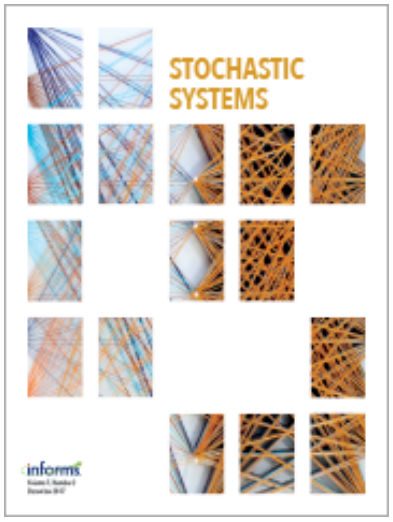


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Join Idle Queue with Service Elasticity: Large-Scale Asymptotics of a Nonmonotone System

Debankur Mukherjee, Alexander Stolyar

To cite this article:
Debankur Mukherjee, Alexander Stolyar (2019) Join Idle Queue with Service Elasticity: Large-Scale Asymptotics of a Nonmonotone System. Stochastic Systems 9(4):338-358. https://doi.org/10.1287/stsy.2019.0030

http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 9, No. 4, December 2019, pp. 338–358
ISSN 1946-5238 (online)

Join Idle Queue with Service Elasticity: Large-Scale Asymptotics of a Nonmonotone System

Debankur Mukherjee,[a] Alexander Stolyar[b]

[a] Georgia Institute of Technology, Atlanta, Georgia 30332; [b] University of Illinois at Urbana–Champaign, Champaign, Illinois 61820
Contact: debankur.mukherjee@isye.gatech.edu, https://orcid.org/0000-0003-1678-4893 (DM); stolyar@illinois.edu, https://orcid.org/0000-0002-1496-9803 (AS)

Received: March 20, 2018
Revised: November 12, 2018; January 16, 2019
Accepted: January 29, 2019
Published Online in Articles in Advance: December 10, 2019

MSC2010 Subject Classification: Primary: 60K25, 60F17; secondary: 68M20

https://doi.org/10.1287/stsy.2019.0030

Abstract. We consider the model of a token-based joint autoscaling and load-balancing strategy, proposed in a recent paper by Mukherjee et al. [Mukherjee D, Dhara S, Borst SC, Van Leeuwaarden JSH (2017) Optimal service elasticity in large-scale distributed systems. *Proc. ACM Measurement Anal. Comput. Systems* 1(1):25:1–25:28.], which offers an efficient scalable implementation and yet achieves asymptotically optimal steady-state delay performance and energy consumption as the number of servers $N \to \infty$. In the aforementioned work, the asymptotic results are obtained *under the assumption that the queues have fixed-size finite buffers*, and therefore, the fundamental question of stability of the proposed scheme with infinite buffers was left open. In this paper, we address this fundamental stability question. The system stability under the usual subcritical load assumption is not automatic. Moreover, the stability may *not* even hold for all N. The key challenge stems from the fact that the process *lacks monotonicity*, which has been the powerful primary tool for establishing stability in load-balancing models. We develop a novel method to prove that the subcritically loaded system is stable for *large enough* N and establish convergence of steady-state distributions to the optimal one as $N \to \infty$. The method goes beyond the state-of-the-art techniques; it uses an induction-based idea and a "weak monotonicity" property of the model. This technique is of independent interest and may have broader applicability.

History: Former designation of this paper was SSY-2018-011.

Funding: D. Mukherjee was supported by The Netherlands Organization for Scientific Research (NWO) through Gravitation Networks [Grant 024.002.003] and TOP-GO [Grant 613.001.012].

Keywords: stability • join idle queue • mean-field limit • fluid limit • load balancing

1. Introduction

1.1. Background and Motivation

Load balancing and autoscaling are two principal pillars in modern-day data centers and cloud networks and, therefore, have gained renewed interest in the past two decades. In its basic setup, a large-scale system consists of a pool of a large number of servers and a single dispatcher to which tasks arrive sequentially. Each task has to be instantaneously assigned to some server or discarded. Load-balancing algorithms primarily concern design and analysis of algorithms to distribute incoming tasks among the servers as evenly as possible while using minimal instantaneous queue-length information. At the same time, a big proportion of the tasks processed by these data centers come with business-critical performance requirements. This forces service providers to increase their capacity at a tremendous rate to cope with the high-demand period in the presence of a time-varying demand pattern. Consequently, the energy consumption by the servers in these huge data centers has risen dramatically and become a dominant factor in managing data center operations and cloud infrastructure platforms. Auto-scaling provides a popular paradigm for automatically adjusting service capacity in response to demand while meeting performance targets.

1.1.1. Load Balancing in Large Systems. The study of load-balancing schemes in large-scale systems has a very rich history, and for decades, a lot of research has been conducted in understanding the fundamental trade-off between delay performance and communication overhead per task. The so-called "power-of-d" schemes, in which

each arrival is assigned to the shortest among d randomly chosen queues, provide surprising improvements for $d \geq 2$ over purely random routing ($d = 1$) while maintaining the communication overhead as low as d per task. This scheme, along with its many variations, has been studied extensively in Aghajani and Ramanan (2017), Bramson et al. (2012, 2013), Eschenfeldt and Gamarnik (2016), Mitzenmacher (2001), Mukherjee et al. (2018), Vvedenskaya et al. (1996), Ying (2017), and many more. Relatively recently, the join-the-idle-queue (JIQ) scheme was proposed by Lu et al. (2011), with which an arriving task is assigned to an idle server (if any), or in case all servers are busy, it is assigned to some queue uniformly at random. The JIQ scheme has a low-cost, token-based implementation that involves only $O(1)$ communication overhead per task. Large-scale asymptotic results by Stolyar (2015, 2017) show that, under Markovian assumptions, the JIQ policy achieves a zero probability of wait for any fixed subcritical load per server in a regime in which the total number of servers grows large. It should be noted that the results by Stolyar (2015, 2017) even hold for considerably more general scenarios, such as decreasing hazard-rate service-time distributions and heterogeneous server pools. Recently, it has been further shown by Foss and Stolyar (2017) that, when the average load per server $\lambda < 1/2$, the large-scale asymptotic optimality of JIQ is preserved even under completely general service-time distributions. Results by Mukherjee et al. (2016) indicate that, under Markovian assumptions, the JIQ policy has the same diffusion limit as the join-the-shortest-queue strategy and, thus, achieves diffusion-level optimality. These results show that the JIQ policy provides asymptotically optimal delay performance in large-scale systems while only involving minimal communication overhead (at most one message per task on average). We refer to Van der Boor et al. (2018) for a recent survey on load-balancing schemes.

1.1.2. Autoscaling with a Centralized Queue. Queue-driven autoscaling techniques have been widely investigated in the literature: Andrew et al. (2010), Urgaonkar et al. (2010), Lin et al. (2012, 2013), Liu et al. (2011a, b, 2012), Wierman et al. (2012), Gandhi et al. (2014), and Pender and Phung-Duc (2016). In systems with a centralized queue, it is very common to put servers to "sleep" while the demand is low because servers in sleep mode consume much less energy than active servers. Under Markovian assumptions, the behavior of these mechanisms can be described in terms of various incarnations of M/M/N queues with setup times. There are several further recent papers that examine on-demand server addition/removal in a somewhat different vein; see Nguyen and Stolyar (2016) and Pang and Stolyar (2016). Generalizations toward non-stationary arrivals and impatience effects have also been considered recently by Pender and Phung-Duc (2016). Unfortunately, data centers and cloud networks with millions of servers are too complex to maintain any centralized queue, and it involves a prohibitively high communication burden to obtain instantaneous system information even for a small fraction of servers.

1.1.3. Joint Load Balancing and Autoscaling in Distributed Systems. Motivated by all these, a token-based, joint load-balancing and autoscaling scheme called token-based auto balance scaling (TABS) was proposed by Mukherjee et al. (2017), and it offers an efficient scalable implementation and yet achieves asymptotically optimal steady-state delay performance and energy consumption as the number of servers $N \to \infty$. Under the TABS scheme, once a server becomes idle, it spends an Exp(μ) time (standby period) before turning off. Each incoming task is forwarded to an idle "on" server if such exists; otherwise, the task goes to a randomly selected "on" (and busy) server and the turning on of one "off" server is triggered; that is, its setup period is started. The setup period takes an Exp(ν) time, after which the server becomes available (idle on). A more detailed description of the TABS scheme is given in Section 2. In Mukherjee et al. (2017), the authors left open a fundamental question: Is the system with a given number N of servers stable under the TABS scheme? The analysis in Mukherjee et al. (2017) bypasses the issue of stability by assuming that each server in the system has a finite buffer capacity. Thus, it remains an important open challenge to understand the stability property of the TABS scheme without the finite-buffer restriction.

1.2. Key Contributions and Our Approach

In this paper, we address the stability issue for systems under the TABS scheme without the assumption of finite buffers and examine the asymptotic behavior of the system as N becomes large. Analyzing the stability of the TABS scheme in the infinite buffer scenario poses a significant challenge because the stability of the finite-N system, that is, the system with a finite number N of servers under the usual subcritical load assumption is not automatic. In fact, as we further discuss in Remark 1 in detail, even under subcritical load, the system may *not* be stable for all N. Our first main result is that, for any fixed subcritical load, the system is

stable for *large enough* N. Further, using this large-N stability result in combination with mean-field analysis, we establish convergence of the sequence of steady-state distributions as $N \to \infty$.

The key challenge in showing large-N stability for systems under the TABS scheme stems from the fact that the occupancy state process lacks monotonicity. It is well-known that monotonicity is a powerful primary tool for establishing stability of load-balancing models; see Bramson et al. (2012), Stolyar (2015, 2017), and Vvedenskaya et al. (1996). In fact, process monotonicity is used extensively not only for stability analysis and not only in queueing literature; for example, many interacting-particle systems' results rely crucially on monotonicity, for example, those for spin systems and exclusion processes; see Liggett (1985). The lack of monotonicity immediately complicates the situation as, for example, in Foss and Stolyar (2017) and Shneer and Stolyar (2018). Specifically, when the service-time distribution is general, it is the lack of monotonicity that has left open the stability questions for the power-of-d scheme when system load $\lambda > 1/4$ (see Bramson et al. 2012) and the JIQ scheme when $\lambda > 1/2$ (see Foss and Stolyar 2017). We develop a novel method for proving *large-N stability* for subcritically loaded systems, and using that, we establish the convergence of the sequence of steady-state distributions as $N \to \infty$. Our method uses an induction-based idea and relies on a "weak monotonicity" property of the model as further detailed herein. To the best of our knowledge, this is the first time both the *traditional fluid limit* (in the sense of large starting state) and the *mean-field fluid limit* (when the number of servers grows large) are used in an intricate manner to obtain large-N stability results.

To establish the large-N stability, we actually prove a stronger statement. We consider an artificial system, in which some of the queues are infinite at all times. Then, loosely speaking, we prove that the following holds for all sufficiently large N: *If the system with N servers contains k servers with infinite queue lengths, $0 \le k \le N$, then* (i) *the subsystem consisting of the remaining (i.e., finite) queues is stable, and* (ii) *when this subsystem is at steady state, the average rate at which tasks join the infinite queues is strictly smaller than that at which tasks depart from them.* Note that the case $k = 0$ corresponds to the desired stability result.

The use of backward induction in k facilitates proving this statement. For a fixed N, first, we introduce the notion of a fluid sample path (FSP) for systems in which some queues might be infinite. The base case of the backward induction is when $k = N$, and assuming the statement for k, we show that it holds for $k - 1$. We use the classical fluid-stability argument (as in Rybko and Stolyar 1992, Dai 1995, and Stolyar 1995) to establish stability for the system in which the number of infinite queues is $k - 1$. As mentioned, here the notion of the traditional FSP is needed to be suitably extended to fit to the systems in which some servers have infinite queue lengths. Loosely speaking, for the fluid stability, the "large queues" behave as "infinite queues" for which the induction statement provides us with the drift estimates. Also, to calculate the drift of a queue in the fluid limit for fixed but large enough N, we use the mean-field analysis. A more detailed heuristic road map of this proof argument is presented in Section 4.1. This technique is of independent interest and potentially has a much broader applicability in proving large-N stability for nonmonotone systems, in which the state-of-the-art results have remained scarce so far.

1.3. Organization of the Paper

The rest of the paper is organized as follows. In Section 2, we present a detailed model description, state the main results, and provide their ramifications along with discussions of several proof heuristics. The full proof of the main results are deferred until Section 3. Section 4 introduces an inductive approach to prove the large-N stability result. We present the proof of the large-scale system (when $N \to \infty$) using mean-field analysis in Section 5. Finally, we make a few brief concluding remarks in Section 6.

2. Model Description and Main Result

In this section, first we describe the system and the TABS scheme in detail and then state the main results and discuss their ramifications.

Consider a system of N parallel queues with identical servers and a single dispatcher. Tasks with unit-mean exponentially distributed service requirements arrive as a Poisson process of rate λN with $\lambda < 1$. Incoming tasks cannot be queued at the dispatcher and must immediately and irrevocably be forwarded to one of the servers at which they can be queued. Each server has an infinite buffer capacity. The service discipline at each server is oblivious to the actual service requirements (e.g., first-come, first-served). A turned-off server takes an exponentially distributed time with mean $1/\nu$ (to be henceforth denoted as $\text{Exp}(\nu)$) time (setup period) to be turned on. We now describe the token-based, joint autoscaling and load-balancing scheme called TABS as introduced by Mukherjee et al. (2017).

2.1. TABS Scheme by Mukherjee et al. (2017)

• When a server becomes idle, it sends a "green" message to the dispatcher, waits for an Exp(μ) time (standby period), and turns itself off by sending a "red" message to the dispatcher (the corresponding green message is destroyed).

• When a task arrives, the dispatcher selects a green message at random if there are any and assigns the task to the corresponding server (the corresponding green message is replaced by a "yellow" message). Otherwise, the task is assigned to an arbitrary busy server (and is lost if there is none), and if, at that arrival epoch, there is a red message at the dispatcher, then it selects one at random, and the setup procedure of the corresponding server is initiated, replacing its red message by an "orange" message. Setup procedure takes Exp(ν) time after which the server becomes active.

• Any server that activates because of the latter event sends a green message to the dispatcher (the corresponding orange message is replaced), waits for an Exp(μ) time for a possible assignment of a task, and again turns itself off by sending a red message to the dispatcher.

As described in Mukherjee et al. (2017), the TABS scheme gives rise to a distributed operation in which servers are in one of four states (busy, idle-on, idle-off, or setup) and advertise their state to the dispatcher via exchange of tokens. Figure 1 illustrates this token-based exchange protocol. Note that setup procedures are never aborted and continued even when idle-on servers do become available.

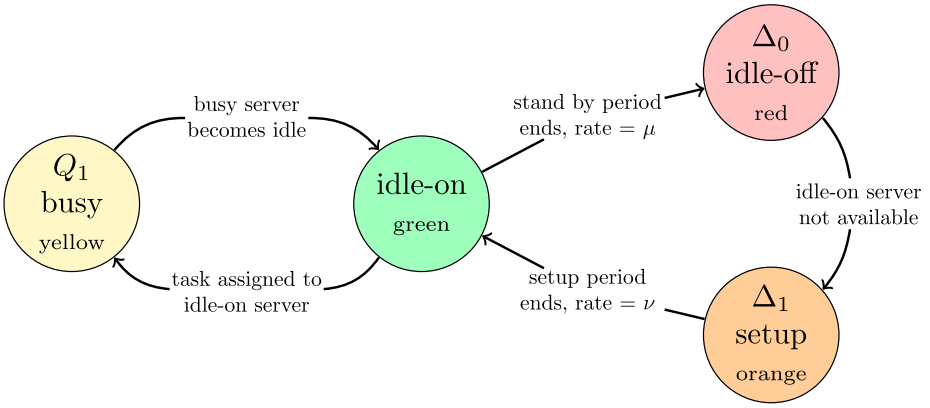
2.2. Notation

For the system with N servers, let $X_j^N(t)$ denote the queue length of server j at time t, $j = 1, 2, \ldots, N$, and $\mathbf{Q}^N(t) := (Q_1^N(t), Q_2^N(t), \ldots)$ denote the system occupancy state, where $Q_i^N(t)$ is the number of servers with queue length greater than or equal to i at time t, including the possible task in service, $i = 1, 2, \ldots$. Also, let $\Delta_0^N(t)$ and $\Delta_1^N(t)$ denote the number of idle-off servers and servers in setup mode at time t, respectively. Note that the process $(\mathbf{Q}^N(t), \Delta_0^N(t), \Delta_1^N(t))_{t\geq 0}$ provides a Markovian state description by virtue of the exchangeablity of the servers. It is easy to see that, for any fixed N, this process is an irreducible countable-state Markov chain. Therefore, its positive recurrence, which we refer to as *stability*, is equivalent to ergodicity and to the existence of unique stationary distribution. Further, let $U^N(t)$ denote the number of idle-on servers at time t. We focus upon an asymptotic analysis, in which the task-arrival rate and the number of servers grow large in proportion. The *mean-field fluid-scaled* quantities are denoted by the respective small letters, such as $q_i^N(t) := Q_i^N(t)/N$, $\delta_0^N(t) = \Delta_0^N(t)/N$, $\delta_1^N(t) = \Delta_1^N(t)/N$, and $u^N(t) := U^N(t)/N$. Notation for the *conventional fluid-scaled* occupancy states for a fixed N are introduced later in Section 3.1. For brevity in notation, we write $\mathbf{q}^N(t) = (q_1^N(t), q_2^N(t), \ldots)$ and $\boldsymbol{\delta}^N(t) = (\delta_0^N(t), \delta_1^N(t))$. Let

$$E = \left\{(\boldsymbol{q}, \boldsymbol{\delta}) \in [0,1]^\infty : q_i \geq q_{i+1},\ \forall i,\ \delta_0 + \delta_1 + q_1 \leq 1\right\},$$

denote the space of all mean-field fluid-scaled occupancy states so that $(\mathbf{q}^N(t), \boldsymbol{\delta}^N(t))$ takes value in E for all t. Endow E with the product topology, and the Borel σ-algebra $\mathscr{E}$, generated by the open sets of E. For any complete separable metric space E, denote by $D_E[0,\infty)$, the set of all E-valued *càdlàg* (right continuous with left limit exists) processes.

Figure 1. Illustration of Server On-Off Decision Rules in the TABS Scheme Along with Message Colors and State Variables as Given in Mukherjee et al. (2017)



Also, throughout the paper, by the symbol "$\xrightarrow{\mathbb{P}}$" we denote convergence in probability for real-valued random variables, and the abbreviation "u.o.c." stands for "*uniformly on compact sets*."

We now present our first main result, which states that, for any fixed choice of the parameters, a subcritically loaded system under the TABS scheme is stable for large enough N.

Theorem 1. *For any fixed $\mu, \nu > 0$, and $\lambda < 1$, the system with N servers under the TABS scheme is stable (positive recurrent) for large enough N.*

Theorem 1 is proved in Section 3.

Remark 1. It is worthwhile to mention that the "large-N" stability as stated in Theorem 1 is the best one can hope for. In fact, for fixed N and λ, there are values of the parameters μ and ν such that the system under the TABS scheme may not be stable. To elaborate further on this point, consider a system with two servers A and B, and $1/2 < \lambda < 1$. Let server A start with a large queue, and the initial queue length at server B be small. In that case, observe that, every time the queue length at server B hits zero with positive probability, it turns idle-off before the next arrival epoch. Once server B is idle-off, the arrival rate into server A becomes $2\lambda > 1$. Thus, before server B turns idle-on again, the expected number of tasks that join server A is given by at least $2\lambda/\nu$, and the expected number of departures is $1/\nu$. Thus, the queue length at server A increases by $(2\lambda - 1)/\nu$, which can be very large if ν is small. Further note that, once server B becomes busy again, both servers receive an arrival rate $\lambda < 1$, and hence, it is more likely that server B will empty out again, repeating the previous scenario. The situation becomes better as N increases. Indeed, for large N, if "too many" servers are idle-off and too many tasks do not find an idle queue to join, the system starts producing servers in setup mode fast enough, and as a result, more and more servers start becoming busy. This heuristic has been illustrated in Figure 2 with examples of three scenarios with small, moderate, and large values of N, respectively.

Remark 2. As Remark 1 demonstrates, the underlying reason for the possible instability of a system with a *fixed finite* number of servers is as follows. The set of all queues may divide into the subsets of "large" and "small" queues. The small queues are in "steady state," and it is such that, with nonzero probability, some or all of those queues are in the idle-off or setup states, which increases the instantaneous arrival rate into the large queues. As a result, the average arrival rate into each large queue may exceed one, giving it a positive drift. The reason why this scenario becomes impossible for all sufficiently large N is as follows. If we consider the subsets of large and small queues, then, when N is large, the steady state of the small queues concentrates so that, with high probability, either almost all servers are busy (in which case the average arrival rate into each large queue is at most λ) or idle-on servers are almost always present (in which case the average arrival rate into each large queue is close to zero). Our large-N asymptotic results are used to show the concentration of the steady state of small queues. We leave the study of the minimum value of N required for stability (with $\lambda < 1$ being fixed) as an interesting open direction.

In the next theorem, we identify the limit of the sequence of stationary distributions of the occupancy processes as $N \to \infty$. In particular, we establish that, under subcritical load, for any fixed $\mu, \nu > 0$, the steady-state occupancy process converges weakly to the unique fixed point. (For the finite buffer scenario, this was proved in Mukherjee et al. 2017, proposition 3.3). Denote by $\boldsymbol{q}^N(\infty)$ and $\boldsymbol{\delta}^N(\infty)$ the random values of $\boldsymbol{q}^N(t)$ and $\boldsymbol{\delta}^N(t)$ in the steady state, respectively.

Theorem 2. *For any fixed $\mu, \nu > 0$, and $\lambda < 1$, the sequence of steady states $(\boldsymbol{q}^N(\infty), \boldsymbol{\delta}^N(\infty))$ converges weakly to the fixed point $(\boldsymbol{q}^\star, \boldsymbol{\delta}^\star)$ as $N \to \infty$, where*

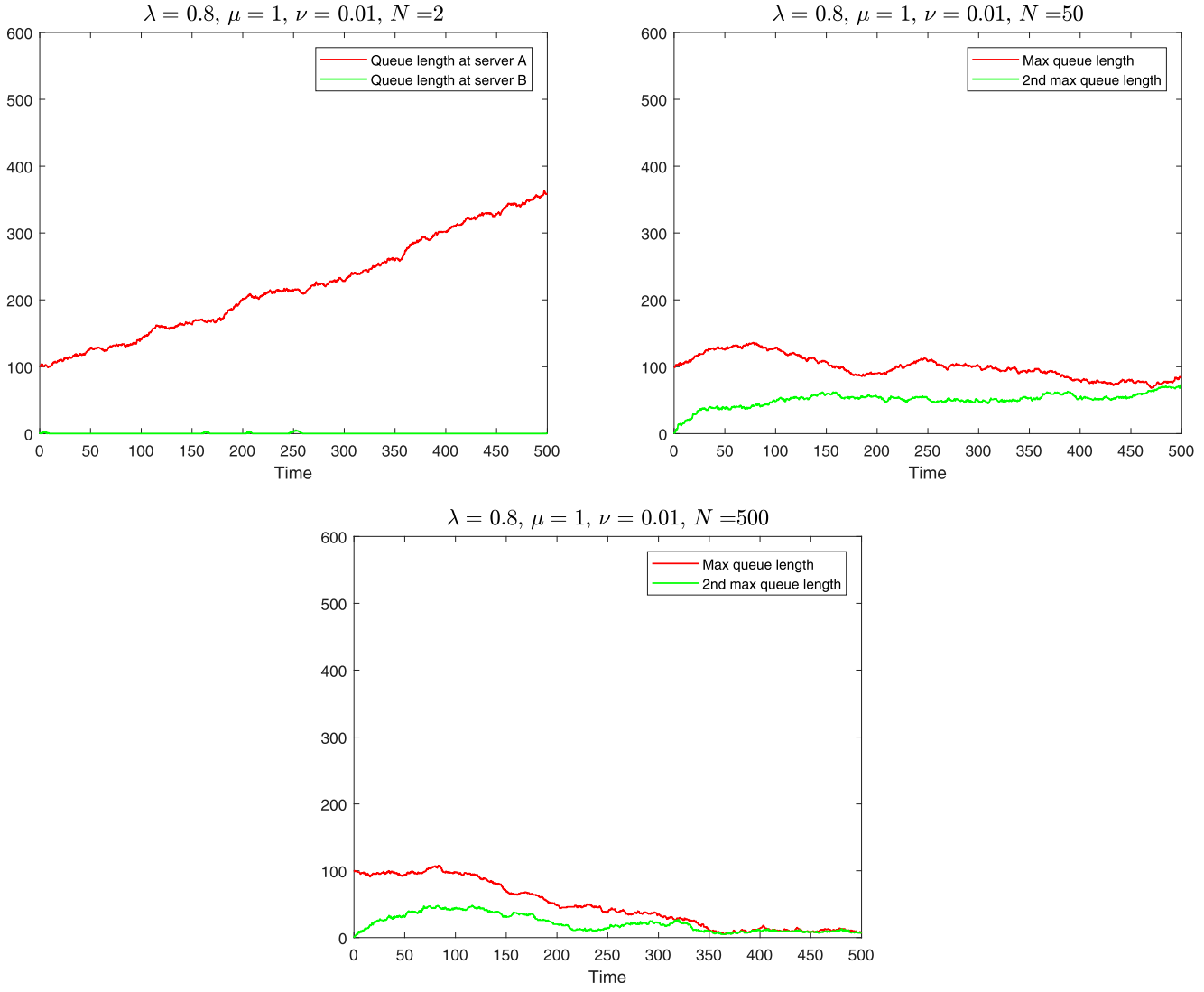
$$\delta_0^\star = 1 - \lambda \quad \delta_1^\star = 0 \quad q_1^\star = \lambda, \quad q_i^\star = 0 \qquad \text{for all} \quad i \geq 2.$$

Theorem 2 has the following corollary.

Corollary 1. *The steady-state probability of an arriving task being immediately assigned to an idle-on server (i.e., its waiting time being zero) converges to one as $N \to \infty$.*

Indeed, the form of the fixed point $(\boldsymbol{q}^\star, \boldsymbol{\delta}^\star)$ is such that, asymptotically, the steady-state rate at which idle-on servers are "generated" by the task departures is λN. It also shows that the steady-state mean time for an idle-on server to stay in that state converges to zero. (This is by Little's law because the fraction of idle-on servers vanishes.) This, in turn, implies that, asymptotically, the steady-state average rate at which idle-on servers are "taken" by new arrivals is λN. This means that, asymptotically, the fraction of new arrivals that immediately go to idle-on servers is one; which is what Corollary 1 asserts. Note further that, because the task arrival

Figure 2. Stability and Instability of Systems under the TABS Scheme



Notes. (Top left) Illustration of instability of the TABS scheme for $N = 2$ via sample paths of the queue length process. (Top right) Sample paths of the maximum and second maximum queue length processes in an intermediate system ($N = 50$) for the same parameter choices. (Bottom). The system becomes stable for a large enough system ($N = 500$).

process is Poisson and PASTA property, Corollary 1 is equivalent to the property that the steady-state probability of *not* having an idle-on server in the system vanishes:

$$\lim_{N\to\infty} \mathbb{P}\big(Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N\big) = 0. \tag{1}$$

(Theorem 2 easily generalizes to far more general arrival processes than Poisson. For more general arrival processes, Corollary 1 still holds, but (1) is not immediate because PASTA property no longer applies.)

Theorem 2 and Corollary 1 show that the TABS scheme is asymptotically optimal as $N \to \infty$ in that (a) a task waiting time vanishes and (b) the fraction of active servers becomes minimum possible.

3. Proofs of the Main Results

In Section 3.1, we introduce the notion of conventional fluid scaling (when the number of servers is fixed) and FSPs, and state Proposition 1 that implies Theorem 1 as an immediate corollary. Section 3.2 contains two key results for sequence of systems with increasing system size, that is, number of servers $N \to \infty$, and proves Theorem 2.

3.1. Conventional Fluid Limit for a System with Fixed N

In this section, first, we introduce a notion of FSP for finite-N systems in which some of the queue lengths are infinite. We emphasize that this is a *conventional fluid limit* in the sense that the number of servers is fixed, but the time and the queue length at each server are scaled by some parameter that goes to infinity.

Loosely speaking, conventional fluid limits are usually defined as follows: For a fixed N, consider a sequence of systems with increasing initial norm (total queue length) R, say. Now scale the queue length process at each server and the time by R. Then any weak limit of this sequence of (space and time) scaled processes is called an FSP. Observe that this definition is inherently not fit if the system has some servers whose initial queue length is infinity. Thus, we introduce a suitable notion of FSP that does not require the scaled norm of the initial state to be one. We now introduce a rigorous notion of FSP for systems with some of the queues being infinite.

3.1.1. Fluid Limit of a System with Some of the Queues Being Infinite.

Consider a system of N servers with indices in $\mathcal{N}$, among which k servers with indices in $\mathcal{H} \subseteq \mathcal{N}$ have infinite queue lengths. Now consider any sequence of systems indexed by R such that $\sum_{i\in\mathcal{N}\setminus\mathcal{H}} X_i^{N,R}(0) < \infty$, and

$$x_i^{N,R}(t) := \frac{X_i^{N,R}(Rt)}{R}, \quad i \in \mathcal{N} \setminus \mathcal{H} \tag{2}$$

be the corresponding scaled processes. For fixed N, the scaling in (2) is henceforth called the *conventional fluid-scaled* queue-length process. Also, for the Rth system, let $A_i^{N,R}(t)$ and $D_i^{N,R}(t)$ denote the cumulative number of arrivals to and departures from server i with $a_i^{N,R}(t) := A_i^{N,R}(Rt)/R$ and $d_i^{N,R}(t) := D_i^{N,R}(Rt)/R$ being the corresponding fluid-scaled processes, $i \in \mathcal{N}$. We often omit the superscript N when it is fixed from the context.

Now for any fixed N, suppose the (conventional fluid-scaled) initial states converge, that is, $x^R(0) \to x(0)$ for some fixed $x(0)$ such that $0 \le \sum_{i\in\mathcal{N}\setminus\mathcal{H}} x_i(0) < \infty$ and $x_i(0) = \infty$ for $i \in \mathcal{H}$. Then a set of uniformly Lipschitz continuous functions $(x_i(t), a_i(t), d_i(t))_{i\in\mathcal{N}}$ on the time interval $[0, T]$ (where T is possibly infinite) with the convention $x_i(\cdot) \equiv \infty$ for all $i \in \mathcal{H}$, is called an FSP starting from $\mathbf{x}(0)$ if, for any subsequence of $\{R\}$ there exists a further subsequence (which we still denote by $\{R\}$) such that, with probability one, along that subsequence, the following convergences hold:

i. For all $i \in \mathcal{N}$, $a_i^R(\cdot) \to a_i(\cdot)$ and $d_i^R(\cdot) \to d_i(\cdot)$, u.o.c.

ii. For $i \in \mathcal{N} \setminus \mathcal{H}$, $x_i^R(\cdot) \to x_i(\cdot)$ u.o.c.

Note that this definition is equivalent to convergence in probability to the unique FSP. For any FSP, almost all points (with respect to Lebesgue measure) are *regular*; that is, for all $i \in \mathcal{N} \setminus \mathcal{H}$, $x_i(t)$ has proper left and right derivatives with respect to t, and for all such regular points,

$$x_i'(t) = a_i'(t) - d_i'(t).$$

3.1.2. Infinite Queues as Part of an FSP.

The arrival and departure functions $a_i(t)$ and $d_i(t)$ are well defined for each queue, including infinite queues. Of course, the derivative $x_i'(t)$ for an infinite queue makes no direct sense (because an infinite queue remains infinite at all times). However, we adopt a convention that $x_i'(t) = a_i'(t) - d_i'(t)$ for all queues, including the infinite ones. For an FSP, $x_i'(t)$ is sometimes referred to as a "drift" of (finite or infinite) queue i at time t.

We are now in a position to state the key result that establishes the large-N stability of the TABS scheme.

Proposition 1. *The following holds for all sufficiently large N. For each $0 \le k \le N$, consider a system in which k servers with indices in $\mathcal{H}$ have infinite queues, and the remaining $N - k$ queues are finite. Then, for each $j = 1, 2, \ldots, N$, there exists $\varepsilon(j) > 0$ such that the following properties hold ($\varepsilon(j)$ and other constants specified herein also depend on N).*

1. *For any $\mathbf{x}(0)$ such that $0 \le \sum_{i\in\mathcal{N}\setminus\mathcal{H}} x_i(0) < \infty$ and $x_i(0) = \infty$ for $i \in \mathcal{H}$, there exists $T(k, \mathbf{x}(0)) < \infty$ and a unique FSP on the interval $[0, T(k, \mathbf{x}(0))]$, for which the following properties hold.*
   i. *If, at a regular point t, $\mathcal{M}(t) := \{i \in \mathcal{N} : x_i(t) > 0\}$ with $|\mathcal{M}(t)| = m > k$, then $x_i'(t) = -\varepsilon(m)$ for all $i \in \mathcal{M}(t)$.*
   ii. *For any $i \in \mathcal{N} \setminus \mathcal{H}$, if $x_i(t_0) = 0$ for some $0 \le t_0 \le T(k, \mathbf{x}(0))$, then $x_i(t) = 0$ for all $t_0 \le t \le T(k, \mathbf{x}(0))$.*
   iii. *$T(k, \mathbf{x}(0)) = \inf\{t : x_i(t) = 0$ for all $i \in \mathcal{N} \setminus \mathcal{H}\}$.*
2. *The subsystem with $N - k$ finite queues is stable.*
3. *When the subsystem with $N - k$ finite queues is in steady state, the average arrival rate into each of the k servers having infinite queue lengths is equal to $1 - \varepsilon(k)$.*

4. *For any* $x(0)$ *such that* $0 \le \sum_{i \in \mathcal{N}\setminus\mathcal{H}} x_i(0) < \infty$ *and* $x_i(0) = \infty$ *for* $i \in \mathcal{H}$, *there exists a unique FSP on the entire interval* $[0,\infty)$. *In* $[0, T(k, x(0))]$, *it is as described in part 1. Starting from* $T(k, x(0))$, *all queues in* $\mathcal{N} \setminus \mathcal{H}$ *stay at zero, and all infinite queues have the drift equal to* $-\varepsilon(k)$.

Although part 2 follows from part 1 and part 4 is stronger than part 1, the statement of Proposition 1 is arranged as it is to facilitate its proof as we see in Section 4 in detail.

Proof of Theorem 1. Note that Theorem 1 is a special case of Proposition 1 when $k = 0$.

3.2. Large-Scale Asymptotics: Auxiliary Results

In this section, we state two crucial lemmas that describe asymptotic properties of the sequence of systems as the number of servers $N \to \infty$ *if stability is given*. Their proofs involve mean-field fluid scaling and limits.

Lemma 1. *There exist* $\varepsilon_1 > 0$ *and* $C_q = C_q(\varepsilon_1) > 0$ *such that the following holds. Consider any sequence of systems with* $N \to \infty$ *and* $k = k(N)$ *infinite queues such that* $k(N)/N \to \kappa \in [0,1]$ *and assume that each of these systems is stable. Then, for all sufficiently large* N,

$$\mathbb{P}\left(q_1^N(\infty) < \varepsilon_1\right) \le \mathrm{e}^{-C_q N}.$$

Lemma 2. *Consider any sequence of systems with* $N \to \infty$ *and* $k = k(N)$ *infinite queues such that* $k(N)/N \to \kappa \in [0,1]$ *and assume that each of these systems is stable. The following statements hold*:

1. *If* $\kappa \ge 1 - \lambda$, *then* $q_1^N(\infty) \xrightarrow{\mathbb{P}} 1$ *as* $N \to \infty$.
2. *If* $\kappa < 1 - \lambda$, *then the limit of the sequence of stationary occupancy states* $(\boldsymbol{q}^N(\infty), \boldsymbol{\delta}^N(\infty))$ *is the distribution concentrated at the unique equilibrium point* $(\boldsymbol{q}^\star(\kappa), \boldsymbol{\delta}^\star(\kappa))$ *such that*

$$q_1^\star(\kappa) = \kappa + \lambda, \quad q_2^\star(\kappa) = \kappa$$
$$\delta_0^\star(\kappa) = 1 - \lambda - \kappa, \quad \delta_1^\star(\kappa) = 0.$$

Lemmas 1 and 2 are proved in Section 5. These results are used to derive necessary large-N bounds on the expected arrival rate into each of the servers having infinite queue lengths when the system is in steady state.

Remark 3. It is also worthwhile to note that Lemmas 1 and 2 can be thought of as a weak monotonicity property of the TABS scheme as mentioned earlier. Loosely speaking, the weak monotonicity requires that no matter where the system starts, in some fixed time, the system arrives at a state with a certain fraction of busy servers. The purpose of Lemmas 1 and 2 is to bound *under the assumption of stability*, the expected rate at which a task arrives to the infinite queues when the subsystem containing the finite queues is in steady state: In this regard,

i. Lemma 2 guarantees high-probability bounds on the total number of busy servers so that, with probability tending to one as $N \to \infty$, the fraction of busy servers in the whole system is at least λ in steady state.

ii. But note that because the arrival rate is λN, when the system has few busy servers (even with an asymptotically vanishing probability), the arrival rate to the infinite servers can become $\Theta(N)$. Thus, we need the exponential bound stated in Lemma 1 to obtain the bound on the expected rate of arrivals to the infinite queues.

In Section 4.4, we see that, as a consequence of Lemmas 1 and 2, we obtain that, for large enough N, under the assumption of stability, the steady-state rate at which tasks join an infinite queue is strictly less than one, and the drift of the infinite queues as defined in Section 3.1 becomes strictly negative. This fact is used in the proof of Proposition 1.

Proof of Theorem 2. Note that, given the large-N stability property proved in Proposition 1 for $k(N) = 0$ and the convergence of stationary distributions under the assumption of stability in Lemma 2, the proof of Theorem 2 is immediate.

4. Proof of Proposition 1: An Inductive Approach

Throughout this section, we prove Proposition 1. The proof consists of several steps and uses both conventional fluid limit and mean-field fluid scaling and limit in an intricate fashion. We first provide a road map of the whole proof argument.

4.1. Proof Idea and the Road Map

The key idea for the proof of Proposition 1 is to use backward induction in k, starting from the base case $k = N$. For $k = N$, all the queues are infinite. In that case, parts 1 and 2 are vacuously satisfied with the convention $T(N, \mathbf{x}(0)) = 0$. Further observe that the TABS scheme does not differentiate between two large queues (in fact, any

two nonempty queues). Thus, when all queues are infinite, because all servers are always busy, each arriving task is assigned uniformly at random, and each server has an arrival rate λ and a departure rate one. Thus, it is immediate that the drift of each server is $-(1-\lambda)<0$, and thus, $\varepsilon(N)=1-\lambda$. This proves 3, and then 4 follows as well.

Now, we discuss the ideas to establish the backward induction step; that is, assume that parts 1–4 hold for $k \geq k(N)+1$ for some $k(N) \in \{0,1,\ldots,N-1\}$ and verify that the statements hold for $k = k(N)$. Rigorous proofs to verify parts 1–4 for $k = k(N)$ are presented in Sections 4.2–4.5. We begin by providing a road map of these four subsections.

4.1.1. Part 1. Recall that we denote by $\mathcal{H}$ the indices of the servers having infinite queue lengths and by $\mathcal{N}$ the set of all server indices. Denote by $x_{(i)}$ the ith largest component of $\mathbf{x}$ (ties are broken arbitrarily). Then for any $\mathbf{x}$ with $m \in \{0,1,\ldots,N-1\}$ infinite components, define

$$T(m,\mathbf{x}) := \frac{x_{(N)}}{\varepsilon(N)} + \sum_{i=1}^{N-m-1} \frac{x_{(N-i)} - x_{(N-i+1)}}{\varepsilon(N-i)} \tag{3}$$

with the convention that $T(N,\mathbf{x}) = 0$ if all components of $\mathbf{x}$ are infinite. For $k(N) \in \{0,1,\ldots,N-1\}$, part 1 is proved with the choice of $T(k,\mathbf{x}(0))$ as given by (3). Indeed, recall that we are at the backward induction step at which there are $k(N)$ infinite queues, and we also know from the hypothesis that parts 1–4 hold if there are $k(N)+1$ or larger infinite queues in the system. Loosely speaking, the idea is that, as long as a conventional fluid-scaled queue length $x_j(t)$ at some server $j \in \mathcal{N} \setminus \mathcal{H}$ is positive, it can be coupled with a system in which the queue length at server j is infinite. Thus, as long as there is at least one server $j \in \mathcal{N} \setminus \mathcal{H}$ with $x_j(t) > 0$, the system can be "treated" as a system with at least $k(N)+1$ infinite queues, in which case, part 4 of the backward induction hypothesis furnishes with the drift of each positive component of the FSP (in turn, which is equal to the drift of each infinite queue for the corresponding system).

Now to explain the choice of $T(m,\mathbf{x})$ in (3), observe that, when all the components of the N-dimensional FSP are strictly positive, each component has a negative drift of $-\varepsilon(N)$. Thus, $x_{(N)}/\varepsilon(N)$ is the time when at least one component of the N-dimensional FSP hits zero. From this time point onwards, each positive component has a drift of $-\varepsilon(N-1)$ and, thus, $x_{(N)}/\varepsilon(N) + (x_{(N-1)} - x_{(N)})/\varepsilon(N-1)$ is the time when two components hit zero. Proceeding this way, one can see that, at time $T(m,\mathbf{x}(0))$, all finite positive components of the FSP hit zero. This argument is formalized in Section 4.2.

4.1.2. Part 1 $\Longrightarrow$ Part 2. To prove part 2, we use the fluid limit technique of proving stochastic stability as in Rybko and Stolyar (1992), Dai (1995), and Stolyar (1995); see, for example, Dai (1995), theorem 4.2, or Stolyar (1995), theorem 7.2, for a rigorous statement. Here we need to show that the sum of the noninfinite queues (of an FSP) drains to zero. This is true because, by part 1, each positive noninfinite queue will have negative drift. The formal proof is in Section 4.3.

4.1.3. Part 2 + Lemmas 1 and 2 $\Longrightarrow$ Part 3. Note that, in the proofs of parts 1 and 2, we have only used the backward induction hypothesis and have not imposed any restriction on the value of N. This is the only part where, in the proof, we use the large-scale asymptotics, in particular, Lemmas 1 and 2. For that reason, in the statement of Proposition 1, we use large-enough N. The idea here is to prove by contradiction. Suppose part 3 does not hold for infinitely many values of N. In that case, it can be argued that there exist a subsequence $\{N\}$ and *some* sequence $\{k(N)\}$ with $k(N) \in \{0,1,\ldots,N-1\}$ such that when the subsystem consisting of $N-k(N)$ finite queues is in the steady state, the average arrival rate into each of the $k(N)$ servers having infinite queue lengths is at least one along the subsequence. Loosely speaking, in that case, Lemmas 1 and 2 together imply that for large enough N, there are "enough" busy servers so that the rate of arrival to each infinite queue is strictly smaller than one, which leads to a contradiction. Note that we can apply Lemmas 1 and 2 here because part 2 ensures the required stability. The rigorous proof is in Section 4.4.

4.1.4. Parts 2, 3 + Time-Scale Separation $\Longrightarrow$ Part 4. We assume that parts 1–3 hold for $k \in \{k(N), k(N)+1,\ldots,N\}$, and we verify part 4 for $k = k(N)$. Observe that it only remains to prove convergence to the FSP on the (scaled) time interval $[T(k,\mathbf{x}(0)),\infty]$. For this, observe that it is enough to consider the sequence of systems for which $\mathbf{x}^R(0) \to \mathbf{x}(0)$, where $x_i(0) = 0$ for all $i \in \mathcal{N} \setminus \mathcal{H}$. In particular, all that remains to be shown is that the drift of each infinite queue is indeed $-\varepsilon(k)$. Recall the conventional fluid scaling and FSP from Section 3.1, and let R be the scaling parameter. The proof consists of two main parts:

i. Let us fix any state z of the *unscaled* process. If the sequence of systems is such that $\mathbf{x}^R(0) \to \mathbf{x}(0)$, where $x_i(0) = 0$ for all $i \in \mathcal{N} \setminus \mathcal{H}$; then because of part 2, for the subsystem consisting of finite queues, the (scaled)

hitting time to the (unscaled) state z converges in probability to zero. Also, because this subsystem is positive recurrent (because of part 2), starting from a fixed (unscaled) state z, its expected (unscaled) return time to the state z is $O(1)$. This allows us to split the (unscaled) time line into independent and identically distributed (i.i.d.) renewal cycles of finite expected lengths. In addition, this also shows that, in the scaled time, the subsystem of finite queues evolves on a faster time scale and achieves "instantaneous stationarity."

ii. From this observation, we can claim that the number of arrivals to any specific infinite queue can be written as a sum of arrivals in the previously defined i.i.d. renewal cycles. Using the strong law of large numbers (SLLN), we can then show that, in the limit $R \to \infty$, the instantaneous rate of arrival to a specific infinite queue is given by the *average arrival rate* when the subsystem with $N-k$ finite queues is in steady state. Therefore, part 3 completes the verification of part 4.

This argument is rigorously carried out in Section 4.5.

4.2. Coupling with Infinite Queues to Verify Part 1

The proof of part 1 is constructive in the sense that we identify the structure of the FSP exactly. Fix any $\mathbf{x}(0)$ such that $0 \le \sum_{i\in\mathcal{N}\setminus\mathcal{H}} x_i(0) < \infty$ and $x_i(0) = \infty$ for $i \in \mathcal{H}$. The case in which $\sum_{i\in\mathcal{N}\setminus\mathcal{H}} x_i(0) = 0$ is trivial, so assume $\sum_{i\in\mathcal{N}\setminus\mathcal{H}} x_i(0) > 0$. Let $\mathcal{H}_1 \subseteq \mathcal{N}\setminus\mathcal{H}$ be the set of server-indices i such that $x_i(0) > 0$ and denote $k_1 = |\mathcal{H}_1| > 0$. Denote $x_{(k+k_1)} = \min_{i\in\mathcal{H}_1} x_i(0)$ and $\tau = x_{(k+k_1)}/\varepsilon(k+k_1) > 0$. First, we show that, in the interval $[0,\tau]$, the sequence of scaled processes converges to the FSP such that $x_i'(t) = -\varepsilon(k+k_1)$ for all $i \in \mathcal{H}\cup\mathcal{H}_1$; this FSP is, of course, such that, at time τ, the number of nonzero queues is strictly less that it was at time zero.

Consider the sequence of processes $(x_i^R(\cdot), a_i^R(\cdot), d_i^R(\cdot))_{i\in\mathcal{N}}$ along any subsequence $\{R\}$. Define the stopping time

$$\tau^R := \inf\{t : x_i^R(t) = 0 \text{ for some } i \in \mathcal{H}_1\}.$$

In the time interval $[0,\tau^R]$, we couple this system with a system with $k+k_1$ infinite queues. Let us label this system Π. Let $(\bar{x}_i^R(\cdot), \bar{a}_i^R(\cdot), \bar{d}_i^R(\cdot))_{i\in\mathcal{N}}$ be the queue length, arrival, and departure processes corresponding to the system Π, and assume that $\bar{x}_i^R(0)$ is infinite for $i \in \mathcal{H}\cup\mathcal{H}_1$. We now couple the two systems in the interval $[0,\tau^R]$ in the natural way, including coupling the arrivals and departures to the queues in $\mathcal{H}_1$. Then the arrival and departure processes, to/from each server, coincide in both systems in $[0,\tau^R]$. Therefore, using the induction hypothesis for systems with $k+k_1 \ge k(N)+1$ infinite queues, with probability one, there exists a further subsequence of $\{R\}$ along which, in the interval $[0,\tau]$,

$$(\bar{x}_i^R(\cdot), \bar{a}_i^R(\cdot), \bar{d}_i^R(\cdot))_{i\in\mathcal{N}} \to (\bar{x}_i(\cdot), \bar{a}_i(\cdot), \bar{d}_i(\cdot))_{i\in\mathcal{N}},$$

where $\bar{x}_i \equiv 0$ for all $i \in \mathcal{N}\setminus(\mathcal{H}\cup\mathcal{H}_1)$ and $\bar{x}_j \equiv \infty$ with $\bar{x}_j' \equiv -\varepsilon(k+k_1) < 0$ for all $j \in \mathcal{H}\cup\mathcal{H}_1$. From here, it easily follows that, with probability one, $\tau^R \to \tau$. Thus, in turn, we obtain the probability one convergence in $[0,\tau]$,

$$(x_i^R(\cdot), a_i^R(\cdot), d_i^R(\cdot))_{i\in\mathcal{N}} \to (x_i(\cdot), \bar{a}_i(\cdot), \bar{d}_i(\cdot))_{i\in\mathcal{N}}$$

with $x_i = \bar{x}_i \equiv 0$ for all $i \in \mathcal{N}\setminus(\mathcal{H}\cup\mathcal{H}_1)$ and $x_i' \equiv -\varepsilon(k+k_1) < 0$ for all $i \in \mathcal{H}\cup\mathcal{H}_1$.

We proved the convergence to the specific FSP in the interval $[0,\tau]$. The state of this FSP at time τ is such that the number of nonzero queues is strictly less than at time zero. We can now repeat this argument for the process starting time τ and so on until all finite queues hit zero, that is, until $\sum_{i\in\mathcal{N}\setminus\mathcal{H}} x_i(t)$ hits zero for the first time, which is easily observed to be equal to $T(k(N),\mathbf{x}(0))$ as given in (3). This completes the verification of part 1.

It is worthwhile to note that to prove part 1 for $k = k(N)$, the induction hypothesis assumes that $\varepsilon(k') > 0$ for all $k' > k$ but not $\varepsilon(k)$ itself. The induction step verifying that $\varepsilon(k) > 0$ is done when we prove part 3 in Section 4.4.

4.3. Conventional Fluid-Limit Stability to Verify Part 2

As mentioned earlier, we use the fluid limit technique of proving stochastic stability as in Rybko and Stolyar (1992), Dai (1995), and Stolyar (1995) to prove part 2; see, for example, Dai (1995), theorem 4.2, or Stolyar (1995), theorem 7.2. Specifically, it suffices to prove Lemma 3. (It is stated in the form that is convenient to use again later on.)

For the queue length vector $\mathbf{X}$, define the norm

$$\|\mathbf{X}\| := \sum_{i\notin\mathcal{H}} X_i \tag{4}$$

to be the total number of tasks at the finite queues.

Lemma 3. *There exist fixed $\gamma \in (0,1)$, $\tau > 0$ and $C > 0$ such that, if $\|\mathbf{X}(0)\| = R \geq C$, then*

$$\mathbb{E}\|X(R\tau)\| \leq (1-\gamma)\|X(0)\|.$$

Lemma 3 says that, if the system starts from an initial state in which the total number of tasks in the finite queues is suitably large, then the time it takes when the expected total number of tasks in the finite queues falls below a certain fraction of the initial number is proportional to itself. The proof of Lemma 3 is fairly straightforward, but is provided for completeness.

Proof of Lemma 3. Consider a sequence of initial states with an increasing norm; that is, $\mathbf{X}^R(0)$ is such that $\|\mathbf{X}^R(0)\| = R$, where $R^{-1}\mathbf{X}^R(0) \to \mathbf{x}(0)$ as $R \to \infty$, where $\|\mathbf{x}(0)\| = 1$. Then, from part 1, we know that, as $R \to \infty$, on the time interval $[0, T(k, \mathbf{x}(0))]$, the process $R^{-1}\mathbf{X}^R(Rt)$ converges in probability to the unique deterministic process $\mathbf{x}(t)$, satisfying

$$\sum_{i \in \mathcal{N}\setminus\mathcal{H}} x_i'(t) \leq -c < 0 \quad \text{whenever} \quad \sum_{i \in \mathcal{N}\setminus\mathcal{H}} x_i(t) > 0. \tag{5}$$

It is also easy to see that $T(k, \mathbf{x}(0)) \geq \tau$ for some $\tau > 0$ and (uniformly on $\|\mathbf{x}(0)\|=1$) $\|\mathbf{x}(\tau)\| \leq 1-\gamma$ for some $\gamma = \gamma(\tau) > 0$.

Now, because the expected number of arrivals into the Rth system up to time t, when scaled by R, is λt for any finite t, we obtain $\mathbb{E}(R^{-1}\|X^R(t)\|) \leq 1 + \lambda t$. Therefore, the convergence in probability also implies the convergence in expectation. Thus, for this choice of γ,

$$\limsup_{R\to\infty} \mathbb{E}\left(\frac{\|\mathbf{X}^R(R\tau)\|}{R}\right) < 1-\gamma.$$

Hence, there exists C such that for all $R \geq C$,

$$\mathbb{E}\left(\frac{\|\mathbf{X}^R(R\tau)\|}{R}\right) = \mathbb{E}\left(\frac{\|\mathbf{X}^R(R\tau)\|}{\|\mathbf{X}^R(0)\|}\right) \leq 1-\gamma.$$

This completes the proof of Lemma 3.

4.4. Large-Scale Asymptotics to Verify Part 3

We proved in part 2 that the system consisting of $N-k$ finite queues is stable. This implies that there exists a well-defined steady-state average rate $1-\varepsilon(k)$ of new arrivals into each infinite queue. We need to verify (for all sufficiently large N) the backward induction step proving that $\varepsilon(k) > 0$. This verification uses a contradiction. Namely, for any fixed N, if part 3 does not hold for some k, then that means $\varepsilon(k) \leq 0$. Now, if we assume that the induction step for part 3 does not hold for *infinitely many* (k, N) pairs (where $k = k(N)$), then, for all of them, $\varepsilon(k(N)) \leq 0$. This is formally stated in Implication 1. In that case, we construct a sequence of systems with increasing N for which we obtain a contradiction using Lemmas 1 and 2. We note that this is the only part in the proof of Proposition 1 in which we use the large-scale (i.e., $N \to \infty$) asymptotic results.

Implication 1. Suppose, for infinitely many N, the induction step to prove part 3 of Proposition 1 does not hold for some $k = k(N)$. Then there exists a subsequence of $\{N\}$ (which we still denote by $\{N\}$) diverging to infinity such that (i) the system with $k(N)$ infinite queues is stable and (ii) $\varepsilon(k(N)) \leq 0$.

We now show that Implication 1 leads to a contradiction; this verifies part 3 of Proposition 1. Suppose Implication 1 is true. Choose a further subsequence $\{N\}$ along which $k(N)/N$ converges to $\kappa \in [0,1]$. As in the statement of Lemma 2, we consider two regimes depending on whether $\kappa \geq 1-\lambda$ or not and arrive at contradictions in both cases. Because all the infinite queues are exchangeable, we use σ to denote a typical infinite queue.

4.4.1. Case 1. First consider the case in which $\kappa \geq 1-\lambda$. Note that the expected steady-state instantaneous rate of arrival to σ is given by

$$\mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)}\mathbf{1}_{[Q_1^N(\infty)+\Delta_0^N(\infty)+\Delta_1^N(\infty)=N]}\right) \leq \mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)}\right) = \mathbb{E}\left(\frac{\lambda}{q_1^N(\infty)}\right) + o(1). \tag{6}$$

Now observe that, for large N, $\lambda/q_1^N(\infty) \le 2\lambda/\kappa$ because $q_1^N(s) \ge \kappa/2 > 0$. Further, from Lemma 2, we know that $q_1^N(\infty) \xrightarrow{\mathbb{P}} 1$ as $N \to \infty$. Consequently, $\mathbb{E}(\lambda/q_1^N(\infty)) \to \lambda$ as $N \to \infty$. Therefore, for large enough N,

$$\mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)}\mathbf{1}_{[Q_1^N(\infty)+\Delta_0^N(\infty)+\Delta_1^N(\infty)=N]}\right) \le \frac{1+\lambda}{2} = 1 - \frac{1-\lambda}{2} < 1, \tag{7}$$

which is a contradiction to part ii of Implication 1.

4.4.2. Case 2. In case $\kappa < 1-\lambda$, first note that the statement in part 3 is vacuously satisfied if $k(N) \equiv 0$. Thus, without loss of generality, assume $k(N) > 0$. Fix ε_1 as in Lemma 1. In that case, (6) becomes

$$\begin{aligned}
&\mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)}\mathbf{1}_{[Q_1^N(\infty)+\Delta_0^N(\infty)+\Delta_1^N(\infty)=N]}\right)\\
&\le \mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)}\mathbf{1}_{[Q_1^N(\infty)+\Delta_0^N(\infty)+\Delta_1^N(\infty)=N,\ Q_1^N(\infty)\ge\varepsilon_1 N]}\right) + \lambda N\mathbb{P}\big(Q_1^N(\infty) < \varepsilon_1 N\big)\\
&\le \frac{\lambda N}{\varepsilon_1 N}\mathbb{P}\big(Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N\big) + \lambda N\mathbb{P}\big(Q_1^N(\infty) < \varepsilon_1 N\big).
\end{aligned}$$

Now, because of part 2 of Lemma 2, we know that

$$\mathbb{P}\big(Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N\big) \to 0,$$

and furthermore, Lemma 1 yields

$$N\mathbb{P}\big(Q_1^N(\infty) < \varepsilon_1 N\big) \to 0 \quad \text{as} \quad N \to \infty.$$

Thus,

$$\mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)}\mathbf{1}_{[Q_1^N(\infty)+\Delta_0^N(\infty)+\Delta_1^N(\infty)=N]}\right) \to 0 \quad \text{as} \quad N \to \infty. \tag{8}$$

In particular, for large enough N, the expected steady-state arrival rate is bounded away from one, which is again a contradiction to part ii of Implication 1. This completes the verification of part 3 of the backward induction hypothesis.

4.5. Time-Scale Separation to Verify Part 4

Assume parts 1–3 hold for all $k \in \{k(N), k(N)+1, \dots, N\}$. Now consider a system containing $k = k(N)$ infinite queues with indices in $\mathcal{H}$, and recall the conventional fluid scaling and FSP from Section 3.1. Also, in this subsection, whenever we refer to the process $\{\mathbf{X}(t)\}_{t\ge 0}$, the components in $\mathcal{H}$ should be taken to be infinite.

Lemma 4 states a hitting time result that is used in verifying part 4.

For any $C > 0$, define the set $\mathcal{C} := \{\|\mathbf{X}\| \le C\}$ and the stopping time $\theta_C := \inf\{t : \mathbf{X}(t) \in \mathcal{C}\}$. For large enough C, the next lemma bounds the expected hitting time to the fixed set $\mathcal{C}$ in terms of the norm of the initial state.

Lemma 4. *There exists $C, C_1 > 0$, such that, if $\|\mathbf{X}(0)\| = R \ge C$, then*

$$\mathbb{E}(\theta_C | \mathbf{X}(0)) \le C_1 \|\mathbf{X}(0)\|.$$

Proof of Lemma 4. Fix $\gamma \in (0,1)$, $\tau > 0$, and $C > 0$ as in Lemma 3. For $i \ge 1$, define the sequence of random variables $T_i := \tau\|\mathbf{X}(T_{i-1})\|$ with the convention that $T_0 = 0$. Now consider the discrete time Markov chain $\{\Phi_i : i \ge 0\}$ adapted to the filtration $\mathcal{F} = \bigcup_{i\ge 0}\mathcal{F}_i$, where $\Phi_i = \mathbf{X}(T_i)$ is the value of the continuous time Markov process sample at times T_is, and $\mathcal{F}_i = \sigma(\Phi_0, \Phi_1, \dots, \Phi_i)$ is the sigma field generated by $\{\Phi_0, \Phi_1, \dots, \Phi_i\}$. Further, for $i \ge 0$, define the stopping time $\hat{\theta}_C := \inf\{j \ge 0 : \Phi_j \le C\}$. Then observe that

$$\theta_C \le \sum_{i=1}^{\hat{\theta}_C} T_i =: \Psi_C.$$

Also define $\alpha_i = \sum_{j=1}^{i} T_j$ for $i \geq 1$, and hence, $\alpha_{\hat{\theta}_C} = \Psi_C$. Then, as a consequence of Dynkin's lemma (Meyn and Tweedie 1993, theorem 11.3.1), using proposition 11.3.2 of Meyn and Tweedie (1993), we have

$$\mathbb{E}(\theta_C) \leq \mathbb{E}(\Psi_C) \leq \frac{\tau}{\gamma}\mathbb{E}\|\mathbf{X}(0)\|.$$

Choosing $C_1 = \tau/\gamma$ completes the proof.

Now we have all the ingredients to verify part 4 of the backward induction hypothesis. Note that we now look at the sequence of conventional fluid-scaled processes starting at (scaled) time $T(k(N), \mathbf{x}(0))$. Let us show that starting from time $T(k(N), \mathbf{x}(0))$, the drift of each of the infinite queues is at most $-\varepsilon(k(N))$. Specifically, we construct a probability space in which the required probability one convergence holds.

To simplify writing, we assume that the system starts at time zero, and thus, it is enough to consider a sequence of initial queue length vectors such that

$$\|\mathbf{x}^R(0)\| \to 0 \quad \text{as} \quad R \to \infty,$$

where R is the parameter in the conventional fluid scaling. Hence, Lemma 4 yields that $R^{-1}\mathbb{E}(\theta_C|\mathbf{X}^R(0)) \to 0$ as $R \to \infty$. Consequently, $R^{-1}\theta_C \xrightarrow{\mathbb{P}} 0$. Thus, the fluid-scaled time to hit the set $\mathscr{C}$ vanishes in probability, which is stated formally in the following claim.

Claim 1. If the sequence of initial states is such that $\|\mathbf{x}^R(0)\| \to 0$ as $R \to \infty$, then $R^{-1}\theta_C \xrightarrow{\mathbb{P}} 0$, as $R \to \infty$.

Now pick any (unscaled) state $z \in \mathscr{C}$ and define the stopping time $\hat{\theta}_z$ as

$$\hat{\theta}_z := \inf\{t \geq 0 : \mathbf{X}(t) = z\}.$$

Because of part 2 of the backward induction hypothesis, the unscaled process $\mathbf{X}(\cdot)$ is irreducible and positive recurrent, we have the following claim.

Claim 2. If the sequence of initial states is such that $\mathbf{x}^R(0) \in \mathscr{C}$, then $R^{-1}\hat{\theta}_z \xrightarrow{\mathbb{P}} 0$, as $R \to \infty$.

Up to time $\hat{\theta}_z$, consider the product topology on the sequence space. Then Claims 1 and 2 yield that, for a sequence of initial states such that $\|\mathbf{x}^R(0)\| \to 0$ as $R \to \infty$, there exists a subsequence $\{R\}$, along which with probability one, $R^{-1}\hat{\theta}_z \to 0$. Starting from the time $\hat{\theta}_z$, along this subsequence, we construct the sequence of processes $\mathbf{x}^R(\cdot)$ on the same probability space as follows.

1. Define the space of an infinite sequence of i.i.d. renewal cycles of the unscaled process $\mathbf{X}(\cdot)$ with the unscaled state z being the renewal state, that is,

$$\left\{\mathbf{X}^{(i)}(t) : 0 \leq t \leq \hat{\theta}_z^{(i)}, \mathbf{X}^{(i)}(0) = z\right\}$$

for $i = 1, 2, \ldots$ are i.i.d. copies, and $\hat{\theta}_z^{(i)}$ are also i.i.d. copies of $\hat{\theta}_z$.

2. Define the process $\mathbf{X}^R(\cdot)$ as

$$\mathbf{X}^R(Rt) = \sum_{i=1}^{\infty} \mathbf{X}^{(i)}(Rt - \Theta(i-1))\mathbf{1}_{[\Theta(i-1) \leq Rt < \Theta(i)]}, \quad \text{where} \quad \Theta(i) := \sum_{j=1}^{i} \hat{\theta}^{(j)}, \ i \geq 0$$

Let $A_n^R(t)$ denote the cumulative number of arrivals up to time t in the Rth system to a fixed server n with infinite queue length when the system starts from the state z. Observe that

$$\sum_{i=0}^{N_\theta^R} A^{(i)} \leq A_n^R(Rt) \leq \sum_{i=0}^{N_\theta^R+1} A^{(i)},$$

where $N_\theta^R := \max\{j : \Theta(j) \leq Rt\}$ and $A^{(i)}$ is the total number of arrivals into queue n during ith renewal cycle. Now, because the subsystem consisting of the finite queues is stable, $\mathbf{X}(\cdot)$ is positive recurrent. Thus, we have $\mathbb{E}(\hat{\theta}_z|\mathbf{X}(0) = z) < \infty$, and hence, with probability one,

$$\frac{N_\theta^R}{R} \to \frac{t}{\mathbb{E}(\hat{\theta}_z|\mathbf{X}(0) = z)}, \quad \text{as} \quad R \to \infty.$$

Thus, using part 3 of the backward induction hypothesis, SLLN yields that, with probability one,

$$\frac{1}{R}A_n^R(Rt) \to (1-\varepsilon(k(N)))t, \quad \text{as} \quad R \to \infty.$$

This is true for any $t \geq 0$, and therefore (because $A_n^R(Rt)$ is nondecreasing in t), the convergence is uniform on compact sets with probability one. Also, because the average departure rate from each of the servers with infinite queue lengths is always one, it easily follows that the fluid-scaled departure process (starting time $\hat{\theta}_z$) with probability one converges (u.o.c.) to the function t. Combining this with the probability one convergence of the time $\hat{\theta}_z$ to zero, we obtain the probability one, u.o.c. convergence to the unique FSP with $a_n(t) = (1-\varepsilon(k(N)))t$, $d_n(t) = t$, and thus, $x_n'(t) = -\varepsilon(k(N))$ for each infinite queue n.

Finally, combining the probability one convergence $\hat{\theta}_z \to 0$ with the construction of the process after $\hat{\theta}_z$ as renewal with i.i.d. cycles, we obtain the probability one, u.o.c., convergence of $x_i^R(t) \to x_i(t) = 0$ on the interval $[T(k(N), \mathbf{x}(0)), \infty)$ for all $i \in \mathcal{N} \setminus \mathcal{H}$.

This completes the verification of part 4 and, hence, of Proposition 1.

5. Mean-Field Analysis for Large-Scale Asymptotics

In this section, we analyze the large-N behavior of the system. In particular, we prove Lemmas 1 and 2. The next proposition is a basic mean-field fluid limit result that we need later. Define

$$E_\kappa := \{(\boldsymbol{q}, \boldsymbol{\delta}) \in [0,1]^\infty : q_i \geq q_{i+1} \geq \kappa, \ \forall i, \ \delta_0 + \delta_1 + q_1 \leq 1\}.$$

Proposition 2. *Assume $k(N)/N \to \kappa \in [0,1]$, and the sequence of initial states $(\boldsymbol{q}^N(0), \boldsymbol{\delta}^N(0))$ converge to a fixed $(\boldsymbol{q}(0), \boldsymbol{\delta}(0)) \in E_\kappa$ as $N \to \infty$, where $q_1(0) > 0$. Then, with probability one, any subsequence of $\{N\}$ has a further subsequence along which $\{(\boldsymbol{q}^N(t), \boldsymbol{\delta}^N(t))\}_{t\geq 0}$ converges uniformly on compact time intervals to some deterministic trajectory $\{(\boldsymbol{q}(t), \boldsymbol{\delta}(t))\}_{t\geq 0}$ satisfying the following equations:*

$$q_i(t) = q_i(0) + \int_0^t \lambda p_{i-1}(\boldsymbol{q}(s), \boldsymbol{\delta}(s), \lambda)ds - \int_0^t (q_i(s) - q_{i+1}(s))ds, \ i \geq 1$$

$$\delta_0(t) = \delta_0(0) + \mu \int_0^t u(s)ds - \xi(t),$$

$$\delta_1(t) = \delta_1(0) + \xi(t) - \nu \int_0^t \delta_1(s)ds,$$

where

$$u(t) = 1 - q_1(t) - \delta_0(t) - \delta_1(t),$$

$$\xi(t) = \int_0^t \lambda(1 - p_0(\boldsymbol{q}(s), \boldsymbol{\delta}(s), \lambda))\mathbf{1}_{[\delta_0(s)>0]}ds.$$

For any $(\boldsymbol{q}, \boldsymbol{\delta}) \in E$, $\lambda > 0$, $(p_i(\boldsymbol{q}, \boldsymbol{\delta}, \lambda))_{i\geq 0}$ are given by

$$p_0(\boldsymbol{q}, \boldsymbol{\delta}, \lambda) = \begin{cases} 1 & \text{if} \quad u = 1 - q_1 - \delta_0 - \delta_1 > 0, \\ \min\{\lambda^{-1}(\delta_1\nu + q_1 - q_2), 1\}, & \text{otherwise}, \end{cases}$$

$$p_i(\boldsymbol{q}, \boldsymbol{\delta}, \lambda) = (1 - p_0(\boldsymbol{q}, \boldsymbol{\delta}, \lambda))(q_i - q_{i+1})q_1^{-1}, \ i \geq 1.$$

This type of result is standard and is obtained using functional strong LLN, for example, as in Mukherjee et al. (2017) and Stolyar (2015, 2017); we omit its proof. Also, we note that, although Proposition 2 is a version of Mukherjee et al. (2017), theorem 3.1, it is different in that it is suitably modified for the case of infinite buffers and some queues being infinite, and it states a somewhat different type of convergence, convenient for the use in this paper. Define a *mean-field fluid sample path* (MFFSP) to be any deterministic trajectory satisfying the properties stated in Proposition 2.

We now provide an intuitive explanation of the mean-field fluid limit stated in Proposition 2. It is similar to that behind Mukherjee et al. (2017), theorem 3.1. The term $u(t)$ corresponds to the asymptotic fraction of idle-on servers in the system at time t, and $\xi(t)$ represents the asymptotic cumulative number of server setups (scaled by N) that have been initiated during $[0,t]$. The coefficient $p_i(\boldsymbol{q}, \boldsymbol{\delta}, \lambda)$ can be interpreted as the

instantaneous fraction of incoming tasks that are assigned to some server with queue length i when the fluid-scaled occupancy state is $(\boldsymbol{q}, \boldsymbol{\delta})$ and the scaled instantaneous arrival rate is λ. Observe that as long as $u > 0$, there are idle-on servers, and hence, all the arriving tasks join idle servers. This explains that, if $u > 0$, $p_0(\boldsymbol{q}, \boldsymbol{\delta}, \lambda) = 1$ and $p_i(\boldsymbol{q}, \boldsymbol{\delta}, \lambda) = 0$ for $i = 1, 2, \ldots$. If $u = 0$, then observe that servers become idle at rate $q_1 - q_2$, and servers in setup mode turn on at rate $\delta_1 \nu$. Thus, the idle-on servers are created at a total rate $\delta_1 \nu + q_1 - q_2$. If this rate is larger than the arrival rate λ, then almost all the arriving tasks can be assigned to idle servers. Otherwise, only a fraction $(\delta_1 \nu + q_1 - q_2)/\lambda$ of arriving tasks join idle servers. The rest of the tasks are distributed uniformly among busy servers, so a proportion $(q_i - q_{i+1})q_1^{-1}$ are assigned to servers having queue length i. For any $i = 1, 2, \ldots$, q_i increases when there is an arrival to some server with queue length $i - 1$, which occurs at rate $\lambda p_{i-1}(\boldsymbol{q}, \boldsymbol{\delta}, \lambda)$, and it decreases when there is a departure from some server with queue length i, which occurs at rate $q_i - q_{i-1}$. Because each idle-on server turns off at rate μ, the fraction of servers in the off mode increases at rate μu. Observe that, if $\delta_0 > 0$, for each task that cannot be assigned to an idle server, a setup procedure is initiated at one idle-off server. As noted, $\xi(t)$ captures the (scaled) cumulative number of setup procedures initiated up to time t. Therefore, the fraction of idle-off servers and the fraction of servers in setup mode decreases and increases by $\xi(t)$, respectively, during $[0, t]$. Finally, because each server in setup mode becomes idle-on at rate ν, the fraction of servers in setup mode decreases at rate $\nu\delta_1$.

5.1. Proof of Lemma 1

Throughout this subsection, we prove Lemma 1. Within this proof, we use the following terminology. Let A^N be an event pertaining to the Nth system. We write $\mathbb{P}(A^N) = \eta(N)$ to mean the following property: *There exist* $C > 0$ *and* $N_1 > 0$ *such that* $\mathbb{P}(A^N) \le \mathrm{e}^{-CN}$ *for all* $N \ge N_1$. If event A^N depends on some parameter p (say, the process initial state), we say that $\mathbb{P}(A^N) = \eta(N)$ *uniformly in* p if the property holds for common fixed $C > 0$ and $N_1 > 0$.

To prove the lemma, clearly, it suffices to prove that, for some fixed $T_0 > 0$ and $\varepsilon_0 > 0$,

$$\mathbb{P}\big(q_1^N(T_0) \le \varepsilon_0\big) = \eta(N), \tag{9}$$

uniformly on the process initial states $(\boldsymbol{q}^N(0), \boldsymbol{\delta}^N(0))$. This is what we do in the rest of the proof.

Fix any $T_0 > 0$; $\varepsilon_0 > 0$ is chosen later. We now prove several claims, which rather simply follow from the process structure and basic large deviation estimates (specifically, Cramer's theorem); they serve as building blocks for the proof argument.

Claim 3. (i) For any $\varepsilon > 0$, uniformly in $\tau \in [0, T_0]$ and uniformly in $q_1^N(0) \ge \varepsilon$,

$$\mathbb{P}\big(q_1^N(\tau) \le (\varepsilon/2)\mathrm{e}^{-T_0}\big) = \eta(N). \tag{10}$$

(ii) For any $\varepsilon > 0$, uniformly in $\tau \in [0, T_0]$ and uniformly in $\delta_1^N(0) \ge \varepsilon$,

$$\mathbb{P}\big(\delta_1^N(\tau) \le (\varepsilon/2)\mathrm{e}^{-\nu T_0}\big) = \eta(N). \tag{11}$$

Indeed, to prove (10), observe that any busy server at time t stays busy in the interval $[t, t + \tau]$ with probability at least $\mathrm{e}^{-\tau} \ge \mathrm{e}^{-T_0}$. It remains to recall that $q_1^N(0) \ge \varepsilon$ corresponds to at least εN busy servers in the unscaled system and apply Cramer's theorem. Statement (ii) is proved analogously.

Claim 4. For any sufficiently small $T_1 > 0$, there exists $\varepsilon_1' > 0$ such that, uniformly in the initial state $(\boldsymbol{q}^N(0), \boldsymbol{\delta}^N(0))$,

$$\mathbb{P}\big(q_1^N(T_1) + \delta_1^N(T_1) \le \varepsilon_1'\big) = \eta(N). \tag{12}$$

Indeed, fix any $T_1 > 0$ such that $\lambda T_1 \le 1/4$. Suppose first that either $q_1^N(0) \ge 1/4$ or $\delta_1^N(0) \ge 1/4$; uniformly on all such initial conditions, the claim follows by using Claim 3. Suppose now that $q_1^N(0) < 1/4$ and $\delta_1^N(0) < 1/4$, and therefore, $\delta_0^N(0) + u^N(0) > 1/2$, where recall that u^N is the fraction of idle-on servers. The (unscaled) number of new customer arrivals in $[0, T_1]$, denote it by $H[0, T_1]$, is Poisson with mean $\lambda T_1 N$; therefore,

$$\mathbb{P}(|H[0, T_1]/N - \lambda T_1| \ge (1/2)\lambda T_1) = \eta(N).$$

This means that, with probability $1 - \eta(N)$, we have $H[0, T_1]/N < \delta_0^N(0) + u^N(0)$, and therefore, each arrival in $[0, T_1]$ creates either a new busy server or a new setup server; furthermore, each of these newly created busy or

setup servers will not change its state until time T_1 with probability at least $e^{-\nu' T_1}$, where $\nu' = \max\{\nu, 1\}$. It remains to choose $\varepsilon_1' \in (0, (1/4)\lambda T_1 e^{-\nu' T_1})$ to obtain the claim.

Claim 5. For any $\varepsilon_1 > 0$ and any $T_2 > 0$, there exists $\varepsilon_2' > 0$ such that, uniformly in $\delta_1^N(0) \geq \varepsilon_1$,

$$\mathbb{P}\left(q_1^N(T_2) + u^N(T_2) \leq \varepsilon_2'\right) = \eta(N). \tag{13}$$

Indeed, at time zero, there are at least $\varepsilon_1 N$ setup servers. Fix any $T_2 > 0$. In $[0, T_2]$, each of them turns into an idle-on server with probability at least $1 - e^{-\nu T_2}$; those servers that do turn into idle-on are either still idle-on or busy at time T_2 with probability at least $e^{-\nu'' T_2}$, where $\nu'' = \max\{\mu, \nu\}$. It remains to choose $\varepsilon_2' \in (0, (1/2)\varepsilon_1 e^{-\nu'' T_2})$ and apply Cramer's theorem.

Claim 6. For any $\varepsilon_2 > 0$ and any sufficiently small $T_3 > 0$, there exists $\varepsilon_3 > 0$ such that, uniformly in $u^N(0) \geq \varepsilon_2$,

$$\mathbb{P}\left(q_1^N(T_3) \leq \varepsilon_3\right) = \eta(N). \tag{14}$$

Indeed, fix T_3 small enough so that $e^{-\mu T_3} > 3/4$ and $\lambda T_3 < \varepsilon_2/2$. At time zero, there are at least $\varepsilon_2 N$ idle-on servers; with probability at least $e^{-\mu T_3} > 3/4$, they will still be idle-on at time T_3 *unless* they are taken by a new arrival. The (unscaled) number of new arrivals in $[0, T_3]$, namely $H[0, T_3]$, is Poisson with mean $\lambda T_3 N$, and therefore, $H[0, T_3]/N$ concentrates at λT_3: $\mathbb{P}(|H[0, T_3]/N - \lambda T_3| \geq (1/2)\lambda T_3) = \eta(N)$. We conclude that, with probability $1 - \eta(N)$, every new arrival in $[0, T_3]$ goes to an idle-on server and turns it into busy; each of those servers, in turn, remain busy until T_3 with probability at least e^{-T_3}. It remains to choose $\varepsilon_3 \in (0, (1/4)\lambda T_3 e^{-T_3})$ to obtain the claim.

With these claims, we are now in position to conclude the proof of the lemma. Choose small $T_1 > 0$ and $\varepsilon_1' > 0$ as in Claim 4, and then $\varepsilon_1 = \varepsilon_1'/2$. For the chosen ε_1, choose small $T_2 > 0$ and $\varepsilon_2' > 0$ as in Claim 5, and then $\varepsilon_2 = \varepsilon_2'/2$. Finally, for the chosen ε_2, choose small $T_3 > 0$ and $\varepsilon_3 > 0$ as in Claim 6. Note that T_1, T_2, T_3 can be small enough so that $T_3' \doteq T_1 + T_2 + T_3 \leq T_0$; let us also denote $T_2' = T_1 + T_2$. Choose $\varepsilon_0 = (1/2)\min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\} e^{-T_0}$.

According to Claim 4, with probability $1 - \eta(N)$, at time T_1, we have either $q_1^N(T_1) \geq \varepsilon_1$ or $\delta_1^N(T_1) \geq \varepsilon_1$. Conditioned on a state at T_1 satisfying $q_1^N(T_1) \geq \varepsilon_1$, we have (9) by applying Claim 3. Therefore, it remains to prove (9) conditioned on a state at T_1 satisfying $\delta_1^N(T_1) \geq \varepsilon_1$. Under this condition at T_1, we obtain from Claim 5 that, with probability $1 - \eta(N)$, at time T_2' we have either $q_1^N(T_2') \geq \varepsilon_2$ or $u^N(T_2') \geq \varepsilon_2$. Then, conditioned on a state at T_2' satisfying $q_1^N(T_2') \geq \varepsilon_2$, we have (9) by once again applying Claim 3. It now remains to prove (9) conditioned on a state at T_2' satisfying $u^N(T_2') \geq \varepsilon_2$. Under this condition at T_2', we obtain from Claim 6 that, with probability $1 - \eta(N)$, at time T_3', we have $q_1^N(T_3') \geq \varepsilon_3$, and conditioned on $q_1^N(T_3') \geq \varepsilon_3$ at T_3', we have (9) by, yet again, Claim 3. The proof is complete.

5.2. Proof of Lemma 2

Throughout this section, we prove Lemma 2. Recall that the stability of the subsystem $\mathcal{N} \setminus \mathcal{H}$ is assumed, and hence, there exists a unique stationary distribution for each N. Recall that we denote by $\boldsymbol{q}^N(\infty)$ the random value of $\boldsymbol{q}^N(t)$ in the steady state. We start by stating a few basic facts about the mean-field limits that facilitate the proof of Lemma 2.

Recall the definition of MFFSP from the paragraph after Proposition 2 and that $u(t) = 1 - q_1(t) - \delta_0(t) - \delta_1(t)$. Also, denote by $y_1(t) = q_1(t) - q_2(t)$ and by (d^+/dt) the right derivative.

Claim 7. For any $\varepsilon > 0$, there exists $\alpha > 0$ such that any MFFSP with $q_1(0) > 0$ satisfies the following properties for all $t \geq 0$:

i. If $y_1(t) \leq \lambda - \varepsilon$ and $u(t) > 0$, then $(d^+/dt)q_1(t) \geq \alpha$.

ii. If $y_1(t) \leq \lambda - \varepsilon$, $u(t) = 0$, and $\delta_1(t) \geq \varepsilon$, then $(d^+/dt)q_1(t) \geq \alpha$.

iii. If $y_1(t) \leq \lambda - \varepsilon$, $u(t) = 0$, $\delta_1(t) = 0$, and $\delta_0(t) > 0$, then $(d^+/dt)\delta_1(t) \geq \varepsilon$.

Fix any $\varepsilon > 0$. First, observe that because $q_1(0) > 0$ and because of Proposition 2, $q_1(0)$ is nondecreasing whenever $q_1(t) - q_2(t) \leq \lambda$, and we have $q_1(t) \geq \min\{q_1(0), \lambda\} > 0$ for all $t \geq 0$. Thus, Proposition 2 can be applied for all $t \geq 0$ throughout the MFFSP. Choose $\alpha = \min\{\varepsilon\nu, \varepsilon\}$.

For Claim 7i, note that, if $y_1(t) \leq \lambda - \varepsilon$ and $u(t) > 0$, then

$$(d^+/dt)q_1(t) = \lambda - (q_1(t) - q_2(t)) \geq \varepsilon \geq \alpha.$$

For Claim 7ii, note that, if $y_1(t) \le \lambda - \varepsilon$, $u(t) = 0$, and $\delta_1(t) \ge \varepsilon$, then, because of Proposition 2,

$$\begin{aligned}(d^+/dt)q_1(t) &= \min\{(\delta_1(t)\nu + q_1(t) - q_2(t)), \lambda\} - (q_1(t) - q_2(t)) \\ &= \min\{\delta_1(t)\nu, \quad \lambda - (q_1(t) - q_2(t))\} \ge \min\{\varepsilon\nu, \quad \varepsilon\} = \alpha.\end{aligned}$$

Finally, for Claim 7iii, note that, from Proposition 2, if $y_1(t) \le \lambda - \varepsilon$, $u(t) = 0$, $\delta_1(t) = 0$, and $\delta_0(t) > 0$, then $(d^+/dt)\delta_1(t) = \lambda - (q_1(t) - q_2(t)) \ge \varepsilon$.

5.2.1. Proof of Statement 1. Note that it is enough to prove the following property of any MFFSP:

Claim 8. Starting from any state $\boldsymbol{q}(0) \in E_\kappa$ with $\kappa \ge 1 - \lambda$ along any MFFSP, we have

$$\lim_{t\to\infty} q_1(t) = 1.$$

Indeed, Claim 8 implies that, under the assumption of stability, asymptotically the stationary distribution of $q_1^N(t)$ must concentrate at $q_1^\star = 1$ as $N \to \infty$.

Proof of Claim 8. We prove by contradiction. Note that, for the case under consideration, $q_i(t) \ge \kappa$ for all $i \ge 1$ and $t \ge 0$. Therefore, throughout the proof of Claim 8, we can assume $q_1(0) \ge \kappa > 0$ and can apply Proposition 2 and Claim 7.

Note that, if $q_1(t) < 1$, we have $q_1(t) - q_2(t) < 1 - \kappa \le \lambda$, and hence, because of Claim 7, $q_1(t)$ is nondecreasing. Thus, if Claim 8 does not hold, then there exists an $\varepsilon > 0$, such that $q_1(t) \le 1 - \varepsilon\nu$ for all $t \ge 0$, and hence,

$$q_1(t) - q_2(t) \le \lambda - \varepsilon\nu \quad \text{for all} \quad t \ge 0. \tag{15}$$

The high-level proof idea is that, if $q_1(t)$ remains below one by a nonvanishing amount for all $t \ge 0$, then the (scaled) rate $q_1(t) - q_2(t)$ of busy servers turning idle-on would not be high enough to match the (scaled) rate λ of incoming jobs. If there are idle-on servers (as in Claim 7i) or sufficiently many servers in setup mode (as in Claim 7ii), then we can still assign incoming tasks to idle-on servers, but this drives up the fraction of busy servers $q_1(t)$ and cannot continue indefinitely because $q_1(t) \le 1 - \varepsilon\nu$ for all $t \ge 0$. This means that we cannot initiate an unbounded number of setup procedures (see Equation (19)). At the same time, as argued, we cannot continue assigning tasks to idle-on servers either. Thus, throughout the MFFSP, a positive fraction of the jobs are assigned to busy servers, which initiates an unbounded (scaled) number of setup procedures and, hence, the contradiction.

Define the subset $\mathscr{K}_\kappa \subseteq E$ as

$$\mathscr{K}_\kappa := \{(\boldsymbol{q}, \boldsymbol{\delta}) \in E_\kappa : q_1 + \delta_0 + \delta_1 = 1, \delta_1\nu + q_1 - q_2 \le \lambda\},$$

and denote by $\mathbf{1}_{\mathscr{K}_\kappa}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))$ the indicator of $(\boldsymbol{q}(s), \boldsymbol{\delta}(s)) \in \mathscr{K}_\kappa$. Observe that, because of Proposition 2, $q_1(t)$ can be written as

$$q_1(t) = q_1(0) + \int_0^t \delta_1(s)\nu \mathbf{1}_{\mathscr{K}_\kappa}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))\mathrm{d}s + \int_0^t [\lambda - q_1(s) + q_2(s)]\mathbf{1}_{\mathscr{K}_\kappa^c}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))\mathrm{d}s. \tag{16}$$

Thus,

$$q_1(t) \ge q_1(0) + \int_0^t [\lambda - q_1(s) + q_2(s)]\mathbf{1}_{\mathscr{K}_\kappa^c}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))\mathrm{d}s,$$

and (15) yields that there exists a positive constant K_1, which may depend on ε such that $\forall\ t \ge 0$

$$\int_0^t \mathbf{1}_{\mathscr{K}_\kappa^c}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))\mathrm{d}s < K_1 \implies \int_0^t \mathbf{1}_{[u(s)>0]}\mathrm{d}s < K_1. \tag{17}$$

Again, from (16), we obtain

$$\begin{aligned}q_1(t) &\ge q_1(0) + \int_0^t \delta_1(s)\nu \mathbf{1}_{\mathscr{K}_\kappa}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))\mathrm{d}s \\ &\ge q_1(0) + \nu\int_0^t \delta_1(s)\mathrm{d}s - (\nu + 1)\int_0^t \mathbf{1}_{\mathscr{K}_\kappa^c}(\boldsymbol{q}(s), \boldsymbol{\delta}(s))\mathrm{d}s.\end{aligned}$$

and, thus, by (15) and (17), there exist positive constants K_2, K_2', which may depend on ε such that, for some K_1', $\forall\, t \geq 0$

$$\int_0^t \delta_1(s)\mathrm{d}s < K_1' \implies \int_0^t \mathbf{1}_{\left[\delta_1(s)>\frac{\varepsilon}{2}\right]}\mathrm{d}s < K_2, \implies \int_0^t \mathbf{1}_{\left[\delta_1(s)>\frac{\varepsilon\nu}{2}\right]}\mathrm{d}s < K_2'. \tag{18}$$

Consequently, because of Proposition 2, because $\delta_1(t) = \delta_1(0) + \xi(t) - \nu\int_0^t \delta_1(s)\mathrm{d}s$, it must be the case that

$$\limsup_{t\to\infty} \xi(t) < \infty. \tag{19}$$

Furthermore, because $q_1(t) \leq 1 - \varepsilon\nu$ for all $t \geq 0$,

$$\mathbf{1}_{[\delta_0(t)=0]} \leq \mathbf{1}_{[u(t)>0]} + \mathbf{1}_{\left[\delta_1(t)\geq\frac{\varepsilon\nu}{2}\right]}.$$

Thus, (17) and (18) yield $\forall\, t \geq 0$,

$$\int_0^t \mathbf{1}_{[\delta_0(s)=0]}\mathrm{d}s \leq K_1 + K_2'. \tag{20}$$

Now, from Proposition 2, observe that

$$\begin{aligned}\xi(t) &= \int_0^t \lambda(1 - p_0(\boldsymbol{q}(s), \boldsymbol{\delta}(s), \lambda))\mathbf{1}_{[\delta_0(s)>0]}\mathrm{d}s \\ &\geq \int_0^t \lambda(1 - p_0(\boldsymbol{q}(s), \boldsymbol{\delta}(s), \lambda))\mathbf{1}_{[\delta_0(s)>0, u(s)=0, \delta_1(s)\leq\varepsilon/2]}\mathrm{d}s,\end{aligned}$$

and on the set $\{s : \delta_0(s) > 0, u(s) = 0, \delta_1(s) \leq \varepsilon/2\}$, we have $p_0(\boldsymbol{q}(s), \boldsymbol{\delta}(s), \lambda) \leq \lambda^{-1}(\varepsilon\nu/2 + q_1(s) - \kappa)$. Therefore,

$$\begin{aligned}\xi(t) &\geq \int_0^t \left(\lambda - \frac{\varepsilon\nu}{2} - q_1(s) + \kappa\right)\mathbf{1}_{[\delta_0(s)>0, u(s)=0, \delta_1(s)\leq\varepsilon/2]}\mathrm{d}s \\ &\geq \int_0^t \left(\lambda - \frac{\varepsilon\nu}{2} - q_1(s) + \kappa\right)\mathrm{d}s - \int_0^t \mathbf{1}_{[\delta_0(s)=0]}\mathrm{d}s - \int_0^t \mathbf{1}_{[u(s)>0]}\mathrm{d}s - \int_0^t \mathbf{1}_{[\delta_1(s)>\varepsilon/2]}\mathrm{d}s,\end{aligned} \tag{21}$$

where the second inequality is because $q_1(s) \geq \kappa$ and $\varepsilon\nu/2 > 0$, and hence, $\lambda - \varepsilon\nu/2 - q_1(s) + \kappa \leq mbda < 1$. Therefore, because $\lambda + \kappa \geq 1$, we have $\lambda - \varepsilon\nu/2 - q_1(s) + \kappa \geq \varepsilon\nu/2 > 0$, and Equations (17), (18), (20), and (21) imply $\liminf_{t\to\infty} \xi(t) = \infty$, which is a contradiction with (19). This completes the proof of Claim 8.

5.2.2. Proof of Statement 2. First, we establish convergence of $q_1^N(\infty)$ as $N \to \infty$, followed by convergence of $q_2^N(\infty)$, $\delta_0^N(\infty)$, and $\delta_1^N(\infty)$.

Convergence of $q_1^N(\infty)$. First, we show that, for all $\varepsilon_2 > 0$,

$$\limsup_{N\to\infty} P(q_1^N(\infty) < \kappa + \lambda - \varepsilon_2) = 0. \tag{22}$$

Because of Lemma 1, note that any limit of stationary distributions is such that, with probability one, $q_1 \geq \varepsilon_1$ for some fixed $\varepsilon_1 > 0$. Therefore, throughout the proof of part 2 of Lemma 2, it is enough to consider MFFSP so that $q_1(0) \geq \varepsilon_1$, and Proposition 2 and Claim 7 can be used. Thus, for (22), it is enough to show that any MFFSP has the following property.

Claim 9. Starting from any state $\boldsymbol{q}(0) \in E_\kappa$ with $q_1(0) \in [\varepsilon_1, \kappa + \lambda)$ along any MFFSP, we have

$$\liminf_{t\to\infty} q_1(t) \geq \kappa + \lambda. \tag{23}$$

Similar arguments as in the proof of Claim 8 can be used to prove Claim 9, for which we omit the details. Claim 9 then implies (22).

Further, observe that, because we have assumed that the system is stable, we have

$$\lim_{N\to\infty} \mathbb{E}(q_1^N(\infty)) \leq \kappa + \lambda. \tag{24}$$

Fix any $\varepsilon_2' > 0$. Now, for all fixed $M > 0$,

$$\begin{aligned}\mathbb{E}(q_1^N(\infty)) - (\kappa+\lambda) &\geq \varepsilon_2' \mathbb{P}(q_1^N(\infty) > \kappa+\lambda+\varepsilon_2') - \frac{\varepsilon_2'}{M}\mathbb{P}\left(\kappa+\lambda-\frac{\varepsilon_2'}{M} \leq q_1^N(\infty) \leq \kappa+\lambda+\varepsilon_2'\right) \\ &\quad - \mathbb{P}\left(q_1^N(\infty) < \kappa+\lambda-\frac{\varepsilon_2'}{M}\right), \\ &\geq \varepsilon_2' \mathbb{P}(q_1^N(\infty) > \kappa+\lambda+\varepsilon_2') - \frac{\varepsilon_2'}{M} - \mathbb{P}\left(q_1^N(\infty) < \kappa+\lambda-\frac{\varepsilon_2'}{M}\right),\end{aligned}$$

and thus, from (22) and (24),

$$\limsup_{N\to\infty} \mathbb{P}(q_1^N(\infty) > \kappa+\lambda+\varepsilon_2') \leq \frac{1}{M} \quad \text{for all } M > 0,$$

which, in conjunction with (22), completes the proof of convergence of $q_1^N(\infty)$.

Convergence of $q_2^N(\infty)$. Note that, given the convergence of $q_1^N(\infty)$ to $\kappa+\lambda$ as $N \to \infty$, the following property of the mean-field limit is sufficient to prove that the sequence of stationary distributions $q_2^N(\infty)$ concentrate at $q_2^\star = \kappa$ as $N \to \infty$:

Claim 10. For any $\varepsilon_3 > 0$, there exists a fixed T_0 and $\varepsilon_4 > 0$ such that, starting from any state $\boldsymbol{q}(0) \in E_\kappa$ with $q_1(0) = \lambda+\kappa$ and $q_2(0) > \kappa+\varepsilon_3$ along any MFFSP, we have $q_1(T_0) \geq \kappa+\lambda+\varepsilon_4$.

Indeed, if the sequence of the stationary distributions were such that

$$\limsup_{N\to\infty} \mathbb{P}\big(q_2^N(\infty) > \kappa+\varepsilon_3\big) > 0,$$

then Claim 10 would imply that $\limsup_{N\to\infty} \mathbb{P}\big(q_1^N(\infty) > \kappa+\lambda+\varepsilon_4/2\big) > 0$, which contradicts the convergence of $q_1^N(\infty)$.

Proof of Claim 10. We prove by contradiction. Note that because $q_1(0) = \lambda+\kappa$ and $q_2(0) > \kappa+\varepsilon_3$ and the rates of change are bounded in a sufficiently small neighborhood $[0, T_0]$ (depending only on ε_3), we have, for all $t \in [0, T_0]$, (i) $q_1(t) \leq \lambda+\kappa+\varepsilon_3/2$, (ii) $q_2(t) \geq \kappa+\varepsilon_3/2$, and

$$\text{(iii) } y_1(t) = q_1(t) - q_2(t) \leq \lambda - \frac{\varepsilon_3}{2}.$$

Because of Claim 7, $q_1(t)$ is nondecreasing in $[0, T_0]$, and it is enough to produce a subinterval of $[0, T_0]$, where the right derivative of $q_1(t)$ is bounded away from zero. Now we consider two cases:

5.2.3. Case 1. There exists $t' \in [0, T_0/2]$ such that $u(t') = 0$ and $\delta_1(t') \leq \varepsilon_3/2$. In this case, $\delta_0(t') > 0$, and in a sufficiently small time interval, almost all points (with respect to Lebesgue measure) are regular for $\delta_1(t)$. Also, because of Proposition 2, because, for $t \geq t'$,

$$\delta_1(t) = \delta_1(t') + \xi(t) - \xi(t') - \nu\int_{t'}^{t} \delta_1(s)\mathrm{d}s,$$

with $(d^+/dt)\xi(t) = \lambda - y_1(t) \geq \varepsilon_3/2$ at $t = t'$, we have, for sufficiently small $t_1 < T_0/4$ (where choice of t_1 does not depend on t'), $\delta_1(t'+t_1) \geq t_1\varepsilon_3/4$. Also, because the rate of decrease of $\delta_1(t)$ is bounded, there exists $t_2 < T_0/4$ (where choice of t_2 does not depend on t' as well) such that,

$$\delta_1(t) \geq \frac{t_1\varepsilon_3}{8} \quad \text{for all} \quad t \in [t'+t_1, t'+t_1+t_2] \subseteq [0, T_0].$$

Thus, because of Claim 7, there exists $\alpha > 0$ such that, during the time interval $[t'+t_1, t'+t_1+t_2]$, $(d^+/dt)q_1(t) \geq \min\{\alpha, \nu t_1\varepsilon_3/8\}$. Consequently,

$$q_1(T_0) \geq q_1(t'+t_1+t_2) \geq \lambda+\kappa+\min\left\{\alpha, \frac{\nu t_1\varepsilon_3}{8}\right\}t_2. \tag{25}$$

It is important to note that the choices of t_1 and t_2 depend only on ε_3 and not on t'.

5.2.4. Case 2. For all $t \in [0, T_0/2]$, either $u(t) > 0$ or $\delta_1(t) > \varepsilon_3/2$. In this case, because of Claims 7i and 7ii, there exists $\alpha > 0$ such that $(d^+/dt)q_1(t) > \alpha$ for all $t \in [0, T_0/2]$. Also, because $q_2(t)$ is nondecreasing in $[0, T_0]$, we obtain

$$q_1(T_0) \geq q_1\left(\frac{T_0}{2}\right) \geq \lambda + \kappa + \frac{\alpha T_0}{2}. \tag{26}$$

Combining the two cases and choosing

$$\varepsilon_4 = \min\left\{\min\left\{\alpha, \frac{\nu t_1 \varepsilon_3}{8}\right\} t_2, \frac{\alpha T_0}{2}\right\} > 0$$

completes the proof of Claim 10.

Convergence of $\delta_1^N(\infty)$ and $\delta_0^N(\infty)$. Given the convergence of $q_1^N(\infty)$ and $q_2^N(\infty)$, the convergence of $\delta_1^N(\infty)$ and $\delta_0^N(\infty)$ can be seen immediately by observing that the mean-field limit has the following property:

Claim 11. Starting from any state $\boldsymbol{q}(0) \in E_\kappa$ with $q_1(0) = \lambda + \kappa$ and $q_2(0) = \kappa$ along any MFFSP, $\delta_1(t) \to 0$ and $\delta_0(t) \to 1 - \lambda - \kappa$ as $t \to \infty$.

The proof of Claim 11 is immediate from the description of the mean-field limit as in Proposition 2 and, hence, is omitted. This completes the proof of Lemma 2.

6. Conclusion

In this paper, we studied the stability of systems under the TABS scheme and established large-scale asymptotics of the sequence of steady states. Understanding stability of stochastic systems is of fundamental importance. Systems under the TABS scheme, as it turned out, may be unstable for some N even under a subcritical load assumption. As in many other cases, the lack of monotonicity makes the stability analysis much more challenging from a methodological standpoint. We developed a novel induction-based method and establish that the TABS scheme is stable for all large enough N. The proof technique is of independent interest and potentially has a much broader applicability. The key model-dependent part of our method is what can be called a weak monotonicity property, which ensures that, for large enough N with high probability, no matter where the system starts, in some fixed amount of time, there will be a certain fraction of busy servers. Both traditional fluid limit (fixed N, initial state goes to infinity) and mean-field limit (for a sequence of processes with the number of queues $N \to \infty$) were used in an intricate manner to establish the results.

References

Aghajani R, Ramanan K (2017) The hydrodynamic limit of a randomized load balancing network. Working paper, University of California, San Diego, San Diego.

Andrew LLH, Lin M, Wierman A (2010) Optimality, fairness, and robustness in speed scaling designs. *ACM SIGMETRICS Performance Evaluation Rev.* 38(1):37–48.

Bramson M, Lu Y, Prabhakar B (2012) Asymptotic independence of queues under randomized load balancing. *Queueing Systems* 71(3):247–292.

Bramson M, Lu Y, Prabhakar B (2013) Decay of tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Probab.* 23(5):1841–1878.

Dai JG (1995) On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* 5(1): 49–77.

Eschenfeldt P, Gamarnik D (2016) Supermarket queueing system in the heavy traffic regime. Short queue dynamics. Working paper, Massachusetts General Hospital, Boston.

Foss SG, Stolyar AL (2017) Large-scale join-idle-queue system with general service times. *J. Appl. Probab.* 54(4):995–1007.

Gandhi A, Doroudi S, Harchol-Balter M, Scheller-Wolf A (2014) Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Systems* 77(2):177–209.

Liggett TM (1985) *Interacting Particle Systems* (Springer, New York).

Lin M, Liu Z, Wierman A, Andrew LLH (2012) Online algorithms for geographical load balancing. *Internat. Green Comput. Conf.* (IEEE, Washington, DC), 1–10.

Lin M, Wierman A, Andrew LLH, Thereska E (2013) Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Trans. Networking* 21(5):1378–1391.

Liu Z, Lin M, Wierman A, Low SH, Andrew LLH (2011a) Geographical load balancing with renewables. *ACM SIGMETRICS Performance Evaluation Rev.* 39(3):62–66.

Liu Z, Lin M, Wierman A, Low SH, Andrew LLH (2011b) Greening geographical load balancing. *Proc. SIGMETRICS '11* (ACM, New York), 233–244.

Liu Z, Chen Y, Bash C, Wierman A, Gmach D, Wang Z, Marwah M, Hyser C (2012) Renewable and cooling aware workload management for sustainable data centers. *ACM SIGMETRICS Performance Evaluation Rev.* 40(1):175–186.

Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.

Meyn SP, Tweedie RL (1993) *Markov Chains and Stochastic Stability* (Springer, London).

Mitzenmacher M (2001) The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distribution Systems* 12(10):1094–1104.

Mukherjee D, Borst SC, Van Leeuwaarden JSH, Whiting PA (2016) Universality of load balancing schemes on the diffusion scale. *J. Appl. Probab.* 53(4):1111–1124.

Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2018) Universality of power-of-d load balancing in many-server systems. *Stochastic Systems* 8(4):265–292.

Mukherjee D, Dhara S, Borst SC, Van Leeuwaarden JSH (2017) Optimal service elasticity in large-scale distributed systems. *Proc. ACM Measurement Anal. Comput. Systems* 1(1):3–3.

Nguyen LM, Stolyar AL (2016) A service system with randomly behaving on-demand Aagents. *ACM SIGMETRICS Performance Evaluation Rev.* 44(1):365–366.

Pang G, Stolyar AL (2016) A service system with on-demand agent invitations. *Queueing Systems* 82(3–4):259–283.

Pender J, Phung-Duc T (2016) A law of large numbers for M/M/c/delayoff-setup queues with nonstationary arrivals. Wittevrongel S, Phung-Duc T, eds. *Proc. ASMTA '16*, Lecture Notes in Computer Science, vol. 9845 (Springer, Cham, Switzerland), 253–268.

Rybko AN, Stolyar AL (1992) On the ergodicity of stochastic processes describing the operation of open queueing networks. *Problems Inform.Transmission* 28(3):199–200.

Shneer S, Stolyar A (2018) Stability conditions for a discrete-time decentralised medium access algorithm. *Ann. Appl. Probab.* 28(6):3600–3628.

Stolyar AL (1995) On the stability of multiclass queueing networks: A relaxed sufficient condition via limiting fluid processes. *Markov Processes Related Fields* 1(4):491–512.

Stolyar AL (2015) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80(4):341–361.

Stolyar AL (2017) Pull-based load distribution among heterogeneous parallel servers: the case of multiple routers. *Queueing Systems* 85(1):31–65.

Urgaonkar R, Kozat UC, Igarashi K, Neely MJ (2010) Dynamic resource allocation and power management in virtualized data centers. *Proc. IEEE Network Oper. Management Sympos.* (*NOMS*) (IEEE, Washington, DC), 479–486.

Van der Boor M, Borst SC, van Leeuwaarden JSH, Mukherjee D (2018) Scalable load balancing in networked systems: Universality properties and stochastic coupling methods. Sirakov B, Ney de Souza P, Viana M, eds. *Proc. Internat. Congress Mathematicians* (World Scientific, Singapore), 3881–3912.

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32(1):20–34.

Wierman A, Andrew LLH, Tang A (2012) Power-aware speed scaling in processor sharing systems: optimality and robustness. *Performance Evaluation* 69(12):601–622.

Ying L (2017) Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Measurement Anal. Comput. Systems* 1(1):Article 12.