–(17). We now return to that general description. First we define

$$\hat{B}^\lambda(t):=\sum_{i\in\mathcal{J}}|\widehat{U}_i^\lambda(t)|\vee\sum_{i\in\mathcal{J}}|\widehat{V}_i^\lambda(t)|, \tag{64}$$

with $\widehat{U}_i^\lambda(t)$ and $\widehat{V}_i^\lambda(t)$ as defined in (16). Then we let $Z_i^\lambda(t):=\sum_{j\in\mathcal{J}}Z_{i,j}^\lambda(t)$ be the number of class i customers in service at time t and define $\hat{Z}_\Sigma^\lambda(t):=\sum_{i\in\mathcal{J}}\hat{Z}_i^\lambda(t)$. The arguments leading to (28) are immediately adapted to show that

$$\hat{X}_i^\lambda(t)=\hat{X}_i^\lambda(0)-\beta_i t-\mu_i\int_0^t\hat{Z}_i^\lambda(s)\,ds-\theta_i\int_0^t\hat{Q}_i^\lambda(s)\,ds+\widehat{M}_i^\lambda(t)+o_P(1), \tag{65}$$

where $\beta_i=\mu_i\sum_{j\in\mathcal{J}}x_{i,j}\gamma_j$ and $\widehat{M}_i^\lambda(t)$ is a square integrable martingale defined through

$$\widehat{M}_i^\lambda(t):=\widehat{M}_{A_i}^\lambda(t)-\sum_{j\in\mathcal{J}}\widehat{M}_{i,j}^\lambda(t)-\widehat{M}_{R_i}^\lambda(t).$$

Redefining $\widehat{M}_\Sigma^\lambda(t):=\sum_{i\in\mathcal{J}}\widehat{M}_i^\lambda(t)$, we have

$$\hat{X}_\Sigma^\lambda(t)=\hat{X}_\Sigma^\lambda(0)-\sum_{i\in\mathcal{J}}\beta_i t-\sum_{i\in\mathcal{J}}\mu_i\int_0^t\hat{Z}_i^\lambda(s)\,ds-\sum_{i\in\mathcal{J}}\theta_i\int_0^t\hat{Q}_i^\lambda(s)\,ds+\widehat{M}_\Sigma^\lambda(t). \tag{66}$$

The proof proceeds through the same stopping argument used in §6.2, where $T^\lambda:=\varsigma^\lambda\wedge T$ and ς^λ is defined as in (30). For the stopped processes, our proof is similar to the proof in §6.2. The main difference between the proofs for the different Conditions (C1) and (C2), is in the choice of the HS processes and the HM equations.

Once these are redefined, all the statements of the theorems, lemmas, and corollaries in §6.2 are the same in this setting with the exception, of course, of replacing Condition (C1) with Condition (C2). Moreover, once the HS processes and HM equations are redefined, all the proofs are adapted from §6.2 with only minor changes which should be clear once the required definitions are in place. Hence, we omit the detailed proofs and only make the required new definitions.

Toward that end, we extend (39) by adding, for each $i \in \mathcal{I}$ and $j \in \mathcal{J}$, the process of class i departures from pool j, given by

$$D^{\lambda}_{i,j}(t) := S_{i,j}\left(\mu_i \int_0^t Z^{\lambda}_{i,j}(s)\,ds\right),$$

as well as the process of cumulative class i departures given by $D^{\lambda}_i(t) := \sum_{j\in\mathcal{J}} D^{\lambda}_{i,j}(t)$. The hydrodynamically scaled processes are then defined as in (40) with the addition of the following: For $m < \sqrt{\lambda}T$, we define

$$D^{\lambda,m}_i(t) := \frac{1}{\sqrt{\lambda}}\left(D^{\lambda}_i\left(\frac{m}{\sqrt{\lambda}} + \frac{t}{\sqrt{\lambda}}\right) - D^{\lambda}_i\left(\frac{m}{\sqrt{\lambda}}\right)\right), \quad i \in \mathcal{I},$$

$$D^{\lambda,m}_{i,j}(t) := \frac{1}{\sqrt{\lambda}}\left(D^{\lambda}_{i,j}\left(\frac{m}{\sqrt{\lambda}} + \frac{t}{\sqrt{\lambda}}\right) - D^{\lambda}_{i,j}\left(\frac{m}{\sqrt{\lambda}}\right)\right), \quad i \in \mathcal{I}, \ \ j \in \mathcal{J},$$

$$D'^{\lambda,m}_{i,j}(t) := D^{\lambda,m}_{i,j}(t) - \frac{1}{\sqrt{\lambda}}\mu_i \int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z^{\lambda}_{i,j}(s)\,ds, \quad i \in \mathcal{I}, \ \ j \in \mathcal{J}.$$

The hydrodynamic model equations for

$$\tilde{X} := \left(\tilde{A}_i(t), \tilde{A}_{ij}(t), \tilde{\Phi}_{ij}(t), \tilde{D}_j(t), \tilde{D}_i(t), \tilde{D}'_{ij}(t), \tilde{R}_i(t), \tilde{\Theta}_i(t), \tilde{Q}_i(t), \tilde{U}_i(t), \tilde{Z}_j(t);\ i \in \mathcal{I},\ j \in \mathcal{J}\right),$$

are given by the first four equations in (41) and (42) with the addition of the following equations:

$$\tilde{Z}_j(t) = \tilde{Z}_j(0) + \sum_{i\in J(i)} \tilde{A}_{ij}(t) + \sum_{i\in J(i)} \tilde{\Phi}_{ij}(t) - \tilde{D}_j(t), \quad j \in \mathcal{J}, \tag{67}$$

$$\tilde{D}_i(t) = a_i t, \quad i \in \mathcal{I}, \tag{68}$$

$$\tilde{D}'_{ij}(t) = 0, \quad i \in \mathcal{I}, \ \ j \in \mathcal{J}, \tag{69}$$

$$\sum_{i\in\tilde{I}_+(t)} \sum_{j\in J(i)} d\tilde{\Phi}_{ij}(t) \ge \sum_{i\in I^+(t)} a_i + c_1, \tag{70}$$

$$\sum_{i\in\tilde{I}^+(t)} d\tilde{U}_i(t) \le -c_2, \quad \text{whenever } \tilde{I}^+(t) \ne \varnothing. \tag{71}$$

Where c_1 and c_2 are strictly positive constants and, as before, $\tilde{I}^+(t) = \{i \in \mathcal{I}\colon \tilde{U}_i(t) > 0\}$, and $J(\tilde{I}^+(t)) = \{j \in \mathcal{J}\colon i \in J(i), \text{ for some } i \in \tilde{I}^+(t)\}$.

We redefine

$$\Omega^{\lambda,k}_3 := \Omega^{\lambda,k}_3(\epsilon) := \left\{\omega \in \Omega\colon \max_{m<\sqrt{\lambda}T^{\lambda}} \|D^{\lambda,m}(t) - at\|^*_{L_k} \le \epsilon\right\}, \tag{72}$$

with $D^{\lambda,m}(t) = (D^{\lambda,m}_1(t), \ldots, D^{\lambda,m}_I(t))$ and add the set

$$\Omega^{\lambda,k}_5 := \Omega^{\lambda,k}_5(\epsilon) := \left\{\omega \in \Omega\colon \max_{(i,j)\in\mathcal{I}\times\mathcal{J}} \max_{m<\sqrt{\lambda}T^{\lambda}} \left\| D^{\lambda,m}_{i,j}(t) - \int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z^{\lambda}_{i,j}(s)\,ds \right\|^*_{L_k} \le \epsilon\right\}. \tag{73}$$

The other subsets $\Omega^{\lambda,k}_1$, $\Omega^{\lambda,k}_2$, and $\Omega^{\lambda,k}_4$ remain as defined before. Finally, we redefine $\mathcal{K}^{\lambda,k} = \bigcap_{i=1}^5 \Omega^{\lambda,k}_i$. With these definitions, all the statements of §6.2 hold in this section without any change, and the adaptation of the proofs are straightforward, leading to the proof of SSC under Condition (C2).

6.4. SSC under Condition (C3). In §5 we observed that, for some special tree networks and choices of the ratio vector $v(\cdot)$, QIR control is equivalent to the GQIR control in Atar [4]. Consequently, the SSC result for these networks follows from Atar [4]. Here we treat more general tree networks. We use the function G in (13).

Using the sample path construction in (14) and applying the martingale decomposition, we write

$$X^{\lambda}_i(t) = X^{\lambda}_i(0) + \lambda_i t - \sum_{j\in J(i)} \mu_{i,j}\bar{x}_{i,j}N^{\lambda}_j - \sum_{j\in J(i)} \mu_{i,j}\int_0^t (Z^{\lambda}_{i,j}(s) - \bar{x}_{i,j}N^{\lambda}_j)\,ds - \theta_i \int_0^t Q^{\lambda}_i(s)\,ds + M^{\lambda}_i(t), \tag{74}$$

where

$$M_i^\lambda(t) := M_{A_i}^\lambda(t) - \sum_{j\in J(i)} M_{i,j}^\lambda(t) - M_{R_i}^\lambda(t), \tag{75}$$

with $M_{i,j}^\lambda(t)$, $M_{A_i}^\lambda(t)$, and $M_{R_i}^\lambda(t)$ as defined in (20). Scaling by $\sqrt{\lambda}$ we have

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) + \beta_i t - \sum_{j\in J(i)} \mu_{i,j} \int_0^t \hat{Z}_{i,j}^\lambda(s)\,ds - \theta_i \int_0^t \hat{Q}_i^\lambda(s)\,ds + \hat{M}_i^\lambda(t) + o_P(1), \tag{76}$$

where

$$\beta_i := \lim_{\lambda\to\infty} \frac{\lambda_i t - \sum_{j\in\mathcal{J}} \bar{x}_{i,j} N_j^\lambda}{\sqrt{\lambda}}$$

and $\hat{M}_i^\lambda(t) := M_i^\lambda(t)/\sqrt{\lambda}$. We now apply (12) to get

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) + \beta_i t - \sum_{j\in J(i)} \mu_{i,j} \int_0^t G_{ij}(\hat{X}^\lambda(s) - \hat{Q}^\lambda(s), -\hat{I}^\lambda(s))\,ds - \theta_i \int_0^t \hat{Q}_i^\lambda(s)\,ds + \hat{M}_i^\lambda(t) + o_P(1). \tag{77}$$

We now proceed using the stopping argument we used in the previous sections. We use $\hat{B}^\lambda(t)$ defined in (29). As in the previous section we first establish SSC assuming that all the processes are stopped at the bounded stopping time $T^\lambda := \varsigma^\lambda \wedge T$ in (30). Having defined the stopping time, the proof of SSC follows the proof of §6.2 very closely. Hence, we will abbreviate all statements and proofs and provide details only when the differences are significant. The most significant modification is in the proof of stochastic boundedness and C-tightness as stated in Lemma 6.7 below. Hence, we first state and prove this lemma. We then proceed to outline the required modifications to the arguments in §6.2.

Toward this end, recall that $\hat{X}^\lambda(t) := (\hat{X}_1^\lambda(t), \ldots, \hat{X}_I^\lambda(t))$ and $\hat{Z}^\lambda(t) := (\hat{Z}_{i,j}^\lambda(t);\ i\in\mathcal{I},\ j\in\mathcal{J})$.

Lemma 6.7 (Stochastic Boundedness of Scaled Queueing Processes). *The sequences* $\{\hat{Q}_\Sigma^\lambda(t\wedge T^\lambda)\}$, $\{\hat{I}_\Sigma^\lambda(t\wedge T^\lambda)\}$, $\{\hat{X}^\lambda(t\wedge T^\lambda)\}$, *and* $\{\hat{Z}^\lambda(t\wedge T^\lambda)\}$ *are stochastically bounded. Also, the sequence of processes* $\{\hat{\bar{X}}_i^\lambda(t\wedge T^\lambda)\colon \lambda>0\}$ *is C-tight.*

Proof. First note that, by the linearity of $G(\cdot,\cdot)$ in (13), we have

$$\begin{aligned} G\big(\hat{X}^\lambda(t) - \hat{Q}^\lambda(t), -\hat{I}^\lambda(t)\big) &= G\big(\hat{X}^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p([\hat{X}_\Sigma^\lambda(t)]^+), -[\hat{X}_\Sigma^\lambda(t)]^- v([\hat{X}_\Sigma^\lambda(t)]^-)\big) \\ &\quad - G\big(\hat{Q}^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p([\hat{X}_\Sigma^\lambda(t)]^+), \hat{I}^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^- v([\hat{X}_\Sigma^\lambda(t)]^-)\big), \end{aligned} \tag{78}$$

where

$$[\hat{X}_\Sigma^\lambda(t)]^+ p([\hat{X}_\Sigma^\lambda(t)]^+) := \big([\hat{X}_\Sigma^\lambda(t)]^+ p_1([\hat{X}_\Sigma^\lambda(t)]^+), \ldots, [\hat{X}_\Sigma^\lambda(t)]^+ p_I([\hat{X}_\Sigma^\lambda(t)]^+)\big),$$

and $[\hat{X}_\Sigma^\lambda(t)]^- v([\hat{X}_\Sigma^\lambda(t)]^-)$ is defined similarly from the individual coordinates. Using Equation (77) as well as the linearity of $G(\cdot,\cdot)$ we then have

$$\|\hat{X}^\lambda\|_t^* \le \|\hat{X}^\lambda(0)\| + c_1 T + c_2 \int_0^t (\|\hat{X}^\lambda\|_s^* + \|\hat{B}^\lambda\|_s^*)\,ds + \|\hat{M}_i^\lambda\|_t^*,$$

for some positive constants c_1 and c_2. Applying Gronwall's inequality (Lemma 4.1 in Pang et al. [23]), we have

$$\|\hat{X}^\lambda(t\wedge T^\lambda)\|_T^* \le c_1(\|\hat{X}^\lambda(0)\| + T + \|\hat{B}^\lambda(\cdot\wedge T^\lambda)\|_T^* + \|\hat{M}_i^\lambda(\cdot\wedge T^\lambda)\|_T^*)e^{c_2 T},$$

for some redefined positive constants c_1 and c_2. By Lemma 6.1, $\hat{M}_i^\lambda(t)$ is stochastically bounded and, by the definition of T^λ and Equation (31), $\hat{B}^\lambda(t\wedge T^\lambda)$ is stochastically bounded. Here we also use the fact that $\hat{B}^\lambda(0)$ is itself stochastically bounded by the assumed convergence of $\hat{X}^\lambda(0)$.

The stochastic boundedness of $\hat{X}^\lambda(t\wedge T^\lambda)$ now follows from it being bounded by a sum of stochastically bounded sequences. Next, $\hat{I}_\Sigma^\lambda(t\wedge T^\lambda)$ and $\hat{Q}_\Sigma^\lambda(t\wedge T^\lambda)$ are stochastically bounded by applying (33) and (34), which still hold with the new definition of $\hat{B}^\lambda$. Then $\hat{Z}^\lambda(\cdot\wedge T^\lambda)$ is stochastically bounded because (under Condition (C3)) it involves a linear mapping of $\hat{Q}$, $\hat{I}$, and $\hat{X}$ (see Equation (12)). The result of the proposition now follows because the sum of stochastically bounded sequences is itself stochastically bounded. Finally, the proof of the C-tightness of $\hat{X}^\lambda(t\wedge T^\lambda)$ follows very similarly to the proof of Lemma 6.2 by applying Gronwall's inequality and making the obvious required modifications. □

Having Lemma 6.7, the proof of SSC follows the proof in §6.2 for Condition (C1) very closely. However, some modifications to the hydrodynamically scaled processes and hydrodynamic limit functions are required. We now outline these modifications:

(i) Augment the state descriptor $\mathbb{X}^\lambda(t)$ by adding the coordinates $Z^\lambda_{i,j}(t)$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Also, replace the processes $D^\lambda_j(t)$ with the processes $D^\lambda_{i,j}(t)$ defined to be the number of service completions of class i customers with pool j agents. Namely,

$$D^\lambda_{i,j}(t) := S_{i,j}\left(\mu_{i,j}\int_0^t Z^\lambda_{i,j}(s)\,ds\right).$$

(ii) Augment the hydrodynamically scaled process $\mathbb{X}^{\lambda,m}(t)$ by adding the processes $Z^{\lambda,m}_{i,j}(t)$ defined by

$$Z^{\lambda,m}_{i,j}(t) := \frac{1}{\sqrt{\lambda}}\left(Z^\lambda_{i,j}\left(\frac{t}{\sqrt{\lambda}}+\frac{m}{\sqrt{\lambda}}\right)-\bar{x}_{i,j}N^\lambda_j\right).$$

Also, replace the process $D^{\lambda,m}_j(t)$ with the processes $D^{\lambda,m}_{i,j}(t)$ defined in the obvious way from $D^\lambda_{i,j}(t)$.

(iii) Augment the hydrodynamic limit function $\tilde{X}$ by adding $\tilde{Z}_{i,j}(t)$ for each $i \in \mathcal{I}$ and $j \in \mathcal{J}$ and replacing $\tilde{D}_j(t)$ with $\tilde{D}_{i,j}(t)$.

(iv) Replace the last equation in (41) with

$$\tilde{Z}_{i,j}(t) = \tilde{Z}_{i,j}(0) + \sum_{i\in\mathcal{I}}\tilde{A}_{ij}(t) + \sum_{i\in\mathcal{I}}\tilde{\Phi}_{ij}(t) - \mu_{i,j}\bar{x}_{i,j}\bar{\nu}_j t, \quad i \in \mathcal{I},\ j \in \mathcal{J}.$$

(v) Replace Equation (43) with

$$\sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)} d\tilde{\Phi}_{ij}(t) = \sum_{j\in J(\tilde{I}^+(t))}\mu_{i,j}\bar{x}_{i,j}\bar{\nu}_j.$$

Equation (44) would then follow by noting that the tree assumption implies that given a set $B \subset \mathcal{I}$, $\sum_{i\in\mathcal{I}} a_i - \sum_{i\in B, j\in\mathcal{J}(i)}\mu_{i,j}\bar{x}_{i,j}\bar{\nu}_j < -c$, for some constant $c > 0$.

(vi) Redefine

$$\Omega^{\lambda,k}_3 := \Omega^{\lambda,k}_3(\epsilon) := \left\{\omega\in\Omega\colon \max_{m<\sqrt{\lambda}T^\lambda}\|D^{\lambda,m}(t) - \mu\bar{x}\bar{\nu}t\|^*_{L_k} \le \epsilon\right\}, \tag{79}$$

where now $D^{\lambda,m}(t) := (D^\lambda_{i,j}(t);\ i \in \mathcal{I},\ j \in \mathcal{J})$ and $\mu\bar{x}\bar{\nu} = (\mu_{i,j}\bar{x}_{i,j}\bar{\nu}_j;\ i \in \mathcal{I},\ j \in \mathcal{J})$.

(vii) In the proof of Lemma 6.3 one replaces the argument for $D^{\lambda,m}_j(t)$ with the corresponding argument for $D^{\lambda,m}_{i,j}(t)$. The arguments follow very similarly with the exception of the following additional step:

$$\sup_{t_1,t_2\le L_k}|D^{\lambda,m}_{i,j}(t_2) - D^{\lambda,m}_{i,j}(t_1)| \le \frac{1}{\sqrt{\lambda}}\sup_{t_1,t_2\le L_k}\left|S_{i,j}\left(\mu_{i,j}\int_0^{t_2/\sqrt{\lambda}} Z^\lambda_{i,j}(s)\,ds\right) - S_{i,j}\left(\mu_{i,j}\int_0^{t_1/\sqrt{\lambda}} Z^\lambda_{i,j}(s)\,ds\right)\right|.$$

In the proof of Lemma 6.3 we use the stochastic boundedness of the process $\hat{Z}^\lambda_j(t)$. Similarly, one should now use Lemma 6.7 and, in particular, the stochastic boundedness of the sequence $\{\hat{Z}^\lambda(\cdot\wedge T^\lambda)\}$.

(viii) Equation (59) should be replaced with

$$\sum_{i\in\tilde{I}^+(t)}\sum_{j\in J(i)}(\Phi^{\lambda,m}_{i,j}(t+\tau') - \Phi^{\lambda,m}_{i,j}(t-\tau')) = \sum_{j\in J(\tilde{I}^+(t))}\sum_{k\in\mathcal{I}}(D^{\lambda,m}_{k,j}(t+\tau') - D^{\lambda,m}_{k,j}(t-\tau')).$$

With the above modifications, one can easily follow §6.2 to construct the complete proof of SSC under Condition (C3).

7. Proofs of the results in §4.

Proof of Theorem 4.1. The limit for $\hat{X}^\lambda_\Sigma$ is obtained through an application of the continuous mapping theorem, e.g., Theorem 3.4.1 of Whitt [30]. In particular, by SSC, we have $\|\hat{X}^\lambda_\Sigma - \hat{Y}^\lambda_\Sigma\| \Rightarrow 0$ in D_- as $\lambda\to\infty$, where

$$\hat{Y}^\lambda_\Sigma(t) = \hat{X}^\lambda_\Sigma(0) - \sum_{j\in\mathcal{J}}\mu_j\gamma_j t + \sum_{j\in\mathcal{J}}\mu_j\int_0^t \tilde{v}_j([\hat{X}^\lambda_\Sigma(s)]^-)\,ds - \sum_{i\in I}\theta_i\int_0^t \tilde{p}_i([\hat{X}^\lambda_\Sigma(s)]^+)\,ds + \hat{M}^\lambda_\Sigma(t),$$

and $\hat{X}^\lambda_\Sigma$ has the representation given in (28). Applying Corollary 6.5, we have that both $\hat{X}^\lambda_\Sigma(t)$ and $\hat{Y}^\lambda_\Sigma(t)$ are C-tight. Consequently, the asymptotic equivalence above in D_- extends to D, and we can write

$$\hat{X}^\lambda_\Sigma(t) = \hat{X}^\lambda_\Sigma(0) - \sum_{j\in\mathcal{J}} \mu_j\gamma_j t + \sum_{j\in\mathcal{J}} \mu_j \int_0^t \tilde{v}_j([\hat{X}^\lambda_\Sigma(s)]^-)\,ds - \sum_{i\in I} \theta_i \int_0^t \tilde{p}_i([\hat{X}^\lambda_\Sigma(s)]^+)\,ds + \widehat{M}^\lambda_\Sigma(t) + o_P(1). \tag{80}$$

By Theorem 4.1 in Pang et al. [23], this integral representation for $\hat{X}^\lambda_\Sigma(t)$ is a measurable continuous mapping from D to itself. In particular, we can get the convergence of $\hat{X}^\lambda_\Sigma(t)$ from the convergence of the basic martingale $\widehat{M}^\lambda_\Sigma(t)$ established in the following lemma.

LEMMA 7.1. *Under the conditions of Theorem* 4.1, $\widehat{M}^\lambda_\Sigma(t) \Rightarrow \sqrt{2}B(t)$ *in* D *as* $\lambda\to\infty$, *where* $\{B(t), t\ge 0\}$ *is a standard Brownian motion.*

The proof of Lemma 7.1 is postponed to the end of this section. We can now apply the continuous-mapping theorem to (80) and use Lemma 7.1 to get the convergence of $\hat{X}^\lambda_\Sigma(t)$. By applying the continuous-mapping theorem again, we can then extend the limit to the vector process

$$\left(\hat{X}^\lambda_\Sigma(t), \tilde{p}_1([\hat{X}^\lambda_\Sigma(t)]^+), \ldots, \tilde{p}_I([\hat{X}^\lambda_\Sigma(t)]^+), \frac{1}{a_1}\tilde{p}_1([\hat{X}^\lambda_\Sigma(t)]^+), \ldots, \right.$$
$$\left. \frac{1}{a_I}\tilde{p}_I([\hat{X}^\lambda_\Sigma(t)]^+), \tilde{v}_1([\hat{X}^\lambda_\Sigma(t)]^-), \ldots, \tilde{v}_J([\hat{X}^\lambda_\Sigma(t)]^-)\right) \quad \text{in } D^{2I+J+1}.$$

Then we can apply the convergence-together theorem again to show that this process has the same limit as the process $(\hat{X}^\lambda_\Sigma(t), \hat{Q}^\lambda_1(t), \ldots, \hat{Q}^\lambda_I(t), \widehat{W}^\lambda_1(t), \ldots, \widehat{W}^\lambda_I(t), \hat{I}^\lambda_1(t), \ldots, \hat{I}^\lambda_J(t))$ on D^{2I+J+1}_-. First, the SSC establishes the connection for the queue-length processes. Then we can apply Puhalskii's [25] first-passage-time argument to extend the result from the queue lengths to treat the waiting times as well; see the corollary in Puhalskii [25], §13.7 of Whitt [30] (especially Theorem 13.7.4), and Corollary B.3 in the appendix of Gurvich et al. [18]. Some care is needed in applying the first-passage-time argument in the model with abandonments. The complete detailed argument follows the proof of Theorem 4.1 in Talreja and Whitt [27] involving carefully constructed lower and upper bounds. We omit the complete argument and refer the reader to Talreja and Whitt [27]. We conclude this proof by noting that Equation (6) has a unique (possibly weak) solution by virtue of established results for one-dimensional SDE's; see, e.g., Remark 5.5.19 and Exercise 5.5.38 in Karatzas and Shreve [20].

PROOF OF COROLLARY 4.2. The proof follows an interchange-of-limits argument, following Halfin and Whitt [19]. It is easy to check (using, for example, Browne and Whitt [7]) that, with the conditions of the corollary, the diffusion process $\hat{X}_\Sigma(t)$, as given in Theorem 4.1, has a unique stationary distribution coinciding with the distribution of $\hat{X}_\Sigma(\infty)$. Now, by Prohorov's Theorem, the tight sequence $\hat{X}^\lambda_\Sigma(\infty)$ has a convergent subsequence $\hat{X}^{\lambda^k}_\Sigma(\infty)$. Let $\hat{X}^{\lambda^k}_\Sigma(0)$ be distributed as $\hat{X}^{\lambda^k}_\Sigma(\infty)$. Then $\hat{X}^{\lambda^k}_\Sigma(t)$ is a strictly stationary process and since we already proved that $\hat{X}^{\lambda^k}_\Sigma(t) \Rightarrow \hat{X}_\Sigma(t)$, $\hat{X}_\Sigma(t)$ will be a process with $\hat{X}_\Sigma(0)$ having the distribution of the limit of $\hat{X}^{\lambda^k}_\Sigma(\infty)$. However, since $\hat{X}^{\lambda^k}_\Sigma(t)$ is stationary for each k, so is $\hat{X}_\Sigma(t)$. Hence, the limit of $\hat{X}^{\lambda^k}_\Sigma(\infty)$ must be the unique stationary distribution of $\hat{X}_\Sigma(t)$. The same argument applies to any convergent subsequence and hence the sequence $\hat{X}^\lambda_\Sigma(\infty)$ itself must converge to this limit (see Theorem 2.3 in Billingsley [5]). The convergence of the moments now follows from uniform integrability. Since $\hat{Q}^\lambda_i(\infty) \le \hat{Q}^\lambda_\Sigma(\infty)$ almost surely, the sequence $\hat{Q}^\lambda_i(\infty)$ is also uniformly integrable for all $i\in\mathcal{I}$. Hence, $E[\hat{Q}^\lambda_i(\infty)] \to E[\tilde{p}_i([\hat{X}_\Sigma(\infty)]^+)]$. Recalling that $\lambda_i = a_i\lambda$, and using Little's law, we obtain $E[\widehat{W}^\lambda_i(\infty)] \Rightarrow (1/a_i)E[\tilde{p}_i([\hat{X}_\Sigma(\infty)]^+)]$.

PROOF OF THEOREM 4.3. The proof is very similar to that of Theorem 4.1 and we only outline the differences. First, similarly to Lemma 7.1, one proves that

$$(\widehat{M}^\lambda_1(t), \ldots, \widehat{M}^\lambda_I(t)) \;\Rightarrow\; \sqrt{2a}B(t), \quad \text{in } D^I, \;\text{ as } \lambda\to\infty, \tag{81}$$

where $\sqrt{2a}B(t) \equiv (\sqrt{2a_1}B_1(t), \ldots, \sqrt{2a_I}B_I(t))$ and $B_i(t)$, $i\in\mathcal{I}$, are independent Brownian motions. Using Equation (65) and replacing $\hat{Z}^\lambda_i(t) = \hat{X}^\lambda_i(t) - \hat{Q}^\lambda_i(t)$, one then applies the SSC from Theorem 3.1 and the continuous mapping theorem, as in the proof of Theorem 4.1, to get the required convergence.

PROOF OF THEOREM 4.4. Again, the proof is very similar to that of Theorem 4.1 and we only outline the differences. First, as in the proof of Theorem 4.3, one proves that (81) holds, but here with $\widehat{M}^\lambda_i(t)$, $i\in\mathcal{I}$, as defined in (75). Using the representation in (77), one then applies the SSC from Theorem 3.1 and the continuous mapping theorem, as in the proof of Theorem 4.1, to get the required convergence. In applying the continuous mapping theorem we use the linearity of the mapping $G(\cdot,\cdot)$ on the domain D_G; see §5.

Proof of Lemma 7.1. Let $\widehat{M}_A^\lambda(t) = (\widehat{M}_{A_1}^\lambda(t), \ldots, \widehat{M}_{A_I}^\lambda(t))$, $\widehat{M}_S^\lambda(t) = (\widehat{M}_1^\lambda(t), \ldots, \widehat{M}_J^\lambda(t))$, and $\widehat{M}_R^\lambda(t) = (\widehat{M}_{R_1}^\lambda(t), \ldots, \widehat{M}_{R_I}^\lambda(t))$. Then, we prove that

$$(\widehat{M}_A^\lambda(t), \widehat{M}_S^\lambda(t), \widehat{M}_R^\lambda(t)) \;\Rightarrow\; (\sqrt{a}B_A(t), \sqrt{\mu\bar{\nu}}B_S^\lambda(t), 0), \quad \text{in } D^{2I+J}, \text{ as } \lambda \to \infty. \tag{82}$$

Here $\hat{B}_A(t)$, $\hat{B}_S(t)$ are, respectively, I and J independent Brownian motions. Also $a := (a_1, \ldots, a_I)$, $\mu\bar{\nu} := (\mu_1\bar{\nu}_1, \ldots, \mu_J\bar{\nu}_J)$, and the square root and the vector product are interpreted componentwise. Also, 0 in (82) is the 0 vector in $\mathbb{R}^I$. Recalling that

$$\widehat{M}_\Sigma^\lambda = \sum_{i\in\mathcal{I}} \widehat{M}_{A_i}^\lambda - \sum_{j\in\mathcal{J}} \widehat{M}_j^\lambda - \sum_{i\in\mathcal{I}} \widehat{M}_{R_i}^\lambda,$$

the result of the lemma then follows from the continuity of the addition operator under continuous limits (see Theorem 12.7.1 in Whitt [30]).

We turn, then, to the proof of (82). The proof is based on a functional central limit theorem for Poisson processes and on a random-time change argument. Specifically, recall that

$$\widehat{M}_{A_i}^\lambda(t) = \frac{A_i(\lambda_i t) - \lambda_i t}{\sqrt{\lambda}}, \qquad \widehat{M}_j^\lambda(t) = \frac{S_j\big(\mu_j \int_0^t Z_j^\lambda(s)\,ds\big) - \mu_j \int_0^t Z_j^\lambda(s)\,ds}{\sqrt{\lambda}}, \quad i \in \mathcal{I}, \;\; j \in \mathcal{J},$$

and

$$\widehat{M}_{R_i}^\lambda(t) = \frac{R_i\big(\theta_i \int_0^t Q_i^\lambda(s)\,ds\big) - \theta_i \int_0^t Q_i^\lambda(s)\,ds}{\sqrt{\lambda}}, \quad i \in \mathcal{I}.$$

Define

$$\widetilde{M}_{A_i}^\lambda(t) = \frac{A_i(\lambda t) - \lambda t}{\sqrt{\lambda}}, \qquad \widetilde{M}_j^\lambda(t) = \frac{S_j(\lambda t) - \lambda t}{\sqrt{\lambda}}, \quad i \in \mathcal{I}, \;\; j \in \mathcal{J},$$

and

$$\widetilde{M}_{R_i}^\lambda(t) = \frac{R_i(\lambda t) - \lambda t}{\sqrt{\lambda}}, \quad i \in \mathcal{I},$$

and let $\widetilde{M}_A^\lambda(t)$, $\widetilde{M}_S^\lambda(t)$, and $\widetilde{M}_R^\lambda(t)$ be the corresponding vector-valued processes. Then, as $A_i(\cdot)$, $R_i(\cdot)$, and $S_i(\cdot)$ are independent unit-rate Poisson processes, we have that

$$(\widetilde{M}_A^\lambda(t), \widetilde{M}_S^\lambda(t), \widetilde{M}_R^\lambda(t)) \;\Rightarrow\; (B_A(t), B_S(t), B_R(t)), \quad \text{in } D^{2I+J}, \text{ as } \lambda \to \infty,$$

where B_A and B_R are independent I-dimensional standard Brownian motions and B_S is a J-dimensional standard Brownian motion (see, for example, Theorem 5.1 in Pang et al. [23]).

The last step of the proof is to apply a random-time change argument. Let

$$\Psi_{S_j}^\lambda(t) := \frac{\mu_j \int_0^t Z_j^\lambda(s)\,ds}{\lambda}, \quad j \in \mathcal{J}, \qquad \text{and} \qquad \Psi_{R_i}^\lambda(t) := \frac{\theta_i \int_0^t Q_i^\lambda(s)\,ds}{\lambda}, \quad i \in \mathcal{I},$$

and $\Psi_{A_i}^\lambda(t) := \lambda_i/\lambda t = a_i t$, $i \in \mathcal{I}$. Let Ψ_A^λ, Ψ_S^λ, and Ψ_R^λ be the corresponding vector-valued processes. By Corollary 6.5, we have that $\hat{I}_j^\lambda(t)$ and $\hat{Q}_i^\lambda(t)$ are stochastically bounded processes for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. This implies that

$$\left(\frac{I_1^\lambda(t)}{\lambda}, \ldots, \frac{I_J^\lambda(t)}{\lambda}, \frac{Q_1^\lambda(t)}{\lambda}, \ldots, \frac{Q_I^\lambda(t)}{\lambda}\right) \;\Rightarrow\; \eta, \quad \text{in } D^{I+J}, \text{ as } \lambda \to \infty,$$

where $\eta(t) \equiv (0, 0, \ldots, 0)$ (see, for example, Lemma 5.10 in Pang et al. [23]). In particular, since $Z_j^\lambda(t) = N_j^\lambda - I_j^\lambda(t)$, we have that

$$\left(\frac{Z_1^\lambda(t)}{\lambda}, \ldots, \frac{Z_J^\lambda(t)}{\lambda}, \frac{Q_1^\lambda(t)}{\lambda}, \ldots, \frac{Q_I^\lambda(t)}{\lambda}\right) \;\Rightarrow\; \eta', \quad \text{in } D^{I+J}, \text{ as } \lambda \to \infty,$$

where $\eta'(t) \equiv (\mu_1\bar{\nu}_1, \ldots, \mu_j\bar{\nu}_j, 0, \ldots, 0)$. We can then apply the continuous mapping theorem with the integral mapping

$$(x_1, \ldots, x_j, y_1, \ldots, y_I) \mapsto \left(\mu_1 \int_0^t x_1(s)\,ds, \ldots, \mu_J \int_0^t x_J(s)\,ds, \theta_1 \int_0^t y_1(s)\,ds, \ldots, \theta_I \int_0^t y_I(s)\,ds\right),$$

to show that

$$(\Psi_A^\lambda(t), \Psi_S^\lambda(t), \Psi_R^\lambda(t)) \;\Rightarrow\; (a, \mu\bar{\nu}t, 0), \quad \text{in } D^{I+J}, \text{ as } \lambda \to \infty.$$

Finally, applying the random-time change theorem (see Theorem 13.2.1 in Whitt [30]), we conclude that

$$\big(\widetilde{M}_A^\lambda(\Phi_A^\lambda(t)), \widetilde{M}_S^\lambda(\Phi_S^\lambda(t)), \widetilde{M}_R^\lambda(\Phi_A^\lambda(t))\big) \;\Rightarrow\; \big(\sqrt{a}B_A(t), \sqrt{\mu\bar{\nu}}B_S^\lambda(t), 0\big), \quad \text{in } D^{2I+J}, \text{ as } \lambda\to\infty.$$

Because, by definition,

$$(\widehat{M}_A^\lambda(t), \widehat{M}_S^\lambda(t), \widehat{M}_R^\lambda(t)) = (\widetilde{M}_A^\lambda(\Phi_A^\lambda(t)), \widetilde{M}_S^\lambda(\Phi_S^\lambda(t)), \widetilde{M}_R^\lambda(\Phi_A^\lambda(t))),$$

the proof is complete.

8. Proofs of the results in §5.

Proof of Lemma 5.1. Fix the sample paths of $A_i(\cdot)$, $i\in\mathcal{I}$, $S_{i,j}(\cdot)$, $i\in\mathcal{I}$, $j\in\mathcal{J}$, and $R_i(\cdot)$, $i\in\mathcal{I}$. We will show that $(X^{\mathrm{GQIR}}(t), Z^{\mathrm{GQIR}}(t))$ and $(X^{\mathrm{QIR}}(t), Z^{\mathrm{QIR}}(t))$ then have the same sample paths. This, in turn, establishes equivalence in distribution.

Recall that j^* is the only agent type with $|I(j^*)|>1$. Fix the vector $v=e_{j^*}$ and note that, by definition, $\sum_{i\in\mathcal{I}}\check{Z}_{i,k}^\lambda(t)=0$, for every $k\neq j^*$ (see Equations (11) and (13)). Moreover, because by Condition (C3), the set $I(k)$ consists of a single class for all $k\neq j^*$, we have that $\check{Z}_{i,k}^\lambda(t)=0$, for all $k\neq j^*$ and $i=I(k)$. This also implies that $-\hat{Z}_{i,k}^\lambda(t)=\hat{I}_k^\lambda\geq 0$, for $k\neq j^*$ and $i=I(k)$. In particular, for all $t\geq 0$ and all $k\neq j^*$ and $i=I(k)$ we have that

$$(\check{Z}_{i,k}^\lambda(t)-\hat{Z}_{i,k}^\lambda(t))^+=\hat{I}_k^\lambda(t). \tag{83}$$

Now, fix a class i, with $Q_i^\lambda(t)=0$. Then, as $\check{Z}_{i,k}^\lambda(t)=0$, for all $k\neq j^*$, we have that

$$\check{Z}_{i,j^*}^\lambda(t)=\hat{X}_i^\lambda(t)\leq\hat{X}_i^\lambda(t)-\sum_{k\neq j^*}\hat{Z}_{i,k}^\lambda(t)=\hat{Z}_{i,j^*}^\lambda(t),$$

and in particular that

$$(\check{Z}_{i,j^*}^\lambda(t)-\hat{Z}_{i,j^*}^\lambda(t))^+=0. \tag{84}$$

Combining (83) and (84), we see that, upon arrival of a class i customer, if there are any idle agents in pool $k\neq j^*$ (and in particular $Q_i^\lambda(t-)=0$), then the decision rule of GQIR is equivalent to routing the customer to agent pool k with $k\in\arg\max_{k\in J(i),\,k\neq j^*}\hat{I}_k^\lambda(t)$. If the only idle agents are in pool j^*, then route the customer to pool j^*.

On the other hand, under Condition (C3), QIR is such that, whenever $\hat{I}_{j^*}(t)>0$, we must have $Q_i^\lambda(t)=0$, $i\in\mathcal{I}$, and $\hat{I}_\Sigma^\lambda(t)=[\hat{X}_\Sigma^\lambda(t)]^-$, by the definition of $\hat{X}_\Sigma^\lambda(t)$. Hence, whenever $\hat{I}_{j^*}^\lambda(t)>0$, we also have $\hat{I}_{j^*}^\lambda(t)\leq\hat{I}_\Sigma^\lambda(t)=[\hat{X}_\Sigma^\lambda(t)]^-$. In particular,

$$j^*=\arg\max_{j\in J(i),\,\hat{I}_j^\lambda(t)>0}\{\hat{I}_j^\lambda(t)-[\hat{X}_\Sigma^\lambda(t)]^-v_j([\hat{X}_\Sigma^\lambda(t)]^-)\},$$

only if $\hat{I}_k^\lambda(t)=0$ for all $k\neq j^*$. Evidently then, the decision rule under QIR reduces to the one given above for GQIR; i.e., route the customer to agent pool k with $k\in\arg\max_{k\in J(i),\,k\neq j^*}\hat{I}_k^\lambda(t)$. Route the customer to pool j^* only if the only idle agents are in pool j^*.

We now show that the decision rule in a service completion epoch is the same under both controls. Trivially, the decision rules are the same under QIR and GQIR when the service completion is in agent pool $k\neq j^*$ because, by Condition (C3), these pools serve a single queue each. Now consider a service completion epoch in pool j^*. Since $v=e_j^*$, for all i such that $\hat{Q}_i^\lambda(t)>0$, we must have that $\hat{Z}_{i,k}^\lambda(t)=0$ (otherwise $\hat{Q}_i^\lambda(t)=0$) for all $k\neq j^*$ and, in particular, $\hat{Z}_{i,j^*}(t)=\hat{X}_i^\lambda(t)-\hat{Q}_i^\lambda(t)-\sum_{k\neq j^*}\hat{Z}_{i,k}^\lambda(t)=\hat{X}_i^\lambda(t)-\hat{Q}_i^\lambda(t)$. Using Equations (11) and (13) as before, we also have $\check{Z}_{i,k}^\lambda(t)=0$ for all $k\in J(i)$, $k\neq j^*$, and $\check{Z}_{i,j^*}(t)=\hat{X}_i^\lambda(t)-[\hat{X}_\Sigma^\lambda(t)]^+p_i([\hat{X}_\Sigma^\lambda(t)]^+)$. Hence, upon service completion in pool j^*,

$$i\in\arg\max_{k:\ \hat{Q}_k^\lambda(t)>0}\big(\hat{Q}_k(t)-[\hat{X}_\Sigma^\lambda(t)]^+p_k([\hat{X}_\Sigma^\lambda(t)]^+)\big)$$

if and only if

$$i\in\arg\max_{k:\ \hat{Q}_k^\lambda(t)>0}\big(\check{Z}_{k,j^*}^\lambda(t)-\hat{Z}_{k,j^*}^\lambda(t)\big)^+.$$

Because we use the same decision rule for breaking ties, both controls will make the same decision in a service completion epoch. We have shown equivalence of the decision rules of QIR and GQIR under Condition (C3). By induction, this, in turn, implies that we will have the same sample paths under both controls.

Proof of Theorem 5.1. We will explain why SSC under these conditions is a consequence of Proposition 1 in Atar [4]. First note that their results are given for a Markov control policy (see Definition 4 in Atar [4]) given by a function $h := (h_1, h_2)$, where $h_i\colon \mathbb{R}^I \mapsto \mathbb{U}$, $i = 1, 2$,

$$\mathbb{U} := \left\{(u, v) \in \mathbb{R}^{I+J}\colon u_i, v_j \geq 0,\ i \in \mathcal{I},\ j \in \mathcal{J}, \sum_{i\in\mathcal{I}} u_i = \sum_{j\in\mathcal{J}} v_j = 1\right\},$$

and the functions h_i are assumed to be locally Hölder continuous away from 0 (with 0 being here the origin of $\mathbb{R}^I$); see part (iii) of Theorem 2 in Atar [4]. Clearly, these conditions apply to a pair of admissible state-dependent ratio functions p and v, as defined in Definition 2.2. Indeed, with two such functions p and v, we can define h for $x \in \mathbb{R}$ by

$$h(x) := \left(p\left(\left[\sum_{i\in\mathcal{I}} x_i\right]^+\right), v\left(\left[\sum_{i\in\mathcal{I}} x_i\right]^-\right)\right),$$

and position ourselves in the framework of Atar [4]. To be able to apply the result (Atar [4]) directly, one additional observation is required. The control proposed in Atar [4] is not precisely GQIR as defined in Definition 5.1. Instead, it is a modification of this control in which the function $h(\cdot)$ is replaced by a different function for all times that are greater than a certain stopping time; see Equation (56) in Atar [4]. However, a careful reading of Atar [4] reveals that, while this modification is required for the large-time estimates in Proposition 2 of Atar [4], it is not used in the proof of Proposition 1 in Atar [4]. Consequently, Proposition 1 in Atar [4] is valid for GQIR as defined in Definition 5.1 and it implies the desired SSC result. Indeed, by translation of notation, the first statement in part (iv) of Proposition 1 in Atar [4] corresponds to the statement

$$\sup_{s\leq u\leq t}\left\{\sum_{i\in\mathcal{I}}\left|\hat{Q}_i^\lambda(u) - [\hat{X}_\Sigma^\lambda(u)]^+ p_i([\hat{X}_\Sigma^\lambda(u)]^+)\right| + \sum_{j\in\mathcal{J}}\left|\hat{I}_j^\lambda(u) - [\hat{X}_\Sigma^\lambda(u)]^- v_j([\hat{X}_\Sigma^\lambda(u)]^-)\right|\right\} \Rightarrow 0; \tag{85}$$

see the definition of $J^n(t)$ in Equations (74), (52), and (53) of Atar [4]. The second part of (iv) corresponds to the statement

$$\sup_{s\leq u\leq t}\{\hat{Q}_\Sigma^\lambda(u) \wedge \hat{I}_\Sigma^\lambda(t)\} \Rightarrow 0; \tag{86}$$

see the definition of $M^n(t)$ in Equation (61) of Atar [4]. But, by our definition of $\hat{X}_\Sigma^\lambda(t)$, Equations (85) and (86) combined imply the state-space-collapse conclusion in Theorem 5.1. Finally, Equation (5) follows directly from part (i) of Proposition 1 in Atar [4].

Acknowledgments. This research is based on the first author's doctoral dissertation at Columbia University. The second author was supported by NSF Grant DMI-0457095.

References

[1] Armony, M. 2005. Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems* **51**(3–4) 287–329.
[2] Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.
[3] Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(2) 271–292.
[4] Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15**(4) 2606–2650.
[5] Billingsley, P. 1968. *Convergence of Probability Measures*. J. Wiley & Sons, New York.
[6] Bramson, M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.
[7] Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. H. Dshalalow, ed. *Advances in Queueing*: *Theory, Methods, and Open Problems*. CRC Press, Inc., Boca Raton, FL, 463–480.
[8] Budhiraja, A., C. Lee. 2008. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.* **34**(1) 45–56.
[9] Dai, J. G. 1995. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.
[10] Dai, J. G., T. Tezcan. 2008. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* **59**(2) 95–134.
[11] Dai, J. G., T. Tezcan. 2009. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* Forthcoming.
[12] Dai, J. G., T. Tezcan. 2009. State space collapse in many-server diffusion limits of parallel server systems. *Math. Oper. Res.* Forthcoming.
[13] Gamarnik, D., A. Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Probab.* **16** 56–90.

[14] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.
[15] Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.
[16] Gurvich, I., W. Whitt. 2007. Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* Forthcoming.
[17] Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management*. **11**(2) 237–253.
[18] Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.
[19] Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.
[20] Karatzas, I., S. E. Shreve. 1991. *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer-Verlag, New York.
[21] Lipster, R. Sh., A. N. Shirayev. 1989. *Theory of Martingales*. Kluwer Academic Publishers, Boston.
[22] Mandelbaum, A., S. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52** 836–855.
[23] Pang, G., R. Talrejea, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* **4** 193–267.
[24] Protter, P. 1992. *Stochastic Integration and Differential Equations—A New Approach*. Springer-Verlag, New York.
[25] Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* **19**(4) 946–954.
[26] Royden, H. L. 1968. *Real Analysis*, 2nd ed. Macmillan, London.
[27] Talreja, R., W. Whitt. 2008. Heavy-traffic limits for waiting times in many-server queues with abandonments. Working paper, Columbia University, New York.
[28] Tezcan, T. 2008. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Oper. Res.* **33**(1) 51–90.
[29] Van der Vaart, A. W. 2006. Martingales, diffusions and financial mathematics—lecture notes. Available at: http://www.math.vu.nl/sto/onderwijs/mdfm/.
[30] Whitt, W. 2002. *Stochastic-Process Limits*: *An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York.
[31] Williams, R. J. 1998. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* **30**(1–2) 27–88.