Performance Evaluation 127–128 (2018) 176–193

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Flexible load balancing with multi-dimensional state-space collapse: Throughput and heavy-traffic delay optimality☆

Xingyu Zhou [a,*], Jian Tan [a], Ness Shroff [a,b]

[a] Department of ECE, The Ohio State University, Columbus, USA
[b] Department of CSE, The Ohio State University, Columbus, USA

ARTICLE INFO

Article history:
Available online 2 November 2018

Keywords:
Heavy-traffic delay optimality
Throughput optimality
Flexible load balancing
Multi-dimensional state-space collapse

ABSTRACT

Heavy traffic analysis for load balancing policies has relied heavily on the condition of state-space collapse onto a single-dimensional line in previous works. In this paper, via Lyapunov-drift analysis, we rigorously prove that even under a multi-dimensional state-space collapse, steady-state heavy-traffic delay optimality can still be achieved for a general load balancing system. This result directly implies that achieving steady-state heavy-traffic delay optimality simply requires that no server is idling while others are busy at heavy loads, thus complementing and extending the result obtained by diffusion approximations. Further, we explore the greater flexibility provided by allowing a multi-dimensional state-space collapse in designing new load balancing policies that are both throughput optimal and heavy-traffic delay optimal in steady state. This is achieved by overcoming various technical challenges, and the methods used in this paper could be of independent interest.

1. Introduction

We consider a discrete-time load balancing system which consists of one dispatcher and N servers, each associated with an infinite buffer queue. The service rate of server n is μ_n. At each time-slot t, the exogenous tasks arrive with rate λ_Σ, and upon arrival each task is immediately dispatched to one of the queues. A load balancing policy is a rule that selects the queue to which a new arrival in each time-slot should be dispatched. In recent years the development of efficient load balancing policies has received significant attention because of their applicability in distributed architectures such as Web service [1], large data storage systems (e.g., HBase [2]), and cloud computing systems [3]. A desirable load balancing policy is often one that is able to improve the average response time while achieving high utilization of resources. To this end, many works in the literature have focused on minimizing the average delay in the heavy-traffic regime where the exogenous arrival rate approaches the boundary of the capacity region, i.e., the heavy-traffic parameter $\epsilon = \sum \mu_n - \lambda_\Sigma$ approaches zero in our system.

At the heart of most heavy-traffic analysis is the notion of *state-space collapse*, which roughly means that the original multi-dimensional system space concentrates around a single dimensional (or generally a lower dimensional) subspace as the heavy-traffic parameter ϵ goes to zero. For instance, via either diffusion approximations [4] or the recently developed drift-based framework [5], it has been shown that under the so-called join-shortest-queue (JSQ) policy, the load balancing system in heavy-traffic would collapse to a one-dimensional line where all the queue lengths are equal. This indicates that

☆ This project has been funded in part through ONR, USA grant N00014-17-1-2417 and National Science Foundation grants CNS-1719371, 1717060 and 1518829.
* Corresponding author.
E-mail addresses: zhou.2055@osu.edu (X. Zhou), tan.252@osu.edu (J. Tan), shroff.11@osu.edu (N. Shroff).

https://doi.org/10.1016/j.peva.2018.10.003
0166-5316/

the system behaves as if there is only a single queue with all the servers pooled together as an aggregated server, which is often called *completeresourcepooling*. This result directly implies that JSQ is asymptotically optimal, i.e., heavy-traffic delay optimal, since the response time in the pooled single-server system is stochastically less than that of a typical load balancing system. The same one-dimensional state-space collapse is also the key in establishing heavy-traffic delay optimality for the so-called power-of-d policy [6,7], where the dispatcher routes the new arrival to a server with the shortest queue length among $d \geq 2$ servers selected uniformly at random. Instead of requiring the information of all the queue lengths as in JSQ, the power-of-d policy is able to achieve asymptotic optimality with partial queue length information, and hence provides greater flexibility and scalability for large-scale distributed systems.

The authors in [8] argue heuristically that state-space collapse to a line where all the queue lengths are equal may not be necessary for showing heavy-traffic delay optimality in load balancing systems. In particular, they proposed a symmetric threshold policy for a load balancing system with *two* homogeneous servers, and conjectured that as long as the threshold satisfies a certain property, the total work process under the threshold policy has the same diffusion limit as that under the optimal JSQ. Nevertheless, the system state in heavy-traffic limit under the threshold policy is now in the two-dimensional positive orthant rather than a single-dimensional line where all the queue lengths are equal. Hence, the authors in [8] argued that the key feature of a heavy-traffic optimal policy is to keep all the servers busy when there is substantial work rather than the strong property of maintaining all the queue lengths equal. This argument is validated in a two-server system with an asymmetric threshold policy proposed in [9], under which the total work process is proven to have the same diffusion limit as that of JSQ while the state space collapses to a *line* where the lengths of two queues are not equal. Note that besides only considering a two-server system, a further limitation in both [8] and [9] is that the asymptotic optimality holds only for a finite time interval. This is because the interchange of limits was not established for diffusion approximations in either [8] or [9]. Therefore, an interesting open problem is whether or not *steady-state* delay optimality in heavy-traffic holds under a *multi-dimensional* state-space collapse for a *general* load balancing system, and if so, how one can design a load balancing policy to achieve it.

In this paper, we take a systematic approach to addressing this problem. First, we extend the recently developed drift-based framework in [5] to rigorously show that even under a multi-dimensional state-space collapse, a load balancing policy is still able to achieve heavy-traffic delay optimality in *steady-state*. This result then allows us to explore the *flexibility* in designing load balancing policies that are not only throughput optimal but also heavy-traffic delay optimal in steady-state. The main contributions of this paper can be summarized as follows:

- We rigorously establish heavy-traffic delay optimality in steady-state under a multi-dimensional state-space collapse for a general load balancing system. More precisely, we consider a symmetric finitely generated cone $\mathcal{K}_\alpha$ parameterized by a nonnegative $\alpha \in [0, 1]$. In particular, when $\alpha = 1$ the cone reduces to the line where all the components are equal, and when $\alpha = 0$ the cone is the nonnegative orthant. Our first main result (cf. Theorem 2) states that given a throughput optimal load balancing policy, if the system state collapses to a cone $\mathcal{K}_\alpha$ with any fixed $\alpha \in (0, 1]$, this policy is heavy-traffic delay optimal in steady-state. The importance of this result is two-fold: (i) it rigorously proves a conjectured insight behind steady-state heavy-traffic delay optimality in load balancing systems. In particular, it shows that to achieve the heavy-traffic optimality in steady-state for a general system, a load balancing policy should also just be able to keep all the servers busy when there is substantial work, rather than the strong requirement of maintaining all the queue lengths equal. This complements and extends the diffusion approximation results in [8,9]. (ii) it enables us to establish heavy-traffic delay optimality under general state-space collapse regions (including even non-convex regions). This can be achieved by showing that the actual state-space collapse region can be covered by a cone $\mathcal{K}_\alpha$ with some $\alpha \in (0, 1]$, which directly implies that the system state also collapses to the cone $\mathcal{K}_\alpha$, and hence heavy-traffic delay optimality follows from Theorem 2.
- By exploiting the key implications of the first result, we are then able to characterize the degree of flexibility (from two different dimensions) in designing new load balancing policies that are both throughput-optimal and heavy traffic delay-optimal in steady-state. (i) The first dimension of flexibility is concerned with the frequency of favoring shorter queues. We find that instead of favoring shorter queues at each time-slot for every system-state, it is sufficient to favor shorter queues only when the system-state is outside a cone $\mathcal{K}_\alpha$ for any fixed $\alpha \in (0, 1]$. This means that whenever the system-state is within the cone $\mathcal{K}_\alpha$, the dispatcher is allowed to use an arbitrary Markovian dispatching distribution, and this flexibility increases as α approaches zero. (ii) The second dimension is related to the intensity with which shorter queues are favored. We find that instead of only joining the shortest queue as in the JSQ policy or having monotone decreasing probabilities from joining the shortest queue to the longest queue in the power-of-d policy, an even weaker intensity of favoring shorter queues is sufficient and this intensity can be characterized by some parameter δ. The above flexibilities from two different dimensions are stitched together in Theorem 3. We also consider the case where these two flexibilities scale with the heavy-traffic parameter ϵ, i.e., both α and δ decrease to zero as ϵ approaches zero. We show that steady-state heavy-traffic delay optimality is preserved as long as $\alpha\delta = \Omega(\epsilon^\beta)$ for any $\beta \in [0, 1)$ (cf. Proposition 4). This result offers us even more flexibility in designing efficient load balancing policies.
- The techniques used in this paper are of independent interest. For example, in order to establish throughput optimality defined in this paper, namely positive recurrence with bounded moments in steady-state, the standard stochastic drift analysis of a suitable Lyapunov function is very difficult in our case because the drift within the cone $\mathcal{K}_\alpha$ can be positive. To address the problem, we combine fluid approximations with stochastic Lyapunov theory. In particular, we show

that the Lyapunov function used in the fluid model, i.e., the sum of the queue lengths, is also a suitable Lyapunov function for the original stochastic system. This connection allows us to carry out the drift analysis in the fluid domain. Then, we come back to the stochastic system and apply the drift-based analysis of the same Lyapunov function in the original system by leveraging the result in the fluid domain to show both positive recurrence and bounded moments. Furthermore, for the result of state-space collapse to a cone, the standard analysis adopted in the single-dimensional state-space collapse fails as well in this case. This is because in contrast to the projection onto a line, the projection onto a convex cone is more complex. In fact, a closed-form formula of the projection onto a polyhedral cone is still an open problem. To circumvent this difficulty, instead of obtaining the exact projection, we find that it is sufficient to establish a monotone property of the projection to show that the system state collapses to a cone. Moreover, the bounds on the moments in the state-space collapse result hold even when the system is not in heavy-traffic, and hence can be independently used as a performance evaluation tool for the pre-limit load balancing systems.

1.1. Related work

The use of state-space collapse to study the delay performance in the heavy-traffic regime was introduced in [4] for two parallel servers. The authors, via diffusion approximations, showed that the two separate servers under the JSQ policy act as a pooled resource in the heavy-traffic limit. Since then, the methodology of diffusion limits combined with state-space collapse has been adopted in a number of papers on parallel servers [10,11,6,12,13]. For example, the author in [10] generalized the results in [4] to the case of renewal arrivals and general service times. In [6], power-of-d was shown to have the same diffusion limit as JSQ in the heavy-traffic limit. The common step in all these works is to show that the diffusion limit in heavy-traffic converges to a one-dimensional Brownian motion, which implies sample-path optimality in finite time. However, in order to capture the behavior in steady-state, an interchange of limit argument needs to be proven, which is often difficult and often not undertaken in the aforementioned works. Some exceptions include works [14,15], in which the authors proved an interchange of limit argument for generalized Jackson networks of single-server queues, thus establishing that the stationary distribution of diffusion limit provides a valid approximation for the steady-state of the original network.

Recently, the authors in [5] proposed a drift-based framework, which is able to establish steady-state heavy-traffic optimality of load balancing policy JSQ and scheduling policy MaxWeight. One of the main features of this framework is that it is able to avoid the interchange-of-limits issue by directly working on the stationary distribution. Due to this nice feature, the drift-based framework has been recently adopted to show steady-state heavy-traffic optimality of several policies in different scenarios. For instance, based on this framework, the authors in [7] established the steady-state heavy-traffic optimality of power-of-d policy. The authors in [16] identified a class of heavy-traffic delay optimal policies. Moreover, it has been shown in [17] that a joint JSQ and MaxWeight policy is heavy-traffic delay optimal for MapReduce clusters under a specific traffic scenario. For all traffic scenarios, a heavy-traffic delay optimal policy called 'local-task-first' policy was proposed in [18] based on this new framework.

However, it is worth noting that the state-space collapse of all the aforementioned heavy-traffic optimal load balancing policies is only one-dimensional. A two-dimensional state-space collapse was considered in [8], in which the authors argued heuristically that heavy-traffic delay optimality is preserved in this case, and hence claimed that the key feature of a heavy-traffic optimal load balancing policy is to keep all the servers busy when there is substantial remaining work, rather than the strong property of maintaining all the queue lengths equal. The authors in [9] validated this claim for a two-server system under an asymmetric threshold policy. In particular, they showed that the diffusion limit of the work process is the same as that under JSQ, while the state-space collapses to a *line* where the queue lengths of two servers are not equal. However, both the results in [8] and [9] hold only for a finite time since the validity of the interchange of limits argument was not established for the diffusion approximations in either paper. Motivated by this, in this paper, we extend the recently developed drift-based framework [5] and successfully establish *steady-state* heavy-traffic optimality under a multi-dimensional state-space collapse for a general load balancing system, hence complementing and extending the diffusion approximation results in [8] and [9].

We would also like to remark that besides load balancing (or scheduling) in parallel servers, state-space collapse result also plays a key role in other heavy-traffic scenarios. For example, given an $n \times n$ input queued switch, it was shown that under the complete resource pooling (CRP) condition (equivalently one-dimensional state-space collapse), the MaxWeight scheduling algorithm is heavy-traffic delay optimal in the sense of diffusion limit [19]. When the CRP condition is not met, the state-space would then collapse to a multi-dimensional space instead of a line. In this case, a diffusion limit has been established in [20]. For the steady-state behavior in this case, via the drift-based framework, it was shown that the MaxWeight scheduling policy can guarantee optimal delay scaling with respect to n [21,22]. However, in contrast to the case of the single-dimensional state-space collapse in [19,5], heavy-traffic delay optimality is not established in [21,22]. Recently, multi-dimensional state-space collapse was also used to show delay insensitivity of the proportionally fair policy in a bandwidth sharing network in heavy-traffic [23]. We finally remark that the heavy-traffic regime considered in this paper and all the aforementioned papers is the conventional heavy-traffic regime, which is different from the Halfin–Whitt heavy-traffic regime (also known as many-server heavy-traffic regime or quality-and-efficiency-driven regime). In this regime, the heavy-traffic parameter ϵ approaches zero and the number of servers N goes to infinity at the same time [24–26].

1.2. Notations

The dot product in $\mathbb{R}^N$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle \triangleq \sum_{n=1}^{N} x_n y_n$. For any $\mathbf{x} \in \mathbb{R}^N$, the l_1 norm is denoted by $\|\mathbf{x}\|_1 \triangleq \sum_{n=1}^{N} |x_n|$ and l_2 norm is denoted by $\|\mathbf{x}\| \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. In general, the l_r norm is denoted by $\|\mathbf{x}\|_r \triangleq (\sum_{n=1}^{N} |x_n|^r)^{1/r}$. Let $\mathcal{N}$ denote the set $\{1, 2, \ldots, N\}$.

2. System model and preliminaries

This section first precisely describes the system model, and then presents several necessary preliminaries.

2.1. System model

We consider a discrete-time load balancing system as follows. There is a central dispatcher and N servers indexed by n, each of which maintains a FIFO (first-in, first-out) infinite buffer size queue denoted by Q_n. In each time-slot, the central dispatcher routes the new task arrivals to one of the servers as in [5,7,17,18,27,16]. Once a task joins a queue, it will remain in that queue until its service is completed.

2.1.1. Arrival and service

Let $A_\Sigma(t)$ denote the number of exogenous tasks that arrive at the beginning of time-slot t. We assume that $A_\Sigma(t)$ is an integer-valued random variable, which is *i.i.d.* across time-slots. The mean and variance of $A_\Sigma(t)$ are denoted by λ_Σ and σ_Σ^2, respectively. We further assume that there is a positive probability for $A_\Sigma(t)$ to be zero and the arrival process has a finite support, i.e., $A_\Sigma(t) \leq A_{\max} < \infty$ for all t. Let $S_n(t)$ denote the amount of service that server n offers for queue n in time-slot t. Note that this is not necessarily equal to the number of tasks that leaves the queue because the queue may be empty. We assume that $S_n(t)$ is an integer-valued random variable, which is *i.i.d.* across time-slots. We also assume that $S_n(t)$ is independent across different servers as well as the arrival process. As before, $S_n(t)$ is also assumed to have a finite support, i.e., $S_n(t) \leq S_{\max} < \infty$ for all t and n. The mean and variance of $S_n(t)$ are denoted as μ_n and ν_n^2, respectively. Let $\mu_\Sigma \triangleq \Sigma_{n=1}^{N} \mu_n$ and $\nu_\Sigma^2 \triangleq \Sigma_{n=1}^{N} \nu_n^2$ denote the mean and variance of the hypothetical total service process $S_\Sigma(t) \triangleq \sum_{n=1}^{N} S_n(t)$.

2.1.2. Queue dynamics

Let $Q_n(t)$ be the queue length of server n at the beginning of time slot t. Let $A_n(t)$ denote the number of tasks routed to queue n at the beginning of time-slot t according to the dispatching decision. Then the evolution of the length of queue n is given by

$$Q_n(t+1) = Q_n(t) + A_n(t) - S_n(t) + U_n(t), n = 1, 2, \ldots, N, \tag{1}$$

where $U_n(t) = \max\{S_n(t) - Q_n(t) - A_n(t), 0\}$ is the unused service due to an empty queue.

2.2. Preliminaries

In this paper, we assume that the dispatching decision in each time-slot can at most depend on $\mathbf{Q}(t)$. Thus, with the system model above, the queue length process $\{\mathbf{Q}(t), t \geq 0\}$ forms a Markov chain. We consider a set of load balancing systems $\{\mathbf{Q}^{(\epsilon)}(t), t \geq 0\}$ parameterized by ϵ such that the mean arrival rate of the exogenous arrival process $\{A_\Sigma^{(\epsilon)}(t), t \geq 0\}$ is $\lambda_\Sigma^{(\epsilon)} = \mu_\Sigma - \epsilon$. Note that the heavy-traffic parameter ϵ characterizes the distance between the arrival rate and the capacity region boundary.

We say that a load balancing system is stable if the Markov chain $\{\mathbf{Q}(t), t \geq 0\}$ is positive recurrent, and then use $\overline{\mathbf{Q}}$ to denote the random vector whose distribution is the same as the steady-state distribution of $\{\mathbf{Q}(t), t \geq 0\}$. Now, we are ready to present the definitions of throughput optimality and steady-state heavy-traffic delay optimality, respectively.

Definition 1 (*Throughput Optimal*)**.** A load balancing policy is said to be throughput optimal if for any arrival rate within the capacity region, i.e., for any $\epsilon > 0$, it can stabilize the system and all the moments of $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ are finite.

Note that this is a stronger definition of throughput optimality than that in [17,18,16], because besides the positive recurrence, it also requires all the moments to be finite in steady state for any arrival rate within capacity region.

In the heavy-traffic analysis, one is interested in the behavior of the queue lengths as ϵ approaches zero. In order to present and understand the definition of steady-state heavy-traffic delay optimality, we will first recall the fundamental lower bound on the expected sum queue lengths under any throughput optimal policy [5].

Lemma 1. *Given any throughput optimal policy and assuming that $(\sigma_\Sigma^{(\epsilon)})^2$ converges to a constant σ_Σ^2 as ϵ decreases to zero, then*

$$\liminf_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] \geq \frac{\zeta}{2}, \tag{2}$$

where $\zeta \triangleq \sigma_\Sigma^2 + \nu_\Sigma^2$.

The right-hand-side of Eq. (2) is the heavy-traffic limit of a hypothetic single-server system with arrival process $A_{\Sigma}^{(\epsilon)}(t)$ and service process $\sum_n^N S_n(t)$ for all $t \geq 0$. This hypothetical single-server queueing system is often called the *resource-pooled system*. Since a task cannot be moved from one queue to another in the load balancing system, it is easy to see that the expected sum queue lengths of the load balancing system is larger than the expected queue length in the resource-pooled system. However, under a certain load balancing policy, the lower bound in Eq. (2) can actually be attained in the heavy-traffic limit and hence based on Little's law this policy achieves the minimum average delay of the system in steady-state. This directly motivates the following definition of steady-state heavy-traffic delay optimality as in [5,7,17,18,27,16].

Definition 2 (*Heavy-traffic Delay Optimality in Steady-state*)**.** A load balancing scheme is said to be heavy-traffic delay optimal in steady-state if the steady-state queue length vector $\overline{\mathbf{Q}}^{(\epsilon)}$ satisfies

$$\limsup_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] \leq \frac{\zeta}{2},$$

where ζ is defined in Lemma 1.

Before we end this section, we will introduce an N-dimensional cone, which will be very useful in our upcoming analysis. In particular, the cone $\mathcal{K}_\alpha$ is finitely generated by a set of N vectors $\{\mathbf{b}^{(n)}, n \in \mathcal{N}\}$, i.e.,

$$\mathcal{K}_\alpha = \left\{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} = \sum_{n \in \mathcal{N}} w_n \mathbf{b}^{(n)}, w_n \geq 0 \text{ for all } n \in \mathcal{N}\right\}, \tag{3}$$

where $\mathbf{b}^{(n)}$ is an N-dimensional vector with the nth component being 1 and α everywhere else for some $\alpha \in [0, 1]$. It follows that, if $\alpha = 0$, the cone $\mathcal{K}_\alpha$ is the non-negative orthant of $\mathbb{R}^N$, and if $\alpha = 1$, the cone $\mathcal{K}_\alpha$ reduces to the single-dimensional line in which all the components are equal. The polar cone $\mathcal{K}_\alpha^\circ$ of the cone $\mathcal{K}_\alpha$ is defined as

$$\mathcal{K}_\alpha^\circ = \left\{\mathbf{x} \in \mathbb{R}^N : \langle \mathbf{x}, \mathbf{y} \rangle \leq 0 \text{ for all } \mathbf{y} \in \mathcal{K}_\alpha\right\},$$

which will also be quite important in our analysis.

3. Main results

In this section, we present our main results. First, we show that a load balancing policy can be heavy-traffic delay optimal in steady-state even under a multi-dimensional state-space collapse. Then, by leveraging the key insight behind this result, we are able to explore the degree of flexibility a load balancing policy can enjoy while guaranteeing both throughput optimality and heavy-traffic delay optimality. Furthermore, a useful generalization of this result is presented at the end of this section.

3.1. Multi-dimensional state-space collapse

We will first introduce the notion of state-space collapse used in this paper, which intuitively means that in steady-state the queue length process concentrates around a region of the state-space in heavy-traffic. As stated before, in most of the previous works on load balancing, the state-space collapse region is a single-dimensional line. In contrast, we are interested in the situation where the state-space collapse region is the N-dimensional cone $\mathcal{K}_\alpha$ defined in Eq. (3), which includes the single-dimensional line as a special case. For a given cone $\mathcal{K}_\alpha$, we decompose $\overline{\mathbf{Q}}^{(\epsilon)}$ into two parts as follows

$$\overline{\mathbf{Q}}^{(\epsilon)} = \overline{\mathbf{Q}}_{\|}^{(\epsilon)} + \overline{\mathbf{Q}}_{\perp}^{(\epsilon)},$$

where $\overline{\mathbf{Q}}_{\|}^{(\epsilon)}$ is the projection onto the cone $\mathcal{K}_\alpha$, referred to as the parallel component, and $\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}$ is the remainder, referred to as the perpendicular component, which is actually the projection onto the polar cone $\mathcal{K}_\alpha^\circ$. Note that this decomposition is well defined and unique since the cones $\mathcal{K}_\alpha$ and $\mathcal{K}_\alpha^\circ$ are both closed and convex, which follows from the fact that $\mathcal{K}_\alpha$ is finitely generated. The norm $\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|$ is the distance between $\overline{\mathbf{Q}}^{(\epsilon)}$ and the cone $\mathcal{K}_\alpha$. We say that the queue length process concentrates around the cone $\mathcal{K}_\alpha$ if the moments of the distance $\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|$ are upper bounded by constants. This motivates the following definition.

Definition 3 (*State-space Collapse to $\mathcal{K}_\alpha$*)**.** Given an $\alpha \in (0, 1]$, we say the state-space of a load balancing system collapses to the cone $\mathcal{K}_\alpha$ if

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^r\right] \leq M_r \tag{4}$$

for all $\epsilon \in (0, \epsilon_0)$, $\epsilon_0 > 0$ and for each $r = 1, 2, \ldots$, in which M_r are constants that are independent of ϵ.

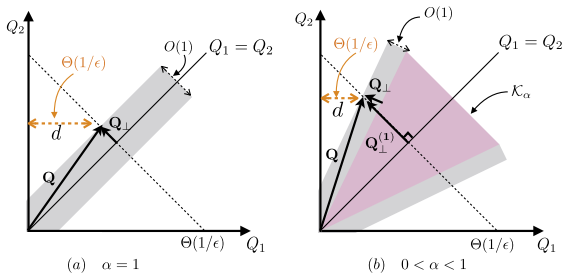

Fig. 1. A geometric illustration of the key insight of steady-state heavy-traffic optimality in load balancing systems. In the figures above, we use the gray area to represent the bounded distance between $\mathbf{Q}$ and the cone $\mathcal{K}_\alpha$. The dashed line represents the total tasks in the system and hence has order $1/\epsilon$. In (a), $\alpha = 1$ and hence the system state collapses to the one-dimensional line. As a result, the queue lengths of two servers are nearly equal. However, Theorem 2 tells us that this is not the key feature of heavy-traffic optimality since in (b) the queue lengths difference between two servers, $\|\mathbf{Q}_\perp^{(1)}\|$, is of the same order of $\|\mathbf{Q}\|$. Rather, the key feature behind heavy-traffic optimality is that it keeps all the servers busy when there is substantial work. This is achieved by keeping the system state far away from the boundary via state-space collapse, as ϵ approaches zero (see the distance d in both (a) and (b)).

Remark 1. It is worth noting that although the result of state-space collapse is often used as the key step in establishing heavy-traffic delay optimality, the upper bound itself holds even when the system is not in the heavy-traffic limit, and this can be of independent interest for analyzing the performance of the system.

Now, we are ready to present our first main result.

Theorem 2. *Given a throughput optimal load balancing policy, if there exists an $\alpha \in (0, 1]$ such that the state-space collapses to the cone $\mathcal{K}_\alpha$, then this policy is heavy-traffic delay optimal in steady-state.*

Proof. See Section 4.1. □

From this theorem, we can make the following important observations regarding heavy-traffic delay optimality in load balancing systems. A geometric illustration is presented in Fig. 1 to facilitate the understanding.

(i) If $\alpha = 1$, then this theorem reduces to previous results on heavy-traffic delay optimality under a single-dimensional state-space. In this case, the state-space can be regarded as if it evolves in a one-dimensional subspace where all queues are equal. This is because the queue-length difference between servers is bounded by a constant and hence is substantially smaller than the queue lengths themselves, which are on the order of $1/\epsilon$. See Fig. 1(a).

(ii) This theorem tells us that in order to be heavy-traffic delay optimal, a policy does not necessarily have to keep all the queues equal as in the case of $\alpha = 1$. This is because Theorem 2 implies that heavy-traffic delay optimality is preserved even when the difference between the various queues is of the same order as the queue lengths themselves, as shown in Fig. 1(b). Therefore, this theorem indicates that the key feature of a heavy-traffic delay optimal policy is that *it keeps all the servers busy when there are substantial tasks in the system.* This is achieved by keeping system states far away from the boundary via state-space collapse as ϵ approaches zero, see the distance d in Fig. 1.

(iii) It should also be pointed out that the cone $\mathcal{K}_\alpha$ in Theorem 2 is not necessarily the actual region that state-space collapses to. In fact, this theorem tells us that for heavy-traffic delay optimality, the actual region of state-space collapse, say $\mathcal{R}$, does not matter as long as it lies within a cone $\mathcal{K}_\alpha$ for some $\alpha \in (0, 1]$. This is because in this case the distance to the cone $\mathcal{K}_\alpha$ is not larger than that to the region $\mathcal{R}$. Thus, once it collapses to the region $\mathcal{R}$, it also collapses to the cone $\mathcal{K}_\alpha$ according to Definition 3, and hence achieves heavy-traffic delay optimality. This nice property may be of independent interest since it enables us to establish heavy-traffic delay optimality even when the multi-dimensional state-space collapse region is non-regular and non-convex.

Now, we turn to provide the high-level intuition on why Theorem 2 holds. To start with, note that the following equation

$$U_n(t)Q_n(t+1) = 0 \tag{5}$$

holds for all n and t. This follows directly from the queue dynamics in Eq. (1). Thus, for the single-server resource-pooled system, when there is positive unused service at time-slot t, the queue must be empty at time-slot $t+1$. In contrast, for a load balancing system, due to the fact that a task cannot be moved from one queue to another queue, there exist situations when one queue, say i, has positive unused service, i.e., $U_i(t) > 0$ and hence $Q_i(t+1) = 0$, while there are remaining tasks in other queues, i.e., $Q_j(t+1) > 0$ for some j. As a result, the average queue length of the resource-pooled system is the lower bound for the load balancing system. Therefore, in order to achieve this lower bound (and hence heavy-traffic delay optimality by definition), a load balancing policy should guarantee that when one queue has positive unused service, all

the other queues should be empty in steady-state. In fact, this is actually the insight behind the sufficient and necessary condition for heavy-traffic delay optimality in Lemma 6, i.e.,

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\| \overline{\mathbf{Q}}^{(\epsilon)}(t+1) \right\|_1 \left\| \overline{\mathbf{U}}^{(\epsilon)}(t) \right\|_1 \right] = 0. \tag{6}$$

Next, let us take a closer look at the condition above. As a result of Eq. (5), the left-hand side of Eq. (6) is always zero when all the queue lengths are positive. This explains why it is not necessary to keep all the queue lengths equal as in the case of single-dimensional state-space collapse. Instead, what really matters in the condition is the situation when $U_i(t) > 0$ (and hence $Q_i(t+1) = 0$) for some i. In this case, the condition requires all the other queue lengths must be zero as well, i.e., $Q_n(t+1) = 0$ for all n. In other words, it requires all the servers be busy when there is substantial work, which is intuitively satisfied when the queue length process collapses to the cone $\mathcal{K}_\alpha$ for any $\alpha \in (0, 1]$. This is because the cone is far away from the states where some queue is empty while another queue is non-empty in heavy traffic. The proof is presented in Section 4.1.

Remark 2. We would remark on the choice of the particular form of cone $\mathcal{K}_\alpha$, which provides further intuitions on Theorem 2. One reason for the choice is that $\mathcal{K}_\alpha$ is finitely generated, and hence it is closed and convex. This guarantees that the projection onto this cone is well-defined and unique. Another more important reason is that $\mathcal{K}_\alpha$ can approach the non-negative orthant while guaranteeing that within the cone there are no *'bad points'* where some queue is empty and another queue is non-empty. This directly implies that once the state-space collapses to the cone $\mathcal{K}_\alpha$, the condition in Eq. (6) is satisfied. One might think of using an 'ice-cream' cone defined below as a substitute of cone $\mathcal{K}_\alpha$,

$$\mathcal{K}_\theta = \left\{ \mathbf{x} \in \mathbb{R}^N : \frac{\|\mathbf{x}_\parallel^{(\mathbf{1})}\|}{\|\mathbf{x}\|} \geq \cos(\theta) \right\},$$

where $\mathbf{x}_\parallel^{(\mathbf{1})}$ is the projection of $\mathbf{x}$ onto the line $\mathbf{1} = (1, 1, \ldots, 1)$. The key problem with such a choice is that in order to exclude all the *'bad points'* from $\mathcal{K}_\theta$ for a large system size N, the cone $\mathcal{K}_\theta$ has to basically reduce to the line $\mathbf{1}$, and hence does not provide us with any further flexibility. To see this, let us start with $N = 2$. In this case, the choice of $\mathcal{K}_\theta$ is fine since it is able to approach the non-negative orthant as θ approaches $\pi/4$, while guaranteeing that there are no *bad points* within the cone. In fact, it is easy to see that in this case θ plays the same role as α in $\mathcal{K}_\alpha$. Now, consider the case $N = 3$. One might choose $\theta < \arccos(1/\sqrt{3})$ in order to exclude the *'bad points'* on the axes from the cone $\mathcal{K}_\theta$. However, all the points on the line $\mathbf{x} = (1, 1, 0)$ are also *'bad points'*. Thus, in order to exclude all of these points, the choice of θ should be $[0, \arccos(\sqrt{2}/\sqrt{3}))$. In general, for the N-dimension case, the choice of θ should be less than $\arccos(\sqrt{N-1}/\sqrt{N})$, which approaches zero for large N. Hence, for a large system size N, the cone $\mathcal{K}_\theta$ has to approach the single-dimensional line $\mathbf{1}$ in order to guarantee heavy-traffic delay optimality, which is not interesting because it does not provide any significant flexibility compared to the single-dimensional line. The insight of keeping all the servers busy by excluding the *'bad points'* when there is substantial work is also useful for scheduling problems. For example, the cone considered in [21] for the scheduling problem in a switch system contains infinitely many *'bad points'*. It is actually due to this problem that the MaxWeight policy in this case can only guarantee optimal scaling rather than heavy-traffic delay optimality under the single-dimensional state-space collapse in [19,5].

In contrast, the $\mathcal{K}_\alpha$ considered in our paper is able to exclude all the *'bad points'* for any $\alpha > 0$ and hence guarantees heavy-traffic delay optimality. This fact not only captures the essence of heavy-traffic delay optimality in load balancing systems via multi-dimensional state-space collapse, but also provides flexibility in analyzing and designing new load balancing policies, which will be explored in the next section.

3.2. Flexible load balancing

In this section, instead of focusing on yet another policy, we step back and explore the possibility provided by Theorem 2 in analyzing and more importantly designing flexible load balancing policies that are both throughput optimal and heavy-traffic delay optimal. This is motivated by the fact that existing policies are often too restrictive and might not be easily adopted to guarantee system performance in scenarios where *data locality* or *inaccurate* information of queue lengths exists, which are common in load balancing systems [28,29].

Before we present our main result, let us first introduce some necessary concepts. Let $P_n(t)$ be the probability that the new arrivals are dispatched to the nth shortest queue at time-slot t. By the Markovian assumption, the dispatching distribution $\mathbf{P}(t)$ can at most depend on $\mathbf{Q}(t)$. Let

$$\Delta(t) = \mathbf{P}(t) - \mathbf{P}_{\text{rand}}(t), \tag{7}$$

where $\mathbf{P}_{\text{rand}}(t)$ is the dispatching distribution under uniform random routing (homogeneous case) or proportional random routing (heterogeneous case), i.e., for homogeneous servers, each component of $\mathbf{P}_{\text{rand}}(t)$ is $1/N$, and for heterogeneous servers the nth component of $\mathbf{P}_{\text{rand}}(t)$ is $\mu_{\sigma_t(n)}/\mu_\Sigma$ where $\sigma_t(n)$ is the index of the nth shortest queue at time-slot t.

To facilitate the understanding of the concepts above, let us look at some examples. Consider a load balancing system with four homogeneous servers. Under uniform random routing, we have $\Delta(t) = (0, 0, 0, 0)$ for each time-slot t. Under the JSQ policy, the dispatcher always assigns the new arrival to the shortest queue, and thus we have

$\Delta(t) = (3/4, -1/4, -1/4, -1/4)$ for each time-slot t. Under the power-of-2 policy, the dispatcher randomly picks two servers and dispatches the new arrivals to the server with the shorter queue length. It easily follows that $\Delta(t) = (1/4, 1/12, -1/12, -1/4)$ for each time-slot t. Note that, from these examples, we can see that a positive value of $\Delta_n(t)$ means that the dispatcher favors the nth shortest queue, while a non-positive value means the dispatcher disfavors the corresponding queue. This is because $\Delta(t)$ equals $(0, 0, 0, 0)$ under uniform random routing, which has no preference over any queues.

Now, we are prepared to present our second main result, which characterizes the degree of flexibility a load balancing policy enjoys while guaranteeing throughput and heavy-traffic delay optimality.

Theorem 3. *Given a load balancing policy, if there exists a cone $\mathcal{K}_\alpha$ with $\alpha \in (0, 1]$ such that for all $\mathbf{Q}(t) \notin \mathcal{K}_\alpha$, there is some $k \in \{2, \ldots, N\}$ such that*

$$\Delta_n(t) \geq 0, n \leq k \text{ and } \Delta_n(t) \leq 0, n \geq k \tag{8}$$

and

$$\min\left(|\Delta_1(t)|, |\Delta_N(t)|\right) \geq \delta \tag{9}$$

for some positive constant δ that is independent of ϵ, then this policy is both throughput and heavy-traffic delay optimal in steady-state.

Proof. See Section 4.2. □

Remark 3. It can be easily seen that previous steady-state heavy-traffic delay optimal policies, namely JSQ and power-of-d, satisfy the conditions in Theorem 3 with $\alpha = 1$.

Before we turn to the technical aspects, let us first elaborate on the key messages behind this theorem. In sum, this theorem characterizes the flexibility in achieving throughput and heavy-traffic delay optimality from the following two dimensions.

(i) The first dimension relates to the frequency of favoring shorter queues. This can be seen from the fact that there are no requirements on $\Delta(t)$ whenever $\mathbf{Q}(t)$ falls in the cone $\mathcal{K}_\alpha$, and α can be arbitrarily close to zero. This is significantly different from previous heavy-traffic delay optimal policies, e.g., JSQ and power-of-d. These policies have to favor shorter queues for *every* time-slot and *every* system state. This is often too restrictive and may not be achievable, especially when considering the data locality problem, since in this case the dispatcher has to place tasks to servers that store the corresponding input data chunks. In contrast, the above theorem tells us that a load balancing policy has the flexibility to adopt any dispatching distribution whenever the queue-length state falls in a region that can be covered by a cone $\mathcal{K}_\alpha$ for some $\alpha \in (0, 1]$. For example, for a load balancing system with heterogeneous servers, the dispatcher can just use uniform random routing when the system state lies within a cone $\mathcal{K}_\alpha$, which provides us a lot of flexibility (e.g., easy implementation and lower message overhead), compared to JSQ policy, and the flexibility increases as α decreases. It is also worth noting that although the delay optimality is preserved in heavy-traffic for any $\alpha \in (0, 1]$, the actual delay performance under medium or low loads might get worse as α decreases in some cases. Thus, the parameter α also captures an important trade-off between flexibility and delay performance under medium loads, which may be an interesting open problem to explore in the future.

(ii) The second dimension is related to the intensity with which shorter queues are favored. This can be seen from the conditions on $\Delta(t)$ in Eqs. (10) and (11). Specifically, instead of joining only the shortest queue as in the JSQ policy or having monotone decreasing probabilities from joining the shortest queue to the longest queue in the power-of-d policy, an even weaker intensity of favoring shorter queues is sufficient, and this intensity can be characterized by the parameter δ. This kind of flexibility is very useful when the queue length information available at the dispatcher may be inaccurate due to communication delay or sampling error.

We now highlight the technical contributions behind this theorem. Since this theorem is proved based on Theorem 2, all we need to show are throughput optimality and state-space collapse to the cone.

For throughput optimality, i.e., positive recurrence and bounded moments in steady-state, the standard drift analysis of a suitable Lyapunov function is very difficult in our case. This is because the drift within the cone $\mathcal{K}_\alpha$ can be positive; hence, it is challenging from renewal theory to find a sufficiently large T such that the drift within T time slots is negative outside a finite set. Our approach is to combine fluid approximations with stochastic Lyapunov theory. In particular, we show that the Lyapunov function used in the fluid model, i.e., the sum of the queue lengths, is also a suitable Lyapunov function for the original stochastic system. This connection allows us to carry out drift analysis in the fluid domain, which makes it easier to find the T. Then, we come back to the stochastic system and apply the drift analysis of the same Lyapunov function in the stochastic system to show both positive recurrence and bounded moments in steady-state. This approach also provides us with a good intuition on throughput optimality in load balancing systems. Informally speaking, if $\mathbf{Q}(t)$ is in the cone $\mathcal{K}_\alpha$, the drift of the sum queue lengths is of order ϵ towards the origin for any dispatching distribution since there is no unused service. If $\mathbf{Q}(t)$ is outside the cone $\mathcal{K}_\alpha$, it is easy to see that the dispatching distribution in Theorem 3 is

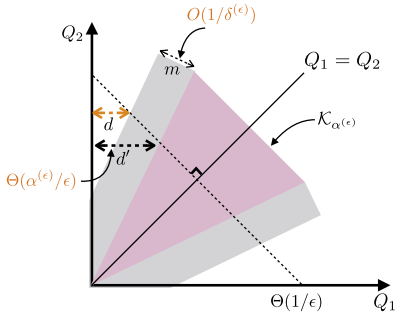

Fig. 2. A geometric illustration of the result in Proposition 4. As before, we use the gray area to represent the distance between $\mathbf{Q}$ and the cone $\mathcal{K}_\alpha$. The dashed line represents the total tasks in the system and hence has order $1/\epsilon$. As in Fig. 1, in order to guarantee that all the servers are busy when there is substantial work in the system, the distance d should be large enough when ϵ goes to zero. To this end, it is sufficient to require that the order of distance m is smaller than that of the distance d'. It is easy to see $\alpha^{(\epsilon)}\delta^{(\epsilon)} = \Omega(\epsilon^\beta)$ for any $\beta \in [0, 1)$ satisfies this requirement.

strictly better than random routing, and hence enjoys a drift towards the origin. Therefore, the sum queue length will not go to infinity in steady-state. It is worth noting that for a continuous-time model, the idea of combining fluid models with Lyapunov drift to show bounded moments was investigated in [30] for a multi-class queueing network with a fixed routing matrix independent of queue lengths. Because the results are obtained for the fixed routing case, they cannot be directly applied to our load balancing case in which routing decisions are based on queue lengths.

For the state-space collapse to the cone $\mathcal{K}_\alpha$, the standard analysis adopted in the single-dimensional collapse also fails in this case. This is because in contrast to the projection onto a line, the projection onto a cone is very difficult. In fact, a closed-form formula of the projection onto a polyhedral cone is still an open problem. To circumvent this difficulty, instead of obtaining the exact projection, we are able to find an important monotone property on the projection, which is sufficient to establish a negative drift independent of ϵ along the direction of $\mathbf{Q}_\perp$ when the queue length state is outside the cone $\mathcal{K}_\alpha$. This in turn indicates that the distance between the system state and the cone $\mathcal{K}_\alpha$ cannot go to infinity as ϵ approaches zero. Therefore, by definition, it establishes the result of state-space collapse to the cone. Combining this with throughput optimality yields heavy-traffic delay optimality according to Theorem 2.

Remark 4. We would like to remark that the techniques used to prove Theorem 3 are of independent interest and may have broader applicability. For example, we are currently investigating whether this technique can be used to design a broader class of heavy-traffic delay optimal *scheduling* policies.

3.3. Generalization

In the last section, we have shown that when it comes to designing a heavy-traffic delay optimal load balancing policy, one has the flexibility of choosing the frequency and intensity of favoring shorter queues, which are parameterized by some fixed positive constants α and δ, respectively. In particular, smaller values of these two constants mean favoring the shorter queues less frequently and with less intensity. In this section, we will show that these two constants can actually approach zero at a certain rate with respect to the heavy-traffic parameter ϵ so that the given policy can still guarantee heavy-traffic delay optimality. As a result, we can exploit this fact to achieve even significant flexibility in designing new policies.

Proposition 4. *Given a throughput optimal load balancing policy, if there exists a cone $\mathcal{K}_{\alpha^{(\epsilon)}}$ such that for all $\mathbf{Q}(t) \notin \mathcal{K}_{\alpha^{(\epsilon)}}$, there is some $k \in \{2, \ldots, N\}$ such that*

$$\Delta_n(t) \geq 0, n \leq k \text{ and } \Delta_n(t) \leq 0, n \geq k \tag{10}$$

and

$$\min\left(|\Delta_1(t)|, |\Delta_N(t)|\right) \geq \delta^{(\epsilon)} \tag{11}$$

for some $\delta^{(\epsilon)}$. Suppose that $\alpha^{(\epsilon)}$ and $\delta^{(\epsilon)}$ satisfy

$$\alpha^{(\epsilon)}\delta^{(\epsilon)} = \Omega(\epsilon^\beta)$$

for some $\beta \in [0, 1)$, then this policy is heavy-traffic delay optimal.

Proof. See Appendix G. □

As before, a geometric view of the result of Proposition 4 is presented in Fig. 2.

4. Proofs

In this section, we present the proofs of Theorems 2 and 3, respectively.

4.1. Proof of Theorem 2

Before we present the proof, we first introduce the following lemma.

Lemma 5. *For any $\epsilon > 0$ and $t \geq 0$, we have*

$$Q_n^{(\epsilon)}(t+1)U_n^{(\epsilon)}(t) = 0.$$

Moreover, if the system has a finite first moment, then we have for some constants c_1 and c_r

$$\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2\right] \leq c_1\epsilon \text{ and } \mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_r^r\right] \leq c_r\epsilon,$$

where $r \in (1, \infty)$.

Proof. According to the queues dynamic in Eq. (1), we can see that when $U_n(t)$ is positive, $Q_n(t+1)$ must be zero, which directly implies the result $Q_n^{(\epsilon)}(t+1)U_n^{(\epsilon)}(t) = 0$ for any $\epsilon > 0$, $1 \leq n \leq N$ and $t \geq 0$. To show the second result, let us consider the Lyapunov function $W_1(\mathbf{Q}(t)) \triangleq \|\mathbf{Q}(t)\|_1$. Since the system has a finite first moment, the mean drift of $W_1(\mathbf{Q})$ is zero in steady state, which gives

$$\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right] = \epsilon.$$

Then, due to the fact that $U_n(t) \leq S_{\max}$ for all $1 \leq n \leq N$ and $t \geq 0$, we have $\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_r^r \leq (S_{\max})^{r-1}\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1$, which implies that $c_r = (S_{\max})^{r-1}$. Note that $\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2 \leq N\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_2^2$, which gives $c_1 = NS_{\max}$. □

Now we are ready to prove Theorem 2.

Proof of Theorem 2. We prove this theorem by combining the following lemma with the condition of state-space collapse to the cone $\mathcal{K}_\alpha$. The proof of the lemma is relegated to Appendix A.

Lemma 6. *For a throughput optimal policy, it is heavy-traffic delay optimal if and only if*

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1\left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0. \tag{12}$$

Next, we will show that under the condition that the state space collapses to a cone $\mathcal{K}_\alpha$ with $\alpha \in (0, 1]$, the condition in Eq. (12) holds. For brevity, we will omit the references t and ϵ, and use $\overline{\mathbf{Q}}^+$ to denote $\overline{\mathbf{Q}}(t+1)$ in the following. First, we have

$$\begin{aligned}
\mathcal{T}^{(\epsilon)} &\triangleq \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1\left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \\
&= \mathbb{E}\left[\sum_{i=1}^N \overline{U}_i\left(\sum_{j=1}^N \overline{Q}_j^+\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^N \overline{U}_i\left(\sum_{j=1}^N \left(\overline{Q}_{\|j}^+ + \overline{Q}_{\perp j}^+\right)\right)\right],
\end{aligned} \tag{13}$$

where $\overline{Q}_{\|j}^+$ is the jth component of $(\overline{\mathbf{Q}}^+)_\|$ and similarly $\overline{Q}_{\perp j}^+$ is the jth component of $(\overline{\mathbf{Q}}^+)_\perp$. For simplicity, we use $\overline{\mathbf{Q}}_\|^+$ to denote $(\overline{\mathbf{Q}}^+)_\|$ and $\overline{\mathbf{Q}}_\perp^+$ to denote $(\overline{\mathbf{Q}}^+)_\perp$, respectively. Since the vector $\overline{\mathbf{Q}}_\|^+$ is in cone $\mathcal{K}_\alpha$ by definition, there exist non-negative weights $w_1, \ldots, w_N$ such that $\overline{\mathbf{Q}}_\|^+ = \sum w_n \mathbf{b}^{(n)}$. Recall that when $U_n(t) > 0$, $Q_n(t+1) = 0$ by Lemma 5. Thus, when $\overline{U}_i(t) > 0$, we have

$$\begin{aligned}
\overline{Q}_i^+ &= 0 \\
\overline{Q}_{\|i}^+ &= -\overline{Q}_{\perp i}^+ \\
\sum w_n b_i^{(n)} &= \overline{Q}_{\|i}^+ \\
\sum w_n b_j^{(n)} &= \overline{Q}_{\|j}^+ \leq \frac{1}{\alpha}\overline{Q}_{\|i}^+ \text{ for all } j \neq i
\end{aligned}$$

The last inequality follows from the definition of vector $\mathbf{b}^{(\eta)}$. Therefore, the term $\mathcal{T}$ in Eq. (13) can be upper bounded as follows.

$$\begin{aligned}
\mathcal{T}^{(\epsilon)} &\overset{(a)}{\leq} \mathbb{E}\left[\sum_i \overline{U}_i\left(-N_1\overline{Q}^+_{\perp i} + \sum_j \overline{Q}^+_{\perp j}\right)\right] \\
&= \mathbb{E}\left[\langle \overline{\mathbf{U}}, -N_1\overline{\mathbf{Q}}^+_{\perp}\rangle\right] + \mathbb{E}\left[\langle \overline{\mathbf{U}}, \langle \mathbf{1}, \overline{\mathbf{Q}}^+_{\perp}\rangle \mathbf{1}\rangle\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\langle \overline{\mathbf{U}}, -N_1\overline{\mathbf{Q}}^+_{\perp}\rangle\right] \\
&\overset{(c)}{\leq} N_1\sqrt{\mathbb{E}\left[\left\|\overline{\mathbf{U}}\right\|^2\right]\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^+_{\perp}\right\|^2\right]}. \\
&\overset{(d)}{\leq} N_1\sqrt{c_2\epsilon M_2}
\end{aligned} \tag{14}$$

where in (a) $N_1 = N/\alpha$; (b) comes from the non-negativity of $\overline{\mathbf{U}}$ and the fact that $\langle \mathbf{1}, \overline{\mathbf{Q}}^+_{\perp}\rangle \leq 0$ since $\mathbf{1} \in \mathcal{K}_\alpha$ and $\overline{\mathbf{Q}}^+_{\perp} \in \mathcal{K}^\circ_\alpha$; (c) is the result of Cauchy–Schwarz inequality for random vectors; (d) holds because of Lemma 5 and the definition of state-space collapse in Eq. (4) combined with the fact that $\mathbf{Q}(t+1)$ and $\mathbf{Q}(t)$ have the same distribution in steady-state. Since c_2, M_2 and N_1 are all constants that are independent of ϵ, we have $\lim_{\epsilon\to 0}\mathcal{T}^{(\epsilon)} = 0$, which establishes the result in Eq. (12), and hence heavy-traffic delay optimality. □

4.2. Proof of Theorem 3

As already pointed out, the proof of Theorem 3 naturally falls into two parts: throughput optimality and state-space collapse. Then, it follows directly from Theorem 2 that the result in Theorem 3 is true. In both proofs, we will use the Lyapunov drift-based approach developed in [5] to derive bounded moments in steady state. The following lemma is a T-step version of Lemmas 2 and 3 in [21]. This lemma could be proved by simply replacing the one-step transition probability to T-step transition probability, and hence we omit the proof here.

Lemma 7. *For an irreducible aperiodic and positive recurrent Markov chain $\{X(t), t \geq 0\}$ over a countable state space $\mathcal{X}$, which converges in distribution to $\overline{X}$, and suppose $V : \mathcal{X} \to \mathbb{R}_+$ is a Lyapunov function. We define the T time-slot drift of V at X as*

$$\Delta V(X) \triangleq [V(X(t_0+T)) - V(X(t_0))]\mathcal{I}(X(t_0) = X),$$

where $\mathcal{I}(.)$ is the indicator function. Suppose for some positive finite integer T, the T time-slot drift of V satisfies the following conditions:

- *(C1) There exists an $\eta > 0$ and a $\kappa < \infty$ such that for any $t_0 = 1, 2, \ldots$ and for all $X \in \mathcal{X}$ with $V(X) \geq \kappa$,*

$$\mathbb{E}[\Delta V(X) \mid X(t_0) = X] \leq -\eta.$$

- *(C2) There exists a constant $D < \infty$ such that for all $X \in \mathcal{X}$,*

$$\mathbb{P}(|\Delta V(X)| \leq D) = 1.$$

Then $\{V(X(t)), t \geq 0\}$ converges in distribution to a random variable $\overline{V}$, and all moments of $\overline{V}$ exist and are finite. More specifically, we have for any $r = 1, 2, \ldots$

$$\mathbb{E}\left[V(\overline{X})^r\right] \leq (2\kappa)^r + (4D)^r\left(\frac{D+\eta}{\eta}\right)^r r!. \tag{15}$$

4.2.1. Throughput optimality

We would prove the following result in this subsection.

Proposition 8. *Under the condition of Theorem 3, the given policy is throughput optimal.*

We would prove this result by combining fluid approximations with stochastic Lyapunov theory. Thus, let us first introduce some necessary notations and useful lemmas. In order to distinguish from stochastic analysis, we define $\mathcal{X} \triangleq (\mathcal{X}(t), t = 0, 1, 2, \ldots)$, in which $\mathcal{X}(t) \triangleq (Q_1(t), Q_2(t), \ldots, Q_n(t))$ in fluid domain. Then under our assumption and the queue-length based policy, $\mathcal{X} = (\mathcal{X}(t), t = 0, 1, 2, \ldots)$ is a discrete-time countable Markov chain. That is, the system state is denoted by $\mathcal{X}$ in the fluid approximation analysis. To establish the fluid model of $\mathcal{X}$, we need several notations. Let us define the norm of $\mathcal{X}(t)$ as $\|\mathcal{X}(t)\|_1 \triangleq \sum_{n=1}^N Q_n(t)$. Let $\mathcal{X}^{(x)}$ denote a process $\mathcal{X}$ with an initial state satisfying

$$\left\|\mathcal{X}^{(x)}(0)\right\|_1 = x. \tag{16}$$

Let $\mathcal{A}_i(t)$ and $\mathcal{D}_i(t)$ denote the *accumulated* arrival and actual departure tasks at queue i up to time-slot t, respectively. $\mathcal{A}_\Sigma(\tau)$ denotes the *accumulated* exogenous arrivals for a given τ units of time-slots. $\mathcal{S}_i(\tau)$ denotes the *accumulated* offered service for queue i during a given τ units of time-slots. Moreover, let $\mathcal{G}_i(t)$ denote the accumulated number of time-slots up to time-slot t in which the new arrivals are routed to queue i, and let $\mathcal{B}_i(t) \triangleq \sum_0^t 1\{Q_i(s) > 0\}$ denote the accumulated number of time-slots up to time-slot t in which queue i is busy. We also adopt the convention that $\mathcal{A}_i(0) = 0$, $\mathcal{D}_i(0) = 0$, $\mathcal{G}_i(0) = 0$ and $\mathcal{B}_i(0) = 0$. Therefore, we have $\mathcal{A}_i(t) = \mathcal{A}_\Sigma(G_i(t)) \leq \mathcal{A}_\Sigma(t)$ and $\mathcal{D}_i(t) = \mathcal{S}_i(\mathcal{B}_i(t)) \leq \mathcal{S}_i(t)$. Then the queue length Q_i can be described in an alternative form as follows

$$Q_i(t) = Q_i(0) + \mathcal{A}_i(t) - \mathcal{D}_i(t). \tag{17}$$

Let us define another process $\mathcal{Y} \triangleq (Q, \mathcal{A}, \mathcal{D}, \mathcal{A}_\Sigma, \mathcal{S}, \mathcal{G}, \mathcal{B})$, i.e., a tuple that denotes a list of processes, and clearly, a sample path of $\mathcal{Y}^{(x)}$ uniquely determines the sample path of $\mathcal{X}^{(x)}$. Then, we extend the definition of $\mathcal{Y}$ to each continuous time $t \geq 0$ as $\mathcal{Y}^{(x)}(t) \triangleq \mathcal{Y}^{(x)}(\lfloor t \rfloor)$. Recall that a sequence of functions $f_n(\cdot)$ is said to converge to a function $f(\cdot)$ uniformly over compact (u.o.c) interval if for all $t \geq 0$, $\lim_{n\to\infty} \sup_{0\leq t'\leq t} \left|f_n(t') - f(t')\right| = 0$. We now consider a sequence of process $\left\{\frac{1}{x_n}\mathcal{Y}^{(x_n)}(x_n\cdot)\right\}$, which is scaled both in time and space, and show the convergence properties of the sequence in the following lemma.

Lemma 9. *With probability one, for any sequence of the process $\{\frac{1}{x_n}\mathcal{Y}^{(x_n)}(x_n\cdot)\}$, where x_n is a sequence of positive integers with $x_n \to \infty$, there exists a subsequence x_{n_k} with $x_{n_k} \to \infty$ as $k \to \infty$ such that the following u.o.c convergences hold:*

$$\frac{1}{x_{n_k}} Q_i^{(x_{n_k})}(x_{n_k}t) \to q_i(t) \tag{18}$$

$$\frac{1}{x_{n_k}} \mathcal{A}_i^{(x_{n_k})}(x_{n_k}t) \to a_i(t) \tag{19}$$

$$\frac{1}{x_{n_k}} \mathcal{D}_i^{(x_{n_k})}(x_{n_k}t) \to d_i(t) \tag{20}$$

$$\frac{1}{x_{n_k}} \mathcal{A}_\Sigma^{(x_{n_k})}(x_{n_k}t) \to a_\Sigma(t) \tag{21}$$

$$\frac{1}{x_{n_k}} \mathcal{S}_i^{(x_{n_k})}(x_{n_k}t) \to s_i(t) \tag{22}$$

$$\frac{1}{x_{n_k}} \mathcal{G}_i^{(x_{n_k})}(x_{n_k}t) \to g_i(t) \tag{23}$$

$$\frac{1}{x_{n_k}} \mathcal{B}_i^{(x_{n_k})}(x_{n_k}t) \to b_i(t) \tag{24}$$

where q_i, a_i, d_i, a_Σ, s_i, g_i and b_i are some Lipschitz continuous functions in $[0, \infty)$. Hence all the functions are differentiable at almost every time $t \in [0, \infty)$, which is called regular time.

Proof. See Appendix B. □

The fluid model of our considered load balancing system is given by the following lemma. Note that the fluid model holds for any work-conserving FIFO and queue-length based policy. By utilizing the random time-change theorem in Chapter 5 of [31], we have the following results.

Lemma 10. *Any fluid limit $(q_i, a_i, d_i, a_\Sigma, s_i, g_i, b_i)$ satisfies the following equations*

$$q_i(t) = q_i(0) + a_i(t) - d_i(t) \tag{25}$$

$$a_i(t) = \lambda g_i(t) \tag{26}$$

$$d_i(t) = \mu_i b_i(t) \tag{27}$$

$$a_\Sigma(t) = \lambda t \tag{28}$$

$$s_i(t) = \mu_i t \tag{29}$$

$$\sum_{i}^{n} g_i(t) = t \tag{30}$$

and for any regular time t, we have

$$q_i'(t) = \begin{cases} \lambda g_i'(t) - \mu_i, & q_i(t) > 0. \\ 0, & q_i(t) = 0. \end{cases} \tag{31}$$

Proof. See Appendix C. □

Now we are well prepared to present the proof of Proposition 8 on throughput optimality.

Proof of Proposition 8. First, recall the permutation $\sigma_t(\cdot)$ of $(1, 2, \ldots, N)$ which satisfies $Q_{\sigma_t(1)}(t) \le Q_{\sigma_t(2)}(t) \le \ldots Q_{\sigma_t(N)}(t)$ and ties are broken randomly. Now we can establish the following claim, the proof of which is relegated to Appendix D.

Claim 1. *If $q_{\sigma_t(1)}(t) = q_{\sigma_t(2)}(t) = q_{\sigma_t(m)}(t) = 0 < q_{\sigma_t(m+1)}(t) \le \ldots \le q_{\sigma_t(N)}(t)$ for some $1 \le m < N$, then*

$$\sum_{n=m+1}^{N} g'_{\sigma_t(n)} = \sum_{n=m+1}^{N} \left(\Delta_n(t) + \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma} \right),$$

and $\Delta(t)$ satisfies the conditions in Eqs. (10) and (11) of Theorem 3.

Now, let us consider a Lyapunov function $V(\mathbf{z}) \triangleq \|z\|_1$. We would like to show that $\dot{V}(\mathbf{q}(t)) \le -l$ for some constant $l > 0$ whenever $V(\mathbf{q}(t)) > 0$. There are two cases to consider.

Case 1: $q_i(t) > 0$ for all $i \in \{0, 1, \ldots, N\}$.

In this case, by Eqs. (31) and (30), we have

$$\dot{V}(\mathbf{q}(t)) = \sum_{n=1}^{N} q_n'(t) = \lambda \Big(\sum_{n=1}^{N} g_n'(t) \Big) - \sum_{n=1}^{N} \mu_i = \lambda - \sum_{n=1}^{N} \mu_n = -\epsilon. \tag{32}$$

Case 2: For some $1 \le m < N$, $q_{\sigma_t(1)}(t) = q_{\sigma_t(2)}(t) = q_{\sigma_t(m)}(t) = 0 < q_{\sigma_t(m+1)}(t) \le \ldots \le q_{\sigma_t(N)}(t)$.

In this case, we have

$$\begin{aligned} \dot{V}(\mathbf{q}(t)) &\overset{(a)}{=} \sum_{n=m+1}^{N} q_n'(t) \\ &\overset{(b)}{=} \sum_{n=m+1}^{N} \lambda \left(\Delta_n(t) + \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma} \right) - \sum_{n=m+1}^{N} \mu_{\sigma_t(n)} \\ &\overset{(c)}{\le} \sum_{n=m+1}^{N} \lambda \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma} - \sum_{n=m+1}^{N} \mu_{\sigma_t(n)} \\ &\overset{(d)}{\le} -\epsilon \frac{\mu_{\min}}{\mu_\Sigma} \end{aligned}$$

where (a) comes from Eq. (31); (b) follows from Claim 1; (c) holds due to the fact that $\sum_{n=m+1}^{N} \Delta_n(t) \le 0$ when $\Delta(t)$ satisfies Eqs. (10) and (11); (d) is true since $\lambda = \mu_\Sigma - \epsilon$ and $\mu_{\min} = \min_{1 \le n \le N}(\mu_n)$.

Therefore, combining the above two cases, yields

$$\dot{V}(\mathbf{q}(t)) \le -l \text{ whenever } V(\mathbf{q}(t)) > 0$$

where $l \triangleq -\epsilon \mu_{\min}/\mu_\Sigma > 0$. This result implies that for any $\gamma \in (0, 1)$, there exists a finite T such that $V(\mathbf{q}(T)) \le \gamma$. Now, consider any fixed sequence of processes $\{\mathcal{X}^{(x)}, x = 1, 2, \ldots\}$ (for simplicity also denoted as $\{x\}$). Then, from the convergence in Lemma 9, we have that for any subsequence $\{x_n\}$ of $\{x\}$, there exists a further (sub)subsequence $\{x_{n_k}\}$ with probability one such that

$$\lim_{k \to \infty} \frac{1}{x_{n_k}} \left\| \mathcal{X}^{(x_{n_k})}(x_{n_k} T) \right\|_1 = \sum_{i}^{n} |q_i(T)| \le \gamma \triangleq 1 - \xi.$$

This further implies that with probability one,

$$\limsup_{x\to\infty}\left[\frac{1}{x}\left\|\mathcal{X}^{(x)}(xT)\right\|_1\right]\le 1-\xi$$

holds, because there is always a subsequence of $\{x\}$ that converges to the same limit as $\limsup_{x\to\infty}\left[\frac{1}{x}\left\|\mathcal{X}^{(x)}(xT)\right\|_1\right]$.

According to Eq. (17), we have $\left\|\mathcal{X}^{(x)}(xT)\right\|_1\le x+\sum_{i=1}^{N}\mathcal{A}_i(xT)$. Hence,

$$\mathbb{E}\left[\frac{1}{x}\left\|\mathcal{X}^{(x)}(xT)\right\|_1\right]\le 1+\lambda T\le\infty.$$

Therefore, from the dominated convergence theorem, we have

$$\limsup_{x\to\infty}\mathbb{E}\left[\frac{1}{x}\left\|\mathcal{X}^{(x)}(xT)\right\|_1\right]=\mathbb{E}\left[\limsup_{x\to\infty}\frac{1}{x}\left\|\mathcal{X}^{(x)}(xT)\right\|_1\right]\le 1-\xi.$$

This result in turn implies that there exists an x_0 such that for all $x=\left\|\mathcal{X}^{(x)}(0)\right\|_1\ge x_0$

$$\mathbb{E}\left[\left\|\mathcal{X}^{(x)}(xT)\right\|_1-\left\|\mathcal{X}^{(x)}(0)\right\|_1\right]\le-\frac{\xi x_0}{2}. \tag{33}$$

Now, let us turn to the stochastic analysis of the Lyapunov drift. In particular, we consider the mean drift of Lyapunov function $V(\mathbf{Q}(t))=\|\mathbf{Q}(t)\|_1$. We need to show that the Lyapunov function $V(.)$ satisfies the conditions (C1) and (C2) in Lemma 7, respectively.

For Condition (C2), we have

$$\begin{aligned}|\Delta V(\mathbf{Q})|&=\left|\|\mathbf{Q}(t_0+T)\|_1-\|\mathbf{Q}(t_0)\|_1\right|\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\&\overset{(a)}{\le}\|\mathbf{Q}(t_0+T)-\mathbf{Q}(t_0)\|_1\,\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\&\overset{(b)}{\le}TN\max(A_{\max},S_{\max})\end{aligned}$$

where (a) follows from the fact that $|\|\mathbf{x}\|_1-\|\mathbf{y}\|_1|\le\|\mathbf{x}-\mathbf{y}\|_1$ holds for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^N$; (b) holds due to the assumptions that the $A_\Sigma(t)\le A_{\max}$ and $S_n(t)\le S_{\max}$ for all $t\ge 0$ and all $1\le n\le N$, and are independent of the queue length. This establishes the condition (C2) in Lemma 7.

For Condition (C1), we have

$$\begin{aligned}&\mathbb{E}\left[\Delta V(\mathbf{Q})\mid\mathbf{Q}(t_0)=\mathbf{Q}\right]\\=&\mathbb{E}\left[\|\mathbf{Q}(t_0+T_1)\|_1-\|\mathbf{Q}(t_0)\|_1\mid\mathbf{Q}(t_0)=\mathbf{Q}\right]\\\overset{(a)}{=}&\mathbb{E}\left[\|\mathbf{Q}(T_1)\|_1-\|\mathbf{Q}(0)\|_1\mid\mathbf{Q}(0)=\mathbf{Q}\right]\\\overset{(b)}{\le}&-\frac{\xi x_0}{2}\end{aligned}$$

where (a) follows from the i.i.d assumption of exogenous arrival and service, and the system is Markovian with respect to the vector of queue lengths; (b) holds for $T_1=x_0T$ and $V(\mathbf{Q}(0))\ge x_0$. This directly comes from Eq. (33) and the fact $\|\mathcal{X}(t)\|_1=\sum_{n=1}^{N}Q_n(t)$. Hence, it establishes the condition (C1) in Lemma 7, and thus throughput optimality. □

4.2.2. State-space collapse to cone

We would prove the following result in this subsection, which combined with the throughput optimality in the last subsection directly implies heavy-traffic delay optimality according to Theorem 2.

Proposition 11. *Under the condition of Theorem 3, the state-space in steady-state collapses to the cone $\mathcal{K}_\alpha$, i.e., there exists $\epsilon_0=\mu_\Sigma\delta/(4N+2\delta)$ such that for all $\epsilon\in(0,\epsilon_0)$*

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|^r\right]\le M_r \tag{34}$$

holds for each $r=1,2,\ldots$, in which M_r are constants that are independent of ϵ.

Before we prove Proposition 11, we first define the following Lyapunov functions and their corresponding drifts.

$$V_\perp(\mathbf{Q})\triangleq\|\mathbf{Q}_\perp\|,\ W(\mathbf{Q})\triangleq\|\mathbf{Q}\|^2\ \text{and}\ W_\parallel(\mathbf{Q})\triangleq\left\|\mathbf{Q}_\parallel\right\|^2$$

with the corresponding one time-slot drift given by

$$\begin{aligned}\Delta V_\perp(\mathbf{Q})&\triangleq[V_\perp(\mathbf{Q}(t_0+1))-V_\perp(\mathbf{Q}(t_0))]\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\\Delta W(\mathbf{Q})&\triangleq[W(\mathbf{Q}(t_0+1))-W(\mathbf{Q}(t_0))]\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\\Delta W_\parallel(\mathbf{Q})&\triangleq[W_\parallel(\mathbf{Q}(t_0+1))-W_\parallel(\mathbf{Q}(t_0))]\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\end{aligned}$$

Now, we are ready to prove Proposition 11.

Proof of Proposition 11. To establish the bounded moments of $\|\mathbf{Q}_\perp\|$, based on Lemma 7, all we need to show is that the drift of Lyapunov function $V_\perp(.)$ satisfies the two conditions for all $\epsilon \in (0, \epsilon_0)$. For condition (C2), we have

$$\begin{aligned}
&|\Delta V_\perp(\mathbf{Q})|\\
=&|\|\mathbf{Q}_\perp(t_0+1)\| - \|\mathbf{Q}_\perp(t_0)\|\,|\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\
\overset{(a)}{\leq}& \|\mathbf{Q}_\perp(t_0+1) - \mathbf{Q}_\perp(t_0)\|\,\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\
\overset{(b)}{\leq}& \|\mathbf{Q}(t_0+1) - \mathbf{Q}(t_0)\|\,\mathcal{I}(\mathbf{Q}(t_0)=\mathbf{Q})\\
\overset{(c)}{\leq}& \sqrt{N}\max(A_{\max}, S_{\max})
\end{aligned} \tag{35}$$

where (a) follows from the fact that $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x}-\mathbf{y}\|$ holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$; (b) follows from the non-expansive property of projection and the fact that $\mathbf{Q}_\perp$ is the projection onto the convex closed cone $\mathcal{K}_\alpha^\circ$. (c) holds due to the assumptions that the $A_\Sigma(t) \leq A_{\max}$ and $S_n(t) \leq S_{\max}$ for all $t \geq 0$ and all $1 \leq n \leq N$, and are both independent of queue lengths. This verifies Condition (C2) in Lemma 7.

For condition (C1), we need the following result, the proof of which is relegated to Appendix E.

Claim 2. *For any $t \geq 0$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\Delta V_\perp(\mathbf{Q}) \mid \mathbf{Q}(t)=\mathbf{Q}\right]\\
\leq&\frac{1}{2\,\|\mathbf{Q}_\perp(t)\|}\mathbb{E}\left[(2\langle \mathbf{Q}_\perp(t), \mathbf{A}(t)-\mathbf{S}(t)\rangle + L) \mid \mathbf{Q}(t)=\mathbf{Q}\right]
\end{aligned}$$

where $L \triangleq N\max(A_{\max}, S_{\max})^2$.

Thus, based on Claim 2, in order to establish condition (C1) for all $\epsilon \in (0, \epsilon_0)$, it suffices to show

$$\mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t)-\mathbf{S}(t)\rangle \mid \mathbf{Q}(t)=\mathbf{Q}\right] \leq -c\,\|\mathbf{Q}_\perp(t)\| \tag{36}$$

holds for all $\epsilon \in (0, \epsilon_0)$, and c is independent of ϵ.

To this end, first recall the permutation $\sigma_t(\cdot)$ of $(1, 2, \ldots, N)$ which satisfies $Q_{\sigma_t(1)}(t) \leq Q_{\sigma_t(2)}(t) \leq \ldots Q_{\sigma_t(N)}(t)$ and ties are broken randomly. In the following, for simplicity of notation, we let $\widehat{\mathbf{Q}}(t) = (Q_{\sigma_t(1)}(t), Q_{\sigma_t(2)}(t), \ldots, Q_{\sigma_t(N)}(t))$, and similarly the arrival process $\widehat{\mathbf{A}}(t) = (A_{\sigma_t(1)}(t), A_{\sigma_t(2)}(t), \ldots, A_{\sigma_t(N)}(t))$ and the service vector $\widehat{\mathbf{S}}(t) = (S_{\sigma_t(1)}(t), S_{\sigma_t(2)}(t), \ldots, S_{\sigma_t(N)}(t))$. Now, the left-hand-side of Eq. (36) can be written as follows.

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t)-\mathbf{S}(t)\rangle \mid \mathbf{Q}(t)=\mathbf{Q}\right]\\
\overset{(a)}{=}&\mathbb{E}\left[\langle \widehat{\mathbf{Q}}_\perp(t), \widehat{\mathbf{A}}(t)-\widehat{\mathbf{S}}(t)\rangle \mid \mathbf{Q}(t)=\mathbf{Q}\right]\\
\overset{(b)}{=}&\sum_{n=1}^N \widehat{Q}_{\perp n}\left[\lambda_\Sigma\left(\Delta_n(t) + \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma}\right) - \mu_{\sigma_t(n)}\right]\\
\overset{(c)}{=}&\sum_{n=1}^N \widehat{Q}_{\perp n}\Delta_n(t)\lambda_\Sigma + \sum_{n=1}^N \widehat{Q}_{\perp,n}\left(-\epsilon\frac{\mu_{\sigma_t(n)}}{\mu_\Sigma}\right)\\
\leq&\sum_{n=1}^N \widehat{Q}_{\perp n}\Delta_n(t)\lambda_\Sigma + \epsilon\left\|\widehat{\mathbf{Q}}_\perp(t)\right\|_1
\end{aligned} \tag{37}$$

where (a) comes from the fact that the cone $\mathcal{K}_\alpha$ is symmetric with respect to the line $\mathbf{1} = (1, 1, \ldots, 1)$; In (b), $\widehat{Q}_{\perp n}$ is the nth component of the vector $\widehat{\mathbf{Q}}_\perp(t)$ and (b) holds because of the definition of $\Delta(t)$ in Eq. (7), and the fact that the service process is independent of queue lengths; (c) follows from the fact that $\lambda_\Sigma = \mu_\Sigma - \epsilon$.

Now, let us focus on the first term of Eq. (37). To establish an upper bound on it, we will first establish the following important monotone property of $\widehat{\mathbf{Q}}_\perp(t)$. That is,

$$\widehat{Q}_{\perp 1}(t) \leq \widehat{Q}_{\perp 2}(t) \leq \cdots \leq \widehat{Q}_{\perp N}(t). \tag{38}$$

First, in the case of $\alpha = 1$, the cone $\mathcal{K}_\alpha$ reduces to the line $\mathbf{1}$. Thus, it can be easily obtained that $\widehat{Q}_{\perp n}(t) = Q_{\sigma_t(n)}(t) - Q_{\text{avg}}(t)$ where $Q_{\text{avg}}(t) = \sum_{n=1}^N Q_n(t)/N$, which satisfies the monotone property. Hence, we are only left with the task of establishing the monotone property for the case of $\alpha \in (0, 1)$.

Note that, since $\widehat{\mathbf{Q}}_\parallel(t) \in \mathcal{K}_\alpha$, we have $\widehat{\mathbf{Q}}_\parallel(t) = \sum_{n=1}^N w_n \mathbf{b}^{(n)}$, where $w_n \geq 0$. Let $\mathcal{I}$ be a subset of $\{1, 2, \ldots, N\}$ such that for any $i \in \mathcal{I}$ $w_i > 0$ and for any $i \notin \mathcal{I}$, $w_i = 0$, i.e., the subset $\mathcal{I}$ contains all the index n such that $w_n > 0$. It suffices to consider

the case when $\mathcal{I}$ is nonempty. This is because when $\mathcal{I}$ is empty, we have $\widehat{\mathbf{Q}}_{\|}(t) = 0$, which directly implies the monotone property since $\widehat{\mathbf{Q}}_{\perp}(t) = \widehat{\mathbf{Q}}(t)$ in this case.

Now consider the case when $\mathcal{I}$ is nonempty. First, we have

$$\langle \widehat{\mathbf{Q}}_{\perp}(t), \mathbf{b}^{(i)} \rangle \leq 0 \tag{39}$$

holds for any $i \in \{1, 2, \ldots, N\}$. Moreover, for any $i \in \mathcal{I}$

$$\langle \widehat{\mathbf{Q}}_{\perp}(t), \mathbf{b}^{(i)} \rangle = 0. \tag{40}$$

The inequality in Eq. (39) follows from the fact that $\mathbf{b}^{(i)} \in \mathcal{K}_{\alpha}$ and $\widehat{\mathbf{Q}}_{\perp}(t) \in \mathcal{K}_{\alpha}^{\circ}$. The equality in Eq. (40) follows from the fact that

$$0 = \langle \widehat{\mathbf{Q}}_{\perp}(t), \widehat{\mathbf{Q}}_{\|}(t) \rangle = \sum_{i \in \mathcal{I}} w_i \langle \widehat{\mathbf{Q}}_{\perp}(t), \mathbf{b}^{(i)} \rangle,$$

along with Eq. (39) and $w_i > 0$ for all $i \in \mathcal{I}$. Eqs. (39) and (40) enable us to establish the following claim, the proof of which is relegated to Appendix F.

Claim 3. *If $m \in \mathcal{I}$ with $1 \leq m < N - 1$, then $m + 1 \in \mathcal{I}$.*

Note that Claim 3 directly implies that there exists an m_0 (which depends on $\mathbf{Q}(t)$) such that for all $i \geq m_0$, $i \in \mathcal{I}$ and for $i < m_0$, $i \notin \mathcal{I}$. Hence, by Eq. (40), for all $i \geq m_0$

$$0 = \langle \widehat{\mathbf{Q}}_{\perp}(t), \mathbf{b}^{(i)} \rangle = \alpha \sum_{n=1}^{N} \widehat{\mathbf{Q}}_{\perp n}(t) + (1 - \alpha) \widehat{Q}_{\perp i}(t),$$

which implies that for all $i \geq m_0$

$$\widehat{\mathbf{Q}}_{\perp i}(t) = c \geq 0 \tag{41}$$

for some constant c. This is because $\sum_{n=1}^{N} \widehat{\mathbf{Q}}_{\perp n}(t) \leq 0$, due to the fact that $\mathbf{1} \in \mathbf{K}_{\alpha}$ and $\widehat{\mathbf{Q}}_{\perp}(t) \in \mathcal{K}_{\alpha}^{\circ}$. On the other hand, for any $i \leq j < m_0$, we have

$$\widehat{\mathbf{Q}}_{\perp i}(t) \leq \widehat{\mathbf{Q}}_{\perp j}(t). \tag{42}$$

This holds since $\widehat{\mathbf{Q}}_{\|i}(t) = \widehat{\mathbf{Q}}_{\|j}(t)$ and $\widehat{Q}_i(t) \leq \widehat{Q}_j(t)$. Moreover, we have

$$\widehat{\mathbf{Q}}_{\perp m_0}(t) \geq \widehat{\mathbf{Q}}_{\perp (m_0 - 1)}(t). \tag{43}$$

This can be shown by contradiction. Suppose $\widehat{\mathbf{Q}}_{\perp (m_0-1)}(t) > \widehat{\mathbf{Q}}_{\perp m_0}(t)$, then

$$\begin{aligned}
\langle \widehat{\mathbf{Q}}_{\perp}(t), \mathbf{b}^{(m_0 - 1)} \rangle &= \alpha \sum_{n=1}^{N} \widehat{\mathbf{Q}}_{\perp n}(t) + (1 - \alpha) \widehat{Q}_{\perp (m_0 - 1)}(t) \\
&> \alpha \sum_{n=1}^{N} \widehat{\mathbf{Q}}_{\perp n}(t) + (1 - \alpha) \widehat{Q}_{\perp m_0}(t) \\
&= \langle \widehat{\mathbf{Q}}_{\perp}(t), \mathbf{b}^{(m_0)} \rangle \\
&= 0
\end{aligned}$$

which contradicts with Eq. (39). Then, combining Eqs. (41) and (42) and (43), yields the fact that $\widehat{\mathbf{Q}}_{\perp N}(t) \geq 0$ and the monotone property in Eq. (38). As a result, we have $\widehat{\mathbf{Q}}_{\perp 1}(t) \leq 0$ since otherwise $\sum_{n=1}^{N} \widehat{\mathbf{Q}}_{\perp, n}(t)$ would be strictly positive.

Having established the monotone property of $\widehat{\mathbf{Q}}_{\perp}(t)$ and auxiliary results that $\widehat{\mathbf{Q}}_{\perp N}(t) \geq 0$ and $\widehat{\mathbf{Q}}_{\perp 1}(t) \leq 0$, we can now proceed to obtain an upper bound on the first term in Eq. (37). In particular, we can first bound it in terms of $|\widehat{\mathbf{Q}}_{\perp 1}(t)|$ and the δ in Eq. (11). In particular, we have

$$\sum_{n=1}^{N} \widehat{Q}_{\perp n} \Delta_n(t) \lambda_{\Sigma} \leq -\lambda_{\Sigma} \delta |\widehat{\mathbf{Q}}_{\perp 1}(t)|. \tag{44}$$

This upper bound can be verified as follows. First, if $\mathbf{Q}(t) \in \mathcal{K}_{\alpha}$, then $\widehat{\mathbf{Q}}_{\perp n}(t) = 0$ for all n, and hence Eq. (44) holds. If $\mathbf{Q}(t) \notin \mathcal{K}_{\alpha}$, then $\Delta(t)$ satisfies the two conditions in Eqs. (10) and (11) in Theorem 3, which specify the construction process of $\Delta(t)$. In particular, each $\Delta(t)$ that satisfies the two conditions can be constructed as follows. To begin with, all the $\Delta_n(t)$ is 0. Then, according to the condition in Eq. (11), we should first decrease $\Delta_N(t)$ by the amount of δ, and increase $\Delta_1(t)$ by the amount of δ. After this, the left-hand-side of Eq. (44) is equal to $\lambda_{\Sigma} \left(\delta \widehat{\mathbf{Q}}_{\perp 1}(t) + (-\delta) \widehat{\mathbf{Q}}_{\perp N}(t) \right)$, which is upper bounded by $-\lambda_{\Sigma} \delta |\widehat{\mathbf{Q}}_{\perp 1}(t)|$, since $\widehat{\mathbf{Q}}_{\perp N}(t) \geq 0$ and $\widehat{\mathbf{Q}}_{\perp 1}(t) \leq 0$. Next, due to the condition in Eq. (10) and the fact that $\sum_{n=1}^{N} \Delta_n(t) = 0$,

any further procedure (if needed) for the construction of $\Delta(t)$ can only take the following way: it decreases some $\Delta_i(t)$ by a certain amount (say c_1) where $i \geq k$, and then increase some $\Delta_j(t)$ by the same amount c_1 where $j \leq k$. We claim that any of this procedure cannot increase the value of the left-hand-side of Eq. (44) due to the monotone property of $\widehat{\mathbf{Q}}_\perp(t)$. To see this, let us denote by $\mathcal{L}_{ij}$ the change of the value of the left-hand-side of Eq. (44) incurred by the procedure above. Thus, we have

$$\mathcal{L}_{ij} = -c_1\widehat{Q}_{\perp i}(t) + c_1\widehat{Q}_{\perp j}(t) \leq 0,$$

which follows from the monotone property of $\widehat{\mathbf{Q}}_\perp(t)$. Therefore, we have verified the upper bound in Eq. (44).

Next, we establish an upper bound on $\|\widehat{\mathbf{Q}}_\perp(t)\|_1$ in terms of $|\widehat{\mathbf{Q}}_{\perp 1}(t)|$ as follows

$$\left\|\widehat{\mathbf{Q}}_\perp(t)\right\|_1 \leq 2N|\widehat{\mathbf{Q}}_{\perp 1}(t)|. \tag{45}$$

This follows from the monotone property of $\widehat{\mathbf{Q}}_\perp(t)$ and the fact that $\sum_{n=1}^N \widehat{\mathbf{Q}}_{\perp n}(t) \leq 0$. Now, combining Eqs. (37) and (44) and (45), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid \mathbf{Q}(t) = \mathbf{Q}\right] \\
&\leq \left(\epsilon - \frac{\lambda_\Sigma \delta}{2N}\right) \left\|\widehat{\mathbf{Q}}_\perp(t)\right\|_1 \\
&\leq -\frac{\mu_\Sigma \delta}{4N} \left\|\widehat{\mathbf{Q}}_\perp(t)\right\|_1 \text{ whenever } \epsilon \leq \frac{\mu_\Sigma \delta}{4N + 2\delta} \\
&\leq -\frac{\mu_\Sigma \delta}{4N} \|\mathbf{Q}_\perp(t)\|
\end{aligned} \tag{46}$$

where the last inequality comes from the fact that $\|\widehat{\mathbf{Q}}_\perp(t)\|_1 = \|\mathbf{Q}_\perp(t)\|_1$ and $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^N$. This establishes the inequality in Eq. (36) with $c = \mu_\Sigma\delta/4N$ and $\epsilon_0 = \mu_\Sigma\delta/(4N + 2\delta)$. Hence, based on Claim 2, we have verified the condition (C1) in Lemma 7, which directly establishes the state-space collapse result in Proposition 11. □

5. Conclusions

We have rigorously shown that even under a multi-dimensional state-space collapse, steady-state heavy-traffic delay optimality can be achieved for a general load balancing system. This result suggests that the insight behind heavy-traffic optimality conveyed by diffusion approximations is still valid in *steady state*, thus complementing and extending the diffusion approximation results in [8,9]. Moreover, our steady-state delay optimality result might also give a possible direction for proving the interchange of limits for the diffusion approximation results in [8,9]. By leveraging this result, we are able to explore the greater flexibility provided by allowing a multi-dimensional state-space collapse in designing new load balancing policies that are both throughput optimal and heavy-traffic delay optimal in steady state. Furthermore, the proof techniques used in this paper are of independent interest as well.

Appendix. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.peva.2018.10.003.

References

[1] V. Gupta, M.H. Balter, K. Sigman, W. Whitt, Analysis of join-the-shortest-queue routing for web server farms, Perform. Eval. 64 (9) (2007) 1062–1081.
[2] L. George, HBase: The Definitive Guide, O'Reilly Media, Inc, 2011.
[3] I. Foster, Y. Zhao, I. Raicu, S. Lu, Cloud computing and grid computing 360-degree compared, in: 2008 Grid Computing Environments Workshop, IEEE, 2008, pp. 1–10.
[4] G. Foschini, J. Salz, A basic dynamic routing problem and diffusion, IEEE Trans. Commun. 26 (3) (1978) 320–327.
[5] A. Eryilmaz, R. Srikant, Asymptotically tight steady-state queue length bounds implied by drift conditions, Queueing Syst. 72 (3–4) (2012) 311–359.
[6] H. Chen, H.-Q. Ye, Asymptotic optimality of balanced routing, Oper. Research 60 (1) (2012) 163–179.
[7] S.T. Maguluri, R. Srikant, L. Ying, Heavy traffic optimal resource allocation algorithms for cloud computing clusters, Perform. Eval. 81 (2014) 20–39.
[8] F. Kelly, C. Laws, Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling, Queueing Syst. 13 (1–3) (1993) 47–86.
[9] Y.-C. Teh, A.R. Ward, Critical thresholds for dynamic routing in queueing networks, Queueing Syst. 42 (3) (2002) 297–316.
[10] M.I. Reiman, Some diffusion approximations with state space collapse, in: Modelling and Performance Evaluation Methodology, Springer, 1984, pp. 207–240.
[11] S.L. Bell, R.J. Williams, Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy, Ann. Appl. Probab. (2001) 608–649.
[12] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks, Queueing Syst. 30 (1–2) (1998) 89–140.
[13] J.M. Harrison, Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies, Ann. Appl. Probab. (1998) 822–848.
[14] D. Gamarnik, A. Zeevi, et al., Validity of heavy traffic steady-state approximations in generalized Jackson networks, Ann. Appl. Probab. 16 (1) (2006) 56–90.
[15] A. Budhiraja, C. Lee, Stationary distribution convergence for generalized Jackson networks in heavy traffic, Math. Oper. Res. 34 (1) (2009) 45–56.

[16] X. Zhou, F. Wu, J. Tan, Y. Sun, N. Shroff, Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms, 2017. arXiv preprint arXiv:1710.04357.
[17] W. Wang, K. Zhu, L. Ying, J. Tan, L. Zhang, Maptask scheduling in mapreduce with data locality: Throughput and heavy-traffic optimality, IEEE/ACM Trans. Netw. 24 (1) (2016) 190–203.
[18] Q. Xie, Y. Lu, Priority algorithm for near-data scheduling: Throughput and heavy-traffic optimality, in: Proceedings of IEEE International Conference on Computer Communications, INFOCOM, 2015, pp. 963–972.
[19] A.L. Stolyar, et al., Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic, Ann. Appl. Probab. 14 (1) (2004) 1–53.
[20] W. Kang, R. Williams, Diffusion approximation for an input-queued packet switch operating under a maximum weight algorithm, Stoch. Syst. (2012).
[21] S.T. Maguluri, R. Srikant, et al., Heavy traffic queue length behavior in a switch under the MaxWeight algorithm, Stoch. Syst. 6 (1) (2016) 211–250.
[22] S.T. Maguluri, S.K. Burle, R. Srikant, Optimal heavy-traffic queue length scaling in an incompletely saturated switch, Queueing Syst. 88 (3–4) (2018) 279–309.
[23] W. Wang, S.T. Maguluri, R. Srikant, L. Ying, Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing, ACM SIGMETRICS Perform. Eval. Rev. 45 (2) (2018) 232–245.
[24] M. Armony, Dynamic routing in large-scale service systems with heterogeneous servers, Queueing Syst. 51 (3–4) (2005) 287–329.
[25] I. Gurvich, W. Whitt, Queue-and-idleness-ratio controls in many-server service systems, Math. Oper. Res. 34 (2) (2009) 363–396.
[26] J. Dai, T. Tezcan, State space collapse in many-server diffusion limits of parallel server systems, Math. Oper. Res. 36 (2) (2011) 271–320.
[27] Q. Xie, A. Yekkehkhany, Y. Lu, Scheduling with multi-level data locality: Throughput and heavy-traffic optimality, in: Proceedings of IEEE International Conference on Computer Communications, INFOCOM, 2016, pp. 1–9.
[28] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, I. Stoica, Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling, in: Proceedings of the 5th European conference on Computer systems, ACM, 2010, pp. 265–278.
[29] G.S. Paschos, M. Leconte, A. Destounis, Routing with blinkers: Online throughput maximization without queue length information, in: 2016 IEEE International Symposium on Information Theory (ISIT), IEEE, 2016, pp. 1436–1440.
[30] J.G. Dai, S.P. Meyn, Stability and convergence of moments for multiclass queueing networks via fluid limit models, IEEE Trans. Automat. Control 40 (11) (1995) 1889–1904.
[31] H. Chen, D.D. Yao, Fundamentals of queueing networks: Performance, asymptotics, and optimization, Vol. 46, Springer Science & Business Media, 2013.

Xingyu Zhou is a Ph.D. student at the Department of Electrical and Computer Engineering of the Ohio State University, Columbus, OH, USA. His research interests lie in the broad area of applied probability, stochastic systems, optimization with applications in data centers, cloud computing, and energy-efficient communications. He received his B.S. degree from Beijing University of Posts and Telecommunications (BUPT) in 2012, Beijing, China, and his M.S. degree from Tsinghua University in 2015, Beijing, China, both in the Department of Electrical Engineering with the highest honor. He has received the student travel grants from ACM Sigmetrics conference and IFIP Performance conference. He is also the recipient of various awards, including the Outstanding Graduate Award of Beijing city in both 2012 and 2015, the National Scholarship of China, and the Academic Rising Star Award in electrical engineering of Tsinghua University. He served as a reviewer for journals including IEEE Journal on Selected Areas in Communications, IEEE Communications Surveys and Tutorials, IEEE Transactions on Network Science and Engineering, Performance Evaluation, IEEE Access, and conference including ACM Sigmetrics, ACM MobiHoc, INFOCOM, ICC, Globecom, GlobalSIP, WiOpt.

Jian Tan received the B.S. degree from University of Science and Technology of China, Hefei, China in 2002, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, USA in 2004 and 2008, respectively, all in Electrical Engineering. His Ph.D. thesis won the Eliahu Jury Award from Columbia University. From 2016, he is an assistant professor of the Department of Electrical and Computer Engineering at The Ohio State University, Columbus, OH, USA. Before that, he was with IBM T. J. Watson Research Center, Yorktown Heights, NY, USA and then the Institute of Data Science and Technology of Alibaba, Seattle, WA, USA. He was a postdoctoral researcher with the Networking and Communications Research Lab, The Ohio State University from 2009 to 2010. He interned with Lucent Bell Laboratories, Murray Hill, NJ, USA, during the summers of 2005 and 2006, and with Microsoft Research, Cambridge, UK, in the winter of 2007. His current research interests focus on stochastic modeling, distributed computing systems, and statistical learning. At IBM Research he focused on using mathematical modeling techniques to analyze large-scale stochastic systems and optimizing the performance for distributed computing systems. At Alibaba, he worked on statistical learning algorithms for recommendations. His work has been integrated into the recommendation engine of T-Mall, a production system deployed at Alibaba.

Ness B. Shroff (S'91-M'93-SM'01-F'07) received the Ph.D. degree in electrical engineering from Columbia University in 1994. He joined Purdue University immediately thereafter as an Assistant Professor with the School of Electrical and Computer Engineering. At Purdue, he became a Full Professor of ECE and the Director of CWSA in 2004, a university-wide center on wireless systems and applications. In 2007, he joined The Ohio State University, where he holds the Ohio Eminent Scholar Endowed Chair in networking and communications, in the departments of ECE and CSE. He holds or has held visiting (chaired) professor positions at Tsinghua University, Beijing, China, Shanghai Jiaotong University, Shanghai, China, and IIT Bombay, Mumbai, India. He is currently an Editor at Large of the IEEE/ACM TRANSACTIONS ON NETWORKING, and has served as Senior Editor of the IEEE TRANSACTIONS ON CONTROL OF NETWORKED SYSTEMS, and a Technical Editor of the IEEE Network Magazine. He has also served as the TPC co-chair and general chair of numerous IEEE and ACM conferences. He currently serves as the chair of the ACM Mobihoc steering committee. He has received numerous best paper awards for his research and is listed in Thomson Reuters' on The World's Most Influential Scientific Minds, and is noted as a Highly Cited Researcher by Thomson Reuters. He also received the IEEE INFOCOM Achievement Award for seminal contributions to scheduling and resource allocation in wireless networks.