

# Optimal Load-Balancing for High-Density Wireless Networks with Flow-Level Dynamics

Bin Li\*, Xiangqi Kong\* and Lei Wang†

\*Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, USA

†School of Software, Dalian University of Technology, China

Email: binli@uri.edu, xqkong@my.uri.edu, lei.wang@dlut.edu.cn

**Abstract**—We consider the load-balancing design for forwarding incoming flows to access points (APs) in high-density wireless networks with both channel fading and flow-level dynamics, where each incoming flow has a certain amount of service demand and leaves the system once its service request is complete (referred as *flow-level dynamic model*). The efficient load-balancing design is strongly needed for supporting high-quality wireless connections in high-density areas. Despite the presence of a variety of earlier works on the design and analysis of the load-balancing schemes in wireless networks, there does not exist a work on the load-balancing design in the realistic flow-level dynamic model.

In this work, we propose a Join-the-Least-Workload (JLW) Algorithm that always forwards the incoming flows to the AP with the smallest workload in the presence of flow-level dynamics. However, our considered flow-level dynamic model differs from traditional queuing model for wireless networks in the following two aspects: (1) the dynamics of the flows is *short-term* and flows will leave the network once they received the desired amount of service; (2) each individual flow faces an independent channel fading. These differences pose significant challenges on the system performance analysis. To tackle these challenges, we perform Lyapunov-drift-based analysis of the stochastic network taking into account sharp flow-level dynamics. Our analysis reveals that our proposed JLW Algorithm not only achieves maximum system throughput, but also minimizes the total system workload in heavy-traffic regimes. Moreover, we observe from both our theoretical and simulation results that the mean total workload performance under the proposed JLW Algorithm does not degrade as the number of APs increases, which is strongly desirable in high-density wireless networks.

## I. INTRODUCTION

With the rapid growth of smart phones, there is a strong need for high-quality wireless local access network (WLAN) connections in high-density areas, such as convention centers, auditoriums, hotel meeting rooms, lecture halls, sports stadiums, and concert halls. These high-speed wireless connections are not only for business and entertainment purposes, but more importantly provide emergency response communications in crowded places in response to unexpected events such as fire, shooting, and terrorist attack. To support such wireless connections in high-density WLANs, multiple access points (APs) are necessary to be deployed for providing satisfactory services for wireless users. However, in conventional WLANs, each user is automatically associated with the AP that has the best channel quality, which causes significant load imbalance among APs and results in poor network performance (e.g. [12]). This raises a natural question in how to develop an

efficient joint load-balancing and scheduling algorithm that first determines which AP an incoming user should associate with, then each individual AP needs to decide which users it serves. The goal of such an algorithm is to maximize system throughput (or equivalently support network users as many as possible) and to minimize average user's delay.

While load-balancing for multiple APs with various fairness criteria (e.g., [3], [8], [13], [1], [5], [20]) have been studied extensively, relatively limited work on the realistic model exists where a mobile user transmits data from a file (or a flow), and either departs or becomes silent for a while, which was observed in prior work (e.g., [2], [12]). In such practical wireless networks, even the design of scheduling algorithms in a single AP case is quite non-trivial, let alone load-balancing design. Indeed, most existing scheduling designs including the well-known MaxWeight-type algorithms (e.g., [21], [22]) implicitly assume that the system consists of a fixed number of persistent users that continuously inject packets into the network and will never leave the network, and thus perform poorly in the presence of dynamic flows (e.g., [23], [24]). The main reason is that the queue-length-based MaxWeight algorithm myopically selects a feasible schedule with the maximal residual size of dynamic flows and hence the flows with small backlogs may stay in the network forever. Subsequent works (e.g. [15], [14], [18]) have developed throughput-optimal scheduling algorithms that do not require any prior knowledge of channels and user demands. Despite these advances in efficient scheduling design for wireless networks with flow-level dynamics, the load-balancing design among multiple APs is far less explored.

On the contrary, the load-balancing schemes have been explored extensively in data centers that distribute arriving jobs across servers with the goal of minimizing queueing delays. The celebrated Join-the-Shortest-Queue (JSQ) policy (e.g., [26], [7]), where all arrivals are forwarded to the shortest queue, has been shown to not only achieve maximum system throughput but also minimize mean delay in the symmetric case [22], or under the heavy-traffic regime (e.g., [7], [6]). There are many variants of JSQ policy, such as Join-the-Least-Loaded-Queue (JLQ) policy (e.g., [9]) instead forwarding incoming jobs to the queue with the smallest amount of remaining work (or workload), and low-communication overhead load-balancing schemes (e.g., Power-of-Two-Choices [25], [16], Batch Sampling [17], [27]). However, all these

load-balancing schemes presume that queueing disciplines are either First-Come-First-Serve (FCFS) (e.g., [7], [26], [25], [16], [17], [27]) or Processor-Sharing (PS) (e.g., [4], [9]), and their performance is unclear in wireless networks with both channel fading and flow-level dynamics, where each individual flow (or job) faces an independent wireless fading channel. This motivates us to investigate efficient load-balancing design in the presence of flow-level dynamics with wireless fading. The following list highlights our contributions as well as the outline of the remainder of the paper:

- In Section II, we formulate the problem of load-balancing design among multiple APs in high-density wireless networks in the presence of flow-level dynamics.

- In Section III, we present two existing policies and show their performance deficiencies.

- In Section IV, we propose an efficient load-balancing scheme for wireless networks with flow-level dynamics, and show that it not only achieves maximum system throughput but also minimizes the mean system workload in heavily loaded conditions.

- We support our analytical results with extensive simulations in Section V, which not only confirm our theoretical findings but also exhibit the excellent performance of our proposed algorithm in general cases.

A note on Notation: We use bold and script font of a variable to denote a vector and a set, respectively. We use  $\langle \mathbf{x}, \mathbf{y} \rangle$  to denote the inner product of two vectors  $\mathbf{x}$  and  $\mathbf{y}$ . Let  $\|\mathbf{x}\|_1$  and  $\|\mathbf{x}\|$  denote the  $l_1$  and  $l_2$  norm of the vector  $\mathbf{x}$ , respectively. We also use  $\mathbf{x} \succeq \mathbf{y}$  to denote that each component of vector  $\mathbf{x}$  is not less than that of vector  $\mathbf{y}$ .

## II. SYSTEM MODEL

We consider a wireless network with  $M$  access points (APs). We assume that the system operates in a *slotted time* manner. Here, we assume that these  $M$  APs operate in orthogonal channels and can serve users (referred as flows in the rest of the paper) at the same time. However, within each AP, due to the wireless interference, at most one flow can be served in each time slot.

Let  $A_\Sigma[t]$  denote the number of flows arriving at the system in time slot  $t$  that is independently and identically distributed (i.i.d.) over time with mean  $\lambda_\Sigma > 0$ , and  $A_\Sigma[t] \leq A_{\max}$  for some positive number  $A_{\max}$ ,  $\forall t \geq 0$ . We use  $F_j[t]$  to denote the number of packets of newly arriving flow  $j$  that follows any probability distribution with mean  $\eta > 0$ , and  $F_j[t] \leq F_{\max}$  for some  $0 < F_{\max} < \infty$ ,  $\forall t \geq 0$ . We use  $N_m[t]$  to denote the number of flows in AP  $m$  in time slot  $t$ . We also use  $\mathcal{A}_\Sigma[t]$  and  $\mathcal{N}_m[t]$  to denote the set of newly arriving flows at the system and the set of existing flows in AP  $m$  in time slot  $t$ , respectively. Let  $R_{m,j}[t]$  be the number of residual packets of flow  $j$  in AP  $m$  in time slot  $t$ .

We assume that each AP has  $K + 1$  possible channel rates  $c_0, c_1, c_2, \dots, c_K$  with  $0 = c_0 < c_1 < c_2 < \dots < c_K = c_{\max}$ , where  $c_k$  is a positive integer number denoting that at most  $c_k$  packets can be delivered in one time slot,  $\forall k = 1, 2, \dots, K$ . We use  $C_{m,j}[t]$  to capture wireless channel fading of each

flow  $j$  in the  $m^{\text{th}}$  AP, which measures the maximum number of packets that can be transmitted in time slot  $t$  if flow  $j$  is scheduled in time slot  $t$ . We assume that  $(C_{m,j}[t])_{j \in \mathcal{N}_m[t]}$  are independently distributed across APs and i.i.d. over both time and flows within each AP with probability distribution

$$\Pr \{C_{m,j}[t] = c_k\} = p_{m,k}, \forall k = 0, 1, 2, \dots, K. \quad (1)$$

Here, we reasonably assume that both probability that the channel for each flow is unavailable and achieves the maximum channel rate are strictly positive, i.e.,  $p_{m,0} > 0$  and  $p_{m,K} > 0$ ,  $\forall m = 1, 2, \dots, M$ .

In order to characterize the underlying dynamics of flows, we introduce following notations. Let  $W_m[t] \triangleq \sum_{j \in \mathcal{N}_m[t]} \lceil R_j[t]/c_{\max} \rceil$  be the total workload in AP  $m$  in time slot  $t$  that measures the minimum number of slots required for completing all existing service requests in AP  $m$ . We use  $\nu_\Sigma[t] \triangleq \sum_{j \in \mathcal{A}_\Sigma[t]} \lceil F_j[t]/c_{\max} \rceil$  and  $\nu_m[t] \triangleq \sum_{j \in \mathcal{A}_m[t]} \lceil F_j[t]/c_{\max} \rceil$  to denote the total amount of new workload arriving at the system and the amount of new workload injected to AP  $m$  under some load-balancing policy in time slot  $t$ , respectively, where  $\mathcal{A}_m[t]$  denotes the set of arriving flows at AP  $m$  in time slot  $t$ . We also use  $A_m[t]$  to represent the number of newly arriving flows at AP  $m$  in time slot  $t$ . Let  $\rho \triangleq \mathbb{E}[\nu_\Sigma[t]] = \lambda_\Sigma w$  be the traffic intensity, where  $w \triangleq \mathbb{E}[\lceil F_j[t]/c_{\max} \rceil]$  denotes the expected minimum number of slots required for serving a newly arriving flow.

We define  $\mu_m[t]$  to be the amount of workload decreasing at AP  $m$  in time slot  $t$ . Since the maximum channel rate is  $c_{\max}$ ,  $\mu_m[t]$  is equal to either 0 or 1. In addition, if at least one of flows in AP  $m$  has the maximum channel rate  $c_{\max}$  in time slot  $t$ , then  $\mu_m[t] = 1$ . Therefore,  $\mu_m[t] \geq \mathbb{1}_{\mathcal{F}_m}$ , where  $\mathcal{F}_m$  denotes the event that at least one of flows in AP  $m$  has the maximum channel rate  $c_{\max}$ . Based on the above setup, the evolution of the workload  $W_m[t]$  at each AP  $m$  can be described as follows:

$$W_m[t + 1] = W_m[t] + \nu_m[t] - \mu_m[t], \forall m = 1, \dots, M. \quad (2)$$

We call AP  $m$  *stable* if its average workload is finite, i.e.,

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[W_m[t]] < \infty.$$

We say that the system is stable if all its APs are stable. The *capacity region*  $\Lambda$  is defined as a maximum set of traffic intensity  $\rho$  for which the system is stable under some policy. It is shown in Appendix A that  $\Lambda = \{\rho : \rho \leq M\}$ , where we recall that  $M$  is the number of APs. Note that  $\rho$  denotes the average amount of incoming workload, which is the expected minimum number of slots required for serving incoming flows. On the other hand,  $M$  is the maximum amount of workload that can decrease in each time slot. In order to make the system stable,  $\rho$  should not be greater than  $M$ .

The *throughput-optimal* algorithm stabilizes the system for any traffic intensity lying strictly inside the capacity region  $\Lambda$ . In this paper, we focus on the performance of load-balancing schemes that determine which AP should serve the newly

incoming flows. We are interested in a load-balancing scheme for high-density wireless networks that not only support flows as many as possible, but also complete service requests of existing flows as fast as possible. The first goal is equivalent to maximizing system throughput, while the second goal can be achieved by minimizing the total mean system workload that measures the expected minimum number of time slots to finish all existing service requests. In the rest of the paper, we consider the following scheduling policy (see [23]) within each AP: in each time slot, each AP  $m$  always serves a flow  $j_m^*$  with the maximum channel rate among all its existing flows, breaking ties uniformly at random, i.e.,  $j_m^* \in \arg \max_{j \in \mathcal{N}_m[t]} C_{m,j}[t]$ . As we show later, our proposed load-balancing algorithm together with this specific scheduling policy achieves both our desired goals.

Next, we discuss the performance deficiencies of existing policies that motivate us for further investigations.

### III. PERFORMANCE DEFICIENCIES OF EXISTING POLICIES

In this section, we present two existing policies and show their performance deficiencies.

#### A. Throughput Deficiency of Best-Channel-First Policy

In this subsection, we consider the policy in conventional WLANs, where each incoming flow joins the AP with the best channel quality (e.g. [12]), which is given as follows:

---

**Best-Channel-First (BCF) Algorithm:** In each time slot  $t$ , forward each incoming flow to the AP with the largest channel rate, breaking ties uniformly at random.

---

Intuitively, more flows go to the AP with the better channel quality under the BCF Algorithm, and thus the AP with the better channel quality will be congested. This leads to the insufficient usage of APs with the relatively worst channel quality and results in the system throughput performance loss, let alone its mean workload performance. To see the throughput inefficiency of the BCF Algorithm, we consider the system with two APs, where flows at both APs face independent ON-OFF channel fading with different distributions. In particular, let  $p_m \triangleq \Pr\{C_{m,j}[t] = 1\}$  denote the probability that flow  $j$  has an available channel at AP  $m$ , where  $m = 1, 2$ . Without loss of generality, we assume that  $p_1 > p_2$ . For each incoming flow  $j$ , the probability that it will join the first AP under the BCF Algorithm is equal to

$$\Pr\{C_{1,j} = 1, C_{2,j} = 0\} + \frac{1}{2} \Pr\{C_{1,j} = 1, C_{2,j} = 1\} + \frac{1}{2} \Pr\{C_{1,j} = 0, C_{2,j} = 0\} = \frac{1}{2}(1 + p_1 - p_2). \quad (3)$$

Therefore, the traffic intensity to the first AP is equal to  $\rho(1 + p_1 - p_2)/2$ . In order to maintain the stability of the first AP,  $\rho(1 + p_1 - p_2)/2 \leq 1$  should be satisfied, since the workload can decrease at most by one in each AP in each time slot. Thus, the BCF Algorithm can at most support the throughput region:  $\{\rho : \rho \leq 2/(1 + p_1 - p_2)\}$ . However, the capacity region in this case is  $\Lambda = \{\rho : \rho \leq 2\}$ . Therefore,

the BCF Algorithm suffers from throughput performance loss by  $(p_1 - p_2)/(1 + p_1 - p_2)$ . For example, the throughput performance loss is 33.33% when  $p_1 = 0.9$  and  $p_2 = 0.4$ . Fig. 1 illustrates the throughput performance loss percentage with respect to the channel quality difference between two APs (i.e.,  $p_1 - p_2$ ) under the BCF Algorithm. We can observe from Fig. 1 that the throughput performance loss can be as high as 50% in an extreme case when the first AP always has perfect channel quality and the second AP has extremely poor channel quality, i.e.,  $p_1 = 1$  and  $p_2 = 0$ . Moreover, as the channel quality difference between two APs becomes large, the BCF Algorithm suffers from greater throughput performance loss. The reason is that if the channel quality between two APs is quite different, then the incoming flows prefer to join the AP with the better channel quality. On one hand, this results in a large number of flows accumulating at the first AP and leads to traffic congestion. On the other hand, the second AP does not have a sufficient number of flows to serve and is underutilized. In fact, a simple randomized policy that simply forwards each incoming flow to each AP uniformly at random can achieve full capacity region but it suffers from poor mean workload performance, as shown in the next subsection.

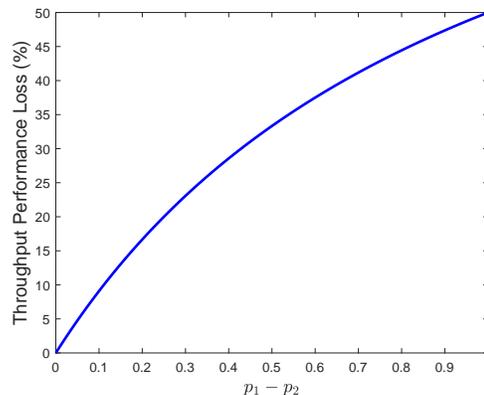


Fig. 1: Throughput loss under the BCF Algorithm

#### B. Mean Workload Deficiency of Randomized Load-Balancing

In this subsection, we consider both throughput and mean workload performance of a randomized load-balancing scheme, which works as follows:

---

**Randomized Load-Balancing (RLB) Algorithm:** In each time slot  $t$ , forward each incoming flow to each AP uniformly at random.

---

The following proposition shows that the RLB Algorithm can indeed achieve the maximum system throughput.

*Proposition 1:* The RLB Algorithm is throughput-optimal, i.e., it stabilizes the system for any traffic intensity  $\rho$  lying strictly inside the capacity region  $\Lambda$ . Moreover, all moments of steady-state workloads are finite.

*Proof:* Under the RLB Algorithm, the number of incoming flows forwarded to each AP in each time slot is i.i.d. with mean  $\lambda_\Sigma/M$  and their flow sizes are also i.i.d. with the same

probability distribution as before. The proof is a special case of that in Proposition 3 in the case with a single AP and the mean flow arrival rate of  $\lambda_\Sigma/M$ , and thus is omitted for simplicity. ■

Note that the RLB Algorithm randomly forwards the incoming flows to each AP and may cause significant workload imbalance across APs especially when the number of APs is relatively large, which is the case in high-density wireless networks. Even though this does not hurt the throughput performance, it results in the large mean workload, which implicitly degrades the mean delay performance of each flow. In order to characterize the mean workload performance, we build on the recently developed approach of using Lyapunov drifts for the steady-state analysis of queueing networks [6]. However, in our considered flow-level dynamic model, flows dynamically arrive at the system and will leave once they receive the desired amount of service, and existing flows suffer from independent channel fading. These two characteristics make our model quite different from the traditional FCFS queueing model that is considered in [6]. Thus, novel techniques are required to analyze the heavy-traffic performance of the RLB Algorithm.

To that end, we consider the workload arrival process  $\{\nu_\Sigma^{(\epsilon)}[t]\}_{t \geq 0}$ , parameterized by  $\epsilon > 0$ , with traffic intensity  $\rho^{(\epsilon)}$  satisfying  $\epsilon = M - \rho^{(\epsilon)} > 0$  and  $\text{Var}(\nu_\Sigma^{(\epsilon)})$ . Here,  $\epsilon$  characterizes the closeness of the traffic intensity to the boundary of the capacity region, and is usually referred as *heavy-traffic parameter*. We are interested in understanding the steady-state workload with vanishing  $\epsilon$ . The next proposition shows that the RLB Algorithm results in the large mean workload even under the Bernoulli flow arrival.

*Proposition 2:* Assume that the number of arriving flows  $A_\Sigma[t]$  follows Bernoulli distribution. Let  $\widetilde{\mathbf{W}}^{(\epsilon)} = (\widetilde{W}_m^{(\epsilon)})_{m=1}^M$  be a random vector with the same distribution as the steady-state distribution of the workload processes under the RLB Algorithm. Consider the heavy-traffic limit  $\epsilon \downarrow 0$ , suppose that the variance  $\text{Var}(\nu_\Sigma^{(\epsilon)})$  of the arrival process  $\{\nu_\Sigma^{(\epsilon)}\}_{t \geq 0}$  converges to a constant  $\sigma^2$ . Then, we have

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m^{(\epsilon)} \right] = \frac{1}{2} (\sigma^2 + M(M-1)). \quad (4)$$

*Proof:* Under the RLB Algorithm, the number of incoming flows joining to the  $m^{\text{th}}$  AP in time slot  $t$  can be represented as follows:

$$A_m[t] = A_\Sigma[t] \mathbb{1}_{\mathcal{H}[t]}, \quad (5)$$

where  $\mathcal{H}[t]$  denotes the event that the incoming flow joins the  $m^{\text{th}}$  AP in time slot  $t$ , which is independent from workload  $W_m[t]$ . In addition, under the RLB Algorithm, the event  $\mathcal{H}[t]$  is i.i.d. with Bernoulli distribution with mean  $1/M$ . Therefore, the workload arriving at AP  $m$  in time slot  $t$  is

$$\nu_m[t] = \nu_\Sigma[t] \mathbb{1}_{\mathcal{H}[t]}, \quad (6)$$

where  $\nu_\Sigma = A_\Sigma[t] \lceil F_j[t]/c_{\max} \rceil$  since  $A_\Sigma[t]$  follows Bernoulli

distribution. Thus, we have

$$\mathbb{E}[\nu_m[t]] = \frac{1}{M} \rho^{(\epsilon)},$$

and  $\text{Var}(\nu_m[t]) = \frac{1}{M} \text{Var}(\nu_\Sigma^{(\epsilon)}) + \frac{M-1}{M^2} (\rho^{(\epsilon)})^2. \quad (7)$

Hence, we have

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} \left[ \widetilde{W}_m^{(\epsilon)} \right] = \frac{1}{2} \left( \frac{1}{M} \sigma^2 + M - 1 \right), \quad (8)$$

which implies the desired result by summing over  $m = 1, 2, \dots, M$ . The proof of (8) is a special case of that in Proposition 4 and Proposition 5 in the case with a single AP and the arrival workload  $\nu_m[t]$ , and thus is omitted for conciseness. ■

From Proposition 2, we can observe that the mean total workload under the RLB Algorithm increases quadratically with the number of APs, which is undesirable in high-density networks even when  $M = 10$ . The main reason lies in that the RLB Algorithm randomly makes the load-balancing decision, and does not fully utilize the precious network resources. This motivates us to develop a throughput-optimal load-balancing scheme under which the mean total workload is minimized and does not suffer from performance loss as the number of APs scales. This property is pronounced in highly-dense wireless networks in the presence of many APs.

#### IV. EFFICIENT LOAD-BALANCING DESIGN

In this section, we first propose a workload-aware load-balancing algorithm. Then, we show that the proposed algorithm not only achieves maximum system throughput but also minimizes the mean total workload in the heavy-traffic regime.

As we discussed in the last section, the inefficiency of both BCF and RLB Algorithms lie in the fact that they are not aware of system workloads and thus cause significant load imbalance among multiple APs. This motivates us to develop a workload-aware load-balancing scheme that can evenly distribute incoming workloads across multiple APs. Motivated by the design of JSQ and JLQ polices in data centers, we propose the following workload-aware load-balancing algorithm that aims to balance workloads across multiple APs in the presence of dynamic flows.

**Join-the-Least-Workload (JLW) Algorithm:** In each time slot  $t$ , given the current workload  $\mathbf{W}[t] = (W_m[t])_{m=1}^M$ , forward all the arriving flows to the AP with the smallest workload, i.e.,

$$\mathbf{A}^*[t] \in \arg \min_{\mathbf{A}=(A_m)_{m=1}^M \geq \mathbf{0}; \sum_{m=1}^M A_m = A_\Sigma[t]} \langle \mathbf{A}, \mathbf{W}[t] \rangle, \quad (9)$$

breaking ties uniformly at random.

In the JLW Algorithm, the central controller broadcasts the ID of the AP with the least workload in each time slot, and thus all arriving flows will join that AP. This is possible in high-density wireless networks, where all APs are interconnected. The main difference between our JLW Algorithm and the

JLQ Algorithm lies in that the JLQ Algorithm is mainly designed for the system with FCFS or PS queueing discipline, while each flow in our scenario faces an independent channel fading and potentially may have different service rate. This poses significant challenges for the performance analysis of the JLW Algorithm. Nevertheless, we can still show that the JLW Algorithm achieves maximum system throughput. Moreover, we can show that all moments of steady-state workload are bounded under the JLW Algorithm, which enables us to analyze the mean workload performance by using the Lyapunov-type approach developed in [6].

*Proposition 3:* The JLW Algorithm is throughput-optimal, i.e., it stabilizes the system for any traffic intensity lying strictly inside the capacity region  $\Lambda$ . Moreover, all moments of steady-state workloads are bounded.

*Proof:* The proof is available in Section VI. ■

Having established the throughput optimality and the moment existence of the steady-state workload of the JLW Algorithm, we are ready to analyze the mean workload performance in the heavy-traffic regime. Similar to the heavy-traffic analysis of the RLB Algorithm, we consider the arrival process  $\{\nu_{\Sigma}^{(\epsilon)}[t]\}_{t \geq 0}$  with heavy-traffic parameter  $\epsilon > 0$  characterizing the closeness of the traffic intensity to the boundary of the capacity region, i.e.,  $\epsilon \triangleq M - \rho^{(\epsilon)} > 0$ .

*Proposition 4:* Let  $\widetilde{\mathbf{W}}^{(\epsilon)} = (\widetilde{W}_m^{(\epsilon)})_{m=1}^M$  be a random vector with the same distribution as the steady-state distribution of the workload processes under the JLW Algorithm. Consider the heavy-traffic limit  $\epsilon \downarrow 0$ , suppose that the variance  $\text{Var}(\nu_{\Sigma}^{(\epsilon)})$  of the arrival process  $\{\nu_{\Sigma}^{(\epsilon)}\}_{t \geq 0}$  converges to a constant  $\sigma^2$ . Then, we have

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m^{(\epsilon)} \right] \leq \frac{\sigma^2}{2}. \quad (10)$$

*Proof:* The proof is available in Section VII. ■

In fact, the upper bound in (10) is also tight. To see this, we provide a generic lower bound for all feasible load-balancing policies by constructing a hypothetical single-server queue  $\{\Phi[t]\}_{t \geq 0}$  with the arrival process  $\{\nu_{\Sigma}^{(\epsilon)}[t]\}_{t \geq 0}$  and the constant service rate  $M$ . The queue-length evolution of this single-server queue can be described as follows:

$$\Phi[t+1] = \max \left\{ \Phi[t] + \nu_{\Sigma}^{(\epsilon)}[t] - M, 0 \right\}. \quad (11)$$

It is easy to see that the constructed single-server queue length  $\{\Phi[t]\}_{t \geq 0}$  is stochastically smaller than the total workload process  $\{\sum_{m=1}^M W_m[t]\}_{t \geq 0}$  of the original system under any feasible policy, since the total system workload can at most decrease by  $M$  in one time slot. Hence, by using [6, Lemma 4] for the constructed single-server queue, we have the following lower bound on the steady-state workload under any feasible policy.

*Proposition 5:* Let  $\widetilde{\mathbf{W}}^{(\epsilon)} = (\widetilde{W}_m^{(\epsilon)})_{m=1}^M$  be a random vector with the same distribution as the steady-state distribution of the workload processes under any feasible load-balancing policy. Consider the heavy-traffic limit  $\epsilon \downarrow 0$ , suppose that the variance  $\text{Var}(\nu_{\Sigma}^{(\epsilon)})$  of the arrival process  $\{\nu_{\Sigma}^{(\epsilon)}\}_{t \geq 0}$  converges to a

constant  $\sigma^2$ . Then,

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m^{(\epsilon)} \right] \geq \frac{\sigma^2}{2}. \quad (12)$$

This together with Proposition 4 establishes the heavy-traffic optimality of our proposed JLW Algorithm. Moreover, compared with the mean workload performance under the RLB Algorithm in the heavy-traffic regime, the mean workload under our JLW Algorithm does not incur any performance loss by increasing the number of APs. This desirable property implies that the JLW Algorithm is suitable for deployment in high-density wireless networks.

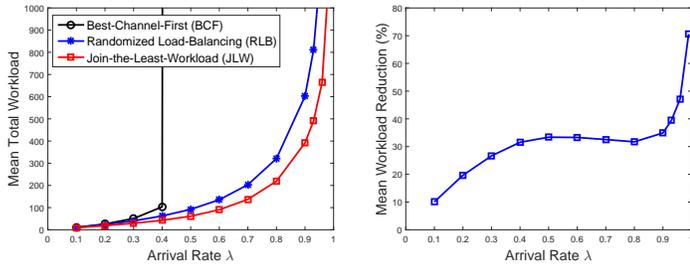
## V. SIMULATION RESULTS

In this section, we provide simulation results for our proposed JLW Algorithm and compare its performance to both BCF and RLB Algorithms. In the simulations, we assume that the number of flows arriving at the system in each time slot follows a Bernoulli distribution with mean  $\lambda$ . Each flow at each AP faces i.i.d. channel fading with rates 0, 1, 5, 10 and corresponding probability 0.1, 0.2, 0.5, 0.2. The flow size  $F$  is equal to  $10 \times \beta$  with probability  $(w-1)/(\beta-1)$  and 10 otherwise, where we recall that  $w = \mathbb{E}[\lceil F/c_{\max} \rceil]$  is the mean newly arriving workload and  $\beta \geq 2$  is some parameter that measures the variance of the newly arriving workload. Indeed, the variance of the newly arriving workload in this setup is equal to  $(w-1)\beta - w(w-1)$ , which linearly increases with the parameter  $\beta$ . We let  $w$  be equal to the number of APs  $M$  and thus the capacity region  $\Lambda$  is  $\{\lambda : 0 < \lambda \leq 1\}$ . We set  $M = 5$  and  $\beta = 20$  in the simulations, unless we specifically mention them.

### A. Throughput Performance

Fig. 2a shows the mean total workload performance versus the mean arrival rate under the BCF Algorithm, the RLB Algorithm and our proposed JLW Algorithm. We can observe from Fig. 2a that both JLW and RLB Algorithms can stabilize the system for any arrival rate  $\lambda$  between 0 and 1, which validate their throughput optimality (cf. Proposition 1 and Proposition 3). In contrast, the BCF Algorithm cannot support the arrival rate of  $\lambda = 0.45$ , where the mean workload blows up. This also matches our discussions about the throughput deficiency of the BCF Algorithm in Section III-A. In addition, we can see that the mean workload under the JLW Algorithm is smaller than that under the RLB Algorithm under different arrival rates. To see it more clearly, Fig. 2b characterizes the mean workload reduction percentage by the JLW Algorithm compared with the RLB Algorithm. From Fig. 2b, we can observe that the mean workload reduction is 10% even when the arrival rate is equal to 0.1, and can reach as high as 70% when the arrival rate is 0.99. The reason is that the RLB Algorithm randomly forwards newly arriving flows to each AP and thus may cause some APs underutilized, while the JLW Algorithm aims to balance the workloads across APs and utilizes network resources more efficiently. Thus, the JLW

Algorithm shows significant performance gain over the RLB Algorithm especially when the arrival rate is high.

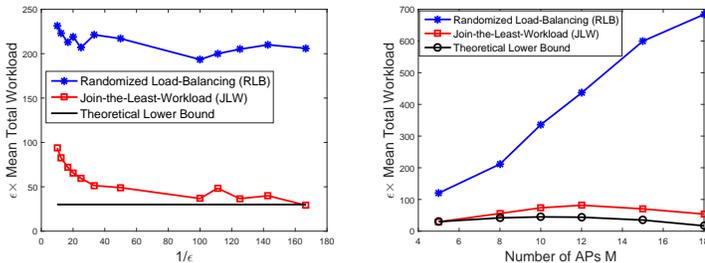


(a) Throughput performance validation (b) Workload reduction by JLW

Fig. 2: The workload performance of the JLW Algorithm

### B. Heavy-Traffic Performance

Fig. 3a shows the impact of heavy-traffic parameter  $\epsilon$  on the mean total workload under both RLB and JLW Algorithms. From Fig. 3a, we can observe that the mean total workload under the JLW Algorithm converges to the theoretical lower bound (equal to 30) derived in Proposition 5, while the RLB Algorithm always keeps it away from the theoretical lower bound. This confirms the heavy-traffic optimality of the JLW Algorithm, i.e., it minimizes the mean total workload as the heavy-traffic parameter  $\epsilon$  diminishes.



(a) Heavy-traffic optimality validation (b) Impact of number of APs

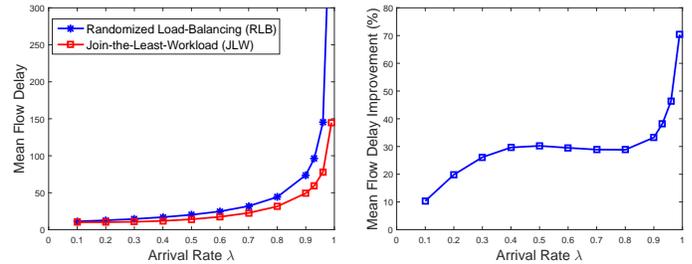
Fig. 3: Mean workload in heavy-traffic regimes

Fig. 3b studies the impact of number of APs  $M$  on the mean total workload under both RLB and JLW Algorithms, where we fix  $\epsilon$  to 0.006 and vary the number of APs from 5 to 18. From Fig. 3b, we can observe that the performance of the JLW Algorithm stays close to the theoretical lower bound, which is equal to  $(21M - 20 - M^2)/2$  from Proposition 5 under our setting. In contrast, the product of mean total workload and heavy-traffic parameter  $\epsilon$  under the RLB Algorithm is equal to  $10M - 10$  from Proposition 2 in the heavy-traffic regime, which linearly increases with the number of APs (also can be observed from Fig. 3b). This renders infeasibility of the RLB Algorithm for high-density wireless networks with many APs.

### C. Mean Delay Performance

In this subsection, we study the mean delay performance of flows under both RLB and JLW Algorithms. From Fig. 4a, we can observe that the JLW Algorithm outperforms the RLB Algorithm in terms of mean delay performance. Moreover, the delay improvement by the JLW Algorithm is very similar to its workload reduction compared with the RLB Algorithm. The reason lies in that the smaller workload implies that each flow

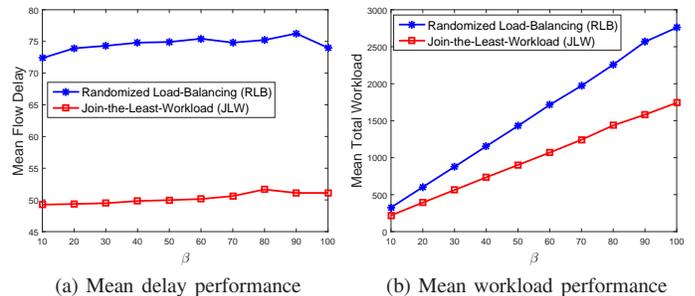
spends the less waiting time in the system and thus experiences smaller delay.



(a) Comparison between RLB and JLW (b) Delay improvement by JLW

Fig. 4: The delay performance of the JLW Algorithm

Next, we study the impact of the variance of the flow size on the system performance. Recall that the parameter  $\beta$  characterizes the variance of the flow size. The larger the  $\beta$ , the higher variance of the flow size. Here, we fix the arrival rate  $\lambda$  to 0.9, and vary the parameter  $\beta$  from 10 to 100. We can see from Fig. 5a that the JLW Algorithm always performs better than the RLB Algorithm, and the parameter  $\beta$  does not affect their mean delay performance. This is referred as the *delay insensitivity* property in queuing literature, where the mean delay performance is insensitive to the flow-size distribution beyond its mean. This is expected, since each AP always serves the flow with the maximum channel rate, and in the extreme non-fading case, flows are served in a preemptive random order, where the mean delay performance exhibits insensitivity property to the flow size distribution (see [11]). However, different from the mean delay performance, the mean total workload under both RLB and JLW Algorithms increases linearly with the parameter  $\beta$ , as shown in Fig. 5b. The reason lies in the fact that the total system workload is lower-bounded by the queue-length of a hypothetical single-server queue  $\{\Phi[t]\}_{t \geq 0}$  with the new workload arrival process  $\{\nu_\Sigma[t]\}_{t \geq 0}$  and the constant service rate  $M$  under any feasible load-balancing policies (also see the discussion in Section IV), where the mean queue-length  $\mathbb{E}[\Phi[t]]$  is sensitive to the variance of  $\nu_\Sigma[t]$ .



(a) Mean delay performance (b) Mean workload performance

Fig. 5: The impact of the flow size distribution

## VI. THROUGHPUT OPTIMALITY ANALYSIS

In this section, we establish the throughput optimality of the JLW algorithm as well as the boundedness of all moments of steady-state workload, where the later property enables us to

analyze the mean workload performance in Section VII. We choose the Lyapunov function

$$V(\mathbf{W}) \triangleq \|\mathbf{W}\|. \quad (13)$$

Then, we consider the conditional expectation of its drift  $\Delta V(\mathbf{W}) \triangleq (V(\mathbf{W}[t+1]) - V(\mathbf{W}[t])) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}}$ .

$$\begin{aligned} & \mathbb{E}[\Delta V(\mathbf{W})|\mathbf{W}[t]=\mathbf{W}] \\ &= \mathbb{E}[V(\mathbf{W}[t+1]) - V(\mathbf{W}[t])|\mathbf{W}[t]=\mathbf{W}] \\ &= \mathbb{E}\left[\sqrt{\|\mathbf{W}[t+1]\|^2} - \sqrt{\|\mathbf{W}[t]\|^2} \middle| \mathbf{W}[t]=\mathbf{W}\right] \\ &\leq \frac{1}{2\|\mathbf{W}\|} \mathbb{E}[L(\mathbf{W}[t+1]) - L(\mathbf{W}[t])|\mathbf{W}[t]=\mathbf{W}], \quad (14) \end{aligned}$$

where the last step is true for  $L(\mathbf{W}) \triangleq \|\mathbf{W}\|^2$  and follows from the fact that  $f(x) = \sqrt{x}$  is concave for  $x \geq 0$  and thus  $f(y) - f(x) \leq f'(x)(y-x) = (y-x)/(2\sqrt{x})$  with  $y = \|\mathbf{W}[t+1]\|^2$  and  $x = \|\mathbf{W}[t]\|^2$ .

Next, we focus on the expected difference in (14), which is just the expected drift of  $L(\mathbf{W})$ . We will omit the time index  $[t]$  after the first step for conciseness.

$$\begin{aligned} \mathbb{E}[\Delta L(\mathbf{W})|\mathbf{W}] &= \mathbb{E}[\|\mathbf{W}[t+1]\|^2 - \|\mathbf{W}[t]\|^2|\mathbf{W}] \\ &= \mathbb{E}[\|\mathbf{W} + \boldsymbol{\nu} - \boldsymbol{\mu}\|^2 - \|\mathbf{W}\|^2|\mathbf{W}] \\ &= \mathbb{E}[2\langle \mathbf{W}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle + \|\boldsymbol{\nu} - \boldsymbol{\mu}\|^2|\mathbf{W}] \\ &\leq 2\mathbb{E}[\langle \mathbf{W}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle|\mathbf{W}] + K_1, \quad (15) \end{aligned}$$

where  $K_1 \triangleq M(\nu_{\max}^2 + 1)$  is bounded and  $\nu_{\max} \triangleq A_{\max}[F_{\max}/c_{\max}]$ .

Next, we focus on  $\mathbb{E}[\langle \mathbf{W}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle|\mathbf{W}]$ . Since the traffic intensity  $\rho$  is strictly inside the capacity region  $\Lambda$ , there exists an  $\epsilon$  such that  $\epsilon = M - \rho > 0$ . We define a hypothetical arrival rate vector  $\boldsymbol{\lambda} = (\lambda_m)_{m=1}^M$  as  $\boldsymbol{\lambda} = \mathbf{1}/w - \epsilon\mathbf{1}/(wM)$ , where we recall that  $w = \mathbb{E}[\lceil F_j[t]/c_{\max} \rceil]$ . Hence, we have  $\sum_{m=1}^M \lambda_m = M/w - \epsilon/w = \lambda_{\Sigma}$ , where we use the fact that  $\rho = \lambda_{\Sigma}w$ . Therefore, we have

$$\begin{aligned} & \mathbb{E}[\langle \mathbf{W}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle|\mathbf{W}] \\ &\stackrel{(a)}{=} \langle \mathbf{W}, w\mathbb{E}[\mathbf{A}|\mathbf{W}] - w\boldsymbol{\lambda} \rangle - \langle \mathbf{W}, \mathbf{1} - w\boldsymbol{\lambda} \rangle - \mathbb{E}[\langle \mathbf{W}, \boldsymbol{\mu} - \mathbf{1} \rangle|\mathbf{W}] \\ &\stackrel{(b)}{\leq} w\langle \mathbf{W}, \mathbb{E}[\mathbf{A}|\mathbf{W}] - \boldsymbol{\lambda} \rangle - \frac{\epsilon}{M}\|\mathbf{W}\|_1 + \mathbb{E}\left[\sum_{m=1}^M W_m \mathbb{1}_{\overline{\mathcal{F}}_m} \middle| \mathbf{W}\right], \quad (16) \end{aligned}$$

where step (a) uses the fact that  $\mathbb{E}[\boldsymbol{\nu}|\mathbf{W}] = w\mathbb{E}[\mathbf{A}|\mathbf{W}]$ ; (b) uses the fact that  $\mu_m \geq \mathbb{1}_{\mathcal{F}_m}$  for all  $m = 1, 2, \dots, M$  and we recall that  $\mathcal{F}_m$  denotes the event that at least one flow has the maximum channel rate  $c_{\max}$  and  $\overline{\mathcal{F}}_m$  is the complement of the event  $\mathcal{F}_m$ .

For the term  $\langle \mathbf{W}, \mathbb{E}[\mathbf{A}|\mathbf{W}] - \boldsymbol{\lambda} \rangle$  in (16), we have

$$\begin{aligned} \langle \mathbf{W}, \mathbb{E}[\mathbf{A}|\mathbf{W}] - \boldsymbol{\lambda} \rangle &= W_{\min} \mathbb{E}[A_{\Sigma}|\mathbf{W}] - \langle \mathbf{W}, \boldsymbol{\lambda} \rangle \\ &= W_{\min} \lambda_{\Sigma} - \sum_{m=1}^M \lambda_m W_m \\ &= \sum_{m=1}^M \lambda_m (W_{\min} - W_m) \\ &\leq 0, \quad (17) \end{aligned}$$

where the first step follows from the definition of the JLW Algorithm.

With regard to the term  $\mathbb{E}\left[\sum_{m=1}^M W_m \mathbb{1}_{\overline{\mathcal{F}}_m} \middle| \mathbf{W}\right]$  in (16), we have

$$\begin{aligned} & \mathbb{E}\left[\sum_{m=1}^M W_m \mathbb{1}_{\overline{\mathcal{F}}_m} \middle| \mathbf{W}\right] = \mathbb{E}\left[\sum_{m=1}^M W_m (1 - p_{m,K})^{N_m} \middle| \mathbf{W}\right] \\ &\stackrel{(a)}{\leq} \mathbb{E}\left[\sum_{m=1}^M W_m (1 - p_K^{\min})^{N_m} \middle| \mathbf{W}\right] \\ &\stackrel{(b)}{\leq} \mathbb{E}\left[\sum_{m=1}^M W_m (1 - p_K^{\min})^{W_m/w_{\max}} \middle| \mathbf{W}\right] \\ &\stackrel{(c)}{=} \mathbb{E}\left[\sum_{m=1}^M W_m (1 - p_K^{\min})^{W_m/w_{\max}} \mathbb{1}_{\{W_m \leq \bar{w}_m\}} \right. \\ &\quad \left. + \sum_{m=1}^M W_m (1 - p_K^{\min})^{W_m/w_{\max}} \mathbb{1}_{\{W_m > \bar{w}_m\}} \middle| \mathbf{W}\right] \\ &\stackrel{(d)}{\leq} \mathbb{E}\left[\sum_{m=1}^M W_m \mathbb{1}_{\{W_m \leq \bar{w}_m\}} + \sum_{m=1}^M \mathbb{1}_{\{W_m > \bar{w}_m\}} \middle| \mathbf{W}\right] \stackrel{(e)}{\leq} K_2, \quad (18) \end{aligned}$$

where step (a) is true for  $p_K^{\min} \triangleq \min_m p_{m,K} > 0$ ; (b) is true for  $w_{\max} = \lceil F_{\max}/c_{\max} \rceil$  and follows from the fact that given the workload  $W_m$ , the number of flows at AP  $m$  is at least  $W_m/w_{\max}$  (i.e.,  $N_m \geq W_m/w_{\max}$ ); (c) is true for some constant  $\bar{w}_m > 0$  such that  $W_m(1 - p_K^{\min})^{W_m/w_{\max}} \leq 1$  holds whenever  $W_m > \bar{w}_m$  (since  $\lim_{W_m \rightarrow \infty} W_m(1 - p_K^{\min})^{W_m/w_{\max}} = 0$ ); (d) uses the fact that  $(1 - p_K^{\min})^{W_m/w_{\max}} \leq 1$  and the definition of  $\bar{w}_m$ ; (e) is true for  $K_2 \triangleq \sum_{m=1}^M (\bar{w}_m + 1)$ .

By combining (16), (17) and (18), and substituting it into (15), we have

$$\begin{aligned} \mathbb{E}[\Delta L(\mathbf{W})|\mathbf{W}] &\leq -\frac{2\epsilon}{M}\|\mathbf{W}\|_1 + K_1 + 2K_2 \\ &\leq -\frac{2\epsilon}{M}\|\mathbf{W}\| + K_1 + 2K_2 \quad (19) \end{aligned}$$

where the last step uses the fact that  $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|$  for any vector  $\mathbf{x}$ .

By substituting (19) into (14), we have

$$\mathbb{E}[\Delta V(\mathbf{W})|\mathbf{W}] \leq -\frac{\epsilon}{M} + \frac{K_1 + 2K_2}{2V(\mathbf{W})}. \quad (20)$$

This implies that when  $V(\mathbf{W})$  is sufficiently large, its condi-

tional expected drift is strictly negative.

Next, we will show that the drift of  $V(\mathbf{W})$  is also bounded, which together with (20) establishes the desired result by [10, Theorem 2.3].

$$\begin{aligned}
|\Delta V(\mathbf{W})| &= \|\|\mathbf{W}[t+1]\| - \|\mathbf{W}[t]\|\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(a)}{\leq} \|\|\mathbf{W}[t+1] - \mathbf{W}[t]\|\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(b)}{\leq} \|\|\mathbf{W}[t+1] - \mathbf{W}[t]\|_1\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\leq M \max_m |W_m[t+1] - W_m[t]| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\leq M(\nu_{\max} + 1), \tag{21}
\end{aligned}$$

where step (a) follows from the triangle inequality for vectors  $\mathbf{x}$  and  $\mathbf{y}$ , i.e.,  $\|\|\mathbf{x}\| - \|\mathbf{y}\|\| \leq \|\mathbf{x} - \mathbf{y}\|$ ; (b) uses the fact that  $\|\|\mathbf{x}\| \leq \|\mathbf{x}\|_1$  for any vector  $\mathbf{x}$ .

## VII. HEAVY-TRAFFIC ANALYSIS

In this section, we provide a proof of Proposition 4. In particular, we show that the proposed JLW Algorithm minimizes the expected workload in the heavy-traffic regime. The proof includes two parts: 1) showing state-space collapse; 2) using the state-space collapse result to obtain an upper bound on the mean workload. Yet, it is worth noting that each flow faces an independent channel fading and may have different service rate and thus its evolution is quite different from traditional FCFS queueing systems. Thus, it calls for a novel technique to establish heavy-traffic optimality of the JLW Algorithm.

### A. State-Space Collapse

In this subsection, we establish a state-space collapse result under the JLW Algorithm. That is, we develop the upper bound for the deviation of steady-state workloads from their average. This state-space collapse happens because JLW routes each arrival to the AP with the smaller workload to balance the workload across all APs.

Let  $\{\mathbf{W}^{(\epsilon)}[t]\}_{t \geq 0}$  be the workload process under the JLW Algorithm, where we recall that the heavy-traffic parameter  $\epsilon$  characterizes the closeness of the traffic intensity  $\rho$  and the boundary of the capacity region  $\Lambda$ , i.e.,  $\epsilon = M - \rho^{(\epsilon)} > 0$ . Proposition 3 shows that all moments of the steady-state workload exist. To that end, we use  $\widetilde{\mathbf{W}}^{(\epsilon)}$  to denote the steady-state workload random vector. Note that the JLW Algorithm tries to equalize the workload across APs, and thus we expect the state space collapses along the direction of a unit vector, all of whose components are equal, i.e.,  $\mathbf{c} \triangleq (1/\sqrt{M})_{m=1}^M$ . Note that  $\mathbf{W}^{(\epsilon)}[t] \Rightarrow \widetilde{\mathbf{W}}^{(\epsilon)}$  due to Proposition 3, where  $\Rightarrow$  denotes convergence in distribution. Then, by the continuous mapping theorem, we have  $\mathbf{W}_{\parallel}^{(\epsilon)}[t] \Rightarrow \widetilde{\mathbf{W}}_{\parallel}^{(\epsilon)}$ , and  $\mathbf{W}_{\perp}^{(\epsilon)}[t] \Rightarrow \widetilde{\mathbf{W}}_{\perp}^{(\epsilon)}$ , where the projection and the perpendicular vector of any given  $M$ -dimensional vector  $\mathbf{I} = (I_m)_{m=1}^M$  with respect to the vector  $\mathbf{c}$  are defined as follows:

$$\mathbf{I}_{\parallel} \triangleq \langle \mathbf{I}, \mathbf{c} \rangle \mathbf{c} = \frac{I_{\Sigma}}{M} \mathbf{1}, \text{ and } \mathbf{I}_{\perp} \triangleq \mathbf{I} - \mathbf{I}_{\parallel} = \left( I_m - \frac{I_{\Sigma}}{M} \right)_{m=1}^M,$$

respectively, and  $I_{\Sigma} \triangleq \sum_{m=1}^M I_m$ , where  $\mathbf{1}$  is  $M$ -dimensional vector of ones.

Next, we will show that under the JLW Algorithm, all moments of  $\widetilde{\mathbf{W}}_{\perp}^{(\epsilon)}$  are bounded by some constants independent of heavy-traffic parameter  $\epsilon > 0$ .

*Proposition 6:* For any  $\delta \in (0, 1/(2w))$ , under the JLW Algorithm, there exists a sequence of finite positive numbers  $\{H_n\}_{n=1,2,\dots}$  such that

$$\mathbb{E} \left[ \|\widetilde{\mathbf{W}}_{\perp}^{(\epsilon)}\|^n \right] \leq H_n, \forall n = 1, 2, \dots \tag{22}$$

for all  $\epsilon \in (0, M/2)$ , where we recall that  $w$  is the mean workload of a newly arriving flow.

*Proof:* In the following proof, we will omit  $\epsilon$  associated with the workload processes for ease of exposition. We consider the Lyapunov function  $V_{\perp}(\mathbf{W}) \triangleq \|\mathbf{W}_{\perp}\|$ , and its drift is defined as

$$\Delta V_{\perp}(\mathbf{W}) \triangleq (V_{\perp}(\mathbf{W}[t+1]) - V_{\perp}(\mathbf{W}[t])) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}}. \tag{23}$$

Since the workload process  $\{\mathbf{W}[t]\}_{t \geq 0}$  has both bounded increments and decrements, we can show that the drift of  $V_{\perp}(\mathbf{W}[t])$  is absolutely bounded by some positive constant for all workload vector  $\mathbf{W}$ . Indeed, we have

$$\begin{aligned}
|\Delta V_{\perp}(\mathbf{W})| &= \|\|\mathbf{W}_{\perp}[t+1]\| - \|\mathbf{W}_{\perp}[t]\|\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(a)}{\leq} \|\|\mathbf{W}_{\perp}[t+1] - \mathbf{W}_{\perp}[t]\|\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(b)}{=} \|\|\mathbf{W}[t+1] - \mathbf{W}[t] - (\mathbf{W}_{\parallel}[t+1] - \mathbf{W}_{\parallel}[t])\|\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(c)}{\leq} \left( \|\|\mathbf{W}[t+1] - \mathbf{W}[t]\| + \left\| (\mathbf{W}[t+1] - \mathbf{W}[t])_{\parallel} \right\| \right) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(d)}{\leq} 2\|\|\mathbf{W}[t+1] - \mathbf{W}[t]\|\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(e)}{\leq} 2\|\|\mathbf{W}[t+1] - \mathbf{W}[t]\|_1\| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\
&\stackrel{(f)}{\leq} 2M \max_m |W_m[t+1] - W_m[t]| \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \leq 2M(\nu_{\max} + 1),
\end{aligned}$$

where step (a) uses the triangle inequality for vectors  $\mathbf{x}$  and  $\mathbf{y}$ , i.e.,  $\|\|\mathbf{x}\| - \|\mathbf{y}\|\| \leq \|\mathbf{x} - \mathbf{y}\|$ ; (b) follows from the definition of  $\mathbf{W}_{\perp} \triangleq \mathbf{W} - \mathbf{W}_{\parallel}$ ; (c) uses the fact that  $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$  for two vectors  $\mathbf{x}$  and  $\mathbf{y}$  and the fact that  $\mathbf{x}_{\parallel} - \mathbf{y}_{\parallel} = (\mathbf{x} - \mathbf{y})_{\parallel}$ ; (d) uses the fact that  $\|\mathbf{x}_{\parallel}\| \leq \|\mathbf{x}\|$ ; (e) uses the fact that  $\|\mathbf{x}\| \leq \|\mathbf{x}\|_1$  for any vector  $\mathbf{x}$ ; (f) is true for  $\nu_{\max} \triangleq A_{\max} \lceil F_{\max}/c_{\max} \rceil$  and follows from (2).

Next, we will show that when  $V_{\perp}(\mathbf{W})$  is sufficiently large, it has a strictly negative drift independent of  $\epsilon$ . This together with the absolute boundedness of the drift establishes the desired result by [10, Theorem 2.3]. However, it is not easy to directly study the drift of  $\|\mathbf{W}_{\perp}\|$ . Instead, it is easier to study the drift of  $\|\mathbf{W}\|^2$  and  $\|\mathbf{W}_{\parallel}\|^2$ , which provides a proper upper bound

on the drift of  $\|\mathbf{W}_\perp\|$ . Indeed,

$$\begin{aligned} \Delta V_\perp(\mathbf{W}) &= (V_\perp(\mathbf{W}[t+1]) - V_\perp(\mathbf{W}[t])) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\ &= \left( \sqrt{\|\mathbf{W}_\perp[t+1]\|^2} - \sqrt{\|\mathbf{W}_\perp[t]\|^2} \right) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\ &\stackrel{(a)}{\leq} \frac{1}{2\|\mathbf{W}_\perp[t]\|} (\|\mathbf{W}_\perp[t+1]\|^2 - \|\mathbf{W}_\perp[t]\|^2) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \\ &\stackrel{(b)}{=} \frac{1}{2\|\mathbf{W}_\perp\|} (\Delta L(\mathbf{W}) - \Delta L_\parallel(\mathbf{W})), \end{aligned} \quad (24)$$

where step (a) follows from the fact that  $f(x) = \sqrt{x}$  is concave for  $x \geq 0$  and thus  $f(y) - f(x) \leq f'(x)(y-x) = (y-x)/(2\sqrt{x})$  with  $y = \|\mathbf{W}_\perp[t+1]\|^2$  and  $x = \|\mathbf{W}_\perp[t]\|^2$ ; (b) uses the fact that  $\|\mathbf{x}_\perp\|^2 = \|\mathbf{x}\|^2 - \|\mathbf{x}_\parallel\|^2$  for any vector  $\mathbf{x}$ , and is true for  $L(\mathbf{W}) \triangleq \|\mathbf{W}\|^2$ ,  $L_\parallel(\mathbf{W}) \triangleq \|\mathbf{W}_\parallel\|^2$ , and

$$\Delta L(\mathbf{W}) \triangleq (L(\mathbf{W}[t+1]) - L(\mathbf{W}[t])) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}} \quad (25)$$

$$\Delta L_\parallel(\mathbf{W}) \triangleq (L_\parallel(\mathbf{W}[t+1]) - L_\parallel(\mathbf{W}[t])) \mathbb{1}_{\{\mathbf{W}[t]=\mathbf{W}\}}. \quad (26)$$

Next, we consider the conditional expectations of  $\Delta L(\mathbf{W})$  and  $\Delta L_\parallel(\mathbf{W})$ , respectively. From (15), (16), and (18), we have

$$\begin{aligned} \mathbb{E}[\Delta L(\mathbf{W})|\mathbf{W}] &\leq 2w\langle \mathbf{W}, \mathbb{E}[\mathbf{A}|\mathbf{W}] - \boldsymbol{\lambda} \rangle - \frac{2\epsilon}{M} \|\mathbf{W}\|_1 \\ &\quad + K_1 + 2K_2, \end{aligned} \quad (27)$$

where  $K_1$  and  $K_2$  are some positive constants.

Next, we consider the term  $\langle \mathbf{W}, \mathbb{E}[\mathbf{A}|\mathbf{W}] - \boldsymbol{\lambda} \rangle$  in (27).

$$\begin{aligned} \langle \mathbf{W}, \mathbb{E}[\mathbf{A}|\mathbf{W}] - \boldsymbol{\lambda} \rangle &\stackrel{(a)}{=} W_{\min} \mathbb{E}[A_\Sigma|\mathbf{W}] - \langle \mathbf{W}, \boldsymbol{\lambda} \rangle \\ &= W_{\min} \lambda_\Sigma - \sum_{m=1}^M \lambda_m W_m \\ &= - \sum_{m=1}^M \lambda_m (W_m - W_{\min}) \\ &\stackrel{(b)}{\leq} - \lambda_{\min} \sum_{m=1}^M |W_m - W_{\min}| \\ &= - \lambda_{\min} \|\mathbf{W} - W_{\min} \mathbf{1}\|_1 \\ &\stackrel{(c)}{\leq} - \lambda_{\min} \|\mathbf{W} - W_{\min} \mathbf{1}\| \\ &\stackrel{(d)}{\leq} - \lambda_{\min} \left\| \mathbf{W} - \frac{1}{M} W_\Sigma \mathbf{1} \right\| \\ &\stackrel{(e)}{\leq} - \delta \|\mathbf{W}_\perp\|, \end{aligned} \quad (28)$$

where step (a) is true for  $W_{\min} \triangleq \min_m W_m$  and follows from the definition of the JLW Algorithm; (b) is true for  $\lambda_{\min} \triangleq \min_m \lambda_m$ ; (c) follows from the fact that  $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|$  for any vector  $\mathbf{x}$ ; (d) uses the fact that  $W_\Sigma/M$  minimizes the convex function  $\|\mathbf{W} - y\mathbf{1}\|$  over  $y \in \mathbb{R}$ ; (e) is true since  $\lambda_{\min} > \delta$  for any  $\delta \in (0, 1/(2w))$  and  $\epsilon \in (0, M/2)$ .

By substituting (28) into (27), we have

$$\mathbb{E}[\Delta L(\mathbf{W})|\mathbf{W}] \leq -\frac{2\epsilon}{\sqrt{M}} \|\mathbf{W}_\parallel\| - 2w\delta \|\mathbf{W}_\perp\| + K_1 + 2K_2. \quad (29)$$

On the other hand, we have

$$\begin{aligned} \mathbb{E}[\Delta L_\parallel(\mathbf{W})|\mathbf{W}] &= \mathbb{E}[\langle \mathbf{c}, \mathbf{W}[t+1] \rangle^2 - \langle \mathbf{c}, \mathbf{W}[t] \rangle^2 | \mathbf{W}] \\ &= \mathbb{E}[\langle \mathbf{c}, \mathbf{W} + \boldsymbol{\nu} - \boldsymbol{\mu} \rangle^2 - \langle \mathbf{c}, \mathbf{W} \rangle^2 | \mathbf{W}] \\ &= \mathbb{E}[2\langle \mathbf{c}, \mathbf{W} \rangle \langle \mathbf{c}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle + \langle \mathbf{c}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle^2 | \mathbf{W}] \\ &\geq 2\langle \mathbf{c}, \mathbf{W} \rangle \langle \mathbf{c}, \mathbb{E}[\boldsymbol{\nu} - \boldsymbol{\mu} | \mathbf{W}] \rangle \\ &\stackrel{(a)}{\geq} 2\|\mathbf{W}_\parallel\| \frac{1}{\sqrt{M}} \sum_{m=1}^M (\mathbb{E}[\nu_m | \mathbf{W}] - 1) \\ &= 2\|\mathbf{W}_\parallel\| \frac{1}{\sqrt{M}} (\mathbb{E}[\nu_\Sigma] - M) \stackrel{(b)}{=} -\frac{2\epsilon}{\sqrt{M}} \|\mathbf{W}_\parallel\|, \end{aligned} \quad (30)$$

where step (a) uses the fact that  $\mu_m \leq 1, \forall m = 1, 2, \dots, M$ ; (b) follows from the facts that  $\mathbb{E}[\nu_\Sigma] = \rho$  and  $\epsilon = M - \rho$ .

By substituting (29) and (30) into (24), we have

$$\begin{aligned} \mathbb{E}[\Delta V_\perp(\mathbf{W})|\mathbf{W}] &\leq \frac{1}{2\|\mathbf{W}_\perp\|} (-2w\delta \|\mathbf{W}_\perp\| + K_1 + 2K_2) \\ &= -w\delta + \frac{K_1 + 2K_2}{2\|\mathbf{W}_\perp\|}. \end{aligned} \quad (31)$$

Hence, when  $V_\perp(\mathbf{W}) = \|\mathbf{W}_\perp\|$  is sufficiently large, its expected drift is strictly negative, independent of heavy-traffic parameter  $\epsilon$ .  $\blacksquare$

## B. Upper Bound Analysis

Having established the state-space collapse result, we are ready to provide the upper bound on the mean workload under the JLW Algorithm in the heavy-traffic regime. In Proposition 3, we have shown that all moments of steady-state workloads are bounded under the JLW Algorithm. This enables us to analyze its heavy-traffic performance by using the methodology of ‘‘setting the drift of a Lyapunov function equal to zero’’ (see [6]).

We will omit the superscript ( $\epsilon$ ) associated with the workload for brevity in the rest of proof. To facilitate the proof, we introduce  $U_m \triangleq 1 - \mu_m$  and thus the evolution of the workload can be rewritten as

$$\mathbf{W}[t+1] = \mathbf{W}[t] + \boldsymbol{\nu}[t] - \mathbf{1} + \mathbf{U}[t], \quad (32)$$

where  $\mathbf{U}[t] \triangleq (U_m[t])_{m=1}^M$ . Note that  $U_m$  is different from the unused service in traditional queues. Indeed, recall that  $\mathbb{1}_{\mathcal{F}_m} \leq \mu_m \leq 1$  and thus  $0 \leq U_m \leq \mathbb{1}_{\overline{\mathcal{F}_m}}$ , where we recall that  $\mathcal{F}_m$  denotes the event that at least one flow in AP  $m$  has the maximum channel rate  $c_{\max}$  and  $\overline{\mathcal{F}_m}$  is the complement of the event  $\mathcal{F}_m$ . If all flows at AP  $m$  do not have the maximum channel rate and the served flow has the size larger than its channel rate, then the workload may not decrease by one (i.e.,  $\mu_m = 0$ ), which implies that  $U_m$  is equal to 1. However, the flow at AP  $m$  does receive the service and does not incur any unused service. This difference causes that the technique in addressing unused service in [6] does not apply and requires additional non-trivial efforts.

In order to derive an upper bound on  $\mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \right]$ , we

need the following fundamental identity [6, Lemma 8]:

$$\begin{aligned} & \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}} \rangle \langle \mathbf{c}, \mathbf{1} - \boldsymbol{\nu} \rangle \right] \\ &= \frac{1}{2} \mathbb{E} \left[ \langle \mathbf{c}, \boldsymbol{\nu} - \mathbf{1} \rangle^2 \right] + \frac{1}{2} \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle^2 \right] \\ & \quad + \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}} + \boldsymbol{\nu} - \mathbf{1} \rangle \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right], \end{aligned} \quad (33)$$

where  $\widetilde{\mathbf{U}}$  is a random vector with the same distribution as the steady-state distribution of the process  $\{\mathbf{U}[t]\}_{t \geq 0}$ . The identity (33) is derived by setting the expected drift of  $\langle \mathbf{c}, \mathbf{W} \rangle^2$  to 0 due to the existence of second moment of steady-state workload under the JLW algorithm by Proposition 3.

First, we consider the left-hand-side (LHS) of (33).

$$\begin{aligned} \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}} \rangle \langle \mathbf{c}, \mathbf{1} - \boldsymbol{\nu} \rangle \right] &= \frac{1}{\sqrt{M}} (M - \mathbb{E}[\nu_\Sigma]) \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}} \rangle \right] \\ &= \frac{\epsilon}{M} \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \right]. \end{aligned} \quad (34)$$

Next, we will provide an upper bound for each individual term on the right-hand side (RHS) of (33). By simply setting the expected drift of  $\langle \mathbf{c}, \mathbf{W} \rangle$  equal to zero, we have

$$\begin{aligned} \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right] &= \mathbb{E} \left[ \langle \mathbf{c}, \mathbf{1} \rangle - \langle \mathbf{c}, \boldsymbol{\nu} \rangle \right] = \frac{1}{\sqrt{M}} (M - \mathbb{E}[\nu_\Sigma]) \\ &= \frac{\epsilon}{\sqrt{M}}, \end{aligned} \quad (35)$$

which implies

$$\mathbb{E} \left[ \sum_{m=1}^M \widetilde{U}_m \right] = \epsilon. \quad (36)$$

For the first term on the RHS of (33), we have

$$\begin{aligned} \frac{1}{2} \mathbb{E} \left[ \langle \mathbf{c}, \boldsymbol{\nu} - \mathbf{1} \rangle^2 \right] &= \frac{1}{2M} \mathbb{E} \left[ (\nu_\Sigma - M)^2 \right] \\ &= \frac{1}{2M} \mathbb{E} \left[ (\nu_\Sigma - \rho - \epsilon)^2 \right] = \frac{1}{2M} (\text{Var}(\nu_\Sigma) + \epsilon^2). \end{aligned} \quad (37)$$

For the second term on the RHS of (33), we have

$$\frac{1}{2} \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle^2 \right] \stackrel{(a)}{\leq} \frac{1}{2} \langle \mathbf{c}, \mathbf{1} \rangle \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right] \stackrel{(b)}{=} \frac{1}{2} \epsilon, \quad (38)$$

where step (a) follows from the fact that  $\widetilde{U}_m \leq 1$ , and (b) uses (35).

For the last term on the RHS of (33), we have

$$\begin{aligned} & \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}} + \boldsymbol{\nu} - \mathbf{1} \rangle \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right] \\ & \stackrel{(a)}{=} \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}}^+ \rangle \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right] - \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle^2 \right] \\ & \leq \mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}}^+ \rangle \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right] \\ & \stackrel{(b)}{=} \mathbb{E} \left[ \langle \widetilde{\mathbf{W}}_\parallel^+, \widetilde{\mathbf{U}}_\parallel \rangle \right] = \mathbb{E} \left[ \langle \widetilde{\mathbf{W}}^+ - \widetilde{\mathbf{W}}_\perp^+, \widetilde{\mathbf{U}} - \widetilde{\mathbf{U}}_\perp \rangle \right] \\ & = \mathbb{E} \left[ \langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}} \rangle + \langle \widetilde{\mathbf{W}}_\perp^+, \widetilde{\mathbf{U}}_\perp \rangle - \langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}}_\perp \rangle - \langle \widetilde{\mathbf{W}}_\perp^+, \widetilde{\mathbf{U}} \rangle \right] \\ & \stackrel{(c)}{=} \mathbb{E} \left[ \langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}} \rangle \right] + \mathbb{E} \left[ \langle -\widetilde{\mathbf{W}}_\perp^+, \widetilde{\mathbf{U}} \rangle \right], \end{aligned} \quad (39)$$

where step (a) is true for vector  $\mathbf{I}^+$  denoting  $\mathbf{I}[t+1]$  and

follows the evolution of the workload  $\mathbf{W}[t]$  (cf. (32)); (b) follows from the fact that  $\widetilde{\mathbf{W}}_\parallel^+$  and  $\widetilde{\mathbf{U}}_\parallel$  are along the same direction  $\mathbf{c}$ ; (c) uses the fact that  $\langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}}_\perp \rangle = \langle \widetilde{\mathbf{W}}_\perp^+, \widetilde{\mathbf{U}}_\perp \rangle + \langle \widetilde{\mathbf{W}}_\parallel^+, \widetilde{\mathbf{U}}_\perp \rangle = \langle \widetilde{\mathbf{W}}_\perp^+, \widetilde{\mathbf{U}}_\perp \rangle$ .

Next, we consider terms in the RHS of (39). For the term  $\mathbb{E} \left[ \langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}} \rangle \right]$ , we have

$$\begin{aligned} & \mathbb{E} \left[ \langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}} \rangle \right] = \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m[t+1] \widetilde{U}_m[t] \right] \\ & \leq \mathbb{E} \left[ \sum_{m=1}^M (\widetilde{W}_m[t] + \nu_m[t]) \widetilde{U}_m[t] \right] \\ & \stackrel{(a)}{\leq} \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \widetilde{U}_m \right] + \sqrt{\mathbb{E} \left[ \sum_{m=1}^M \nu_m^2 \right] \mathbb{E} \left[ \sum_{m=1}^M \widetilde{U}_m^2 \right]} \\ & \stackrel{(b)}{\leq} \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \widetilde{U}_m \right] + \sqrt{\mathbb{E} \left[ \left( \sum_{m=1}^M \nu_m \right)^2 \right] \mathbb{E} \left[ \sum_{m=1}^M \widetilde{U}_m \right]} \\ & \stackrel{(c)}{=} \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \widetilde{U}_m \right] + \sqrt{\epsilon \mathbb{E}[\nu_\Sigma^2]}, \end{aligned} \quad (40)$$

where step (a) follows from Cauchy–Schwarz inequality; (b) uses the fact that  $\widetilde{U}_m \leq 1$ ; (c) uses (36).

For the term  $\mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \widetilde{U}_m \right]$  in (40), we have

$$\begin{aligned} & \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \widetilde{U}_m \right] \stackrel{(a)}{\leq} \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m (1 - p_{m,K})^{\widetilde{N}_m} \right] \\ & = \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m (1 - p_{m,K})^{\frac{\widetilde{N}_m}{2}} (1 - p_{m,K})^{\frac{\widetilde{N}_m}{2}} \right] \\ & \stackrel{(b)}{\leq} \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m (1 - p_K^{\min})^{\frac{\widetilde{W}_m}{2w_{\max}}} (1 - p_{m,K})^{\frac{\widetilde{N}_m}{2}} \right] \\ & \stackrel{(c)}{\leq} (\hat{w}_{\max} + 1) \mathbb{E} \left[ \sum_{m=1}^M (1 - p_{m,K})^{\frac{\widetilde{N}_m}{2}} \right] \\ & \stackrel{(d)}{\leq} (\hat{w}_{\max} + 1) \left( \mathbb{E} \left[ \sum_{m=1}^M (1 - p_{m,K})^{d\widetilde{N}_m} \right] \right)^{\frac{1}{2d}} M^{\frac{2d-1}{2d}} \\ & \stackrel{(e)}{\leq} (\hat{w}_{\max} + 1) \left( \mathbb{E} \left[ \sum_{m=1}^M (p_{m,0})^{\widetilde{N}_m} \right] \right)^{\frac{1}{2d}} M^{\frac{2d-1}{2d}} \\ & \stackrel{(f)}{\leq} (\hat{w}_{\max} + 1) \left( \mathbb{E} \left[ \sum_{m=1}^M \widetilde{U}_m \right] \right)^{\frac{1}{2d}} M^{\frac{2d-1}{2d}} \\ & \stackrel{(g)}{=} (\hat{w}_{\max} + 1) \epsilon^{\frac{1}{2d}} M^{\frac{2d-1}{2d}}, \end{aligned} \quad (41)$$

where step (a) uses the fact that  $U_m = 1 - \mu_m \leq \mathbb{1}_{\mathcal{F}_m}$ ; (b) is true for  $p_K^{\min} \triangleq \min_m p_{m,K} > 0$  and  $w_{\max} = \lceil F_{\max}/c_{\max} \rceil$  and the fact that the number of flows at AP  $m$  is at least  $\widetilde{W}_m/w_{\max}$  (i.e.,  $\widetilde{N}_m \geq \widetilde{W}_m/w_{\max}$ ); (c) is true for  $\hat{w}_{\max} \triangleq \max_{m=1,2,\dots,M} \hat{w}_m$  and  $\hat{w}_m$  is some positive constant such that  $\widetilde{W}_m (1 - p_K^{\min})^{\widetilde{W}_m/(2w_{\max})} \leq 1$  whenever  $\widetilde{W}_m > \hat{w}_m$

(since  $\lim_{\widetilde{W}_m \rightarrow \infty} \widetilde{W}_m(1 - p_K^{\min})\widetilde{W}_m/(2w_{\max}) = 0$ ), and is derived using similar steps in (18); (d) is true for some  $d > 1$  such that  $(1 - p_{m,K})^d \leq p_{m,0}$  (which is the possible due to the assumptions that  $p_{m,0} > 0$  and  $p_{m,K} > 0$ ) and follows from Hölder's inequality; (e) uses the definition of the constant  $d$ ; (f) uses the fact that  $U_m \geq \mathbb{1}_{\mathcal{G}_m}$  and  $\mathcal{G}_m$  denotes the event that all flows at AP  $m$  do not have available channels, i.e.,  $\mu_m = 0$ ; (g) uses (36).

By substituting (41) into (40), we have

$$\mathbb{E} \left[ \langle \widetilde{\mathbf{W}}^+, \widetilde{\mathbf{U}} \rangle \right] \leq (\hat{w}_{\max} + 1) \epsilon^{\frac{1}{2d}} M^{\frac{2d-1}{2d}} + \sqrt{\epsilon \mathbb{E}[\nu_{\Sigma}^2]}. \quad (42)$$

For the term  $\mathbb{E} \left[ \langle -\widetilde{\mathbf{W}}_{\perp}^+, \widetilde{\mathbf{U}} \rangle \right]$  in (39), we have

$$\begin{aligned} \mathbb{E} \left[ \langle -\widetilde{\mathbf{W}}_{\perp}^+, \widetilde{\mathbf{U}} \rangle \right] &\stackrel{(a)}{\leq} \sqrt{\mathbb{E} \left[ \|\widetilde{\mathbf{W}}_{\perp}^+\|^2 \right] \mathbb{E} \left[ \|\widetilde{\mathbf{U}}\|^2 \right]} \\ &\stackrel{(b)}{\leq} \sqrt{\mathbb{E} \left[ \|\widetilde{\mathbf{W}}_{\perp}^+\|^2 \right] \mathbb{E} \left[ \sum_{m=1}^M \widetilde{U}_m \right]} \\ &\stackrel{(c)}{\leq} \sqrt{H_2 \epsilon}, \end{aligned} \quad (43)$$

where step (a) uses Cauchy–Schwarz inequality; (b) uses the fact that  $U_m \leq 1$ ; (c) uses the state-space collapse result (cf. Proposition 6) and (36).

By substituting (42) and (43) into (39), we have

$$\begin{aligned} &\mathbb{E} \left[ \langle \mathbf{c}, \widetilde{\mathbf{W}} + \boldsymbol{\nu} - \mathbf{1} \rangle \langle \mathbf{c}, \widetilde{\mathbf{U}} \rangle \right] \\ &\leq (\hat{w}_{\max} + 1) \epsilon^{\frac{1}{2d}} M^{\frac{2d-1}{2d}} + \sqrt{\epsilon \mathbb{E}[\nu_{\Sigma}^2]} + \sqrt{H_2 \epsilon} \triangleq G(\epsilon). \end{aligned} \quad (44)$$

By substituting (34), (37), (38), and (44) into (33), we have

$$\epsilon \mathbb{E} \left[ \sum_{m=1}^M \widetilde{W}_m \right] \leq \frac{1}{2} (\text{Var}(\nu_{\Sigma}) + \epsilon^2) + \frac{1}{2} M \epsilon + M G(\epsilon),$$

which implies the desired result as  $\epsilon \downarrow 0$ .

## VIII. CONCLUSIONS

In this paper, we studied the optimal load-balancing design in high-density wireless networks with both channel fading and flow-level dynamics. We discussed the performance deficiencies of existing policies and developed a workload-aware load-balancing scheme in the presence of dynamic flows. We showed that our proposed load-balancing algorithm not only achieves maximum system throughput, but also minimizes the mean total workload in heavy-traffic regimes. In addition, our analysis implies that the mean total workload performance under our proposed algorithm is robust to the number of APs, which is strongly desirable in high-density wireless networks. Finally, extensive simulations were performed to confirm our theoretical results.

## APPENDIX A

### CHARACTERIZATION OF CAPACITY REGION

(1) (Necessity) Assume that  $\rho > M$  is true. Consider the Lyapunov function  $J(\mathbf{W}) \triangleq \sum_{m=1}^M W_m$ . Then, we have

$$\begin{aligned} &\mathbb{E} [J(\mathbf{W}[t+1]) - J(\mathbf{W}[t]) | \mathbf{W}[t] = \mathbf{W}] \\ &\stackrel{(a)}{=} \sum_{m=1}^M \mathbb{E} [\nu_m[t] - \mu_m[t] | \mathbf{W}[t] = \mathbf{W}] \\ &= \mathbb{E} [\nu_{\Sigma}[t]] - \sum_{m=1}^M \mathbb{E} [\mu_m[t] | \mathbf{W}[t] = \mathbf{W}] \\ &\stackrel{(b)}{\geq} \rho - M \stackrel{(c)}{>} 0, \end{aligned} \quad (45)$$

where step (a) uses the dynamic of workload (cf. (2)); (b) uses  $\nu_{\Sigma}[t] = \sum_{m=1}^M \nu_m[t]$ ,  $\mathbb{E} [\nu_{\Sigma}[t]] = \rho$ , and the fact that  $\mu_m[t] \leq 1$ ; (c) uses our contradictory assumption.

Thus, by [19, Theorem 3.3.10], no policy can stabilize the system.

(2) (Sufficiency) Proposition 3 in Section IV shows that any arrival traffic intensity  $\rho$  strictly inside  $\Lambda$  (i.e.,  $\rho < M$ ) can be supported by the policy proposed in Section IV. This together with the necessity proof establishes the desired result.

## REFERENCES

- [1] G. Athanasiou, P. Weeraddana, C. Fischione, and L. Tassiulas. Optimizing client association for load balancing and fairness in millimeter-wave wireless networks. *IEEE/ACM Transactions on Networking*, 23(3):836–850, 2015.
- [2] A. Balachandran, G. Voelker, P. Bahl, and P. Rangan. Characterizing user behavior and network performance in a public wireless lan. In *ACM SIGMETRICS Performance Evaluation Review*, volume 30, pages 195–205. ACM, 2002.
- [3] Y. Bejerano, S. Han, and L. Li. Fairness and load balancing in wireless lans using association control. In *Proceedings of the 10th annual international conference on Mobile computing and networking*, pages 315–329. ACM, 2004.
- [4] F. Bonomi. On job assignment for a parallel system of processor sharing queues. *IEEE Transactions on Computers*, 39(7):858–869, 1990.
- [5] M. Dwijaksara, W. Jeon, and D. Jeong. A joint user association and load balancing scheme for wireless lans supporting multicast transmission. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 688–695. ACM, 2016.
- [6] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72:311–359, 2012.
- [7] G. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26(3):320–327, 1978.
- [8] H. Gong and J. Kim. Dynamic load balancing through association control of mobile users in wifi networks. *IEEE Transactions on Consumer Electronics*, 54(2), 2008.
- [9] V. Gupta, M. Harchol-Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007.
- [10] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14:502–525, 1982.
- [11] W. Henderson and P. Taylor. Insensitivity in discrete time queues with a moving server. *Queueing systems*, 11(3):273–297, 1992.
- [12] D. Kotz and K. Essien. Analysis of a campus-wide wireless network. *Wireless Networks*, 11(1-2):115–133, 2005.
- [13] W. Li, S. Wang, Y. Cui, X. Cheng, R. Xin, M. Al-Rodhaan, and A. Al-Dhelaan. Ap association for proportional fairness in multirate wlans. *IEEE/ACM Transactions on Networking (TON)*, 22(1):191–202, 2014.

- [14] S. Liu, L. Ying, and R. Srikant. Scheduling in multichannel wireless networks with flow-level dynamics. In *Proc. ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, New York, June 2010.
- [15] S. Liu, L. Ying, and R. Srikant. Throughput-optimal opportunistic scheduling in the presence of flow-level dynamics. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Sam Diego, CA, March 2010.
- [16] M. Mitzenmacher. *The power of two choices in randomized load balancing*. Ph.D. Thesis, University of California at Berkeley, 1996.
- [17] K. Ousterhout, P. Wendell, M. Zaharia, and I. Stoica. Sparrow: distributed, low latency scheduling. In *Proc. ACM Symposium on Operating Systems Principles (SOSP)*, Pennsylvania, PA, USA, November 2013.
- [18] B. Sadiq and G. de Veciana. Throughput optimality of delay-driven maxweight scheduler for a wireless system with flow dynamics. In *Proc. Allerton Conference on Communications, Control and Computing (Allerton)*, Monticello, IL, October 2009.
- [19] R. Srikant and L. Ying. *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.
- [20] L. Sun, L. Wang, Z. Qin, Z. Ma, and Z. Yuan. A novel on-line association algorithm in multiple-ap wireless lan. In *International Conference on Wireless Algorithms, Systems, and Applications*, pages 890–902. Springer, 2017.
- [21] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 36:1936–1948, December 1992.
- [22] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, 39:466–478, March 1993.
- [23] P. van de Ven, S. Borst, and S. Shneer. Instability of maxweight scheduling algorithms. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Rio de Janeiro, Brazil, April 2009.
- [24] P. van de Ven, S. Borst, and L. Ying. Inefficiency of maxweight scheduling in spatial wireless networks. *Computer Communications*, 36(12):1350–1359, 2013.
- [25] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.
- [26] W. Whitt. Deciding which queue to join: some counterexamples. *Operations Research*, 34:55–62, 1986.
- [27] L. Ying, R. Srikant, and X. Kang. The power of slightly more than one sample in randomized load balancing. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Hong Kong, April 2015.