Degree of Queue Imbalance: Overcoming the Limitation of Heavy-traffic Delay Optimality in Load Balancing Systems

XINGYU ZHOU*, The Ohio State University, USA
FEI WU*, The Ohio State University, USA
JIAN TAN, The Ohio State University, USA
KANNAN SRINIVASAN, The Ohio State University, USA
NESS SHROFF, The Ohio State University, USA

Heavy-traffic delay optimality is considered to be an important metric in evaluating the delay performance of load balancing schemes. In this paper, we argue that heavy-traffic delay optimality is a coarse metric that does not necessarily imply good delay performance. Specifically, we show that any load balancing scheme is heavy-traffic delay optimal as long as it satisfies a fairly weak condition. This condition only requires that in the long-term the dispatcher favors, even slightly, shorter queues over longer queues. Hence, although a load balancing scheme could be heavy-traffic delay optimal, the empirical delay performance of heavy-traffic delay optimal schemes can range from very good (that of join-shortest-queue) to very bad (arbitrarily close to the performance of random routing). To overcome this limitation, we introduce a new metric called *degree of queue imbalance*, which measures the queue length difference between all the servers in steady-state. Given a heavy-traffic delay optimal load balancing scheme, we can characterize the resultant *degree of queue imbalance*. This, in turn, allows us to explicitly differentiate between good and poor load balancing schemes. Thus, this paper suggests that when designing good load balancing schemes, they should not only be heavy-traffic delay optimal, but also have a low degree of queue imbalance.

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

Additional Key Words and Phrases: Heavy-traffic delay optimality; Limitation; New metric

ACM Reference Format:
Xingyu Zhou, Fei Wu, Jian Tan, Kannan Srinivasan, and Ness Shroff. 2018. Degree of Queue Imbalance: Overcoming the Limitation of Heavy-traffic Delay Optimality in Load Balancing Systems. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 1, Article 21 (March 2018), 41 pages. https://doi.org/10.1145/3179424

21

*Co-primary author.

This work has been supported in part by ONR grant N00014-17-1-2417 and NSF CNS-1719371, CNS-1717060.
Authors' addresses: Xingyu Zhou, zhou.2055@osu.edu, The Ohio State University, Department of ECE, 2015 Neil Ave. Columbus, OH, 43210, USA; Fei Wu, wu.1973@osu.edu, The Ohio State University, Department of CSE, 2015 Neil Ave. Columbus, OH, 43210, USA; Jian Tan, tan.252@osu.edu, The Ohio State University, Department of ECE, 2015 Neil Ave. Columbus, OH, 43210, USA; Kannan Srinivasan, kannan@cse.ohio-state.edu, The Ohio State University, Department of CSE, 2015 Neil Ave. Columbus, OH, 43210, USA; Ness Shroff, shroff.11@osu.edu, The Ohio State University, Departments of ECE and CSE, 2015 Neil Ave. Columbus, OH, 43210, USA.

2476-1249/2018/3-ART21 $15.00
https://doi.org/10.1145/3179424

1 INTRODUCTION

Load balancing is a common approach to task assignment in distributed architectures such as Web service [9], large data stores (e.g., HBase [8]), and cloud computing [6], etc. In such designs there is a dispatcher that seeks to balance the assignment of jobs across the servers in the system so that the queueing delay is minimized.

To provide delay performance guarantees and to design effective load balancing schemes, it is imperative to develop analytical tools to evaluate the system performance under different load balancing schemes. To that end, one important line of research has focused on the *heavy-traffic* regime when the load ρ approaches 1. For instance, it has been shown that the well-known load balancing policies join-shortest-queue (JSQ) and power-of-d (for any $d \geq 2$) are delay optimal in the heavy-traffic sense [4] [13].

Despite both JSQ and power-of-d being heavy-traffic delay optimal, it has been observed empirically that the delay of power-of-d (for small d) is non-negligibly larger than that of JSQ. This raises two interesting questions: (1) How large can the difference in the resultant delay be when using different heavy-traffic delay optimal load balancing schemes? (2) Can we characterize this difference, and explicitly differentiate between good and poor load balancing schemes that may all be heavy-traffic delay optimal?

In this paper, we take a systematic approach in answering the above questions. To this end, the main contributions of this paper are summarized as follows:

- We show that heavy traffic delay optimality is a coarse metric and does not necessarily imply good delay performance. Specifically, we show that any load balancing scheme is heavy-traffic delay optimal as long as it satisfies a fairly weak condition. This condition only requires that in the long-term (instead of each time-slot), the dispatcher favors, even slightly, shorter queues over longer queues. As a result, empirical delay performance of heavy-traffic delay optimal schemes can range from very good (that of join-shortest-queue) to very bad (arbitrarily close to the performance of random routing and far worse than that of power-of-d).
- To overcome the fundamental limitation of heavy-traffic delay optimality, we introduce a new metric called *degree of queue imbalance*, which measures the queue length difference between all the servers in steady-state. Given a heavy traffic delay optimal load balancing scheme, we can characterize the resultant *degree of queue imbalance*. Specifically, we first capture the essence of a load balancing scheme by the notion of *degree of dispatching preference*, which measures the extent to which the dispatcher favors shorter queues over long queues in the long term. We show that the new metric i.e., *degree of queue imbalance*, is inversely proportional to the square of the *degree of dispatching preference* of the load balancing scheme. This also allows us to derive a non-asymptotic upper bound on the average delay under a given heavy-traffic delay optimal scheme. In this way, via the *degree of queue imbalance*, we reveal how the load balancing scheme would affect the system performance. This, in turn, enables us to explicitly differentiate between good and poor heavy traffic optimal load balancing schemes.
- We further investigate the degree of dispatching preference in the large system regime under power-of-d policy. While this is different from the heavy-traffic regime, our earlier results relating the degree of dispatching preference to the queueing performance motivate this study. In particular, for power-of-d, we derive the limiting value and convergence rate of the degree of dispatching preference under different growth rates of d when the number of servers goes to infinity.

1.1 Related work

Characterizing the exact delay performance of load balancing systems is known to be very difficult in general. Hence, most of the works in the literature have focused on the asymptotic analysis.

Heavy-traffic analysis has been an important tool for characterizing the performance of queueing systems, e.g., [5] [17] [2] [20] [22]. For load balancing systems, it has been shown that, under two homogeneous servers, JSQ has the same behavior as that of a standard $M/M/2$ system in heavy traffic limit via diffusion approximation [5] [17]. Recently, based on the Lyapunov drift condition, an alternative method [4] has been proposed to prove heavy-traffic delay optimality of routing and scheduling policies. It has been used to show heavy-traffic delay optimality of several specific policies for distributed computing systems. For instance, using the Lyapunov drift-based approach, the power-of-d policy proposed in [14] has been shown to be heavy-traffic delay optimal for any $d \geq 2$ [13]. Under power-of-d, the dispatcher probes d servers uniformly at random and dispatches new arrivals to the server with the shortest queue among the d servers. Moreover, it has been shown [13] that a joint JSQ and MaxWeight policy is heavy-traffic delay optimal when jobs are preemptive for homogeneous servers. This result was extended to MapReduce clusters for a specific traffic scenario in [21]. For all traffic scenarios, a heavy-traffic delay optimal policy called 'local-task-first' policy is proposed in [23] under two-level data locality. To address multi-level data locality scenario, a heavy-traffic delay optimal policy was proposed in [24].

In addition, a pull-based policy has been recently proposed and investigated [11] [7]. It has superior delay performance over power-of-d under medium traffic loads. Nevertheless, its performance degrades under high traffic loads [15]. As a result, it is not heavy-traffic delay optimal for general load balancing systems [19]. We finally remark that the heavy-traffic regime considered in this and other papers in the literature [4, 13, 21, 23, 24] is different from the Halfin-Whitt heavy-traffic regime, in which the load ρ approaches one and the number of servers goes to infinity at the same time.

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 $\|\mathbf{x}\| \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Let $\mathbf{1}_N \triangleq \frac{1}{\sqrt{N}}(1, 1, \ldots, 1)$. Then the parallel and perpendicular component of any vector $\mathbf{x}$ in $\mathbb{R}^N$ with respect to the vector $\mathbf{1}_N$ is denoted by $\mathbf{x}_{\|} \triangleq \langle \mathbf{1}_N, \mathbf{x} \rangle \mathbf{1}_N$ and $\mathbf{x}_{\perp} \triangleq \mathbf{x} - \mathbf{x}_{\|}$, respectively.

2 SYSTEM MODEL AND PRELIMINARIES

We consider a discrete-time model for a load balancing system that has one central dispatcher and N parallel servers, indexed by $1, 2, \ldots, N$. Each server n has a FIFO (first-in, first-out) queue denoted by Q_n. We use $\mathbf{Q}(t) = (Q_1(t), Q_2(t), \ldots, Q_N(t))$ to denote the queue lengths at the beginning of time-slot t. Same as [4, 13, 21, 23, 24], the tasks arrived at the beginning each time-slot are immediately dispatched to one of the servers. Once a task joins a queue, it will remain in that queue until its service is completed.

2.1 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 as λ_{Σ} and σ_{Σ}^2, respectively. We further assume that there is a positive probability for $A_{\Sigma}(t)$ to be zero and the arrival process has a finite support, i.e., $A_{\Sigma}(t) \leq A_{\max} < \infty$ for all t.

Let $S_n(t)$ denote the amount of service that server n offers for queue n in time-slot t. 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, and has a finite support, i.e., $S_n(t) \leq S_{\max} < \infty$ for all t and n. The mean and variance of $S_n(t)$ are denoted as μ_n and ν_n^2, respectively. Let $\mu_\Sigma \triangleq \Sigma_{n=1}^N \mu_n$ and $\nu_\Sigma^2 \triangleq \Sigma_{n=1}^N \nu_n^2$.

2.2 Load Balancing Schemes

At the beginning of time-slot t, the dispatcher routes the newly arrived tasks according to some load balancing policy $\eta(t)$, which is a rule that selects the queue to which a new arrival in time-slot t should be dispatched. We allow a load balancing scheme to take different policies, such as JSQ, Power-of-d and random routing, in different time-slots. That is, $\eta(t)$ may evolve over time. Specifically, a load balancing scheme is modeled by a Markov chain as follows.

Definition 2.1 (Load Balancing Scheme). A *load balancing scheme* is a Markov chain $\{\eta(t), t \geq 0\}$, which is composed of the load balancing policy $\eta(t)$ adopted in each time-slot t with a pre-defined transition matrix.

In this paper, we assume that for a load balancing scheme, the corresponding policy Markov chain $\{\eta(t), t \geq 0\}$ is ergodic and has a finite number of states, i.e., the dispatcher may switch between a finite number of load balancing policies. A load balancing scheme is said to be feasible, if it leads to a stable queueing system for any $\lambda_\Sigma < \mu_\Sigma$.

Let $A_n(t)$ denote the number of tasks routed to queue n at the beginning of time-slot t. According to the above definitions, $A_n(t)$ depends on both $\mathbf{Q}(t)$ and $\eta(t)$.

2.3 Queueing Dynamics

In each time-slot, three events take place in order as follows. First, tasks arrive at the beginning of time-slot t. Then, based on the policy $\eta(t)$ and the queue length information (maybe partial) about $\mathbf{Q}(t)$, the dispatcher decides $A_n(t)$ and routes the newly arrived tasks to the servers. Last, the tasks at the queues are processed by the corresponding servers. Therefore, the queue length at each server n, satisfies the following dynamics for $n = 1, 2, \ldots, N$.

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

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

2.4 Heavy-traffic Delay Optimality

Note that in this paper the queueing system can be described by a Markov chain $\{Z(t) = (\mathbf{Q}(t), \eta(t)), t \geq 0\}$. We consider a system $\{Z^{(\epsilon)}(t), t \geq 0\}$ 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 ϵ characterizes the distance between the arrival rate and the capacity region boundary. In heavy-traffic analysis, one is interested in the steady-state queue lengths values as ϵ approaches zero. To introduce heavy-traffic delay optimality, we first restate the general lower bound given in [4].

Lemma 2.2. *Given any feasible load balancing scheme, let $\overline{\mathbf{Q}}^{(\epsilon)}$ be a random vector which is equal in distribution to the queue length $\mathbf{Q}(t)$ in the steady state. Assume $(\sigma_\Sigma^{(\epsilon)})^2$ converges to a constant σ_Σ^2 as ϵ decreases to zero, then*

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

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

This result can be proved by constructing a hypothetical single-server queueing 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*. It is easy to show that for any feasible load balancing scheme the sum queue-length process $\{\sum_{n=1}^N \overline{Q}_n^{(\epsilon)}(t), t \geq 0\}$ is stochastically larger than the resource-pooled system.

Motivated by the universal lower bound above, as in [4, 13, 21, 23, 24], *heavy-traffic delay optimality* of a load balancing scheme is defined as follows.

Definition 2.3. A load balancing scheme is said to be heavy-traffic delay optimal 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 2.2.

3 PREVIEW OF THE MAIN RESULTS

In this section, let us first have a preview of the key concept and main results in this paper. Some statements are informal and the rigorous versions are presented in subsequent sections.

3.1 Heavy-traffic Optimality is a Coarse Metric

Given a policy $\eta(t)$ at time-slot t, the corresponding *dispatching distribution* is denoted by a vector $\mathbf{P}_{\eta(t)}(t)$, where its n-th component corresponds to the probability that the new arrivals are dispatched to the n-th shortest queue at time-slot t. Based on this, we introduce the following key concept.

Dispatching Preference: The dispatching preference for a given policy $\eta(t)$ is given by

$$\Delta_{\eta(t)}(t) \triangleq \mathbf{P}_{\eta(t)}(t) - \mathbf{P}_{\text{rand}},$$

where $\mathbf{P}_{\text{rand}}$ corresponds to the dispatching distribution of random routing.

For example, under JSQ, since the new arrivals are always dispatched to the shortest queue, we have $\mathbf{P}_{\text{JSQ}}(t) = (1, 0, \cdots, 0)$. Under random routing, since the arrivals are dispatched to each server with equal probabilities, it follows that $\mathbf{P}_{\text{rand}}(t) = \left(\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N}\right)$. Thus, by definition, we have $\Delta_{\text{JSQ}}(t) = \left(1 - \frac{1}{N}, -\frac{1}{N}, \cdots, -\frac{1}{N}\right)$.

By Definition 2.1, we allow the load balancing scheme to adopt different policies in different time-slots, and the policies can be time-correlated. Nevertheless, we show that heavy-traffic delay optimality only requires the load balancing scheme to satisfy a fairly weak condition in the long term.

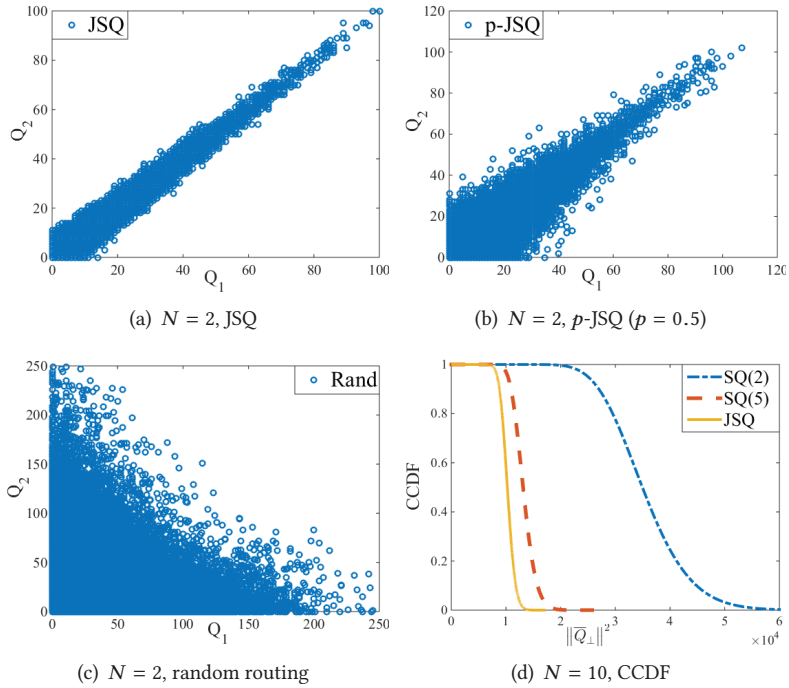
Main Result 1: A load balancing scheme is heavy-traffic delay optimal if it satisfies the Long-term Dispatching Preference Condition (LDPC), which is defined as follows.

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N,$$

where $\widetilde{\Delta} \triangleq \mathbb{E}\left[\overline{\Delta}\right]$ and $\overline{\Delta}$ is a random vector which is equal in distribution to $\Delta(t)$ in the steady state.

This result suggests that heavy-traffic delay optimality only requires that in the long-term (instead of each time-slot), the dispatcher favors, even slightly, shorter queues over longer queues. Therefore, *there could be a large difference in the delay between different heavy-traffic delay optimal load balancing schemes.*

The rigorous results and detailed discussions on this part are presented in Section 4.



(a) $N = 2$, JSQ (b) $N = 2$, p-JSQ ($p = 0.5$)

(c) $N = 2$, random routing (d) $N = 10$, CCDF

Fig. 1. Scatter plots of queue length distributions of load balancing schemes. The policy power-of-d is given by SQ(d) in the figures. Under p-JSQ policy, the dispatcher adopts JSQ with probability p at each time-slot, otherwise it uses random routing.

3.2 A Refined Metric: Degree of Queue Imbalance

To overcome the limitation of heavy-traffic delay optimality, we introduce a new metric called degree of queue imbalance, which measures the queue length difference between all the servers in steady-state, i.e., how well the loads are balanced across the servers in the system.

Degree of Queue Imbalance: The degree of queue imbalance in a load balancing system with a steady-state queue length vector $\overline{\mathbf{Q}}$ is given by $\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}\right\|^2\right]$, where $\mathbf{Q}_{\perp} \triangleq \mathbf{Q}(t) - \langle \mathbf{Q}, \mathbf{1}_N \rangle \mathbf{1}_N$.

The intuition behind the new metric is two-fold.

First, from the scatter plots of the steady-state queue length distributions of different load balancing schemes as shown in Fig. 1, it can be observed that different load balancing schemes lead to different degree of queue imbalance even though they may all be heavy-traffic delay optimal, e.g., JSQ, SQ(2) and SQ(5). As expected, the larger the degree of queue imbalance, the worse the empirical delay performance. It can also be seen that the stronger the preference on shorter queues, the lower the degree of queue imbalance.

Furthermore, the degree of queue imbalance is also motivated by the following non-asymptotic upper bound on the average delay. Under any heavy-traffic delay optimal scheme, the average delay is upper bounded by

$$D_{\text{avg}}^{(\epsilon)} \leq \frac{\zeta^{(\epsilon)}}{2\lambda_{\Sigma}^{(\epsilon)}} \cdot \frac{1}{\epsilon} + \frac{M}{\lambda_{\Sigma}^{(\epsilon)}} \cdot \sqrt{\frac{\text{Degree of Queue Imbalance}}{\epsilon}},$$

where $\zeta^{(\epsilon)} \triangleq (\sigma_\Sigma^{(\epsilon)})^2 + \nu_\Sigma^2$, and M is some positive constant.

Hence, both empirical and theoretical results suggest that the lower the degree of queue imbalance that results from the load balancing scheme, the better the delay performance.

Therefore, the key challenge is to understand how the load balancing scheme would impact the degree of queue imbalance. In the following, (i) we first give a characterization of the extent to which the dispatcher favors the shorter queues, and (ii) we then show a fundamental relationship between the degree of dispatching preference and the resultant degree of queue imbalance.

Degree of Dispatching Preference: The degree of dispatching preference for a given load balancing scheme $\{\eta(t), t \geq 0\}$ is given by the l_1 norm of the long-term dispatching preference, i.e., $\left\|\widetilde{\Delta}\right\|_1$.

Main Result 2: For any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|^2\right] = \Theta\left(\frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\right).$$

It is worth noting that the degree of dispatching preference $\left\|\widetilde{\Delta}\right\|_1$ corresponds to the *total variation distance* between the long-term dispatching distribution of the scheme $\widetilde{\mathbf{P}}$ and the dispatching distribution of random routing $\mathbf{P}_{\text{rand}}$, i.e.,

$$\left\|\widetilde{\Delta}\right\|_1 = 2\delta\left(\widetilde{\mathbf{P}}, \mathbf{P}_{\text{rand}}\right),$$

where $\delta(\,.\,)$ denotes the total variation distance between two finite-space distributions. From this, we can see that the definition of dispatching preference is quite intuitive, i.e., the closer to random routing, the smaller the preference on shorter queues.

Based on this result, the degree of queue imbalance can be approximated by the degree of dispatching preference as follows.

$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}.$$

In this way, via the *degree of queue imbalance*, we reveal how the load balancing scheme would affect the system performance. Therefore, *the degree of queue imbalance enables us to explicitly differentiate between good and poor heavy-traffic delay optimal load balancing schemes.*

The rigorous results and detailed discussions on this part are presented in Section 5.

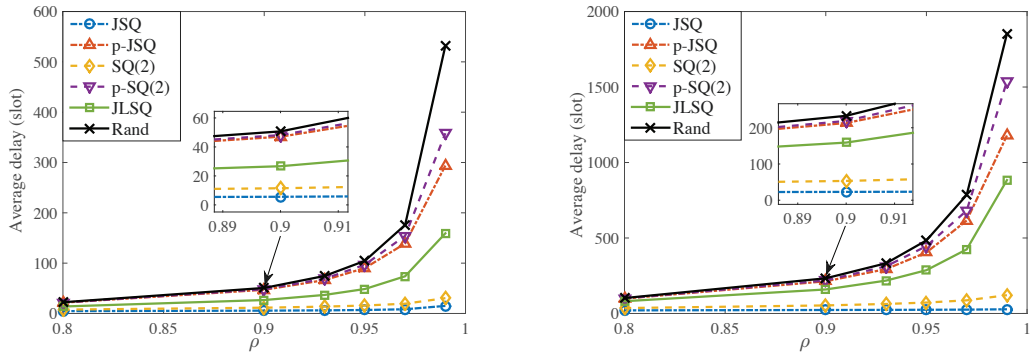
Combining the above two main results, our paper suggests that when designing good load balancing schemes, they should not only be heavy-traffic delay optimal, but also have low degree of queue imbalance.

4 HEAVY-TRAFFIC DELAY OPTIMALITY IS NOT ENOUGH

In this section, we show that, the delay performance of a large class of heavy traffic optimal load balancing schemes can be quite poor in practice. In order to conduct the analysis, we introduce the notion of dispatching preference, based on which we can show the limitations of heavy-traffic delay optimality.

4.1 Dispatching Preference

For a load balancing scheme $\{\eta(t), t \geq 0\}$, given an increasing permutation of queues from the shortest queue to the longest queue, we define the *dispatching distribution* in time-slot t as a vector $\mathbf{P}_{\eta(t)}(t)$, where its n-th component denotes the probability that the new arrivals are dispatched to the n-th shortest queue in time-slot t. When there are ties in the queue lengths, the dispatcher may



(a) $N = 10$ (b) $N = 50$

Fig. 2. Delay performance of load balancing schemes. By Theorem 4.3, all the schemes in the figures except for random routing are heavy-traffic delay optimal. However, their actual delay performance can range from very good (that of join-shortest-queue) to very bad (very close to the performance of random routing).

break ties arbitrarily. By the assumption, $\mathbf{P}_{\eta(t)}(t)$ only depends on the policy $\eta(t)$ in time-slot t, and may vary with $\eta(t)$ over different time-slots. For ease of presentation, we assume that the servers are homogeneous with rate μ. In Section 8, we will discuss how to extend the results to the case with heterogeneous servers.

Based on the dispatching distribution, we can define the notion of dispatching preference as follows.

Definition 4.1 (Dispatching Preference). The *dispatching preference* for a given policy $\eta(t)$ in time-slot t is given by

$$\Delta_{\eta(t)}(t) \triangleq \mathbf{P}_{\eta(t)}(t) - \mathbf{P}_{\text{rand}},$$

where $\mathbf{P}_{\text{rand}}$ corresponds to the dispatching distribution of random routing.

To better understand dispatching distribution and dispatching preference, let us look at a load balancing system with 3 servers. Under random routing, the new arrivals are dispatched to each server with equal probabilities. Thus,

$$\mathbf{P}_{\text{rand}}(t) = \left(\frac{1}{N}, \cdots, \frac{1}{N}\right) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

Under JSQ, the new arrivals are always dispatched to the shortest queue, and thus

$$\mathbf{P}_{\text{JSQ}}(t) = (1, 0, 0)\,.$$

Under power-of-2, the dispatcher randomly picks two servers and dispatches the new arrivals to the server with shorter queue length. It easily follows that

$$\mathbf{P}_{\text{power-of-2}}(t) = \left(\frac{2}{3}, \frac{1}{3}, 0\right).$$

Hence, by definition, we have

$$\Delta_{\text{JSQ}}(t) = \left(\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}\right),$$
$$\Delta_{\text{power-of-2}}(t) = \left(\frac{1}{3}, 0, -\frac{1}{3}\right).$$

Remark 1. For a given load balancing scheme, i.e., the Markov chain $\{\eta(t), t \geq 0\}$, there is a corresponding Markov chain $\{\mathbf{P}(t), t \geq 0\}$ or $\{\Delta(t), t \geq 0\}$ by definition. Moreover, $\{\Delta(t), t \geq 0\}$ has a steady-state vector $\overline{\Delta}$ for any given load balancing scheme with an ergodic and finite-state policy Markov chain $\{\eta(t), t \geq 0\}$, which is the case considered in this paper.

4.2 Limitations of Heavy-traffic Optimality

To see the limitations of heavy-traffic delay optimality, we first introduce the following key condition.

Definition 4.2 (Long-term Dispatching Preference Condition (LDPC)). A Markov chain $\{\eta(t), t \geq 0\}$ is said to satisfy the LDPC condition if the corresponding Markov chain of dispatching preference $\{\Delta(t), t \geq 0\}$ is independent of ϵ and satisfies

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N,$$

where $\widetilde{\Delta} = \mathbb{E}\left[\overline{\Delta}\right]$ and $\overline{\Delta}$ is a random vector which is equal in distribution to $\Delta(t)$ in the steady state.

Then, using the condition above, we are able to establish the following main result.

Theorem 4.3. *Consider a load balancing scheme with $\{\eta(t), t \geq 0\}$ being a finite-space, ergodic Markov chain. If $\{\eta(t), t \geq 0\}$ satisfies the Long-term Dispatching Preference Condition (LDPC), then the load balancing scheme is heavy-traffic delay optimal.*

Proof. See Section 7.1. □

Theorem 4.3 suggests that heavy-traffic delay optimality is limited in the following sense: *For a large class of heavy-traffic delay optimal load balancing schemes, their delay performance can be poor in practice.* To better illustrate this point, in the following we first present two basic properties of LDPC, and then present two specific examples to facilitate the understanding of the limitations of heavy-traffic delay optimality.

The class of load balancing schemes satisfying LDPC is closed under the following two linear operations.

(1) (Linear multiplication) Given a load balancing scheme with $\{\eta_1(t), t \geq 0\}$ satisfying LDPC, the following new load balancing scheme $\{\eta_2(t), t \geq 0\}$ also satisfies LDPC. Fix any $p > 0$, in each time-slot t under the new scheme, the dispatcher adopts $\eta_1(t)$ with probability p and uses random routing otherwise. This property holds since $\widetilde{\Delta}_{\eta_2} = p \cdot \widetilde{\Delta}_{\eta_1}$.
(2) (Linear combination) Given two load balancing schemes with $\{\eta_1(t), t \geq 0\}$ and $\{\eta_2(t), t \geq 0\}$ both satisfying LDPC, the following new load balancing scheme $\{\eta_3(t), t \geq 0\}$ also satisfies LDPC. Fix any $p_1, p_2 > 0$ with $p_1 + p_2 = 1$, in each time-slot t, the dispatcher adopts $\eta_1(t)$ with probability p_1 and adopts $\eta_2(t)$ with probability p_2. This property results from the fact that $\widetilde{\Delta}_{\eta_3} = p_1 \cdot \widetilde{\Delta}_{\eta_1} + p_2 \cdot \widetilde{\Delta}_{\eta_2}$.

Based on the properties above, we present two load balancing schemes that have very poor empirical delay performance whereas they are still heavy-traffic delay optimal.

Example 4.4. (p-JSQ and p-SQ(d)) In each time-slot, the dispatcher adopts JSQ (power-of-d) with probability p and uses random routing otherwise. It is easy to check that JSQ and power-of-d satisfy LDPC. By Property (1), it follows that p-JSQ and p-SQ(d) also satisfy LDPC, and hence are heavy-traffic delay optimal by Theorem 4.3.

Key observation: Despite heavy-traffic delay optimality, since p can be arbitrarily close to zero, one can expect that the delay performance of p-JSQ and p-SQ(d) can be arbitrarily close to that of random routing as p approaches zero. As shown in Fig. 2, when $p = 0.01$, the delay performance of

p-JSQ and p-SQ(d) is much worse than that of JSQ and power-of-d (SQ(d) in figures), and is close to that of random routing, even when the load is as high as 0.99.

Notice that all the heavy-traffic delay optimal schemes discussed so far statistically favor shorter queues in every time-slot. In practice, *due to data locality* and other constraints, it is possible that *the dispatcher has to route new arrivals to longer queues in certain time-slots*. It is important to know *whether a load balancing scheme in this case can be heavy-traffic delay optimal*. Theorem 4.3 suggests that a load balancing scheme is heavy-traffic delay optimal as long as it favors shorter queues in the long-term instead of each time-slot. In the sequel, we present a simple example to demonstrate this idea.

Example 4.5. (Join longer or shorter queue (JLSQ) scheme) In each time-slot, with probability $\hat{p}$ the new arrivals are dispatched to a queue which is chosen uniformly at random from the longest N_1 queues. Otherwise, the new arrivals are routed to a queue which is chosen uniformly at random from the shortest $N - N_1$ queues. It is easy to see that if $\hat{p} < \frac{N_1}{N}$, then the corresponding dispatching preference of JLSQ satisfies LDPC, hence by Theorem 4.3, JLSQ is heavy-traffic delay optimal when $\hat{p} < \frac{N_1}{N}$.

Key observation: In a special case $N_1 = \frac{N}{2}$, in order to achieve heavy-traffic delay optimality, it is fine to join longer queues temporarily, as long as the time devoted to it in the long term is slightly less than the time devoted to joining shorter queues. As a result, despite heavy-traffic delay optimality, the empirical delay performance of JLSQ can be very bad when $\hat{p}$ approaches $\frac{N_1}{N}$.

After observing the limitation of heavy-traffic delay optimality through examples and simulations, we now turn to the key insights behind the proof of Theorem 4.3, which theoretically explain the limitation of heavy-traffic delay optimality.

Key insights in the proof of Theorem 4.3:

(1) From Lemma 2.2, sum of the queue lengths of a load balancing system is lower bounded by its corresponding resource-pooled system. This is because in a load balancing system there exists the situation when one queue is empty with a positive unused service while there are still waiting jobs in other queues. Thus, to achieve the same average delay of the resource-pooled system in the heavy-traffic limit, i.e., heavy-traffic optimality, it is required for the load balancing scheme to guarantee that when one queue is empty with a positive unused service, all the other queues are close to empty, i.e., there is no waste of service. This is actually the intuition behind the sufficient and necessary condition for heavy-traffic delay optimality proved in Proposition 7.4, i.e.,

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0. \tag{2}$$

It is worth pointing out that this condition is fairly weak in the following sense: To be heavy-traffic delay optimal, it is not necessary to balance the queues all the time (e.g., JSQ), rather, it is only required that the queue lengths are all close to zero when one queue becomes zero with a positive unused service. Based on this fact, it is not surprising that a weak condition, e.g., LDPC, is able to guarantee heavy-traffic delay optimality.

(2) The left-hand-side of Eq. (2) can be upper bounded as follows.

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \le N\sqrt{C\epsilon\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}(t)\right\|^2\right]},$$

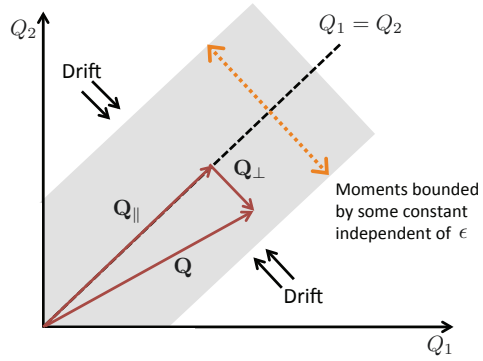


Fig. 3. Insights behind Theorem 4.3.

where C is a constant independent of ϵ. This upper bound suggests that to be heavy-traffic delay optimal, it suffices to guarantee that the second moment of $\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|$ is bounded by any constant that is independent of ϵ. This condition is often called *state-space collapse*.

(3) To show the bounded moment of $\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|$ via Lyapunov-drift condition, the main idea is to show that there is a strictly negative drift independent of ϵ along the direction perpendicular to the line $\mathbf{1}_N \triangleq \frac{1}{\sqrt{N}}(1, 1, \ldots, 1)$ when $\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|$ is large enough. Note that, it is not necessary to have a strictly negative drift for each time-slot. Instead, it only requires that there exists a finite T such that the drift within the T slots is strictly negative. In addition, it is easy to see that under random routing, i.e., $\Delta(t) = 0$ for all t, the drift is zero at each time-slot. Hence, it follows that to be heavy-traffic delay optimal, a load balancing scheme only needs to have a slight preference on shorter queues. This is actually the key insight behind the LDPC, which further demonstrates the limitation of heavy-traffic delay optimality.

Remark 2. Note that in some of the previous works, showing the state-space collapse is also a key step in establishing heavy-traffic delay optimality. However, they suffer from two main limitations: (1) Their analysis is policy by policy. It is not known in general under which condition a load balancing scheme can lead to state-space collapse. (2) Moreover, the previous works only focus on the load balancing schemes consisting of one specific policy, which does not change over time. In contrast, we allow the load balancing scheme to adopt different policies in different time-slots, and the policies can be time-correlated. Thus, our contribution is that for this general class of load balancing schemes, we show a fairly weak condition for heavy-traffic delay optimality: Any load balancing scheme is heavy-traffic delay optimal as long as it satisfies the Long-term Dispatching Preference Condition (LDPC), which suggests that heavy-traffic delay optimality is a coarse metric.

4.3 Numerical Results

Now we use simulations to show the limitation of heavy-traffic delay optimality. In particular, we show that for fixed load ρ, the delay performance of heavy-traffic delay optimal schemes can be very poor. In the simulations, as shown in Fig. 2, we consider the previously discussed heavy-traffic delay optimal schemes, i.e., JSQ, p-JSQ (p=0.01), power-of-2 (SQ(2)), p-SQ(2) (p=0.01), JLSQ ($N_1 = N/2$, $\hat{p} = 0.49$), as well as random routing which is not heavy-traffic delay optimal [5]. In each time-slot t, the exogenous arrival $A_\Sigma(t)$ and service $S_n(t)$ are drawn from a Poisson distribution with rate λ_Σ and $\mu_n = \mu = 1$. We consider two cases with different number of servers, i.e., $N = 10$ and $N = 50$, respectively.

From Fig. 2 (a,b), it can be seen that for both cases $N = 10$ and $N = 50$, the delay performance of heavy-traffic delay optimal schemes can range from very good (that of JSQ) to very bad (quite close to the performance of random routing). Although by the definition of heavy-traffic delay optimality, the delay of all these schemes (except for random routing) should scale at the same order as JSQ when $\rho \to 1$, the average delays of these schemes are very different in the simulation results even when $\rho = 0.99$. For example, when $\rho = 0.99$, the average delay of p-SQ(2) is almost 23 and 54 times larger than that of JSQ for $N = 10$ and $N = 50$, respectively. Moreover, the differences of the delay between these schemes become larger as the number of servers increases.

To sum up, all the heavy-traffic delay optimal schemes have the optimal asymptotic delay performance as $\rho \to 1$. However, from all the discussions above, we have shown that for any fixed load, the actual delay performance of heavy-traffic delay optimal schemes can range from very good to very poor. This motivates the following important question: *How can we explicitly differentiate between good and poor load balancing schemes that may all be heavy-traffic delay optimal?*

5 A REFINED METRIC

In this section, we introduce a new metric called *degree of queue imbalance*, which is able to differentiate between good and poor heavy-traffic delay optimal load balancing schemes.

5.1 Degree of Queue Imbalance

To formally define this new metric, we first denote

$$\mathbf{Q}_{\parallel}(t) \triangleq \langle \mathbf{Q}, \mathbf{1}_N \rangle \mathbf{1}_N,$$

and

$$\mathbf{Q}_{\perp}(t) \triangleq \mathbf{Q}(t) - \mathbf{Q}_{\parallel}(t) = \mathbf{Q}(t) - Q_{\text{avg}}(t)\mathbf{1},$$

in which $\mathbf{1}_N \triangleq \frac{1}{\sqrt{N}}(1, 1, \ldots, 1)$, $\mathbf{1} \triangleq (1, 1, \ldots, 1)$ and $Q_{\text{avg}}(t)$ is the average queue length at time-slot t.

Definition 5.1 (Degree of Queue Imbalance). The degree of queue imbalance for a load balancing system with a steady-state queue length vector $\overline{\mathbf{Q}}$ is given by $\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}\right\|^2\right]$.

From the definition, we can see that the degree of queue imbalance measures the queue length difference between all the servers in steady-state. From the scatter plots of the queue length distribution shown in Fig. 1 (a-c), it can be observed that a good load balancing scheme, such as JSQ, is able to statistically maintain a good balance between the two queues, i.e., the expected queue lengths difference between the servers is kept small. This is achieved by dispatching the new arrivals to the shorter queue, and thus reducing the queue length difference between the servers. On the other hand, a poor load balancing scheme, such as random routing, results in a large queue length difference between the servers, because the dispatcher has no (or little) preference on the shorter queue. This insight remains true when there are more than two servers, as shown in Fig. 1 (d).

Furthermore, the degree of queue imbalance is also motivated by the following non-asymptotic upper bound on the average delay.

Lemma 5.2. *Under any heavy-traffic delay optimal scheme satisfying LDPC, for any $\epsilon > 0$, the average delay is upper bounded by*

$$D_{avg}^{(\epsilon)} \leq \frac{\zeta^{(\epsilon)}}{2\lambda_{\Sigma}^{(\epsilon)}} \cdot \frac{1}{\epsilon} + \frac{M}{\lambda_{\Sigma}^{(\epsilon)}} \cdot \sqrt{\frac{\textit{Degree of Queue Imbalance}}{\epsilon}},$$

where $\zeta^{(\epsilon)} \triangleq (\sigma_{\Sigma}^{(\epsilon)})^2 + v_{\Sigma}^2$, and M is some positive constant.

Proof. See Appendix G. □

Hence, it is suggested that the smaller the queue imbalance generated by a load balancing scheme, the better its delay performance.

Therefore, the key challenge is to understand how the load balancing scheme would impact the degree of queue imbalance. In the following, (i) we first give a characterization of the extent to which the dispatcher favors the shorter queues, and (ii) we then show a fundamental relationship between the degree of dispatching preference and the resultant degree of queue imbalance.

5.2 Differentiating Good and Poor Schemes

Given a load balancing scheme, we capture the extent to which the dispatcher favors shorter queues through the notion *degree of dispatching preference*, which is defined as follows.

Definition 5.3 (Degree of Dispatching Preference). The *degree of dispatching preference* for a given load balancing scheme $\{\eta(t), t \geq 0\}$ is given by the l_1 norm of the long-term dispatching preference, i.e., $\left\|\widetilde{\Delta}\right\|_1$.

It is worth noting that the degree of dispatching preference $\left\|\widetilde{\Delta}\right\|_1$ corresponds to the *total variation distance* between the long-term dispatching distribution of the scheme $\widetilde{\mathbf{P}}$ and the dispatching distribution of random routing $\mathbf{P}_{\text{rand}}$, i.e.,

$$\left\|\widetilde{\Delta}\right\|_1 = 2\delta\left(\widetilde{\mathbf{P}}, \mathbf{P}_{\text{rand}}\right),$$

where $\delta(\,.\,)$ denotes the total variation distance between two finite-space distributions. This means that for a load balancing scheme, the smaller the degree of dispatching preference on the shorter queues, the closer it is to random routing, which is quite intuitive.

In the following, we show a fundamental relationship between the degree of queue imbalance and the degree of dispatching preference.

Theorem 5.4. *Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of*

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] = \Theta\left(\frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\right).$$

This theorem shows that the smaller the dispatching preference, the larger the degree of queue imbalance. Take p-JSQ and p-SQ(d) for example. By the linear multiplication property of LDPC, it can be seen that the degree of queue imbalance for both p-JSQ and p-SQ(d) is on the order of $\Theta\left(\frac{1}{p^2}\right)$ with respect to p. Thus, although they still remain heavy-traffic delay optimal for any $p > 0$, the degree of queue imbalance goes to infinity as p approaches 0.

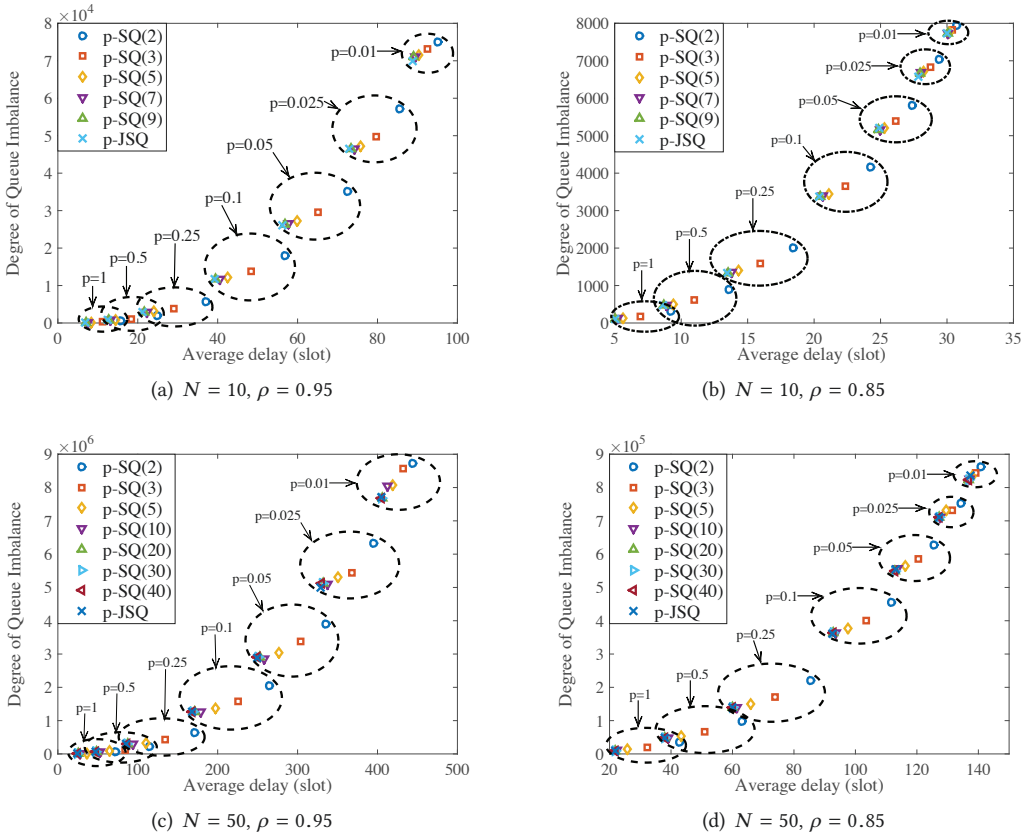
A Delay Upper Bound via Degree of Queue Imbalance:

A non-asymptotic upper bound on the average delay follows directly from Theorem 5.4 and Lemma 5.2.

Corollary 5.5. *Given a load balancing scheme satisfying LDPC, the average delay $D_{avg}^{(\epsilon)}$ for all $\epsilon \leq \epsilon_0$ is upper bounded by*

$$D_{avg}^{(\epsilon)} \leq \frac{\zeta^{(\epsilon)}}{2\lambda_\Sigma^{(\epsilon)}} \cdot \frac{1}{\epsilon} + \frac{M'}{\left\|\widetilde{\Delta}\right\|_1 \cdot \lambda_\Sigma^{(\epsilon)}} \cdot \sqrt{\frac{1}{\epsilon}}, \tag{3}$$

where $\zeta^{(\epsilon)} \triangleq (\sigma_\Sigma^{(\epsilon)})^2 + v_\Sigma^2$, $\epsilon_0 \triangleq \frac{\mu_\Sigma}{2}$, and M' is some positive constant.



(a) $N = 10$, $\rho = 0.95$ (b) $N = 10$, $\rho = 0.85$

(c) $N = 50$, $\rho = 0.95$ (d) $N = 50$, $\rho = 0.85$

Fig. 4. Degree of Queue Imbalance vs. Average delay of load balancing schemes. The parameter p ranges from 0.01 to 1.

The right-hand-side of Eq. (3) is related to a load balancing scheme by its degree of dispatching preference. The larger the degree of dispatching preference of the scheme, the smaller the upper bound of the average delay. Although this result only gives an upper bound of the delay performance, it suggests that the delay performance is poor if the load balancing scheme has a large degree queue imbalance. This insight is also verified by extensive numerical results, as shown in Fig. 4 (a-d) which will be discussed later.

Having shown the main result and its implications, we can now elaborate on the technical details behind it. In particular, Theorem 5.4 is obtained via two steps. First, we derive an upper bound on the degree of queue imbalance as follows.

Proposition 5.6. *Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is upper bounded by*

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \leq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_1,$$

where $M_1 \triangleq 2\max\left(\frac{16N^2(2K+LT)}{T\mu_\Sigma}, 4\sqrt{2}DN\left(1+\frac{16DN}{T\mu_\Sigma}\right)\right)^2$, *in which* $K \triangleq \mu_\Sigma NT^2\max(A_{\max}, S_{\max})$, $L \triangleq N\max(A_{\max}, S_{\max})^2$ *and* $D \triangleq 2T\sqrt{N}\max(A_{\max}, S_{\max})$.

Proof. See Appendix F □

Remark 3. Note that in the previous works [4] [13], it has been shown that under some specific heavy-traffic delay optimal schemes (e.g., JSQ and power-of-*d*), the moment of $\|\mathbf{Q}_\perp\|$ are bounded by some constant independent of ϵ in heavy-traffic. However, it is still unknown: (1) Whether there exists an explicit form of the upper bound? (2) How the upper bound is related to the different load balancing schemes. In a recent work [12], the authors derived an upper bound of the moment of $\|\mathbf{Q}_\perp\|$ for Max-Weight scheduling in a switch system. But it was not clear how a upper bound of the moment of $\|\mathbf{Q}_\perp\|$ can be derived for different load balancing schemes in a load balancing system, which is settled down by the proposition above.

Second, we derive a lower bound on the degree of queue imbalance.

Proposition 5.7. *Under a heavy-traffic delay optimal load balancing scheme such that* $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ *has bounded second moment for all* $\epsilon > 0$, *the degree of queue imbalance is lower bounded by*

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}M_2,$$

where $M_2 = \frac{(N-1)^2\left(\sigma_\Sigma^2+\mu_\Sigma^2+\nu_\Sigma^2\right)^2}{4N^2\mu_\Sigma^2}$, *which is a constant independent of* ϵ.

Proof. See Section 7.2 □

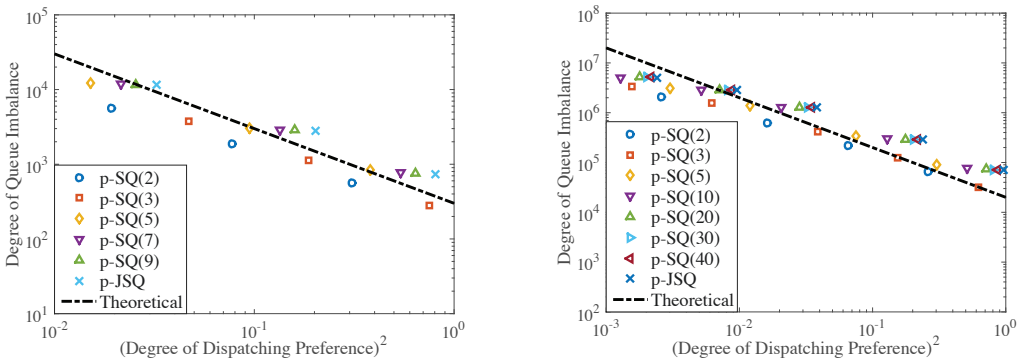
Remark 4. Technically, it is very challenging to derive a meaningful lower bound on the degree of queue imbalance. In particular, the conventional way to bound moments in [1] fails for the following reason. It is required to derive a uniform lower bound of the Lyapunov-drift for all the queue length states. However, due to the boundary constraint that $Q_n(t) \geq 0$ for n and t, some queue length states on the boundary has a much smaller Lyapunov-drift than the other states, which makes the uniform lower bound meaningless. To overcome this difficulty, we develop a novel approach: We obtain a universal equality, i.e., Eq. (5), which characterizes the steady state of the system for *any* heavy-traffic delay optimal load balancing schemes. This enables us to derive the lower bound for any heavy-traffic delay optimal load balancing scheme. The proof sketch is presented as follows.

Proof sketch of Proposition 5.7. We consider the following particular Lyapunov function

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

Under the assumption that the second moment of $\|\mathbf{Q}\|$ is finite, then the mean drift of $V_1(\cdot)$ is zero in steady state, which yields

$$\begin{aligned}
&2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1\left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right]\\
&=\underbrace{2N\mathbb{E}\left[\langle\overline{\mathbf{Q}}_\perp^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)} - \overline{\mathbf{S}}^{(\epsilon)}\rangle\right]}_{\mathcal{T}_1^{(\epsilon)}} - \underbrace{\sum_{i=1}^N\sum_{j>i}^N\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right]}_{\mathcal{T}_2^{(\epsilon)}} + \underbrace{\sum_{i=1}^N\sum_{j>i}^N\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right]}_{\mathcal{T}_3^{(\epsilon)}}
\end{aligned} \tag{4}$$



(a) $N = 10$, $\rho = 0.95$ (b) $N = 50$, $\rho = 0.95$

Fig. 5. Degree of Queue Imbalance vs. Degree of Dispatching Preference.

Note that for any heavy-traffic delay optimal scheme, the left-hand-side (LHS) of Eq. (4) is zero as $\epsilon \to 0$ by the necessary condition in Proposition 7.4. As a result, we obtain the following universal steady-state equality for all heavy-traffic delay optimal schemes.

$$\lim_{\epsilon \downarrow 0} \mathcal{T}_1^{(\epsilon)} = \lim_{\epsilon \downarrow 0} \mathcal{T}_2^{(\epsilon)} - \lim_{\epsilon \downarrow 0} \mathcal{T}_3^{(\epsilon)}. \tag{5}$$

Then, we can characterize each term respectively. To begin with, it is easy to see that $\mathcal{T}_2^{(\epsilon)}$ approaches 0 as $\epsilon \to 0$ since it describes the unused service. Furthermore, we can show that $\mathcal{T}_3^{(\epsilon)}$ converges to some constant K independent of ϵ when $\epsilon \to 0$. For the term $\mathcal{T}_1^{(\epsilon)}$, we can simplify it to

$$\mathcal{T}_1^{\epsilon} = 2\lambda_{\Sigma}^{(\epsilon)} N \mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \widetilde{\Delta} \rangle\right],$$

where the n-th component of the vector $\mathbf{Q}_{\sigma_t}(t)$ is the n-th shortest queue length at time-slot t. Therefore, Eq. (5) can be simplified as

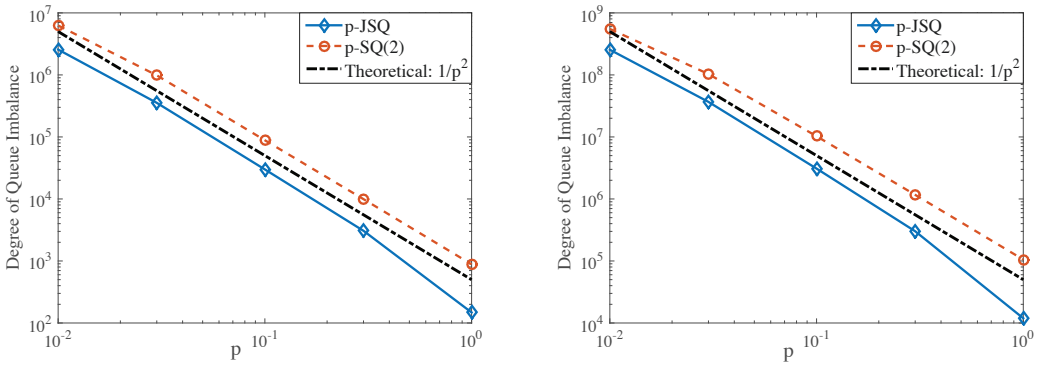
$$\lim_{\epsilon \downarrow 0} 2\mu_{\Sigma} N \mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \widetilde{\Delta} \rangle\right] = -K. \tag{6}$$

Hence, according to Eq. (6), the required result follows directly from Cauchy-Schwartz inequality. □

5.3 Numerical Results

We conduct extensive simulations to demonstrate the effectiveness of the new metric, i.e., degree of queue imbalance, and also verify the theoretical results. The distributions of the arrival and service processes are set to be the same as Section 4. All the data points are collected over 10^7 time-slots.

In Fig. 4 (a-d), the average delay performance is plotted with respect to the degree of queue imbalance under different load balancing schemes, for different loads, and for different number of servers. It can be observed that in all cases, i.e., different loads and different number of servers, the delay performance becomes worse if the load balancing scheme results in a larger degree of queue imbalance. Thus, the numerical results suggest that the degree of queue imbalance is empirically a good metric to differentiate between good and poor load balancing schemes, verifying the insight from Corollary 5.5.



(a) $N = 10$, $\epsilon = 0.001$ (b) $N = 50$, $\epsilon = 0.001$

Fig. 6. Degree of Queue Imbalance vs. p.

To verify Theorem 5.4, in Fig. 5 (a-b), the degree of queue imbalance is plotted with respect to the degree of dispatching preference. From Fig. 5, it can be seen that the degree of queue imbalance increases when the degree of dispatching preference decreases. Specifically, for both cases $N = 10$ and $N = 50$, the growth rate of the degree of queue imbalance with respect to the degree of dispatching preference matches well to the theoretical predictions given by Theorem 5.4. In Fig. 6 (a,b), the degree of queue imbalance under p-JSQ and p-SQ(d) is plotted with respect to p. Recall that for both p-JSQ and p-SQ(d) the degree of queue imbalance scales on the order of $\Theta\left(\frac{1}{p^2}\right)$ by Theorem 5.4. In Fig. 6, the numerical results match well with the theoretical predictions for both cases $N = 10$ and $N = 50$. Hence, our theoretical results are validated by the numerical results. Furthermore, it is very important to note that although Theorem 5.4 is stated in the heavy-traffic regime, from Fig. 5 (a,b) and Fig. 6 (a,b), the results still hold empirically in the non-asymptotic regime, e.g., $\rho < 1$.

6 DEGREE OF DISPATCHING PREFERENCE IN THE LARGE SYSTEM REGIME

In Section 5, we have shown the fundamental relationship between system performance and the degree of dispatching preference in the heavy-traffic regime. In this section, we further study the degree of dispatching preference in the large system regime. In particular, we focus on power-of-d, which is the most well known heavy-traffic delay optimal policy with a low communication overhead. The choice of d has a substantial impact on the delay performance and the message overhead, i.e., a larger d improves the delay performance at the cost of a larger message overhead. Since power-of-d is heavy-traffic optimal for any $d \geq 2$, the existing heavy-traffic analytical tool provides limited insights. In the following, we study this fundamental trade-off between the delay performance and the message overhead via our new analytical framework. Specifically, we study the important question: What is the minimum growth rate of d with respect to the number of servers N to achieve a good degree of dispatching preference, which in turn implies good delay performance and low message overhead?

For ease of presentation, the degree of dispatching preference for power-of-d is denoted by $\left\|\Delta^{(d)}\right\|_1$. We first show the large system limit of $\left\|\Delta^{(d)}\right\|_1$ under different growth rates of d.

Proposition 6.1. *The large system limit for the degree of dispatching preference* $\left\|\Delta^{(d)}\right\|_1$ *under different growth rates of* d *when* $N \to \infty$ *can be characterized by*

(1) For any fixed $d \geq 2$, we have

$$2\frac{d-1}{d}\left(\frac{1}{d^2}\right)^{\frac{1}{d-1}} \leq \lim_{N\to\infty}\left\|\Delta^{(d)}\right\|_1 \leq 2\left(\frac{1}{d}\right)^{1/d}.$$

(2) If $\lim_{N\to\infty} d(N) = \infty$, then we have

$$\lim_{N\to\infty}\left\|\Delta^{(d)}\right\|_1 = \lim_{N\to\infty}\left\|\Delta^{(N)}\right\|_1 = 2$$

Proof. See Appendix H. □

Remark 5. From (1) of Proposition 6.1, it can be seen that for any fixed $d \geq 2$, the degree of dispatching preference will always be strictly larger than zero even when $N \to \infty$. This agrees with the result that power-of-d leads to a substantial delay improvement over random routing. From (2) of Proposition 6.1, we can see that as long as $d(N) \to \infty$, the limiting value of $\left\|\Delta^{(d)}\right\|_1$ converges to that of JSQ. This insight behind this result agrees with the previous result [16], which shows that stochastic optimality of the JSQ policy can be preserved at the fluid-level as long as $d = \omega(1)$.

The above result shows that the growth rate of d does not have any impact on the limiting value of $\left\|\Delta^{(d)}\right\|_1$. However, in the next proposition, we show that the convergence rate of $\left\|\Delta^{(d)}\right\|_1$ depends on the growth rate of d.

Proposition 6.2. *If $\lim_{N\to\infty} d(N) = \infty$ and $d(N) = o(N)$,*

$$\left|2 - \left\|\Delta^{(d)}\right\|_1\right| = \Theta\left(\frac{\log d}{d}\right)$$

Proof. See Appendix I. □

Remark 6. Although Proposition 6.1 suggests that $\left\|\Delta^{(d)}\right\|_1$ converges to 2 as long as $\lim_{N\to\infty} d(N) = \infty$, Proposition 6.2 reveals that the corresponding convergence rate is on the order of $\Theta\left(\frac{\log d}{d}\right)$.

Corollary 6.3. *If $\lim_{N\to\infty} d(N) = \infty$ and $d(N) = o(N)$,*

(1) For some $K > 0$ and $d = K\frac{1}{\epsilon}\log\frac{1}{\epsilon}$, there exists an ϵ^* such that $\left|2 - \left\|\Delta^{(d)}\right\|_1\right| \leq \epsilon$ for all $\epsilon < \epsilon^*$.

(2) Given any $\gamma > 0$, if $d = O\left(\frac{1}{\epsilon}\left(\log\frac{1}{\epsilon}\right)^{(1-\gamma)}\right)$, then there exits an ϵ^* such that $\left|2 - \left\|\Delta^{(d)}\right\|_1\right| > \epsilon$ for all $\epsilon < \epsilon^*$.

Proof. See Appendix J. □

Remark 7. Corollary 6.3 suggests that, to keep the degree of dispatching preference within an ϵ-neighborhood of the optimal value, it is sufficient and necessary to let d grow on the order of $\frac{1}{\epsilon}\log\frac{1}{\epsilon}$.

7 PROOF OF MAIN RESULTS

Let us first define a permutation $\sigma_t(\cdot)$ of $(1, 2, \ldots, N)$ which satisfies $Q_{\sigma_t(1)}(t) \leq Q_{\sigma_t(2)}(t) \leq \ldots \leq Q_{\sigma_t(N)}(t)$, i.e., the queues are sorted according to their queues in a non-decreasing order with ties broken arbitrarily.

7.1 Proof of Theorem 4.3

Before we present the proof of Theorem 4.3, we would first introduce the following useful results.

Lemma 7.1. *Consider a time-slot t_0 and a positive integer T. Then for any t with $t_0 \le t \le t_0 + T$ and n with $1 \le n \le N$,*

$$|Q_{\sigma_t(n)}(t) - Q_{\sigma_{t_0}(n)}(t_0)| \le T \max(A_{\max}, S_{\max}).$$

Proof. See Appendix B.1 □

Lemma 7.2. *For any t_0 and $1 \le n \le N$, we have*

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0)\right] = \widetilde{\Delta}_n$$

Proof. See Appendix B.2 □

Now we are ready to present the proof of Theorem 4.3.

Proof of Proposition 4.3. The proof will be divided into three steps.

Step 1: We will show that under a load balancing scheme satisfying LDPC, the system is stable with bounded moments.

Proposition 7.3. *Under a load balancing scheme satisfying LDPC, the system is stable, i.e., the Markov chain $\{Z^{(\epsilon)}(t), t \ge 0\}$ is positive recurrent for any $\epsilon > 0$. Moreover, all the moments of $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ are bounded for any $\epsilon > 0$.*

Proof. See Appendix C □

Step 2: We will present a sufficient and necessary condition for heavy-traffic delay optimality when the system is stable with a bounded second moment.

Proposition 7.4. *Consider a load balancing system with a load balancing scheme such that the second moment of $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ is bounded for any $\epsilon > 0$, then the load balancing policy is heavy-traffic delay optimal if and only if*

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

Proof. See Appendix D □

Step 3: We will show that under a load balancing scheme satisfying LDPC, Eq. (7) holds, which establishes heavy-traffic delay optimality. This is achieved by showing that all the moments of $\|\mathbf{Q}_\perp\|$ are bounded by a constant independent of ϵ, i.e., *state-space collapse.*

Proposition 7.5. *Under a load balancing satisfying LDPC, the sufficient and necessary condition in Eq. (7) holds.*

Proof. See Appendix E □

The result in Theorem 4.3 follows directly from the three steps.

□

7.2 Proof of Proposition 5.7

Proof of Proposition 5.7. 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 $\mathbf{Q}^+ \triangleq \mathbf{Q}(t+1)$.

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

in which (a) follows from the fact that $Q_n(t+1)U_n(t) = 0$ for all n and $t > 0$ as shown in Lemma D.1; (b) comes from the definition of $\mathbf{Q}_\perp$.

Since $\|\mathbf{Q}\|^2$ has bounded moment in steady state, 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 implies that

$$\begin{aligned}
&2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left((\overline{Q}_i^+)^{(\epsilon)}\overline{U}_j^{(\epsilon)} + (\overline{Q}_j^+)^{(\epsilon)}U_i^{(\epsilon)}\right)\right]\\
=&2N\mathbb{E}\left[\langle\overline{\mathbf{Q}}_\perp^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)} - \overline{\mathbf{S}}^{(\epsilon)}\rangle\right] - \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] + \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right]
\end{aligned} \tag{8}$$

Now we will further simplify each term in Eq. (8). First, the left-hand-side (LHS) of it can be rewritten as follows

$$LHS = 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1\left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \triangleq 2\overline{\mathcal{B}}^{(\epsilon)}, \tag{9}$$

which holds because $Q_n(t+1)U_n(t) = 0$ for all n and $t \geq 0$ as shown in Lemma D.1. Then we turn to simplify each term on the right-hand-side of Eq. (8).

For the first term on the right-hand-side of Eq. (8), we can rewrite it as follows

$$\begin{aligned}\mathcal{T}_1^{(\epsilon)} &\triangleq 2N\mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\perp}^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)} - \overline{\mathbf{S}}^{(\epsilon)}\rangle\right] \\ &\overset{(a)}{=} 2N\mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\perp}^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)}\rangle\right] \\ &\overset{(b)}{=} 2\lambda_{\Sigma}^{(\epsilon)} N\mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \widetilde{\Delta}\rangle\right], \end{aligned} \tag{10}$$

where (a) follows from that the service are independent of queue lengths and are homogeneous; (b) is true since $\overline{\Delta}$ is independent of $\overline{\mathbf{Q}}$ and the nth element of $\overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}$ is $\overline{Q}_{\overline{\sigma}(n)} - \overline{Q}_{\text{avg}}$.

For the second term on the right-hand-side of Eq. (8), we have

$$\mathcal{T}_2^{(\epsilon)} \triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \leq \epsilon(N-1)S_{\max}, \tag{11}$$

which follows from the fact that $0 \leq U_n \leq S_{\max}$ for all n and $\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right] = \epsilon$ as shown in Lemma D.1.

The third term on the right-hand-side of Eq. (8) can be simplified as follows

$$\begin{aligned}\mathcal{T}_3^{(\epsilon)} &\triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \\ &\overset{(a)}{=} \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)}\right)^2 - \left(\overline{S}_i^{(\epsilon)} - \overline{S}_j^{(\epsilon)}\right)^2\right] \\ &\overset{(b)}{=} (N-1)\left(\left(\sigma_{\Sigma}^{(\epsilon)}\right)^2 + \left(\lambda_{\Sigma}^{(\epsilon)}\right)^2 + \nu_{\Sigma}^2\right), \end{aligned} \tag{12}$$

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.

Let $\overline{\mathcal{B}}^{(\epsilon)} \triangleq \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right]$. Now combining Eqs. (8), (10), (11) and (12), we obtain

$$2\overline{\mathcal{B}}^{(\epsilon)} - \mathcal{T}_3^{(\epsilon)} \leq \mathcal{T}_1^{(\epsilon)} \leq 2\overline{\mathcal{B}}^{(\epsilon)} - \mathcal{T}_3^{(\epsilon)} + \epsilon(N-1)S_{\max}. \tag{13}$$

Since we assume heavy-traffic delay optimality and bounded second moment of $\|\mathbf{Q}\|$, it follows from Proposition 7.4 that $\lim_{\epsilon\downarrow 0}\overline{\mathcal{B}}^{(\epsilon)} = 0$. Consequently, from Eq. (13), we have

$$\hat{M} = \lim_{\epsilon\downarrow 0}\mathcal{T}_1^{(\epsilon)} = \lim_{\epsilon\downarrow 0} 2\mu_{\Sigma}N\mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \widetilde{\Delta}\rangle\right],$$

in which $\hat{M} \triangleq -\lim_{\epsilon\downarrow 0}\mathcal{T}_3^{(\epsilon)} = -(N-1)\left(\sigma_{\Sigma}^2 + \mu_{\Sigma}^2 + \nu_{\Sigma}^2\right)$ assuming $\left(\sigma_{\Sigma}^{(\epsilon)}\right)^2$ converges to a constant σ_{Σ}^2. Applying Cauchy-Schwartz inequality to the equality above, yields,

$$\begin{aligned}\hat{M}^2 &= \lim_{\epsilon\downarrow 0}\left(2\mu_{\Sigma}N\mathbb{E}\left[\langle \overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \widetilde{\Delta}\rangle\right]\right)^2 \\ &\leq 4N^2\mu_{\Sigma}^2 \lim_{\epsilon\downarrow 0}\left\|\widetilde{\Delta}\right\|^2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right]. \end{aligned}$$

Therefore, we have

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{\hat{M}^2}{4N^2\mu_{\Sigma}^2\left\|\widetilde{\Delta}\right\|^2} = \frac{1}{\left\|\widetilde{\Delta}\right\|^2} M_2 \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_2,$$

where $M_2 \triangleq \frac{\hat{M}^2}{4N^2\mu_{\Sigma}^2}$. Hence, it completes the proof. □

8 RESULTS FOR THE GENERAL CASE

In this section, we show how to extend our main results in Sections 4 and 5 to the general case when the servers are heterogeneous. Basically, we show that the main insights that have been shown in the homogeneous server case still hold in a weaker sense for the heterogeneous server case, under some additional conditions.

Definition 8.1. The dispatching preference in the general case is a function of $Z(t)$, the nth element of which is given by

$$\Delta_{Z(t),n}(t) \triangleq \mathbf{P}(t) - \frac{\mu_{\sigma_t(n)}}{\mu_{\Sigma}},$$

where the n-th component of $\mathbf{P}(t)$ is again the probability of joining the n-th shortest queue at time-slot t, and $\sigma_t(n)$ represents the server that has the n-th shortest queue at time-slot t.

Based on this definition, we can obtain the following two results for the general case. The first result is the extension of Theorem 4.3 and Proposition 5.6 to the general case.

Proposition 8.2. *Consider a load balancing system under a load balancing scheme such that there is a steady-state vector $\overline{Z}^{(\epsilon)}$ for all $\epsilon > 0$. Let $\widetilde{\Delta}^{(\epsilon)} \triangleq \mathbb{E}\left[\Delta_{\overline{Z}^{(\epsilon)}}\right]$ and suppose that for all $\epsilon > 0$*

$$\widetilde{\Delta}_1^{(\epsilon)} \geq \widetilde{\Delta}_2^{(\epsilon)} \geq \ldots \geq \widetilde{\Delta}_N^{(\epsilon)} \quad \text{and} \quad \widetilde{\Delta}_1^{(\epsilon)} \neq \widetilde{\Delta}_N^{(\epsilon)}.$$

Then the second moment of $\left\|\overline{\mathbf{Q}}_{\perp}\right\|$ in heavy-traffic is bounded, i.e.,

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \leq \frac{1}{\left\|\widetilde{\Delta}^{(\epsilon)}\right\|_1^2} \hat{M}_1,$$

where $\hat{M}_1$ is independent of ϵ. Moreover, if $\left\|\widetilde{\Delta}^{(\epsilon)}\right\|_1^2 = \omega(\epsilon)$, then the load balancing scheme is heavy-traffic delay optimal.

Proof. See Appendix L □

The second result is an extension of Proposition 5.7 to the general case.

Proposition 8.3. *Consider a load balancing system under a heavy-traffic delay optimal load balancing scheme such that there is a steady-state vector $\overline{Z}^{(\epsilon)}$ with a bounded second moment of $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ for all $\epsilon > 0$. Let $\overline{\Delta}^{(\epsilon)} \triangleq \Delta_{\overline{Z}^{(\epsilon)}}$. If the following two conditions hold*

(1) The first moment of $\left\|\mathbf{Q}_{\perp}^{(\epsilon)}\right\|$ is $o(\frac{1}{\epsilon})$.

(2) $\mathbb{E}\left[\langle \mu_{\overline{\sigma},\perp}^{(\epsilon)}, \overline{\mathbf{P}}^{(\epsilon)} \rangle\right]$ converges to a constant that is independent of ϵ as ϵ approaches zero.

Then the second moment of $\|\mathbf{Q}_{\perp}\|$ in heavy-traffic limit can be lower bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\mathbb{E}\left[\left\|\overline{\Delta}^{(\epsilon)}\right\|_1^2\right]} \hat{M}_2.$$

where $\hat{M}_2$ is a constant independent of ϵ.

Proof. See Appendix M □

9 DISCUSSION

This paper motivates the following interesting open questions.

1. How can we generalize the framework in this paper to load balancing systems with multiple dispatchers and data locality?

2. Does heavy-traffic delay optimality suffer from similar limitations for scheduling in distributed systems? If so, can one develop a new metric similar to degree of queue imbalance to overcome it?

3. Is the LDPC condition necessary for heavy-traffic delay optimality? We conjecture it is not, because it is argued heuristically in [10] that a policy with a multi-dimensional state-space collapse (rather than the single-dimensional collapse in this paper) is still heavy-traffic delay optimal, yet it does not satisfy LDPC.

4. How can we derive a tighter characterization of the delay performance via degree of queue imbalance? In this paper, we have only derived an upper bound on the average delay in terms of degree of queue imbalance.

5. In a load balancing system with heterogeneous servers, a perfect balance of the queue lengths may not be good. As shown in [18], another policy called the shortest-expected-delay (SED) policy is able to achieve a lower delay than that of JSQ. Different from JSQ, SED tries to maintain a balance of the queue lengths which are weighted by the corresponding service rate. Also, the authors in [3] showed that it helps to decrease the mean task slowdown by unbalancing the load (e.g., SITA-V policy) when the task size distribution is heavy-tailed. Nevertheless, little has been known on how to design delay optimal load balancing schemes with heterogeneous servers.

10 CONCLUSIONS

In this paper, we show that any load balancing scheme is heavy-traffic delay optimal as long as in the long-term, the dispatcher favors, even slightly, shorter queues over longer queues. Thus, the actual delay performance of a heavy-traffic delay optimal scheme could vary between being very good to being very bad. In contrast, our proposed new metric called *degree of queue imbalance* enables us to explicitly differentiate between good and poor load balancing schemes. Thus, when designing good load balancing schemes, they should not only be heavy-traffic delay optimal, but also have a low degree of queue imbalance.

REFERENCES

[1] Dimitris Bertsimas, David Gamarnik, and John N Tsitsiklis. 2001. Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Annals of Applied Probability* (2001), 1384–1428.

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

[3] Mark E Crovella, Mor Harchol-Balter, and Cristina D Murta. 1998. Task assignment in a distributed system (extended abstract): improving performance by unbalancing load. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 26. ACM, 268–269.

[4] 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.

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

[6] Ian Foster, Yong Zhao, Ioan Raicu, and Shiyong Lu. 2008. Cloud computing and grid computing 360-degree compared. In *2008 Grid Computing Environments Workshop*. Ieee, 1–10.

[7] David Gamarnik, John N Tsitsiklis, and Martin Zubeldia. 2016. Delay, Memory, and Messaging Tradeoffs in Distributed Service Systems. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. ACM, 1–12.

[8] Lars George. 2011. *HBase: the definitive guide.* " O'Reilly Media, Inc.".
[9] Varun Gupta, Mor Harchol Balter, Karl Sigman, and Ward Whitt. 2007. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation* 64, 9 (2007), 1062–1081.
[10] FP Kelly and CN Laws. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing systems* 13, 1 (1993), 47–86.
[11] 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.
[12] Siva Theja Maguluri, R Srikant, et al. 2016. Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stochastic Systems* 6, 1 (2016), 211–250.
[13] 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.
[14] Michael Mitzenmacher. 2001. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems* 12, 10 (2001), 1094–1104.
[15] Michael Mitzenmacher. 2016. Analyzing distributed Join-Idle-Queue: A fluid limit approach. In *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on.* IEEE, 312–318.
[16] Debankur Mukherjee, Sem C Borst, Johan SH van Leeuwaarden, and Philip A Whiting. 2016. Universality of Power-of-d Load Balancing in Many-Server Systems. *arXiv preprint arXiv:1612.00723* (2016).
[17] Martin I Reiman. 1984. Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology.* Springer, 207–240.
[18] Jori Selen, Ivo Adan, Stella Kapodistria, and Johan van Leeuwaarden. 2016. Steady-state analysis of shortest expected delay routing. *Queueing Systems* 84, 3-4 (2016), 309–354.
[19] Alexander L Stolyar. 2015. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80, 4 (2015), 341–361.
[20] Alexander L Stolyar et al. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* 14, 1 (2004), 1–53.
[21] 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.
[22] Ruth J Williams. 1998. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems* 30, 1 (1998), 27–88.
[23] Qiaomin Xie and Yi Lu. 2015. Priority algorithm for near-data scheduling: Throughput and heavy-traffic optimality. In *Computer Communications (INFOCOM), 2015 IEEE Conference on.* IEEE, 963–972.
[24] Qiaomin Xie, Ali Yekkehkhany, and Yi Lu. 2016. Scheduling with multi-level data locality: Throughput and heavy-traffic optimality. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on.* IEEE, 1–9.

A MOMENTS BOUND FROM DRIFT CONDITIONS

In this paper, we will use the Lyapunov drift conditions based method developed in [4] to derive bounded moments in steady state. The following lemma is a T-step version of Lemmas 2 and 3 in [12]. This lemma could be proved by simply replacing the one-step transition probability to T-step transition probability, and hence we omit the proof here.

Lemma A.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 T time-slot drift of V at X as*

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

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

- *(C1) There exists an $\eta > 0$ and a $\kappa < \infty$ such that for any $t_0 = 1, 2, \ldots$ and for all $X \in \mathcal{X}$ with $V(X) \geq \kappa$,*
$$\mathbb{E}\left[\Delta V(X) \mid X(t_0) = X\right] \leq -\eta.$$
- *(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 $\overline{V}$, and all moments of $\overline{V}$ exist and are finite. More specifically, we have for any $r = 1, 2, \ldots$

$$\mathbb{E}\left[V(\overline{X})^r\right] \leq (2\kappa)^r + (4D)^r \left(\frac{D+\eta}{\eta}\right)^r r!. \tag{14}$$

B PROOFS OF LEMMA 7.1 - LEMMA 7.2

B.1 Proof of Lemma 7.1

Proof. Let us consider any two consecutive time-slots t and $t+1$. We claim that $|Q_{\sigma_{t+1}(n)}(t+1) - Q_{\sigma_t(n)}(t)| \leq M$ holds for any $1 \leq n \leq N$, where $M = \max(A_{\max}, S_{\max})$. We prove this claim by contradiction.

First, by the boundedness of arrival and service, we can easily see that $Q_{\sigma_{t+1}(N)}(t+1) - Q_{\sigma_t(N)}(t) \leq M$. Assume that there exists a k with $1 \leq k < N-1$ such that $Q_{\sigma_{t+1}(k)}(t+1) - Q_{\sigma_t(k)}(t) > M$. Then it directly implies that $(\sigma_{t+1}(1), \ldots, \sigma_{t+1}(k))$ cannot be a permutation of $(\sigma_t(1), \ldots, \sigma_t(k))$. This is true since for any n with $1 \leq n \leq k$, we have

$$Q_{\sigma_t(n)}(t+1) \leq Q_{\sigma_t(n)}(t) + M \leq Q_{\sigma_t(k)}(t) + M, \tag{15}$$

which contradicts the assumption. Hence there is at least one element of $(\sigma_t(1), \ldots, \sigma_t(k))$ is in $(\sigma_{t+1}(k+1), \ldots, \sigma_{t+1}(N))$. However, this cannot hold due to Eq. (15) and the fact that for any n with $k+1 \leq n \leq N$

$$Q_{\sigma_{t+1}(n)}(t+1) \geq Q_{\sigma_{t+1}(k)}(t+1) > Q_{\sigma_t(k)}(t) + M$$

Therefore, we have for any k with $1 \leq k \leq N$, $Q_{\sigma_{t+1}(k)}(t+1) - Q_{\sigma_t(k)}(t) \leq M$.

Similarly, we can easily see that $Q_{\sigma_{t+1}(1)}(t+1) - Q_{\sigma_t(1)}(t) \geq -M$. Assume that there exists a k with $1 < k \leq N$ such that $Q_{\sigma_{t+1}(k)}(t+1) - Q_{\sigma_t(k)}(t) < -M$. Then it implies that $(\sigma_{t+1}(k), \ldots, \sigma_{t+1}(N))$ cannot be a permutation of $(\sigma_t(k), \ldots, \sigma_t(N))$. This holds since for any $k \leq n \leq N$

$$Q_{\sigma_t(n)}(t+1) \geq Q_{\sigma_t(n)}(t) - M \geq Q_{\sigma_t(k)}(t) - M, \tag{16}$$

which contradicts the assumption. Hence there is at least one element of $(\sigma_t(k), \ldots, \sigma_t(N))$ is in $(\sigma_{t+1}(1), \ldots, \sigma_{t+1}(k-1))$. However, this cannot hold due to Eq. (16) and the fact that for any n with $1 \leq n \leq k-1$

$$Q_{\sigma_{t+1}(n)}(t+1) \leq Q_{\sigma_{t+1}(k)}(t+1) < Q_{\sigma_t(k)}(t) - M$$

Therefore, we have for any k with $1 \leq k \leq N$, $Q_{\sigma_{t+1}(k)}(t+1) - Q_{\sigma_t(k)}(t) \geq -M$. Hence, we have proved the claim, which directly implies the conclusion of this lemma. □

B.2 Proof of Lemma 7.2

Proof. Let $f_T \triangleq \frac{1}{T}\left(\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0)\right)$, then we have

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0)\right]$$

$$= \lim_{T\to\infty} \mathbb{E}\left[f_T\right] \overset{(a)}{=} \widetilde{\Delta}_n$$

where (a) is the result of dominated convergence theorem since $f_T \to \widetilde{\Delta}_n$ almost surely by the ergodicity of the Markov chain $\{\eta(t), t \geq 0\}$ (independent of $\mathbf{Q}(t)$) and $|f_T| \leq 1$ as $|\Delta_n(t)| \leq 1$. □

C PROOF OF PROPOSITION 7.3

Before we prove Proposition 7.3, we first introduce the following lemma.

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

$$\|\mathbf{Q}(t+1)\|^2 - \|\mathbf{Q}(t)\|^2 \leq 2\langle \mathbf{Q}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L \tag{17}$$

where L is a finite constant.

Proof. Consider the left-hand-side (LHS) of Eq. (17).

$$\begin{aligned} LHS &= \|\mathbf{Q}(t) + \mathbf{A}(t) - \mathbf{S}(t) + \mathbf{U}(t)\|^2 - \|\mathbf{Q}(t)\|^2 \\ &\overset{(a)}{\leq} \|\mathbf{Q}(t) + \mathbf{A}(t) - \mathbf{S}(t)\|^2 - \|\mathbf{Q}(t)\|^2 \\ &= 2\langle \mathbf{Q}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + \|\mathbf{A}(t) - \mathbf{S}(t)\|^2 \\ &\overset{(b)}{\leq} 2\langle \mathbf{Q}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L \end{aligned}$$

where inequality (a) holds as $[\max(a, 0)]^2 \leq a^2$ for any $a \in \mathbb{R}$; in inequality (b), $L \triangleq N \max(A_{\max}, S_{\max})^2$ holds due to the assumptions that $A_\Sigma(t) \leq A_{\max}$ and $S_n(t) \leq S_{\max}$ for all $t \geq 0$ and all $1 \leq n \leq N$, and the fact that they are both independent of the queue lengths. □

Now we are ready to present the proof of Proposition 7.3.

Proof of Proposition 7.3. We first show that the Markov chain $\{Z(t) = (\mathbf{Q}(t), \eta(t)), t \geq 0\}$ is irreducible and aperiodic. Let $\mathcal{M}$ denote the finite set for the policy spaces of the irreducible and aperiodic chain $\{\eta(t), t \geq 0\}$. Let $Z(0) = (\mathbf{Q}(0), \eta(0)) = (0_{N\times 1}, \eta_0)$ for some $\eta_0 \in \mathcal{M}$, and the state space $\mathcal{S} \subset \mathbb{N}^N \times \mathcal{M}$ consists of all the states that can be reached from the initial state, where $\mathbb{N}$ is the set of nonnegative integers. Then the chain is irreducible because for any state Z in the state space, the Markov chain is capable of reaching the initial state within a finite step if there are no exogenous arrivals and all the service is at least one during each time-slot, which can happen with a positive probability under our assumptions. The chain is also aperiodic since the transition probability from the initial state to itself is positive.

To prove positive recurrence, we will adopt the Foster-Lyapunov theorem. Specifically, it is sufficient to find a Lyapunov function and a positive constant T such that the expected drift in T time-slots is bounded within a finite subset of the state space and negative outside this subset.

To that end, it is advantageous to work with quadratic Lyapunov function $W(Z) \triangleq \|\mathbf{Q}\|^2$ in this case since $\{\eta(t), t \geq 0\}$ is a finite Markov chain. Then for any time-slot t_0, the T time-slot

conditional mean drift of $W(Z)$ is given by

$$\begin{aligned}
&\mathbb{E}\left[W(Z(t_0+T)) - W(Z(t_0)) \mid Z(t_0)\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t_0+T)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0)\right] \\
=&\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} (\|\mathbf{Q}(t+1)\|^2 - \|\mathbf{Q}(t)\|^2) \mid Z(t_0)\right] \\
\overset{(a)}{\leq}&\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} 2\langle \mathbf{Q}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L \mid Z(t_0)\right] \\
\overset{(b)}{=}&2\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{Q}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t)\right] \mid Z(t_0)\right] + LT \\
\overset{(c)}{=}&2\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\sum_{n=1}^{N} Q_{\sigma_n(t)}(t)\left(\Delta_n(t)\lambda_\Sigma - \epsilon\frac{1}{N}\right) \mid Z(t_0)\right] + LT \\
\overset{(d)}{\leq}&2\mathbb{E}\left[\sum_{n=1}^{N} Q_{\sigma_n(t_0)}(t_0)\sum_{t=t_0}^{t_0+T-1}\left(\Delta_n(t)\lambda_\Sigma - \epsilon\frac{1}{N}\right) \mid Z(t_0)\right] + K_1
\end{aligned} \tag{18}$$

where (a) comes from Lemma C.1; (b) follows from the tower property of conditional expectation and the fact that $\mathbf{Q}(t)$, $\mathbf{A}(t)$ and $\mathbf{S}(t)$ are conditionally independent of $Z(t_0)$ when given $Z(t)$; (c) follows from the assumption that the servers are homogeneous with rate μ; (d) follows from Lemma 7.1 and $K_1 \triangleq LT + 2\mu NT^2 \max(A_{\max}, S_{\max})(N+1)$.

Now from Lemma 7.2, we can conclude that for any $\epsilon_1 > 0$, there exists a T_n for each n such that for any $T > T_n$

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0)\right] \leq \widetilde{\Delta}_n + \epsilon_1.$$

Hence, choose $\epsilon_1 < \frac{\epsilon}{N^2\mu}$ and for $T > \max(T_1, T_2, \ldots, T_N)$, Eq. (18) can be further upper bounded as follows

$$\begin{aligned}
&\mathbb{E}\left[W(Z(t_0+T)) - W(Z(t_0)) \mid Z(t_0)\right] \\
\leq&2T\left(\lambda_\Sigma \sum_{n=1}^{N} Q_{\sigma_n(t_0)}(t_0)\widetilde{\Delta}_n + \lambda_\Sigma\epsilon_1 \|\mathbf{Q}(t_0)\|_1 - \frac{\epsilon}{N}\|\mathbf{Q}(t_0)\|_1\right) + K_1 \\
\overset{(a)}{\leq}&2T\|\mathbf{Q}(t_0)\|_1\left(\lambda_\Sigma\epsilon_1 - \frac{\epsilon}{N}\right) + K_1 \\
\overset{(b)}{=}&-2\theta T\|\mathbf{Q}(t_0)\|_1 + K_1 \\
\overset{(c)}{\leq}&-2\theta T\|\mathbf{Q}(t_0)\| + K_1
\end{aligned} \tag{19}$$

where (a) comes from $\sum_{n=1}^{N} Q_{\sigma_{t_0}(n)}(t_0)\widetilde{\Delta}_n \leq 0$. This holds due to the monotonicity of $Q_{\sigma_n(t_0)}(t_0)$ and $\widetilde{\Delta}_n$, and the fact that $\sum_{n=1}^{N}\widetilde{\Delta}_n = 0$; (b) follows from $\theta \triangleq \frac{\epsilon}{N} - \lambda_\Sigma\epsilon_1$, which is positive by our choice of ϵ_1; (c) holds since $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^N$. Therefore, by the Foster-Lyapunov theorem, we conclude that the Markov chain $\{Z(t), t \geq 0\}$ is positive recurrent, and hence ergodic.

Having established the ergodicity, we can now turn to apply Lemma A.1 to show that all the moments of $\|\mathbf{Q}\|$ are upper bounded. In particular, let us consider the Lyapunov function $V(Z) \triangleq \|\mathbf{Q}\|$, and try to show that it satisfies the two conditions (C1) and (C2) in Lemma A.1.

For condition (C1), we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta V(Z) \mid Z(t_0) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t_0+T)\| - \|\mathbf{Q}(t_0)\| \mid Z(t_0) = Z\right] \\
=&\mathbb{E}\left[\sqrt{\|\mathbf{Q}(t_0+T)\|^2} - \sqrt{\|\mathbf{Q}(t_0)\|^2} \mid Z(t_0) = Z\right] \\
\overset{(a)}{\leq}&\frac{1}{2\|\mathbf{Q}(t_0)\|}\mathbb{E}\left[\|\mathbf{Q}(t_0+T)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0) = Z\right] \\
\overset{(b)}{\leq}&-T\theta + \frac{K_1}{2\|\mathbf{Q}(t_0)\|}
\end{aligned}$$

where (a) follows from the fact that $f(x) = \sqrt{x}$ is concave; (b) comes from the upper bound in Eq. (19). Hence, (C1) in Lemma A.1 is verified.

For Condition (C2), we have

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

where (a) follows from the fact that $|\,\|\mathbf{x}\| - \|\mathbf{y}\|\,| \leq \|\mathbf{x}-\mathbf{y}\|$ holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$; (b) holds due to the assumptions that the $A_\Sigma(t) \leq A_{\max}$ and $S_n(t) \leq S_{\max}$ for all $t \geq 0$ and all $1 \leq n \leq N$, and independent of the queue length. This verifies Condition (C2) and hence completes the proof of Proposition 7.3. □

D PROOF OF PROPOSITION 7.4

Before we present the proof of Proposition 7.4, we first introduce the following useful result on unused service in load balancing system.

Lemma D.1. *For any $\epsilon > 0$ and $t \geq 0$, we have*

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

Moreover, if the system has a finite first moment, then we have for some constants c_1 and c_2

$$\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2\right] \leq c_1\epsilon \text{ and } \quad \mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|^2\right] \leq c_2\epsilon$$

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

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

Then, due to the fact that $U_n(t) \leq S_{\max}$ for all $1 \leq n \leq N$ and $t \geq 0$, we have $\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|^2 \leq S_{\max}\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1$, which implies that $c_2 = S_{\max}$. Note that $\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2 \leq N\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|^2$, which gives $c_1 = NS_{\max}$. □

Now we are well prepared for the proof of Proposition 7.4.

Proof of Proposition 7.4. Let us consider the Lyapunov function $V_1(Z(t)) \triangleq \|\mathbf{Q}(t)\|_1^2$, and the corresponding conditional mean drift is given by

$$\begin{aligned}
&\mathbb{E}\left[V_1(Z(t+1)) - V_1(Z(t)) \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t+1)\|_1^2 - \|\mathbf{Q}(t)\|_1^2 \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[\left(\|\mathbf{Q}(t)\|_1 + \|\mathbf{A}(t)\|_1 - \|\mathbf{S}(t)\|_1 + \|\mathbf{U}(t)\|_1\right)^2 \mid Z(t) = Z\right] \\
&- \mathbb{E}\left[\|\mathbf{Q}(t)\|_1^2 \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[2\|\mathbf{Q}\|_1 \left(\|\mathbf{A}\|_1 - \|\mathbf{S}\|_1\right) + \left(\|\mathbf{A}\|_1 - \|\mathbf{S}\|_1\right)^2\right. \\
&\left. + 2\left(\|\mathbf{Q}\|_1 + \|\mathbf{A}\|_1 - \|\mathbf{S}\|_1\right)\|\mathbf{U}\|_1 + \|\mathbf{U}\|_1^2 \mid Z(t) = Z\right] \\
=&\mathbb{E}\left[2\|\mathbf{Q}\|_1 \left(\|\mathbf{A}\|_1 - \|\mathbf{S}\|_1\right) + \left(\|\mathbf{A}\|_1 - \|\mathbf{S}\|_1\right)^2\right. \\
&\left. + 2\|\mathbf{Q}(t+1)\|_1 \|\mathbf{U}\|_1 - \|\mathbf{U}\|_1^2 \mid Z(t) = Z\right] \qquad (20)
\end{aligned}$$

Under the assumption that the second moment of $\|\mathbf{Q}\|$ is bounded in steady state, we have the mean drift of $V_1(Z)$ is zero in steady state. Taking expectation of both sides of Eq. (20) with respect to the steady-state distribution $\overline{Z}^{(\epsilon)}$, yields

$$\epsilon\mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] = \frac{\zeta^{(\epsilon)}}{2} + \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] - \frac{1}{2}\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1^2\right] \qquad (21)$$

where $\zeta^{(\epsilon)} = (\sigma_\Sigma^{(\epsilon)})^2 + \nu_\Sigma^2 + \epsilon^2$. Then by utilizing the property of unused service shown in Lemma D.1, we have

$$\frac{\zeta^{(\epsilon)}}{2} + \overline{\mathcal{B}}^{(\epsilon)} - \frac{1}{2}c_1\epsilon \le \epsilon\mathbb{E}\left[\sum_{n=1}^{N} \overline{Q}_n^{(\epsilon)}\right] \le \frac{\zeta^{(\epsilon)}}{2} + \overline{\mathcal{B}}^{(\epsilon)},$$

in which $\overline{\mathcal{B}}^{(\epsilon)} \triangleq \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right]$. Since $\zeta^{(\epsilon)}$ converges to ζ, from the inequality above and the definition of heavy-traffic delay optimality, we can easily see that the sufficient and necessary condition is $\lim_{\epsilon\downarrow 0} \overline{\mathcal{B}}^{(\epsilon)} = 0$, which completes the proof. □

E PROOF OF PROPOSITION 7.5

Before we prove Proposition 7.5, we first define the following Lyapunov functions and their corresponding drifts.

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

with the corresponding T time-slot drift given by

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

Note that Lemma C.1 provides a upper bound on the drift of $W(Z)$ for $T = 1$. Similarly, the following lemma provides a lower bound on the drift of $W_\parallel(Z)$ for $T = 1$.

Lemma E.1. *For any $t \ge 0$, we have*

$$\|\mathbf{Q}_\parallel(t+1)\|^2 - \|\mathbf{Q}_\parallel(t)\|^2 \ge 2\langle \mathbf{Q}_\parallel(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle.$$

Proof.

$$
\begin{aligned}
&\|\mathbf{Q}_{\|}(t+1)\|^2 - \|\mathbf{Q}_{\|}(t)\|^2 \\
=&2\langle \mathbf{Q}_{\|}(t), \mathbf{Q}_{\|}(t+1) - \mathbf{Q}_{\|}(t)\rangle + \|\mathbf{Q}_{\|}(t+1) - \mathbf{Q}_{\|}(t)\|^2 \\
\geq&2\langle \mathbf{Q}_{\|}(t), \mathbf{Q}_{\|}(t+1) - \mathbf{Q}_{\|}(t)\rangle \\
=&2\langle \mathbf{Q}_{\|}(t), \mathbf{Q}(t+1) - \mathbf{Q}(t)\rangle - 2\langle \mathbf{Q}_{\|}(t), \mathbf{Q}_{\perp}(t+1) - \mathbf{Q}_{\perp}(t)\rangle \\
\overset{(a)}{\geq}&2\langle \mathbf{Q}_{\|}(t), \mathbf{Q}(t+1) - \mathbf{Q}(t)\rangle \\
\overset{(b)}{\geq}&2\langle \mathbf{Q}_{\|}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle
\end{aligned}
$$

where the inequality (a) is true because $\langle \mathbf{Q}_{\|}(t), \mathbf{Q}_{\perp}(t)\rangle = 0$ and $\langle \mathbf{Q}_{\perp}(t+1), \mathbf{Q}_{\|}(t)\rangle = 0$; (b) follows from the fact that all the components of $\mathbf{Q}_{\|}(t)$ and $\mathbf{U}(t)$ are nonnegative. □

Based on the bounds on $W(Z)$ and $W_{\|}(Z)$, we are able to bound the drift of $V_{\perp}(Z)$ as follows.

Lemma E.2. *For any $t_0 \geq 0$ and $Z \in \mathcal{S}$, we have*

$$
\begin{aligned}
&\mathbb{E}\left[\Delta V_{\perp}(Z) \mid Z(t_0) = Z\right] \\
\leq&\frac{1}{2\|\mathbf{Q}_{\perp}(t_0)\|}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\left(2\langle \mathbf{Q}_{\perp}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z(t_0) = Z\right]
\end{aligned}
$$

Proof. Let us define $\Psi(t) \triangleq \|\mathbf{Q}(t+1)\|^2 - \|\mathbf{Q}(t)\|^2$ and $\Psi_{\|}(t) \triangleq \|\mathbf{Q}_{\|}(t+1)\|^2 - \|\mathbf{Q}_{\|}(t)\|^2$, then

$$
\begin{aligned}
&\mathbb{E}\left[\Delta V_{\perp}(Z) \mid Z(t_0) = Z\right] \\
\overset{(a)}{\leq}&\frac{1}{2\|\mathbf{Q}_{\perp}(t_0)\|}\mathbb{E}\left[\Delta W(Z) - \Delta W_{\|}(Z) \mid Z(t_0) = Z\right] \\
=&\frac{1}{2\|\mathbf{Q}_{\perp}(t_0)\|}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\Psi(t) - \Psi_{\|}(t) \mid Z(t_0 = Z)\right] \\
\overset{(b)}{\leq}&\frac{1}{2\|\mathbf{Q}_{\perp}(t_0)\|}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\left(2\langle \mathbf{Q}_{\perp}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle + L\right) \mid Z(t_0) = Z\right]
\end{aligned}
$$

where (a) is a natural extension of Lemma 7 in [4], which can be easily proved via the concavity of root function; (b) follows directly from Lemmas C.1 and E.1. □

Now we get ready to present the proof of Proposition 7.5.

Proof of Proposition 7.5. To show the bounded moments of $\|\mathbf{Q}_{\perp}\|$, we will again verify the two conditions in Lemma A.1 when applied to the Lyapunov function $V_{\perp}(Z) \triangleq \|\mathbf{Q}_{\perp}\|$.

For condition (C2), we have

$$
\begin{aligned}
&|\Delta V_{\perp}(Z)| \\
=&|\,\|\mathbf{Q}_{\perp}(t_0+T)\| - \|\mathbf{Q}_{\perp}(t_0)\|\,|\mathcal{I}(Z(t_0)=Z) \\
\overset{(a)}{\leq}& \|\mathbf{Q}_{\perp}(t_0+T) - \mathbf{Q}_{\perp}(t_0)\|\,\mathcal{I}(Z(t_0)=Z) \\
=&\|\mathbf{Q}(t_0+T) - \mathbf{Q}_{\parallel}(t_0+T) - \mathbf{Q}(t_0) + \mathbf{Q}_{\parallel}(t_0)\|\,\mathcal{I}(Z(t_0)=Z) \\
\overset{(b)}{\leq}& \|\mathbf{Q}(t_0+T) - \mathbf{Q}(t_0)\| + \|\mathbf{Q}_{\parallel}(t_0+T) - \mathbf{Q}_{\parallel}(t_0)\|\,\mathcal{I}(Z(t_0)=Z) \\
\overset{(c)}{\leq}& 2\,\|\mathbf{Q}(t_0+T) - \mathbf{Q}(t_0)\|\,\mathcal{I}(Z(t_0)=Z) \\
\overset{(d)}{\leq}& 2T\sqrt{N}\max(A_{\max}, S_{\max})
\end{aligned}
\tag{22}
$$

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

For condition (C1), by Lemma E.2, we will first focus on the inner product between $\mathbf{Q}_{\perp}(t)$ and $\mathbf{A}(t) - \mathbf{S}(t)$.

$$
\begin{aligned}
&\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_{\perp}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t_0)=Z\right] \\
\overset{(a)}{=}& \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{Q}_{\perp}(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t)\right] \mid Z(t_0)=Z\right] \\
\overset{(b)}{=}& \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{Q}_{\perp}(t), \mathbf{A}(t)\rangle \mid Z(t)\right] \mid Z(t_0)=Z\right] \\
\overset{(c)}{=}& \lambda_{\Sigma} \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\sum_{n=1}^{N} \left(Q_{\sigma_n(t)}(t) - Q_{\text{avg}}(t)\right)\left(\Delta_n(t) + \frac{1}{N}\right) \mid Z(t_0)=Z\right] \\
\overset{(d)}{=}& \lambda_{\Sigma}\mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=t_0}^{t_0+T-1} Q_{\sigma_n(t)}(t)\Delta_n(t) \mid Z(t_0)=Z\right] \\
\overset{(e)}{\leq}& \lambda_{\Sigma}\mathbb{E}\left[\sum_{n=1}^{N} Q_{\sigma_n(t_0)}(t_0)\sum_{t=t_0}^{t_0+T-1}\Delta_n(t) \mid Z(t_0)=Z\right] + K \\
\overset{(f)}{=}& \lambda_{\Sigma}\sum_{n=1}^{N}\left(Q_{\sigma_n(t_0)}(t_0) - Q_{\text{avg}}(t_0)\right)\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\Delta_n(t) \mid Z(t_0)=Z\right] + K,
\end{aligned}
$$

where (a) follows from the tower property of conditional expectation and the fact that $\mathbf{Q}(t)$, $\mathbf{A}(t)$ and $\mathbf{S}(t)$ are conditionally independent of $Z(t_0)$ when given $Z(t)$. (b) holds because the servers are assumed to be homogeneous; (c) follows from the definition of the dispatching preference $\Delta(t)$; (d) holds since $\sum_{n=1}^{N}(Q_{\sigma_n(t)}(t) - Q_{\text{avg}}(t)) = 0$ and $\sum_{n=1}^{N}\Delta_n(t) = 0$; (e) follows from Lemma 7.1 and the constant $K \triangleq \mu_{\Sigma}NT^2\max(A_{\max}, S_{\max})$; (f) is true since $\sum_{n=1}^{N}\Delta_n(t) = 0$ for all $t \geq 0$ and $Q_{\text{avg}}(t_0)$ is the average queue length at time-slot t_0.

Now from Lemma 7.2, we have for any $\epsilon_1 > 0$ there exists a T_n for each n such that for all $T > T_n$

$$\widetilde{\Delta}_n - \epsilon_1 \le \frac{1}{T}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0)\right] \le \widetilde{\Delta}_n + \epsilon_1. \tag{23}$$

Let $Q_\perp^{(n)}(t_0) \triangleq Q_{\sigma_n(t_0)}(t_0) - Q_{\text{avg}}(t_0)$ and $T > \max(T_1, T_2, \ldots, T_N)$, then

$$\begin{aligned}
&\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t_0) = Z\right] \\
\le &\lambda_\Sigma \sum_{n=1}^{N} Q_\perp^{(n)}(t_0)\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0) = Z\right] + K \\
\overset{(a)}{\le} &\lambda_\Sigma T \sum_{n=1}^{N} Q_\perp^{(n)}(t_0)\left(\widetilde{\Delta}_n + \left(2\mathcal{I}\left(Q_\perp^{(n)}(t_0) \ge 0\right) - 1\right)\epsilon_1\right) + K \\
\le &\lambda_\Sigma T\left(\sum_{n=1}^{N} Q_\perp^{(n)}(t_0)\widetilde{\Delta}_n + \epsilon_1 \left\|\mathbf{Q}_\perp(t_0)\right\|_1\right) + K,
\end{aligned} \tag{24}$$

in which (a) comes from Eq. (23). Now, let us turn to analyze the term $\sum_{n=1}^{N} Q_\perp^{(n)}(t_0)\widetilde{\Delta}_n$. Note that LDPC condition, i.e., $\widetilde{\Delta}_1 \ge \widetilde{\Delta}_2 \ge \cdots \ge \widetilde{\Delta}_N$ and $\widetilde{\Delta}_1 \ne \widetilde{\Delta}_N$, implies that $\widetilde{\Delta}_1 > 0$ and $|\widetilde{\Delta}_N| > 0$ since $\sum_{n=1}^{N} \widetilde{\Delta}_n = 0$. Moreover, we have

$$\begin{aligned}
\sum_{n=1}^{N} Q_\perp^{(n)}(t_0)\widetilde{\Delta}_n &\overset{(a)}{=} \sum_{n=1}^{N} Q_{\sigma_{t_0}(n)}(t_0)\widetilde{\Delta}_n \\
&\overset{(b)}{\le} -\min\left(|\widetilde{\Delta}_1|, |\widetilde{\Delta}_N|\right)\left(Q_{\sigma_{t_0}(N)}(t_0) - Q_{\sigma_{t_0}(1)}(t_0)\right) \\
&\overset{(c)}{\le} -\delta \left\|\mathbf{Q}_\perp(t_0)\right\|_1
\end{aligned} \tag{25}$$

where (a) follows $\sum_{n=1}^{N} \widetilde{\Delta}_n = 0$; (b) comes from the monotonicity of both $Q_{\sigma_{t_0}(n)}(t_0)$ and $\widetilde{\Delta}_n$, as well as the fact that $\sum_{n=1}^{N} \widetilde{\Delta}_n = 0$; (c) is true since $\left\|\mathbf{Q}_\perp(t_0)\right\|_1 \le N\left(Q_{\sigma_{t_0}(N)}(t_0) - Q_{\sigma_{t_0}(1)}(t_0)\right)$ and $\delta \triangleq \frac{1}{N}\min\{|\widetilde{\Delta}_1|, |\widetilde{\Delta}_N|\} > 0$, which is independent of ϵ. Now combining Eq. (24) with Eq. (25), yields

$$\begin{aligned}
&\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t_0) = Z\right] \\
\le &\lambda_\Sigma T\left(-\delta \left\|\mathbf{Q}_\perp(t_0)\right\|_1 + \epsilon_1 \left\|\mathbf{Q}_\perp(t_0)\right\|_1\right) + K \\
\le &-\lambda_\Sigma T\left(\delta - \epsilon_1\right)\left\|\mathbf{Q}_\perp(t_0)\right\| + K
\end{aligned}$$

which holds since $\|\mathbf{x}\|_1 \ge \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^N$. Choose $\epsilon_1 = \frac{\delta}{2} > 0$, then for all $\epsilon \le \frac{\mu_\Sigma}{2}$, we can find a finite T such that

$$\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t_0) = Z\right] \le -\eta_1 \left\|\mathbf{Q}_\perp(t_0)\right\| + K,$$

in which $\eta_1 \triangleq \frac{T\delta\mu_\Sigma}{4}$. Now substituting the inequality above into the upper bound in Lemma E.2, we conclude that

$$\mathbb{E}\left[\Delta V_\perp(Z) \mid Z(t_0) = Z\right] \le -\eta_1 + \frac{K_2}{\left\|\mathbf{Q}_\perp(t_0)\right\|} \tag{26}$$

where $K_2 \triangleq K + \frac{LT}{2}$ and η_1 are both independent of ϵ. Hence, by Lemma A.1, we have all the moments of $\left\|\mathbf{Q}_\perp^{(\epsilon)}\right\|$ in steady state are bounded by a constant independent of ϵ when $\epsilon \le \frac{\mu_\Sigma}{2}$.

Finally, we will show that bounded second moment of $\|\mathbf{Q}_\perp\|$ implies that Eq. (7) holds. Note that the left-hand-side of Eq. (7) can be upper bounded as follows

$$\begin{aligned}
&\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \\
&\overset{(a)}{=} N\mathbb{E}\left[\langle \overline{\mathbf{U}}^{(\epsilon)}(t), -\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t+1)\rangle\right] \\
&\overset{(b)}{\le} N\sqrt{\mathbb{E}\left[\left\|\overline{\mathbf{U}}_\perp^{(\epsilon)}(t)\right\|^2\right]\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t+1)\right\|^2\right]} \\
&\overset{(c)}{=} N\sqrt{\mathbb{E}\left[\left\|\overline{\mathbf{U}}_\perp^{(\epsilon)}(t)\right\|^2\right]\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t)\right\|^2\right]} \\
&\overset{(d)}{\le} N\sqrt{c_2\epsilon\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t)\right\|^2\right]}
\end{aligned} \tag{27}$$

where (a) comes from the property $Q_n^{(\epsilon)}(t+1)U_n^{(\epsilon)}(t) = 0$ for all $1 \le n \le N$ and all $t \ge 0$ in Lemma D.1 and the definition of $\mathbf{Q}_\perp$; (b) holds due to Cauchy-Schwartz inequality; (c) is true since the distributions of $\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t+1)$ and $\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t)$ are the same in steady state. Therefore, given that $\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}(t)\right\|^2\right]$ is bounded by a constant independent of ϵ, Eq. (7) directly holds, which establishes heavy-traffic delay optimality. □

F PROOF OF PROPOSITION 5.6

Proof. Recall Eq. (26) in the proof of Proposition 7.5, we have for all $\epsilon \le \frac{\mu_\Sigma}{2}$

$$\mathbb{E}\left[\Delta V_\perp(Z) \mid Z(t_0) = Z\right] \le -\eta_1 + \frac{K_2}{\|\mathbf{Q}_\perp(t_0)\|}$$

where $K_2 \triangleq K + \frac{LT}{2}$ and $\eta_1 \triangleq \frac{T\delta\mu_\Sigma}{4}$. Now we take a closer look at $\delta = \frac{1}{N}\min\{|\widetilde{\Delta}_1|, |\widetilde{\Delta}_N|\}$, which can be further lower bounded by $\frac{\left\|\widetilde{\Delta}\right\|_1}{2N^2}$. This is true since $\left\|\widetilde{\Delta}\right\|_1 \le 2N|\widetilde{\Delta}_1|$ and $\left\|\widetilde{\Delta}\right\|_1 \le 2N|\widetilde{\Delta}_N|$, which comes from the fact that $\widetilde{\Delta}_1 \ge \widetilde{\Delta}_2 \ge \cdots \ge \widetilde{\Delta}_N$ and $\sum_{n=1}^N \widetilde{\Delta}_n = 0$.

Therefore, based on Eq. (22) and the equation above, we can conclude that the drift of $V_\perp(Z)$ satisfies conditions (C1) and (C2) in Lemma A.1 with

$$\begin{aligned}
\kappa &= \frac{2K_2}{\eta_1} \\
\eta &= \frac{\left\|\widetilde{\Delta}\right\|_1 T\mu_\Sigma}{16N^2} \\
D &= 2T\sqrt{N}\max(A_{\max}, S_{\max})
\end{aligned}$$

Hence, substituting these three equalities into Eq. (14) in Lemma A.1, yields

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] &\le (2\kappa)^2 + (4D)^2\left(\frac{D+\eta}{\eta}\right)^2 2! \\ &\le \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\left(4H_1^2 + 2H_2^2\left(D + \frac{T\mu_\Sigma}{16N}\right)^2\right) \\ &\le \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_1 \end{aligned} \tag{28}$$

for all $\epsilon \le \frac{\mu_\Sigma}{2}$, in which $H_1 \triangleq \frac{8N^2(2K+LT)}{T\mu_\Sigma}$ and $H_2 \triangleq \frac{64DN^2}{T\mu_\Sigma}$ and $M_1 \triangleq 2\max\left(\frac{16N^2(2K+LT)}{T\mu_\Sigma}, 4\sqrt{2}DN\left(1+\frac{16DN}{T\mu_\Sigma}\right)\right)^2$. Hence M_1 is independent of ϵ, which completes the proof. □

G PROOF OF LEMMA 5.2

Proof. By Proposition 7.3, we have the second moment of $\|\mathbf{Q}\|$ in steady-state is bounded under LDPC. Hence, it follows from Eqs. (20) and (21) that

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{N}\overline{Q}_n^{(\epsilon)}\right] &\le \frac{\zeta^{(\epsilon)}}{2\epsilon} + \frac{1}{\epsilon}\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1\left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \\ &\overset{(a)}{\le} \frac{\zeta^{(\epsilon)}}{2\epsilon} + \frac{N}{\epsilon}\sqrt{c_2\epsilon\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}(t)\right\|^2\right]} \\ &\overset{(b)}{=} \frac{\zeta^{(\epsilon)}}{2\epsilon} + M\sqrt{\frac{\text{Degree of Queue Imbalance}}{\epsilon}} \end{aligned} \tag{29}$$

where (a) follows from Eq. (27) and $\zeta^{(\epsilon)} \triangleq (\sigma_\Sigma^{(\epsilon)})^2 + \nu_\Sigma^2$; (b) results from the definition of degree of queue imbalance and $M \triangleq \sqrt{c_2N^2}$. Note that, the Corollary 5.5 follows directly from Eq. (28) and Little's Law. □

H PROOF OF PROPOSITION 6.1

Proof of (1). By definition, we have $\Delta_n^{(d)} = P_n^{(d)} - \frac{1}{N}$, in which $P_n^{(d)}$ of power-of-d is given by

$$P_n^{(d)} = \binom{N-n}{d-1}\Big/\binom{N}{d} = \frac{d}{N}\prod_{j=1}^{d-1}\left(1-\frac{n-1}{N-j}\right).$$

For the upper bound, we are interested in the index n such that $\Delta_n^{(d)} \ge 0$, i.e., $P_n^{(d)} \ge \frac{1}{N}$. Note that $P_n^{(d)}$ is non-increasing and $P_n^{(d)} \ge \frac{d}{N}\left(1-\frac{n}{N-d}\right)^d$ for all n. Then it follows that $\Delta_n^{(d)} \ge 0$ for all $n \le n^*$, where n^* satisfies

$$d\left(1-\frac{n^*}{N-d}\right)^d = 1,$$

which gives $n^* = \left(1-d^{-1/d}\right)(N-d)$. Hence, by the property that $\sum_{n=1}^{N}\Delta_n^{(d)} = 0$ and $\min(\Delta_n^{(d)}, 1 \le n \le N) = -\frac{1}{N}$, we have

$$\left\|\Delta^{(d)}\right\|_1 \le 2\frac{1}{N}(N-n^*) = 2\left(\frac{1}{d}\right)^{1/d} + \frac{2}{N}\left(1-\left(\frac{1}{d}\right)^{1/d}\right)d, \tag{30}$$

which converges to $2\left(\frac{1}{d}\right)^{1/d}$ as $N \to \infty$.

For the lower bound, we are interested in the index n such that $P_n^{(d)} \leq \frac{\varepsilon}{N}$ for some positive constant $\varepsilon < 1$. Note that $P_n^{(d)} \leq \frac{d}{N}\left(1 - \frac{n-1}{N-1}\right)^{d-1}$ and it is non-increasing. It follows that $P_n^{(d)} \leq \frac{\varepsilon}{N}$ for all $n \geq \hat{n}$, where $\hat{n}$ satisfies equation

$$d\left(1 - \frac{\hat{n}-1}{N-1}\right)^{d-1} = \varepsilon,$$

which gives $\hat{n} = (N-1)\left(1 - \left(\frac{\varepsilon}{d}\right)^{\frac{1}{d-1}}\right) + 1$. Hence, since $\Delta_n^{(d)}$ is non-increasing and its sum is zero, we have

$$\left\|\Delta^{(d)}\right\|_1 \geq 2\frac{1-\varepsilon}{N}(N-\hat{n}) = 2\left(\frac{\varepsilon}{d}\right)^{\frac{1}{d-1}}(1-\varepsilon) - 2\frac{\left(\frac{\varepsilon}{d}\right)^{\frac{1}{d-1}}}{N}(1-\varepsilon), \tag{31}$$

which converges to $2\left(\frac{\varepsilon}{d}\right)^{\frac{1}{d-1}}(1-\varepsilon)$ as $N \to \infty$. This lower bound obtains its maximal value $2\frac{d-1}{d}\left(\frac{1}{d^2}\right)^{\frac{1}{d-1}}$ at $\varepsilon = \frac{1}{d}$. Note that $\varepsilon = \frac{1}{d} < 1$ for all $d \geq 2$. Hence, this completes the proof for the assertion in (1). □

Proof of (2). First, it is easy to see that $\lim_{N\to\infty}\left\|\Delta^{(N)}\right\|_1 = 2$. This is simply true since $\left\|\Delta^{(N)}\right\|_1 = \frac{2(N-1)}{N}$.

To show that $\lim_{N\to\infty}\left\|\Delta^{(d)}\right\|_1 = 2$, we are interested in the index n such that $P_n^{(d)} \leq \frac{\varepsilon}{N}$ for some fixed $0 < \varepsilon < 1$ when N is large enough. And again note that $P_n^{(d)} \leq \frac{d}{N}\left(1 - \frac{n-1}{N-1}\right)^{d-1}$, it suffices to consider the n such that

$$d\left(1 - \frac{n-1}{N-1}\right)^{d-1} \leq \varepsilon \tag{32}$$

holds when N is large enough. In particular, let us consider the case where n is a linear function of N that is defined as $n(\alpha) = \alpha N + 1 - \alpha$ for some constant $0 < \alpha < 1$. Substituting it into the left-hand-side of Eq. (32), yields

$$d\left(1 - \frac{n(\alpha)-1}{N-1}\right)^{d-1} = d(1-\alpha)^{d-1},$$

which converges to zero as $N \to \infty$ since $d(N) \to \infty$. As a result, for a sufficient large N, $P_n^{(d)} \leq \frac{\varepsilon}{N}$ for all $n \geq n(\alpha)$ since $P_n^{(d)}$ is non-increasing. Hence, by the property that the sum of $\Delta_n^{(d)}$ is zero, we obtain for a sufficient large N

$$\left\|\Delta^{(d)}\right\|_1 \geq 2(N - n(\alpha))\frac{1-\varepsilon}{N} = 2((1-\alpha)N + \alpha - 1)\frac{1-\varepsilon}{N}, \tag{33}$$

which converges to $2(1-\alpha)(1-\varepsilon)$ as $N \to 0$. Since Eq. (33) holds for arbitrary $0 < \varepsilon < 1$ and arbitrary $0 < \alpha < 1$, it directly follows $\lim_{N\to\infty}\left\|\Delta^{(d)}\right\|_1 = 2$, which finishes the proof for assertion (2). □

I PROOF OF PROPOSITION 6.2

The following two lemmas will be the basis for the proof of Proposition 6.2.

Lemma I.1. *If* $\lim_{N\to\infty} d(N) = \infty$ *and* $d(N) = o(N)$,

$$\log\left(\frac{\left\|\Delta^{(d)}\right\|_1 - 2\alpha}{2 - 2\alpha}\right) \leq \Theta\left(-\frac{\log d}{d}\right)$$

holds for any $\alpha \in (0,1)$.

Proof. See Appendix K.1 □

Lemma I.2. *If* $\lim_{N\to\infty} d(N) = \infty$,

$$\log\left(\frac{\left\|\Delta^{(d)}\right\|_1}{2}\right) \geq \Theta\left(-\frac{\log d}{d}\right).$$

Proof. See Appendix K.2 □

Now, we are ready to prove Proposition 6.2.

Proof of Proposition 6.2. By Lemma I.1, for a large enough d, we have

$$\log\left(\frac{\left\|\Delta^{(d)}\right\|_1 - 2\alpha}{2 - 2\alpha}\right) \leq -c_1\frac{\log d}{d},$$

holds for some constant $c_1 > 0$. This implies that

$$\frac{\left\|\Delta^{(d)}\right\|_1 - 2\alpha}{2 - 2\alpha} \leq 1 - c_1\frac{\log d}{d} + o\left(-\frac{\log d}{d}\right).$$

Multiplying both sides by $2 - 2\alpha$ and rearranging it, yields,

$$2 - \left\|\Delta^{(d)}\right\|_1 \geq (2 - 2\alpha)\left(c_1\frac{\log d}{d} + o\left(\frac{\log d}{d}\right)\right) = \Theta\left(\frac{\log d}{d}\right).$$

By Lemma I.2, for a large enough d, we have

$$\frac{\left\|\Delta^{(d)}\right\|_1}{2} \geq \left(1 - c_2\frac{\log d}{d} + o\left(-\frac{\log d}{d}\right)\right),$$

which directly implies

$$2 - \left\|\Delta^{(d)}\right\|_1 \leq \Theta\left(\frac{\log d}{d}\right).$$

Hence, this completes the proof. □

J PROOF OF COROLLARY 6.3

Proof of (1). By Proposition 6.2, we have

$$2 - \left\|\Delta^{(d)}\right\|_1 \leq C\frac{\log d}{d}$$

for some constant $C > 0$ when $d \geq d_0$. Let $K = 2C$ and $d = K\frac{1}{\epsilon}\log\frac{1}{\epsilon}$, then there exists a ϵ_0 such that $d \geq d_0$ for all $\epsilon \leq \epsilon_0$ and hence

$$2 - \left\|\Delta^{(d)}\right\|_1 \leq \epsilon\frac{\log K + \log\frac{1}{\epsilon} + \log\log\frac{1}{\epsilon}}{2\log\frac{1}{\epsilon}}.$$

Thus, there exists a ϵ_1 such that for any $\epsilon < \epsilon_1$, the right-hand-side of the inequality above is strictly less than one. Take $\epsilon^* = \min(\epsilon_0, \epsilon_1)$, we have completed the proof of (1). □

Proof of (2). Similarly, by Proposition 6.2, we have

$$2 - \left\|\Delta^{(d)}\right\|_1 \geq C\frac{\log d}{d}$$

for some constant $C > 0$ when $d \geq d_0$. If $d = O\left(\frac{1}{\epsilon}\left(\log\frac{1}{\epsilon}\right)^{(1-\gamma)}\right)$, then there exists a K and $\epsilon_1 > 0$ such that $d \leq K\frac{1}{\epsilon}\log\frac{1}{\epsilon}^{(1-\gamma)}$ for all $\epsilon < \epsilon_1$. Choose $\epsilon_2 < \epsilon_1$ such that for all $\epsilon < \epsilon_2$, $d \geq d_0$. Then, we have for any $\epsilon < \epsilon_2$,

$$2 - \left\|\Delta^{(d)}\right\|_1 \geq \epsilon\frac{C(\log K + \log\frac{1}{\epsilon} + (1-\gamma)\log\log\frac{1}{\epsilon})}{K(\log\frac{1}{\epsilon})^{(1-\gamma)}},$$

which is strictly larger than ϵ for some $\epsilon_3 > 0$. Take $\epsilon^* = \min(\epsilon_2, \epsilon_3)$, we have completed the proof of (2). □

K PROOF OF LEMMA I.1 AND LEMMA I.2

K.1 Proof of Lemma I.1

Proof. From Eq. (30), we have

$$\left\|\Delta^{(d)}\right\|_1 \leq 2\left(\frac{d}{N}\left(1 - \left(\frac{1}{d}\right)^{1/d}\right) + \left(\frac{1}{d}\right)^{1/d}\right)$$

For any $\alpha \in (0,1)$, since $d = o(N)$, we can find a N_0 and d_0 such that for all $N \geq N_0$ and $d \geq d_0$, $\frac{d}{N} \leq \alpha$ is always true. Hence, we have

$$\left\|\Delta^{(d)}\right\|_1 - 2\alpha \leq 2\left(\frac{1}{d}\right)^{1/d}(1-\alpha),$$

which directly implies the result, hence completing the proof. □

K.2 Proof of Lemma I.2

Proof. From Eq. (31), we have

$$\left\|\Delta^{(d)}\right\|_1 \geq 2\left(\frac{\varepsilon}{d}\right)^{\frac{1}{d-1}}(1-\varepsilon)\left(1 - \frac{1}{N}\right).$$

The right-hand-side achieves its maximal value at $\epsilon = \frac{1}{d}$, which gives

$$\begin{aligned}\left\|\Delta^{(d)}\right\|_1 &\geq 2\left(\frac{1}{d^2}\right)^{\frac{1}{d-1}}\left(1 - \frac{1}{d}\right)\left(1 - \frac{1}{N}\right)\\ &\geq 2\left(\frac{1}{d^2}\right)^{\frac{1}{d-1}}\left(1 - \frac{1}{d}\right)^2,\end{aligned}$$

which comes from the fact that $2 \leq d \leq N$. Thus, we have

$$\begin{aligned}\log\left(\frac{\left\|\Delta^{(d)}\right\|_1}{2}\right) &\geq 2\log\left(1 - \frac{1}{d}\right) - \frac{2}{d-1}\log d\\ &\overset{(a)}{=} 2\left(-\frac{1}{d} + o\left(-\frac{1}{d}\right)\right) - \frac{2}{d-1}\log d\\ &= \Theta\left(-\frac{\log d}{d}\right)\end{aligned} \tag{34}$$

where (a) comes from the fact that $d(N) \to \infty$. □

L PROOF OF PROPOSITION 8.2

Proof. The proof of Proposition 8.2 follows the same idea in the proof of Theorem 4.3. In particular, to show that the second moment of $\left\|\overline{\mathbf{Q}}_\perp\right\|$ in heavy-traffic limit is bounded, we would again verify the two conditions (C1) and (C2) in Lemma A.1 for $V_\perp(Z) \triangleq \|\mathbf{Q}_\perp\|$.

For condition (C2), it follows exactly the same as in the proof of Proposition 7.5, that is,

$$|\Delta V_\perp(Z)| \le 2T\sqrt{N}\max(A_{\max}, S_{\max}). \tag{35}$$

For condition (C1), we have

$$\begin{aligned}
&\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t_0) = Z\right] \\
&= \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t)\right] \mid Z(t_0) = Z\right] \\
&\overset{(a)}{=} \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\sum_{n=1}^{N} Q_\perp^{(n)}(t)\left(\lambda_\Sigma \Delta_n(t) - \epsilon\frac{\mu_{\sigma_n(t)}}{\mu_\Sigma}\right) \mid Z(t_0) = Z\right] \\
&\le \sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\sum_{n=1}^{N} Q_\perp^{(n)}(t)\lambda_\Sigma \Delta_n(t) + \epsilon\left\|\mathbf{Q}_\perp(t)\right\|_1 \mid Z(t_0) = Z\right] \\
&\overset{(b)}{\le} \lambda_\Sigma \sum_{n=1}^{N} Q_\perp^{(n)}(t_0)\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Delta_n(t) \mid Z(t_0) = Z\right] + T\sqrt{N}\epsilon\left\|\mathbf{Q}_\perp(t_0)\right\| + \hat{K}, \\
&\overset{(c)}{\le} -T\left(\lambda_\Sigma\left(\frac{\left\|\widetilde{\Delta}\right\|_1}{2N^2} - \epsilon_1\right) - \sqrt{N}\epsilon\right)\left\|\mathbf{Q}_\perp(t_0)\right\|_1 + \hat{K}
\end{aligned}$$

where (a) follows from the definition of $\Delta(t)$ for the general case and $Q_\perp^{(n)}(t) \triangleq Q_{\sigma_n(t)}(t) - Q_{\text{avg}}(t)$; (b) results from the first inequality in Eq. (24) and the bounded drift in condition (C2). As a result, $\hat{K} \triangleq \mu_\Sigma NT^2\max(A_{\max}, S_{\max}) + 2\epsilon NT^2\max(A_{\max}, S_{\max})$; (c) comes from Eqs. (24) and (25) and the fact that $\|\mathbf{x}\| \le \|\mathbf{x}\|_1$ for all $\mathbf{x}$ in $\mathbb{R}^N$.

We choose $\epsilon_1 = \frac{\left\|\widetilde{\Delta}\right\|_1}{4N^2} > 0$, then for all $\epsilon \le \epsilon_0 = \frac{\mu_\Sigma\left\|\widetilde{\Delta}\right\|_1}{2\left(\left\|\widetilde{\Delta}\right\|_1 + 4N^2\sqrt{N}\right)}$ we have

$$\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_\perp(t), \mathbf{A}(t) - \mathbf{S}(t)\rangle \mid Z(t_0) = Z\right] \le -\hat{\eta}_1\left\|\mathbf{Q}_\perp(t_0)\right\| + \hat{K},$$

in which $\hat{\eta}_1 \triangleq \frac{\mu_\Sigma T\left\|\widetilde{\Delta}\right\|_1}{8N^2}$. The rest of the proof follows exactly the same as in proof of Proposition 5.6. In particular, we have for all $\epsilon \le \epsilon_0$

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|^2\right] \le \frac{1}{\left\|\widetilde{\Delta}^{(\epsilon)}\right\|_1^2}\hat{M}_1 \tag{36}$$

in which $\hat{M}_1$ is independent of ϵ but $\widetilde{\Delta}$ is not in general, which highlights the difference.

In order to show heavy-traffic delay optimality, we would again apply the sufficient and necessary condition in Proposition 7.4. Thus, we need first prove that $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ is bounded for all $\epsilon > 0$. Similarly,

we check the conditions (C1) and (C2) for $V(Z) \triangleq \|\mathbf{Q}\|$. Since (C2) is exactly the same as before, we are only interested in (C1) for the general case.

Substituting the definition of $\Delta(t)$ for the general case into Eqs. (18) and (19), yields

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{Q}(t_0+T)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0)\right] \\
\leq & 2T\left(\lambda_\Sigma \sum_{n=1}^{N} Q_{\sigma_n(t_0)}(t_0)\widetilde{\Delta}_n + \lambda_\Sigma \epsilon_1 \|\mathbf{Q}(t_0)\|_1 - \frac{\epsilon\mu_{\min}}{\mu_\Sigma}\|\mathbf{Q}(t_0)\|_1\right) + \hat{K}_1 \\
\leq & 2T\|\mathbf{Q}(t_0)\|_1\left(\lambda_\Sigma\epsilon_1 - \frac{\epsilon\mu_{\min}}{\mu_\Sigma}\right) + \hat{K}_1 \\
\leq & -2\hat{\theta}T\|\mathbf{Q}(t_0)\| + \hat{K}_1,
\end{aligned}$$

where $\mu_{\min} \triangleq \min_n \mu_n$, $\hat{\theta} \triangleq \frac{\epsilon\mu_{\min}}{\mu_\Sigma} - \lambda_\Sigma\epsilon_1$, which is positive when choosing $\epsilon_1 < \frac{\epsilon\mu_{\min}}{(\mu_\Sigma)^2}$ and $\hat{K}_1 \triangleq 2LT + 2NT^2\max(A_{\max}, S_{\max})(\mu_\Sigma + \mu_{\max})$ and $\mu_{\max} \triangleq \max_n \mu_n$.

Then for condition (C1), we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta V(Z) \mid Z(t_0) = Z\right] \\
=&\mathbb{E}\left[\|\mathbf{Q}(t_0+T)\| - \|\mathbf{Q}(t_0)\| \mid Z(t_0) = Z\right] \\
=&\mathbb{E}\left[\sqrt{\|\mathbf{Q}(t_0+T)\|^2} - \sqrt{\|\mathbf{Q}(t_0)\|^2} \mid Z(t_0) = Z\right] \\
\leq & \frac{1}{2\|\mathbf{Q}(t_0)\|}\mathbb{E}\left[\|\mathbf{Q}(t_0+T)\|^2 - \|\mathbf{Q}(t_0)\|^2 \mid Z(t_0) = Z\right] \\
\leq & -T\hat{\theta} + \frac{\hat{K}_1}{2\|\mathbf{Q}(t_0)\|}.
\end{aligned}$$

This verifies (C1), hence establishing that all the moments of $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$ are bounded for any $\epsilon > 0$.

Now we are able to check the condition in Proposition 7.4. Note that as shown in Eq. (27), we have

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \leq N\sqrt{c_2\epsilon\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}_\perp(t)\right\|^2\right]}$$

Then according to Eq.(36), we have for all $\epsilon \leq \epsilon_0$

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \leq N\sqrt{c_2\hat{M}_1\epsilon\frac{1}{\left\|\widetilde{\Delta}^{(\epsilon)}\right\|_1^2}}$$

which approaches zero as ϵ goes to zero when $\left\|\widetilde{\Delta}^{(\epsilon)}\right\|_1^2 = \omega(\epsilon)$, hence establishing heavy-traffic delay optimality. □

M PROOF OF PROPOSITION 8.3

Proof. As in the proof of Proposition 5.7, we have the following equality in steady state for any $\epsilon > 0$.

$$\begin{aligned}
&2\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left((\overline{Q}_i^+)^{(\epsilon)}\overline{U}_j^{(\epsilon)} + (\overline{Q}_j^+)^{(\epsilon)}U_i^{(\epsilon)}\right)\right] \\
=&2N\mathbb{E}\left[\langle\overline{\mathbf{Q}}_\perp^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)} - \overline{\mathbf{S}}^{(\epsilon)}\rangle\right] - \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \\
&+\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \qquad (37)
\end{aligned}$$

First, as before, the left-hand-side (LHS) of Eq. (37) can be rewritten as

$$LHS = 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1\left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \triangleq 2\overline{\mathcal{B}}^{(\epsilon)},$$

which holds because $Q_n(t+1)U_n(t) = 0$ for all n and $t \geq 0$ as shown in Lemma D.1.

The main task is to simplify each term on the right-hand-side of Eq. (37) in the general case. For the first term on the right-hand-side of Eq. (37), we can rewrite it as follows

$$\begin{aligned}
\hat{\mathcal{T}}_1^{(\epsilon)} &\triangleq 2N\mathbb{E}\left[\langle\overline{\mathbf{Q}}_\perp^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)} - \overline{\mathbf{S}}^{(\epsilon)}\rangle\right] \\
&= 2N\mathbb{E}\left[\mathbb{E}\left[\langle\overline{\mathbf{Q}}_\perp^{(\epsilon)}, \overline{\mathbf{A}}^{(\epsilon)} - \overline{\mathbf{S}}^{(\epsilon)}\rangle \mid \overline{Z}^{(\epsilon)}\right]\right] \\
&= 2N\mathbb{E}\left[\langle\overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \lambda_\Sigma^{(\epsilon)}\left(\overline{\Delta}^{(\epsilon)} + \frac{\mu_{\overline{\sigma}}^{(\epsilon)}}{\mu_\Sigma}\right) - \mu_{\overline{\sigma}}^{(\epsilon)}\rangle\right] \\
&= 2N\lambda_\Sigma^{(\epsilon)}\mathbb{E}\left[\langle\overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \overline{\Delta}^{(\epsilon)}\rangle\right] + 2N\mathbb{E}\left[\langle\overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, -\frac{\epsilon}{\mu_\Sigma}\mu_{\overline{\sigma}}^{(\epsilon)}\rangle\right],
\end{aligned}$$

which implies that

$$\lim_{\epsilon\downarrow 0}\hat{\mathcal{T}}_1^{(\epsilon)} = \lim_{\epsilon\downarrow 0} 2N\mu_\Sigma^{(\epsilon)}\mathbb{E}\left[\langle\overline{\mathbf{Q}}_{\overline{\sigma},\perp}^{(\epsilon)}, \overline{\Delta}^{(\epsilon)}\rangle\right], \qquad (38)$$

under the assumption (1) that the first moment of $\left\|\overline{\mathbf{Q}}_\perp^{(\epsilon)}\right\|$ is $o(1/\epsilon)$.

For the second term on the right-hand-side of Eq. (37), we have

$$\hat{\mathcal{T}}_2^{(\epsilon)} \triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \leq \epsilon(N-1)S_{\max}$$

which follows from the fact that $0 \leq U_n \leq S_{\max}$ for all n and $\mathbb{E}\left[\left\|\overline{\mathbf{U}}^{(\epsilon)}\right\|_1\right] = \epsilon$ as shown in Lemma D.1.

The third term on the right-hand-side of Eq. (37) can be simplified as follows

$$\begin{aligned}
\hat{\mathcal{T}}_3^{(\epsilon)} &\triangleq \sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \\
&= \mathcal{L}^{(\epsilon)} - 2\lambda_\Sigma^{(\epsilon)}N\mathbb{E}\left[\langle\mu_{\overline{\sigma},\perp}^{(\epsilon)}, \overline{\mathbf{P}}^{(\epsilon)}\rangle\right]
\end{aligned}$$

where $\mathcal{L}^{(\epsilon)} \triangleq (N-1)\left(\left(\sigma_\Sigma^{(\epsilon)}\right)^2 + \left(\lambda_\Sigma^{(\epsilon)}\right)^2 + \nu_\Sigma^2\right) + \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\mu_i - \mu_j\right)^2$.

Now by the sufficient and necessary condition in Proposition 7.4, we have $\lim_{\epsilon \downarrow 0} \overline{\mathcal{B}}^{(\epsilon)} = 0$. Consequently, from Eq. (37), we can obtain

$$\lim_{\epsilon \downarrow 0} \hat{\mathcal{T}}_1^{(\epsilon)} = -\lim_{\epsilon \downarrow 0} \hat{\mathcal{T}}_3^{(\epsilon)} \triangleq C \tag{39}$$

where C is a constant independent of ϵ under the assumption (2) in Proposition 8.3. Thus, combining Eqs. (38) and (39) and applying Cauchy-Schwartz inequality, yields the required result, hence completing the proof. □