Join-Idle-Queue with Service Elasticity

Debankur Mukherjee[⋆] and Alexander Stolyar[†]

[⋆]Eindhoven University of Technology, The Netherlands
[†]University of Illinois at Urbana-Champaign

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 [4] (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 $N \to \infty$. 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 $N \to \infty$. 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.

Keywords

Load balancing, auto-scaling, Join-Idle-Queue, many-server asymptotics, stability, mean-field limit, fluid limit

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 servers as evenly as possible, while using minimal instantaneous queue length information, and auto-scaling provides a popular paradigm for automatically adjusting service capacity in response to demand while meeting performance targets.

Queue-driven auto-scaling techniques have been widely investigated in the literature [3, 6] and many more, see [4] for a detailed discussion. 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. 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.

Motivated by the above, a token-based joint load balancing and auto-scaling scheme called TABS was proposed in [4], 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 [4], the authors left open a fundamental question: Is the system with a given number N of servers stable under TABS scheme? The analysis in [4] bypasses the issue of stability by assuming that each server in the system has a finite buffer capacity. Thus, it remained an important open challenge to understand the stability properties of the TABS scheme in the case of infinite buffers.

Key contributions and our approach. In this paper we address (a) The stability issue for systems under the TABS scheme with infinite buffers and (b) 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, even under subcritical load, the system may *not* be stable for all N (see [5, Remark 1]). Our first main result is that for any fixed subcritical load, the system is stable for *large enough* N. Further, in conjunction with this large-N stability result, which in particular involves 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 corresponding Markov process lacks monotonicity. It is well-known that monotonicity is a powerful primary tool for establishing stability of load balancing models [1, 8] and many more. We develop a novel method for proving *large-N sta-*

MAMA 2018 Workshop Irvine, California, USA

bility for subcritically loaded systems, and simultaneously establishing 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.

A detailed heuristic roadmap of the above proof argument is presented in Section 3. 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.

2. MAIN RESULTS

In this section, first we will describe the system and the TABS scheme, and then state the main results and discuss their ramifications. Detailed proof of all the results below can be found in [5].

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. Each server has an infinite buffer capacity. Under the TABS scheme, once a server becomes idle, it spends an $\mathrm{Exp}(\mu)$ 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 turning on of one 'off' server is triggered, i.e., its setup period is started. The setup period takes an $\mathrm{Exp}(\nu)$ time, after which the server becomes available (idle on). For the detailed token-based mechanism of the TABS scheme we refer to [4, 5].

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 $Q_i^N(t)$ denote 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. 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. The *mean-field fluid-scaled* quantities are denoted by the respective small letters, viz. $q_i^N(t) := Q_i^N(t)/N$, $\delta_0^N(t) = \Delta_0^N(t)/N$, and $\delta_1^N(t) = \Delta_1^N(t)/N$. We write $\mathbf{q}^N(t) = (q_1^N(t), q_2^N(t), \dots)$ and $\boldsymbol{\delta}^N(t) = (\delta_0^N(t), \delta_1^N(t))$. By the symbol '$\xrightarrow{\mathbb{P}}$' we denote convergence in probability for real-valued random variables.

We now present our main results.

Theorem 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.*

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 μ, $\nu > 0$, 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 $\delta_0^\star = 1 - \lambda$, $\delta_1^\star = 0$, $q_1^\star = \lambda$, $q_i^\star = 0$, for all $i \geq 2$.*

Note that the fixed point $(\boldsymbol{q}^\star, \boldsymbol{\delta}^\star)$ 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.

3. PROOFS OF THE MAIN RESULTS

First let us 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.

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 $x_i^{N,R}(t) := X_i^{N,R}(Rt)/R$, for $i \in \mathcal{N} \setminus \mathcal{K}$ be the corresponding scaled processes. 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 $0 \leq \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) < \infty$ 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.

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. *The following holds for all sufficiently large N. For each $0 \leq k \leq N$, 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 $0 \leq \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) < \infty$ 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* $t \geq t_0$*.*

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

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

(4) *For any* $x(0)$ *such that* $0 \leq \sum_{i \in \mathcal{N} \setminus \mathcal{K}} x_i(0) < \infty$ *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 is arranged as it is to facilitate its proof (see the proof idea below).

Proof of Theorem 1. Note that Theorem 1 is a special case of Proposition 3 when $k = 0$. □

Next we will state a lemma that describes asymptotic properties of sequence of systems as the number of servers $N \to \infty$, *if stability is given.* Its proof involves mean-field fluid scaling and limits.

Lemma 4. *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* $\kappa \geq 1 - \lambda$*, 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* $q_1^\star(\kappa) = \kappa + \lambda, \quad q_2^\star(\kappa) = \kappa, \delta_0^\star(\kappa) = 1 - \lambda - \kappa, \quad \delta_1^\star(\kappa) = 0$*.*

Proof of Theorem 2. Given the large-N stability property proved in Proposition 3 for $k(N) = 0$, Theorem 2 is immediate from Lemma 4. □

Proof idea for Proposition 3. The key idea is to use backward induction in k, starting from the base case $k = N$. For $k = N$, all the queues are infinite, and Parts (1) and (2) are vacuously satisfied with the convention $T(N, \mathbf{x}(0)) = 0$. Further observe that 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, 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 $k \geq k(N) + 1$ for some $k(N) \in \{0, 1, \ldots, N-1\}$ and verify that the statements hold for $k = k(N)$.

Part (1): The idea is that as long as a conventional fluid-scaled queue length at some server is positive, it can be coupled with a system where the corresponding queue length is infinite. Thus, as long as there is at least one server with positive fluid-scaled queue length, 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 components of the FSP (in turn, which is equal to the drift of each infinite queue for the corresponding system).

Part (1) $\Longrightarrow$ Part (2): To prove Part 2, we use the fluid limit technique of proving stochastic stability as in [2, 7]. Here we 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.

Part (2) + Lemma 4 $\Longrightarrow$ Part (3): This is the only part where in the proof we use the large-scale asymptotics, in particular, Lemma 4. 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 exists 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. Lemma 4 is then used to arrive at a contradiction. Note that we can apply Lemma 4 here, because Part (2) ensures the required stability.

Parts (2), (3) + Time-scale separation $\Longrightarrow$ Part (4): For this we first claim that the number of arrivals to any specific infinite queue can be written as a sum of arrivals in finite-length i.i.d. renewal cycles. Using the strong law of large numbers (SLLN) we can then show that in the limit $R \to \infty$ (recall that R is the fluid scaling parameter), 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). □

4. REFERENCES

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

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

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

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

[5] D. Mukherjee and A.L. Stolyar. Join-Idle-Queue with service elasticity: Large-scale asymptotics of a non-monotone system. *arXiv:1803.07689*, 2018.

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

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

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