Optimal Control of Distributed Parallel Server Systems
Under the Halfin and Whitt Regime

Tolga Tezcan
Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign,
Urbana, Illinois 61801, ttezcan@uiuc.edu

We consider a distributed parallel server system that consists of multiple server pools and a single customer class. We show that the minimum-expected-delay faster-server-first (MED-FSF) routing policy asymptotically minimizes the stationary distribution of the total queue length and the stationary delay probability in the Halfin and Whitt regime. We propose the minimum-expected-delay load-balancing (MED-LB) routing policy to balance the utilizations of all the servers in a distributed system with no unnecessary idling. We show that this policy balances both the long-run and finite-time average utilizations over all the server pools in the Halfin and Whitt regime. We next show that, under either the MED-FSF or the MED-LB policy, a distributed system performs as well as the corresponding inverted V-system. Finally, we show that, operating under the MED-LB policy, both the distributed system and the inverted V-system have similar performances to a corresponding M/M/n system. We illustrate the quality of our asymptotic results for several parallel server systems via simulation experiments.

Key words: large-scale service networks; scheduling and routing control; heavy traffic; quality and efficiency driven; distributed systems

MSC2000 subject classification: Primary: 60K25; secondary: 90B22, 60F17
OR/MS subject classification: Primary: queues; secondary: diffusion models, limit theorems, optimization
History: Received November 4, 2005; revised September 7, 2006, and December 21, 2006.

1. Introduction. The processing capacity of a company may have to be divided among geographically dispersed centers because of outsourcing and intentional decentralizing. Parallel server systems with distributed server pools have been commonly used to model such systems; see Teh and Ward [36] and Kogan et al. [26], among others. In this paper, we focus on distributed parallel server systems with many servers. These models arise from applications in the service sector, in particular, from the call center operations of multinational and multistate companies.

The use of call centers as the customer service interface has become standard practice as an effective business tool and as a convenience expected by customers. They are a primary way that companies around the world interact with their customers. They have applications in many industries as they can provide customers a single access point to a set of diverse services. In the past decade, the number of call centers has grown dramatically. There are now about 70,000 call centers in United States alone, employing six million agents (Utchitelle [39]). The call center industry has had growth rates more than 20% annually. Annual expenditures on call centers are estimated to be between $100 billion and $300 billion with anywhere from 50% to 75% of this cost being devoted to labor. See Aksin and Harker [1] and Gans et al. [19], and references in the latter paper for more background on call centers.

As the volume and types of calls handled in call centers increased, the policies and procedures governing the use of call centers have become increasingly varied and sophisticated. As a result, the management of call centers has become more and more challenging. One of the complications of managing call centers that arose over time is that many companies have created multiple call centers in diverse locations, either as a result of deliberate decision to decentralize or through mergers and acquisitions. Multinational and multistate companies with operations in various countries and states often need multioffice call centers that are geographically dispersed to meet the requirements of customers in different markets. A multilocation call center is also an ideal solution for companies that want to offshore some or all of their work to areas with cheaper labor. Now, such companies are realizing that there are substantial efficiencies to be gained from operating these distributed centers as a single logical center.

Several different methods are being used by companies with geographically dispersed call centers to route arriving calls to individual call centers. One of the methods is to hold calls centrally in a queue and route individual calls to call centers as servers become free. Another method employs dynamic routing policies that routes a customer based on the state of the system at the time of his arrival to a call center, where he queues when all the servers in that center is busy. The combinations of these methods with periodic load balancing (LB) are also being used Gans et al. [19]. For large call centers with many server pools, having a central queue is not cost effective because it requires that each call center be connected to this central queue (Kogan et al. [26]). Besides, numerical tests in Borst et al. [9] show that dynamic routing policies perform well so that
“a multiple site system approaches quite closely the performance of a single virtual facility.” In this paper, we provide effective routing policies and prove that under these policies, the performance of a dispersed call center is comparable to the performance of the system with all the servers residing in the same location.

It is common in call center practice to assume that arrival rates are constant over short periods of time during the day. Fifteen-, 30-, and 60-minute intervals are commonly used in the industry. Given the arrival rate in an interval, and the structure of a call center, two issues that must be addressed are staffing and routing. Staffing decisions are usually made in advance and routing decisions can be made in real time. In this paper, we address the problem of optimal routing for a given staffing level. However, the approach in Armony and Mandelbaum [5] for the ∧-systems can be used to address the optimal staffing problem in this setting as well.

For a call center with a single customer class and a single server pool, the Erlang C formula is the most common tool used in practice to calculate the required staffing levels to provide a certain level of service quality. This formula dates back to the work of Erlang; see Brockmeyer et al. [11] for a compilation of his studies. The main assumptions are that the arrival process is Poisson, service times are exponential and the system reaches steady state fast enough. Such a fast convergence allows the use of steady-state probabilities to calculate the desired performance measures. The most common performance measure is known as the service level, which is stated in terms of the percentage of customers who have to wait more than a certain amount of time before their service starts. Given the arrival and service rates and the service-level target, the Erlang C formula can be used to find the minimum number of servers required. The Erlang C formula has been extended to include customer abandonments (Bacelli and Hebuterne [6]), and finite trunk (buffer) space (Jagerman [25]), which are commonly known as the Erlang A and the Erlang B formulas, respectively.

Another approach to analyze large service systems is based on the asymptotic regime proposed by Halfin and Whitt [23]. In Halfin and Whitt [23], they consider a sequence of GI/M/n systems with \( \rho_n \to 1 \) as \( n \to \infty \), where \( \rho_n = \lambda_n / (n \mu) \) denotes the traffic intensity, \( \lambda_n \) denotes the arrival rate to the system with \( n \) servers, and \( \mu \) is the service rate of each server. They show that the steady-state probability that all the servers are busy converges to a limit \( 0 < \alpha < 1 \) if and only if \( \sqrt{n}(1 - \rho_n) \to \theta \), for some \( \theta > 0 \), as \( n \to \infty \). Let \( X^n(t) \) denote the total number of customers, including those in service, in the \( n \)th system at time \( t \). They show that the scaled queue length process, \( \hat{X}^n(t) = (X^n(t) - n) / \sqrt{n} \), converges to a diffusion process if \( \hat{X}^n(0) \) converges weakly. They also show that the steady-state distribution of the queue length converges to the steady-state distribution of the limiting diffusion process. The results of Halfin and Whitt [23] also provide a simple expression for the required number of servers, which is known as the square-root safety staffing principle. One can infer from their results that to achieve the service-level target \( \alpha \), the number of servers \( n \) must be set to

\[
n = R + \theta \sqrt{R},
\]

where \( R = \lambda / \mu \) is called the offered load and \( \theta \) is a deterministic function of \( \alpha \).

This asymptotic regime is also known as the quality and efficiency driven (QED) regime since only a small percentage of customers are delayed in queue before their service starts even though servers experience high utilization. Diffusion limits in the QED regime have also been used to analyze more complicated parallel server systems. We refer the interested reader to Borst et al. [8], Armony and Maglaras [3, 4], Armony and Mandelbaum [5], and Gurvich et al. [22], among others.

In this paper, we consider a distributed parallel server system that consists of multiple server pools and a single customer class. Each server pool consists of multiple homogeneous servers and has its own dedicated queue. Customers arrive at the system according to a Poisson process. Each customer must be routed to a server pool or a queue at his arrival time following a routing policy. Under a routing policy, when there are idle servers in the pool, a customer arriving at that instant may be routed for service immediately; otherwise, the customer is routed to the selected queue waiting to be served later. Note that in a system with a single queue, if all the servers are busy at the time of a customer’s arrival, he is routed to a server pool when a server idles. Once a customer receives a service, he leaves the system. The service times at each server pool are assumed to be i.i.d. and exponentially distributed. Our main goal is to construct efficient routing policies so that despite customers are routed upon arrival, the system works as if there is only one queue where customers wait for service. This fact is referred to as the complete resource pooling in the conventional heavy traffic literature; see, for example, Reiman [33].

The MED routing policy is a widely studied and used policy in different applications. Under the MED routing policy, when a customer finds all the servers busy at the time of his arrival, he is routed to the queue with the minimum delay; otherwise, he is routed to one of the server pools with idle servers. The MED policy has been widely studied in the literature. Kogan et al. [26] demonstrates numerically that the distributed pools
system with the MED policy is not inferior to the corresponding system with a central queue, with respect to the stationary waiting time distribution. The MED policy (and a simpler version, join the shortest queue policy) has been shown to achieve complete resource pooling in distributed parallel server systems under conventional heavy traffic; see Foschini and Salz [17], Reiman [33], and Laws [28]. Winston [46] and Weber [41] showed that when service times are exponential or have increasing failure rate, then it is optimal to assign the job to the shortest queue. Whitt [42], however, showed that for other service-time distributions, the shortest queue rule need not be optimal. As noted in §5.3 of Gans et al. [19], “while there is a fairly extensive literature on LB, little of it appears to be directly applicable to distributed parallel server systems with many servers.”

In this paper, we focus on two routing policies: the minimum-expected-delay faster-server-first (MED-FSF) policy and the minimum-expected-delay load-balancing (MED-LB) policy. Under both policies, if all servers are busy when a customer arrives at the system, that customer is routed to the queue that has the minimum expected delay. If there is an idle server at his arrival time, then under the MED-FSF policy, the customer is routed to the fastest available pool and under the MED-LB to the least utilized available pool.

Distributed parallel server systems we study is closely related to the inverted-V-systems studied in Armony [2]. An inverted-V-system, or ∧-system, also consists of multiple server pools and a single customer class. Unlike a distributed server pool system, in an ∧-system there is only one queue and each server can serve a customer waiting in that queue. Because there is only one queue, there is no routing decision to make in an ∧-system when an arriving customer finds all servers busy. Thus the MED-FSF and MED-LB routing policies in our distributed system reduces to the faster-server-first (FSF) and load-balancing (LB) policies in the corresponding ∧-system.

Armony [2] shows that the faster FSF policy is asymptotically optimal in the QED regime in the sense that it minimizes the stationary distribution of the waiting time and queue length processes in the limit as the arrival rate goes to infinity. In this paper, we show that the MED-FSF policy is also asymptotically optimal in the QED regime for our distributed systems. Our optimality result is weaker than that in Armony [2], where they proved that the MED-FSF policy achieves the minimum stationary queue length distribution and minimum stationary probability that a customer gets delayed in the queue before his service starts.

Under the MED-FSF policy, all the servers in our distributed system, except those with the lowest service rate, experience 100% utilization. However, this implies that the most efficient servers are busy almost all the time, whereas the slowest servers enjoy significantly more idle time. It is well known that employee burnout is aggravated by unfairness in the work environment and performance objectives that are not feasible such as 100% utilization; see Levin [29]. Therefore, a common goal in call center management is to have all servers to be utilized fairly. The MED-LB policy achieves this objective. We show that the MED-LB policy asymptotically balances the load of the servers; i.e., the utilizations of all the servers in the system become equal as the offered load gets large. We further show that, operating under the MED-LB policy, the distributed system achieves the resource pooling effect in that the stationary distribution of the total queue length and waiting time processes are approximated by those in an $M/M/n$ system. The arrival rate and total number of servers in the latter system are the same as those in the distributed system. However, the service rate for each server is equal to the average rate among all servers in the distributed system. The performance of distributed systems is usually approximated by this single server pool system in practice. It is shown here for the first time that this approximation is asymptotically correct. The same result also holds for ∧-systems under the LB policy. One can come up with other policies that would yield balanced utilizations across server pools in our distributed system but the MED-LB policy is easy to implement and its performance can be accurately estimated using the Erlang C formula. We also discuss how the MED-LB policy can be modified to distribute the total available idle time in desired proportions among all server pools.

Using our asymptotic results, we derive approximations for the performances of distributed systems under the MED-FSF and MED-LB policies. Because all our results are asymptotic, we conduct simulation experiments to illustrate the accuracy of our results. We compare the performance of distributed systems under the MED-FSF and MED-LB policies with that of the corresponding ∧-systems and test the accuracy of the approximations obtained from the asymptotic analysis in several distributed systems. We conduct additional simulation experiments to test if the MED-LB and LB policies balance the utilizations of servers in relatively small systems. The simulation results show that our asymptotic results are also observed in systems with sizes comparable to existing call centers and asymptotic approximations provide accurate estimates for the stationary delay probability and expected waiting time.

The main results of this paper can be summarized as follows:

(i) Under the MED-FSF policy, the stationary delay probability and stationary queue lengths are asymptotically minimized among all admissible policies. Furthermore, under the MED-FSF policy, the distributed system performs as well as a corresponding ∧-system in terms of waiting times and queue lengths in this system.
(ii) The MED-LB policy in a distributed system asymptotically balances the utilizations of the server pools. Also, under the MED-LB policy, a distributed system performs as well as the corresponding \( \land \)-system. Both systems perform as well as a corresponding \( M/M/n \) system.

(iii) As the above results are derived through many server diffusion limits, we obtain formulas for approximate performance analysis of a distributed system under both the MED-FSF and MED-LB policies.

While establishing the diffusion limits, we use the framework built in Dai and Tezcan [15] to prove state space collapse (SSC) results in the many-server heavy-traffic analysis of parallel server systems. In that paper, the SSC framework of Bramson [10] is extended from the conventional heavy-traffic limiting regime to the many-server asymptotic regime. Dai and Tezcan [15] provided a systematic way to prove SSC results for parallel server systems. This paper also presents an application of their result.

Gamarnik and Zeevi [18] recently showed that the stationary distribution of generalized Jackson networks converges to that of the limiting diffusion process in the conventional heavy-traffic regime. We use their results, in particular, Theorem 5 in that paper, to validate the so-called “interchange-of-limits” to prove the convergence of the stationary distribution to that of the limiting diffusion process in this setting.

We close this section with a brief summary of the related literature. As described above, the QED regime provided a general framework for the asymptotic analysis of large service systems. The analysis in Halfin and Whitt [23] is extended to include customer abandonment in Garnett et al. [20]. They have also explored two more parametric regimes in addition to the QED regime: efficiency driven and quality driven (QD) regimes. For efficiency and QD systems \( \alpha \approx 1 \) and \( \alpha \approx 0 \), respectively, where \( \alpha \) is the limiting stationary probability that an arriving customer is delayed in the queue. Puhalskii and Reiman [32] established the diffusion limit of a \( G/PH/n \) system, where \( PH \) stands for a phase-type service time distribution. Whitt [45] studies the many-server diffusion limit of a \( G/H_\infty^n/m \) system, where \( H_\infty^j \) indicates that the service time distribution is an extremal distribution among the class of hyperexponential distributions. He later uses this analysis in Whitt [44] to approximate \( G/GI/n/m \) systems.

Optimal staffing and control of large service systems under the QED regime is a new and dynamic research field. Armony and Maglaras [4, 3] study an \( M/M/n \) system with customers having the option to choose to be called back within a certain amount of time instead of waiting in the queue. They show that a threshold policy that changes the priority of the classes based on queue lengths is asymptotically optimal in the QED regime in the sense that it minimizes real-time delays while guaranteeing the delay bound of the postponed service. Armony [2] studies \( \land \)-parallel server systems under the QED regime. As explained above, she shows that the FSF policy is asymptotically optimal. In Armony and Mandelbaum [5], they show that the FSF policy with a staffing rule similar to the square-root safety rule is asymptotically optimal in the QED regime. In Gurvich et al. [22] (also see Gurvich [21]), they consider \( \lor \)-systems. They show that a threshold-type policy with square-root safety rule is asymptotically optimal with respect to several objectives under the QED regime. Recently, Tezcan and Dai [38] show that a greedy scheduling rule is asymptotically optimal for \( N \)-systems in the QED regime.

There is a substantial literature on the analysis of \( \land \)-systems under conventional heavy traffic with one or more customer classes, e.g., see Van Mieghem [40], Harrison and López [24], Stolyar [34], Mandelbaum and Stolyar [30], and the references in the last paper. Distributed systems with multiple customer classes under conventional heavy traffic are studied by Stolyar [35]. He shows, under the MinDrift routing rule that these systems perform as well as the corresponding \( \land \)-system operating under an optimal routing policy with multiple customer classes.

The rest of this paper is organized as follows. In §2, we present the details of the queueing model we consider. We also present the asymptotic regime we consider and explain the details of the control policies we analyze. We summarize our main results in §3. In §4, we present the results of the simulation experiments conducted. We prove our main results in §5. In §6, we discuss how the MED-LB policy can be extended to distribute the total available idle time among the server pools in desired proportions.

### 2. The queueing model and the asymptotic framework.

As described in the previous section, we consider a distributed parallel server system that consists of \( J \geq 2 \) server pools and a single customer class. Server pool \( j \) consists of \( N_j \) homogeneous servers and has its own dedicated queue. Customers arrive at the system according to a Poisson process with rate \( \lambda \). Each customer must be routed to a server pool or a queue at his arrival time following a certain routing policy. Under a routing policy, an arriving customer may be routed to an idle server if there are idle servers at the time of his arrival or to a queue waiting to be served later. Once a customer receives service, he leaves the system. The service time of each server in pool \( j \) is assumed to be exponentially distributed with rate \( \mu_j \). A server residing in server pool \( j \) can only handle customers who are routed to the \( j \)th queue. Once a customer joins a queue, he cannot swap to other queues nor can he renege. Customers in the same queue
are served on a first-in-first-out (FIFO) basis. We refer to such a system as a distributed parallel server system or a distributed system. The corresponding ∨-system of a distributed system has the same parameters with the distributed system, except that it has only one queue, and a customer who finds all the servers busy at the time of his arrival, is routed to this queue waiting to be served by one of the servers in the system later.

The customers who are routed to the jth queue or server pool are called class j customers. For notational convenience, we define \( \mathcal{J} = \{1, \ldots, J\} \) the set of server pools. Because each customer class is associated with a unique server pool, this set will also give the set of indices for the customer classes.

Without loss of generality, we assume that

\[
\mu_1 \leq \mu_2 \leq \cdots \leq \mu_J
\]

and set \( \mu = (\mu_1, \ldots, \mu_J) \). We set \( N = (N_1, \ldots, N_J) \) and use \( |N| = \sum_{j=1}^J N_j \) to denote the total number of servers in the system.

We use \( Z_j(t) \) to denote the total number of customers being served by a server in pool \( j \), \( Q_j(t) \) to denote the total number of customers in queue \( j \), and \( X(t) \) to denote the total number of customers in the system at time \( t \). We assume that nonidling condition holds at time 0, i.e., \( Q_j(0)|Z_j(0)| = 0 \). We set \( Z_j = \{Z_j(t), t \geq 0\} \), \( Q_j = \{Q_j(t), t \geq 0\} \), and \( X = \{X(t), t \geq 0\} \). When we need to express the dependence of a stochastic process \( Y \) arising in a queueing system on the routing policy used, say \( \pi \), we write \( Y(\cdot, \pi) \).

The dynamics of a parallel server system are controlled by a routing policy. A routing policy must specify how to dispatch a server to serve a customer. Unless stated otherwise, for all the systems considered in this paper, we assume that when a server finishes service and there are customers waiting in the queue, he cannot idle and serves the longest waiting customer in the queue. A routing policy is said to be head of the line (HL) if it satisfies the latter condition. The routing policy also determines how an arriving job is routed to a queue or a server pool. A routing policy is said to be nonidling if it routes an arriving customer to one of the idle servers if there are any available at the time of the customer. We only consider nonpreemptive policies. However, we note that allowing preemption under a policy will not change the average waiting time in the system.

A routing policy is said to be admissible if while making a decision at time \( t \), the policy only uses the information captured by \( \{Q(t), Z(t)\} \). We focus on policies that are admissible. The class of policies we consider can be extended as discussed in Remark 5.1.

We focus on two nonidling routing policies: the MED-LB and the MED-FSF policies. Under the MED policy when a customer arrives to the system to find all the servers busy, he is routed to the queue with the minimum expected delay (or waiting time). The expected waiting time in queue \( j \) at time \( t \) is defined by \( Q_j(t)/|N_j \mu_j \). We will show this quantity agrees with the actual waiting time of a customer who sees \( Q_j(t) \) customers at the time of his arrival when the offered load is large. Under both the MED-LB and MED-FSF policies, if all servers are busy when a customer arrives at the system, the customer is routed to one of the queues according to the MED policy. If there is an idle server at his arrival time, then under the MED-FSF policy, the customer is routed to the fastest available pool, and under the MED-LB policy, to the least utilized available pool, where the utilization of the server pool \( j \) is given by \( Z_j(t)/N_j \). We assume that ties are broken arbitrarily.

In an ∨-system, there is only one queue, hence there is no routing decision to make when an arriving customer finds all servers busy. Thus the MED-FSF and MED-LB routing policies in our distributed system reduces to the FSF and LB policies in the corresponding ∨-system.

For the asymptotic analysis of the these systems, we consider a sequence of systems indexed by \( r \). The arrival rate to the \( r \)th system, \( \lambda^r \), is equal to \( r \). We append "\( r \)\" to the processes that are associated with the \( r \)th system, e.g., \( Q_j^r(t) \) is used to denote the number of class \( j \) customers in the queue in the \( r \)th system at time \( t \).

The service rates are held fixed for all the systems in this sequence. The number of servers in each pool in the \( r \)th system is given by \( N^r = (N^r_1, \ldots, N^r_J) \) and as before, the total number of servers in the \( r \)th system is denoted by \( |N^r| \). For simplicity, we assume that

\[
N^r_j = \lceil \beta_j|N^r| \rceil \quad \text{for all } j \in \mathcal{J},
\]

where \( \beta_j > 0 \) is given for each \( j \in \mathcal{J} \) with \( \sum_{j=1}^J \beta_j = 1 \), and for a real number \( x \), \( \lceil x \rceil \) is the least integer greater than or equal to \( x \). (It is actually enough to assume that \( N^r_j/|N^r| \to \beta_j \) as \( r \to \infty \) for all \( j \in \mathcal{J} \) for the results in this paper to hold.) We define the average service rate \( \mu \) across all the servers by

\[
\mu = \sum_{j=1}^J \beta_j \mu_j.
\]
Let the traffic intensity for the \( r \)th system be defined by
\[
\rho' = \frac{\lambda'}{\sum_{j \in J} N'_i \mu_j}.
\]
We assume that
\[
\sqrt{|N'|}(1 - \rho') \to \frac{\theta}{\sqrt{\mu}} \quad (4)
\]
for some \( \theta > 0 \). Assumption (4) implies that the system reaches heavy traffic as \( r \to \infty \).

We define the diffusion-scaled queue length, \( \hat{Q}'_j(t) \), the diffusion-scaled number of customers in service, \( \hat{Z}'_j(t) \), and the diffusion-scaled total number of customers in the system, \( \hat{X}'(t) \), by
\[
\hat{Q}'_j(t) = \frac{Q'_j(t)}{\sqrt{|N'|}}, \quad \hat{Z}'_j(t) = \frac{Z'_j(t) - N'_j}{\sqrt{|N'|}}, \quad \text{and} \quad \hat{X}'(t) = \frac{X'(t) - |N'|}{\sqrt{|N'|}} \quad (5)
\]
for all \( j \in J \). We analyze the weak limits of \( \hat{Q}'_j, \hat{Z}'_j, \) and \( \hat{X}'(t) \) as \( r \to \infty \).

Let \( W'(t) \) denote the amount of time a customer will wait before his service starts if he arrives at time \( t \) and \( W' = \{ W'(t) : t \geq 0 \} \). The process \( W' \) is known as the virtual waiting time process. We define the diffusion-scaled virtual waiting time process \( \hat{W}'(t) \) by
\[
\hat{W}'(t) = \sqrt{|N'|}W'(t). \quad (6)
\]

We are also interested in the asymptotic behavior of the stationary distribution of \( (\hat{Q}', \hat{Z}', \hat{X}', \hat{W}') \) as \( r \to \infty \). For a routing policy \( \pi \), we denote the stationary probability distribution of \( (\hat{Q}', \hat{Z}', \hat{X}', \hat{W}') \) by \( \mathbb{P}_\pi \) when it exists. For notational convenience, we denote by
\[
(\hat{Q}'(\infty, \pi), \hat{Z}'(\infty, \pi), \hat{X}'(\infty, \pi), \hat{W}'(\infty, \pi))
\]
a random variate that has distribution \( \mathbb{P}_\pi \). We call \( \mathbb{P}(W'(\infty) > 0) \) the stationary delay probability in the \( r \)th system. If a stationary distribution of a process \( Y \) does not exist, we set
\[
\mathbb{P}[Y(\infty) > x] = 1 \quad (7)
\]
for all \( x \in \mathbb{R} \).

3. Main results. Our main results are based on the asymptotic analysis of the stochastic process \( (\hat{Q}', \hat{Z}', \hat{X}', \hat{W}') \) and its stationary behavior as \( r \to \infty \). The proofs of the results in this section are presented in §5.6.

We first focus on the MED-FSF policy and show that it minimizes the stationary distribution of the queue lengths and the stationary delay probability among all admissible policies as described in the following theorem.

THEOREM 3.1. Consider a sequence of MED-FSF distributed server systems. Assume that (2) and (4) hold. Then, for any admissible routing policy \( \pi \)
\[
\lim_{r \to \infty} \mathbb{P}[\hat{X}'(\infty, MED-FSF) > x] \leq \lim_{r \to \infty} \inf \mathbb{P}[\hat{X}'(\infty, \pi) > x] \quad (8)
\]
for all \( x \in \mathbb{R} \) and
\[
\lim_{r \to \infty} \mathbb{P}[\hat{W}'(\infty, MED-FSF) > 0] \leq \lim_{r \to \infty} \inf \mathbb{P}[\hat{W}'(\infty, \pi) > 0]. \quad (9)
\]

In Theorem 3.1, we only require that \( \pi \) is admissible and do not assume that it is nonidling or serves customers on a FIFO basis. Theorem 3.1 is proved by comparing the limit of the sequence of the stationary distributions of the distributed systems with that of the corresponding \( ^\wedge \) -system. We show that the MED-FSF policy achieves the same asymptotic performance as it does in an identical \( ^\wedge \) -system. Using this result and the asymptotic optimality of the FSF policy in an \( ^\wedge \) -system, we prove that the MED-FSF policy is asymptotically optimal as described in Theorem 3.1.

Let \( X'_\cdot(t) \) be the number of customers in the corresponding \( ^\wedge \) -system at time \( t \) and
\[
\hat{X}'\cdot(t) = \frac{X'_\cdot(t) - |N'|}{\sqrt{|N'|}}. \quad (10)
\]
We denote the virtual waiting time process in these systems by \( W'_\cdot \) and define the diffusion-scaled waiting process by \( \hat{W}'\cdot(t) = \sqrt{|N'|}W'_\cdot(t) \). As a by-product of the proof of Theorem 3.1, we have the following result.
**Theorem 3.2.** Consider a sequence of MED-FSF distributed server systems and the sequence of corresponding FSF systems. Assume that (2) and (4) hold. Then,

\[ \lim_{r \to \infty} \mathbb{P}\{ \hat{X}^r(\infty, \text{MED-FSF}) > x \} = \lim_{r \to \infty} \mathbb{P}\{ \hat{X}^r_{\text{FSF}}(\infty, \text{FSF}) > x \} = F(x) \]  

for all \( x \in \mathbb{R} \) and

\[ \lim_{r \to \infty} \mathbb{P}\{ \hat{W}^r(\infty, \text{MED-FSF}) > w \} = \lim_{r \to \infty} \mathbb{P}\{ \hat{W}^r_{\text{FSF}}(\infty, \text{FSF}) > w \} = F(\mu w) \]  

for all \( w \geq 0 \), where \( F(x) = \int_{-\infty}^{x} f(u)du \) and \( f \) is the density function defined by

\[
f(x) = \begin{cases} \frac{\theta}{\sqrt{\mu}} \exp\{-\theta x/\sqrt{\mu}\} \alpha, & \text{if } x \geq 0 \\ \sqrt{\frac{\mu_1}{\mu}} \phi\left(\frac{\mu_1}{\mu} x + \frac{\theta}{\sqrt{\mu_1}}\right) (1 - \alpha), & \text{if } x < 0, \end{cases}\]  

where

\[
\alpha = \left[ 1 + \frac{\theta/\sqrt{\mu_1} \Phi(\theta/\sqrt{\mu_1})}{\phi(\theta/\sqrt{\mu_1})} \right]^{-1}. \]  

**Remark 3.1.** Armony [2] shows that the FSF routing policy is asymptotically optimal for the systems in the sense that

\[ \lim_{r \to \infty} \mathbb{P}\{ \hat{W}^r(\infty, \text{FSF}) > w \} \leq \liminf_{r \to \infty} \mathbb{P}\{ \hat{W}^r(\infty, \pi) > w \} \]  

for all \( w \geq 0 \) and any admissible HL policy \( \pi \). Note that this is stronger than (9) since our result only holds for \( w = 0 \). The main reason is that for systems working under an HL policy, (8) and (9) imply (15) since customers are served on a FIFO basis. In a distributed system, a server can idle even though there are customers in other queues, and a customer arriving at that instant overtakes the customers that are already in queue and starts his service before them. Therefore, customers in a distributed system under, for example, a nonidle routing policy are not served on a FIFO basis. Hence (8) and (9) do not imply (15) for distributed systems.

Next, we obtain approximations for the performance of the \( r \)th system under the FSF and MED-FSF policies using (11) and (12).

**Corollary 3.1.** Consider a sequence of MED-FSF distributed server systems. Under the assumptions of Theorem 3.2

\[ \mathbb{P}\{ W^r(\infty) > 0 \} \to \alpha \quad \text{and} \quad \mathbb{E}\{ \hat{W}^r(\infty) \} \to \frac{\alpha}{\sqrt{\mu_1} \theta} \]  

as \( r \to \infty \), where \( \alpha \) is given by (14) and \( \mu \) is given by (3).

**Remark 3.2.** Armony [2] also hold for the sequence of corresponding FSF systems.

Now, we focus on the distributed systems operating under the MED-LB policy and the corresponding systems. Note that in order for \( \hat{Q}^r \) to have a meaningful limit

\[ \frac{Z^r_i(\cdot)}{N^r_i} \to 1 \quad \text{a.s. u.o.c.} \]  

as \( r \to \infty \), as otherwise one can show that \( \hat{Q}^r(\cdot) \to \infty \) a.s. u.o.c. as \( r \to \infty \). Therefore, under any policy satisfying (17)

\[ \left| \frac{Z^r_i(\cdot)}{N^r_i} - \frac{Z^r_j(\cdot)}{N^r_j} \right| \to 0 \quad \text{a.s. u.o.c.} \]  

as \( r \to \infty \) for \( i, j \in \mathcal{J} \).

The result (17) barely sheds any light on the actual differences between the utilizations of the servers as it is illustrated in the simulation experiments in the next section. Therefore we are interested in establishing a stronger result. We first define what we mean by LB in mathematical terms. For consistency with our preceding analysis, we focus on the steady state of a distributed system but a similar definition for finite time intervals can also be considered. We say that a policy asymptotically balances the steady-state utilizations of the server pools if

\[ \lim_{r \to \infty} \sqrt{|\mathcal{J}|} \mathbb{E}\left[ \left| \frac{Z^r_i(\infty)}{N^r_i} - \frac{Z^r_j(\infty)}{N^r_j} \right| \right] = 0 \]  

for all pairs \( i, j \in \mathcal{J} \).
Next, we show that the MED-LB policy asymptotically balances the load of the servers in a distributed system.

**Theorem 3.3.** Consider a sequence of MED-LB distributed server systems. Assume that (2) and (4) hold. Then, (18) holds. If in addition

\[(\hat{Q}'(0), \hat{Z}'(0)) \to (\hat{Q}(0), \hat{Z}(0)),\]

as \(r \to \infty\) for a random vector \((\hat{Q}(0), \hat{Z}(0))\), and

\[\left| \frac{\hat{Z}'_i(0)}{\beta_i} - \frac{\hat{Z}'_j(0)}{\beta_j} \right| \to 0\]

for all \(i, j \in \mathcal{J}\) in probability as \(r \to \infty\), then for any \(T > 0\),

\[\sqrt{|N|} \left\| \frac{Z'_i(t)}{N'_i} - \frac{Z'_j(t)}{N'_j} \right\|_T \to 0 \quad \text{in probability as } r \to \infty\]

in probability as \(r \to \infty\) for \(i, j \in \mathcal{J}\).

**Remark 3.3.** It can be shown that the results of Theorem 3.3 also hold for the sequence of corresponding LB \(\wedge\)-systems.

Next, we look at the asymptotic differences between the utilizations under the MED-FSF policy to put the previous result in perspective.

**Theorem 3.4.** Consider a sequence of MED-FSF distributed server systems. Under the assumptions of Theorem 3.2, for \(i = 2, 3, \ldots, J\)

\[\lim_{r \to \infty} \sqrt{|N|} \mathbb{E} \left[ \frac{Z'_i(\infty)}{N'_i} - \frac{Z'_j(\infty)}{N'_j} \right] = \mathbb{E}[(X(\infty))^+],\]

where \(x^- = \min\{x, 0\}\) and \(X(\infty)\) is a random variable with distribution \(F\) defined in Theorem 3.2.

We next prove that, under the MED-LB policy, a distributed system performs as well as the corresponding \(\wedge\)-system and both systems perform as well as a corresponding \(M/M/n\) system.

Consider a sequence of \(M/M/n\) systems with the arrival rate and the number of servers in the \(r\)th system is equal to those of the \(r\)th distributed system. Assume that the service rate of each server in this system is equal to the average service rate \(\mu\) in the distributed system given by (3). Let \(X'(t)\) denote the total number of customers in the \(r\)th \(M/M/n\) system at time \(t\). We define the diffusion-scaled total number of customers process in these systems by

\[\hat{X}'(t) = \frac{X'(t) - |N'|}{\sqrt{|N'|}}.\]

We use \(\hat{X}'(\infty)\) to denote the weak limit of \(\hat{X}'(t)\) as \(t \to \infty\), which exists for each \(r\) by (4) and standard results on the existence of a stationary distribution of an \(M/M/n\) system. Let \(\hat{W}'(t)\) denote the virtual waiting time for the \(r\)th single server system; \(\hat{W}'(t) = \sqrt{|N'|}W'(t)\) and \(\hat{W}'(\infty)\) denote the weak limit of \(\hat{W}'(t)\) as \(t \to \infty\).

**Theorem 3.5.** Consider a sequence of MED-LB distributed systems and the sequence of corresponding LB \(\wedge\)-systems. Assume that (2) and (4) hold. Let \(X'\) and \(W'\) be defined as above. Then,

\[\lim_{r \to \infty} \mathbb{P}\{\hat{X}'(\infty) > x\} = \lim_{r \to \infty} \mathbb{P}\{\hat{X}'(\infty) > x\} = \lim_{r \to \infty} \mathbb{P}\{\hat{X}'(\infty) > x\} = F(x)\]

for all \(x \in \mathbb{R}\) and

\[\lim_{r \to \infty} \mathbb{P}\{\hat{W}'(\infty) > w\} = \lim_{r \to \infty} \mathbb{P}\{\hat{W}'(\infty) > w\} = \lim_{r \to \infty} \mathbb{P}\{\hat{W}'(\infty) > w\} = F(\mu w)\]

for all \(w \geq 0\), where \(F(x) = \int_{-\infty}^{x} f(u)du\) and \(f\) is the density function defined by

\[f(x) = \begin{cases} \frac{\theta}{\sqrt{\mu}} \exp\{-\theta x/\sqrt{\mu}\} \alpha, & \text{if } x \geq 0 \\ \phi \left( x + \frac{\theta}{\sqrt{\mu}} \right) - \Phi \left( \frac{\theta}{\sqrt{\mu}} \right), & \text{if } x < 0 \end{cases}\]
where
\[ \alpha = \left[ 1 + \frac{(\theta / \sqrt{\mu})\Phi(\theta / \sqrt{\mu})}{\phi(\theta / \sqrt{\mu})} \right]^{-1}. \]

**Remark 3.4.** One can show by using our results that the results of Corollary 3.1 also hold for the distributed systems operating under the MED-LB policy and the corresponding \(^{-}\)-systems with \(\alpha\) given by (27).

### 4. Simulation experiments.
Because the results presented in §3 are asymptotical results, in this section, we conduct simulation experiments to evaluate the quality of those results. We consider five cases. In each case, we simulate a distributed system and the corresponding \(^{-}\)-system. The parameters of all five cases are displayed in Table 1 (the time unit is taken to be one minute.)

The first four cases correspond to systems that have three server pools, and the last case corresponds to a system that has eight server pools. The parameters of the first three cases are selected to investigate the effect of the offered load, defined by \(\lambda / \mu\), on the quality of our results. We set the arrival rate in the second and third cases to be 10 and 40 times the arrival rate in the first case, respectively, to observe this effect. Balanced server assignment among all pools may affect the quality of our asymptotic results. To observe the effect of unbalanced server staffing, in the fourth case, one of the server pools is set to have significantly fewer servers than the other pools. Finally, in the last case, we consider a system with eight pools to observe the effect of the number of pools on the quality of our results.

In all the simulation experiments, each performance measure estimate is presented with its 95% confidence interval. The length of each simulation run is selected to allow 12 million customers to arrive to the system. Also, a warm-up period of 1.2 million customer arrivals is used. We divide the total simulation length to 10 time intervals of equal length to apply batch means technique; see Law and Kelton [27]. The confidence intervals that are reported along with the estimates are obtained using the batch means. Also, when two or more policies are compared, the simulations are run using common random number generators, so that the service requirement of the \(n\)th customers and interarrival times in all the simulations are the same for \(n = 1, 2, \ldots\).

#### 4.1. Simulation results of the MED-FSF policy.
In this section, we focus on the FSF policy. Theorem 3.2 says that when the offered load is high, a distributed system operating under the MED-FSF routing policy has a similar performance to the corresponding \(^{-}\)-system operating under the FSF routing policy. In this section, we test this result in five systems with different parameters.

Table 2 displays simulation results as well as analytical approximations for all five cases. The results under the MED-FSF policy are displayed in the left half of Table 2. (The results under the MED-LB policy, displayed in the right half of Table 2, will be discussed in the next section.) For each case, the simulation estimates of the delay probability \((P(W > 0))\), in percentage, and the average waiting time \((E[W])\) in seconds are presented for the distributed system, in row DS, and for the corresponding \(^{-}\)-system, in row \(^{-}\). The half-widths of confidence intervals are presented in parentheses next to the simulation estimates. The differences between the estimates from the distributed system and the ones from the corresponding \(^{-}\)-system are presented in row DS – \(^{-}\).

We first note that as the offered load gets high, the performance of the distributed system under the MED-FSF becomes very close to that of the corresponding \(^{-}\)-system. In the first case, the percentage difference between the estimated delay probability in these two models is around 3.6%, in the second case, it decreases to 1.5%, and decreases to less than 1% in the third case. The differences of the expected waiting times are even smaller.

By comparing the results of the fourth case with those of the second case, we observe that having a significantly smaller server pool does not affect the percentage differences. However, the number of server pools in the system has a big impact. The percentage difference between the delay probabilities are twice as much as the difference in the first case, which is the second largest difference among all the cases.

Formulas in (16) give analytical approximations for delay probability and average waiting time. For each parameter case, these approximate values are presented in row Approx. The differences between the analytical and simulation results are given in columns DS – Approx.

<table>
<thead>
<tr>
<th>Case</th>
<th>(J)</th>
<th>(\lambda)</th>
<th>(N)</th>
<th>(\mu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>50</td>
<td>(13, 7, 9)</td>
<td>(1.48, 1.77, 2.4)</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>500</td>
<td>(125, 63, 89)</td>
<td>(1.48, 1.77, 2.4)</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2,000</td>
<td>(497, 255, 347)</td>
<td>(1.48, 1.77, 2.4)</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>500</td>
<td>(195, 95, 22)</td>
<td>(1.48, 1.77, 2.4)</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>500</td>
<td>(45, 45, 68, 70, 76, 81, 112, 126)</td>
<td>(0.72, 0.95, 0.85, 0.8, 0.86, 0.9, 0.88, 0.67)</td>
</tr>
</tbody>
</table>
approximations and simulation estimates from the distributed systems are presented in row DS–Approx. In the first three cases, the percentage difference of delay probability is at most 0.33%. Even in Case 4 when the server pools are not balanced, the approximation for delay probability is quite close to the simulation estimate (the difference is 1.2%). In Case 5 when there are eight server pools, and the arrival rate is equal to 500 as in the second case, the approximation performs significantly worse, 5.1% in percentage difference. We observed in the simulation results that the probability of overtaking, defined in Remark 3.1, increases with the number of server pools. Hence the performance of the approximation for the delay probability diminishes as the number of servers increase. However, we note that the difference in the average waiting times is only 0.05 seconds. Therefore, unlike the delay probability, the performance of the approximation for the waiting time does not seem to suffer from increasing the number of server pools.

From the simulation results, it is clear that the distributed systems operating under the MED-FSF policy do perform as well as the corresponding ∧-systems in terms of the delay probability and average waiting time in the queue.

4.2. Simulation results of the MED-LB policy. In this section, we focus on the MED-LB routing policy. Recall that the simulation results for the distributed systems operating under the MED-LB routing policy and for the corresponding ∧-systems operating under the LB routing policy are presented in the right half of Table 2.

Theorem 3.5 asserts that when the offered load is high, the performance of a distributed system operating under the MED-LB policy is similar to that of the corresponding ∧-system under the LB policy. This is clearly observed in the first four cases; the largest percentage difference between delay probabilities is 1.5% and the largest difference between average waiting times is 1.14 seconds. In Cases 2–4, the difference in waiting times is less than 0.1 seconds. The number of server pools has a big impact on the percentage difference of the delay probability for the systems under MED-LB as well; the percentage difference is 6.2% in Case 5, four times more than the second largest difference. However, the difference in waiting time is again very small; only 0.01 seconds.

Next, we assess the quality of the approximations provided by Remark 3.4 and (16) using the simulation results. First, we note that in Cases 2–5, the differences between the average waiting times estimated for the distributed system and the ones provided by the approximation are less than 0.1 seconds. This difference is
1.14 seconds in the first experiment. Therefore, in terms of the average waiting time, approximations provide very accurate estimates. The percentage difference of the delay probabilities is less than 1.7% in the first four cases, but in Case 5, it is significantly larger (6.4%). Again, this is because of the increase in the overtaking probability when the number of server pools increases.

We conclude that the distributed systems operating under the MED-LB policy performs as well as the corresponding ∨-systems operating under the LB policy.

Theorem 3.5 asserts that the difference between the utilizations of the servers is $o(1/\sqrt{|N|^r})$ when the offered load is high for the distributed systems operating under the MED-LB policy. To evaluate the quality of this asymptotic result, we conduct additional simulation experiments. Also, we simulate the same systems under the MED-FSF policy for comparison.

We consider three cases with two server pools. The parameters of these experiments are selected to investigate the effects of different service rates and unbalanced staffing levels on the difference of utilizations. The values of the parameters for these cases are displayed in Table 3. The first case is a homogeneous system in the sense that the number of servers in each pool is the same and the service rates of all the servers are equal. In the second case, we set the service rate of the first server pool to be two-thirds of the service rate of the second server pool. In the third case, we set the number of servers in the first pool to be significantly less than that in the second pool, and set the service rates as in the second case. The first system is not simulated under the FSF policy because it would yield similar results with the LB policy under a reasonable tie breaker rule because the service times are equal.

Table 4 presents the results of the simulation experiments. We also consider the corresponding ∨-model in each case. We display the percentage difference between the average utilization of the first and second server pools in ∨-systems under the ∨ column and for the distributed systems under the DS column. As before, the left side of the tables is reserved for the MED-FSF policy, and the right side for MED-LB policy. For all estimates, we show the half-width of the 95% confidence intervals in parentheses next to the simulation estimates.

Observe that under the LB and MED-LB policies, the percentage differences between the utilizations of the servers in all the systems are very small; less than 0.8%. Hence, even for these relatively small systems, the LB policy seems to balance the load of the server pools. We observe that the differences in the utilizations are more in the systems with different service rates and unbalanced staffing levels than that in the homogenous system. Also, the differences are slightly lower in ∨-systems than those in the distributed systems. When we compare the percentage differences under the FSF and MED-FSF policies with the LB and MED-LB policies, we see that they increase by about 10 times. This verifies once again how the FSF policy may be unfair in routing calls to server pools.

### 5. Proofs of the main results.

In this section, we prove the results stated in §3. In §5.1, we discuss the dynamics of distributed systems to mathematically characterize the behavior of these systems. The notation used in the rest of the paper is introduced in §5.2.

First, we provide an outline of the proofs. To prove our main results, we first establish the weak limits of $\hat{Q}'$ and $\bar{Z}'$. Once we prove weak convergence, then we establish the convergence of stationary distributions.

To prove weak convergence, we need to prove SSC results under each policy. We use the framework in Dai and Tezcan [15] to prove these results. This framework consists of three steps. First, we establish the invariant state of the fluid limits of distributed systems under each policy. Then, we focus on the hydrodynamic limits of

<table>
<thead>
<tr>
<th>Case</th>
<th>$J$</th>
<th>$\lambda$</th>
<th>$N$</th>
<th>$\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>97</td>
<td>(50, 50)</td>
<td>(1, 1)</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>97</td>
<td>(50, 50)</td>
<td>(0.8, 1.2)</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>97</td>
<td>(29, 64)</td>
<td>(0.8, 1.2)</td>
</tr>
</tbody>
</table>

Table 4. The percentage differences between utilizations.

<table>
<thead>
<tr>
<th>Case</th>
<th>MED-FSF</th>
<th>MED-LB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\wedge$</td>
<td>$\Delta$</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>5.20 (0.11)</td>
<td>5.15 (0.11)</td>
</tr>
<tr>
<td>8</td>
<td>6.93 (0.15)</td>
<td>6.81 (0.15)</td>
</tr>
</tbody>
</table>
distributed systems under each policy and show that corresponding SSC results hold in hydrodynamic limits. In the last step, we show that \( \hat{Q} \) and \( \hat{Z} \) satisfy a compact containment condition.

The remainder of this section is organized as follows. In §5.3, we provide an asymptotic upper bound for \( Q'(t) \) and an asymptotic lower bound for \( Z'_r(t) - N'_r \) in a sequence of distributed systems operating under a non-idling routing policy. These bounds are later used to show that \( \hat{Q}(t) \) and \( \hat{Z}(t) \) satisfy a compact containment condition. They also play an important role in proving the convergence of stationary distributions.

In §5.4, we analyze the limits of the fluid scaled processes, \( Q'(\cdot)/|N'| \) and \( Z'(\cdot)/|N'| \), as \( r \to \infty \) and establish the invariant states of the fluid limits. In §5.5, we study the weak limits of the diffusion scaled processes \( \hat{Q}(\cdot) \) and \( \hat{Z}(\cdot) \) of a distributed system operating under the MED-FSF or the MED-LB policies as \( r \to \infty \). The second and third steps to prove an SSC result are included in this section. We then establish the convergence of the stationary distribution of \( \hat{X} \) as \( r \to \infty \). Finally, we provide the proofs of Theorems 3.1, 3.2, 3.3, and 3.5 and Corollary 3.1 in §5.6 using the results in §§5.3–5.5.

### 5.1. The dynamics of distributed parallel server systems.

In this section, we give the details of the dynamics of the distributed parallel server systems. We also present the equations that must be satisfied by the systems operating under a non-idling routing policy, and the MED-FSF and MED-LB policies. This section lays the mathematical framework for the analysis of the distributed systems.

We follow the notation used in Dai and Tezcan [15]. Specifically, we use \( A'(t) \) to denote the total number of customer arrivals by time \( t \) in the \( r \)th system and set \( A' = [A'(t), t \geq 0] \). Let \( A_{j'}^{q'\cdot}(t) \) to denote the total number of customers who are routed to queue \( j \) at the time of their arrival and who had to wait in the queue before their service started before time \( t \) in the \( r \)th system and \( A_{j'}^{q'\cdot} = (A_{j'}^{q'\cdot}(t), t \geq 0) \). We use \( B'_{j'}^{q'\cdot}(t) \) to denote the total number of customers who have been routed to queue \( j \) and started their service immediately after their arrival at server pool \( j \) before time \( t \) in the \( r \)th system and set \( B_{j'}^{q'\cdot} = (B_{j'}^{q'\cdot}(t), t \geq 0) \). Unlike the general setting in Dai and Tezcan [15], in our distributed parallel server systems, there is only one arrival stream, so we omit the subscripts from the arrival processes associated with the arrival type. We set \( A_{j'}^{\bar{q}'\cdot} = (A_{j'}^{q'\cdot}, \ldots, A_{j'}^{q'\cdot}) \) and \( A_{j'}^{\bar{q}'\cdot} = (A_{j'}^{q'\cdot}, \ldots, A_{j'}^{q'\cdot}) \).

Let \( T_{j'}'(t) \) denote the total time spent serving customers by all \( N_j \) servers of pool \( j \) and \( Y_{j'}'(t) \) denote the total idle time experienced by the servers of pool \( j \) up to time \( t \) in the \( r \)th system, \( T_{j'}' = \{T_{j'}'(t), t \geq 0\} \) and \( Y_{j'}' = \{Y_{j'}'(t), t \geq 0\} \). Recall that we use \( Z_{j'}'(t) \) to denote the total number of customers being served by a server in pool \( j \) and \( Q_{j'}'(t) \) to denote the total number of customers in queue \( j \) at time \( t \) in the \( r \)th system. We denote by \( D_{j'}'(t) \) the total number of customers whose service is completed by a server in pool \( j \) by time \( t \) in the \( r \)th system and set \( D_{j'}' = \{D_{j'}'(t), t \geq 0\} \). We use \( B_{j'}'(t) \) to denote the total number of customers who are delayed in the queue and whose service started in pool \( j \) before time \( t \) in the \( r \)th system and set \( B_{j'}' = \{B_{j'}'(t), t \geq 0\} \). We set \( Z_{j'}' = \{Z_{j'}'(t), t \geq 0\} \), \( Q_{j'}' = \{Q_{j'}'(t), t \geq 0\} \), \( Z' = \{Z_1', \ldots, Z_j'\} \), \( Q' = \{Q_1', \ldots, Q_j'\} \), \( D' = \{D_1', \ldots, D_j'\} \) and \( B' = \{B_1', \ldots, B_j'\} \). Let \( \{S_j, j = 1, \ldots, J\} \) be a set of independent Poisson processes with each process \( S_j \) having rate \( \mu_j > 0 \). We set \( S = \{S_j, \ldots, S_j\} \).

We set \( X_{\pi} = (A', A_{\bar{q}'\cdot}^{q'\cdot}, A_{\bar{q}'\cdot}^{q'\cdot}, Z', T', Y', B', D') \), where \( \pi \) is a nonidling routing policy. We call \( X_{\pi} \) the performance process of the \( \pi \) distributed server pool system. Observe that each component of \( (A', A_{\bar{q}'\cdot}^{q'\cdot}, A_{\bar{q}'\cdot}^{q'\cdot}, Z', T', Y', B', D') \) is nondecreasing and each component of \( (Q', Z') \) is nonnegative. Given \( \pi \), a system as described above should satisfy Equations (2.2)–(12.12) in Dai and Tezcan [15]. In addition, we assume that \( A' \) and \( S \) satisfy the conditions in §2.2 in Dai and Tezcan [15].

**Remark 5.1.** As discussed in Dai and Tezcan [15], the process \( X_{\pi} \), which is called the perturbed system there, is pathwise different from the performance process of the underlying distributed system described in §2. If \( \pi \) is an admissible policy, then it can be shown, similar to Theorem 2.1 in Tezcan [37], that \( X_{\pi} \) and the performance process associated with the \( N \)-system have the same distribution; see also Dai and Tezcan [15] for more details. We note that the class of admissible policies can be extended to cover all policies under which these two systems have the same distribution.

We next present the additional equations that must be satisfied by the distributed systems operating under the MED-FSF and MED-LB policies. First, we focus on the nonidling routing policies.

For a nonidling routing policy \( \pi \), in addition to Equations (2.2)–(12.12) in Dai and Tezcan [15], \( X_{\pi} \) must also satisfy the following condition:

\[
\int_0^\infty \frac{1}{\pi} \left\{ \sum_{i=1}^J Z_{(i)}'(s) < N_{(i)}' \right\} dA_{\pi}(s) = 0.
\]

This condition implies that an arriving customer who finds idle servers will be routed to one of the idle servers. In addition to (28), we assume that for \( \pi \in \Pi \), there exists \( a'_{\pi} > 0 \) for each \( r > 0 \) such that

\[
A_{j'}^{q'\cdot}(t) \text{ can only increase when } Q_{j'}'(t) \leq a'_{\pi} Q_{j'}'(t) \quad \text{for all } j' \in J.
\]
This implies that as long as the number of customers in one of the queues is more than $a'_r$ times the queue length of another queue, arriving customers are not routed to the former queue. Note that the MED-FSF and MED-LB policies satisfy (29).

Recall that we assume that the service rates are increasing with the index of the server pool as stated in Assumption (1). Under the MED-FSF policy, the following must hold:

$$A_j^{q,r}(t) \text{ can only increase when } \frac{Q_j'(t) + Z_j'(t) - N_j'}{N_j' \mu_j} \leq \min_{j \in J} \left\{ \frac{Q_j'(t) + Z_j'(t) - N_j'}{N_j' \mu_j} \right\}$$

and

$$A_j^{q,r}(t) \text{ can only increase when } \sum_{j'=1}^{j} (Z_{j'}'(t) - N_{j'}') = 0 \quad \text{for } j \in J.$$ 

By the nonidling condition (28), $A_j^{q,r}(t)$ can only increase when all the servers are busy. Hence (30) is invoked when there are no idle servers in the system. And if $Z_j'(t) = N_j'$, then $Q_j'(t)/(N_j' \mu_j)$ gives the expected delay time of a customer joining the $j$th queue before his service starts. The condition (31) implies that customers can be routed to server pool $j$ only when all the faster servers in the system, servers in pools $j+1$ through $J$, are busy.

Under the MED-LB policy, the following must hold in addition to (30):

$$A_j^{q,r}(t) \text{ can only increase when } \frac{Z_j'(t)}{N_j'} \leq \min_{j \in J} \left\{ \frac{Z_j'(t)}{N_j'} \right\}.$$ 

In this case, $Z_j'(t)/N_j'$ gives the proportion of busy servers. Hence (32) implies that the server pool with the lowest proportion of busy servers receives the arrival. The ties in (30) and (32) are broken arbitrarily.

5.2. 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 $[0, \infty)$. We use $\{x_n\}$ to denote a sequence whose $n$th term is $x_n$. For a function $f : \mathbb{R} \rightarrow \mathbb{R}$, 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 [16]. 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 $J_d$ topology and the weak convergence, which we denote by “$\Rightarrow$,” in this space is considered with respect to this topology. For a sequence of functions $\{x_n\} \subset \mathbb{D}^d[0, \infty)$, the sequence is said to converge uniformly on compact sets to $x \in \mathbb{D}^d[0, \infty)$ as $n \to \infty$, denoted by $x_n \Rightarrow x$ u.o.c. if for each $T > 0$,

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

We assume that all the random variables are defined in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A stochastic process can be viewed as a function from $\Omega \times [0, \infty)$ to $\mathbb{R}$. In several occasions, we will need to analyze the sample paths of stochastic processes. In those cases, we will explicitly express the dependence by writing $Y(\cdot, \omega)$ for the sample path associated with $\omega \in \Omega$ of a stochastic process $Y$. If the sample paths of a subset of $\Omega$ is analyzed, we omit $\omega$ from the notation.

We set $\mu_{\min} = \min_{j \in J} \{\mu_j\}$, $N_{\min}^r = \min_{j \in J} \{N_j^r\}$,

$$S_j(t) = S_j(t) - t \quad \text{and} \quad \dot{S}_j(t) = \dot{A}_j'(t) = A_j'(t) - \dot{A}'t.$$ 

For $T > 0$, we define

$$\mathcal{C}_T = \bigcap_{r=1}^{\infty} \{ \|A'(t) - A'(t-\cdot)\|_T \leq 1 \} \bigcap_{j=1}^{J} \{ \|S_j(t) - S_j(t-\cdot)\|_{N_j^r, T} \leq 1 \}. $$

By Lemma D.1,

$$\mathbb{P}(\mathcal{C}_T) = 1 \quad \text{for any } T > 0.$$ 


Let
\[ \lambda = \lim_{r \to \infty} \lambda'/|N'|. \]

By (4),
\[ \lambda = \sum_{j=1}^{J} \beta_j \mu_j = \mu. \]

We denote the set of nonidling routing policies that satisfy (29) by \( \Pi \), and with a slight abuse of terminology, we also refer to these routing policies as nonidling routing policies. We use the convention
\[ \inf \{ \emptyset \} = \infty \quad \text{and} \quad \sup \{ \emptyset \} = -\infty. \]

We also use the big and little-o notation: For two sequences, \( \{x_n\} \) and \( \{y_n\} \), we say \( x_n = O(y_n) \) and write \( x_n = O(y_n) \) if there exists \( n_0 \) and \( M \) such that \( |x_n| \leq M|y_n| \) for \( n > n_0 \). We say \( x_n = o(y_n) \) and write \( x_n = o(y_n) \) if \( \lim_{n \to \infty} |x_n|/|y_n| = 0 \). For two random variables \( T_1 \) and \( T_2 \), \( T_1 \sim T_2 \) means they have the same distribution.

### 5.3. Asymptotic bounds on \( Q' \) and \( Z' \) under a nonidling routing policy.

In this section, we derive asymptotic bounds on \( Q' \) and \( Z' \). These bounds are used to show that \( Q' \) and \( Z' \) are stochastically bounded in each finite interval \([0, T]\), which is required in §5.5.1 to prove our SSC results. They are also used to define Lyapunov functions that are used to prove the convergence of stationary distributions in §5.5.

Let \( x = (x_1, \ldots, x_J) \in \mathbb{R}^J \). We define \( \varphi'_i: \mathbb{R}^J \to \mathbb{R} \) for \( i = 1, 2 \) by
\[
\varphi'_i(x) = \sum_{j=1}^{J} (N'_j - x_j) \quad \text{and} \quad \varphi''_i(x) = \sum_{j=1}^{J} x_j.
\]

Clearly, \( \varphi'_i(Z'_j(t)) \geq 0 \) and \( \varphi''_i(Q'_j(t)) \geq 0 \) for all \( t \geq 0 \). We present bounds for \( \varphi'_i(Z'_j(t)) \) and \( \varphi''_i(Q'_j(t)) \) in terms of their initial states and certain functions of primitive processes. The proofs of the results in this section are placed in Appendix A. First, we present bounds for \( \varphi'_i(Z'_j(t)) \). Recall that \( M \cap \emptyset \) is defined by (34).

**Theorem 5.1.** Let \( \mathcal{X} \) be a distributed parallel server system operating under a nonidling routing policy. Assume that (2) and (4) hold. Then, there exists \( r_0 > 0 \) such that for every \( t_0 > 0 \), \( \omega \in M \cap \emptyset \), and \( r > r_0 \) if
\[
\varphi'_i(Z'_j(t_0)) \geq \frac{4\theta \sqrt{\lambda'}(t_0 \vee 1)}{\mu_{\min} \wedge 1},
\]
then
\[
\varphi'_i(Z'_j(t)) \leq \varphi'_i(Z'_j(t_0)) - \theta \sqrt{\lambda'} + a(\sqrt{|N'|}) + 2 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{N'[0]} + 2\|A'(t) - \lambda't\|_{0},
\]
otherwise
\[
\varphi'_i(Z'_j(t)) \leq 2J + \varphi'_i(Z'_j(t_0)) + \theta \sqrt{\lambda'} - \theta(\sqrt{|N'|}) + 2 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{N'[0]} + 2\|A'(t) - \lambda't\|_{0}.
\]

Next, we present bounds for \( \varphi''_i(Q'_j(t)) \). One of the terms in these bounds is \( \zeta' \) that is defined by
\[
\zeta'_i(t_0) = \sup_{s_1 \leq s_2 \in [0, t_0]} \left\{ (\theta \sqrt{\lambda'} - N''_i \mu_{\min})(s_2 - s_1) + [\tilde{A}'(s_2) - \tilde{A}'(s_1)] + \sum_{j=1}^{J} \tilde{S}_j(|N'| (v_j + (s_2 - s_1))) - \tilde{S}_j(|N'| v_j) \right\},
\]
where \( \tilde{S}_j \) and \( \tilde{A} \) are given by (33).

**Theorem 5.2.** Let \( \mathcal{X} \) be a distributed parallel server system operating under a nonidling routing policy. Assume that (2) and (4) hold. Then, there exists \( r_0 > 0 \) such that for every \( t_0 > 0 \), \( \omega \in M \cap \emptyset \), and \( r > r_0 \) if
\[
\varphi''_i(Q'(t)) > \theta \sqrt{\lambda'}(t_0 \vee 1),
\]
then

$$
\varphi'_2(Q'(t_0)) \leq \varphi'_2(Q'(0)) - \theta \sqrt{N'} t_0 + |o(\sqrt{N'})| + 2 t_0 (\mu_{\max} \lor 1) (J + \xi'(t_0)) \\
+ 4 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{N' \mid t_0} + 2 \|A'(t) - \lambda' t\|_{t_0}.
$$

(42)

otherwise

$$
\varphi'_2(Q'(t_0)) \leq \varphi'_2(Q'(0)) + 2J + (t_0 \lor 1)(\mu_{\max} \lor 1) (J + \xi'(t_0)) + |o(\sqrt{N'})| \\
+ 4 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|_{N' \mid t_0} + 2 \|A'(t) - \lambda' t\|_{t_0}.
$$

(43)

**Remark 5.2.** In an \(\land\)-system, there is only one queue, hence the processes \(A, A_g, Q\) are one dimensional, but otherwise all the components of \((A', A'_g, A'_j, Q', Z', T', Y', B', D')\) have the same interpretations they have for the distributed systems.

The results of Theorems 5.1 and 5.2 also hold for an \(\land\)-system \(\mathbb{X}' = (A', A'_g, A'_j, Q', Z', T', Y', B', D')\) operating under a nonidling and HL routing policy with \(\varphi'_2 : \mathbb{R} \rightarrow \mathbb{R}\) is defined by \(\varphi'_2(x) = x\).

**5.4. Fluid limits.** In this section, we study the fluid limits of the distributed systems. The results in this section are needed to show that the diffusion scaling introduced in \(\S 2\) is properly defined. By using the result on the invariant states of the fluid limits in this section, we verify that when (19) holds, the fluid limits are time invariant. The results in this section are used in the proofs of our SSC results to verify that Assumption 1 of Dai and Tezcan [15] holds and in proving the weak convergence of \(\hat{X}'\).

The fluid scaling \(\mathbb{X}'(\cdot)\) is defined by \(\mathbb{X}'(\cdot) = \mathbb{X}'(\cdot)/\sqrt{N'}\). The following notation and definitions are introduced in Dai and Tezcan [15], but we repeat them here for completeness. The process \(\mathbb{X}'(\cdot)\) is called a fluid limit of \(\{X'\}\) if there exists a sequence \(\{r_i\}\), with \(r_i \rightarrow \infty\) as \(i \rightarrow \infty\), and \(\omega \in \mathcal{C}\) such that \(\mathbb{X}'(\cdot, \omega)\) converges u.o.c. to \(\mathbb{X}(\cdot, \omega)\), where \(\mathcal{C} \subset \Omega\) is taken as in Theorem A.1 in Dai and Tezcan [15] and \(\mathbb{P}[\mathcal{C}] = 1\). The existence of the fluid limits are established and the fluid model equations that are satisfied by every fluid limit are presented in Theorem A.1 in Dai and Tezcan [15]. We call the vector \((q, z)\) an invariant state of the fluid limits if for any fluid limit \(\mathbb{X}'\), \(Q'(0) = q\) and \(Z'(0) = z\) implies \(\hat{Q}(t) = q\) and \(\hat{Z}(t) = z\) for all \(t > 0\).

The following result characterizes the invariant states of the fluid limits of the MED-FSF and MED-LB distributed server pool systems.

**Lemma 5.1.** Let \(\{X'\}\) be a sequence of MED-FSF or MED-LB distributed server pool systems. Assume that (2) and (4) hold and that \(\{Q'(0)\}\) is bounded a.s. as \(r \rightarrow \infty\). Let \(q_1(a) = a, a \geq 0, q_1(a) = a\mu, \beta_j / (\mu_1 \beta_j), q(a) = (q_1(a), q_2(a), \ldots, q_J(a)), z_j = \beta_j,\) and \(z = (z_1, \ldots, z_J)\). Then, \(\mathcal{M} = \{(q(a), z): a \geq 0\}\) is the set of all the invariant states of the fluid limits of \(\{X'\}\).

A proof is presented in Appendix B.

**Remark 5.3.** It can similarly be proved that if \(\{X'\}\) is a sequence of LB or FSF \(\land\)-systems and \(\{Q'(0)\}\) is bounded a.s. as \(r \rightarrow \infty\), then \(\mathcal{M} = \{(a, z): a \geq 0\}\) is the set of all the invariant states of the fluid limits of \(\{X'\}\).

**5.5. Diffusion limits.** In this section, we establish the weak limits of \(\hat{Q}', \hat{Z}',\) and \(\hat{W}'\) as \(r \rightarrow \infty\). By (19), \(\hat{Q}'(0) \rightarrow 0\) and \(\hat{Z}'(0) \rightarrow z\) as \(r \rightarrow \infty\), where \(z\) is given as in Lemma 5.1. Hence the diffusion scalings defined in (5) give the fluctuations around the fluid limits.

In the following section, we establish two SSC results for the distributed systems operating under the MED-FSF and MED-LB policies. Then, in \(\S 5.5.2\), we establish the diffusion limits using these SSC results. We focus on the stationary distributions of these processes in \(\S 5.5.3\) and 5.5.4.

**5.5.1. SSC.** We first give an intuitive explanation of our SSC results and illustrate the results in a distributed system with two server pools. These results are proven to hold for systems with an arbitrary number of server pools. The proofs of the propositions in this section are presented in Appendix C.1.

The MED policy routes the customers to the queue with the minimum expected delay when all the servers are busy, where the expected delay of a queue; say \(j\), at time \(t\) is defined by \(Q'_j(t)/\mu_j N'_j\). Assume that all the servers are busy at time \(t\) and

$$
Q'_1(t)/\mu_1 N'_1 < Q'_2(t)/\mu_2 N'_2.
$$

(44)

Because the arrival rate is greater than the total service rate of servers in the first pool, one would expect to see that the number of customers in queue 1 will increase and the number of customers in queue 2 will decrease starting from time \(t\), barring, of course, some stochastic fluctuations. Hence the value of \(Q'_2(t)/\mu_2 N'_2\) —
$Q_i(t)/(\mu_i N_i^r)$ is expected to decrease and as long as (44) holds. Under the MED-FSF policy, if a server in pool $2$ becomes available in a two-pool distributed system at time $t$, then he receives the next arriving customer after time $t$ (recall that we assume $\mu_1 < \mu_2$). Since $\lambda' > \mu_2 N_i^r$ for $r$ is large enough, the idle time of servers in higher priority pools becomes very small for $r$ large enough. In general, we have the following result.

**Proposition 5.1.** Let $\{\mathcal{X}'\}$ be a sequence of MED-FSF distributed server pool systems. Assume that (2), (4), and (19) hold. Then, for some $L' = o(\sqrt{NT})$ with $L' \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$
P \left\{ \sup_{L' \sqrt{N} \leq T} \left| \frac{Q_j(t)}{\mu_j} - \frac{\hat{Q}_j(t)}{\mu_j} \right| \right\} \to 0 \quad (45)$$

as $r \to \infty$.

If in addition,

$$
\left| \sum_{j=2}^{\infty} \hat{Z}_j'(0) \right| \to 0 \quad (46)
$$

and

$$
\left| \frac{\hat{Q}_j'(0)}{\beta_j} - \frac{\hat{Q}_j'(0)}{\beta_j} \right| \to 0 \quad (47)
$$

in probability as $r \to \infty$ for $j, j' \in \mathcal{J}$, then for every $T > 0$,

$$
\left| \frac{\hat{Q}_j'(t)}{\beta_j} - \frac{\hat{Q}_j'(t)}{\beta_j} \right| \to 0 \quad (48)
$$

for all $j, j' \in \mathcal{J}$ in probability as $r \to \infty$.

**Remark 5.4.** It can be similarly shown that if $\{\mathcal{X}'\}$ is a sequence of FSF $\wedge$-systems and (2), (4), and (19) hold, then for some $L' = o(\sqrt{NT})$ with $L' \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$
P \left\{ \sup_{L' \sqrt{N} \leq T} \left| \sum_{j=2}^{\infty} \hat{Z}_j'(t) \right| > \epsilon \right\} \to 0 \quad (49)
$$

as $r \to \infty$. If in addition (46) holds, then for every $T > 0$,

$$
\left| \sum_{j=2}^{\infty} \hat{Z}_j'(t) \right| \to 0 \quad (50)
$$

in probability as $r \to \infty$.

Next, we consider the SSC under the MED-LB policy. In a distributed parallel server system with two server pools under the MED-LB policy if the percentage of busy servers in the first pool is less than that in the second pool, the differences will decrease because the server pool with higher percentage of busy servers will not receive any arrivals. We have the following result for the distributed systems under the MED-LB policy.

**Proposition 5.2.** Let $\{\mathcal{X}'\}$ be a sequence of MED-LB distributed server pool systems. Assume that (2), (4), and (19) hold. Then, for some $L' = o(\sqrt{NT})$ with $L' \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$
P \left\{ \sup_{L' \sqrt{N} \leq T} \left| \frac{Q_j(t)}{\mu_j} - \frac{\hat{Q}_j(t)}{\mu_j} \right| \right\} \to 0 \quad (50)$$

as $r \to \infty$.

If in addition (20) and (47) hold, then for every $T > 0$,

$$
\left| \frac{\hat{Q}_j(t)}{\beta_j} - \frac{\hat{Q}_j(t)}{\beta_j} \right| \to 0 \quad (51)
$$

for all $j, j' \in \mathcal{J}$ in probability as $r \to \infty$. 
Remark 5.5. It can be similarly shown that if \( \{X'\} \) is a sequence of LB \&-systems and (2), (4), and (19) hold, then for some \( L' = o(\sqrt{N'}) \) with \( L' \rightarrow \infty \) as \( r \rightarrow \infty \), and for every \( T > 0 \) and \( \varepsilon > 0 \),

\[
\mathbb{P}\left\{ \sup_{L'/\sqrt{N'} \leq t \leq T} \left| \frac{\hat{Z}_j'(t)}{\beta_j} - \frac{\hat{Z}_{j'}'(t)}{\beta_{j'}} \right| > \varepsilon \right\} \rightarrow 0
\]
as \( r \rightarrow \infty \). If in addition (20) holds, then for every \( T > 0 \),

\[
\left\| \frac{\hat{Z}_j'(t)}{\beta_j} - \frac{\hat{Z}_{j'}'(t)}{\beta_{j'}} \right\|_T \rightarrow 0
\]
for all \( j, j' \in \mathcal{J} \) in probability as \( r \rightarrow \infty \).

5.5.2. Diffusion limits of the total queue length and the virtual waiting time processes. The SSC results established in (48) and (51) reveal that under the MED-FSF and MED-LB policies, the individual queue lengths and number of customers in service in each pool can be estimated from the total number of customers in the system with an error that goes to zero in probability as \( r \rightarrow \infty \). Hence it is enough to focus on the total number of customers in the system instead of analyzing each queue and number of customers in service in a server pool separately. To this end, let \( \hat{X}'(t) \) be defined as in (5) and \( \hat{X}' = [X'(t): t \geq 0] \). We have the following weak limits for \( \hat{X}' \) under the MED-FSF and MED-LB policies. The proofs of these results are presented in Appendix C.2.

Proposition 5.3. Let \( \{X'\} \) be a sequence of MED-FSF distributed server pool systems. Assume that (2), (4), (19), and (48) hold. Then,

\[ \hat{X}' \Rightarrow \bar{X}, \quad \text{as } r \rightarrow \infty, \]

where \( \bar{X} \) is the unique solution to the following stochastic differential equation (SDE):

\[
d\bar{X}(t) = h(\bar{X})dt + \sqrt{2\mu}d\bar{b}(t),
\]

where \( \bar{b} \) is a standard Brownian motion and

\[
h(x) = \begin{cases} 
-\theta \sqrt{\mu} & \text{if } x \geq 0, \\
-\theta \sqrt{\mu} - \mu x & \text{if } x < 0.
\end{cases}
\]

Remark 6.6. By Theorem 11.4.5 in Whitt [43] and by Proposition 5.1, under the conditions of Proposition 5.3,

\[ (\hat{Q}', \hat{Z}') \Rightarrow (\hat{Q}, \hat{Z}), \]

where \( \hat{Q} = (\hat{Q}_1, \ldots, \hat{Q}_j) \) and \( \hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_j) \) with

\[
\hat{Q}_j(t) = \frac{\mu_j \beta_j}{\sum_{j'=1}^{J} \mu_j \beta_{j'}} (\bar{X}(t))^+ \quad \text{for } j \in \mathcal{J},
\]

\[
\hat{Z}_j(t) = (\bar{X}(t))^+ \quad \text{and} \quad \hat{Z}_j(t) = 0 \quad \text{for } 2 \leq j \leq J.
\]

Proposition 5.4. Let \( \{X'\} \) be a sequence of MED-LB distributed server pool systems. Assume that (2), (4), (19), and (51) hold. Then,

\[ \hat{X}' \Rightarrow \tilde{X} \quad \text{as } r \rightarrow \infty, \]

where \( \tilde{X} \) is the unique solution to the following SDE:

\[
d\tilde{X}(t) = h(\tilde{X})dt + \sqrt{2\mu}d\tilde{b}(t),
\]

where \( \tilde{b} \) is a standard Brownian motion and

\[
h(x) = \begin{cases} 
-\theta \sqrt{\mu} & \text{if } x \geq 0, \\
-\theta \sqrt{\mu} - \mu x & \text{if } x < 0.
\end{cases}
\]
Remark 5.7. By Theorem 11.4.5 in Whitt [43] and Proposition 5.2, under the conditions of Proposition 5.4,
\[(\hat{Q}', \hat{Z}') \Rightarrow (\hat{Q}, \hat{Z}),\]
where \(\hat{Q} = (\hat{Q}_1, \ldots, \hat{Q}_j)\) and \(\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_j)\) with
\[
\begin{align*}
\hat{Q}_j(t) &= \frac{\mu_j \beta_j}{\sum_{j'=1}^n \mu_j \beta_j'} (\hat{X}(t))^+ \quad \text{for } j \in \mathcal{J} \\
\hat{Z}_j(t) &= \frac{\beta_j}{\sum_{j'=1}^n \beta_j'} (\hat{X}(t))^− \quad \text{for } j \in \mathcal{J}.
\end{align*}
\]

Remark 5.8. It can be similarly shown that
(i) if \(\{X'\}\) is a sequence of FSF \(\wedge\)-systems satisfying (2), (4), (19), and (49), then
\[
\hat{X}_r' \Rightarrow \hat{X}, \quad \text{as } r \to \infty,
\]
and (54) holds, where \(\hat{X}_r'\) is given by (10) and \(\hat{X}\) is the unique solution to the SDE (52).
(ii) if \(\{X'\}\) is a sequence of LB \(\wedge\)-systems satisfying (2), (4), (19), and (52), then
\[
\hat{X}_r' \Rightarrow \hat{X}, \quad \text{as } r \to \infty,
\]
and (58) holds, where \(\hat{X}_r'\) is given by (10) and \(\hat{X}\) is the unique solution to the SDE (55).

The result for the FSF \(\wedge\)-systems has also been proved in Armony [2].

Next, we focus on the virtual waiting time process. Let \(W'_j(t)\) be the virtual waiting time for queue \(j\) at time \(t\) in the \(r\)th system, i.e., the time a customer would wait before its service is started if he joins queue \(j\) at time \(t\) under an HL routing policy, and \(W'_j = \{W'_j(t): t \geq 0\}\) be the virtual waiting time process for the \(j\)th queue. Then,
\[
W'_j(t) = \inf \{s \geq 0: D_j'(s + t) \geq Q'_j(0) + Z'_j(0) + A'_j\epsilon_s(t) + A'_j(t) - (N'_j - 1)\}.
\]
Let \(\kappa'(t)\) denote the index of the server pool or the queue that a customer arriving at time \(t\) would be routed to. Obviously, \(\kappa'(t)\) depends on the routing policy. For example, under the MED-LB policy,
\[
\kappa'(t) = \begin{cases} 
\{l: Q'_l(t)/(\mu_l N'_l) < Q'_j(t)/(\mu_j N'_j) \text{ for all } l \notin \mathcal{J} \} & \text{if } \sum_{j \in \mathcal{J}} Z'_j(t) = |N'|, \\
\{j: Z'_j(t)/N'_j < Z'_j(t)/N'_j \text{ for all } l \notin \mathcal{J} \} & \text{if } \sum_{j \in \mathcal{J}} Z'_j(t) < |N'|.
\end{cases}
\]

From the definition of \(\kappa'(t)\), it follows that
\[
W'_r(t) = W'_{\kappa'(t)}(t).
\]

We show that weak limit of \(W'_r\) can be expressed as a simple function of \(X\).

Theorem 5.3. Let \(\{X'\}\) be a sequence of MED-FSF (MED-LB) distributed server pool systems. Under the conditions of Proposition 5.3 (resp. Proposition 5.4),
\[
\tilde{W}' \Rightarrow \frac{[\hat{X}]^+}{\mu},
\]
as \(r \to \infty\), where \(\hat{X}\) is the unique solution to the SDE (52) (resp. the SDE (55)).

Remark 5.9. It can be similarly shown that
(i) if \(\{X'\}\) is a sequence of FSF \(\wedge\)-systems that satisfies the conditions of the first part of Remark 5.8 then (59) holds with \(\tilde{W}'\) replaced by \(\tilde{W}'_r\).
(ii) if \(\{X'\}\) is a sequence of LB \(\wedge\)-systems that satisfies the conditions of the second part of Remark 5.8 then (59) holds with \(\tilde{W}'\) replaced by \(\tilde{W}'_r\).

The result for the FSF \(\wedge\)-systems has also been proved in Armony [2].

5.5.3. Stationary distributions of the diffusion limits. Our asymptotic optimality and equivalence results are stated in terms of the stationary distributions. The main reason is that the staffing decisions in call centers
are usually made using stationary values of the performance measures. Hence it is of practical value to study the convergence of the stationary probabilities of the queue length and waiting time processes. In this section, we present the steady-state probabilities for \( \hat{X} \), the limiting diffusion process in Propositions 5.3 and 5.4.

**Theorem 5.4.** Let \( \hat{X}(\cdot) \) be the diffusion process that is the unique solution to the SDE (52). Then, the steady-state distribution of \( \hat{X}(\cdot) \) has a density \( f \) given by (13).

**Theorem 5.5.** Let \( \hat{X}(\cdot) \) be the diffusion process that is the unique solution to the SDE (55). Then, the steady-state distribution of \( \hat{X}(\cdot) \) has a density \( f \) given by (26).

Theorems 5.4 and 5.5 can be proven similarly to Proposition 3.5 in Armony [2].

### 5.5.4. Convergence of stationary distributions.

In this section, we show that \( \hat{X}'(\infty) \) converges weakly to the stationary distribution of its weak limit under the MED-FSF and MED-LB policies as \( r \to \infty \). To prove the convergence, we first show that the stationary distribution exists and then show that the sequence of stationary distributions are tight.

To show the existence of the stationary distribution of \( \hat{X}' \) for each \( r \), we consider the stability of a distributed server pool system under a nonidling routing policy. We show that \( (Q', Z') \) has a stationary distribution whenever a natural traffic condition is satisfied.

**Theorem 5.6.** Let \( \pi \in \Pi \) and \( X'_{\pi} \) be a \( \pi \) distributed server pool system. If

\[
\chi' < \sum_{j \in J} \mu_j N_j',
\]

then the process \( (Q', Z') \) has a unique stationary distribution.

We present a proof in Appendix C.3.1. The proof is based on the relationship established in Dai [13] between the stability of the corresponding conventional fluid limit and the positive recurrence of the underlying Markov chain.

Next, we show that the sequence of the stationary distributions of the process \( \{\hat{X}', \hat{Z}'\} \) is tight under any nonidling routing policy. Recall that a sequence of random variables, \( \{L'_r\} \), taking values in a metric space \((\mathcal{F}, \rho)\) is said to be tight if for every \( \epsilon > 0 \), there exists a compact set \( \mathcal{K} \subset \mathcal{F} \) such that \( \inf P[L'_r \in \mathcal{K}] > 1 - \epsilon \) (Ethier and Kurtz [16]).

**Theorem 5.7.** Let \( \pi \in \Pi \) and \( \{X'_{\pi}\} \) be a sequence of \( \pi \) distributed server pool systems. If (2) and (4) hold, then the sequence \( \{\hat{Q}'(\infty, \pi), \hat{Z}'(\infty, \pi)\} \) is tight.

A proof is presented in Appendix C.3.2. The proof is based on the results of Gamarnik and Zeevi [18]. In particular, we define two functions and show that they are geometric Lyapunov functions for these systems. Then, we use Theorem 5 in their paper to conclude the proof.

Recall that \( \hat{X}'(\infty) \) and \( \hat{W}'(\infty) \) denote the stationary distribution of the processes \( \hat{X}' \) and \( \hat{W}' \), respectively.

**Theorem 5.8.** Let \( \{X'\} \) be a sequence of MED-FSF distributed server pool systems. If (2) and (4) hold, then

\[
\hat{X}'(\infty) \Rightarrow \hat{X}(\infty) \quad \text{and} \quad \hat{W}'(\infty) \Rightarrow \frac{[X(\infty)]^+}{\mu},
\]

where \( \hat{X}(\infty) \) has the density given by (13).

**Theorem 5.9.** Let \( \{X'\} \) be a sequence of MED-LB distributed server pool systems. If (2) and (4) hold, then

\[
\hat{X}'(\infty) \Rightarrow \hat{X}(\infty) \quad \text{and} \quad \hat{W}'(\infty) \Rightarrow \frac{[X(\infty)]^+}{\mu},
\]

where \( \hat{X}(\infty) \) has the density given by (26).

The proofs of Theorems 5.8 and 5.9 are presented in Appendix C.3.3.

**Remark 5.10.** It can be similarly shown that

(i) if \( \{X'\} \) is a sequence of FSF \( \wedge \)-systems that satisfy (2) and (4), then (61) and (62) hold with \( \hat{X}' \) and \( \hat{W}' \) are replaced by \( \hat{X}'_{\wedge} \) and \( \hat{W}'_{\wedge} \), respectively;

(ii) if \( \{X'\} \) is a sequence of LB \( \wedge \)-systems that satisfy (2) and (4), then (63) and (64) hold with \( \hat{X}' \) and \( \hat{W}' \) are replaced by \( \hat{X}'_{\wedge} \) and \( \hat{W}'_{\wedge} \), respectively.

The result for the FSF \( \wedge \)-systems has also been proved in Armony [2].
5.6. Proofs of the results in §3. In this section, we prove Theorems 3.1, 3.2, 3.3, and 3.5 and Corollary 3.1.

Proof of Theorem 3.1. Consider a sequence of MED-FSF distributed server systems described in §2. Assume that (2) and (4) hold.

Fix an admissible routing policy $\pi$ and consider a sequence of $\pi$ distributed server systems and a sequence of $\land$-systems with the $r$th systems in both sequences having the same arrival and service rates and number of servers in each pool. Let $Q'_r(t)$ denote the number of customers in the queue and $Z'_{j,\land}(t)$ denote the number of customers in service in the $r$th pool at time $t$ in the $r$th $\land$-system. Recall that $Q' = (Q'_1, \ldots, Q'_r)$ and $Z' = (Z'_1, \ldots, Z'_r)$ are the number of customers in the queue and in the service processes in a distributed system.

We claim that there exists an admissible routing policy $\pi'$ for the $\land$-systems such that for a distributed system operating under the policy $\pi$ and the corresponding $\land$-system operating under the policy $\pi'$

$$Q'_r(\infty, \pi') = \sum_{j \in \mathcal{J}} Q'_j(\infty, \pi) \quad \text{and} \quad Z'_{j,\land}(\infty, \pi') = Z'_j(\infty, \pi) \quad \text{for all } j \in \mathcal{J}. \quad (65)$$

The policy $\pi'$ is constructed from the policy $\pi$ as follows. Consider the distributed system and the corresponding $\land$-system, and assume that the interarrival times of the customers to each system is equal and the service requirement of the $k$th customer arriving to each system is the same. The routing policy $\pi$ dictates the order customers are served in the distributed system. Assume that the system is initially empty. The customers in the $\land$-system can be served in the same order and in the same server pool as follows; start the service of the $k$th arriving customer in the $\land$-system when the $k$th arriving customer’s service in the distributed system starts and in both systems route the customer to the same server pool. Denote the routing policy in the $\land$-system by $\pi'$. Then,

$$Q'_r(\cdot, \pi') = \sum_{j \in \mathcal{J}} Q'_j(\cdot, \pi) \quad \text{and} \quad Z'_{j,\land}(\cdot, \pi') = Z'_j(\cdot, \pi) \quad \text{for all } j \in \mathcal{J} \text{ a.s.} \quad (66)$$

Hence (65) holds. Also, $\pi'$ is admissible since $\pi$ is.

Let

$$\hat{X}'_r(t) = \left( Q'_r(t) + \sum_{j \in \mathcal{J}} (Z'_{j,\land}(t) - N'_j) \right) / \sqrt{|N'|}$$

and $\hat{X}'_r(\infty, \pi')$ be the weak limit of $\hat{X}'_r(t, \pi')$ as $t \to \infty$ if it exists, and taken as in (7) otherwise. Then, by (66), $\hat{X}'_r(\infty, \pi') \sim \hat{X}'(\infty, \pi)$.

Now, consider the preemptive FSF (P-FSF) policy in the same sequence of $\land$-systems. A preemptive policy allows a customer to be handed off to another server, who will resume the service from the point it has been discontinued. Under the P-FSF policy in an $\land$-system if an arriving customer finds more than one available server, he is served by the faster one. Also, slower servers hand off a customer whenever a faster server becomes available.

By Proposition 3.1 of Armony [2] in an $\land$-system,

$$\mathbb{P}\{X'_r(\infty, \text{P-FSF}) > x\} \leq \mathbb{P}\{X'_r(\infty, \pi') > x\}$$

for any admissible policy $\pi'$ and for all $x \in \mathbb{R}$. (In Armony [2], $\pi'$ is assumed to be HL, but it is not required in the proof of their Proposition 3.1.) By Proposition 4.5 of Armony [2], the argument above, and Theorem 5.8, we get (8).

Observe from Theorem 5.8 that

$$\lim_{r \to \infty} \mathbb{P}\{W'_r(\infty, \text{MED-FSF}) > 0\} = \lim_{r \to \infty} \mathbb{P}\{X'_r(\infty, \text{MED-FSF}) > 0\},$$

since $\hat{X}(\infty)$ is a continuous random variable. We get (9) by combining the PASTA property of an admissible policy (Wolff [47]) with (8).

Proof of Theorem 3.2. The result follows from Proposition 5.3, Remark 5.8, Theorem 5.3, Remark 5.9, Theorem 5.4, Theorem 5.8, and Remark 5.10.

Proof of Corollary 3.1. Note that by (12), we have that

$$\mathbb{P}(W'(\infty) > 0) \to \alpha$$

as $r \to \infty$, where $\alpha$ is given by (14). Now, consider $E[Q'(\infty)]$, the expected queue length in the steady state. By Theorem 5.8 and (C.43),

$$E[Q'(\infty)\sqrt{|N'|}] \to E[(\hat{X}(\infty))^+] \quad (67)$$
as \( r \to \infty \), where
\[
\mathbb{E}[(\tilde{X}(\infty))^+] = \alpha \frac{\sqrt{\mu}}{\theta} \tag{68}
\]
by Theorem 5.4. By combining (67), (68), (59), (2), and (4), we obtain
\[
\left| \mathbb{E}[\bar{W}^r(\infty)] - \frac{\sqrt{|N^r|\mathbb{E}[(\tilde{X}(\infty))^+]}}{\sqrt{r}} \right| = \left| \mathbb{E}[\bar{W}^r(\infty)] - \alpha \frac{\sqrt{|N^r|\mu}}{\sqrt{r\theta}} \right| \to 0,
\]
where \( \alpha \) is given by (14). □

**Proof of Theorem 3.3.** The convergence in (21) follows from Proposition 5.2. Next, we prove (18). For \( i, j \in \mathcal{J} \), define
\[
\mathcal{T}_{ij} = \left| \frac{\tilde{Z}_i(\infty)}{\beta_i} - \frac{\tilde{Z}_j(\infty)}{\beta_j} \right|.
\]
Observe that by Proposition 5.2 and Theorem 5.7,
\[
\mathcal{T}_{ij} \to 0 \quad \text{in probability as } r \to \infty.
\]
By (C.43), the sequence \{\mathcal{T}_{ij}\} is uniformly integrable. This yields (18) by (69) and Theorem 4.5.4 of Chung [12]. □

**Proof of Theorem 3.4.** For \( i \geq 2 \), by Proposition 5.1 and Theorem 5.7,
\[
\left| \frac{\tilde{Z}_i(\infty)}{\beta_i} \right| \to 0 \quad \text{in probability as } r \to \infty
\]
and
\[
|\tilde{Z}_i(\infty)| \to (\tilde{X}(\infty))^+ \quad \text{in probability as } r \to \infty.
\]
By (C.43), the sequence
\[
\left\{ \left| \frac{\tilde{Z}_i(\infty)}{\beta_i} \right| \right\}
\]
is uniformly integrable. This yields (22) by Theorem 4.5.4 of Chung [12]. □

**Proof of Theorem 3.5.** The last equality in (24) follows from Theorem 1 of Halfin and Whitt [23] and the last equality in (25) can be proved using arguments similar to those in the proof of Theorem 3 in Garnett et al. [20]. The first and the second equalities in (24) and (25) follows from Proposition 5.4, Remark 5.8, Theorem 5.3, Remark 5.9, Theorem 5.5, Theorem 5.9, and Remark 5.10. □

**6. Concluding remarks.** Under the FSF policy, all the servers except those with the lowest service rate are utilized 100\%, and under the LB policy the utilizations of all the servers are equal. The LB policy can be modified to distribute the available percentage of idle time, which is equal to \((1-r')\) in the \( r \)th system, in desired proportions among all the server pools. To illustrate this, let \( d = (d_1, \ldots, d_j) \). Under the modified LB with parameter \( d \) (MLB\(_d\)) policy (also under the MED-MLB\(_d\) policy) when there are idle servers in the system, a customer arriving to the system at time \( t \) is routed to the server pool with minimum
\[
\frac{Z'_j(t) - N'_j}{d_j N'_j}.
\]
Note that if \( d_1 = d_2 = \cdots = d_j \), this policy reduces to the original LB policy. If \( d_j < d_k \) for \( k, j \in \mathcal{J} \), the utilization of the server pool \( j \) will be more than the utilization of the server pool \( k \). Therefore the utilizations of all servers can be controlled by assigning appropriate values to \( d \). One can show similar to (21) that
\[
\sqrt{|N^r|} \left\| \frac{Z'_j(t) - N'_j}{d_j N'_j} - \frac{Z'_i(t) - N'_i}{d_i N'_i} \right\|_T \to 0 \quad \text{in probability as } r \to \infty.
\]
Similar results that are established for the systems operating under the LB and MED-LB routing policies in §3 can also be shown to hold under the MLB\(_d\) and MED-MLB\(_d\) routing policies. In particular, under the MLB\(_d\) and MED-MLB\(_d\) routing policies, (16) hold with
\[
\alpha = \left[ 1 + \frac{\theta / \sqrt{\mu} \Phi(\theta / \sqrt{\mu})}{\phi(\theta / \sqrt{\mu})} \right]^{-1}.
\]
where
\[ \tilde{\mu} = \frac{\sum_{j=1}^{J} d_j \beta_j \mu_j}{\sum_{j=1}^{J} \beta_j \mu_j}. \]

Let the limiting stationary delay probability under a policy \( \pi \) be denoted by \( \alpha_x \). If \( 1 = d_1 \geq d_2 \geq \cdots \geq d_J \), then it can easily be shown under the assumption (1) that
\[ \alpha_{\text{FSF}} \leq \alpha_{\text{MLB}}, \leq \alpha_{\text{LB}}. \]  

If \( \mu_1 < \mu_2 \leq \mu_3 \leq \cdots \leq \mu_J \) and \( 1 = d_1 > d_2 \geq \cdots \geq d_J \), then the inequalities in (70) are strict and as \( d_2 \downarrow 0 \), \( \alpha_{\text{MLB}} \downarrow \alpha_{\text{FSF}} \), and as \( d_j \uparrow 1 \), \( \alpha_{\text{MLB}} \uparrow \alpha_{\text{LB}} \).

Our results can be extended to the case when each distributed center is an \( \land \)-system. In that case, the definition of expected delay in a queue should be modified in an obvious way. In some applications, one or more pools have significantly fewer servers than the other pools. In this case, our approximations may perform poorly because our results are based on the asymptotic analysis when all the server pools have significantly many servers. Also, as illustrated by simulation results, when the number of parallel server pools is moderately large, the quality of approximations diminishes. For such systems, a dynamic LB-scheme in addition to a dynamic routing policy, may be needed to improve the performance of distributed systems.

**Appendix A. Proofs of the results in §5.3.**

**Proof of Theorem 5.1.** For notational simplicity, we set \( \varphi_r(t) = \varphi_r(Z'(t)) \). Fix \( t_0 > 0, r > 0 \) and \( \omega \in \mathcal{M}^\infty \). (1) Assume that (37) holds. Let \( t_1(\omega) = \inf \{ t : \varphi_t(\omega) < \varphi_0(\omega)/2 \} \). We investigate two possible cases, \( t_1(\omega) \leq t_0 \) and \( t_1(\omega) > t_0 \), separately. We omit \( \omega \) from the notation in the rest of the proof.

**Case 1.** First, assume that \( t_1 \leq t_0 \). Set \( s_i = t_1 \) and define
\[ s_{2i+1} = \inf \{ t > s_{2i} : \varphi_r(t) = 0 \} \quad \text{and} \quad s_{2i+2} = \inf \{ t > s_{2i+1} : \varphi_r(t) > 0 \} \quad \text{for } i = 0, 1, \ldots. \]  

For any \( t \in [s_{2i+1}, s_{2i+2}] \), \( \varphi_r(t) = 0 \), so assume that \( t_0 \in [s_{2i}, s_{2i+1}) \) for some \( i \). Note that
\[ \varphi_r(t_0) \leq \varphi_r(s_2) - (A'(t_0) - A'(s_2)) + \sum_{j=1}^{J} (S_j(B'_j(t_0)) - S_j(B'_j(s_2))), \]

since all arrivals are routed to one of the queues that have idle servers. Hence
\[ \varphi_r(t_0) \leq \varphi_r(s_2) - (A'(t_0) - A'(s_2)) + \sum_{j=1}^{J} (\bar{S}_j(B'_j(t_0)) - \bar{S}_j(B'_j(s_2))) + \sum_{j=1}^{J} \mu_j \int_{0}^{t_0} Z'_j(s) \, ds - \lambda'(t_0 - s_2) \]
\[ \leq \varphi_r(s_2) + \theta \sqrt{\lambda'(t_0 - t_1)} + |\alpha(\sqrt{N'})| + 2 \sum_{j=1}^{J} \| S_j(t) - \mu_j t \|_{N'/0} + 2 \| A'(t) - \lambda' t \|_{N'/0}. \]  

Since by Lemma D.1, \( \varphi_r(s_2) < (\varphi_r(0)/2) \lor J \) and \( \varphi_r(0)/2 > J \) for \( r \) large enough,
\[ \varphi_r(t_0) \leq \varphi_r(0)/2 + \theta \sqrt{\lambda'(t_0 - t_1)} + |\alpha(\sqrt{N'})| + 2 \sum_{j=1}^{J} \| S_j(t) - \mu_j t \|_{N'/0} + 2 \| A'(t) - \lambda' t \|_{N'/0}. \]

Now, add and subtract \( \varphi_r(0)/2 \) to the right-hand side (RHS) above to get (38).

**Case 2.** Now, assume that \( t_1 > t_0 \), then
\[ \varphi_r(t_0) \leq \varphi_r(0) - A'(t_0) + \sum_{j=1}^{J} S_j(B'_j(t_0)) \]
\[ = \varphi_r(0) - A'(t_0) + \sum_{j=1}^{J} \bar{S}_j(B'_j(t_0)) + \sum_{j=1}^{J} \mu_j \int_{0}^{t_0} Z'_j(s) \, ds - \lambda' t_0 \]
\[ \leq \varphi_r(0) - (\mu_{\min} \varphi_r(0)/2 - \sqrt{\lambda' \theta}) t_0 + |\alpha(\sqrt{N'})| + 2 \sum_{j=1}^{J} \| S_j(t) - \mu_j t \|_{N'/0} + 2 \| A'(t) - \lambda' t \|_{N'/0}. \]

By (37), the last inequality gives (38).
(2) Now, assume that
\[ \varphi_j'(Z'(0)) \leq \frac{4\theta \sqrt{\lambda'} (t_0 \vee 1)}{\mu_{\min} \wedge 1}. \]
First, assume that \( \varphi_j'(0, \omega) > 0 \). Set \( s_0 = 0 \) and define \( s_{2i+1} \) and \( s_{2i+2} \) as in (A.1). For any \( t \in [s_{2i+1}, s_{2i+2}] \), \( \varphi_j'(t) = 0 \), so assume that \( t_0 \in [s_{2i}, s_{2i+1}] \) for some \( i \). Then, we have that (A.2) holds with \( t_1 = 0 \). Since \( \varphi_j'(s_0) < \varphi_j'(0) \vee J \), (A.2) yields (39). If \( \varphi_j'(0, \omega) = 0 \), then define \( t_1 = \inf \{ t > 0 : \varphi_j'(t, \omega) > 0 \} \). Since \( \varphi_j'(t_1, \omega) < 2J \), we get the result from a similar discussion used in the case with \( \varphi_j'(0, \omega) > 0 \) above. □

**Proof of Theorem 5.2.** For notational simplicity, we set \( \varphi_j = \varphi_j(Q'(t)) \). Fix \( t_0 > 0, r > 0 \) and \( \omega \in \mathbb{R}^\mathcal{F} \). (1) Assume that \( \varphi_j'(0) > \theta \sqrt{\lambda'} (t_0 \vee 1) \).

Let
\[ t_1 = \inf \{ t \geq 0 : \varphi_j'(t) = 0 \} \quad \text{and} \quad t_2 = \inf \{ t \geq 0 : \varphi_j'(t) = 0 \}. \]

We will study the following three possible cases separately: (1) \( t_1 \leq t_0 \leq t_2 \), (2) \( t_1 > t_0 \) and \( t_2 > t_0 \), and (3) \( t_2 \leq t_0 \).

**Case 1.** Assume that \( t_1 \leq t_0 \leq t_2 \). Let \( s_0 = t_1 \) and define
\[ s_{2i+1} = \inf \{ t > s_2 : \varphi_j'(t) > 0 \} \quad \text{and} \quad s_{2i+2} = \inf \{ t > s_{2i+1} : \varphi_j'(t) = 0 \} \quad \text{for } i = 0, 1, \ldots. \]

One of the following must hold: \( s_1 > t_0 \) or \( s_{2k+2} < t_0 < s_{2k+3} \) or \( s_{2k+1} \leq t \leq s_{2k+2} \) for some \( k \geq 0 \).

If \( s_1 > t_0 \), then
\[ \varphi_j'(t_0) \leq \varphi_j'(t_1) + (A'(t_0) - A'(t_1)) + \sum_{j=1}^J (S_j(t_0) - S_j(t_1)) \]
\[ \leq \varphi_j'(t_1) - \sqrt{\lambda'} \theta (t_0 - t_1) + 2 \sum_{j=1}^J \| S_j(t) - \mu_j t \|_{N_J^r} + 2 \| A'(t) - \lambda' t \|_{N_J^r}. \] (A.3)

By definition of \( t_1 \),
\[ \varphi_j'(t_1) \leq \varphi_j'(0) - N_{\min} r_{\min} t_1 + 2 \sum_{j=1}^J \| S_j(t) - \mu_j t \|_{N_J^r}. \] (A.4)

For \( r \) large enough \( N_{\min} r_{\min} / \sqrt{\lambda'} > \theta / \mu_{\min} \), hence for such \( r \), we get (42) by combining (A.3) and (A.4).

Now, assume that either \( s_{2k+2} < t_0 < s_{2k+3} \) or \( s_{2k+1} \leq t \leq s_{2k+2} \) for some \( k \geq 0 \). We define \( \Delta_i = [s_{2i+1}, s_{2i+2}] \) for \( i = 0, 1, \ldots \). For any \( i \geq 0 \) and \( t \in [s_{2i+1}, s_{2i+2}] \), there exists at least one \( j_i \in \mathcal{J} \) such that \( Q_j(t) < N_j^r \).

Define
\[ a_j' = \inf \{ t \geq s_{2i+1} : Q_j'(t) < N_j^r \} \wedge s_{2i+2}. \]

We assume for simplicity that \( s_1 = a_1' \leq a_1' \leq \cdots \leq a_1' \leq s_{2i+2} \). If \( Q_j'(t) < N_j^r \) for some \( t \in [s_{2i+1}, s_{2i+2}] \), then \( Q_j'(s_{2i+2}) < N_j^r + 2 \), by Lemma D.1 and since no arrivals join queue \( j \) during \([s_{2i+1}, s_{2i+2}] \) when \( Q_j'(t) \geq N_j^r \).

Hence, if \( a_j' < s_{2i+2} \), then \( \varphi_j'(s_{2i+2}) < 2J \). Now, assume that \( a_j' = s_{2i+2} \). Then, for any \( t \in [s_{2i+1}, s_{2i+2}] \),
\[ \varphi_j'(t) \leq \varphi_j'(s_{2i+1}) - (A'(t) - A'(s_{2i+1})) + \sum_{j=1}^{J-1} (S_j(B_j'(t)) - S_j(B_j'(s_{2i+1}))) \]
\[ \leq 2J - \tilde{A}'(t) - \tilde{A}'(s_{2i+1}) + \sum_{j=1}^{J-1} (\tilde{S}_j(B_j'(t)) - \tilde{S}_j(B_j'(s_{2i+1}))) + (\theta \sqrt{\lambda'} - N_{\min} r_{\min}) (t - s_{2i+1}) \]
\[ \leq 2J + \tilde{Z}'(t_0). \] (A.5)

Choose \( l \leq k \), so that \( \Delta_i \) is the last interval such that \( a_j' < s_{2i+2} \wedge t_0 \) for all \( j \in \mathcal{J} \) if there exists such \( l \) or set \( l = k + 1 \). Hence \( \Delta_i \) is the last interval during which all the queues become empty at least once.

(a) Assume that \( l \leq k \). If \( s_{2i+1} \leq t_0 \), then
\[ \varphi_j'(t_0) < 2J. \] (A.6)
If $s_{2j+2} < t_0$, then (A.5) holds for all $t \in [s_{2j+2}, t_0]$ since if $t \in [s_{2j+2}, s_{2j+3}]$, then $\varphi'_j(t) = 0$ and otherwise we have from the discussion above that (A.5) holds. Hence, for $Q'_j(t) = \sum_{j' \in J} (Q'_j(t) + Z'_j(t)),$

$$Q'_j(t_0) - Q'_j(s_{2j+2}) \leq \tilde{\lambda}'(t_0) - \tilde{\lambda}'(s_{2j+2}) - \sum_{j' \in J} (S(B'_j(t_0)) - S(B'_j(s_{2j+2})))$$

$$\leq (\mu_{\text{max}}(2J + \tilde{\zeta}'(t_0)) - \sqrt{\tilde{\lambda}'(t_0)}(t_0 - s_{2j+2}) + 2 \sum_{j' \in J} ||S_j(t) - \mu_j t||_{N'_{00}} + 2||\tilde{A}'(t) - \tilde{\lambda}'t||_{t_0}.$$  

Since $\varphi'_j(t_0) \leq 2J + \tilde{\zeta}'(t_0)$, we have from the last inequality that

$$\varphi'_j(t_0) - \varphi'_j(s_{2j+2}) \leq (\mu_{\text{max}}(2J + \tilde{\zeta}'(t_0)) - \sqrt{\tilde{\lambda}'(t_0)}(t_0 - s_{2j+2})$$

$$+ 2 \sum_{j' \in J} ||S_j(t) - \mu_j t||_{N'_{00}} + 2||\tilde{A}'(t) - \tilde{\lambda}'t||_{t_0} + 2J + \tilde{\zeta}'(t_0).$$ (A.7)

Since $\varphi'_j(s_{2j+2}) < 2J$ by definition of $l$, we get (42) from (A.7).

(b) If $l = k + 1$, then (A.5) holds for all $t \in [t_1, t_0]$. Hence

$$\varphi'_j(t_0) - \varphi'_j(t_1) \leq (\mu_{\text{max}}(2J + \tilde{\zeta}'(t_0)) - \sqrt{\tilde{\lambda}'(t_0)}(t_0 - t_1)$$

$$+ 2 \sum_{j' \in J} ||S_j(t) - \mu_j t||_{N'_{00}} + 2||\tilde{A}'(t) - \tilde{\lambda}'t||_{t_0} + 2J + \tilde{\zeta}'(t_0).$$ (A.8)

For $r$ large enough $N_{\text{min}}' / \sqrt{\tilde{\lambda}} > \theta / \mu_{\text{min}}$ hence for such $r$, we get (42) by combining (A.8) and (A.4).

Case 2. Now, assume that $t_1 > t_0$ and $t_2 > t_0$. Then, none of the arrivals join a queue that has customers waiting during $[0, t_0]$. Let $a_j = \inf\{t > 0: Q'_j(t) < N'_j\}$. As in Case 1, if $a_j \leq t_0$ for all $j \in J$, then $\varphi'_j(t_0) < 2J$. If $a_j > t_0$ for some $j \in J$, then

$$\varphi'_j(t_0) - \varphi'_j(0) \leq -N'_{\text{min}}\mu_{\text{min}}t_0 + 2 \sum_{j \in J} ||S_j(t) - \mu_j t||_{N'_{00}}.$$ (A.9)

Hence, for $r$ large enough, we get (42).

Case 3. If $t_2 < t_0$, define

$$\tilde{t}'_2 = \sup\{t > t_0: \varphi'_j(t) > 0\}. $$

If $\tilde{t}'_2 \geq t_0$, then $\varphi'_j(t_0) = 0$, so assume that $\tilde{t}'_2 < t_0$. We have that $0 < \varphi'_j(\tilde{t}'_2) < 2J$. We get (42) from the discussion below.

(2) Now, assume that

$$\varphi'_j(Q'_j(0)) \leq \theta \sqrt{\lambda'}(t_0 \vee 1),$$

and define

$$\tilde{t} = \sup\{t_0 \geq t \geq 0: \varphi'_j(t) = 0\}.$$ 

If $\tilde{t} = -\infty$, set $\tilde{t} = 0$. Observe that $\theta \sqrt{\lambda'}(t_0 \vee 1) \geq \varphi'_j(\tilde{t}) > 0$ and $\varphi'_j(\tilde{t}) > 0$ for all $[\tilde{t}, t_0]$. By considering the path from $\tilde{t}$ to $t_0$, one can use the same arguments above, but this time only Cases 1 and 2 will apply. It is obvious that (A.3)–(A.9) hold for this case as well, since they do not depend on the initial value of $\varphi'_j$. Hence we get (43). □

Appendix B. Proofs of the results in §5.4. We first establish the additional fluid model equations that must be satisfied by the fluid limits of the MED-FSF and MED-LB distributed server pool systems. Then, using these equations, we determine the set of invariant states of the fluid limits for both systems.

Lemma B.1. Let $\{X_t\}$ be a sequence of MED-FSF distributed server pool systems. Assume that $\{Q'_j(0)\}$ is bounded a.s. as $r \to \infty$. Every fluid limit $\bar{X}$ of $\{X_t\}$ satisfies the following equations in addition to the fluid model equations (A.2)–(A.11) in Dai and Tezcan [15]. For every $j \in J$ and a regular point $t$ of $\bar{X}$,

$$\dot{\bar{A}}^j_j(t) = 0 \quad \text{when} \quad \frac{\dot{Q}'_j(t)}{\beta_j \mu_j} > \frac{\dot{Q}'_j(t)}{\beta_j \mu_j'} \quad \text{for some} \quad j' \in J \quad \text{and}$$

$$\sum_{j' = j}^{j} \dot{\bar{A}}^j_j(t) = \lambda \quad \text{if} \sum_{j' = j}^{j} \bar{Z}_{j_j}(t) - \beta_j < 0. \quad \text{(B.1)}$$

For every $j \in J$ and a regular point $t$ of $\bar{X}$,
Proof. Let \( \bar{X} \) be a fluid limit. Fix \( t > 0 \) and assume that \( \frac{\bar{Q}_j(t)}{\bar{\beta}_j \mu_j} > \frac{\bar{Q}_j(t)}{\bar{\beta}_j \mu_j} \) for some \( j' \in \mathcal{J} \). By continuity of \( \bar{X} \), there exists \( \delta > 0 \) such that

\[
\frac{\bar{Q}_j(s)}{\beta_j \mu_j} > \frac{\bar{Q}_j(s)}{\beta_j \mu_j}
\]

for all \( s \in [t - \delta, t + \delta] \). Let \( \omega \in \mathcal{A} \) for \( \mathcal{A} \) given as in §4.4. Assume for notational simplicity that \( \bar{X}^\prime(\cdot, \omega) \) converges u.o.c. to \( \bar{X}(\cdot, \omega) \). Note that by (2),

\[
N^\prime_\ell = \beta_\ell |N^\prime| + o(|N^\prime|) \quad \text{for all } \ell \in \mathcal{J}.
\]

Hence, for \( r \) large enough

\[
\frac{\bar{Q}_j(s)}{\beta_j \mu_j + o(|N^\prime|)/|N^\prime|} > \frac{\bar{Q}_j(s)}{\beta_j \mu_j - o(|N^\prime|)/|N^\prime|}.
\]

so

\[
\frac{\bar{Q}_j(s)/|N^\prime|}{\beta_j \mu_j + o(|N^\prime|)/|N^\prime|} > \frac{\bar{Q}_j(s)/|N^\prime|}{\beta_j \mu_j - o(|N^\prime|)/|N^\prime|}.
\]

Thence

\[
\frac{Q'_j(s)}{N'_j \mu_j} \geq \frac{Q'_j(s)}{|N^\prime| \beta_j \mu_j + o(|N^\prime|)} > \frac{Q'_j(s)}{|N^\prime| \beta_j \mu_j - o(|N^\prime|)} > \frac{Q'_j(s)}{N'_j \mu_j}.
\]

for large enough \( r \) for all \( s \in [t - \delta, t + \delta] \). This implies by (30) for \( r \) large enough that \( A^{\prime, \prime}_j(t) \) is flat on \([t - \delta, t + \delta]\). Hence \( \bar{X}^\prime(t) = 0 \). Fluid limit Equation (B.2) is proved similarly.

Lemma B.2. Let \( \{X^\prime\} \) be a sequence of MED-LB distributed server pool systems. Assume that \( \{\bar{Q}^\prime(0)\} \) is bounded a.s. as \( r \to \infty \). Every fluid limit \( \bar{X} \) of \( \{X^\prime\} \) satisfies (B.1) and the following equation in the fluid model equations (A.2)–(A.8) in Dai and Tezcan [15]. For every \( j \in \mathcal{J} \) and a regular point \( t \) of \( \bar{X} \),

\[
\dot{A}_{j}(t) = 0 \quad \text{when } \frac{\bar{Z}_j(t)}{\beta_j} > \frac{\bar{Z}_{j'}(t)}{\beta_{j'}} \quad \text{for some } j' \in \mathcal{J}. \tag{B.4}
\]

The proof is similar to that of Lemma B.1.

Proof of Lemma 5.1. Let \( \{X^\prime\} \) be a sequence of MED-FSF distributed server pool systems and \( h_1(t) = \sum_{j=1}^{J} |\bar{Z}_j(t) - \beta_j| \). Note that if \( h_1(t) > 0 \) and \( t \) is a regular point of \( \bar{X} \), then \( h_1(t) \leq 0 \) from the fluid model Equation (B.2), and Equations (A.2) and (A.9) in Dai and Tezcan [15]. Hence, if \( h_1(0) = 0 \), then \( h_1(t) = 0 \) for all \( t \geq 0 \) by virtue of Lemma 2.4.5 of Dai [14].

Now, let \( h_1(t) = \max_{j \in \mathcal{J}} \frac{Q_j(t)}{\beta_j} - \min_{j \in \mathcal{J}} \frac{Q_j(t)}{\beta_j} \) and assume that \( \bar{Z}_j(0) = \beta_j \) for all \( j \in \mathcal{J} \). If \( t \) is a regular point of \( \bar{X} \) and \( h_1(t) > 0 \), then \( \bar{h}_1(t) \leq 0 \), by (B.1), (B.2), (A.2)–(A.11) in Dai and Tezcan [15] and Lemma 2.8.6 of Dai [14]. Hence, if \( h_1(0) = 0 \), then \( h_1(t) = 0 \) for all \( t \geq 0 \). So, if \( h_1(0) = h_1(0) = 0 \), then \( (\bar{Q}(0), \bar{Z}(0)) = (\bar{Q}(0), \bar{Z}(0)) \). Note that \( h_1(0) = h_2(0) = 0 \) if and only if \( (q, z) \in \mathcal{M} \). For the MED-LB policy, the result is proved similarly using (B.4). \( \square \)

Appendix C. Proofs of the results in §5.5.

C.1. Proofs of Propositions 5.1 and 5.2. We use the framework of Dai and Tezcan [15]. We need to check that Assumptions 1–4 in that paper are satisfied by the sequence of MED-FSF and MED-LB distributed systems. It can easily be checked by using (2) and (4) that the static planning problem (2.20) in Dai and Tezcan [15] has a unique optimal solution with \( x_{j'}^* = 1 \) for all \( j' \in \mathcal{J} \), and so by (4), Assumption 1 in that paper is satisfied. Assumption 2 in Dai and Tezcan [15] is satisfied by both policies by Lemma 5.1 and by Theorem A.1 in the same paper. So, we focus on Assumptions 2 and 4. In this section, we first define the appropriate SSC functions, see §4.1 of Dai and Tezcan [15] for more details on SSC functions, then we show that Assumption 3 in that paper is satisfied by the MED-LB and MED-FSF distributed systems.

Hydrodynamic scaling and hydrodynamic limits are introduced in Dai and Tezcan [15]. They showed that hydrodynamic limits satisfy a set of equations that are called hydrodynamic model equations. To check Assumption 4 in Dai and Tezcan [15], one needs to show that the hydrodynamic model solutions, which are solutions of the hydrodynamic model equations, satisfy certain conditions. We start with characterizing the additional hydrodynamic model equations for the MED-FSF and MED-LB policies; see Lemmas C.1 and C.2. Then, we show that hydrodynamic model equations satisfy Assumption 4 in Dai and Tezcan [15] in Propositions C.1 and C.2. This gives us the multiplicative SSC results. To prove the strong SSC results stated in Propositions 5.1
and 5.2, we show in Theorem C.1 that condition (4.15) in Dai and Tezcan [15] is satisfied by \( \hat{Q} \) and \( \hat{Z} \) under any completely nonidling policy.

Recall that \( \lambda = \lim_{r \to \infty} \frac{\lambda'}{|N'|} = \mu \).

**Lemma C.1.** Let \( \{X^r\} \) be a sequence of MED-FSF distributed server pool systems. Every hydrodynamic limit \( \tilde{X} \) of \( \{X^r\} \) satisfies the following equations in addition to Equations (4.9)–(4.14) in Dai and Tezcan [15]. For every \( j \in \mathcal{J} \),

\[
\dot{\hat{X}}_j^r(t) = 0 \quad \text{when} \quad \frac{\tilde{Q}_j(t)}{\beta_j \mu_j} > \frac{\tilde{Q}_j'(t)}{\beta_j' \mu_j} \quad \text{for some} \quad j' \in \mathcal{J} \quad \text{and} \quad (C.1)
\]

\[
\dot{\hat{A}}_j^r(t) = \lambda \quad \text{when} \quad \sum_{l=j}^\infty \hat{Z}_l(t) < 0, \quad (C.2)
\]

where \( \dot{\hat{A}}_j^r(t) = \sum_{l=j}^\infty \dot{\hat{A}}_l^r(t) \).

**Proof.** Let \( X^{r,m} \) be the hydrodynamically scaled version of \( X^r \) as defined in §5.1 in Dai and Tezcan [15]. Then, by (30) and (31), \( X^{r,m} \) satisfies the following equations:

\[
A_j^{q,r,m}(t) \text{ can only increase when } \frac{Q_j^{r,m}(t)}{N'_j \mu_j} \leq \frac{Q_j^{r,m}(t)}{N'_j \mu_j} \quad \text{for all } j' \in \mathcal{J} \quad \text{and} \quad (C.3)
\]

\[
A_j^{q,r,m}(t) \text{ can only increase when } \sum_{l=j+1}^\infty Z_j^{r,m}(t) = 0. \quad (C.4)
\]

Let \( \tilde{X} \) be a cluster point of \( \{X^{r,m}\} \) for some \( L > 0 \).

Fix \( T > 0 \). By definition of a hydrodynamic limit, given \( \epsilon > 0 \), one can choose \( (r, m, \omega) \) such that

\[
\| \tilde{X}(t) - X^{r,m}(t, w) \|_\infty \leq \epsilon. \quad (C.5)
\]

The rest of the proof is similar to the proof of the first part of Lemma 5.1. Fix \( L > t > 0 \) and \( j \in \mathcal{J} \). Assume that

\[
\frac{\tilde{Q}_j(t)}{\beta_j \mu_j} > \frac{\tilde{Q}_j'(t)}{\beta_j' \mu_j} \quad \text{for some} \quad j' \in \mathcal{J}. \]

By continuity of \( \tilde{X} \), there exists a \( \delta > 0 \) such that

\[
\frac{\tilde{Q}_j(s)}{\beta_j \mu_j} \geq \frac{\tilde{Q}_j'(s)}{\beta_j' \mu_j} \quad \text{for all} \quad s \in [t - \delta, t + \delta]. \]

Since \( \epsilon \) is arbitrary, by (C.5), one can choose \( (r, m, \omega) \) such that

\[
\frac{Q_j^{r,m}(s)}{\beta_j \mu_j} \geq \frac{Q_j^{r,m}(s)}{\beta_j \mu_j} \quad \text{for all} \quad s \in [t - \delta, t + \delta]. \]

Since \( \beta_j = N'_j / |N'| - o(N')/|N'| \), this implies, similar to (B.3), by (C.3) that \( A_j^{q,r,m}(s) \) is flat on \( [t - \delta, t + \delta] \). Hence \( A^q(t) \) cannot increase on \( [t - \delta, t + \delta] \). The second hydrodynamic Equation (C.2) is proved similarly using (C.4). \( \Box \)

**Lemma C.2.** Let \( \{X^r\} \) be a sequence of MED-LB distributed server pool systems. Then, in addition to Equations (4.9)–(4.14) in Dai and Tezcan [15], every hydrodynamic limit \( \tilde{X} \) of \( \{X^r\} \) satisfies (C.1) and

\[
\dot{\hat{A}}_j(t) = 0 \quad \text{when} \quad \frac{\tilde{Z}_j(t)}{\beta_j} > \frac{\tilde{Z}_j'(t)}{\beta_j} \quad \text{for some} \quad j' \in \mathcal{J}. \quad (C.6)
\]

The proof is similar to that of Lemma C.1.

Next, we define the SSC function for the MED-FSF policy. Let \( \hat{f}_j : \mathbb{R}^{2J} \to \mathbb{R} \) be defined by

\[
\hat{f}_j(q, z) = \frac{q_j}{\mu_j \beta_j}.
\]
for \( j \in \mathcal{J} \). The SSC function, \( \hat{g} : \mathbb{R}^{2J} \to \mathbb{R} \), for the MED-FSF policy is defined as follows:

\[
\hat{g}(q, z) = \max_{j \in \mathcal{J}} \{ f_j(q, z) \} - \min_{j \in \mathcal{J}} \{ f_j(q, z) \} + \sum_{j=2}^{J} \hat{Z}_j(t). \tag{C.7}
\]

It is obvious that \( \hat{g} \) is continuous and \( \hat{g}(aq, \alpha z) = \alpha \hat{g}(q, z) \) for all \((q, z) \in \mathbb{R}^{2J}\). Hence \( \hat{g} \) satisfies Assumption 3 in Dai and Tezcan [15]. Also, \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) = 0 \) if and only if

\[
\frac{\hat{Q}_j(t)}{\beta_j \mu_j} = \frac{\hat{Q}_j(t)}{\beta_j \mu_j} \quad \text{and} \quad \sum_{j=2}^{J} \hat{Z}_j(t) = 0 \quad \text{for all} \quad j, j' \in \mathcal{J}.
\]

Therefore \( \hat{g} \) is the desired SSC function. Next, we show that the hydrodynamic model solutions and \( \hat{g} \) satisfy Assumption 4 in Dai and Tezcan [15].

**Proposition C.1.** Let \( \{X'\} \) be a sequence of MED-FSF distributed server pool systems. Let \( \hat{g} \) be defined as in (C.7). For any hydrodynamic model solution \( \tilde{X} \),

\[
\hat{g}(\hat{Q}(t), \hat{Z}(t)) \leq H(t) \quad \text{for} \quad t \geq 0
\]

with \( H(t) = (\hat{g}(\hat{Q}(0), \hat{Z}(0)) - (\mu_{\min} \beta_{\min} \wedge 1)) t \wedge 0 \). Whenever \( \hat{g}(\hat{Q}(0), \hat{Z}(0)) = 0 \), then \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) = 0 \) for \( t \geq 0 \). In particular, the hydrodynamic model solutions of \( \{X'\} \) and \( \hat{g} \) satisfy Assumption 4 in Dai and Tezcan [15].

**Proof.** Let \( \tilde{X} \) be a hydrodynamic model solution and \( t \geq 0 \) a regular point of \( \tilde{X} \). Assume that \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) > 0 \).

Let

\[
\mathcal{U}_{\min}(t) = \{ j \in \mathcal{J} : \hat{f}_j(\hat{Q}(t), \hat{Z}(t)) \leq \hat{f}_j(\hat{Q}(t), \hat{Z}(t)) \quad \text{for all} \quad j' \in \mathcal{J} \} \quad \text{and} \quad \mathcal{U}_{\max}(t) = \{ j \in \mathcal{J} : \hat{f}_j(\hat{Q}(t), \hat{Z}(t)) \geq \hat{f}_j(\hat{Q}(t), \hat{Z}(t)) \quad \text{for all} \quad j' \in \mathcal{J} \}.
\]

Since \( t \) is a regular point of \( \tilde{X} \), by Lemma 2.8.6 of Dai [14],

\[
\hat{g}(\hat{Q}(t), \hat{Z}(t)) = \hat{f}_i(t) - \hat{f}_i(t) - \sum_{j=2}^{J} \hat{Z}_j(t) \quad \text{for all} \quad i \in \mathcal{U}_{\max}(t) \quad \text{and} \quad j \in \mathcal{U}_{\min}(t).
\]

First, assume that \( \sum_{j=2}^{J} \hat{Z}_j(t) < 0 \). Then,

\[
\sum_{j=2}^{J} \hat{Z}_j(t) \geq \lambda - \sum_{j=2}^{J} \mu_j \beta_j \geq \mu_{\min} \beta_{\min}
\]

by (C.2) and (4.5) in Dai and Tezcan [15]. Also, by (C.1) and (4.8) in Dai and Tezcan [15], \( \hat{f}_j(t) = -1 \) or 0 and \( \hat{f}_j(t) = 0 \) for all \( i \in \mathcal{U}_{\max}(t) \) and \( j \in \mathcal{U}_{\min}(t) \). Hence \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) \leq -\mu_{\min} \beta_{\min} \).

Now, assume that \( \sum_{j=2}^{J} \hat{Z}_j(t) = 0 \). Then, \( \sum_{j=2}^{J} \hat{Z}_j(t) = 0 \) since \( \sum_{j=2}^{J} \hat{Z}_j(t) > 0 \) whenever \( \sum_{j=2}^{J} \hat{Z}_j(t) < 0 \). Hence

\[
\hat{g}(\hat{Q}(t), \hat{Z}(t)) = \hat{f}_i(t) - \hat{f}_i(t).
\]

We get as in the proof of Proposition C.2 below that \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) \leq -(-\mu_{\min} \wedge 1) \). This gives the first claim. Second claim follows from the fact that \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) < 0 \) if \( \hat{g}(\hat{Q}(t), \hat{Z}(t)) > 0 \). \( \square \)

Next, we define the SSC function for the MED-LB policy. Let \( q = (q_1, \ldots, q_J) \in \mathbb{R}^J \), \( z = (z_1, \ldots, z_J) \in \mathbb{R}^J \), and \( f_j : \mathbb{R}^{2J} \to \mathbb{R} \) be defined by

\[
f_j(q, z) = \frac{q_j}{\mu_j \beta_j} + \frac{z_j}{\beta_j}
\]

for \( j \in \mathcal{J} \). The SSC function, \( g : \mathbb{R}^{2J} \to \mathbb{R} \), for MED-LB policy is defined by

\[
g(q, z) = \max_{j \in \mathcal{J}} \{ f_j(q, z) \} - \min_{j \in \mathcal{J}} \{ f_j(q, z) \}. \tag{C.8}
\]
It is easily checked that \( g \) is continuous and \( g(\alpha q, \alpha z) = \alpha g(q, z) \) for all \((q, z) \in \mathbb{R}^{2j}\). Hence \( g \) satisfies Assumption 3 in Dai and Tezcan [15]. Also,
\[
g(\tilde{Q}(t), \tilde{Z}(t)) = 0
\]
if and only if
\[
\frac{\tilde{Q}_j(t)}{\beta_j \mu_j} = \frac{\tilde{Q}_{j'}(t)}{\beta_{j'} \mu_{j'}} \quad \text{and} \quad \frac{\tilde{Z}_j(t)}{\beta_j} = \frac{\tilde{Z}_{j'}(t)}{\beta_{j'}} \quad \text{for all } j, j' \in \mathcal{J}.
\]
Therefore, \( g \) is the desired SSC function. Next, we show that \( g \) satisfies Assumption 4 in Dai and Tezcan [15].

**Proposition C.2.** Let \( \{X'\} \) be a sequence of MED-LB distributed server pool systems. Let \( g \) be defined as in (C.8). For any hydrodynamic model solution \( \tilde{X} \),
\[
g(\tilde{Q}(t), \tilde{Z}(t)) \leq H(t) \quad \text{for } t \geq 0
\]
with \( H(t) = (g(\tilde{Q}(0), \tilde{Z}(0)) - (\mu_{\text{min}} \wedge 1) t) \wedge 0 \). Whenever \( g(\tilde{Q}(0), \tilde{Z}(0)) = 0 \), then \( g(\tilde{Q}(t), \tilde{Z}(t)) = 0 \) for \( t \geq 0 \).

In particular, \( \{X'\} \) with \( g \) satisfy Assumption 4 in Dai and Tezcan [15].

**Proof.** Let \( \tilde{X} \) be a hydrodynamic limit and \( t \geq 0 \) be a regular point of \( \tilde{X} \).
\[
\text{Assume that } g(\tilde{Q}(t), \tilde{Z}(t)) > 0.
\]
Let
\[
\mathcal{U}_{\text{min}}(t) = \{ j \in \mathcal{J} : f_j(\tilde{Q}(t), \tilde{Z}(t)) \leq f_j(\tilde{Q}(t), \tilde{Z}(t)) \text{ for all } j' \in \mathcal{J} \} \quad \text{and}
\]
\[
\mathcal{U}_{\text{max}}(t) = \{ j \in \mathcal{J} : f_j(\tilde{Q}(t), \tilde{Z}(t)) \geq f_j(\tilde{Q}(t), \tilde{Z}(t)) \text{ for all } j' \in \mathcal{J} \}.
\]

Since \( t \) is a regular point of \( \tilde{X} \), by Lemma 2.8.6 of Dai [14],
\[
\dot{g}(\tilde{Q}(t), \tilde{Z}(t)) = \dot{f}_i(\tilde{Q}(t), \tilde{Z}(t)) - \dot{f}_j(\tilde{Q}(t), \tilde{Z}(t)) \quad \text{for all } i \in \mathcal{U}_{\text{max}}(t) \text{ and } j \in \mathcal{U}_{\text{min}}(t).
\]

Also, observe that \( \mathcal{U}_{\text{max}}(t) \cap \mathcal{U}_{\text{min}}(t) = \emptyset \) since \( g(\tilde{Q}(t), \tilde{Z}(t)) > 0 \).

We first show that if \( g(\tilde{Q}(t), \tilde{Z}(t)) > 0 \), then
\[
\dot{f}_i(\tilde{Q}(t), \tilde{Z}(t)) \leq - (\mu_{\text{min}} \wedge 1) \quad \text{for all } i \in \mathcal{U}_{\text{max}}(t).
\]

First, observe that if \( g(\tilde{Q}(t), \tilde{Z}(t)) > 0 \) and \( i \in \mathcal{U}_{\text{max}}(t) \)
\[
\frac{\tilde{Q}_j(t)}{\mu_j \beta_j} > \frac{\tilde{Q}_{j'}(t)}{\mu_{j'} \beta_{j'}} \quad \text{or} \quad \frac{\tilde{Z}_j(t)}{\beta_j} > \frac{\tilde{Z}_{j'}(t)}{\beta_{j'}}
\]
for some \( j \in \mathcal{U}_{\text{min}}(t) \). Therefore, by (C.1), (C.6), Equations (4.2) and (4.5) in Dai and Tezcan [15]
\[
\dot{\tilde{Q}}_i(t) + \dot{\tilde{Z}}_i(t) \leq - \mu_{\text{min}} \beta_{\text{min}}.
\]

In addition, either \( \dot{\tilde{Q}}_i(t) = 0 \) or \( \dot{\tilde{Z}}_j(t) = 0 \) by (4.7) in Dai and Tezcan [15]. This gives (C.12).

Next, we show that if \( g(\tilde{Q}(t), \tilde{Z}(t)) > 0 \), then
\[
\dot{f}_i(\tilde{Q}(t), \tilde{Z}(t)) \geq 0 \quad \text{for all } i \in \mathcal{U}_{\text{min}}(t).
\]

By (4.1), (4.2), and (4.5) in Dai and Tezcan [15],
\[
\sum_{i \in \mathcal{U}_{\text{min}}(t)} \dot{\tilde{Q}}_i(t) + \dot{\tilde{Z}}_i(t) = \lambda - \sum_{i \in \mathcal{U}_{\text{max}}(t)} \mu_i \beta_i > 0,
\]
where the last inequality follows from the fact that \( \mathcal{U}_{\text{min}}(t) \neq \emptyset \). By (C.11), we get (C.13). Combining (C.12) with (C.13) gives (C.9). The second claim immediately follows from Lemma 2.8.6 of Dai [14] and the fact that \( \dot{g}(\tilde{Q}(t), \tilde{Z}(t)) < 0 \) whenever \( g(\tilde{Q}(t), \tilde{Z}(t)) > 0 \) as shown above. \( \square \)

Next, we show that under a nonidling routing policy, the sequence of distributed systems satisfy condition (4.15) in Dai and Tezcan [15].
THEOREM C.1. Let $\pi \in \Pi$ be a nonidling routing policy and assume that (19) holds. Then, for every $T > 0$,
\[
\lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}[\|\hat{L}(t)\|_T \lor \|\hat{L}'(t)\|_T > C] = 0.
\] (C.14)

PROOF. Fix $\pi \in \Pi$, $T > 0$, and assume that (19) holds. Observe that (19) implies
\[
\lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}[\|\hat{L}(0)\| \lor |\hat{L}'(0)| > C] = 0.
\] (C.15)

By (35), (38), and (39), for $C > 0$,
\[
\mathbb{P}[\|\hat{L}(t)\|_T > C] \leq \mathbb{P}\left\{\|\hat{L}(0)\| + \theta \sqrt{N'/|N'|}T + |o(\sqrt{|N'|})|/\sqrt{|N'|}
\right.
\]
\[
+ 2 \sum_{j=1}^{J} \left\{\frac{\|J_{tj}(|N'|t) - |N'|\mu_j\|^2}{\sqrt{|N'|}} + 2 \frac{\|A'(t) - \lambda't\|_T}{\sqrt{|N'|}}\right\} > C/4
\]
\[
\leq \mathbb{P}[\|\hat{L}(0)\| > C/4] + \mathbb{P}\left[\theta \sqrt{N'/|N'|}T + |o(\sqrt{|N'|})|/\sqrt{|N'|} > C/4\right]
\]
\[
+ \mathbb{P}\left\{2 \sum_{j=1}^{J} \left\{\frac{\|J_{tj}(|N'|t) - |N'|\mu_j\|^2}{\sqrt{|N'|}}\right\} > C/4\right\}
\]
For any $j \in J$, $(J_{tj}(|N'|t) - |N'|\mu_j)/\sqrt{|N'|}$ converges weakly to a Brownian motion with variance $\mu_j$, and $(A'(t) - \lambda't)/\sqrt{|N'|}$ converges weakly to a Brownian motion with variance $\lambda$. Hence, by the continuous mapping theorem
\[
\lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}\left\{2 \sum_{j=1}^{J} \frac{\|J_{tj}(|N'|t) - |N'|\mu_j\|^2}{\sqrt{|N'|}} > C/4\right\}
\]
\[
= \lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}\left\{2 \frac{\|A'(t) - \lambda't\|_T}{\sqrt{|N'|}} > C/4\right\} = 0.
\] (C.16)

Thus, by (C.15) and (C.16),
\[
\lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}[\|\hat{L}(t)\|_T > C] = 0.
\]

By (35), (42), and (43), for $C > 0$,
\[
\mathbb{P}[\|\hat{L}(t)\|_T > C] \leq \mathbb{P}[\|\hat{L}(0)\| > C/5] + \mathbb{P}\left[\theta \sqrt{N'/|N'|}T + |o(\sqrt{|N'|})|/\sqrt{|N'|} > C/5\right]
\]
\[
+ \mathbb{P}\left\{2T(\mu \max \lor 1)\zeta(T)/\sqrt{|N'|} > C/5\right\}
\]
\[
+ \mathbb{P}\left\{4 \sum_{j=1}^{J} \frac{\|J_{tj}(|N'|t) - |N'|\mu_j\|^2}{\sqrt{|N'|}} > C/5\right\}
\]
By (C.15) and (C.16) it is enough to show that for any $T > 0$,
\[
\lim_{C \to \infty} \lim_{r \to \infty} \mathbb{P}\left\{2T(\mu \max \lor 1)\zeta(T)/\sqrt{|N'|} > C/5\right\} = 0
\] (C.17)
to complete the proof of (C.14). For notational simplicity, assume that $\mu \max > 1$ and choose $r$ large enough so that $N_{\max} - \theta \sqrt{N} > 2|N'|B_j$ for some $B_j > 0$. Observe that
\[
\mathbb{P}\left\{2 \mu \max T \frac{\zeta(T)}{\sqrt{|N'|}} > C\right\}
\]
\[
\leq \mathbb{P}\left\{2 \mu \max T \sup_{0 \leq s_1 \leq s_2 \leq T} \left\{-\sqrt{|N'|}B_j(s_2 - s_1) + \frac{|A'(s_2) - \hat{A}'(s_1)|}{\sqrt{|N'|}}\right\} > C/2\right\}
\]
\[
+ \sum_{j=1}^{J} \mathbb{P}\left\{2 \mu \max T \sup_{0 \leq s_1 \leq s_2 \leq T} \left\{-\sqrt{|N'|}B_j(s_2 - s_1) + \frac{|J_{tj}(|N'|s_2) - \hat{J}_{tj}(|N'|s_1)|}{\sqrt{|N'|}}\right\} > C/(2J)\right\}
\]
\[
\leq \mathbb{P}\left\{4 \mu \max T \frac{\|A'(t) - \lambda't\|_T}{\sqrt{|N'|}} > C/2\right\} + \sum_{j=1}^{J} \mathbb{P}\left\{4 \mu \max T \frac{\|J_{tj}(|N'|t) - |N'|\mu_j\|^2}{\sqrt{|N'|}} > C/(2J)\right\}.
\]
We get (C.17), again, by virtue of the continuous mapping theorem. □
PROOF OF PROPOSITION 5.1. Let \( \{X^r\} \) be a sequence of MED-FSF distributed server pool systems. Assume that (2), (4), and (19) hold. We showed above that this sequence with \( \hat{g} \) defined as in (C.7) satisfy Assumptions 1–4 in Dai and Tezcan [15]. So, we conclude by Theorem 4.2 in the same paper that for some \( L' = o(\sqrt{|N'|}) \) with \( L' \to \infty \) as \( r \to \infty \), and for every \( T > 0 \) and \( \epsilon > 0 \),

$$
P \left\{ \sup_{L'/\sqrt{|N'|}, t \leq T} \left| \frac{\hat{g}(\hat{Q}^r(t), \hat{Z}^r(t))}{\sqrt{|N'|}} \right| > \epsilon \right\} \to 0, \tag{C.18}$$

as \( r \to \infty \). Combining (C.18) with Theorem C.1 and using Remark 4.4 in Dai and Tezcan [15] and Theorem C.1 above that (48) holds. □

Proposition 5.2 is proved similarly by using \( g \), defined in (C.8), instead of \( \hat{g} \).

C.2. Proofs of the results in \$5.5.2\$.

PROOF OF PROPOSITION 5.3. Let \( \{X^r\} \) be a sequence of MED-FSF distributed server pool systems. Assume that (2), (4), (19), and (48) hold.

By (48),

$$
\left\| \frac{\hat{Q}^r_j(t)}{\bar{\beta}_j \mu_j} - \frac{\hat{Q}^r_j(t)}{\bar{\beta}_j \mu_j} \right\|_T \simeq \left\| \frac{\bar{\beta}_j}{\bar{\beta}_j} \right\|_T \leq \epsilon(r), \tag{C.19}
$$

for \( j, j' \in J \), where \( \epsilon(r) \to 0 \) as \( r \to \infty \) in probability. This gives

$$
\sum_{j=1}^{J} \mu_j \int_{0}^{t} \hat{Z}^r_j(s) \, ds = \mu_i \int_{0}^{t} \hat{Z}^r_i(s) \, ds + \epsilon(r)
$$

and

$$
(\hat{X}^r(t))^+ = -\hat{Z}^r_i(t) + J \epsilon(r). \tag{C.20}
$$

Observe that

$$
\hat{X}^r(t) = \hat{X}^r(0) + \left( \frac{A'(t) - \lambda^r t}{\sqrt{|N'|}} \right) - \sum_{j=1}^{J} \left( \frac{S_j(|N'| \int_{0}^{t} \hat{Z}^r_j(s) \, ds)}{|N'|} \right) \left( \frac{1}{\sqrt{|N'|}} \right)^2 - \sum_{j=1}^{J} \mu_j \int_{0}^{t} \hat{Z}^r_j(s) \, ds + \frac{(\lambda^r - \sum_{j=1}^{J} \mu_j N_j^r)}{|N'|} t. \tag{C.21}
$$

By (2), (4) and (C.19)–(C.21),

$$
\sum_{j=1}^{J} \mu_j \int_{0}^{t} \hat{Z}^r_j(s) \, ds = \mu_i \int_{0}^{t} \hat{Z}^r_i(s) \, ds + J \epsilon(r) = -\mu_i \int_{0}^{t} (\hat{X}^r(s))^+ \, ds + 2J \epsilon(r). \tag{C.22}
$$

Also,

$$
\left( \frac{A'(t) - \lambda^r t}{\sqrt{|N'|}} \right) \Rightarrow W_u \quad \text{and} \quad \sum_{j=1}^{J} \left( \frac{S_j(|N'| \int_{0}^{t} \hat{Z}^r_j(s) \, ds)}{|N'|} \right) \left( \frac{1}{\sqrt{|N'|}} \right)^2 \Rightarrow W_d. \tag{C.23}
$$

where \( W_u \) and \( W_d \) are Brownian motions with variances equal to \( \mu_d \), by Lemma 5.1 and since \( \lambda^r / |N'| \to \mu_d \), as \( r \to \infty \). We combine (C.22)–(C.24) and appeal to the continuous mapping theorem to complete the proof. □

PROOF OF PROPOSITION 5.4. The proof is similar to the proof of Proposition 5.3 above. Let \( \{X^r\} \) be a sequence of MED-LB distributed server pool systems. Assume that (2), (4), and (19) hold. Let \( \bar{Z}^r_z(t) = \sum_{j=1}^{J} \bar{Z}^r_j(t) \).

By Proposition 5.2,

$$
\left\| \frac{\hat{Q}^r_j(t)}{\beta_j \mu_j} - \frac{\hat{Q}^r_j(t)}{\beta_j \mu_j} \right\|_T \simeq \left\| \frac{\beta_j}{\bar{\beta}_j} \frac{\bar{Z}^r_j(t) - \bar{Z}^r_z(t)}{T} \right\|_T \leq \epsilon(r)
$$

where \( \beta_j \) and \( \bar{\beta}_j \) are Brownian motions with variances equal to \( \mu_j \).
for all \( j, j' \in J \), where \( \epsilon(r) \to 0 \) as \( r \to \infty \) in probability. This gives
\[
(\hat{X'}(t))^+ = -\hat{Z}_j(t) + J\epsilon(r)
\]
and
\[
\sum_{j=1}^{J} \mu_j \int_0^t \hat{Z}_j(s) \, ds = \sum_{j=1}^{J} \mu_j \beta_j \int_0^t \hat{Z}_j(t) \, ds = \mu \int_0^t \hat{Z}_j(t) \, ds + J\epsilon(r).
\]
The other arguments in the previous proof can be repeated verbatim to conclude the proof. □

**Proof of Theorem 5.3.** Let \( \{X'\} \) be a sequence of MED-FSF distributed server pool systems. Assume that (2), (4), and (19) hold. We prove the theorem for \( J = 2 \), the proof for an arbitrary \( J \) is similar. By Propositions 5.1 and 5.3, and Theorem 11.4.5 of Whitt [43],
\[
(\hat{Q}'_j(t), \hat{Z}'_j(t), \hat{Q}'_j(t), \hat{Z}'_j(t)) \Rightarrow \left( \frac{\mu_1\beta_1}{\mu_1\beta_1 + \mu_2\beta_2} (X(t))^+, -(X(t))^-, \frac{\mu_2\beta_2}{\mu_1\beta_1 + \mu_2\beta_2} (X(t))^+, 0 \right), \tag{C.25}
\]
as \( r \to \infty \) in \( C([0, \infty)) \). Let
\[
\hat{D}'_j(t) = \frac{D_j'(t) - \mu_jN_j't}{\sqrt{|N|}}.
\]
Hence
\[
\hat{D}'_j(t) = \sqrt{|N|} \left( \frac{S_j(|N| \int_0^t \hat{Z}'_j(s) \, ds)}{|N|} - \mu_j \int_0^t \hat{Z}'_j(s) \, ds \right) + \mu_j \int_0^t \hat{Z}'_j(s) \, ds.
\]
By virtue of Theorem 11.5.1 of Whitt [43] and the continuous mapping theorem,
\[
\left\{ \mu_j \int_0^t \hat{Z}'_j(s) \, ds \right\} \tag{C.26}
\]
converges weakly to a continuous limit. Also, the sequence
\[
\left\{ \sqrt{|N|} \left( \frac{S_j(|N| \int_0^t \hat{Z}'_j(s) \, ds)}{|N|} - \mu_j \int_0^t \hat{Z}'_j(s) \, ds \right) \right\} \tag{C.27}
\]
converges weakly by Lemma 5.1 and the convergence together theorem, hence it is tight. From (C.26), (C.27), and Theorem 11.6.7 of Whitt [43], we have that \( \{\hat{D}'_j(t)\} \) is tight in uniform topology. Thence we have again from Theorem 11.6.7 of Whitt, (C.25), and the tightness of the scaled departure processes that the sequence
\[
\{(\hat{Q}_j^0, \hat{Z}_j^0, \hat{D}_j^0, \hat{Q}'_j, \hat{Z}'_j, \hat{D}'_j)\}
\]
is tight. Thus, there exists a subsequence \( r_k \) such that
\[
(\hat{Q}_j^0, \hat{Z}_j^0, \hat{D}_j^0, \hat{Q}'_j, \hat{Z}'_j, \hat{D}'_j) \Rightarrow (\hat{Q}_1, \hat{Z}_1, \hat{D}_1, \hat{Q}_2, \hat{Z}_2, \hat{D}_2)
\]
as \( k \to \infty \) for some process \((\hat{Q}_j, \hat{Z}_j, \hat{D}_j, \hat{Q}'_j, \hat{Z}'_j, \hat{D}'_j)\). Let \( a'_j(t) = A_j^q(t) + A_j^r(t) \) be the total number of arrivals to the \( j \)th pool by time \( t \). We define the diffusion scaled arrival process, \( \hat{a}'_j \), to queue \( j \) by
\[
\hat{a}'_j(t) = \sqrt{|N|} \left( \frac{a'_j(t) - \mu_jN_j't}{|N|} \right).
\]
Since
\[
\hat{a}'_j(t) = -\hat{Q}'_j(0) - \hat{Z}'_j(0) + \hat{Q}'_j(t) + \hat{Z}'_j(t) + \hat{D}'_j(t),
\]
we have by the continuous mapping theorem that
\[
(\hat{a}'_j, \hat{D}'_j, \hat{a}'_j, \hat{D}'_j) \Rightarrow (\hat{a}_1, \hat{D}_1, \hat{a}_2, \hat{D}_2),
\]
where \( \hat{a}_i(t) = \hat{Q}_i(t) + \hat{Z}_i(t) + \hat{D}_i(t) - \hat{Q}_i(0) - \hat{Z}_i(0) \) for \( i = 1, 2 \).

Note that the processes \( \hat{a}_i \) and \( \hat{D}_i \), for \( i = 1, 2 \), are continuous a.s. since all the tightness results above hold in uniform topology and \( \hat{a}_i(0) = 0 \) and \( \hat{D}_i(0) = 0 \), \( i = 1, 2 \). By using the corollary in Puhalskii [31], we have that
\[
(\sqrt{|N|} \hat{W}_1^0, \sqrt{|N|} \hat{W}_2^0) \Rightarrow \left( \hat{W}_1, \hat{W}_2 \right),
\]
where
\[
\hat{W}_i(t) = \frac{\hat{X}^+}{\mu} \text{ for } i = 1, 2 \tag{C.28}
\]
We note that for all \(i\) and \(j\) of the waiting time processes for each pool. To prove the convergence of \(W'\), we note that
\[
(W'_i(t) \wedge W'_j(t)) \leq W'(t) \leq (W'_i(t) \vee W'_j(t)) \quad \text{a.s.}
\]
for all \(r\) and \(t \geq 0\). Hence
\[
0 \leq \sqrt{|N^r|} (W'_i(t) \vee W'_j(t)) - \sqrt{|N^r|} W'(t) \\
\leq \sqrt{|N^r|} (W'_i(t) \wedge W'_j(t)) - \sqrt{|N^r|}(W'_i(t) \vee W'_j(t)) \quad \text{a.s.}
\]
The last term converges to zero by the continuous mapping theorem and from (C.28). But weak convergence to a deterministic limit implies convergence in probability, see, for example, Billingsley [7]. Therefore, \(\sqrt{|N^r|} (W'_i(t) \vee W'_j(t)) - \sqrt{|N^r|} W'(t)\) converges to zero in probability. We have the convergence of \(W'\) to \([X]'/\mu\) by virtue of Theorem 3.1. of Billingsley [7].

\[\square\]

C.3. Proofs of the results in §5.5.4.

C.3.1. Proof of Theorem 5.6. Fix a nonidling routing policy \(\pi \in \Pi\) and let \(\Xi^\pi_{\infty} (\cdot) = \Xi^\pi_{\infty} (n \cdot)/n\). This scaling is known as the conventional fluid scaling. (These are not related to the fluid scaling that are discussed in §5.4 and will not be used elsewhere in this paper outside this section.) Similar to Dai [14], \(\Xi^\pi_{\infty} \in \mathbb{D}^{2}\) is said to be a fluid limit of \(\{\Xi^\pi_{\infty}\}\) if there exists a subsequence \(\{n_k\}\) of \(\{n\}\) and \(\omega \in \Omega\), satisfying
\[
\lim_{k \to \infty} E'(\lambda' t)/t = \lambda' \quad \text{and} \quad \lim_{k \to \infty} S_j(t)/t = \mu_j
\]
for all \(j \in \mathcal{J}\) such that
\[
\lim_{k \to \infty} \Xi^\pi_{\infty} (\cdot, \omega) = \Xi^\pi_{\infty} (\cdot)
\]
for all \(j \in \mathcal{J}\). It can be shown as in Dai [13] that fluid limits for queueing systems with multiple servers exist and satisfy the following equations for all \(i \geq 0\):
\[
\lambda' i = \sum_{j=1}^{J} (\tilde{A}'_{ij}(t) + 

\tilde{A}'_{ij}(t) = \tilde{Q}'(0) + 

\tilde{Q}'(t) + \tilde{Y}'(t) = N^r_j \mu_j \quad \text{for all } j \in \mathcal{J}

\tilde{Y}'(t) \text{ can only increase when } \tilde{Q}'(t) = 0 \quad \text{for all } j \in \mathcal{J}

\tilde{A}'_{ij}, \tilde{A}'_{ij}^q, \tilde{T}'_j, \text{ and } \tilde{Y}'_j \text{ are nondecreasing for all } j \in \mathcal{J}.

We note that \(\tilde{Z}'_j(t) = 0\) for all \(t \geq 0\), since \(\tilde{Z}'_{ij}(t) \leq |N^r|/n_j\) and so goes to zero as \(k \to \infty\) for fixed \(r\). It is clear from these equations that every fluid limit is absolutely continuous, hence differentiable almost everywhere.

In this section, we show that the fluid model of \(\Xi^\pi_{\infty}\) is stable (see Definition 4.1 of Dai [13]) when \(\pi \in \Pi\) and (60) holds. Then, we appeal to Theorem 4.2 in Dai [13]. This theorem is applicable only to single server systems but can be extended to cover the systems with multiple servers. We omit the proof because it follows straightforwardly from the analysis in Dai [13].

\textbf{Proof.} Fix a routing policy \(\pi \in \Pi\) and \(r > 0\). Let \(\Xi'\) be a fluid limit of \(\{\Xi^\pi_{\infty}\}\). Fix a regular point \(t > 0\).

We first show that fluid limits of \(\{\Xi^\pi_{\infty}\}\) satisfy
\[
\tilde{A}'_{ij}(t) + \tilde{A}'_{ij}(t) = 0 \quad \text{when } \tilde{Q}'(t) > 0 \text{ and } \tilde{Q}'(t) = 0 \quad \text{for some } j' \in \mathcal{J}
\]
for any \(j \in \mathcal{J}\). To prove this, assume that \(\tilde{Q}'(t) > 0\) and \(\tilde{Q}'(t) = 0\). By continuity of \(\tilde{Q}'\), there exists a \(\delta > 0\) such that
\[
\tilde{Q}'(s) > 2\varepsilon \quad \text{and} \quad \tilde{Q}'(s) < \varepsilon/(2\alpha') \quad \text{for all } s \in [t - \delta, t + \delta] \text{ and for some } \varepsilon > 0.
\]
Let \(\Xi^\pi_{\infty} (\cdot, \omega) \rightarrow \Xi^\pi_{\infty} (\cdot) u.o.c.\) as \(k \to \infty\). Then, for \(k\) large enough,
\[
\tilde{Q}'_{ij}(s) > \varepsilon \quad \text{and} \quad \tilde{Q}'_{ij}(s) < \varepsilon/\alpha' \quad \text{for all } s \in [t - \delta, t + \delta].
Hence $Q'_j(n_s) > a'_r(t)Q'_j(n_s)$ for all $s \in [r - \delta, t + \delta]$. Therefore, by (29), $A''_j$ is flat on $s \in [n_k(t - \delta), n_k(t + \delta)]$. Note that $A''_j$ is flat on $s \in [n_k(t - \delta), n_k(t + \delta)]$ since $Q'_j(n_s) > 0$. This gives (C.30).

Let $\tilde Q'_j(t) = \sum_{j' \in J} \tilde Q'_j(t)$ and assume that $\tilde Q'_j(t) > 0$. First, assume that there exists $j' \in \mathcal{J}$ such that $\tilde Q'_j(t) = 0$. Then, since $\tilde Q'_j$ is absolutely continuous, and differentiable at time $t$ and attains a minimum at $t > 0$, $\tilde Q'_j(t) = 0$. Hence

$$\frac{\dot{\tilde Q}'(t)}{\tilde Q'_j(t)} = \sum_{j \in \mathcal{J}} \frac{\dot{\tilde Q}'(t)}{\tilde Q'_j(t)} + \sum_{j \in \mathcal{J}} \frac{\dot{\tilde Q}'(t)}{\tilde Q'_j(t)} \leq -\mu_{\text{min}} N'_{\text{min}}$$

by (C.29) and (C.30).

If $\tilde Q'_j(t) > 0$ for all $j \in \mathcal{J}$, then

$$\frac{\dot{\tilde Q}'(t)}{\tilde Q'_j(t)} = \lambda' - \sum_{j \in \mathcal{J}} \mu_j N'_{j} < -\epsilon$$

for some $\epsilon > 0$ by (60) and (C.29).

Hence, if $\tilde Q'_j(t) > 0$ and $t$ is a regular point, then $\tilde Q'_j(t) < -(\epsilon \wedge \mu_{\text{min}} N'_{\text{min}})$ so the fluid model of $\mathcal{X}_\pi$ is stable by Lemma 5.2 of Dai [13]. We conclude the existence of a stationary distribution of $(Q', Z')$ by Theorem 4.2 of Dai [13].

**C.3.2. Proof of Theorem 5.7.** Choose $t_0 > 0$ such that for large enough $r$

$$\mathbb{E}_x[\exp\{-\theta \sqrt{\lambda' / |N'| \sqrt{t_0}}\}]\mathbb{E}_x\left[\exp\left\{2 \frac{\|A'(t) - \lambda' t\|_{l_0}}{\sqrt{|N'| t_0}}\right\}\right]$$

$$\cdot \prod_{j=1}^{J} \mathbb{E}_x\left[\exp\left\{2 \frac{\sum_{j=0}^{J} \|S_j(t) - \mu_j t\|_{l_0}}{\sqrt{|N'| t_0}}\right\}\right] < 1/2.$$  

(C.31)

Note that the existence of such $t_0$ and $r$ is guaranteed by Lemma D.2.

Let $x_i \in \mathbb{R}^J$ for $i = 1, 2$ and $x = (x_1, x_2)$. We define $\Phi'_i(x) : \mathbb{R}^J \rightarrow \mathbb{R}$ by

$$\Phi'_i(x) = \exp\{(|N'| t_0)^{-1/2} \phi'_i(x_2)\},$$

(C.32)

where $t_0$ is as chosen in (C.31) and $\phi'_i$ is defined as in (36). We show using Theorem 5.1 that $\Phi'_i$ is a geometric Lyapunov function; see Definition 2 in Gamarnik and Zeevi [18], then we appeal to Theorem 5 in the same paper to complete the proof.

Recall that $\mathbb{P}_\pi'$ denotes the stationary distribution of $(Q', Z')$ under the routing policy $\pi$. We denote the expectation operator with respect to this distribution by $\mathbb{E}_\pi'$. We set

$$\mathbb{E}_x[\cdot] = \mathbb{E}_x[\cdot | Q'(0) = x_1, Z'(0) = x_2]$$

for $x = (x_1, x_2)$. $x_i = (x_{i1}, \ldots, x_{ij}) \in \mathbb{R}^J$ for $i = 1, 2$, with $x_1 \geq 0$, $0 \leq x_2 \leq N'_j$ and $x_i(N'_j - x_2) = 0$ for all $j \in \mathcal{J}$.

**Proposition C.3.** Let $\Phi'_i$ be defined as in (C.32). There exists $t_0 > 1$ and $0 < \gamma < 1$ such that for $r$ large enough,

$$\sup_{x \in \mathbb{R}^J : \Phi'_i(x) > \kappa} \mathbb{E}_x[\Phi'_i(Q'(t_0), Z'(t_0)) / \Phi'_i(x)] \leq \gamma \quad \text{and} \quad \gamma > (\mu_{\text{min}} \wedge 1).$$

(C.33)

where $\kappa = \exp\{4\theta \sqrt{\lambda' / |N'| \sqrt{t_0}}/(\mu_{\text{min}} \wedge 1)\}.$

**Proof.** Fix a $t_0 > 1$ that satisfies (C.31). Note that if $\Phi'_i(x) > \exp\{4\theta \sqrt{\lambda' / |N'| \sqrt{t_0}}/(\mu_{\text{min}} \wedge 1)\}$, then $\phi'_i(Z'(0)) > 4\theta \sqrt{N_0}/(\mu_{\text{min}} \wedge 1)$. Hence, by (38) for $r$ large enough,

$$\sup_{x \in \mathbb{R}^J : \Phi'_i(x) > \kappa} \left[\mathbb{E}_x[\Phi'_i(Q'(t_0), Z'(t_0)) / \Phi'_i(x)]\right] \leq 2 \mathbb{E}_x[\exp\{-\theta \sqrt{\lambda' / |N'| \sqrt{t_0}}\}]\mathbb{E}_x\left[\exp\left\{2 \frac{\|A'(t) - \lambda' t\|_{l_0}}{\sqrt{|N'| t_0}}\right\}\right]$$

$$\cdot \prod_{j=1}^{J} \mathbb{E}_x\left[\exp\left\{2 \frac{\sum_{j=0}^{J} \|S_j(t) - \mu_j t\|_{l_0}}{\sqrt{|N'| t_0}}\right\}\right].$$

This gives (C.33) by (C.31).
If $\Phi'_0(x) \leq \exp(4\theta/|N'|/\sqrt{t_0}/(\mu_{\min} \land 1))$, then $\Phi'_0(Q'(0)) \leq 4\theta\sqrt{t_0}/(\mu_{\min} \land 1)$. Hence, by (39),

$$\sup_{x \in \mathbb{R}^{2J}, \Phi'_0(x) > \kappa} \{ E_x [\Phi'_0(Q'(t_0), Z'(t_0))/\Phi'_0(x)] \} \leq \exp\{\theta \lambda'/|N'|(\sqrt{t_0})\}.$$ (C.38)

We get (C.34) by virtue of Lemma D.2.

**Lemma C.3.** There exist $t_0 > 0$ and $r_0$ such that for $r > r_0$,

$$\exp[-\theta \lambda'/|N'|(\sqrt{t_0}) + \sqrt{t_0}2(\mu_{\max} \lor 1)J/\sqrt{|N'|}]$$

$$\cdot \left( E_x [\exp\{4\sqrt{t_0}(\mu_{\max} \lor 1)\xi'(t_0)/\sqrt{|N'|})] ight)$$

$$+ E_x \left[ \exp \left( \frac{4 \|A'(t) - \lambda' t\|}{\sqrt{|N'|t_0}} \right) \right] \frac{J}{\prod_{j=1}^{J} E_x \left[ \exp \left( \frac{8 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|}{\sqrt{|N'|t_0}} \right) \right]} < B_1/2.$$ (C.35)

**Proof.** By Lemma D.2, there exists $r_1 > 0$,

$$E_x \left[ \exp \left( \frac{4 \|A'(t) - \lambda' t\|}{\sqrt{|N'|t_0}} \right) \right] \frac{J}{\prod_{j=1}^{J} E_x \left[ \exp \left( \frac{8 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|}{\sqrt{|N'|t_0}} \right) \right]} < B_1/2,$$ (C.36)

for some $J + 2 < B_1 < \infty$ and all $r > r_1$. Now, for large enough $t_2 > t_1$, $r_2$, and $r > r_2 > r_1$,

$$\exp[-\theta \lambda'/|N'|(\sqrt{t_0}) + \sqrt{t_0}2(\mu_{\max} \lor 1)J/\sqrt{|N'|}] < \frac{1}{4B_1}.$$ (C.37)

By Lemma D.3, we can choose $r_3$ large enough, so that for $r > r_3 > r_2$,

$$E_x \left[ \exp\{4\sqrt{t_2}(\mu_{\max} \lor 1)\xi'(t_0)/\sqrt{|N'|})] < B_1/2.$$ (C.38)

We get (C.35) by combining (C.36)–(C.38).

**Proposition C.4.** Let $\Phi'_0$ be defined as in (C.39). There exists $t_0 > 1$ and $0 < \gamma < 1$ such that for $r$ large enough,

$$\sup_{x \in \mathbb{R}^{2J}, \Phi'_0(x) > \kappa} \{ E_x [\Phi'_0(Q'(t_0), Z'(t_0))/\Phi'_0(x)] \} \leq \gamma \quad \text{and} \quad \gamma \equiv \sup_{x \in \mathbb{R}^{2J}} \{ E_x [\Phi'_0(Q'(t_0), Z'(t_0))/\Phi'_0(x)] \} < \infty,$$ (C.40)

where $\kappa = \exp[\theta \lambda'/|N'|(\sqrt{t_0})].$

**Proof.** Choose $t_0$ and $r_0$ as in Lemma C.3. Note that if $\Phi'_0(x) > \exp[\theta \lambda'/|N'|(\sqrt{t_0})], then $\Phi'_0(Q'(0)) > \theta \lambda'/t_0$. Hence, by (42),

$$\sup_{x \in \mathbb{R}^{2J}, \Phi'_0(x) > \kappa} \{ E_x [\Phi'_0(Q'(t_0), Z'(t_0))/\Phi'_0(x)] \}$$

$$\leq \exp\{-\theta \lambda'/|N'|(\sqrt{t_0}) + \sqrt{t_0}2(\mu_{\max} \lor 1)J/\sqrt{|N'|}\}$$

$$\cdot \left( E_x [\exp\{4\sqrt{t_0}(\mu_{\max} \lor 1)\xi'(t_0)/\sqrt{|N'|})] ight)$$

$$+ E_x \left[ \exp \left( \frac{4 \|A'(t) - \lambda' t\|}{\sqrt{|N'|t_0}} \right) \right] \frac{J}{\prod_{j=1}^{J} E_x \left[ \exp \left( \frac{8 \sum_{j=1}^{J} \|S_j(t) - \mu_j t\|}{\sqrt{|N'|t_0}} \right) \right]} < 1,$$ (C.41)

where the last inequality follows from Lemma C.3. This gives (C.40).
Now, assume that $\Phi_2(x) \leq \exp\{\theta \sqrt{\lambda/|N|}\sqrt{t_0}\}$. By (43),

$$\sup_{x \in \mathbb{R}^2: \Phi_2(x) > R} [E_x \{\Phi_2(Q'(t_0), Z'(t_0))/\Phi_2(x)\}]$$

$$\leq \exp\{2J + \theta \sqrt{\lambda/|N|}/\sqrt{t_0} + \sqrt{t_0}(\mu_{\max} \lor 1)J/\sqrt{|N|}\}$$

$$\cdot \left( E_x \left[ \exp\left\{4\sqrt{t_0}(\mu_{\max} \lor 1)z'(t_0)/\sqrt{|N|}\right\} \right] \right.$$ + $\left. E_x \left[ \exp\left\{4\sqrt{|N|}z_j(t)/\sqrt{|N|}\right\} \right] \right)^j \left[ \prod_{j=1}^J \left[ \exp\left\{8\sqrt{|N|/t_0} \right\} \right] \right)$.  

This gives (C.41) by Lemmas D.2 and D.3. □

**Proof of Theorem 5.7.** Let $\pi \in \Pi$ be a nonidling routing policy. We claim that for $r$ large enough,

$$\Pr_{\pi^r} \left\{ \sum_{j=1}^J Q_j(0)/\sqrt{|N|} > s \right\} \leq c_1 \exp\{-c_2s\} \quad \text{and}$$

$$\Pr_{\pi^r} \left\{ \sum_{j=1}^J N_j - Z'_j(0)/\sqrt{|N|} > s \right\} \leq c_1 \exp\{-c_2s\}$$

for some $c_1, c_2 > 0$.

Let $t_0$ be given as in Proposition C.3. By Theorem 5 of Game and Zeevi [18] and Proposition C.3,

$$E_{\pi^r} \{\Phi_1(Q'(0), Z'(0))\} \leq \frac{Q_1(t_0)\kappa}{1 - \gamma} \leq c_0 \exp \left\{ \frac{4\theta \sqrt{\lambda/|N|}/\sqrt{t_0}}{\mu_{\min} \land 1} \right\},$$

for some $c_0 > 0$. By Markov’s inequality,

$$\Pr_{\pi^r} \left\{ \exp\left\{\Phi_1(Q'(0))/\sqrt{|N|}t_0\right\} > \exp\{s\} \right\} \leq \exp\{-s\} E_{\pi^r} \{\Phi_1(Q'(0))\} < c_1 \exp\{-s\}$$

for some $c_0 > 0$. This gives (C.42). The second inequality (C.43) is proved similarly using Proposition C.4.

Theorem 5.7 immediately follows from (C.42) and (C.43) since both $Q_j(0)/\sqrt{|N|}$ and $(N_j - Z'_j(0))/\sqrt{|N|}$ are nonnegative. □

**C.3.3. Proofs of Theorems 5.8 and 5.9.** The proof is similar to that of Theorem 8 in Game and Zeevi [18]. Assume that (2) and (4) hold.

By Theorem 5.6, $(Q'(\infty), Z'(\infty))$ exists for each $r$, and by Theorem 5.7, the sequence $\{(\hat{Q}'(\cdot), \hat{Z}'(\cdot))\}$ is tight. Therefore, every subsequence of $\{(\hat{Q}'(\cdot), \hat{Z}'(\cdot))\}$ has a convergent subsequence. Hence it is enough to show that each convergent subsequence of $\{(\hat{Q}'(\cdot), \hat{Z}'(\cdot))\}$ converges to the same limit (Billingsley [7, p. 59]) and this limit has the same distribution with the stationary distribution of $(\hat{Q}, \hat{Z})$ that is given by

$$\hat{Q}_j(\cdot) = \frac{\mu_j \beta_j}{\sum_{j=1}^J \mu_j \beta_j} (\hat{X}(\cdot))^+, \quad \hat{Z}_j(\cdot) = (\hat{X}(\cdot))^-$$.  

for $j \geq 2$ by Remark 5.6, where $\hat{X}(\cdot)$ has the density given by (13).

For notational simplicity, let $\{(\hat{Q}'(\cdot), \hat{Z}'(\cdot))\}$ be a convergent subsequence and denote the weak limit by $(\hat{Q}'(\cdot), \hat{Z}'(\cdot))$. Let $\{\hat{X}'\}$ be a sequence of MED-FSP distributed systems with

$$\hat{Q}'(0), \hat{Z}'(0) \sim (\hat{Q}'(\cdot), \hat{Z}'(\cdot)),$$

i.e., $\{(\hat{Q}'(0), \hat{Z}'(0))\}$ has the stationary distribution. Then, by Proposition 5.1, for some $L' = o(\sqrt{|N|})$ with $L' \to \infty$ as $r \to \infty$, and for every $T > 0$ and $\epsilon > 0$,

$$\Pr \left\{ \sup_{L'/\sqrt{|N|} \leq t \leq T} \left| \frac{\hat{Q}'(t)}{\beta_j \mu_j} - \frac{\hat{Q}'_j(t)}{\beta_j \mu_j} \right| \lor \left| \sum_{j=1}^J \hat{Z}_j(t) \right| > \epsilon \right\} \to 0$$

as $r \to \infty$. 

Let
\[(q'(\cdot), z'(\cdot)) = (Q'(\cdot + L'/\sqrt{|N'|}), Z'(\cdot + L'/\sqrt{|N'|})).\]

By (C.44),
\[(q'(0), z'(0)) \sim (\hat{Q}'(\infty), \hat{Z}'(\infty)),\]  
(C.46)
since \((\hat{Q}'(\infty), \hat{Z}'(\infty))\) is the unique stationary distribution. Therefore, \{(q'(0), z'(0))\} satisfies the conditions of Proposition 5.3 by (C.45) and (C.46). Hence
\[(q', z') \Rightarrow (\hat{Q}, \hat{Z}),\]  
(C.47)
where \(\hat{Q}\) and \(\hat{Z}\) are given by (53) and (54), respectively, and \((\hat{Q}, \hat{Z}(0)) \sim (\hat{Q}'(\infty), \hat{Z}'(\infty))\). Fix \(t > 0\). Then, \((q'(t), z'(t)) \sim (\hat{Q}'(\infty), \hat{Z}'(\infty))\), again by stationarity of \((\hat{Q}'(\infty), \hat{Z}'(\infty))\). Since \{(q'(t), z'(t))\} converges weakly to \((\hat{Q}(t), \hat{Z}(t))\) by (C.47), \((\hat{Q}(t), \hat{Z}(t)) \sim (\hat{Q}'(\infty), \hat{Z}'(\infty))\). Hence \((\hat{Q}'(\infty), \hat{Z}'(\infty))\) is the unique stationary distribution of \((\hat{Q}, \hat{Z})\).

The weak convergence of \(W'(\infty)\) to \(\hat{X}(\infty)/\mu\) can be proved similarly by starting each process in its steady state and repeating the arguments in the proof of Theorem 5.3.

The proof of Theorem 5.9 is similar. \(\square\)

**Appendix D. Auxiliary results.**

**Lemma D.1.** Let \(M\) be a renewal process with interarrival times given by the sequence of i.i.d. random variables \(\{m(i); i = 1, 2, \ldots\}\). Assume that \(\mathbb{P}\{m(1) = 0\} = 0\). For \(t_0 > 0\), let
\[
\mathcal{M} = \bigcap_{r=1}^{\infty} \{\|M(t) - M(t-)\|_{n'} \leq 1\},
\]
where \(\{n'\}\) is a sequence of real numbers with \(n' = O(|N'|)\). Then, \(\mathbb{P}(\mathcal{M}) = 1\).

**Proof.** Fix \(r > 0\) and \(t_0 > 0\). Then, \(\mathbb{P}\{M(n't) < n'\} = 1\). Let \(\mathcal{U} = \bigcup_{i=1}^{\infty} \{m(i) > 0\}\). By the assumption of the lemma \(\mathbb{P}(\mathcal{U}) = 1\). Define for \(k = 1, 2, \ldots\),
\[
\mathcal{M}'_k = \{M(n't_0) < k\} \quad \text{and} \quad \mathcal{M}' = \{\|M(t) - M(t-)\|_{n't_0} \leq 1\}.
\]
Observe that
\[
\bigcup_{k=1}^{\infty} (\mathcal{M}'_k \cap \mathcal{U}) \subset \mathcal{M}'.
\]
Hence
\[
\mathbb{P}(\mathcal{M}') \geq \mathbb{P}\left(\bigcup_{k=1}^{\infty} (\mathcal{M}'_k \cap \mathcal{U})\right) = 1.
\]
Since \(\mathcal{M} = \bigcap_{n=1}^{\infty} \mathcal{M}'\), \(\mathbb{P}(\mathcal{M}) = 1\). \(\square\)

**Lemma D.2.** Let \(M\) be a Poisson process with rate \(\gamma > 0\), \(\{n'\}\) be a sequence of nonnegative real numbers such that \(n' = O(|N'|)\), and \(\alpha > 0\). Then, there exists \(B_1 < \infty\) such that for every \(0 < t_0 < \infty\),
\[
\limsup_{r \to \infty} \left[\exp \left\{\alpha \sup_{0 \leq t \leq n't_0}\frac{|M(t) - \gamma t|}{\sqrt{|N'|t_0}}\right\}\right] < B_1.
\]
(D.1)

**Remark D.1.** If \(M\) is a Poisson process with rate 1, then \(M'(\cdot) = M(\gamma \cdot)\) is a Poisson process with rate \(\gamma\), hence (D.1) also holds for the process \(M(\gamma \cdot)\).

**Proof.** Fix \(\alpha > 0\) and \(t_0 > 0\). Since \(n' = O(|N'|)\), \(n'/|N'| < a\) for \(r\) large enough and for some \(a > 0\).
As in the proof of Lemma 1 in Gamarnik and Zeevi [18],
\[
\limsup_{n \to \infty} \mathbb{E} \left[ \sup_{0 \leq t \leq n} \exp \left\{ 2\alpha \sqrt{a} n^{-1/2} |M(t) - \gamma t| \right\} \right] < B_2
\]
for some \( B_2 < \infty \). We have
\[
\mathbb{E} \left[ \exp \left\{ \alpha \sup_{0 \leq t \leq t_0} \frac{|M(t) - \gamma t|}{\sqrt{|N'|}} \right\} \right] \leq \mathbb{E} \left[ \exp \left\{ \alpha \sqrt{a} \sup_{0 \leq t \leq |N'|t_0} \frac{|M(t) - \gamma t|}{\sqrt{|N'|}} \right\} \right]
\leq \mathbb{E} \left[ \exp \left\{ 2\alpha \sqrt{a} \sup_{0 \leq t \leq |a|N'|t_0} \frac{|M(t) - \gamma t|}{\sqrt{|N'|}} \right\} \right]
= \mathbb{E} \left[ \sup_{0 \leq t \leq |a|N'|t_0} \exp \left\{ 2\alpha \sqrt{a} \frac{|M(t) - \gamma t|}{\sqrt{|a|N'|t_0}} \right\} \right].
\]

This with (D.2) gives (D.1).

**Lemma D.3.** Let \( \zeta' \) be defined as in (40). For every \( t_0 > 0 \) and \( \alpha > 0 \), there exists \( r_0 > 0 \) such that for \( r > r_0 \),
\[
\mathbb{E} \left[ \exp \left\{ 4(\mu_{\max} \vee 1)t_0 \sqrt{\frac{\zeta'(t_0)}{|N'|}} \right\} \right] < (J + 1) + 2\alpha.
\]

**Proof.** For notational simplicity, we assume that \( \mu_{\max} > 1 \). Choose \( r \) large enough, so that \( N_{\min}^r \geq \theta \sqrt{N} > 2|N'|B_1 \) for some \( B_1 > 0 \). Then, for such \( r \),
\[
\mathbb{E} \left[ \exp \left\{ 4\mu_{\max} t_0 \sqrt{\frac{\zeta'(t_0)}{|N'|}} \right\} \right] \leq \mathbb{E} \left[ \exp \left\{ 8\mu_{\max} t_0 \sup_{0 \leq s_1 \leq s_2 \leq 0} \left\{ -\sqrt{|N'|}B_3 (s_2 - s_1) + \frac{|\tilde{A}'(s_2) - \tilde{A}'(s_1)|}{\sqrt{|N'|}} \right\} \right\} \right]
+ \mathbb{E} \left[ \exp \left\{ 8\mu_{\max} t_0 \sup_{0 \leq s_1 \leq s_2 \leq 0, v_1, \ldots, v_9: v_1^2 + \ldots + v_9^2 \leq s_2 - s_1} \left\{ -\sqrt{|N'|}B_3 (s_2 - s_1)ight. \right.
+ \left. \sum_{j=1}^9 \left| \tilde{s}_j((v_j + (s_2 - s_1)|N'|) - \tilde{s}_j(v_j|N'|) \right| \right\} \right\} \right].
\]

We show that the first term on the RHS above is bounded by \( 1 + \alpha \), it can similarly be shown that the second term is bounded by \( J + \alpha \). First, observe that for any \( \epsilon > 0 \),
\[
\mathbb{E} \left[ \exp \left\{ 8\mu_{\max} t_0 \sup_{0 \leq s_1 \leq s_2 \leq 0} \left\{ -\sqrt{|N'|}B_3 (s_2 - s_1) + \frac{|\tilde{A}'(s_2) - \tilde{A}'(s_1)|}{\sqrt{|N'|}} \right\} \right\} \right] \leq \mathbb{E} \left[ \exp \left\{ 8\mu_{\max} t_0 \left( -\sqrt{|N'|}B_3 + \sup_{0 \leq s_1 \leq s_2 \leq 0} \left\{ |\tilde{A}'(s_2) - \tilde{A}'(s_1)| \right\} \right) \right\} \right]
+ \mathbb{E} \left[ \exp \left\{ 8\mu_{\max} t_0 \left( \sup_{0 \leq s_1 \leq s_2 \leq 0, |s_1 - s_2| < \epsilon} \left\{ |\tilde{A}'(s_2) - \tilde{A}'(s_1)| \right\} \right) \right\} \right].
\]

We next show that we can choose \( \epsilon > 0 \), so that the terms on the RHS of (D.4) are bounded by \( 1 + \alpha \). Let \( \tilde{A}'(t) = \tilde{A}'(t)/\sqrt{|N'|} \equiv (A'(t) - \lambda^2 t)/\sqrt{|N'|} \). Then, \( A' \Rightarrow W_a \), as \( r \to \infty \), where \( W_a \) is a Brownian motion with variance \( \lambda \). By the continuous mapping theorem,
\[
\mathbb{P} \left\{ \sup_{0 \leq s_1 \leq s_2 \leq 0, |s_1 - s_2| < \epsilon} \frac{|\tilde{A}'(s_2) - \tilde{A}'(s_1)|}{\sqrt{|N'|}} > \frac{\log u}{8\mu_{\max} t_0} \right\} \to \mathbb{P} \left\{ \sup_{0 \leq s_1 \leq s_2 \leq 0, |s_1 - s_2| < \epsilon} |W_a(s_2) - W_a(s_1)| > \frac{\log u}{8\mu_{\max} t_0} \right\}.
\]
Hence, by virtue of the dominated convergence theorem, we have that

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \left\{ \frac{\hat{A}'(s_2) - \hat{A}'(s_1)}{\sqrt{|N'|}} \right\} \right) \right\} \right] \\
= \mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \{|W_a(s_2) - W_a(s_1)|\} \right) \right\} \right] \quad \text{(D.5)}
\]

as \( r \to \infty \). By a.s. continuity of a Brownian motion,

\[
\exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \{|W_a(s_2) - W_a(s_1)|\} \right) \right\} \to 1
\]
a.s. as \( \varepsilon \to 0 \). Since for every \( \varepsilon > 0 \),

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \{|W_a(s_2) - W_a(s_1)|\} \right) \right\} \right] \\
\leq \mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \{|W_a(s_2) - W_a(s_1)|\} \right) \right\} \right] < \infty.
\]

Another application of the dominated convergence theorem yields that

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \{|W_a(s_2) - W_a(s_1)|\} \right) \right\} \right] \\
\to 1
\]
as \( \varepsilon \to 0 \). Thus, for every \( \alpha > 0 \), we can find \( \varepsilon > 0 \) such that

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \{|W_a(s_2) - W_a(s_1)|\} \right) \right\} \right] < 1 + \alpha/4.
\]

Hence, from (D.5), we can find \( r_0 > 0 \) such that for all \( r > r_0 \),

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \left( \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0 \atop |\xi_1 - \xi_2| < \varepsilon} \left\{ \frac{\hat{A}'(s_2) - \hat{A}'(s_1)}{\sqrt{|N'|}} \right\} \right) \right\} \right] < 1 + \alpha/3. \quad \text{(D.6)}
\]

Next, observe that

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0} \left\{ \frac{\hat{A}'(s_2) - \hat{A}'(s_1)}{\sqrt{|N'|}} \right\} \right\} \right] \\
\leq \mathbb{E}\left[ \exp\left\{ 16\mu_{\max}t_0 \sup_{0 \leq \xi_0} \left\{ \frac{\hat{A}'(t) - \hat{A}'(s)}{\sqrt{|N'|}} \right\} \right\} \right]. \quad \text{(D.7)}
\]

Note that, for large enough \( r \), the term on the RHS of (D.7) is bounded by Lemma D.2. Hence, by selecting \( r \) large enough, we can make the first term on the RHS of (D.4) arbitrarily small for any \( \varepsilon > 0 \).

Fix \( \varepsilon > 0 \) and choose \( r_1 \) large enough, so that (D.6) holds for every \( r > r_1 \). Now, for this choice of \( \varepsilon > 0 \), choose \( r_2 \) large enough, so that for \( r > r_2 \), the first term on the RHS of (D.4) is bounded by \( \alpha/4 \). Therefore, for every \( r > r_2 \),

\[
\mathbb{E}\left[ \exp\left\{ 8\mu_{\max}t_0 \sup_{0 \leq \xi_1, \xi_2 \leq \xi_0} \left\{ -\sqrt{|N'|}B(s_2 - s_1) + \frac{\hat{A}'(s_2) - \hat{A}'(s_1)}{\sqrt{|N'|}} \right\} \right\} \right] < 1 + \alpha
\]
by (D.4).
Next, we outline the details how the second term on the RHS of (D.3) is handled. Observe that

$$
\mathbb{E} \left[ \exp \left( 8 \mu_{\text{max}} t_0 \sum_{0 \leq s_1 \leq s_2 \leq 0} \exp \left\{ - \frac{\sqrt{N}}{|N'|} B_3 (s_2 - s_1) + \sum_{j=1} J \left[ \frac{\sqrt{N}}{|N'|} (s_j - s_1) \right] \right\} \right] \right] \leq \sum_{j=1} J \mathbb{E} \left[ \exp \left( 8 \mu_{\text{max}} t_0 \sum_{0 \leq s_1 \leq s_2 \leq 0} \exp \left\{ - \frac{\sqrt{N}}{|N'|} B_3 (s_2 - s_1) + \frac{\sqrt{N}}{|N'|} (s_j - s_1) \left[ \frac{\sqrt{N}}{|N'|} (s_j - s_1) \right] \right\} \right] \right].
$$

It can be shown as above that for large enough \( r \),

$$
\mathbb{E} \left[ \exp \left( 8 \mu_{\text{max}} t_0 \sum_{0 \leq s_1 \leq s_2 \leq 0} \exp \left\{ - \frac{\sqrt{N}}{|N'|} B_3 (s_2 - s_1) + \frac{\sqrt{N}}{|N'|} (s_j - s_1) \left[ \frac{\sqrt{N}}{|N'|} (s_j - s_1) \right] \right\} \right] < 1 + \alpha / J
$$

for each \( j \in J \). □

Acknowledgments. The research presented in this paper forms part of the author’s doctoral dissertation whose research is supported in part by the National Science Foundation grant DMI-0300599 and an IBM Faculty Award to Dr. Jim Dai. The author thanks Dr. Dai for his guidance, insight, and help during this project. His extensive assistance made this paper better structured and more readable. The author also thanks Jim Luedtke for his helpful comments on an earlier draft.

References


