Priority Algorithm for Near-data Scheduling: Throughput and Heavy-Traffic Optimality

Qiaomin Xie, Yi Lu
{qxie3, yilu4}@illinois.edu
University of Illinois at Urbana-Champaign

Abstract

The prevalence of data-parallel applications has made near-data scheduling an important problem. An example is the map task scheduling in the map-reduce framework. Wang et. al. [14] was the first to identify its capacity region and proposed a throughput-optimal algorithm based on MaxWeight. However, the study of the algorithm's delay performance revealed that it is only heavy-traffic optimal for a very special traffic scenario, where all traffic concentrates on a subset of servers. We propose a simple "local-tasks first" priority algorithm and show that it is throughput-optimal and heavy-traffic optimal for *all* traffic scenarios, i.e., it asymptotically minimizes the average delay as the arrival rate vector approaches the boundary of the capacity region. So far, it is the only known heavy-traffic optimal algorithm for this setting. As the algorithm is based on pre-determined priority, a direct application of the Lyapunov drift technique does not work. The main proof ideas are the construction of an *ideal load decomposition* and the separate treatment of two subsystems based on their ideal load. To the best of our knowledge, this is the only setup of affinity scheduling where a simple priority algorithm is shown to be heavy-traffic optimal. Simulation shows that our algorithm also significantly outperforms existing algorithms at loads away from the boundary of the capacity region.

1 Introduction

The collection of large data sets by online social networks, search engines, scientific research and the health-care industry have made traditional methods of data processing inadequate. The timely extraction of useful patterns and information from the large data set has motivated data-parallel processing. A popular example of data-parallel processing is the map-reduce framework, first used by Google [6], and widely adopted in various open-source and proprietary versions of Hadoop [4]. The *map* stage of the framework assigns the tasks to process the input data set, hence requires a near-data scheduler.

In order to facilitate parallel computation, a large data set is divided into small data chunks. Each data chunk is replicated on a few servers to increase availability. For each task, we call a server a *local server* for the task if the data chunk associated with the task is stored locally, and we call this task a *local task* for the server; otherwise, the server is called a *remote server* for the task and this task is called a *remote task* for the server. While assigning tasks, it is a critical consideration to schedule a task on a local server [15, 2, 3, 11]. Scheduling in this setting is called the *near-data* scheduling problem, or scheduling with *data locality*.

1.1 The Problem

The near-data scheduling problem is a special case of affinity scheduling [5, 13, 12, 9, 10] where each type of tasks has different processing rates on different subsets of servers. It has two unique features:

1. It is impractical to have one queue for each type of tasks.

The type of a particular task is determined by the location of its data chunks. To achieve high availability and yet avoid excessive amount of storage, each data chunk is typically replicated on a small number of servers. For instance, the number is 3 for the map-reduce cluster.

When a server breaks down and its disks are replaced, all its local data chunks are restored by copying from their respective replicas. To avoid excessive amount of traffic from any single server, which can disrupt

its service, it is desirable to distribute the replicas over a large number of servers. Hence, the current practice in a map-reduce cluster is for each data chunk to uniformly sample 3 servers. This makes the number of types *cubic* in the number of servers in a cluster, which itself can be as large as tens of thousands. As a result, it is impractical to have one queue for each type of tasks.

2. Local servers are faster (on average) than remote servers.

At a remote server, the data chunk for a particular task needs to be retrieved over the network before processing. It is measured in [15] that a remote server on average takes twice as much time as a local server. From the perspective of each task, the cluster is divided into two subsets, one with a faster rate than the other, although the subsets vary for different tasks. This regularity in service rates calls for a simple optimal algorithm. There can also be an intermediate rate in map-reduce clusters called rack-local. We do not include it in our model, although it is straightforward to extend our algorithm to this setting. We do not consider pre-fetching in this paper.

In addition, the algorithm should not assume any knowledge of arrival rates in order to be robust with load variations.

1.2 Previous Work

The existing work on affinity scheduling [5, 13, 12, 9, 10] requires a queue for each type of tasks, hence does not apply in this setting.

Wang et. al. [14] was the first to formulate this problem from a stochastic network perspective and identified its capacity region. They proposed a new queueing structure and a scheduling algorithm consisting of the Join the Shortest Queue (JSQ) together with the MaxWeight policy. The JSQ step distributes the load into local and remote queues: a task is pre-assigned as remote if its local queues are longer than the remote queue. The MaxWeight step stabilizes the queues with a threshold-based priority policy.

The JSQ-MaxWeight algorithm was shown to be throughput-optimal. However, it was shown in [14] that it is heavy-traffic optimal only for a very special traffic scenario, where all traffic concentrates on a subset of servers. In particular, some servers receive *zero* local tasks and only provide remote service; and any server with non-zero local tasks is overloaded (with load exceeding 1) and *requires* remote service as a result.

There are a number of heuristics [4, 15, 11] proposed for near-data scheduling, but their fundamental throughput and delay properties are not known.

1.3 Our Approach

We use the same formulation as [14]. The cluster is modeled as a time-slotted system, in which tasks arrive at the beginning of each time slot according to some stochastic process. Each task processes one data chunk. Within each time slot, a task is completed with probability α at a local server, or with probability γ ($\gamma < \alpha$) at a remote server. We consider two traffic scenarios that require distinct proof techniques (although our algorithm does not distinguish between them as we assume no knowledge of arrival rates):

Evenly loaded. This is the case where with appropriate load balancing, each server can accommodate its load locally. No remote service is necessary in this scenario.

Locally overloaded (hotspots). More often, the data requested by the incoming traffic are skewed towards a subset of servers [3] and exceed their capacity. We call these servers *beneficiaries* as they require remote service to remain stable, and call the servers with spare capacity *helpers*. This includes the special scenario in [14] for which the JSQ-MaxWeight algorithm is shown to be heavy-traffic optimal, and is more general as it allows non-zero local traffic at helpers, as well as traffic that is local to both a helper and a beneficiary.

We use a simple queueing structure where each server has a queue storing its local tasks. We propose an algorithm where a newly arrived task is routed to a local server with the shortest queue; each server processes tasks from its local queue as long as it is non-empty, and when its local queue is empty, the server processes a remote task from the longest queue in the system. We establish the following results:

- We prove that our algorithm is throughput optimal, i.e., it can stabilize any arrival rate vector strictly within the capacity region identified in [14]. Since the algorithm has a predetermined priority of "local-tasks first", existing techniques using the L_2 norm Lyapunov drift, such as in [14], do not apply: There exist states with arbitrarily large L_2 norm where the drift remains positive. The main idea is the

construction of the *ideal load decomposition* for each arrival vector, which separates the servers into helpers and beneficiaries. The stability of the helper subsystem (which by itself is not Markovian) is established first, and the spare capacity helps stabilize the beneficiary subsystem.

- In addition, we prove that our algorithm is heavy-traffic optimal for both the evenly loaded and locally overloaded scenarios, i.e., it asymptotically minimizes the average delay as the arrival rate vector approaches the boundary of the capacity region. Since [14] shows heavy-traffic optimality only for a special traffic scenario, our algorithm is so far the only known heavy-traffic optimal algorithm. Further, to the best of our knowledge, this is the only setting of affinity scheduling where a "local-tasks first" algorithm is shown to be heavy-traffic optimal, which can be of separate interest.

  The locally overloaded case is the more challenging of the two. The proof first establishes *state-space collapse*, where we show that the helper subsystem has uniformly bounded moments independent of the heavy-traffic parameter, and the beneficiary subsystem reduces to a single dimension where all queue lengths are equal. We remark that this result depends on our "local-tasks first" policy as the helper queues are drained first, *independent* of the beneficiaries. In contrast, JSQ-MaxWeight results in helper queues growing proportionally with the beneficiaries. The proof uses construction of *ideal* processes to bound the dependence between helpers and beneficiaries through shared local arrivals and remote services.

Finally, simulation shows that our algorithm significantly outperforms the JSQ-MaxWeight algorithm at loads away from the boundary of the capacity region. Delay improvement up to a factor of 4 is observed.

2 System Model

We consider a discrete-time model for a computing cluster with M parallel servers, indexed by $m \in \{1, 2, \cdots, M\}$. Each data chunk is replicated on a set $\bar{L}$ of servers. As each task processes one data chunk, it has $|\bar{L}|$ local servers. We define the *type* of a task as the set $\bar{L}$ of its local servers. For instance, with $|\bar{L}| = 3$ the task type $\bar{L}$ is defined as:

$$\bar{L} \in \{(m_1, m_2, m_3) \in \mathcal{M}^3, m_1 < m_2 < m_3\},$$

where m_1, m_2, m_3 are the indices of the three local servers. We use $m \in \bar{L}$ to denote that server m is a local server for type $\bar{L}$ tasks. We denote by $\mathcal{L}$ the set of task types and $N = |\mathcal{L}|$.

Arrivals. Let $A_{\bar{L}}(t)$ denote the number of type $\bar{L}$ tasks that arrive at the beginning of time slot t. We assume that the arrival process of type $\bar{L}$ tasks is i.i.d. with rate-$\lambda_{\bar{L}}$. We denote the arrival rate vector by $\lambda = (\lambda_{\bar{L}} : \bar{L} \in \mathcal{L})$. The number of total arrivals in one time slot is assumed to be bounded, i.e., $\sum_{\bar{L} \in \mathcal{L}} A_{\bar{L}}(t) \leq C_A$. We further assume bounded second moment of the arrival processes.

Services. For each task, we assume that its service time follows a geometric distribution with mean $1/\alpha$ if processed at a local server, and with mean $1/\gamma$ at a remote server. On average, a task is processed faster at a local server. So we assume $\alpha > \gamma$. At most one task is being processed at each server at any time and all services are non-preemptive.

2.1 Algorithm

Our proposed algorithm is illustrated in Fig. 1. The central scheduler maintains a set of M queues within the scheduler, where the m-th queue, denoted by Q_m, only receives tasks local to server m. We call it a local queue for tasks of type $\bar{L}$ if $m \in \bar{L}$. Note that there can be tasks local to server m but buffered at Q_n, $n \neq m$, where server n is another local server for the tasks. Let the vector $Q(t) = (Q_1(t), Q_2(t), \cdots, Q_M(t))$ denote the queue lengths at time t. At the beginning of each time slot t, the central scheduler routes new arrivals to one of the queues and schedules a new task for an idle server as follows:

Figure 1: The proposed algorithm

Load balancing: When a task arrives, the scheduler compares the lengths of the task's local queues, $\{Q_m | m \in \bar{L}\}$, and inserts the task into the shortest queue. Ties are broken randomly. Let $A_{\bar{L},m}(t)$ denote

the number of type $\bar{L}$ tasks that are routed to Q_m. The total number of tasks that join queue Q_m, denoted by $A_m(t)$, is given by

$$A_m(t) = \sum_{\bar{L}:m\in\bar{L}} A_{\bar{L},m}(t).$$

Prioritized scheduling: Let $f_m(t)$ denote the working status of server m at time slot t.

$$\begin{cases} f_m(t) = -1 & \text{if server } m \text{ is idle} \\ f_m(t) = n & \text{if server } m \text{ serves a task from queue } n \end{cases}$$

When server m completes a task at the end of time slot $t-1$, i.e., $f_m(t^-) = -1$, it is available for a new task at time slot t. $f_m(t) = m$ indicates that server m is working on a local task, and $f_m(t) = n$, where $n \neq m$, indicates that server m is working on a remote task. The scheduling decision is based on the working status vector $f(t) = (f_1(t), f_2(t), \cdots, f_M(t))$ and queue length vector $Q(t)$.

- **Local tasks first.** When server m becomes idle, the scheduler sends the head-of-line task from Q_m.

- **Remote tasks.** When server m becomes idle and Q_m is empty, the scheduler sends a remote task to server m from the longest queue in the system, if the length of the longest queue, denoted by Q^{max}, exceeds the threshold $T_s = \alpha/\gamma$. The threshold is to ensure that the remote task will experience a smaller completion time in expectation, since the mean processing time at a remote server is $1/\gamma$, and the mean waiting time plus processing time at a local server is Q^{max}/α.

Let $\eta_m(t)$ denote the scheduling decision for server m at time slot t, which is the index of the queue server m is scheduled to serve. Note that $\eta_m(t) = f_m(t)$ for all busy servers, and when $f_m(t^-) = -1$, i.e., server m is idle, $\eta_m(t)$ is determined by the scheduler according to the algorithm.

2.2 Queue dynamics

Let $S^l_m(t)$ and $R_m(t)$ denote the local and remote service provided by server m respectively, where $S^l_m(t) \sim \text{Bern}(\alpha I_{\{\eta_m(t)=m\}})$ and $R_m(t) \sim \text{Bern}(\gamma I_{\{\eta_m(t)\neq m\}})$ are two Bernoulli random variables with varying probability: $S^l_m(t) \sim \text{Bern}(\alpha)$ when server m is scheduled to the local queue, and Bern(0) otherwise; $R_m(t) \sim \text{Bern}(\gamma)$ when server m is scheduled to a remote queue, and Bern(0) otherwise.

Note that the local service *received* by server m is also $S^l_m(t)$, whereas the remote service *received* by server m is $S^r_m(t) \equiv \sum_{n:n\neq m} R_n(t) I_{\{\eta_n(t)=m\}}$, which is the sum of all remote service provided by server n to sever m. Let $S_m(t) \equiv S^l_m(t) + S^r_m(t)$ denote the departure process for queue m. Hence the queue lengths satisfy the following equation:

$$Q_m(t+1) \quad = \quad Q_m(t) + A_m(t) - S_m(t) + U_m(t),$$

where $U_m(t) = \max\{0, S_m(t) - A_m(t) - Q_m(t)\}$ is the unused service. As the service times follow geometric distributions, $Q(t)$ together with the working status vector $f(t)$ form an irreducible and aperiodic Markov chain $\{Z(t) = (Q(t), f(t)), t \geq 0\}$.

3 Ideal Load Decomposition

A key component of the proof of both throughput and heavy-traffic optimality is a construction we call the ideal load decomposition. It is ideal in the sense that it *minimizes* the work in the system by locally serving as many tasks as possible. The construction serves two purposes: 1) The ideal load obtained for each server is used as an intermediary in the proofs of stability and state-space collapse; 2) The construction uniquely identifies two subsystems, helpers and beneficiaries, which have very different behavior and require distinct treatment in the proofs.

Helpers and Beneficiaries

A server is a helper if it is *not overloaded*, *provides* remote service and its local queue does *not receive* remote service under the ideal load decomposition. In contrast, a server is a beneficiary if it is *overloaded*, does *not provide* remote service, and its local queue *receives* remote service from the helpers. We will define an overloaded server in a more precise manner in 3.2. While pure helpers and beneficiaries do not exist in

a real system, the ideal load decomposition approximately depicts the load distribution in the heavy-traffic regime.

In the rest of the section, we construct the ideal load decomposition. We start from a new definition of the capacity region, which is equivalent to that identified in [14], but uses a more refined decomposition appropriate for our algorithm. The ideal load decomposition is constructed from this refined decomposition in two steps: 1) Identify the overloaded servers; 2) Construct the decomposition that produces helpers and beneficiaries.

3.1 An Equivalent Capacity Region

Let Λ be the set of arrival rates such that each element has a decomposition satisfying the following condition:

$$\begin{aligned}\Lambda \quad = \quad & \{\lambda = (\lambda_{\bar{L}} : \bar{L} \in \mathcal{L}) | \ \exists \ (\lambda_{\bar{L},n,m}) \ s.t. \\ & \lambda_{\bar{L},n,m} \geq 0, \forall \bar{L} \in \mathcal{L}, \forall n \in \mathcal{L}, m \in \mathcal{M}, \\ & \lambda_{\bar{L}} = \sum_{n:n\in\bar{L}} \sum_{m=1}^{M} \lambda_{\bar{L},n,m}, \forall \bar{L} \in \mathcal{L}, \\ & \sum_{\bar{L}:m\in\bar{L}} \sum_{n:n\in\bar{L}} \frac{\lambda_{\bar{L},n,m}}{\alpha} + \sum_{\bar{L}:m\notin\bar{L}} \sum_{n:n\in\bar{L}} \frac{\lambda_{\bar{L},n,m}}{\gamma} < 1, \qquad (1) \\ & \forall m \in \mathcal{M}\},\end{aligned}$$

where Eqn (1) states that the sum of the local and remote load at each server is less than 1.

Lemma 1 *The capacity region Λ is equivalent to the capacity region in [14].*

The proof is straightforward. The rate $\lambda_{\bar{L}}$ is decomposed in [14] into $\lambda_{\bar{L},m}$, which is the rate of type-$\bar{L}$ arrival allocated to server m. We further refine the decomposition by simply writing $\lambda_{\bar{L},m} \equiv \sum_n \lambda_{\bar{L},n,m}$, where n is the index of the queue at which a task is *queued* till processed at server m. Observe that $\lambda_{\bar{L},n,m} = 0$ if $n \notin \bar{L}$, since tasks only join their local queues with the proposed algorithm.

3.2 Overloaded Servers

Let $\nu_{n,m}$ denote the total rate of arrivals routed to Q_n, and eventually processed at server m, $\nu_{n,m} \equiv \sum_{\bar{L}:n\in\bar{L}} \lambda_{\bar{L},n,m}$.

Figure 2: Ideal load decomposition.

Figure 2 illustrates $(\nu_{n,m})$ where the m-th sub-queue at Q_n denotes the arrivals routed to Q_n but processed at server m. Note that the sub-queues are only a part of the construction, and do not exist in the actual data structure.

Let $\mathcal{S} \subseteq \mathcal{M}$ be a subset of servers. We denote by $\mathcal{L}_{\mathcal{S}}$ the set of task types *only local* to servers in $\mathcal{S}$

Lemma 2 *For any arrival rate vector $\lambda \in \Lambda$, there exists a decomposition $\{\tilde{\lambda}_{\bar{L},n,m}\}$ which satisfies Eqn. (1) and $\forall n \in \mathcal{D} = \{n \in \mathcal{M} : \sum_{\bar{L}:n\in\bar{L}} \sum_m \tilde{\lambda}_{\bar{L},n,m} \geq \alpha\}$, where $\mathcal{D}$ denotes the overloaded set with arrival rate greater than α,*

$$\tilde{\lambda}_{\bar{L},n,m} = 0, \quad \forall \bar{L} \notin \mathcal{L}_{\mathcal{D}}, m \in \mathcal{M}, \qquad (2)$$

Note that the decomposition $\{\tilde{\lambda}_{\bar{L},n,m}\}$ is such that for any overloaded set $\mathcal{D}$, it only receives non-zero arrivals from task types that are only local to $\mathcal{D}$. In other words, any task type that is also local to some server m not in the overloaded set, will be directed to m. This ensures that the set $\mathcal{D}$ is truly overloaded as no load balancing with the rest of the system will reduce its load. The decomposition also minimizes the total load in the system as the amount of local load is maximized. Note that $\mathcal{D}$ is unique for a given arrival

vector λ, although the decomposition $\{\tilde{\lambda}_{\bar{L},n,m}\}$ is not unique. When $\mathcal{D}$ is non-empty, we call the system *locally overloaded*.

The proof takes a decomposition $\left\{\lambda_{\bar{L},n,m}\right\}$ satisfying Eqn. (1), and iteratively moves an appropriate amount of load from overloaded ($\sum_m \nu_{n,m} \geq \alpha$) queues to underloaded ($\sum_m \nu_{n,m} < \alpha$) queues. This is possible whenever an overloaded queue receives local arrivals that are also local to some underloaded queue. At the end of each step, either there is no more shared local load between the two queues, or they have both become underloaded or overloaded. It can be shown that at each step, the decomposition continues to satisfy Eqn. (1) and reduces the total load in the system.

The full proof is provided in Appendix.

3.3 Ideal Load Decomposition

Lemma 3 *For any arrival rate vector $\lambda \in \Lambda$, there exists a decomposition $\left\{{\lambda^*}_{\bar{L},n,m}\right\}$ satisfying Eqn. (1) and for $\forall m \in \mathcal{M}$, either $m \in \mathcal{H}$ or $m \in \mathcal{B}$, where*

$$\begin{aligned}
\mathcal{H} &= \{n \in \mathcal{M} | \sum_{\bar{L}:n\in\bar{L}} \sum_m \lambda^*_{\bar{L},n,m} < \alpha, \\
&\quad \textit{and } {\lambda^*}_{\bar{L},n,m} = 0,\ \forall \bar{L} \in \mathcal{L}, \forall m \neq n\}, \\
\mathcal{B} &= \{n \in \mathcal{M} | \sum_{\bar{L}:n\in\bar{L}} \sum_m \lambda^*_{\bar{L},n,m} \geq \alpha, \\
&\quad \textit{and } \lambda^*_{\bar{L},n,m} = 0,\ \forall \bar{L} \notin \tilde{\mathcal{L}}_b, m \in \mathcal{M}, \\
&\quad \textit{and } \lambda^*_{\bar{L},m,n} = 0,\ \forall \bar{L} \in \mathcal{L}, \forall m \neq n\}.
\end{aligned}$$

Lemma 3 states that for any arrival vector, there exists an ideal load decomposition, under which a server is either a helper or a beneficiary. A helper server $n \in \mathcal{H}$ receives no remote service, hence $\nu_{n,m} = 0$ for all $m \neq n$. The Q_1 and Q_2 in Fig. 2 belong to such servers. Only the local sub-queue has non-zero rate, denoted by $\nu_{n,n}$. A beneficiary server $m \in \mathcal{B}$, provides no remote service, but receives remote service from helpers. The Q_M in Fig. 2 illustrates such a situation. Note that Q_M receives remote service from server 1 and 2.

The proof constructs the ideal load decomposition iteratively from $\{\tilde{\lambda}_{\bar{L},n,m}\}$ given in Lemma 2. The main idea is that if an underloaded server receives remote service, it can process this work locally while reducing the remote service it provides, until it becomes a helper; if an overloaded server provides remote service, it can instead use this service towards its local load while reducing the remote service it receives, until it becomes a beneficiary. The full proof can be found in Appendix A.

4 Throughput Optimality

We devote this section to the proof of the following theorem:

Theorem 1 (Throughput Optimality) *The proposed algorithm is throughput optimal. That is, it stabilizes any arrival rate vector strictly within the capacity region.*

By Lemma 1, it is equivalent to prove that the proposed algorithm stabilizes any arrival rate vector within Λ, defined in 3.1. The standard approach using a quadratic Lyapunov function does not apply in our setting, as a remote queue, despite its large queue length, can continue to grow, while a shorter queue receives local service, hence increasing the quadratic drift. Although the remote queue will be served after the local queue is empty, the time taken to obtain a negative drift will depend on the system state.

To address the challenge, we treat the helper and beneficiary subsystems, as defined in Lemma 3, separately. The proof has three main steps. First, we show that the helper subsystem is stable using an extension of Lemma 1 in [7]. If the beneficiary subsystem is empty, this alone proves Theorem 1. In the case where the *beneficiary* subsystem is non-empty, we show that the beneficiary queues are either all stable or none of them is stable. This allows us to show the stability of the *beneficiary* subsystem by contradiction.

Let M_h and M_b denote the number of helpers and beneficiaries, respectively. For simplicity, assume that $\mathcal{H} = \{1, 2, \cdots, M_h\}$, and $\mathcal{B} = \{M_h + 1, \cdots, ..., M\}$. Let $Q^{(H)}(t)$ and $Q^{(B)}(t)$ denote the vector of helper queues and beneficiary queues, respectively.

4.1 Stability of Helper Subsystem

We have the following lemma for the stability of helper subsystem.

Lemma 4 *For any arrival rate vector $\lambda \in \Lambda$, the helper queues defined by its ideal load decomposition will be stabilized with the proposed algorithm.*

Throughout this section, notations with superscript $^{\mathcal{H}}$ are used to denote the corresponding vectors for helpers. Since the arrivals and services for helpers depend on the state of beneficiaries, $\{Z^{(H)}(t) = (Q^{(H)}(t), f^{(H)}(t))\}_{t\geq 0}$ itself is not a Markov chain. We use an extension of Lemma 1 in [7], which can be derived from [8].

It states that the subsystem is stable if there exists a positive inter T and a Lyapunov function V defined on the subsystem only whose T time slot drift satisfies the following two conditions: (i) finite drift with probability 1; (ii) negative drift for sufficiently large V.

To prove T-period drift of $V_h(Z(t))$ satisfying condition (ii), we need the following lemmas. The main idea is to use the ideal decomposition as a potential set of arrival rates, and show that

1) The actual load arriving at Q_m with the proposed load balancing step is dominated, in an appropriate sense, by the ideal decomposition;

2) The local service at a helper server is sufficient to accommodate all load arriving at its queue according to the ideal decomposition;

We defer the proof of the lemmas to Appendix B.

Lemma 5 (Arrival.) *Consider any arrival rate vector $\lambda \in \Lambda$ and $\mathcal{H}$ is the corresponding helper set defined in Lemma 3. Then under the proposed algorithm, for any $t \geq t_0$,*

$$\mathbb{E}\left[\langle Q^{(H)}(t), A^{(H)}(t)\rangle - \langle Q^{(H)}(t), \lambda^{*H}\rangle | Z(t_0)\right] \leq 0. \tag{3}$$

The lemma states that the actual arrival rate on the left-hand-side of the equation is dominated by the arrival rates in the ideal decomposition, in a dot product with the queue lengths. It indicates that the proposed algorithm keeps the number of tasks at least as balanced as the ideal decomposition. The main idea of the proof is to regroup the arrival rates according to the *types* of tasks, and use the fact that an incoming task always joins the shortest queue.

Lemma 6 (Local service.) *Consider any arrival rate vector $\lambda \in \Lambda$ and $\mathcal{H}$ is the corresponding helper set defined in Lemma 3. Then under the proposed algorithm, there exists $T_1 > 0$ such that for any $T > T_1$ and any t_0,*

$$\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\left(\langle Q^{(H)}(t), \lambda^{*(H)}\rangle - \langle Q^{(H)}(t), S^{(H)}(t)\rangle\right)|Z(t_0)\right] \leq -\theta_1 |Q^{(H)}(t_0)|_1 + C_1, \tag{4}$$

where $\theta_1 > 0$ and C_1 are constants independent of $Z(t_0)$.

Note that λ_m^* is the rate of local arrivals for server m according to the ideal decomposition. Note that $\forall m \in \mathcal{H}$, $\lambda_m^* = \lambda_m^{*l}$, i.e., all local arrivals are served locally under the ideal decomposition. Thus Lemma 6 indicates that all servers are able to accommodate their local load assigned by the ideal decomposition. The proof uses the fact that the local service rate is always α as long as there is local tasks present, and the inequality is obtained using the definition of the capacity region.

Proof for Lemma 4:

Consider Lyapunov functions:

$$V_h(Z(t)) = ||Q^{(H)}(t)||, \quad W_h(Z(t)) = ||Q^{(H)}(t)||^2.$$

The corresponding T-period drifts are denoted by:

$$\begin{aligned}\Delta V_h(Z(t_0)) &:= \mathbb{E}\left[V(t_0+T)-V(t_0)|Z(t_0)\right]\\ \Delta W_h(Z(t_0)) &:= \mathbb{E}\left[W(t_0+T)-W(t_0)|Z(t_0)\right]\end{aligned}$$

Observe that $V_h(Z(t)) = \sqrt{W_h(Z(t))}$. By the concavity of the square root function, we have

$$\Delta V_h(Z(t_0)) \leq \frac{1}{2V_h(Z(t_0))}\mathbb{E}\left[W(t_0+T)-W(t_0)|Z(t_0)\right] = \frac{\Delta W_h(Z(t_0))}{2V_h(Z(t_0))}.$$

We start with analyzing $\Delta W_h(Z(t_0))$.

$$\begin{aligned}\Delta W_h(Z(t_0)) &= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} W(t+1)-W(t)|Z(t_0)\right]\\ &= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} ||Q(t+1)||^2-||Q(t)||^2|Z(t_0)\right]\\ &= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(2\langle Q(t),A(t)-S(t)\rangle + 2\langle Q(t),U(t)\rangle + ||A(t)-S(t)+U(t)||^2\right)|Z(t_0)\right]\end{aligned}$$

By Lemma **??**, the term $\langle Q(t),U(t)\rangle \leq M^2$. Since both the arrival vector $A(t)$ and the service vector $S(t)$ are bounded, so as the unused vector $U(t)$, the term $||A(t)-S(t)+U(t)||^2$ can be bounded by a constant. Thus the T-time slot drift can be bounded as

$$\Delta W_h(Z(t_0)) = \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \langle Q(t),A(t)-S(t)\rangle|Z(t_0)\right] + C$$

For any arrival rate vector $\lambda \in \Lambda$ and the corresponding ideal decomposition λ_m^*, we split the expectation term into two terms using λ_m^*:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \langle Q(t),A(t)-S(t)\rangle|Z(t_0)\right] &= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\langle Q(t),A(t)\rangle - \langle Q(t),\lambda^*\rangle\right)|Z(t_0)\right] \quad (5)\\ &+ \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\langle Q(t),\lambda^{*l}\rangle - \langle Q(t),S^l(t)\rangle\right)|Z(t_0)\right] \quad (6)\end{aligned}$$

By Lemma 5, $(5) \leq 0$.

By Lemma 6 $(6) \leq -\theta_1|Q(t_0)|_1 + C_1$, where θ_1 and C_1 are two positive constants independent of $Z(t_0)$.

Together, we have

$$\Delta W_h(Z(t_0)) \leq -\theta_1|Q(t_0)|_1 + C,$$

where $C > 0$ is a constant.

Therefore,

$$\Delta V_h(Z(t_0)) \leq \frac{\Delta W_h(Z(t_0))}{2V_h(Z(t_0))} \leq \frac{-\theta_1|Q(t_0)|_1 + C}{2|Q(t_0)|} \leq -\theta_1 + \frac{C}{2|Q(t_0)|}.$$

The last inequality comes from the fact that L^2 norm of a non-negative vector is always less than its L^1 norm. This means that we have negative drift for sufficiently large $V_h(Z)$.

In addition, by the boundedness of arrivals and service, we have

$$\left|||Q^{(H)}(t)|| - ||Q^{(H)}(t_0)||\right| \leq ||Q^{(H)}(t) - Q^{(H)}(t_0)|| \leq T\sqrt{M_h}\max\{M, C_A\}. \quad (7)$$

That is, the drift of $V_h(Z)$ is finite with probability 1. Using extended Lemma 1 in [7], $V_h(Z(t)), t \geq 0$ converges in distribution to a random variable $\hat{V}_h$, and there exist constants θ^* and $C*$ with $\theta^* > 0$ such that $\mathbb{E}\left[e^{\theta^* \hat{V}_h}\right] \leq C^*$. Thus the helper subsystem is stable. ■

Consider the helper subsystem in steady state. Observe that the total arrival rate for this subsystem is at most $\sum_{\bar{L} \in \mathcal{L}^*_{\mathcal{H}}} \lambda_{\bar{L}}$, where $\mathcal{L}^*_{\mathcal{H}}$ is the set of task types that have at least one local server in $\mathcal{H}$. Since arrivals at helper subsystem can be processed remotely, the total amount of local service provided by helper servers is no greater than $\sum_{\bar{L} \in \mathcal{L}^*_{\mathcal{H}}} \lambda_{\bar{L}}$. Hence the total amount of remote service provided by all helpers in steady state, denoted by $R_{\mathcal{H}}$, can be lower bounded as

$$R_{\mathcal{H}} \geq \gamma \left(M_h - \frac{1}{\alpha} \sum_{\bar{L} \in \mathcal{L}^*_{\mathcal{H}}} \lambda_{\bar{L}} \right). \tag{8}$$

4.2 Stability of Beneficiary Subsystem

Assume the beneficiaries subsystem is non-empty. We will prove the following important property of the set $\mathcal{B}$.

Lemma 7 *For any arrival rate vector $\lambda \in \Lambda$, either all queues in $\mathcal{B}$ are stable or none of them is stable.*

Proof. We prove this lemma by contradiction. Let $\mathcal{S}$ and $\mathcal{S}^c$ denote the set of stable and unstable beneficiaries, respectively. Assume that $\mathcal{S} \neq \emptyset$ and $\mathcal{S}^c \neq \emptyset$. By Lemma 4, helper queues $\mathcal{H}$ are stable. Consider the system with queues of $\mathcal{S} \cup \mathcal{H}$ in steady state. Since queues in $\mathcal{S}^c$ grow with time, the probability that the maximum queue is among $\mathcal{S} \cup \mathcal{H}$ is arbitrarily small. Hence the amount of remote service offered by helpers and devoted to queues $\mathcal{S} \cup \mathcal{H}$ can be arbitrarily small, denoted by $\delta > 0$. Thus $\forall m \in \mathcal{S}$, the amount of service it receives satisfies $\mathbb{E}[S_m(t)] \leq \alpha + \delta$.

Consider arrivals for $\mathcal{S}$. By a similar argument, the amount of tasks shared among $\mathcal{S}$ and $\mathcal{S}^c$ (if such task types exist) that join $\mathcal{S}^c$ can be arbitrarily small, denoted by $\sigma \geq 0$. Hence $\exists k \in \mathcal{S}$ such that $\mathbb{E}[A_k(t)] \geq \mathbb{E}\left[\sum_{m \in \mathcal{S}} A_m(t)\right] / |\mathcal{S}| > \alpha - \sigma/|\mathcal{S}|$. Thus there exits a constant $\theta > 0$ s.t. $\mathbb{E}[A_k(t)] \geq \alpha - \sigma/|\mathcal{S}| + \theta$. Choosing sufficiently small δ and σ, we can have $\mathbb{E}[S_k(t)] < \mathbb{E}[A_k(t)]$, which contradicts the assumption that beneficiary k is stable. ■

Next we show the stability of the beneficiary subsystem.

Lemma 8 *For any arrival rate vector $\lambda \in \Lambda$, all queues in $\mathcal{B}$ will be stabilized under the proposed algorithm.*

Proof. Again, we prove the statement by contradiction. By Lemma 7, we can assume that all beneficiaries are unstable. Using a similar argument as for Lemma 7, we can show that at most δ remote service is devoted to queues in $\mathcal{H}$. For all $n \in \mathcal{B}$, its instability implies $\mathbb{P}(Q_n(t) = 0) = 0$, hence $\mathbb{E}\left[S^l_m(t)\right] = \alpha$. Then we have $\mathbb{E}\left[\sum_{m \in \mathcal{B}} S_m(t)\right] \geq M_b \alpha + R_{\mathcal{H}} - \delta$.

And the total arrival rate for $\mathcal{B}$ is given by $\mathbb{E}[\sum_{m \in \mathcal{B}} A_m(t)] = \sum_{\bar{L} \in \mathcal{L}_{\mathcal{B}}} \lambda_{\bar{L}} + \sigma$, where σ is the amount of arrivals local to a helper but have joined a beneficiary. It can be made arbitrarily small as beneficiary queues become sufficiently large.

Define $\epsilon = \alpha M_b + \gamma(M_h - \frac{1}{\alpha} \sum_{\bar{L} \in \mathcal{L}^*_{\mathcal{H}}} \lambda_{\bar{L}}) - \sum_{\bar{L} \in \mathcal{L}_{\mathcal{B}}} \lambda_{\bar{L}}$. Since the ideal load decomposition satisfies Eqn. (1), ϵ is a positive constant. Select small δ and σ such that $\delta < \frac{\epsilon}{4}$, and $\sigma < \frac{\epsilon}{4}$. Then $\exists T > 0$ such that for $\forall t > T$,

$$\mathbb{E}\left[\sum_{m \in \mathcal{B}} S_m(t)\right] > \mathbb{E}\left[\sum_{m \in \mathcal{B}} A_m(t)\right] + \frac{\epsilon}{2}.$$

This contradicts the assumption that all queues in $\mathcal{B}$ are unstable. This completes the proof. ■

5 Heavy Traffic Optimality

In this section, we show that the proposed algorithm achieves queue length optimality in the heavy-traffic limit. We consider the two cases separately: locally overloaded and evenly loaded. The proof follows the

Lyapunov drift-based approach recently developed in [7]. However, as the "local-tasks first" policy excludes the use of a quadratic Lyapunov function, the main challenge is to prove the state-space collapse and derive a matching upper bound for the locally overloaded case. The main idea is to first show uniform boundedness for the helper queues, and analyze the Lyapunov drift for the beneficiary subsystem with a steady-state helper subsystem, and bound the amount of remote service received by helpers and the amount of helper traffic routed to beneficiaries. The proof for the evenly loaded case is considerably simpler, and we will only briefly state the results.

5.1 Locally Overloaded Traffic

With locally overloaded traffic, there exist a set of beneficiary queues and all helper queues have bounded loads. Under the ideal load decomposition, all task types in $\mathcal{L}^*_{\mathcal{H}}$, i.e., types that have at least one local server in helpers, are all routed to helper queues. Using the same notation as in 4, there are M_h helpers and M_b beneficiaries.

Consider the heavy-traffic regime where the total local load on helpers

$$\sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} \lambda_{\bar{L}} \equiv \Phi\alpha. \tag{9}$$

For any $\lambda \in \Lambda$, it is easy to see that $\sum_{\bar{L}\in\mathcal{L}} \lambda_{\bar{L}} < M_b\alpha + \Phi\alpha + \gamma(M_h - \Phi)$. We assume that

$$\sum_{\bar{L}\in\mathcal{L}} \lambda_{\bar{L}} = M_b\alpha + \Phi\alpha + \gamma(M_h - \Phi) - \epsilon, \tag{10}$$

where $\epsilon > 0$ characterizes the distance between the arrival rate vector and the capacity boundary.

Consider any arrival processes $\{A^{(\epsilon)}_{\bar{L}}(t), t \geq 0\}_{\bar{L}\in\mathcal{L}}$ with arrival rate vector $\lambda^{(\epsilon)}$ satisfying $\mathcal{B} \neq \emptyset$ and Eqs. (9)-(10). Note that the variance of $\{A^{(\epsilon)}_{\bar{L}}(t), t \geq 0\}_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}}$ is independent of ϵ. And the variance of the number of arrivals that are only local to beneficiaries is given by $Var(\sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} A^{(\epsilon)}_{\bar{L}}(t)) = (\sigma^{(\epsilon)}_b)^2$, which converges to σ^2_b as $\epsilon \to 0$.

The corresponding Markov chain $\{Z^{(\epsilon)}(t) = (Q^{(\epsilon)}(t), f^{(\epsilon)}(t)), t \geq 0\}$ has been shown to be positive recurrent. All theorems in this section will concern the *steady-state* queue lengths with $0 < \epsilon < \Phi\alpha + \gamma(M_h - \Phi)$. Due to space constraint, we defer all proofs to [1] and only highlight the important steps.

Theorem 2 *(***Helper queues***)*

$$\lim_{\epsilon\to 0^+} \epsilon\mathbb{E}\left[\sum_{m\in\mathcal{H}} Q^{(\epsilon)}_m(t)\right] = 0,$$

Theorem 2 states that the expected queue length of $Q^{(\epsilon)H}(t)$ is bounded and independent of ϵ. The theorem follows from the same Lyapunov function for Lemma 4, which, with the positive recurrence of the entire system, implies that all moments of $Q^{(\epsilon)H}(t)$ are bounded according to Lemma 1 in [7]. Therefore, we only need to consider the beneficiary queue lengths in the rest of 5.1.

5.1.1 Lower Bound

Consider a hypothetical single server system with arrival process $\{a^{(\epsilon)}(t), t \geq 0\}$ and service process $\{\beta^{(\epsilon)}(t), t \geq 0\}$, where

$$a^{(\epsilon)}(t) = \sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} A^{(\epsilon)}_{\bar{L}}(t), \quad \beta^{(\epsilon)}(t) = \sum_{i\in\mathcal{B}} X_i(t) + \sum_{j\in\mathcal{H}} Y_j(t).$$

Here $\{X_i(t)\}_{i\in\mathcal{B}}$ and $\{Y_j(t)\}_{j\in\mathcal{H}}$ are independent and each process is temporally i.i.d. For any $i \in \mathcal{B}$, let $X_i(t) \sim Bern(\alpha)$. And $\forall j \in \mathcal{H}$, $Y_j(t) \sim Bern(\gamma(1-\rho_j))$, where ρ_j is the proportion of time helper j spends on local tasks in steady state. Hence $\mathbb{E}\left[\sum_{j\in\mathcal{H}} Y_j(t)\right]$ represents the total amount of remote service provided by helpers. We denote $Var(\beta^{(\epsilon)}(t))$ by $(\nu^{(\epsilon)}_b)^2$, which converges to a constant ν^2_b as $\epsilon \to 0$. Let $\{\Phi(t)\}$ denote the corresponding queue-length process. Then in steady state, $\{\Phi(t)\}$ is stochastically smaller than the total

beneficiary queue-length process $\{\sum_{m\in\mathcal{B}} Q_m^{(\epsilon)}(t)\}$ of the original system. Using Lemma 4 in [7] to derive a lower bound on $\mathbb{E}[\Phi(t)]$, we obtain the following theorem.

Theorem 3 *(***Lower Bound***)*

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m^{(\epsilon)}(t)\right] \geq \frac{(\sigma_b^{(\epsilon)})^2 + (\nu_b^{(\epsilon)})^2 + \epsilon^2}{2\epsilon} - \frac{M}{2}.$$

Therefore, in the heavy traffic limit as $\epsilon \downarrow 0$*,*

$$\liminf_{\epsilon\to 0^+} \epsilon\mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m^{(\epsilon)}(t)\right] \geq \frac{\sigma_b^2 + \nu_b^2}{2}. \tag{11}$$

5.1.2 State Space Collapse

Throughout this section, we use notations with superscript $^{(B)}$ to denote the corresponding vectors for beneficiaries. Denote the queueing and working status process for beneficiaries as $\{Z^{(B)}(t) = (Q^{(B)}(t), f^{(B)}(t))\}$. Define

$$c_b = \frac{1}{\sqrt{M_b}}(\underbrace{1, 1, \cdots, 1}_{B}). \tag{12}$$

Then the parallel and perpendicular components of the queue length vector $Q^{(B)}$ with respect to c are given by:

$$Q_{||}^{(B)} = \langle c_b, Q^{(B)}\rangle c_b, \quad Q_{\perp}^{(B)} = Q^{(B)} - Q_{||}^{(B)}.$$

We will establish state-space collapse of $Q^{(B)}$ along the direction c_b, by showing that $Q_{\perp}^{(B)}$ is bounded and independent of the heavy-traffic parameter ϵ.

Remark. With the bounded moments of the helper queue lengths, the whole queue length vector Q collapses to the following direction c:

$$c = \frac{1}{\sqrt{M_b}}(\underbrace{0, 0, \cdots, 0}_{M_h}, \underbrace{1, 1, \cdots, 1}_{M_b}). \tag{13}$$

To establish state space collapse, we need to show that when the arrival rate vector λ satisfies the above heavy traffic assumption, there exists ideal load decomposition $\{\lambda^*_{\bar{L},m,n}\}$ satisfying the following propositions in addition to Lemma 3.

Lemma 9 *There exists a positive constant* λ_0 *not depending on* ϵ *such that:*

1 . $\forall m \in \mathcal{M}$,

$$\sum_{\bar{L}:m\in\bar{L}} \sum_{n:n\in\bar{L}} \frac{\lambda^*_{\bar{L},n,m}}{\alpha} + \sum_{\bar{L}:m\notin\bar{L}} \sum_{n:n\in\bar{L}} \frac{\lambda^*_{\bar{L},n,m}}{\gamma} = 1 - \epsilon_0;$$

2 . $\forall \bar{L} \in \mathcal{L}_{\mathcal{B}}, \forall m \in \bar{L}$, $\sum_{n\in\mathcal{H}} \lambda^*_{\bar{L},m,n} \geq \lambda_0$*;*

where $\epsilon_0 = \frac{\epsilon}{\alpha M_b + \gamma M_h}$.

Next, we consider the Lyapunov function

$$V(Z^{(B)}) = ||Q_{\perp}^{(B)}||.$$

By the extended version of Lemma 1 in [7], it is sufficient to show that the T-period drift of $V(Z^{(B)})$ is always finite and is negative for sufficient large V. The analysis follows similar steps in [14], and we highlight the steps that deal with shared traffic between helpers and beneficiaries, and remote service received by helpers. Full proof can be found in Appendix C.

We bound the drift of $V(Z^{(B)})$ as

$$\mathbb{E}\left[\Delta V(Z^{(B)})|Z^{(B)}(t_0)\right] \leq \frac{\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0)\right] + C}{||Q_{\perp}^{(B)}||}$$

where C is a constant and $G(t) = \langle Q^{(B)}(t), A^{(B)}(t) - S^{(B)}(t)\rangle - \langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle$.

Consider the following random variables

$$t_m^* = \min\{\tau : \tau \geq t_0, f_m(\tau^-) = -1\}, m \in \mathcal{M}, \quad (14)$$
$$t^* = \max_{m\in\mathcal{M}} t_m^*. \quad (15)$$

By the time slot t^*, all servers have been available at least once.

To bound $\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0)\right]$, we decompose the probability space by condition on t^*. Let $T = JK$. The key step is to bound $\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} G(t)|t^* < t_0 + K, Z^{(B)}(t_0)\right]$, which is different from that in [14]. We break $G(t)$ into four terms and obtain bound for each term.

For each $m \in \mathcal{B}$, define $\hat{A}_m(t) = \sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} A_{\bar{L},m}$, i.e., $\hat{A}$ excludes arrivals that are also local to helpers from beneficiaries. And define $\lambda_m^{*l} = \sum_{\bar{L}:m\in\mathcal{L}} \lambda_{\bar{L},m,m}^*$ and $\lambda_m^{*r} = \sum_{\bar{L}:m\in\mathcal{L}} \sum_{n:n\neq m} \lambda_{\bar{L},m,n}^*$ from the ideal load decomposition $\{\lambda_{\bar{L},m,n}^*\}$. Let $T_d = t_0 + T - t^*$, and $D(Q(t_0)) = M_b Q^{max}(t_0) - \sum_{m\in\mathcal{B}} Q_m(t_0)$, where $Q^{max}(t)$ is the maximum queue length at time t. We use F_i, $i \in \mathcal{N}$, to denote a positive constant not depending on ϵ. For ease of exposition, we temporarily omit the superscript $^{(B)}$ in the following argument.

Lemma 10 *(**Arrivals**)*

$$\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\Big(\langle Q(t), \hat{A}(t)\rangle - \langle Q(t), \lambda^{*l}\rangle \right.$$
$$-\langle Q^{max}(t)\sqrt{M_b}c_b, \lambda^{*r} - \lambda_0\sqrt{M_b}c_b\rangle$$
$$\left.-\langle Q(t), \lambda_0\sqrt{M_b}c_b\rangle\Big)\,|t^* < t_0 + K, Z^{(B)}(t_0)\right] \leq 0.$$

Lemma 11 *(**Local service**)*

$$\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\Big(\langle Q(t) - Q^{max}(t)\sqrt{M_b}c_b, \lambda_0\sqrt{M_b}c_b\rangle\right.$$
$$\left.+\langle Q(t), \lambda^{*l}\rangle - \langle Q(t), S^l(t)\rangle\Big)\,|t^* < t_0 + K, Z^{(B)}(t_0)\right]$$
$$\leq -T_d\left[\lambda_0 D(Q(t_0)) + \alpha\epsilon_0\sum_{m\in\mathcal{B}} Q_m(t_0)\right] + F_1$$

Lemma 12 *(**Remote service**)*

$$\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\Big(\langle Q^{max}(t)\sqrt{M_b}c_b, \lambda^{*r}\rangle - \langle Q(t), S^r(t)\rangle\right.$$
$$\left.-\langle c_b, Q(t)\rangle\langle c_b, \hat{A}(t) - S(t)\rangle\Big)\,|t^* < t_0 + K, Z^{(B)}(t_0)\right]$$
$$\leq T_d\left[\frac{\lambda_0}{4}D(Q(t_0)) + \alpha\epsilon_0\sum_{m\in\mathcal{B}} Q_m(t_0)\right] + F_2.$$

Lemma 13 *(**Extra arrivals to beneficiaries**)*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\Big(-\langle c_b, Q(t)\rangle\langle c_b, A(t)-\hat{A}(t)\rangle\right.\\
&\left.+\langle Q(t), A(t)-\hat{A}(t)\rangle\Big)\,|t^* < t_0+K, Z^{(B)}(t_0)\right]\\
\leq\ & T_d\frac{\lambda_0}{4}D(Q(t_0)) + F_3.
\end{aligned}$$

Observe that Lemma 10 allows us to distribute the remote service among Q^{max} and the remaining beneficiary queues. This ensures that the remote service provided by helpers will suffice even if a small proportion of it is devoted to helpers. The key observation is that the helper subsystem becomes decoupled from the beneficiaries as $Q^{(B)}$ becomes large, which allows us to bound the amount of remote service received by helpers and the arrivals local to a helper but joining a beneficiary queue, as in Lemma 12 and 13. Combining inequalities from Lemma 10-13, we can obtain the following.

Theorem 4 *(**State space collapse**) There exists a sequence of finite numbers $\{C_r : r \in \mathcal{N}\}$ such that for each positive integer r,*

$$\mathbb{E}\left[\left|\left|Q_\perp\right|\right|^r\right] \leq C_r,$$

where $Q_\perp$ is the component of Q perpendicular to the direction c.

5.1.3 Upper Bound

We use the Lyapunov drift-based moment bounding technique developed in [7]. The main difficulty arises from the fact that the total amount of service received at beneficiary queues, $\sum_{m\in\mathcal{B}} S_m(t)$, depends on the queueing process $Q(t)$: for any $m \in \mathcal{B}$, the local service provided by server m, $\{S_m^l(t)\}$ is neither i.i.d, nor independent of $Q_m(t)$; the amount of remote service $\mathcal{B}$ received, $\sum_{m\in\mathcal{B}} S_m^r(t)$, relies on the occurrence of system states that the maximum queue is among $\mathcal{B}$. In addition, the existence of tasks types shared among $\mathcal{H}$ and $\mathcal{B}$ makes total arrivals for $\mathcal{B}$, $\sum_{m\in\mathcal{B}} A_m(t)$, depend on $Q(t)$ as well. Hence we define the following ideal processes to decouple the dependence.

Ideal local service process $\hat{S}^l(t)$:

$$\hat{S}_m^l(t) = \begin{cases} X_m^l(t) & \text{if } m \in \mathcal{B} \\ S_m^l(t) & \text{if } m \in \mathcal{H} \end{cases}$$

where the processes $\{X_m^l(t), t \geq 0\}_{m\in\mathcal{B}}$ is coupled with $\{S_m^l(t), t \geq 0\}_{m\in\mathcal{B}}$ in the following way: If $\eta_m(t) = m$, $X_m^l(t) = S_m^l(t)$; if $\eta_m(t) \neq m$, $X_m^l(t) = 1$ when $R_m(t) = 1$, and $X_m^l(t) \sim Bern(\frac{\alpha-\gamma}{1-\gamma})$ when $R_m(t) = 0$. Hence $\forall m \in \mathcal{B}$, $\{X_m^l(t), t \geq 0\}$ is i.i.d. with $X_m^l(t) \sim Bern(\alpha)$.

Ideal remote service process $\hat{R}(t)$:

$$\hat{R}_m(t) = \begin{cases} 0 & \text{if } m \in \mathcal{B} \\ X_m^r(t) & \text{if } m \in \mathcal{H} \end{cases}$$

where the processes $\{X_m^r(t), t \geq 0\}_{m\in\mathcal{H}}$ is coupled with $\{R_m(t), t \geq 0\}_{m\in\mathcal{H}}$ in the following way: If $\eta_m(t) \neq m$, $X_m^r(t) = R_m(t)$; if $\eta_m(t) = m$, $X_m^l(t) \sim Bern(\gamma)$. Hence for $m \in \mathcal{H}$, $\{X_m^r(t), t \geq 0\}$ is i.i.d. with $X_m^r(t) \sim Bern(\gamma)$.

Ideal scheduling decision process $\hat{\eta}(t)$: For any $m \in \mathcal{B}$, $\hat{\eta}_m(t) = m$. For any $m \in \mathcal{H}$, $\hat{\eta}_m(t) = \eta_m(t)$ if $\eta_m(t) = m$; when $f_m(t^-) = -1$ and $Q_m(t) = 0$, $\hat{\eta}_m(t) = argmax_{n\in\mathcal{B}}\{Q_n(t)\}$. That is, idle helper server with empty local queue is scheduled to serve the maximum beneficiary queue under the *ideal scheduling*.

Ideal remote service received $\hat{S}^r(t)$:

$$\hat{S}_n^r(t) = \begin{cases} \sum\limits_{m\in\mathcal{H}} \hat{R}_m(t)\cdot I_{\{\hat{\eta}_m(t)=n\}} & \text{if } n \in \mathcal{B} \\ 0 & \text{if } n \in \mathcal{H} \end{cases}$$

And the *ideal departure* for queue m is given by $\hat{S}_m(t) = \hat{S}^l_m(t) + \hat{S}^r_m(t)$.

Ideal arrival process $\hat{A}(t)$: all shared task types are routed to helpers and distributed evenly among the local helper servers.

For $m \in \mathcal{B}$, let

$$\hat{A}_m(t) = \sum_{\bar{L} \in \mathcal{L}_\mathcal{B}: m \in \bar{L}} A_{\bar{L},m} = A_m(t) - \sum_{\bar{L} \notin \mathcal{L}_\mathcal{B}: m \in \bar{L}} A_{\bar{L},m}$$

For $m \in \mathcal{H}$, let

$$\hat{A}_m(t) = A_m(t) + \sum_{\bar{L} \notin \mathcal{L}_\mathcal{B}: m \in \bar{L}} \frac{\sum_{n \in \bar{L} \cap \mathcal{B}} A_{\bar{L},n}}{|\{k : k \in \bar{L} \cap \mathcal{H}\}|}$$

Then we can rewrite the queue dynamics as

$$Q(t+1) = Q(t) + \hat{A}(t) - \hat{S}(t) + \hat{U}(t),$$

where $\hat{U}(t) = \hat{S}(t) - S(t) + A(t) - \hat{A}(t) + U(t)$. We will use this queue dynamics to represent the Lapunov drift.

Lemma 14 *For the map task scheduling system, consider any arrival process with an arrival rate vector strictly within the capacity region. Suppose the queueing process is in steady state under the proposed algorithm. Then for any direction c,*

$$\begin{aligned} 2\mathbb{E}\left[\langle c, Q(t)\rangle \langle c, \hat{S}(t) - \hat{A}(t)\rangle\right] &= \mathbb{E}\left[\langle c, \hat{A}(t) - \hat{S}(t)\rangle^2\right] + \mathbb{E}\left[\langle c, \hat{U}(t)\rangle^2\right] && (16) \\ &\quad + 2\mathbb{E}\left[\langle c, Q(t) + \hat{A}(t) - \hat{S}(t)\rangle \langle c, \hat{U}(t)\rangle\right] && (17) \end{aligned}$$

We can obtain an upper bound on $\mathbb{E}[\langle c, Q(t)\rangle]$ by bounding each of the above terms.

Theorem 5 *(***Upper Bound***) For the map-scheduling system, consider an arrival process $\{A^{(\epsilon)}_{\bar{L}}(t), t \geq 0\}_{\bar{L} \in \mathcal{L}}$ parameterized by $\epsilon > 0$, with mean $\{\lambda^{(\epsilon)}_{\bar{L}}\}_{\bar{L} \in \mathcal{L}}$ satisfying $\mathcal{B} \neq \emptyset$ and Eqs. (9)- (10). Assuming that the Markov chain $\{(Q^{(\epsilon)}(t), f^{(\epsilon)}(t))\}$ is in steady sate under the proposed algorithm. Then for any t and any ϵ with*

$$0 < \epsilon < \Phi\alpha + \gamma(M_h - \Phi),$$

then the expected beneficiary queue lengths in steady-state can be upper bounded as

$$\mathbb{E}\left[\sum_{m \in \mathcal{B}} Q^{(\epsilon)}_m(t)\right] \leq \frac{(\sigma^{(\epsilon)}_b)^2 + (\nu^{(\epsilon)}_b)^2}{2\epsilon} + D^{(\epsilon)}, \tag{18}$$

where $D^{(\epsilon)} = o(\frac{1}{\epsilon})$, i.e., $\lim_{\epsilon \to 0^+} \epsilon D^{(\epsilon)} = 0$.

Therefore, in the heavy-traffic limit, the upper bound becomes,

$$\limsup_{\epsilon \to 0^+} \epsilon \mathbb{E}\left[\sum_{m \in \mathcal{B}} Q^{(\epsilon)}_m(t)\right] \leq \frac{\sigma_b^2 + \nu_b^2}{2} \tag{19}$$

This upper bound under heavy-traffic limit coincides with the lower bound (11), which establishes the first moment heavy-traffic optimality of the proposed algorithm.

5.2 Evenly Loaded Traffic

We consider the heavy-traffic regime where the arrival rate vector $\lambda \in \Lambda$ satisfies $\mathcal{H} = \mathcal{M}$. There exists $\epsilon > 0$ such that

$$\sum_{\bar{L} \in \mathcal{L}} \lambda_{\bar{L}} = M\alpha - \epsilon. \tag{20}$$

First we consider a special case that any two servers $(\hat{m}, m)$ in the system are connected, in the sense that there exists a sequence of servers $(\hat{m}, m_1, \cdots, m_k, m)$ such that for any consecutive servers in the sequence, there exists a task type $\bar{L} \in \mathcal{L}$ with $\lambda_{\bar{L}} > 0$ local to both servers. That is, all servers are connected by the arrivals.

Consider any arrival processes $\{A^{(\epsilon)}_{\bar{L}}(t), t \geq 0\}_{\bar{L} \in \mathcal{L}}$, parameterized by $\epsilon > 0$, with arrival rate vector $\lambda^{(\epsilon)}$ connecting all servers and satisfying Eqn. (20). Denote the variance of the arrival process as $(\sigma^{(\epsilon)})^2$. For all queue lengths in steady state, and $0 < \epsilon < M\alpha$, we obtain the three theorems analogous to the locally overloaded case. We defer full proof to Appendix D.

Theorem 6 *(**Lower bound**)*

$$\mathbb{E}\left[\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)\right] \geq \frac{(\sigma^{(\epsilon)})^2 + \nu^2 + \epsilon^2}{2\epsilon} - \frac{M}{2},$$

where ν^2 is the variance of the service process. Therefore, in the heavy traffic limit, we have

$$\liminf_{\epsilon \to 0^+} \epsilon \mathbb{E}\left[\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)\right] \geq \frac{\sigma^2 + \nu^2}{2}.$$

Theorem 7 *(**State space collapse**) Let*

$$c_u = \frac{1}{\sqrt{M}}(\underbrace{1, 1, \cdots, 1}_{M}),$$

there exists a sequence of finite numbers $\{C'_r : r \in \mathcal{N}\}$ such that for each positive integer r,

$$\mathbb{E}\left[\left|\left|Q_\perp\right|\right|^r\right] \leq C'_r,$$

where $Q_\perp$ is the component of Q perpendicular to c_u.

Theorem 8 *(**Upper bound**)*

$$\mathbb{E}\left[\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)\right] \leq \frac{(\sigma^{(\epsilon)})^2 + \nu^2}{2\epsilon} + D^{(\epsilon)}_u,$$

where $D^{(\epsilon)}_u = o(\frac{1}{\epsilon})$, i.e., $\lim_{\epsilon \to 0^+} \epsilon D^{(\epsilon)}_u = 0$. Therefore, in the heavy-traffic limit, we have

$$\limsup_{\epsilon \to 0^+} \epsilon \mathbb{E}\left[\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)\right] \leq \frac{\sigma^2 + \nu^2}{2}$$

The heavy-traffic optimality of the proposed algorithm follows by the coincidence of lower and upper bounds.

Remark: If the arrival rate vector λ makes some server pairs $(\hat{m}, m)$ isolated from each other, we can always decompose servers into disjoint groups, such that servers within each group are connected, while isolated from servers outside. For each connected group $\mathcal{H}_i$, we can obtain the corresponding upper bound on $\mathbb{E}\left[\sum_{m \in \mathcal{H}_i} Q^{(\epsilon)}(t)\right]$ as Theorem 8. Together they give an upper bound on $\mathbb{E}\left[\sum_m Q^{(\epsilon)}(t)\right]$, which coincides with the lower bound in the heavy traffic limit.

6 Evaluation

We compare the mean task completion time of the proposed algorithm against the JSQ-MaxWeight algorithm. We simulate a system of $M = 1000$ servers with local service rate $\alpha = 1$ and remote service rate $\gamma = 0.5$, which corresponds to a mean slowdown of 2, consistent to the measurements in [15]. At each task arrival, a set of three servers are chosen to be its local servers according to the distribution of the requested data. We consider two cases:

1. Evenly loaded traffic. The set of three servers are sampled uniformly randomly from all M servers. This simulates the scenario when the data requested by the incoming traffic are evenly distributed on all servers.

2. Locally overloaded traffic. At each task arrival, with probability σ, the task samples a set of three servers uniformly randomly from a subset of N servers, and with probability $1 - \sigma$, it samples from the remaining $M - N$ servers. When σ is large enough, the N servers form a hot-spot in the system. We report the results for $\sigma = 0.8$ and $N/M = 0.5$.

(a) Evenly loaded traffic

(b) Locally overloaded traffic

Figure 3: Average task completion time.

Figure 3 compares JSQ-MaxWeight and our proposed algorithm. We separate the figures into two load regions and use different vertical scales to make the comparison more visible. For both evenly loaded and locally overloaded traffic, the proposed algorithm has similar performance as JSQ-MaxWeight at low load, and achieves up to 4-fold improvement over medium to high load.

7 Conclusion

We proposed a near-data scheduling algorithm and proved its throughput and heavy-traffic optimality. The algorithm is also shown to have superior performance in simulation.

References

[1] Priority algorithm for near-data scheduling: Throughput and heavy-traffic optimality. *https://www.dropbox.com/s/tpscj222z3sqd8r/near-data-long.pdf*.

[2] C. Abad, Y. Lu, and R. Campbell. Dare: Adaptive data replication for efficient cluster scheduling. In *IEEE Cluster*, 2011.

[3] G. Ananthanarayanan, S. Agarwal, S. Kandula, A. Greenberg, I. Stoica, D. Harlan, and E. Harris. Scarlett: Coping with skewed popularity content in MapReduce clusters. In *Proc. Eur. Conf. Comput. Syst. (EuroSys)*, 2011.

[4] Apache Hadoop, June 2011.

[5] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Annals of Applied Probability*, 11, 2001.

[6] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *Proc. of OSDI*, 2004.

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

[8] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14(3):pp. 502–525, 1982.

[9] J. M. Harrison. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *Annals of Applied Probability*, 1998.

[10] J. M. Harrison and M. J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Syst. Theory Appl.*, 33(4), Apr. 1999.

[11] M. Isard, V. Prabhakaran, J. Currey, U. Wieder, K. Talwar, and A. Goldberg. Quincy: Fair scheduling for distributed computing clusters. In *Proc. of SOSP*, 2009.

[12] D. Y. M. Squillante, C. Xia and L. Zhang. Threshold-based priority policies for parallel-server systems with affinity scheduling. In *Proceedings of the IEEE American Control Conference*, 2001.

[13] A. Mandelbaum and A. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, 52, 2004.

[14] W. Wang, K. Zhu, L. Ying, J. Tan, and L. Zhang. A throughput optimal algorithm for map task scheduling in mapreduce with data locality. *SIGMETRICS Perform. Eval. Rev.*, 40(4), Apr. 2013.

[15] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica. Delay scheduling: A simple technique for achieving locality and fairness in cluster scheduling. In *Proc. Eur. Conf. Comput. Syst. (EuroSys)*, 2010.

Appendix A

We first prove Lemma 2 and then use the special decomposition from Lemma 2 to prove Lemma 3.

7.1 Proof of Lemma 2.

Proof.

Given $\lambda \in \Lambda$, there exists a decomposition $\left\{\lambda_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$ satisfying Eqn. (1). We apply an iterative approach to construct $\left\{\tilde{\lambda}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$ from $\left\{\lambda_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$.

Let $\left\{\lambda^{(l)}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}, l \geq 0$ denote the decomposition after the l-th processing iteration. Let $\mathcal{M}_h^{(l)}$ and $\mathcal{M}_b^{(l)}$ denote the corresponding locally underloaded queues and locally overloaded queues, respectively. And $\mathcal{L}_b^{(l)}$ is used to denote the set of tasks types that are only local to $\mathcal{M}_b^{(l)}$, $\mathcal{L}_s^{(l)}$ the set of task types shared among $\mathcal{M}_h^{(l)}$ and $\mathcal{M}_b^{(l)}$. Initialize $\left\{\lambda^{(0)}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$ as the given decomposition $\left\{\lambda_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$. If there exists $\bar{L} \in \mathcal{L}_s^{(l)}$ such that $\lambda^{(l)}_{\bar{L},n_1,m} > 0$ for some $n_1 \in \mathcal{M}_b^{(l)}$, $m \in \mathcal{M}$, $\left\{\lambda^{(l+1)}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$ will be updated as follows. Otherwise, the iterative processing ends up with $\left\{\tilde{\lambda}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\} = \left\{\lambda^{(l)}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$.

The $l+1$-th iteration will re-distribute $\lambda^{(l)}_{\bar{L},n_1,m}$ from temporal overloaded queue n_1 to temporal underloaded queue n_2 which is also local to $\bar{L}$. The amount of removal depends on the . Consider the following four cases.

Case (i): $\lambda^{(l)}_{n_1} - \lambda^{(l)}_{\bar{L},n_1,m} \geq \alpha$, $\lambda^{(l)}_{n_2} + \lambda^{(l)}_{\bar{L},n_1,m} < \alpha$

Let

$$\begin{aligned}
\lambda^{(l+1)}_{\bar{L},n_1,m} &= 0, \\
\lambda^{(l+1)}_{\bar{L},n_2,m} &= \lambda^{(l)}_{\bar{L},n_2,m} + \lambda^{(l)}_{\bar{L},n_1,m},
\end{aligned}$$

All other components $\lambda^{(l+1)}_{\bar{L},n,m'}$ remains the same as previous iteration. Hence after $l+1$-th iteration, n_1 is still overloaded, while n_1 is still underloaded. Observe that for $\forall m' \in \mathcal{M}, m' \neq m$ Eqn. (1) still holds under $\left\{\lambda^{(l+1)}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$. Meanwhile, for m, the total amount of remote service provided remains the same as the l-th iteration, which ensures the correctness of Eqn. (1) for m.

Case (ii): $\lambda^{(l)}_{n_1} - \lambda^{(l)}_{\bar{L},n_1,m} < \alpha$, $\lambda^{(l)}_{n_2} + \lambda^{(l)}_{\bar{L},n_1,m} < \alpha$

Update $\left\{\lambda^{(l+1)}_{\bar{L},n,m} : \bar{L} \in \mathcal{L}, n \in \bar{L}, m \in \mathcal{M}\right\}$ as case (i). Thus the $l+1$-th iteration redistributes the shared load between n_1 and n_2, making n_1 underloaded as n_2. Again it is obvious that the rate decomposition after the $l+1$-th iteration satisfies Eqn. (1).

Case (iii):: $\lambda^{(l)}_{n_1} - \lambda^{(l)}_{\bar{L},n_1,m} \geq \alpha$, $\lambda^{(l)}_{n_2} + \lambda^{(l)}_{\bar{L},n_1,m} \geq \alpha$

Let

$$\begin{aligned}
\delta &= \min\{\lambda^{(l)}_{\bar{L},n_1,m}, \frac{\lambda^{(l)-\lambda^{(l)}_{n_2}}_{n_1}}{2}\}, \\
\lambda^{(l+1)}_{\bar{L},n_1,m} &= \lambda^{(l)}_{\bar{L},n_1,m} - \delta, \\
\lambda^{(l+1)}_{\bar{L},n_2,m} &= \lambda^{(l)}_{\bar{L},n_2,m} + \delta.
\end{aligned}$$

Keep all other components $\lambda^{(l+1)}_{\bar{L},n,m'}$ as previous iteration. Observe that such a removal makes n_2 be overloaded as n_1 and minimizes the local load difference between n_1 and n_2. Again Eqn. (1) holds for $\forall m \in \mathcal{M}$ after $l+1$-th iteration.

Case (iv):: $\lambda_{n_1}^{(l)} - \lambda_{\bar{L},n_1,m}^{(l)} < \alpha$, $\lambda_{n_2}^{(l)} + \lambda_{\bar{L},n_1,m}^{(l)} \geq \alpha$

Follow the same update procedure as case (iii).

If $\lambda_{n_1} + \lambda_{n_2} \geq 2\alpha$, the update turns n_2 into an overloaded queue like case (iii).

If $\lambda_{n_1} + \lambda_{n_2} < 2\alpha$, $\delta = \frac{\lambda_{n_1}^{(l)} - \lambda_{n_2}^{(l)}}{2} < \lambda_{\bar{L},n_1,m}^{(l)}$. Thus $\lambda_{n_1}^{(l+1)} < \alpha$ and $\lambda_{n_2}^{(l+1)} < \alpha$, i.e., both n_1 and n_2 are underloaded after the $l+1$-th iteration.

It is easy to verify that $\rho(\lambda^{(l+1)}) < \rho(\lambda^{(l)})$. Note that any arrival exchange among underloaded queues only or among overloaded queues only won't decrease the total virtual load. When all shared type tasks join underloaded queue, the corresponding total load is minimized as there is no possible arrival exchange that will reduce the total load. This ensures the convergence of the above iterative approach. And the decomposition after the algorithm stops gives the desired decomposition. This completes the proof for Lemma 2. ■

7.2 Proof of Lemma 3.

Proof.

We construct the ideal decomposition iteratively from $\left\{\tilde{\lambda}_{\bar{L},n,m}\right\}$ given in Lemma 2 by exchanging remote load for local load in each buffer. First consider exchange for $\mathcal{H}$. Define

$$\psi_h(\tilde{\lambda}) = \sum_{n\in\mathcal{H}} \sum_{m:m\neq n} \nu_{n,m}$$

as the total amount of remote service *received* by $\mathcal{H}$ with the decomposition $\{\tilde{\lambda}_{\bar{L},n,m}\}$. Whenever there exists some remote sub-queue of queue $n \in \mathcal{H}$ with non-zero load, for instance $\nu_{n,m} > 0 (m \neq n)$, we move $\nu_{n,m}$ to local sub-queue $\nu_{n,n}$, and move $\min \nu_{n,m}, \sum_{k\neq n} \nu_{k,n}$ amount of load from remote sub-queues at the n-th column to the corresponding sub-queues at the m-th column. It is easy to see that such exchange maintains validity of Eqn. (1), and reduces $\psi_h(\tilde{\lambda})$ by $\nu_{n,m}$ at least. The iterative process ends when no remote load left in the buffers of $\mathcal{H}$, i.e., Eqn. (3) satisfied.

Then we start load exchange for $\mathcal{B}$. Define

$$\phi_b(\tilde{\lambda}) = \sum_{m_1\in\mathcal{B}} \sum_{\substack{m_2\in B \\ m_2\neq m_1}} \nu_{m_2,m_1}$$

as the total amount of remote service *offered* by $\mathcal{B}$ with the updated decomposition $\{\tilde{\lambda}_{\bar{L},n,m}\}$ satisfying Eqn. (3). If some beneficiary buffer m_1 offers remote service, i.e., $\exists \nu_{m_2,m_1} > 0$ where $m_2 \in B$ and $m_2 \neq m_1$, we can reassign the offered remote service to servers from which queue m_1 receives remote service as follows: move ν_{m_2,m_1} amount of load from some remote sub-queues $\nu_{m_1,k}$ within Q_{m_1} to the local sub-queue ν_{m_1,m_1}, and replace same amount of load from ν_{m_2,m_1} at the corresponding sub-queues $\nu_{m_2,k}$ within Q_{m_2}. Note that such removal won't increase remote service offered by other beneficiaries. Hence ψ_b is reduced by ν_{m_2,m_1} at least. Again Eqn. (1) holds for all $m \in \mathcal{M}$ with such removal. Similarly, the iterative process ends when $\psi_b = 0$, i.e., Eqn (3) satisfied. ■

Appendix B

For ease of exposition, in this section, we temporally omit the superscript $^{(H)}$.

7.3 Proof of Lemma 5.

Under the proposed algorithm, every arriving task at the beginning of each time slot will join its shortest local queue. For $\forall \bar{L} \in \mathcal{L}^*_{\mathcal{H}}$, define $Q^*_{\bar{L}}(t) = \min_{m \in \bar{L} \cup \mathcal{H}} \{Q_m(t)\}$. For any task type that is only local to $\mathcal{H}$, i.e., $\bar{L} \in \mathcal{L}_{\mathcal{H}}$, it will be routed to queue $Q^*_{\bar{L}}(t)$ at the beginning of time slot t. Meanwhile, tasks shared among $\mathcal{B}$ and $\mathcal{H}$ might join $Q^*_{\bar{L}}(t)$ or the shortest local queue in $\mathcal{B}$.

$$
\begin{aligned}
\mathbb{E}\left[\langle Q^{(H)}(t), A^{(H)}(t)\rangle | Z(t)\right] &= \mathbb{E}\left[\sum_{m\in\mathcal{H}} Q_m(t) A_m(t) | Z(t)\right] \\
&= \mathbb{E}\left[\sum_{m\in\mathcal{H}} \sum_{\bar{L}: m\in\bar{L}} Q_m(t) A_{\bar{L},m}(t) | Z(t)\right] \\
&= \mathbb{E}\left[\sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} \sum_{m\in\bar{L}\cap\mathcal{H}} Q_m(t) A_{\bar{L},m}(t) | Z(t)\right] \\
&\le \sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} \mathbb{E}\left[Q^*_{\bar{L}}(t) A_{\bar{L}}(t) | Z(t)\right] \\
&= \sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} Q^*_{\bar{L}}(t) \lambda_{\bar{L}} \\
&= \sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} Q^*_{\bar{L}}(t) \sum_{m\in\bar{L}\cap\mathcal{H}} \sum_{n=1}^{M} \lambda^*_{\bar{L},m,n} \qquad (21) \\
&= \sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} \sum_{m\in\bar{L}\cap\mathcal{H}} \lambda^*_{\bar{L},m,m} Q^*_{\bar{L}}(t) \qquad (22) \\
&\le \sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} \sum_{m\in\bar{L}\cap\mathcal{H}} \lambda^*_{\bar{L},m,m} Q_m(t) \\
&= \sum_{m\in\mathcal{H}} \lambda^*_m Q_m(t) \\
&= \mathbb{E}\left[\langle Q^{(H)}(t), \lambda^{*(H)}\rangle | Z(t_0)\right], \qquad (23)
\end{aligned}
$$

where Eq.(21) and (22) follow from the ideal decomposition. ■

Note that

$$
\begin{aligned}
& \mathbb{E}\left[\langle Q^{(H)}(t), A^{(H)}(t)\rangle - \langle Q^{(H)}(t), \lambda^{*(H)}\rangle | Z(t_0)\right] \\
= \; & \mathbb{E}\left[\mathbb{E}\left[\langle Q^{(H)}(t), A^{(H)}(t)\rangle - \langle Q^{(H)}(t), \lambda^{*(H)}\rangle | Z(t)\right] | Z(t_0)\right] \\
\le \; & 0
\end{aligned}
$$

This completes the proof for Lemma 5.

7.4 Proof of Lemma 6.

Proof. Consider the following random variables

$$
\begin{aligned}
& t^*_m = \min\{\tau : \tau \ge t_0, f_m(\tau) = -1\}, m \in \mathcal{M}, \\
& t^* = \max_{1\le m\le M} t^*_m.
\end{aligned}
$$

So server m makes the first scheduling decision after t_0 at t_m^*. And t^* is the first time slot that every server has made at least one scheduling decision after t_0. Let $T = JK$, where $J > 0$ and $K > 0$. We then decompose the probability space into two parts by using t^*: $A_1 = \{t^* > t_0 + K | Z(t_0)\}$ and $A_2 = \{t^* \leq t_0 + K | Z(t_0)\}$. Let B_i denote the expectation term that is further conditioned on A_i, $i = 1, 2$, i.e.,

$$B_i = \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\langle Q^{(H)}(t), \lambda^{*(H)}\rangle - \langle Q^{(H)}(t), S^{(H)}(t)\rangle\right) | Z(t_0) = Z, A_i\right]$$

Thus the expectation term in (4) is broken down into two parts: $B_1\mathbb{P}[A_1]$ and $B_2\mathbb{P}[A_2]$. That is,

$$\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\langle Q^{(H)}(t), \lambda^{*(H)}\rangle - \langle Q^{(H)}(t), S^{(H)}(t)\rangle\right) | Z(t_0) = Z\right] = B_1\mathbb{P}[A_1] + B_2\mathbb{P}[A_2]$$

Note that we assume bounded arrivals and departures for the system. Without loss of generosity, assume that there exists $C_A > 0$ and $C_D > 0$ such that for $\forall t_1, t \in [t_0 t_0 + T]$, where $t_1 < t$,

$$Q_m(t) \leq Q_m(t_1) + (t - t_1)C_A$$
$$Q_m(t) \geq Q_m(t_1) - (t - t_1)C_D.$$

As $\lambda \in \Lambda$, there exists $\vartheta > 0$ such that for $\forall m \in \mathcal{M}$, the decomposition satisfies

$$\sum_{\bar{L}: m \in \bar{L}} \frac{{\lambda^*}_{\bar{L},m,m}}{\alpha} + \sum_{\bar{L}: m \notin \bar{L}} \sum_{n: n \in \bar{L}} \frac{{\lambda^*}_{\bar{L},n,m}}{\gamma} \leq \frac{1}{1+\vartheta}. \tag{24}$$

Thus $\sum_{\bar{L}: m \in \bar{L}} {\lambda^*}_{\bar{L},m,m} < \alpha$. Together with the bounded difference between $Q_m(t_0)$ and $Q_m(t)$, we can bound B_1 as

$$\begin{aligned}
B_1 &= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\langle Q^{(H)}(t), \lambda^{*(H)}\rangle - \langle Q^{(H)}(t), S^{(H)}(t)\rangle\right) | Z(t_0) = Z, A_1\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\sum_{m \in \mathcal{H}} Q_m(t)\lambda_m^*\right) | Z(t_0) = Z, A_1\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \left(\sum_{m \in \mathcal{H}} Q_m(t) \sum_{\bar{L}: m \in \bar{L}} \lambda^*_{\bar{L},m,m}\right) | Z(t_0) = Z, A_1\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \sum_{m \in \mathcal{H}} \alpha Q_m(t) | Z(t_0) = Z, A_1\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \sum_{m \in \mathcal{H}} \alpha(Q_m(t_0) + (t - t_0)C_A) | Z(t_0) = Z, A_1\right] \\
&\leq \alpha T \sum_{m \in \mathcal{H}} Q_m(t_0) + \alpha T^2 M C_A
\end{aligned} \tag{25}$$

To bound the term B_2, we divide the summation into two parts: from $t = t_0$ to $t = t^*$ and from $t = t^* + 1$ to $t = t_0 + T - 1$. The first part can be bounded in a similar way as term B_1.

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{t=t_0}^{t^*}\left(\langle Q^{(H)}(t),\lambda^{*(H)}\rangle - \langle Q^{(H)}(t),S^{(H)}(t)\rangle\right)|Z(t_0)=Z,t^*\le t_0+K\right] \\
\le\ & \mathbb{E}\left[\sum_{t=t_0}^{t^*}\left(\sum_{m\in\mathcal{H}}Q_m(t)\sum_{\bar{L}:m\in\bar{L}}\lambda^*_{\bar{L},m,m}\right)|Z(t_0)=Z,t^*\le t_0+K\right] \\
\le\ & \mathbb{E}\left[\sum_{t=t_0}^{t^*}\sum_{m\in\mathcal{H}}\alpha Q_m(t)|Z(t_0)=Z,t^*\le t_0+K\right] \\
\le\ & \mathbb{E}\left[\sum_{t=t_0}^{t^*}\sum_{m\in\mathcal{H}}\alpha(Q_m(t_0)+(t-t_0)C_A)|Z(t_0)=Z,t^*\le t_0+K\right] \\
\le\ & \alpha(t^*-t_0)\sum_{m\in\mathcal{H}}Q_m(t_0)+\alpha(t^*-t_0)TMC_A. \qquad (26)
\end{aligned}
$$

For the second part, we first let it condition on t^*, and then further conditioned on $Z(t)$. Note that $\forall t\in(t^*,t_0+T)$ and $m\in\mathcal{H}$,

$$
\begin{aligned}
& \mathbb{E}\left[\left(\langle Q^{(H)}(t),\lambda^{*(H)}\rangle - \langle Q^{(H)}(t),S^{(H)}(t)\rangle\right)|Z(t_0)=Z,t^*\le t_0+K\right] \\
=\ & \mathbb{E}\left[\sum_{m\in\mathcal{H}}(Q_m(t)\lambda^*_m - Q_m(t)S^l_m(t)-Q_m(t)S^r_m(t))|Z(t_0)=Z,t^*\le t_0+K\right] \\
\le\ & \mathbb{E}\left[\sum_{m\in\mathcal{H}}(Q_m(t)\lambda^*_m - Q_m(t)S^l_m(t))|Z(t_0)=Z,t^*\le t_0+K\right] \\
=\ & \sum_{m\in\mathcal{H}}\mathbb{E}\left[Q_m(t)\left(\sum_{\bar{L}:m\in\bar{L}}\lambda^*_{\bar{L},m,m}-\alpha I_{\{\eta_m(t)=m\}}\right)|Z(t_0),t^*\right] \\
=\ & \mathbb{E}\left[\mathbb{E}\left[Q_m(t)\left(\sum_{\bar{L}:m\in\bar{L}}\lambda^*_{\bar{L},m,m}-\alpha I_{\{\eta_m(t)=m\}}\right)|Z(t_0),Z(t),t^*\right]|Z(t_0),t^*\right]. \qquad (27)
\end{aligned}
$$

As $t>t^*$, given $Z(t)$, $\eta(t)$ is independent of all the previous system state. Thus we have

$$
\begin{aligned}
& \mathbb{E}\left[Q_m(t)\left(\sum_{\bar{L}:m\in\bar{L}}\lambda^*_{\bar{L},m,m}-\alpha I_{\{\eta_m(t)=m\}}\right)||Z(t_0),Z(t),t^*\right] \\
=\ & Q_m(t)\sum_{\bar{L}:m\in\bar{L}}\lambda^*_{\bar{L},m,m}-Q_m(t)\alpha\mathbb{E}\left[I_{\{\eta_m(t)=m\}}|Z(t)\right]. \qquad (28)
\end{aligned}
$$

Note that $\eta_m(t)$ is conditionally independent of $Q(t)$ given $Z(t)$.

Consider the following random variables

$$\tau^t_m := \max\{\tau:\tau\le t, f_m(\tau)=-1\}, m\in\mathcal{M}.$$

Hence τ^t_m is the last moment before t at which server m makes a scheduling decision. Therefore the status of server m remains the same from time τ^t_m to t, i.e., $\eta_m(t)=\eta_m(\tau^t_m)$. Observe that $\eta_m(\tau^t_m)=m$ if $Q_m(\tau^t_m)>0$, as local tasks will be scheduled first. If $Q_m(\tau^t_m)=0$, $\eta_m(\tau^t_m)\neq m$. Thus,

$$Q_m(\tau^t_m)\mathbb{E}\left[I_{\{\eta_m(t)=m\}}|Z(\tau^t_m)\right]=Q_m(\tau^t_m). \qquad (29)$$

Using the bounded difference between $Q_m(t_0)$, $Q_m(\tau_m^t)$ and $Q_m(t)$, we have

$$\begin{aligned}
& Q_m(t)\mathbb{E}\left[I_{\{\eta_m(t)=m\}}|Z(t)\right] \\
= \;& \mathbb{E}\left[Q_m(t)\mathbb{E}\left[I_{\{\eta_m(t)=m\}}|Z(\tau_m^t)\right]|Z(t)\right] \\
\geq \;& \mathbb{E}\left[(Q_m(\tau_m^t)-TC_D)\mathbb{E}\left[I_{\{\eta_m(t)=m\}}|Z(\tau_m^t)\right]|Z(t)\right] \\
\geq \;& \mathbb{E}\left[Q_m(\tau_m^t)\mathbb{E}\left[I_{\{\eta_m(t)=m\}}|Z(\tau_m^t)\right]|Z(t)\right]-TC_D \\
= \;& \mathbb{E}\left[Q_m(\tau_m^t)|Z(t)\right]-TC_D \\
\geq \;& \mathbb{E}\left[(Q_m(t)-TC_A|Z(t)\right]-TC_D \\
= \;& \mathbb{E}\left[Q_m(t)|Z(t)\right]-TC_A-TC_D
\end{aligned} \tag{30}$$

As $\sum_{\bar{L}:m\in\bar{L}}\lambda^*{}_{\bar{L},m,m}\leq\frac{\alpha}{1+\vartheta}$, together with Eqn. (30), we can bound the term (28) by

$$\sum_{m\in\mathcal{H}}Q_m(t)(\frac{\alpha}{1+\vartheta}-\alpha)+\alpha T(C_A+C_D)\leq-\frac{\alpha\vartheta}{1+\vartheta}\sum_{m\in\mathcal{H}}Q_m(t_0)+\alpha T(C_A+C_D). \tag{31}$$

Hence we can obtain the bound for the summation from $t=t^*+1$ to $t=t_0+T-1$:

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\left(\sum_{m\in\mathcal{H}}Q_m(t)\sum_{\bar{L}:m\in\bar{L}}\lambda^*_{\bar{L},m,m}-\sum_{m\in\mathcal{H}}Q_m(t)\alpha I_{\{\eta_m(t)=m\}}\right)|Z(t_0)=Z,t^*\leq t_0+K\right] \\
\leq \;& -(t_0+T-t^*)\frac{\alpha\vartheta}{1+\vartheta}\sum_{m\in\mathcal{H}}Q_m(t_0)+C.
\end{aligned} \tag{32}$$

Now we can bound the term B_2 by combining the bounds for two summations in Eqn. (26) and (32):

$$B_2\leq K\alpha(1-\frac{(J-1)\vartheta}{1+\vartheta})\sum_{m\in\mathcal{H}}Q_m(t_0)+C.$$

Let $\zeta=\frac{\vartheta}{1+\vartheta}$, and $J_0=1+\frac{1}{\zeta}$. Pick any $J>J_0$, then $K\alpha(1-\frac{(J-1)\vartheta}{1+\vartheta})<0$.
From Lemma **??**, we have

$$\begin{aligned}
&\mathbb{P}(t^*\leq t_0+K|Z(t_0)\geq(1-(1-\gamma)^K)^M \\
&\mathbb{P}(t^*\geq t_0+K|Z(t_0)\leq 1-(1-(1-\gamma)^K)^M.
\end{aligned}$$

Applying the bound for B_1 and B_2, together with the above two inequalities, we can obtain

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\left(\langle Q^{(H)}(t),\lambda^{*(H)}\rangle-\langle Q^{(H)}(t),S^{(H)}(t)\rangle\right)|Z(t_0)=Z\right] \\
\leq \;& \alpha T(\sum_{m\in\mathcal{H}}Q_m(t_0))(1-(1-(1-\gamma)^K)^M)+K\alpha(1-\frac{(J-1)\vartheta}{1+\vartheta})\sum_{m\in\mathcal{H}}Q_m(t_0)(1-(1-\gamma)^K)^M+C \\
= \;& \alpha T\left(H_1(K)+\frac{1}{J}(1+\zeta)H_2(K)-\zeta H_2(K)\right)\sum_{m\in\mathcal{H}}Q_m(t_0)+C,
\end{aligned} \tag{33}$$

where $H_1(K)=1-(1-(1-\gamma)^K)^M$, $H_2(K)=(1-(1-\gamma)^K)^M$, and C is a constant independent of $Z(t_0)$.

Next we select K and J to make the coefficient of $\sum_{m\in\mathcal{H}}Q_m(t_0)$ negative. First pick any $\theta\in(0,\zeta)$. Note that $H_1(K)\to 0$ as $K\to\infty$, there exists K_1 such that $\forall K>K_1$, $H_1(K)\leq\frac{\zeta-\theta}{3}$. Since $H_2(K)\to 1$ as $K\to\infty$, there exists K_2 such that $\forall K>K_2$, $H_2(K)\geq 1-\frac{\zeta-\theta}{3\zeta}$. Let $J_2=\frac{3(1+\zeta)}{\zeta-\theta}$, then $\forall J>J_2$, $\frac{1}{J}(1+\zeta)H_2(K)<\frac{\zeta-\theta}{3}$. Thus, by picking $K>\max\{K_1,K_2\}$ and $J>\max\{J_1,J2\}$, we obtain

$$
\begin{aligned}
& H_1(K) + \frac{1}{J}(1+\zeta)H_2(K) - \zeta H_2(K) \\
\leq \quad & \frac{\zeta-\theta}{3} + \frac{\zeta-\theta}{3} - \zeta(1 - \frac{\zeta-\theta}{3\zeta}) \\
= \quad & -\theta.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} \Big(\langle Q^{(H)}(t), \lambda^{*(H)}\rangle - \langle Q^{(H)}(t), S^{(H)}(t)\rangle\Big) \,|Z(t_0) = Z\right] \\
\leq \quad & -\theta\alpha T|Q^{(H)}(t_0)|_1 + C, \\
= \quad & -\theta_1|Q^{(H)}(t_0)|_1 + C.
\end{aligned}
$$

where $\theta_1 = \theta\alpha T$ and C are independent of $Z(t_0)$ This proves Lemma 6. ■

Appendix C

7.5 Proof of Lemma 9

Lemma 15 *There exists a decomposition* $\left\{\lambda^*{}_{\bar{L},n,m}\right\}$ *satisfying Lemma 3 and the following conditions*

1 $\forall m \in \mathcal{B}$,

$$\sum_{\bar{L}:m\in\bar{L}}\sum_{n:n\in\bar{L}}\frac{\lambda^*_{\bar{L},n,m}}{\alpha}+\sum_{\bar{L}:m\notin\bar{L}}\sum_{n:n\in\bar{L}}\frac{\lambda^*_{\bar{L},n,m}}{\gamma}=1-\epsilon_b;$$

2 $\forall m \in \mathcal{H}$,

$$\sum_{\bar{L}:m\in\bar{L}}\sum_{n:n\in\bar{L}}\frac{\lambda^*_{\bar{L},n,m}}{\alpha}+\sum_{\bar{L}:m\notin\bar{L}}\sum_{n:n\in\bar{L}}\frac{\lambda^*_{\bar{L},n,m}}{\gamma}<1;$$

3 $\forall m \in \mathcal{B}$, $\exists \bar{L} \in \mathcal{L}_{\mathcal{B}}$, *s.t.* $\sum_{n\in\mathcal{H}}\lambda^*_{\bar{L},m,n}\geq\lambda_0$;

where ϵ_b *is a constant satisfying* $0<\epsilon_b<\frac{\epsilon}{\alpha M_b}$, *and* λ_0 *is a positive constant not depending on* ϵ.

Proof.

We will prove this lemma by constructing a decomposition that meets the three conditions.

Consider a decomposition $\left\{\lambda_{\bar{L},n,m}\right\}$ that satisfies Lemma 3. We fix the decomposition of $\mathcal{L}_{\mathcal{H}}$ over $\mathcal{H}$ and define

$$\rho^h_{max}=\max\lim_{m\in\mathcal{H}}\{\sum_{\bar{L}:m\in\bar{L}}\sum_{n:n\in\bar{L}}\frac{\lambda^*_{\bar{L},n,m}}{\alpha}\}$$

.

Note that with $\left\{\lambda_{\bar{L},n,m}\right\}$, only task types $\mathcal{L}_{\mathcal{B}}$ contribute to the arrival for $\mathcal{B}$.

In the following argument, we will focus on modifying the decomposition of $\mathcal{L}_{\mathcal{B}}$ over $\mathcal{B}$ to achieve the goal. In particular, for ease of exposition, we model the relationship between the task types $\mathcal{L}_{\mathcal{B}}$ and the beneficiaries $\mathcal{B}$ by a bipartite graph $\mathbb{G}=(\mathcal{X},\mathcal{Y},\mathcal{E})$. Each vertex $x\in\mathcal{X}$ corresponds to a task type $\bar{L}\in\mathcal{L}_{\mathcal{B}}$ and we assign x a budget $b(x)=\lambda_{\bar{L}}$. Each vertex $y\in\mathcal{Y}$ represents a server $m\in\mathcal{B}$. If sever m is local to task type $\bar{L}$, we put an edge xy in $\mathcal{E}$. For any vertex v in the graph, we denote the set of its neighbor vertices by $\mathcal{N}(v)$. And let $\mathcal{N}(\mathcal{V})=\cup_{v\in\mathcal{V}}\mathcal{N}(v)$ for any vertex set $\mathcal{V}$. Consider the weight function

$$\begin{aligned} w:\quad & \mathcal{E}\to[0,+\infty)\\ & xy\to w(xy)\end{aligned}$$

Let $w(x)=\sum_{y\in\mathcal{N}(x)}w(xy)$ and $w(y)=\sum_{x\in\mathcal{N}(y)}w(xy)$. If a weight function w satisfies that $\forall x\in\mathcal{X}$, $w(x)=b(x)$ and $w(y)\geq\alpha$, it is said to be a *proper weight function*. Let $\mathcal{W}$ be the set of proper weight functions. Then $\mathcal{W}$ is nonempty by Lemma 3.

Observe that with the fixed decomposition of $\mathcal{L}_{\mathcal{H}}$ over $\mathcal{H}$, for any proper weight function, we can further decompose $w(xy)=\lambda_{\bar{L},m}$ into $\{\tilde{w}(xy,z)\}_{z\in\mathcal{M}}=\{\lambda_{\bar{L},m,n}\}_{n\in\mathcal{M}}$ to satisfy Eq. (1) and (3). For any such refined decomposition, let $u(y)=\sum_{\bar{L}:m\in\bar{L}}\lambda_{\bar{L},m,m}=\sum_{x\in\mathcal{N}(y)}w(xy,y)$, which denotes the rate of arrivals that are served locally at server m. Then $w(y)-u(y)$ is the rate of arrivals served remotely by helpers $\mathcal{H}$. Hence in the rest of the proof we only consider proper weight functions. To prove the lemma, it suffices to find a weight function w and its refined decomposition $\tilde{w}$ such that

$$\forall x\in\mathcal{X},w(x)=b(x), \tag{34}$$

$$\forall y\in\mathcal{Y},w(y)\geq\alpha,u(y)=\alpha-\epsilon_b, \tag{35}$$

$$\forall x\in\mathcal{X},\forall y\in\mathcal{Y},\forall z\in\mathcal{H},\tilde{w}(xy,z)>\lambda_0. \tag{36}$$

Step 1.

Define

$$\kappa_1 = \min_{\mathcal{G}\subseteq\mathcal{B}} \left\{ \sum_{\bar{L}\in\mathcal{L}(\mathcal{G})} \lambda_{\bar{L}} - |\mathcal{G}|\alpha \right\}.$$

From the assumption for the heavy traffic, $\kappa_1 > 0$, and for any $\mathcal{G} \subseteq \mathcal{B}$,

$$\sum_{\bar{L}\in\mathcal{L}(\mathcal{G})} \lambda_{\bar{L}} \geq |\mathcal{G}|\alpha + \kappa_1. \tag{37}$$

First we obtain a proper weight function w such that for any $y \in \mathcal{Y}$, $w(y) \geq \alpha + \frac{\kappa'}{M_b}$, where

$$\kappa' = \min\left\{ \kappa_1, \frac{\gamma(1-\rho^h_{max})}{|\mathcal{L}_{\mathcal{B}}|}, \min_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} \frac{\lambda_{\bar{L}}}{|\bar{L}|+1} \right\}. \tag{38}$$

We have the following claim.

Claim 1. For any proper weight function $w \in \mathcal{M}$, if there exists $y_0 \in \mathcal{Y}$ with $w(y_0) < \alpha + \frac{\kappa'}{M_b}$, then there exists a path $\mathcal{P} = y_0x_0y_1x_1\cdots y_k$ such that $w(x_iy_{i+1}) > 0$ for $i = 0, 1, ..., k-1$, and $w(y_i) \leq \alpha + \frac{\kappa'}{M_b}$ for $i = 1, ..., k-1$ and $w(y_k) > \alpha + \frac{\kappa'}{M_b}$.

Proof of the Claim.
If there exists $x_0 \in \mathcal{N}(y_0)$ and $y_1 \in \mathcal{N}(x_0)$ such that $w(x_0y_1) > 0$ and $w(y_1) > \alpha + \frac{\kappa'}{M_b}$, then let $\mathcal{P} = y_0x_0y_1$ and it is done. Otherwise $\forall x \in \mathcal{N}(y_0)$ and $y \in \mathcal{N}(x)$, either $w(xy) = 0$ or $w(y) \leq \alpha + \frac{\kappa'}{M_b}$. Consider the sets $\mathcal{X}_0 = \mathcal{N}(y_0)$ and $\mathcal{Y}_1 = \mathcal{N}(\mathcal{X}_0)$. Let $\mathcal{Z} = \{y \in \mathcal{Y} | \exists x \in \mathcal{N}(y), s.t.\ x \notin \mathcal{X}_0\}$, which is the set of vertices that have neighbours outside $\mathcal{X}_0$. Note that for any $y \in \mathcal{Y}_1\backslash\mathcal{Z}$, $\exists x \in \mathcal{X}_0$ such that $w(xy) > 0$, otherwise $w(y) = 0$, which contradicts with $w(y) \geq \alpha$. Moreover, $\mathcal{Z} \neq \emptyset$ and $\exists x_0 \in \mathcal{X}_0$, $y_1 \in \mathcal{Z}$ such that $w(x_0y_1) > 0$, since if this is not case then

$$\sum_{x\in\mathcal{X}_0} b(x) = \sum_{x\in\mathcal{X}_0} w(x) = \sum_{y\in\mathcal{Y}_1\backslash\mathcal{Z}} w(y) = \sum_{y\in\mathcal{Y}_1\backslash\mathcal{Z}} w(y) < |\mathcal{Y}_1\backslash\mathcal{Z}|(\alpha + \frac{\kappa'}{M_b}) \leq |\mathcal{Y}_1\backslash\mathcal{Z}|\alpha + \kappa',$$

which contradicts with the heavy traffic assumption in (37). Thus there exists $\exists x_0 \in \mathcal{X}_0$, $y_1 \in \mathcal{Z}$ such that $w(x_0y_1) > 0$. If $\exists x_1 \in \mathcal{N}(y_1)$, and $\exists y_2 \in \mathcal{N}(x_1)$ such that $w(x_1y_2) > 0$ and $w(y_2) > \alpha + \frac{\kappa'}{M_b}$, then let $\mathcal{P} = y_0x_0y_1x_1y_2$ and we are done. Otherwise, let $\mathcal{X}_1 = \mathcal{N}(\{y_0, y_1\})$, and $\mathcal{Y}_2 = \mathcal{N}(\mathcal{X}_1)$. Arguing similarly, we can find $x_1 \in \mathcal{X}_1$ and $y_2 \in \mathcal{Y}_2$ such that y_2 has neighbors outside of $\mathcal{X}_1$ and $w(x_1y_2) > 0$. Then if $\exists x_2 \in \mathcal{N}(y_2)$, and $\exists y_3 \in \mathcal{N}(x_2)$ such that $w(x_2y_3) > 0$ and $w(y_3) > \alpha + \frac{\kappa'}{M_b}$, then let $\mathcal{P} = y_0x_0y_1x_1y_2x_2y_3$ and we are done. Otherwise we can continue to consider $\mathcal{X}_2 = \mathcal{N}(\{y_0, y_1, y_2\})$, and $\mathcal{Y}_2 = \mathcal{N}(\mathcal{X}_2)$. The procedure will end in finite steps for the following reason. In the connected component $(\mathcal{X}', \mathcal{Y}', \mathcal{E}')$ that contains y_0, there exists at least $y \in \mathcal{Y}'$ such that $w(y) > \alpha + \frac{\kappa'}{M_b}$. Otherwise

$$\sum_{x\in\mathcal{X}'} b(x) = \sum_{x\in\mathcal{X}'} w(x) = \sum_{y\in\mathcal{Y}':\exists x\in\mathcal{X}' \ s.t.\ w(xy)>0} w(y) < |\mathcal{Y}'|(\alpha + \frac{\kappa'}{M_b}) \leq |\mathcal{Y}'|\alpha + \kappa',$$

which contradicts with the heavy traffic assumption.

Following the above procedure, we obtain a sequence $\mathcal{Y}_0 \subsetneq \mathcal{Y}_1 \subsetneq \cdots$ in $\mathcal{Y}'$. The procedure ends when the sequence hits some $y \in \mathcal{Y}'$ with $w(y) > \alpha + \frac{\kappa'}{M_b}$. So it takes at most $|\mathcal{Y}'|$ steps. This completes the proof for the claim.

■

Consider a proper weight function w such that $\min_{y\in\mathcal{Y}} w(y)$ is maximized. Then for any $y \in \mathcal{Y}$, $w(y) \geq \alpha + \frac{\kappa'}{M_b}$. If w does not satisfy this condition, let $y_0 \in \arg\min_{y\in\mathcal{Y}} w(y)$. Then $\alpha \leq w(y_0) < \alpha + \frac{\kappa'}{M_b}$. From Claim 1, there exists a path $\mathcal{P} = y_0x_0y_1x_1\cdots y_k$ such that $w(x_iy_{i+1}) > 0$ for $i = 0, 1, ..., k-1$, and $w(y_i) \leq \alpha + \frac{\kappa'}{M_b}$ for $i = 1, ..., k-1$ and $w(y_k) > \alpha + \frac{\kappa'}{M_b}$. Let $\kappa_2 = \min\{w(x_0y_1), w(x_1y_2), ..., w(x_{k-1}y_k), w(y_k) - (\alpha + \frac{\kappa'}{M_b})\}$. Then $\kappa_2 > 0$. We modify w to get another weight function $\tilde{w}$ as follows

$$\tilde{w}(xy) = \begin{cases} w(xy) + \kappa_2 & \text{if } x = x_i, y = y_i, \text{ where } i = 0, 1, ..., k \\ w(xy) - \kappa_2 & \text{if } x = x_i, y = y_{i+1}, \text{ where } i = 0, 1, ..., k-1 \\ w(xy) & \text{otherwise.} \end{cases}$$

By the definition of κ_2, $\tilde{w}(xy) \geq 0$ for any $xy \in \mathcal{E}$. And for any $x \in \mathcal{X}$, $b(x) = \tilde{w}(x)$. For any $y \in \mathcal{Y}, y \neq y_0, y_k$, $\tilde{w}(y) = w(y) \geq \alpha$. And $\tilde{w}(y_k) \geq \alpha + \frac{\kappa'}{M_b}$, $\tilde{w}(y_0) = w(y_0) + \kappa_2 > w(y_0) \geq \alpha$. We then modify other vertices in $\arg\min_{y\in\mathcal{Y}} w(y)$ using similar method, which results a proper weight function $\hat{w}$. Then $\min_{y\in\mathcal{Y}} \hat{w}(y) > \min_{y\in\mathcal{Y}} w(y)$, which contradicts with the assumption that w maximize $\min_{y\in\mathcal{Y}} w(y)$.

Step 2.

Next we modify the refined decomposition $w(xy, z)$ of weight function w to $w'(xy, z)$ such that $w'(xy, z)$ satisfies condition (35).

Let $\epsilon_b = \min_{y\in\mathcal{Y}}(1 - \frac{u(y)}{\alpha})$. Note that $w(xy, z)$ satisfies Eq. (1), hence

$$\begin{aligned} \sum_{y\in\mathcal{Y}} u(y) &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} w(xy, y) \\ &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} w(xy) - \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{H}} w(xy, z) \\ &= \sum_{\bar{L}\in\mathcal{L}_\mathcal{B}} \lambda_{\bar{L}} - \sum_{\bar{L}\in\mathcal{L}_\mathcal{B}}\sum_{y:y\in\bar{L}}\sum_{z\in\mathcal{H}} \lambda_{\bar{L},y,z} \\ &< \sum_{\bar{L}\in\mathcal{L}_\mathcal{B}} \lambda_{\bar{L}} - \gamma(M_h - \sum_{\bar{L}\in\mathcal{L}^*_\mathcal{H}} \lambda_{\bar{L}}) \\ &= \frac{\epsilon}{\alpha} \end{aligned}$$

Hence

$$\epsilon_b \leq \frac{\sum_{y\in\mathcal{Y}}(\alpha - u(y))}{M_b\alpha} \leq \frac{\epsilon}{M_b\alpha}.$$

Note that $w(y) \geq \alpha + \frac{\kappa'}{M_b}$. For any $y \in \mathcal{Y}$ with $\alpha - u(y) > \alpha\epsilon_b$, it is easy to increase $u(y)$ to $\alpha(1 - \epsilon_b)$ by reducing appropriate amount of remote load on each helper. Observe that such modification will maintain the correctness of Eq. (1). Let $w'(xy, z)$ denote the refined weight function with this modification.

To obtain a weight function that satisfies condition (36), we take the following steps to modify w'.

Step 3.

For any $x \in \mathcal{X}$, let $f(x) = \sum_{y\in\mathcal{N}(x)} w(xy, y)$, which is the arrival rate of tasks with type $\bar{L}$ that are served locally. We first modify w' to w_1 such that w_1 satisfies

$$b(x) - f_1(x) \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b}, \quad \forall x \in \mathcal{X}. \tag{39}$$

Let $w^l(xy) = w(xy, y)$ and $w^r(xy) = w(xy) - w^l(xy) = w(xy, y)$.

Consider any $x \in \mathcal{X}$. By the pigeon hole principle, either $b(x) - f(x) \geq \frac{\lambda_{\bar{L}}}{|\bar{L}|+1}$ or there exists an edge xy such that $w(xy, y) \geq \frac{\lambda_{\bar{L}}}{|\bar{L}|+1}$. That is, either $b(x) - f(x) \geq \frac{\lambda_{\bar{L}}}{|\bar{L}|+1} \geq \kappa' \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b}$. For the case $b(x) - f(x) <$

$\frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b} < \frac{\lambda_{\bar{L}}}{|\bar{L}|+1}$, there exists an edge xy such that $w(xy,y) \geq \frac{\lambda_{\bar{L}}}{|\bar{L}|+1}$. Let $w_1^r(xy) = \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b}$. Denote the amount of change for $w^r(xy)$ by $\delta = \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b} - w^r(xy)$. And let $w_1^l(xy) = w^l(xy) - \delta \geq 0$. Note that $\sum_{x\in\mathcal{X}} w^r(xy) = w(y) - u(y) \geq \alpha + \frac{\kappa'}{M_b} - \alpha(1-\epsilon_b) > \frac{\kappa'}{M_b}$. Again by the pigeon hole principle, for each such edge xy, there exists $x' \in \mathcal{X}$ such that $w^r(xy) > \frac{\kappa'}{|\mathcal{L}_\mathcal{B}| M_b}$. To keep $u(y) = \sum_{x\in\mathcal{X}} w^l(xy)$ unchanged, let $w_1^r(x'y) = w^r(x'y) - \delta$, and $w_1^l(x'y) = w^l(x'y) + \delta$. Then $w_1^r(x'y) > \frac{\kappa'}{|\mathcal{L}_\mathcal{B}| M_b} - \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b} > \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b}$, which won't violate the condition (39) for x'. And keep other $w^l(xy)$ and $w^r(xy)$ unchanged. Then $w_1(xy)$ satisfies (39).

Step 4. Next we modify w_1 to w_2 such that w_2 satisfies the following condition

$$w_2^r(xy) \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3}, \quad \forall x \in \mathcal{X}, \quad \forall y \in \mathcal{Y}. \tag{40}$$

For any edge xy with $w_1^r(xy) < \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3}$, let $w_2^r(xy) = \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3}$. For each such edge, again by the pigeon hole principle, there exists $y' \in \mathcal{Y}$ such that $w_1^r(xy') > \frac{b(x)-f_1(x)}{|\bar{L}|} \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b |\bar{L}|} \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^2}$. Let $w_2^r(xy') = w_1^r(xy') - (\frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3} - w_1^r(xy)) \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3}$. On other edges let the values of w_2 and w_1 be equal. It is easy to verify that w_2 is still a proper weight function, since for each $y \in \mathcal{Y}$, $u_2(y) = \alpha(1-\epsilon_b)$ remains unchanged, and $w_2(y) \geq w_1(y) - \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3} |\mathcal{L}_\mathcal{B}| M_b \geq \alpha + \frac{\kappa'}{M_b} - \frac{\kappa'}{|\mathcal{L}_\mathcal{B}| M_b^2} \geq \alpha$.

Step 5. At the last step, we modify w_2 to w'' such that w'' satisfies condition (36). In particular, let $\lambda_0 = \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3 M_h}$.

For any edge xy, let $w''(xy,y) = w_2^l(xy)$. Then $u''(y) = u_2(y) = \alpha(1-\epsilon_b)$. Next we focus on the decomposition of $w_2^r(xy)$ into $\{w''(xy,z)\}_{z\in\mathcal{H}}$. We start with an intermediate function $\{w_3(xy,z)\}_{z\in\mathcal{H}}$ with $w_3(xy,z) = \lambda_0$ for any $z \in \mathcal{H}$. Then for any $z \in \mathcal{H}$, $\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} w_3(xy,z) = \sum_{x\in\mathcal{X}} \lambda_0 |\bar{L}| < \sum_{x\in\mathcal{X}} M_b \lambda_0 = \frac{\kappa'}{|\mathcal{L}_\mathcal{B}| M_b^2 M_h}$. By the definition of κ',

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{w_3(xy,z)}{\gamma} + \frac{\lambda_z^l}{\alpha} \leq \frac{1-\rho_{max}^h}{|\mathcal{L}_\mathcal{B}| M_b^2 M_h} + \frac{\lambda_z^l}{\alpha} < 1 - \rho_{max}^h + \frac{\lambda_z^l}{\alpha} < 1. \tag{41}$$

Hence for any $z \in \mathcal{H}$, the initial assignment $\{w_3(xy,z)\}$ and the decomposition of $\mathcal{L}_\mathcal{H}^*$ make Eq.(1) satisfied.

For any edge xy, $w_2^r(xy) - w_3^r(xy) = w_2^r(xy) - sum_{z\in\mathcal{H}} w_3(xy,z) \geq \frac{\kappa'}{|\mathcal{L}_\mathcal{B}|^2 M_b^3} - M_h \lambda_0 = 0$. Next we distribute the remaining $w_2^r(xy) - w_3^r(xy)$ over helpers $\mathcal{H}$ and ensure Eq. (1) hold. Let $\epsilon_h = \frac{\epsilon - M_b \alpha \epsilon_b}{M_h \gamma}$. Then

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} w_2^r(xy) = \sum_{x\in\mathcal{X}} w(x) - \sum_{y\in\mathcal{Y}} u(y) = \sum_{\bar{L}\in\mathcal{L}_\mathcal{B}} \lambda_{\bar{L}} - M_b \alpha(1-\epsilon_b) = M_h\gamma - \frac{\gamma}{\alpha} \sum_{\bar{L}\in\mathcal{L}_\mathcal{H}^*} \lambda_{\bar{L}} - M_h \gamma \epsilon_h. \tag{42}$$

Hence if there exists edge xy with $w_2^r(xy) - w_3^r(xy) > 0$,

$$\sum_{z\in\mathcal{H}} \left(\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{w_3(xy,z)}{\gamma} + \frac{\lambda_z^l}{\alpha} \right) < M_h - M_h \epsilon_h.$$

By the pigeon hole principle, there exists $z \in \mathcal{H}$ such that

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{w_3(xy,z)}{\gamma} + \frac{\lambda_z^l}{\alpha} < 1 - \epsilon_h. \tag{43}$$

Then we increase $w_3(xy,z)$ by an appropriate amount such that either $w_2^r(xy) - w_3^r(xy) = 0$ or

$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{w_3(xy,z)}{\gamma}+\frac{\lambda_z^l}{\alpha}=1-\epsilon_h$. If $w_2^r(xy)-w_3^r(xy)>0$, we repeat the above increasing procedure for $w_3(xy,z)$. For any $w(xy,z)$, the procedure results in $w_2^r(xy)-w_3^r(xy)=0$ and maintains the validity of Eq.(41). It takes no more than M_h steps.

By letting $w''(xy,z)=w_3(xy,z)$, we obtain a refined weight function w'' that satisfies conditions (34)-(36). This completes the proof.

■

7.6 Proof of Lemma 10-13

7.6.1 Proof of Lemma 10

The proof is similar to the proof of Lemma 5.

7.6.2 Proof of Lemma 11

The proof is similar to the proof of Lemma 6.

7.6.3 Proof of Lemma 12

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\langle Q^{max}(t)\sqrt{M_b}c_b,\lambda^{*r}\rangle|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
=\;&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\sum_{m\in\mathcal{B}}\left(Q^{max}(t)\lambda_m^{*r}\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
\leq\;&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\sum_{m\in\mathcal{B}}\left((Q^{max}(t_0)+TC_A)\lambda_m^{*r}\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
\leq\;&(t_0+T-t^*)Q^{max}(t_0)\sum_{m\in\mathcal{B}}\lambda_m^{*r}+C,
\end{aligned}$$

where C is a constant.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\langle Q(t),S^r(t)\rangle|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
=\;&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\sum_{m\in\mathcal{B}}\left(Q_m(t)\sum_{n:n\neq m}R_n(t)I_{\{\eta_n(t)=m\}}\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
=\;&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\sum_{n\in\mathcal{M}}R_n(t)\left(\sum_{m\in\mathcal{B}:m\neq n}Q_m(t)I_{\{\eta_n(t)=m\}}\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
\geq\;&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\sum_{n\in\mathcal{H}}\gamma\left(\sum_{m\in\mathcal{B}}Q_m(t)I_{\{\eta_n(t)=m\}}\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]
\end{aligned}\tag{44}$$

Similar to the proof for Lemma 6, consider the following random variables

$$\tau_n^t:=\max\{\tau:\tau\leq t,f_n(\tau)=-1\},n\in\mathcal{M}.$$

Hence τ_n^t is the last moment before t at which server n makes a scheduling decision. Therefore the status of server m remains the same from time τ_n^t to t, i.e., $\eta_n(t)=\eta_n(\tau_n^t)$. Note that if a remote task is scheduled for server n, it must comes from the longest queue. Hence,

$$\begin{aligned}\sum_{m\in\mathcal{B}}\mathbb{E}\left[Q_m(\tau_n^t)I_{\{\eta_n(t)=m\}}|Z(\tau_n^t)\right] &= \sum_{m\in\mathcal{B}}\mathbb{E}\left[Q_m(\tau_n^t)I_{\{\eta_n(\tau_n^t)=m\}}|Z(\tau_n^t)\right]\\ &= Q^{max}(\tau_n^t)I_{\{\eta_n(\tau_n^t)\in\mathcal{B}\}} = Q^{max}(\tau_n^t)I_{\{\eta_n(t)\in\mathcal{B}\}}\end{aligned}$$

Applying the bounded difference between $Q(t_0)$, $Q(\tau_n^t)$ and $Q(t)$ yields

$$\begin{aligned}&\mathbb{E}\left[\sum_{m\in\mathcal{B}}Q_m(t)I_{\{\eta_n(t)=m\}}|Z^{(B)}(t_0)\right]\\ =\;&\mathbb{E}\left[\mathbb{E}\left[\sum_{m\in\mathcal{B}}Q_m(t)I_{\{\eta_n(t)=m\}}|Z(t)\right]|Z^{(B)}(t_0)\right]\\ =\;&\mathbb{E}\left[\sum_{m\in\mathcal{B}}Q_m(t)\mathbb{E}\left[I_{\{\eta_n(t)=m\}}|Z(t)\right]|Z^{(B)}(t_0)\right]\\ =\;&\mathbb{E}\left[\sum_{m\in\mathcal{B}}Q_m(t)\mathbb{E}\left[\mathbb{E}\left[I_{\{\eta_n(t)=m\}}|Z(\tau_n^t)\right]|Z(t)\right]|Z^{(B)}(t_0)\right]\\ \geq\;&\mathbb{E}\left[\sum_{m\in\mathcal{B}}(Q_m(\tau_n^t)-TM)\mathbb{E}\left[\mathbb{E}\left[I_{\{\eta_n(t)=m\}}|Z(\tau_n^t)\right]|Z(t)\right]|Z^{(B)}(t_0)\right]\\ \geq\;&\mathbb{E}\left[\mathbb{E}\left[\sum_{m\in\mathcal{B}}\mathbb{E}\left[Q_m(\tau_n^t)I_{\{\eta_n(t)=m\}}|Z(\tau_n^t)\right]|Z(t)\right]|Z^{(B)}(t_0)\right]-TM\\ =\;&\mathbb{E}\left[\mathbb{E}\left[Q^{max}(\tau_n^t)I_{\{\eta_n(t)\in\mathcal{B}\}}|Z(t)\right]|Z^{(B)}(t_0)\right]-TM\\ \geq\;&\mathbb{E}\left[Q^{max}(t_0)\mathbb{E}\left[I_{\{\eta_n(t)\in\mathcal{B}\}}|Z(t)\right]|Z^{(B)}(t_0)\right]-2TM\\ \geq\;&Q^{max}(t_0)\mathbb{E}\left[I_{\{\eta_n(t)\in\mathcal{B}\}}|Z^{(B)}(t_0)\right]+C. \qquad (45)\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\left(\langle Q^{max}(t)\sqrt{M_b}c_b,\lambda^{*r}\rangle-\langle Q(t),S^r(t)\rangle\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\ \leq\;&(t_0+T-t^*)Q^{max}(t_0)(\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\mathbb{E}\left[\sum_{n\in\mathcal{H}}I_{\{\eta_n(t)\in\mathcal{B}\}}|Z^{(B)}(t_0)\right])+C.\end{aligned}$$

By the boundedness of arrivals and service, we have

$$\langle c_b,Q(t_0)\rangle-\frac{TM}{\sqrt{M_b}}\leq\langle c_b,Q(t)\rangle\leq\langle c_b,Q(t_0)\rangle+\frac{TC_A}{\sqrt{M_b}}$$

Hence

$$\begin{aligned}&\mathbb{E}\left[\langle c_b,Q(t)\rangle\langle c_b,\hat{A}(t)-S(t)\rangle|t^*<t_0+K,Z^{(B)}(t_0)\right]\\ \geq\;&\mathbb{E}\left[\langle c_b,Q(t_0)\rangle\langle c_b,\hat{A}(t)-S(t)\rangle-\frac{TC_A}{\sqrt{M_b}}\langle c_b,S(t)\rangle|t^*<t_0+K,Z^{(B)}(t_0)\right]\\ \geq\;&\langle c_b,Q(t_0)\rangle\frac{1}{\sqrt{M_b}}\mathbb{E}\left[\sum_{m\in\mathcal{B}}(\hat{A}_m(t)-S_m(t))|t^*<t_0+K,Z^{(B)}(t_0)\right]-C\end{aligned}$$

Observe that

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{A}_m(t)|t^*<t_0+K,Z^{(B)}(t_0)\right]=\mathbb{E}\left[\sum_{\bar{L}\in\mathcal{L}_\mathcal{B}}A_{\bar{L}}(t)|t^*<t_0+K,Z^{(B)}(t_0)\right]=\sum_{\bar{L}\in\mathcal{L}_\mathcal{B}}\lambda_{\bar{L}}$$

And

$$\begin{aligned}
&\mathbb{E}\left[\sum_{m\in\mathcal{B}}S_m(t)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
=\;&\mathbb{E}\left[\sum_{m\in\mathcal{B}}(S_m^l(t)+\sum_{n:n\neq m}R_n(t)I_{\{\eta_n(t)=m\}})|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
=\;&\mathbb{E}\left[\sum_{m\in\mathcal{B}}(\alpha I_{\{\eta_m(t)=m\}}+\gamma I_{\{\eta_m(t)\in\mathcal{B}\}})+\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
\leq\;&\alpha M_b+\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\langle c_b,Q(t)\rangle\langle c_b,\hat{A}(t)-S(t)\rangle|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
\geq\;&\langle c_b,Q(t_0)\rangle\frac{1}{\sqrt{M_b}}\left(\sum_{\bar{L}\in\mathcal{L}_\mathcal{B}}\lambda_{\bar{L}}-\alpha M_b+\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]\right)-C
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1}\left(\langle Q^{max}(t)\sqrt{M_b}c_b,\lambda^{*r}\rangle-\langle Q(t),S^r(t)\rangle-\langle c_b,Q(t)\rangle\langle c_b,\hat{A}(t)-S(t)\rangle\right)|t^*<t_0+K,Z^{(B)}(t_0)\right]\\
\leq\;&(t_0+T-t^*)Q^{max}(t_0)\left(\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\mathbb{E}\left[\sum_{n\in\mathcal{H}}I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]\right)\\
&-(t_0+T-t^*)\langle c_b,Q(t_0)\rangle\frac{1}{\sqrt{M_b}}\left(\sum_{\bar{L}\in\mathcal{L}_\mathcal{B}}\lambda_{\bar{L}}-\alpha M_b+\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]\right)+C\\
=\;&(t_0+T-t^*)\left[\alpha\epsilon_0\sum_{m\in\mathcal{B}}Q_m(t_0)+\frac{1}{M_b}\left(\sum_{m\in\mathcal{B}}Q_m(t_0)-M_bQ^{max}(t_0)\right)\right.\\
&\left.\cdot\left(\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]-\sum_{m\in\mathcal{B}}\lambda_m^{*r}\right)\right]+C
\end{aligned}\tag{46}$$

We will show that $\forall\epsilon<\frac{M_b\lambda_0}{4}$, there exists a constant $L_r>0$ not depending on ϵ such that $\forall Z^{(B)}(t_0)$ with $||Z_\perp^{(B)}(t_0)||\geq L_r$

$$\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]\geq\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\frac{M_b\lambda_0}{4}\tag{47}$$

We prove Eq. (47) by contradiction.

Assume that $\exists\epsilon<\frac{M_b\lambda_0}{4}$, $\forall L_1>0$, there exists $||Z_\perp^{(B)}(t_0)||>L_1$ such that

$$\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|t^*<t_0+K,Z^{(B)}(t_0)\right]<\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\frac{M_b\lambda_0}{4}.$$

Then we can bound the total amount of service received by beneficiaries when $Z^{(B)}$ hits the state $Z^{(B)}(t_0)$ as

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{B}}S_m(t)|Z^{(B)}(t_0)\right] &= \mathbb{E}\left[\sum_{m\in\mathcal{B}}(\alpha I_{\{\eta_m(t)=m\}}+\gamma I_{\{\eta_m(t)\in\mathcal{B}\}})+\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|Z^{(B)}(t_0)\right]\\ &< M_b\alpha+\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\frac{M_b\lambda_0}{4}\\ &< M_b\alpha+\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\epsilon.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{B}}A_m(t)\right]\geq\sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}}\lambda_{\bar{L}} &= M_b\alpha+\Phi\alpha+\gamma(M_h-\Phi)-\epsilon\\ &= M_b\alpha+\sum_{m\in\mathcal{B}}\lambda_m^{*r}-\epsilon\\ &> \mathbb{E}\left[\sum_{m\in\mathcal{B}}S_m(t)|Z^{(B)}(t_0)\right].\end{aligned}$$

That is, when $Z^{(B)}$ hits the state $Z^{(B)}(t_0)$, the amount of service beneficiaries receive is insufficient for the arrival. Arguing similar as the proof for stability of beneficiary system, all beneficiary queues will grow together. Then shared arrivals will join helper queues, i.e., the helper subsystem receives arrivals with maximum rate $\sum_{\bar{L}\in\mathcal{L}_{\mathcal{H}}^*}\lambda_{\bar{L}}$. From Section 4, the helper subsystem will be stable with such arrivals and any moment of $|Q^(H)|$ is bounded. Consider $\hat{Z}^{(B)}$ with $||\hat{Z}_\perp^{(B)}||>||Z_\perp^{(B)}(t_0)||\geq L_r$ and $\hat{Q}_m>Q_m(t_0)$ for any $m\in\mathcal{B}$, then $\hat{Q}_B^{max}\geq\frac{1}{\sqrt{M_b}}||\hat{Z}_\perp^{(B)}||=L_2$. Note that we can make $\texttt{Pr}[Q_H^{max}>\hat{Q}_B^{max}]$ arbitrarily small by selecting sufficiently large L_2. An upper bound on the amount of remote service provided by helpers and devoted to helpers, denoted by δ_{HH}, is given by $\delta_{HH}\leq\gamma\texttt{Pr}[Q_H^{max}>\hat{Q}_B^{max}]$, which can be arbitrarily small. Hence $\exists L_2>0$ such that $\delta_{HH}<\frac{M_b\lambda_0}{4}$.

Thus we can obtain an lower bound on the amount of remote service provided by helpers and devoted to beneficiaries

$$\mathbb{E}\left[\sum_{n\in\mathcal{H}}\gamma I_{\{\eta_n(t)\in\mathcal{B}\}}|\hat{Z}^{(B)}\right]\geq R_{\mathcal{H}}-\frac{M_b\lambda_0}{4}.$$

This contradicts with the assumption. Note that L_2 does not depend on ϵ as λ_0 is independent of ϵ.
This finishes the proof for Lemma 12.

7.6.4 Proof of Lemma 13

For each $m\in\mathcal{B}$, define

$$A_m^e(t)=\sum_{\bar{L}:\bar{L}\notin\mathcal{L}_{\mathcal{B}},m\in\bar{L}}A_{\bar{L},m}(t)$$

For any $L > 0$,

$$\begin{aligned}
& \mathbb{E}\left[\langle Q(t), A(t) - \hat{A}(t)\rangle | Z^{(B)}(t_0)\right] \\
= \; & \mathbb{E}\left[\sum_{m\in\mathcal{B}} \left(Q_m(t) \sum_{\bar{L}:\bar{L}\notin\mathcal{L}_{\mathcal{B}}, m\in\bar{L}} A_{\bar{L},m}(t)\right) | Z^{(B)}(t_0)\right] \\
= \; & \mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m(t) A^e_m(t) | Z^{(B)}(t_0)\right] \\
= \; & \mathbb{E}\left[\sum_{m\in\mathcal{B}} \left(Q_m(t) I_{\{Q_m(t)<L\}} A^e_m(t) + Q_m(t) I_{\{Q_m(t)\geq L\}} A^e_m(t)\right) | Z^{(B)}(t_0)\right] \\
\leq \; & \mathbb{E}\left[L \sum_{m\in\mathcal{B}} A^e_m(t) + \sum_{m\in\mathcal{B}} Q_m(t) I_{\{Q_m(t)\geq L\}} A^e_m(t) | Z^{(B)}(t_0)\right] \\
\leq \; & C + \mathbb{E}\left[Q^{max}(t) \sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A^e_m(t) | Z^{(B)}(t_0)\right] \\
\leq \; & C + Q^{max}(t_0)\mathbb{E}\left[\sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A^e_m(t) | Z^{(B)}(t_0)\right].
\end{aligned}$$

Note that

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} \langle c_b, Q(t)\rangle\langle c_b, A(t) - \hat{A}(t)\rangle | t^* < t_0 + K, Z^{(B)}(t_0)\right] \\
= \; & \mathbb{E}\left[\frac{\sum_{m\in\mathcal{B}} Q_m(t)}{M_b} \sum_{m\in\mathcal{B}} A^e_m(t) | t^* < t_0 + K, Z^{(B)}(t_0)\right] \\
\geq \; & \frac{\sum_{m\in\mathcal{B}} Q_m(t_0)}{M_b} \mathbb{E}\left[\sum_{m\in\mathcal{B}} \left(I_{\{Q_m(t)<L\}} A^e_m(t) + I_{\{Q_m(t)\geq L\}} A^e_m(t)\right) | t^* < t_0 + K, Z^{(B)}(t_0)\right] - C \\
\geq \; & \frac{\sum_{m\in\mathcal{B}} Q_m(t_0)}{M_b} \mathbb{E}\left[\sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A^e_m(t) | t^* < t_0 + K, Z^{(B)}(t_0)\right] - C
\end{aligned} \tag{48}$$

Hence

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} \left(-\langle c_b, Q(t)\rangle\langle c_b, A(t) - \hat{A}(t)\rangle + \langle Q(t), A(t) - \hat{A}(t)\rangle\right) | t^* < t_0 + K, Z^{(B)}(t_0)\right] \\
\leq \; & -\frac{1}{M_b}\left(\sum_{m\in\mathcal{B}} Q_m(t_0) - M_b Q^{max}(t_0)\right) \mathbb{E}\left[\sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A^e_m(t) | Z^{(B)}(t_0)\right] + C.
\end{aligned} \tag{49}$$

Next we will show that $\forall \epsilon < \frac{M_b\lambda_0(\alpha-\gamma)}{4\alpha}$, there exists $L_a > 0$ not depending on ϵ, $\forall L > L_a > 0$,

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A^e_m(t) | Z^{(B)}(t_0)\right] \leq \frac{M_b\lambda_0}{4}. \tag{50}$$

Similar to the proof for Eq.(47), we prove Eq.(50) by contradiction.
Assume that $\exists \epsilon < \frac{M_b\lambda_0(\alpha-\gamma)}{4\alpha}$, such that $\forall L > 0$, $\exists Z^{(B)}$ such that

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A^e_m(t) | Z^{(B)}(t_0)\right] > \frac{M_b\lambda_0}{4}.$$

Then total arrival for $\mathcal{B}$ is bounded as

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{B}} A_m(t)|Z^{(B)}(t_0)\right] &\geq \mathbb{E}\left[\sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} A_{\bar{L}}(t)|Z^{(B)}(t_0)\right] + \mathbb{E}\left[\sum_{m\in\mathcal{B}} I_{\{Q_m(t)\geq L\}} A_m^e(t)|Z^{(B)}(t_0)\right] \\ &> \sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} \lambda_{\bar{L}} + \frac{M_b\lambda_0}{4}\sum_{\bar{L}\in\mathcal{L}_{\mathcal{B}}} \lambda_{\bar{L}} + \frac{\alpha}{\alpha-\gamma}\epsilon \\ &\geq \mathbb{E}\left[\sum_{m\in\mathcal{B}} S_m(t)|Z^{(B)}(t_0)\right].\end{aligned}$$

Thus all beneficiaries grow together when sub-system hits the state $Z^{(B)}(t_0)$. Again shared arrivals will join helper queues. Consider stable helper subsystem with maximum arrival rate $\sum_{\bar{L}\in\mathcal{L}^*_{\mathcal{H}}} \lambda_{\bar{L}}$. Note that there exists uniform bound for stable helper subsystem. The bounded moments of $||Q^{(H)}||$ ensure that $\texttt{Pr}[Q_H^{max} > L_a]$ can be arbitrarily small with sufficiently large L_a. Hence $\exists L_a > 0$ such that the amount of shared arrivals that join $\mathcal{B}$, denoted by a^e, is upper bounded as

$$a^e \leq C_A \texttt{Pr}[Q_H^{max} > Q_B^{min} | Q^{min} > L_a] < \frac{M_b\lambda_0}{4}. \tag{51}$$

Note that L_a does not depend on ϵ. This contradicts with the assumption.
This completes the proof.

7.7 Proof of Theorem 4

We need the following lemmas to prove this theorem. For ease of exposition, we temporarily omit the superscript $^{(\epsilon)}$.

Lemma 16 *For any t,*

$$\langle Q(t), U(t)\rangle \leq M^2.$$

Lemma 17 *Let c be a vector with unit norm in $\mathbb{R}^{M_b}$. Then for any $t \geq 0$,*

$$\left|\left|Q_{||}^{(B)}(t+1)\right|\right|^2 - \left|\left|Q_{||}^{(B)}(t)\right|\right|^2 \geq 2\langle c, Q^{(B)}(t)\rangle\langle c, A^{(B)}(t) - S^{(B)}(t)\rangle,$$

where $Q_{||}^{(B)}$ is the parallel component of the beneficiary queue length vector $Q^{(B)}$ with respect to the direction c.

Lemma 18 *Consider a time slot t_0 and a positive integer T. Let c be a vector with unit norm in $\mathbb{R}^{M_b}$. Then for any t with $t_0 \leq t < t_0 + T$,*

$$\left|\left|Q_{\perp}^{(B)}(t)\right|\right| - \left|\left|Q_{\perp}^{(B)}(t_0)\right|\right| \leq T\sqrt{M_b}\max\{M, C_A\}, \tag{52}$$

where $Q_{\perp}^{(B)}$ is the perpendicular component of the beneficiary queue length vector $Q^{(B)}$ with respect to the direction c.

Lemma 19 *Consider a time slot t_0 and a positive integer T. For any t with $t_0 \leq t < t_0 + T$, let $G(t) = \langle Q^{(B)}(t), A^{(B)}(t) - S^{(B)}(t)\rangle - \langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle$. Then $G(t) \leq h\left|\left|Q_{\perp}^{(B)}(t_0)\right|\right| + F_0$, where $h = \sqrt{M_b}\max\{M, C_A\}$ and $F_0 = M_bT(\max\{M, C_A\})^2$ are constants.*

Proof for Theorem 4.
Consider the Lyapunov function

$$V(Z^{(B)}) = \left|\left|Q_{\perp}^{(B)}\right|\right|.$$

By the extended version of Lemma 1 in [7], it is sufficient to show that the T-period drift of $V(Z^{(B)})$ is always finite and is negative for sufficient large V. Fix an ϵ within the range specified in the theorem. Note that the T time slot drift of V is given by $\Delta V(Z^{(B)}) = [V(Z^{(B)}(t_0+T) - V(Z^{(B)}(t_0)))]I(Z^{(B)}(t_0) = Z)$.

First we show that $\Delta V(Z^{(B)})$ satisfies finite condition. From lemma 18, we can see that $Pr(\Delta V(Z^{(B)}) \leq C) = 1$ with $C = T\sqrt{M_b}\max\{M, C_A\}$.

Next we focus on the negative drift condition. Consider the following Lyapunov functions:

$$W(Z^{(B)}) = \left|\left|Q^{(B)}\right|\right|^2, W_{||}(Z^{(B)}) = \left|\left|Q_{||}^{(B)}\right|\right|^2.$$

Then $V(Z^{(B)}) = \sqrt{W(Z^{(B)}) - W_{||}(Z^{(B)})}$. Due to the concavity of the square root function, the drift of $V(Z^{(B)})$ satisfies the following inequality (Lemma 7 in [7]):

$$\Delta V(Z^{(B)}) \leq \frac{1}{2\left|\left|Q_{\perp}^{(B)}\right|\right|}(\Delta W(Z^{(B)}) - \Delta W_{||}(Z^{(B)})), \tag{53}$$

where $\Delta W(Z^{(B)})$ and $\Delta W_{||}(Z^{(B)})$ are the T time slot drifts for $W(Z^{(B)})$ and $W_{||}(Z^{(B)})$ respectively.

As it is hard to study the drift of $V(Z^{(B)})$ directly, we will get started with the drifts of $W(Z^{(B)})$ and $W_{||}(Z^{(B)})$.

For the drift $\Delta W(Z^{(B)})$, by the boundedness of arrivals and service and the property of the unused service in Lemma 16, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta W(Z^{(B)}(t_0))|Z^{(B)}(t_0)\right] &= \mathbb{E}\left[\left|\left|Q^{(B)}(t_0+T)\right|\right|^2 - \left|\left|Q^{(B)}(t_0)\right|\right|^2|Z^{(B)}(t_0)\right] \\
&= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\left(\left|\left|Q^{(B)}(t+1)\right|\right|^2 - \left|\left|Q^{(B)}(t)\right|\right|^2\right)|Z^{(B)}(t_0)\right] \\
&= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\Big(2\langle Q^{(B)}(t), A^{(B)}(t) - S^{(B)}(t)\rangle + 2\langle Q^{(B)}(t), U^{(B)}(t)\rangle + \left|\left|A^{(B)}(t) - S^{(B)}(t) + U^{(B}\right.\right. \\
&\leq 2\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\langle Q^{(B)}(t), A^{(B)}(t) - S^{(B)}(t)\rangle|Z^{(B)}(t_0)\right] + C_2,
\end{aligned}$$

where $C_2 > 0$ is a constant.

For the drift $\Delta W_{||}(Z^{(B)})$, by Lemma 17,

$$\begin{aligned}
\mathbb{E}\left[\Delta W_{||}(Z^{(B)}(t_0))|Z^{(B)}(t_0)\right] &= \mathbb{E}\left[\left|\left|Q_{||}^{(B)}(t_0+T)\right|\right|^2 - \left|\left|Q_{||}^{(B)}(t_0)\right|\right|^2|Z^{(B)}(t_0)\right] \\
&\geq 2\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1}\langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle|Z^{(B)}(t_0)\right]
\end{aligned}$$

Let $G(t) = \langle Q^{(B)}(t), A^{(B)}(t) - S^{(B)}(t)\rangle - \langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle$. Combining the above two inequalities gives us:

$$\mathbb{E}\left[\Delta W(Z^{(B)}(t_0)) - \Delta W_{||}(Z^{(B)}(t_0))|Z^{(B)}(t_0)\right] \leq 2\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G^{(B)}(t)|Z^{(B)}(t_0)\right] + C_2$$

To bound $\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0)\right]$, similar to the proof for throughput optimality, we decompose the probability space into two parts by two parts by using t^*: $D_1 = \{t^* \geq t_0 + K|Z(t_0)\}$ and $D_2 = \{t^* < t_0 + K|Z(t_0)\}$. Let $T = JK$,where J and K are positive integers. Then

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0)\right] &= \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0), t^* \geq t_0+K\right] Pr(t^* \geq t_0+K|Z(t_0)) \quad (54)\\ &+ \mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0), t^* < t_0+K\right] Pr(t^* < t_0+K|Z^{(B)}(t_0)). \quad (55)\end{aligned}$$

For the term 54, by Lemma 19 we have $\mathbb{E}\left[\sum_t G(t)|Z^{(B)}(t_0), t^* \geq t_0+K\right] \leq hT\left|\left|Q_\perp^{(B)}(t_0)\right|\right| + F_0T$.

For the term 55, we divide the summation into two parts: from $t = t_0$ to $t = t^*$ and from $t = t^*+1$ to $t = t_0+T-1$. The first part can be bounded in a similar way as term 54 by Lemma 19: $\mathbb{E}\left[\sum_{t=t_0}^{t^*} G(t)|Z^{(B)}(t_0), t^* < t_0+K\right] \leq Kh\left|\left|Q_\perp^{(B)}(t_0)\right|\right| + KF_0$. For the second part, by Lemmas 10-**??**, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} G(t)|Z^{(B)}(t_0), t^* < t_0+K\right]\\ \leq\ & -(t_0+T-t^*)\frac{\lambda_0}{2}\left(M_bQ^{max}(t_0) - \sum_{m\in\mathcal{B}} Q_m(t_0)\right) + F_1+F_2+F_3\\ =\ & -(t_0+T-t^*)\frac{\lambda_0}{2}||Q^{max}(t_0)\sqrt{M_b}c_b - Q^{(B)}(t_0)||_1 + F_4,\\ \leq\ & -(t_0+T-t^*)\frac{\lambda_0}{2}||Q^{max}(t_0)\sqrt{M_b}c_b - Q^{(B)}(t_0)|| + F_4\end{aligned}$$

where $F_4 = F_1+F_2+F_3$, and $||\cdot||_1$ is the L^1 norm. The last inequality follows by the fact that the L^1 norm of a vector is no smaller than its L^2 norm.

As $\langle c_b, Q^{(B)}(t)\rangle$ minimizes the convex function $||xc_b - Q^{(B)}(t_0)||$ over $x \in \mathbb{R}$, i.e.,

$$||Q^{max}(t_0)\sqrt{M_b}c_b - Q^{(B)}(t_0)|| \geq ||\langle c_b, Q^{(B)}(t)\rangle c_b - Q^{(B)}(t_0)|| = ||Q_\perp^{(B)}(t_0)||.$$

Hence

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} G(t)|Z^{(B)}(t_0), t^* < t_0+K\right] &\leq -(t_0+T-t^*)\frac{\lambda_0}{2}||Q_\perp^{(B)}(t_0)|| + F_4\\ &\leq -(J-1)K\frac{\lambda_0}{2}||Q_\perp^{(B)}(t_0)|| + F_4\end{aligned}$$

Denote $Pr(t^* \geq t_0+K|Z(t_0))$ by P_K. Combining the bounds we have

$$\begin{aligned}\mathbb{E}\left[\Delta V(Z^{(B)})|Z^{(B)}(t_0)\right] &\leq \frac{1}{\left|\left|Q_\perp^{(B)}(t_0)\right|\right|}(\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G(t)|Z^{(B)}(t_0)\right] + C_2)\\ &\leq -F_5 + \frac{F_6}{\left|\left|Q_\perp^{(B)}(t_0)\right|\right|}, \quad (56)\end{aligned}$$

where $F_5 = P_KhT + (1-P_K)hK - \lambda_0(1-P_K)(J-1)K$ and $F_6 = C_2 + P_KF_1T + (1-P_K)F_0K + (1-P_K)F_4$. As $\lim_{K,J\to\infty} F_5 = +\infty$, for any $\theta > 0$, there exist large enough K and J such that $-F_5 < \theta$. Pick any δ with

$0 < \delta < \theta$ and let $\zeta = \frac{F_6}{\theta - \delta}$. Then $\mathbb{E}\left[\Delta V(Z^{(B)}) | Z^{(B)}(t_0)\right] \leq -\delta$ for all $Z^{(B)}$ with $V(Z^{(B)}) \geq \zeta$. This means that the drift of $V(Z^{(B)})$ is negative for sufficiently large $V(Z^{(B)})$, as the constants δ and ζ do not depend on ϵ. Therefore there exists a sequence of constants $\{C_r : r \in \mathbb{N}\}$ such that $\mathbb{E}\left[\left|\left|Q_{\perp}^{(\epsilon)(B)}\right|\right|^r\right] \leq C_r$ for each $r = 1, 2, \cdots$.

■

7.8 Proof of Lemma 16-19

7.8.1 Proof of Lemma 16

By the definition of $U_m(t)$, if $U_m(t) > 0$, the number of tasks in queue m must be less than the number of available servers scheduled to this queue at time slot t. Since $S_m(t) \leq M$, $Q_m(t) \leq Q_m(t) + A_m(t) < M$. Note that $U_m(t) \leq M$ If $U_m(t) = 0$, $Q_m(t)U_m(t) = 0$. Hence $Q_m(t)U_m(t) < MU_m(t)$. Therefore, $\langle Q(t), U(t)\rangle < \sum_{m\in\mathcal{M}} MU_m(t) = M^2$ ■

7.8.2 Proof of Lemma 17

By the queue dynamics,

$$\begin{aligned} & \left|\left|Q_{||}^{(B)}(t+1)\right|\right|^2 - \left|\left|Q_{||}^{(B)}(t)\right|\right|^2 \\ = \; & \langle c_b, A^{(B)}(t) - S^{(B)}(t)) + U^{(B)}(t)\rangle^2 + 2\langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle + 2\langle c_b, Q^{(B)}(t)\rangle\langle c_b, U^{(B)}(t)\rangle \\ \geq \; & 2\langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle \end{aligned}$$

7.8.3 Proof of Lemma 18

Note that $(Q^{(B)}(t) - Q^{(B)}(t_0))_{\perp} = Q_{\perp}^{(B)}(t) - Q_{\perp}^{(B)}(t_0)$. By the boundedness of arrivals and service, we have

$$\left|\left|\left|Q_{\perp}^{(B)}(t)\right|\right| - \left|\left|Q_{\perp}^{(B)}(t_0)\right|\right|\right| \leq \left|\left|Q_{\perp}^{(B)}(t) - Q_{\perp}^{(B)}(t_0)\right|\right| \leq ||Q^{(B)}(t) - Q^{(B)}(t_0)|| \leq T\sqrt{M_b}\max\{M, C_A\}.$$

7.8.4 Proof of Lemma 19

By Cauchy-Schwartz inequality,

$$\begin{aligned} G(t) &= \langle Q^{(B)}(t), A^{(B)}(t) - S^{(B)}(t)\rangle - \langle c_b, Q^{(B)}(t)\rangle\langle c_b, A^{(B)}(t) - S^{(B)}(t)\rangle \\ &= \langle Q_{\perp}^{(B)}(t), A_{\perp}^{(B)}(t) - S_{\perp}^{(B)}(t)\rangle \\ &\leq ||Q_{\perp}^{(B)}(t)|| \cdot ||A_{\perp}^{(B)}(t) - S_{\perp}^{(B)}(t)|| \\ &\leq ||Q_{\perp}^{(B)}(t)|| \cdot ||A^{(B)}(t) - S^{(B)}(t)|| \end{aligned} \tag{57}$$

By the boundedness of arrivals and service, $||A^{(B)}(t) - S^{(B)}(t)|| \leq \sqrt{M_b}\max\{M, C_A\}$. From Lemma 18, we have

$$\begin{aligned} G(t) &\leq (Q^{(B)}(t_0) + T\sqrt{M_b}\max\{M, C_A\}) \cdot \sqrt{M_b}\max\{M, C_A\} \\ &= h\left|\left|Q_{\perp}^{(B)}(t_0)\right|\right| + F_0, \end{aligned} \tag{58}$$

where $h = \sqrt{M_b}\max\{M, C_A\}$ and $F_0 = M_b T(\max\{M, C_A\})^2$ are two constants.

7.9 Proof of Theorem 5

Proof.

Under the *ideal arrival process*, shared type tasks that join beneficiaries queues are re-distributed among its helper local servers evenly.

Hence

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{A}_m(t)\right] = \mathbb{E}\left[\sum_{\bar{L}\in\mathcal{L}_b}A_{\bar{L}}(t)\right] = \sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}} = B\alpha + H\gamma - \gamma\sum_{m\in\mathcal{H}}\rho_m^* - \epsilon$$

Let ρ_m denote the proportion of time server m spends on serving local queue m in steady state. Then

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{S}_m(t)\right] = M_b\alpha + \sum_{m\in\mathcal{H}}\gamma(1-\rho_m) = M_b\alpha + H\gamma - \gamma\sum_{m\in\mathcal{H}}\rho_m$$

$$\mathbb{E}\left[\sum_{m\in\mathcal{H}}\hat{S}_m(t)\right] = \sum_{m\in\mathcal{H}}\alpha\rho_m = \alpha\sum_{m\in\mathcal{H}}\rho_m$$

And

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{S}_m(t) - \sum_{m\in\mathcal{B}}\hat{A}_m(t)\right] = \epsilon + \gamma(\sum_{m\in\mathcal{H}}\rho_m^* - \sum_{m\in\mathcal{H}}\rho_m) = \epsilon + \delta,$$

where $\delta = \gamma(\sum_{m\in\mathcal{H}}\rho_m^* - \sum_{m\in\mathcal{H}}\rho_m)$.

For convenience, we temporarily omit the superscript (ϵ). For any time slot t, we analyze each term in Lemma 14 with respect to the collapse direction c defined in (13).

$$\begin{aligned}
& \mathbb{E}\left[\langle c, Q(t)\rangle\langle c, \hat{S}(t) - \hat{A}(t)\rangle\right] \\
= \quad & \frac{1}{M_b}\mathbb{E}\left[\left(\sum_{m\in\mathcal{B}}Q_m(t)\right)\left(\sum_{m\in\mathcal{B}}\hat{S}_m(t) - \sum_{m\in\mathcal{B}}\hat{A}_m(t)\right)\right] \\
= \quad & \frac{1}{M_b}\mathbb{E}\left[\sum_{m\in\mathcal{B}}Q_m(t)\right]\mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{S}_m(t) - \sum_{m\in\mathcal{B}}\hat{A}_m(t)\right] \qquad (59) \\
= \quad & \frac{\epsilon+\delta}{M_b}\mathbb{E}\left[\sum_{m\in\mathcal{B}}Q_m(t)\right] \qquad (60)
\end{aligned}$$

where (59) follows from the fact that the total arrivals of tasks that are only local to beneficiaries are do not depend on the beneficiary queue-lengths, so as the ideal service process for beneficiary queues; (60) follows from the definition of the ideal arrivals and service.

It is easy to verify that

$$Var(\sum_{m\in\mathcal{B}}\hat{A}_m(t)) = Var(\sum_{\bar{L}\in\mathcal{L}_b}A_{\bar{L}}(t)) = (\sigma_b^{(\epsilon)})^2 \qquad (61)$$

$$Var(\sum_{m\in\mathcal{B}}\hat{S}_m(t)) = M_b\alpha(1-\alpha) + \sum_{m\in\mathcal{H}}\gamma(1-\rho_m)(1-\gamma(1-\rho_m)) + g(\delta) = (\nu_b^{(\epsilon)})^2. \qquad (62)$$

As $\{\hat{A}(t)\}$ and $\{\hat{S}(t)\}$ are independent, and

$$\mathbb{E}\left[< c, \hat{S} >\right] - \mathbb{E}\left[< c, \hat{A} >\right] = \frac{\epsilon+\delta}{\sqrt{M_b}}, \qquad (63)$$

we have

$$\mathbb{E}\left[\langle c, \hat{A}(t) - \hat{S}(t)\rangle^2\right] = \frac{1}{M_b}((\sigma_b^{(\epsilon)})^2 + (\nu_b^{(\epsilon)})^2 + (\epsilon + \delta_b)^2) \tag{64}$$

Since $Q(t)$ is in steady state, $\mathbb{E}\left[\langle c, \hat{A}(t) - \hat{S}(t) + \hat{U}(t)\rangle\right] = \mathbb{E}\left[\langle c, Q(t+1)\rangle - \langle c, Q(t)\rangle\right] = 0$. Thus $\mathbb{E}\left[\langle c, \hat{U}(t)\rangle\right] = \mathbb{E}\left[\langle c, \hat{S}(t) - \hat{A}(t)\rangle\right] = \frac{\epsilon+\delta}{\sqrt{M_b}}$. Meanwhile,

$$\langle c, \hat{U}(t)\rangle = \langle c, \hat{S}(t) - S(t) + A(t) - \hat{A}(t) + U(t)\rangle \leq \langle c, \hat{S}(t) + A(t) + U(t)\rangle \leq \frac{2M + C_A}{\sqrt{M_b}}$$

By the coupling of $\{\hat{S}(t), t \geq 0\}$ and $\{S(t), t \geq 0\}$, $\langle c, \hat{S}(t) - S(t)\rangle \geq 0$. In addition, $\langle c, A(t) - \hat{A}(t)\rangle \geq 0$. Hence $\langle c, \hat{U}(t)\rangle \geq 0$.

Therefore

$$\mathbb{E}\left[\langle c, \hat{U}(t)\rangle^2\right] \leq \frac{2M + C_A}{\sqrt{M_b}}\mathbb{E}\left[\langle c, \hat{U}(t)\rangle\right] = \frac{(2M + C_A)(\epsilon + \delta)}{M_b}$$

Finally we bound the term (17).

$$\begin{aligned}\mathbb{E}\left[\langle c, Q(t) + \hat{A}(t) - \hat{S}(t)\rangle\langle c, \hat{U}(t)\rangle\right] &= \mathbb{E}\left[\langle c, Q(t)\rangle\langle c, \hat{U}(t)\rangle\right] + \mathbb{E}\left[\langle c, \hat{A}(t) - \hat{S}(t)\rangle\langle c, \hat{U}(t)\rangle\right] \\ &\leq \mathbb{E}\left[\langle c, Q(t)\rangle\langle c, \hat{U}(t)\rangle\right] + \frac{(2M + C_A)(\epsilon + \delta)}{M_b} \end{aligned} \tag{65}$$

Note that

$$\begin{aligned}\langle c, Q(t)\rangle\langle c, \hat{U}(t)\rangle &= \langle Q(t), \hat{U}(t)\rangle - \langle Q_\perp(t), \hat{U}_\perp(t)\rangle \\ &= \langle Q(t), \hat{S}(t) - S(t)\rangle + \langle Q(t), A(t) - \hat{A}(t)\rangle + \langle Q(t), U(t)\rangle - \langle Q_\perp(t), \hat{U}_\perp(t)\rangle \end{aligned} \tag{66}$$

The following lemmas bound the terms in Eq. (66).

Lemma 20 $\mathbb{E}\left[||\hat{U}(t)||^2\right] \leq R_3\epsilon$, *where* R_3 *is a constant that doesn't depend on* ϵ.

Lemma 21 $\mathbb{E}\left[\langle Q(t), \hat{S}(t) - S(t)\rangle\right] \leq o(\epsilon) + R_1\sqrt{M}\mathbb{E}\left[\langle e, \hat{S}(t) - S(t)\rangle\right]$, *where* $e = \frac{1}{\sqrt{M}}\underbrace{(1, 1, \cdots, 1)}_{M}$ *and* $R_1 > 0$ *is a constant.*

Lemma 22 $\langle Q(t), A(t) - \hat{A}(t)\rangle \leq 0$

Lemma 23 $\langle Q(t), U(t)\rangle \leq M\sqrt{M}\langle e, U(t)\rangle$, *where* $e = \frac{1}{\sqrt{M}}\underbrace{(1, 1, \cdots, 1)}_{M}$.

Let $R_2 = \max\{R_1, M\}$, then

$$\begin{aligned}\langle Q(t), \hat{S}(t) - S(t)\rangle + \langle Q(t), U(t)\rangle &\leq R_1'\epsilon + R_2\sqrt{M}\mathbb{E}\left[\langle e, \hat{S}(t) - S(t)\rangle + \langle e, U(t)\rangle\right] \\ &= o(\epsilon) + R_2\sqrt{M}\mathbb{E}\left[\langle e, \hat{S}(t) - S(t) + U(t)\rangle\right] \\ &= o(\epsilon) + R_2(\epsilon - \frac{\alpha - \gamma}{\alpha}\delta) \end{aligned}$$

The last equality follows from the fact that $Q(t)$ is in steady state. Thus $\mathbb{E}\left[\langle e, A(t) - S(t) + U(t)\rangle\right] = \mathbb{E}\left[\langle E, Q(t+1)\rangle - \langle E, Q(t)\rangle\right] = 0$. This implies that $\mathbb{E}\left[\langle e, \hat{S}(t) - S(t) + U(t)\rangle\right] = \mathbb{E}\left[\langle e, \hat{S}(t) - A(t)\rangle\right] = \frac{1}{\sqrt{M}}(\epsilon - \frac{\alpha-\gamma}{\alpha}\delta)$.

We use the state space collapse result to bound $-\langle Q_\perp(t), \hat{U}_\perp(t)\rangle$.

$$\mathbb{E}\left[-\langle Q_\perp(t), \hat{U}_\perp(t)\rangle\right] \leq \sqrt{\mathbb{E}\left[||Q_\perp(t)||^2\right]\mathbb{E}\left[||\hat{U}(t)||^2\right]} \quad (67)$$

$$\leq \sqrt{C_2\mathbb{E}\left[||\hat{U}(t)||^2\right]} \quad (68)$$

$$\leq \sqrt{C_2R_3\epsilon} \quad (69)$$

Combining these inequalities, we can bound the term (17) as

$$\begin{aligned}&\mathbb{E}\left[\langle c, Q(t) + \hat{A}(t) - \hat{S}(t)\rangle\langle c, \hat{U}(t)\rangle\right]\\ \leq\ & (\epsilon+\delta)\frac{2M+C_A}{M_b} + o(\epsilon) + R_2(\epsilon - \frac{\alpha-\gamma}{\alpha}\delta) + \sqrt{C_2R_3\epsilon}\end{aligned}$$

Now we revive the superscript $^{(\epsilon)}$. Combining the inequalities we have

$$\begin{aligned}2\frac{\epsilon+\delta}{M_b}\mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m(t)\right] \leq\ & \frac{1}{M_b}((\sigma_b^{(\epsilon)})^2 + (\nu_b^{(\epsilon)})^2 + (\epsilon+\delta)^2)\\ & +(\epsilon+\delta)\frac{2M+C_A}{M_b} + o(\epsilon) + R_2(\epsilon - \frac{\alpha-\gamma}{\alpha}\delta) + \sqrt{C_2R_3\epsilon}\end{aligned}$$

Note that $\delta = \gamma(\sum_{m\in\mathcal{H}}\rho_m - \frac{1}{\alpha}\sum_{\bar{L}} \in \mathcal{L}_{\mathcal{H}}^*\lambda_{\bar{L}}) > 0$, and $\delta = o(\epsilon)$. Thus,

$$2\frac{\epsilon}{M_b}\mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m(t)\right] \leq 2\frac{\epsilon+\delta}{M_b}\mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m(t)\right]$$

Therefore

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}} Q_m(t)\right] \leq \frac{((\sigma_b^{(\epsilon)})^2 + (\nu_b^{(\epsilon)})^2}{2(\epsilon+\delta)} + D^{(\epsilon)},$$

where

$$D^{(\epsilon)} = \frac{\epsilon+\delta}{2} + \frac{(2M+C_A)}{2} + \frac{M_b o(\epsilon)}{2\epsilon} + M_bR_2\frac{\epsilon}{2(\epsilon+\delta)} - \frac{M_bR_2(\alpha-\gamma)}{2\alpha}\frac{\delta}{\epsilon+\delta} + M_b\sqrt{C_2R_3}\frac{\sqrt{\epsilon}}{2(\epsilon+\delta)}$$

As $\epsilon \downarrow 0^+$, $\delta \downarrow 0^+$, we have $\limsup_{\epsilon\to 0^+} D^{(\epsilon)} = 0$. Thus $D^{(\epsilon)} = o(\frac{1}{\epsilon})$.

■

7.10 Proof of Lemma 21 -23

7.10.1 Proof of Lemma 20.

Proof.

We will show that

$$\mathbb{E}\left[||\hat{S}(t) - S(t)||^2\right] \leq C_1\epsilon$$
$$\mathbb{E}\left[||A(t) - \hat{A}(t)||^2\right] \leq C_2\epsilon$$
$$\mathbb{E}\left[||U(t)||^2\right] \leq C_3\epsilon,$$

where C_1, C_2, C_3 are constants independent of ϵ.

Thus

$$\begin{aligned}\mathbb{E}\left[||\hat{U}(t)||^2\right] &= \mathbb{E}\left[||\hat{S}(t)-S(t)+A(t)-\hat{A}(t)+U(t)||^2\right]\\ &\leq \mathbb{E}\left[||\hat{S}(t)-S(t)||^2\right]+\mathbb{E}\left[||A(t)-\hat{A}(t)||^2\right]+\mathbb{E}\left[||U(t)||^2\right]\\ &\leq (C_1+C_2+C_3)\epsilon\end{aligned}$$

For $\forall m \in \mathcal{B}$,

$$\begin{aligned}&\hat{S}_m(t)-S_m(t)\\ =\;& X_m^l(t)+\sum_{n\in\mathcal{H}} X_n^r(t)\cdot I_{\{\hat{\eta}_n(t)=m\}}-S_m^l(t)-\sum_{n:n\neq m} R_n(t)I_{\{\eta_n(t)=m\}}\\ =\;& X_m^l(t)(1-I_{\{\eta_m(t)=m\}})+\sum_{n\in\mathcal{H}} X_n^r(t)(I_{\{\hat{\eta}_n(t)=m\}}-I_{\{\eta_n(t)=m\}})-\sum_{\substack{n\in\mathcal{B}\\ n\neq m}} R_n(t)I_{\{\eta_n(t)=m\}}\end{aligned}$$

It can be seen that $-(M_b+1)\leq \hat{S}_m(t)-S_m(t)\leq 1+H$. So $|\hat{S}_m(t)-S_m(t)|\leq M$.
For $\forall m\in\mathcal{H}$,

$$\hat{S}_m(t)-S_m(t)=-\sum_{n:n\neq m} R_n(t)I_{\{\eta_n(t)=m\}}.$$

It is obvious that $0\geq \hat{S}_m(t)-S_m(t)\geq -(M-1)>-M$.

Let $N_{bb}(t)(N_{bh}(t))$ denote the number of servers in $\mathcal{B}$ that are scheduled to serve remote beneficiary(helper) queues at time slot t. Similarly define $N_{hb}(t)(N_{hb}(t))$ as the number of servers in $\mathcal{H}$ that are scheduled to serve remote beneficiary(helper) queues at time slot t.

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{B}}(\hat{S}_m(t)-S_m(t))|Z(t)\right] &= \alpha(N_{bb}(t)+N_{bh}(t))+\gamma N_{hh}(t)-\gamma N_{bb}(t) \qquad (70)\\ \mathbb{E}\left[\sum_{m\in\mathcal{H}}(\hat{S}_m(t)-S_m(t))|Z(t)\right] &= -\gamma N_{bh}(t))-\gamma N_{hh}(t)\end{aligned}$$

And

$$\mathbb{E}\left[\sum_{m=1}^{M}(\hat{S}_m(t)-S_m(t))|Z(t)\right]=(\alpha-\gamma)(N_{bb}(t)+N_{bh}(t)). \qquad (71)$$

By taking expectation over $Z(t)$ on both sides of Eq. (70) and (71), we have

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{B}} S_m(t)\right] &= \mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{S}_m(t)\right]-(\alpha-\gamma)\mathbb{E}\left[N_{bb}(t)\right]-\alpha\mathbb{E}\left[N_{bh}(t)\right]-\gamma\mathbb{E}\left[N_{hh}(t)\right]\\ \mathbb{E}\left[\sum_{m=1}^{M} S_m(t)\right] &= \mathbb{E}\left[\sum_{m=1}^{M}\hat{S}_m(t)\right]-(\alpha-\gamma)\left(\mathbb{E}\left[N_{bb}(t)\right]+\mathbb{E}\left[N_{bh}(t)\right]\right)\end{aligned}$$

Since we consider steady state,

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{B}} S_m(t)\right] &= \mathbb{E}\left[\sum_{m\in\mathcal{B}} A_m(t)\right]+\mathbb{E}\left[\sum_{m\in\mathcal{B}} U_m(t)\right]\geq \mathbb{E}\left[\sum_{m\in\mathcal{B}} A_m(t)\right]\geq \sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}}\\ \mathbb{E}\left[\sum_{m=1}^{M} S_m(t)\right] &= \mathbb{E}\left[\sum_{m=1}^{M} A_m(t)\right]+\mathbb{E}\left[\sum_{m=1}^{M} U_m(t)\right]\geq \mathbb{E}\left[\sum_{m=1}^{M} A_m(t)\right]=\sum_{\bar{L}\in\mathcal{L}}\lambda_{\bar{L}}\end{aligned}$$

Then

$$(\alpha-\gamma)\mathbb{E}\left[N_{bb}(t)\right]+\alpha\mathbb{E}\left[N_{bh}(t)\right]+\gamma\mathbb{E}\left[N_{hh}(t)\right]\leq\mathbb{E}\left[\sum_{m\in\mathcal{B}}\hat{S}_m(t)\right]-\sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}}=\epsilon+\delta \tag{72}$$

$$(\alpha-\gamma)\left(\mathbb{E}\left[N_{bb}(t)\right]+\mathbb{E}\left[N_{bh}(t)\right]\right)\leq\mathbb{E}\left[\sum_{m=1}^{M}\hat{S}_m(t)\right]-\sum_{\bar{L}\in\mathcal{L}}\lambda_{\bar{L}}=\epsilon-\frac{\alpha-\gamma}{\gamma}\delta \tag{73}$$

Eliminate δ by multiplying Eq. (73) with $\frac{\gamma}{\alpha-\gamma}$ and adding it to Eq. (72), we have

$$(\alpha+\gamma)\mathbb{E}\left[N_{bh}(t)\right]+\alpha\mathbb{E}\left[N_{bb}(t)\right]+\gamma\mathbb{E}\left[N_{hh}(t)\right]\leq\frac{\alpha}{\alpha-\gamma}\epsilon$$

Therefore

$$\begin{aligned}\mathbb{E}\left[N_{bh}(t)\right]&\leq\frac{\alpha}{(\alpha+\gamma)(\alpha-\gamma)}\epsilon\\ \mathbb{E}\left[N_{bb}(t)\right]&\leq\frac{\epsilon}{\alpha-\gamma}\\ \mathbb{E}\left[N_{hh}(t)\right]&\leq\frac{\alpha}{\gamma(\alpha-\gamma)}\epsilon\end{aligned}$$

Let $\mathcal{M}_{bn}(t)=\{m\in\mathcal{B}:\hat{S}_m(t)-S_m(t)\leq 0\}$ and $\mathcal{M}_{bp}(t)=\{m\in\mathcal{B}:\hat{S}_m(t)-S_m(t)>0\}$.
Note that

$$\begin{aligned}&\mathbb{E}\left[\sum_{m\in\mathcal{M}_{bn}}(\hat{S}_m(t)-S_m(t))|Z(t)\right]\\ \geq\;&\mathbb{E}\left[\sum_{m\in\mathcal{M}_{bn}}\sum_{\substack{n\in\mathcal{B}\\ n\neq m}}-R_n(t)I_{\{\eta_n(t)=m\}}|Z(t)\right]\\ \geq\;&-\mathbb{E}\left[\sum_{n\in\mathcal{B}}R_n(t)I_{\{\eta_n(t)\neq n\ \text{ and }\ \eta_n(t)\in\mathcal{B}\}}|Z(t)\right]\\ =\;&-\gamma N_{bb}(t)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}(\hat{S}_m(t)-S_m(t))^2|Z(t)\right]&\leq\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}M|\hat{S}_m(t)-S_m(t)||Z(t)\right]\\ &=M\mathbb{E}\left[\sum_{m\in\mathcal{M}_{bp}}(\hat{S}_m(t)-S_m(t))-\sum_{m\in\mathcal{M}_{bn}}(\hat{S}_m(t)-S_m(t))|Z(t)\right]\\ &=M\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}(\hat{S}_m(t)-S_m(t))-2\sum_{m\in\mathcal{M}_{bn}}(\hat{S}_m(t)-S_m(t))|Z(t)\right]\\ &\leq M(\alpha-\gamma)(N_{bb}(t)+N_{bh}(t))+2M\gamma N_{bb}(t)\\ &=M(\alpha+\gamma)N_{bb}(t)\end{aligned}$$

As for $\forall m\in\mathcal{H}$, $M\leq\hat{S}_m(t)-S_m(t)\leq 0$,

$$\begin{aligned}\mathbb{E}\left[\sum_{m\in\mathcal{M}_h}(\hat{S}_m(t)-S_m(t))^2|Z(t)\right]&\leq\mathbb{E}\left[\sum_{m\in\mathcal{M}_h}-M(\hat{S}_m(t)-S_m(t))|Z(t)\right]\\ &=M\gamma(N_{bh}(t)+N_{hh}(t))\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\sum_{m=1}^{M}(\hat{S}_m(t)-S_m(t))^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{m=1}^{M}(\hat{S}_m(t)-S_m(t))^2\right]|Z(t)\right] \\ &\leq M(\alpha+\gamma)\mathbb{E}\left[N_{bb}(t)\right]+M\gamma(\mathbb{E}\left[N_{bh}(t)\right]+\mathbb{E}\left[N_{hh}(t)\right]) \\ &\leq C_1\epsilon, \end{aligned} \tag{74}$$

where $C_1=\frac{2\alpha^2+\gamma^2+4\alpha\gamma}{\alpha^2-\gamma^2}$ is a constant that doesn't depend on ϵ.

Next we will show that $\mathbb{E}\left[||A(t)-\hat{A}(t)||^2\right]\leq C_2\epsilon$. Note that $\forall m\in\mathcal{M}$, $|A_m(t)-\hat{A}_m(t)|\leq C_A$. In particular, for $\forall m\in\mathcal{B}$, $A_m(t)-\hat{A}_m(t)\geq 0$ and for $\forall m\in\mathcal{H}$, $A_m(t)-\hat{A}_m(t)\leq 0$. Let $A_s^b(t)$ denote the total amount of shared tasks that are routed to beneficiary queues at time slot t. Then we have

$$\begin{aligned}\mathbb{E}\left[||A(t)-\hat{A}(t)||^2\right] &= \mathbb{E}\left[\sum_{m=1}^{M}(A_m(t)-\hat{A}_m(t))^2\right] \\ &\leq \mathbb{E}\left[\sum_{m=1}^{M}C_A|A_m(t)-\hat{A}_m(t)|\right] \\ &= C_A\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}(A_m(t)-\hat{A}_m(t))-\sum_{m\in\mathcal{M}_h}(A_m(t)-\hat{A}_m(t))\right] \\ &= 2C_A\mathbb{E}\left[A_s^b(t)\right]\end{aligned}$$

We will give expressions for the expected amount of service received by helpers and beneficiaries respectively in terms of $N_{bh}(t),N_{bb}(t),N_{hb}(t),N_{hh}(t)$ given $Z(t)$.

$$\mathbb{E}\left[\sum_{m\in\mathcal{M}_h}S_m(t)|Z(t)\right]=\alpha(M_h-N_{hh}(t)-N_{hb}(t))+\gamma N_{hh}(t)+\gamma N_{bh}(t)\leq\alpha M_h-\alpha N_{hb}(t))+\gamma N_{bh}(t) \tag{75}$$

$$\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}S_m(t)|Z(t)\right]=\alpha(M_b-N_{bh}(t)-N_{bb}(t))+\gamma N_{hb}(t)+\gamma N_{bb}(t)\leq\alpha M_b-\alpha N_{bh}(t))+\gamma N_{hb}(t) \tag{76}$$

By eliminating $N_{hb}(t)$ on the right hand sides of Eq. (75) and (76), we have

$$\frac{\gamma}{\alpha}\mathbb{E}\left[\sum_{m\in\mathcal{M}_h}S_m(t)|Z(t)\right]+\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}S_m(t)|Z(t)\right]\leq\alpha M_b+\gamma M_h-\frac{1}{\alpha}(\alpha^2-\gamma^2)N_{bh}(t)\leq\alpha B+\gamma H$$

Thus

$$\frac{\gamma}{\alpha}\mathbb{E}\left[\sum_{m\in\mathcal{M}_h}S_m(t)\right]+\mathbb{E}\left[\sum_{m\in\mathcal{M}_b}S_m(t)\right]\leq\alpha B+\gamma H$$

As we consider steady state,

$$\mathbb{E}\left[\sum_{m\in\mathcal{H}}S_m(t)\right]=\mathbb{E}\left[\sum_{m\in\mathcal{H}}A_m(t)\right]+\mathbb{E}\left[\sum_{m\in\mathcal{H}}U_m(t)\right]\geq\mathbb{E}\left[\sum_{m\in\mathcal{H}}A_m(t)\right]=\sum_{\bar{L}\in\mathcal{L}_h\cup\mathcal{L}_s}\lambda_{\bar{L}}-\mathbb{E}\left[A_s^b(t)\right]$$

$$\mathbb{E}\left[\sum_{m\in\mathcal{B}}S_m(t)\right]=\mathbb{E}\left[\sum_{m\in\mathcal{B}}A_m(t)\right]+\mathbb{E}\left[\sum_{m\in\mathcal{B}}U_m(t)\right]\geq\mathbb{E}\left[\sum_{m\in\mathcal{B}}A_m(t)\right]=\sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}}+\mathbb{E}\left[A_s^b(t)\right]$$

Combining the inequalities, we have

$$\begin{aligned}\alpha M_b+\gamma M_h &\geq \frac{\gamma}{\alpha}\sum_{\bar{L}\in\mathcal{L}_h\cup\mathcal{L}_s}\lambda_{\bar{L}}-\frac{\gamma}{\alpha}\mathbb{E}\left[A_s^b(t)\right]+\sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}}+\mathbb{E}\left[A_s^b(t)\right] \\ &= \alpha M_b+\gamma M_h-\epsilon+\frac{\alpha-\gamma}{\alpha}\mathbb{E}\left[A_s^b(t)\right]\end{aligned}$$

Hence

$$\mathbb{E}\left[A_s^b(t)\right] \leq \frac{\alpha-\gamma}{\alpha}\epsilon.$$

Therefore

$$\mathbb{E}\left[||A(t)-\hat{A}(t)||^2\right] \leq 2C_A \cdot \frac{\alpha-\gamma}{\alpha}\epsilon = C_2\epsilon,$$

where $C_2 = 2C_A\frac{\alpha-\gamma}{\alpha}$ is a constant.

Now consider the term $\mathbb{E}\left[||U(t)||^2\right]$.

Since $0 \leq U_m(t) \leq M$, $\mathbb{E}\left[||U(t)||^2\right] \leq M\mathbb{E}\left[\sum_{m=1}^{M} U_m(t)\right]$. As the queueing process is in steady state,

$$\mathbb{E}\left[\sum_{m=1}^{M} U_m(t)\right] = \mathbb{E}\left[\sum_{m=1}^{M} S_m(t)\right] - \mathbb{E}\left[\sum_{m=1}^{M} A_m(t)\right]$$

From Eq. (76), we have

$$\begin{aligned}
\mathbb{E}\left[N_{hb}(t)\right] &\geq \frac{1}{\gamma}\left(\mathbb{E}\left[\sum_{m\in\mathcal{M}_b} S_m(t)\right] - \alpha M_b + \alpha\mathbb{E}\left[N_{bh}(t)\right]\right) \\
&\geq \frac{1}{\gamma}\left(\sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}} - \alpha M_b + \alpha\mathbb{E}\left[N_{bh}(t)\right]\right)
\end{aligned}$$

By Eq. (75) and (76),

$$\begin{aligned}
\mathbb{E}\left[\sum_{m=1}^{M} S_m(t)\right] &\leq \alpha M_b + \gamma M_h - (\alpha-\gamma)\mathbb{E}\left[N_{bh}(t)\right] - (\alpha-\gamma)\mathbb{E}\left[N_{hb}(t)\right] \\
&\leq \alpha M_b + \gamma M_h - (\alpha-\gamma)\mathbb{E}\left[N_{bh}(t)\right] - \frac{\alpha-\gamma}{\gamma}\left(\sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}} - \alpha M_b + \alpha\mathbb{E}\left[N_{bh}(t)\right]\right) \\
&\leq \alpha M_b + \gamma M_h - \frac{\alpha-\gamma}{\gamma}\left(\sum_{\bar{L}\in\mathcal{L}_b}\lambda_{\bar{L}} - \alpha M_b\right) \\
&= \sum_{\bar{L}\in\mathcal{L}}\lambda_{\bar{L}} + \frac{\alpha}{\gamma}\epsilon
\end{aligned}$$

Therefore

$$\mathbb{E}\left[||U(t)||^2\right] \leq \frac{\alpha M}{\gamma}\epsilon. \tag{77}$$

■

7.10.2 Proof of Lemma 21.

$$
\begin{aligned}
& \langle Q(t), \hat{S}(t) - S(t) \rangle \\
= & \sum_{m \in \mathcal{H}} Q_m(t)(\hat{S}_m(t) - S_m(t)) + \sum_{m \in \mathcal{B}} Q_m(t)(\hat{S}_m(t) - S_m(t)) \\
= & \sum_{m \in \mathcal{H}} Q_m(t) \left(- \sum_{n:n \neq m} R_n(t) I_{\eta_n(t)=m} \right) \\
+ & \sum_{m \in \mathcal{B}} Q_m(t) \left(X_m^l(t) + \sum_{n \in \mathcal{H}} X_n^r(t) \cdot I_{\{\hat{\eta}_n(t)=m\}} - S_m^l(t) - \sum_{n:n \neq m} R_n(t) I_{\{\eta_n(t)=m\}} \right) \\
= & \sum_{m \in \mathcal{H}} \left(X_m^r(t) \sum_{n \in \mathcal{B}} Q_n(t) I_{\{\hat{\eta}_m(t)=n\}} - R_m(t) \sum_{n:n \neq m} Q_n(t) I_{\{\eta_m(t)=n\}} \right) \qquad (78) \\
+ & \sum_{m \in \mathcal{B}} \left(Q_m(t)(X_m^l(t) - S_m^l(t)) - R_m(t) \sum_{n:n \neq m} Q_n(t) I_{\{\eta_m(t)=n\}} \right) \qquad (79)
\end{aligned}
$$

By the coupling of $\{X_m^r(t), t \geq 0\}$ with $\{R_m(t), t \geq 0\}$, the expectation of term (78) can be simplified as

$$
\gamma \sum_{m \in \mathcal{H}} \mathbb{E} \left[\sum_{n \in \mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}} - I_{\{\eta_m(t)=n\}}) - \sum_{\substack{n:n \neq m \\ n \in \mathcal{H}}} Q_n(t) I_{\{\eta_m(t)=n\}} \right]
$$

Consider the random variable τ_m^t, which is the last time slot before t at which server m makes a scheduling decision. Then

$$
\begin{aligned}
& \mathbb{E} \left[\sum_{n \in \mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}} - I_{\{\eta_m(t)=n\}}) - \sum_{\substack{n:n \neq m \\ n \in \mathcal{H}}} Q_n(t) I_{\{\eta_m(t)=n\}} \right] \qquad (80) \\
= & \sum_{n=1}^{t} \mathbb{E} \left[\sum_{n \in \mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}} - I_{\{\eta_m(t)=n\}}) - \sum_{\substack{n:n \neq m \\ n \in \mathcal{H}}} Q_n(t) I_{\{\eta_m(t)=n\}} | \tau_m^t = t - n \right] \cdot Pr(\tau_m^t = t - n)
\end{aligned}
$$

For a particular $\tau_m^t = t - n$, we decompose the probability space based on $Q_m(\tau_m^t)$.

Case (i): $Q_m(\tau_m^t) > 0$
Under the proposed algorithm, $\eta_m(\tau_m^t) = m$ when $Q_m(\tau_m^t) > 0$. Hence

$$
\mathbb{E} \left[\sum_{n \in \mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}} - I_{\{\eta_m(t)=n\}}) - \sum_{\substack{n:n \neq m \\ n \in \mathcal{H}}} Q_n(t) I_{\{\eta_m(t)=n\}} | Q_m(\tau_m^t) > 0, \tau_m^t = t - n \right] = 0.
$$

Case (ii): $Q_m(\tau_m^t) = 0$
Let $Q_b^{max}(t) = \max_{m:m \in \mathcal{B}} \{Q_m(t)\}$ and $Q_h^{max}(t) = \max_{m:m \in \mathcal{H}} \{Q_m(t)\}$. Under the proposed algorithm, $\eta_m(\tau_m^t) = \arg\max_{n:n \neq m} \{Q_n(\tau_m^t)\}$ if $Q_m(\tau_m^t) = 0$. We further decompose the probability space based on the values of $Q_b^{max}(\tau_m^t)$ and $Q_h^{max}(\tau_m^t)$.

Observe that if $Q_b^{max}(\tau_m^t) > Q_h^{max}(\tau_m^t)$, $\hat{\eta}_m(\tau_m^t) = \eta_m(\tau_m^t) = Q_b^{max}(\tau_m^t)$. Hence the expectation in Eq.(80) is equal to zero under this case.

If $Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t)$, $\eta_m(\tau_m^t) = Q_b^{max}(\tau_m^t)$ and $\hat{\eta}_m(\tau_m^t) = Q_b^{max}(\tau_m^t)$. By the boundedness of arrivals and departures, we have

$$\begin{aligned}
& \mathbb{E}\left[\sum_{n\in\mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}} - I_{\{\eta_m(t)=n\}}) - \sum_{\substack{n:n\neq m\\ n\in\mathcal{H}}} Q_n(t) I_{\{\eta_m(t)=n\}} | Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t\right] \\
\leq\ & \mathbb{E}\left[\sum_{n\in\mathcal{B}} (Q_n(\tau_m^t) + nC_A) I_{\{\hat{\eta}_m(t)=n\}} - \sum_{\substack{n:n\neq m\\ n\in\mathcal{H}}} (Q_n(\tau_m^t) - nM) I_{\{\eta_m(t)=n\}} | Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t\right] \\
=\ & \mathbb{E}\left[Q_b^{max}(\tau_m^t) + nC_A - Q_h^{max}(\tau_m^t) + nM | Q_b^{max}(\tau_m^t) > Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t\right] \\
\leq\ & n(C_A + M)
\end{aligned}$$

Then we can bound the term (80) as

$$\begin{aligned}
& \mathbb{E}\left[\sum_{n\in\mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}} - I_{\{\eta_m(t)=n\}}) - \sum_{\substack{n:n\neq m\\ n\in\mathcal{H}}} Q_n(t) I_{\{\eta_m(t)=n\}}\right] \\
=\ & \sum_{n=1}^{t} \gamma n(C_A + M) \cdot Pr(Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t) | Q_m(\tau_m^t) = 0, \tau_m^t = t-n) \cdot Pr(Q_m(\tau_m^t) = 0 | \tau_m^t = t-n) \\
& \cdot Pr(\tau_m^t = t-n) \\
=\ & \sum_{n=1}^{t} \gamma n(C_A + M) \cdot Pr(Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t = t-n) \qquad (81)
\end{aligned}$$

The event $(Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t = t-n)$ is equivalent to the event that at time slot $t-n$, server m is idle and is scheduled to the maximum helper queue. Let $k = \arg\max_{n\in\mathcal{H}}\{Q_n(\tau_m^t)\}$. And for any time slot between $t-n$ and t, the working status of server m is equal to k. Hence

$$\begin{aligned}
& (Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t = t-n) \\
=\ & (f_m((t-n)^-) = 0, \eta_m(t-n) = k, f_m(t-n+1) = k, ..., f_m(t) = k)
\end{aligned}$$

So

$$\begin{aligned}
& Pr(Q_b^{max}(\tau_m^t) \leq Q_h^{max}(\tau_m^t), Q_m(\tau_m^t) = 0, \tau_m^t = t-n) \\
=\ & Pr(f_m((t-n)^-) = 0, \eta_m(t-n) = k, f_m(t-n+1) = k, ..., f_m(t) = k) \\
=\ & Pr(f_m(t-n+1) = k, ..., f_m(t) = k | f_m((t-n)^-) = 0, \eta_m(t-n) = k) Pr(f_m((t-n)^-) = 0, \eta_m(t-n) = k)
\end{aligned}$$

Let Y_m denote the event that when server m becomes idle, it is scheduled to the maximum helper queue. As we consider the steady state,

$$Pr(f_m((t-n)^-) = 0, \eta_m(t-n) \in \mathcal{H}) = Pr(Y_m)$$

By using the chain rule of probability, we have

$$
\begin{aligned}
& Pr(f_m(t-n+1)=k,...,f_m(t)=k|f_m((t-n)^-)=0,\eta_m(t-n)=k) \\
= \;& Pr(f_m(t-n+1)=k,...,f_m(t)=k|f_m((t-n)^-)=0,\eta_m(t-n)=k) \\
= \;& Pr(f_m(t-n+1)=k,|f_m((t-n)^-)=0,\eta_m(t-n)=k) \\
& \cdot \sum_{i=0}^{n-2} Pr(f_m(t-i)=k|f_m(t-i-1)=k,...,f(t-n+1)=k,f_m((t-n)^-)=0,\eta_m(t-n)=k)
\end{aligned}
$$

Given that server m is scheduled to serve a remote task from another helper queue at time slot $t-n$, the working status $f_m(t-n+1)$ is determined by the random variable $R_m(t-n+1) \sim Bern(\gamma)$. Hence

$$Pr(f_m(t-n+1)=k,|f_m((t-n)^-)=0,\eta_m(t-n)=k)=1-\gamma.$$

Similarly, for any $i=0,1,...,n-2$, given $f_m(t-i-1)=k$, $f_m(t-i)$ is determined by $R_m(t-i-1) \sim Bern(\gamma)$, thus

$$Pr(f_m(t-i)=k|f_m(t-i-1)=k,...,f(t-n+1)=k,f_m((t-n)^-)=0,\eta_m(t-n)=k)=1-\gamma$$

By utilizing these inequalities, we can bound the term (80) as

$$
\begin{aligned}
& \gamma\mathbb{E}\left[\sum_{n\in\mathcal{B}} Q_n(t)(I_{\{\hat{\eta}_m(t)=n\}}-I_{\{\eta_m(t)=n\}})-\sum_{\substack{n:n\neq m\\ n\in\mathcal{H}}} Q_n(t)I_{\{\eta_m(t)=n\}}\right] \\
\leq \;& \sum_{n=1}^{t}\gamma n(C_A+M)(1-\gamma)^n\cdot Pr(Y_m) \\
\leq \;& \gamma(C_A+M)(1-\gamma)\cdot Pr(Y_m)\sum_{n=1}^{\infty} n(1-\gamma)^{n-1} \\
= \;& \frac{(C_A+M)(1-\gamma)}{\gamma}\cdot Pr(Y_m)
\end{aligned}
$$

Next we will bound the expectation of term (79) in a similar way. Again consider the random variable τ_m^t and first decompose the probability space based on τ_m^t.

For a particular $\tau_m^t=t-n$, consider $Q_m(\tau_m^t)$.

Case (i): $Q_m(\tau_m^t)>0$

Under the proposed algorithm, $\eta_m(t)=\eta_m(\tau_m^t)=m$ with $Q_m(\tau_m^t)>0$. And $X_m^l(t)=S_m^l(t)$. Hence the term (79) is equal to zero.

Case (ii): $Q_m(\tau_m^t)=0$

When $Q_m(\tau_m^t)=0$, $\eta_m(t)=\eta_m(\tau_m^t)=\arg\max_{n\neq m}\{Q_n(\tau_m^t)\}$. By the bounded difference between $Q_m(t)$ and $Q_m(\tau_m^t)$, we have:

$$
\begin{aligned}
& \mathbb{E}\left[Q_m(t)(X_m^l(t)-S_m^l(t))-R_m(t)\sum_{n:n\neq m} Q_n(t)I_{\{\eta_m(t)=n\}}|Q_m(\tau_m^t)=0,\tau_m^t=t-n\right] \\
\leq \;& \mathbb{E}\left[(Q_m(\tau_m^t)+nC_A)\alpha(1-I_{\{\eta_m(t)=m\}})-\gamma\sum_{n:n\neq m}(Q_n(\tau_m^t)-nM)I_{\{\eta_m(t)=n\}}|Q_m(\tau_m^t)=0,\tau_m^t=t-n\right] \\
= \;& \mathbb{E}\left[nC_A\alpha-\gamma(Q^{max}(\tau_m^t)-nM)|Q_m(\tau_m^t)=0,\tau_m^t=t-n\right]Pr(Q_m(\tau_m^t)=0|\tau_m^t=t-n) \\
\leq \;& n(\alpha C_A+\gamma M)
\end{aligned}
$$

Thus we can bound the term (79) as

$$\begin{aligned}
&\mathbb{E}\left[Q_m(t)(X_m^l(t)-S_m^l(t))-R_m(t)\sum_{n:n\neq m}Q_n(t)I_{\{\eta_m(t)=n\}}\right]\\
=&\sum_{n=1}^{t}\mathbb{E}\left[Q_m(t)(X_m^l(t)-S_m^l(t))-R_m(t)\sum_{n:n\neq m}Q_n(t)I_{\{\eta_m(t)=n\}}|\tau_m^t=t-n\right]\cdot Pr(\tau_m^t=t-n)\\
=&\sum_{n=1}^{t}\Bigg(\mathbb{E}\left[Q_m(t)(X_m^l(t)-S_m^l(t))-R_m(t)\sum_{n:n\neq m}Q_n(t)I_{\{\eta_m(t)=n\}}|Q_m(\tau_m^t)=0,\tau_m^t=t-n\right]\\
&\cdot Pr(Q_m(\tau_m^t)=0|\tau_m^t=t-n)\cdot Pr(\tau_m^t=t-n)+0\cdot Pr(Q_m(\tau_m^t)>0|\tau_m^t=t-n)\cdot Pr(\tau_m^t=t-n)\Big)\\
\leq&\sum_{n=1}^{t}n(\alpha C_A+\gamma M)\cdot Pr(Q_m(\tau_m^t)=0|\tau_m^t=t-n)\cdot Pr(\tau_m^t=t-n) \qquad (82)
\end{aligned}$$

For $\forall m\in\mathcal{B}$, let Y_m denote the event that when server m is idle, its local queue is empty so it is scheduled to serve the maximum queue in the system. In steady state,

$$Pr(Q_m(\tau_m^t)=0,\tau_m^t=t-n)=Pr(Y_m).$$

Similar to the analysis for the term (81), we can bound the term (82) by

$$\frac{(\alpha C_A+\gamma M))(1-\gamma)}{\gamma^2}\cdot Pr(Y_m)$$

Therefore,

$$\mathbb{E}\left[\langle Q(t),\hat{S}(t)-S(t)\rangle\right]\leq\frac{(C_A+M))(1-\gamma)}{\gamma}\sum_{n\in\mathcal{H}}Pr(Y_m)+\frac{(\alpha C_A+\gamma M))(1-\gamma)}{\gamma^2}\sum_{n\in\mathcal{B}}Pr(Y_m) \qquad (83)$$

On the other hand

$$\begin{aligned}
&\sqrt{M}\langle e,\hat{S}(t)-S(t)\rangle\\
=&\sum_{m\in\mathcal{H}}(\hat{S}_m(t)-S_m(t))+\sum_{m\in\mathcal{B}}(\hat{S}_m(t)-S_m(t))\\
=&\sum_{m\in\mathcal{H}}\left(-\sum_{n:n\neq m}R_n(t)I_{\eta_n(t)=m}\right)+\sum_{m\in\mathcal{B}}\left(X_m^l(t)+\sum_{n\in\mathcal{H}}X_n^r(t)\cdot I_{\{\hat{\eta}_n(t)=m\}}-S_m^l(t)-\sum_{n:n\neq m}R_n(t)I_{\{\eta_n(t)=m\}}\right)\\
=&\sum_{m\in\mathcal{H}}\left(X_m^r(t)I_{\{\hat{\eta}_m(t)\neq m\}}-R_m(t)I_{\{\eta_m(t)\neq m\}}\right)+\sum_{m\in\mathcal{B}}\left(X_m^l(t)-S_m^l(t)-R_m(t)I_{\{\eta_m(t)\neq m\}}\right)\\
=&\sum_{m\in\mathcal{B}}\left(X_m^l(t)(1-I_{\{\eta_m(t)=m\}})-R_m(t)I_{\{\eta_m(t)\neq m\}}\right)
\end{aligned}$$

For $\forall m\in\mathcal{B}$,

$$\mathbb{E}\left[X_m^l(t)(1-I_{\{\eta_m(t)=m\}})-R_m(t)I_{\{\eta_m(t)\neq m\}}\right]=(\alpha-\gamma)Pr(I_{\{\eta_m(t)\neq m\}})=(\alpha-\gamma)Pr(Y_m) \qquad (84)$$

Thus

$$\mathbb{E}\left[\langle e,\hat{S}(t)-S(t)\rangle\right]\geq(\alpha-\gamma)\sum_{m\in\mathcal{B}}Pr(Y_m)$$

From proof of Lemma 20, we have

$$\mathbb{E}\left[N_{bh}(t)\right] \leq \frac{\alpha}{(\alpha+\gamma)(\alpha-\gamma)}\epsilon$$
$$\mathbb{E}\left[N_{bb}(t)\right] \leq \frac{\epsilon}{\alpha-\gamma},$$

i.e., the number of beneficiary servers that provide remote service is of order $o(\epsilon)$. This implies that $\sum_{m\in\mathcal{B}} Pr(Y_m) = o(\epsilon)$.

Therefore

$$\begin{aligned}
& \mathbb{E}\left[\langle Q(t), \hat{S}(t) - S(t)\rangle\right] \\
\leq\ & \frac{(C_A+M))(1-\gamma)}{\gamma}\sum_{n\in\mathcal{H}} Pr(Y_m) + \frac{(\alpha C_A+\gamma M))(1-\gamma)}{\gamma^2(\alpha-\gamma)}\mathbb{E}\left[\langle e, \hat{S}(t) - S(t)\rangle\right] \\
\leq\ & o(\epsilon) + R_1\sqrt{M}e\langle e, \hat{S}(t) - S(t)\rangle
\end{aligned}$$

where $R_1 = \frac{(\alpha C_A+\gamma M))(1-\gamma)}{\gamma^2(\alpha-\gamma)\sqrt{M}}$ is a constant not depending on ϵ. ■

7.10.3 Proof of Lemma 22 .

Proof.

By the definition of $\hat{A}$:

$$\begin{aligned}
\langle Q, A - \hat{A}\rangle &= \sum_{m\in\mathcal{B}}\left(Q_m(t)\sum_{\bar{L}\in\mathcal{L}_s: m\in\bar{L}} A_{\bar{L},m}\right) - \sum_{m\in\mathcal{H}}\left(Q_m(t)\sum_{\bar{L}\in\mathcal{L}_s: m\in\bar{L}}\frac{\sum_{n\in\bar{L}\cap\mathcal{B}} A_{\bar{L},n}}{|\{k: k\in\bar{L}\cap\mathcal{H}\}|}\right) \\
&= \sum_{\bar{L}\in\mathcal{L}_s}\left[\sum_{n\in\bar{L}\cap\mathcal{B}} Q_n(t)A_{\bar{L},n} - \sum_{m\in\bar{L}\cap\mathcal{H}} Q_m(t)\frac{\sum_{n\in\bar{L}\cap\mathcal{B}} A_{\bar{L},n}}{|\{k: k\in\bar{L}\cap\mathcal{H}\}|}\right] \\
&= \sum_{\substack{\bar{L}=(m_1,m_2,m_3)\\ m_1,m_2\in\mathcal{H}, m_3\in\mathcal{B}}}\left[Q_{m_3}(t)A_{\bar{L},m_3} - (Q_{m_1}(t)+Q_{m_2}(t))\frac{A_{\bar{L},m_3}}{2}\right] \\
&+ \sum_{\substack{\bar{L}=(m_1,m_2,m_3)\\ m_1\in\mathcal{H}, m_2,m_3\in\mathcal{B}}}\left[Q_{m_2}(t)A_{\bar{L},m_2} + Q_{m_3}(t)A_{\bar{L},m_3} - Q_{m_1}(t)(A_{\bar{L},m_2}+A_{\bar{L},m_3})\right]
\end{aligned}$$

Case (i): $\bar{L} = (m_1, m_2, m_3)$ with $m_1, m_2 \in \mathcal{H}, m_3 \in \mathcal{B}$

As arriving tasks are routed to the shortest local queue, $A_{\bar{L},m_3} > 0$ only if $Q_{m_3} \leq Q_{m_1}(t)$ and $Q_{m_3} \leq Q_{m_2}(t)$. Hence $A_{\bar{L},m_3}(Q_{m_3} - (Q_{m_1}(t) + Q_{m_2}(t))/2) \leq 0$.

Case (ii): $\bar{L} = (m_1, m_2, m_3)$ with $m_1 \in \mathcal{H}, m_2, m_3 \in \mathcal{B}$

Similarly, $A_{\bar{L},m_2} > 0$ only if $Q_{m_2} \leq Q_{m_1}(t)$. Hence $A_{\bar{L},m_2}(Q_{m_2} - Q_{m_1}(t)) \leq 0$. The same goes for $A_{\bar{L},m_3}(Q_{m_3} - Q_{m_1}(t)) \leq 0$.

Therefore

$$\langle Q, A - \hat{A}\rangle \leq 0$$

■

7.10.4 Proof of Lemma 23.

Proof. $\langle Q(t), U(t)\rangle = \sum_{m\in\mathcal{M}} Q_m(t)U_m(t)$. By the definition of $U_m(t)$, if $U_m(t) > 0$, the number of tasks in queue m must be less than the number of available servers scheduled to this queue at time slot t. Since $S_m(t) \leq M$, $Q_m(t) \leq Q_m(t) + A_m(t) < M$. If $U_m(t) = 0$, $Q_m(t)U_m(t) = 0$. Hence $Q_m(t)U_m(t) < MU_m(t)$. Therefore, $\langle Q(t), U(t)\rangle < \sum_{m\in\mathcal{M}} MU_m(t) = M\sqrt{M}\langle e, U(t)\rangle$, where $e = \frac{1}{\sqrt{M}}\underbrace{(1, 1, \cdots, 1)}_{M}$.

■

Appendix D

The proof for uniform traffic case follows exactly the same three steps for the skewed traffic case. The symmetry brought about by the uniform ideal load for all queues will significantly simplify the proof. Analogue to Lemma 9, we have the following lemma for the ideal load decomposition with uniform traffic, which is essential for proving state space collapse.

Lemma 24 *There exists a positive constant λ_{min} not depending on ϵ such that:*

1 . $\forall m \in \mathcal{M}$,

$$\sum_{\bar{L}:m\in\bar{L}} \sum_{n:n\in\bar{L}} \frac{\lambda^*_{\bar{L},n,m}}{\alpha} + \sum_{\bar{L}:m\notin\bar{L}} \sum_{n:n\in\bar{L}} \frac{\lambda^*_{\bar{L},n,m}}{\gamma} = 1-\epsilon_0$$

;

3 . $\forall \bar{L} \in \mathcal{L}, \forall m \in \bar{L}$, $\lambda^*_{\bar{L},m,m} \geq \lambda_{min}$.

The proof of this lemma is the same as that of Lemma 9.

7.11 Lower Bound

Consider a single server system with arrival process $\{a^{(\epsilon)}(t), t \geq 0\}$ and service process $\{\beta^{(\epsilon)}(t), t \geq 0\}$, where

$$a^{(\epsilon)}(t) = \sum_{\bar{L}\in\mathcal{L}_{\mathcal{M}}} A^{(\epsilon)}_{\bar{L}}(t),\ \beta^{(\epsilon)}(t) = \sum_{m=1}^{M} X_m(t).$$

Here $\{X_m(t), t \geq 0\}_{m\in\mathcal{M}}$ are independent, and each process is temporally i.i.d. with $X_m(t) \sim Bern(\alpha)$. Assume that the mean of $a^{(\epsilon)}(t)$ is $M\alpha - \epsilon$ and the variance is denoted by $(\sigma^{(\epsilon)})^2$. Then the corresponding queue-length process is stochastically smaller than $\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)$. Hence

$$\mathbb{E}\left[\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)\right] \geq \frac{(\sigma^{(\epsilon)})^2 + \nu^2 + \epsilon^2}{2\epsilon} - \frac{M}{2},$$

where ν^2 is the variance for $\{\beta^{(\epsilon)}(t)\}$.

Therefore, in the heavy traffic limit, we have

$$\liminf_{\epsilon\to 0^+} \epsilon\mathbb{E}\left[\sum_{m=1}^{M} Q^{(\epsilon)}_m(t)\right] \geq \frac{\sigma^2+\nu^2}{2}. \tag{85}$$

7.12 State Space Collapse

We will show that the queue length vector Q collapses to the direction c_u, where

$$c_u = \frac{1}{\sqrt{M}}(\underbrace{1, 1, \cdots, 1}_{M}). \tag{86}$$

Let $Q_{||}$ and $Q_{\perp}$ be the components of Q parallel and perpendicular to the direction c_u. Consider the Lyapunov functions $V(Z) = \left|\left|Q_{\perp}\right|\right|$. We again show that the T-period drift of $V(Z)$ satisfies two conditions. The analysis follows the same steps for locally overloaded traffic, but it is significantly simplified due to the symmetric load for each server.

We need lemmas analogue to Lemmas 17-19, with beneficiary queue length vector Q replaced by all queue length vector Q.

Lemma 25 *Let c be a vector with unit norm in $\mathbb{R}^M$. Then for any $t \geq 0$,*

$$\left|\left|Q_{||}(t+1)\right|\right|^2 - \left|\left|Q_{||}(t)\right|\right|^2 \geq 2\langle c_u, Q(t)\rangle\langle c_u, A(t) - S(t)\rangle,$$

where $Q_{||}$ is the parallel component of the beneficiary queue length vector Q with respect to the direction c.

Lemma 26 *Consider a time slot t_0 and a positive integer T. Let c be a vector with unit norm in $\mathbb{R}^M$. Then for any t with $t_0 \leq t < t_0 + T$,*

$$\left|\left|Q_{\perp}(t)\right|\right| - \left|\left|Q_{\perp}(t_0)\right|\right| \leq T\sqrt{M}\max\{M, C_A\}, \tag{87}$$

where $Q_{\perp}$ is the perpendicular component of the beneficiary queue length vector Q with respect to the direction c.

Lemma 27 *Consider a time slot t_0 and a positive integer T. For any t with $t_0 \leq t < t_0 + T$, let $G_e(t) = \langle Q(t), A(t) - S(t)\rangle - \langle c_b, Q(t)\rangle\langle c_b, A(t) - S(t)\rangle$. Then $G_e(t) \leq h\left|\left|Q_{\perp}(t_0)\right|\right| + F_0$, where $h' = \sqrt{M}\max\{M, C_A\}$ and $F_0' = MT(\max\{M, C_A\})^2$ are constants.*

Following the same steps for analyzing drift of $V(Z^{(B)})$ in Appendix C yields

$$\mathbb{E}\left[\Delta V(Z)|Z(t_0)\right] \leq \frac{\mathbb{E}\left[\sum_{t=t_0}^{t_0+T-1} G_e(t)|Z(t_0)\right] + C}{||Q_{\perp}||}$$

where C is a constant and $G_e(t) = \langle Q(t), A(t) - S(t)\rangle - \langle c_u, Q(t)\rangle\langle c_u, A(t) - S(t)\rangle$.

The key step is again to bound $\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} G_e(t)|Z(t_0), t^* < t_0 + K\right]$. Since we consider the entire system, we do not have the shared arrival issue here. In addition, as the local load for each server approaches 1, each server devotes to serving its local queue. Hence the remote service vanishes as $\lambda^{(\epsilon)}$ is close to the capacity boundary, which enable us to get rid of the remote service terms in bounding corresponding $G_e(t)$. We have the inequalities analogue to Lemma 10-12.

Lemma 28 *For any $t^* < t < t_0 + T$*

$$\mathbb{E}\left[\langle Q(t), A(t)\rangle - \langle Q(t), \lambda^*\rangle \,|\, t^*, Z(t_0)\right] \leq -\lambda_{min}||Q_{\perp}(t_0)|| + F_1', \tag{88}$$

where F_1' is a constant not depending on ϵ.

Lemma 29

$$\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} \left(\langle Q(t), \lambda^*\rangle - \langle Q(t), S(t)\rangle\right) |t^*, Z(t_0)\right] \leq -(t_0 + T - t^*)\frac{\epsilon}{M}\sum_m Q_m(t_0) + F_2',$$

where F_2' is a positive constant not depending on ϵ.

Lemma 30 *For any $t^* < t < t_0 + T$*

$$\mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, A(t) - S(t)\rangle|t^*, Z(t_0)\right] \geq -\frac{\epsilon}{M}\sum_m Q_m(t_0) - F_3',$$

where F_3' is a positive constant not depending on ϵ.

Together they give us an upper bound on $\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} G_e(t)|Z(t_0), t^* < t_0+K\right]$.

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t^*+1}^{t_0+T-1} G_e(t)|Z(t_0), t^* < t_0+K\right] &\leq -(t_0+T-t^*)\lambda_{min}||Q_\perp(t_0)|| + F_4' \\ &\leq -(J-1)K\lambda_{min}||Q_\perp(t_0)|| + F_4',\end{aligned}$$

where $F_4' = F_1' + F_2' + F_3'$ is constant independent of ϵ.

Utilizing Lemma 27 and the above inequality yields an upper bound on the drift of $V(Z)$, which is similar to Eqn (56). Hence the drift of $V(Z)$ is negative for sufficiently large $V(Z)$. Moreover, Lemma 26 implies finite drift of $V(Z)$. Note that the Markov chain $Z^{(\epsilon)}(t) = (Q^{(\epsilon)}(t), f^{(\epsilon)}(t)), t \geq 0$ is positive recurrent. Therefore, by the extended lemma 1 in [7], all moments of $V(Z)$ are finite and independent of ϵ. State space collapse of Q along the direction c_u follows.

7.13 Proof of Lemma 28-30

Proof of Lemma 28

The proof is similar to that of Lemma 5.

For $\forall \bar{L} \in \mathcal{L}$, define $Q^*_{\bar{L}}(t) = \min_{m\in\bar{L}}\{Q_m(t)\}$. Thus tasks of type $\bar{L}$ will be routed to queue $Q^*_{\bar{L}}(t)$ at the beginning of time slot t.

$$\begin{aligned}\mathbb{E}\left[\langle Q(t), A(t)\rangle|Z(t)\right] &= \mathbb{E}\left[\sum_{m=1}^{M} Q_m(t)A_m(t)|Z(t)\right] \\ &= \mathbb{E}\left[\sum_m \sum_{\bar{L}:m\in\bar{L}} Q_m(t)A_{\bar{L},m}(t)|Z(t)\right] \\ &= \sum_{\bar{L}} \mathbb{E}\left[Q^*_{\bar{L}}(t)A_{\bar{L}}(t)|Z(t)\right] \\ &= \sum_{\bar{L}} Q^*_{\bar{L}}(t)\lambda_{\bar{L}} \\ &= \sum_{\bar{L}} Q^*_{\bar{L}}(t) \sum_{m:m\in\bar{L}} \sum_{n=1}^{M} \lambda^*_{\bar{L},m,n} \qquad (89) \\ &= \sum_{\bar{L}} \sum_{m:m\in\bar{L}} \lambda^*_{\bar{L},m,m} Q^*_{\bar{L}}(t) \qquad (90)\end{aligned}$$

where Eq.(89) and (90) follow from the ideal decomposition, since $\mathcal{H} = \mathcal{M}$ with evenly loaded traffic.

Then we have

$$\begin{aligned}&\mathbb{E}\left[\langle Q(t), A(t)\rangle - \langle Q(t), \lambda^*\rangle|\, Z(t)\right] = \sum_{\bar{L}} \sum_{m:m\in\bar{L}} \lambda^*_{\bar{L},m,m} Q^*_{\bar{L}}(t) - \sum_m \lambda^*_m Q_m(t) \\ =\ & \sum_{\bar{L}} \sum_{m:m\in\bar{L}} \lambda^*_{\bar{L},m,m} Q^*_{\bar{L}}(t) - \sum_m \sum_{\bar{L}:m\in\bar{L}} \lambda^*_{\bar{L},m,m} Q_m(t) \\ =\ & -\sum_{\bar{L}} \sum_{m:m\in\bar{L}} \lambda^*_{\bar{L},m,m}(Q_m(t) - Q^*_{\bar{L}}(t)) \\ \leq\ & -\lambda_{min} \sum_{\bar{L}} \sum_{m:m\in\bar{L}} \lambda^*_{\bar{L},m,m}(Q_m(t) - Q^*_{\bar{L}}(t)), \qquad (91)\end{aligned}$$

where the last inequality follows by Lemma 24 and the fact that $Q_m(t) \geq Q^*_{\bar{L}}(t)$ for any $\bar{L} \in \mathcal{L}$, $\forall m \in \bar{L}$.

Assume that $m_1 = \arg\max_{m\in\mathcal{M}}\{Q_m(t)\}$, and $m_k = \arg\min_{m\in\mathcal{M}}\{Q_m(t)\}$. Denote the maximum queue length at time slot t by $Q^{max}(t)$. That is, $Q^{max}(t) = Q_{m_1}(t)$. Note that for any $\bar{L}\mathcal{L}$ such that $m_k \in \bar{L}$, $Q^*_{\bar{L}}(t) = Q_{m_k}(t)$ as Q_{m_k} is the minimum queue at time t.

Since any pair of servers in the system are connected, there exit a sequence of servers $(m_1, m_2, ..., m_{k-1}, m_k)$ such that for any consecutive servers in the sequence, say (m_i, m_{i+1}), there exits a task type $\bar{L}_{i,i+1} \in \mathcal{L}$ with $\lambda_{\bar{L}_{i,i+1}} > 0$ local to both server m_i and m_{i+1}. For the summation in (91), we keep terms of types $\bar{L}_{1,2}, \bar{L}_{2,3}, ..., \bar{L}_{k-1,k}$, and for each task type $\bar{L}_{i,i+1}$, we only keep $m = m_i$ term. All other terms are discarded.

Hence

$$\begin{aligned}
& \mathbb{E}\left[\langle Q(t), A(t)\rangle - \langle Q(t), \lambda^*\rangle \,|\, Z(t)\right] \\
\leq\ & -\lambda_{min}\left(Q_{m_1}(t) - Q^*_{\bar{L}_{1,2}} + Q_{m_2}(t) - Q^*_{\bar{L}_{2,3}} + \cdots + Q_{m_{k-1}}(t) - Q^*_{\bar{L}_{k-1,k}}\right) \\
\leq\ & -\lambda_{min}\left(Q_{m_1}(t) - Q^*_{\bar{L}_{k-1,k}}\right) \\
=\ & -\lambda_{min}\left(Q^{max}(t) - Q^{min}(t)\right) \\
=\ & -\frac{\lambda_{min}}{\sqrt{M}}\sqrt{M(Q^{max}(t) - Q^{min}(t))^2} \\
\leq\ & -\frac{\lambda_{min}}{\sqrt{M}}\sqrt{\sum_m \left(Q_m(t) - \frac{\sum_i Q_i(t)}{M}\right)^2} \\
=\ & -\frac{\lambda_{min}}{\sqrt{M}}||Q_\perp(t)|| \\
\leq\ & -\frac{\lambda_{min}}{\sqrt{M}}||Q_\perp(t_0)|| + F_1'
\end{aligned}$$

where the last inequality comes from Lemma 25.

This completes the proof.

■

Proof of Lemma 29 is the same as that of Lemma 6. Observe that for any $m \in \mathcal{M}$, $\sum_{\bar{L}:m\in\bar{L}} \lambda_{\bar{L},m,m} = \alpha(1 - \frac{\epsilon}{M})$ by Lemma 24. Replacing $\frac{\alpha}{1+\vartheta}$ with $\alpha(1 - \frac{\epsilon}{M})$ gives the part on the right hand side in Lemma 29.

Proof of Lemma 30

By the boundedness of arrivals and service, we have

$$\langle c_u, Q(t_0)\rangle - \frac{TM}{\sqrt{M}} \leq \langle c_u, Q(t)\rangle \leq \langle c_u, Q(t_0)\rangle + \frac{TC_A}{\sqrt{M}}$$

Hence

$$\begin{aligned}
& \mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, A(t) - S(t)\rangle | t^*, Z(t_0)\right] \\
\geq\ & \mathbb{E}\left[\langle c_u, Q(t_0)\rangle\langle c_u, A(t) - S(t)\rangle - \frac{TC_A}{\sqrt{M}}\langle c_u, S(t)\rangle | t^*, Z(t_0)\right] \\
\geq\ & \langle c_u, Q(t_0)\rangle\frac{1}{\sqrt{M}}\mathbb{E}\left[\sum_{m\in\mathcal{M}}(A_m(t) - S_m(t)) | t^*, Z(t_0)\right] - F_3',
\end{aligned}$$

where F_3' is a constant.

Note that

$$\mathbb{E}\left[\sum_{m\in\mathcal{M}} A_m(t)\right] = \mathbb{E}\left[\sum_{\bar{L}\in\mathcal{L}} A_{\bar{L}}(t)\right] = \sum_{\bar{L}\in\mathcal{L}} \lambda_{\bar{L}} = M\alpha - \epsilon.$$

$$\mathbb{E}\left[\sum_{m\in\mathcal{M}} S_m(t)|t^*, Z(t_0)\right] = \mathbb{E}\left[\sum_{m\in\mathcal{M}} (\alpha I_{\{\eta_m(t)=m\}} + \gamma I_{\{\eta_m(t)\neq m\}})|t^*, Z(t_0)\right] \leq M\alpha$$

Combining the above inequalities yields

$$\mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, A(t) - S(t)\rangle|t^*, Z(t_0)\right] \geq \langle c_u, Q(t_0)\rangle\frac{1}{\sqrt{M}}(-\epsilon) - F_3' = -\frac{\epsilon}{M}\sum_m Q_m(t_0) - F_3' \tag{92}$$

■

7.14 Upper Bound

Again we construct an *ideal service process* $\{\hat{S}(t), t \geq 0\}$ that makes $\{\sum_m \hat{S}_m(t)\}$ independent of $\sum_m Q_m(t)$. In particular, $\forall m \in \mathcal{M}$, its *ideal local service process* $\hat{S}_m^l(t)$ is defined in the same way as that for beneficiaries in the skewed traffic case.

Ideal local service process $\hat{S}^l(t)$:

$$\hat{S}_m^l(t) = X_m^l(t), \forall m \in \mathcal{M}$$

where the processes $\{X_m^l(t), t \geq 0\}_{m\in\mathcal{M}}$ is coupled with $\{S_m(t), t \geq 0\}_{m\in\mathcal{M}}$ in the following way: If $\eta_m(t) = m$, $X_m^l(t) = S_m^l(t)$; if $\eta_m(t) \neq m$, $X_m^l(t) = 1$ when $R_m(t) = 1$, and $X_m^l(t) \sim Bern(\frac{\alpha-\gamma}{1-\gamma})$ when $R_m(t) = 0$. Hence $\forall m \in \mathcal{M}$, $\{X_m^l(t), t \geq 0\}$ is i.i.d. with $X_m^l(t) \sim Bern(\alpha)$.

Ideal remote service process $\hat{R}(t)$: For any $m \in \mathcal{M}$, $\hat{R}_m(t) = 0$.

Ideal scheduling decision process $\hat{\eta}(t)$: For any $m \in \mathcal{M}$, $\hat{\eta}_m(t) = m$.

Since the total amount of arrivals for the system is independent of queue-length process, we do not need to define *ideal arrival process* here.

Observe that

$$\mathbb{E}\left[\sum_{m\in\mathcal{M}} \hat{S}_m(t)\right] = M\alpha$$
$$Var(\sum_{m\in\mathcal{M}} \hat{S}_m(t)) = M\alpha(1-\alpha)$$

Denote the variance of total ideal service process $\{\sum_{m\in\mathcal{M}} \hat{S}_m(t)\}$ by ν^2.

And

$$\mathbb{E}\left[\sum_{m\in\mathcal{M}} \hat{S}_m(t) - \sum_{m\in\mathcal{M}} A_m(t)\right] = \epsilon$$

Then we can rewrite the queue dynamics as

$$Q(t+1) = Q(t) + A(t) - \hat{S}(t) + \hat{U}(t), \tag{93}$$

where $\hat{U}(t) = \hat{S}(t) - S(t) + U(t)$. Again setting the drift of $W_{||}(Z) = ||Q_{||}||^2$ to zero gives the following equation, which is similar to that in Lemma 14.

$$\begin{aligned} & 2\mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, \hat{S}(t) - A(t)\rangle\right] \\ = \; & \mathbb{E}\left[\langle c_u, A(t) - \hat{S}(t)\rangle^2\right] + \mathbb{E}\left[\langle c_u, \hat{U}(t)\rangle^2\right] \qquad (94) \\ & +2\mathbb{E}\left[\langle c_u, Q(t) + A(t) - \hat{S}(t)\rangle\langle c_u, \hat{U}(t)\rangle\right] \qquad (95) \end{aligned}$$

Again we obtain the upper bound by bounding each term in Eqs. (94)-(95).
For the term on the left side of the equation, we have

$$\begin{aligned}\mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, \hat{S}(t) - A(t)\rangle\right] &= \frac{1}{M}\mathbb{E}\left[\left(\sum_{m\in\mathcal{M}} Q_m(t)\right)\left(\sum_{m\in\mathcal{M}} \hat{S}_m(t) - \sum_{m\in\mathcal{M}} A_m(t)\right)\right] \\ &= \frac{1}{M}\mathbb{E}\left[\sum_{m\in\mathcal{M}} Q_m(t)\right]\mathbb{E}\left[\sum_{m\in\mathcal{M}} \hat{S}_m(t) - \sum_{m\in\mathcal{M}} A_m(t)\right] \\ &= \frac{\epsilon}{M}\mathbb{E}\left[\sum_{m\in\mathcal{M}} Q_m(t)\right]\end{aligned}$$

We first bound the two terms in (94).
By the definition of ideal service process, it is easy to verify that

$$\mathbb{E}\left[\langle c_u, A(t) - \hat{S}(t)\rangle^2\right] = \frac{(\sigma^{(\epsilon)})^2 + \nu^2 + \epsilon^2}{M}$$

Since $Q(t)$ is in steady state, $\mathbb{E}\left[\langle c_u, A(t) - \hat{S}(t) + \hat{U}(t)\rangle\right] = \mathbb{E}\left[\langle c_u, Q(t+1)\rangle - \langle c_u, Q(t)\rangle\right] = 0$. Thus $\mathbb{E}\left[\langle c_u, \hat{U}(t)\rangle\right] = \mathbb{E}\left[\langle c_u, \hat{S}(t) - A(t)\rangle\right] = \frac{\epsilon}{\sqrt{M}}$. Meanwhile,

$$\langle c_u, \hat{U}(t)\rangle = \langle c_u, \hat{S}(t) - S(t) + U(t)\rangle \le \langle c_u, \hat{S}(t) + U(t)\rangle \le 2\sqrt{M}$$

By the coupling of $\{\hat{S}(t), t \ge 0\}$ and $\{S(t), t \ge 0\}$, $\langle c_u, \hat{S}(t) - S(t)\rangle \ge 0$.
Therefore

$$\mathbb{E}\left[\langle c_u, \hat{U}(t)\rangle^2\right] \le 2\sqrt{M}\mathbb{E}\left[\langle c_u, \hat{U}(t)\rangle\right] = 2\epsilon$$

Finally we bound the term (95).

$$\begin{aligned}\mathbb{E}\left[\langle c_u, Q(t) + A(t) - \hat{S}(t)\rangle\langle c_u, \hat{U}(t)\rangle\right] &= \mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, \hat{U}(t)\rangle\right] + \mathbb{E}\left[\langle c_u, A(t) - \hat{S}(t)\rangle\langle c_u, \hat{U}(t)\rangle\right] \\ &\le \mathbb{E}\left[\langle c_u, Q(t)\rangle\langle c_u, \hat{U}(t)\rangle\right] + 2\sqrt{M}\epsilon\end{aligned}$$

Note that

$$\begin{aligned}\langle c_u, Q(t)\rangle\langle c_u, \hat{U}(t)\rangle &= \langle Q(t), \hat{U}(t)\rangle - \langle Q_\perp(t), \hat{U}_\perp(t)\rangle \\ &= \langle Q(t), \hat{S}(t) - S(t)\rangle + \langle Q(t), A(t) - A(t)\rangle + \langle Q(t), U(t)\rangle - \langle Q_\perp(t), \hat{U}_\perp(t)\rangle \quad (96)\end{aligned}$$

The following lemma, which is analogue to Lemma 21 for locally overloaded traffic, bounds the first term in Eq. (96).

Lemma 31 $\mathbb{E}\left[\langle Q(t), \hat{S}(t) - S(t)\rangle\right] \le R_1''\sqrt{M}\mathbb{E}\left[\langle c_u, \hat{S}(t) - S(t)\rangle\right]$, *where* $R_1'' > 0$ *is a constant.*

Let $R_2' = \max\{R_1'', M\}$, then

$$\begin{aligned}\langle Q(t), \hat{S}(t) - S(t)\rangle + \langle Q(t), U(t)\rangle &\le R_2'\sqrt{M}\mathbb{E}\left[\langle c_u, \hat{S}(t) - S(t)\rangle + \langle c_u, U(t)\rangle\right] \\ &= R_2'\sqrt{M}\mathbb{E}\left[\langle c_u, \hat{S}(t) - S(t) + U(t)\rangle\right] \\ &= R_2'\epsilon\end{aligned}$$

To bound the last term in Eqn. (96), we will first bound $\mathbb{E}\left[||\hat{U}(t)||^2\right]$ as Lemma 23. By following the analysis steps in proof of Lemma 23, we can show that

$$\mathbb{E}\left[||\hat{U}(t)||^2\right] \leq R_3'\epsilon$$

,

where R_3' is a constant that doesn't depend on ϵ.

We use the state space collapse result to bound $-\langle Q_\perp(t), \hat{U}_\perp(t)\rangle$.

$$\begin{aligned}
\mathbb{E}\left[-\langle Q_\perp(t), \hat{U}_\perp(t)\rangle\right] &\leq \sqrt{\mathbb{E}\left[\left|\left|Q_\perp(t)\right|\right|^2\right]\mathbb{E}\left[\left|\left|\hat{U}(t)\right|\right|^2\right]} &(97)\\
&\leq \sqrt{C_2'\mathbb{E}\left[\left|\left|\hat{U}(t)\right|\right|^2\right]} &(98)\\
&\leq \sqrt{C_2'R_3'\epsilon} &(99)
\end{aligned}$$

Combining these inequalities, we can bound the term (95) as

$$\begin{aligned}
&\mathbb{E}\left[\langle c, Q(t) + A(t) - \hat{S}(t)\rangle\langle c, \hat{U}(t)\rangle\right]\\
\leq\ & 2\epsilon + R_2'\epsilon + \sqrt{C_2'R_3'\epsilon}
\end{aligned}$$

Now we revive the superscript $^{(\epsilon)}$. Combining the inequalities we have

$$2\frac{\epsilon}{M}\mathbb{E}\left[\sum_{m\in\mathcal{M}} Q_m(t)\right] \leq \frac{(\sigma^{(\epsilon)})^2 + \nu^2 + \epsilon^2}{M} + 2\epsilon + R_2'\epsilon + \sqrt{C_2'R_3'\epsilon}.$$

Hence

$$\mathbb{E}\left[\sum_m Q_m^{(\epsilon)}(t)\right] \leq \frac{(\sigma_b^{(\epsilon)})^2 + \nu^2}{2\epsilon} + D_u^{(\epsilon)}, \tag{100}$$

where

$$D_u^{(\epsilon)} = \frac{\epsilon}{2} + M + R_2' + \frac{M\sqrt{C_2'R_3'}}{2\sqrt{\epsilon}}.$$

Note that $D_u^{(\epsilon)} = o(\frac{1}{\epsilon})$, i.e., $\lim_{\epsilon\to 0^+} \epsilon D_u^{(\epsilon)} = 0$.

Therefore, in the heavy-traffic limit, we have

$$\limsup_{\epsilon\to 0^+} \epsilon\mathbb{E}\left[\sum_m Q_m^{(\epsilon)}(t)\right] \leq \frac{\sigma^2 + \nu^2}{2}$$

Then the first moment heavy-traffic optimality of the proposed algorithm follows by the coincidence of lower and upper bound.