MATHEMATICS OF OPERATIONS RESEARCH
Vol. 36, No. 2, May 2011, pp. 271–320
ISSN 0364-765X | EISSN 1526-5471 | 11 | 3602 | 0271

DOI 10.1287/moor.1110.0494

State Space Collapse in Many-Server Diffusion Limits of Parallel Server Systems

J. G. Dai
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, dai@gatech.edu

Tolga Tezcan
Simon Graduate School of Business, University of Rochester, Rochester, New York 14627, tolga.tezcan@simon.rochester.edu

We consider a class of queueing systems that consist of server pools in parallel and multiple customer classes. Customer service times are assumed to be exponentially distributed. We study the asymptotic behavior of these queueing systems in a heavy traffic regime that is known as the Halfin-Whitt many-server asymptotic regime. Our main contribution is a general framework for establishing state space collapse results in this regime for parallel server systems. In our work, state space collapse refers to a decrease in the dimension of the processes tracking the number of customers in each class waiting for service and the number of customers in each class being served by various server pools. We define and introduce a "state space collapse" function, which governs the exact details of the state space collapse. We show that a state space collapse result holds in many-server heavy traffic if a corresponding deterministic hydrodynamic model satisfies a similar state space collapse condition. Unlike the single-server heavy traffic setting for multiclass queueing network, our hydrodynamic model is different from the standard fluid model for many-server queues. Our methodology is similar in spirit to that in Bramson [Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.], which focuses on the single-server heavy traffic regime. We illustrate the applications of our results by establishing state space collapse results in many-server diffusion limits for V-model systems under static-buffer-priority policy and the threshold policy proposed in the literature.

Key words: parallel server systems; large scale systems; scheduling and routing control; heavy traffic
MSC2000 subject classification: Primary: 60K25, 68M20; secondary: 90B22, 60F17.
OR/MS subject classification: Primary: queues secondary: diffusion models, limit theorems
History: Received March 20, 2010; revised December 29, 2010, and February 23, 2011.

1. Introduction and literature review. Multiclass queueing networks have been extensively used to model queueing systems arising in manufacturing and service industries (Harrison [28]). A special class of these networks, parallel server systems, are of current interest. They are commonly used to model service systems with many servers; see Gans et al. [21], Maglaras and Zeevi [33, 34, 35], and Randhawa and Kumar [41] for different applications. In a parallel server system, customers are handled by a set of server pools and leave the system after service. We also restrict our attention to systems with exponential service times. Similar to multiclass queueing networks, exact analysis of parallel server systems is limited to a few special cases. Even when available, the results from exact analysis provide limited insight on the general properties of performances of these systems and rarely can be used for optimization purposes.

As an alternative, parallel server systems have been analyzed by using diffusion approximations under two different heavy traffic regimes; see Chen and Yao [12], Whitt [48], Gans et al. [21], among others. In this paper we focus on the *many-server heavy traffic* regime that is similar to that proposed in Halfin and Whitt [26]. Under the many-server heavy traffic regime, arrival rates and number of servers in each pool grow to infinity in such a way that the nominal load converges to one at a certain rate. As stated in §4 of Gans et al. [21], many-server asymptotic analysis is one of the most promising research directions in the analysis of parallel server systems with many servers, particularly for the analysis and control of call centers.

Central to any heavy traffic analysis, either for performance analysis or optimal control, is some heavy traffic limit theorem that states that a certain diffusion-scaled performance process converges to a diffusion process in heavy traffic. Often, the key to the proof of such a limit theorem is a state space collapse (SSC) result. In two companion papers, Bramson [10] and Williams [51], sufficient conditions are given under which a *conventional heavy traffic* limit theorem holds for a general class of multiclass queueing networks. Unlike the many-server heavy traffic regime, in the conventional heavy traffic regime, an increase in the arrival rate is matched by an increase in the service rate while keeping the number of servers in each server pool fixed. (Equivalently, the conventional heavy traffic can be achieved by employing diffusion-scaling in time and space while keeping the arrival and service rate fixed; see, for example, Williams [51].)

It is shown in Bramson [10] that to prove an SSC result in conventional heavy traffic limit it is enough to show that a similar SSC result holds for the *hydrodynamic model*. His framework has been used to show SSC results

in conventional heavy traffic limits for multiclass queueing networks and, more generally, stochastic processing networks; see Bramson and Dai [11], Mandelbaum and Stolyar [37], Dai and Lin [15]. These SSC results are then used to establish the heavy traffic diffusion limits of the systems under consideration.

It is apparent from the current literature that SSC results in many-server heavy traffic are also crucial for establishing diffusion limits; see Armony [1], Armony and Maglaras [2, 3], Tezcan and Dai [46], Gurvich et al. [25], and Tezcan [45]. However, results similar to those in Bramson [10] and Williams [51] have not been established in the many-server heavy traffic regime yet. Due to conspicuous differences between the conventional and many-server heavy traffic regimes (see Gans et al. [21, §4] for more details), the results of Williams and Bramson cannot be readily extended to the many-server asymptotic analysis.

In this paper we present a general framework for establishing SSC results in the many-server regime for parallel server systems. Specifically, we extend the framework in Bramson [10] to show that *multiplicative SSC* results in many-server diffusion limits can be established by verifying that the associated hydrodynamic model satisfies certain conditions. The hydrodynamic model is defined by a set of deterministic equations that are similar to, but different from, the standard set of fluid model equations; see the last four paragraphs of §4.1 for a detailed discussion on the differences of these two models. We illustrate our main result by establishing SSC results for V-parallel server systems under two different policies. Our results have also been instrumental in the analysis of distributed systems in Tezcan [45] and general parallel server systems in Dai and Tezcan [16].

Our approach to establishing the framework to prove SSC results in many-server heavy traffic is similar to that of Bramson [10]. We use the hydrodynamic scaling that is obtained by slowing down the time in the many-server diffusion scaling. Using this scaling, the events that happen instantaneously in the diffusion limits can be observed in more detail. By utilizing the connection between the hydrodynamic limits and the diffusion limits, we show that for a SSC result to hold for diffusion limits it must hold eventually for the *hydrodynamic limits*. The general structure and definition of hydrodynamic limits are complicated. It is not clear as to how one can check the required condition on hydrodynamic limits by using the definition directly. We overcome this hurdle by showing that the hydrodynamic limits of a general parallel server system must satisfy a set of deterministic equations that we call the hydrodynamic model equations. These equations possess some of the nice properties of the standard fluid model equations, but they are different. We illustrate how fluid model tools can be used to show the SSC results for the hydrodynamic limits in the V-systems in §7.

Our results differ from Bramson's in the following aspects. As described above, we focus on the many-server heavy traffic regime whereas Bramson focuses on systems under conventional heavy-traffic. The hydrodynamic model in conventional heavy traffic coincides with the standard fluid model that is obtained from the diffusion scaling by further scaling the space to obtain a strong-law-of-large-numbers scaling. However, we show that in the many-server heavy traffic, hydrodynamic limits satisfy a set of deterministic equations that are similar to but different from the fluid model equations. In addition, we introduce the notion of an SSC function to formulate our SSC results. Bramson, on the other hand, focused on establishing a relationship between a low-dimensional workload process and a high-dimensional queue length process.

We illustrate our approach by establishing SSC results for two systems. The first system we focus on is a static buffer priority (SBP) V-parallel server system. In a V-parallel server system, there is only one server pool handling several customer classes. Under an SBP policy, the classes are assigned fixed priorities. When a server switches from one customer to another, the new customer will be taken from the highest priority nonempty class. We show, using our framework, that under an SBP scheduling policy, the system achieves the following SSC: all of the buffers, except the one with the lowest priority, are always empty in the many-server diffusion limit. A special case of our SSC result for the V-model has appeared in Gurvich et al. [25], and a slightly different model has been analyzed in Puhalskii and Reiman [40].

The second system we study is also a V-model with two customer classes but under a different policy; we focus on the threshold policy proposed in Armony and Maglaras [3]. Under this policy, customers in class 1 have (nonpreemptive) priority over customers in class 2 unless the number of customers in the second queue exceeds a threshold value. When this occurs, second customer class is given (nonpreemptive) priority. Under this policy, we show that one can obtain the number of customers in either queue, given the total number of customers in the system in the diffusion limit. More specifically, we show that the number of customers in the second class will never exceed the threshold value in the limit. If, in the limit, the total number of customers in the queue is less than the threshold, there are no customers in the first queue. When the total number of customers exceeds the threshold, the number of customers in the second queue is equal to the threshold, and the rest will be in the first queue. This SSC result is instrumental in proving that this policy is asymptotically optimal in minimizing the wait time of the customers in the first class subject to a waiting time constraint for the second class.

The remainder of this paper is organized as follows. In the rest of this section, we review the related literature and present the notation and terminology used. In §2, we first give the technical details of the systems that are considered. We present a set of equations that these systems must satisfy, and we define the primitive processes. In §3, we formulate a static planning problem that is similar to that in Harrison [29]. Using this formulation we characterize a general many-server heavy traffic condition. Then we define the diffusion-scaling that will be studied in the subsequent sections. In §4, we present the general framework to prove a SSC result in the diffusion limits. Section 5 is devoted to the proof of our main results presented in §4. We present extensions to our main results in §6. We establish an SSC result under the static priority system for V-models in §7. In §8 we establish the SSC result for the V-model under the policy proposed in Armony and Maglaras [3].

1.1. Literature review. Standard references on conventional heavy traffic analysis include Harrison [28], Chen and Yao [12], and Whitt [48]. Results from classical queueing theory for the analysis of parallel server systems can be found in several textbooks; see, for example, Ross [43] and Gross and Harris [23]. Early many-server diffusion approximations for $GI/G/n$ systems have appeared in Borovkov [9], Iglehart [31], and Whitt [47], with the limiting traffic intensity of the system converging to a value less than one. Halfin and Whitt [26] studied the $GI/M/n$ system in the many-server heavy traffic regime that is the focus of this paper. We restrict the rest of our review to the literature on the Halfin-Whitt many-server asymptotic analysis.

The analysis of Halfin and Whitt has been extended in several directions. Garnett et al. [22] studied the asymptotic analysis of an $M/M/n$ system with impatient customers, and they established results similar to those in Halfin and Whitt. Puhalskii and Reiman [40] established the diffusion limit of a $G/Ph/n$ system, where Ph stands for a phase-type service time distribution. They also established the many-server diffusion limits of a V-model parallel server system under a static priority policy. To the best of our knowledge, the first SSC result in the Halfin-Whitt regime appeared in Puhalskii and Reiman [40]. Whitt [50] studied the many-server diffusion limit of a $G/H_2^*/n/m$ system, where H_2^* indicates that the service time distribution is an extremal distribution among the class of hyperexponential distributions. He later used this analysis in Whitt [49] to approximate $G/GI/n/m$ systems. In Dai et al. [17], many-server diffusion limits are established for critically loaded $G/Ph/n+GI$ systems and for overloaded $G/Ph/n+M$ systems. Reed [42] established one-dimensional limits for the number of customers in $G/G/n$ systems, and Mandelbaum and Momčilović [36] extended this result to $G/G/n+G$ systems.

After the initial version of this paper was completed, Gurvich and Whitt [24] used an approach similar to ours to study the diffusion limits of queue-and-idleness-ratio (QIR) controls in many-server systems. There is a subtle difference between our approach and theirs. We provide a general framework to prove *multiplicative* SSC results. Once a multiplicative SSC result is proved, there is an extra step needed to prove an SSC result. The extra step is to prove a stochastic bound for the diffusion-scaled processes. An application of our framework for a general parallel server system under static priority policies can be found in Dai and Tezcan [16]. In Gurvich and Whitt [24], instead of trying to establish a multiplicative SSC result, by using a stopping time argument and hydrodynamic scaling, they prove an SSC result for QIR policies directly. The SSC function under their policies does not satisfy Assumption 4.1 in our paper, except when the holding cost is linear. Therefore, our framework cannot be used to obtain their SSC result for queues operating under QIR policies and incurring a general holding cost. It is an interesting research direction to identify whether their approach can be generalized to a general class of control policies.

Armony and Maglaras [2, 3] studied an $M/M/n$ system with two customer classes. The hydrodynamic scaling we introduce in this study is based on the hydrodynamic scaling in Bramson [10] and is similar to the scaling that is used in Armony and Maglaras [3]. The scaling in Armony and Maglaras [3] has also been used earlier in conventional heavy traffic in Maglaras [32] and in many-server heavy traffic in Fleming et al. [20]. Gurvich et al. [25] studied a V-parallel server system with impatient customers; they show that a static buffer priority policy with a threshold policy is asymptotically optimal. Armony [1] studied an inverted-V-parallel server system and shows that the faster-server-first (FSF) policy is asymptotically optimal. The SSC results established in Gurvich et al. [25] and Armony [1] can easily be proved by using our framework. In Tezcan and Dai [46] and Dai and Tezcan [16], we showed that a greedy policy is asymptotically optimal for N-systems and parallel serve systems, respectively. In our proofs, the framework established in this paper plays a pivotal role.

In conventional heavy traffic, Brownian control problems have been formulated to devise good policies (see Harrison [27]). Extending this idea, Harrison and Zeevi [30] and Atar et al. [7] formulated a diffusion control problem to study a V-parallel server system with impatient customers in many-server heavy traffic. Atar et al. [7] proved that the policies obtained from this approach are asymptotically optimal. Atar [5, 6] followed an approach similar to that in Atar et al. [7] to find asymptotically optimal policies for tree-like systems. In Mandelbaum

et al. [38], they used a uniform acceleration technique to obtain the fluid and diffusion limit of a Markovian service network. Included in their modeling framework are networks of $M_t/M_t/n_t$ queues with abandonment and retrials in many-server heavy traffic.

1.2. Notation. The set of nonnegative integers is denoted by $\mathbb{N}$. For an integer $d \geq 1$, the d-dimensional Euclidean space is denoted by $\mathbb{R}^d$, and $\mathbb{R}_+$ denotes $[0, \infty)$. Unless stated otherwise, for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we will use the norm $|x| = \max_{\{i=1,\ldots,n\}} |x_i|$. For the norm of an $n \times m$ matrix A, we will use $|A| = \max_{\{i=1,\ldots,n\}} |A_i|$, where A_i is the ith row of A. We use $\{x^r\}$ to denote a sequence whose rth term is x^r. For $x \in \mathbb{R}$, $x^- = (-x) \vee 0$ and $x^+ = x \vee 0$. For a function $f\colon \mathbb{R} \to \mathbb{R}^d$ with d being some positive integer, we say that t is a regular point of f if f is differentiable at t and use $\dot{f}(t)$ to denote its derivative at t.

For each positive integer d, $\mathbb{D}^d[0, \infty)$ denotes the d-dimensional Skorohod path space (see Ethier and Kurtz [19]). All of the processes considered in this paper have sample paths in this space. For $x, y \in \mathbb{D}^d[0, \infty)$ and $T > 0$ we set

$$\|x(t) - y(t)\|_T = \sup_{0 \leq t \leq T} |x(t) - y(t)|.$$

The space $\mathbb{D}^d[0, \infty)$ is endowed with the J_1 topology, and the weak convergence in this space is considered with respect to this topology. For a sequence of functions $\{x^r\}$ in $\mathbb{D}^d[0, \infty)$, the sequence is said to converge uniformly on compact (u.o.c.) sets to $x \in \mathbb{D}^d[0, \infty)$ as $r \to \infty$, denoted by $x^r \to x$ u.o.c., if for each $T > 0$

$$\|x^r(t) - x(t)\|_T \to 0 \quad \text{as } n \to \infty.$$

The term FSLLN stands for functional strong law of large numbers, and FCLT stands for functional central limit theorem (see Chen and Yao [12] for details).

A sequence of random variables $\{x^r\}$ is said to satisfy the compact containment condition if

$$\lim_{C \to \infty} \limsup_{r \to \infty} P\{|x^r| > C\} = 0. \tag{1}$$

A sequence of stochastic processes $\{X^r(\cdot)\}$ is said to satisfy the compact containment condition if $\|X^r(t)\|_T$ satisfies the compact containment condition for every $T > 0$.

2. Parallel server systems and the asymptotic framework. We consider a system with parallel server pools and several customer classes. A server pool consists of several servers whose service capacities and capabilities are the same. Customers arrive at the system exogenously, and upon arrival they are routed to one of the buffers (or queues). Two customers that are routed to the same buffer are said to be in the same class. Each customer is served by one of the servers. Once the service of a customer is completed by one of the servers, the customer leaves the system. We assume that with each customer there is an associated patience time. A customer abandons the system without getting any service if the waiting time in the queue exceeds the customer's patience. Once a customer joins a queue, he cannot swap to other queues, and once his service starts, he cannot abandon the system. We refer to these systems as *parallel server systems*.

We use I to denote the number of arrival streams, J to denote the number of server pools, and K to denote the number of customer classes. For notational convenience, we define $\mathcal{I} = \{1, \ldots, I\}$ as the set of arrival streams, $\mathcal{J} = \{1, \ldots, J\}$ as the set of server pools, and $\mathcal{K} = \{1, \ldots, K\}$ as the set of customer classes.

The customers arriving in the ith stream are called type i customers. The arrivals of type i customers follow a delayed renewal process with rate λ_i; see (18) below. Upon arrival, each arriving customer is assigned to one of the buffers according to a routing policy (more details are provided below). Customers that are routed to buffer k are said to be class k customers. We assume that the set of pools that can handle class k customers is fixed and denote it by $\mathcal{J}(k)$. Similarly, we assume that the set of queues that servers in pool j can handle is fixed and denote it by $\mathcal{K}(j)$.

After a customer is routed to a queue, say, queue k, he proceeds directly to service if there is an available server in one of the pools in $\mathcal{J}(k)$. Otherwise, the customer joins the queue, waiting to be served later. We assume that the service time of a class k customer by a server in pool j is exponentially distributed with rate $\mu_{jk} > 0$ for all $k \in \mathcal{K}(j)$ and that customers in class k have exponentially distributed patience with rate γ_k. We denote the number of servers in pool j by N_j for $j \in \mathcal{J}$ and set $N = (N_1, \ldots, N_j)$. The total number of servers in the system is denoted by $|N|$.

To operate a multiclass parallel server system, control policies must also be given. A control policy must specify a routing policy that can be used to route an arriving customer to one of the buffers and a scheduling

policy that can be used to dispatch a server to serve a customer. Such dispatching is needed in two circumstances: first, whenever a server completes the service of a customer and there exist multiple customers in different classes that the server can handle and, second, whenever a customer arrives and there exist one or more idle servers who can handle that customer class. We restrict our attention to control policies that are head-of-the-line and nonidling. A scheduling policy is said to be *nonidling* if a server never idles when there is a customer waiting in one of the queues that can be served by that server and *head-of-the-line* (HL) if each server can only serve one customer at any given time and the customers in the same queue proceed to service on a FIFO basis. We assume that our control policies are *nonpreemptive*; under such a policy once the service of a customer starts, it can not be interrupted before it is finished. Although our results still hold in some special cases when preemptions are allowed, in general they cannot be used with preemptions (see Remark 5.1 for more details).

We assume also that rerouting of customers is not allowed. We call a control policy nonidling and HL if the associated scheduling policy is nonidling and HL. The nonidling assumption can be relaxed. Our asymptotic results also hold when the nonidling condition is only assumed to hold in the limit (see §2.2 for more details).

We also focus on control policies that are Markovian in the sense that they only use information on the queue length and number of customers in service to make routing decisions at the time of an arrival or to allocate servers to customer classes at the time of an arrival or a departure (note that Markovian policies are nonanticipative). We define a strictly increasing sequence $\{\sigma_l\}_{l=0}^{\infty}$ that specifies the successive times at which an arrival occurs to, or a departure occurs from, some class in the network. These time points depend, of course, on the policy that the system operates under and can be constructed as described below. We assume that the policy takes actions only when the state of the system is changed via an arrival or a departure and, hence, the server allocations remain constant between $[\sigma_n, \sigma_{n+1})$ for $n \geq 1$. The new allocations for the next interval $[\sigma_{n+1}, \sigma_{n+2})$ are assigned based on the state of the system during the previous interval $[\sigma_n, \sigma_{n+1})$ and the events that happen at time σ_{n+1}.

Let $Q(t) = (Q_k(t);\ k \in \mathcal{K})$ and $Z(t) = (Z_{jk}(t);\ j \in \mathcal{J},\ k \in \mathcal{K}(j))$, where $Q_k(t)$ denotes the number of class k customers in queue (not including those in service) at time t; and $Z_{jk}(t)$ denotes the number of class k customers served by a server in server pool j at time t. To specify the allocation scheme we assume that associated with each policy π there exists a function $f_\pi\colon \mathbb{N}^I \times \mathbb{N}^{J\times K} \times \mathfrak{E} \to \mathbb{N}^I \times \mathbb{N}^{J\times K}$ such that

$$f_\pi(Q(\sigma_n-), Z(\sigma_n-), e_n) = (Q(\sigma_n), Z(\sigma_n)) \tag{2}$$

gives the new allocations, where $\mathfrak{E}$ is the set of possible events, and e_n is an event at time σ_n. When more than one event occurs at time σ_n (which we show can only happen with probability zero), these events are ordered arbitrarily, and the policy π makes successive allocations via (2) for each event e_n. The function f_π must satisfy a set of constraints, such as nonidling and the capacity constraint. For example, the latter constraint says that the system cannot allocate more servers from a server pool than the number of servers available. Rather than spelling out all constraints explicitly, these constraints will be formulated implicitly through a set of system equations in the next section. We call f_π the transition function for policy π, and we say that a policy is *admissible* if it is nonidling, HL, nonpreemptive, and has the Markovian structure described in (2).

2.1. The dynamics of parallel server systems. In this section we describe the dynamics of a parallel server system. Actually, we will describe in detail the dynamics of a "perturbed" system. The perturbed system is closely related to the parallel server system, and it allows us to write down queueing network equations that are similar to the ones in the standard multiclass queueing networks. The equivalence of these two systems, under the exponential service and patience time assumption, will be discussed at the end of this section; note that a control policy for routing and server scheduling is needed to operate the perturbed system. Like the parallel server system, we assume that each control policy for the perturbed system is admissible. We denote a generic admissible control policy by π.

The perturbed system is identical to the parallel server system, except that its service and abandonment mechanisms are modified as follows. At any given time, when $n \geq 1$ servers in pool j serve n class k customers, the n servers work on a single class k customer. Note that in the original system a server can only serve one customer at a time. The remaining $n-1$ customers are said to be *locked for service*; they do not receive any service, even though they have left queue k. The single customer in service, called the *active customer*, can be chosen arbitrarily among the n customers. We assume that the service efforts from the n servers are *additive* in that service of the active customer is completed when the total time spent by all servers on the customer reaches the service requirement of the customer. When the service of the active customer is completed, the customer departs the system, and one of the servers working on that customer is freed. At this point, the remaining $n-1$

servers choose a new active customer from one of the $n-1$ locked customers in an arbitrary fashion, and the freed server is either assigned to a class, say k', or stays idle following a nonidling scheduling policy. In the former case, the server locks a class k' customer, with a given service requirement, for service. If there is an active customer that is currently being served by n' servers in pool j that are working on class k' customers, the new server joins the service efforts of these n' servers on the active customer. Otherwise, the locked class k' becomes an active customer, served by the new server.

The abandonment mechanism is modified similarly to the service mechanism. The mathematical characterization is given below by (4) and (5). We assume that whenever there are customers waiting in a queue, only the first customer in that queue may abandon. We assume that with each customer class, there is an associated remaining patience time. The remaining patience time associated with class k is set equal to the value of a new exponential random variable with rate γ_k at time 0 and at each time point a customer abandons from that class. Also, at time t the remaining patience time decreases with rate q_k, where q_k is the number of class k customers waiting in the queue at time t.

The object of study in this paper is a stochastic process $\mathbb{X} = (A, A_q, A_s, Q, Z, R, G, T, B, D)$, where $\mathbb{X}$ is defined via the perturbed system, and each of its components is explained in the next few paragraphs. The notation used in this section is inspired by that used in Puhalskii and Reiman [40] and Armony [1]. The first component is $A = (A_i\colon i \in \mathcal{I})$, where $A_i(t)$ denotes the total number of arrivals by time t for type i customers. We give more details about the structure of the arrival process in the next section. Here, we just mention that it is a delayed renewal process (see Ross [43]). The second component is $A_q = (A_{ik};\ i \in \mathcal{I},\ k \in \mathcal{K})$, where $A_{ik}(t)$ denotes the total number of type i arrivals by time t who are routed to queue k at the time of their arrival and who had to wait in the queue before their service started. The third component is $A_s = (A_{ijk};\ i \in \mathcal{I},\ k \in \mathcal{K},\ j \in \mathcal{J}(k))$, where $A_{ijk}(t)$ denotes the total number of type i customers who have been routed to queue k and locked for service immediately after their arrival at server pool j by time t. The component B is $(B_{jk}\colon j \in \mathcal{J},\ k \in \mathcal{J}(k))$, where $B_{jk}(t)$ denotes the total number of class k customers who are delayed in the queue and whose service started in pool j before time t. The components Z and Q are $(Z_{jk}\colon j \in \mathcal{J},\ k \in \mathcal{K}(j))$ and $Q = (Q_k\colon k \in \mathcal{K})$, respectively, where we use $Z_{jk}(t)$ to denote the total number of servers in pool j that serve class k customers and $Q_k(t)$ to denote the total number of customers in queue k at time t. The components T and D are $(T_{jk}\colon j \in \mathcal{I},\ k \in \mathcal{J}(j))$ and $(D_{jk}\colon j \in \mathcal{J},\ k \in \mathcal{K}(j))$, respectively, where $T_{jk}(t)$ denotes the total time spent serving class k customers by all N_j servers of pool j, and $D_{jk}(t)$ denotes the total number of class k customers whose service is completed by a server in pool j by time t. The component R is $(R_k\colon k \in \mathcal{K})$, where $R_k(t)$ denotes the number of customers who have abandoned queue k by time t.

Let $\{S_{jk},\ j \in \mathcal{J},\ k \in \mathcal{K}\}$ be a set of independent Poisson processes with each process S_{jk} having rate $\mu_{jk} > 0$. We set $S = (S_{jk})$ and $\mu = (\mu_{jk})$. For the perturbed system, we model the total number of class k customers whose service is completed by servers in pool j via

$$D_{jk}(t) = S_{jk}(T_{jk}(t)), \quad t \geq 0. \tag{3}$$

Similarly, let $F_k = \{F_k(t)\colon t \geq 0\}$ be a Poisson processes with rate γ_k for $k \in \mathcal{K}$. We define

$$G_k(t) = \int_0^t Q_k(s)\,\mathrm{d}s \quad t \geq 0 \tag{4}$$

for all $k \in \mathcal{K}$. For the perturbed system

$$R_k(t) = F_k(G_k(t)), \quad t \geq 0, \tag{5}$$

for all $k \in \mathcal{K}$.

The process $\mathbb{X}$ depends on the control policy used in the perturbed system. To emphasize the dependence on the control policy π used, we use $\mathbb{X}_\pi$ to denote the process. Clearly, each component of A, A_q, A_s, B, T, and D is a nondecreasing process, and each component of Q and Z is nonnegative. Furthermore, the process $\mathbb{X}_\pi$ satisfies the following equations in addition to (3)–(5) for all $t \geq 0$.

$$A_i(t) = \sum_{k \in \mathcal{K}} A_{ik}(t) + \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}(k)} A_{ijk}(t), \quad \text{for all } i \in \mathcal{I}, \tag{6}$$

$$Q_k(t) = Q_k(0) + \sum_{i \in \mathcal{I}} A_{ik}(t) - \sum_{j \in \mathcal{J}(k)} B_{jk}(t) - R_k(t), \quad \text{for all } k \in \mathcal{K}, \tag{7}$$

$$Z_{jk}(t) = Z_{jk}(0) + \sum_{i \in \mathcal{I}} A_{ijk}(t) + B_{jk}(t) - D_{jk}(t), \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{8}$$

$$\sum_{k\in\mathcal{K}(j)} Z_{jk}(t) \le N_j, \quad \text{for all } j\in\mathcal{J}, \tag{9}$$

$$T_{jk}(t) = \int_0^t Z_{jk}(s)\,\mathrm{d}s, \quad \text{for all } j\in\mathcal{J} \text{ and } k\in\mathcal{K}(j), \tag{10}$$

$$Q_k(t)\Big(\sum_{j\in\mathcal{J}(k)}\Big(N_j - \sum_{l\in\mathcal{K}(j)} Z_{jl}(t)\Big)\Big) = 0, \quad \text{for all } k\in\mathcal{K}, \tag{11}$$

$$\int_0^t \sum_{j\in\mathcal{J}(k)}\Big(N_j - \sum_{l\in\mathcal{K}(j)} Z_{jl}(s-)\Big)\,\mathrm{d}A_{ik}(s) = 0, \quad \text{for all } i\in\mathcal{I} \text{ and } k\in\mathcal{K}, \tag{12}$$

$$\text{Equations associated with the control policy } \pi. \tag{13}$$

The interpretation of (10) is that the busy time for server pool j working on class k at time t accumulates with rate equal to the total number of servers from pool j working on class k customers at time t. Equations (11) and (12) are based on the assumed nonidling property of a control policy. Equation (11) implies that there can be customers in the queue only when all the servers that can serve that queue are busy. Equation (12) implies that an arriving customer is delayed in the queue only if there is no idle server that can serve to that customer at the time of his arrival. Equation (13) forces the routing and scheduling decisions to be made according to the selected routing and scheduling policies. Other conditions are self-explanatory.

We call $\mathbb{X}_\pi$ the π-parallel server system process (or just π-parallel server system), although $\mathbb{X}_\pi$ is a process defined through the perturbed system. We note that each component of $\mathbb{X}_\pi$ is an element of the Skorohod space with the appropriate dimension, and so is $\mathbb{X}_\pi$.

Note that for a given admissible control policy π, it can be applied to both the parallel server system and the perturbed system. For the parallel system, one can define the corresponding process

$$\mathbb{X}'_\pi = (A', A'_q, A'_s, Q', Z', R', G', T', B', D') \tag{14}$$

with each component having the same interpretation as in the perturbed system. Clearly, $A' = A$. Yet, careful readers have noticed that the corresponding Equations (5) and (3) for the abandonment and departure processes, respectively, do *not* hold. Indeed, $\mathbb{X}'_\pi$ is *sample pathwise* different from the corresponding process $\mathbb{X}_\pi$, although $\mathbb{X}'_\pi$ satisfies all Equations (6)–(13), except for (3) and (5). For the admissible policies described in the previous section, under the assumptions that our service times and patience times are exponentially distributed, $\mathbb{X}_\pi$ is equal to $\mathbb{X}'_\pi$ in distribution when they are given the same initial condition.

THEOREM 2.1. *Under an admissible policy π, $\mathbb{X}_\pi$ is equal to $\mathbb{X}'_\pi$ in distribution when they are given the same initial condition.*

The proof is presented in Appendix A.

2.2. Primitive processes. The main goal of this paper is to study the SSC results in many-server diffusion limits. Therefore, we analyze a sequence of systems indexed by r such that the arrival rates grow to infinity as $r\to\infty$. The number of servers also grows to infinity to meet the growing demand. We append "r" to the processes that are associated with the rth system, e.g., $Q^r_k(t)$ is used to denote the number of class k customers in the queue in the rth system at time t. The arrival rate for the ith arrival stream in the rth system is given by λ^r_i, and we set $\lambda^r = (\lambda^r_1, \dots, \lambda^r_I)$. We assume that

$$\lambda^r_i \to \infty, \quad i\in\mathcal{I}, \tag{15}$$

as $r\to\infty$.

Let $\{S_{jk},\ j\in\mathcal{J},\ k\in\mathcal{K}\}$ be the Poisson process defined as in the previous section and $\{v_{jk}(l)\colon\ l=1,2,\dots\}$ be the corresponding sequence of independent and identically distributed (iid) exponential random variables. Because S_{jk} is a Poisson process, $v_{jk}(l)$ has exponential distribution with rate μ_{jk}. We define $V_{jk}\colon \mathbb{N}\to\mathbb{R}$ by

$$V_{jk}(m) = \sum_{l=1}^m v_{jk}(l), \quad m\in\mathbb{N},$$

where, by convention, empty sums are set to be zero. The term $V_{jk}(m)$ is the total service requirement of the first m class k customers that are served by pool j servers, and V_{jk} is known as the cumulative service time process. By the duality of S_{jk} and V_{jk}, one has

$$S_{jk}(t) = \max\{m\colon\ V_{jk}(m)\le t\}, \quad t\ge 0.$$

It follows from (3) that

$$V_{jk}(D^r_{jk}(t)) \le T^r_{jk}(t) \le V_{jk}(D^r_{jk}(t)+1), \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j). \tag{16}$$

This condition is identical to the HL condition in a standard multiclass queueing network, where each station has a single server (see, for example, Dai [13]).

Let $\{F_k,\ k \in \mathcal{K}\}$ be the Poisson process defined as in the previous section and $\{v_k(l)\colon\ l = 1, 2, \dots\}$ be the corresponding sequence of iid random variables. Each random variable has an exponential distribution with rate γ_k. We define $\Upsilon_k\colon \mathbb{N} \to \mathbb{R}$ by

$$\Upsilon_k(m) = \sum_{l=1}^{m} v_k(l), \quad m \in \mathbb{N}.$$

The process $\Upsilon_k(m)$ gives the total waiting time needed for m customers from class k to abandon. Similar to the discussion for the service times, by the duality of F_k and Υ_k, one has

$$F_k(t) = \max\{m\colon\ \Upsilon_k(m) \le t\}, \quad t \ge 0.$$

It follows from (5) that

$$\Upsilon_k(R^r_k(t)) \le G^r_k(t) \le \Upsilon_k(R^r_k(t)+1), \quad \text{for all } k \in \mathcal{K}. \tag{17}$$

Next, we give the details of the arrival processes. Let $\chi_i = \{E_i(t)\colon\ t \ge 0\}$ be a delayed renewal process with rate 1 and $\chi = \{\chi_i\colon\ i \in \mathcal{I}\}$. We assume that χ_i's are independent. Let

$$A^r_i(t) = \chi_i(\lambda^r_i t). \tag{18}$$

Let $\{u_i(l)\colon\ l = 1, 2, \dots\}$ be the sequence of interarrival times that are associated with the process χ_i. Note that they are independent and that $\{u_i(l)\colon\ l = 2, 3, \dots\}$ are identically distributed. We define $U_i\colon \mathbb{N} \to \mathbb{R}$ by

$$U_i(m) = \sum_{l=1}^{m} u_i(l), \quad m \in \mathbb{N},$$

and so

$$\chi_i(t) = \max\{m\colon\ U_i(m) \le t\}.$$

We require that the interarrival times of the arrival processes satisfy the following condition, which is similar to condition (3.4) in Bramson [10]:

$$\mathbb{E}[u_i(2)^{2+\epsilon}] < \infty, \quad \text{for all } i \in \mathcal{I} \text{ and for some } \epsilon > 0. \tag{19}$$

Condition (19) is automatically satisfied by the service times because they are assumed be exponentially distributed. This condition is needed in Lemma C.2, which is an integral part of our analysis. For the rest of the paper, we assume that the primitive processes of a parallel server system satisfy (19). We also assume that $Q^r(0)$, $Z^r(0)$, E, F, and S are independent.

We require that the number of servers in the rth system in each pool is selected so that

$$\lim_{r\to\infty} \frac{N^r_j}{|N^r|} = \beta_j, \quad \text{for all } j \in \mathcal{J} \text{ and for some } \beta_j > 0 \qquad \text{and} \tag{20}$$

$$\lim_{r\to\infty} \frac{\lambda^r_i}{|N^r|} = \lambda_i, \quad \text{for all } i \in \mathcal{I} \text{ and for some } 0 < \lambda_i < \infty. \tag{21}$$

We set $\lambda = (\lambda_1, \dots, \lambda_I)$ and assume that $\{|N^r|\}$ is strictly increasing in r. Conditions (15) and (21) imply that $|N^r| \to \infty$ as $r \to \infty$.

A policy is said to be *asymptotically nonidling* if there exists a sequence $\{\varsigma^r\} \subset \mathbb{R}_+$ such that

$$Q^r_k(t) > 0 \quad \text{only when} \quad \left(\sum_{j\in\mathcal{J}(k)} \left(N_j - \sum_{l\in\mathcal{K}(j)} Z_{jl}(t)\right)\right) < \varsigma^r, \tag{22}$$

$$A_{ik}(t_2) - A_{ik}(t_1) = 0 \quad \text{if} \quad \sum_{j\in\mathcal{J}(k)} \left(N_j - \sum_{l\in\mathcal{K}(j)} Z_{jl}(s)\right) > \varsigma^r, \tag{23}$$

for all $i \in \mathcal{I}$ and $k \in \mathcal{K}$, and $s \in [t_1, t_2]$, and

$$\frac{s^r}{\sqrt{|N^r|}} \to 0 \tag{24}$$

as $r \to \infty$. It can be shown that hydrodynamic and fluid limits of nonidling policies and asymptotically nonidling policies satisfy the same nonidling conditions. Hence, our framework can also be used to study SSC results under an asymptotically nonidling policy. Hydrodynamic and fluid limits are introduced in §4 and Appendix B, respectively.

3. The static planning problem and asymptotic framework. The static planning problem (SPP) has been used in the literature to determine the optimal nominal allocations of servers' capacities for the service of customer classes (see Harrison [29], Dai and Lin [14], among others). Nominal allocations determine the long-run proportion of servers' effort allocated to each class. We take a similar approach to determine the nominal proportion of servers in a server pool that will be allocated to serve each class.

The static planning problem is defined as

$$\begin{aligned} \min \quad & \rho \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} \alpha_{ik} = \lambda_i, \quad \text{for all } i \in \mathcal{I}, \\ & \sum_{j\in\mathcal{J}(k)} \beta_j \mu_{jk} x_{jk} = \sum_{i\in\mathcal{I}} \alpha_{ik}, \quad \text{for all } k \in \mathcal{K}, \\ & \sum_{k\in\mathcal{K}(j)} x_{jk} \le \rho, \quad \text{for all } j \in \mathcal{J}, \\ & x_{jk}, \alpha_{ik} \ge 0, \quad \text{for all } j \in \mathcal{J},\ k \in \mathcal{K}, \text{ and } i \in \mathcal{I}. \end{aligned} \tag{25}$$

The quantity α_{ik}/λ_i can be thought of as the long-run proportion of type i customers that are routed to queue k and x_{jk} as the average long-run fraction of time for pool j servers working on class k customers. We set $\alpha = \{\alpha_{ik}\colon i \in \mathcal{I},\ j \in \mathcal{K}\}$ and $x = \{x_{jk}\colon j \in \mathcal{J},\ k \in \mathcal{K}\}$.

The objective of the SPP is to minimize the average utilization of the "busiest" server pool. From this formulation it is clear that referring to x as the "fraction of time" is a misnomer because $\sum_{k\in\mathcal{K}(j)} x_{jk}$ may be greater than 1. We use the term "fraction of time" because of Assumption 3.1 below.

The main difference between our formulation of the SPP and the one in Harrison [29] is that we model routing of customers to queues explicitly, as in Stolyar [44]. We pay the price by having one more constraint than Harrison's formulation. The main constraint is to be able to serve all of the incoming customers. This is formulated in the first and the second constraints. The first constraint ensures that all the arriving customers are routed to one of the queues, and the second constraint is needed to guarantee that enough service capacity is allocated to all customer classes.

Let (ρ^*, x^*, α^*) be an optimal solution to the SPP. If $\rho^* > 1$, it can be easily shown that the queue length process is not bounded under the fluid limit for r large enough (fluid limits are defined in Appendix B). We will assume for the rest of this paper that $\rho^* \le 1$.

Now, consider the sequence of parallel server systems described in the previous section and the associated SPP with the rth system:

$$\begin{aligned} \min \quad & \rho^r \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} \alpha^r_{ik} = \lambda^r_i, \quad \text{for all } i \in \mathcal{I}, \\ & \sum_{j\in\mathcal{J}(k)} N^r_j \mu_{jk} x^r_{jk} = \sum_{i\in\mathcal{I}} \alpha^r_{ik}, \quad \text{for all } k \in \mathcal{K}, \\ & \sum_{k\in\mathcal{K}(j)} x^r_{jk} \le \rho^r, \quad \text{for all } j \in \mathcal{J}, \\ & x^r_{jk}, \alpha^r_{ik} \ge 0, \quad \text{for all } j \in \mathcal{J},\ k \in \mathcal{K}, \text{ and } i \in \mathcal{I}. \end{aligned} \tag{26}$$

Let $(\rho^{r,*}, x^{r,*}, \alpha^{r,*})$ be an optimal solution to (26). Note that the SPP (25) has at least one solution because the objective function is continuous, and the constraints define a compact set. Next, we formulate the many-server heavy traffic condition.

ASSUMPTION 3.1. *For each optimal solution* (ρ^*, x^*, α^*) *of the SPP* (25) *with* λ *given by* (21) *and* β *given by* (20)*, we have* $\rho^* = 1$ *and* $\sum_{k\in\mathcal{K}(j)} x^*_{jk} = 1$ *for all* $j \in \mathcal{J}$*. Moreover, for any sequence of optimal solutions* $\{x^{r,*}\}$ *of* (26) *we have*

$$x^{r,*} \to x^*,$$

as $r \to \infty$ *for some optimal solution* x^* *of* (25)*.*

Even when the SPP (25) has an optimal solution with $\rho^* \le 1$, it is not a trivial task to come up with a control policy that will achieve the optimal allocations in the long run. If ρ^* is close to one, small deviations from the optimal allocations may again cause the queue length to grow without a bound. This phenomenon is closely related to the stability of a control policy in a multiclass queueing network setting. In this paper we only consider control policies that satisfy the following assumption.

ASSUMPTION 3.2. *For a control policy* π*,*

$$Q^r(\cdot)/|N^r| \to 0 \qquad \text{and} \qquad Z^r(\cdot)/|N^r| \to z \quad \text{u.o.c. a.s.,} \tag{27}$$

as $r \to \infty$*, if* $(Q^r(0)/|N^r|, Z^r(0)/|N^r|) \to (0, z)$ *a.s., as* $r \to \infty$*, where* $z = (z_{jk},\ j \in \mathcal{J},\ k \in \mathcal{K}(j))$*; and* $z_{jk} = \beta_j x^*_{jk}$ *for an optimal solution* (ρ^*, x^*, α^*) *of* (25)*.*

We provide a fluid model framework that can be used to ensure that a control policy satisfies Assumption 3.2 in Appendix B. Assumption 3.1 is fairly standard in heavy-traffic analysis (usually, uniqueness of x^* is also assumed). Assumption 3.2 is on a control policy π. Under a control policy, when Assumption 3.2 is satisfied, the fluid limits exist and do not blow up, even though they are critically loaded. Clearly, this condition must be satisfied by any policy that has a "reasonable" performance (see also the discussion at the end of §4.1 for the importance of this assumption). We assume that

$$Q^r(0)/|N^r| \to 0 \qquad \text{and} \qquad Z^r(0)/|N^r| \to z \quad \text{a.s.,} \tag{28}$$

as $r \to \infty$, where z is given as in Assumption 3.2. Under Assumption 3.2, condition (28) implies that

$$Q^r(\cdot)/|N^r| \to 0 \qquad \text{and} \qquad Z^r(\cdot)/|N^r| \to z(\cdot) \quad \text{u.o.c. a.s.,}$$

as $r \to \infty$, where $z(t) = z$, for $t \ge 0$. In general, diffusion limits are introduced as a refinement of the fluid limits. Under condition (28) and Assumption 3.2, we define the diffusive scaling as follows:

$$\hat{Q}^r(t) = \frac{Q^r(t)}{\sqrt{|N^r|}} \qquad \text{and} \qquad \hat{Z}^r_{jk}(t) = \frac{Z^r_{jk}(t) - x^*_{jk}N^r_j}{\sqrt{|N^r|}}, \quad \text{for } t \ge 0. \tag{29}$$

4. Main results.

In this section, we present a general framework to prove a SSC result in the many-server diffusion limit of a π-parallel server system process. We first introduce the hydrodynamic model equations. The solutions of these equations play an important role in the general SSC framework. We present our main results in §4.2. Naturally, some of the hydrodynamic equations depend on the policy used. Examples of SSC results and hydrodynamic equations will be discussed in §§7 and 8. The proofs of the results in this section are presented in §5.

4.1. Hydrodynamic model equations.

Consider the process $\tilde{\mathbb{X}}_\pi = (\tilde{A}, \tilde{A}_q, \tilde{A}_s, \tilde{Q}, \tilde{Z}, \tilde{B})$ and the following set of equations:

$$\lambda_i t = \sum_{k\in\mathcal{K}} \tilde{A}_{ik}(t) + \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \tilde{A}_{ijk}(t), \quad \text{for all } i \in \mathcal{I}, \tag{30}$$

$$\tilde{Q}_k(t) = \tilde{Q}_k(0) + \sum_{i\in\mathcal{I}} \tilde{A}_{ik}(t) - \sum_{j\in\mathcal{J}(k)} \tilde{B}_{jk}(t), \quad \text{for all } k \in \mathcal{K}, \tag{31}$$

$$\tilde{A}_{ik}, \tilde{A}_{ijk}, \tilde{B}_{jk} \text{ are nondecreasing for all } i \in \mathcal{I},\ j \in \mathcal{J}, \text{ and } k \in \mathcal{K}, \tag{32}$$

$$\tilde{Z}_{jk}(t) = \tilde{Z}_{jk}(0) + \sum_{i\in\mathcal{I}} \tilde{A}_{ijk}(t) + \tilde{B}_{jk}(t) - \mu_{jk}\tilde{T}_{jk}(t), \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{33}$$

$$\tilde{T}_{jk}(t) = \int_0^t z_{jk}\,ds = z_{jk}t, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{34}$$

$$\tilde{Q}_k(t) \geq 0, \quad \text{for all } k \in \mathcal{K} \qquad \text{and} \qquad \sum_{k \in \mathcal{K}(j)} \tilde{Z}_{jk}(t) \leq 0, \quad \text{for all } j \in \mathcal{J}, \tag{35}$$

$$\tilde{Q}_k(t)\left(\sum_{j \in \mathcal{J}(k)} \sum_{l \in \mathcal{K}(j)} \tilde{Z}_{jl}(t) \right) = 0, \quad \text{for all } k \in \mathcal{K}, \tag{36}$$

$$\int_0^t \sum_{j \in \mathcal{J}(k)} \left(\sum_{l \in \mathcal{K}(j)} \tilde{Z}_{jl}(s) \right) \mathrm{d}\tilde{A}_{ik}(s) = 0, \quad \text{for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K}, \tag{37}$$

$$\text{Additional equations associated with the control policy } \pi, \tag{38}$$

where λ_i is defined as in (21) and z_{ij} as in Assumption 3.2. Equations (30)–(38) are called the hydrodynamic model equations, and they define the *hydrodynamic model* of the π-parallel server system. Any process $\tilde{\mathbb{X}}_\pi$ satisfying (30)–(38) for all $t \geq 0$ is called a hydrodynamic model solution. The differences between the fluid and hydrodynamic models are discussed at the end of this section.

Hydrodynamic model solutions are similar to the fluid model solutions; they are deterministic and absolutely continuous, hence, almost everywhere differentiable. Absolute continuity follows from the following result.

Proposition 4.1. *Any process $\tilde{\mathbb{X}}_\pi$ satisfying (30)–(38) for all $t \geq 0$ is Lipschitz continuous.*

Proof. Assume that $\tilde{\mathbb{X}}_\pi$ satisfies (30)–(38) for all $t \geq 0$. First note that $\tilde{A}_{ik}$, and $\tilde{A}_{ijk}$ are Lipschitz for all $i \in \mathcal{I}$, $j \in \mathcal{J}$ and $k \in \mathcal{K}$ by (30) and (32). Next, we show that $\tilde{B}_{jk}$ is Lipschitz continuous.

Let $t_1 < t_2$. If $\tilde{Q}_k(t) > 0$ for all $t \in [t_1, t_2]$, then by (33), (36), and the fact that $\tilde{A}_{ijk}$ is nondecreasing,

$$\sum_{k' \in \mathcal{K}(j)} (\tilde{B}_{jk'}(t_2) - \tilde{B}_{jk'}(t_1)) \leq \sum_{k' \in \mathcal{K}(j)} \mu_{jk'} z_{jk'}(t_2 - t_1) \tag{39}$$

for all $j \in \mathcal{J}(k)$. Because $\tilde{B}_{jk}$s are nondecreasing, this implies

$$\tilde{B}_{jk}(t_2) - \tilde{B}_{jk}(t_1) \leq \sum_{k' \in \mathcal{K}(j)} \mu_{jk'} z_{jk'}(t_2 - t_1) \tag{40}$$

for all $j \in \mathcal{J}(k)$.

Now assume that $\tilde{Q}_k(t) = 0$ for some $t \in [t_1, t_2]$ and let $t_0 = \inf\{t \in [t_1, t_2]\colon \tilde{Q}_k(t) = 0\}$. We assume that $t_0 > t_1$, because otherwise the proof follows from (44) below by replacing t_0 with t_1. We prove below that

$$\lim_{t \uparrow t_0} \tilde{Q}_k(t) = \tilde{Q}_k(t_0) = 0. \tag{41}$$

The continuity of $\tilde{A}_{ik}$, Equation (31), condition (41), and the fact that $\tilde{B}_{jk}$s are nondecreasing implies that

$$\lim_{t \uparrow t_0} \tilde{B}_{jk}(t) = \tilde{B}_{jk}(t_0) \tag{42}$$

for all $j \in \mathcal{J}(k)$.

By definition of t_0, we have that $\tilde{Q}_k(t) > 0$ for all $t \in [t_1, t_0)$. Hence, similar to (40),

$$\tilde{B}_{jk}(t) - \tilde{B}_{jk}(t_1) \leq \sum_{k' \in \mathcal{K}(j)} \mu_{jk'} z_{jk'}(t - t_1)$$

for all $j \in \mathcal{J}(k)$ and $t \in [t_1, t_0)$. By taking limit $t \uparrow t_0$ and using (42), we have

$$\tilde{B}_{jk}(t_0) - \tilde{B}_{jk}(t_1) \leq \sum_{k' \in \mathcal{K}(j)} \mu_{jk'} z_{jk'}(t_0 - t_1). \tag{43}$$

Also, by (31) and the fact that $\tilde{Q}_k(t_0) = 0$,

$$\sum_{i \in \mathcal{I}} \left(\tilde{A}_{ik}(t_2) - \tilde{A}_{ik}(t_0)\right) \geq \sum_{j \in \mathcal{J}(k)} \left(\tilde{B}_{jk}(t_2) - \tilde{B}_{jk}(t_0)\right). \tag{44}$$

Then, it follows from (42), (43), (44) and the fact that $\tilde{B}_{jk}$s are nondecreasing that

$$\tilde{B}_{jk}(t_2) - \tilde{B}_{jk}(t_1) \leq \sum_{i \in \mathcal{I}} (\tilde{A}_{ik}(t_2) - \tilde{A}_{ik}(t_1)) + \sum_{k' \in \mathcal{K}(j)} \mu_{jk'} z_{jk'}(t_2 - t_1). \tag{45}$$

Hence, $\tilde{B}$ is Lipschitz. The Lipschitz continuity of $\tilde{Z}_{jk}$ and $\tilde{Q}_k$ now follow from this, (31), and (33).

To complete the proof of the lemma, it remains to prove (41). We now prove

$$\lim_{t\uparrow t_0} \tilde{Q}_k(t) = 0. \tag{46}$$

The result

$$\tilde{Q}_k(t_0) = 0 \tag{47}$$

is proved similarly by considering t_0, instead of t_n in the proof below.

Assume on the contrary that (46) does not hold. Then there exists an increasing sequence $\{t_n\}$ such that $t_n \uparrow t_0$ and $\lim_{n\to\infty} \tilde{Q}_k(t_n) > \delta$ for some $\delta > 0$. By the definition of t_0, for any $\epsilon > 0$ there exist $\tilde{t}_0$ and n large enough such that $\tilde{Q}_k(t_n) > \delta$, $\tilde{Q}_k(\tilde{t}_0) = 0$ and $\tilde{t}_0 - t_n < \epsilon$. By (31), this implies that, for $\epsilon > 0$ small enough,

$$\tilde{B}_{j'k}(\tilde{t}_0) - \tilde{B}_{j'k}(t_n) \geq a\delta/2 \tag{48}$$

for $a = 1/|\mathcal{J}(k)|$ and for some $j' \in \mathcal{J}(k)$.

Because $\tilde{Q}_k(t_n) > 0$, by (36), $\sum_{k'\in\mathcal{K}(j)} \tilde{Z}_{jk'}(t_n) = 0$ for all $j \in \mathcal{J}(k)$. Hence, by (35)

$$\sum_{k'\in\mathcal{K}(j')} \tilde{Z}_{j'k'}(\tilde{t}_0) - \sum_{k'\in\mathcal{K}(j')} \tilde{Z}_{j'k'}(t_n) \leq 0. \tag{49}$$

This inequality and (33) imply that

$$\sum_{i\in\mathcal{I}} \sum_{k'\in\mathcal{K}(j')} (\tilde{A}_{ij'k'}(\tilde{t}_0) - \tilde{A}_{ij'k'}(t_n)) + \sum_{k'\in\mathcal{K}(j')} (\tilde{B}_{j'k'}(\tilde{t}_0) - \tilde{B}_{j'k'}(t_n)) \leq \sum_{k'\in\mathcal{K}(j')} \mu_{jk} z_{jk}(\tilde{t}_0 - t_n).$$

Because $\tilde{A}_{ijk}$ is nondecreasing, by selecting ϵ small and n large enough, we have

$$\sum_{k'\in\mathcal{K}(j')} (\tilde{B}_{j'k'}(\tilde{t}_0) - \tilde{B}_{j'k'}(t_n)) \leq a\delta/4. \tag{50}$$

Inequalities (48) and (50), and the fact that B_{jk}s are nondecreasing imply that

$$\tilde{B}_{j'k''}(\tilde{t}_0) - \tilde{B}_{j'k''}(t_n) \leq -a\delta/(4|\mathcal{K}(j')|). \tag{51}$$

for some $k'' \in \mathcal{K}(j')$, which contradicts with the fact that $\tilde{B}_{jk}$s are nondecreasing. Therefore, (46) is proved. □

It will be proved in Proposition 5.4 in §5 that the hydrodynamic model equations are satisfied by hydrodynamic limits under certain general assumptions; these limits are obtained from the hydrodynamically scaled sequences such as

$$\left\{\left(\frac{1}{\sqrt{|N^r|}} Q_k^r(t/\sqrt{|N^r|}),\ \frac{1}{\sqrt{|N^r|}}(Z_{jk}^r(t/\sqrt{|N^r|}) - x_{jk}^* N_j^r)\right),\ t \geq 0\right\} \quad r = 1, 2, \ldots; \tag{52}$$

see §5 for details. Equation (38) is obtained from the policy π. It has to be justified mathematically that the hydrodynamic limits satisfy this equation. We demonstrate this for two systems in §§7 and 8.

Differences between fluid and hydrodynamic models. A fluid model is introduced in Appendix B. It is defined by fluid model Equations (B3)–(B10). Unlike hydrodynamic model equations, fluid model equations are satisfied by fluid limits obtained from fluid-scaled sequences

$$\left\{\left(\frac{1}{|N^r|} Q_k^r(t),\ \frac{1}{|N^r|} Z_{jk}^r(t)\right),\ t \geq 0\right\}, \quad r = 1, 2, \ldots .$$

The fluid model and fluid limits developed in Appendix B provide a practical tool for one to verify Assumption 3.2. The fluid-scaling keeps the diffusion time scale but reduces the space resolution by a factor of $1/\sqrt{|N^r|}$, whereas the hydrodynamic scaling slows down the diffusion scaling in (29) by a factor of $1/\sqrt{|N^r|}$ (see (52)).

A comparison of hydrodynamic model Equations (30)–(38) with fluid model Equations (B3)–(B10) reveals major differences. The most important difference is between hydrodynamic model Equation (34) and fluid model Equation (B6). The fluid model Equation (B6) is intuitive and is a direct consequence of system dynamic

Equation (10). The hydrodynamic model Equation (34) is subtle. It holds under Assumption 3.2 and is justified through a hydrodynamic limit procedure (see (71) and the derivation of (C27) in Appendix C.3.1). Even under Assumption 3.2, when the initial condition $\bar{Z}_{jk}(0) \neq z_{jk}$, it will take some time for $\bar{Z}_{jk}(t)$ to converge to z_{jk} in the fluid model. Therefore, when $\bar{Z}_{jk}(0) = \tilde{Z}_{jk}(0) \neq z_{jk}$, the fluid model dynamics and the hydrodynamic model dynamics are different.

Without Assumption 3.2, one may still attempt to develop a hydrodynamic model. In this case, (34) is false in general, and we do not know what dynamic equation can be used to replace (34). Without an additional dynamic equation on $\tilde{T}_{jk}$ such as (34), the hydrodynamic model can hardly be analyzed, and thus the entire hydrodynamic framework is likely useless.

In the single-server multiclass queueing network setting, Bramson [10] uses Equations (2.15)–(2.20) in his paper to define a deterministic model. Using his Equations (5.10)–(5.15) for hydrodynamically scaled processes, he shows that each hydrodynamic limit is a solution to (2.15)–(2.20). Therefore, following the terminology of this paper, Bramson's deterministic model should be a hydrodynamic model, although he calls it a fluid model. The confusion is partly justified because his deterministic model (2.15)–(2.20) is also identical to the (nondelayed) fluid model studied in Dai [13] for multiclass queueing networks. Because both scalings and both models are simultaneously used in this paper, to avoid possible confusion, we purposely choose two different labels for these two scalings and two models. The term hydrodynamic scaling is consistent with the usage in Bramson [10] (see §5.1 for more details).

4.2. SSC in the diffusion limits. We need a machinery to define a state space collapse in mathematical terms, for this we use a function with the following properties. Let g: $\mathbb{R}^{K+d_z} \to \mathbb{R}^+$, where $d_z = \sum_{j\in\mathcal{J}} |\mathcal{K}(j)|$, be a nonnegative function that satisfies the following *homogeneity* condition:

$$g(\alpha x) = \alpha^c g(x), \tag{53}$$

for some $c > 0$, for all $x \in \mathbb{R}^{K+d_z}$, and for all $0 \leq \alpha \leq 1$. Recall that d_z is the dimension of the process Z. We call g an SSC-function. Nonnegativity assumption is made for notational convenience, and one can always consider $|g|$ in order to have a nonnegative function if g can take negative values. We make the following assumption about the SSC function.

Assumption 4.1. *The function g: $\mathbb{R}^{K+d_z} \to \mathbb{R}^+$ satisfies (53) and is continuous on $\mathbb{R}^{K+d_z}$.*

Assumption 4.1 is needed for a simple reason; we will consider a sequence of stochastic processes that converges to another one, and we would like to show that the sequence that consists of the values of g evaluated for each process converges to the value of g evaluated at the limiting process. Assumption 4.1 makes this possible by virtue of the continuous mapping theorem (see Chen and Yao [12]). Condition (53) will be needed when we translate the results from hydrodynamic scaled processes to diffusion-scaled processes (see Proposition 5.6). The class of functions that satisfy Assumption 4.1 is large enough for most purposes; however, this class can be extended as discussed in §6. Examples of SSC functions are presented in §§7 and 8.

As the machinery to state an SSC result has been set, we are ready to state the conditions on the hydrodynamic model solutions that imply that an SSC result holds in the diffusion limit. The following assumption is analogous to Bramson [10, Assumption 3.2].

Assumption 4.2. *Let g be a function that satisfies Assumption 4.1. There exists a function $H(t)$ with $H(t) \to 0$ as $t \to \infty$ such that*

$$g(\tilde{Q}(t), \tilde{Z}(t)) \leq H(t) \quad \text{for all } t \geq 0 \tag{54}$$

for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ satisfying $|(\tilde{Q}(0), \tilde{Z}(0))| \leq 1$. Furthermore, for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ with $g(\tilde{Q}(0), \tilde{Z}(0)) = 0$ and $|(\tilde{Q}(0), \tilde{Z}(0))| \leq 1$, $g(\tilde{Q}(t), \tilde{Z}(t)) = 0$ for $t \geq 0$.

We are ready to state the main result of this paper.

Theorem 4.1. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes. Suppose that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, the hydrodynamic model of π-parallel server system satisfies Assumption 4.2, and*

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \quad \text{in probability} \tag{55}$$

as $r \to \infty$. Then, for each $T > 0$,

$$\frac{\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_T}{(\|\hat{Q}^r(t)\|_T \vee \|\hat{Z}^r(t)\|_T \vee 1)^c} \to 0 \quad \text{in probability}, \tag{56}$$

as $r \to \infty$, where $c > 0$ is given as in (53).

REMARK 4.1. The result of Theorem 4.1 is still valid if it is only assumed that hydrodynamic limits, not the hydrodynamic model, satisfy Assumption 4.2. This relaxes the assumption because it will be shown that every hydrodynamic limit over a finite time interval $[0, L]$, for some $L > 0$, is a hydrodynamic model solution on $[0, L]$. The set of hydrodynamic model solutions may contain processes that are not hydrodynamic limits.

REMARK 4.2. The SSC result, as stated in Theorem 4.1, is called the *multiplicative SSC*. If $(\hat{Q}^r, \hat{Z}^r)$ also satisfies the compact containment condition, then one can use this property to remove the denominator from (56) and obtain a (*strong*) SSC that is more suitable for applications.

The condition (55) can be relaxed as in Bramson [10, Theorem 3] to only require that $\hat{Q}^r(0)$ and $\hat{Z}^r(0)$ satisfy the compact containment condition. The SSC result in this case, however, does not hold initially at time 0.

THEOREM 4.2. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes. Suppose that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, the hydrodynamic model of π-parallel server system satisfies Assumption 4.2, and $|\hat{Q}^r(0)| \vee |\hat{Z}^r(0)|$ satisfies the compact containment condition. Then, for some $L^r = o(\sqrt{|N^r|})$ with $L^r \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,*

$$P\left\{\frac{\sup_{L^r/\sqrt{|N^r|}\le t\le T}|g(\hat{Q}^r(t),\hat{Z}^r(t))|}{\sup_{L^r/\sqrt{|N^r|}\le t\le T}(|\hat{Q}^r(t)|\vee|\hat{Z}^r(t)|\vee 1)^c} > \epsilon\right\} \to 0, \tag{57}$$

as $r \to \infty$, where $c > 0$ is given as in (53).

REMARK 4.3. Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes that satisfy the conditions of Theorem 4.2. If in addition H, given as in Assumption 4.2, is bounded, then

$$\lim_{C\to\infty}\limsup_{r\to\infty} P\left\{\|g(\hat{Q}^r(t),\hat{Z}^r(t))\|_{L^r/\sqrt{|N^r|}} > C\right\} = 0. \tag{58}$$

The result (58) may be used to verify that

$$\lim_{C\to\infty}\limsup_{r\to\infty} P\left\{\sup_{L^r/\sqrt{|N^r|}\le t\le T}\left(|\hat{Q}^r(t)|\vee|\hat{Z}^r(t)|\right) > C\right\} = 0. \tag{59}$$

Then, similar to Remark 4.2, one can deduce a strong state space collapse result from Theorem 4.2 using (59).

5. SSC framework.

In this section we prove Theorem 4.1. We begin with introducing the hydrodynamic scaling that will be used to define the hydrodynamic limits. Once we establish the relationship between the hydrodynamic scaled processes and the hydrodynamic limits, we translate condition (54) to a condition on the diffusion-scaled processes. We finally show that this latter condition implies the desired SSC result in the diffusion limit.

5.1. Hydrodynamic scaling and bounds.

The hydrodynamic scaling is used by Bramson [10] to establish a relationship between the hydrodynamic and the diffusion limits in conventional heavy traffic asymptotic analysis. We consider a similar time scaling that slows the process down. This allows us to analyze the events that happen instantaneously in the diffusive scale in more detail. This can be achieved by using a scaling similar to the diffusion scaling as given in (29) but also scaling the time by $1/\sqrt{|N^r|}$. However, this scaling is not suitable for our purposes.

We need the more refined scaling, which we call the *hydrodynamic scaling*. We divide the interval $[0, T]$ into $T\sqrt{|N^r|}$ intervals of length $1/\sqrt{|N^r|}$ and analyze the processes in each intervals. We index the intervals by m. For a nonnegative integer m, let

$$x_{r,m} = \left|Q^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right|^2 \vee \left|Z^r\left(\frac{m}{\sqrt{|N^r|}}\right) - \vec{N}^r x^*\right|^2 \vee |N^r|, \tag{60}$$

where $\vec{N}^r$ is a diagonal matrix with $\vec{N}^r_{jj'} = N^r_j$ if $j = j'$ and 0 otherwise for $j \in \mathcal{J}$ and $x^* = (x_{jk}, j \in \mathcal{J}, k \in \mathcal{K})$ is given as in Assumption 3.2. Hence, $Z^r(t) - \vec{N}^r x^*$ is a $J \times K$ matrix with its (j,k)th entry equal to $Z^r_{jk}(t) - x^*_{jk}N^r_j$ if $k \in \mathcal{K}(j)$ and zero otherwise. Note that the square root of the first two terms of $x_{r,m}$ gives the deviations of these processes from their fluid limits.

We define the hydrodynamic scaling by shifting and scaling the processes of $\mathbb{X}^r$ as follows. For a process X^r associated with the rth process, we denote the hydrodynamic scaled version by $X^{r,m}$. For A^r, A^r_s, A^r_q, B^r, D^r, T^r, and R^r, the hydrodynamic scaling is defined for $t \in [0, L]$ for some $L > 0$ by

$$X^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(X^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - X^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right), \tag{61}$$

Hydrodynamic scaled version of Q^r and Z^r are defined as follows;

$$\begin{aligned} Q^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(Q^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right)\right), \qquad \text{and} \\ Z^{r,m}(t) &= \frac{1}{\sqrt{x_{r,m}}}\left(Z^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) - N^r x^*\right). \end{aligned} \tag{62}$$

Note that $Q^{r,m}(\cdot)$ and $Z^{r,m}(\cdot)$ keep track of the deviation of processes $Q^r(\cdot)$ and $Z^r(\cdot)$, respectively, from their respective fluid limits during the time interval $[m/\sqrt{|N^r|}, \sqrt{x_{r,m}}L/|N^r| + m/\sqrt{|N^r|}]$, which is also scaled with their initial value at time $m/\sqrt{|N^r|}$. Observe that $x_{r,m}$ must be in the order of $\sqrt{|N^r|}$ for Q^r and Z^r to have meaningful diffusion limits. Also, if $x_{r,m}$ is in the order of $\sqrt{|N^r|}$, then $Q^{r,m}(\cdot)$ and $Z^{r,m}(\cdot)$ is very similar to the diffusion scaling. Although our results hold no matter how $x_{r,m}$ behaves, provided conditions of Theorem 4.1 hold, this reveals the relationship between the hydrodynamic and diffusion scaling that will be used to translate a SSC result from hydrodynamic limits to diffusion limits.

For notational convenience, with a slight abuse of notation, we set

$$V^{r,m}_{jk}(D^{r,m}_{jk}(t), b) = \frac{1}{\sqrt{x_{r,m}}}\left(V_{jk}\left(D^r_{jk}\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) + b_1\right) - V_{jk}\left(D^r_{jk}\left(\frac{m}{\sqrt{|N^r|}}\right) + b_2\right)\right), \tag{63}$$

$$Y^{r,m}_{k}(R^{r,m}_{k}(t), b) = \frac{1}{\sqrt{x_{r,m}}}\left(Y_{k}\left(R^r_{k}\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) + b_1\right) - Y_{k}\left(R^r_{k}\left(\frac{m}{\sqrt{|N^r|}}\right) + b_2\right)\right), \tag{64}$$

and for $b = (b_1, b_2) \in \mathbb{R}^2$. By (16) and (17),

$$V^{r,m}_{jk}(D^{r,m}_{jk}(t), (0,1)) \le T^{r,m}_{jk}(t) \le V^{r,m}_{jk}(D^{r,m}_{jk}(t), (1,0)) \qquad \text{and} \tag{65}$$

$$Y^{r,m}_{k}(R^{r,m}_{k}(t), (0,1)) \le G^{r,m}_{k}(t) \le Y^{r,m}_{k}(R^{r,m}_{k}(t), (1,0)). \tag{66}$$

Let $\mathbb{X}^{r,m} = t(A^{r,m}, A^{r,m}_s, A^{r,m}_q, B^{r,m}, T^{r,m}, Q^{r,m}, Z^{r,m}, R^{r,m}, G^{r,m})$. We refer to $\mathbb{X}^{r,m}$ as the hydrodynamic scaled process. From the definition of $x_{r,m}$ we have that

$$|\mathbb{X}^{r,m}(0)| \le 1.$$

It can easily be checked that $\mathbb{X}^{r,m}$ satisfies the following equations for all $t \ge 0$.

$$A^{r,m}_i(t) = \sum_{k\in\mathcal{K}} A^{r,m}_{ik}(t) + \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}(k)} A^{r,m}_{ijk}(t), \quad \text{for all } i \in \mathcal{I}, \tag{67}$$

$$Q^{r,m}_k(t) = Q^{r,m}_k(0) + \sum_{i\in\mathcal{I}} A^{r,m}_{ik}(t) - \sum_{j\in\mathcal{J}(k)} B^{r,m}_{jk}(t) - R^{r,m}_k(t), \quad \text{for all } k \in \mathcal{K}, \tag{68}$$

$$Z^{r,m}_{jk}(t) = Z^{r,m}_{jk}(0) + \sum_{i\in\mathcal{I}} A^{r,m}_{ijk}(t) + B^{r,m}_{jk}(t) - D^{r,m}_{jk}(t), \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{69}$$

$$D^{r,m}_{jk}(t) = \frac{S_{jk}(\sqrt{x_{r,m}}T^{r,m}_{jk}(t) + T^r_{jk}(m/\sqrt{|N^r|})) - S_{jk}(T^r_{jk}(m/\sqrt{|N^r|}))}{\sqrt{x_{r,m}}}, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{J}(k), \tag{70}$$

$$T^{r,m}_{jk}(t) = \frac{|N^r_j|x^*_{jk}}{|N^r|}t + \frac{\sqrt{x_{r,m}}}{|N^r|}\int_0^t Z^{r,m}_{jk}(s)\,\mathrm{d}s, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{71}$$

$$R^{r,m}_k(t) = \frac{F_k(\sqrt{x_{r,m}}G^{r,m}_k(t) + G^r_k(m/\sqrt{|N^r|})) - F_k(G^r_k(m/\sqrt{|N^r|}))}{\sqrt{x_{r,m}}}, \quad \text{for all } k \in \mathcal{K}, \tag{72}$$

$$G_k^{r,m}(t) = \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^t Q_k^{r,m}(s)\,\mathrm{d}s, \quad \text{for all } k \in \mathcal{K}, \tag{73}$$

$$Q_k^{r,m}(t)\left(\sum_{j\in\mathcal{J}(k)}\sum_{k'\in\mathcal{K}(j)} Z_{jk'}^{r,m}(t)\right) = 0, \quad \text{for all } k \in \mathcal{K}, \tag{74}$$

$$\int_0^t \sum_{j\in\mathcal{J}(k)}\left(\sum_{k'\in\mathcal{K}(j)} Z_{jk'}^{r,m}(s-)\right)\mathrm{d}A_{ik}^{r,m}(s) = 0, \quad \text{for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K}. \tag{75}$$

We have the following estimates that are similar to those established in Proposition 5.1 in Bramson [10].

PROPOSITION 5.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Assume that (20) and (21) hold and π satisfies Assumption 3.2. Fix $\epsilon > 0$, $L > 0$ and $T > 0$. Then, for large enough r, there exists $N > 0$ such that*

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|A^{r,m}(t) - \frac{\lambda^r}{|N^r|}t\right\|_L > \epsilon\right\} \le \epsilon, \tag{76}$$

$$P\left\{\max_{m<\sqrt{|N^r|}T}\sup_{t_1,t_2\le L}|D^{r,m}(t_1) - D^{r,m}(t_2)| > N|t_1 - t_2| + \epsilon\right\} \le \epsilon, \quad \text{and} \tag{77}$$

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V_{jk}^{r,m}(D_{jk}^{r,m}(t), b) - \frac{1}{\mu_{jk}}D_{jk}^{r,m}(t)\right\|_L > \epsilon\right\} \le \epsilon, \tag{78}$$

for all $j \in \mathcal{J}, k \in \mathcal{K}(j)$, and $b = (1,0)$ or $(0,1)$.

The proof is given in Appendix C.2.1.

PROPOSITION 5.2. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. Fix $\epsilon > 0$, $L > 0$ and $T > 0$. Then, for large enough r,*

$$P\left\{\max_{m<\sqrt{||N^r||}T}\|R_k^{r,m}(t)\|_L > \epsilon\right\} \le \epsilon, \tag{79}$$

$k \in \mathcal{K}$.

The proof is given in Appendix C.2.2.

Using these two propositions, one can show that $\mathbb{X}^{r,m}$ is *almost* Lipschitz, as described in the next proposition. In this section and for the remainder of this paper N without a superscript is reused to denote a general constant.

PROPOSITION 5.3. *Let $\{\mathbb{X}^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. Fix $\epsilon > 0$, $L > 0$, and $T > 0$. Then, for large enough r,*

$$P\left\{\max_{m<\sqrt{|N^r|}T}\sup_{t_1,t_2\le L}|\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)| > N|t_1 - t_2| + \epsilon\right\} \le \epsilon, \tag{80}$$

where $N < \infty$ and only depends on (I, J, K, λ).

The proof is similar to that of Proposition 5.2 in Bramson [10] and presented in Appendix C.2.3. For convenience, we assume for the rest of the paper that $N \ge 1$ and $L \ge 1$. Let

$$\mathcal{H}_0^r = \left\{\max_{m<\sqrt{|N^r|}T}\sup_{t_1,t_2\le L}|\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)| \le N|t_1 - t_2| + \epsilon(r)\right\}, \tag{81}$$

where L, N, and T are fixed as before and $\epsilon(r)$ with $\epsilon(r) \to 0$ as $r \to \infty$ is a sequence of real numbers. Similarly, we can replace ϵ in (76), (78), and (79) by $\epsilon(r)$. We denote these new inequalities obtained from (76), (78), and (79) by (76′), (78′), and (79′). Let $\mathcal{H}^r$ denote the intersection of $\mathcal{H}_0^r$ with the complements of the events in (76′), (78′), and (79′). As in Bramson [10], when $\epsilon(r) \to 0$ sufficiently slowly as $r \to \infty$, one can show that $P(\mathcal{H}^r) \to 1$ as $r \to \infty$.

We summarize the above discussion in the following corollary for future reference.

COROLLARY 5.1. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. Fix $L > 0$ and $T > 0$ and choose N and $\epsilon(r)$ as above. Then, for $\mathcal{H}^r$ defined as above*

$$\lim_{r\to\infty} P(\mathcal{H}^r) = 1. \tag{82}$$

REMARK 5.1. If preemptions are allowed, (80) does not have to hold in general. More specifically, the problem is bounding the number of customers whose service start, the term $B_{jk}^{r,m}$, in a given interval. When preemptions are not allowed $B_{jk}^{r,m}(t_2) - B_{jk}^{r,m}(t_1)$ is bounded by the maximum of total number of arrivals to this class in this interval, $\sum_i (A_{ik}^{r,m}(t_2) - A_{ik}^{r,m}(t_1))$, and the total number of departures from server pool j in this interval, $\sum_k D_{jk}^{r,m}(t_2) - D_{jk}^{r,m}(t_1))$. If preemptions are allowed, this bound is not necessarily valid any more.

5.2. Hydrodynamic limits of π-parallel server systems. In this section, we define the hydrodynamic limits of π-parallel server systems. First, we define a set of functions that contains all of the hydrodynamic limits. The following definitions are similar to those in Bramson [10, §6], and the notation is adapted from that paper.

Fix $L > 0$. Let $\tilde{E}$ be the set of right continuous functions with left limits, $x\colon [0, L] \to \mathbb{R}^d$. Let E' denote those $x \in \tilde{E}$ that satisfies

$$|x(0)| \le 1$$

and

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

where constant N is chosen as in Proposition 5.3. We set

$$E^r = \left\{\mathbb{X}^{r,m},\ m < \sqrt{|N^r|}T,\ \omega \in \mathcal{H}^r\right\}$$

and

$$\mathcal{E} = \{E^r\colon\ r \in \mathbb{N}\},$$

where T is fixed, and $\mathcal{H}^r$ is defined as in the previous section. (These quantities are not correlated to the external arrival processes E introduced in §2.2.)

We define a hydrodynamic limit x of $\mathcal{E}$ to be a point $x \in \tilde{E}$ such that for all $\epsilon > 0$ and $r_0 \in \mathbb{N}$, there exist $r \ge r_0$ and $y \in E^r$, with $\|x(\cdot) - y(\cdot)\|_L < \epsilon$.

Because

$$|\mathbb{X}^{r,m}(0)| \le 1 \tag{83}$$

for all $m < \sqrt{|N^r|}T$ and $r \in \mathbb{N}$, the following result is a corollary in Bramson [10, Proposition 4.1] and is similar to Proposition 6.1 in that paper. It shows that the hydrodynamic limits are "rich" in the sense that for r large enough, every hydrodynamic scaled process is close to a hydrodynamic limit.

COROLLARY 5.2. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds, π satisfies Assumption 3.2. Let $\tilde{E}$, E^r, and $\mathcal{E}$ be as specified above. Fix $\epsilon > 0$, $L > 0$, and $T > 0$, and choose r large enough. Then, for $\omega \in \mathcal{H}^r$ and any $m < \sqrt{|N^r|}T$*

$$\|\mathbb{X}^{r,m}(\cdot) - \tilde{\mathbb{X}}(\cdot)\|_L \le \epsilon \tag{84}$$

for some hydrodynamic limit $\tilde{\mathbb{X}}(\cdot)$ of $\mathcal{E}$ with $\tilde{\mathbb{X}}(\cdot) \in E'$.

The next result is mainly needed to translate the condition on the hydrodynamic model solutions to hydrodynamic limits given in Assumption 4.2. It also reveals the origin of hydrodynamic model equations.

PROPOSITION 5.4. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. Choose $L > 0$ and let $\tilde{\mathbb{X}}_\pi$ be a hydrodynamic limit of $\mathcal{E}$ over $[0, L]$. $\tilde{\mathbb{X}}_\pi$ satisfies the hydrodynamic model Equations (30)–(38) on $[0, L]$.*

The proof is given in Appendix C.3.1.

Observe that by (53) and definitions of hydrodynamic and diffusion scalings

$$|g(Q^{r,0}(0), Z^{r,0}(0))| \le |g(\hat{Q}^r(0), \hat{Z}^r(0))|. \tag{85}$$

If condition (55) holds, (85) implies that $g(Q^{r,0}(0), Z^{r,0}(0)) \to 0$ in probability as $r \to \infty$. Therefore, we can choose $\epsilon(r)$ with $\epsilon(r) \to 0$ as $r \to \infty$ such that for $\mathcal{L}^r = \mathcal{H}^r \cap \mathcal{G}^r$, where

$$\mathcal{G}^r = \{|g(Q^{r,0}(0), Z^{r,0}(0))| \le \epsilon(r)\},$$

we have

$$\lim_{r\to\infty} P(\mathcal{L}^r) = 1. \tag{86}$$

We set

$$E_g^r = \{\mathbb{X}^{r,0}(\cdot, \omega),\ \omega \in \mathcal{L}^r\}$$

and

$$\mathcal{E}_g = \{E_g^r,\ r \in \mathbb{N}\}.$$

The following proposition is similar to Bramson [10, Proposition 6.4], a proof is presented in Appendix C.3.2.

PROPOSITION 5.5. *Let $\{\mathbb{X}_\pi^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, and the hydrodynamic model of the π-parallel server system satisfies Assumption 4.2. Fix $\epsilon > 0$, $L > 0$, and $T > 0$, and assume that r is large. Then, for $\omega \in \mathcal{H}^r$,*

$$g\big(Q^{r,m}(t), Z^{r,m}(t)\big) \le H(t) + \epsilon \tag{87}$$

for all $t \in [0, L]$, and $m < \sqrt{|N^r|}T$, with $H(\cdot)$ is given in Assumption 4.2.

Furthermore, for $\omega \in \mathcal{L}^r$

$$\|g(Q^{r,0}(t), Z^{r,0}(t))\|_L \le \epsilon. \tag{88}$$

If, in addition, condition (55) holds, then (86) holds.

5.3. SSC in the diffusion limits. In this section we change the scaling from hydrodynamic to diffusion to prove Theorem 4.1. Once the scaling is changed, a few complications needs to be dealt with regarding the change in the range of the time variable.

We begin with changing the scaling. One can check by employing (29) and (62) that

$$\begin{aligned} Q_k^{r,m}(t) &= \sqrt{\frac{|N^r|}{x_{r,m}}}\hat{Q}_k^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) = \frac{1}{y_{r,m}}\hat{Q}_k^r\left(\frac{1}{\sqrt{|N^r|}}(y_{r,m}t + m)\right) \quad \text{and} \\ Z_{jk}^{r,m}(t) &= \sqrt{\frac{|N^r|}{x_{r,m}}}\hat{Z}_{jk}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|} + \frac{m}{\sqrt{|N^r|}}\right) = \frac{1}{y_{r,m}}\hat{Z}_{jk}^r\left(\frac{1}{\sqrt{|N^r|}}(y_{r,m}t + m)\right), \end{aligned} \tag{89}$$

where

$$y_{r,m} = \sqrt{\frac{x_{r,m}}{|N^r|}} = \left|\hat{Q}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right| \vee \left|\hat{Z}^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right| \vee 1. \tag{90}$$

By changing the scaling in Proposition 5.5 as above, we can rephrase (87) and (88). However, the domain of the time scales will change and the domain $0 \le t \le L$ for the arguments on the left-hand side of (89) will correspond to

$$\frac{m}{\sqrt{|N^r|}} \le t \le \frac{1}{\sqrt{|N^r|}}(y_{r,m}L + m) \tag{91}$$

for the arguments on the right.

PROPOSITION 5.6. *Let $\{\mathbb{X}^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, and the hydrodynamic model of the π-parallel server system satisfies Assumption 4.2. Fix $\epsilon > 0$, $L > 0$, and $T > 0$, and assume that r is large. Then, for $\omega \in \mathcal{H}^r$ and for $H(\cdot)$ given as in Assumption 4.2*

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \le y^c_{r,m} H\left(\frac{1}{y_{r,m}}(\sqrt{|N^r|}t - m)\right) + \epsilon y^c_{r,m} \tag{92}$$

for all $t \in [0, T]$ and m satisfying (91). Also

$$\left\| g(\hat{Q}^r(t), \hat{Z}^r(t)) \right\|_{Ly_{r,0}/\sqrt{|N^r|}} \le \epsilon y^c_{r,0} \tag{93}$$

for $\omega \in \mathcal{L}^r$.

PROOF. The bounds (92) and (93) are obtained from (87) and (88), respectively, by applying (89) and using (53). □

If we can show that $(\sqrt{|N^r|}t - m)/y_{r,m}$ is large, where $|N^r|$ is the total number of servers in the rth system, we can conclude the proof of Theorem 4.1 by using the convergence property of $H(\cdot)$, as given in Assumption 4.2. It will be shown that it is enough to have $\sqrt{|N^r|}t - m$ and L large.

Because the value of L is a matter of choice, we can take L sufficiently large and redefine $\mathcal{H}^r$ with the reselected L. Let H be given as in Assumption 4.2. Because $H(t) \to 0$ as $t \to \infty$, independent of L, for $\epsilon > 0$ fixed, there exists $s^*(\epsilon) > 1$ such that for $t > s^*(\epsilon)$, $H(t) < \epsilon$. We assume for the rest of the paper that

$$L \ge 6Ns^*(\epsilon), \tag{94}$$

where N is chosen as in (81).

To make $\sqrt{|N^r|}t - m$ large, for a fixed $t \in [0, T]$, we take the smallest m that satisfies (91), which we denote by $m_r(t)$. We need the following lemmas, whose proofs are given in Appendix C.4, to show that $\sqrt{|N^r|}t - m_r(t)$ is large.

LEMMA 5.1. *Let $\{\mathbb{X}^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. For fixed $L > 0$ and $T > 0$, and large enough r*

$$y_{r,m+1} \le 3Ny_{r,m} \tag{95}$$

for $\omega \in \mathcal{H}^r$ and $m < \sqrt{|N^r|}T$, with the constant N chosen as in (81).

Let $y_r(m_r(t)) = y_{r,m_r(t)}$.

LEMMA 5.2. *Let $\{\mathbb{X}^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. For fixed $L > 0$ and $T > 0$, and large enough r*

$$\sqrt{|N^r|}t - m_r(t) \ge Ly_r(m_r(t))/6N \tag{96}$$

for $\omega \in \mathcal{H}^r$ and $t \in (Ly_{r,0}/\sqrt{|N^r|}, T]$, with the constant N chosen as in (81).

PROOF OF THEOREM 4.1. Assume that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, the hydrodynamic model of the π-parallel server system satisfies Assumption 4.2, and condition (55) holds.

Fix $\xi > 0$. By (82) and (86), there exists $r_0 > 0$ such that

$$P(\mathcal{H}^r) \ge P(\mathcal{L}^r) > 1 - \xi/2 \tag{97}$$

for all $r > r_0$. Fix $\epsilon > 0$ and take $L \ge 6Ns^*(\epsilon)$. Then, by (92) and Lemma 5.2, for $\omega \in \mathcal{H}^r$, $t \in (Ly_{r,0}/\sqrt{|N^r|}, T]$, and r large enough

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \le 2\epsilon(y_r(m_r(t)))^c. \tag{98}$$

However, by (90),

$$y_r(m_r(t)) = \left|\hat{Q}^r\left(\frac{m_r(t)}{\sqrt{|N^r|}}\right)\right| \vee \left|\hat{Z}^r\left(\frac{m_r(t)}{\sqrt{|N^r|}}\right)\right| \vee 1 \le \|\hat{Q}^r(\cdot)\|_T \vee \|\hat{Z}^r(\cdot)\|_T \vee 1. \tag{99}$$

From (93) and (99), for $t \in [0, Ly_{r,0}/\sqrt{|N^r|}]$ and $\omega \in \mathcal{L}^r$

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \le \epsilon(y_{r,0})^c \le \epsilon\big(\|\hat{Q}^r(\cdot)\|_T \vee \|\hat{Z}^r(\cdot)\|_T \vee 1\big)^c. \tag{100}$$

Combining (98), (99), and (100) gives

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \le 2\epsilon\big(\|\hat{Q}^r(\cdot)\|_T \vee \|\hat{Z}^r(\cdot)\|_T \vee 1\big)^c \tag{101}$$

for all $t \in [0, T]$ and $\omega \in \mathcal{L}^r$. Finally, by (97) and (101), for large enough r,

$$P\left\{\frac{\|g(\hat{Q}^r(\cdot), \hat{Z}^r(\cdot))\|_T}{(\|\hat{Q}^r(\cdot)\|_T \vee \|\hat{Z}^r(\cdot)\|_T \vee 1)^c} > 2\epsilon\right\} < \xi.$$

This clearly implies (56) because $\epsilon > 0$ and $\xi > 0$ are arbitrary. □

Proof of Theorem 4.2. Suppose that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, the hydrodynamic model of the π-parallel server system satisfies Assumption 4.2, and $|\hat{Q}^r(0)| \vee |\hat{Z}^r(0)|$ satisfies the compact containment condition:

Let

$$u^{r,\max} = \max_{i \in \mathcal{I}}\{u_i(m)\colon U_i(m-1) \le 2|N^r||\lambda|L,\ m = 1, 2, \dots\},$$

where $\lambda = (\lambda_1, \dots, \lambda_I)$ is given by (21). In words, $u^{r,\max}$ is an upper bound, for r large enough, for the maximum interarrival time for those events that started before time L of the process $\{A_i\colon i \in \mathcal{I}\}$, because $\lambda_i^r < 2|N^r||\lambda|$ for large enough r. Assume for the moment that

$$u^{r,\max}\big/\sqrt{|N^r|} \to 0 \quad \text{in probability as } r \to \infty \tag{102}$$

and that for some sequence $\{L^r\}$ that satisfies the conditions given in the theorem

$$\left|g\left(\hat{Q}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right), \hat{Z}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right)\right)\right| \to 0 \quad \text{in probability as } r \to \infty. \tag{103}$$

Consider the sequence of processes $\{\mathbb{Y}^r\}$ defined by $\mathbb{Y}^r(\cdot) = \mathbb{X}^r(L^r/\sqrt{|N^r|} + \cdot)$. Then, $\{\mathbb{Y}^r\}$ satisfies (55) by (103). Also by (102), distributions of the first interarrival times of the processes A and S after $L^r/\sqrt{|N^r|}$ satisfy the conditions needed for Proposition 5.1 to be valid. Because the other conditions of Theorem 4.1 are satisfied by $\{\mathbb{Y}^r\}$, the proof above can be repeated to show that (56) holds for $\{\mathbb{Y}^r\}$. However, this shows that (57) holds for $\{\mathbb{X}^r\}$. Hence, it suffices to show that (102) and (103) hold.

The limits (102) are proven in Lemma C.2.

Next we prove (103). We show that there exists a sequence $\{L^r\}$ with $L^r \to \infty$ as $r \to \infty$ and $L^r = o(\sqrt{|N^r|})$ such that for any $\epsilon > 0$ and $\xi > 0$, there exists r' such that

$$P\left\{\left|\left(\hat{Q}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right), \hat{Z}^r\left(\frac{L^r}{\sqrt{|N^r|}}\right)\right)\right| > \epsilon\right\} < \xi, \tag{104}$$

for all $r > r'$.

Set $\delta_n = 1/n$ and $\tilde{L}^n = (N^n)^{1/4}$ for all $n = 1, 2, \dots$. Define $\mathcal{K}^r_{\tilde{L}^n}$ as in §5.1; see (81) and the discussion succeeding it, with L being replaced with $\tilde{L}^n$. Note that by the definition of $\mathcal{K}^r_{\tilde{L}^n}$ and Proposition 5.6, there exists r_n such that for $r > r_n$

$$P\{\mathcal{K}^r_{\tilde{L}^n}\} > 1 - 1/n \tag{105}$$

and

$$g\big(\hat{Q}^r(t), \hat{Z}^r(t)\big) \le y^c_{r,m} H\left(\frac{1}{y_{r,m}}\big(\sqrt{|N^r|}t - m\big)\right) + \delta_n y^c_{r,m} \tag{106}$$

holds for all $t \in [0, T]$ and m satisfying (91).

Set $L^r = \tilde{L}^1$ and $\tilde{\mathcal{K}}^r = \mathcal{K}^r_{\tilde{L}^1}$ for $r \le r_2$, $L^r = \tilde{L}^n$, and $\tilde{\mathcal{K}}^r = \mathcal{K}^r_{\tilde{L}^n}$ for $r \in (r_n, r_{n+1}]$, and for $n = 2, 3, \ldots$. Note that $L^r = o(\sqrt{|N^r|})$, and $L^r \to \infty$ as $r \to \infty$. Furthermore,

$$\lim_{r\to\infty} P(\tilde{\mathcal{K}}^r) = 1.$$

Fix $\epsilon, \xi > 0$. Let

$$\mathcal{U}^r_C = \{|\hat{Q}^r(0)| \vee |\hat{Z}^r(0)| < C\}. \tag{107}$$

Choose r_0 and $C > 1$ such that for $r \ge r_0$

$$P(\mathcal{U}^r_C) > 1 - \xi/2.$$

We fix C to this value for the rest of the proof.

Let r_1' be the smallest integer greater than r_0 that satisfies $\delta_{r_1'} < \epsilon/(2C^c)$. Choose $r_2' > r_1'$ such that for all $r > r_2'$, $L^r > 2s^*(\delta_{r_1'})C$, where s^* is defined as in (94).

For $t \in [C^{-1}L^r y_{r,0}/\sqrt{|N^r|}, L^r y_{r,0}/\sqrt{|N^r|}]$, $m_r(t) = 0$ from (91) and

$$\sqrt{|N^r|}t \ge C^{-1}L^r y_{r,0}.$$

Hence, for $r > r_2'$, by (106),

$$g(\hat{Q}^r(t), \hat{Z}^r(t)) \le 2\delta_{r_1'} y^c_{r,0} < \epsilon \tag{108}$$

for all $t \in [C^{-1}L^r y_{r,0}/\sqrt{|N^r|}, L^r y_{r,0}/\sqrt{|N^r|}]$ and $\omega \in \tilde{\mathcal{K}}^r \cap \mathcal{U}^r_C$.

Now observe that for $\omega \in \tilde{\mathcal{K}}^r \cap \mathcal{U}^r_C$, $L^r/\sqrt{|N^r|} \in [C^{-1}L^r y_{r,0}/\sqrt{|N^r|}, L^r y_{r,0}/\sqrt{|N^r|}]$ for all $r \ge 1$. Hence, by (5.3) and (108) there exits $r' > r_2'$ such that for $r > r'$

$$P\{g(\hat{Q}^r(L^r/\sqrt{|N^r|}), \hat{Z}^r(L^r/\sqrt{|N^r|})) > \epsilon\} < \xi.$$

This gives (104), thus completes the proof of (103). □

Proof of Remark 4.3. Assume that $\{\mathbb{X}^r_\pi\}$ is a sequence of π-parallel server system processes that satisfy the conditions of Theorem 4.2. Also, assume that $g(\hat{Q}^r(0), \hat{Z}^r(0))$ satisfy the compact containment condition and H is bounded.

Fix $L > 0$. By assumption there exists a constant $B_0 > 0$ such that $\sup_{t\in[0,\infty)} H(t) < B_0$. By (87)

$$g(Q^{r,m}(t), Z^{r,m}(t)) < 2B_0$$

for all $t \in [0, L]$. This implies, similar to (92), that

$$\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{Ly_{r,0}/\sqrt{|N^r|}} < 2B_0 y^c_{r,0}. \tag{109}$$

for all $\omega \in \mathcal{K}^r$. Let $\mathcal{U}^r_C$ be defined as in (107). Since $|\hat{Q}^r(0)| \vee |\hat{Z}^r(0)|$ satisfy the compact containment condition by assumption, for $\epsilon > 0$ fixed, there exists $C > 0$ and $r_1 > 0$ such that $P(\mathcal{U}^r_C) > 1 - \epsilon$, for all $r > r_1$. For each fixed L, on $\mathcal{U}^r_C \cap \mathcal{K}^r$

$$\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{L/\sqrt{|N^r|}} \le \|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{Ly_{r,0}/\sqrt{|N^r|}} \le R.$$

Now choose the sequence $\{L^r\}$ as in the previous proof. Then,

$$\limsup_{r\to\infty} P\{\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_{L^r/\sqrt{|N^r|}} > C\} < \epsilon.$$

Since ϵ and C is arbitrary, this completes the proof. □

6. Extensions. In this section we present two extensions of our main result. In the first extension, we weaken the homogeneity assumption (53) on the SSC function. In the second extension, we only assume that the SSC function is continuous but make an additional assumption that $\hat{Q}^r$ and $\hat{Z}^r$ satisfy the compact containment condition.

6.1. A weaker homogeneity condition. Theorem 4.1 can be generalized by relaxing condition (53) on the class of SSC functions. We replace condition (53) with the following condition: there exist $c_1 > 0$ and $c_2 > 0$ such that

$$\alpha^{c_1} g(x) \le g(\alpha x) \le \alpha^{c_2} g(x) \tag{110}$$

for all $x \in \mathbb{R}^{K+d_z}$ and $0 \le \alpha \le 1$.

COROLLARY 6.1. *Under condition* (110), *Theorem* 4.1 *holds with* c *replaced with* c_1.

The proof is placed in Appendix D.1.

6.2. When the homogeneity condition does not hold. When the SSC function g does not satisfy homogeneity assumption (53) or (110), the framework in Theorem 4.1 does not directly apply. In this section we present a framework without requiring the homogeneity assumption on the SSC function g. We show that when g is only continuous but (1) holds, a relationship between SSC in hydrodynamic model solutions and diffusion limits still exists.

One of the models we are interested in studying is presented in Armony and Maglaras [3], where a threshold type policy is proposed. For this kind of policies conditions (53) or (110) prove to be too strong. More details will be discussed in §8.

Now we consider a general parallel server system and extend our main result, Theorem 4.1, to prove SSC results when the SSC function satisfies a weaker condition. For the extension we only assume that the SSC function satisfies the following condition.

ASSUMPTION 6.1. *The SSC function* $g\colon \mathbb{R}^{K+d_z} \to \mathbb{R}^+$ *is a nonnegative and continuous function.*

When the SSC function g only satisfies Assumption 6.1 but not Assumption 4.1 we need the following compact containment condition on the queue length and number of busy server processes.

ASSUMPTION 6.2. *For every* $T > 0$, (1) *holds for the sequence of random variables* $l\{(\|\hat{Q}^r(\cdot)\|_T \vee \|\hat{Z}^r(\cdot)\|_T)\}$.

When the SSC function g does not satisfy (53), we replace Assumption 4.2 with the following stronger assumption.

ASSUMPTION 6.3. *There exists a constant* C_0 *such that for every* $C > C_0$, *there exists a function* $H_C(t)$ *with* $H_C(t) \to 0$ *as* $t \to \infty$ *such that*

$$g\big(C(\tilde{Q}(t), \tilde{Z}(t))\big) \le H_C(t) \quad \textit{for all } t \ge 0 \tag{111}$$

for each hydrodynamic model solution $\tilde{\mathbb{X}}_\pi$ *satisfying* $|(\tilde{Q}(0), \tilde{Z}(0))| \le 1$.

Furthermore, for each hydrodynamic hydrodynamic solution $\tilde{\mathbb{X}}_\pi$ *with* $|(\tilde{Q}(0), \tilde{Z}(0))| \le 1$ *and* $g(C(\tilde{Q}(0), \tilde{Z}(0))) = 0$, $g(C(\tilde{Q}(t), \tilde{Z}(t))) = 0$ *for* $t \ge 0$.

We are ready to state the main result of this section.

THEOREM 6.1. *Let* $\{\mathbb{X}^r_\pi\}$ *be a sequence of* π*-parallel server system processes. Suppose that Assumption* 1 *holds,* π *satisfies Assumption* 3.2, g *satisfies Assumption* 6.1, *Assumption* 6.2 *holds, the hydrodynamic model of* π*-parallel server system satisfies Assumption* 6.3, *and*

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \quad \textit{in probability} \tag{112}$$

as $r \to \infty$. *Then, for each* $T > 0$,

$$\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_T \to 0 \quad \textit{in probability}, \tag{113}$$

as $r \to \infty$.

The proof is presented in Appendix D.2.

REMARK 6.1. Because we will use a slightly different hydrodynamic scaling in the proof of Theorem 6.1, the hydrodynamic Equations (38) used in Assumption 6.3 can be different from those in Assumption 4.2. We remark that an equation can be added into the hydrodynamic model only when it is satisfied by each hydrodynamic limit. See the proof of Theorem 6.1 for more details.

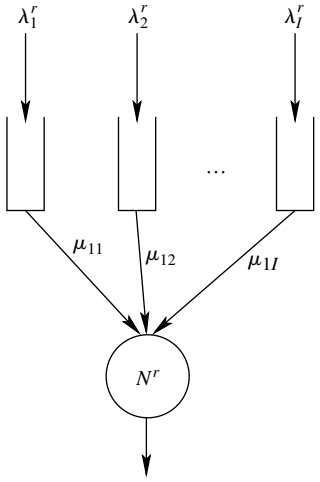

FIGURE 1. A V-parallel server system.

7. SSC for V-parallel server systems. In this and the following section we illustrate the applications of our main results. Additional (more complex) applications of our main results can be found in Tezcan [45] and Dai and Tezcan [16]. The first system we consider is known as the V-parallel server system. A V-parallel server system consists of multiple customer classes and a single server pool (see Figure 1 for an example). We study this model when the scheduling decisions are made according to a static buffer priority (SBP) policy and we call the resulting queueing system the SBP V-parallel server system. We show that in the diffusion limit of an SBP V-parallel server system all the buffers except the one with the lowest priority is empty. This model has recently been studied by Gurvich et al. [25]. They prove an SSC result that is similar to our main result of this section (Theorem 7.1 below) by assuming that the service rates of all classes are the same; see Proposition 3.2 in that paper. A slightly different model with phase-type service time distributions is studied by Puhalskii and Reiman [40].

As explained above, in a V-parallel server system there are I arrival streams and a single server pool. The number of customer classes is equal to the number of arrival streams, hence $I = K$. Upon arrival, a type i customer joins queue i, so there is no routing decision to be made.

We make the following assumptions about the service and the arrival rates. For the arrival rates, we assume that (15) holds. For the number of servers, condition (20) is automatically satisfied. We assume (21) holds with $\rho^* = \sum_i \lambda_i/\mu_{1i} = 1$. Furthermore, we assume that

$$\sqrt{|N^r|}\left(\frac{\lambda_i^r}{|N^r|} - \lambda_i\right) \to b_i \quad \text{as } r \to \infty, \tag{114}$$

for $b_i \in \mathbb{R}$, which implies that

$$\lim_{r\to\infty} \sqrt{|N^r|}\left(1 - \sum_{i\in\mathcal{I}} \frac{\lambda_i^r}{|N^r|\mu_{1i}}\right) = \theta \tag{115}$$

for some $\theta \in \mathbb{R}$. Let the traffic intensity ρ^r be defined by

$$\rho^r = \sum_{i=1}^{I} \frac{\lambda_i^r}{|N^r|\mu_{1i}}.$$

Condition (115) implies that $\rho^r \to \rho^* = 1$ as $r \to \infty$. Clearly (25) has a unique solution with $x_{1i}^* = \lambda_i/\mu_{1i}$, $\rho^* = 1$, and $\{\rho^r\}$ satisfies Assumption 3.1.

Under an SBP policy, each classes is assigned a fixed priority. When a server needs to choose a new customer to serve, that server chooses the longest waiting customer in the highest priority nonempty class. We assume that there is no tie in the priorities and for simplicity that every class has priority over all the classes that have a lower index. The following result shows that the diffusion limits of all the queue length processes except the one with the lowest priority is zero.

THEOREM 7.1. *Let* $\{\mathbb{X}_{SBP}^r\}$ *be a sequence of SBP V-parallel server system processes. Assume that* (15), (21), *and* (114) *hold. Also assume that* (28) *holds and*

$$(\hat{Q}^r(0), \hat{Z}^r(0)) \;\Rightarrow\; (\hat{Q}(0), \hat{Z}(0)) \tag{116}$$

for a random vector $(\hat{Q}(0), \hat{Z}(0))$ *and*

$$\sum_{i=2}^{I} \hat{Q}_i^r(0) \to 0 \quad \textit{in probability} \tag{117}$$

as $r \to \infty$. *Then, for each* $T > 0$, *as* $r \to \infty$

$$\left\| \sum_{i=2}^{I} \hat{Q}_i^r(t) \right\|_T \to 0 \quad \textit{in probability}. \tag{118}$$

We provide a proof in §7.1.

REMARK 7.1. If (116) is replaced by the weaker condition

$$|\hat{Q}^r(0)| \vee |\hat{Z}^r(0)| \quad \text{satisfies the compact containment condition} \tag{119}$$

then (118) still holds because (**??**) is only used in the proof to show that (119) holds.

If assumptions (116) and (117) are replaced by (119), the following holds by Theorem 4.2; for any $T > 0$

$$\sup_{\tau^r \le t \le T} \left| \sum_{i=2}^{I} \hat{Q}_i^r(t) \right| \to 0 \quad \text{in probability}, \tag{120}$$

where $\{\tau^r\}$ is a sequence of real numbers with $\tau^r > 0$ and $\tau^r \to 0$ as $r \to \infty$. Note that this result is weaker than (118) because the SSC result does not hold initially but holds after time zero.

7.1. Establishing the SSC results for the SBP V-parallel server systems. In this section we present the steps involved in proving Theorem 7.1 and provide a proof at the end of the section. The proofs of the results presented in this section are placed in Appendix E. As discussed above, by (15), (21), and (114), Assumption 3.1 holds. To use Theorem 4.1, we verify Assumptions 3.2–4.2 hold below.

Next we show that under the SBP policy, Assumption 3.2 is satisfied.

PROPOSITION 7.1. *Let* $\{\mathbb{X}_{SBP}^r\}$ *be a sequence of SBP V-parallel server system processes. Assume that* (15), (21), *and* (114) *hold. Then* $\{\mathbb{X}_{SBP}^r\}$ *satisfies Assumption* 3.2.

Next, we present the SSC function and the additional hydrodynamic equations for the SBP V-parallel server systems. Let

$$g(q, z) = \sum_{i=2}^{I} |q_i|, \tag{121}$$

where $q, z \in \mathbb{R}^I$. Note that, $g(\hat{Q}^r(t), \hat{Z}^r(t)) = \sum_{i=2}^{I} |\hat{Q}_i^r(t)| = \sum_{i=2}^{I} \hat{Q}_i^r$, and hence g is the desired SSC function. Clearly g is continuous and satisfies Assumption 4.1 with $c = 1$.

We now verify that Assumption 4.2 holds. First we need to identify the additional equations satisfied the hydrodynamic limits of the SBP V-parallel server systems. For an SBP V-parallel server system, if there are customers waiting in a queue whose class has priority over another class, then all the customers in the lower priority class have to wait until those customers of higher priority class are served. This property is also preserved in the hydrodynamic limits as given by (122) in the following result.

PROPOSITION 7.2. *Let* $\{\mathbb{X}_{SBP}^r\}$ *be a sequence of SBP V-parallel server systems described as above that satisfies the conditions of Theorem* 7.1. *Fix* $L > 0$. *Then, in addition to* (30)–(37), *each hydrodynamic limit* $\tilde{\mathbb{X}}_{SBP}$ *of* $\{\mathbb{X}_{SBP}^r\}$ *on* $[0, L]$ *satisfies*

$$\dot{\tilde{A}}_{k1k}^{\ominus}(t) + \dot{\tilde{B}}_{1k}^{\ominus}(t) = 0 \qquad \textit{when} \qquad \tilde{Q}_{k+1}^{\oplus}(t) > 0, \tag{122}$$

for all $t \in [0, L]$ *and* $k \in \mathcal{K}$, *where*

$$\tilde{A}_{k1k}^{\ominus}(t) = \sum_{l=1}^{k} \tilde{A}_{l1l}(t), \qquad \tilde{B}_{1k}^{\ominus}(t) = \sum_{l=1}^{k} \tilde{B}_{1l}(t), \qquad \textit{and} \qquad \tilde{Q}_k^{\oplus}(t) = \sum_{j=k}^{I} \tilde{Q}_j(t).$$

We next show that Assumption 4.2 is satisfied by the hydrodynamic model of the SBP V-parallel server systems. Note that by Proposition 7.2, the hydrodynamic model equations consist of (30)–(37) and (122).

PROPOSITION 7.3. *Let $\{\mathbb{X}^r_{SBP}\}$ be a sequence of SBP V-parallel server system processes. Assume that (15), (21), and (114) hold. The hydrodynamic model of the SBP V-parallel server system satisfies Assumption 4.2 with $H(t) = (I - \mu_{11} z_{11} t)^+$.*

Now we are ready to prove a multiplicative SSC result for the SBP V-parallel server systems using Theorem 4.1.

THEOREM 7.2. *Let $\{\mathbb{X}^r_{SBP}\}$ be a sequence of SBP V-parallel server system processes. Assume that (15), (21), and (114) hold. Then, for each $T > 0$,*

$$\frac{\|\sum_{i=2}^{I} \hat{Q}^r_i(t)\|_T}{(\|\hat{Q}^r(t)\|_T \vee \|\hat{Z}^r(t)\|_T \vee 1)} \to 0 \quad \textit{in probability} \tag{123}$$

as $r \to \infty$.

PROOF. Let $\{\mathbb{X}^r_{\mathrm{SBP}}\}$ be a sequence of SBP V-parallel server system processes. Assume that (15), (21) with $\rho^* = 1$, and (114) hold.

It can easily be checked using (15) and (21) that Assumption 3.1 holds. Assumption 3.2 holds by Proposition 7.1. By definition, g given by (121) satisfies Assumption 4.1. Assumption 4.2 holds by Proposition 7.3. By virtue of Theorem 4.1, we have that (123) holds. □

Following Remark 4.2, we obtain Theorem 7.1 by virtue of the following result.

LEMMA 7.1. *Let $\{\mathbb{X}^r_{SBP}\}$ be a sequence of SBP V-parallel server system processes and assume that the conditions of Theorem 7.2 hold. Then, $\hat{Z}^r$ and $\hat{Q}^r$ satisfies the compact containment condition.*

8. Armony-Maglaras threshold policy. In this section we focus on the model studied in Armony and Maglaras [3]. Our purpose is to illustrate the extension of our main result presented in §6.2. In Armony and Maglaras [3], a V-model system has been used to study a contact center with two channels; one for real-time telephone service and another for a postponed call-back service offered with a guarantee on the maximum delay until a reply is received. We assume that the second customer class consists of those customers who call for the call-back option.

Armony and Maglaras [3] proposed the following policy.

Threshold Rule. If $Q_2(t) > \sqrt{|N^r|}\theta$, give priority to class 2, otherwise give priority to class 1.

Let

$$\lambda^r = |N^r|\mu\left(1 - \frac{\beta}{\sqrt{|N^r|}}\right). \tag{124}$$

We assume that the arrival rates for each customer class is given according to

$$\lambda^r_1 = \eta\lambda^r \qquad \text{and} \qquad \lambda^r_2 = (1-\eta)\lambda^r. \tag{125}$$

for some $\eta \in (0,1)$. Let

$$\hat{X}^r(t) = \hat{Q}^r_1(t) + \hat{Q}^r_2(t) + \hat{Z}^r_{11}(t) + \hat{Z}^r_{21}(t) \tag{126}$$

and assume that

$$(\hat{Q}^r(0), \hat{Z}^r(0)) \;\Rightarrow\; (\hat{Q}(0), \hat{Q}(0)), \tag{127}$$

as $r \to \infty$. By Theorem 2 in Halfin and Whitt [26], $\hat{X}^r$ converges weakly to a diffusion process X as $r \to \infty$.

We show that the following SSC result holds.

PROPOSITION 8.1. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (127), (124), and (125) hold and*

$$\big(\hat{Q}^r_1(0), \hat{Q}^r_2(0)\big) \;\Rightarrow\; \big((X(0)-\theta)^+, (X(0)^+ \wedge \theta)\big) \tag{128}$$

as $r \to \infty$. Then

$$\big(\hat{Q}^r_1(\cdot), \hat{Q}^r_2(\cdot)\big) \;\Rightarrow\; \big((X(\cdot)-\theta)^+, (X(\cdot)^+ \wedge \theta)\big)$$

as $r \to \infty$.

We provide a proof in §8.1. Proposition 8.1 was first presented in Armony and Maglaras [3]; see Proposition 3.1 there. The proof presented in Armony and Maglaras [3] contains a step that cannot be rigorously justified, see inequality (29) in that paper. In this section, we will present an alternative proof using Theorem 6.1. Using Proposition 8.1 one can prove the asymptotic optimality of the threshold policy; see Proposition 3.4 in Armony and Maglaras [3].

8.1. Establishing the state space collapse results for the V-systems under the threshold policy. In this section we prove Proposition 8.1 and illustrate the steps involved proving this result. The proofs of the results presented in this section are placed in Appendix F. We use Theorem 6.1 to prove Proposition 8.1. By (124) and (125) Assumption 3.1 holds. Below we verify Assumptions 3.2, 6.1, 6.2, and 6.3 hold.

First we show that Assumption 3.2 holds.

PROPOSITION 8.2. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (127), (124), and (125) hold. Then $\{\mathbb{X}^r\}$ satisfies Assumption 3.2.*

Next we define the SSC function for this setting. Let $q=(q_1,q_2)\in\mathbb{R}^2$, $z=(z_1,z_2)\in\mathbb{R}^2$, $x=q_1+q_2+z_1+z_2$, and $g\colon \mathbb{R}^4\to\mathbb{R}$ be defined by

$$g(q,z)=q_1-(x-\theta)^+ \tag{129}$$

Clearly $|g|$ is continuous but it does not satisfy Assumption 4.1 but satisfies Assumption 6.1. Therefore, we use Theorem 6.1.

Next, we show that Assumption 6.2 holds.

PROPOSITION 8.3. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (127), (124), and (125) hold. Then,*

$$\lim_{R\to\infty}\lim_{r\to\infty} P\{\|\hat{Q}^r(t)\|_T\vee\|\hat{Z}^r(t)\|_T>R\}=0, \tag{130}$$

i.e., $\{\mathbb{X}^r\}$ satisfies Assumption 1.

Next we present the additional hydrodynamic equations. For $R>0$ and $T>0$ and let $\mathcal{A}_R^r(T)$ be defined by

$$\mathcal{A}_R^r(T)=\{(\|\hat{Q}^r(\cdot)\|_T\vee\|\hat{Z}^r(\cdot)\|_T)\le R\}. \tag{131}$$

PROPOSITION 8.4. *Let $\{\mathbb{X}^r\}$ be a sequence of V-systems under the threshold policy described above that satisfies the conditions of Theorem 8.1. Fix $T>0$, $R>0$, and $L>0$. Then, in addition to (30)–(37), each hydrodynamic limit $\tilde{\mathbb{X}}$ of $\{\mathbb{X}^r\}$ on $\{\mathcal{A}_R^r(T)\}$ satisfies*

$$\dot{\tilde{B}}_{11}(t)=\mu \quad \textit{when} \quad \tilde{g}(R(\tilde{Q}(t),\tilde{Z}(t)))>0 \quad \textit{and} \quad \tilde{Q}_1(t)>0 \tag{132}$$

$$\dot{\tilde{B}}_{12}(t)=\mu \quad \textit{when} \quad \tilde{g}(R(\tilde{Q}(t),\tilde{Z}(t)))<0 \quad \textit{and} \quad \tilde{Q}_2(t)>0 \tag{133}$$

for $t\in[0,L]$.

Now we are ready to prove Proposition 8.1.

PROOF OF PROPOSITION 8.1. Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (124), (125), (127), and (128) holds.

From Propositions 8.2 and 8.3 and the definition of g in (129), to invoke Theorem 6.1, it is enough to show that Assumption 6.3 holds.

We prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big|g(\hat{Q}^r(t),\hat{Z}^r(t))\big|<0 \tag{134}$$

for every regular point t of $|g|$ whenever $|g(\hat{Q}^r(t),\hat{Z}^r(t))|>0$, which implies Assumption 6.3.

Let $\tilde{X}_i(t)=\tilde{Q}_i(t)+\tilde{Z}_{1i}(t)$ and $\tilde{X}(t)=\tilde{X}_1(t)+\tilde{X}_2(t)$. Then, by (33) and (33), $\tilde{X}_i(t)=\tilde{X}_i(0)$ for all $t\ge 0$, hence

$$\dot{\tilde{X}}(t)=0 \quad \text{for all } t\ge 0. \tag{135}$$

First assume that $g_1(R(\tilde{Q}(t),\tilde{Z}(t)))>0$. Then, by (132)

$$\dot{\tilde{Q}}_1(t)=\lambda_1-\mu=-(1-\eta)t.$$

Hence,

$$\dot{g}_1(R(\tilde{Q}(t),\tilde{Z}(t)))=R\dot{\tilde{Q}}_1(t)-\frac{\mathrm{d}}{\mathrm{d}t}(R\tilde{X}(t)-\theta)^+=-(1-\eta)t.$$

by (135).

Similarly, if $g_1(R(\tilde{Q}(t),\tilde{Z}(t)))<0$, then

$$\dot{g}_1\big(R(\tilde{Q}(t),\tilde{Z}(t))\big)=\dot{\tilde{Q}}_1(t)=\eta t.$$

This proves (134). □

Appendix A. Equivalence of the original and perturbed systems. This appendix is devoted to the proof of Theorem 2.1. The proof will be presented at the end of this section. In Pang et al. [39, Lemma 2.1], the authors show a similar result for $M/M/n+M$ systems. Our setting is significantly more complicated because arrivals are general, and scheduling decisions affect the system evolution.

For notational simplicity we focus on systems with no routing and abandonment. We assume that a type i customer will be automatically routed to queue i at the time of his arrival. Therefore, $I = K$, and we omit the subscript k from the notation. For the rest of this section we fix a policy π. Recall that each policy is associated with a transaction function f_π. For simplicity, we assume that f_π is a deterministic function, but it can be taken as a random variable that only depends on the state of the system at the decision instant.

To prove the equivalence of the original system with the perturbed system we model both systems as piecewise-deterministic Markov processes (PDMP's) that are introduced by Davis [18]. In §A.1 we give a brief overview of PDMP's. In §A.2, we construct a PDMP for our parallel server system. In §A.3, we construct a PDMP for our perturbed system and complete the proof of Theorem 2.1.

A.1. Piecewise-deterministic Markov processes (PDMP) and parallel server systems. A thorough treatment of the subject and examples of how to model $M/G/1$ and $G/G/1$ queues as PDMPs can be found in Davis [18].

For our purposes, it is enough to define a PDMP on a state space $E \subset \mathbb{R}^p$ that is closed in $\mathbb{R}^p$ for some positive integer p. A portion of the state space, denoted by E_δ, is designated as the topological boundary. Then, $E_o = E \backslash E_\delta$ is the "interior" of E. We let $\mathscr{E}$ denote the Borel subsets of E, and we will let $P(E)$ be the space of probability measures on the measurable space $(E, \mathscr{E})$; the space $P(E)$ is endowed with the topology of weak convergence. Under suitable regularity conditions a PDMP can be uniquely determined by a function $h\colon E \to \mathbb{R}^p$, an intensity function $\gamma\colon E \to \mathbb{R}_+$, and a transition measure $\varpi\colon E \to P(E)$. We assume that $\varpi(E_o \mid x) = 1$ for each $x \in E$.

By convention, each sample path $\{x(t),\ t \geq 0\}$ of a PDMP is right continuous on $[0, \infty)$ and has left limits in $(0, \infty)$. For each time $t \geq 0$, the state $x(t)$ always lives in the "interior" E_o of the state space E. For $t > 0$, we use $x(t-)$ to denote the left limit at t; namely, $x(t-) = \lim_{s \uparrow t} x(s) \in E_\delta$. It is possible that $x(t-)$ goes outside of the "interior." When $x(t-) \in E_\delta$, a jump occurs at time t, moving the state instantaneously into the "interior." While $x(t)$ is in the "interior," it can also make jumps. Such a jump is governed by an exponential clock with rate $\gamma(x)$ when $x(t) = x$, independently of the process history. Between jumps $x(t)$ obeys $dx(t)/dt = h(x(t))$. If a jump occurs at time t with either $x(t) = x \in E_o$ or $x(t-) = x \in E_\delta$, the process is transferred immediately to a new state in E_o that is randomly chosen following probability measure $\varpi(dx \mid x)$. We use σ_n to denote the nth jump time of the PDMP process $x(t)$. Let $N(t) = \sum_{i=1}^{\infty} I_{\sigma_i \leq t}$. Under the assumption that

$$\mathbb{E}[N(t)] < \infty \quad \text{for all } t, \tag{A1}$$

it can be shown that $\{x(t), t \geq 0\}$ is a strong Markov process, see Davis [18]

For parallel server systems, function h will be used to model the fact that once an interarrival time is generated the remaining interarrival time will decrease linearly at rate 1 until it reaches zero. Once it reaches zero a new interarrival time is generated. This will be modeled as defining the boundary E_δ and the transition measure ϖ appropriately. The intensity function γ defines the service rate at each instant. The transition measure ϖ will be defined to govern the behavior of the system when a new customers arrives to the system or a service is completed. When a new customer arrives, a new interarrival time for that class is generated, and the system's state is updated to this new state according to the interarrival time distribution. If there are idle servers in the system, the arriving customer may be assigned to one of these servers according to the scheduling policy. When a service is completed by a server, the server starts a new service from a customer waiting in queue (if there are any) according to the scheduling policy or stays idle. Both types of jumps are modeled by defining ϖ properly.

A.2. Construction of parallel server systems. In this section we construct the processes associated with a parallel server system. Recall that arrivals to class i are given by a delayed renewal process A_i (see §2). For notational simplicity we assume A_i is a renewal process and we assume that the interarrival times of A_i are given by the i.i.d. sequence $\{u_i(m)\colon m = 1, 2, \dots\}$ for each $i \in \mathcal{I}$. We also assume that $\mathbb{P}\{u_i(1) > 0\} = 1$ and that the probability of having two or more simultaneous arrivals is zero. The latter assumption holds, e.g., when each $u_i(1)$ has a density.

Let $X(t) = (Q(t), Z(t), b(t))$ denote the state of the system at time t, where Q and Z have the same interpretations as before (see §2) and $b(t) = (b_1(t), \dots, b_I(t))$ with $b_i(t)$ is the remaining time before the next

class i customer will arrive at time t. Although we appended "$\prime$" in §2 to processes associated with the original system, we ignore it in this section for notational simplicity.

Note that $Q(t) \in \mathbb{N}_+^I$, $Z(t) \in \mathbb{N}^{I\times J}$, and $b(t) \in \mathbb{R}_+^I$. Hence $X(t) \in \mathbb{N}_+^I \times \mathbb{N}^{I\times J} \times \mathbb{R}_+^I$. In terms of the PDMP characterization in §A.1, we define the state space E for X by $E = \mathbb{R}^I \times \mathbb{R}^{I\times J} \times \mathbb{R}_+^I$, with the boundary $E_\delta = \mathbb{R}^I \times \mathbb{R}^{I\times J} \times \partial\mathbb{R}_+^I$, where

$$\partial\mathbb{R}_+^I = \{x = (x_1, \ldots, x_I) \in \mathbb{R}_+^I \colon x_i = 0 \text{ for some } i = 1, \ldots, I\}.$$

Let $E_o = E \backslash E_\delta$. It is clear that $E_o = \mathbb{R}^I \times \mathbb{R}^{I\times J} \times \mathbb{R}_{++}^I$, where

$$\mathbb{R}_{++}^I = \{x \in \mathbb{R}^I \colon x_i > 0 \text{ for } i = 1, \ldots, I\}.$$

We assume for simplicity that $X(0) \in E_o$, which means $b_i(0) > 0$ for $i = 1, \ldots, I$. We also assume $b_i(0) \neq b_\ell(0)$ for $i \neq \ell$. Recall that, we use $\{\sigma_n\}$ to denote the increasing sequence of event times, either an arrival to or departure from the system.

Next we explain how we can model a parallel server system as a PDMP. Between jump points

$$dQ(t)/dt = dZ(t)/dt = 0 \qquad \text{and} \qquad db(t)/dt = -e, \tag{A2}$$

where e is a I-dimensional vector of ones. In terms of the PDMP characterization in §A.1, the function $h\colon E \to \mathbb{R}^I \times \mathbb{R}^{I\times J} \times \mathbb{R}^I$ is given by $h = (h_1, h_2)$ with $h_1\colon E \to \mathbb{R}^I \times \mathbb{R}^{I\times J}$, and $h_2\colon E \to \mathbb{R}^I$, where

$$h_1(x) = 0 \qquad \text{and} \qquad h_2(x) = -e, \tag{A3}$$

for any $x \in E$.

The boundary E_δ is reached when one of the remaining interarrival times $b_i(t)$ reaches zero. At that instant, an arrival to class i occurs. A new interarrival time, u_i, for class i is generated following interarrival distribution F_i. At this time t, b_i jumps with $b_i(t) = u_i$ and the other components of b do not change at time t. Also, Q and Z are updated at time time according to the scheduling policy, in our setting according to function f_π (see §2 for a definition), as explained next.

Because service times are assumed to be exponentially distributed, the total service rate at any time is equal to summation of service rates of all customers in service at that instant. Therefore, for any $x = (q, z, b) \in E_o$, the intensity function is given by

$$\gamma(x) = \sum_{k\in\mathcal{K},\, j\in\mathcal{J}(k)} z_{jk}\mu_{jk}. \tag{A4}$$

To complete the formal definition of the PDMP for the parallel server system, we need to specify the transition measure ϖ. Recall that the transition measure must be defined in two circumstances; (i) when the system reaches the boundary at t, i.e., $X(t-) \in E_\delta$, (ii) when a jump occurs at t in the interior, i.e., $X(t) \in E_o$.

First, we focus on the case when the system reaches the boundary at time t. Denote $x = (q, z, b) = X(t-)$. It is necessarily true that $x \in E_\delta$ or equivalently one of components of b is zero. Assume $b_i = 0$ for some i. We define the probability measure ϖ on $(\mathbb{N}^I \times \mathbb{N}^{J\times K} \times \mathbb{R}_+^K) \times (\mathbb{N}^K \times \mathbb{N}^{I\times J} \times \mathbb{R}_+^K)$ as follows:

$$\varpi(B_1 \times B_2 \mid x) = 1_{f_\pi(q,z,e_i)}(B_1)\mu_{F_i}(B_2), \tag{A5}$$

where $B_1 \subset \mathbb{N}^I \times \mathbb{N}^{I\times J}$, $B_2 \in \mathcal{B}(\mathbb{R}^I)$, e_i denotes the event that a customer arrived to class i, and μ_{F_i} is the measure associated with interarrival distribution F_i. The distribution given in (A5) specifies the behavior of the system when there is an arrival to class i.

We next specify the transition measure when a jump occurs at t in the interior. Let $x = (q, z, b) = X(t)$. This time $\varpi(\cdot \mid x)$ is a discrete distribution: at (q^{ji}, z^{ji}, b), it has mass

$$\frac{z_{ji}\mu_{ji}}{\sum_{k\in\mathcal{I},\, \ell\in\mathcal{J}(k)} z_{\ell k}\mu_{\ell k}} \tag{A6}$$

where $(q^{ji}, z^{ji}) = f_\pi(q, z, e_{ji})$, where e_{ji} denotes the event that a server in pool j finishes serving a class i customer. The right-hand side of (A6) gives the probability that the service of a class i customer is completed in server pool j, given that there is a service completion in the system. In addition, once a service is completed, the remaining service time of each customer who is still in service can be generated according to that customer's service time distribution. Because service times are exponentially distributed, it is possible to define a parallel server system this way.

It is clear that $X(t) = (Q(t), Z(t), b(t))$ is a PDMP with intensity function γ, transition measure ϖ, and evolution Equation (A2). Also, because $\mathbb{E}[A_k(t)] < \infty$ for all $t \geq 0$ and remaining service times at event times have exponential distribution, $\mathbb{E}[N(t)] < \infty$ for all $t \geq 0$. Therefore, (A1) is satisfied and $\{X(t), t \geq 0\}$ is a regular strong PDMP.

A.3. Construction of perturbed systems. Next, we focus on our perturbed system and show that it has the same intensity function, evolution equation, and the transition measure with the parallel server system described in the previous section. We again fix an admissible policy π.

Note that in the perturbed system, the total service rate is equal to the number of customers in service times their service rate. Also, service times are still exponential, hence, the remaining service times are also exponential after an arrival or a service completion. Therefore, the service rate is given by (A4). In addition, (A5) still holds because arrivals are governed by the same process as in the parallel server systems. Also, (A6) still holds because service times are exponential.

Proof of Theorem 2.1. Note that the perturbed system has the same PDMP characterization as the original system; that is, they have the same transition measure, intensity function, and evolution equation. Also, by an argument similar to that at the end of the last section, $\mathbb{E}[N(t)] < \infty$ for all $t \geq 0$ for the perturb system, too. Hence, in Davis [18, Theorem 5.5] both Markov processes, the parallel server system and the perturb system have the same generator. Therefore, they have the same finite-dimensional distributions in Ethier and Kurtz [19, Proposition 1.6, Chapter 4]. □

Appendix B. Fluid limits. In this section we study the fluid limits and present the fluid model equations of parallel server systems. Also, we establish a general framework that can be used to check whether Assumption 3.2 is satisfied by a control policy.

Let $\mathcal{A} \subset \Omega$ be such that $\{\bar{Q}^r(0)\}$ is bounded and the following FSLLN holds:

$$\frac{E(|N^r|\cdot)}{|N^r|} \to \nu(\cdot), \qquad \frac{S_{jk}(|N^r|\cdot)}{|N^r|} \to \alpha_{jk}(\cdot), \qquad \text{and} \qquad \frac{F_k(|N^r|\cdot)}{|N^r|} \to \Gamma_k(\cdot) \quad \text{u.o.c.} \tag{B1}$$

as $r \to \infty$, where $\alpha_{jk}(t) = \mu_{jk}t$, $\Gamma_k(t) = \gamma_k t$, for all $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$, and $\nu(t) = te$, where e is the I-dimensional row vector of ones. Note that we can take $P(\mathcal{A}) = 1$ from (15). For the rest of paper we only consider sample paths in $\mathcal{A}$.

Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes and

$$\bar{\mathbb{X}}^r_\pi(t) = \mathbb{X}^r_\pi(t)/|N^r|. \tag{B2}$$

We call this scaling the fluid scaling and $\bar{\mathbb{X}}^r_\pi$ the fluid scaled process. $\bar{\mathbb{X}}_\pi$ is called a fluid limit of $\{X^r_\pi\}$ if there exists an $\omega \in \mathcal{A}$ and a sequence $\{r_l\}$ with $r_l \to \infty$ as $l \to \infty$, such that $\bar{\mathbb{X}}^{r_l}_\pi(\cdot, \omega)$ converges u.o.c. to $\bar{\mathbb{X}}_\pi$ as $l \to \infty$. The following theorem is analogous to Dai [13, Theorem 4.1]. Its proof is given at the end of this section.

Theorem B.1. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes. Assume that (20) and (21) hold and $\{\bar{Q}^r(0)\}$ is bounded a.s. as $r \to \infty$. Then, $\{\bar{\mathbb{X}}^r_\pi\}$ is a.s. precompact (i.e., every subsequence has a convergent subsequence) in the Skorohod space $\mathbb{D}^d[0,\infty)$ endowed with the u.o.c. topology. Thus, the fluid limits exist, and each fluid limit, $\bar{\mathbb{X}}_\pi$, of $\{\bar{\mathbb{X}}^r_\pi\}$ satisfies the following equations for all $t \geq 0$:*

$$\lambda_i t = \sum_{k\in\mathcal{K}} \bar{A}_{ik}(t) + \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(k)} \bar{A}_{ijk}(t), \quad \text{for all } i \in \mathcal{I}, \tag{B3}$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \sum_{i\in\mathcal{I}} \bar{A}_{ik}(t) - \sum_{j\in\mathcal{J}(k)} \bar{B}_{jk}(t) - \gamma_k \int_0^t \bar{Q}_k(s)\,\mathrm{d}s, \quad \text{for all } k \in \mathcal{K}, \tag{B4}$$

$$\bar{Z}_{jk}(t) = \bar{Z}_{jk}(0) + \sum_{i\in\mathcal{I}} \bar{A}_{ijk}(t) + \bar{B}_{jk}(t) - \mu_{jk}\bar{T}_{jk}(t), \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{B5}$$

$$\bar{T}_{jk}(t) = \int_0^t \bar{Z}_{jk}(s)\,\mathrm{d}s, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j), \tag{B6}$$

$$\bar{Y}_j(t) = \beta_j t - \sum_{k\in\mathcal{K}(j)} \bar{T}_{jk}(t), \quad \text{for all } j \in \mathcal{J}, \tag{B7}$$

$$\bar{Q}_k(t)\left(\sum_{j\in\mathcal{J}(k)} \left(\beta_j - \sum_{l\in\mathcal{K}(j)} \bar{Z}_{jl}(t)\right)\right) = 0, \quad \text{for all } k \in \mathcal{K}, \tag{B8}$$

$$\int_0^t \sum_{k\in\mathcal{K}(j)} \bar{Q}_k(s)\,\mathrm{d}\bar{Y}_j(s) = 0, \quad \text{for all } j \in \mathcal{J}, \tag{B9}$$

$$\int_0^t \sum_{j\in\mathcal{J}(k)} \left(\beta_j - \sum_{l\in\mathcal{K}(j)} \bar{Z}_{jl}(s)\right) \mathrm{d}\bar{A}_{ik}(s) = 0, \quad \text{for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K}, \tag{B10}$$

$$\bar{A}, \bar{A}_q, \bar{A}_s, \bar{T}, \text{ and } \bar{B} \text{ are nondecreasing}, \tag{B11}$$

$$\bar{Q}(t) \geq 0, \bar{Z}_{jk}(t) \geq 0, \qquad \text{and} \qquad \sum_{k \in \mathcal{K}(j)} \bar{Z}_{jk}(t) \leq \beta_j, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j). \tag{B12}$$

Definition B.1. We call the vector (q, z) a steady state of the fluid limits if for any fluid limit $\bar{\mathbb{X}}_\pi$, $\bar{Q}(0) = q$ and $\bar{Z}(0) = z$ implies $\bar{Q}(t) = q$ and $\bar{Z}(t) = z$ for all $t > 0$.

We denote the set of all the steady states of the fluid limits of $\{\mathbb{X}^r_\pi\}$ by $\mathcal{M}_\pi$. The following result presents a condition that is equivalent to Assumption 3.2.

Lemma B.1. *Let $\{\mathbb{X}^r_\pi\}$ be a sequence of π-parallel server system processes that satisfy the conditions of Theorem B.1 and Assumption 3.1. A control policy π satisfies Assumption 3.2 if $(0, z) \in \mathcal{M}_\pi$, where $z_{jk} = \beta_j x^*_{jk}$ and x^* is given as in Assumption 3.1.*

Proof. Assume that $(Q^r(0)/|N^r|, Z^r(0)/|N^r|) \to (0, z)$ a.s. as $r \to \infty$ and $(0, z) \in \mathcal{M}_\pi$. We prove the result by contradiction. Assume that Assumption 3.2 does not hold. Then we can find $\omega \in \mathcal{A}$ and a subsequence, denoted again by r for notational simplicity, such that $\bar{\mathbb{X}}^r_\pi$ is convergent. Because every fluid limit satisfies the fluid model equations this implies $\bar{Q}^r(\cdot) \to 0$ and $\bar{Z}^r(\cdot) \to z$ as $r \to \infty$ u.o.c., which contradicts our initial assumption. □

Proof of Theorem B.1. Assume that (20) and (21) hold. Consider a sequence of numbers that we again denote, with a slight abuse of notation, by $\{r\}$. We show that $\{\bar{\mathbb{X}}^r(\cdot, \omega)\}$ has a convergent subsequence, for all $\omega \in \mathcal{A}$. We fix $\omega \in \mathcal{A}$ for the rest of the proof.

Observe that

$$\left| \frac{T^r(t_2, \omega)}{|N^r|} - \frac{T^r(t_1, \omega)}{|N^r|} \right| \leq |t_2 - t_1|,$$

for all $0 \leq t_1 < t_2$. Hence, $\{\bar{T}^r(\cdot, w)\}$ is tight; there exists a subsequence $\{r_l\}$ such that $\bar{T}^{r_l}(\cdot, \omega)$ converges u.o.c. to some continuous function $\bar{T}$. We define the fluid scaled idle time process $\bar{Y}^r_j$ for the jth server by

$$\bar{Y}^r_j(t) = \frac{N^r_j}{|N^r|} t - \sum_{k \in \mathcal{K}(j)} \bar{T}^r_{jk}(t) \tag{B13}$$

for all $t \geq 0$ and set $\bar{Y}^r = (\bar{Y}^r_1, \ldots, \bar{Y}^r_j)$.

Now, because $\bar{D}^{r_l}_{jk}(t) = (1/N^r) S_{jk}(N^r \bar{T}^{r_l}_{jk}(t))$,

$$\bar{D}^{r_l}_{jk}(\cdot) \text{ converges u.o.c. to } \bar{D}_{jk}(\cdot), \tag{B14}$$

where $\bar{D}_{jk}(t) = \mu_{jk} \bar{T}(t)$ in Ata and Kumar [4, Lemma 11]. By (B1), $\bar{A}^{r_l}(\cdot, \omega)$ converges u.o.c. to $\bar{A}(t) = \lambda t$; hence, it is precompact in $D^I[0, \infty)$. However, for all $0 \leq t_1 < t_2$ and $j \in \mathcal{J}$,

$$\bar{A}^{r_l}_{ik}(t_2, \omega) - \bar{A}^{r_l}_{ik}(t_1, \omega) \leq \bar{A}^{r_l}(t_2, \omega) - \bar{A}^{r_l}(t_1, \omega).$$

By Billingsley [8, Theorem 12.3], this implies that $\{\bar{A}^{r_l}_{ik}(\cdot, \omega)\}$ is also precompact. By the same argument, so is $\{\bar{A}^{r_l}_{ijk}(\cdot, \omega)\}$, for all $i \in \mathcal{I}$, $k \in \mathcal{K}$, and $j \in \mathcal{J}(k)$.

Fix $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$, and $0 \leq t_1 \leq t_2$. We omit ω from the notation below. Note that $B^{r_l}_{jk}$ can only increase when a service in pool j is completed. On the other hand, some of the servers in pool j will receive customers right after they arrive at the system. Hence, by (8), we have that

$$\bar{B}^{r_l}_{jk}(t_2) - \bar{B}^{r_l}_{jk}(t_1) \leq \sum_{k' \in \mathcal{K}(j)} (\bar{D}^{r_l}_{jk'}(t_2) - \bar{D}^{r_l}_{jk'}(t_1)). \tag{B15}$$

Because $\bar{B}^{r_l}_{jk}$ is nondecreasing, by (B14) and again by Billingsley [8, Theorem 12.3], (B15) implies that $\{\bar{B}^{r_l}_{jk}(\cdot, \omega)\}$ is precompact. This implies by (8) and (B14) that $\{\bar{Z}^{r_l}_{jk}(\cdot, \omega)\}$ is precompact.

Assume without loss of generality that

$$\begin{aligned} &\big(\bar{A}^{r_l}(\cdot), \bar{A}^{r_l}_s(\cdot), \bar{A}^{r_l}_q(\cdot), \bar{D}^{r_l}(\cdot), \bar{T}^{r_l}(\cdot), \bar{Y}^{r_l}(\cdot), \bar{B}^{r_l}, \bar{Z}^{r_l}(\cdot)\big) \\ &\quad \to \big(\bar{A}(\cdot), \bar{A}_s(\cdot), \bar{A}_q(\cdot), \bar{D}(\cdot), \bar{T}(\cdot), \bar{Y}(\cdot), \bar{B}, \bar{Z}(\cdot)\big) \end{aligned} \tag{B16}$$

u.o.c. and

$$\bar{Q}^{r_l}(0) \to \bar{Q}(0) \tag{B17}$$

as $l \to \infty$. We next show that $\{\bar{Q}^{r_l}(\cdot)\}$ is precompact. Fix $T > 0$ and choose r_l large enough so that

$$\|\bar{A}^{r_l}(\cdot)\|_T < M$$

for $r > r_l$ and $M < \infty$. Note that such M exists by (B16). Then, for $0 \le s \le t \le T$

$$\int_0^t \bar{Q}_k^{r_l}(u)\,du - \int_0^s \bar{Q}_k^{r_l}(u)\,du \le (t-s)M.$$

for all $k \in \mathcal{K}$. Hence, the sequence of processes

$$\left\{\int_0^{\cdot} \bar{Q}_k^{r}(u)\,du\right\} \tag{B18}$$

is precompact for all $k \in \mathcal{K}$. This, by (B1) and Ata and Kumar [4, Lemma 11], implies that the sequence

$$\left\{\frac{F_k(|N^{r_l}|\int_0^{\cdot} \bar{Q}_k^{r_l}(s)\,\mathrm{d}s)}{|N^{r_l}|}\right\} \tag{B19}$$

is precompact for all $k \in \mathcal{K}$. By (B16)–(B19) and (7), we conclude that

$$\{\bar{Q}^{r_l}(\cdot)\}$$

is precompact.

Next, we show that every fluid limit satisfies (B3)–(B10). Let $\bar{\mathbb{X}}$ be a fluid limit and for notational convenience assume that

$$\bar{\mathbb{X}}^r(\cdot,\omega) \to \mathbb{X}(\cdot) \tag{B20}$$

u.o.c. as $r \to \infty$ for some $\omega \in \mathcal{A}$. $\bar{\mathbb{X}}$ satisfies (B3) by (6), the convergence of $\bar{A}_i^r(\cdot,\omega)$ to $\bar{A}_i(t) = \lambda_i t$ and the fact that $\bar{A}_{ik}^r(\cdot,\omega)$ and $\bar{A}_{ijk}^r(\cdot,\omega)$ are both convergent.

Note that because $\bar{Q}^r(\cdot) \to \bar{Q}(\cdot)$ u.o.c. as $r \to \infty$,

$$\int_0^{\cdot} \bar{Q}_k^{r_l}(s)\,\mathrm{d}s \to \int_0^{\cdot} \bar{Q}_k(s)\,\mathrm{d}s \quad \text{u.o.c. as } r \to \infty$$

by Lemma 11 in [4] and so by (B1) and again Lemma 11 in [4]

$$\frac{F_k(|N^{r_l}|\int_0^{\cdot} \bar{Q}_k^{r_l}(s)\,\mathrm{d}s)}{|N^{r_l}|} \to \gamma \int_0^{\cdot} \bar{Q}_k(s)\,\mathrm{d}s \quad \text{u.o.c. as } r \to \infty.$$

Therefore, fluid limits satisfy (B4) by (7) and (B20). Equation (B5) follows from (B14), the convergence of $\bar{Z}_{jk}^r(0,\omega)$, $\bar{A}_{ijk}^r(\cdot,\omega)$, and $\bar{B}_{jk}^r(\cdot,\omega)$. Equation (B6) follows from (10) and the convergence of $\bar{Z}_{jk}^r(\cdot,\omega)$ to $\bar{Z}_{jk}(\cdot)$ u.o.c.

We next show that $\bar{\mathbb{X}}$ satisfies (B8). Fix $t > 0$. If $\bar{Q}_k(t) = 0$, then (B8) is satisfied trivially, so assume that $\bar{Q}_k(t) > 0$. By the continuity of $\bar{Q}_k$, there exist $t > \delta > 0$ and $\varepsilon > 0$ such that $\bar{Q}_k(s) > \varepsilon$ for all $s \in [t-\delta, t+\delta]$. Because $\bar{Q}_k^r$ converges u.o.c. to $\bar{Q}_k$, for large enough r

$$\bar{Q}_k^r(s,\omega) > \varepsilon/4 \quad \text{for all } s \in [t-\delta, t+\delta].$$

By (11), this gives

$$\sum_{j\in\mathcal{J}(k)} \left(\frac{N_j^r}{|N^r|} - \sum_{l\in\mathcal{K}(j)} \bar{Z}_{jl}^r(s,\omega)\right) = 0 \quad \text{for all } s \in [t-\delta, t+\delta].$$

Using the fact that $\bar{\mathbb{X}}^r$ converges u.o.c. to $\bar{\mathbb{X}}$ again, we have that

$$\sum_{j\in\mathcal{J}(k)} \left(\beta_j - \sum_{l\in\mathcal{K}(j)} \bar{Z}_{jl}(s)\right) = 0 \quad \text{for all } s \in [t-\delta, t+\delta],$$

thus proving (B8). It can be shown similarly that $\bar{\mathbb{X}}$ satisfies (B9) and (B10). □

Remark B.1. It follows from (B3)–(B6) and the proof of Theorem B.1 that each component of $\bar{\mathbb{X}}$ is Lipschitz continuous, and so they are absolutely continuous and differentiable almost everywhere with respect to the Lebesgue measure on $[0,\infty)$.

Appendix C. Proofs of the results in §5

C.1. Bounds on hydrodynamically scaled processes. We need the following lemma to prove that the departure process of a hydrodynamic limit satisfies the associated hydrodynamic model equation and the number of customers abandoning from the queues are negligible in the hydrodynamic limits. Recall that we denote by $\mathcal{A}$ the subset of Ω whose elements satisfy (B1) and $P(\mathcal{A}) = 1$.

LEMMA C.1. *Let $\{\mathbb{X}^r\}$ be a sequence of π-parallel server system processes. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. Fix $\varepsilon > 0$, $L > 0$, and $T > 0$. Then, for large enough r and $\omega \in \mathcal{A}$*

$$\max_{m<\sqrt{|N^r|}T} \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^L |Z_{jk}^{r,m}(s)|\,\mathrm{d}s < \varepsilon, \quad \forall\, j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j).$$

and

$$\max_{m<\sqrt{|N^r|}T} \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^L |Q_k^{r,m}(s)|\,\mathrm{d}s < \varepsilon, \quad \forall\, k \in \mathcal{K}.$$

PROOF. Fix $\omega \in \mathcal{A}$ and $\varepsilon > 0$. Let z be given as in Assumption 3.2. For $m < \sqrt{|N^r|}T$, by (60),

$$\frac{\sqrt{x_{r,m}}}{|N^r|} \le \left\| \frac{Q^r(t)}{|N^r|} \right\|_T \vee \left\| \frac{Z^r(t)}{|N^r|} - \frac{\vec{N}^r}{|N^r|} x^* \right\|_T \vee \frac{1}{\sqrt{|N^r|}}.$$

Because π is assumed to satisfy Assumption 3.2, Lemma B.1 implies $\limsup_{r\to\infty} \|Q^r(t)/|N^r|\|_T \vee \|Z^r(t)/|N^r| - \vec{N}^r/|N^r|x^*\|_T = 0$. Hence, for r large enough and $m < \sqrt{|N^r|}T$,

$$\frac{\sqrt{x_{r,m}}}{|N^r|} \le \varepsilon. \tag{C1}$$

Similarly, for r large enough,

$$\left\| \frac{Q^r(t)}{|N^r|} \right\|_{L\varepsilon+T} \vee \left\| \frac{Z^r(t)}{|N^r|} - \frac{\vec{N}^r}{|N^r|} x^* \right\|_{L\varepsilon+T} < \frac{\varepsilon}{L}. \tag{C2}$$

Choose r large enough so that (C1) and (C2) hold. Then, for such r and for $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$

$$\max_{m<\sqrt{|N^r|}T} \frac{1}{|N^r|} \int_0^L \left| Z_{jk}^r\left(\frac{\sqrt{x_{r,m}}}{|N^r|} s + \frac{m}{\sqrt{|N^r|}} \right) - x_{jk}^* N_j^r \right| \mathrm{d}s \le L \left\| \frac{Z_{jk}^r(t)}{|N^r|} - x_{jk}^* \frac{N_j^r}{|N^r|} \right\|_{L\varepsilon+T} < \varepsilon$$

and

$$\max_{m<\sqrt{|N^r|}T} \frac{1}{|N^r|} \int_0^L \left| Q_k^r\left(\frac{\sqrt{x_{r,m}}}{|N^r|} s + \frac{m}{\sqrt{|N^r|}} \right) \right| \mathrm{d}s \le L \left\| \frac{Q^r(t)}{|N^r|} \right\|_{L\varepsilon+T} < \varepsilon. \quad \square$$

REMARK C.1. Let $\{\epsilon(r)\}$ be a sequence with $\epsilon(r) \to 0$ as $r \to \infty$. Define

$$\Theta^r = \left\{ \max_{m<\sqrt{|N^r|}T} \left(\frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^T |Z^{r,m}(s)|\,\mathrm{d}s \vee \frac{\sqrt{x_{r,m}}}{|N^r|} \int_0^T |Q^{r,m}(s)|\,\mathrm{d}s \right) < \epsilon(r) \right\}. \tag{C3}$$

For $\epsilon(r) \to 0$ slowly enough as $r \to \infty$, $\lim_{r\to\infty} P(\Theta^r) = 1$, by Lemma C.1. Hence,

$$\lim_{r\to\infty} P(\Theta^r \cap \mathcal{X}^r) = 1, \tag{C4}$$

by Corollary 5.1, where $\mathcal{X}^r$ defined as in §5.1. With a slight abuse of notation, we set $\mathcal{X}^r = \Theta^r \cap \mathcal{X}^r$ for simplicity.

C.2. Proofs of the results in §5.1

C.2.1. Proof of Proposition 5.1. We observe as in Bramson [10] that it is enough to investigate the processes with index $m = 0$ and then to multiply the ensuing error bounds by the number of processes in each case; $\sqrt{|N^r|}T$. To see this, note that

$$A_i^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(\chi_i\left(\frac{\lambda_i^r}{|N^r|}\left(\sqrt{x_{r,m}}t + \sqrt{|N^r|}m\right)\right) - \chi_i\left(\frac{\lambda_i^r}{|N^r|}\sqrt{|N^r|}m\right)\right).$$

Let $u_i^{r,m}(1)$ be the first residual interarrival time of χ_i after time $\lambda_i^r/|N^r|\sqrt{|N^r|}m$.

$$\begin{aligned}
&P\left\{\max_{m<\sqrt{|N^r|}T}\left\|A_i^{r,m}(t) - \frac{\lambda_i^r}{|N^r|}t\right\|_L > 2\epsilon L\right\}\\
&= P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\frac{1}{\sqrt{x_{r,m}}}\left(\chi_i\left(t+\sqrt{|N^r|}m\frac{\lambda_i^r}{|N^r|}\right) - \chi_i\left(\frac{\lambda_i^r}{|N^r|}\sqrt{|N^r|}m\right)\right) - \frac{t}{\sqrt{x_{r,m}}}\right\|_{L_{\sqrt{x_{r,m}}\lambda_i^r/|N^r|}} > 2\epsilon L\right\}\\
&\le P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\chi_i\left(t+\sqrt{|N^r|}m\frac{\lambda_i^r}{|N^r|}\right) - \chi_i v\left(\frac{\lambda_i^r}{|N^r|}\sqrt{|N^r|}m\right) - (t-u_i^{r,m}(1))\right\|_{L_{\sqrt{x_{r,m}}\lambda_i^r/|N^r|}} > \sqrt{x_{r,m}}\epsilon L\right\}\\
&\quad + P\left\{\max_{m<\sqrt{|N^r|}T}\frac{u_i^{r,m}(1)}{\sqrt{x_{r,m}}} > \epsilon L\right\}. \qquad \text{(C5)}
\end{aligned}$$

To show that is enough to focus on $m = 0$ first we claim that the second term goes to zero as $r \to \infty$. This follows by Lemma C.2, because it is enough to show that $\{u_i^{r,m}(1),\ i \in \mathcal{I}\}$ satisfies

$$u_i^{r,m}(1)\big/\sqrt{|N^r|} \to 0 \quad \text{in probability as } r \to \infty, \text{ for all } i \in \mathcal{I} \qquad \text{(C6)}$$

for all $m < \sqrt{|N^r|}T$. Note that

$$\chi_i\left(t+\sqrt{|N^r|}m\frac{\lambda_i^r}{|N^r|}\right) - \chi_i\left(\frac{\lambda_i^r}{|N^r|}\sqrt{|N^r|}m\right) = \chi_i\left(t+\sqrt{|N^r|}m\frac{\lambda_i^r}{|N^r|}\right) - \chi_i\left(\frac{\lambda_i^r}{|N^r|}\sqrt{|N^r|}m + (t-u_i^{r,m}(1))\right)$$

Also, by definition of A_i^r (see (18)) and because π is an admissible policy, $\chi_i(t+\sqrt{|N^r|}m\lambda_i^r/|N^r|) - \chi_i(\lambda_i^r/|N^r|\sqrt{|N^r|}m)$ is independent of $x_{r,m}$. Hence, for a random variable $\tilde{x}_{r,m}$ which has the same distribution as $x_{r,m}$ and $\tilde{u}_i^{r,m}$, which has the same distribution as $u_i^{r,m}$, and both of which are independent of $\mathbb{X}^r$, and because χ_i is a renewal process, we have

$$\begin{aligned}
&P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\chi_i\left(t+\sqrt{|N^r|}m\frac{\lambda_i^r}{|N^r|}\right) - \chi_i\left(\frac{\lambda_i^r}{|N^r|}\sqrt{|N^r|}m + (t-u_i^{r,m}(1))\right) - (t-u_i^{r,m}(1))\right\|_{L_{\sqrt{x_{r,m}}\lambda_i^r/|N^r|}} \ge \sqrt{x_{r,m}}\epsilon L\right\}\\
&= Pl\left\{\max_{m<\sqrt{|N^r|}T} l\left\|\chi_i(t+u_i^{r,0}(1)) - \chi_i(u_i^{r,0}(1)+\tilde{u}_i^{r,m}(1)) - (t-\tilde{u}_i^{r,m}(1))r\right\|_{L_{\sqrt{\tilde{x}_{r,m}}\lambda_i^r/|N^r|}} - \sqrt{\tilde{x}_{r,m}}\epsilon L \ge 0r\right\}. \qquad \text{(C7)}
\end{aligned}$$

Also, observe that

$$\begin{aligned}
&P\left\{\max_{m<\sqrt{|N^r|}T}\left\|\chi_i(t+u_i^{r,0}(1)) - \chi_i(u_i^{r,0}(1)+\tilde{u}_i^{r,m}(1)) - (t-\tilde{u}_i^{r,m}(1))\right\|_{L_{\sqrt{\tilde{x}_{r,m}}\lambda_i^r/|N^r|}} - \sqrt{\tilde{x}_{r,m}}\epsilon L \ge 0\right\}\\
&\le \sum_{m\le\sqrt{|N^r|}T} P\left\{\|\chi_i(t)-t\|_{L\sqrt{\tilde{x}_{r,m}}\lambda_i^r/|N^r|} > \sqrt{\tilde{x}_{r,m}}\epsilon L\right\} + P\left\{\|\chi_i(t+\epsilon^r)-\chi_i(t)\|_{2|\lambda||N^r|T} > \sqrt{|N^r|}L\epsilon\right\}, \qquad \text{(C8)}
\end{aligned}$$

for $\epsilon^r \to 0$ in probability by Lemma C.2. Let $\hat{\chi}_i^r(t) = \chi_i(|N^r|t)/\sqrt{|N^r|} - 1/\sqrt{|N^r|}$. Then by FCLT, $\hat{\chi}_i^r$ converges weakly to a Brownian motion as $r \to \infty$. The second term in (C8) converges to 0 as $r \to \infty$. To conclude the proof, we show below that for each m

$$P\left\{\|\chi_i(t)-t\|_{L\sqrt{\tilde{x}_{r,m}}\lambda_i^r/|N^r|} > \sqrt{\tilde{x}_{r,m}}\epsilon L\right\} < \epsilon/(L\sqrt{|N^r|})$$

Similarly, for each departure process too, it is enough to investigate the processes with index $m = 0$, because we have

$$D_{jk}^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(S_{jk}\left(T_{jk}^{r}\left(\frac{\sqrt{x_{r,m}}}{|N^r|}t + \frac{m}{\sqrt{|N^r|}}\right)\right) - S_{jk}\left(T_{jk}^{r}\left(\frac{m}{\sqrt{|N^r|}}\right)\right)\right) \tag{C9}$$

for $j \in \mathcal{J}$, $k \in \mathcal{K}(j)$. Hence, by restarting the process at time $m/\sqrt{|N^r|}$, we have that the only condition to be checked is whether the residual time of the first arrival for S_{jk} after time $T_{jk}^{r}(m/\sqrt{|N^r|}) \in [0, |N^r|T]$ satisfies a similar condition to (C6).

The following lemma, taken from Bramson [10], shows that (C6) holds. Let

$$u_i^{r,T,\max} = \max\{|u_i(l)|\colon U_i(l-1) \le 2|\lambda||N^r|T\}, \quad \text{for all } i \in \mathcal{I} \qquad \text{and}$$
$$v_{jk}^{r,T,\max} = \max\{|v_{jk}(l)|\colon V_{jk}(l-1) \le |N^r|T\}, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j).$$

Lemma C.2. *Assume that $u_i(1)/\sqrt{N^r} \to 0$ in probability as $r \to \infty$ and (19) holds and that $\lambda^r/|N^r|$ is bounded. Then, for given T,*

$$u_i^{r,T,\max}\Big/\sqrt{|N^r|} \to 0 \quad \text{in probability as } r \to \infty, \quad \text{for all } i \in \mathcal{I} \qquad \text{and}$$
$$v_{jk}^{r,T,\max}\Big/\sqrt{|N^r|} \to 0 \quad \text{in probability as } r \to \infty, \quad \text{for all } j \in \mathcal{J} \text{ and } k \in \mathcal{K}(j).$$

Proof. The proofs immediately follow by taking $r = \sqrt{2|\lambda||N^r|}$ and $r = \sqrt{|N^r|}$, respectively, in Bramson [10, Lemma 5.1]. □

Fix $\epsilon > 0$, $L > 0$, and $T > 0$. We prove each bound separately.

Proof of (76). Fix $i \in \mathcal{I}$. Because $x_{r,0} \ge |N^r|$ by (60), similar to (5.31) in Bramson [10], using Lemma C.2, for given $\epsilon > 0$ and large enough r,

$$\begin{aligned} P\big(\|\chi_i(t) - t\|_{2|\lambda|L\sqrt{x_{r,0}}} \ge 2|\lambda|\epsilon L\sqrt{x_{r,0}}\big) &= \int_0^\infty P\big(\|\chi_i(t) - t\|_{2|\lambda|L\sqrt{x}} \ge 2|\lambda|\epsilon L\sqrt{x}\big)\, dF_{x_{r,0}}(x) \\ &\le \frac{\epsilon}{|2\lambda|L\sqrt{|N^r|}}. \end{aligned}$$

The first inequality follows from the fact that arrivals after a time point t are independent of the state of the system at that point. Again for r large enough,

$$\begin{aligned} \frac{1}{\sqrt{x_{r,0}}}\|\chi_i(t) - t\|_{2|\lambda|L\sqrt{x_{r,0}}} &\ge \left\| \frac{A_i^r(\sqrt{x_{r,0}}t/|N^r|)}{\sqrt{x_{r,0}}} - \frac{\lambda_i^r}{|N^r|}t \right\|_L \\ &= \left\| A_i^{r,0}(t) - \frac{\lambda_i^r}{|N^r|}t \right\|_L. \end{aligned}$$

Hence,

$$P\left\{ \left\| A_i^{r,0}(t) - \frac{\lambda_i^r}{|N^r|}t \right\|_L > 2\epsilon L \right\} \le \frac{2\epsilon}{L\sqrt{|N^r|}}$$

and so

$$P\left\{ \left\| A^{r,0}(t) - \frac{\lambda^r}{|N^r|}t \right\|_L > \epsilon L \right\} \le \frac{2I\epsilon}{2|\lambda|L\sqrt{|N^r|}}.$$

Multiplying the error bounds by $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ by a factor of $2I(L \vee T)$ we obtain (76). □

Proof of (77). Fix $j \in \mathcal{J}$ and $k \in \mathcal{K}(j)$. We first show that for r large enough

$$P\left\{ \sup_{0 \le t_1 \le t_2 \le L} (S_{jk}(2\beta_j\sqrt{x_{r,0}}t_2) - S_{jk}(2\beta_j\sqrt{x_{r,0}}t_1)) \ge 2\beta_j\sqrt{x_{r,0}}\frac{(t_2 - t_1)}{\mu_{jk}} + 4\sqrt{x_{r,0}}\beta_j L\epsilon \right\} \le \frac{4\epsilon}{\beta_j L\sqrt{|N^r|}}. \tag{C10}$$

By Proposition 4.3 of Bramson [10] and Lemma C.2, for large enough r,

$$P\left\{ \left\| S_{jk}(t) - \frac{t}{\mu_{jk}} \right\|_{2\beta_j L\sqrt{x_{r,0}}} \ge 2\beta_j L\sqrt{x_{r,0}}\epsilon \right\} \le 2\frac{\epsilon}{\beta_j L\sqrt{|N^r|}}.$$

Then,

$$P\left\{\sup_{0\le t_1\le t_2\le L}\left(\left(S_{jk}(2\beta_j\sqrt{x_{r,0}}t_2)-\frac{2\beta_j\sqrt{x_{r,0}}t_2}{\mu_{jk}}\right)-\left(S_{jk}(2\beta_j\sqrt{x_{r,0}}t_1)-\frac{2\beta_j\sqrt{x_{r,0}}t_1}{\mu_{jk}}\right)\right)\ge 4\sqrt{x_{r,0}}\beta_j L\epsilon\right\}$$
$$\le\frac{4\epsilon}{\beta_j L\sqrt{|N^r|}}.$$

This gives (C10). Next, we show that

$$P\left\{\sup_{0\le t_1\le t_2\le L}\left(S_{jk}\left(T^r_{jk}\left(\frac{\sqrt{x_{r,0}}t_2}{|N^r|}\right)\right)-S_{jk}\left(T^r_{jk}\left(\frac{\sqrt{x_{r,0}}t_1}{|N^r|}\right)\right)\right)\ge\beta_j\sqrt{x_{r,0}}\frac{(t_2-t_1)}{\mu_{jk}}+5\sqrt{x_{r,0}}\beta_j L\epsilon\right\}$$
$$\le\frac{5\epsilon}{L\beta_j\sqrt{|N^r|}}. \tag{C11}$$

We prove (C11) by showing that the event in (C11) is included in (C10). Assume that for $\omega\in\Omega$

$$S_{jk}\left(T^r_{jk}\left(\frac{\sqrt{x_{r,0}}t_2}{|N^r|}\right)\right)-S_{jk}\left(T^r_{jk}\left(\frac{\sqrt{x_{r,0}}t_1}{|N^r|}\right)\right)\ge\beta_j\sqrt{x_{r,0}}\frac{(t_2-t_1)}{\mu_{jk}}+4\sqrt{x_{r,0}}\beta_j L\epsilon \tag{C12}$$

for some $0\le t_1\le t_2\le L$. Let

$$\tau_l=T^r_{jk}\left(\frac{\sqrt{x_{r,0}}t_l}{|N^r|},\omega\right)$$

for $l=1,2$. Then, for r large enough

$$0\le\tau_1\le\tau_2\le 2L\sqrt{x_{r,0}}\beta_j\quad\text{and} \tag{C13}$$
$$\tau_2-\tau_1\le 2\sqrt{x_{r,0}}\beta_j(t_2-t_1). \tag{C14}$$

By (C12) and (C14)

$$S_{jk}\left(2\beta_j\sqrt{x_{r,0}}\frac{\tau_2}{2\beta_j\sqrt{x_{r,0}}}\right)-S_{jk}\left(2\beta_j\sqrt{x_{r,0}}\frac{\tau_1}{2\beta_j\sqrt{x_{r,0}}}\right)$$
$$\ge 2\beta_j\sqrt{x_{r,0}}\frac{\tau_2/(2\beta_j\sqrt{x_{r,0}})-\tau_1/(2\sqrt{\beta_j x_{r,0}})}{\mu_{jk}}+4\sqrt{x_{r,0}}\beta_j L\epsilon.$$

By (C13), $0\le\tau_1/(2\beta_j\sqrt{x_{r,0}})\le\tau_2/(2\beta_j\sqrt{x_{r,0}})\le L$. Using this and (C14), we get that ω also satisfies the inequality in (C5). Thus we have (C11). By (C9), this implies, by reselecting ϵ, that

$$P\left\{\sup_{t_1,t_2\in[0,L]}|D^{r,0}(t_2)-D^{r,0}(t_1)|\ge N|t_2-t_1|+\epsilon\right\}\le\frac{\epsilon}{\sqrt{|N^r|}}, \tag{C15}$$

with $N=\max_{j\in\mathcal{J},k\in\mathcal{K}(j)}\{\beta_j/\mu_{jk}\}$. Multiplying the exceptional probability by $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ appropriately we obtain (77). □

Proof of (78). By setting $\epsilon=1$, $t_2=L$, and $t_1=0$ in (C15), we have that

$$P\left\{D^r_{jk}\left(\frac{\sqrt{x_{r,0}}}{|N^r|}L\right)\ge 2NL\sqrt{x_{r,0}}\right\}\le\frac{\epsilon}{\sqrt{|N^r|}}. \tag{C16}$$

Off of the exceptional set given in (C16)

$$D^r_{jk}\left(\frac{\sqrt{x_{r,0}}}{|N^r|}L\right)+1\le 3NL\sqrt{x_{r,0}}.$$

Let $a=0$ or 1. It follows from Bramson [10, Proposition 4.2] that for large enough n

$$P\left\{\left\|V_{jk}(l)-\frac{l}{\mu_{jk}}\right\|_n\ge\epsilon n\right\}\le\frac{\epsilon}{n}.$$

By setting $n = 3NL\sqrt{x_{r,0}}$, we get

$$P\left\{\left\|V_{jk}(D^r_{jk}(t)+a)-\frac{D^r_{jk}(t)}{\mu_{jk}}\right\|_{(\sqrt{x_{r,0}}/|N^r|)L}\geq 3NL\sqrt{x_{r,0}}\epsilon\right\}\leq B_2\frac{\epsilon}{\sqrt{|N^r|}}$$

for $B_2 \geq 2/3NL$. By enlarging ϵ appropriately, we get for $\tilde{b} = (1,0)$ or $(0,0)$

$$P\left\{\left\|V^{r,0}_{jk}(D^{r,0}_{jk}(t),\tilde{b})-\frac{D^{r,0}_{jk}(t)}{\mu_{jk}}\right\|_L\geq\epsilon\right\}\leq\frac{\epsilon}{\sqrt{|N^r|}}.$$

Multiplying the exceptional probability by $\lceil\sqrt{|N^r|}T\rceil$ and enlarging ϵ appropriately, we obtain

$$P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V^{r,m}_{jk}(D^{r,m}_{jk}(t),\tilde{b})-\frac{D^{r,m}_{jk}(t)}{\mu_{jk}}\right\|_L\geq\epsilon\right\}\leq\epsilon. \tag{C17}$$

For $b = (0,1)$ and $\tilde{b} = (0,0)$, by (63)

$$\begin{aligned}&P\left\{\max_{m<\sqrt{|N^r|}T}\left\|V^{r,m}_{jk}(D^{r,m}_{jk}(t),\tilde{b})-V^{r,m}_{jk}(D^{r,m}_{jk}(t),b)\right\|_L\geq\epsilon\right\}\\&\quad\leq P\left\{\max_{m<\sqrt{|N^r|}T}\left|V_{jk}\left(D^r_{jk}\left(\frac{m}{\sqrt{|N^r|}}\right)\right)-V_{jk}\left(D^r_{jk}\left(\frac{m}{\sqrt{|N^r|}}\right)+1\right)\right|\geq\sqrt{x_{r,m}}\epsilon\right\}.\end{aligned} \tag{C18}$$

Observe that, by (16), $V_{jk}(D^r_{jk}(m/\sqrt{|N^r|})) \leq |N^r|T$ and by Lemma C.2

$$P\left\{v^{r,T,\max}_{jk}\geq\sqrt{x_{r,m}}\epsilon\right\}\leq\epsilon \tag{C19}$$

for large enough r. Thus, we get (78) by combining (16) with (C17)–(C19). □

C.2.2. Proof of Proposition 5.2. Assume that Assumption 3.1 holds and π satisfies Assumption 3.2. Fix $\epsilon > 0$, $L > 0$ and $T > 0$. Recall that

$$R^{r,m}_k(t)=\frac{1}{\sqrt{x_{r,m}}}\left(F_k\left(\sqrt{x_{r,m}}G^{r,m}_k(t)+G^r_k(m/\sqrt{|N^r|})\right)-F_k\left(G^r_k(m/\sqrt{|N^r|})\right)\right).$$

By Assumption 3.2 and Lemma B.1

$$\frac{\|Q^r(t)\|_T}{N^r}\to 0$$

in probability as $N^r \to \infty$. Hence, for

$$\mathscr{C}^r=\{\|Q^r(t)\|_T\leq N^r\},$$

$P\mathscr{C}^r \to 1$ as $r \to \infty$. For the rest of the proof we only consider $\omega \in \mathscr{C}^r$.

Now, on $\mathscr{C}^r$,

$$\max_{m\leq\sqrt{|N^r|}T}\left|G^r_k(m/\sqrt{|N^r|})\right|\leq G^r_k(T)\leq|N^r|T$$

Let

$$v^{r,T,\max}_k=\max\left\{\left|v_i(l)\right|\colon\Upsilon_k(l-1)\leq|N^r|T\right\}.$$

Then, similar to Lemma C.2, we have that

$$v^{r,T,\max}_k/\sqrt{|N^r|}\to 0\quad\text{in probability as } r\to\infty. \tag{C20}$$

Hence, similar to the proof of (76) we can focus on $R_k^{r,0}$. Then, for Θ^r defined as in Remark C.1,

$$P\left(\Theta^r \cap \left\{\left\|\frac{F(\sqrt{x_{r,0}}G_k^{r,0}(t))}{\sqrt{x_{r,0}}}\right\|_L > \epsilon\right\}\right) \le P\left\{\left\|\frac{F(\sqrt{x_{r,0}}\epsilon(r))}{\sqrt{x_{r,0}}}\right\|_L > \epsilon\right\}$$
$$= P\left\{\left\|F(\epsilon(r)t) - \gamma_k\epsilon(r)t\right\|_{L\sqrt{x_{r,0}}} + \gamma_k\epsilon(r)T\sqrt{x_{r,0}} > \epsilon\sqrt{x_{r,0}}\right\} < \frac{\epsilon}{\sqrt{|N^r|}} \tag{C21}$$

in Bramson [10, Proposition 4.2] and the fact that $\epsilon(r) \to 0$ as $r \to \infty$.

Hence,

$$P\left\{\max_{m<\sqrt{\|N^r\|T}} \|R_k^{r,m}(t)\|_L > \epsilon\right\} \le \epsilon(r) + P\left(\Theta^r \cap \left\{\max_{m<\sqrt{\|N^r\|T}} \|R_k^{r,m}(t)\|_L > \epsilon\right\}\right)$$
$$\le \epsilon(r) + \sum_{m<\sqrt{\|N^r\|T}} P\left(\Theta^r \cap \{\|R_k^{r,m}(t)\|_L > \epsilon\}\right)$$
$$\le \epsilon(r) + 2T\epsilon,$$

where the last inequality follows from (C20) and (C21).

C.2.3. Proof of Proposition 5.3. We use the bounds established in (76)–(78) and (79). Fix L, T, and $\epsilon > 0$. Choose r large enough so that (76)–(78) and (79) hold with $\epsilon/(3d)$. Let $\mathcal{V}^r$ be the intersection of the complements of the events given in (76)–(78), so $P\{\mathcal{V}^r\} > 1 - \epsilon$. We show that for r large enough and all $\omega \in \mathcal{V}^r$

$$\max_{m<\sqrt{|N^r|T}} \sup_{t_1,t_2\le L} \left|\mathbb{X}^{r,m}(t_1) - \mathbb{X}^{r,m}(t_2)\right| \le \tilde{N}|t_1 - t_2| + \epsilon \tag{C22}$$

for some $\tilde{N}$ that only depends on (I, J, K, λ). We fix $\omega \in \mathcal{V}^r$ for the rest of the proof and so omit it from the notation. Let $t_1, t_2 \in [0, T]$ and $m \ge 0$. We first show that for any $j \in \mathcal{J}$, and $k \in \mathcal{K}(j)$

$$\left|Z_{jk}^{r,m}(t_2) - Z_{jk}^{r,m}(t_1)\right| \le N_0|t_2 - t_1| + \epsilon \tag{C23}$$

for some $N_0 > 0$. Because $B_{jk}^{r,m}$ is nondecreasing similar to (B15) we have by (61) and (69) that

$$0 \le B_{jk}^{r,m}(t_2) - B_{jk}^{r,m}(t_1) \le \sum_{l\in\mathcal{K}(j)} \left(D_{jl}^{r,m}(t_2) - D_{jl}^{r,m}(t_1)\right). \tag{C24}$$

Combining (C24) with (69) yields

$$\left|Z_{jk}^{r,m}(t_2) - Z_{jk}^{r,m}(t_1)\right| \le K\left|D^{r,m}(t_2) - D^{r,m}(t_1)\right| + I\left|A^{r,m}(t_2) - A^{r,m}(t_1)\right|.$$

By (76), $|A^{r,m}(t_2) - A^{r,m}(t_1)| < 2|\lambda||t_2 - t_1| + \epsilon$ for r large enough. By setting $N_0 = KN + 2I|\lambda|$ and using (77), we get (C23). Equation (C24) gives that

$$\left|B_{jk}^{r,m}(t_2) - B_{jk}^{r,m}(t_1)\right| \le N_0|t_2 - t_1| + \epsilon.$$

By (79)

$$|R_k^{r,m}(t)|_L \le \epsilon.$$

Combining this with (68) gives

$$\left|Q_k^{r,m}(t_2) - Q_k^{r,m}(t_1)\right| \le N_1|t_2 - t_1| + N_1\epsilon,$$

for $N_1 = (I + J)N_0$. Observe that for any $i \in \mathcal{I}$, $k \in \mathcal{K}$, and $j \in \mathcal{J}(k)$

$$\left|A_{ik}^{r,m}(t_2) - A_{ik}^{r,m}(t_1)\right| \le \left|A_i^{r,m}(t_2) - A_i^{r,m}(t_1)\right|,$$
$$\left|A_{ijk}^{r,m}(t_2) - A_{ijk}^{r,m}(t_1)\right| \le \left|A_i^{r,m}(t_2) - A_i^{r,m}(t_1)\right|.$$

Also, for any $j \in \mathcal{J}$ and $k \in \mathcal{K}(j)$ and for r large enough, by (66),

$$\left\|T_{jk}^{r,m}(t_2) - T_{jk}^{r,m}(t_1)\right\|_L \le 2\beta_j|t_2 - t_1| \qquad \text{and} \qquad \left\|Y_j^{r,m}(t_2) - Y_j^{r,m}(t_1)\right\|_L \le 2\beta_j|t_2 - t_1|.$$

Note that, by the definition of $\mathcal{V}^r$, the inequalities above hold for all $m < \sqrt{|N^r|T}$. This shows that (C22) holds, for r large enough, with $\tilde{N} = N_1 \vee 2$.

C.3. Proofs of the results in §5.2

C.3.1. Proof of Proposition 5.4. Proof is similar to that in Bramson [10, Proposition 6.2]. Assume that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, and (55) holds.

Fix $\omega \in \mathcal{K}^r$ and let $\mathbb{X}^{r,m}$ be given as in §5.1. By (76′), we have for large enough r that

$$\left\| A^{r,m}(t) - \frac{\lambda^r}{|N^r|} t \right\|_L \leq \epsilon(r). \tag{C25}$$

Combining (65) with (78′) gives

$$\left\| T_{jk}^{r,m}(t) - \frac{1}{\mu_{jk}} D_{jk}^{r,m}(t) \right\|_L \leq \epsilon(r). \tag{C26}$$

Recall that $z_{jk} = \beta_j x_{jk}^*$. Using (C4), (71) and Remark C.1 gives

$$\left\| T_{jk}^{r,m}(t) - z_{jk} t \right\|_L \leq \epsilon(r). \tag{C27}$$

By (78′)

$$|R_k^{r,m}(t)|_L \leq \epsilon(r). \tag{C28}$$

Now select any hydrodynamic limit $\tilde{\mathbb{X}}$ of $\mathcal{C}$. For given $\delta > 0$, choose (r, m) so that, $\epsilon(r) \leq \delta$,

$$\left\| \tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t, w) \right\|_L \leq \delta, \tag{C29}$$

and

$$\left| \frac{\lambda^r}{|N^r|} - \lambda \right| \leq \delta. \tag{C30}$$

It follows from (C25) and (C30) that

$$\left\| \tilde{A}(t) - \lambda t \right\|_L \leq 2\delta \tag{C31}$$

and from (C26) and (C27) that

$$\left\| \tilde{D}_{jk}(t) - \mu_{jk} z_{jk} t \right\|_L \leq 2\delta. \tag{C32}$$

By (C28)

$$\|\tilde{R}_k(t)\|_L \leq 2\delta. \tag{C33}$$

By combining (C29), (C31), (C33), (67), and (68), we get

$$\left\| \lambda_i - \sum_{k \in \mathcal{K}} \tilde{A}_{ik}(t) - \sum_{k \in \mathcal{K},\, j \in \mathcal{K}(k)} \tilde{A}_{ijk}(t) \right\|_L \leq 2KJ\delta \qquad \text{and} \tag{C34}$$

$$\left\| \tilde{Q}_k(t) - \tilde{Q}_k(0) - \sum_{i \in \mathcal{I}} \tilde{A}_{ik}(t) + \sum_{j \in \mathcal{J}(k)} \tilde{B}_{jk}(t) \right\|_L \leq 4IJ\delta. \tag{C35}$$

By combining (C29) with (C32) and (69), we get

$$\left\| \tilde{Z}_{jk}(t) - \tilde{Z}_{jk}(0) - \sum_{i \in \mathcal{I}} \tilde{A}_{ijk}(t) - \tilde{B}_{jk}(t) + \mu_{jk} z_{jk} t \right\|_L \leq 6I\delta. \tag{C36}$$

Equations (C32)–(C36) show that the hydrodynamic limits satisfy (30), (31), and (33). Equations (35) and (32) are clearly satisfied by the hydrodynamic limits.

That the hydrodynamic limits satisfy (36) and (37) is proved similarly to the fact that the fluid limits satisfy the fluid analogs of those equations. Hence, we only illustrate the proof of (36).

Fix a hydrodynamic limit $\tilde{\mathbb{X}}$. By the definition of a hydrodynamic limit, there exists a sequence (r_l, m_l, ω_l), with $\omega_l \in \mathcal{H}^l$ for all $l \geq 0$, such that

$$\mathbb{X}^{r_l, m_l}(\cdot, \omega_l) \to \tilde{\mathbb{X}}(\cdot) \tag{C37}$$

u.o.c. as $l \to \infty$. Fix $t > 0$. If $\tilde{Q}_k(t) = 0$, (36) holds trivially. Now we assume that $\tilde{Q}_k(t) > a$ for some $a > 0$. By (C37), there exists an l_0 such that

$$Q_k^{r_l, m_l}(t, \omega_l) > a/2$$

for all $l > l_0$. This implies, by (74), that

$$\sum_{j \in \mathcal{J}(k)} \sum_{l \in \mathcal{H}(j)} \tilde{Z}_{jl}^{r_l, m_l}(t, \omega_l) = 0.$$

Hence,

$$Q_k^{r_l, m_l}(t, \omega_l) \sum_{j \in \mathcal{J}(k)} \sum_{l \in \mathcal{H}(j)} \tilde{Z}_{jl}^{r_l, m_l}(t, \omega_l) = 0. \tag{C38}$$

Convergence in (C37) implies that

$$Q_k^{r_l, m_l}(t, \omega_l) \sum_{j \in \mathcal{J}(k)} \sum_{l \in \mathcal{H}(j)} \tilde{Z}_{jl}^{r_l, m_l}(t, \omega_l) \to \tilde{Q}_k(t) \sum_{j \in \mathcal{J}(k)} \sum_{l \in \mathcal{H}(j)} \tilde{Z}_{jl}(t) \quad \text{as } l \to \infty.$$

This gives (36) by (C38).

C.3.2. Proof of Proposition 5.5. Assume that Assumption 3.1 holds, π satisfies Assumption 3.2, g satisfies Assumption 4.1, and the hydrodynamic model of the π-parallel server systems satisfies Assumption 4.2.

Fix $L > 0$ and let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathcal{C}$. Note that because $\tilde{\mathbb{X}}$ is a limit of hydrodynamically scaled processes $|\tilde{\mathbb{X}}(0)| \leq 1$ by (83). Also, by Proposition 5.4, $\tilde{\mathbb{X}}$ satisfies the hydrodynamic model Equations (30)–(37) for all $t \in [0, L]$. Thus, using (30), (31), (33), and the fact that $|\tilde{\mathbb{X}}(0)| \leq 1$, one can show that there exists $R_L > 0$ such that

$$\|\tilde{\mathbb{X}}(t)\|_L \leq R_L. \tag{C39}$$

Fix $\epsilon > 0$. Because g is continuous, there exists $\delta > 0$ such that

$$|g(x) - g(y)| < \epsilon \tag{C40}$$

if $|x - y| < \delta$ and $x, y \in [-2R_L, 2R_L]^{I+d_z}$.

Fix $0 < \delta < R_L$ as given above and choose r large enough so that (84) holds for all $\omega \in \mathcal{H}^r$ and any $m < \sqrt{|N^r|}T$, that is,

$$\|\mathbb{X}^{r,m}(\cdot) - \tilde{\mathbb{X}}(\cdot)\|_L \leq \delta \tag{C41}$$

for some hydrodynamic limit $\tilde{\mathbb{X}}$ of $\mathcal{C}$. Hence, by (C39),

$$\|\mathbb{X}^{r,m}(t)\|_L \leq 2R_L. \tag{C42}$$

By (C39)–(C42) and Assumption 4.2 we have for all $t \in [0, L]$ that

$$g\big(Q^{r,m}(t), Z^{r,m}(t)\big) \leq H(t) + \epsilon.$$

Result (88) is proven similarly. Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathcal{C}_g$. Then, there exits a sequence $\{\mathbb{X}^{r_k,0}\}$, where $\{r_k\}$ is a subsequence of $\{r\}$, such that

$$\|\mathbb{X}^{r_k,0}(\cdot) - \tilde{\mathbb{X}}(\cdot)\| \to 0 \tag{C43}$$

as $k \to \infty$. But by definition of $\mathcal{C}_g$, $g(\tilde{Q}^{r_k,0}(0), \tilde{Z}^{r_k,0}(0)) \to 0$. This implies by the continuity of g and (C43) that $g(\tilde{Q}(0), \tilde{Z}(0)) = 0$. Thus, by the last statement in Assumption 4.2,

$$\|g(\tilde{Q}(t), \tilde{Z}(t))\|_L = 0. \tag{C44}$$

This shows that (C44) holds for every hydrodynamic limit of $\mathscr{E}_g$.

One can show as in Corollary 5.2 that hydrodynamic limits of $\mathscr{E}_g$ are rich in $\mathscr{E}_g$. Hence, for large enough r and $\omega \in \mathscr{L}^r$,

$$\left\| \mathbb{X}^{r,0}(\cdot) - \tilde{\mathbb{X}}(\cdot) \right\|_L \le \delta$$

for some hydrodynamic limit $\tilde{\mathbb{X}} \in \mathscr{E}$ of $\mathscr{E}_g$. Using (C40) we have

$$g\left(Q^{r,0}(t), Z^{r,0}(t)\right) \le \epsilon$$

for all $t \in [0, L]$.

The validity of (86) when (55) holds is already proved before Proposition 5.5.

C.4. Proofs of the results in §5.3

C.4.1. Proof of Lemma 5.1. For $\omega \in \mathscr{H}^r$ and r chosen large enough it follows from (81) that

$$\left| Q^{r,m}(t_2) - Q^{r,m}(t_1) \right| \le N|t_2 - t_1| + 1$$

for $t_1, t_2 \in [0, L]$ and $m < \sqrt{|N^r|}T$. Setting $t_1 = 0$ and $t_2 = 1/y_{r,m}$ and applying (90) to the above inequality gives

$$\left| Q^r\left(\frac{m+1}{\sqrt{|N^r|}}\right) - Q^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \le \sqrt{x_{r,m}}\frac{N}{y_{r,m}} + \sqrt{x_{r,m}},$$

and so

$$\left| \hat{Q}^r\left(\frac{m+1}{\sqrt{|N^r|}}\right) \right| - \left| \hat{Q}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \le N + y_{r,m} \le 2Ny_{r,m}.$$

The same argument gives

$$\left| \hat{Z}^r\left(\frac{m+1}{\sqrt{|N^r|}}\right) \right| - \left| \hat{Z}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \le 2Ny_{r,m}.$$

Hence,

$$y_{r,m+1} \le \left(\left| \hat{Q}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \vee \left| \hat{Z}^r\left(\frac{m}{\sqrt{|N^r|}}\right) \right| \vee 1 \right) + 2Ny_{r,m} \le 3Ny_{r,m},$$

which yields (95).

C.4.2. Proof of Lemma 5.2. Let $t \in (Ly_{r,0}/\sqrt{|N^r|}, T]$. It follows from the definition of $m_r(t)$ that $m_r(t) \ge 1$. So,

$$\sqrt{|N^r|}t - (m_r(t) - 1) \ge Ly_r(m_r(t) - 1).$$

Setting $m = m_r(t) - 1$ in Lemma 5.1, one has

$$\sqrt{|N^r|}t - m_r(t) \ge Ly_r(m_r(t) - 1) - 1 \ge \frac{L}{3N}y_r(m_r(t)) - 1 \ge \frac{L}{6N}y_r(m_r(t))$$

assuming $L \ge 6N$ as in (94).

Appendix D. Proofs of the results in §6

D.1. Proof of Corollary 6.1. A careful review of the proof of Theorem 4.1 reveals that (53) is needed to show two results. The first result is the inequality in (85). If we only assume that (110) holds, this result still holds because

$$g(Q^{r,0}(0), Z^{r,0}(0)) \le \left(\sqrt{\frac{N^r}{x_{r,0}}}\right)^{c_2} g(\hat{Q}^r(0), \hat{Z}^r(0)) \le g(\hat{Q}^r(0), \hat{Z}^r(0)),$$

as $|N^r|/x_{r,0} \le 1$ by (60). The second result, for which (53) is needed, is Proposition 5.6. By using (110) instead of (53) we get

$$\begin{aligned} g(Q^{r,m}g(t), Z^{r,m}(t)) &= g\left(\sqrt{N^r/x_{r,0}}\left(\hat{Q}^r\left(\frac{\sqrt{x_{r,m}}t}{N^r} + \frac{m}{\sqrt{N^r}}\right), Z^r\left(\frac{\sqrt{x_{r,m}}t}{N^r} + \frac{m}{\sqrt{N^r}}\right)\right)\right) \\ &\ge (\sqrt{N^r/x_{r,0}})^{c_1} g\left(\hat{Q}^r\left(\frac{\sqrt{x_{r,m}}t}{N^r} + \frac{m}{\sqrt{N^r}}\right), Z^r\left(\frac{\sqrt{x_{r,m}}t}{N^r} + \frac{m}{\sqrt{N^r}}\right)\right). \end{aligned}$$

Observe that this gives (92) and (93) for all $t \in [0, T]$ and m satisfying (91) with c replaced with c_1. The rest of the proof can be repeated verbatim to show that Theorem 4.1 holds when c is replaced with c_1.

D.2. Proof of Theorem 6.1. In the rest of this section we assume that Assumption 1 holds, π satisfies Assumption 3.2, g satisfies Assumption 6.1, the hydrodynamic limits of π-parallel server system satisfies Assumption 6.3, Assumption 6.2 holds and

$$g(\hat{Q}^r(0), \hat{Z}^r(0)) \to 0 \quad \text{in probability}$$

as $r \to \infty$.

D.2.1. Modified hydrodynamic limits. Fix $T > 0$ and $\epsilon > 0$. We will show that for r large enough

$$P\{\|g(\hat{Q}^r(t), \hat{Z}^r(t))\|_T > \epsilon\} < \eta,$$

where $\eta > 0$ is arbitrary. Note that this implies the conclusion of Theorem 6.1. Hence, we also fix $\eta > 0$ for the rest of the proof.

By Assumption 6.2, for each $C > 0$ large enough there exists an $r_0 > 0$ that may depend on C such that for all $r > r_0$

$$P(\mathcal{A}_C^r(T)) > 1 - \eta/2,$$

where, for $T > 0$ and $C > 0$, $\mathcal{A}_C^r(T)$ is defined in (131).

To introduce the modified hydrodynamic limits, we change the hydrodynamic scaling slightly. For any nonnegative integer $m < \sqrt{|N^r|}T$, let

$$x_{r,m} = \left|Q^r\left(\frac{m}{\sqrt{|N^r|}}\right)\right|^2 \vee \left|Z^r\left(\frac{m}{\sqrt{|N^r|}}\right) - N^r x^*\right|^2 \vee (C^2|N^r|)$$

The difference between this definition and the definition (60) is the last term. We note that

$$x_{r,m} = C^2|N^r| \tag{D1}$$

on $\mathcal{A}_C^r(T)$ for $m < \sqrt{|N^r|}T$. We define the hydrodynamic scaling as in (62) and (63). Observe that Equations (67)–(75) are still valid. Fix $L > 0$. The results in Propositions 5.1 and 5.2 still hold, hence so does the result in Proposition 5.3. We redefine $\mathcal{H}^r(T)$ to be the intersection of $\mathcal{H}_0^r$ in (81) with $\mathcal{A}_C^r(T)$ and the complements of the events in (76′), (78′), and (79′). The latter three events are modified from three events in (76), (78), and (79) as explained in the passage immediately below (81). As in Corollary 5.1

$$\lim_{r\to\infty} P(\mathcal{H}^r(T)) > 1 - \eta/2.$$

Let

$$E^r = \{\mathbb{X}^{r,m}, m < \sqrt{|N^r|}T, \omega \in \mathcal{K}^r(T)\}.$$

Because (83) holds, Corollary 5.2 holds on $\mathcal{K}^r(T)$ with E^r defined as above and

$$\mathcal{E} = \{E^r\colon r \in \mathbb{N}\}.$$

We call the hydrodynamic limits in this case the hydrodynamic limits on $\{\mathcal{A}_C^r(T)\}$. Observe that the hydrodynamic limits on $\{\mathcal{A}_C^r(T)\}$ also satisfy hydrodynamic model Equations (30)–(37) by Proposition 5.4. However, the policy depend Equation (38) can be different. Thus, the hydrodynamic model equations can be different from those in §4.1.

Next we establish a similar result to Proposition 5.5. First note that on $\mathcal{K}^r(T)$

$$g\left(\hat{Q}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right),\hat{Z}^r\left(\frac{\sqrt{x_{r,m}}t}{|N^r|}+\frac{m}{\sqrt{|N^r|}}\right)\right) = g\left(\sqrt{\frac{x_{r,m}}{|N^r|}}(Q^{r,m}(t),Z^{r,m}(t))\right)$$
$$= g(C(Q^{r,m}(t),Z^{r,m}(t))) \tag{D2}$$

for $m < \sqrt{|N^r|}T$ by (D1) since $\mathcal{K}^r(T) \subset \mathcal{A}_C^r(T)$.

Therefore,

$$g(\hat{Q}^r(0),\hat{Z}^r(0)) = g(C(Q^{r,0}(0),Z^{r,0}(0)))$$

on $\mathcal{K}^r(T)$.

Let $\mathcal{L}^r(T) = \mathcal{K}^r(T) \cap \mathcal{G}^r$ where

$$\mathcal{G}^r = \{|g(C(Q^{r,0}(0),Z^{r,0}(0)))| \le \epsilon(r)\},$$

with $\epsilon(r) \to 0$ slowly enough as $r \to 0$ so that

$$\lim_{r\to\infty} P(\mathcal{L}^r(T)) > 1-\eta/2.$$

As in Proposition 5.5, using (6.3) and the continuity of g we have for $C > C_0$ and $r > r_0$ large enough that

$$g(C(Q^{r,m}(t),Z^{r,m}(t))) \le H_{C,T}(t)+\epsilon,\ \ t\in[0,L] \tag{D3}$$

on $\mathcal{K}^r(T)$. Using the second part of Assumption 6.3, similar to (88) we have

$$\|g(C(Q^{r,0}(t),Z^{r,0}(t)))\|_L \le \epsilon \tag{D4}$$

on $\mathcal{L}^r(T)$ for r large enough.

D.2.2. SSC in diffusion limits. Let

$$y_{r,m} = \sqrt{\frac{x_{r,m}}{|N^r|}} \tag{D5}$$

We begin with changing the scaling using (D2). As in Proposition 5.6 we have from (D3) and (D2) that

$$g(\hat{Q}^r(t),\hat{Z}^r(t)) \le H_{C,T}\left(\frac{1}{y_{r,m}}(\sqrt{|N^r|}t-m)\right)+\epsilon$$

for $\omega \in \mathcal{K}^r(T)$, r large enough and

$$\frac{m}{\sqrt{|N^r|}} \le t \le \frac{1}{\sqrt{|N^r|}}(y_{r,m}L+m). \tag{D6}$$

Also by (D4) we have that

$$\|g(\hat{Q}^r(t),\hat{Z}^r(t))\|_{Ly_{r,0}/\sqrt{|N^r|}} \le \epsilon$$

on $\mathcal{L}^r(T)$ for r large enough.

Let $m_r(t)$ be the smallest m that satisfies (D6) with t and $y_r(m_r(t)) = y_{r,m_r(t)}$. Note that on $\mathcal{K}^r(T)$

$$Ly_{r,n} = Ly_{r,m} \quad \text{for all } n, m < \sqrt{|N^r|}T,$$

by (D1) and (D5). Now observe that if $t \in [Ly_{r,0}/\sqrt{|N^r|}, T]$ then $m_r(t) \geq 1$ hence

$$\sqrt{|N^r|}t - (m_r(t) - 1) > Ly_r(m_r(t) - 1) = Ly_r(m_r(t)).$$

Therefore

$$\sqrt{|N^r|}t - m_r(t) > Ly_r(m_r(t)) - 1 > \frac{L}{2} y_r(m_r(t)).$$

for $L > 2$.

Because the value of L is a matter of choice, we can take L sufficiently large and redefine $\mathcal{K}^r(T)$ with the reselected L. Let $H_{C,T}$ be given as in Assumption 6.3. Because $H_{C,T}(t) \to 0$ as $t \to \infty$ is independent of L, for $\epsilon > 0$ fixed, there exists $s^*(\epsilon) > 1$ such that for $t > s^*(\epsilon)$, $H_{C,T}(t) < \epsilon$. So we set

$$L \geq 6s^*(\epsilon).$$

The proof is completed similarly to the proof of Theorem 4.1.

Appendix E. Proofs of the results in §7. In this section we provide the proofs of the results in §7.1.

E.1. Proof of Proposition 7.1. We start by presenting the additional equations satisfied by the SBP V-parallel server systems. For each class k, we denote by $Q_k^{\oplus}(t)$ the total number of customers in the queue whose priorities are at least as great as k and by $B_{1k}^{\ominus}(t)$ the total number of customers who got delayed in the queue and whose service started before time t and whose priorities are at most as large as k. We define $A_{k1k}^{\ominus}(t)$ similarly. Hence,

$$Q_k^{\oplus}(t) = \sum_{j=k}^{I} Q_j^r(t), \qquad A_{k1k}^{\ominus}(t) = \sum_{l=1}^{k} A_{l1l}(t) \qquad \text{and} \qquad B_{1k}^{\ominus}(t) = \sum_{l=1}^{k} B_{1k}(t). \tag{E1}$$

The SBP policy entails that

$$B_{1k}^{\ominus}(t) + A_{k1k}^{\ominus}(t) \quad \text{can only increase when } Q_{k+1}^{\oplus}(t) = 0. \tag{E2}$$

The following proposition characterizes the fluid limits of the SBP V-parallel server systems.

Proposition E.1. *Let $\{\mathbb{X}_{SBP}^r\}$ be a sequence of SBP V-parallel server system processes. Assume that the conditions of Theorem B.1 are satisfied.*

(i) *In addition to the fluid limit Equations (B3)–(B10), each fluid limit $\bar{\mathbb{X}}_{SBP}$ of $\mathbb{X}_{SBP}^r$ satisfies*

$$\dot{\bar{A}}_{k1k}^{\ominus}(t) + \dot{\bar{B}}_{1k}^{\ominus}(t) = 0 \qquad \textit{when} \qquad \bar{Q}_{k+1}^{\oplus}(t) > 0 \tag{E3}$$

or equivalently

$$\dot{\bar{A}}_{k1k}^{\oplus}(t) + \dot{\bar{B}}_{1k}^{\oplus}(t) = \sum_{k=1}^{K} \dot{\bar{D}}_{1k}(t) \qquad \textit{when} \qquad \bar{Q}_k^{\oplus}(t) > 0, \tag{E4}$$

where $\bar{A}_{k1k}^{\ominus}(t)$ and $\bar{B}_{1k}^{\ominus}(t)$ are defined as in (E1) with processes in the original scale being replaced by their fluid limit counterparts, and

$$\bar{A}_{k1k}^{\oplus}(t) = \sum_{l=k}^{I} \bar{A}_{l1l}(t) \qquad \textit{and} \qquad \bar{B}_{1k}^{\oplus}(t) = \sum_{l=k}^{I} \bar{B}_{1l}(t).$$

(ii) *Let $\vec{q}_r = (q_1, \ldots, q_I)$, where $q_1 = r \geq 0$ and $q_i = 0$ for $i = 2, \ldots, I$. Let $z = \{z_{1i}, \ldots, z_{1I}\}$, where $z_{1i} = \lambda_i/\mu_{1i}$ for $i = 1, \ldots, I$. Then, $\mathcal{M}_{SBP} = \{(\vec{q}_r, z)\colon r \geq 0\}$ is the set of all the steady states of the fluid limits of $\mathbb{X}_{SBP}^r$.*

Proof. We prove the proposition in two parts.

(i) Let $\bar{\mathbb{X}}$ be a fluid limit and for notational convenience assume that $\{\bar{\mathbb{X}}^r(\cdot, \omega)\}$, for some $\omega \in \mathcal{A}$, where $\mathcal{A}$ is defined as in the proof of Theorem B.1, converges u.o.c. to $\bar{\mathbb{X}}$. Assume that $\bar{Q}_k^{\oplus}(t) > 0$.

By the continuity of $\bar{Q}$ there exists $\varepsilon > 0$ and $\delta > 0$ such that $\bar{Q}_k^+(s) > \varepsilon$ for all $s \in [t-\delta, t+\delta]$. Because $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$ converges u.o.c. to $\bar{\mathbb{X}}$, $\sum_{l=k}^K \bar{Q}_l^r(s) > \varepsilon/4$ for all $s \in [t-\delta, t+\delta]$ and r large enough. Hence, $A_{l1l}^r(\cdot,\omega)$ and $B_{1l}^r(\cdot,\omega)$ is flat on $[t-\delta, t+\delta]$ for all $l < k$ by (E2). Hence

$$\dot{\bar{A}}_{l1l}(t) + \dot{\bar{B}}_{1l}(t) = 0 \quad \text{for all } l < k. \tag{E5}$$

This implies (E3). By (11), $\sum_{k=1}^K \bar{Z}_{1k}^r(s) = N^r$ for all $s \in [t-\delta, t+\delta]$, since $\bar{Q}_k^+(s) > \varepsilon$. Hence,

$$\sum_{k=1}^K \dot{\bar{Z}}_{1k}(s) = 0.$$

But

$$\sum_{k=1}^K \dot{\bar{Z}}_{1k}(s) = \sum_{k=1}^K \dot{\bar{A}}_{k1k}(t) + \dot{\bar{B}}_{1k}(t),$$

by (B5). This yields (E4) by (E5).

(ii) Fix $(\vec{q}_r, z) \in \mathcal{M}_{\text{SBP}}$. We show that if $\bar{Q}(0) = \vec{q}_r$ and $\bar{Z}(0) = z$ then $\bar{Q}(t) = \vec{q}_r$ and $\bar{Z}(t) = z$ for all $t > 0$. So assume that $\bar{Q}(0) = \vec{q}_r$ and $\bar{Z}(0) = z$ for a fluid model solution.

We first prove the result for $K = 2$. We start by showing that $\bar{Z}_{12}(t) \geq z_{12}$. Let $f_1(t) = (\bar{Z}_{12}(t) - z_{12})^-$. It is enough to show, by virtue of Dai [13, Lemma 5.2], that $\dot{f}_1(t) \leq 0$ whenever $f_1(t) > 0$ for a regular point $t > 0$. Assume that f_1 is differentiable at time $t > 0$ and that $f_1(t) > 0$, i.e., $\bar{Z}_{12}(t) < z_{12}$. Note that by (B5),

$$\dot{\bar{Z}}_{12}(t) = \dot{\bar{B}}_{12}(t) + \dot{\bar{A}}_{212}(t) - \mu_{12}\bar{Z}_{12}(t).$$

If $\bar{Q}_2(t) > 0$, then by (E5), (B5), and (B9), $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \dot{\bar{D}}_{11}(t) + \dot{\bar{D}}_{12}(t) = \mu_1\bar{Z}_{11}(t) + \mu_2\bar{Z}_{21}(t)$. Also $\bar{Q}_2(t) > 0$ implies $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) = 1$, by (B8). Hence, $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) > \mu_{12}\bar{Z}_{12}(t)$, which implies $\dot{\bar{Z}}_{12}(t) > 0$ and $\dot{f}_1(t) < 0$. If $\bar{Q}_2(t) = 0$, then we claim that $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_2$. If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) < 1$, this trivially follows from (B8). If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) = 1$, then we use the fact that $\dot{\bar{Q}}_2(t) = 0$, because it achieves its minimum at t. This implies by (B4) that $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_2$. Hence, if $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) < z_{12}$, then $\dot{f}_1(t) \leq 0$. Hence, if $\bar{Z}_{12}(0) \geq z_{12}$ then $\bar{Z}_{12}(t) \geq z_{12}$ for all $t \geq 0$.

Next, we show that if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) \geq z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) \geq z_{12}$ for all $t \geq 0$. Assume that $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) \geq z_{12}$ and that $\bar{Q}_2(t) > 0$. By the previous argument we have that $\bar{Z}_{12}(t) \geq z_{12}$. By (B4) and (B5), $\dot{\bar{Q}}_2(t) + \dot{\bar{Z}}_{12}(t) \leq \lambda_1 - \mu_{12}\bar{Z}_{12}(t) \leq 0$. By (E5), $\dot{\bar{Z}}_{12}(t) \geq 0$ when $\bar{Q}_2(t) > 0$. Hence, if $\bar{Q}(t) > 0$ and it is differentiable at $t > 0$, then $\dot{\bar{Q}}(t) \leq 0$. Hence, if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) \geq z_{12}$, then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) \geq z_{12}$ for all $t \geq 0$.

Now we are ready to show that if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) = z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) = z_{12}$ for all $t \geq 0$. Assume that $\bar{Q}_2(0) = 0$, $\bar{Z}_{12}(0) = z_{12}$, and $\bar{Z}_{12}(t) > z_{12}$ for a regular point $t > 0$. By the previous paragraph $\bar{Q}_2(t) = 0$, hence $\dot{\bar{Q}}_2(t) = 0$ by a similar argument above. If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) = 1$, $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_1$ by the fact that $\dot{\bar{Q}}_2(t) = 0$, and by Equations (B3) and (B4). If $\bar{Z}_{11}(t) + \bar{Z}_{12}(t) < 1$, then $\dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) = \lambda_1$ by (B3), (B4), (B6), and (B10). Because $\dot{\bar{Z}}_{12}(t) = \dot{\bar{A}}_{212}(t) + \dot{\bar{B}}_{12}(t) - \mu_{12}\bar{Z}_{12}(t)$, by (B5), $\dot{\bar{Z}}_{12}(t) < 0$ if $\bar{Z}_{12}(t) > z_{12}$. Hence, if $\bar{Q}_2(0) = 0$ and $\bar{Z}_{12}(0) = z_{12}$ then $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) = z_{12}$ for all $t \geq 0$.

Next we show that if $\bar{Q}_2(0) = 0$, $\bar{Z}_{12}(0) = z_{12}$, and $\bar{Z}_{11}(t) = z_{11}$ then $\bar{Q}_2(t) = 0$, $\bar{Z}_{12}(t) = z_{12}$, and $\bar{Z}_{11}(t) = z_{11}$ for all $t \geq 0$. Let $t > 0$ be a regular point. By the arguments above, we have that $\bar{Q}_2(t) = 0$ and $\bar{Z}_{12}(t) = z_{12}$. Hence, $\bar{Z}_{11}(t) \leq z_{11}$ by the definition of the fluid scaling. So assume that $\bar{Z}_{11}(t) < z_{11}$. This implies $\bar{Q}_{11}(t) = 0$. Hence, $\dot{\bar{A}}_{111}(t) + \dot{\bar{B}}_{11}(t) = \lambda_1$. This gives that $\dot{\bar{Z}}_{11}(t) > 0$, because $\dot{\bar{Z}}_{11}(t) = \dot{\bar{A}}_{111}(t) + \dot{\bar{B}}_{11}(t) - \mu_{11}\bar{Z}_{11}(t)$.

Finally, we show that if $\bar{Q}_2(0) = 0$, $\bar{Z}_{12}(0) = z_{12}$, $\bar{Q}_1(0) = r$, and $\bar{Z}_{11}(0) = z_{11}$ then $\bar{Q}_2(t) = 0$, $\bar{Z}_{12}(t) = z_{12}$, $\bar{Q}_1(t) = r$, and $\bar{Z}_{11}(t) = z_{11}$ for all $t \geq 0$. Let $t > 0$ be a regular point. By the arguments above, we have that $\bar{Q}_2(t) = 0$, $\bar{Z}_{12}(t) = z_{12}$, and $\bar{Z}_{11}(t) = z_{11}$. Assume that $\bar{Q}_1(t) > r$. Then $\dot{\bar{Q}}_1(t) = \dot{\bar{A}}_{11}(t) - \dot{\bar{B}}_{11}(t) = 0$, since by (B3) $\dot{\bar{A}}_{11}(t) + \dot{\bar{A}}_{111}(t) = \lambda_1$ and by (B5) $\dot{\bar{A}}_{111}(t) + \dot{\bar{B}}_{11}(t) = \mu_{11}z_{11} = \lambda_1$, when $\bar{Z}_{11}(t) = z_{11}$, $\bar{Z}_{12}(t) = z_{12}$, and $\bar{Q}_{12}(t) = 0$. This completes the proof for the case when $K = 2$.

If $K > 2$, the argument above can be repeated first to prove that if $\bar{Q}_K(0) = 0$ and $\bar{Z}_{K1}(0) = z_{1K}$ then $\bar{Q}_K(t) = 0$ and $\bar{Z}_{K1}(t) = z_{1K}$ for all $t \geq 0$. The same argument now can be repeated for $K-1$ because $\bar{Z}_K$ is a constant function. Proceeding inductively, the same result can be shown to hold for all $k > 1$. The last step, for $k = 1$, is same as above. □

Next we complete the proof of Proposition 7.1 using Proposition E.1. By (28), $\bar{Q}^r(0) \to 0$ and $\bar{Z}^r(0) \to z$ as $r \to \infty$, for $z = (z_{11}, \ldots, z_{1K})$, where $z_{1k} = \lambda_k/\mu_{1k}$. Hence, $\mathbb{X}_{\text{SBP}}^r$ satisfies Assumption 3.2 by Lemma B.1 and Proposition E.1.

E.2. Proof of Proposition 7.2. Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\{\mathbb{X}^{r,m}_{\mathrm{SBP}}\}$. Fix $t>0$. Assume that $\tilde{Q}^{\oplus}_k(t)>0$. Then, by the continuity of $\tilde{Q}$, there exists an $\varepsilon>0$ and a $\tau>0$ such that $\tilde{Q}^{\oplus}_k(s)>\varepsilon$ for all $s\in[t-\tau,t+\tau]$. Fix $0<\delta<\epsilon/2$ and choose r large enough, together with an integer m and $\omega\in\mathscr{K}^r$, so that $\delta<\epsilon(r)$, for $\epsilon(r)$ selected as in Corollary 5.1. It follows from Corollary 5.2 that $(Q^{r,m}_k)^+(s)>\varepsilon/4$ for all $s\in[t-\tau,t+\tau]$.

We obtain from (E2) that

$$(B^{r,m}_{1k})^{\ominus}(t)+(A^{r,m}_{k1k})^{\ominus}(t)\quad\text{can only increase when}\quad(Q^{r,m}_{k+1})^{\oplus}(t)=0, \tag{E6}$$

where the processes $(B^{r,m}_{1k})^{\ominus}$, $(A^{r,m}_{k1k})^{\ominus}$, and $(Q^{r,m}_{k+1})^{\oplus}$ are defined as in (E1), with the hydrodynamically scaled processes replacing the corresponding original processes. By (E6), $B^{r,m}_{1l}(\cdot)$ is flat on $[t-\tau,t+\tau]$ for all $l<k$, and because

$$\left|\left(B^{r,m}_{1l}(t+\tau)-B^{r,m}_{1l}(t-\tau)\right)-\left(\tilde{B}_{1l}(t+\tau)-\tilde{B}_{1l}(t-\tau)\right)\right|<2\delta$$

and δ is arbitrary, $\tilde{B}_{1l}(\cdot)$ is also flat on $[t-\tau,t+\tau]$ for all $l<k$. Therefore

$$\dot{\tilde{B}}_{1l}(t)=0 \tag{E7}$$

for all $l<k$. Similarly

$$\dot{\tilde{A}}_{l1l}(t)=0 \tag{E8}$$

for all $l<k$. This yields (122).

E.3. Proof of Proposition 7.3

Remark E.1. Note that if $\tilde{Q}^{\oplus}_k(t)>0$, then $\sum_{k=1}^{K}\tilde{Z}_{1k}(s)=0$ for every $s\in[t-\tau,t+\tau]$ by (36), where τ is selected as in the previous proof. Then, by (33)

$$\sum_{k=1}^{K}\dot{\tilde{Z}}_{1k}(s)=\sum_{l=1}^{K}\left(\dot{\tilde{A}}_{l1l}(t)+\dot{\tilde{B}}_{1l}(t)-\mu_{1k}z_{1k}\right)=0. \tag{E9}$$

Combining (E7), (E8), and (E9) gives that

$$\dot{\tilde{A}}^{\oplus}_{k1k}(t)+\dot{\tilde{B}}^{\oplus}_{1k}(t)=\sum_{k=1}^{K}\mu_{1k}z_{1k}\qquad\text{when}\qquad\tilde{Q}^{\oplus}_k(t)>0 \tag{E10}$$

for all $t\in[0,L]$, where

$$\tilde{A}^{\oplus}_{k1k}(t)=\sum_{l=k}^{I}\tilde{A}_{l1l}(t),\qquad\tilde{B}^{\oplus}_{1k}(t)=\sum_{l=k}^{I}\tilde{B}_{1l}(t),\qquad\text{and}\qquad\tilde{Q}^{\oplus}_k(t)=\sum_{j=k}^{I}\tilde{Q}_j(t).$$

Let $\tilde{\mathbb{X}}_{\mathrm{SBP}}$ be a hydrodynamic model solution to the hydrodynamic model of the SBP V-parallel server system. Let $f(t)=\sum_{i=2}^{I}\tilde{Q}_i(t)$. By definition of the hydrodynamic scaling we have $|\tilde{\mathbb{X}}(0)|\le 1$. Hence $f(0)<I$. We use Lemma 5.2 of Dai [13] to show that $f(t)=0$ for $t\ge I/(\mu_{11})$. Assume that $f(t)>0$ and $\tilde{\mathbb{X}}_{\mathrm{SBP}}$ is differentiable at time t. Observe that by (31)

$$\sum_{i=2}^{I}\dot{\tilde{Q}}_i(t)=\sum_{i=2}^{I}\dot{\tilde{A}}_{ii}(t)-\sum_{i=2}^{I}\dot{\tilde{B}}_{1i}(t) \tag{E11}$$

and by (30)

$$\sum_{i=2}^{I}\dot{\tilde{A}}_{ii}(t)+\sum_{i=2}^{I}\dot{\tilde{A}}_{i1i}(t)=\sum_{i=2}^{I}\lambda_i=\sum_{i=2}^{I}\mu_{1i}z_{1i}. \tag{E12}$$

Because $\sum_{i=2}^{I}\tilde{Q}_i(t)>0$, $\sum_{i=1}^{I}\dot{\tilde{Z}}_i(t)=0$ by (36) and the continuity of $\tilde{Q}$. Hence, by (33),

$$\sum_{i=1}^{I}\dot{\tilde{A}}_{i1i}(t)+\sum_{i=1}^{I}\dot{\tilde{B}}_{1i}(t)=\sum_{i=1}^{I}\mu_{1i}z_{1i}. \tag{E13}$$

Combining (E11)–(E13) with (E10) gives

$$\dot{f}(t)=\sum_{i=2}^{I}\mu_{1i}z_{1i}-\sum_{i=1}^{I}\dot{\tilde{D}}_{1i}(t)\le-\mu_{11}z_{11}.$$

E.4. Proof of Lemma 7.1. The proof is similar to that of Lemma 3.2 of Puhalskii and Reiman [40]. We only give a sketch of the proof here. Note that

$$Q_k^r(t) + Z_{1k}^r(t) = Q_k^r(0) + Z_{1k}^r(0) + A_k^r(t) - S_{1k}\left(\int_0^t Z_{1k}^r(s)\,\mathrm{d}s\right).$$

Let

$$X_k^r(t) = \frac{Q_k^r(t) + Z_{1k}^r(t) - N^r\lambda_k/\mu_{1k}}{\sqrt{N^r}}.$$

Then,

$$\begin{aligned} X_k^r(t) = X_k^r(0) &+ \left(\frac{A_k^r(t) - \lambda_k^r t}{\sqrt{N^r}}\right) - \sqrt{N^r}\left(\frac{S_{1k}(N^r\int_0^t \bar{Z}_{1k}^r(s)\,\mathrm{d}s)}{N^r} - \mu_{1k}\int_0^t \bar{Z}_{1k}^r(s)\,\mathrm{d}s\right) \\ &+ \sqrt{N^r}t\left(\frac{\lambda_k^r}{N^r} - \lambda_k\right) - \mu_{1k}\int_0^t \hat{Z}_{1k}^r(s)\,\mathrm{d}s. \end{aligned} \tag{E14}$$

Observe that

$$|\hat{Z}_{1k}^r(t)| \le |X_k^r(t)| + \left(\sum_{i=1}^K X_i^r(t)\right)^+. \tag{E15}$$

The proof is completed using Gronwall's inequality and arguments similar to that used in the proof of Lemma 3.2 and Puhalskii and Reiman [40, p. 589] with (114), (E14), and (E15).

Appendix F. Proofs of Results in §8

F.1. Proof of Proposition 8.2. Let $\{\mathbb{X}^r\}$ be a sequence of V-systems working under the Armony-Maglaras threshold policy. We start our analysis by presenting the additional equations that must be satisfied by $\mathbb{X}^r$.

Because class 2 jobs get priority when the number of class 2 jobs in the queue exceeds $\sqrt{|N^r|}\theta$,

$$B_{11}^r(t) + A_{111}^r(t) \quad \text{can only increase when} \quad Q_2(t)^r < \sqrt{|N^r|}\theta. \tag{F1}$$

Also,

$$B_{21}^r(t) \quad \text{can only increase when} \quad Q_2(t)^r \ge \sqrt{|N^r|}\theta. \tag{F2}$$

The following proposition characterizes the fluid limits of the V-parallel server systems working under the Armony-Maglaras threshold policy.

Proposition F.1. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that the conditions of Theorem B.1 are satisfied.*

(i) *In addition to the fluid limit Equations (B3)–(B10), each fluid limit $\bar{\mathbb{X}}$ of $\mathbb{X}^r$ satisfies*

$$\dot{\bar{A}}_{111}(t) + \dot{\bar{B}}_{11}(t) = 0 \qquad \text{when} \qquad \bar{Q}_2(t) > 0.$$

(ii) *Let $\vec{q}_r = (q_1, q_2)$, where $q_1 = r \ge 0$ and $q_2 = 0$ and $z = \{z_{11}, z_{12}\}$, where $z_{1i} = \lambda_i/\mu_i$ for $i = 1, 2$. Then, $\mathcal{M} = \{(\vec{q}_r, z)\colon r \ge 0\}$ is the set of all the invariant states of the fluid limits of $\mathbb{X}^r$.*

By (127),

$$(\bar{Q}^r(0), \bar{Z}^r(0)) \;\Rightarrow\; (0, z), \tag{F3}$$

where $z = (\lambda_1/\mu_1, \lambda_2/\mu_2)$, hence $\mathbb{X}^r$ satisfies Assumption 3.2 by Proposition F.1. Note that, $\mathbb{X}^r$ satisfies Assumption 3.1 by (124) and (125).

Proof of Proposition F.1. We prove the proposition in two parts.

(i) Let $\bar{\mathbb{X}}$ be a fluid limit and for notational convenience assume that $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$, for some $\omega \in \mathcal{A}$, where $\mathcal{A}$ is defined as in proof of Theorem B.1, converges u.o.c. to $\bar{\mathbb{X}}$. Assume that $\bar{Q}_2(t) > 0$.

By the continuity of $\bar{Q}$ there exists $\varepsilon > 0$ and $\delta > 0$ such that $\bar{Q}_2(s) > \varepsilon$ for all $s \in [t-\delta, t+\delta]$. Since $\{\bar{\mathbb{X}}^r(\cdot,\omega)\}$ converges u.o.c. to $\bar{\mathbb{X}}$, $\bar{Q}_2^r(s) > \varepsilon/4$ for all $s \in [t-\delta, t+\delta]$ and r large enough. Hence, $A_{11}^r(\cdot,\omega)$ and $B_{11}^r(\cdot,\omega)$ are flat on $[t-\delta, t+\delta]$ by (F1). Hence

$$\dot{\bar{A}}_{111}(t) + \dot{\bar{B}}_{11}(t) = 0. \tag{F4}$$

(ii) The proof is similar to that of part (ii) Proposition E.1. □

F.2. Proof of Proposition 8.3. The following result is established by Halfin and Whitt [26].

THEOREM F.1. *Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy and $\hat{X}^r$ be defined as in (126). Assume that (127), (124), and (125) hold. Then*

$$\hat{X}^r(\cdot) \;\Rightarrow\; \hat{X}(\cdot),$$

where

$$\hat{X}(t) = \hat{X}(t) + W(t) - \beta t - \mu \int_0^t (\hat{X}(s))^- \, ds$$

and W is a driftless Brownian motion with variance 2μ.

It can be easily showed using Theorem F.1 that

$$\lim_{R\to\infty} \lim_{r\to\infty} P\{\|\hat{X}^r(t)\|_T > R\} = 0. \tag{F5}$$

Proof of Proposition 8.3 is similar to that of Lemma 3.2 of Puhalskii and Reiman [40]. Let

$$\hat{X}^r_k(t) = \frac{Q^r_k(t) + Z^r_{1k}(t) - |N^r|\lambda_k/\mu}{\sqrt{|N^r|}} \tag{F6}$$

for $k = 1, 2$. We claim that

$$|\hat{Z}^r_{1k}(t)| \le |\hat{X}^r_k(t)| + (\hat{X}^r(t))^+. \tag{F7}$$

To prove this, assume that $\hat{Z}^r_{1k}(t) < 0$; otherwise, the result is obvious. If $\hat{Z}^r_{11}(t) + \hat{Z}^r_{12}(t) < 0$, then $\hat{Q}^r_k(t) = 0$, so the result follows. Assume that $\hat{Z}^r_{11}(t) + \hat{Z}^r_{12}(t) = 0$. Without loss of generality we can assume that $k = 1$. Because $\hat{Z}^r_{11}(t) < 0$, $\hat{Z}^r_{12}(t) = -\hat{Z}^r_{11}(t)$ and $\hat{Q}^r_2(t) \ge 0$, so (F7) follows.

By (F6), for $k = 1, 2$

$$\begin{aligned}\hat{X}^r_k(t) = \hat{X}^r_k(0) &+ \left(\frac{A^r_k(t) - \lambda^r_k t}{\sqrt{|N^r|}}\right) - \sqrt{|N^r|}\left(\frac{S_{1k}(|N^r|\int_0^t \bar{Z}^r_{1k}(s)\,\mathrm{d}s)}{|N^r|} - \mu_{1k}\int_0^t \bar{Z}^r_{1k}(s)\,\mathrm{d}s\right)\\ &+ \sqrt{|N^r|}t\left(\frac{\lambda^r_k}{|N^r|} - \lambda_k\right) - \mu_{1k}\int_0^t \hat{Z}^r_{1k}(s)\,\mathrm{d}s.\end{aligned}$$

Let

$$\hat{A}^r_k(t) = \frac{A^r_k(t) - \lambda^r_k t}{\sqrt{|N^r|}} \qquad \text{and} \qquad c^r_k(t) = \sqrt{|N^r|}\left(\frac{S_{1k}(|N^r|\int_0^t \bar{Z}^r_{1k}(s)\,\mathrm{d}s)}{|N^r|} - \mu\int_0^t \bar{Z}^r_{1k}(s)\,\mathrm{d}s\right).$$

Note that

$$\hat{A}^r_k(\cdot) \;\Rightarrow\; W^a_k(\cdot) \qquad \text{and} \qquad \hat{S}^r_k(t) \;\Rightarrow\; W^s_k(\cdot) \tag{F8}$$

as $r \to \infty$ by Proposition 8.2, (F3), and Donsker's theorem, see Billingsley [8], where W^a_k and W^b_k are Brownian motions with zero drift and variance λ_k.

Now observe that

$$\begin{aligned}&|X^r_1(t)| + |X^r_2(t)|\\ &\quad\le |\hat{X}^r_1(0)| + |\hat{X}^r_2(0)| + |\hat{A}^r_1(t) + \hat{S}^r_1(t)| + |\hat{A}^r_2(t) + \hat{S}^r_2(t)| + \mu\int_0^t \big(|\hat{Z}^r_{11}(s)| + |\hat{Z}^r_{21}(s)|\big)\,\mathrm{d}s\\ &\quad\le |\hat{X}^r_1(0)| + |\hat{X}^r_2(0)| + |\hat{A}^r_1(t)| + |\hat{A}^r_2(t)| + |\hat{S}^r_1(t)| + |\hat{S}^r_2(t)| + 3\mu\int_0^t \big(|X^r_1(s)| + |X^r_2(s)|\big)\,\mathrm{d}s,\end{aligned}$$

where the last inequality follows from (F7). This with Gronwall's inequality and (F8) gives

$$\lim_{R\to\infty} \lim_{r\to\infty} P\big\{\|\hat{X}^r_1(t)\|_T \vee \|\hat{X}^r_2(t)\|_T > R\big\} = 0.$$

This gives (130) because $\hat{Q}^r_i(t) \ge 0$ for all $t \ge 0$, $r \ge 0$, and $k = 1, 2$.

F.3. Proof of Proposition 8.4. Let $\{\mathbb{X}^r\}$ be a sequence of V-parallel server system processes working under the Armony-Maglaras threshold policy. Assume that (124), (125), (127), and (128) hold.

Note that, by (F1) and (F2)

$$B^r_{11}(t)+A^r_{111}(t) \quad \text{can only increase when} \quad g\big(\hat{Q}^r(t),\hat{Z}^r(t)\big)=\hat{Q}^r_1(t)-(\hat{X}^r(t)-\theta)^+>0 \tag{F9}$$

and

$$B^r_{21}(t) \quad \text{can only increase when} \quad g(\hat{Q}^r(t),\hat{Z}^r(t))=\hat{Q}^r_1(t)-(\hat{X}^r(t)-\theta)^+\le 0, \tag{F10}$$

because, if $\hat{Q}^r_1(t)>(\hat{X}^r(t)-\theta)^+$, then $\hat{X}^r(t)=\hat{Q}^r_1(t)+\hat{Q}^r_2(t)$, because the policy is nonidling. Therefore, $\hat{Q}^r_2(t)\le\theta$ in this case. Similarly, if $\hat{Q}^r_1(t)\le(\hat{X}^r(t)-\theta)^+$ and $\hat{Q}^r_2(t)>0$, then $\hat{Q}^r_2(t)\ge\theta$.

Equations (F9) and (F10) imply that

$$B^{r,m}_{11}(t)+A^{r,m}_{111}(t) \quad \text{can only increase when} \quad g\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\big(Q^{r,m}(t),Z^{r,m}(t)\big)\right)>0 \qquad \text{and} \tag{F11}$$

$$B^{r,m}_{21}(t) \quad \text{can only increase when} \quad g\left(\sqrt{\frac{x_{r,m}}{|N^r|}}\big(Q^{r,m}(t),Z^{r,m}(t)\big)\right)\le 0, \tag{F12}$$

where the hydrodynamic scaled process $\mathbb{X}^{r,m}$ is defined as in §D.2.

Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit on $\mathcal{A}^r_R(T)$. Note that $\tilde{\mathbb{X}}$ satisfies (30)–(37) (see §D.2 for more details). We next characterize the additional equations associated with the policy. We claim that for $t\in[0,T]$

$$\dot{\tilde{B}}_{11}(t)=\mu \ \text{ when } \ g\big(R(\tilde{Q}(t),\tilde{Z}(t))\big)>0 \qquad \text{and} \qquad \tilde{Q}_1(t)>0 \tag{F13}$$

$$\dot{\tilde{B}}_{12}(t)=\mu \ \text{ when } \ g\big(R(\tilde{Q}(t),\tilde{Z}(t))\big)<0 \qquad \text{and} \qquad \tilde{Q}_2(t)>0. \tag{F14}$$

To show this, assume that

$$g\big(R(\tilde{Q}(t),\tilde{Z}(t))\big)>2\epsilon \qquad \text{and} \qquad \tilde{Q}_1(t)>2\epsilon \tag{F15}$$

for some $\epsilon>0$. By continuity of g and $\tilde{\mathbb{X}}$ there exists $\delta>0$ such that

$$g\big(R(\tilde{Q}(s),\tilde{Z}(s))\big)>\epsilon \qquad \text{and} \qquad \tilde{Q}_1(s)>\epsilon,$$

for all $s\in[t-\delta,t+\delta]$.

Pick r large enough together with an integer m and $\omega\in\mathcal{A}^r_R(T)$ so that

$$\big\|\tilde{\mathbb{X}}(t)-\mathbb{X}^{r,m}(t)\big\|<\epsilon/2.$$

This gives that

$$g\big(R(Q^{r,m}(s),Z^{r,m}(s))\big)>\epsilon/2 \qquad \text{and} \qquad Q^{r,m}_1(s)>\epsilon/2,$$

because $\sqrt{x_{r,m}/|N^r|}=R$ on $\mathcal{A}^r(T)$ (see §D.2). By (F11)

$$B^{r,m}_{12}(t+\delta)-B^{r,m}_{12}(t-\delta)=0,$$

and so

$$\dot{\tilde{B}}_{12}(t)=0,$$

Now, by (33)

$$\dot{\tilde{Z}}_{12}(t)=-(1-\eta)\mu \qquad \text{and} \qquad \dot{\tilde{Z}}_{11}(t)=\dot{\tilde{B}}_{11}(t)-\mu\eta\mu.$$

Equations (33), (37), and (F15) give that

$$\dot{\tilde{Z}}_{11}(t)+\dot{\tilde{Z}}_{11}(t)=0.$$

Hence,

$$\dot{\tilde{B}}_{11}(t)=\mu.$$

Condition (133) is proved similarly.

Acknowledgments. J. G. Dai's research was supported in part by National Science Foundation Grants DMI-0300599, CMMI-0727400, CMMI-0825840, CMMI-1030589, and by an IBM Faculty Award. T. Tezcan's research was supported in part by National Science Foundation Grant CMMI-0954126.

References

[1] Armony, M. 2005. Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems* **51** 287–329.
[2] Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52** 527–545.
[3] Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Oper. Res.* **52** 271–292.
[4] Ata, B., S. Kumar. 2005. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.* **15** 331–391.
[5] Atar, R. 2005. A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* **15** 820–852.
[6] Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15** 2606–2650.
[7] Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic. *Ann. Appl. Probab.* **14** 1084–1134.
[8] Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.
[9] Borovkov, A. A. 1967. On limit laws for service processes in multi-channel systems. *Siberian Math. J.* **8** 983–1004.
[10] Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.
[11] Bramson, M., J. G. Dai. 2001. Heavy traffic limits for some queueing networks. *Ann. Appl. Probab.* **11** 49–90.
[12] Chen, H., D. D. Yao. 2001. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York.
[13] Dai, J. G. 1995. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.
[14] Dai, J. G., W. Lin. 2005. Maximum pressure policies in stochastic processing networks. *Oper. Res.* **53** 197–218.
[15] Dai, J. G., W. Lin. 2008. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.* **18** 2239–2299.
[16] Dai, J. G., T. Tezcan. 2008. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* **59** 95–134.
[17] Dai, J. G., S. He, T. Tezcan. 2010. Many-server diffusion limits for $G/Ph/n+GI$ queues. *Ann. Appl. Probab.* **20** 1854–1890.
[18] Davis, M. H. A. 1984. Piecewise-deterministic Markov processes: A general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B* **46**(3) 353–388.
[19] Ethier, S. N., T. G. Kurtz. 1986. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, New York.
[20] Fleming, P., A. L. Stolyar, B. Simon. 1994. Heavy traffic limit for a mobile phone system model. *Proc. 2nd Internat. Conf. Telecommunication Systems, Modeling Anal., Nashville, TN*, 317–327.
[21] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.
[22] Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **48** 566–583.
[23] Gross, D., C. M. Harris. 1998. *Fundamentals of Queueing Theory*, 3rd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.
[24] Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* **58**(2) 316–328.
[25] Gurvich, I., M. Armony, A. Mandelbaum. 2005. Service level differentiation in call centers with fully flexible servers. *Management Sci.* **54** 279–294.
[26] Halfin, S., W. Whitt 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.
[27] Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. John Wiley & Sons, New York.
[28] Harrison, J. M. 1988. Brownian models of queueing networks with heterogeneous customer populations. W. Fleming, P. L. Lions, eds. *Stochastic Differential Systems, Stochastic Control Theory and Their Applications*, Vol. 10. The IMA Volumes in Mathematics and Its Applications, Springer-Verlag, New York, 147–186.
[29] Harrison, J. M. 2000. Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab.* **10** 75–103.
[30] Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime. *Oper. Res.* **52** 243–257.
[31] Iglehart, D. L. 1973. Weak convergence of compound stochastic process. *Stochastic Processes Appl.* **1** 11–31.
[32] Maglaras, C. 2000. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Ann. Appl. Probab.* **10**(3) 897–929.
[33] Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49** 1018–1038.
[34] Maglaras, C., A. Zeevi. 2004. Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels. *Math. Oper. Res.* **29**(4) 786–813.
[35] Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53** 242–262.
[36] Mandelbaum, A., P. Momčilović. 2009. Queues with many servers and impatient customers. Technical report, Technion, Haifa, Israel.

[37] Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Oper. Res.* **52** 836–855.
[38] Mandelbaum, A., W. A. Massey, M. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* **30** 149–201.
[39] Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probab. Surveys* **4** 193–267.
[40] Puhalskii, A., M. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595.
[41] Randhawa, R. S., S. Kumar. 2009. Multiserver loss systems with subscribers. *Math. Oper. Res.* **34**(1) 142–179.
[42] Reed, J. 2009. The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* **19** 2211–2269.
[43] Ross, S. 1996. *Stochastic Processsses*. John Wiley & Sons, New York.
[44] Stolyar, A. L. 2005. Optimal routing in output-queued flexible server systems. *Probab. Engrg. Informational Sci.* **19** 141–189.
[45] Tezcan, T. 2008. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Oper. Res.* **33** 51–90.
[46] Tezcan, T., J. G. Dai. 2010. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* **58**(1) 94–110.
[47] Whitt, W. 1982. On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Probab.* **14** 171–190.
[48] Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.
[49] Whitt, W. 2004. A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* **52** 922–941.
[50] Whitt, W. 2005. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* **30** 1–27.
[51] Williams, R. J. 1998. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* **30** 27–88.