

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

Itay Gurvich\*      Ward Whitt †

October 15, 2007

## Abstract

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 in two subsequent papers to solve important staffing and control problems for large-scale service systems.

## 1 Introduction

**Parallel-Server Systems.** In this paper we focus on a family of multi-class queueing networks known as *Parallel-Server Systems* (PSS's). 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

---

\*Columbia Business School, 41 Uris Hall, 3022 Broadway, New York, NY 10027. (ig2126@columbia.edu)

†IEOR Department, Columbia University, 304 S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699. (ww2040@columbia.edu)

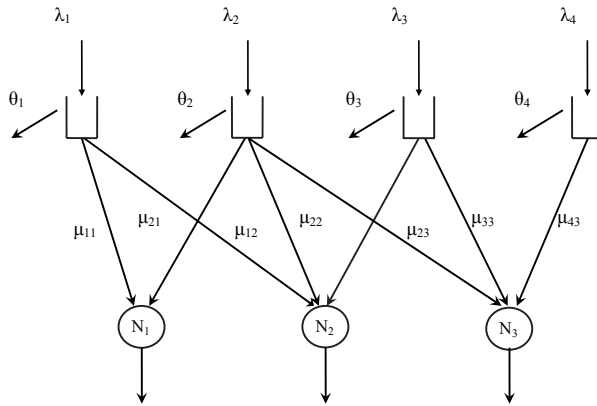


Figure 1: A PSS and its corresponding routing graph

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. 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 PSS's. Customers arrive exogenously according to independent Poisson processes, with one for each class. The class- $i$  arrival rate is  $\lambda_i$ . Customers from each class enter service in order of arrival. 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  $\mu_{i,j}$ . For some of the results here, we will assume in addition that  $\mu_{i,j}$  depends only upon  $i$ , or only upon  $j$ , or upon 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  $i$ .

PSS's are used to model various manufacturing and service systems, especially call centers; see Gans et al. [16]. 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 are especially interested in PSS's with a large number of agents.

**The Queue-and-Idleness-Ratio (QIR) Rule.** It remains to specify how the customers are assigned to agents, which is what this paper is mostly about. 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.

We now explain the QIR control. It uses two vector functions  $p(\cdot) := (p_1(\cdot), \dots, p_I(\cdot))$  and  $v(\cdot) := (v_1(\cdot), \dots, v_J(\cdot))$ , 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  $x \in \mathbb{R}_+$ .

The QIR rule 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*

$$\hat{Q}_i^\lambda(t) - \hat{Q}_\Sigma^\lambda(t) p_i(\hat{Q}_\Sigma^\lambda(t)), \quad (1)$$

and by routing an arriving customer at time  $t$  to the agent pool with the greatest *idleness imbalance*

$$\hat{I}_j^\lambda(t) - \hat{I}_\Sigma^\lambda(t) p_j(\hat{I}_\Sigma^\lambda(t)), \quad (2)$$

where  $\hat{Q}_i^\lambda(t)$  and  $\hat{I}_j^\lambda(t)$  are, respectively, the properly scaled class- $i$  queue length and type- $j$  number of idle agents at time  $t$  in the  $\lambda^{th}$  system. The remaining variables with  $\Sigma$  subscripts –  $\hat{Q}_\Sigma^\lambda(t) = \sum_{i \in \mathcal{I}} \hat{Q}_i^\lambda(t)$  and  $\hat{I}_\Sigma^\lambda(t) = \sum_{j \in \mathcal{J}} \hat{I}_j^\lambda(t)$  – are the corresponding aggregate quantities. The aim of QIR, then, is to drive these imbalances toward 0. We will show that 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 than given in equations (1) and (2); 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. For example, 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 in order 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.

**The QED Many-Server Heavy-Traffic Limiting Regime.** In order to establish these asymptotic-proportionality results for many-server PSS's, we work in the QED many-server heavy-traffic limiting regime, first formalized by Halfin and Whitt [21] for the  $M/M/N$  queue. For the  $M/M/N$  model, the QED regime is obtained by letting the aggregate arrival rate  $\lambda$  and the number  $N$  of agents grow indefinitely, while holding the service rate fixed, so that the utilization,  $\rho$ , approaches its critical value 1 in an appropriate manner. Specifically, considering a sequence of systems indexed by the aggregate arrival rate  $\lambda$ , it is assumed that

$$\sqrt{\lambda}(1 - \rho^\lambda) \rightarrow \beta \text{ as } \lambda \rightarrow \infty, \quad (3)$$

where  $-\infty < \beta < \infty$ . Halfin and Whitt showed that the limit in (3) holds 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 (3), 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 PSS's considered here. In particular, the QED limiting regime is much more difficult to specify for multi-class multi-type queues. We will use mathematical programs for that purpose; see §2.

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  $\rho$  approach one.

**State-Space Collapse (SSC).** It should be evident that PSS's, like many other multi-class queueing networks, are quite complex, making them very challenging to analyze. To address this complexity, recent research on queueing networks has aimed to establish special asymptotic techniques in order to obtain more elementary descriptions of these complex models. The seminal papers by Bramson [8] and Williams ([34]) provide a magnificent example of such a simplification by showing how complex multi-class queueing networks can be approximated by more elementary diffusion models, known as semi-martingale 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 PSS's 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 PSS's were based on ad-hoc proofs for specific models and controls. Examples include Armony [1], Gurvich et. al. [18], 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 [31] and Dai and Tezcan [13, 14].

We contribute to this SSC literature by establishing that our proposed QIR controls produce an important SSC for a large class of PSS's 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 [8], lie at the heart of the SSC proofs here.

As a consequence of our SSC results here, asymptotically, the multi-dimensional queue-length and idleness processes will be completely determined by the two-dimensional process that contains only the aggregate queue-length and idleness information. This notion of SSC is somewhat weaker than often seen in the conventional heavy-traffic regime, where the multi-dimensional queue-length process actually collapses into a one-dimensional process; e.g., see Mandelbaum and Stolyar [25]. However, we too obtain that strong one-dimensional form of SSC in the many-server setting when we assume that the service rates are pool-dependent; see §5.

In closing this discussion of SSC, we mention the important paper by Atar [5], which focuses on establishing asymptotic optimality in heavy-traffic for PSS's in the QED regime. An SSC result can be deduced as a corollary of his analysis. In some cases QIR will be equivalent to Atar's control. In those cases our SSC result will build on his results.

**The Value of QIR and SSC.** The power of QIR for controlling PSS's is demonstrated by our two subsequent papers: First, in [19], 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 pre-determined Quality-of-Service (QoS) constraints. The control component of our proposed solution is a

special case of QIR called *Fixed-Queue-Ratio* (FQR) routing.

Second, in [20] 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- $c\mu$  ( $Gc\mu$ ) rule. By doing this, we are able to partially extend the important results of Mandelbaum and Stolyar [25] to the QED regime.

The  $Gc\mu$  rule is a great example of a simple and intuitive control that is applicable to very general network structures. The  $Gc\mu$  rule was first introduced by Van-Meighem [33] for the multi-class and *single-server* model, and then was generalized to more complicated PSS's by Mandelbaum and Stolyar [25]. Mandelbaum and Stolyar consider a queueing system with multiple customer classes and multi-skilled *single-server* service stations in parallel. The service rates  $\mu_{i,j}$  are allowed to depend on both the customer class  $i$  and the server  $j$ . If  $Q_i$  is the queue length of class- $i$  customers, then this queue is assumed to incur cost at the rate of  $C_i(Q_i)$ , where  $C_i$  is an increasing convex real-valued function. Mandelbaum and Stolyar show that the following rule is asymptotically optimal for finite-horizon problems in the conventional heavy-traffic regime: When becoming free, server  $j$  chooses for service the customer from the head of the line (who has waited the longest) from the class- $i$  queue (from among the eligible queues), where  $i$  yields the maximum value of  $C'_i(Q_i)\mu_{i,j}$ , with  $C'_i$  being the derivative of  $C_i$ . The most appealing aspect of  $Gc\mu$  is its decentralized nature that allows decisions to be made locally for each server, based only on the queue lengths of the classes that this server can serve. Thus there is no need for central control.

The QIR controls are reminiscent of  $Gc\mu$  in terms of its simplicity and non-centralized nature. This powerful aspect of QIR is further underscored by the understanding that, in contrast to the single-server setting of [25], the many-server setting requires a much greater care with respect to the cross-occupancy levels, i.e, the number of type- $j$  agents giving service to class- $i$  customers for each  $i$  and  $j$ .

**Summary.** The main result of this paper, then, is, to identify conditions under which the QIR control will yield the desired SSC result. In addition, we also establish diffusion limits for the case in which service rates are pool-dependent and obtain some auxiliary results that can be useful in the application of QIR to specific problems, as in [19] and [20]. For the two problems examined in [19] and [20], we need to find optimal or nearly-optimal decision rules specifying how customers should be assigned to agents. We show that QIR is such a decision rule. In both applications, the SSC that is obtained through QIR is crucial. Since QIR is a

parametric family of controls, we expect that, with appropriately chosen parameters, it can be used in other applications as well.

**Organization of the Paper.** The PSS and the QIR controls are defined in §2. The main result of the paper, the SSC result, is stated in §3 and proved in §4. Finally, §5 establishes stochastic-process limits for the case of pool-dependent service rates. In addition, §5 provides some auxiliary results relating SSC to performance measures that are of interest in applications.

## 2 The Model

We consider PSS's with a set  $\mathcal{I} := \{1, \dots, I\}$  of customer classes and a set  $\mathcal{J} := \{1, \dots, J\}$  of agent pools. Pool  $j$  contains  $N_j$  agents. Service times are assumed to be exponentially distributed. The mean handling (service) time of a class- $i$  customer by a type- $j$  agent is  $1/\mu_{i,j}$ ; equivalently,  $\mu_{i,j}$  is the rate a type- $j$  agent can serve a class- $i$  customer. Whenever type- $j$  agents do not have the required skill to serve class- $i$  customers,  $\mu_{i,j} = 0$ .

Customers of class  $i$  arrive according to a Poisson process with rate  $\lambda_i$ . The aggregate arrival rate is  $\lambda := \sum_{i \in \mathcal{I}} \lambda_i$ . Finally, class- $i$  customers have an exponential patience with rate  $\theta_i$ . That is, each class- $i$  customer will abandon the system (and be removed from the queue) if his waiting time exceeds his patience. Infinite patience is obtained by setting  $\theta_i = 0$ .

The possible routing for this 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:  $E := \{(i, j) \in \mathcal{I} \times \mathcal{J} : \mu_{i,j} > 0\}$ . 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 sub-graph will depend on the solution of a linear program (LP); see more below. Figure 1 depicts a PSS routing graph.

Let  $Q_i(t)$  be the queue length of class- $i$  customers, and  $I_j(t)$  the number of idle agent 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  $X_i^\lambda(t) := Q_i(t) + \sum_{j \in \mathcal{J}} Z_{i,j}$ . Finally, we

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  $\lambda$ . We add the superscript  $\lambda$  to express the dependence on the index. Thus,  $Q_i^\lambda(t)$  stands for the class- $i$  queue length at time  $t$  in the  $\lambda^{th}$  system. The routing graph and service rates  $\{\mu_{i,j}, i \in \mathcal{I}, j \in \mathcal{J}\}$  do not change with  $\lambda$ .

**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 [36]. 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,  $\{Y^\lambda : \lambda > 0\}$  with sample paths in  $D^d$ , we will be showing that  $Y^\lambda \Rightarrow 0$  in  $D^d$  as  $\lambda \rightarrow \infty$ ; 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  $\mathbb{R}$ .

Since 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  $\|B\|_{s,T}^* = \sup_{s \leq t \leq T} \|B(t)\|$ , 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 now  $\|B(t)\| = \sum_{k=1}^d \sum_{l=1}^m |B_{k,l}(t)|$ . We omit the subscripts whenever we refer to vectors or to vector processes. For example,  $N^\lambda$  and  $X^\lambda(t)$ , will stand, respectively, for the vector in  $\mathbb{Z}^J$  whose components are  $N_j^\lambda$  and the vector in  $\mathbb{R}^I$  whose components are  $X_i^\lambda(t)$ .

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 some conventions for vector products: For two vectors  $x, y \in \mathbb{R}^d$ ,  $xy$  is the component wise product ( $(xy)_i = x_i y_i$ ). Clearly, 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  $j^{\text{th}}$  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 (3), but that is not enough.

**Assumption 2.1 (heavy-traffic conditions)** *There are constant  $a_i > 0$ ,  $i \in \mathcal{I}$ , and  $\nu_j > 0$ ,  $j \in \mathcal{J}$ , such that, as  $\lambda \rightarrow \infty$ ,*

$$\frac{\lambda_i}{\lambda} \rightarrow a_i > 0, \quad i \in \mathcal{I}, \quad \text{and} \quad \frac{N_j^\lambda}{\lambda} \rightarrow \nu_j > 0, \quad j \in \mathcal{J}.$$

*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  $\lambda \rightarrow \infty$ ,*

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

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

$$\begin{aligned} & \text{Minimize} && \rho \\ & \text{Subject to:} && \sum_{j \in \mathcal{J}} \mu_{i,j} \nu_j x_{i,j} = a_i, \quad i \in \mathcal{I}, \\ & && \sum_{i \in \mathcal{I}} x_{i,j} \leq \rho, \quad j \in \mathcal{J}, \\ & && \rho \geq 0, x_{i,j} \geq 0, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}. \end{aligned} \tag{4}$$

Since the constants of Assumption 2.1 are specified, the mathematical program in (4) is an LP. An optimal solution is a vector  $(x, \rho) \in \mathbb{R}_+^{I \times J} \times \mathbb{R}_+$ . We impose a critical-loading assumption that requires that the selected optimal solution does not yield an underloaded or overloaded system. That is clearly necessary to put us in the QED regime.

**Assumption 2.2 (critical loading)** *For any optimal solution  $(\bar{x}, \bar{\rho})$ , we have that  $\sum_{i \in \mathcal{I}} \bar{x}_{i,j} = 1$  for all  $j \in \mathcal{J}$  and, consequently, 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, this is imposed to prevent the fluid from drifting into an underloaded state and away from the QED regime.

With Assumption 2.2, it suffices to denote the selected optimal solution by its  $x$  coordinate, since the value of  $\rho$  will necessarily be 1. Hence, fix an optimal solution  $\bar{x}$  for (4) satisfying Assumption 2.2. We now indicate how the chosen optimal solution  $\bar{x}$  is used. Since 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  $\{(i, j) \in \mathcal{I} \times \mathcal{J} : \bar{x}_{i,j} > 0\}$ . 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  $x_{i,j} \leq 0$  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 should have a clear notion of 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 this basic feature, the first important structural assumption we impose is the following:

**Assumption 2.3 (connected routing graph)** *The selected optimal solution ( $\bar{x}$ ) for (4) produces a routing graph determined by the edges  $\mathcal{E}(\bar{x}) := \{(i, j) \in \mathcal{I} \times \mathcal{J} : \bar{x}_{i,j} > 0\}$  that is connected.*

**Assumptions 2.1-2.3 are assumed to hold throughout the rest of the paper.** By saying that the graph is connected, we follow common graph-theory terminology: A graph is connected if there exists a path between every two nodes in the graph. This connected-graph assumption is crucial for the ability to instantaneously balance the system asymptotically; see Section 2.7 of Atar [5] for elaboration. We will actually need a finer characterization of the network graph beyond its connectedness. More specifically, connected graphs can be cyclic or acyclic (a graph is acyclic if there is a unique path between each pair of

nodes). This distinction will be important for our results, because QIR will work well with cyclic networks only when certain parametric conditions hold. The importance of this distinction is made explicit in our SSC results; see Theorem 3.1 and Remark 3.1. Given an optimal solution  $(\bar{x})$  to (4),  $J(i)(\bar{x})$  for  $i \in \mathcal{I}$  is defined to be the set of agent pools connected to customer class  $i$ , i.e.,  $J(i)(\bar{x}) = \{j \in \mathcal{J} : \bar{x}_{i,j} > 0\}$ . 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.,  $I(j)(\bar{x}) = \{i \in \mathcal{I} : \bar{x}_{i,j} > 0\}$ . 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 (4) for which the edges in  $\mathcal{E}(\bar{x})$  induce a connected routing graph. Then, we define the following scaled processes:

$$\begin{aligned} \hat{X}_\Sigma^\lambda(t) &:= \frac{X_\Sigma^\lambda(t) - N_\Sigma^\lambda}{\sqrt{\lambda}}; & \hat{I}_\Sigma^\lambda(t) &:= \frac{I_\Sigma^\lambda(t)}{\sqrt{\lambda}}; & \hat{X}_i^\lambda(t) &:= \frac{X_i^\lambda(t) - \sum_{j \in \mathcal{J}} \bar{x}_{i,j} N_j^\lambda}{\sqrt{\lambda}}, \quad i \in \mathcal{I}; \\ \hat{Q}_i^\lambda(t) &:= \frac{Q_i^\lambda(t)}{\sqrt{\lambda}}, \quad i \in \mathcal{I}; & \hat{I}_j^\lambda(t) &:= \frac{I_j^\lambda(t)}{\sqrt{\lambda}}, \quad j \in \mathcal{J}; & \hat{Z}_{i,j}^\lambda(t) &:= \frac{Z_{i,j}^\lambda(t) - \bar{x}_{i,j} N_j^\lambda}{\sqrt{\lambda}}, \quad (i,j) \in \mathcal{I} \times \mathcal{J}. \end{aligned}$$

## 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

$$\|f(x) - f(y)\| \leq C_K \|x - y\|^\alpha \quad \text{for all } x, y \in K, \quad (5)$$

where  $\|\cdot\|$  is a chosen norm inducing the usual Euclidean topology, which we take to be the  $\mathbb{L}^1$  norm:  $\|x\| := \sum_i |x_i|$ . With that definition, we are ready to define our new class of admissible state-dependent ratio functions.

**Definition 2.2 (an admissible state-dependent ratio function)** For an integer  $d > 0$ , a vector valued function  $r \equiv r(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}_+^d$ , 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  $r_k : \mathbb{R}_+ \mapsto \mathbb{R}_+$  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 (4) is fixed and the routing graph  $\mathcal{E}(\bar{x})$  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**

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

*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**

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

*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

$$(\hat{Q}^\lambda, \hat{Z}^\lambda) := (\hat{Q}_i^\lambda(t), \hat{Z}_{i,j}^\lambda(t); i \in \mathcal{I}, j \in \mathcal{J}) \quad (6)$$

is a CTMC with stationary transition probabilities.

**Remark 2.1 (simplification under fixed queue ratios)** We point out that if  $p(\cdot) \equiv (p_1, \dots, p_I)$  with  $p_i > 0$  for all  $i \in \mathcal{I}$ , QIR is equivalently given by choosing upon service completion to serve the customer from the head of queue  $i^*$  where

$$i^* \equiv i^*(t) \in \operatorname{argmax}_{i \in \mathcal{I}(j), \hat{Q}_i^\lambda(t) > 0} \left\{ \frac{\hat{Q}_i^\lambda(t)}{p_i} \right\}.$$

Hence, with positive fixed ratios the QIR control is significantly simplified. ■

**Remark 2.2 (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

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

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

### 3 State-Space Collapse Under QIR

Our general SSC result under QIR is:

#### Theorem 3.1 (SSC under QIR)

Fix an optimal solution  $\bar{x}$  for (4) 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 Definitions 2.3. Suppose that at least one of the following conditions holds with respect to  $\bar{x}$ :

- **C-1 Only one pool has 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  $v = e_{j^*}$ .
- **C-2 The service rates depend only on the agent type:** For all  $(i, j) \in \mathcal{E}(\bar{x})$ ,  $\mu_{i,j} = \mu_j$ .
- **C-3 The service rates depend only on the customer class:** For all  $(i, j) \in \mathcal{E}(\bar{x})$ ,  $\mu_{i,j} = \mu_i$ .

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

$$\hat{Q}_i^\lambda(t) - \hat{Q}_\Sigma^\lambda(t) p_i \left( \hat{Q}_\Sigma^\lambda(t) \right) \Rightarrow 0 \quad \text{in } D_- \quad \text{as } \lambda \rightarrow \infty, \quad i \in \mathcal{I}, \quad (7)$$

and

$$\hat{I}_j^\lambda(t) - \hat{I}_\Sigma^\lambda(t) v_j \left( \hat{I}_\Sigma^\lambda(t) \right) \Rightarrow 0 \quad \text{in } D_- \quad \text{as } \lambda \rightarrow \infty, \quad j \in \mathcal{J}. \quad (8)$$

The convergence in (7) and (8) is strengthened to convergence in  $D$  if we assume that

$$\hat{Q}_i^\lambda(0) - \hat{Q}_\Sigma^\lambda(0) p_i \left( \hat{Q}_\Sigma^\lambda(0) \right) \Rightarrow 0, \quad i \in \mathcal{I}, \quad \text{and} \quad \hat{I}_j^\lambda(0) - \hat{I}_\Sigma^\lambda(0) v_j \left( \hat{I}_\Sigma^\lambda(0) \right) \Rightarrow 0, \quad j \in \mathcal{J}. \quad (9)$$

Finally, if condition C-1 holds, then,

$$\frac{1}{\sqrt{\lambda}} \hat{Z}_{ij}^\lambda(t) \Rightarrow 0 \quad \text{in } D \quad \text{as } \lambda \rightarrow \infty, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}. \quad (10)$$

**Remark 3.1 (network-graph characterization)** To interpret the conditions of Theorem 3.1, note that under either condition C-2 or C-3 arbitrary connected graphs are allowed. In particular, cyclic graphs are allowed. Condition C-1, however, rules out cyclic structures and requires that the network graph be a special type of a tree. There are some basic, but important, models that do satisfy the structural requirement of condition C-1: The V model satisfies the condition C-1 trivially. So do the N, M and  $\wedge$  (inverted- V) models as depicted in *Figure 2*. The SSC result under condition C-1 is a consequence of Atar [5]: With the special tree structure imposed by condition C-1, QIR becomes equivalent to the control constructed in [5], which we denote by GQIR (see §4.2).

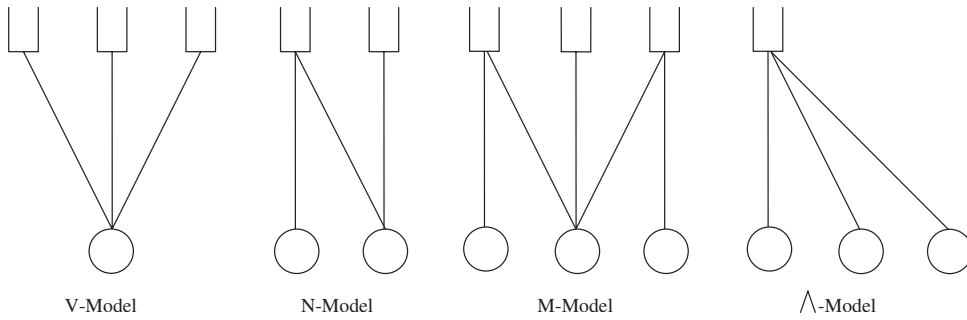


Figure 2: The V, N, M and  $\wedge$  Models

If neither of the conditions C-2 nor C-3 is satisfied and if the graph is a more general tree than required by condition C-1, then our theorem does not apply. In that case, one needs to use a different control - see Theorem 4.2 below. But even with Theorem 4.2, we have not covered all network graphs. There are some very basic structures, like the  $X$  and  $W$  models, that are excluded. ■

**Remark 3.2 (class-dependent service rates)** It seems reasonable that service rates should depend more on the customer class than on the agent type. Thus, the fact that QIR performs as desired under condition C-3 in Theorem 3.1 is good news. ■

**Remark 3.3 (class-pool occupancy processes)** Why does equation (10) hold under condition C-1 and not under any of the other conditions? The reason is the acyclic-tree structure imposed by condition C-1. In a tree-like model, controlling the queues and idleness processes uniquely determines the class-pool occupancy processes  $Z_{i,j}^\lambda(t)$  through a linear mapping (see equation (38) below). This 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 either conditions C-2 or C-3 hold. In the presence of any 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  $\sum_{j \in \mathcal{J}} \mu_j \bar{v}_j$  under condition C-2, or  $\sum_{i \in \mathcal{I}} \mu_i a_i$  under condition C-3. ■

While state-space collapse is an asymptotic property, systems of medium size already exhibit this phenomenon. The following example is an illustration.

**Example 3.1 (SSC in a two-class model)** To illustrate that SSC is evident in systems of medium size, 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  $\bar{v}_1 = 1/4$ ,  $\bar{v}_2 = 7/24$  and  $\gamma_1 = \gamma_2 = 0$ , so that an optimal solution for (4) 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 2).

We simulate this resulting  $N$  model to see if there is approximate state-space collapse. 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 as depicted in Figure 3 show that  $Q_2(t)$  and  $2 \cdot Q_1(t)$  are hardly distinguishable. ■

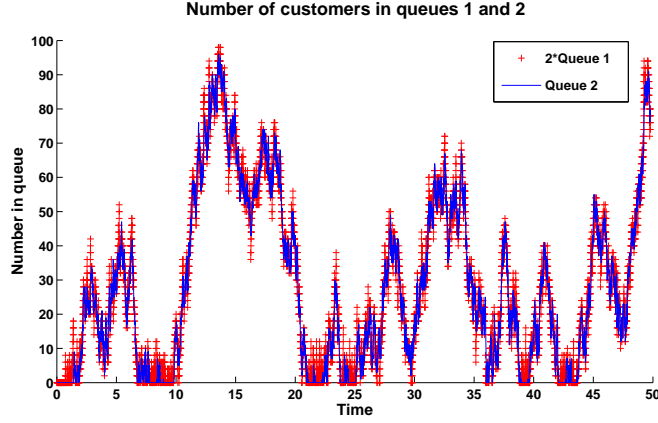


Figure 3: The N Model for the Numerical Experiment

## 4 Proof of Theorem 3.1

### 4.1 Sample-Path Construction, Martingales and Other 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. [26] for an overview.

#### 4.1.1 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

$$X_i^\lambda(t) := X_i^\lambda(0) + A_i(\lambda t) - \sum_{j \in \mathcal{J}} S_{i,j} \left( \mu_{i,j} \int_0^t Z_{i,j}^\lambda(s) ds \right) - R_i \left( \theta_i \int_0^t Q_i^\lambda(s) ds \right), \quad t \geq 0. \quad (11)$$

By direct construction, the stochastic process  $\{(X_1^\lambda(t), \dots, X_I^\lambda(t)) : t \geq 0\}$  has the correct distribution and is a process in  $D^I$ . (A formal argument would follow the proof of Lemma 2.1 in [26].) Clearly, equation (11) does not yet fully define the system dynamics because the dynamics of the queue-length processes  $Q_i^\lambda(t)$  and the busy-agent processes  $Z_{i,j}^\lambda(t)$  are not yet specified. To complete the definition, let  $A_{ij}^\lambda(t)$ ,  $(i, j) \in \mathcal{I} \times \mathcal{J}$  be the cumulative number of class- $i$  customers routed to an idle type- $j$  agent immediately upon arrival by time  $t$ ; let  $\Phi_{i,j}^\lambda(t)$ ,  $(i, j) \in \mathcal{I} \times \mathcal{J}$ , be the cumulative number of class- $i$

customers assigned to a type- $j$  agent after waiting in queue, by time  $t$ . Then, we write

$$Z_{i,j}^\lambda(t) = Z_{i,j}^\lambda(0) + A_{i,j}^\lambda(t) + \Phi_{i,j}^\lambda(t) - S_{i,j} \left( \mu_{i,j} \int_0^t Z_{i,j}^\lambda(s) ds \right), \quad (12)$$

$$Q_i^\lambda(t) = Q_i^\lambda(0) + A_i^\lambda(t) - \sum_{j \in \mathcal{J}} A_{i,j}^\lambda(t) - \sum_{j \in \mathcal{J}} \Phi_{i,j}^\lambda(t) - R_i \left( \theta_i \int_0^t Q_i^\lambda(s) ds \right), \quad (13)$$

$$I_j^\lambda(t) = N_j^\lambda - \sum_{i \in \mathcal{I}} Z_{i,j}^\lambda(t), \quad (14)$$

and we also let  $D_j^\lambda(t) := \sum_{k \in \mathcal{I}} S_{k,j} \left( \mu_{k,j} \int_0^t Z_{k,j}^\lambda(s) ds \right)$ , 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 incorporate these specifics, let

$$\hat{U}_i^\lambda(t) := \hat{Q}_i^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p_i([\hat{X}_\Sigma^\lambda(t)]^+), \quad i \in \mathcal{I}, \quad (15)$$

and

$$\hat{V}_j^\lambda(t) := \hat{I}_j^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^- v_j([\hat{X}_\Sigma^\lambda(t)]^-), \quad j \in \mathcal{J}. \quad (16)$$

Then we can write

$$A_{i,j}^\lambda(t) = \int_0^t 1 \left\{ \hat{I}_j^\lambda(s-) > 0, j \in \operatorname{argmax}_{j \in J(i)} \hat{V}_j^\lambda(s-) \right\} dA_i^\lambda(s), \quad (17)$$

and

$$\Phi_{i,j}^\lambda(t) = \int_0^t 1 \left\{ \hat{Q}_i^\lambda(s-) > 0, i \in \operatorname{argmax}_{i \in I(j)} \hat{U}_i^\lambda(s-) \right\} dD_j^\lambda(s). \quad (18)$$

These processes are well (uniquely) defined since we have assumed that ties are broken in a consistent manner.

### 4.1.2 Martingales

We now develop a martingale representation; again see [26] for background. Given the processes defined above, we define the  $\sigma$ -algebras

$$\begin{aligned} \mathcal{F}^\lambda(t) &:= \sigma \left\{ A_i(\lambda_i s), X_i^\lambda(s), Q_i^\lambda(s), Z_{i,j}^\lambda(s), R_i \left( \theta_i \int_0^s Q_i^\lambda(u) du \right), \right. \\ &\quad \left. S_{i,j} \left( \mu_{i,j} \int_0^s Z_{i,j}^\lambda(u) du \right) : i \in \mathcal{I}, j \in \mathcal{J}, s \leq t \right\}, \quad t \geq 0, \end{aligned} \quad (19)$$

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

We now exploit a **martingale decomposition**. In doing so, we follow the terminology used in [26]. First, as in [26], we will assume that  $E[Q_{\Sigma}^\lambda(0)] < \infty$  for all  $\lambda$ , but we note that this assumption is made without loss of generality, because it can be later relaxed, as in §6.3 of [26]. A straightforward adaptation of Lemmas 3.2 and 3.4 in [26] to our setting establishes that the processes  $A_i(\lambda_i t)$ ,  $R_i \left( \theta_i \int_0^t Q_i^\lambda(s) ds \right)$  and  $S_{i,j} \left( \mu_{i,j} \int_0^t Z_{i,j}^\lambda(s) ds \right)$  admit martingale decompositions with respect to the filtration  $\mathbb{F}^\lambda$ . Specifically, the processes:

$$M_{i,j}^\lambda(t) := S_{i,j} \left( \mu_{i,j} \int_0^t Z_{i,j}^\lambda(s) ds \right) - \mu_{i,j} \int_0^t Z_{i,j}^\lambda(s) ds, \quad i \in \mathcal{I}, j \in \mathcal{J}, \quad (20)$$

$$M_{A_i}^\lambda(t) := A_i(\lambda_i t) - \lambda_i t, \quad i \in \mathcal{I}, \quad (21)$$

$$M_{R_i}^\lambda(t) := R_i \left( \theta_i \int_0^t Q_i^\lambda(s) ds \right) - \theta_i \int_0^t Q_i^\lambda(s) ds, \quad i \in \mathcal{I}, \quad (22)$$

where we note that  $M_{R_i}^\lambda(t) = 0$ ,  $t \geq 0$  for all  $i$  with  $\theta_i = 0$ , are square-integrable martingales with predictable quadratic variations defined by

$$\langle M_{i,j}^\lambda \rangle(t) := \mu_{i,j} \int_0^t Z_{i,j}^\lambda(s) ds, \quad i \in \mathcal{I}, j \in \mathcal{J}, \quad (23)$$

$$\langle M_{A_i}^\lambda \rangle(t) := \lambda_i t, \quad i \in \mathcal{I}, \quad (24)$$

$$\langle M_{R_i}^\lambda \rangle(t) := \theta_i \int_0^t Q_i^\lambda(s) ds, \quad i \in \mathcal{I}. \quad (25)$$

As integrals of predictable processes with respect to counting processes, the processes

$$A_{i,j}^\lambda(t), (i, j) \in \mathcal{I} \times \mathcal{J}, \text{ and } \Phi_{i,j}^\lambda(t), (i, j) \in \mathcal{I} \times \mathcal{J},$$

also have corresponding martingale decompositions. For example,

$$M_{A_{i,j}}^\lambda(t) := \int_0^t 1 \left\{ \hat{I}_j^\lambda(s-) > 0, j \in \operatorname{argmax}_{j \in J(i)} \hat{V}_j^\lambda(s-) \right\} d \left( A_i^\lambda(s) - \lambda_i s \right), \quad (26)$$

is a square-integrable martingale with predictable quadratic variation

$$\langle M_{A_{i,j}}^\lambda(t) \rangle = \int_0^t 1 \left\{ \hat{I}_j^\lambda(s-) > 0, j \in \operatorname{argmax}_{j \in J(i)} \hat{V}_j^\lambda(s-) \right\} \lambda_i ds. \quad (27)$$

This is a consequence of the preservation of the martingale property under stochastic integration. Specifically, fixing any  $T > 0$ , the martingale  $M_{A_i}^\lambda(t \wedge T) = A_i(\lambda_i(t \wedge T)) - \lambda_i(t \wedge T)$  is square integrable and in particular it is in the space  $\mathcal{H}^2$  defined in page 124 of Protter [27]. Since the integrand in (26) is left continuous (because it is defined through the left limits), it is predictable. Since the integrand is also bounded, we can apply Theorem IV.2.11 (page 129) of [27] that guarantees that the stochastic integral in (26) is itself a square integrable martingale. Finally, item (ii) of Lemma 5.77 (page 85) in Van der Vaart [32] guarantees that the predictable quadratic variation process of (26) is the one given in (27). A similar argument is used for the process  $\Phi_{i,j}^\lambda(t)$ .

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

### 4.1.3 Stochastic Boundedness and Tightness

Our proofs will make extensive use of the concepts of tightness and stochastic boundedness; again see [26] 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  $Y^\lambda : \{Y^\lambda(t) : t \geq 0\}$  in  $D^d$  is **stochastically bounded** if, for all  $T > 0$ ,

$$\lim_{A \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} P\{\|Y^\lambda\|_T^* > A\} = 0, \quad (28)$$

where the norm  $\|\cdot\|_T^*$  is as defined in §2.

We make especial use of  $C$ -tightness. A sequence of processes,  $\{Y^\lambda\}$ , 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 [6] is very useful in establishing  $C$ -tightness. We restate it here:

**Theorem 4.1 (C-tightness)**

Consider a sequence of stochastic processes  $Y^\lambda : \{Y^\lambda(t) : t \geq 0\}$  in  $D^d$ . Suppose that

$$\lim_{A \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} P\{\|Y^\lambda(0)\| > A\} = 0. \quad (29)$$

Suppose further that, for each  $\epsilon > 0$  and  $T > 0$ ,

$$\lim_{\delta \rightarrow 0} \limsup_{\lambda \rightarrow \infty} P\{w_{Y^\lambda}(\delta, T) \geq \epsilon\} = 0, \quad (30)$$

where

$$w_x(\delta, T) := \sup_{0 \leq s < t \leq T: |t-s| \leq \delta} \{|x(t) - x(s)|\}. \quad (31)$$

Then,  $\{Y^\lambda\}$  is  $C$ -tight; i.e., it is a tight sequence in  $D^d$  and the limit of every convergent subsequence is almost surely continuous.

We will want to establish stochastic boundedness and  $C$ -tightness for various martingale processes. We use the general notation  $M^\lambda(t)$ , or  $\hat{M}^\lambda(t)$  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 4.1** (properties of the martingales) *Let  $\hat{M}^\lambda(t)$  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, for all  $\epsilon > 0$ ,*

$$\lim_{\delta \rightarrow 0} \limsup_{\lambda \rightarrow \infty} P\{w_{\langle \hat{M}^\lambda \rangle}(\delta, T) \geq \epsilon\} = 0. \quad (32)$$

*In particular, the process  $\langle \hat{M}^\lambda \rangle(t)$  is  $C$ -tight. Also, the scaled martingale  $\hat{M}^\lambda(t)$  is stochastically bounded.*

That is, for every fixed  $T > 0$ ,

$$\lim_{A \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{0 \leq t \leq T} |\hat{M}^\lambda(t)| > A \right\} = 0. \quad (33)$$

Finally, for any  $\epsilon > 0$ ,

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{0 \leq t \leq T/\sqrt{\lambda}} |\hat{M}^\lambda(t)| > \epsilon \right\} = 0. \quad (34)$$

**Proof:** We begin with equation (32). Let  $\langle \hat{M}^\lambda \rangle(t)$  be the scaled version of any of the predictable-quadratic-variation processes defined in (23)-(25). Then,

$$|\langle \hat{M}^\lambda \rangle(t) - \langle \hat{M}^\lambda \rangle(s)| \leq c(t-s) \frac{1}{\lambda} \max \left\{ \lambda, \sum_{j \in \mathcal{J}} N_j^\lambda, \int_s^t Q_\Sigma^\lambda(u) du \right\}, \quad (35)$$

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

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \frac{\int_s^t Q_\Sigma^\lambda(u) du}{\lambda} > c_2(t-s) \right\} = 0, \quad (36)$$

for some constant  $c_2$  large enough. Plugging this back into (35) we conclude that

$$\lim_{\delta \rightarrow 0} \limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{0 \leq s < t \leq T: |t-s| \leq \delta} |\langle \hat{M}^\lambda \rangle(t) - \langle \hat{M}^\lambda \rangle(s)| > \epsilon \right\} = 0, \quad (37)$$

and, consequently, that (32) holds.

Equations (33) and (34) follow from Lemma 5.8 of [26], which is based on the Lenglart-Rebolledo inequality, stated as Lemma 5.7 there. The extension to finite sum follows from basic properties of the quadratic variation processes (see Problem 1.5.7 in Karatzas and Shreve [22] and using the inequality  $2\langle M_1, M_2 \rangle \leq (\langle M_1 \rangle + \langle M_2 \rangle)$  (see Problem 1.8.9 in Lipster and Shirayev [24]).  $\blacksquare$

We now move to the actual proof of Theorem 3.1. The proof decomposes into three separate proofs, each corresponding to one of the three conditions. We dedicate a separate subsection to each of the conditions, so that subsections 4.2, 4.3 and 4.4 are dedicated to conditions C-1, C-2 and C-3, respectively.

## 4.2 State-Space Collapse under C-1

The proof under condition C-1 builds on the control proposed in Atar [5] and the result of his analysis. State-space collapse results are not, however, stated as such in [5]. Hence, we start by defining his control, which we denote by Generalized QIR (GQIR). While we use the term Generalized QIR, we emphasize that 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, however, to settings in which  $\mathcal{E}(\bar{x})$  contains cycles and to which QIR can be applied as long as the service rates satisfy conditions C-2 or C-3.

Towards the construction of GQIR, we need to 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 [5]):

$$\sum_{j \in \mathcal{J}} z_{i,j} = \alpha_i, \quad \sum_{i \in \mathcal{I}} z_{i,j} = \beta_j, \quad (38)$$

with no additional constraints. Atar [5] shows that the assumed tree structure (condition C-4 in Theorem 4.2 below) implies that there is indeed a unique solution. Thus, the mapping  $G(\cdot, \cdot)$  does indeed define a linear mapping on the domain

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

Moreover, the mapping  $G$  is such that (suppressing the time argument  $t$ )

$$\hat{Z}_{ij}^\lambda := G_{ij}(\hat{X}^\lambda - \hat{Q}^\lambda, -\hat{I}^\lambda), \quad (39)$$

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

$$\check{Z}_{ij}^\lambda := G_{ij} \left( \hat{X}^\lambda - [\hat{X}_\Sigma^\lambda]^+ p \left( [\hat{X}_\Sigma^\lambda]^+ \right), -[\hat{X}_\Sigma^\lambda]^- v \left( [\hat{X}_\Sigma^\lambda]^- \right) \right), \quad i \in \mathcal{I}, j \in \mathcal{J}. \quad (40)$$

We are now ready to introduce the GQIR control as it was constructed in [5]; see Sections 2.5 and 2.6 of [5].

### Definition 4.1 (Generalized QIR: GQIR)

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

agent pool

$$j := j(t) \in \operatorname{argmax}_{k \in J(i), I_k^\lambda > 0} (\check{Z}_{ik}^\lambda(t) - \hat{Z}_{ik}^\lambda(t))^+;$$

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 non-empty queue**

$$i := i(t) \in \operatorname{argmax}_{k \in I(j), \hat{Q}_k^\lambda(t) > 0} (\check{Z}_{kj}^\lambda(t) - \hat{Z}_{kj}^\lambda(t))^+;$$

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

$$(\hat{Q}^\lambda, \hat{Z}^\lambda) := (\hat{Q}_i^\lambda(t), \hat{Z}_{i,j}^\lambda(t); i \in \mathcal{I}, j \in \mathcal{J}) \quad (41)$$

is a CTMC with stationary transition probabilities.

We are now ready to establish the corresponding state-space collapse, which is a consequence of parts (i) and (iv) in Proposition 1 of Atar [5]:

**Theorem 4.2 (state-space collapse under GQIR)** *Fix two admissible state-dependent ratio functions  $p$  and  $v$  and an optimal solution  $\bar{x}$  for (4). Suppose that GQIR is used as defined in Definition 4.1, and that the following condition holds:*

- **C-4** *Under  $\bar{x}$ , the graph induced by the edges  $\mathcal{E}(\bar{x})$  is a tree.*

If, in addition  $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}(0))$  in  $\mathbb{R}^{I+J}$ , then all the conclusions of Theorem 3.1 hold, including (10).

**Proof:** Here we will be explaining why the state-space collapse under these conditions is a consequence of Proposition 1 in Atar [5]. First note that the results in [5] are given for a Markov control policy (see Definition 4 there) as given by a function  $h := (h_1, h_2)$ , where  $h_i : \mathbb{R}^I \mapsto \mathbb{U}$ ,  $i = 1, 2$ ,

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

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 [5]. 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

$$h(x) = \left( p \left( \left[ \sum_{i \in \mathcal{I}} x_i \right]^+ \right), v \left( \left[ \sum_{i \in \mathcal{I}} x_i \right]^- \right) \right), \quad (42)$$

and position ourselves in the framework of [5]. To be able to apply the result [5] directly, one additional observation is required. The control proposed in [5] is not precisely GQIR as defined in 4.1. Rather 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 [5]. A careful reading of [5] reveals, however, that while this modification is required for the large-time-estimates in Proposition 2 there, it is not used in the proof of Proposition 1 in [5]. Consequently, Proposition 1 in [5] is valid for GQIR as defined in 4.1 and it applies the desired state-space collapse result. Indeed, by translation of notation, the first statement in part (iv) of Proposition 1 in [5] corresponds to the statement

$$\sup_{s \leq u \leq t} \left\{ \sum_{i \in \mathcal{I}} \left| \hat{Q}_i^\lambda(u) - [\hat{X}_\Sigma^\lambda(u)]^+ p_i \left( [\hat{X}_\Sigma^\lambda(u)]^+ \right) \right| + \sum_{j \in \mathcal{J}} \left| \hat{I}_j^\lambda(u) - [\hat{X}_\Sigma^\lambda(u)]^- v_j \left( [\hat{X}_\Sigma^\lambda(u)]^- \right) \right| \right\} \Rightarrow 0; \quad (43)$$

see the definition of  $J^n(t)$  in equations (74), (52) and (53) of [5]. The second part of (iv) corresponds to the statement

$$\sup_{s \leq u \leq t} \left\{ \hat{Q}_\Sigma^\lambda(u) \wedge \hat{I}_\Sigma^\lambda(t) \right\} \Rightarrow 0; \quad (44)$$

see the definition of  $M^n(t)$  in equation (61) of [5]. But, by the definition of  $\hat{X}_\Sigma^\lambda(t)$  here, equations (43) and (44) combined imply the state-space-collapse conclusion in Theorem 4.2. Finally, equation (10) follows directly from part (i) of Proposition 1 in [5]. ■

Theorem 4.2 established state-space collapse under GQIR using the results of Atar [5]. Hence, state-space collapse under QIR (with condition C-1) will be established if we show that, given a common initial condition  $(X^\lambda(0), Z^\lambda(0))$ , the process  $(X^\lambda(t), Z^\lambda(t))$  has the same probability law under QIR as it does for GQIR, as long as we use the same rule to break ties. Toward that end, we show that given fixed sample paths of  $A_i(\cdot)$ ,  $i \in \mathcal{I}$ ,  $S_{i,j}(\cdot)$ ,  $i \in \mathcal{I}, j \in \mathcal{J}$  and  $R_i(\cdot)$ ,  $i \in \mathcal{I}$ , and initial condition  $(X^\lambda(0), Z^\lambda(0))$ , the process  $(X^\lambda(t), Z^\lambda(t))$  has the same sample path under with QIR or GQIR. This, in turn, establishes equivalence in probability law.

To be concrete, 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 (38) and (40)). Moreover, since by condition C-1, 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

$$(\check{Z}_{i,k}^\lambda(t) - \hat{Z}_{i,k}^\lambda(t))^+ = \hat{I}_k^\lambda(t). \quad (45)$$

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

$$(\check{Z}_{i,j^*}^\lambda(t) - \hat{Z}_{i,j^*}^\lambda(t))^+ = 0. \quad (46)$$

Combining (45) and (46), we have that upon arrival of a class- $i$  customer, if there are any idle agents in pool  $k \neq j^*$  (and in particular  $Q_i^\lambda(t-) = 0$ ), the decision rule of GQIR is equivalent to routing the customer to agent pool  $k$  with  $k \in \operatorname{argmax}_{k \in J(i), k \neq j^*} \hat{I}_k^\lambda(t)$ . If the only idle agents are in pool  $j^*$  route the customer to pool  $j^*$ .

On the other hand, under condition C-1, QIR is such that whenever  $\hat{I}_{j^*}(t) > 0$  we must have that  $Q_i^\lambda(t) = 0$ ,  $i \in \mathcal{I}$  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,

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

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. That is, route the customer to agent pool  $k$  with  $k \in \operatorname{argmax}_{k \in J(i), k \neq j^*} \hat{I}_k^\lambda(t)$ . 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  $k \neq j^*$  as, by condition C-1, these pools serve a single queue each. Now consider a service completion epoch

in pool  $j^*$ . Note that, as  $v = e_j^*$ , 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 that  $\hat{Z}_{i,j^*}^\lambda(t) = \hat{X}_i^\lambda(t) - \hat{Q}_i^\lambda(t) - \sum_{k \neq j^*} \hat{Z}_{i,k}^\lambda(t) = \hat{X}_i^\lambda(t) - \hat{Q}_i^\lambda(t)$ . Using equations (38) and (40) as before we also have that  $\check{Z}_{i,k}^\lambda(t) = 0$  for all  $k \in J(i), k \neq j^*$ , and  $\check{Z}_{i,j^*}^\lambda(t) = \hat{X}_i^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p_i([\hat{X}_\Sigma^\lambda(t)]^+)$ . Hence, upon service completion in pool  $j^*$ ,

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

if and only if

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

Since 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 C-1. This, in turn, implies, by induction, that we will have the same sample paths under both controls.  $\blacksquare$

### 4.3 State-Space Collapse Under C-2

For the purpose of this proof, it suffices to consider a somewhat less detailed construction of the system dynamics than the one introduced in §4.1. Specifically, we keep (13) and (14) but instead of (12), we write

$$Z_j^\lambda(t) = Z_j^\lambda(0) + \sum_{i \in \mathcal{I}} A_{i,j}^\lambda(t) + \sum_{i \in \mathcal{I}} \Phi_{i,j}^\lambda(t) - S_j \left( \mu_j \int_0^t Z_j^\lambda(s) ds \right), \quad (47)$$

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

$$D_j^\lambda(t) = S_j \left( \mu_j \int_0^t Z_j^\lambda(s) ds \right) \quad (48)$$

Also, instead of the martingales  $M_{i_j}^\lambda(t)$  in (20), we use the martingales

$$M_j^\lambda(t) := S_j \left( \mu_j \int_0^t Z_j^\lambda(s) ds \right) - \mu_j \int_0^t Z_j^\lambda(s) ds, \quad j \in \mathcal{J}. \quad (49)$$

Finally, instead of (11), we have

$$X_{\Sigma}^{\lambda}(t) = X_{\Sigma}^{\lambda}(0) + \sum_{i \in \mathcal{I}} A_i(\lambda_i t) - \sum_{j \in \mathcal{J}} S_j \left( \mu_j \int_0^t Z_j^{\lambda}(s) ds \right) - \sum_{i \in \mathcal{I}} R_i \left( \theta_i \int_0^t Q_i^{\lambda}(s) ds \right). \quad (50)$$

Following §4.1 with respect to the martingale decomposition and applying some algebraic manipulations we write

$$X_{\Sigma}^{\lambda}(t) = X_{\Sigma}^{\lambda}(0) + (\lambda - \sum_{j \in \mathcal{J}} \mu_j N_j^{\lambda})t + \sum_{j \in \mathcal{J}} \mu_j \int_0^t I_j^{\lambda}(s) ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t Q_i^{\lambda}(s) ds + M_{\Sigma}^{\lambda}(t), \quad (51)$$

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

By Assumption 2.1,

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

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

$$\hat{X}_{\Sigma}^{\lambda}(t) = \hat{X}_{\Sigma}^{\lambda}(0) - \beta t + \sum_{j \in \mathcal{J}} \mu_j \int_0^t \hat{I}_j^{\lambda}(s) ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t \hat{Q}_i^{\lambda}(s) ds + \hat{M}_{\Sigma}^{\lambda}(t) + o(1), \text{ as } \lambda \rightarrow \infty. \quad (52)$$

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

The proof now proceeds through a stopping argument. Specifically, let

$$\hat{B}^{\lambda}(t) := \sum_{i \in \mathcal{I}} \hat{U}_i^{\lambda}(t), \quad (53)$$

with  $\hat{U}_i^{\lambda}(t)$  as defined in (15). We first establish state-space collapse assuming that all the processes are stopped at the bounded stopping time  $T^{\lambda} := \sigma^{\lambda} \wedge T$ , where

$$\sigma^{\lambda} := \inf\{t \geq 0 \mid \hat{B}^{\lambda}(t) \geq 2\hat{B}^{\lambda}(0) \vee 1\}. \quad (54)$$

Since all the involved processes are assumed to be right continuous, the stopped processes are well defined (see Propositions 1.1.13 and 1.2.18 of [22]). We note here that the value of  $\hat{B}^{\lambda}(T^{\lambda})$  can be greater than  $2\hat{B}^{\lambda}(0) \vee 1$  as there can be a jump at time  $T^{\lambda}$ . Still, as all arrival and service-completion processes have

jumps of size 1, there exists a constant  $K$  such that

$$\hat{B}^\lambda(T^\lambda) \leq 2\hat{B}^\lambda(0) \vee 1 + K/\sqrt{\lambda} \quad (55)$$

The idea of the stopping argument is that, while it is hard to characterize the limits on the interval  $[0, T]$  directly, it is simpler to establish these limits for the stopped processes. Once these limits are established, showing that  $\sigma^\lambda \xrightarrow{P} \infty$  will imply that the same limiting behavior holds on  $[0, T]$ . For the stopped processes our proof consists of two main modules. First, in Proposition 4.2, we establish 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 4.3 establishes state-space collapse for the stopped processes using the results of Proposition 4.2 and the Bramson [8] and Dai and Tezcan [12] state-space collapse framework. Finally, the state-space collapse result is extended to the whole interval  $[0, T]$  in Proposition 4.6. Throughout it is assumed that the conditions of Theorem 3.1 hold in addition to condition C-2.

**Lemma 4.2 (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, i.e.,*

$$\lim_{A \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} P \left\{ \|\hat{Q}_\Sigma^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{I}_\Sigma^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{X}_\Sigma^\lambda(\cdot \wedge T^\lambda)\|_T^* > A \right\} = 0. \quad (56)$$

Also, the sequence of processes  $\{\hat{X}_\Sigma^\lambda(t \wedge T^\lambda) : \lambda > 0\}$  is C-tight.

**Proof:** By the definition of  $\hat{U}_i^\lambda(t)$  in (15) and Definition 2.2 for state-dependent ratio functions,

$$\hat{B}^\lambda(t) = \sum_{i \in \mathcal{I}} \hat{U}_i^\lambda(t) = \hat{Q}_\Sigma^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+, \quad (57)$$

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

$$\hat{Q}_\Sigma^\lambda(t) = [\hat{X}_\Sigma^\lambda(t)]^+ + \hat{B}^\lambda(t) \leq |\hat{X}_\Sigma^\lambda(t)| + \hat{B}^\lambda(t). \quad (58)$$

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

$$\hat{I}_\Sigma^\lambda(t) \leq |\hat{X}_\Sigma^\lambda(t)| + \hat{Q}_\Sigma^\lambda(t) \leq 2|\hat{X}_\Sigma^\lambda(t)| + \hat{B}^\lambda(t). \quad (59)$$

Plugging these into equation (52) and applying Gronwall's inequality (see Theorem 4.1 and Lemmas 4.1 and 5.6 in [26] or Problem 5.2.7 in [22]), we have

$$\|\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)\|_T^* \leq c_1 \left( |\hat{X}_\Sigma^\lambda(0)| + |\beta|T + \|\hat{B}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{M}_\Sigma^\lambda(\cdot \wedge T^\lambda)\|_T^* \right) e^{c_2 T}, \quad (60)$$

for some positive constants  $c_1$  and  $c_2$ . By Lemma 4.1,  $\hat{M}_\Sigma^\lambda(t)$  is stochastically bounded and, by the definition of  $T^\lambda$  and equation (55),  $\hat{B}^\lambda(t \wedge T^\lambda)$  is stochastically bounded. Here we also use the fact that  $\hat{B}^\lambda(0)$  is itself stochastically bounded by the assumed convergence of  $\hat{X}^\lambda(0)$ .

Finally, the sequence  $\hat{X}^\lambda(0)$  is stochastically bounded because it converges; see Corollary 5.2 in [26]. The stochastic boundedness of  $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$  now follows from it being bounded by a sum of stochastically bounded sequences; see Lemma 5.5 in [26]. Finally,  $\hat{I}_\Sigma^\lambda(t \wedge T^\lambda)$  and  $\hat{Q}_\Sigma^\lambda(t \wedge T^\lambda)$  are now stochastically bounded by applying (58) and (59). The result of the proposition now follows as the sum of stochastically bounded sequences is itself stochastically bounded.

It remains to establish the claimed C-tightness of  $\hat{X}^\lambda(t \wedge T^\lambda)$ . Using (52) once more we have that

$$|\hat{X}_\Sigma^\lambda(t) - \hat{X}_\Sigma^\lambda(s)| \leq |\beta|(t-s) + \sum_{j \in \mathcal{J}} \mu_j \int_s^t \hat{I}_j^\lambda(s) ds + \sum_{i \in \mathcal{I}} \theta_i \int_s^t \hat{Q}_i^\lambda(s) ds + |\hat{M}_\Sigma^\lambda(t) - \hat{M}_\Sigma^\lambda(s)|, \quad (61)$$

and, in particular, that

$$|\hat{X}_\Sigma^\lambda(t) - \hat{X}_\Sigma^\lambda(s)| \leq c_3 \left( |\beta|(t-s) + (t-s) \|\hat{I}_j^\lambda(\cdot \wedge T^\lambda)\|_T^* + (t-s) \|\hat{Q}_\Sigma^\lambda(\cdot \wedge T^\lambda)\|_T^* + |\hat{M}_\Sigma^\lambda(t) - \hat{M}_\Sigma^\lambda(s)| \right), \quad (62)$$

For all  $0 \leq s < t \leq T^\lambda$ . Using (58), (59) and Gronwall's inequality, we have that

$$|\hat{X}_\Sigma^\lambda(t) - \hat{X}_\Sigma^\lambda(s)| \leq c_3 \left( |\beta|(t-s) + |\hat{M}_\Sigma^\lambda(t) - \hat{M}_\Sigma^\lambda(s)| + (t-s) \|\hat{B}^\lambda(\cdot \wedge T^\lambda)\|_T^* \right) e^{c_4 T}, \quad (63)$$

for  $0 \leq s < t \leq T^\lambda$  and for some positive constants  $c_3$  and  $c_4$ . With the definitions in Theorem 4.1,

$$w_{\hat{X}_\Sigma^\lambda(\cdot \wedge T^\lambda)}(\delta, T) \leq c_5 \delta + c_6 \delta \|\hat{B}^\lambda(\cdot \wedge T^\lambda)\|_T^* + c_7 w_{\hat{M}_\Sigma^\lambda(\cdot \wedge T^\lambda)}(\delta, T), \quad (64)$$

for positive constants  $c_5$ ,  $c_6$  and  $c_7$ . As  $\hat{B}^\lambda(t \wedge T^\lambda)$  is stochastically bounded by the definition of  $T^\lambda$ , the C-tightness of  $\hat{X}^\lambda(t \wedge T^\lambda)$  is established if we prove the tightness of  $\hat{M}_\Sigma^\lambda(t \wedge T^\lambda)$ . This, however, follows immediately from Theorem 5.6 in [26] which allows us to deduce the C-tightness of the sequence of

martingales  $\hat{M}_\Sigma^\lambda(t \wedge T^\lambda)$  from the C-tightness of the sequence of predictable-quadratic-variation processes,  $\langle \hat{M}_\Sigma^\lambda \rangle(\cdot \wedge T^\lambda)$  which, in turn, follows from Lemma 4.1. Finally,  $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$  is C-tight by (64). ■

We are now ready to prove state-space collapse for the stopped processes. With the definition of the processes  $\hat{U}_i^\lambda(t)$  and  $\hat{V}_j^\lambda(t)$  in (15) and (16), state-space collapse for the stopped processes is equivalent to the following theorem, as we show in the corollary below:

**Theorem 4.3 (state-space collapse for the stopped-processes)** *For any  $\epsilon > 0$  and  $s, 0 < s < T$ ,*

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{s \leq t \leq T^\lambda} \sum_{i \in \mathcal{I}} |\hat{U}_i^\lambda(t)| > \epsilon \right\} = 0, \text{ and } \limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{s \leq t \leq T^\lambda} \sum_{j \in \mathcal{J}} |\hat{V}_j^\lambda(t)| > \epsilon \right\} = 0. \quad (65)$$

*If, in addition, equation (9) holds then*

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{0 \leq t \leq T^\lambda} \sum_{i \in \mathcal{I}} |\hat{U}_i^\lambda(t)| > \epsilon \right\} = 0, \text{ and } \limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{0 \leq t \leq T^\lambda} \sum_{j \in \mathcal{J}} |\hat{V}_j^\lambda(t)| > \epsilon \right\} = 0 \quad (66)$$

The fact that Theorem 4.3 implies the desired form of state-space collapse follows from the following simple Corollary:

**Corollary 4.4** *Theorem 4.3 implies that*

$$\hat{Q}_i^\lambda(t \wedge T^\lambda) - \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) p_i \left( \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) \right) \Rightarrow 0 \text{ as } \lambda \rightarrow \infty \text{ for all } i \in \mathcal{I},$$

*and*

$$\hat{I}_j^\lambda(t \wedge T^\lambda) - \hat{I}_\Sigma^\lambda(t \wedge T^\lambda) v_j \left( \hat{I}_\Sigma^\lambda(t \wedge T^\lambda) \right) \Rightarrow 0 \text{ as } \lambda \rightarrow \infty \text{ for all } j \in \mathcal{J},$$

*where the convergence is in  $D$  or  $D_-$  depending, as before, on whether or not equation (9) holds.*

**Proof:** We prove the result for the queue processes. The proof for the idleness processes is similar. Given (15), Theorem 4.3 implies, in particular, that  $\hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) - [\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)]^+ \Rightarrow 0$ , where the convergence is in  $D$  if equation (9) holds, and its  $D_-$  otherwise. In turn, using the continuity of the state-dependent ratio

functions and applying the continuous mapping theorem (see §3.4 of Whitt [36]), we have that

$$p_i \left( \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) \right) - p_i \left( [\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)]^+ \right) \Rightarrow 0, \text{ as } \lambda \rightarrow \infty.$$

Consequently,

$$\hat{U}_i^\lambda(t \wedge T^\lambda) - \left( \hat{Q}_i^\lambda(t \wedge T^\lambda) - \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) p_i \left( \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) \right) \right) \Rightarrow 0,$$

implying finally that

$$\hat{Q}_i^\lambda(t \wedge T^\lambda) - \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) p_i \left( \hat{Q}_\Sigma^\lambda(t \wedge T^\lambda) \right) \Rightarrow 0,$$

where the convergence is in  $D$  or  $D_-$  depending, as before, on whether or not equation (9) holds.  $\blacksquare$

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

In preparation for the proof of Theorem 4.3, we first introduce some of the required framework. Our state-space collapse proof will be based on the general state-space collapse framework of Bramson [8], which was recently extended to the many-server setting by Dai and Tezcan [12]. Since some of the assumptions in [12] do not apply in our setting, we give an independent proof.

The framework of [8] 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 state-space collapse, 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 [8] 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

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

We will also use  $D_j^\lambda(t)$ , which was defined in (48). We start with the basic processes indexed by  $\lambda$ :

$$\mathbb{X}^\lambda(t) := \left( A_i^\lambda(t), A_{i,j}^\lambda(t), \Phi_{i,j}^\lambda(t), D_j^\lambda(t), R_i^\lambda(t), \Theta_i^\lambda(t), Q_i^\lambda(t), U_i^\lambda(t), Z_j^\lambda(t); i \in \mathcal{I}, j \in \mathcal{J} \right). \quad (67)$$

Then, for any fixed  $L > 0$ , and every non-negative integer  $m$  with  $m < \sqrt{\lambda}L$ , 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:

$$A_i^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( A_i^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - A_i^\lambda \left( \frac{m}{\sqrt{\lambda}} \right) \right), \quad (68)$$

$$A_{i,j}^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( A_{i,j}^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - A_{i,j}^\lambda \left( \frac{m}{\sqrt{\lambda}} \right) \right), \quad (69)$$

$$\Phi_{i,j}^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( \Phi_{i < j}^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - \Phi_{i,j}^\lambda \left( \frac{m}{\sqrt{\lambda}} \right) \right), \quad (70)$$

$$D_j^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( D_j^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - D_j^\lambda \left( \frac{m}{\sqrt{\lambda}} \right) \right), \quad (71)$$

$$R_i^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( R_i^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - R_i^\lambda \left( \frac{m}{\sqrt{\lambda}} \right) \right), \quad (72)$$

$$\Theta_i^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( \Theta_i^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - \Theta_i^\lambda \left( \frac{m}{\sqrt{\lambda}} \right) \right), \quad (73)$$

$$Q_i^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( Q_i^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) \right), \quad (74)$$

$$U_i^{\lambda,m}(t) := \left( \hat{U}_i^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) \right), \text{ and} \quad (75)$$

$$Z_j^{\lambda,m}(t) := \frac{1}{\sqrt{\lambda}} \left( Z_j^\lambda \left( \frac{t}{\sqrt{\lambda}} + \frac{m}{\sqrt{\lambda}} \right) - N_j^\lambda \right). \quad (76)$$

**Remark 4.1 (the time and space scaling)** The HS processes in equation (68)-(76) have a more elementary time-and-space scaling than in [8] and [12] - see §6.1 of [12] and, in particular, equation (6.1) there. The latter, more complex, scaling is required if we can not prove a priori that the queue and idleness processes are stochastically bounded. In the absence of stochastic boundedness, the resulting notion of state-space collapse is the multiplicative state-space collapse, defined in Theorem 5.1 and Remark 5.3 of [12]. However, when stochastic boundedness is available, state-space collapse and multiplicative state-space collapse are equivalent. Then both can be proved using the more elementary scaling, as illustrated in the proof of Theorem 7.3 in [12]. Since our state-space collapse argument will focus initially on the stopped processes,

the stochastic boundedness established in Lemma 4.2 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

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

satisfying the following hydrodynamic model (HM) equations. However, the following HM equations, in general, do not determine the HL functions uniquely. Thus we will be showing that there is some HL function satisfying the HM equations that is suitably close to the HS process above. Since 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 (77)–(86) below:

$$\tilde{A}_i(t) = a_i t, \quad i \in \mathcal{I}, \quad (77)$$

$$\tilde{\Theta}_i(t) = \tilde{R}_i(t) = 0, \quad \text{for all } t \geq 0, \quad (78)$$

$$\tilde{Q}_i(t) = \tilde{Q}_i(0) + \tilde{A}_i(t) - \sum_{j \in \mathcal{J}} \tilde{A}_{ij}(t) - \sum_{j \in \mathcal{J}} \tilde{\Phi}_{ij}(t), \quad i \in \mathcal{I}, \quad (79)$$

$$\tilde{Q}_i(t) \geq 0, \quad i \in \mathcal{I}, \quad (80)$$

$$\tilde{A}_{ij}(t), \tilde{\Phi}_{ij}(t), \quad i \in \mathcal{I}, j \in \mathcal{J} \text{ are non-decreasing,} \quad (81)$$

$$\tilde{Z}_j(t) = \tilde{Z}_j(0) + \sum_{i \in \mathcal{I}} \tilde{A}_{ij}(t) + \sum_{i \in \mathcal{I}} \tilde{\Phi}_{ij}(t) - \mu_j \bar{\nu}_j t, \quad j \in \mathcal{J}. \quad (82)$$

Equations (77)–(82) appear in the hydrodynamic model of an arbitrary policy. Letting  $\tilde{I}^+(t) := \{i \in \mathcal{I} : \tilde{U}_i(t) > 0\}$ , and  $J(\tilde{I}^+(t)) := \{j \in \mathcal{J} : i \in J(i), \text{ for some } i \in \tilde{I}^+(t)\}$ , we also define the following HM

equations that are specific to FQR

$$\sum_{i \in \mathcal{I}} \tilde{U}_i(t) \geq 0, \quad (83)$$

$$\tilde{U}_i(t) = \tilde{U}_i(0) + \tilde{A}_i(t) - \sum_{j \in \mathcal{J}} \tilde{A}_{ij}(t) - \sum_{j \in \mathcal{J}} \tilde{\Phi}_{ij}(t), \quad i \in \mathcal{I}, \quad (84)$$

$$\sum_{i \in \tilde{I}^+(t)} \sum_{j \in \mathcal{J}(i)} d\tilde{\Phi}_{ij}(t) = \sum_{j \in \mathcal{J}(\tilde{I}^+(t))} \mu_j \bar{\nu}_j, \quad \text{and in particular,} \quad (85)$$

$$\sum_{i \in \tilde{I}^+(t)} d\tilde{U}_i(t) \leq -c, \quad \text{whenever } \tilde{I}^+(t) \neq \emptyset \text{ for some constant } c > 0. \quad (86)$$

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  $|x(0)| \leq k$  and

$$|x(t_2) - x(t_1)| \leq N|t_2 - t_1| \quad \text{for all } t_1, t_2 \in [0, L_k].$$

By the Arzela-Ascoli theorem, p. 221 of Billingsley [6], the set  $E'$  is a compact subset of  $C^d$  and thus of  $D^d$  for appropriate  $d$ . Since the HL functions are Lipschitz, they are absolutely continuous; e.g., see §5.4 of Royden [30].

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 4.4 and 4.5 below. Note that the final HM equation, equation (86), 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 (83); see Lemma 5.2 of Dai [11] for technical support. We will use this fact in the proof of Theorem 4.3.

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  $\lambda$  large enough the HS process is close enough to some HL function  $\tilde{X}$ , satisfying the HM equations (77)-(86).

**Theorem 4.5 (uniform approximation by HL functions)** *For any  $T > 0$ ,  $\delta > 0$  and  $\epsilon > 0$ , there exist  $k := k(\delta, \epsilon, T)$ ,  $L_k$ ,  $\lambda_0$  and subsets  $\mathcal{K}^{\lambda, k}$  of the underlying probability space  $\Omega$ , such that for all  $\lambda \geq \lambda_0$ :*

$$1. \quad \|\hat{U}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{Z}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{Q}^\lambda(\cdot \wedge T^\lambda)\|_T^* \leq k \quad \text{on } \mathcal{K}^{\lambda, k},$$

2. 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)  $\tilde{X}$ , depending on  $\lambda$  and  $m$ , such that

$$\|\mathbb{X}^{\lambda,m} - \tilde{X}\|_{L_k}^* \leq \epsilon.$$

3. Finally,  $P\{\mathcal{K}^{\lambda,k}\} \geq 1 - \delta$ .

Theorem 4.5 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  $\lambda$  sufficiently large, the HS process is close enough to some HL function satisfying the HM equations. We postpone the proof of Theorem 4.5 and apply it now to prove Theorem 4.3.

**Proof of Theorem 4.3:** The proof of the first conclusion consists of two steps. First, we focus on an HL function, fix  $\epsilon$  and  $k$  and show that there exists some finite time  $s^* := s^*(k, \epsilon)$  such that

$$\tilde{U}_i(t) = 0 \quad \text{for } t \geq s^* \quad \text{and for all } i$$

provided that the HL function  $\tilde{X}$  satisfies the HM equations (77)-(86) and  $\sum_{i \in \mathcal{I}} |\tilde{U}_i(0)| \leq k + \epsilon$ .

Property (86) implies that the HL function  $\sum_{i \in \mathcal{I}} [\tilde{U}_i(t)]^+$  decreases at a rate of at least  $c$  until it reaches 0, after which it stays there; see Lemma 5.2 of Dai [11] for technical support. Property (86) directly controls only the positive part, but condition (83) implies that the negative part is dominated by the positive part in absolute value. Hence, when  $\sum_{i \in \mathcal{I}} [\tilde{U}_i(t)]^+ = 0$ , we also have  $\tilde{U}_i(t) = 0$  for all  $i$  by virtue of condition (83). And these functions must remain 0 thereafter, because the positive part cannot increase.

Thus we have the existence of  $s^* := s^*(k, \epsilon)$  such that  $\tilde{U}_i(t) = 0$  for  $t \geq s^*$  for all  $i$  for any HL function  $\tilde{X}$  satisfying the HM equations (77)-(86) with  $\sum_{i \in \mathcal{I}} |\tilde{U}_i(0)| \leq k + \epsilon$ . 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  $\lambda_0$  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

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

Then  $t \in \left[ \frac{m^\lambda(t)}{\sqrt{\lambda}}, \frac{m^\lambda(t)+L_k}{\sqrt{\lambda}} \right]$ , and, by definition of the HS process,

$$\hat{U}_i^\lambda(t) = \hat{U}_i^{\lambda, m^\lambda(t)} \left( \sqrt{\lambda}t - m^\lambda(t) \right).$$

Now, by Theorem 4.5, for  $\lambda$  large enough, there exists an HL function  $\tilde{X}$  that satisfies the HM equations (77)-(86) with

$$|U_i^{\lambda, m^\lambda(t)} - \tilde{U}_i|_{L_k^*} \leq \epsilon.$$

In particular, we have that  $\sum_{i \in \mathcal{I}} |\tilde{U}_i(0)| \leq k + \epsilon$ , which by our previous argument implies that

$$\sum_{i \in \mathcal{I}} |\tilde{U}_i(t)| = 0, \quad t \geq s^*.$$

Since  $L_k \geq \sqrt{\lambda}t - m^\lambda(t) \geq s^*$ , we then have that

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

Combining these relations, we then have

$$\sum_{i \in \mathcal{I}} \left| \hat{U}_i^\lambda(t) \right| \leq \epsilon.$$

Hence,

$$\sup_{s^*/\sqrt{\lambda} \leq t \leq T^\lambda} \sum_{i \in \mathcal{I}} |\hat{U}_i^\lambda(t)| \leq \epsilon \quad \text{on } \mathcal{K}^{\lambda, k}, \quad (87)$$

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

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

from which (65) readily follows.

For the second conclusion, assume that (9) holds and let

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

Then, (9) implies

$$\lim_{\lambda \rightarrow \infty} P\{\tilde{\Omega}^\lambda\} = 1.$$

Consider the process  $U_i^{\lambda,0}(t)$  and its corresponding approximation  $\tilde{U}_i$  from Theorem 4.5. Then, on  $\mathcal{K}^{\lambda,k} \cap \tilde{\Omega}^\lambda$ , we must have that  $\sum_{i \in \mathcal{I}} |\tilde{U}_i(0)| \leq \epsilon$ , and repeating the same argument we used above we will have that

$$\sum_{i \in \mathcal{I}} |\tilde{U}_i(t)| \leq \epsilon, \text{ for all } t \geq 0.$$

In particular,  $\hat{U}_i^\lambda(t) \leq 2\epsilon$  for all  $t \leq T^\lambda \wedge s^*/\sqrt{\lambda}$ . Adding this to (87), we have that on  $\mathcal{K}^{\lambda,k} \cap \tilde{\Omega}^\lambda$ ,

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

As for all  $\lambda$  large enough  $P\left\{\mathcal{K}^{\lambda,k} \cap \tilde{\Omega}^\lambda\right\} \geq 1 - 2\delta$  we have established that

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

implying (66). ■

**Proof of Theorem 4.5** The proof is similar to corresponding proofs in Bramson [8] and in Dai and Tezcan [12] and, specifically, it parallels the proofs in §C.2-C.3 of [12]. However, there are some minor differences. Hence, we write out the proofs, but abbreviate whenever the proof closely follows either [8] or [12].

The proof is divided into three lemmas. Lemma 4.3 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 4.4 to establish the uniform approximation by cluster points. Together with Lemmas 4.1 and 4.2 and Proposition 4.1 of Bramson [8], Lemmas 4.3 and 4.4 here imply the uniform approximation by a Lipschitz function. Bramson [8] elaborates on the compactness and cluster-point structure. Finally, Lemma 4.5 below establishes that each cluster point satisfies equations (77)-(86). That primarily means the last two equations: (85) and (86).

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

$$\Omega_1^{\lambda,k} := \left\{ \omega \in \Omega : \|\hat{U}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{Z}^\lambda(\cdot \wedge T^\lambda)\|_T^* + \|\hat{Q}^\lambda(\cdot \wedge T^\lambda)\|_T^* \leq k \right\}, \quad (88)$$

$$\Omega_2^{\lambda,k} := \Omega_2^{\lambda,k}(\epsilon) := \left\{ \omega \in \Omega : \max_{m < \sqrt{\lambda}T^\lambda} \|A^{\lambda,m}(t) - at\|_{L_k}^* \leq \epsilon \right\}, \quad (89)$$

$$\Omega_3^{\lambda,k} := \Omega_3^{\lambda,k}(\epsilon) := \left\{ \omega \in \Omega : \max_{m < \sqrt{\lambda}T^\lambda} \|D^{\lambda,m}(t) - \mu\bar{v}t\|_{L_k}^* \leq \epsilon \right\}, \quad (90)$$

where  $A^{\lambda,m}(t) = (A_1^{\lambda,m}(t), \dots, A_I^{\lambda,m}(t))$ ,  $a = (a_1, \dots, a_I)$ ,  $D^{\lambda,m}(t) = (D_1^{\lambda,m}(t), \dots, D_J^{\lambda,m}(t))$  and  $\mu\bar{v} = (\mu_1\bar{v}_1, \dots, \mu_J\bar{v}_J)$ . Finally, let

$$\Omega_4^{\lambda,k} := \Omega_4^{\lambda,k}(\epsilon, N) := \left\{ \omega \in \Omega : \max_{m < \sqrt{\lambda}T^\lambda} \sup_{t_1, t_2 \leq L_k} \|\mathbb{X}^{\lambda,m}(t_2) - \mathbb{X}^{\lambda,m}(t_1)\| \leq N|t_2 - t_1| + \epsilon \right\}, \quad (91)$$

where  $N$  is some fixed constant that depends only on the vectors  $a, \bar{v}$  as well as  $I$  and  $J$  and whose specific value will be made explicit within the proof of the following lemma. Set  $\mathcal{K}^{\lambda,k} = \bigcap_{i=1}^4 \Omega_i^{\lambda,k}$ . The following lemma is the analogue of Propositions 6.2 and 6.3 in [12].

**Lemma 4.3** *For any  $\epsilon$  and  $\delta > 0$ , there exist  $k$  and  $L_k$  so that*

$$\liminf_{\lambda \rightarrow \infty} P\{\mathcal{K}^{\lambda,k}\} \geq 1 - \delta$$

for  $\mathcal{K}^{\lambda,k}$  defined above.

**Proof:** By Lemma 4.2,  $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$ ,  $\hat{Q}_\Sigma^\lambda(t \wedge T^\lambda)$  and  $\hat{I}_\Sigma^\lambda(t \wedge T^\lambda)$  are stochastically bounded. As  $\hat{Z}_j(t) = -\hat{I}_j^\lambda(t)$  and, since by definition,

$$\sum_{i \in \mathcal{I}} |\hat{U}_i^\lambda(t)| \leq \hat{Q}_\Sigma^\lambda(t) + |\hat{X}_\Sigma^\lambda(t)|,$$

we have that

$$\lim_{k \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} P\left\{ \Omega - \Omega_1^{\lambda,k} \right\} = 0. \quad (92)$$

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  $t \leq L_k$ ,

$$\begin{aligned} D_j^{\lambda,m}(t) &= \frac{1}{\sqrt{\lambda}} \left( S_j \left( \mu_j \int_0^{(m+t)/\sqrt{\lambda}} Z_j^\lambda(s) ds \right) - S_j \left( \mu_j \int_0^{m/\sqrt{\lambda}} Z_j^\lambda(s) ds \right) \right) \\ &\stackrel{d}{=} \frac{1}{\sqrt{\lambda}} \left( S_j \left( \mu_j \int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_j^\lambda(s) ds \right) \right), \end{aligned} \quad (93)$$

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

$$\left\| D_j^{\lambda,m}(t) - \frac{1}{\sqrt{\lambda}} \mu_j \int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_j^\lambda(s) ds \right\|_{L_k}^* \stackrel{d}{=} \frac{1}{\sqrt{\lambda}} \left\| S_j(\mu_j N_j^\lambda t) - \mu_j N_j^\lambda t \right\|_{\psi^\lambda}^*, \quad (94)$$

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 [8], 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

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \left\| D_j^{\lambda,m}(t) - \frac{1}{\sqrt{\lambda}} \mu_j \int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_j^\lambda(s) ds \right\|_{L_k}^* > \epsilon L_k \right\} \leq \frac{\epsilon \lambda T}{\mu_j N_j^\lambda L_k} \leq \frac{2\epsilon T}{\mu_j \bar{\nu}_j L_k}, \quad (95)$$

for all  $\lambda$  large enough, since  $N_j^\lambda/\lambda \rightarrow \bar{\nu}_j$  as  $\lambda \rightarrow \infty$ . We can then replace  $\epsilon$  with  $\epsilon/L_k$  and choose  $L_k$  large enough so that the probability in (95) is bounded by  $\delta/8J$ . Finally, since  $\hat{I}_\Sigma^\lambda(t \wedge T^\lambda)$  is stochastically bounded (by Lemma 4.2), and since  $N_j^\lambda = \bar{\nu}_j \lambda + \gamma_j \sqrt{\lambda} + o(\sqrt{\lambda})$ , we have

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

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

have, for all  $\lambda$  large enough,

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

so that

$$P\{\Omega_3^{\lambda, k}\} \geq 1 - \delta/4. \quad (96)$$

The similar but easier argument for  $\Omega_2^{\lambda, k}$  shows that

$$P\{\Omega_2^{\lambda, k}\} \geq 1 - \delta/4, \quad (97)$$

for all  $\lambda$  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  $Z_j^\lambda(s) \leq N_j^\lambda$  for all  $s \geq 0$ , we have that

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

For  $t_1, t_2 \leq L_k$ ,

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

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

$$\begin{aligned} P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |D_j^{\lambda, m}(t_2) - D_j^{\lambda, m}(t_1)| > \mu_j \frac{N_j^\lambda}{\lambda} |t_2 - t_1| + \epsilon' \right\} \\ \leq \sqrt{\lambda} T \cdot P \left\{ \sup_{t \leq \frac{L_k}{\sqrt{\lambda}}} |S_j(\mu_j N_j^\lambda t) - \mu_j N_j^\lambda t| > \frac{\epsilon'}{2} \sqrt{\lambda} \right\}. \end{aligned} \quad (98)$$

Fixing  $\delta' > 0$ , applying Proposition 4.3 of [8] to the right-hand side of (98), and using the fact that  $N_j^\lambda / \lambda \rightarrow$

$\bar{\nu}_j$  as  $\lambda \rightarrow \infty$ , we have, for fixed  $\delta' > 0$ ,

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |D_j^{\lambda, m}(t_2) - D_j^{\lambda, m}(t_1)| > \mu_j \bar{\nu}_j |t_2 - t_1| + \epsilon' \right\} \leq \delta', \quad (99)$$

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

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |A_i^{\lambda, m}(t_2) - A_i^{\lambda, m}(t_1)| > a_i |t_2 - t_1| + \epsilon' \right\} \leq \delta', \quad (100)$$

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 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) and (18) to get, for all  $i \in \mathcal{I}, j \in \mathcal{J}$ ,

$$|A_{i,j}^{\lambda, m}(t_2) - A_{i,j}^{\lambda, m}(t_1)| \leq |A_i^{\lambda, m}(t_2) - A_i^{\lambda, m}(t_1)|, \quad (101)$$

and

$$|\Phi_{i,j}^{\lambda, m}(t_2) - \Phi_{i,j}^{\lambda, m}(t_1)| \leq |D_j^{\lambda, m}(t_2) - D_j^{\lambda, m}(t_1)|. \quad (102)$$

Combining (99)–(102), we obtain

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |A_{i,j}^{\lambda, m}(t_2) - A_{i,j}^{\lambda, m}(t_1)| > N' |t_2 - t_1| + \epsilon' \right\} \leq \delta', \quad i \in \mathcal{I}, j \in \mathcal{J}, \quad (103)$$

and

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |\Phi_{i,j}^{\lambda, m}(t_2) - \Phi_{i,j}^{\lambda, m}(t_1)| > N' |t_2 - t_1| + \epsilon' \right\} \leq \delta', \quad i \in \mathcal{I}, j \in \mathcal{J}, \quad (104)$$

for all  $\lambda$  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,

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

Hence,

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

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

$$\sup_{t \leq L_k} \left| R_i^{\lambda,m}(t) - \frac{1}{\sqrt{\lambda}} \int_{(m+t_1)/\sqrt{\lambda}}^{(m+t_2)/\sqrt{\lambda}} Q_i^\lambda(s) ds \right| \leq_{st} \frac{1}{\sqrt{\lambda}} \sup_{t \leq L_k/\sqrt{\lambda}} \left| R_i(\theta_i k \sqrt{\lambda} t) - \theta_i k \sqrt{\lambda} t \right|.$$

Applying Proposition 4.3 of [8] once again, we have, for all  $\lambda$  large enough,

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

Since  $\theta_i L_k / \sqrt{\lambda} \leq \epsilon'/2$  and  $P\{(\Omega_1^{\lambda,k})^c\} \leq \delta'/2$  for all  $\lambda$  large enough,

$$P \left\{ \sup_{t_1, t_2 \leq L_k} |R_i^{\lambda,m}(t_2) - R_i^{\lambda,m}(t_1)| > \epsilon' \right\} \leq \delta', \quad (105)$$

for all  $\lambda$  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 that choose new values of  $\delta'$  and  $\epsilon'$  and combine (100)–(105) to obtain

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |Q_i^{\lambda,m}(t_2) - Q_i^{\lambda,m}(t_1)| > N'' |t_2 - t_1| + \epsilon'' \right\} \leq \delta'', \quad i \in \mathcal{I}, \quad (106)$$

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

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |Z_j^{\lambda,m}(t_2) - Z_j^{\lambda,m}(t_1)| > N'' |t_2 - t_1| + \epsilon'' \right\} \leq \delta'', \quad j \in \mathcal{J}, \quad (107)$$

for all  $\lambda$  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,  $\|\hat{X}_\Sigma^\lambda\|_{T^\lambda}^* \leq k$ , and the Hölder condition on the ratio function (see Definition 2.2) implies that there exists constants  $c_k$  and  $\alpha_k$ , depending

on  $k$  and  $\epsilon'$  but not on  $\lambda$ , so that for all  $0 < t_1 \leq t_2 \leq T^\lambda$ ,

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

see for example equation (118) in [5]. Hence,

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |\Theta_i^{\lambda, m}(t_2) - \Theta_i^{\lambda, m}(t_1)| > \epsilon' \right\} \leq P\{(\Omega_1^{\lambda, k})^{c_1}\} \\ + P \left\{ \sup_{0 \leq t_1 < t_2 \leq T^\lambda: |t_2 - t_1| \leq L_k / \sqrt{\lambda}} c_k |\hat{X}_\Sigma^\lambda(t_2) - \hat{X}_\Sigma^\lambda(t_1)|^{\alpha_k} > \epsilon'/2 \right\}$$

Since we can choose  $k$  large enough so that for any  $\lambda$  large enough  $P\{\Omega_1^{\lambda, k}\} \geq 1 - \delta'/2$  for any  $\lambda$  large enough, since  $\hat{X}_\Sigma^\lambda(t \wedge T^\lambda)$  is C-tight by Lemma 4.2, and since we can move the exponent  $\alpha_k$ , first outside the supremum and then to the other side by raising to the reciprocal power, we can conclude that

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |\Theta_i^{\lambda, m}(t_2) - \Theta_i^{\lambda, m}(t_1)| > \epsilon' \right\} \leq \delta'. \quad (108)$$

Combining (106) and (108) and using the fact that

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

we can choose new values of  $\epsilon'$  and  $\delta'$  so that

$$P \left\{ \max_{m < \sqrt{\lambda} T^\lambda} \sup_{t_1, t_2 \leq L_k} |U_i^{\lambda, m}(t_2) - U_i^{\lambda, m}(t_1)| > N'' |t_2 - t_1| + \epsilon'' \right\} \leq \delta'', \quad (109)$$

for all  $\lambda$  large enough. Now set  $N = 8(I + J + IJ)$ . Then, we can combine (99), (100) and (104)–(109) and choose new values  $\delta'$ ,  $\delta''$ ,  $\epsilon'$  and  $\epsilon''$  appropriately to conclude that

$$P\{\Omega_4^{\lambda, k}\} \geq 1 - \delta/4, \quad (110)$$

for all  $\lambda$  large enough. Finally, (92), (96), (97), and (110) are combined to conclude that

$$P\{\mathcal{K}^{\lambda, k}\} \geq 1 - \delta,$$

for some  $k, L_k, \lambda_0$  and for all  $\lambda \geq \lambda_0$ . ■

Having proved that the family of HS processes is approximately Lipschitz, 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 [8]. The proof is exactly as in [8] and is hence omitted.

**Lemma 4.4** *For all  $\epsilon > 0$ , there exists  $k > 0, L_k > 0$  and  $\lambda_0$ , so that, for all  $\lambda > \lambda_0, \omega \in \mathcal{K}^{\lambda,k}$  and  $m < \sqrt{\lambda T^\lambda}$ , there exists an HL function  $\tilde{\mathbb{X}}$  (a Lipschitz function satisfying the HM equations) such that*

$$\|\mathbb{X}^{\lambda,m} - \tilde{\mathbb{X}}\|_{L_k}^* \leq \epsilon. \quad (111)$$

Lemmas 4.3 and 4.4 combined show that, given  $\epsilon > 0$  and  $\delta > 0$ , we can choose a set  $\mathcal{K}^{\lambda,k}$  (which also depends on  $\epsilon$ ) 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 [8] imply that the approximating HL function is Lipschitz.

To establish Theorem 4.5, it remains only to show that all the hydrodynamic limits satisfy equations (77)-(86). That is done in the following lemma. The lemma below is mostly an analogue of Proposition 6.6 in [12]. The major difference is to show that the QIR-specific equations (83)-(86) hold for any hydrodynamic limit  $\tilde{X}$ .

**Lemma 4.5** *Fix  $k > 0, L_k > 0$  and let  $\tilde{X}$  be a hydrodynamic limit of the family of HS processes  $\mathbb{X}^{\lambda,m}$  over  $[0, L_k]$ . Then  $\tilde{X}$  satisfies equations (77)-(86).*

**Proof:** Using the definitions of the HL function and the set  $\mathcal{K}^{\lambda,k}$ , it is immediate that any HL function satisfies equations (77)-(84); see the proof of Proposition 6.6 in [12]). We turn, then, to prove that any hydrodynamic limit  $\tilde{X}$  satisfies (85).

Toward that end, recall the definition  $\tilde{I}^+(t) = \{i \in \mathcal{I} : \tilde{U}_i(t) > 0\}$  and consider  $t \geq 0$  with  $\tilde{I}^+(t) \neq \emptyset$ . Let

$$\tilde{\epsilon}(t) = \min_{i \in \tilde{I}^+(t)} \tilde{U}_i(t).$$

Since every HL function  $x$  is absolutely continuous (and thus continuous), we must have an interval  $[t - \tau, t + \tau]$  such that for all  $u \in [t - \tau, t + \tau]$ ,  $\min_{i \in \tilde{I}^+(u)} \tilde{U}_i(u) \geq \tilde{\epsilon}(t)/2$ . Moreover, by the continuity of  $\tilde{U}_i(t)$ , the interval can be chosen so that  $\tilde{U}_i(u) \leq \tilde{\epsilon}(t)/8$  for all  $i \notin \tilde{I}^+(t)$  and  $u \in [t - \tau, t + \tau]$ . Next, since  $\tilde{X}$  is an HL function, we can fix  $k$  and argue that there exists  $\lambda$  large enough,  $\omega \in \mathcal{K}^{\lambda, k}$  and  $m < \sqrt{\lambda}T^\lambda$ , so that,

$$\|U_i^{\lambda, m} - \tilde{U}_i\|_{L_k}^* \leq \epsilon,$$

for  $\epsilon \leq \tilde{\epsilon}(t)/8$ . In particular, we can choose  $L_k$  (and appropriately choose anew value of  $\lambda$ ) so that there exists a neighborhood  $[t - \tau', t + \tau']$  of  $t$  such that for all  $u \in [t - \tau', t + \tau']$ ,

$$U_i^{\lambda, m}(u) \geq \frac{3}{8}\tilde{\epsilon}(t), \quad i \in \tilde{I}^+(t), \quad \text{and} \quad U_i^{\lambda, m}(u) \leq \frac{2}{8}\tilde{\epsilon}(t), \quad i \notin \tilde{I}^+(t).$$

Then, for any  $j \in J(\tilde{I}^+(t)) = \{j \in \mathcal{J} : i \in I(j) \text{ for some } i \in \tilde{I}^+(t)\}$ ,

$$\operatorname{argmax}_{i \in I(j)} U_i^{\lambda, m}(u) \in \tilde{I}^+(t) \tag{112}$$

for all  $u \in [t - \tau', t + \tau']$ . In turn, by the definition of  $\Phi_{i,j}^\lambda(t)$  (see equation (18)) we have that

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

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,  $\|D_j^{\lambda, m}(t) - \mu_j \bar{\nu}_j t\|_{L_k}^* \leq \epsilon$  on  $\mathcal{K}^{\lambda, k}$ . Hence,

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

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

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

Since  $\epsilon$  was chosen arbitrarily and the bound holds for any  $\tilde{\tau} \leq \tau'$ ,

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

Equation (85) now follows. As for equation (86), note that

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

where the last inequality follows from (85). We now observe that, due to Assumptions 2.3 and 2.2, there exists  $c > 0$  such that for any strict subset  $B \subset \mathcal{I}$ ,

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

Equation (86) 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 (77)-(86) and the proof of the lemma is complete.  $\blacksquare$

With Lemma 4.5 we have completed the proofs of Theorems 4.5 and 4.3 for the stopped processes. Theorem 4.3, under condition C-2, is then established by showing in the following lemma that  $\sigma \Rightarrow \infty$  as  $\lambda \rightarrow \infty$ . That implies the state-space collapse extends to the whole interval  $[0, T]$ .

**Lemma 4.6** *for each  $T > 0$ ,*

$$\lim_{\lambda \rightarrow \infty} P\{\sigma^\lambda \leq T\} = 0. \quad (113)$$

**Proof:** By definition,  $\hat{B}^\lambda(t) = \sum_{i \in \mathcal{I}} \hat{U}_i^\lambda(t) \geq 0$ . Using (54), we can write

$$P\{\sigma^\lambda \leq T\} = P\left\{|\hat{B}^\lambda|_{\sigma^\lambda \wedge T}^* \geq 2\hat{B}^\lambda(0) \vee 1\right\} = P\left\{\sup_{0 \leq t \leq T^\lambda} \sum_{i \in \mathcal{I}} \hat{U}_i^\lambda(t) > 2\hat{B}^\lambda(0) \vee 1\right\}, \quad (114)$$

where we use the equivalence of the events  $\{\sigma^\lambda \leq T\}$  and  $\{\hat{B}^\lambda(T^\lambda) \geq 2\hat{B}^\lambda(0) \vee 1\}$  as follows from the definition of  $\sigma^\lambda$  in equation (54). In order to establish that  $P\{\sigma^\lambda \leq T\} \rightarrow 0$  as  $\lambda \rightarrow \infty$ , it suffices to prove that the right hand side of (114) converges to 0. Toward that end, fix  $\delta > 0$ . Then, there exists  $k$  such that  $P\{\mathcal{K}^{\lambda,k}\} \geq 1 - \delta$  for all  $\lambda$  large enough. Fix  $0 < \epsilon < 1/2$ . By Theorem 4.5, for all  $\lambda$  large enough and  $\omega \in \mathcal{K}^{\lambda,k}$ , there exists  $\tilde{U}_i(t)$  that satisfies (83)-(86) such that

$$\|U^{\lambda,0}(t) - \tilde{U}(t)\|_{L_k}^* \leq \epsilon$$

By equation (86),

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

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

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

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

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

Choosing  $L_k \geq 2\lceil s^* \rceil$  with  $s^*$  as defined in the proof of Theorem 4.3 and repeating the arguments in that proof, we have

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

so that

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

We conclude by noting that

$$\begin{aligned} P \left\{ \sup_{0 \leq t \leq T^\lambda} \sum_{i \in \mathcal{I}} \hat{U}_i^\lambda(t) > 2\hat{B}^\lambda(0) \vee 1 \right\} &\leq P \left\{ \sup_{0 \leq t \leq L_k/\sqrt{\lambda}} \sum_{i \in \mathcal{I}} \hat{U}_i^\lambda(t) > 2\hat{B}^\lambda(0) \vee 1 \right\} \\ &+ P \left\{ \sup_{L_k/\sqrt{\lambda} \leq t \leq T^\lambda} \sum_{i \in \mathcal{I}} \hat{U}_i^\lambda(t) > 2\hat{B}^\lambda(0) \vee 1 \right\} \leq 2\delta. \end{aligned} \quad (115)$$

Since  $\delta$  was arbitrary the proof is complete. ■

With Lemma 4.6 we have completed the proof of Theorem 4.3 and in turn the proof of Theorem 3.1 under condition C-2. Specifically, we have shown that, for all  $0 < s < T$ ,

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{s \leq t \leq T} \sum_{i \in \mathcal{I}} |\hat{U}_i^\lambda(t)| > \epsilon \right\} = 0. \quad (116)$$

If, in addition, equation (9) holds then

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \sup_{0 \leq t \leq T} \sum_{i \in \mathcal{I}} |\hat{U}_i^\lambda(t)| > \epsilon \right\} = 0. \quad (117)$$

The proof of state-space collapse for the processes  $\hat{V}_j^\lambda(t)$  follows similarly. Before turning to the proof of Theorem 3.1 under condition C-3, we state the following corollary that will be of use in §5.

**Corollary 4.6** *The sequences  $\hat{Q}_\Sigma^\lambda(t)$ ,  $\hat{I}_\Sigma^\lambda(t)$  and  $\hat{X}_\Sigma^\lambda(t)$  are stochastically bounded. i.e.,*

$$\lim_{A \rightarrow \infty} \limsup_{\lambda \rightarrow \infty} P \left\{ \|\hat{Q}_\Sigma^\lambda\|_T^* + \|\hat{I}_\Sigma^\lambda\|_T^* + \|\hat{X}_\Sigma^\lambda\|_T^* > A \right\} = 0. \quad (118)$$

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

**Proof:** These results have already been proved for the stopped processes in Lemma 4.2. This additional result then follows from Lemma 4.6. ■

#### 4.4 State-Space Collapse Under C-3

In §4.3, due to the pool-dependent service rates, we could use a somewhat less detailed description of the system dynamics than the general description given (11)-(18). We now return to that general description. In addition to the process  $\hat{B}^\lambda(t)$  defined in (53), 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  $\hat{Z}_\Sigma^\lambda(t) := \sum_{i \in \mathcal{I}} \hat{Z}_i^\lambda(t)$ . The arguments leading to (52) are immediately adapted to show that

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) - \beta_i t - \mu_i \int_0^t \hat{Z}_i^\lambda(s) ds - \theta_i \int_0^t \hat{Q}_i^\lambda(s) ds + \hat{M}_i^\lambda(t) + o(1) \quad \text{as } \lambda \rightarrow \infty, \quad (119)$$

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

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

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

$$\hat{X}_\Sigma^\lambda(t) = \hat{X}_\Sigma^\lambda(0) - \sum_{i \in \mathcal{I}} \beta_i t - \sum_{i \in \mathcal{I}} \mu_i \int_0^t \hat{Z}_i^\lambda(s) ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t \hat{Q}_i^\lambda(s) ds + \hat{M}_\Sigma^\lambda(t) + o(1), \quad (120)$$

again as  $\lambda \rightarrow \infty$ .

The proof proceeds through the same stopping argument used in §4.3, where  $T^\lambda := \sigma^\lambda \wedge T$  and  $\sigma^\lambda$  is defined as in (54). For the stopped processes, our proof is similar to the proof in §4.3. The main difference between the proofs for the different conditions, C-2 and C-3, 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 §4.3 are the same in this setting with the exception, of course, of replacing condition C-2 with condition C-3. Moreover, once the HS processes and HM equations are redefined, all the proofs are adapted from 4.3 with only minor changes, that 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 (67) by adding, for each  $i \in \mathcal{I}$  and  $j \in \mathcal{J}$ , the process of class- $i$  departures from pool  $j$ , given by

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

as well as the process of cumulative class- $i$  departures given by  $D_i^\lambda(t) := \sum_{j \in \mathcal{J}} D_{i,j}^\lambda(t)$ . The hydrodynamically-scaled processes are then defined as in (68)-(76) with the addition of the following: For  $m < \sqrt{\lambda}T$ , we define

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

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

and

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

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 equations (77)-(81) and (83)-(84) with the addition of the following equations:

$$\tilde{Z}_j(t) = \tilde{Z}_j(0) + \sum_{i \in J(i)} \tilde{A}_{ij}(t) + \sum_{i \in J(i)} \tilde{\Phi}_{ij}(t) - \tilde{D}_j(t), \quad j \in \mathcal{J}, \quad (121)$$

$$\tilde{D}_i(t) = a_i t, \quad i \in \mathcal{I}, \quad (122)$$

$$\tilde{D}'_{ij}(t) = 0, \quad i \in \mathcal{I}, j \in \mathcal{J}, \quad (123)$$

$$\sum_{i \in \tilde{I}^+(t)} \sum_{j \in J(i)} d\tilde{\Phi}_{ij}(t) \geq \sum_{i \in \tilde{I}^+(t)} a_i + c_1, \quad (124)$$

$$\sum_{i \in \tilde{I}^+(t)} d\tilde{U}_i(t) \leq -c_2, \quad \text{whenever } \tilde{I}^+(t) \neq \emptyset, \quad (125)$$

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

We redefine

$$\Omega_3^{\lambda,k} := \Omega_3^{\lambda,k}(\epsilon) := \left\{ \omega \in \Omega : \max_{m < \sqrt{\lambda} T^\lambda} \|D^{\lambda,m}(t) - at\|_{L_k}^* \leq \epsilon \right\}, \quad (126)$$

with  $D^{\lambda,m}(t) = (D_1^{\lambda,m}(t), \dots, D_I^{\lambda,m}(t))$  and add the set

$$\Omega_5^{\lambda,k} := \Omega_5^{\lambda,k}(\epsilon) := \left\{ \omega \in \Omega : \max_{(i,j) \in \mathcal{I} \times \mathcal{J}} \max_{m < \sqrt{\lambda} T^\lambda} \left\| D_{i,j}^{\lambda,m}(t) - \int_{m/\sqrt{\lambda}}^{(m+t)/\sqrt{\lambda}} Z_{i,j}^\lambda(s) ds \right\|_{L_k}^* \leq \epsilon \right\}. \quad (127)$$

The other subsets  $\Omega_1^{\lambda,k}$ ,  $\Omega_2^{\lambda,k}$  and  $\Omega_4^{\lambda,k}$  remain as defined in (88), (89) and (91), respectively. Finally, we redefine  $\mathcal{K}^{\lambda,k} = \bigcap_{i=1}^5 \Omega_i^{\lambda,k}$ .

With this definitions, all the statements of §4.3 are carried to this section without any change, and the adaptation of the proofs are straightforward, leading to the proof of state-space collapse under condition C-3.

This section completes the proof of Theorem 3.1. We now turn to prove some auxiliary stochastic-process limits under the assumption that condition C-2 holds, i.e., that service rates are pool dependent.

## 5 Auxiliary Results

In this section we establish some auxiliary results that build on the state-space collapse results in the previous sections. In §5.1 we establish stochastic-process limits under either conditions C-2 and C-3. The limits under these two conditions are of interest as they relate the complicated SBR model to much simpler models, namely, the single-class multi-type inverted-V model and the multi-class single-type V model; see Figure 2. In §5.2 we deduce the convergence of important performance measures from that of the sequence  $\hat{X}_\Sigma^\lambda(t)$ . The results in this section are used in our two subsequent papers [19] and [20] but are of interest in their own right.

Toward these ends, we define  $\hat{W}_i^\lambda(t) := \sqrt{\lambda}W_i^\lambda(t)$  to be the scaled virtual waiting time process of class- $i$  customers in the  $\lambda^{th}$  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  $x \geq 0$ :

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

### 5.1 Stochastic-process limits

The following stochastic-process limit shows the consequence of the state-space collapse; 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 processes imply the heavy-traffic Little's law discussed after Definition 2.1 in the main paper. 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 5.1 (diffusion limit under condition C-2)** *Under the assumptions of Theorem 3.1 with condition C-2, we have the joint convergence*

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

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

$$\hat{X}_{\Sigma}(t) = \hat{X}_{\Sigma}(0) - \sum_{j \in \mathcal{J}} \mu_j \gamma_j t + \sum_{j \in \mathcal{J}} \mu_j \int_0^t \tilde{v}_j([\hat{X}_{\Sigma}(s)]^-) ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t \tilde{p}_i([\hat{X}_{\Sigma}(s)]^+) ds + \sqrt{2}B(t), \quad (128)$$

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

$$\hat{X}_{\Sigma}^{\lambda}(t) \Rightarrow \hat{X}_{\Sigma}(t), \text{ in } D, \text{ as } \lambda \rightarrow \infty$$

**Remark 5.1 (when C-2 fails to hold)** The result of Theorem 5.1 is rather strong. While the state of the PSS is characterized by the multi-dimensional process  $(\hat{Q}_i^{\lambda}(t), \hat{Z}_{i,j}^{\lambda}(t); i \in \mathcal{I}, j \in \mathcal{J})$  for each  $\lambda$ , asymptotically as  $\lambda \rightarrow \infty$  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 C-2 is really necessary to obtain this result. Specifically, observe that under condition C-2, 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, in order to know the departure rate from the system. Once this observation is made, Theorem 5.1 follows from state-space collapse, as the latter allows us to control the proportions of idle agents. In the absence of condition C-2, we need more detailed information in order 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 state-space collapse. ■

**Remark 5.2 (equivalence with the single-class model)** Note that whenever  $\mu_j \equiv \mu$  and  $\theta_j \equiv \theta$ , 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 [17]. If, in addition  $\theta = 0$ , then the same replacement of scaling leads to the limit process given in Theorem 2 of [21] 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  $\lambda$ , service rate  $\mu$  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  $\lambda^{\text{th}}$   $M/M/N$

queue and let

$$Y^\lambda(t) = \frac{X^\lambda(t) - N^\lambda}{\sqrt{\lambda}}.$$

Then, Theorem 2 of [21] 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

$$Y(t) = Y(0) - \mu \sum_{j \in \mathcal{J}} \gamma_j t + \mu \int_0^t [Y(s)]^- ds + \sqrt{2}B(t), \quad (129)$$

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 5.3 (equivalence with the inverted-V model)** Note that 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, same service rates  $\{\mu_j, j \in \mathcal{J}\}$  and same staffing levels  $\{N_j^\lambda, j \in \mathcal{J}\}$ , but with a single customer class having arrival rate  $\lambda$ . 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 [19] and [20]. ■

**Proof of Theorem 5.1:** The limit for  $\hat{X}_\Sigma^\lambda$  is obtained through an application of the continuous mapping theorem, e.g., Theorem 3.4.1 of [36]. In particular, by state-space collapse, we have

$$\|\hat{X}_\Sigma^\lambda - \hat{Y}_\Sigma^\lambda\| \Rightarrow 0 \quad \text{in } D_-, \text{ as } \lambda \rightarrow \infty, \quad (130)$$

where

$$\hat{Y}_\Sigma^\lambda(t) = \hat{X}_\Sigma^\lambda(0) - \sum_{j \in \mathcal{J}} \mu_j \gamma_j t + \sum_{j \in \mathcal{J}} \mu_j \int_0^t \tilde{v}_j([\hat{X}_\Sigma^\lambda(s)]^-) ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t \tilde{p}_i([\hat{X}_\Sigma^\lambda(s)]^+) ds + \hat{M}_\Sigma^\lambda(t), \quad (131)$$

and  $\hat{X}_\Sigma^\lambda$  has the representation given in (52). Applying Corollary 4.6, we have that both  $\hat{X}_\Sigma^\lambda(t)$  and  $\hat{Y}_\Sigma^\lambda(t)$  are C-Tight. Consequently, the convergence in (130) is extended to  $D$ , and we can write

$$\hat{X}_\Sigma^\lambda(t) = \hat{X}_\Sigma^\lambda(0) - \sum_{j \in \mathcal{J}} \mu_j \gamma_j t + \sum_{j \in \mathcal{J}} \mu_j \int_0^t \tilde{v}_j([\hat{X}_\Sigma^\lambda(s)]^-) ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t \tilde{p}_i([\hat{X}_\Sigma^\lambda(s)]^+) ds + \hat{M}_\Sigma^\lambda(t) + o(1). \quad (132)$$

By Theorem 4.1 in [26], this integral representation for  $\hat{X}_\Sigma^\lambda(t)$  is a measurable continuous mapping from  $D$  to itself. In particular, we can get the convergence of  $\hat{X}_\Sigma^\lambda(t)$  from the convergence of  $\hat{M}_\Sigma^\lambda(t)$  that is established in the following lemma.

**Lemma 5.1** *Under the conditions of Theorem 5.1,*

$$\hat{M}_\Sigma^\lambda(t) \Rightarrow \sqrt{2}B(t), \text{ in } D, \text{ as } \lambda \rightarrow \infty, \quad (133)$$

where  $\{B(t), t \geq 0\}$  is a standard Brownian motion.

The proof of Lemma 5.1 is postponed to the end of this section. We can now apply the continuous-mapping theorem to (132) and use Lemma 5.1 to get the convergence of  $\hat{X}_\Sigma^\lambda(t)$ . By applying the continuous-mapping theorem again, we can then extend the limit to the vector process

$$\left( \hat{X}_\Sigma^\lambda(t), \tilde{p}_1([\hat{X}_\Sigma^\lambda(t)]^+), \dots, \tilde{p}_I([\hat{X}_\Sigma^\lambda(t)]^+), \frac{1}{a_1} \tilde{p}_1([\hat{X}_\Sigma^\lambda(t)]^+), \dots, \right. \\ \left. \frac{1}{a_I} \tilde{p}_I([\hat{X}_\Sigma^\lambda(t)]^+), \tilde{v}_1([\hat{X}_\Sigma^\lambda(t)]^-), \dots, \tilde{v}_J([\hat{X}_\Sigma^\lambda(t)]^-) \right) \text{ in } D^{2I+J+1}. \quad (134)$$

Then we can apply the convergence-together theorem again to show that this process has the same limit as the process  $(\hat{X}_\Sigma^\lambda(t), \hat{Q}_1^\lambda(t), \dots, \hat{Q}_I^\lambda(t), \hat{W}_1^\lambda(t), \dots, \hat{W}_I^\lambda(t), \hat{I}_1^\lambda(t), \dots, \hat{I}_J^\lambda(t))$  on  $D_-^{2I+J+1}$ . First, the state-space collapse establishes the connection for the queue-length processes. Then we can apply Puhalskii's [28] first-passage-time argument to extend the result from the queue lengths to treat the waiting times as well; see the Corollary in [28], §13.7 of [36] (especially Theorem 13.7.4) and Corollary B.3 in the appendix of [18]. The fact that equation (128) has a unique (possibly weak) solution follows from known results for one-dimensional SDE's; see e.g. Remark 5.5.19 and Exercise 5.5.38 in [22].  $\blacksquare$

The diffusion limits given in the Theorem 5.1 characterize the asymptotic behavior on bounded time periods. The next natural step is to try and 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 [21], [17], [18], and also in the more complicated case of Tezcan [31]. Establishing such an interchange is, however, extremely hard in general. Gamarnik and Zeevi [15] and Budhiraja and Lee [10] 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 [31]. 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 5.2 below.

**Corollary 5.2 (steady-state limits)** *Assume that steady state exists for each  $\lambda$  large enough and that the sequence  $(\hat{Q}_\Sigma^\lambda(\infty), \hat{I}_\Sigma^\lambda(\infty))$  is tight. Then we have that, as  $\lambda \rightarrow \infty$ ,*

$$\hat{X}_\Sigma^\lambda(\infty) \Rightarrow \hat{X}_\Sigma(\infty). \quad (135)$$

Moreover,

$$\hat{I}_j^\lambda(\infty) \Rightarrow \tilde{v}_j([\hat{X}_\Sigma(\infty)]^-), \quad j \in \mathcal{J}, \quad (136)$$

$$\hat{Q}_i^\lambda(\infty) \Rightarrow \tilde{p}_i([\hat{X}_\Sigma(\infty)]^+), \quad \text{and} \quad \hat{W}_i^\lambda(\infty) \Rightarrow \frac{1}{a_i} \tilde{p}_i([\hat{X}_\Sigma(\infty)]^+), \quad i \in \mathcal{I}. \quad (137)$$

If, in addition, the sequence  $\hat{Q}_\Sigma^\lambda(\infty)$  is uniformly integrable then the convergence in (137) holds also in expectation.

**Proof:** The proof follows an interchange-of-limits argument, following [21]. It is easy to check (using for example Browne and Whitt [9]) that, with the conditions of the corollary, the diffusion process  $\hat{X}_\Sigma(t)$ , as given in Theorem 5.1, has a unique stationary distribution coinciding with the distribution of  $\hat{X}_\Sigma(\infty)$ . Now, by Prohorov's Theorem the tight sequence  $\hat{X}_\Sigma^\lambda(\infty)$  has a convergent subsequence  $\hat{X}_\Sigma^{\lambda^k}(\infty)$ . Let  $\hat{X}_\Sigma^{\lambda^k}(0)$  be distributed as  $\hat{X}_\Sigma^{\lambda^k}(\infty)$ . Then  $\hat{X}_\Sigma^{\lambda^k}(t)$  is a strictly stationary process and since we already proved that  $\hat{X}_\Sigma^{\lambda^k}(t) \Rightarrow \hat{X}_\Sigma(t)$ ,  $\hat{X}_\Sigma(t)$  will be a process with  $\hat{X}_\Sigma(0)$  having the distribution of the limit of  $\hat{X}_\Sigma^{\lambda^k}(\infty)$ . However, since  $\hat{X}_\Sigma^{\lambda^k}(t)$  is stationary for each  $k$ , so is  $\hat{X}_\Sigma(t)$ . Hence, the limit of  $\hat{X}_\Sigma^{\lambda^k}(\infty)$  must be the unique stationary distribution of  $\hat{X}_\Sigma(t)$ . The same argument applies to any convergent subsequence and hence the sequence  $\hat{X}_\Sigma^\lambda(\infty)$  itself must converge to this limit (see Theorem 2.3 in [6]). The convergence of the moments now follows from uniform integrability. Since  $\hat{Q}_i^\lambda(\infty) \leq \hat{Q}_\Sigma^\lambda(\infty)$  almost surely, the sequence  $\hat{Q}_i^\lambda(\infty)$  is also uniformly integrable for all  $i \in \mathcal{I}$ . Hence,

$$E[\hat{Q}_i^\lambda(\infty)] \rightarrow E\left[\tilde{p}_i([\hat{X}_\Sigma(\infty)]^+)\right].$$

Recalling that  $\lambda_i = a_i \lambda$ , and using Little's law, we obtain

$$E \left[ \hat{W}_i^\lambda(\infty) \right] \Rightarrow \frac{1}{a_i} E \left[ \tilde{p}_i([\hat{X}_\Sigma(\infty)]^+) \right].$$

■

The following is a diffusion-limit result under condition 3. We refer the reader to §4.4 for the definition of  $\beta_i$  and the construction of the processes  $\hat{X}_i^\lambda(t)$ .

**Theorem 5.3 (diffusion limit under condition C-3)** *Under the assumptions of Theorem 3.1 with condition C-3, we have the joint convergence*

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

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

$$\hat{X}_i(t) = \hat{X}_i(0) - \beta_i t - \mu_i \int_0^t \hat{X}_i(s) ds - (\theta_i - \mu_i) \int_0^t \tilde{p}_i([\hat{X}_\Sigma(s)]^+) ds + \sqrt{2a_i} B_i(t), \quad (138)$$

with  $\{B_i(t), t \geq 0\}$  for  $i \in \mathcal{I}$  being standard independent Brownian motions.

**Proof:** The proof is very similar to that of Theorem 5.1 and we only outline the differences. First, similarly to Lemma 5.1 one proves that

$$(\hat{M}_1^\lambda(t), \dots, \hat{M}_I^\lambda(t)) \Rightarrow \sqrt{2a} B(t), \text{ in } D^I, \text{ as } \lambda \rightarrow \infty,$$

where  $\sqrt{2a} B(t) \equiv (\sqrt{2a_1} B_1(t), \dots, \sqrt{2a_I} B_I(t))$  and  $B_i(t)$ ,  $i \in \mathcal{I}$  are independent Brownian motions. Using equation (119) and replacing  $\hat{Z}_i^\lambda(t) = \hat{X}_i^\lambda(t) - \hat{Q}_i^\lambda(t)$ , one then applies the state-space collapse from Theorem 3.1 and the Continuous Mapping Theorem, as in the proof of Theorem, 5.1 to get the required convergence. ■

**Remark 5.4 (equivalence with the V model)** Note that the limit we obtain in Theorem 5.3 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 agents pools but a single customer class; see Remark 5.3. ■

## 5.2 Convergence of performance measures

Define the empirical averages

$$\bar{W}^{\lambda, T} = \frac{\sum_{i=1}^I \sum_{k=1}^{A_i(T)} w_{i,k}^\lambda}{A(T)} \text{ and } \bar{F}_i^{\lambda, T}(y) = \frac{\sum_{k=1}^{A_i^\lambda(T)} \mathbf{1}\{w_{i,k}^\lambda > y\}}{A_i^\lambda(T)}, \quad (139)$$

where  $w_{i,k}^\lambda$  is the realized waiting time of the  $k^{\text{th}}$  class- $i$  customer to arrive to the system after time 0 and  $\mathbf{1}B$  is the indicator of the event  $B$ , which is equal to 1 if  $B$  occurs and 0 otherwise. The following Proposition identifies limits for these averages.

**Proposition 5.1** *Under the conditions of Theorem 3.1.*

$$\frac{1}{T} \int_0^T \hat{Q}_\Sigma^\lambda(t) dt \Rightarrow \frac{1}{T} \int_0^T [\hat{X}_\Sigma(t)]^+ dt \quad \text{in } \mathbb{R} \quad \text{as } \lambda \rightarrow \infty. \quad (140)$$

Also, for every  $y \geq 0$  and  $T > 0$ ,

$$\bar{F}_i^{\lambda, T}(y/\sqrt{\lambda}) \Rightarrow \frac{1}{T} \int_0^T \mathbf{1} \left\{ \frac{1}{a_i} \tilde{p}_i([\hat{X}_\Sigma(t)]^+) > y \right\} dt \quad \text{in } [0, 1] \quad \text{as } \lambda \rightarrow \infty. \quad (141)$$

Finally,

$$\sqrt{\lambda} \bar{W}^{\lambda, T} \Rightarrow \frac{1}{T} \int_0^T [\hat{X}_\Sigma(t)]^+ dt \quad \text{in } \mathbb{R} \quad \text{as } \lambda \rightarrow \infty, \quad (142)$$

where  $\hat{X}_\Sigma(t)$  is the limit process from Theorem 5.1.

**Proof:** First we note that under any of the conditions C-1-C-3 the sequence  $\hat{X}_\Sigma^\lambda(t)$  is C-Tight and converges to a limit with almost surely continuous sample path. Indeed, for conditions C-2 and C-3 this follows from Theorems 5.1 and Theorem 5.3 respectively. The convergence under condition C-1 follows from the equivalence of QIR and GQIR as proved in §4.2 and the convergence results in Proposition 1 of Atar [5]. Hence, without loss of generality we let  $\hat{X}_\Sigma(t)$  be the limit of  $\hat{X}_\Sigma^\lambda(t)$ .

Now, equation (140) follows immediately from Theorem 5.1 and the continuity of the integration mapping  $x \mapsto \int_0^T x(t)dt/T$ . The argument for equation (141) has two parts. In the first part we prove a result for the virtual waiting process. In the second part we extend this result to the waiting-time time averages by showing that both are asymptotically equivalent.

First, applying Puhalskii's corollary [28] we have that

$$\left(\hat{W}_1^\lambda(t), \dots, \hat{W}_I^\lambda(t)\right) \Rightarrow \left(\tilde{p}_1([\hat{X}_\Sigma(t)]^+)/a_1, \dots, \tilde{p}_1([\hat{X}_\Sigma(t)]^+)/a_1\right), \text{ in } D_-, \text{ as } \lambda \rightarrow \infty$$

The first part of the argument then follows the proof of Corollary 3 in [21], applying the continuous mapping

$$\psi(x, a, T) = \frac{1}{T} \int_0^T \mathbf{1}_{\{x(t) \geq a\}} dt,$$

to  $(-\hat{W}_i^\lambda(t), -y/\sqrt{\lambda}, T)$ . Following the argument in Corollary 3 of [21] we observe that the mapping is indeed continuous because the limit  $\hat{W}_i(t) = \tilde{p}_i([\hat{X}_\Sigma(t)]^+)/a_i$ , has a density for each  $t$ . The existence of a density for the process  $\hat{X}_\Sigma(t)$  is guaranteed by the continuity of the drift and the constant diffusion coefficients - see for example pages 368-369 in [22]. Hence, we have

$$\begin{aligned} \frac{1}{T} \int_0^T \mathbf{1} \left\{ W_i^\lambda(t) \leq y/\sqrt{\lambda} \right\} dt &= \frac{1}{T} \int_0^T \mathbf{1} \left\{ -W_i^\lambda(t) \geq -y/\sqrt{\lambda} \right\} dt \\ &\Rightarrow \frac{1}{T} \int_0^T \mathbf{1} \left\{ -\frac{1}{a_i} \tilde{p}_i([\hat{X}_\Sigma(t)]^+) \geq -y \right\} dt, \end{aligned} \quad (143)$$

where the convergence is in  $[0, 1]$  as  $\lambda \rightarrow \infty$ . Some care is needed above as the convergence of  $\hat{W}_i^\lambda(t)$  to  $\tilde{p}_i([\hat{X}_\Sigma(t)]^+)$  holds only on  $D_-$ . But this is easy to overcome by considering first the integral on  $[\delta, T]$  for  $\delta > 0$  and then taking  $\delta$  to be small. As a consequence,

$$\begin{aligned} \frac{1}{T} \int_0^T \mathbf{1} \left\{ W_i^\lambda(t) > y/\sqrt{\lambda} \right\} dt &= 1 - \frac{1}{T} \int_0^T \mathbf{1} \left\{ W_i^\lambda(t) \leq y/\sqrt{\lambda} \right\} dt \\ &\Rightarrow \frac{1}{T} \int_0^T \mathbf{1} \left\{ \frac{1}{a_i} \tilde{p}_i([\hat{X}_\Sigma(t)]^+) > y \right\} dt. \end{aligned} \quad (144)$$

For the second part of the argument, note that

$$\bar{F}_i^{\lambda, T}(y/\sqrt{\lambda}) = \frac{1}{A_i^\lambda(T)} \int_0^T \mathbf{1} \{ W_i^\lambda(t-) > y/\sqrt{\lambda} \} dA_i^\lambda(t).$$

We claim that

$$\frac{1}{A_i^\lambda(T)} \int_0^T \mathbf{1}\{W_i^\lambda(t-) > y/\sqrt{\lambda}\} dA_i^\lambda(t) - \frac{1}{T} \int_0^T \mathbf{1}\{W_i^\lambda(t) > y/\sqrt{\lambda}\} dt \Rightarrow 0 \text{ in } \mathbb{R} \text{ as } \lambda \rightarrow \infty. \quad (145)$$

Indeed, following the argument given after (27), we have that

$$\tilde{M}_i^\lambda(t) := \int_0^t \mathbf{1}\{W_i^\lambda(t-) > y/\sqrt{\lambda}\} d(A_i^\lambda(t) - \lambda_i t),$$

is a square-integrable martingale, with predictable quadratic variation that is bounded by  $\lambda_i t$ . By Lemma 5.8 of [26],  $\tilde{M}_i^\lambda(t)/\sqrt{\lambda_i}$  is a stochastically bounded sequence. In particular, by Lemma 5.9 in [26],

$$\frac{\tilde{M}_i^\lambda(t)}{\lambda_i} \Rightarrow 0 \quad \text{in } D \quad \text{as } \lambda \rightarrow \infty. \quad (146)$$

Together with the strong law of large numbers for renewal processes, we then have

$$\bar{F}_i^{\lambda,T}(y/\sqrt{\lambda}) - \frac{1}{T} \int_0^T \mathbf{1}\{W_i^\lambda(t) > y/\sqrt{\lambda}\} dt = \frac{\lambda_i}{A_i^\lambda(t)} \frac{\tilde{M}_i^\lambda(T)}{\lambda_i T} \Rightarrow 0 \text{ in } \mathbb{R} \text{ as } \lambda \rightarrow \infty.$$

Equation (141) now follows from (144) and (145) and the convergence together theorem (see for example Theorem 11.4.7 of [36]).

For equation (142), a very similar argument is used, but some extra care is required as the integrand will now be  $\hat{W}_i^\lambda(s)$  itself rather than an indicator function, and this process is not bounded. However, by Theorem 5.1,  $\hat{W}_i^\lambda(t)$  is a convergent sequence and it is consequently stochastically bounded. This will suffice for the proof. Specifically, replace the integrand  $\hat{W}_i^\lambda(t)$  with  $\hat{W}_i^\lambda(t) \wedge K$  for some  $K > 0$ . For this bounded integrand we may repeat step by step the argument that we used of  $\bar{F}_i^{\lambda,T}(\cdot)$ , to have that

$$\frac{1}{A_i^\lambda(T)} \int_0^T (\hat{W}_i^\lambda(t-) \wedge K) dA_i^\lambda(t) - \frac{1}{T} \int_0^T (\hat{W}_i^\lambda(t-) \wedge K) dt \Rightarrow 0 \text{ in } \mathbb{R} \text{ as } \lambda \rightarrow \infty. \quad (147)$$

Since  $\hat{W}_i^\lambda(t)$  is stochastically bounded for every  $\epsilon > 0$  we can choose  $K$  large enough so that  $P\{\|\hat{W}_i^\lambda\|_T^* > K\} \leq \epsilon$ . It then follows that

$$\frac{1}{A_i^\lambda(T)} \int_0^T \hat{W}_i^\lambda(t-) dA_i^\lambda(t) - \frac{1}{T} \int_0^T \hat{W}_i^\lambda(t-) dt \Rightarrow 0 \text{ in } \mathbb{R} \text{ as } \lambda \rightarrow \infty. \quad (148)$$

We then apply the convergence together theorem and use the convergence of  $\hat{W}_i^\lambda(t)$  as given in Theorem 5.1. Finally, the argument is concluded by using the fact that

$$\sqrt{\lambda}\bar{W}^{\lambda,T} = \sum_{i \in \mathcal{I}} \frac{1}{A_i^\lambda(T)} \int_0^T \hat{W}_i^\lambda(t-) dA_i^\lambda(t).$$

■

The following corollary might be of interest for applications. It is indeed used in our subsequence paper [20]. The corollary allow one to conclude convergence of integrals starting from 0 despite the fact that state-space collapse holds only in  $D_-$ .

**Proposition 5.2** Fix  $T > 0$ . Let  $f_i : \mathbb{R} \mapsto \mathbb{R}$  be continuous functions. Then, under the conditions of Theorem 3.1

$$\left( \int_0^T f_1(\hat{Q}_i^\lambda(t)) dt, \dots, \int_0^T f_I(\hat{Q}_i^\lambda(t)) dt \right) \Rightarrow \left( \int_0^T f_1(\tilde{p}_1([\hat{X}_\Sigma(t)]^+)) dt, \dots, \int_0^T f_I(\tilde{p}_I([\hat{X}_\Sigma(t)]^+)) dt \right), \text{ in } \mathbb{R} \text{ as } \lambda \rightarrow \infty, \quad (149)$$

where  $\hat{X}_\Sigma(t)$  is the diffusion process from Theorem 5.1.

**Proof:** Fix  $\delta, \epsilon > 0$ . Under any of the conditions of Theorem 3.1 we have that the sequence  $\hat{Q}_\Sigma^\lambda(t)$  is stochastically bounded. For conditions C-2 and C-3 this follows from Corollary 4.6 and its analogue in §4.4. For condition C-1 this follows from Proposition 1 in [5] and the equivalence of QIR and GQIR under this condition; see §4.2. Hence, we may choose  $K$ , such that

$$\limsup_{\lambda \rightarrow \infty} P\{\|\hat{Q}_\Sigma^\lambda\|_T^* > K\} \leq \frac{\epsilon}{2}. \quad (150)$$

By the continuity of the functions  $f_i(\cdot)$ , there exists  $\tilde{\delta} > 0$  small enough so that

$$\sum_{i \in \mathcal{I}} \left( \sup_{x, y \leq K, |x-y| \leq \tilde{\delta}} |f_i(x) - f_i(y)| \right) \leq \frac{\delta}{2}. \quad (151)$$

As in the proof of Theorem 4.3, given  $\delta > 0$ , we have  $s^*$  such that

$$P \left\{ \sup_{s^*/\sqrt{\lambda} \leq t \leq T} \sum_{i \in \mathcal{I}} |\hat{Q}_i^\lambda(t) - \tilde{p}_i([\hat{X}_\Sigma^\lambda(t)]^+)| > \tilde{\delta} \right\} \rightarrow 0, \text{ as } \lambda \rightarrow \infty. \quad (152)$$

We now write

$$\int_0^T f_i(\hat{Q}_i^\lambda(t)) dt = \int_0^{s^*/\sqrt{\lambda}} f_i(\hat{Q}_i^\lambda(t)) dt + \int_{s^*/\sqrt{\lambda}}^T f_i(\hat{Q}_i^\lambda(t)) dt.$$

The stochastic boundedness of  $\hat{Q}_i^\lambda$  follows from that of  $\hat{Q}_\Sigma^\lambda$  and together with the continuity of  $f_i(\cdot)$ , we then have that

$$\sum_{i \in \mathcal{I}} \left| \int_0^{s^*/\sqrt{\lambda}} f_i(\hat{Q}_i^\lambda(t)) dt \right| \xrightarrow{P} 0, \text{ as } \lambda \rightarrow \infty. \quad (153)$$

As  $[\hat{X}_\Sigma^\lambda(t)] \leq \hat{Q}_\Sigma^\lambda(t)$ , the stochastic boundedness of  $\hat{Q}_\Sigma^\lambda$  implies that of  $\hat{X}_\Sigma^\lambda(t)$  and, consequently, we have that

$$\int_0^{s^*/\sqrt{\lambda}} \left| f_i(\tilde{p}_i([\hat{X}_\Sigma^\lambda(t)]^+)) \right| dt \xrightarrow{P} 0, \text{ as } \lambda \rightarrow \infty. \quad (154)$$

Now, define the event

$$\mathcal{O}^\lambda(\delta) := \left\{ \omega \in \Omega : \sum_{i \in \mathcal{I}} \left| \int_{s^*/\sqrt{\lambda}}^T f_i(\hat{Q}_i^\lambda(t)) dt - \int_{s^*/\sqrt{\lambda}}^T f_i(\tilde{p}_i([\hat{X}_\Sigma^\lambda(t)]^+)) dt \right| > \frac{\delta}{2} \right\}.$$

Then,

$$P \left\{ \mathcal{O}^\lambda(\delta) \right\} \leq P \left\{ \mathcal{O}^\lambda(\delta); \|\hat{Q}_\Sigma^\lambda\|_T^* \leq K \right\} + P \left\{ \mathcal{O}^\lambda(\delta); \|\hat{Q}_\Sigma^\lambda\|_T^* > K \right\}.$$

By equations (150), (151) and (152) we then have that

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \mathcal{O}^\lambda(\delta) \right\} \leq \frac{\epsilon}{2},$$

and, together with (153) and (154), that

$$\limsup_{\lambda \rightarrow \infty} P \left\{ \sum_{i \in \mathcal{I}} \left| \int_0^T f_i(\hat{Q}_i^\lambda(t)) dt - \int_0^T f_i(\tilde{p}_i([\hat{X}_\Sigma^\lambda(t)]^+)) dt \right| > \delta \right\} \leq \epsilon. \quad (155)$$

Using Theorem 5.1, the continuity of the integral (see Theorem 11.5.1 in [36]) and the Continuous Mapping Theorem we have the convergence

$$\int_0^T f_i(\tilde{p}_i([\hat{X}_\Sigma^\lambda(t)]^+)) dt \Rightarrow \int_0^T f_i(\tilde{p}_i([\hat{X}_\Sigma(t)]^+)) dt, \quad (156)$$

where  $\hat{X}_\Sigma(t)$  is the diffusion process from Theorem 5.1. Finally, the result of the corollary is obtained by using (155), (156) and applying the Convergence Together Theorem. ■

We conclude this section with the proof of Lemma 5.1.

**Proof of Lemma 5.1.** Let  $\hat{M}_A^\lambda(t) = (\hat{M}_{A_1}^\lambda(t), \dots, \hat{M}_{A_I}^\lambda(t))$ ,  $\hat{M}_S^\lambda(t) = (\hat{M}_1^\lambda(t), \dots, \hat{M}_J^\lambda(t))$ , and  $\hat{M}_R^\lambda(t) = (\hat{M}_{R_1}^\lambda(t), \dots, \hat{M}_{R_I}^\lambda(t))$ . Then, we prove that

$$\left( \hat{M}_A^\lambda(t), \hat{M}_S^\lambda(t), \hat{M}_R^\lambda(t) \right) \Rightarrow \left( \sqrt{a}B_A(t), \sqrt{\mu\bar{\nu}}B_S^\lambda(t), 0 \right), \text{ in } D^{2I+J}, \text{ as } \lambda \rightarrow \infty. \quad (157)$$

Here  $\hat{B}_A(t)$ ,  $\hat{B}_S(t)$  are, respectively,  $I$  and  $J$  independent Brownian motions and  $a = (a_1, \dots, a_I)$ ,  $\mu\bar{\nu} = (\mu_1\bar{\nu}_1, \dots, \mu_J\bar{\nu}_J)$ ; the square-root and the vector product should be interpreted componentwise. Also, 0 in (157) is the 0 vector in  $\mathbb{R}^I$ . Recalling that

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

the result of the lemma then follows from the continuity of the addition operator under continuous limits (see Theorem 12.7.1 in [36]).

We turn, then, to the proof of (157). The proof is based on a functional central limit theorem for Poisson processes and on a random-time change argument. Specifically, recall that

$$\hat{M}_{A_i}^\lambda(t) = \frac{A_i(\lambda t) - \lambda t}{\sqrt{\lambda}}, \quad \hat{M}_j^\lambda(t) = \frac{S_j \left( \mu_j \int_0^t Z_j^\lambda(s) ds \right) - \mu_j \int_0^t Z_j^\lambda(s) ds}{\sqrt{\lambda}}, \quad i \in \mathcal{I}, j \in \mathcal{J},$$

and

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

Define

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

and

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

and let  $\tilde{M}_A^\lambda(t)$ ,  $\tilde{M}_S^\lambda(t)$  and  $\tilde{M}_R^\lambda(t)$  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

$$\left( \tilde{M}_A^\lambda(t), \tilde{M}_S^\lambda(t), \tilde{M}_R^\lambda(t) \right) \Rightarrow \left( B_A(t), B_S(t), B_R(t) \right), \text{ in } D^{2I+J}, \text{ as } \lambda \rightarrow \infty,$$

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 [26]).

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

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

and  $\Psi_{A_i}^\lambda(t) := \lambda_i / \lambda t = a_i t$ ,  $i \in \mathcal{I}$ . Let  $\Psi_A^\lambda$ ,  $\Psi_S^\lambda$  and  $\Psi_R^\lambda$  be the corresponding vector valued processes. By Corollary 4.6, we have that  $\hat{I}_j^\lambda(t)$  and  $\hat{Q}_i^\lambda(t)$  are stochastically bounded process for all  $i \in \mathcal{I}$  and  $j \in \mathcal{J}$ .

This implies that

$$\left( \frac{I_1^\lambda(t)}{\lambda}, \dots, \frac{I_j^\lambda(t)}{\lambda}, \frac{Q_1^\lambda(t)}{\lambda}, \dots, \frac{Q_I^\lambda(t)}{\lambda} \right) \Rightarrow \eta, \quad \text{in } D^{I+J}, \quad \text{as } \lambda \rightarrow \infty, \quad (158)$$

where  $\eta(t) \equiv (0, 0, \dots, 0)$  (see for example Lemma 5.10 in [26]). In particular, since  $Z_j^\lambda(t) = N_j^\lambda - I_j^\lambda(t)$ , we have that

$$\left( \frac{Z_1^\lambda(t)}{\lambda}, \dots, \frac{Z_j^\lambda(t)}{\lambda}, \frac{Q_1^\lambda(t)}{\lambda}, \dots, \frac{Q_I^\lambda(t)}{\lambda} \right) \Rightarrow \eta', \quad \text{in } D^{I+J}, \quad \text{as } \lambda \rightarrow \infty, \quad (159)$$

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

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

to show that

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

Finally, applying the random-time-change theorem (see Theorem 13.2.1 in [36]), we conclude that

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

Since, by definition,

$$\left( \hat{M}_A^\lambda(t), \hat{M}_S^\lambda(t), \hat{M}_R^\lambda(t) \right) = \left( \tilde{M}_A^\lambda \left( \Phi_A^\lambda(t) \right), \tilde{M}_S^\lambda \left( \Phi_S^\lambda(t) \right), \tilde{M}_R^\lambda \left( \Phi_A^\lambda(t) \right) \right),$$

the proof is complete. ■

**Acknowledgments:** The authors are grateful to Avi Mandelbaum and Mor Armony for fruitful discussions, and to Zohar Feldman and Ohad Perry for contributions to the simulation, including the use of their codes. The second author was partially 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. On customer contact centers with a call-back option: customer decisions, routing rules and system design. *Operations Research* **52**(2) 271-292.
- [3] Armony M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527-545.
- [4] Asmussen, S. 2003. *Applied Probability and Queues*, second ed., Springer, New York.
- [5] Atar R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15**(4) 2606-2650.
- [6] Billingsley P. 1968. *Convergence of Probability Measures*. J. Wiley & Sons, New York.
- [7] Borst S., A. Mandelbaum A., M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17-34.
- [8] Bramson M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* **30**(1-2) 89-148.
- [9] Browne S., W. Whitt. 1995. Piecewise-Linear Diffusion Processes. J. Dshalalow, ed. *Advances in Queueing*. CRC Press, Boca Raton, FL, 463-480.
- [10] Budhiraja, A., C. Lee. 2007. Stationary distribution convergence for generalized Jackson networks in heavy traffic. Working paper, The University of North Carolina at Chapel Hill.
- [11] Dai J.G. 1995. On positive Harris Recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* **5** 49-77.
- [12] Dai J.G., T. Tezcan. 2005. State space collapse in many server diffusion limits of parallel server systems. Working Paper, Georgia Institute of Technology, Atlanta, GA.
- [13] Dai J.G, T. Tezcan. 2006. Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Working Paper. Georgia Institute of Technology, Atlanta, GA.
- [14] Dai J.G, T. Tezcan. 2007. Optimal control of parallel server systems with many servers in heavy traffic. Working Paper. Georgia Institute of Technology, Atlanta, GA.
- [15] Gamarnik D., A. Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Annals of Applied Probability* **16** 56-90.
- [16] Gans, N., G. Koole, G., A. Mandelbaum. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5**(2), 79–141.

- [17] Garnett O., A. Mandelbaum, M. Reiman. 2003. Designing a Call Center with Impatient Customers. *Manufacturing Service Oper. Management*, **4**(3), 208-227.
- [18] Gurvich I, M. Armony and A. Mandelbaum. 2006. Service-level differentiation in call Centers with fully flexible servers. *Management Science* forthcoming.
- [19] Gurvich I., W. Whitt., Service-level differentiation in many-server service systems: a solution based on fixed-queue-ratio routing. Working paper, Columbia University, New York, NY.
- [20] Gurvich I., W. Whitt W. 2006. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* forthcoming.
- [21] Halfin S., W. Whitt. 1981. Heavy-traffic Limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567-587.
- [22] Karatzas I., S.E. Shreve. 1991. *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer-Verlag, New York.
- [23] Khas'minskii R.Z. 1960. Ergodic properties of recurrent diffusion processes and stabilization of the solution to the cauchy problem for parabolic equations. *Theory of Probability and its Applications* **5**(2) 179-196.
- [24] Lipster R. Sh., A. N. Shirayev. 1989. *Theory of Martingales*. Kluwer Acad. Publ., Boston.
- [25] Mandelbaum, A., A. Stolyar. 2004. Scheduling flexible servers with convex delay Costs: heavy-traffic optimality of the generalized  $c\mu$ -Rule. *Oper. Res.* **52**(6) 836 - 855.
- [26] Pang, G., R. Talreja, W. Whitt. 2006. Martingale proofs of many-server heavy-traffic limits for Markovian queues. Working Paper. Columbia University, New York.
- [27] Protter P. 1992. *Stochastic integration and differential equations - A new approach*. Springer-Verlag, New York.
- [28] Puhalskii A. 1994. On the invariance principle for the first passage time, *Math. Oper. Res.* **19**(4) 946 - 954.
- [29] Rebolledo R. 1980. Central limit theorems for local martingales. *Probability Theory and Related Fields* **51**(3) 269-286.
- [30] Royden, H. L. 1968. *Real Analysis*, second edition, Macmillan, London.
- [31] Tezcan T. 2006. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Oper. Res.* forthcoming.
- [32] Van der Vaart A.W. 2006. Martingales, diffusions and financial mathematics - Lecture Notes. Available at: <http://www.math.vu.nl/sto/onderwijs/mdfm/>.
- [33] Van Mieghem J.A. 1995. Dynamic scheduling with convex delay costs: the generalized  $c\mu$  rule. *Ann. Appl. Probab.* **5**(3) 809-833.
- [34] Williams R.J. 1998. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse, *Queueing Systems*, **30**(1-2), 27-88.
- [35] Whitt W. 1991. A Review of  $L = \lambda W$  and extensions. *Queueing Systems* **9**(3) 235-268.

- [36] Whitt W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York.