Join-idle-queue with service elasticity: large-scale asymptotics of a non-monotone system

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

Debankur Mukherjee[*1] and Alexander Stolyar[2]

[1]*Eindhoven University of Technology, The Netherlands*
[2]*University of Illinois at Urbana-Champaign*

March 22, 2018

Abstract

We consider the model of a token-based joint auto-scaling and load balancing strategy, proposed in a recent paper by Mukherjee, Dhara, Borst, and van Leeuwaarden [20] (SIGMETRICS '17), which offers an efficient scalable implementation and yet achieves asymptotically optimal steady-state delay performance and energy consumption as the number of servers N → ∞. In the above 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 → ∞. 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.

1 Introduction

Background and motivation. Load balancing and auto-scaling are two principal pillars in modern-day data centers and cloud networks, and therefore, have gained renewed interest in past two decades. In its basic setup, a large-scale system consists of a pool of large number of servers and a single dispatcher, where 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

---

[*]Email: `d.mukherjee@tue.nl`

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 up 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.

Load balancing in large systems. The study of load balancing schemes in large-scale systems have a very rich history, and for decades, a lot of research have been conducted in understanding the fundamental trade-off between delay-performance and communication overhead per task. The so-called 'power-of-d' schemes, where each arrival is assigned to the shortest among d randomly chosen queues, provide surprising improvements for $d \geqslant 2$ over purely random routing ($d = 1$) while maintaining the communication overhead as low as d per task. This scheme along with its many variations have been studied extensively in [1, 3, 4, 6, 17, 19, 31, 33] and many more. Relatively recently, join-the-idle queue (JIQ) scheme was proposed in [15], where 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 in [27, 28] show that under Markovian assumptions, the JIQ policy achieves a zero probability of wait for any fixed subcritical load per server in a regime where the total number of servers grows large. It should be noted that the results in [27, 28] even hold for considerably more general scenarios, viz. decreasing hazard rate service time distributions and heterogeneous servers pools. Recently, it is further shown [7] 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 in [18] indicate that under Markovian assumptions, the JIQ policy has the same diffusion-limit as the Join-the-Shortest-Queue (JSQ) 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 [30] for a recent survey on load balancing schemes.

Auto-scaling with a centralized queue. Queue-driven auto-scaling techniques have been widely investigated in the literature [2, 8, 10–14, 23, 29, 32]. In systems with a centralized queue it is very common to put servers to 'sleep' while the demand is low, since 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 which examine on-demand server addition/removal in a somewhat different vein [21, 22]. Generalizations towards

non-stationary arrivals and impatience effects have also been considered recently [23]. Unfortunately, data centers and cloud networks with millions of servers are too complex to maintain any centralized queue, and it involves prohibitively high communication burden to obtain instantaneous system information even for a small fraction of servers.

Joint load balancing and auto-scaling in distributed systems. Motivated by all the above, a token-based joint load balancing and auto scaling scheme called TABS was proposed in [20], that 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 [20], the authors left open a fundamental question: Is the system with a given number N of servers stable under TABS scheme? The analysis in [20] 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.

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, i.e., the system with finite number N of servers under the usual subcritical load assumption is not automatic. In fact, as we will further discuss in Remark 1 below 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 [3, 27, 28, 31]. 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; see e.g. [9]. The lack of monotonicity immediately complicates the situation, as for example in [7, 25]. 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 $\lambda > 1/4$ [3], and the JIQ scheme when $\lambda > 1/2$ [7]. 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 below. 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, where 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,* $0 \leqslant k \leqslant N$, *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 the above statement. For a fixed N, first we introduce the notion of a fluid sample path (FSP) for systems where 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 [5, 24, 26]) in order to establish stability for the system where the number of infinite queues is $k - 1$. As mentioned above, here the notion of the traditional FSP is needed to be suitably extended to fit to the systems where some servers have infinite queue lengths. Loosely speaking, for the fluid-stability, the 'large queues' behave as 'infinite queues' for which 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 roadmap of the above proof argument is presented in Subsection 4.1. This technique is of independent interest, and potentially has a much broader applicability in proving large-N stability for non-monotone systems, where the state-of-the-art results have remained scarce so far.

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 till Section 3. Section 4 introduces an inductive approach to prove 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 will 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 where 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., FCFS). 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 auto-scaling and load balancing
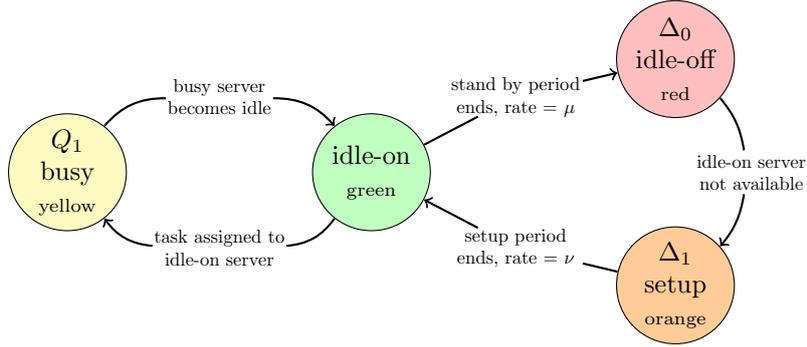
Figure 1: Illustration of server on-off decision rules in the TABS scheme, along with message colors and state variables as given in [20].

scheme called TABS (Token-based Auto Balance Scaling), as introduced in [20].

TABS scheme [20].

- When a server becomes idle, it sends a 'green' message to the dispatcher, waits for an $\text{Exp}(\mu)$ 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 $\text{Exp}(\nu)$ time after which the server becomes active.

- Any server which activates due to the latter event, sends a green message to the dispatcher (the corresponding orange message is replaced), waits for an $\text{Exp}(\mu)$ time for a possible assignment of a task, and again turns itself off by sending a red message to the dispatcher.

As described in [20], the TABS scheme gives rise to a distributed operation in which servers are in one of four states (busy, idle-on, idle-off, or standby), and advertize 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.

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 $(\mathbf{Q}^N(t), \Delta_0^N(t), \Delta_1^N(t))_{t \geqslant 0}$ 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 will focus upon an asymptotic analysis, where 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, viz. $q_i^N(t) := Q_i^N(t)/N$, $\delta_0^N(t) = \Delta_0^N(t)/N$, $\delta_1^N(t) = \Delta_1^N(t)/N$, and $u^N(t) := U^N(t)/N$. Notation for the *conventional fluid-scaled* occupancy states for a fixed N will be introduced later in Subsection 3.1. For brevity in notation, we will 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

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

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 $\mathcal{E}$, 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. By the symbol '$\xrightarrow{\mathbb{P}}$' we denote convergence in probability for real-valued random variables.

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

Theorem 2.1. *For any fixed* μ, $\nu > 0$, *and* $\lambda < 1$, *the system with* N *servers under the TABS scheme is stable (positive recurrent) for large enough* N.

Theorem 2.1 is proved in Section 3.

Remark 1. It is worthwhile to mention that the 'large-N' stability as stated in Theorem 2.1 above 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 2 servers A and B, and $1/2 < \lambda < 1$. Let server A start with a large queue, while the initial queue length at server B be small. In that case, observe that every time the queue length at server B hits 0, 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 $2\lambda/\nu$, while 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 above scenario. The situation becomes better as N increases. Indeed for large N, if 'too
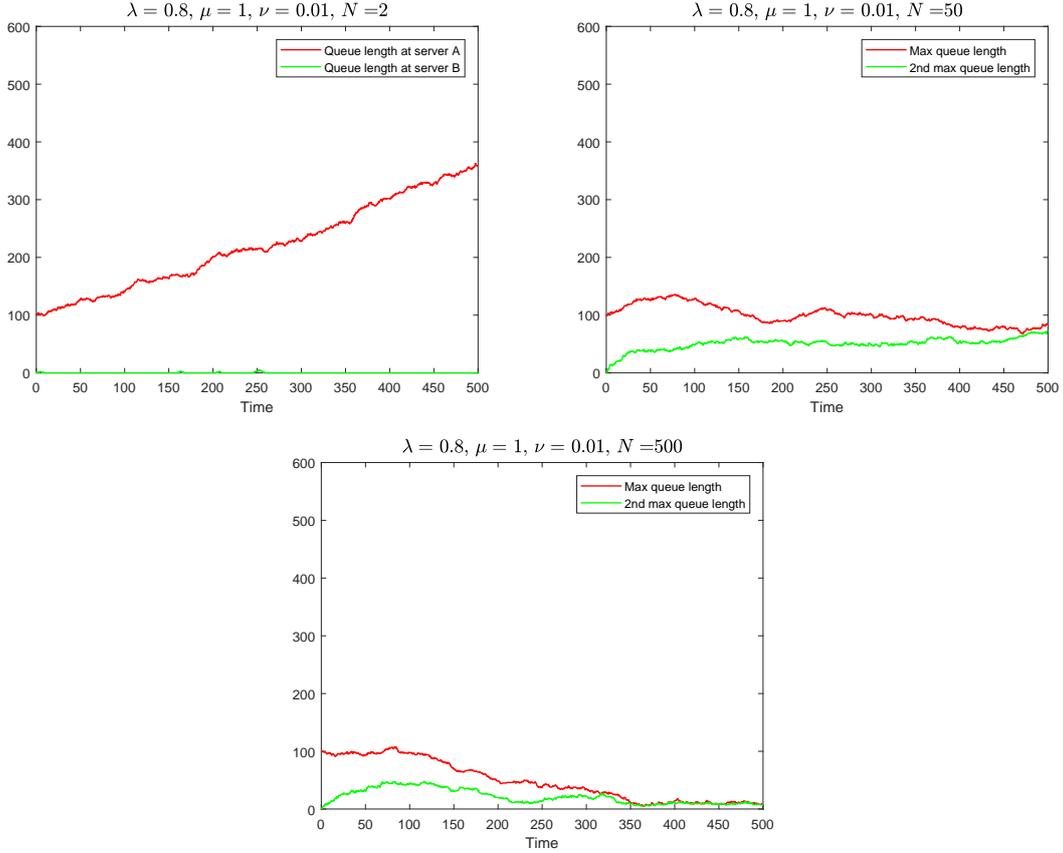
Figure 2: (Top left) Illustration of instability of the TABS scheme for N = 2 via sample paths of the queue length process. (Top right) Sample paths of the maximum and second maximum queue length processes in an intermediate system (N = 50) for the same parameter choices. (Bottom) The system becomes stable for a large enough system (N = 500).

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. The above heuristic has been illustrated in Figure 2 with examples of three scenarios with small, moderate, and large values of N, respectively.

In the next theorem we will identify the limit of the sequence of stationary distributions of the occupancy processes as $N \to \infty$. In particular, we will establish that under sub-critical load, for any fixed μ, $\nu > 0$, the steady-state occupancy process converges weakly to the unique fixed point. (For the finite buffer scenario this was proved in [20, Proposition 3.3].) Denote by $\mathbf{q}^N(\infty)$ and $\boldsymbol{\delta}^N(\infty)$ the random values of $\mathbf{q}^N(t)$ and $\boldsymbol{\delta}^N(t)$ in the steady-state, respectively.

Theorem 2.2. *For any fixed μ, $\nu > 0$, and $\lambda < 1$, the sequence of steady states $(\mathbf{q}^N(\infty), \boldsymbol{\delta}^N(\infty))$*

converges weakly to the fixed point $(\mathbf{q}^\star, \boldsymbol{\delta}^\star)$ *as* $N \to \infty$, *where*

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

Note that the fixed point $(\mathbf{q}^\star, \boldsymbol{\delta}^\star)$ is such that the probability of wait vanishes as $N \to \infty$ and the asymptotic fraction of active servers is minimum possible, and in this sense, the fixed point is optimal. Thus, Theorem 2.2 implies that the TABS scheme provides fluid-level optimality for large-scale systems in terms of delay performance and resource utilization, while involving only $O(1)$ communication overhead per task.

3 Proofs of the main results

In Subsection 3.1 we introduce the notion of conventional fluid scaling (when the number of servers is fixed) and fluid sample paths (FSP), and state Proposition 3.1 that implies Theorem 2.1 as an immediate corollary. Subsection 3.2 contains two key results for sequence of systems with increasing system size, i.e., number of servers $N \to \infty$, and proves Theorem 2.2.

3.1 Conventional fluid limit for a system with fixed N

In this subsection first we will introduce a notion of fluid sample path (FSP) for finite-N systems where some of the queue lengths are infinite. We emphasize that this is *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 1. We now introduce a rigorous notion of FSP for systems with some of the queues being infinite.

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 $\mathcal{K} \subseteq \mathcal{N}$ have infinite queue lengths. Now consider any sequence of systems indexed by R such that $\sum_{i \in \mathcal{N} \setminus \mathcal{K}} X_i^{N,R}(0) < \infty$, and

$$x_i^{N,R}(t) := \frac{X_i^{N,R}(Rt)}{R}, \quad i \in \mathcal{N} \setminus \mathcal{K} \tag{3.1}$$

be the corresponding scaled processes. For fixed N, the scaling in (3.1) will henceforth be called as the *conventional fluid-scaled* queue length process. Also, for the R-th 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 will 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, i.e., $x^R(0) \to x(0)$, for some fixed $x(0)$ such that $0 \leqslant \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) < \infty$ and $x_i(0) = \infty$ for $i \in \mathcal{K}$. 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 $i \in \mathcal{K}$, is called a *fluid sample path* (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 1, 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 $i \in \mathcal{N} \setminus \mathcal{K}$, $x_i^R(\cdot) \to x_i(\cdot)$ u.o.c.

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

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

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 3.1. *The following holds for all sufficiently large N. For each $0 \leqslant k \leqslant N$, consider a system where k servers with indices in $\mathcal{K}$ 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 below, also depend on N).*

(1) *For any $\mathbf{x}(0)$ such that $0 \leqslant \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) < \infty$ and $x_i(0) = \infty$ for $i \in \mathcal{K}$, there exists $T(k, \mathbf{x}(0)) < \infty$ and a unique FSP on the interval $[0, T(k, \mathbf{x}(0))]$, which has the following properties:*

   (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 $i \in \mathcal{N} \setminus \mathcal{K}$, if $x_i(t_0) = 0$ for some t_0, then $x_i(t) = 0$ for all $t \geqslant t_0$.*

   (iii) *$T(k, \mathbf{x}(0)) = \inf \{t : x_i(t) = 0 \text{ for all } i \in \mathcal{N} \setminus \mathcal{K}\}$.*

(2) *The subsystem with $N - k$ finite queues is stable.*

9

(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 at most $1 - \varepsilon(k)$.*

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

Although Part 2 follows from Part 1, and Part 4 is stronger than Part 1, the statement of Proposition 3.1 is arranged as it is to facilitate its proof, as we will see in Section 4 in detail.

Proof of Theorem 2.1. Note that Theorem 2.1 is a special case of Proposition 3.1 when $k = 0$. $\qquad \square$

3.2 Large-scale asymptotics: auxiliary results

In this subsection we will state two crucial lemmas that describe asymptotic properties of 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 3.2. *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,*

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

Lemma 3.3. *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 $\kappa \geqslant 1 - \lambda$, 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 $(\mathbf{q}^N(\infty), \boldsymbol{\delta}^N(\infty))$ is the distribution concentrated at the unique equilibrium point $(\mathbf{q}^\star(\kappa), \boldsymbol{\delta}^\star(\kappa))$, such that*

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

Consequently,
$$\lim_{N \to \infty} \mathbb{P}\left(Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N \right) = 0. \tag{3.2}$$

Lemmas 3.2 and 3.3 are proved in Section 5. These results will be 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 2. It is also worthwhile to note that Lemmas 3.2 and 3.3 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 3.2 and 3.3 is to bound *under the assumption of stability*, the expected rate at which task arrives to the infinite queues when the subsystem containing the finite queues is in steady-state: In this regard

(i) Lemma 3.3 guarantees high probability bounds on the total number of busy servers, so that with probability tending to 1 as $N \to \infty$, the fraction of busy servers in the whole system is at least λ in steady state.

(ii) But note that since 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 3.2 in order to obtain bound on the expected rate of arrivals to the infinite queues.

In Subsection 4.4 we will see that as a consequence of Lemmas 3.2 and 3.3, 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 1, and the drift of the infinite queues as defined in Subsection 3.1 becomes strictly negative. This fact will be used in the proof of Proposition 3.1.

Proof of Theorem 2.2. Note that given the large-N stability property proved in Proposition 3.1 for $k(N) = 0$, and the convergence of stationary distributions under the assumption of stability in Lemma 3.3, the proof of Theorem 2.2 is immediate. □

4 Proof of Proposition 3.1: An inductive approach

Throughout this section we will prove Proposition 3.1. The proof consists of several steps and uses both conventional fluid limit and mean-field fluid scaling and limit in an intricate fashion. Below we first provide a roadmap of the whole proof argument.

4.1 Proof idea and the roadmap

The key idea for the proof of Proposition 3.1 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 TABS scheme does not differentiate between two large queues (in fact, any two non-empty queues). Thus, when all queues are infinite, since 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 1. 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, i.e., assume that Parts (1)–(4) hold for $k \geqslant k(N) + 1$ 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 Subsections 4.2–4.5. We begin by providing a roadmap of these four subsections.

Part (1). Recall that we denote by $\mathcal{K}$ the indices of the servers having infinite queue lengths, and by $\mathcal{N}$ the set of all server indices. Denote by $x_{(i)}$ the i-th 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

$$T(m, \mathbf{x}) := \frac{x_{(N)}}{\varepsilon(N)} + \sum_{i=1}^{N-m-1} \frac{x_{(N-i)} - x_{(N-i+1)}}{\varepsilon(N-i)} \tag{4.1}$$

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 (4.1). Indeed, recall that we are at the backward induction step where 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 $j \in \mathcal{N} \setminus \mathcal{K}$ is positive, it can be coupled with a system where the queue length at server j is infinite. Thus, as long as there is at least one server $j \in \mathcal{N} \setminus \mathcal{K}$ 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 (4.1), 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 0. 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 0. Proceeding this way, one can see that at time $T(m, \mathbf{x}(0))$ all finite positive components of the FSP hit 0. The above argument is formalized in Subsection 4.2.

Part (1) $\implies$ Part (2). To prove Part 2, we will use the fluid limit technique of proving stochastic stability as in [5, 24, 26], see for example [5, Theorem 4.2] or [26, Theorem 7.2] for a rigorous statement. Here we need to show that the sum of the non-infinite queues (of an FSP) drains to 0. This is true, because by Part (1) each positive non-infinite queue will have negative drift. The formal proof is in Subsection 4.3.

Part (2) + Lemmas 3.2 and 3.3 $\implies$ 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 3.2 and 3.3. For that reason, in the statement of Proposition 3.1 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 1, along the subsequence. Loosely speaking, in that case, Lemmas 3.2 and 3.3 together imply that for large enough N, there are 'enough' busy servers, so that the rate of arrival to each infinite queues is strictly smaller than 1, which leads to a contradiction. Note that we can apply Lemmas 3.2 and 3.3 here, because Part (2) ensures the required stability. The rigorous proof is in Subsection 4.4.

Parts (2), (3) + Time-scale separation $\implies$ Part (4). We assume that Parts (1) – (3) hold for $k \in \{k(N), k(N) + 1, \ldots, N\}$, and we will 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 $i \in \mathcal{N} \setminus \mathcal{K}$. 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 Subsection 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 $i \in \mathcal{N} \setminus \mathcal{K}$, then due to Part (2), for the subsystem consisting of finite queues, the (scaled) hitting time to the (unscaled) state z converges in probability to 0. Also, since this subsystem is positive recurrent (due to Part (2)), starting from a fixed (unscaled) state z, its expected (unscaled) return time to the state z is $O(1)$. This will allow us to split the (unscaled) time line into 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 the above observation we can claim that the number of arrivals to any specific infinite queue can be written as a sum of arrivals in the above 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 an 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).

The above argument is rigorously carried out in Subsection 4.5.

4.2 Coupling with infinite queues to verify Part (1)

To prove Part (1), fix any $\mathbf{x}(0)$ such that $0 \leqslant \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) < \infty$ and $x_i(0) = \infty$ for $i \in \mathcal{K}$. Let $\mathcal{K}_1 \subseteq \mathcal{N} \setminus \mathcal{K}$ be the set of server-indices i, such that $x_i(0) > 0$. We will first show that when $\sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) > 0$ with $|\mathcal{M}(t)| = m \geqslant k(N) + 1$, then it has a negative drift $-\varepsilon(m)$ for all $i \in \mathcal{M}(t)$, thus proving Part (1.i). Since $\varepsilon(m)$'s are positive, this will then also imply Part (1.ii). Now assume $\sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) > 0$. In that case we have that $|\mathcal{K}_1| =: k_1 > 0$. Now 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
$$T^R := \inf\left\{t : X_i^R(t) = 0 \text{ for some } i \in \mathcal{K}_1\right\},$$

and $\tau^R = T^R/R$. In the time interval $[0, T^R]$, we will couple this system with a system, let us label it Π, with $k + k_1$ infinite queues. 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 $i \in \mathcal{K} \cup \mathcal{K}_1$. Now couple each arrival to and departure from i-th server in both systems, $i \in \mathcal{N}$. Since the scheme does not distinguish among servers with positive queue lengths, observe that up to time T^R both systems evolve according to their own statistical laws. Also, up to time T^R, the queue length processes at the servers in $\mathcal{N} \setminus (\mathcal{K} \cup \mathcal{K}_1)$ in both systems are identical. Thus, in the (scaled) time interval $[0, \tau^R]$, $a_i^R \equiv \bar{a}_i^R$ and $d_i^R \equiv \bar{d}_i^R$ for all $i \in \mathcal{N}$, and $x_i^R \equiv \bar{x}_i^R$ for all $i \in \mathcal{N} \setminus \mathcal{K}$. Therefore, using induction hypothesis for systems with $k + k_1 \geqslant k(N) + 1$ infinite queues, there exists a subsequence $\{R\}$ along which with probability 1,
$$(\bar{x}_i^R(\cdot), \bar{a}_i^R(\cdot), \bar{d}_i^R(\cdot))_{i \in \mathcal{N}} \to (\bar{x}_i(\cdot), \bar{a}_i(\cdot), \bar{d}_i(\cdot))_{i \in \mathcal{N}},$$

where $\bar{x}_i \equiv 0$ for all $i \in \mathcal{N} \setminus (\mathcal{K} \cup \mathcal{K}_1)$, and $\bar{x}_j \equiv \infty$ with $\bar{x}_j' \equiv -\varepsilon(k + k_1) < 0$ for all $j \in \mathcal{K} \cup \mathcal{K}_1$. Consequently, in the time interval $[0, \tau]$, along that subsequence with probability 1,
$$(x_i^R(\cdot), a_i^R(\cdot), d_i^R(\cdot))_{i \in \mathcal{N}} \to (x_i(\cdot), \bar{a}_i(\cdot), \bar{d}_i(\cdot))_{i \in \mathcal{N}}$$

with $x_i = \bar{x}_i \equiv 0$ for all $i \in \mathcal{N} \setminus (\mathcal{K} \cup \mathcal{K}_1)$ and $x_i' \equiv -\varepsilon(k + k_1) < 0$ for all $i \in \mathcal{K} \cup \mathcal{K}_1$, where $\tau = x_{(k+k_1)}/\varepsilon(k + k_1) > 0$. Observe that the above argument can be extended till the time $\sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(t)$ hits zero. Furthermore, following the argument as above, this time is given by $T(k(N), \mathbf{x}(0))$ as given in (4.1). This completes the proof of Part 1 (iii).

4.3 Conventional fluid-limit stability to verify Part (2)

As mentioned earlier, we will use the fluid limit technique of proving stochastic stability as in [5, 24, 26] to prove Part (2). Consider a sequence of initial states with increasing norm R, i.e., $\sum_{i \in \mathcal{N} \setminus \mathcal{K}} X_i^R(0) = R$ and $X_i^R(0) = \infty$ for $i \in \mathcal{K}$. Then from Part (1.iii), we know that for any sequence there exists a further subsequence $\{R\}$ along which with probability 1, the fluid-scaled occupancy process $(x_i^R(\cdot))_{i \in \mathcal{N}}$ converges to the process $(x_i(\cdot))_{i \in \mathcal{N}}$ for which $\sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(t)$ hits 0 in finite time $T(k(N), \mathbf{x}(0))$, and stays at 0 afterwards. This verifies the fluid-limit stability condition in [5, Theorem 4.2] and [26, Theorem 7.2], and thus completes the verification of Part (2).

4.4 Large-scale asymptotics to verify Part (3)

The verification of the backward induction step for Part (3) uses contradiction. Namely, assuming that the induction step for Part (3) does not hold, we will construct a sequence of systems with increasing N, for which we obtain a contradiction using Lemmas 3.2 and 3.3. We note that this is the only part in the proof of Proposition 3.1, where we use the large-scale (i.e., $N \to \infty$) asymptotic results.

Observe that we have already argued in Subsection 4.1 that *for all* N, Parts (1) – (4) hold for $k = N$. Now, if for some N, Part (3) does not hold for some $k(N) \in \{0, 1, \dots, N-1\}$ while Parts (1)–(4) hold for all $k \geqslant k(N) + 1$, then from the proofs of Parts (1) and (2), note that Parts (1) and (2) hold for $k = k(N)$ as well. Consequently, the subsystem with $N - k(N)$ finite queues is stable. Thus we have the following implication.

Implication 1. *Suppose, for infinitely many* N, *the induction step to prove Part (3) of Proposition 3.1 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) The steady-state arrival rate into each infinite queue is at least 1.*

We will now show that Implication 1 leads to a contradiction – this will prove Part (3) of Proposition 3.1. 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 3.3 we will consider two regimes depending on whether $\kappa \geqslant 1 - \lambda$ or not, and arrive at contradictions in both cases. Since all the infinite queues are exchangeable, we will use σ to denote a typical infinite queue.

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

$$\mathbb{E}\Big(\frac{\lambda N}{Q_1^N(\infty)} \mathbb{1}_{[Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N]}\Big) \leqslant \mathbb{E}\Big(\frac{\lambda N}{Q_1^N(\infty)}\Big) = \mathbb{E}\Big(\frac{\lambda}{q_1^N(\infty)}\Big) + o(1). \qquad (4.2)$$

Now observe that for large N, $\lambda/q_1^N(\infty) \leqslant 2\lambda/\kappa$, since $q_1^N(s) \geqslant \kappa/2 > 0$. Further from Lemma 3.3 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,

$$\mathbb{E}\Big(\frac{\lambda N}{Q_1^N(\infty)} \mathbb{1}_{[Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N]}\Big) \leqslant \frac{1 + \lambda}{2} = 1 - \frac{1 - \lambda}{2} < 1, \qquad (4.3)$$

which is a contradiction to Part (ii) of Implication 1.

Case 2. In case $\kappa < 1 - \lambda$, first note that the statement in Part (3) is vacuously satisfied if $k(N) \equiv 0$ for all large enough N. Thus without loss of generality, assume $k(N) > 0$. Fix ε_1 as in Lemma 3.2. In that case (4.2) becomes

$$\mathbb{E}\Big(\frac{\lambda N}{Q_1^N(\infty)} \mathbb{1}_{[Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N]}\Big)$$

$$\leqslant \mathbb{E}\Big(\frac{\lambda N}{Q_1^N(\infty)} \mathbb{1}_{[Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N, \, Q_1^N(\infty) \geqslant \varepsilon_1 N]}\Big) + \lambda N \mathbb{P}\left(Q_1^N(\infty) < \varepsilon_1 N\right)$$

$$\leqslant \frac{\lambda N}{\varepsilon_1 N} \mathbb{P}\left(Q_1^N(\infty) + \Delta_0^N(\infty) + \Delta_1^N(\infty) = N\right) + \lambda N \mathbb{P}\left(Q_1^N(\infty) < \varepsilon_1 N\right).$$

Now, due to Part (2) of Lemma 3.3, we know that

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

15

and furthermore, Lemma 3.2 yields

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

Thus,

$$\mathbb{E}\left(\frac{\lambda N}{Q_1^N(\infty)} \mathbb{1}_{[Q_1^N(\infty)+\Delta_0^N(\infty)+\Delta_1^N(\infty)=N]}\right) \to 0 \quad \text{as} \quad N \to \infty. \tag{4.4}$$

In particular, for large enough N, the expected steady-state arrival rate is bounded away from 1, 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 $k \in \{k(N), k(N) + 1, \ldots, N\}$. Now consider a system containing $k = k(N)$ infinite queues with indices in $\mathcal{K}$, and recall the conventional fluid scaling and FSP from Subsection 3.1. Also, in this subsection whenever we refer to the process $\{\mathbf{X}(t)\}_{t \geqslant 0}$, the components in $\mathcal{K}$ should be taken to be infinite.

For the queue length vector $\mathbf{X}$, define the norm $\|\mathbf{X}\| := \sum_{i \notin \mathcal{K}} X_i$ to be the total number of tasks at the finite queues. Lemmas 4.1 and 4.2 state two hitting time results that will be used in verifying Part (4).

Lemma 4.1. *For any fixed $\gamma \in (0,1)$, there exists $\tau = \tau(\gamma)$ and $C = C(\gamma)$, such that if $\|\mathbf{X}(0)\| = R \geqslant C$, then*

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

Lemma 4.1 says that if the system starts from an initial state where 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 4.1 is fairly straightforward, but is provided below for completeness.

Proof of Lemma 4.1. Consider a sequence of initial states with an increasing norm, i.e., $\mathbf{X}^R(0)$ is such that $\|\mathbf{X}^R(0)\| = R$ where $R^{-1}\mathbf{X}^R(0) \to \mathbf{x}(0)$ as $R \to \infty$. Then from Part 1 we know that as $R \to \infty$, on the time interval $[0, T(m, \mathbf{x}(0))]$ the process $R^{-1}\mathbf{X}^R(Rt)$ converges in probability to the unique deterministic process $\mathbf{x}(t)$ satisfying

$$\sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i'(t) < -c(k(N)) \quad \text{whenever} \quad \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(t) > 0, \tag{4.5}$$

where $c(m) = \min\left\{k\varepsilon(k) : k(N) + 1 \leqslant k \leqslant N\right\} > 0$. We also know that for any $i \in \mathcal{N} \setminus \mathcal{K}$, if $x_i(t_0) = 0$ for some t_0, then $x_i(t) = 0$ for all $t \geqslant t_0$. Consequently, since $c(k(N))$ is positive, there exists $\tau = \tau(\gamma) < \infty$, such that

$$\sup_{\|\mathbf{x}\|=1} \left\{\|\mathbf{x}(\tau)\| : \mathbf{x}(0) = \mathbf{x} \in [0,1]^{N-k(N)} \times \{\infty\}^{k(N)}\right\} < 1-\gamma.$$

16

Now since the expected number of arrivals into the R-th system up to time t, when scaled by R, is λt for any finite t, we obtain $\mathbb{E}(R^{-1}\|\mathbf{X}^R(t)\|) \leqslant 1 + \lambda t$. Therefore, the convergence in probability also implies the convergence in expectation. Thus for the above 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 $R \geqslant C$,

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

This completes the proof of Lemma 4.1. □

For any $C > 0$, define the set $\mathcal{C} := \{\|\mathbf{X}\| \leqslant C\}$, 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 4.2. *There exists $C, C_1 > 0$, such that if $\|\mathbf{X}(0)\| = R \geqslant C$, then*

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

Proof of Lemma 4.2. Fix any $\gamma \in (0,1)$, and take $\tau = \tau(\gamma)$ and $C = C(\gamma)$ as in Lemma 4.1. For $i \geqslant 1$, 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 $\{\Phi_i : i \geqslant 0\}$ adapted to the filtration $\mathcal{F} = \bigcup_{i \geqslant 0} \mathcal{F}_i$, where $\Phi_i = \mathbf{X}(T_i)$ is the value of the continuous time Markov process sample at times T_i's, and $\mathcal{F}_i = \sigma(\Phi_0, \Phi_1, \ldots, \Phi_i)$ is the sigma field generated by $\{\Phi_0, \Phi_1, \ldots, \Phi_i\}$. Further, for $i \geqslant 0$ define the stopping time $\hat{\theta}_C := \inf\{j \geqslant 0 : Z_j \leqslant C\}$. Then observe that

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

Also define $\alpha_i = \sum_{j=1}^i T_j$ for $i \geqslant 1$, and hence $\alpha_{\hat{\theta}_C} = \Psi_C$. Then as a consequence of Dynkin's lemma [16, Theorem 11.3.1], using [16, Proposition 11.3.2] we have

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

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 $T(k(N), \mathbf{x}(0))$. From the verification of Part (1) we already know that $x_i(t) = 0$ for all $t \geqslant T(k(N), \mathbf{x}(0))$, $i \in \mathcal{N} \setminus \mathcal{K}$. Thus, it only remains to 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 will construct a probability space where the required probability 1 convergence holds.

In order to simplify writing, we assume that the system starts at time 0, 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 4.2 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 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 $z \in C$, and define the stopping time $\hat{\theta}_z$ as

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

Since due to 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 $\mathbf{x}^R(0) \in C$, 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 sub-sequence $\{R\}$, along which with probability 1, $R^{-1}\hat{\theta}_z \to 0$. Starting from the time $\hat{\theta}_z$, along the above 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, i.e.,

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

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

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

Let $A(t)$ denote the cumulative number of arrivals up to time t to a fixed server with infinite queue length when the system starts from the state z. Now, in order to calculate the drift of each of the infinite queues, observe that cumulative number of arrivals up to time Rt to server $n \in \mathcal{K}$ in the R-th system can be written as

$$A_n^R(Rt) = \sum_{i=1}^{N_\theta^R} A_n^{(i)} + B_n(t - \Theta(N_\theta^R)), \quad \text{where} \quad N_\theta^R := \max\{j : \Theta(j) \leqslant Rt\}.$$

18

$A_n^{(i)}$'s are i.i.d. copies of the random variable $A(\hat{\theta}_z)$, $B_n(\cdot)$ is distributed as $A(t)$, and $A_n^{(i)}$'s and $B_n(\cdot)$ are independent of the random variable N_θ^R. Now, since due to Part (2) of the backward induction hypothesis the subsystem consisting of the finite queues is stable, $\mathbf{X}(\cdot)$ is irreducible and positive recurrent. Thus, we have $\mathbb{E}(\hat{\theta}_z | \mathbf{X}(0) = z) < \infty$, and hence, with probability 1,

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

Thus, using Part (3) of the backward induction hypothesis, SLLN yields, with probability 1,

$$\frac{1}{R} A_n^R(Rt) = \frac{1}{R} \sum_{i=1}^{N_\theta^R} A_n^{(i)} + \frac{B_n(t - \Theta(N_\theta^R))}{R} \to \hat{a}t, \quad \text{as} \quad R \to \infty,$$

for some $\hat{a} \leqslant 1 - \varepsilon(k(N))$. Therefore, in the conventional fluid limit, $a_n(t) \leqslant (1 - \varepsilon(k(N)))t$. Also, since the departure rate from each of the servers with infinite queue lengths is always 1, it can be seen that in the conventional fluid limit, $d_n(t) = t$, and thus, the drift of the n-th infinite queue is given by at most $-\varepsilon(k(N))$. Combining the probability 1 convergence of the time $\hat{\theta}_z$ to 0, and the probability space constructed after time $\hat{\theta}_z$, we obtain that along the subsequence $\{R\}$ with probability 1, the fluid-scaled processes converges to a limit where each infinite queue has drift at most $-\varepsilon(k(N))$. This completes the verification of Part (4), and hence of Proposition 3.1.

5 Mean-field analysis for large-scale asymptotics

In this section we will analyze the large-N behavior of the system. In particular, we will prove Lemmas 3.2 and 3.3. The next proposition is a basic mean-field fluid limit result that we need later. Define

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

Proposition 5.1. *Assume* $k(N)/N \to \kappa \in [0,1]$ *and the sequence of initial states* $(\mathbf{q}^N(0), \boldsymbol{\delta}^N(0))$ *converge to a fixed* $(\mathbf{q}(0), \boldsymbol{\delta}(0)) \in E_\kappa$, *as* $N \to \infty$, *where* $q_1(0) > 0$. *Then, with probability 1, any subsequence of* $\{N\}$ *has a further subsequence along which* $\{(\mathbf{q}^N(t), \boldsymbol{\delta}^N(t))\}_{t \geqslant 0}$ *converges, uniformly on compact time intervals, to some deterministic trajectory* $\{(\mathbf{q}(t), \boldsymbol{\delta}(t))\}_{t \geqslant 0}$ *satisfying the following equations:*

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

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

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

where

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

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

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

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

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

This type of result is standard and is obtained using Functional Strong LLN, for example as in [20, 27, 28]; we omit its proof. Also, we note that, while Proposition 5.1 is a version of [20, 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 *mean-field fluid sample path* (MFFSP) to be any deterministic trajectory satisfying the properties stated in Proposition 5.1.

We now provide an intuitive explanation of the mean-field fluid limit stated in Proposition 5.1. It is similar to that behind [20, 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 $p_i(\mathbf{q}, \boldsymbol{\delta}, \lambda)$ 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 $(\mathbf{q}, \boldsymbol{\delta})$ 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 will join idle servers. This explains that if $u > 0$, $p_0(\mathbf{q}, \boldsymbol{\delta}, \lambda) = 1$ and $p_i(\mathbf{q}, \boldsymbol{\delta}, \lambda) = 0$ 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 $\lambda p_{i-1}(\mathbf{q}, \boldsymbol{\delta}, \lambda)$, and it decreases when there is a departure from some server with queue length i, which occurs at rate $q_i - q_{i-1}$. Since 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 above, $\xi(t)$ 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, since 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 3.2

Throughout this subsection we will prove Lemma 3.2. Within this proof we will use the following terminology. Let A^N be an event pertaining to N-th system. We will write $\mathbb{P}\left(A^N\right) = \eta(N)$ to mean the following property: *There exist $C > 0$ and $N_1 > 0$ such that $\mathbb{P}\left(A^N\right) \leqslant e^{-CN}$ for all $N \geqslant N_1$.* If event A^N depends on some parameter p (say, the process initial state), we say that $\mathbb{P}\left(A^N\right) = \eta(N)$ *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 $\varepsilon_0 > 0$

$$\mathbb{P}\left(q_1^N(T_0) \leqslant \varepsilon_0\right) = \eta(N), \tag{5.1}$$

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

Fix any $T_0 > 0$; $\varepsilon_0 > 0$ will be chosen later. We now prove several claims, which rather simply follow from the process structure and basic large deviations estimates (specifically, Cramer's theorem) – they will 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 $q_1^N(0) \geqslant \varepsilon$,*

$$\mathbb{P}\left(q_1^N(\tau) \leqslant (\varepsilon/2)e^{-T_0}\right) = \eta(N). \tag{5.2}$$

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

$$\mathbb{P}\left(\delta_1^N(\tau) \leqslant (\varepsilon/2)e^{-\nu T_0}\right) = \eta(N). \tag{5.3}$$

Indeed, to prove (5.2), observe that any busy server at time t stays busy in the interval $[t, t+\tau]$ with probability at least $e^{-\tau} \geqslant e^{-T_0}$. It remains to recall that $q_1^N(0) \geqslant \varepsilon$ 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 $(\mathbf{q}^N(0), \boldsymbol{\delta}^N(0))$,*

$$\mathbb{P}\left(q_1^N(T_1) + \delta_1^N(T_1) \leqslant \varepsilon_1'\right) = \eta(N). \tag{5.4}$$

Indeed, fix any $T_1 > 0$ such that $\lambda T_1 \leqslant 1/4$. Suppose first that either $q_1^N(0) \geqslant 1/4$ or $\delta_1^N(0) \geqslant 1/4$; 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,

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

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 $\delta_1^N(0) \geqslant \varepsilon_1$,*

$$\mathbb{P}\left(q_1^N(T_2) + u^N(T_2) \leqslant \varepsilon_2'\right) = \eta(N). \tag{5.5}$$

Indeed, at time 0 there are at least $\varepsilon_1 N$ setup servers. Fix any $T_2 > 0$. In $[0, T_2]$ each of them tuns into an idle-on server with probability at least $1 - e^{-\nu T_2}$; those servers that do turn into idle-on will be either still be 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 $u^N(0) \geqslant \varepsilon_2$,*

$$\mathbb{P}\left(q_1^N(T_3) \leqslant \varepsilon_3\right) = \eta(N). \tag{5.6}$$

Indeed, fix T_3 small enough so that $e^{-\mu T_3} > 3/4$ and $\lambda T_3 < \varepsilon_2/2$. At time 0 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: $\mathbb{P}\left(|H[0, T_3]/N - \lambda T_3| \geqslant (1/2)\lambda T_3\right) = \eta(N)$. We conclude that with probability $1 - \eta(n)$ every new arrival in $[0, T_3]$ will go to an idle-on server and turn it into busy; each of those servers, in turn, will 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 small enough so that $T_3' \doteq T_1 + T_2 + T_3 \leqslant T_0$; 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 $q_1^N(T_1) \geqslant \varepsilon_1$ or $\delta_1^N(T_1) \geqslant \varepsilon_1$. Conditioned on a state at T_1 safisfying $q_1^N(T_1) \geqslant \varepsilon_1$, we have (5.1) by applying Claim 3. Therefore, it remains to prove (5.1) conditioned on a state at T_1 satisfying $\delta_1^N(T_1) \geqslant \varepsilon_1$. Under this condition at T_1, we obtain from Claim 5 that, with probability $1 - \eta(N)$, at time T_2' we have either $q_1^N(T_2') \geqslant \varepsilon_2$ or $u^N(T_2') \geqslant \varepsilon_2$. Then, conditioned on a state at T_2' satisfying $q_1^N(T_2') \geqslant \varepsilon_2$, we have (5.1) by once again applying Claim 3. It now remains to prove (5.1) conditioned on a state at T_2' satisfying $u^N(T_2') \geqslant \varepsilon_2$. Under this condition at T_2', we obtain from Claim 6 that, with probability $1 - \eta(N)$, at time T_3' we have $q_1^N(T_3') \geqslant \varepsilon_3$; and conditioned on $q_1^N(T_3') \geqslant \varepsilon_3$ at T_3', we have (5.1) by, yet again, Claim 3. The proof is complete.

5.2 Proof of Lemma 3.3

Throughout this subsection we will prove Lemma 3.3. Recall that the stability of the sub-system $\mathcal{N} \setminus \mathcal{K}$ is assumed, and hence there exists a unique stationary distribution for each N. Recall that we denote by $q^N(\infty)$ the random value of $q^N(t)$ in the steady-state. We will start by stating a few basic facts about the mean-field limits that will facilitate the proof of Lemma 3.3.

Recall the definition of MFFSP from the paragraph after Proposition 5.1, 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 $t \geqslant 0$:*

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

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

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

Proof. Fix any $\varepsilon > 0$. First observe that since $q_1(0) > 0$ and due to Proposition 5.1, $q_1(0)$ is nondecreasing whenever $q_1(t) - q_2(t) \leqslant \lambda$, we have $q_1(t) \geqslant \min\{q_1(0), \lambda\} > 0$ for all $t \geqslant 0$. Thus, Proposition 5.1 can be applied for all $t \geqslant 0$, throughout the MFFSP. Choose $\alpha = \min\{\varepsilon\nu, \ \varepsilon\}$.

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

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

For (ii), note that if $y_1(t) \leqslant \lambda - \varepsilon$, $u(t) > 0$, and $\delta_1(t) \geqslant \varepsilon$, then due to Proposition 5.1,

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

Finally, for (iii), note that from Proposition 5.1 if $y_1(t) \leqslant \lambda - \varepsilon$, $u(t) = 0$, $\delta_1(t) = 0$, and $\delta_0(t) > 0$, then $(d^+/dt)\delta_1(t) = \lambda - (q_1(t) - q_2(t)) \geqslant \varepsilon$. $\qquad\square$

Proof of statement (1). Note that it is enough to prove the following property of any MFFSP:

Claim 8. *Starting from any state $\mathbf{q}(0) \in E_\kappa$ with $\kappa \geqslant 1 - \lambda$ and $q_1(0) \in [\kappa, 1)$, 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 will prove by contradiction. Note that for the case under consideration, $q_i(t) \geqslant \kappa$ for all $i \geqslant 1$ and $t \geqslant 0$. Therefore, throughout the proof of Claim 8 we can assume $q_1(0) \geqslant \kappa > 0$, and can apply Proposition 5.1 and Claim 7.

Note that if $q_1(t) < 1$, we have $q_1(t) - q_2(t) < 1 - \kappa \leqslant \lambda$, and hence due to Claim 7, $q_1(t)$ is non-decreasing. Thus if Claim 8 does not hold, then there exists an $\varepsilon > 0$, such that $q_1(t) \leqslant 1 - \varepsilon\nu$ for all $t \geqslant 0$, and hence

$$q_1(t) - q_2(t) \leqslant \lambda - \varepsilon\nu \quad \text{for all} \quad t \geqslant 0. \tag{5.7}$$

The high-level proof idea is that if $q_1(t)$ remains below 1 by a non-vanishing amount for all $t \geqslant 0$, 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 7.(i)) or sufficiently many servers in setup mode (as in Claim 7.(ii)), 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 since $q_1(t) \leqslant 1 - \varepsilon \nu$ for all $t \geqslant 0$. This means that we cannot initiate an unbounded number of setup procedures (see Equation (5.11)). At the same time, as argued above, 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 $\mathcal{X}_\kappa \subseteq E$ as

$$\mathcal{X}_\kappa := \Big\{ (\mathbf{q}, \boldsymbol{\delta}) \in E_\kappa : q_1 + \delta_0 + \delta_1 = 1, \delta_1 \nu + q_1 - q_2 \leqslant \lambda \Big\},$$

and denote by $\mathbb{1}_{\mathcal{X}_\kappa}(\mathbf{q}(s), \boldsymbol{\delta}(s))$ the indicator of $(\mathbf{q}(s), \boldsymbol{\delta}(s)) \in \mathcal{X}_\kappa$. Observe that due to Proposition 5.1, $q_1(t)$ can be written as

$$q_1(t) = q_1(0) + \int_0^t \delta_1(s) \nu \mathbb{1}_{\mathcal{X}_\kappa}(\mathbf{q}(s), \boldsymbol{\delta}(s)) ds + \int_0^t [\lambda - q_1(s) + q_2(s)] \mathbb{1}_{\mathcal{X}_\kappa^c}(\mathbf{q}(s), \boldsymbol{\delta}(s)) ds. \quad (5.8)$$

Thus,

$$q_1(t) \geqslant q_1(0) + \int_0^t [\lambda - q_1(s) + q_2(s)] \mathbb{1}_{\mathcal{X}_\kappa^c}(\mathbf{q}(s), \boldsymbol{\delta}(s)) ds,$$

and (5.7) yields there exists positive constant K_1, which may depend on ε such that $\forall\, t \geqslant 0$

$$\int_0^t \mathbb{1}_{\mathcal{X}_\kappa^c}(\mathbf{q}(s), \boldsymbol{\delta}(s)) ds < K_1 \implies \int_0^t \mathbb{1}_{[u(s)>0]} ds < K_1. \quad (5.9)$$

Again from (5.8) we obtain

$$q_1(t) \geqslant q_1(0) + \int_0^t \delta_1(s) \nu \mathbb{1}_{\mathcal{X}_\kappa}(\mathbf{q}(s), \boldsymbol{\delta}(s)) ds$$

$$\geqslant q_1(0) + \nu \int_0^t \delta_1(s) ds - (\nu + 1) \int_0^t \mathbb{1}_{\mathcal{X}_\kappa^c}(\mathbf{q}(s), \boldsymbol{\delta}(s)) ds.$$

and thus, by (5.7) and (5.9), there exist positive constants K_2, K_2' which may depend on ε such that $\forall\, t \geqslant 0$

$$\int_0^t \delta_1(s) ds < K_1 \implies \int_0^t \mathbb{1}_{[\delta_1(s) > \frac{\varepsilon}{2}]} ds < K_2, \implies \int_0^t \mathbb{1}_{[\delta_1(s) > \frac{\varepsilon \nu}{2}]} ds < K_2'. \quad (5.10)$$

Consequently, due to Proposition 5.1, since $\delta_1(t) = \delta_1(0) + \xi(t) - \nu \int_0^t \delta_1(s) ds$, it must be the case that

$$\limsup_{t \to \infty} \xi(t) < \infty. \quad (5.11)$$

24

Furthermore, since $q_1(t) \leqslant 1 - \varepsilon\nu$ for all $t \geqslant 0$,

$$\mathbb{1}_{[\delta_0(t)=0]} \leqslant \mathbb{1}_{[u(t)>0]} + \mathbb{1}_{[\delta_1(t) \geqslant \frac{\varepsilon\nu}{2}]}.$$

Thus, (5.9) and (5.10) yield $\forall\, t \geqslant 0$,

$$\int_0^t \mathbb{1}_{[\delta_0(s)=0]} ds \leqslant K_1 + K_2'. \tag{5.12}$$

Now from Proposition 5.1 observe that

$$\xi(t) = \int_0^t \lambda(1 - p_0(\mathbf{q}(s), \boldsymbol{\delta}(s), \lambda)) \mathbb{1}_{[\delta_0(s)>0]} ds$$
$$\geqslant \int_0^t \lambda(1 - p_0(\mathbf{q}(s), \boldsymbol{\delta}(s), \lambda)) \mathbb{1}_{[\delta_0(s)>0, u(s)=0, \delta_1(s) \leqslant \varepsilon/2]} ds,$$

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

$$\xi(t) \geqslant \int_0^t \left(\lambda - \frac{\varepsilon\nu}{2} - q_1(s) + \kappa\right) \mathbb{1}_{[\delta_0(s)>0, u(s)=0, \delta_1(s) \leqslant \varepsilon/2]} ds$$
$$\geqslant \int_0^t \left(\lambda - \frac{\varepsilon\nu}{2} - q_1(s) + \kappa\right) ds - \int_0^t \mathbb{1}_{[\delta_0(s)=0]} ds - \int_0^t \mathbb{1}_{[u(s)>0]} ds - \int_0^t \mathbb{1}_{[\delta_1(s)>\varepsilon/2]} ds, \tag{5.13}$$

where the second inequality is due to the fact that $\lambda - \varepsilon\nu/2 - q_1(s) \leqslant \lambda < 1$. Therefore, since $\lambda + \kappa \geqslant 1$, we have $\lambda - \varepsilon\nu/2 - q_1(s) + \kappa \geqslant \varepsilon\nu/2 > 0$, and Equations (5.9), (5.10), (5.12), and (5.13) implies $\liminf_{t\to\infty} \xi(t) = \infty$, which is a contradiction with (5.11). This completes the proof of Claim 8. $\qquad\square$

Proof of statement (2). First we will 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 will show that for all $\varepsilon_2 > 0$,

$$\limsup_{N\to\infty} P(q_1^N(\infty) < \kappa + \lambda - \varepsilon_2) = 0. \tag{5.14}$$

Due to Lemma 3.2, note that any limit of stationary distributions is such that with probability 1, $q_1 \geqslant \varepsilon_1$ for some fixed $\varepsilon_1 > 0$. Therefore, throughout the proof of Part (2) of Lemma 3.3, it is enough to consider MFFSP so that $q_1(0) \geqslant \varepsilon_1$, and Proposition 5.1 and Claim 7 can be used. Thus, for (5.14), it is enough to show that any MFFSP has the following property:

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

$$\liminf_{t\to\infty} q_1(t) \geqslant \kappa + \lambda. \tag{5.15}$$

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 (5.14).

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

$$\lim_{N\to\infty} \mathbb{E}(q_1^N(\infty)) \leqslant \kappa + \lambda. \tag{5.16}$$

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

$$\mathbb{E}(q_1^N(\infty)) - (\kappa + \lambda) \geqslant \varepsilon_2' \mathbb{P}(q_1^N(\infty) > \kappa + \lambda + \varepsilon_2') - \frac{\varepsilon_2'}{M} \mathbb{P}\left(\kappa + \lambda - \frac{\varepsilon_2'}{M} \leqslant q_1^N(\infty) \leqslant \kappa + \lambda + \varepsilon_2'\right)$$

$$- \mathbb{P}\left(q_1^N(\infty) < \kappa + \lambda - \frac{\varepsilon_2'}{M}\right),$$

$$\geqslant \varepsilon_2' \mathbb{P}(q_1^N(\infty) > \kappa + \lambda + \varepsilon_2') - \frac{\varepsilon_2'}{M} - \mathbb{P}\left(q_1^N(\infty) < \kappa + \lambda - \frac{\varepsilon_2'}{M}\right),$$

and thus, from (5.14) and (5.16) above,

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

which in conjunction with (5.14) completes the proof of convergence of $q_1^N(\infty)$.

Convergence of $q_2^N(\infty)$. Note that given the above 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 $q(0) \in E_\kappa$ with $q_1(0) = \lambda + \kappa$ and $q_2(0) > \kappa + \varepsilon_3$, along any MFFSP we have $q_1(T_0) \geqslant \kappa + \lambda + \varepsilon_4$.*

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

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

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

Proof of Claim 10. We will prove by contradiction. Note that since $q_1(0) = \lambda + \kappa$ 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) $q_1(t) \leqslant \lambda + \kappa + \varepsilon_3/2$, (ii) $q_2(t) \geqslant \kappa + \varepsilon_3/2$, and

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

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

26

Case 1: There exists $t' \in [0, T_0/2]$, such that $u(t') = 0$ and $\delta_1(t') \leqslant \varepsilon_3/2$. 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, due to Proposition 5.1, since for $t \geqslant t'$,

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

with $(d^+/dt)\xi(t) = \lambda - y_1(t) \geqslant \varepsilon_3/2$ at $t = t'$, we have for sufficiently small $t_1 < T_0/4$ (where choice of t_1 does not depend on t'), $\delta_1(t' + t_1) \geqslant t_1\varepsilon_3/4$. Also, since 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,

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

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

$$q_1(T_0) \geqslant q_1(t' + t_1 + t_2) \geqslant \lambda + \kappa + \min\left\{\alpha, \frac{\nu t_1\varepsilon_3}{8}\right\}t_2. \tag{5.17}$$

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

Case 2: For all $t \in [0, T_0/2]$, either $u(t) > 0$ or $\delta_1(t) > \varepsilon_3/2$. In this case, due to Claim 7 (i) and (ii), there exists $\alpha > 0$, such that $(d^+/dt)q_1(t) > \alpha$ for all $t \in [0, T_0/2]$. Also, since $q_2(t)$ is non-decreasing in $[0, T_0]$. we obtain

$$q_1(T_0) \geqslant q_1\left(\frac{T_0}{2}\right) \geqslant \lambda + \kappa + \frac{\alpha T_0}{2}. \tag{5.18}$$

Combining the two cases above, 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. $\qquad\square$

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 $\mathbf{q}(0) \in E_\kappa$ 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 5.1, and hence is omitted.

The proof of the statement in (3.2) follows by using the convergence of steady states and the PASTA property. This completes the proof of Lemma 3.3.

27

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 sub-critical 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.

Acknowledgement

References

[1] Aghajani, R. and Ramanan, K. (2017). The hydrodynamic limit of a randomized load balancing network. *arXiv:1707.02005*.

[2] Andrew, L. L. H., Lin, M., and Wierman, A. (2010). Optimality, fairness, and robustness in speed scaling designs. *ACM SIGMETRICS Perform. Eval. Rev.*, 38(1):37–48.

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

[4] Bramson, M., Lu, Y., and Prabhakar, B. (2013). Decay of tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Probab.*, 23(5).

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

[6] Eschenfeldt, P. and Gamarnik, D. (2016). Supermarket queueing system in the heavy traffic regime. Short queue dynamics. *arXiv: 1610.03522*.

[7] Foss, S. and Stolyar, A. (2017). Large-scale Join-Idle-Queue system with general service times. *J. Appl. Probab.*, 54(4):995–1007.

[8] Gandhi, A., Doroudi, S., Harchol-Balter, M., and Scheller-Wolf, A. (2013). Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. In *Proc. SIGMETRICS '13*, volume 41, page 153.

[9] Liggett, T. M. (1985). *Interacting Particle Systems*. Springer New York.

[10] Lin, M., Liu, Z., Wierman, A., and Andrew, L. L. H. (2012). Online algorithms for geographical load balancing. In *Proc. IGCC'12*, pages 1–10, Washington, DC.

[11] Lin, M., Wierman, A., Andrew, L. L. H., and Thereska, E. (2013). Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Trans. Netw.*, 21(5):1378–1391.

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

[13] Liu, Z., Lin, M., Wierman, A., Low, S. H., and Andrew, L. L. H. (2011a). Geographical load balancing with renewables. *ACM SIGMETRICS Perform. Eval. Rev.*, 39(3):62–66.

[14] Liu, Z., Lin, M., Wierman, A., Low, S. H., and Andrew, L. L. H. (2011b). Greening geographical load balancing. In *Proc. SIGMETRICS '11*, pages 233–244.

[15] Lu, Y., Xie, Q., Kliot, G., Geller, A., Larus, J. R., and Greenberg, A. (2011). Join-idle-queue: a novel load balancing algorithm for dynamically scalable web services. In *Perform. Eval.*, volume 68, pages 1056–1071.

[16] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer London.

[17] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.*, 12(10):1094–1104.

[18] Mukherjee, D., Borst, S. C., van Leeuwaarden, J. S. H., and Whiting, P. A. (2016a). Universality of load balancing schemes on the diffusion scale. *J. Appl. Probab.*, 53(4).

[19] Mukherjee, D., Borst, S. C., van Leeuwaarden, J. S. H., and Whiting, P. A. (2016b). Universality of power-of-d load balancing in many-server systems. *arXiv:1612.00723*.

[20] Mukherjee, D., Dhara, S., Borst, S. C., and van Leeuwaarden, J. S. (2017). Optimal service elasticity in large-scale distributed systems. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):25.

[21] Nguyen, L. M. and Stolyar, A. L. (2016). A service system with randomly behaving on-demand Aagents. *ACM SIGMETRICS Perform. Eval. Rev.*, 44(1):365–366.

[22] Pang, G. and Stolyar, A. L. (2016). A service system with on-demand agent invitations. *Queueing Syst.*, 82(3-4):259–283.

[23] Pender, J. and Phung-Duc, T. (2016). A law of large numbers for M/M/c/delayoff-setup queues with nonstationary arrivals. *Proc. ASMTA '16*, pages 253–268.

[24] Rybko, A. N. and Stolyar, A. L. (1992). Ergodicity of stochastic processes describing the operation of open queueing networks. *Problems of Information Transmission*, 28:199–200.

[25] Shneer, S. and Stolyar, A. (2017). Stability conditions for a discrete-time decentralised medium access algorithm. *arXiv: 1707.01548*.

[26] Stolyar, A. L. (1995). On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes Relat.*, 1(4):491–512.

[27] Stolyar, A. L. (2015). Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Syst.*, 80(4):341–361.

[28] Stolyar, A. L. (2017). Pull-based load distribution among heterogeneous parallel servers: the case of multiple routers. *Queueing Syst.*, 85(1):31–65.

[29] Urgaonkar, R., Kozat, U. C., Igarashi, K., and Neely, M. J. (2010). Dynamic resource allocation and power management in virtualized data centers. In *Proc. IEEE/NOMS 2010*, pages 479–486.

[30] van der Boor, M., Borst, S. C., van Leeuwaarden, J. S. H., and Mukherjee, D. (2018). Scalable load balancing in networked systems: Universality properties and stochastic coupling methods. In *Proc. ICM '18*.

[31] Vvedenskaya, N. D., Dobrushin, R. L., and Karpelevich, F. I. (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34.

[32] Wierman, A., Andrew, L. L. H., and Tang, A. (2012). Power-aware speed scaling in processor sharing systems: optimality and robustness. *Perform. Eval.*, 69(12):601–622.

[33] Ying, L. (2017). Stein's method for mean field approximations in light and heavy traffic regimes. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):12.