Heavy-traffic Delay Optimality in Pull-based Load Balancing Systems: Necessary and Sufficient Conditions

XINGYU ZHOU, The Ohio State University
JIAN TAN, The Ohio State University
NESS SHROFF, The Ohio State University

In this paper, we consider a load balancing system under a general pull-based policy. In particular, each arrival is randomly dispatched to one of the servers with queue length below a threshold; if none exists, this arrival is randomly dispatched to one of the entire set of servers. We are interested in the fundamental relationship between the threshold and the delay performance of the system in heavy traffic. To this end, we first establish the following necessary condition to guarantee heavy-traffic delay optimality: the threshold will grow to infinity as the exogenous arrival rate approaches the boundary of the capacity region (i.e., the load intensity approaches one) but the growth rate should be slower than a polynomial function of the mean number of tasks in the system. As a special case of this result, we directly show that the delay performance of the popular pull-based policy Join-Idle-Queue (JIQ) lies strictly between that of any heavy-traffic delay optimal policy and that of random routing. We further show that a sufficient condition for heavy-traffic delay optimality is that the threshold grows logarithmically with the mean number of tasks in the system. This result directly resolves a generalized version of the conjecture by Kelly and Laws.

CCS Concepts: • **Mathematics of computing** → *Queueing theory*; • **Networks** → *Network performance modeling*; *Network performance analysis*;

Additional Key Words and Phrases: Heavy-traffic delay optimality; Pull-based; Load balancing; Necessary and sufficient conditions

ACM Reference format:
Xingyu Zhou, Jian Tan, and Ness Shroff. 2019. Heavy-traffic Delay Optimality in Pull-based Load Balancing Systems: Necessary and Sufficient Conditions. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 3, Article 44 (January 2019), 33 pages.
https://doi.org/10.1145/3287323

1 INTRODUCTION

We consider a classical load balancing system that consists of a central dispatcher and N servers, each associated with an infinite buffer queue and a service rate μ_n. The exogenous tasks arrive with rate λ_Σ, and upon arrival they must be immediately dispatched to one of the queues. A key to the performance of such a system is the load balancing policy it uses since it directly determines which queue the arriving tasks should join.

To design effective load balancing policies and hence provide good delay performance, it is imperative to develop analytical tools to evaluate the system performance under different load

This work has been funded in part through ONR grant N00014-17-1-2417 and NSF grants CNS-1719371, 1717060, and 1518829.

2476-1249/2019/1-ART44 $15.00
https://doi.org/10.1145/3287323

balancing policies. Towards that goal, one important line of research has focused on the so-called 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. An attractive property of the heavy-traffic regime, as pointed out in [15], is that *'the important features of good control policies are often displayed in the sharpest relief'*. It has been shown that well-known policies such as Join-Shortest-Queue (JSQ) and Power-of-d can achieve asymptotically optimal delay performance in the heavy-traffic regime [5, 7, 8, 19]. Under these two policies, an incoming task is assigned to a server with the shortest queue among $d \geq 2$ servers ($d = N$ for JSQ) sampled uniformly at random.

However, due to the sampling process, the amount of communication overhead is $2d$ per arrival (d for query and d for response), which is undesirable for a large value of d, especially in the JSQ policy when $d = N$. More importantly, since the dispatching decision can only be made after collecting the queue length feedback, there exists a non-zero dispatching delay, which contributes to an increase in the response time. To avoid these drawbacks, an alternative approach, often called pull-based load balancing, has received significant recent attention. Instead of actively sending queries to servers and waiting for responses, the dispatcher under a pull-based load balancing scheme passively listens to the reports from the servers. In particular, each server will report its ID to the dispatcher when it satisfies a certain condition (e.g., its queue length drops below a threshold from above). Then, upon task arrival, the dispatcher checks its record. If it is not empty, the dispatcher randomly removes one ID and sends the arrival to the corresponding server; otherwise, it just randomly selects a queue to join. The classical pull-based policy is the Join-Idle-Queue (JIQ) policy investigated in [16, 22], under which the dispatcher maintains a record of IDs of the idle servers (i.e., the reporting threshold is one). JIQ has been shown to enjoy a low message overhead (at most one per arrival), zero dispatching delay, and better delay performance than Power-of-d under medium loads. Nevertheless, under high loads, its delay performance degrades substantially due to the lack of idle servers. This directly suggests that a varying reporting threshold with respect to the load is necessary to guarantee good delay performance in heavy traffic. Motivated by this observation, in a recent work [30], the authors propose a specific way to update the reporting threshold in a pull-based policy, which is proven to be heavy-traffic delay optimal, while still enjoying many of the nice features of JIQ.

In this paper, instead of focusing on another specific way of determining the reporting threshold, we step back and work towards answering the following fundamental question: *How would different reporting thresholds affect the (heavy traffic) delay performance of a pull-based policy?* To address this question, we take a systematic approach and summarize the main contributions as follows.

- We first present a necessary condition on the reporting threshold for the delay optimality of a pull-based policy in heavy-traffic. In particular, we show that to achieve heavy-traffic delay optimality, the reporting threshold r should grow to infinity as the heavy-traffic parameter ϵ approaches zero, however, it cannot grow too fast (slower than a polynomial function: see Theorem 3.2). An important corollary of Theorem 3.2 is that the delay performance of the JIQ policy (i.e., constant threshold $r = 1$) in heavy traffic lies *strictly* between that of any heavy-traffic delay optimal policies (e.g., JSQ) and that of random routing. This result is somewhat counter-intuitive, since at first glance one may guess that JIQ would degenerate to random routing in heavy traffic since there are hardly any idle servers in the system. However, it turns out that it is not true, and allows us to get a sharp characterization of the JIQ policy in heavy traffic.
- We then establish a sufficient condition on the reporting threshold for heavy-traffic delay optimality of pull-based policies. Specifically, we show that a logarithmic growth rate of the reporting threshold with respect to the mean number of tasks in the system is sufficient to

guarantee the steady-state delay optimality in heavy traffic (see Theorem 3.3). This result directly resolves a conjecture by Kelly and Laws in [15]. In particular, the authors in [15] consider a two-server system with Poisson arrivals and exponential service under a varying reporting threshold. They conjecture that as long as the threshold is greater than a specified constant times the logarithm of the mean number of tasks in the system, then asymptotic delay optimality holds in heavy traffic. Thus, our result not only resolves the conjecture but generalizes it to any fixed finite number of servers with general arrival and service distributions. It is also worthing noting that the asymptotic delay optimality achieved in our paper is in steady-state while the result in [15] holds only for a finite time interval.

- The techniques introduced in this paper may be of independent interest for the analysis of general load balancing policies. More precisely, the key to establishing heavy-traffic delay optimality in this paper is a notion of state-space collapse, which is different from the state-space collapse result often adopted in previous works. As a result, it requires us to develop a new Lyapunov function to conduct the drift analysis. More importantly, due to this new type of state-space collapse, we have to devise a new approach to relate the state-space collapse result to the final heavy-traffic delay optimality.

1.1 Related Work

The investigation of queueing delay in heavy traffic with dynamic routing dates back to [8], in which the authors considered a two-server system under the JSQ policy, and they showed that the two separate servers under JSQ act as a pooled resource in heavy traffic via diffusion approximations. Since then, the methodology of diffusion approximations has been adopted in a number of works on parallel queues [3, 5, 13, 14, 21, 26]. For example, the author in [21] generalized the results in [8] to the case of renewal arrivals and general service times. The functional central limit theorems for the JSQ policy in a load balancing system with multiple servers was derived in [13]. In [5], the Power-of-d policy was shown to have the same diffusion limit as JSQ in the heavy-traffic limit. Many of the works based on the diffusion approximation method rely on showing that a scaled version of queue lengths converges to a regulated Brownian motion. This result typically leads to a sample-path optimality in a finite time interval. However, showing the convergence to the steady-state distribution requires the additional validation of the interchange of limits, which is often not taken (some exceptions include [4, 9], in which the authors proved an interchange of limit argument for generalized Jackson networks with a fixed routing matrix). Motivated by this, the authors in [7] proposed a Lyapunov drift-based approach, which is able to establish steady-state heavy-traffic optimality of the 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. This approach has been utilized to show steady-state heavy-traffic delay optimality of Power-of-d in [19]. Moreover, based on this approach, it has been shown in [25] that a joint JSQ and MaxWeight policy is heavy-traffic delay optimal for MapReduce clusters.

As discussed in the introduction, while JSQ and Power-of-d enjoy heavy-traffic delay optimality, they both have non-zero dispatching delay, and a relatively high message overhead. Motivated by this, a pull-based design of load balancing policies has gained significant recent popularity. The main feature of pull-based load balancing is the introduction of local memory at the dispatcher, which maintains a record of servers satisfying a pre-defined condition (e.g., its queue length is below a threshold in most cases). The dispatching decision is made purely based on the local memory: if it is nonempty, randomly choosing a server in memory to join; otherwise, randomly choosing a server from all the servers. For instance, one illustrative example is the JIQ policy proposed and studied in [16, 22], under which the local memory maintains all the idle servers. As a result, the

arrival is always dispatched to one of the idle servers if there are any; otherwise, it is dispatched randomly. It has been shown that JIQ has a low message overhead (at most one per arrival), zero dispatching delay, and better performance compared to Power-of-2 in medium loads. Nevertheless, since only the idle servers are stored in memory, when the loads become high, its performance degrades substantially because the memory is empty and hence random routing is adopted most of the time. Therefore, this directly suggests that a varying threshold is necessary to guarantee good performance in heavy traffic for a pull-based policy.

To this end, in a recent work [30], the authors successfully propose a pull-based policy with a varying threshold, which is proven to be heavy-traffic delay optimal in steady state while keeping the nice features of JIQ. This naturally raises the question about the fundamental relationship between the choice of the threshold and the delay performance, which is the main focus of this paper. In particular, our work is mainly motivated by the seminal paper [15], in which Kelly and Laws give a conjecture regarding the choice of the threshold that is able to guarantee delay optimality in heavy traffic. More precisely, they consider a two-server system with Poisson arrivals and exponential service. The arrival is dispatched randomly, except when one queue is below the threshold r and the other is above, in which case the arrival is dispatched to the shorter one. Note that this dynamic policy can be exactly implemented by a pull-based load balancing scheme with a threshold r. Kelly and Laws conjecture that as long as the threshold r is greater than a specific constant times the logarithm of the mean number of tasks in the system, then the sum queue lengths process under this threshold policy has the same diffusion limit as that under JSQ. Therefore, the logarithmic growth rate result in our sufficient conditions (see Theorem 3.3) not only directly resolves the conjecture in [15], but generalizes it to systems with any fixed finite number of servers as well as general arrival and service distributions. Moreover, the diffusion limit result conjectured in [15] only gives the optimality in a finite time interval while our heavy traffic optimality result obtained by Lyapunov drift-based approach is in steady state.

It is also worth noting that a logarithmic growth in the threshold is not a coincidence, and has been found in a wide range of scenarios. For example, the authors in [23] consider an asymmetric threshold policy for a two-server case. In that setting, only one server has a threshold r (say server 2). The arrivals are always dispatched to server 1 unless the queue length of server 2 is less than the threshold, in which case the arrival is sent to server 2. One of the main contributions in [23] is that a logarithmic growth rate of r is sufficient to guarantee that this threshold policy achieves the same diffusion limit as that under JSQ in heavy traffic. This result can be seen as a first attempt to resolve the conjecture in [15] with a simpler model. In particular, since there is only one threshold in [23], the network can be characterized by a one-dimensional reflected Brownian motion in heavy traffic. In contrast, the limit process in [15] is a two-dimensional Brownian motion, which is harder to rigorously prove optimality. Besides dynamic routing, a logarithmic growth rate of the threshold also critically affects the performance of scheduling policies in [2, 14]. Both authors considered a system of two parallel servers with dedicated arrivals to each of the queues. One server can only process tasks in its own queue, while a 'super-server' can process tasks from both queues. A threshold policy is proposed in which the 'super-server' processes tasks from its own queue when the other server's queue length is below a threshold, and otherwise the 'super-server' processes the tasks from the other queue. This policy can be viewed as the scheduling counterpart of the asymmetric routing policy considered in [23]. In a 'discrete review' setting, the author in [14] proved that a sufficient condition for the asymptotic optimality of this threshold policy is that the threshold must grow as a constant times the average number of tasks in the system. The same result was generalized to a 'continuous review' setting with more general arrival and service distributions in [2]. As in the paper by Kelly and Laws [15], the asymptotic optimality in [2, 14, 23] holds in a finite time interval since the convergence to the stationary distribution is not validated for the

diffusion approximations. Considering the similarity between the scheduling policies in [2, 14] and the routing policy in [23], our approach developed in this paper might be applied to establish heavy-traffic delay optimality in steady state for dynamic scheduling policies as well.

We shall finally point out that the heavy-traffic regime considered in this paper and all the aforementioned papers assumes that the number of servers is a constant, which is different from the Halfin-Whitt heavy-traffic regime (also known as many-server heavy-traffic regime or quality-and-efficiency-driven regime) [12]. In the latter regime, the heavy-traffic parameter ϵ approaches zero and the number of servers N goes to infinity at the same time [1, 6, 10, 20]. For example, it has been shown that, on any finite time interval, the limiting process under the JIQ policy is indistinguishable from that under the JSQ policy in the Halfin-Whitt heavy-traffic regime [20]. In contrast, in the conventional heavy-traffic regime considered in this paper, its delay performance is strictly between that of JSQ and random routing as shown by Theorem 3.2.

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 describes the system model and assumptions considered in this paper. Then, several necessary preliminaries are presented.

2.1 System model

We consider a discrete-time load balancing system consisting of a central dispatcher and N servers. Each server maintains an infinite capacity FIFO queue. At the central dispatcher, there is also a local memory denoted as $m(t)$, through which the dispatcher can have limited information about the system. In each time-slot, the central dispatcher routes the new incoming tasks to one of the servers, immediately upon arrival as in [7, 19, 25, 27, 28, 30]. Once a task joins a queue, it will remain in that queue until its service is completed. Each server is assumed to be work conserving: a server is idle if and only if its corresponding queue is empty.

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. 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. 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 $S_\Sigma(t) \triangleq \sum_{n=1}^{N} S_n(t)$. To illustrate the key ideas behind the results, we first assume that both the arrival and service processes have a finite support, i.e., $A_\Sigma(t) \leq A_{max} < \infty$ and $S_n(t) \leq S_{max} < \infty$ for all t and n. However, the main results still hold when the support is infinite, as discussed in Section 4.

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 are interested in a general pull-based policy formally defined as follows. In words, under this policy, the arrival is randomly dispatched to one of the servers whose queue lengths are below a threshold r, if there are any; Otherwise, it is dispatched to one of N queues randomly.

Definition 2.1. Join-Below-Threshold (JBT) policy is composed of the following components:

(a) Each server is initialized with an empty queue, and a corresponding ID in the local memory of the dispatcher.
(b) Upon new arrivals at the beginning of each time-slot, the dispatcher checks the available IDs in memory. If one or more IDs exist, it removes one uniformly at random, and sends all the new arrivals to the corresponding server. Otherwise, all the new arrivals are dispatched uniformly at random to one of the servers in the system.
(c) Each server reports its ID to the dispatcher at the end of each time-slot if its queue length is below the threshold, and the dispatcher does not contain its ID (see the remark below on this condition).
(d) For the case of heterogeneous servers, in (c) each server also sends its μ_n to the dispatcher and in (b) instead of choosing the ID uniformly at random, the dispatcher selects the ID in proportion to the service rate. Specifically, if the ID of server i is in $m(t)$, the probability for server i to be chosen is $\mu_i / \sum_{j \in m(t)} \mu_j$.

Remark 1. It is easy to see that JIQ is a special case of JBT with $r = 1$. Morevoer, note that in (c) the server can easily know whether or not its own ID exists at the dispatcher. This is because whenever there are new arrivals to a server, the server immediately knows that its own ID at the dispatcher (if exists) has just been removed in order to dispatch the new arrivals. In addition, after each successful report, the server knows that the dispatcher has just added its ID in the memory. Of course, in the analysis of JBT, we can simply assume that the set of servers whose queue lengths are below the threshold are known at the dispatcher without worrying about the implementational details.

The considered load balancing system under JBT can be modeled as a discrete-time Markov chain $\{Z(t) = (\mathbf{Q}(t), m(t)), t \geq 0\}$ with state space $\mathcal{Z}$, using the queue length vector $\mathbf{Q}(t)$ together with the memory state $m(t)$. We consider a set of load balancing systems $\{Z^{(\epsilon)}(t), t \geq 0\}$ 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 parameter ϵ characterizes the distance between the arrival rate and the boundary of the capacity region. We are interested in the throughput performance and more importantly the steady-state delay performance in the heavy-traffic regime under the JBT policy.

Recall that a load balancing system is stable if the Markov chain $\{Z(t), t \geq 0\}$ is positive recurrent, and $\overline{Z} = \{\overline{\mathbf{Q}}, \overline{m}\}$ denotes the random vector whose distribution is the same as the steady-state distribution of $\{Z(t), t \geq 0\}$. We have the following definition.

Definition 2.2 (Throughput Optimality). A load balancing policy is said to be throughput optimal if for any arrival rate within the capacity region, i.e., for any $\epsilon > 0$, the system is positive recurrence 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 [25, 28, 30], because besides the positive recurrence, it also requires all the moments to be finite in steady state for any arrival rate within the capacity region.

To characterize the steady-state average delay performance in the heavy-traffic regime when ϵ approaches zero, by Little's law, it is sufficient to focus on the summation of all the queue lengths. First, recall the following fundamental lower bound on the expected sum queue lengths in a load balancing system under any throughput optimal policy [7].

LEMMA 2.3. *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 [7, 19, 25, 27, 28, 30].

Definition 2.4 (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 2.3.

In the analysis of the delay performance of JBT, the following region $\mathcal{R}^{(r)}$ in $\mathbb{R}^N$ plays an instrumental role by the virtue of the JBT policy.

$$\mathcal{R}^{(r)} = \mathcal{R}_l^{(r)} \cup \mathcal{R}_u^{(r)}, \tag{3}$$

where $r \geq 1$ and

$$\mathcal{R}_l^{(r)} \triangleq \left\{\mathbf{x} \in \mathbb{R}_+^N : x_n \leq r \text{ for all } n \in \mathcal{N}\right\}$$
$$\mathcal{R}_u^{(r)} \triangleq \left\{\mathbf{x} \in \mathbb{R}_+^N : x_n \geq r \text{ for all } n \in \mathcal{N}\right\}.$$

By the definition of the JBT policy, we have that whenever the queue length vector is within the region $\mathcal{R}^{(r)}$, then JBT reduces to (proportionally) random routing. On the other hand, when the queue lengths vector is outside the region $\mathcal{R}^{(r)}$, shorter queues are preferred over longer queues.

3 MAIN RESULTS

In this section, we present both necessary and sufficient conditions on the threshold r for the JBT policy to be heavy-traffic delay optimal in steady-state. We first establish throughput optimality of the JBT policy, which serves as a basis for the analysis of heavy-traffic delay optimality.

3.1 Throughput optimality

We first prove the following result, which establishes that a load balancing system under the JBT policy is stable with bounded moments on the queue lengths for any threshold $r \geq 1$.

Lemma 3.1. *JBT is throughput optimal with the p-th moment of* $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ *being* $O(1/\epsilon^p)$ *for any threshold* $r \geq 1$ *and integer* $p \geq 1$.

Proof. See Appendix A □

Besides throughput optimality, another important aspect of this lemma is that it serves as the basis for the discussions on heavy-traffic delay optimality in the following sections. This is because, firstly, a load balancing policy that cannot stabilize the system is incapable of being heavy-traffic delay optimal at all. Second, the bounded moments result allows us to set the mean drift of Lyapunov functions concerning queue lengths to be zero in steady state, which plays a pivotal part in the framework of Lyapunov drift-based heavy-traffic analysis.

3.2 Necessary condition

In this section, we show that a necessary condition for the JBT policy to achieve heavy-traffic delay optimality is that the threshold r should grow to infinity as the heavy-traffic parameter ϵ approaches zero. However, as we show it cannot grow too fast. Formally, it is presented in the following theorem.

Theorem 3.2. *Consider a load balancing system with homogeneous servers under the JBT policy.*

(1) Suppose the threshold r is any constant in $[1, \infty)$, then we have

$$\liminf_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] > \frac{\zeta}{2} \tag{4}$$

and

$$\limsup_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] < \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_{n,Rand}^{(\epsilon)}\right], \tag{5}$$

where $\overline{\mathbf{Q}}_{Rand}^{(\epsilon)}$ is the steady-state vector under random routing policy.

(2) Suppose the threshold $r^{(\epsilon)} = (1/\epsilon)^{1+\alpha}$ for any constant $\alpha > 0$, then we have

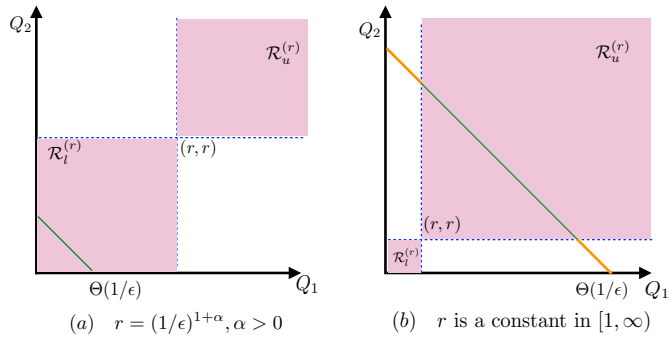
$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_{n,Rand}^{(\epsilon)}\right]. \tag{6}$$

Proof. See Section 5.1 □

Now, we will present the high-level intuitions behind the necessary condition with the illustration in Fig. 1. These intuitions can not only facilitate understanding of the results, but also motivate the sufficient condition in the next section.

To start with, let us consider case (2) when $r^{(\epsilon)} = (1/\epsilon)^{1+\alpha}$ for any $\alpha > 0$. In this case, all the queue lengths are below the threshold r for high loads since the sum queue length in the system is only on the order of $1/\epsilon$. As a result, in case (2), the JBT policy completely degenerates to random routing, which is not heavy-traffic delay optimal [8]. An illustration of case (2) for a two-server system is presented in Fig. 1(a).

Then, we turn to case (1) for which the threshold is a constant. In particular, combing Eqs. (4) and (5) yields that the delay performance of JBT under any constant r in heavy-traffic lies *strictly* between that of a heavy-traffic delay optimal policy (e.g., JSQ) and that of random routing. This



(a) $r = (1/\epsilon)^{1+\alpha}, \alpha > 0$ (b) r is a constant in $[1, \infty)$

Fig. 1. Geometric illustrations of the necessary condition.

reveals an interesting and kind of counter-intuitive insight about the JBT policy under a constant threshold. For example, consider the special case $r = 1$, i.e., the JIQ policy. At first glance, one might expect that the delay performance of JIQ would downgrade to that of random routing in the heavy-traffic limit, since in this case there are hardly any idle servers, and hence the dispatcher under JIQ would just randomly choose one server when allocating arrivals, as in random routing. However, it turns out that this is not true as shown in Eq. (5). That is, the performance of JIQ is still strictly better than that of random routing even in the heavy-traffic limit. This demonstrates that JIQ is able to achieve *partial* resource pooling due to the fact that it adopts queue length information to prefer shorter queues whenever possible. To see this, note that by positive recurrence, there always exists some time when the queue length vector is outside the region $\mathcal{R}^{(r)}$ and hence shorter queues are preferred (i.e., the orange line in Fig. 1(b)), even though it is much less than the time within the region $\mathcal{R}^{(r)}$ (i.e., the green line in Fig. 1(b)). This is totally different from the case in Fig. 1(a) in which the queue-length state always completely remains within the $\mathcal{R}^{(r)}$ for high loads, and hence JBT would downgrade to random routing in the limit.

On the other hand, to explain the liminf result in Eq. (4), we will utilize the following result. That is, the necessary (and sufficient) condition for the JBT policy to be heavy-traffic delay optimal is given by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0. \tag{7}$$

This is a direct application of the results in [29]. Note that since $Q_n(t+1)U_n(t) = 0$, the above condition basically means that the key for JBT to be heavy-traffic delay optimal is that it should guarantee that no server is idling while other servers are busy with high loads. In the case when r is a constant, the event that one queue is zero while others with high loads (denoted by E_{bad}) happens with a non-negligible probability since the axes are close to the region $\mathcal{R}_u^{(r)}$. As a result, the left-hand side of Eq. (7) is strictly positive, and hence JBT is not heavy-traffic delay optimal for a constant r. The intuition that we should guarantee that the event E_{bad} occurs very rarely in heavy-traffic also motivates our sufficient condition in the next section where we let the threshold r grow in a certain rate to guarantee that the axes are far away from the region $\mathcal{R}_u^{(r)}$.

Remark 2. It is worth noting that in [30], a similar result as Eq. (4) has been established for the JIQ policy (i.e., the special case $r = 1$ of JBT) in a two-server system under the constraints that the service processes are constant and the variance of arrival process should be larger than a particular value. Thus, our contribution is to generalize the result in [30] to any constant $r \geq 1$

and any finite number of servers without the constraints on service and arrival process as required in [30]. More importantly, we provide new results given by Eqs. (5) and (6), which give us a sharper understanding of general pull-based policies.

3.3 Sufficient condition

In this section, we now investigate the sufficient condition. In particular, we show that if the threshold in JBT grows at a logarithmic rate with respect to the average sum queue lengths, i.e., $r^{(\epsilon)} \geq K \log(1/\epsilon)$ for some specified constant K, then the JBT policy is heavy-traffic delay optimal in steady state, which is formally presented in the following theorem.

THEOREM 3.3. *Consider a load balancing system under the JBT policy. Suppose that the threshold r satisfies* $r^{(\epsilon)} \geq K \log(1/\epsilon)$ *and* $r^{(\epsilon)} = o(1/\epsilon)$*, where the constant* $K = 2(1+\alpha)/\theta^*$ *for any* $\alpha > 0$ *and* θ^* *is the constant in Eq.* (9)*, then JBT is heavy-traffic delay optimal in steady state.*

PROOF. See Section 5.2 □

The main contributions of this result can be summarized as follows. First, it directly resolves and generalizes a conjecture in [15]. More precisely, the authors in [15] consider a two-server system with Poisson arrivals and exponential service under a threshold policy that has the same implementation as JBT, and conjecture that as long as the threshold is greater than a specified constant times $\log(1/\epsilon)$, the heavy-traffic asymptotic optimality of the threshold routing strategy holds. Thus, our result resolves this conjecture and also generalizes it to any finite number of servers case with general arrival and service distributions. More importantly, the asymptotic optimality defined in [15] holds only for a finite time interval since the convergence to steady-state distribution is not touched. In contrast, our result directly gives the steady-state characterization of the delay optimality in heavy-traffic of the JBT policy.

The key step in establishing the sufficient condition in Theorem 3.3 is the notion of state-space collapse. In words, it says that in heavy traffic the system state under the JBT policy would concentrate around the region $\mathcal{R}^{(r)}$ as defined Eq. (3). To that end, we need the following property of the distance to the region $\mathcal{R}^{(r)}$. The distance of a point $\mathbf{x}$ to the region $\mathcal{R}^{(r)}$ is related to the distances to the regions $\mathcal{R}_l^{(r)}$ and $\mathcal{R}_u^{(r)}$ as follows.

$$d_{\mathcal{R}^{(r)}}(\mathbf{x}) = \min\left(d_{\mathcal{R}_l^{(r)}}(\mathbf{x}), d_{\mathcal{R}_u^{(r)}}(\mathbf{x})\right), \tag{8}$$

where the distance of a point $\mathbf{x}$ to a set $\mathcal{A}$ in $\mathbb{R}^N$ is defined as

$$d_{\mathcal{A}}(\mathbf{x}) \triangleq \inf_{\mathbf{y}\in\mathcal{A}} \{\|\mathbf{x}-\mathbf{y}\|\}.$$

This equality (8) can be established by contradiction. Suppose that

$$\min\left(d_{\mathcal{R}_l^{(r)}}(\mathbf{x}), d_{\mathcal{R}_u^{(r)}}(\mathbf{x})\right) = d_{\mathcal{R}^{(r)}}(\mathbf{x}) + \alpha$$

for some $\alpha > 0$, then there exists a $\mathbf{y}^* \in \mathcal{R}^{(r)}$ such that

$$d_{\mathcal{R}^{(r)}}(\mathbf{x}) \leq \|\mathbf{x}-\mathbf{y}^*\| < \min\left(d_{\mathcal{R}_l^{(r)}}(\mathbf{x}), d_{\mathcal{R}_u^{(r)}}(\mathbf{x})\right).$$

However, since $\mathbf{y}^* \in \mathcal{R}^{(r)} = \mathcal{R}_l^{(r)} \cup \mathcal{R}_u^{(r)}$, this leads to a contradiction to the right-hand side of the inequality above.

We say that the system state concentrates around the region $\mathcal{R}^{(r)}$ if all the moments of the distance $d_{\mathcal{R}^{(r)}}(\overline{\mathbf{Q}})$ are upper bounded by constants. Formally, we have the following definition.

Definition 3.4 (State-space collapse to $\mathcal{R}^{(r)}$). Suppose that the system process converges in distribution to a steady-state random vector $\overline{\mathbf{Q}}^{(\epsilon)}$. Then, we say that the state-space of a load balancing system collapses to the region $\mathcal{R}^{(r)}$ if there exist some positive constants ϵ_0, θ^* and C^* such that for all $\epsilon \in (0, \epsilon_0)$

$$\mathbb{E}\left[e^{\theta^* d_{\mathcal{R}^{(r)}}\left(\overline{\mathbf{Q}}^{(\epsilon)}\right)}\right] \leq C^*, \tag{9}$$

where both θ^* and C^* are independent of ϵ.

Note that this notion of state-space collapse is different from previous works, as will be explained later. For any constant threshold r, Eq. (9) trivially holds since the distance to the region $\mathcal{R}^{(r)}$ is always bounded by a constant. Thus, in the following we only consider the interesting case when r grows to infinity, which is also required by the necessary condition in Theorem 3.2. In this case, we have the following result regarding state space collapse of the JBT policy, which plays a key role in the proof of Theorem 3.3.

Proposition 3.5. *Consider a load balancing system under the JBT policy. Suppose that the threshold satisfies $\lim_{\epsilon \downarrow 0} r^{(\epsilon)} = \infty$, then the system state-space collapses to the region $\mathcal{R}^{(r)}$.*

Proof. See Section 5.3 □

Remark 3. It should be noted that besides being a key step in proving the sufficient conditions in Theorem 3.3, Proposition 3.5 has its own contributions. (i) First, the region of state-space collapse in this paper, i.e., $\mathcal{R}^{(r)}$ is not a single dimensional line as in [7, 19, 25, 27, 28, 30], nor a multi-dimensional convex cone as in [17, 18, 24, 29]. This not only brings new challenges in proving state-space collapse itself, but also requires new methods to relate the collapse result to heavy-traffic delay optimality. More specifically, on the one hand, in order to prove state-space collapse result, we need to handle the non-convexity of $\mathcal{R}^{(r)}$ by choosing the minimum of two distances as the Lyapunov function. The techniques suggested in [29] to handle the non-convex region cannot apply here since the region $\mathcal{R}^{(r)}$ cannot be covered by the cone defined in [29]. On the other hand, in order to utilize the state-space collapse result to conclude heavy-traffic delay optimality, the conventional decompositions of parallel and perpendicular components of the queue length vector $\mathbf{Q}$ would not work. Instead, we need to carefully divide the system state and then apply Chernoff bound on the random variable $d_{\mathcal{R}^{(r)}}\left(\overline{\mathbf{Q}}^{(\epsilon)}\right)$, which is possible by the state-space collapse result in Eq. (9). (ii) Second, the upper bound result in Eq. (9) holds even when the system is not at the heavy-traffic limit, and hence it is of independent interest for analyzing the system performance in the pre-limit regime, especially when combined with optimization techniques.

Now, we turn to provide the high-level intuitions on Proposition 3.5 and Theorem 3.3 with the help of Fig. 2. This will facilitate the understanding of the results as well as their proofs.

To start with, note that by virtue of the JBT policy, when the queue-length state $\mathbf{Q}$ is outside the region $\mathcal{R}^{(r)}$, there always exists a positive drift towards the region $\mathcal{R}^{(r)}$. This is because in this case there exists a positive drift towards the lower region $\mathcal{R}_l^{(r)}$ and a positive drift towards the upper region $\mathcal{R}_u^{(r)}$, respectively (see Fig. 2(a) for an illustration). This provides the key intuition as to why the system state would concentrate around the region $\mathcal{R}^{(r)}$ since suppose there is no drift (e.g., under random routing) the expected distance to the region $\mathcal{R}^{(r)}$ would go to infinity as $r^{(\epsilon)}$ goes to infinity (assuming that the growth rate of $r^{(\epsilon)}$ is not too fast). In contrast, under the JBT policy, the distance remains constant (as shown by the gray color in Fig. 2(b)). This is the reason why we call it a state-space collapse result, which is different from much of previous works where

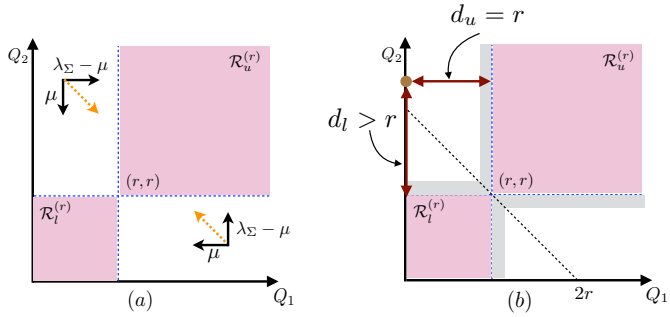


Fig. 2. Geometric illustrations of the sufficient condition.

the system state collapses to a lower dimensional space (e.g., a line or a convex cone) while our state-space collapse region $\mathcal{R}^{(r)}$ is of the same dimension as the original queue-length state vector. Hence, we need to develop new methods to apply this new type of state-space collapse result to achieve heavy-traffic delay optimality of the JBT policy, as in Theorem 3.3.

To this end, we will utilize the sufficient (and necessary) condition in Eq. (7) again. As discussed before, it basically requires us to guarantee that no server is idling while other servers are busy under high loads. To achieve this, a logarithmic growth rate as in Theorem 3.3 is sufficient. For an illustration of the main ideas behind the proof, let us consider a simple two-server case. In this case, Eq. (7) reduces to

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{Q}_1^{(\epsilon)}(t+1)\overline{U}_2^{(\epsilon)} + \overline{Q}_2^{(\epsilon)}(t+1)\overline{U}_1^{(\epsilon)}\right] = 0. \tag{10}$$

Take the second term above for example, it can be rewritten as the summation of the following terms (for simplicity we omit the superscript $^{(\epsilon)}$)

$$\overline{Q}_2(t+1)\overline{U}_1 \mathcal{I}\left(\overline{Q}_2(t+1) \le 2r, \overline{Q}_1(t+1) = 0\right) \tag{11}$$

$$\overline{Q}_2(t+1)\overline{U}_1 \mathcal{I}\left(\overline{Q}_2(t+1) > 2r, \overline{Q}_1(t+1) = 0\right), \tag{12}$$

where we use the fact that $Q_n(t+1)U_n(t) = 0$ again. The expectation of Eq. (11) can be upper bounded by $2r^{(\epsilon)}\epsilon$ since $\mathbb{E}\left[\overline{U}_1\right] \le \epsilon$. For the expectation of Eq. (12), we first apply Cauchy-Schwartz inequality and hence obtain its upper bound as

$$C\frac{1}{\epsilon^2}\mathbb{P}\left(\overline{Q}_2(t+1) > 2r, \overline{Q}_1(t+1) = 0\right),$$

where C is a constant independent of ϵ. Now, we can apply the state-space collapse result (i.e., Eq. (9)) combined with Chernoff bound to show that the probability that one queue is empty and another queue length is larger than $2r$ has an exponential decay rate. In particular, we have

$$\mathbb{P}\left(\overline{Q}_2(t+1) > 2r, \overline{Q}_1(t+1) = 0\right) \overset{(a)}{\le} \mathbb{P}\left(d_{\mathcal{R}^{(r)}}\left(\overline{\mathbf{Q}}^{(\epsilon)}\right) \ge r\right) \overset{(b)}{\le} \frac{C^*}{e^{\theta^* r}},$$

where (a) holds since in this case the distance to the region $\mathcal{R}^{(r)}$ is r (see Fig. 2(b) for an illustration); (b) follows directly from state-space collapse result and Chernoff bound. Therefore, combining the expectations of Eqs. (11) and (12), yields

$$\mathbb{E}\left[\overline{Q}_2^{(\epsilon)}(t+1)\overline{U}_1^{(\epsilon)}\right] \le 2r^{(\epsilon)}\epsilon + C_1\frac{1}{\epsilon^2}\frac{1}{e^{\theta^* r^{(\epsilon)}}},$$

which approaches zero whenever $r^{(\epsilon)} = o(\frac{1}{\epsilon})$ and $r^{(\epsilon)} \ge K\log(1/\epsilon)$ where $K = 2(1+\alpha)/\theta^*$ for any $\alpha > 0$. By the same arguments, we can establish the same result for the expectation of the first term

in Eq. (10). Therefore, we have reached the sufficient condition for heavy-traffic delay optimality in Theorem 3.3.

4 GENERALIZATIONS

For the illustration of the key ideas, the main results in the last section are obtained under the assumptions that both arrival and service processes have finite support. However, it is worth pointing out that the same results still hold (with only a change in constants) when the support is infinite. More specifically, we need the following weak condition on arrival and service processes, which requires that the tails of both arrival and service processes have an exponential decay.

Condition A (Weaker condition on arrival and service). The i.i.d arrival process $A_\Sigma(t)$ and service process $S_n(t)$ satisfy

$$\mathbb{E}\left[e^{\theta_1 A_\Sigma(t)}\right] \leq D_1 \text{ and } \mathbb{E}\left[e^{\theta_2 S_n(t)}\right] \leq D_2,$$

for each n where the constants $\theta_1 > 0$, $\theta_2 > 0$, $D_1 < \infty$ and $D_2 < \infty$ are all independent of ϵ.

In order to obtain the same main results under the weaker condition above, we should make some mild changes in our proofs. In the following, we will highlight the key steps involved in this process.

(i) First, note that in order to establish condition (C1) in Lemma 5.1, we would use the following upper bound in our proofs based on the finite support assumptions.

$$\mathbb{E}\left[\|\mathbf{A}(t_0) - \mathbf{S}(t_0)\|^2 \mid Z(t_0)\right] \leq L \triangleq N \max(A_{max}, S_{max})^2.$$

However, under the weaker Condition A, we can still bound the left-hand side by a constant independent of ϵ. This directly follows from the fact that all the moments of a random variable are finite if its moment generating function is finite in an open interval containing zero.

(ii) Second, we should now replace condition (C2) in Lemma 5.1 with the following weak stochastic domination condition (C2′),

- (C2′) $[\Delta V(X) \mid X(t_0) = X] \prec W$ for all t_0 and $\mathbb{E}\left[e^{\theta W}\right] = D$ is finite for some $\theta > 0$.

This condition holds under the weaker Condition A since the arrival and service processes both have an exponentially bounded tail by the finiteness of their moment generating functions. As shown by Theorem 2.3 in [11], the combination of (C1) and (C2′) is sufficient to guarantee bounded moments as required in the proof of our main results.

(iii) Third, we now should take a careful treatment of the unused service. For example, the following result plays a key role in establishing the necessary and sufficient condition in Eq. (7)

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2\right] = 0.$$

Under the assumption of finite support for the service process, the left-hand side can be easily bounded above by $NS_{max}\epsilon$, which approaches zero as $\epsilon \to 0$. Now, under the weak condition, we need to adopt the truncation trick to handle the unbounded service. More specifically, let us consider any $n \in \mathcal{N}$, we have for any $t \geq 0$ and constant S'

$$\begin{aligned} U_n^2(t) &\leq U_n(t) S_n(t) \\ &= U_n(t) S_n(t) \mathcal{I}\left(S_n(t) \leq S'\right) + U_n(t) S_n(t) \mathcal{I}\left(S_n(t) > S'\right) \\ &\leq U_n(t) S' + S_n^2(t) \mathcal{I}\left(S_n(t) > S'\right). \end{aligned}$$

In steady state, we have

$$\begin{aligned}\mathbb{E}\left[\overline{U}_n^2\right] &\leq \mathbb{E}\left[\overline{U}_n\right] S' + \mathbb{E}\left[S_n^2(\infty)\mathcal{I}\left(S_n(\infty) > S'\right)\right] \\ &\overset{(a)}{\leq} \epsilon S' + \mathbb{E}\left[S_n^2(0)\mathcal{I}\left(S_n(0) > S'\right)\right] \\ &\overset{(b)}{\leq} \epsilon S' + \beta,\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right] = \epsilon$ and service process is *i.i.d.*; in (b), we choose S' such that $\mathbb{E}\left[S_n^2(0)\mathcal{I}\left(S_n(0) > S'\right)\right] \leq \beta$, which is possible by the exponential decay rate of $S_n(0)$ under the weak condition. Thus, we have

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{U}_n^2\right] \leq \beta,$$

for any $\beta > 0$. Hence, we have $\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\overline{U}_n^2\right] = 0$ for each n.

Remark 4. The three highlighted key steps could also demonstrate their generalization power in previous works where the Lyapunov drift-based framework is adopted under the assumption of finite supports for the arrival and service processes.

5 PROOFS

In this paper, we will adopt the Lyapunov drift-based approach developed in [7] to derive bounded moments in steady state. In particular, the following lemma, which follows directly from Lemmas 2 and 3 in [18], will be the main tool in our proofs.

Lemma 5.1. *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 drift of V at X as*

$$\Delta V(X) \triangleq [V(X(t_0 + 1)) - V(X(t_0))]\mathcal{I}(X(t_0) = X),$$

where $\mathcal{I}(.)$ is the indicator function. Suppose the 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 $V(X) \geq \kappa$,*

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

- *(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 $\overline{V}$ for which there exists a $\theta^* > 0$ and a $C^* < \infty$ such that

$$\mathbb{E}\left[e^{\theta^* \overline{V}}\right] \leq C^*,$$

which directly implies that all the moments of $\overline{V}$ exist and are finite. More specifically, we have for any $p = 1, 2, \ldots$

$$\mathbb{E}\left[V(\overline{X})^p\right] \leq (2\kappa)^p + (4D)^p \left(\frac{D + \eta}{\eta}\right)^p p!. \tag{13}$$

We would also utilize the following useful result in our proofs.

Lemma 5.2. *For the JBT policy with threshold $r \geq 1$, it is heavy-traffic delay optimal if and only if*

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0. \tag{14}$$

This lemma is a direct application of the results in [29], which establishes that Eq. (14) is the sufficient and necessary condition for any load balancing policy to be heavy-traffic delay optimal if the system is stable with bounded moments. By Lemma 3.1, we have that the JBT policy is throughput optimal with all the moments being bounded for any $r \geq 1$, and hence the above lemma holds.

5.1 Proof of Theorem 3.2

Before we present our proof, we first give the following useful result, which can be established by setting the mean drift a chosen Lyapunov function to zero in steady state. For completeness, the proof is given at Appendix B.

Lemma 5.3. *Consider a load balancing system with homogeneous servers under the JBT policy. For any threshold $r \geq 1$, we have*

$$2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left((\overline{Q}_i^+)^{(\epsilon)}\overline{U}_j^{(\epsilon)} + (\overline{Q}_j^+)^{(\epsilon)}U_i^{(\epsilon)}\right)\right] = \mathcal{T}_1^{(\epsilon)} + \mathcal{T}_2^{(\epsilon)} - \mathcal{T}_3^{(\epsilon)},$$

where

$$\begin{aligned}
\mathcal{T}_1^{(\epsilon)} &\triangleq 2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i^{(\epsilon)} - \overline{Q}_j^{(\epsilon)}\right)\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)}\right)\right]\\
\mathcal{T}_2^{(\epsilon)} &\triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right]\\
\mathcal{T}_3^{(\epsilon)} &\triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right]\\
\overline{\mathbf{Q}}^+ &\triangleq \overline{\mathbf{Q}}(t+1)
\end{aligned}$$

and $\overline{A}_i^{(\epsilon)}$ and $\overline{U}_i^{(\epsilon)}$ are dependent of $\overline{\mathbf{Q}}$ for each i and $\epsilon > 0$.

Now, we are ready to present the proof of Theorem 3.2.

Proof of Theorem 3.2. To start with, we first note that the sufficient and necessary condition in Lemma 5.2 can be rewritten as follows under the JBT policy.

$$\begin{aligned}
&2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right]\\
\overset{(a)}{=}&2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left((\overline{Q}_i^+)^{(\epsilon)}\overline{U}_j^{(\epsilon)} + (\overline{Q}_j^+)^{(\epsilon)}U_i^{(\epsilon)}\right)\right]\\
\overset{(b)}{=}&4\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left((\overline{Q}_i^+)^{(\epsilon)}\overline{U}_j^{(\epsilon)}\right)\right]\\
\overset{(c)}{=}&4\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j^{(\epsilon)}\mathbb{P}\left((\overline{Q}_i^+)^{(\epsilon)} = k, (\overline{Q}_j^+)^{(\epsilon)} = 0, \overline{U}_j^{(\epsilon)} \geq 1\right)\right),
\end{aligned} \tag{15}$$

in which (a) and (c) follow from the fact $Q_i(t+1)U_i(t) = 0$ for each i and $t \geq 0$; (b) holds by the symmetry property of JBT policy for homogeneous servers.

Thus, by Lemma 5.2, Lemma 5.3 and the above equation, in order to analyze heavy-traffic delay optimality of JBT under any constant threshold, all we need to do is to focus on terms $\mathcal{T}_1^{(\epsilon)}$, $\mathcal{T}_2^{(\epsilon)}$ and $\mathcal{T}_3^{(\epsilon)}$, respectively.

Now, let us first focus on case (1) in Theorem 3.2.

For $\mathcal{T}_1^{(\epsilon)}$, we have

$$\begin{aligned}
\mathcal{T}_1^{(\epsilon)} &\triangleq 2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i^{(\epsilon)} - \overline{Q}_j^{(\epsilon)}\right)\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)}\right)\right] \\
&= 2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i - \overline{Q}_j\right)\left(\overline{A}_i - \overline{A}_j\right)\mathcal{I}\left(\overline{Q}_i \geq r, \overline{Q}_j \geq r\right)\right] \\
&\quad + 2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i - \overline{Q}_j\right)\left(\overline{A}_i - \overline{A}_j\right)\mathcal{I}\left(\overline{Q}_i < r, \overline{Q}_j < r\right)\right] \\
&\quad + 4\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i - \overline{Q}_j\right)\left(\overline{A}_i - \overline{A}_j\right)\mathcal{I}\left(\overline{Q}_i \geq r, \overline{Q}_j < r\right)\right] \\
&\overset{(a)}{=} 4\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i - \overline{Q}_j\right)\left(\overline{A}_i - \overline{A}_j\right)\mathcal{I}\left(\overline{Q}_i \geq r, \overline{Q}_j < r\right)\right] \\
&\overset{(b)}{\geq} -4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\sum_{k=r}^{\infty}(k-m)\mathbb{P}\left(\overline{Q}_i = k, \overline{Q}_j = m\right) \\
&\overset{(c)}{=} -4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\sum_{k=r}^{\infty}(k-m)\mathbb{P}\left(\overline{Q}_i^+ = k, \overline{Q}_j^+ = m\right),
\end{aligned} \tag{16}$$

where (a) follows from the definition of the JBT policy, i.e., when both queues are in memory or both queues are not in memory, they have the same probability to be selected in the homogeneous case; (b) is true since when the ID of server j is in $m(t)$ while the ID of server i is not, we have $A_i(t) = 0$ and $A_j(t) \leq A_\Sigma(t)$ by the definition of the JBT policy; (c) holds since $\overline{\mathbf{Q}}(t+1)$ has the same distribution as $\overline{\mathbf{Q}}(t)$ in steady state.

In order to further simplify the term $\mathcal{T}_1^{(\epsilon)}$, we need to define the following events in which $k \geq r$ and $1 \leq m \leq r-1$.

$$\begin{aligned}
E_{(k,m)} &\triangleq \left\{\overline{Q}_i^+ = k, \overline{Q}_j^+ = m\right\} \\
E_{(k,m)}^+ &\triangleq \left\{\overline{Q}_i(t+2) = k, \overline{Q}_j(t+2) = m\right\} \\
E_{(k,0,0)} &\triangleq \left\{\overline{Q}_i^+ = k, \overline{Q}_j^+ = 0, \overline{U}_j = 0\right\} \\
E_{(k,0,0)}^+ &\triangleq \left\{\overline{Q}_i(t+2) = k, \overline{Q}_j(t+2) = 0, \overline{U}_j^+ = 0\right\} \\
E_{(k,0,\geq 1)} &\triangleq \left\{\overline{Q}_i^+ = k, \overline{Q}_j^+ = 0, \overline{U}_j \geq 1\right\} \\
E_{(k,0,\geq 1)}^+ &\triangleq \left\{\overline{Q}_i(t+2) = k, \overline{Q}_j(t+2) = 0, \overline{U}_j^+ \geq 1\right\}.
\end{aligned}$$

Note that by the assumptions of arrival and service processes, there exists a positive probability $\hat{p}$ (independent of ϵ) such that there is no arrival during one time-slot and meanwhile the potential

service of all the servers are d for some d between 1 and S_{max}. For ease of exposition, we take $d = 1$ in the following proof, and the same techniques apply for the case where $d \neq 1$. Now, for each occurrence of event $E_{(k,m)}$, there exists a positive probability $\hat{p}$ such that $E^+_{(k-1,m-1)}$ will happen. Therefore, we have

$$\mathbb{P}\left(E_{(k-1,m-1)}\right) \overset{(a)}{=} \mathbb{P}\left(E^+_{(k-1,m-1)}\right) \geq \hat{p}\mathbb{P}\left(E_{(k,m)}\right), \tag{17}$$

where (a) holds due to the fact that both events are defined in steady state. Similarly, we have

$$\mathbb{P}\left(E_{(k-1,0,0)}\right) = \mathbb{P}\left(E^+_{(k-1,0,0)}\right) \geq \hat{p}\mathbb{P}\left(E_{(k,1)}\right) \tag{18}$$

$$\mathbb{P}\left(E_{(k-1,0,\geq 1)}\right) = \mathbb{P}\left(E^+_{(k-1,0,\geq 1)}\right) \geq \hat{p}\mathbb{P}\left(E_{(k,0,0)}\right). \tag{19}$$

Now, we can further simplify $\mathcal{T}_1^{(\epsilon)}$ as follows

$$\begin{aligned}
\mathcal{T}_1^{(\epsilon)} &\overset{(a)}{\geq} -4\lambda_\Sigma \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=r}^{\infty} k\mathbb{P}\left(E_{(k,0,\geq 1)}\right) + \frac{1}{\hat{p}}\sum_{k=r}^{\infty} k\mathbb{P}\left(E_{(k-1,0,\geq 1)}\right)\right) \\
&\quad - 4\lambda_\Sigma \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{m=1}^{r-1}\sum_{k=r}^{\infty}\frac{1}{\hat{p}^{m+1}}(k-m)\mathbb{P}\left(E_{(k-m-1,0,\geq 1)}\right)\right) \\
&= -4\lambda_\Sigma \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=0}^{r}\sum_{h=r-l}^{\infty}\frac{1}{\hat{p}^l}h\mathbb{P}\left(E_{(h,0,\geq 1)}\right)\right) \\
&\quad - 4\lambda_\Sigma \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=1}^{r}\sum_{h=r-l}^{\infty}\frac{1}{\hat{p}^l}\mathbb{P}\left(E_{(h,0,\geq 1)}\right)\right) \\
&\overset{(b)}{\geq} -4\lambda_\Sigma \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=0}^{r}\frac{1}{\hat{p}^l}\sum_{h=0}^{\infty}h\overline{U}_j\mathbb{P}\left(E_{(h,0,\geq 1)}\right)\right) \\
&\quad - 4\lambda_\Sigma \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=1}^{r}\frac{1}{\hat{p}^l}\epsilon\right),
\end{aligned} \tag{20}$$

where (a) follows from eqs. (17) to (19); (b) holds since $U_j(t) \geq 1$ and $\mathbb{E}\left[\overline{U}_j\right] \leq \mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right] = \epsilon$. The latter fact can be easily obtained by setting mean drift of $\hat{V}(Z(t)) \triangleq \|\mathbf{Q}(t)\|_1$ to be zero in steady state, which is true since all the moments of $\left\|\overline{\mathbf{Q}}\right\|$ are bounded.

For $\mathcal{T}_2^{(\epsilon)}$, we can simplify it as follows.

$$\begin{aligned}
\mathcal{T}_2^{(\epsilon)} &\triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \\
&\overset{(a)}{=} \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)}\right)^2 + \left(\overline{S}_i^{(\epsilon)} - \overline{S}_j^{(\epsilon)}\right)^2\right] \\
&\overset{(b)}{=} (N-1)\left(\left(\sigma_\Sigma^{(\epsilon)}\right)^2 + \left(\lambda_\Sigma^{(\epsilon)}\right)^2 + \nu_\Sigma^2\right),
\end{aligned} \tag{21}$$

where (a) holds since the arrival and service are independent and the servers are homogeneous; (b) is true because $A_i(t)A_j(t) = 0$ for all $i \neq j$ and $t \geq 0$, and the service is independent and homogeneous.

For $\mathcal{T}_3^{(\epsilon)}$, we can simplify it as follows.

$$\begin{aligned}\mathcal{T}_3^{(\epsilon)} &\triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)}-\overline{U}_j^{(\epsilon)}\right)^2\right]\\ &\overset{(a)}{\leq}(N-1)\,\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2\right]\\ &\overset{(b)}{\leq}\epsilon\,(N-1)\,S_{max},\end{aligned}\tag{22}$$

where (a) follows from the fact that $U_n(t)\geq 0$ for any $n\in\mathcal{N}$; (b) holds because of $U_n(t)\leq S_{max}$ for any $n\in\mathcal{N}$ and the fact $\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right]=\epsilon$.

Now, substituting Eqs. (15), (20), (21) and (22) into the equation in Lemma 5.3, yields

$$\begin{aligned}&4\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j^{(\epsilon)}\mathbb{P}\left((\overline{Q}_i^{+})^{(\epsilon)}=k,(\overline{Q}_i^{+})^{(\epsilon)}=0,\overline{U}_j^{(\epsilon)}\geq 1\right)\right)\\ =&4\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j\mathbb{P}\left(E_{(k,0,\geq 1)}\right)\right)\\ \geq&-4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=0}^{r}\frac{1}{\hat{p}^l}\sum_{h=0}^{\infty}h\overline{U}_j\mathbb{P}\left(E_{(h,0,\geq 1)}\right)\right)-4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=1}^{r}\frac{1}{\hat{p}^l}\epsilon\right)\\ &+(N-1)\left(\left(\sigma_\Sigma^{(\epsilon)}\right)^2+\left(\lambda_\Sigma^{(\epsilon)}\right)^2+\nu_\Sigma^2\right)-S_{max}\,(N-1)\,\epsilon,\end{aligned}$$

which can be simplified as

$$\begin{aligned}&\left(4+4\lambda_\Sigma\sum_{l=0}^{r}\frac{1}{\hat{p}^l}\right)\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j\mathbb{P}\left(E_{(k,0,\geq 1)}\right)\right)\geq-S_{max}\,(N-1)\,\epsilon\\ &\quad-4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{l=1}^{r}\frac{1}{\hat{p}^l}\epsilon\right)+(N-1)\left(\left(\sigma_\Sigma^{(\epsilon)}\right)^2+\left(\lambda_\Sigma^{(\epsilon)}\right)^2+\nu_\Sigma^2\right).\end{aligned}$$

Then taking liminf on both sides gives

$$\liminf_{\epsilon\downarrow 0}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j\mathbb{P}\left(E_{(k,0,\geq 1)}\right)\right)\geq\frac{(N-1)\left(\sigma_\Sigma^2+\mu_\Sigma^2+\nu_\Sigma^2\right)}{4+4\mu_\Sigma\sum_{l=0}^{r}\frac{1}{\hat{p}^l}}>0\tag{23}$$

which holds since threshold r is a constant and $\hat{p}$ would not vanish as $\epsilon\to 0$. Therefore, by Lemma 5.2 and Eq. (15), we have

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

where ζ is the constant defined as in Lemma 2.3.

To establish the inequality (5) in Theorem 3.2, note that the term $\mathcal{T}_1^{(\epsilon)}$ is equal to 0 for any $\epsilon>0$ under random routing, and $\mathcal{T}_2^{(\epsilon)}$ and $\mathcal{T}_3^{(\epsilon)}$ converge to the same constant for both random routing and JBT. Thus, based on Lemma 5.2 and Lemma 5.3, all we need to show is that under the JBT policy $\limsup_{\epsilon\downarrow 0}\mathcal{T}_1^{(\epsilon)}<0$. To this end, we can upper bound it as follows by reusing the equation

(a) in Eq. (16).

$$\begin{aligned}\mathcal{T}_1^{(\epsilon)} &= 4\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i-\overline{Q}_j\right)\left(\overline{A}_i-\overline{A}_j\right)\mathcal{I}\left(\overline{Q}_i\geq r,\overline{Q}_j<r\right)\right]\\ &\overset{(a)}{\leq} -\frac{4\lambda_\Sigma}{N-1}\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\sum_{k=r}^{\infty}(k-m)\mathbb{P}\left(\overline{Q}_i=k,\overline{Q}_j=m\right)\\ &\leq -\frac{4\lambda_\Sigma}{S_{max}(N-1)}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j\mathbb{P}\left(E_{(k,0,\geq 1)}\right)\right),\end{aligned}$$

where (a) holds since when $Q_i(t)\geq r$ and $Q_j(t)<r$, the lower bound on the probability of server j being chosen under JBT is $1/(N-1)$. Now, taking limsup on both sides, yields

$$\begin{aligned}\limsup_{\epsilon\downarrow 0}\mathcal{T}_1^{(\epsilon)} &\leq -\frac{4\lambda_\Sigma}{S_{max}(N-1)}\liminf_{\epsilon\downarrow 0}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\sum_{k=1}^{\infty}k\overline{U}_j\mathbb{P}\left(E_{(k,0,\geq 1)}\right)\right)\\ &< 0,\end{aligned}$$

where the last inequality follows directly from Eq. (23). Hence, we have completed the proof of the first case in Theorem 3.2.

Now, let us turn to case (2) in Theorem 3.2. Based on the discussions above, in order to show that the JBT policy with $r^{(\epsilon)}=(1/\epsilon)^{1+\alpha}$ and $\alpha>0$ achieves the same limit as random routing, all we need to show is that $\lim_{\epsilon\downarrow 0}\mathcal{T}_1^{(\epsilon)}=0$. Again, using the equation (a) in Eq. (16), we obtain

$$\begin{aligned}\mathcal{T}_1^{(\epsilon)} &= 4\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{Q}_i-\overline{Q}_j\right)\left(\overline{A}_i-\overline{A}_j\right)\mathcal{I}\left(\overline{Q}_i\geq r,\overline{Q}_j<r\right)\right]\\ &\geq -4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\mathbb{E}\left[(\overline{Q}_i-m)\mathcal{I}\left(\overline{Q}_i\geq r,\overline{Q}_j=m\right)\right]\\ &\geq -4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\mathbb{E}\left[\overline{Q}_i\mathcal{I}\left(\overline{Q}_i\geq r,\overline{Q}_j=m\right)\right]\\ &\geq -4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\sqrt{\mathbb{E}\left[\overline{Q}_i^2\right]\mathbb{P}\left(\overline{Q}_i\geq r,\overline{Q}_j=m\right)}\\ &\overset{(a)}{\geq} -4\lambda_\Sigma\sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{m=0}^{r-1}\sqrt{\frac{M'}{\epsilon^2}\frac{e^{\theta^*(1/\epsilon)}}{e^{\theta^* r}}},\end{aligned}$$

where (a) follows from the bounded moments in Lemma 3.1 and Chernoff bound based on Eq. (32) in the proof of Lemma 3.1. Thus, if $r^{(\epsilon)}=(1/\epsilon)^{1+\alpha}$ for any constant $\alpha>0$, we have $\lim_{\epsilon\downarrow 0}\mathcal{T}_1^{(\epsilon)}=0$. Hence, we have established the second case in Theorem 3.2. □

5.2 Proof of Theorem 3.3

Proof of Theorem 3.3. Based on the result in Lemma 5.2, in order to prove Theorem 3.3, we need just focus on the left-hand side of Eq. (14). Let us first define

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

in which for brevity we omit the references t and ϵ, and use $\overline{\mathbf{Q}}^+$ to denote $\overline{\mathbf{Q}}(t+1)$. Thus, all we need to show is that $\lim_{\epsilon \downarrow 0} \mathcal{T}^{(\epsilon)} = 0$ under the assumptions of Theorem 3.3. Since $\overline{U}_i \overline{Q}_i^+ = 0$ by the queue-length dynamic in Eq. (1), we have for each $i \in \mathcal{N}$,

$$\begin{aligned}&\mathbb{E}\left[\overline{U}_i \left(\sum_{j=1}^{N} \overline{Q}_j^+\right)\right] \\ =&\mathbb{E}\left[\overline{U}_i \left(\sum_{j=1}^{N} \overline{Q}_j^+\right) \mathcal{I}\left(\overline{Q}_i^+ = 0\right)\right] \\ =&\mathbb{E}\left[\overline{U}_i \left(\sum_{j=1}^{N} \overline{Q}_j^+\right) \mathcal{I}\left(\overline{Q}_i^+ = 0, \max_j \overline{Q}_j^+ \le r\sqrt{N-1} + r\right)\right] \qquad (24) \\ &+ \mathbb{E}\left[\overline{U}_i \left(\sum_{j=1}^{N} \overline{Q}_j^+\right) \mathcal{I}\left(\overline{Q}_i^+ = 0, \max_j \overline{Q}_j^+ > r\sqrt{N-1} + r\right)\right]. \qquad (25)\end{aligned}$$

Now, it remains to show that both Eqs. (24) and (25) approach 0 as $\epsilon \to 0$. To start with, we can bound Eq. (24) as follows.

$$\begin{aligned}&\mathbb{E}\left[\overline{U}_i \left(\sum_{j=1}^{N} \overline{Q}_j^+\right) \mathcal{I}\left(\overline{Q}_i^+ = 0, \max_j \overline{Q}_j^+ \le r\sqrt{N-1} + r\right)\right] \\ \le& r(N-1)(\sqrt{N-1}+1)\mathbb{E}\left[\overline{U}_i\right] \\ \le& r(N-1)(\sqrt{N-1}+1)\epsilon,\end{aligned}$$

where the last inequality follows from the fact $\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right] = \epsilon$. Thus, Eq. (24) approaches 0 as $\epsilon \to 0$ since $r^{(\epsilon)} = o(1/\epsilon)$.

Then, we can turn to bound Eq. (25) in the following way.

$$\begin{aligned}&\mathbb{E}\left[\overline{U}_i \left(\sum_{j=1}^{N} \overline{Q}_j^+\right) \mathcal{I}\left(\overline{Q}_i^+ = 0, \max_j \overline{Q}_j^+ > \sqrt{N-1}r + r\right)\right] \\ \overset{(a)}{\le}& S_{max}\mathbb{E}\left[\left\|\overline{\mathbf{Q}}\right\|_1 \mathcal{I}\left(\overline{Q}_i = 0, \max_j \overline{Q}_j > \sqrt{N-1}r + r\right)\right] \\ \overset{(b)}{\le}& S_{max}\sqrt{\mathbb{E}\left[\left\|\overline{\mathbf{Q}}\right\|_1^2\right] \mathbb{P}\left(d_{\mathcal{R}^{(r)}}\left(\overline{\mathbf{Q}}\right) \ge r\right)} \\ \overset{(c)}{\le}& S_{max}\sqrt{M_2 \frac{1}{\epsilon^2} \frac{C^*}{e^{\theta^* r}}},\end{aligned}$$

where (a) follows from the fact that $U_i(t) \leq S_i(t) \leq S_{max}$ for any $i \in \mathcal{N}$ and $t \geq 0$; (b) holds due to Cauchy-Schwartz inequality and the following facts. For any system state $Z(t)$ that satisfies $Q_i(t) = 0$ for some i and $\max_j \overline{Q}_j > \sqrt{N-1}r + r$, we have

$$d_{\mathcal{R}_l^{(r)}}(\mathbf{Q}(t)) > r\sqrt{N-1}$$
$$r \leq d_{\mathcal{R}_u^{(r)}}(\mathbf{Q}(t)) \leq r\sqrt{N-1}.$$

Thus,

$$d_{\mathcal{R}^{(r)}}(\mathbf{Q}(t)) = \min\{d_{\mathcal{R}_l^{(r)}}(\mathbf{Q}(t)), d_{\mathcal{R}_u^{(r)}}(\mathbf{Q}(t))\} \geq r,$$

and hence we have (b). The inequality (c) comes from the Chernoff bound, the moments bound in Lemma 3.1 and state-space collapse in Proposition 3.5, in which the constants M_2, C^* and θ^* are all independent of ϵ. Now, under the condition that $r^{(\epsilon)} \geq K\log(1/\epsilon)$ where $K = 2(1+\alpha)/\theta^*$ and $\alpha > 0$, we have that Eq. (25) approaches zero as $\epsilon \to 0$. Hence, we have completed the proof of Theorem 3.3.

□

5.3 Proof of Proposition 3.5

Before we present the proof, let us first introduce some useful results. First, let us define

$$V_\perp(Z(t)) \triangleq d_{\mathcal{R}^{(r)}}(\mathbf{Q}(t))$$
$$V_{\perp l}(Z(t)) \triangleq d_{\mathcal{R}_l^{(r)}}(\mathbf{Q}(t))$$
$$V_{\perp u}(Z(t)) \triangleq d_{\mathcal{R}_u^{(r)}}(\mathbf{Q}(t)).$$

By Eq. (8), we have $V_\perp(Z(t)) = \min\{V_{\perp l}(Z(t)), V_{\perp u}(Z(t))\}$. As a result, the drift of $V_\perp(Z)$ has the following four cases.

Case 1: $\Delta V_\perp(Z) = \Delta V_{\perp l}(Z)$
Case 2: $\Delta V_\perp(Z) = \Delta V_{\perp u}(Z)$
Case 3: $\Delta V_\perp(Z) = [V_{\perp l}(Z(t_0+1)) - V_{\perp u}(Z(t_0))]\mathcal{I}(Z(t_0) = Z)$
Case 4: $\Delta V_\perp(Z) = [V_{\perp u}(Z(t_0+1)) - V_{\perp l}(Z(t_0))]\mathcal{I}(Z(t_0) = Z)$.

Note that the drift in Case 3 can be upper bounded by $\Delta V_{\perp u}(Z)$ and the drift in Case 4 can be upper bounded by $\Delta V_{\perp l}(Z)$. Thus, in order to establish upper bounds on the drift of $V_\perp(Z)$, we only need to focus on the first two cases. *In the following, we might omit the superscript $^{(r)}$ for ease of exposition, and revive it when necessary.*

Let us also define

$$\mathcal{R}_l' \triangleq \mathcal{R}_l^{(r)} - \mathbf{r} \text{ and } \mathcal{R}_u' \triangleq \mathcal{R}_u^{(r)} - \mathbf{r}.$$

where $\mathbf{r} = r\mathbf{1}$. Correspondingly, we shift the queue-length vector in the same direction. That is, we let

$$\mathbf{Q}' = \mathbf{Q} - \mathbf{r}. \tag{26}$$

The main motivation behind this shifting process is that it allows us to decompose queue-length vector into parallel and perpendicular components. In particular, given a queue length vector $\mathbf{Q}$, we have the following decompositions

$$\mathbf{Q}' = \mathbf{Q}'_{\parallel\mathcal{R}_l'} + \mathbf{Q}'_{\perp\mathcal{R}_l'}$$
$$\mathbf{Q}' = \mathbf{Q}'_{\parallel\mathcal{R}_u'} + \mathbf{Q}'_{\perp\mathcal{R}_u'},$$

where $\mathbf{Q}'_{\parallel\mathcal{R}'_l}$ and $\mathbf{Q}'_{\parallel\mathcal{R}'_u}$ are the projections of $\mathbf{Q}'$onto $\mathcal{R}'_l$ and $\mathcal{R}'_u$, referred as parallel components. $\mathbf{Q}'_{\perp\mathcal{R}'_l}$ and $\mathbf{Q}'_{\perp\mathcal{R}'_u}$ are the corresponding remainders, referred as perpendicular components. Note that the two decompositions are well defined and unique because $\mathcal{R}'_l$ and $\mathcal{R}'_u$ are both closed and convex. Moreover, we have

$$V_{\perp l}(Z(t)) = \left\|\mathbf{Q}'_{\perp\mathcal{R}'_l}\right\| \text{ and } V_{\perp u}(Z(t)) = \left\|\mathbf{Q}'_{\perp\mathcal{R}'_u}\right\|. \tag{27}$$

This follows directly from the fact that the shifting process would not change the distance.

Now, we are ready to present our proof.

Proof of Proposition 3.5. Since the chain $\{Z(t), t \geq 0\}$ is ergodic under JBT for any $r \geq 1$ by Lemma 3.1, we can apply Lemma 5.1 to establish bounded moments of $\overline{V}_\perp$. In particular, all we need to do is to check the drift conditions (C1) and (C2), respectively. As discussed above, we should only focus on the drifts $\Delta V_{\perp l}(Z)$ and $\Delta V_{\perp u}(Z)$.

For condition (C2), we have the following result, the proof of which is relegated to Appendix C.

Claim 1. *For any $t \geq 0$, we have*

$$|\Delta V(Z(t))| \leq \sqrt{N}\max(A_{max}, S_{max}).$$

This directly verifies condition (C2) in Lemma 5.1. Now, we turn to check condition (C1) for $V_\perp(Z)$. To this end, we need the following result, the proof of which is relegated to Appendix D.

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

$$\begin{aligned}&\mathbb{E}\left[\Delta V_{\perp l}(Z) \mid Z(t) = Z\right]\\ \leq&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_l}(t)\|}\mathbb{E}\left[\left(2\langle\mathbf{Q}'_{\perp\mathcal{R}'_l}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z(t) = Z\right]\end{aligned} \tag{28}$$

and

$$\begin{aligned}&\mathbb{E}\left[\Delta V_{\perp u}(Z) \mid Z(t) = Z\right]\\ \leq&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\|}\mathbb{E}\left[\left(2\langle\mathbf{Q}'_{\perp\mathcal{R}'_u}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z(t) = Z\right]\end{aligned} \tag{29}$$

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

From Claim 2, we can see that the upper bounds on the mean drifts of $\Delta V_{\perp l}(Z)$ and $\Delta V_{\perp u}(Z)$ have the same formula. Thus, we can rewrite it in a compact way as follows.

$$\begin{aligned}&\mathbb{E}\left[\Delta V_{\perp s}(Z) \mid Z(t) = Z\right]\\ \leq&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_s}(t)\|}\mathbb{E}\left[\left(2\langle\mathbf{Q}'_{\perp\mathcal{R}'_s}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z(t) = Z\right]\end{aligned} \tag{30}$$

where the subscript $s \in \{l, u\}$. To upper bound the right-hand side of Eq. (30), we resort to the following result, the proof of which is relegated to Appendix E.

Claim 3. *For $s \in \{l, u\}$ and any system state $Z(t)$ with $V_\perp(Z(t)) > 0$, we have*

$$\mathbb{E}\left[\langle\mathbf{Q}'_{\perp\mathcal{R}'_s}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t) = Z\right] \leq -\frac{\mu_\Sigma\delta}{2N}\left\|\mathbf{Q}'_{\perp\mathcal{R}'_s}(t)\right\|,$$

whenever $\epsilon \leq \frac{\mu_\Sigma\delta}{2N+\delta}$, in which

$$\delta = \frac{\mu_{min}\mu_{min,2}}{\mu_\Sigma(\mu_\Sigma - \mu_{min})},$$

where $\mu_{min} = \min_{n\in\mathcal{N}} \mu_n$, i.e., the smallest service rate among all servers. $\mu_{min,2}$ is the second smallest service rate among all the servers. Hence, δ is a constant independent of ϵ.

Now substituting the upper bound in Claim 3 into Eq. (30), yields

$$\begin{aligned}
&\mathbb{E}\left[\Delta V_{\perp s}(Z) \mid Z(t) = Z\right] \\
\leq &\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_s}(t)\|}\mathbb{E}\left[\left(2\langle \mathbf{Q}'_{\perp\mathcal{R}'_s}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z(t) = Z\right] \\
\leq &-\frac{\mu_\Sigma\delta}{2N} + \frac{L}{2V_{\perp s}(Z)} \text{ whenever } \epsilon \leq \frac{\mu_\Sigma\delta}{2N+\delta} \\
\leq &-\frac{\mu_\Sigma\delta}{4N},
\end{aligned}$$

for $s \in \{l, u\}$ and for any $Z(t)$ such that $V_\perp(Z(t)) > 0$ and $V_{\perp s}(Z(t)) \geq \frac{2NL}{\mu_\Sigma\delta}$.

Therefore, since the drift of $V_\perp(Z(t))$ is either upper bounded by the drift of $V_{\perp l}(Z(t))$ or the drift $V_{\perp u}(Z(t))$, and $V_\perp(Z(t)) = \min\{V_{\perp l}(Z(t)), V_{\perp u}(Z(t))\}$, we have

$$\mathbb{E}\left[\Delta V_\perp(Z) \mid Z(t) = Z\right] \leq -\frac{\mu_\Sigma\delta}{4N} \text{ whenever } V_\perp(Z(t)) \geq \frac{2NL}{\mu_\Sigma\delta}$$

for any $\epsilon \leq \epsilon_0 \triangleq \frac{\mu_\Sigma\delta}{2N+\delta}$.

Thus, condition (C1) in Lemma 5.1 is validated with $\kappa = \frac{2NL}{\mu_\Sigma\delta}$ and $\eta = \frac{\mu_\Sigma\delta}{4N}$, both of which are independent of ϵ (since δ is independent of ϵ by Claim 3). Having established conditions (C1) and (C2) for the Lyapunov function $V_\perp(Z)$, by Lemma 5.1, we have that there exist some positive constants ϵ_0, θ^* and C^* such that for all $\epsilon \in (0, \epsilon_0)$

$$\mathbb{E}\left[e^{\theta^* d_{\mathcal{R}^{(r)}}\left(\overline{\mathbf{Q}}^{(\epsilon)}\right)}\right] \leq C^*,$$

where both θ^* and C^* are independent of ϵ. Hence, we have completed the proof of Proposition 3.5. □

6 CONCLUSION

We have investigated the performance of load balancing systems under a general pull-based policy with a varying threshold. In particular, we have shown that a necessary condition for steady-state heavy-traffic delay optimality is that the threshold must grow to infinity as the load intensity approaches one but its growth rate should be slower than a certain polynomial function of the mean number of tasks in the system. We then showed that a sufficient condition to guarantee steady-state heavy-traffic delay optimality in pull-based load balancing systems is that the threshold must grow logarithmically with the mean number of tasks in the system, which directly resolves a generalized version of the conjecture by Kelly and Laws [15]. Both of the necessary and sufficient conditions are achieved by overcoming various technical challenges, and the methods developed in this paper could be of independent interest. In particular, the methods developed in this paper might provide new directions on establishing steady-state delay optimality for dynamic threshold based scheduling policies in [2, 14].

We finally conjecture that a logarithmic growth rate of the threshold is also necessary for heavy-traffic delay optimality in pull-based load balancing systems, and one possible future work is to extend the current proof of Theorem 3.2 to prove this result, hence providing a tighter characterization of general pull-based load balancing schemes in heavy traffic.

REFERENCES

[1] Mor Armony. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51, 3-4 (2005), 287–329.

[2] Steven L Bell and Ruth J Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Annals of Applied Probability* (2001), 608–649.

[3] Maury Bramson. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* 30, 1-2 (1998), 89–140.

[4] Amarjit Budhiraja and Chihoon Lee. 2009. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research* 34, 1 (2009), 45–56.

[5] Hong Chen and Heng-Qing Ye. 2012. Asymptotic optimality of balanced routing. *Operations research* 60, 1 (2012), 163–179.

[6] JG Dai and Tolga Tezcan. 2011. State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research* 36, 2 (2011), 271–320.

[7] Atilla Eryilmaz and R Srikant. 2012. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72, 3-4 (2012), 311–359.

[8] G Foschini and J. Salz. 1978. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications* 26, 3 (1978), 320–327.

[9] David Gamarnik and Assaf Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability* 16, 1 (2006), 56–90.

[10] Itay Gurvich and Ward Whitt. 2009. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* 34, 2 (2009), 363–396.

[11] Bruce Hajek. 1982. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability* (1982), 502–525.

[12] Shlomo Halfin and Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* 29, 3 (1981), 567–588.

[13] Zhang Hanqin and Wang Rongxin. 1989. Heavy traffic limit theorems for a queueing system in which customers join the shortest line. *Advances in Applied Probability* 21, 2 (1989), 451–469.

[14] J Michael Harrison. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of applied probability* (1998), 822–848.

[15] FP Kelly and CN Laws. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing systems* 13, 1-3 (1993), 47–86.

[16] Yi Lu, Qiaomin Xie, Gabriel Kliot, Alan Geller, James R Larus, and Albert Greenberg. 2011. Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68, 11 (2011), 1056–1071.

[17] Siva Theja Maguluri, Sai Kiran Burle, and R Srikant. 2018. Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing Systems* 88, 3-4 (2018), 279–309.

[18] Siva Theja Maguluri and R. Srikant. 2016. Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stochastic Systems* 6, 1 (2016), 211–250.

[19] Siva Theja Maguluri, R Srikant, and Lei Ying. 2014. Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Performance Evaluation* 81 (2014), 20–39.

[20] Debankur Mukherjee, Sem C Borst, Johan SH Van Leeuwaarden, and Philip A Whiting. 2016. Universality of load balancing schemes on the diffusion scale. *Journal of Applied Probability* 53, 4 (2016), 1111–1124.

[21] Martin I Reiman. 1984. Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology*. Springer, 207–240.

[22] Alexander L Stolyar. 2015. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80, 4 (2015), 341–361.

[23] Yih-Choung Teh and Amy R Ward. 2002. Critical thresholds for dynamic routing in queueing networks. *Queueing Systems* 42, 3 (2002), 297–316.

[24] Weina Wang, Siva Theja Maguluri, R Srikant, and Lei Ying. 2018. Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *ACM SIGMETRICS Performance Evaluation Review* 45, 2 (2018), 232–245.

[25] Weina Wang, Kai Zhu, Lei Ying, Jian Tan, and Li Zhang. 2016. MapTask scheduling in MapReduce with data locality: Throughput and heavy-traffic optimality. *IEEE/ACM Transactions on Networking* 24, 1 (2016), 190–203.

[26] Ruth J Williams. 1998. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems* 30, 1 (1998), 27–88.

[27] Qiaomin Xie and Yi Lu. 2015. Priority algorithm for near-data scheduling: Throughput and heavy-traffic optimality. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*. 963–972.

[28] Qiaomin Xie, Ali Yekkehkhany, and Yi Lu. 2016. Scheduling with multi-level data locality: Throughput and heavy-traffic optimality. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*. 1–9.

[29] Xingyu Zhou, Jian Tan, and Ness Shroff. 2018. Flexible Load Balancing with Multi-dimensional State-space Collapse: Throughput and Heavy-traffic Delay Optimality. *arXiv preprint arXiv:1806.02939* (2018).

[30] Xingyu Zhou, Fei Wu, Jian Tan, Yin Sun, and Ness Shroff. 2017. Designing Low-Complexity Heavy-Traffic Delay-Optimal Load Balancing Schemes: Theory to Algorithms. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1, 2 (2017), 39.

APPENDIX

A PROOF OF LEMMA 3.1

Proof. To begin with, we first show that the Markov chain $\{Z(t) = (\mathbf{Q}(t), m(t)), t \geq 0\}$ is irreducible and aperiodic. Let the initial state be $Z(0) = (\mathbf{Q}(0), m(0)) = (0_{1\times N}, m_0)$ where m_0 is the memory state in which all the N IDs of servers are in the memory. The Markov chain is irreducible since for any state Z in the state space, the Markov chain is able to reach the initial state within a finite step. This happens when there are no exogenous arrivals and all the offered service is at least one during each time-slot, which has a positive probability under our assumptions. The aperiodicity of the Markov chain $\{Z(t) = (\mathbf{Q}(t), m(t)), t \geq 0\}$ follows from the fact that the transition probability from the initial state to itself is positive. In order to show positive recurrence, we adopt the Foster-Lyapunov theorem. In particular, we only need to consider the Lyapunov function $W(Z) \triangleq \|\mathbf{Q}\|^2$ since the memory state is finite. Now for any t_0, the one-step drift is given by

$$\begin{aligned}
&\mathbb{E}\left[W(Z(t_0+1)) - W(Z(t_0)) \mid Z(t_0)\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t_0) + \mathbf{A}(t_0) - \mathbf{S}(t_0) + \mathbf{U}(t_0)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0)\right] \\
\overset{(a)}{\leq}&\mathbb{E}\left[\|\mathbf{Q}(t_0) + \mathbf{A}(t_0) - \mathbf{S}(t_0)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0)\right] \\
=&\mathbb{E}\left[2\langle \mathbf{Q}(t_0), \mathbf{A}(t_0) - \mathbf{S}(t_0)\rangle + \|\mathbf{A}(t_0) - \mathbf{S}(t_0)\|^2 \mid Z(t_0)\right] \\
\overset{(b)}{\leq}&\mathbb{E}\left[2\langle \mathbf{Q}(t_0), \mathbf{A}(t_0) - \mathbf{S}(t_0)\rangle \mid Z(t_0)\right] + L \\
\overset{(c)}{\leq}&2\sum_{n=1}^{N} Q_n(t_0)\left(-\epsilon\frac{\mu_n}{\mu_\Sigma}\right) + L \\
\overset{(d)}{\leq}&-2\epsilon\frac{\mu_{min}}{\mu_\Sigma}\|\mathbf{Q}(t_0)\| + L,
\end{aligned} \tag{31}$$

where (a) follows from the facts that $Q_n(t)+A_n(t)-S_n(t)+U_n(t) = \max(Q_n(t)+A_n(t)-S_n(t), 0)$ for any $t \geq 0$, and $(\max(a, 0))^2 \leq a^2$ for any $a \in \mathbb{R}$; (b) holds since both the arrival and service processes have finite supports and $L = N\max(A_{max}, S_{max})^2$; (c) is true since under the JBT policy the worst case is when (proportionally) random routing is adopted, which happens if the ID in memory is either empty or full; (d) comes from the fact that $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^N$. Therefore, by the Foster-Lyapunov theorem, the Markov chain $\{Z(t) = (\mathbf{Q}(t), m(t)), t \geq 0\}$ is positive recurrent.

Having established the fact that $\{Z(t) = (\mathbf{Q}(t), m(t)), t \geq 0\}$ is irreducible, aperiodic and positive recurrent, we are now ready to apply Lemma 5.1 to show bounded moments of $\left\|\overline{\mathbf{Q}}\right\|$. Let us consider the Lyapunov function $V(Z) = \|\mathbf{Q}\|$, and check the two conditions (C1) and (C2) in Lemma 5.1, respectively.

For condition (C1), we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta V(Z) \mid Z(t_0) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t_0+1)\| - \|\mathbf{Q}(t_0)\| \mid Z(t_0) = Z\right] \\
=&\mathbb{E}\left[\sqrt{\|\mathbf{Q}(t_0+1)\|^2} - \sqrt{\|\mathbf{Q}(t_0)\|^2} \mid Z(t_0) = Z\right] \\
\overset{(a)}{\leq}&\frac{1}{2\|\mathbf{Q}(t_0)\|}\mathbb{E}\left[\|\mathbf{Q}(t_0+1)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0) = Z\right] \\
\overset{(b)}{\leq}&-\epsilon\frac{\mu_{min}}{\mu_\Sigma} + \frac{L}{2\|\mathbf{Q}(t_0)\|},
\end{aligned}$$

where (a) follows from the fact that $f(x) = \sqrt{x}$ is concave; (b) comes from Eq. (31). Thus, condition (C1) is valid with $\kappa = \frac{L\mu_\Sigma}{\epsilon\mu_{min}}$ and $\eta = \frac{\epsilon\mu_{min}}{2\mu_\Sigma}$.

For condition (C2), we have

$$\begin{aligned}
|\Delta V(Z)| &= \left|\|\mathbf{Q}(t_0+1)\| - \|\mathbf{Q}(t_0)\|\right| \mathcal{I}\left(Z(t_0) = Z\right) \\
&\overset{(a)}{\leq} \|\mathbf{Q}(t_0+1) - \mathbf{Q}(t_0)\| \, \mathcal{I}\left(Z(t_0) = Z\right) \\
&\overset{(b)}{\leq} \sqrt{N}\max(A_{max}, S_{max}),
\end{aligned}$$

where (a) holds since $|\,\|\mathbf{x}\| - \|\mathbf{y}\|\,| \leq \|\mathbf{x}-\mathbf{y}\|$ for each $\mathbf{x}$, $\mathbf{y}$ in $\mathbb{R}^N$; (b) follows from the assumptions that $A_\Sigma(t) \leq A_{max}$ and $S_n(t) \leq S_{max}$ for any $t \geq 0$ and $n \in \mathcal{N}$. Thus, condition (C2) is valid with $D = \sqrt{N}\max(A_{max}, S_{max})$.

Therefore, according to Eq. (13) in Lemma 5.1, we get for $p = 1, 2, \ldots,$

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|^p\right] &\leq \frac{1}{\epsilon^p}\left(\frac{2L\mu_\Sigma}{\mu_{min}}\right)^p + \frac{1}{\epsilon^p}\left(\frac{8D\mu_\Sigma}{\mu_{min}}\right)^p (D+\mu_{min})^p p! \\
&\leq \frac{M_p}{\epsilon^p},
\end{aligned}$$

where the constant $M_p = \left(\frac{2L\mu_\Sigma}{\mu_{min}}\right)^p + p!\left(\frac{8D\mu_\Sigma}{\mu_{min}}\right)^p (D+\mu_{min})^p$.

In addition, if we apply Theorem 2.3 in [11], we can obtain that

$$\mathbb{E}\left[e^{\theta^*\|\overline{\mathbf{Q}}^{(\epsilon)}\|}\right] \leq K_1 e^{\theta^* K_2/\epsilon}, \tag{32}$$

where the positive constants θ^*, K_1 and K_2 are all independent of ϵ. □

B PROOF OF LEMMA 5.3

Proof. Let us consider the following Lyapunov function:

$$V_1(Z) \triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\left(Q_i - Q_j\right)^2.$$

We start with the conditional mean drift of $V_1(Z)$. Note that we shall omit the time reference (t) after the first step and $\mathbf{Q}^+ \triangleq \mathbf{Q}(t+1)$.

$$\begin{aligned}
&\mathbb{E}\left[V_1(Z(t+1)) - V_1(Z(t)) \mid Z(t) = Z\right] \\
&= \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbb{E}\left[\left(Q_i(t+1) - Q_j(t+1)\right)^2 - \left(Q_i(t) - Q_j(t)\right)^2 \mid Z(t) = Z\right] \\
&= \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbb{E}\left[2\left(Q_i - Q_j\right)\left(A_i - A_j - S_i + S_j\right) - \left(U_i - U_j\right)^2 \mid Z\right] \\
&\quad + \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbb{E}\left[\left(A_i - A_j - S_i + S_j\right)^2 + 2\left(Q_i^+ - Q_j^+\right)\left(U_i - U_j\right) \mid Z\right] \\
&\overset{(a)}{=} \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbb{E}\left[2\left(Q_i - Q_j\right)\left(A_i - A_j\right) - \left(U_i - U_j\right)^2 \mid Z\right] \\
&\quad + \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbb{E}\left[\left(A_i - A_j - S_i + S_j\right)^2 - 2\left(Q_i^+ U_j + Q_j^+ U_i\right) \mid Z\right],
\end{aligned}$$

in which (a) follows from the fact that the service is independent of queue lengths and homogeneous, as well as $Q_n(t+1)U_n(t) = 0$ for all n and $t > 0$.

Since $\|\mathbf{Q}\|$ has a finite second moment in steady state under JBT by Lemma 3.1, the steady-state mean $\mathbb{E}\left[V_1(\overline{Z}^{(\epsilon)})\right]$ is finite for any $\epsilon > 0$. As a result, the mean drift of $V_1(\cdot)$ is zero in steady state, which directly implies the result in Lemma 5.3. □

C PROOF OF CLAIM 1

Proof. For any $t_0 \geq 0$, we have

$$\begin{aligned}
&|\Delta V_{\perp l}(Z)| \\
&\overset{(a)}{=} \left|\|\mathbf{Q}'_{\perp \mathcal{R}'_l}(t_0+1)\| - \|\mathbf{Q}'_{\perp \mathcal{R}'_l}(t_0)\|\right| \mathcal{I}\left(Z(t_0) = Z\right) \\
&\overset{(b)}{\leq} \|\mathbf{Q}'_{\perp \mathcal{R}'_l}(t_0+1) - \mathbf{Q}'_{\perp \mathcal{R}'_l}(t_0)\| \mathcal{I}\left(Z(t_0) = Z\right) \\
&\overset{(c)}{\leq} \|\mathbf{Q}'(t_0+1) - \mathbf{Q}'(t_0)\| \mathcal{I}\left(Z(t_0) = Z\right) \\
&\overset{(d)}{=} \|\mathbf{Q}(t_0+1) - \mathbf{Q}(t_0)\| \mathcal{I}\left(Z(t_0) = Z\right) \\
&\overset{(e)}{\leq} \sqrt{N}\max(A_{max}, S_{max}),
\end{aligned}$$

where (a) follows from Eq. (27); (b) comes from the fact that $|\,\|\mathbf{x}\| - \|\mathbf{y}\|\,| \leq \|\mathbf{x} - \mathbf{y}\|$ holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$; (c) is due to the non-expansive property of projection and the fact that $\mathbf{Q}'_{\perp \mathcal{R}'_l}$ is the projection of $\mathbf{Q}'$ onto the polar cone of $\mathcal{R}'_l$; (d) follows from the definition of $\mathbf{Q}'$ in Eq. (26); (e) 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 $1 \leq n \leq N$. With the same arguments, we can establish that

$$|\Delta V_{\perp u}(Z)| \leq \sqrt{N}\max(A_{max}, S_{max}).$$

Since the drift of $V_\perp(Z)$ is either upper bounded by $\Delta V_{\perp l}(Z)$ or $\Delta V_{\perp u}(Z)$, we finally get

$$|\Delta V_\perp(Z)| \le \sqrt{N}\max(A_{max}, S_{max}).$$

□

D PROOF OF CLAIM 2

Proof. We first start with inequality (29) in Claim 2. Let us define

$$\Delta W(Z) = \left[\|\mathbf{Q}'(t+1)\|^2 - \|\mathbf{Q}'(t)\|^2\right] \mathcal{I}(Z(t) = Z)$$
$$\Delta W_{\|u}(Z) = \left[\|\mathbf{Q}'_{\|\mathcal{R}'_u}(t+1)\|^2 - \|\mathbf{Q}'_{\|\mathcal{R}'_u}(t)\|^2\right] \mathcal{I}(Z(t) = Z).$$

.

Then, the mean drift of $\Delta V_{\perp u}(Z)$ can be decomposed as follows.

$$\begin{aligned}
&\mathbb{E}\left[\Delta V_{\perp u}(Z) \mid Z(t) = Z\right] \\
\overset{(a)}{=}&\mathbb{E}\left[\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t+1)\| - \|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\| \mid Z(t) = Z\right] \\
=&\left[\sqrt{\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t+1)\|^2} - \sqrt{\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\|^2}\right] \mathcal{I}(Z(t) = Z) \\
\overset{(b)}{\le}&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\|}\mathbb{E}\left[\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t+1)\|^2 - \|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\|^2 \mid Z(t) = Z\right] \\
\overset{(c)}{=}&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\|}\mathbb{E}\left[\Delta W(Z) - \Delta W_{\|u}(Z) \mid Z(t) = Z\right]
\end{aligned} \tag{33}$$

where (a) follows from Eq. (27); (b) holds due to the concavity of function $f(x) = \sqrt{x}$ for $x \ge 0$; (c) comes from the Pythagorean theorem. Next, we will bound each term in Eq. (33), respectively. To begin with, we have an upper bound for the first term as follows.

$$\begin{aligned}
&\mathbb{E}\left[\Delta W(Z) \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}'(t+1)\|^2 - \|\mathbf{Q}'(t)\|^2 \mid Z(t) = Z\right] \\
\overset{(a)}{=}&\mathbb{E}\left[\|\mathbf{Q}(t+1) - \mathbf{r}\|^2 - \|\mathbf{Q}(t) - \mathbf{r}\|^2 \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t) + \mathbf{A}(t) - \mathbf{S}(t) + \mathbf{U}(t) - \mathbf{r}\|^2 - \|\mathbf{Q}(t) - \mathbf{r}\|^2 \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t) + \mathbf{A}(t) - \mathbf{S}(t) - \mathbf{r}\|^2 - \|\mathbf{Q}(t) - \mathbf{r}\|^2 \mid Z(t) = Z\right] \\
&+ \mathbb{E}\left[\|\mathbf{U}(t)\|^2 + 2\langle\mathbf{Q}(t+1) - \mathbf{r} - \mathbf{U}(t), \mathbf{U}(t)\rangle \mid Z(t) = Z\right] \\
\overset{(b)}{\le}&\mathbb{E}\left[2\langle\mathbf{Q}'(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + \|\mathbf{A}(t) - \mathbf{S}(t)\|^2 - 2\langle\mathbf{r}, \mathbf{U}(t)\rangle \mid Z(t) = Z\right] \\
\overset{(c)}{\le}&\mathbb{E}\left[2\langle\mathbf{Q}'(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle - 2\langle\mathbf{r}, \mathbf{U}(t)\rangle \mid Z(t) = Z\right] + L,
\end{aligned} \tag{34}$$

where (a) follows from Eq. (26); (b) holds because of $\langle\mathbf{Q}(t+1), \mathbf{U}(t)\rangle = 0$ and the dropping of $-\|\mathbf{U}(t)\|^2$; in (c), $L = N\max(A_{max}, S_{max})^2$, which is true since both the arrival and service processes have finite support.

We now turn to provide a lower bound on the second term in Eq. (33) as follows.

$$\begin{aligned}
&\mathbb{E}\left[\Delta W_{\|u}(Z) \mid Z(t)=Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}'_{\|\mathcal{R}'_u}(t+1)\|^2 - \|\mathbf{Q}'_{\|\mathcal{R}'_u}(t)\|^2 \mid Z(t)=Z\right] \\
=&\mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{Q}'_{\|\mathcal{R}'_u}(t+1) - \mathbf{Q}'_{\|\mathcal{R}'_u}(t)\rangle \mid Z\right] \\
&+\mathbb{E}\left[\left\|\mathbf{Q}'_{\|\mathcal{R}'_u}(t+1) - \mathbf{Q}'_{\|\mathcal{R}'_u}(t)\right\|^2 \mid Z\right] \\
\geq&\mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{Q}'_{\|\mathcal{R}'_u}(t+1) - \mathbf{Q}'_{\|\mathcal{R}'_u}(t)\rangle \mid Z\right] \\
=&2\mathbb{E}\left[\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{Q}'(t+1) - \mathbf{Q}'(t)\rangle \mid Z\right] \\
&-2\mathbb{E}\left[\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{Q}'_{\perp\mathcal{R}'_u}(t+1) - \mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\rangle \mid Z\right] \\
\overset{(a)}{\geq}&\mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{Q}'(t+1) - \mathbf{Q}'(t)\rangle \mid Z\right] \\
\overset{(b)}{\geq}&\mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z\right],
\end{aligned} \tag{35}$$

where (a) holds because $\langle \mathbf{Q}'_{\|\mathcal{R}'_u}(t), \mathbf{Q}'_{\perp\mathcal{R}'_u}(t)\rangle = 0$ and $\langle \mathbf{Q}'_{\perp\mathcal{R}'_u}(t+1), \mathbf{Q}'_{\|\mathcal{R}'_u}(t)\rangle \leq 0$ since $\mathbf{Q}'_{\perp\mathcal{R}'_u}(t+1)$ is in the polar cone of $\mathcal{R}'_u$; (b) follows from Eq. (26) and the fact that all the components of $\mathbf{Q}'_{\|\mathcal{R}'_u}(t)$ and $\mathbf{U}(t)$ are nonnegative. Thus, substituting Eqs. (34) and (35) into Eq. (33), yields

$$\begin{aligned}
&\mathbb{E}\left[\Delta V_{\perp l}(Z) \mid Z(t)=Z\right] \\
\leq&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_l}(t)\|}\mathbb{E}\left[\left(2\langle \mathbf{Q}'_{\perp\mathcal{R}'_l}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) - 2\langle \mathbf{r}, \mathbf{U}(t)\rangle \mid Z\right] \\
\overset{(a)}{\leq}&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_l}(t)\|}\mathbb{E}\left[\left(2\langle \mathbf{Q}'_{\perp\mathcal{R}'_l}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z\right]
\end{aligned}$$

where (a) holds since all the components of $\mathbf{r}$ and $\mathbf{U}(t)$ are nonnegative. Thus, we have the bound in Eq. (29) of Claim 2.

Next, we turn to the bound in inequality (28). Let us define

$$\Delta W_{\|l}(Z) = \left[\|\mathbf{Q}'_{\|\mathcal{R}'_l}(t+1)\|^2 - \|\mathbf{Q}'_{\|\mathcal{R}'_l}(t)\|^2\right]\mathcal{I}(Z(t)=Z).$$

With the same arguments as in Eq. (33), the mean drift of $\Delta V_{\perp l}(Z)$ can be decomposed into two terms.

$$\begin{aligned}
&\mathbb{E}\left[\Delta V_{\perp l}(Z) \mid Z(t)=Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}'_{\perp\mathcal{R}'_l}(t+1)\| - \|\mathbf{Q}'_{\perp\mathcal{R}'_l}(t)\| \mid Z(t)=Z\right] \\
\leq&\frac{1}{2\|\mathbf{Q}'_{\perp\mathcal{R}'_l}(t)\|}\mathbb{E}\left[\Delta W(Z) - \Delta W_{\|l}(Z) \mid Z(t)=Z\right].
\end{aligned} \tag{36}$$

The first term can be upper bounded as in Eq. (34). The second term can be lower bonded in a similar way as in Eq. (35) except the last step.

$$\begin{aligned}
&\mathbb{E}\left[\Delta W_{\|l}(Z) \mid Z(t) = Z\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_l}(t), \mathbf{Q}'(t+1) - \mathbf{Q}'(t)\rangle \mid Z\right] \\
&\overset{(b)}{=} \mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_l}(t), \mathbf{A}(t) - \mathbf{S}(t) + \mathbf{U}(t)\rangle \mid Z\right] \\
&\overset{(c)}{\geq} \mathbb{E}\left[2\langle \mathbf{Q}'_{\|\mathcal{R}'_l}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle - 2\langle \mathbf{r}, \mathbf{U}(t)\rangle \mid Z\right], \qquad (37)
\end{aligned}$$

where (a) follows from the same arguments as in Eq. (35); (b) comes from the definition of $\mathbf{Q}'$ in Eq. (26); (c) is true since any component of $\mathbf{Q}'_{\|\mathcal{R}'_l}(t)$ is greater or equal to $-r$ by the definition of $\mathcal{R}'_l$. Thus, substituting Eqs. (34) and (37) into Eq. (36) yields the bound in Eq. (28) of Claim 2. Hence, we complete the proof of Claim 2. □

E PROOF OF CLAIM 3

Proof. In order to analyze the inner product in Eq. (30), it is advantageous to reorder the queue-length vector $\mathbf{Q}(t)$. More precisely, let $\sigma_t(\cdot)$ be a permutation of $(1, 2, \ldots, N)$ such that $Q_{\sigma_t(1)}(t) \leq Q_{\sigma_t(2)}(t) \leq \ldots \leq Q_{\sigma_t(N)}(t)$ and ties are broken randomly. We define the permutation vectors as follows

$$\begin{aligned}
\widehat{\mathbf{Q}}(t) &\triangleq (Q_{\sigma_t(1)}(t), Q_{\sigma_t(2)}(t), \ldots, Q_{\sigma_t(N)}(t)) \\
\widehat{\mathbf{A}}(t) &\triangleq (A_{\sigma_t(1)}(t), A_{\sigma_t(2)}(t), \ldots, A_{\sigma_t(N)}(t)) \\
\widehat{\mathbf{S}}(t) &\triangleq (S_{\sigma_t(1)}(t), S_{\sigma_t(2)}(t), \ldots, S_{\sigma_t(N)}(t)).
\end{aligned}$$

Let $p_n(t)$ be the probability that the new arrivals are dispatched to queue n at time-slot t, and $\widehat{\mathbf{P}}(t) = (p_{\sigma_t(1)}(t), p_{\sigma_t(2)}(t), \ldots, p_{\sigma_t(N)}(t))$, i.e., the i-th component of $\widehat{\mathbf{P}}(t)$ is the probability of dispatching arrivals to the i-th shortest queue at time-slot t. We define

$$\Delta(t) = \widehat{\mathbf{P}}(t) - \widehat{\mathbf{P}}_{\text{rand}}(t), \qquad (38)$$

where $\widehat{\mathbf{P}}_{\text{rand}}(t)$ denotes the permutation of the dispatching distribution $\mathbf{p}(t)$ under proportionally random routing, i.e., the i-th component of $\widehat{\mathbf{P}}_{\text{rand}}(t)$ is $\mu_{\sigma_t(i)}/\mu_\Sigma$.

As before, we let $\widehat{\mathbf{Q}}'(t) = \widehat{\mathbf{Q}}(t) - \mathbf{r}$. By the symmetry of $\mathcal{R}'_s$ with respect to the line $\mathbf{1} = (1, 1, \ldots, 1)$, we have that the permutation of the perpendicular component $\mathbf{Q}'_{\perp\mathcal{R}'_s}(t)$ is equal to the perpendicular component of the permutation of $\mathbf{Q}'(t)$, which is denoted by $\widehat{\mathbf{Q}}'_{\perp s}(t)$. That is, $\widehat{\mathbf{Q}}'_{\perp s}(t) = \widehat{\mathbf{Q}}'(t) - \widehat{\mathbf{Q}}'_{\|\mathcal{R}'_s}(t)$ in which $\widehat{\mathbf{Q}}'_{\|\mathcal{R}'_s}(t)$ is the projection of the vector $\widehat{\mathbf{Q}}'(t)$ onto $\mathcal{R}'_s$ and $s \in \{l, u\}$.

Based on the notions introduced above, the inner product in Eq. (30) can be rewritten as follows.

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{Q}'_{\perp \mathcal{R}'_s}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t) = Z\right] \\
\overset{(a)}{=}&\mathbb{E}\left[\langle \widehat{\mathbf{Q}}'_{\perp s}(t), \widehat{\mathbf{A}}(t) - \widehat{\mathbf{S}}(t)\rangle \mid Z(t) = Z\right] \\
\overset{(b)}{=}&\sum_{n=1}^{N} \widehat{Q}'_{\perp s,n}(t)\left[\lambda_\Sigma\left(\Delta_n(t) + \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma}\right) - \mu_{\sigma_t(n)}\right] \\
\overset{(c)}{=}&\sum_{n=1}^{N} \widehat{Q}'_{\perp s,n}(t)\Delta_n(t)\lambda_\Sigma + \sum_{n=1}^{N} \widehat{\mathbf{Q}}'_{\perp s,n}(t)\left(-\epsilon\frac{\mu_{\sigma_t(n)}}{\mu_\Sigma}\right) \\
\leq&\sum_{n=1}^{N} \widehat{Q}'_{\perp s,n}(t)\Delta_n(t)\lambda_\Sigma + \epsilon\left\|\widehat{\mathbf{Q}}'_{\perp s}(t)\right\|_1,
\end{aligned} \tag{39}$$

where (a) follows from the fact inner product remains the same under permutation and the fact that the permutation of $\mathbf{Q}'_{\perp \mathcal{R}'_s}(t)$ is equal to $\widehat{\mathbf{Q}}'_{\perp s}(t)$ as shown above; (b) holds due to the definition of $\Delta(t)$ and $\widehat{Q}'_{\perp s,n}(t)$ is the n-th component of $\widehat{\mathbf{Q}}'_{\perp s}(t)$; (c) simply follows from $\lambda_\Sigma = \mu_\Sigma - \epsilon$.

In order to further analyze Eq. (39), we need the following results, which are proved at the end of this proof.

Claim 4. *Regarding the vectors $\widehat{\mathbf{Q}}'_{\perp s}(t)$ and $\Delta(t)$ in Eq. (39), we have the following properties for any system state $Z(t)$ such that $V_\perp(Z(t)) > 0$.*

(a) The vector $\widehat{\mathbf{Q}}'_{\perp s}(t)$ satisfies $\widehat{Q}'_{\perp s,1}(t) \leq \widehat{Q}'_{\perp s,2}(t) \leq \ldots \leq \widehat{Q}'_{\perp s,N}(t)$ and $\widehat{Q}'_{\perp s,1}(t) \leq 0, \widehat{Q}'_{\perp s,N}(t) \geq 0$, where $s \in \{l, u\}$. More precisely, we have

$$\widehat{Q}'_{\perp l,1}(t) = 0 \text{ and } \widehat{\mathbf{Q}}'_{\perp l,N}(t) > 0 \tag{40}$$

$$\widehat{Q}'_{\perp u,1}(t) < 0 \text{ and } \widehat{\mathbf{Q}}'_{\perp u,N}(t) = 0. \tag{41}$$

(b) The vector $\Delta(t)$ satisfies for some $k \in \{2, 3, \ldots, N\}$

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

and

$$\min\left(|\Delta_1(t)|, |\Delta_N(t)|\right) \geq \delta,$$

for some constant δ that is independent of ϵ.

Based on Claim 4, we can bound the first term in Eq. (39) for any system state $Z(t)$ such that $V_\perp(Z(t)) > 0$ as follows

$$\sum_{n=1}^{N} \widehat{Q}'_{\perp s,n}(t)\Delta_n(t)\lambda_\Sigma \leq -\lambda_\Sigma\delta\left(|\widehat{Q}'_{\perp s,1}(t)| + |\widehat{Q}'_{\perp s,N}(t)|\right). \tag{42}$$

This inequality can be verified as follows. Since $\Delta(t)$ satisfies the property (b) in Claim 4, it can be constructed in the following way. To start with, all the $\Delta_n(t)$ is equal to 0. Then, we decrease $\Delta_N(t)$ and increase $\Delta_1(t)$ by the same amount of δ. After this process, the left-hand side of Eq. (42) is equal to $\lambda_\Sigma(\delta\widehat{Q}'_{\perp s,1}(t) - \delta\widehat{Q}'_{\perp s,N}(t))$, which is equivalent to the right-hand side of Eq. (42) because of $\widehat{Q}'_{\perp s,1}(t) \leq 0$, $\widehat{Q}'_{\perp s,N}(t) \geq 0$ in (a) of Claim 4. Then, due to the first condition in (b) of Claim 4 and the fact that $\sum_{n=1}^{N}\Delta_n(t) = 0$, any further construction (if necessary) for $\Delta(t)$ can only take the following way: it decreases some amount (say β) from $\Delta_i(t)$ where $i \geq k$, and then increase the same amount, i.e., β for some $\Delta_j(t)$ where $j < k$. Through this process, the left-hand side of Eq. (42)

can only further decrease due to the monotone nondecreasing property of $\widehat{\mathbf{Q}}'_{\perp s}(t)$ in (a) of Claim 4. As a result, we have established the upper bound in Eq. (42).

Next, we can further bound the right-hand side of Eq. (42) in terms of $\|\widehat{\mathbf{Q}}'_{\perp s}(t)\|_1$. First, consider the case when $s = l$, we have

$$\begin{aligned}\sum_{n=1}^{N} \widehat{Q}'_{\perp l,n}(t)\Delta_n(t)\lambda_\Sigma &\le -\lambda_\Sigma\delta\left(|\widehat{Q}'_{\perp l,1}(t)| + |\widehat{Q}'_{\perp l,N}(t)|\right)\\ &\le -\lambda_\Sigma\delta|\widehat{Q}'_{\perp l,N}(t)|\\ &\overset{(a)}{\le} \frac{-\lambda_\Sigma\delta}{N}\left\|\widehat{\mathbf{Q}}'_{\perp l}(t)\right\|_1 \end{aligned} \tag{43}$$

where (a) holds since $\|\widehat{\mathbf{Q}}'_{\perp l}(t)\|_1 \le N|\widehat{Q}'_{\perp l,N}(t)|$ by the monotone nondecreasing property of $\widehat{\mathbf{Q}}'_{\perp s}(t)$ and Eq. (40) in (a) of Claim 4. Similarly, when $s = u$, we have

$$\begin{aligned}\sum_{n=1}^{N} \widehat{Q}'_{\perp u,n}(t)\Delta_n(t)\lambda_\Sigma &\le -\lambda_\Sigma\delta\left(|\widehat{Q}'_{\perp u,1}(t)| + |\widehat{Q}'_{\perp u,N}(t)|\right)\\ &\le -\lambda_\Sigma\delta|\widehat{Q}'_{\perp u,1}(t)|\\ &\overset{(a)}{\le} \frac{-\lambda_\Sigma\delta}{N}\left\|\widehat{\mathbf{Q}}'_{\perp u}(t)\right\|_1 \end{aligned} \tag{44}$$

where (a) holds since $\|\widehat{\mathbf{Q}}'_{\perp u}(t)\|_1 \le N|\widehat{Q}'_{\perp l,N}(t)|$ by the monotone nondecreasing property of $\widehat{\mathbf{Q}}'_{\perp s}(t)$ and Eq. (41) in (a) of Claim 4.

Therefore, based on Eqs. (43) and (44), the left-hand side of Eq. (42) can be upper bounded in terms of $\|\widehat{\mathbf{Q}}'_{\perp s}(t)\|_1$ as follows.

$$\sum_{n=1}^{N} \widehat{Q}'_{\perp s,n}(t)\Delta_n(t)\lambda_\Sigma \le \frac{-\lambda_\Sigma\delta}{N}\left\|\widehat{\mathbf{Q}}'_{\perp s}(t)\right\|_1 \tag{45}$$

for $s \in \{l, u\}$ and any system state $Z(t)$ with $V_\perp(Z(t)) > 0$. Now, substituting Eq. (45) into Eq. (39), yields

$$\begin{aligned}&\mathbb{E}\left[\langle \mathbf{Q}'_{\perp\mathcal{R}'_s}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t) = Z\right]\\ &\le \left(\epsilon - \frac{\lambda_\Sigma\delta}{N}\right)\left\|\widehat{\mathbf{Q}}'_{\perp s}(t)\right\|_1\\ &\le -\frac{\mu_\Sigma\delta}{2N}\left\|\widehat{\mathbf{Q}}'_{\perp s}(t)\right\|_1 \text{ whenever } \epsilon \le \frac{\mu_\Sigma\delta}{2N+\delta}\\ &\le -\frac{\mu_\Sigma\delta}{2N}\left\|\mathbf{Q}'_{\perp\mathcal{R}'_s}(t)\right\|,\end{aligned}$$

for $s \in \{l, u\}$ and any system state $Z(t)$ with $V_\perp(Z(t)) > 0$, in which the last inequality follows from the fact $\|\mathbf{Q}'_{\perp\mathcal{R}'_s}(t)\|_1 = \|\widehat{\mathbf{Q}}'_{\perp s}(t)\|_1$ and $\|\mathbf{x}\|_1 \ge \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^N$. Hence, we establish the result in Claim 3.

Now, we give the proof of Claim 4.

For (a), by the definition of $\widehat{\mathbf{Q}}'(t)$, we have $\widehat{Q}'_1(t) \le \widehat{Q}'_2(t) \le \ldots \le \widehat{Q}'_N(t)$. The projection of $\widehat{\mathbf{Q}}'(t)$ onto $\mathcal{R}'_u$, which is equal to $\widehat{\mathbf{Q}}'_{\perp l}(t)$, is given by

$$\widehat{\mathbf{Q}}'_{\perp l}(t) = \widehat{\mathbf{Q}}'_{\|u}(t) = \max\left(\widehat{\mathbf{Q}}'(t), \mathbf{0}\right). \tag{46}$$

As a result, we have

$$\widehat{\mathbf{Q}}'_{\perp u}(t) = \widehat{\mathbf{Q}}'(t) - \widehat{\mathbf{Q}}'_{\|u}(t) = \min\left(\widehat{\mathbf{Q}}'(t), \mathbf{0}\right). \tag{47}$$

Therefore, we have $\widehat{Q}'_{\perp s,1}(t) \leq \widehat{Q}'_{\perp s,2}(t) \leq \ldots \leq \widehat{Q}'_{\perp s,N}(t)$ for $s \in \{l, u\}$. Moreover, since $V_\perp(Z(t)) > 0$, we have $\mathbf{Q}(t) \notin \mathcal{R}^{(r)}$, which implies that $\mathbf{Q}'(t) \notin \mathcal{R}'_l$ and $\mathbf{Q}'(t) \notin \mathcal{R}'_u$. Thus, we have there exist queues i and j such that $Q'_i(t) < 0$ and $Q'_j(t) > 0$, which further gives $\widehat{Q}'_1(t) < 0$ and $\widehat{Q}'_N(t) > 0$. As a result, by Eqs. (46) and (47), we have

$$\widehat{Q}'_{\perp l,1}(t) = 0 \text{ and } \widehat{\mathbf{Q}}'_{\perp l,N}(t) > 0$$
$$\widehat{Q}'_{\perp u,1}(t) < 0 \text{ and } \widehat{\mathbf{Q}}'_{\perp u,N}(t) = 0,$$

which establishes $\widehat{Q}'_{\perp s,1}(t) \leq 0$ and $\widehat{Q}'_{\perp s,N}(t) \geq 0$, where $s \in \{l, u\}$. Hence, we have completed the proof of (a) in Claim 4.

Now let us consider (b) in Claim 4. First, since $V_\perp(Z(t)) > 0$, we have $\mathbf{Q}(t) \notin \mathcal{R}^{(r)}$, which implies that there exists queues i and j such that $Q_i(t) < r$ and $Q_j(t) > r$. This means that the number of IDs in memory denoted by $|m(t)|$ is between 1 and $N-1$. Suppose $|m(t)| = M \in \{1, 2, \ldots, N-1\}$, then we have

$$\Delta_n(t) > 0, n < k \text{ and } \Delta_n(t) < 0, n \geq k,$$

where $k = M + 1$. This is because for $n < k$

$$\Delta_n(t) \overset{(a)}{=} \frac{\mu_{\sigma_t(n)}}{\sum_{i=1}^{M} \mu_{\sigma_t(i)}} - \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma} \overset{(b)}{>} 0,$$

and for $n \geq k$

$$\Delta_n(t) \overset{(c)}{=} 0 - \frac{\mu_{\sigma_t(n)}}{\mu_\Sigma} < 0,$$

where (a) and (c) follow from the definition of $\Delta(t)$ in Eq. (38) and the JBT policy; (b) holds due to $\mu_\Sigma = \sum_{i=1}^{N} \mu_{\sigma_t(i)}$ and $M < N$. Moreover, with simple calculations, we get

$$\min\left(|\Delta_1(t)|, |\Delta_N(t)|\right) \geq \frac{\mu_{min}\mu_{min,2}}{\mu_\Sigma(\mu_\Sigma - \mu_{min})},$$

where $\mu_{min} = \min_{n \in \mathcal{N}} \mu_n$, i.e., the smallest service rate among all servers. $\mu_{min,2}$ is the second smallest service rate among all the servers. Hence, we complete the proof of Claim 4. □