Optimal Scheduling for Unmanned Aerial Vehicle Networks with Flow-Level Dynamics

Xiangqi Kong, *Student Member, IEEE,* Ning Lu, *Member, IEEE,* Bin Li, *Member, IEEE*

Abstract—Unmanned Aerial Vehicle (UAV) Networks have recently attracted great attention as being able to provide convenient and fast wireless connections. One central question is how to allocate a limited number of UAVs to provide wireless services across a large number of regions, where each region has dynamic arriving flows and flows depart from the system once they receive the desired amount of service (referred to as the flow-level dynamic model). In this paper, we propose a MaxWeight-type scheduling algorithm taking into account sharp flow-level dynamics that efficiently redirect UAVs across a large number of regions. However, in our considered model, each flow experiences an independent fading channel and will immediately leave the system once it completes its service, which makes its evolution quite different from the traditional queueing model for wireless networks. This poses significant challenges in our performance analysis. Nevertheless, we incorporate sharp flow-dynamic into the Lyapunov-drift analysis framework, and successfully establish both throughput and heavy-traffic optimality of the proposed algorithm. Extensive simulations are performed to validate the effectiveness of our proposed algorithm.

Index Terms—Wireless UAV networks, Flow-level dynamics, Scheduling design, Throughput, Mean delay, Heavy-traffic analysis.

1 Introduction

Recent advances in Unmanned Aerial Vehicles (UAVs) technologies have demonstrated enormous potential for the development of airborne communication networks in both military and civilian domains. UAVs can be deployed to quickly form a mobile network to provide wireless data connections for users on the ground in situations such as traffic monitoring, remote sensing, and disaster recovery [34]. Especially, a flying UAV equipped with an access point (AP) can provide "connectivity from the sky" [6] for users that require the access to the backbone/core network, where UAVs are connected to the core network by means of such as point-to-point satellite relaying [14] or point-to-multipoint microwave/mmWave backhauling [4]. This has emerged as a promising solution to agile cellular/Internet services provisioning in areas of high/urgent network demand without having to pre-install any wireless access infrastructure [11].

The main advantage of UAV networks is that their deployment can be *agile* and *re-configurable* due to the flexible mobility of UAVs. Recently, the deployment of UAVs (serving as APs or small cell base stations) has attracted many research attention to address challenges such as 3-D deployment [5], spectral efficiency improvement [11], coverage optimization [22], service time maximization [7], optimal placement considering energy efficienty [20], and offloading cellular networks [4], [12], [31]. Besides, UAV trajectory design considering various communication and networking constraints is extensively studied [10], [30], [32], and the resource allocation such as channel assignment and power control is also considered [24], [33]. In addition, existing research has demonstrated the feasibility of UAV-based small cells [21]. To prolong the service duration or even provide persistent service during the mission, advanced charging technologies have been proposed. The state-of-the art solutions include equipping UAVs with solar panels [2], charging the battery during night by using a high energy laser beam [23], and charging the battery by more powerful fixed-wing UAVs via wireless power transfer [1]. However, none of the existing works focuses on the design of optimal scheduling/repositioning algorithms for UAVs in response to the dynamics of network traffic, where users dynamically arrive and depart from the network once they receive the desired amount of service (referred to as *flow-level dynamic model*). To this end, in this work, we concentrate on an important problem of efficiently allocating a limited number of UAVs to serve dynamic network users across a certain geographic area.

Scheduling wireless traffic in the presence of flow-level dynamics has received great attention in recent years, since it is more accurate to characterize the dynamics of real wireless traffic than the traditional wireless networks (e.g., [25], [26], [27]) consisting of a fixed number of persistent user that continuously inject packets into the network and would never leave. In [28], the authors pointed out that the traditional queue-length-based MaxWeight scheduling for persistent users is no longer throughput-optimal when dealing with dynamic flows over time-varying channels. Even in the absence of time-varying channels, the MaxWeight scheduling still fails to achieve the maximum throughput under certain network settings [29]. Subsequent works (e.g., [3], [15], [16], [17], [18], [19]) have developed throughput-optimal scheduling algorithms for dynamic flows in various scenarios. The closest one to our work is the flow-aware CSMA Algorithm developed in [3] that only addressed the throughput performance in non-fading scenarios, which does not meet the desired performance of UAV networks in

- *Xiangqi Kong (xqkong@uri.edu) and Bin Li (binli@uri.edu) are with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI, 02881 USA.*
- *Ning Lu is with the Department of Electrical and Computer Engineering, Queen's University Walter Light Hall, Room 401, 19 Union St W, Kingston, ON K7L 3N9 (e-mail address: ning.lu@queensu.ca).*

the presence of dynamic flows. Note that the reason why the CSMA Algorithm works in the presence of flow-level dynamics without wireless fading lies in that each flow maintains an exponential clock and thus the chance of a region served by a UAV is proportional to the number of flows in the region, which mimics the MaxWeight Algorithm [26]. Here, we use the fact that the minimum of n independently and identically distributed exponential random variables with rate 1 is a exponential random variable with rate n. However, in the presence of general channel fading, this fact does not take the channel rate into account and thus it is unclear that how to generalize the CSMA Algorithm to the case with wireless fading.

In this paper, we consider the problem of scheduling a limited number of UAVs, each equipped with an AP, to serve data traffic flows in a specific geographic area partitioned into a number of small regions, each of which can be fully covered by a UAV. In particular, we consider a time-slotted system and adopt a flow-level dynamic model: each flow dynamically arrives at a particular region, experiences an independent fading channel, and immediately leaves the system once it completes its service. The assignment of UAVs to the regions is made every T consecutive time slots guided by the scheduling algorithm in response to the flow-level dynamics. The objective of this paper is to develop an optimal scheduling algorithm to achieve the optimal system performance, i.e., the maximum system throughput (supporting flows as many as possible) and the minimum system workload (reducing flow latency as much as possible).

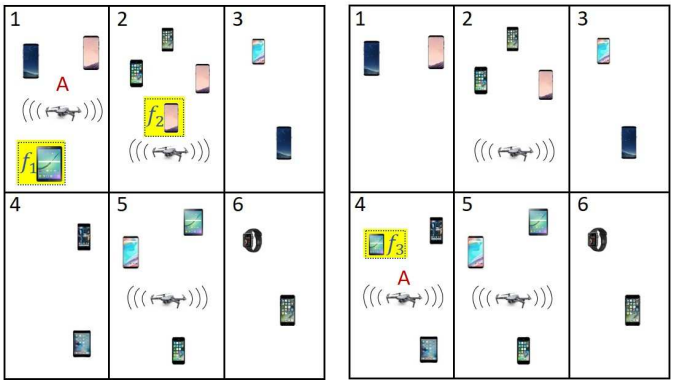
The main contributions of this paper are listed as follows:

- We formulate a problem of wireless scheduling design for UAVs in the presence of dynamic flow-level traffic.
- We propose a MaxWeight-type scheduling algorithm for deploying UAVs to serve dynamic flows, which not only achieves maximum throughput but also minimizes total mean system workload in the heavy-traffic regime.
- We conduct extensive simulations that not only confirm our analytical results but also demonstrate the excellent performance of our proposed algorithm in general setups.

The reminder of this paper is organized as follows. In Section 2, we formulate the problem and state our model. In Section 3, we present a MaxWeight-type scheduling algorithm to find the optimal scheduling decision. In Section 4, we show extensive simulation results to confirm our analytical observations. In Section 5, we show the proof of two equivalent capacity regions. In Section 6 and 7, we provide detailed proofs of our algorithm. In Section 8, we conclude our paper and discuss future work.

2 System Model

We consider a UAV network with M UAV-based wireless Access Points (APs) serving dynamic flows across N ($N > M$) different regions, where flows dynamically arrive at each region and depart once they receive the desired amount of network service. Here, a flow models an active network service session. It can be a service requested by a new user or an existing user that keeps silent for a while. We assume that the system operates in a time-slotted manner. Fig. 1 shows two snapshots of a system with six regions and three UAVs at two different time instances.



(a) Snapshot at time t_1 (b) Snapshot at time t_2 ($t_2 > t_1$)

Fig. 1: A system with six regions and three UAVs at two different time instants: flow f_1 in region 1 and f_2 in region 2 receive services at time t_1, and complete their services before time t_2, while flow f_3 joins region 4. In the meanwhile, UAV A moves from Region 1 to Region 4 by the time instant t_2.

Note that it could be energy-consuming to frequently redirect UAVs across regions on a time slot basis and therefore we assume that UAV redirection decisions are made every T time slots. To this end, we group every T consecutive time slots into a frame, as shown in Fig 2. Note that the UAV hovering usually consumes much less energy when it flies cross the regions. Therefore, by allowing a UAV to fly across regions every T time slots, the larger the T, the less energy the UAV consumes. Let $S_n[k] = 1$ if a UAV is hovering over region n in time frame k, and $S_n[k] = 0$ otherwise. We assume that the traveling time of a UAV to each region is negligible compared to the length of a time frame. We assume that flows randomly arrive at the beginning of each time slot. In particular, we let $\mathcal{A}_n[k;t]$ be the set of flows arriving at region n in t^{th} time slot of the frame k. We use $A_n[k;t]$ to denote the cardinality of the set $\mathcal{A}_n[k;t]$, i.e., $A_n[k;t] = |\mathcal{A}_n[k;t]|$. We assume that $\{A_n[k;t], t = 1, 2, \ldots, T, k \geq 0\}$ are independently and identically distributed (i.i.d.) over time with mean $\lambda_n > 0$, and $A_n[k;t] \leq A_n^{\max}, \forall n, t, k$, for some $A_n^{\max} > 0$. We use $F_{n,f}[k;t]$ to denote the number of packets of a newly arriving flow f (also referred to as the file size of flow f) in region n in the t^{th} time slot of frame k, which is i.i.d. over time with mean $\eta_n > 0$ and $F_{n,f}[k;t] \leq F_n^{\max}, \forall n, t, k$, for some $F_n^{\max} > 0$. In addition, the number of flows and their file sizes are independently distributed across regions.

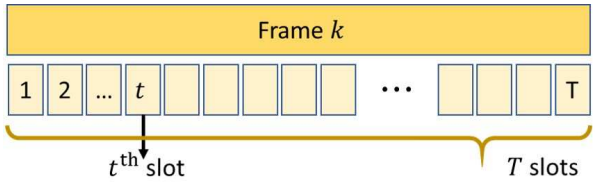


Fig. 2: Relationship between each time slot and one frame

Flows in region n can receive network service only when there is a UAV hovering over that region. Due to wireless

interference, without loss of generality, we assume that at most one flow can be served in each region in each time slot. We also assume that each region at most has one UAV in each frame, and all M UAVs in different regions do not interfere with each other and thus can serve their flows simultaneously. We call a set of regions that can be served by UAVs simultaneously in frame k a *feasible schedule* denoted by $\mathbf{S}[k] \triangleq (S_n[k])_{n=1}^N$, where exact M of its components equal to 1. Let $\mathcal{S}$ be the collection of all feasible schedules. All UAVs are assumed directly connected to a central controller via point-to-multipoint micro/millimeter-wave backhauling with sufficient bandwidth [4].

Due to a limited number of modulation and coding schemes, each flow has a finite number of transmission rates. We use $C_{n,f}[k;t]$ to denote the channel rate (in unit of packets per time slot) of flow f in region n in the t^{th} time slot of frame k, which is i.i.d. with the maximum and minimum channel rates of $c_n^{\max} > 0$ and 0, respectively. Here, we also assume that both probability of each flow having the maximum and minimum channel rates are strictly positive, i.e., $p_n^{\max} \triangleq \Pr\{C_{n,f}[k;t] = c_n^{\max}\} > 0$ and $p_n^{\min} \triangleq \Pr\{C_{n,f}[k;t] = 0\} > 0$. Note that the traffic characteristics being considered depend on the region where the traffic is generated, which we believe is a proper model to characterize real-world location-specific applications. Let $R_{n,f}[k;t]$ denotes the number of residual packets of flow f in region n in the t^{th} time slot of frame k.

TABLE 1: Notations for System Model

Symbol	Description
M	number of UAVs
N	number of Regions
k	time frame index
t	time slot index within a frame, $t = 1, 2, \ldots, T$
T	number of slots in each frame
$\mathcal{A}_n[k;t]$	set of flows arrived at region n in the t^{th} slot of frame k
$A_n[k;t]$	cardinality of $\mathcal{A}_n[k;t]$
λ_n	mean of $A_n[k;t]$, $\lambda_n > 0$
$A_{\max}$	the maximum possible value for $A_n[k;t]$
$F_{n,f}[k;t]$	file size of flow f in region n at t^{th} slot of frame k
$F_n^{\max}$	the maximum possible value for $F_{n,f}[k;t]$
$S_n[k]$	service decision for region n at frame k, $S_n[k] \in \{0, 1\}$
$\mathbf{S}[k]$	feasible schedule in frame k, $\mathbf{S}[k] \triangleq (S_n[k])_{n=1}^N$
$\mathcal{S}$	set of all feasible schedules
$C_{n,f}[k;t]$	channel rate of flow f at region n in the t^{th} slot of frame k
$c_n^{\max}$	the maximum possible value for $C_{n,f}[k;t]$
$p_n^{\max}$	probability that $C_{n,f}[k;t]$ achieves the maximum rate
$p_n^{\min}$	probability that $C_{n,f}[k;t]$ has value 0
$R_{n,f}[k;t]$	residual packets flow f at region n in the t^{th} slot of frame k
$\nu_n[k;t]$	workload arrived at region n at the t^{th} slot of frame k
$\nu_n[k]$	total workload arrived at region n in frame k
ρ_n	mean workload of arriving flow in region n at each slot
$\boldsymbol{\rho}$	mean workload vector, $\boldsymbol{\rho} \triangleq (\rho_n)_{n=1}^N$, or traffic intensity vector
$W_n[k;t]$	total workload at region n at the t^{th} slot of frame k
$W_n[k]$	total workload in region n at the beginning of frame k
$\mathbf{W}[k]$	workload vector, $W_n[k] \triangleq (W_n[k])_{n=1}^N$
$\mathcal{N}_n[k;t]$	set of flows in region n at the t^{th} slot of frame k
$\mu_n[k;t]$	workload decreased in region n at the t^{th} slot of frame k
$\mu_n[k]$	total workload decreased in region n in frame k
Λ	capacity region

To characterize the underlying dynamic of flows, we use *workload* to measure the minimum time slots needed to serve the newly arriving (or existing) flows. In particular, let $\nu_n[k;t] \triangleq \sum_{f\in\mathcal{A}_n[k;t]} \lceil F_{n,f}[k;t]/c_n^{\max} \rceil$ and $W_n[k;t] \triangleq \sum_{f\in\mathcal{N}_n[k;t]} \lceil R_{n,f}[k;t]/c_n^{\max} \rceil$ denote the newly arriving workload and the total workload in region n in the t^{th} time slot of frame k, respectively, where $\mathcal{N}_n[k;t]$ denotes the set of flows in region n in the t^{th} time slot of frame k. We use $\mu_n[k;t]$ to denote the amount of workload decreasing in region n in the t^{th} time slot of time frame k. Here, $\mu_n[k;t]$ depends on whether there is a flow in region n that receives service from a UAV and its residual file size and associated channel rate if there is. Note that in each time slot, there is at most one UAV hovering over one region and each UAV can at most serve one flow. Thus, $\mu_n[k;t]$ is equal to either 0 or 1. *We use $W_n[k] \triangleq W_n[k;1]$ to denote the total workload in region n at the beginning of frame k*. Let $\nu_n[k] \triangleq \sum_{t=1}^T \nu_n[k;t]$ and $\mu_n[k] \triangleq \sum_{t=1}^T \mu_n[k;t]$ be the total workload of newly arriving flows and the total amount of workload decreasing in frame k, respectively. Therefore, the evolution of workload can be described as follows:

$$W_n[k+1] = W_n[k] + \nu_n[k] - S_n[k]\mu_n[k], \tag{1}$$

holds for $n = 1, 2, \ldots, N$.

We call region n *stable* if its average workload is finite, i.e.,

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}[W_n[k]] < \infty. \tag{2}$$

We call the *system stable* if all its regions are stable. The capacity region Λ is the maximum set of mean arriving workload under which the system can be stabilized by some policy, and can be represented as

$$\Lambda \triangleq \left\{ \boldsymbol{\rho} = (\rho_n)_{n=1}^N : \sum_{n=1}^N \rho_n \le M \text{ and } 0 \le \rho_n \le 1, \forall n \right\},$$

where $\rho_n \triangleq \mathbb{E}[\nu_n[k;t]]$ is the mean workload of all newly arriving flows in region n in each time slot (also referred to as traffic intensity). In each time slot, the workload can at most reduce by 1 in each region and thus the workload of newly arriving flows in each region cannot be greater than 1, i.e., $\rho_n \le 1, \forall n = 1, 2, \ldots, N$. Besides, there are at most M UAVs in the system and therefore total system workload can at most decrease by M in each time slot and hence we have $\sum_{n=1}^N \rho_n \le M$ to maintain the system stability. We say an algorithm is *throughput-optimal* if it stabilizes the system for any traffic intensity vector strictly within the capacity region Λ. Notations for system model are listed in Table 1.

In this paper, we are interested in developing a scheduling algorithm that efficiently allocates M UAVs to serve N regions in each time frame and each UAV in each region needs to decide how to serve its flows in each time slot with the following two goals: (i) maximizing the system throughput (i.e., supporting flows as many as possible); (ii) minimizing the system workload (i.e., reducing the flow latency as much as possible).

3 Optimal Scheduling Design

In this section, we first develop a workload-aware scheduling algorithm. Then, we prove that the proposed algorithm

not only achieves throughput optimality but also minimizes the mean total workload in the heavily-loaded conditions, where the latency is most pronounced.

Motivated by the well-known MaxWeight algorithm (see [26]) that prioritizes the users with high congestion levels, the efficient algorithm should send UAVs to serve the most heavily-loaded regions, which is described as follows.

UAV-Optimal-Scheduling (UOS) Algorithm: In each time frame k, given the current workload vector $\mathbf{W}[k] \triangleq (W_n[k])_{n=1}^N$,

(1) Send M UAVs to serve the top M congested regions, i.e. select $\mathbf{S}^*[k] \triangleq (S_n^*[k])_{n=1}^N$ such that

$$\mathbf{S}^*[k] \in \arg\max_{\mathbf{S}\in\mathcal{S}} \langle \mathbf{W}[k], \mathbf{S}\rangle, \tag{3}$$

breaking ties uniformly at random.

(2) In t^{th} time slot of time frame k, each UAV with $S_n^*[k] = 1$ serves a flow with the maximum channel rate, breaking ties uniformly at random.

In the UOS Algorithm, the central controller collects workload information from all regions in each time frame, and sends M UAVs to serve the M most congested regions. The proposed algorithm differs from the traditional queue-length-based MaxWeight algorithm in that the traditional queue-length-based MaxWeight algorithm is designed for the system with First-Come-First-Served queueing discipline, while each flow in our considered scenario suffers from an independent channel fading and thus has a time-varying service rate. Although the authors in [3] proposed a similar algorithm dealing with the flow-level dynamic model, they only established throughput optimality of the proposed algorithm in non-fading setups. In this paper, we show that our proposed algorithm can achieve desired performance in the general setups. Next, we first show that the proposed algorithm can achieve maximum system throughput as well as the boundedness of all steady-state workloads. Note that the proposed UOS Algorithm has a computational complexity of at most $O(\binom{N}{M})$, since the algorithm needs to traverse at most $\binom{N}{M}$ schedules to find the maximum weight.

Proposition 1. *The UOS Algorithm is throughput-optimal, i.e. it stabilizes the system for any arrival traffic intensity vector strictly within the capacity region Λ. Moreover, all moments of steady-state workload are strictly bounded.*

Proof: The proof is available in Section 6. □

Having established the throughput optimality and the moment existence of the steady-state workload of the UOS Algorithm, we can proceed to analyze the mean workload performance in the heavy-traffic regime.

We consider the workload process $\{\mathbf{W}^{(\epsilon)}[k]\}_{k\geq 0}$ with arrival process $\boldsymbol{\nu}^{(\epsilon)}[k] = (\nu_n^{(\epsilon)}[k])_{n=1}^N$ parameterized with $\epsilon > 0$. Here, ϵ is called the *heavy-traffic parameter* that measures the Euclidean distance of traffic intensity vector $\boldsymbol{\rho}^{(\epsilon)}$ to the hyperplane $\mathcal{H}^{(\mathbf{d})} \triangleq \{\boldsymbol{\rho} : \langle \boldsymbol{\rho}, \mathbf{d}\rangle = b_{\mathbf{d}}\}$ with normal vector $\mathbf{d}$ and $b_{\mathbf{d}} > 0$, where $\mathcal{H}^{(\mathbf{d})}$ lies on the boundary of capacity region Λ. We illustrate all these notations in a three-dimensional capacity region in the case of two UAVs serving three regions, as shown in Fig. 3. In this work, we analyze

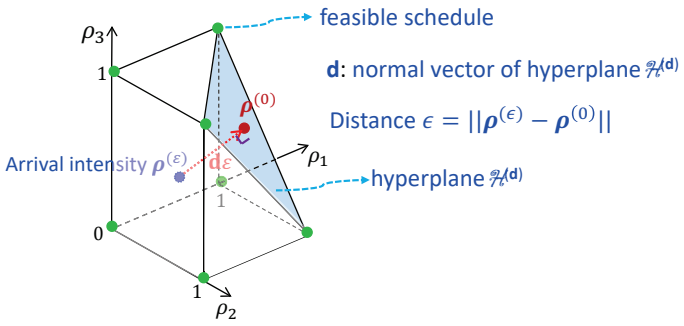


Fig. 3: Geometric structure of the capacity region

the heavy-traffic performance of our UOS Algorithm as $\epsilon \downarrow 0$, i.e., the traffic intensity vector $\boldsymbol{\rho}^{(\epsilon)}$ approaches $\boldsymbol{\rho}^{(0)}$ (i.e., $\boldsymbol{\rho}^{(0)} = \boldsymbol{\rho}^{(\epsilon)} + \epsilon\mathbf{d}$) strictly lying inside hyperplane $\mathcal{H}^{(\mathbf{d})}$. Let $(\sigma_n^{(\epsilon)})^2$ denote the variance of the arrival process $\{\nu_n^{(\epsilon)}[k;t], t = 1, 2, \ldots, T, k \geq 0\}$. Let $\boldsymbol{\sigma}^{(\epsilon)} \triangleq (\sigma_n^{(\epsilon)})_{n=1}^N$. Notations for heavy-traffic analysis are listed in Table 2.

Proposition 2. *Let $\widetilde{\mathbf{W}}^{(\epsilon)} \triangleq (\widetilde{W}_n^{(\epsilon)})_{n=1}^N$ be a random vector with the same distribution as the steady-state distribution of the workload evolution under the UOS Algorithm. Consider the heavy-traffic limit that $\epsilon \downarrow 0$ and then suppose the variance vector $\left(\boldsymbol{\sigma}^{(\epsilon)}\right)^2$ converges to a constant vector $\boldsymbol{\sigma}^2$. Then,*

$$\lim_{\epsilon\downarrow 0} \epsilon \mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{W}}^{(\epsilon)}\rangle] \leq \frac{\langle \mathbf{d}^2, \boldsymbol{\sigma}^2\rangle}{2}. \tag{4}$$

Proof: The proof is available in Section 7. □

Notice that upper bound given in Proposition 2 is also the lower bound of the system under any feasible scheduling policy. In particular, we can show the bound is tight by constructing a single server queue $\{\Phi[k]\}_{k\geq 0}$ with total arrival workload $\nu_\Sigma[k] \triangleq \langle \mathbf{d}, \boldsymbol{\nu}^{(\epsilon)}[k]\rangle$ to approach the arrival rate on $\mathcal{H}^{(\mathbf{d})}$. The single server queue evolution is as follows:

$$\Phi[k+1] = \max\left\{\Phi[k] + \nu_\Sigma[k] - Tb_{\mathbf{d}}, 0\right\}. \tag{5}$$

Then, it is easy to show that $\langle \mathbf{d}, \mathbf{W}[k]\rangle \geq \Phi[k], \forall k \geq 0$. This combines with the lemma in [9, Lemma 4], leading to the following result.

Proposition 3. *Let $\widetilde{\mathbf{W}}^{(\epsilon)}$ be a random vector with the same distribution as the steady-state distribution of the workload evolution under any scheduling algorithm. Consider the heavy-traffic limit that $\epsilon \downarrow 0$ and then suppose the variance vector $(\boldsymbol{\sigma}^{(\epsilon)})^2$ converge to a constant vector $\boldsymbol{\sigma}^2$. Then,*

$$\lim_{\epsilon\downarrow 0} \epsilon \mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{W}}^{(\epsilon)}\rangle] \geq \frac{\langle \mathbf{d}^2, \boldsymbol{\sigma}^2\rangle}{2}. \tag{6}$$

By Proposition 2 and 3, we can conclude that the proposed UOS Algorithm is heavy-traffic optimal.

4 Simulation Results

In this section, we conduct various simulations to verify the efficiency of our purposed UOS Algorithm. In the simulation, we assume that the number of flows arriving at each region n in each time slot follows Bernoulli distribution with mean λ_n. The file size of each arriving flow is equal to $5c_{\max}$ with probability $1/4$ and $c_{\max}$ otherwise. Then, the mean workload of a newly arriving flow is equal to 2.

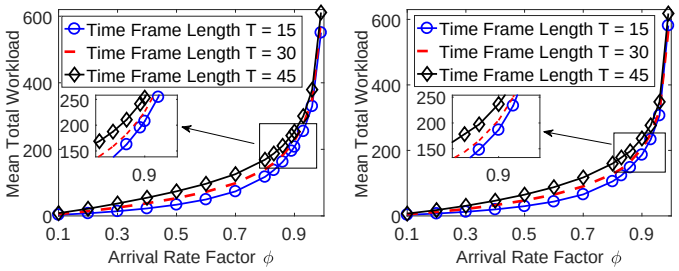
TABLE 2: Notations for Heavy-traffic Analysis

Symbol	Description
ϵ	heavy-traffic parameter
$\boldsymbol{\rho}^{(\epsilon)}$	traffic intensity vector with heavy-traffic parameter ϵ
$\mathbf{d}$	normal vector
$\mathcal{H}^{(\mathbf{d})}$	hyperplane with normal vector $\mathbf{d}$
$\boldsymbol{\rho}^{(0)}$	heavy-traffic intensity vector that strictly lying inside of $\mathcal{H}^{(\mathbf{d})}$
$b_{\mathbf{d}}$	a constant value that describes the hyperplane $\mathcal{H}^{(\mathbf{d})}$
$(\sigma_n^{(\epsilon)})^2$	variance of arrival process $\nu_n^{(\epsilon)}[k;t]$
$(\boldsymbol{\sigma}^{(\epsilon)})^2$	vector of $(\sigma_n^{(\epsilon)})^2$
$\boldsymbol{\sigma}^2$	the limit of $(\boldsymbol{\sigma}^{(\epsilon)})^2$ when $\epsilon \downarrow 0$
$\widetilde{W}_n^{(\epsilon)}$	random variable with the same distribution as the steady-state distribution of the workload evolution under the UOS Algorithm
$\widetilde{\mathbf{W}}^{(\epsilon)}$	vector of $\widetilde{W}_n^{(\epsilon)}$
$\nu_n^{(\epsilon)}[k]$	arriving workload parameterized with ϵ arrived at region n in frame k
$\boldsymbol{\nu}^{(\epsilon)}[k]$	vector of arrival process parameterized with ϵ in frame k
ν_Σ	total workload of arriving flows in frame k
$\Phi[k]$	constructed single server queue with arrival workload $\nu_\Sigma[k]$
$\boldsymbol{\lambda}$	vector of λ_n, $\boldsymbol{\lambda} \triangleq (\lambda_n)_{n=1}^N$

Each flow in each region experiences i.i.d. channel fading with rates of $10, 5, 1, 0$ and corresponding probabilities of $0.2, 0.6, 0.1, 0.1$. In our simulations, We consider $M = 3$ UAVs serving $N = 10$ regions, unless specified otherwise.

4.1 Throughput Performance

We consider the following two arrival cases: symmetric arrival $\boldsymbol{\lambda} = \phi \times (0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15)$, and asymmetric case $\boldsymbol{\lambda} = \phi \times (0.05, 0.1, 0.05, 0.05, 0.35, 0.05, 0.25, 0.15, 0.15, 0.3)$, where $\phi \in (0, 1)$ is the arrival load factor. Fig. 4 shows the mean total workload performance versus ϕ under our proposed UOS Algorithm with different time frame lengths. From Fig. 4, We can observe that the UOS Algorithm can stabilize the system for any $\phi \in (0, 1)$ in both symmetric and asymmetric arrival cases, no matter what the time frame length is. This verifies that the proposed UOS Algorithm indeed achieves maximum system throughput (cf. Proposition 1) and indicates that the time frame length does not affect the throughput performance. However, we can see that the UOS Algorithm with shorter time frame has smaller mean workload. This is because that the UOS Algorithm with the shorter time frame makes the UAV redirection decisions more frequently and thus can quickly response to the fast flow dynamics, yielding the better network performance.

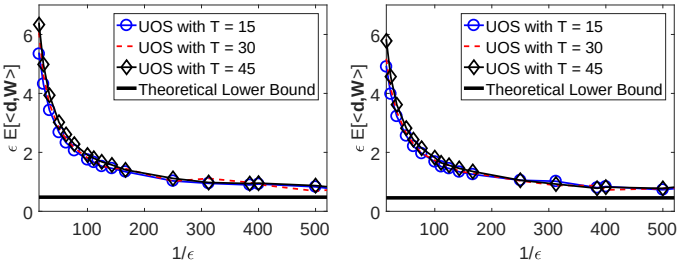


(a) Symmetric arrivals (b) Asymmetric arrivals

Fig. 4: The workload performance of the UOS Algorithm

4.2 Heavy-Traffic Performance

In this subsection, we study the impact of the heavy-traffic parameter ϵ on the mean workload performance. In particular, we consider both symmetric arrival rate vector $\boldsymbol{\lambda} = (0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15) - \epsilon\mathbf{d}$ and asymmetric arrival rate vector $\boldsymbol{\lambda} = (0.05, 0.1, 0.05, 0.05, 0.35, 0.05, 0.25, 0.15, 0.15, 0.3) - \epsilon\mathbf{d}$, where $\mathbf{d} = (1/\sqrt{10})_{n=1}^{10}$ is the normal vector of one of plane facets of the capacity region. From Fig. 5, we can observe that under our proposed UOS Algorithm with different frame lengths, the mean workloads converge to the same theoretical lower bound as the heavy-traffic parameter ϵ diminishes to zero in both symmetric and asymmetric arrival cases. This validates the heavy-traffic optimality of our proposed algorithm (cf. Propositions 2 and 3) and indicates that the frame length does not affect the heavy-traffic optimality.

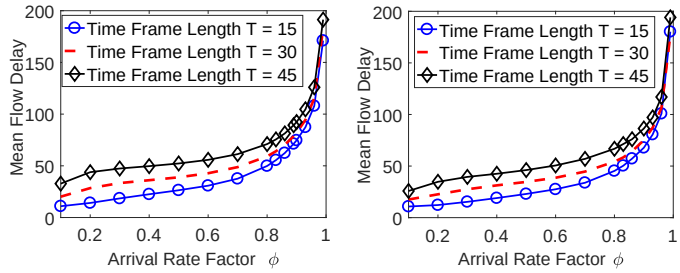


(a) Symmetric heavy traffic (b) Asymmetric heavy traffic

Fig. 5: The performance of UOS Algorithm under heavy traffic arrival processes

4.3 Mean Delay Performance

In this subsection, we study the mean delay performance of proposed UOS Algorithm with different frame lengths under both symmetric and asymmetric arrival processes. We can see from Fig. 6 that the frame length has a similar impact on the mean delay performance as that on the mean workload performance (cf. Fig. 4). The reason lies in that the smaller system workload leads to the shorter waiting time of each flows.



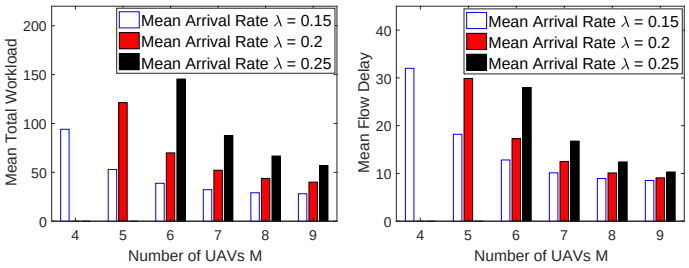
(a) Symmetric arrivals (b) Asymmetric arrivals

Fig. 6: The mean delay performance of the UOS Algorithm

4.4 Impact of the Number of UAVs

In this subsection, we study the impact of the number of UAVs on the network performance. From Fig. 7, we can observe that the system requires a larger number of UAVs to support a larger traffic intensity. For example, the minimum number of UAVs required to support the arrival rate of $\lambda = 0.2$ is 5 while it is 6 when $\lambda = 0.25$. Moreover, the addition of one UAV can dramatically improve the system

performance and further increasing the number of UAVs brings marginal performance improvement. Indeed, when $\lambda = 0.2$, deploying 6 UAVs can improve the mean workload by 42.37% and the mean delay by 42.21% compared with the case with 5 UAVs, while the additional one UAV only leads to additional 14.65% workload improvement and 16.01% delay improvement.

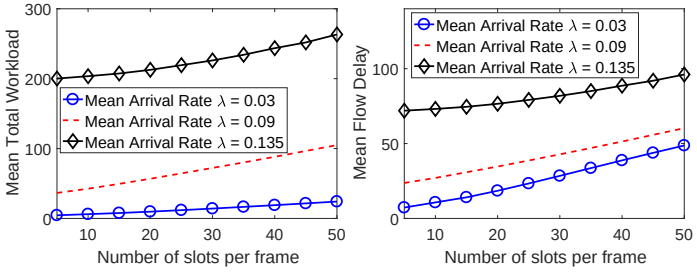


(a) Workload performance (b) Delay performance

Fig. 7: Impact of the number of UAVs

4.5 Impact of the Length of Time Frame

We study the impact of the number of the time slots per frame on the network performance in this subsection. As shown in Fig. 8, both the mean total workload and mean flow delay increase as the number of slots per frame increases under different traffic intensity values. This observation matches our simulation results in both Fig. 4 that the proposed UOS Algorithm with shorter time frame size redirects UAVs more frequently and quickly responses to the sharp flow dynamics, which is at the cost of increasing the energy consumption of the UAVs. Moreover, we can observe from Fig. 8 that the time frame length more significantly affects the mean delay performance when the traffic intensity is lower.



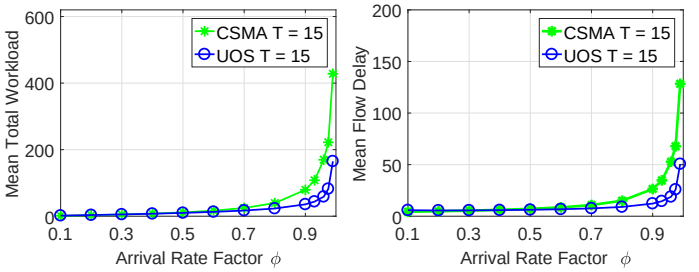
(a) Workload performance (b) Delay performance

Fig. 8: Impact of the length of time frame

4.6 Comparison with the CSMA Algorithm

In this subsection, we compare the CSMA Algorithm and the proposed UOS Algorithm in the non-fading case. Since the CSMA Algorithm only considers the scheduling decision, we run CSMA to obtain flow scheduling decisions and allocate UAVs to serve their corresponding regions. We let the number of UAVs $M = 3$ and the number of regions $N = 4$. The arrival rate vector is set as $\boldsymbol{\lambda} = \phi \times (0.375, 0.375, 0.375, 0.375)$, where $\phi \in (0, 1)$ is the arrival load factor.

We can see from the Fig. 9 that both the UOS Algorithm and the CSMA Algorithm achieve the full capacity region and thus are throughput-optimal. However, its mean



(a) Workload performance (b) Delay performance

Fig. 9: Performance comparison between CSMA and UOS Algorithm in a non-fading case

workload and flow latency performance is worse than that of the UOS Algorithm. This is because the proposed UOS Algorithm aims at serving as many users as possible in the fastest way, which can be interpreted as serving as many user as possible.

5 Proof of Equivalent Capacity Regions

In this section, we show the following two capacity regions are equivalent, where

$$\Lambda_1 \triangleq \left\{ \boldsymbol{\rho} = (\rho_n)_{n=1}^N : \sum_{n=1}^N \rho_n \leq M \text{ and } 0 \leq \rho_n \leq 1, \forall n \right\}$$

$$\Lambda_2 \triangleq \left\{ \boldsymbol{\rho} = (\rho_n)_{n=1}^N : \exists \alpha(\mathbf{S}), \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) = 1, \right.$$
$$\left. \rho_n \leq \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) S_n, \forall n \right\}.$$

First, we show $\Lambda_2 \subseteq \Lambda_1$. Let $\boldsymbol{\rho} \in \Lambda_2$, we have

$$\rho_n \leq \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) S_n \overset{(a)}{\leq} \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) = 1$$

where (a) uses the fact that $S_n \leq 1, \forall n = 1, 2, \ldots, N$. Besides,

$$\sum_{n=1}^N \rho_n \leq \sum_{n=1}^N \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) S_n$$
$$= \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) \sum_{n=1}^N S_n \overset{(b)}{\leq} \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) M = M.$$

where (b) uses the fact that $\sum_{n=1}^N S_n \leq M$. Hence, we have $\boldsymbol{\rho} \in \Lambda_1$ and thus $\Lambda_2 \subseteq \Lambda_1$.

Next, we show $\Lambda_1 \subseteq \Lambda_2$ by contradiction. Assume that $\boldsymbol{\rho} \in \Lambda_1$ and $\boldsymbol{\rho} \notin \Lambda_2$, then there exists ρ_n such that

$$\rho_n > \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S}) S_n, \forall \alpha(\mathbf{S}), \mathbf{S} \in \mathcal{S},$$

then select $\alpha(\mathbf{S}') = 1$ for the schedule $\mathbf{S}'$ with $S'_n = 1$, and we obtain that $\rho_n > 1$, which leads to the contradiction. Thus, $\Lambda_1 \subseteq \Lambda_2$.

6 Proof of Throughput Optimality

We consider the Lyapunov function $V(k) \triangleq \|\mathbf{W}[k]\|$ and its conditional expected drift

$$\begin{aligned}
&\mathbb{E}[\Delta V(\mathbf{W})|\mathbf{W}[k]=\mathbf{W}]\\
&\triangleq \mathbb{E}[\|\mathbf{W}[k+1]\| - \|\mathbf{W}[k]\| \,|\, \mathbf{W}[k]=\mathbf{W}]\\
&= \mathbb{E}[\sqrt{\|\mathbf{W}[k+1]\|^2} - \sqrt{\|\mathbf{W}[k]\|^2} \,|\, \mathbf{W}[k]=\mathbf{W}]\\
&\leq \frac{1}{2\|\mathbf{W}[k]\|}\mathbb{E}[\|\mathbf{W}[k+1]\|^2 - \|\mathbf{W}[k]\|^2 \,|\, \mathbf{W}[k]=\mathbf{W}], \quad (7)
\end{aligned}$$

where the last step follows from the fact that $f(x_2)-f(x_1) \leq (x_2-x_1)f'(x_1) = (x_2-x_1)/2\sqrt{x_1}$ due to the concavity of the function $f(x) \triangleq \sqrt{x}$ for $x>0$. For an easier exposition, let $L(k) \triangleq \|\mathbf{W}[k]\|^2$. In the rest of the proof, we omit the frame index $[k]$ for conciseness.

$$\begin{aligned}
&\mathbb{E}[\Delta L(\mathbf{W}) \,|\, \mathbf{W}] \triangleq \mathbb{E}\left[\|\mathbf{W}[k+1]\|^2 - \|\mathbf{W}[k]\|^2 \,|\mathbf{W}[k]=\mathbf{W}\right]\\
&= \mathbb{E}\left[\sum_{n=1}^{N}(W_n+\nu_n-S_n^*\mu_n)^2 - \sum_{n=1}^{N}W_n^2 \middle| \mathbf{W}\right]\\
&= \mathbb{E}\left[\sum_{n=1}^{N}\left(2W_n(\nu_n-S_n^*\mu_n)+(\nu_n-S_n^*\mu_n)^2\right)\middle|\mathbf{W}\right]\\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{n=1}^{N}\left(2W_n\sum_{t=1}^{T}(\nu_n[k;t]-S_n^*\mu_n[k;t])+(\nu_n-S_n^*\mu_n)^2\right)\middle|\mathbf{W}\right]\\
&\overset{(b)}{\leq} 2T\sum_{n=1}^{N}W_n\rho_n - 2\mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=1}^{T}W_nS_n^*\mu_n[k;t]\middle|\mathbf{W}\right] + B_1,
\end{aligned} \quad (8)$$

where step (a) follows from the definitions of ν_n and μ_n; (b) is true for $B_1 \triangleq N(\nu_{\max}^2+T^2)$, $\nu_{\max} \triangleq T\max_n A_n^{\max}\lceil F_n^{\max}/c_n^{\max}\rceil$ and due to the fact that the newly arriving workload is independent of the current system state.

Next, we consider the first term on the right-hand-side (RHS) of (8).

$$\begin{aligned}
2T\sum_{n=1}^{N}W_n\rho_n &\overset{(a)}{\leq} 2T\sum_{n=1}^{N}W_n\left(\sum_{\mathbf{S}\in\mathcal{S}}\alpha(\mathbf{S})S_n - \epsilon/T\right)\\
&= -2\epsilon\sum_{n=1}^{N}W_n + 2T\sum_{\mathbf{S}\in\mathcal{S}}\alpha(\mathbf{S})\sum_{n=1}^{N}W_nS_n\\
&\overset{(b)}{\leq} -2\epsilon\sum_{n=1}^{N}W_n + 2T\sum_{n=1}^{N}W_nS_n^*
\end{aligned} \quad (9)$$

where step (a) follows from the fact that $\boldsymbol{\rho} = (\rho_n)_{n=1}^N$ is strictly inside the capacity region Λ and hence there exists an $\epsilon>0$ and probability distribution $\{\alpha(\mathbf{S})\}_{\mathbf{S}\in\mathcal{S}}$ such that $\rho_n \leq \sum_{\mathbf{S}\in\mathcal{S}}\alpha(\mathbf{S})S_n - \epsilon/T$ (see Section 5 for the proof), and (b) follows from the definition of the UOS Algorithm and the fact that $\sum_{\mathbf{S}\in\mathcal{S}}\alpha(\mathbf{S})=1$.

Concerning the second term on the RHS of (8), we have

$$\begin{aligned}
&2\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{E}[W_n\mu_n[k;t]S_n^*|\mathbf{W}]\\
&\overset{(a)}{\geq} 2\sum_{n=1}^{N}\mathbb{E}\left[\sum_{t=1}^{T}\left(1-(1-p_n^{\max})^{|\mathcal{N}_n[k;t]|}\right)W_nS_n^*\middle|\mathbf{W}\right]\\
&\overset{(b)}{\geq} 2\sum_{n=1}^{N}\mathbb{E}\left[T\left(1-(1-p_n^{\max})^{\overline{D}_n}\right)W_nS_n^*\mathbb{1}_{\{|\mathcal{N}_n^{\min}|>\overline{D}_n\}}\middle|\mathbf{W}\right]\\
&\overset{(c)}{=} 2\left(1-\frac{1}{2T}\epsilon\right)\sum_{n=1}^{N}\mathbb{E}\left[W_nS_n^*T\mathbb{1}_{\{|\mathcal{N}_n^{\min}|>\overline{D}_n\}}\middle|\mathbf{W}\right]\\
&= 2\left(1-\frac{1}{2T}\epsilon\right)\sum_{n=1}^{N}\mathbb{E}\left[W_nS_n^*T(1-\mathbb{1}_{\{|\mathcal{N}_n^{\min}|\leq\overline{D}_n\}})\middle|\mathbf{W}\right]\\
&\overset{(d)}{\geq} 2\left(1-\frac{1}{2T}\epsilon\right)\sum_{n=1}^{N}W_nS_n^*T - B_2\\
&\geq -\epsilon\sum_{n=1}^{N}W_n + 2T\sum_{n=1}^{N}W_nS_n^* - B_2,
\end{aligned} \quad (10)$$

where step (a) uses the fact that in each slot of each time frame, the workload reduces by one in region n if at least one of its flows has the maximum channel rate; (b) is true for some $\overline{D}_n>0$ with $(1-p_n^{\max})^{\overline{D}_n} = \epsilon/2T$ and follows from the fact that $(1-p_n^{\max})^x$ decreases in $x\in[0,\infty)$ as $x\uparrow$ and $|\mathcal{N}_n^{\min}| \triangleq \min_{t=1,2,\ldots,T}|\mathcal{N}_n[k;t]|$; (c) uses the definition of $\overline{D}_n$; (d) is true for $B_2 \triangleq 2T\lceil F_n^{\max}/c_n^{\max}\rceil\sum_{n=1}^{N}(\overline{D}_n + TA_n^{\max})$.

By substituting (9) and (10) into (8), we have

$$\begin{aligned}
\mathbb{E}[\Delta L(\mathbf{W})|\mathbf{W}] &\leq -\epsilon\sum_{n=1}^{N}W_n + B_1 + B_2\\
&\leq -\epsilon\|\mathbf{W}\| + B_1 + B_2,
\end{aligned} \quad (11)$$

where the last step is true because $\|\mathbf{W}\|_1 \geq \|\mathbf{W}\|$.

By substituting (11) into (7), we have

$$\mathbb{E}[\Delta V(\mathbf{W})|\mathbf{W}] \leq -\frac{\epsilon}{2} + \frac{B_1+B_2}{2\|\mathbf{W}\|}. \quad (12)$$

This implies when the value of Lyapunov function $V(\mathbf{W}) = \|\mathbf{W}\|$ is sufficient large, its conditional expected drift is strictly negative.

Next, we show that the absolute drift of $V(\mathbf{W})$ is bounded by some constant.

$$\begin{aligned}
|\Delta V(\mathbf{W})| &= \left|\|\mathbf{W}[k+1]\| - \|\mathbf{W}[k]\|\right|\mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}}\\
&\overset{(a)}{\leq} \|\mathbf{W}[k+1]-\mathbf{W}[k]\|\mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}}\\
&\overset{(b)}{\leq} \|\mathbf{W}[k+1]-\mathbf{W}[k]\|_1\mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}}\\
&\leq N\max_n|\nu_n - S_n\mu_n|\\
&\leq N(\nu_{\max}+T),
\end{aligned} \quad (13)$$

where step (a) follows from the triangle inequality for vectors, and (b) uses the fact that the l_2-norm of a vector is less or equal to its l_1-norm.

In the end, (13) together with (12) implies the desired result by using [13, Theorem 2.3].

7 Proof of Heavy-traffic Optimality

In this section, we prove that our proposed UOS Algorithm minimizes the mean workload in the heavy-traffic regime. To this end, we first prove that the steady-state workload processes collapse into the normal vector of the hyperplane that the traffic intensity vector approaches (referred to as *state-space collapse*), in the sense that the deviation of the

steady-state workload vector away from the normal vector has all its moments bounded. Then based on this state-space collapse result, we obtain the desired result via the steady-state analysis of stochastic networks taking into account the sharp flow-level dynamics.

7.1 State Space Collapse

In this subsection, we show that the steady-state workloads collapse to the normal vector $\mathbf{d}$, i.e., the deviation of steady-state workload away from the vector $\mathbf{d}$ is uniformly bounded, independently of the heavy-traffic parameter ϵ. For a vector $\mathbf{x}$, we use $\mathbf{x}_\parallel$ and $\mathbf{x}_\perp$ to represent its projection and perpendicular vectors, respectively, i.e.,

$$\mathbf{x}_\parallel \triangleq \langle \mathbf{x}, \mathbf{d}\rangle \mathbf{d}, \text{ and } \mathbf{x}_\perp \triangleq \mathbf{x} - \mathbf{x}_\parallel. \tag{14}$$

Due to Proposition 1, we have $\mathbf{W}^{(\epsilon)}[k] \Rightarrow \widetilde{\mathbf{W}}^{(\epsilon)}$, where $\Rightarrow$ denotes convergence in probability. According to the continuous mapping theorem (see [8]), we have $\mathbf{W}_\parallel^{(\epsilon)}[k] \Rightarrow \widetilde{\mathbf{W}}_\parallel^{(\epsilon)}$ and $\mathbf{W}_\perp^{(\epsilon)}[k] \Rightarrow \widetilde{\mathbf{W}}_\perp^{(\epsilon)}$. Next, we prove that all moments of $\|\widetilde{\mathbf{W}}_\perp\|$ are bounded by some constants, independent from the heavy traffic parameter ϵ.

Proposition 4. *For any arrival intensity vector $\boldsymbol{\rho}^{(\epsilon)}$ parameterized with heavy-traffic parameter ϵ that is strictly inside the capacity region Λ, there exists a sequence of finite constants $\{K_r\}_{r=1,2,..}$ that are independent from ϵ such that $\mathbb{E}\left[\|\widetilde{\mathbf{W}}_\perp\|^r\right] \le K_r$.*

Proof: In the rest of the proof, we omit ϵ associated with the workload processes and drop time index for brevity. We consider the Lyapunov function $V_\perp(\mathbf{W}) \triangleq \|\mathbf{W}_\perp\|$, and we begin with showing the absolute drift of $V_\perp(\mathbf{W})$ is strictly bounded by some positive constant, then prove that its conditional expected drift becomes negative when $V_\perp(\mathbf{W})$ is sufficiently large. These lead to the desired result by [13, Theorem 2.3].

First, we consider the absolute drift of $V_\perp(\mathbf{W})$.

$$\begin{aligned}
&\left|V_\perp(\mathbf{W})\right| \triangleq \left|V_\perp(\mathbf{W}[k+1]) - V_\perp(\mathbf{W}[k])\right| \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&= \left|\left\|\mathbf{W}_\perp[k+1]\right\| - \left\|\mathbf{W}_\perp[k]\right\|\right| \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\overset{(a)}{\le} \left\|\mathbf{W}_\perp[k+1] - \mathbf{W}_\perp[k]\right\| \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&= \left\|\mathbf{W}[k+1] - \mathbf{W}[k] - \left(\mathbf{W}_\parallel[k+1] + \mathbf{W}_\parallel[k]\right)\right\| \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\overset{(b)}{\le} \left(\left\|\mathbf{W}[k+1] - \mathbf{W}[k]\right\| + \left\|\left(\mathbf{W}[k+1] - \mathbf{W}[k]\right)_\parallel\right\|\right) \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\le 2\left\|\mathbf{W}[k+1] - \mathbf{W}[k]\right\| \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\overset{(c)}{\le} 2\left\|\mathbf{W}[k+1] - \mathbf{W}[k]\right\|_1 \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\le 2N \max_n \left|W_n[k+1] - W_n[k]\right| \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\le 2N(\nu_{\max} + T),
\end{aligned} \tag{15}$$

where step (a) uses the fact that $|\|\mathbf{x}_1\| - \|\mathbf{x}_2\|| \le \|\mathbf{x}_1 - \mathbf{x}_2\|$; (b) is true because of $\|\mathbf{x}_1 - \mathbf{x}_2\| \le \|\mathbf{x}_1\| + \|\mathbf{x}_2\|$; (c) follows from the fact that $\|\mathbf{x}\| \le \|\mathbf{x}\|_1$ for any vector $\mathbf{x}$.

Next, we show that the conditional expected drift of $V_\perp(\mathbf{W})$ is strictly negative whenever $V_\perp(\mathbf{W})$ is large enough. However, we note that it is hard to prove this directly. Instead, we use the drifts of $\|\mathbf{W}\|$ and $\|\mathbf{W}_\parallel\|$ to upper-bound that of $V_\perp(\mathbf{W})$.

$$\begin{aligned}
\Delta V_\perp(\mathbf{W}) &= \left(\sqrt{\|\mathbf{W}_\perp[k+1]\|^2} - \sqrt{\|\mathbf{W}_\perp[k]\|^2}\right) \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}} \\
&\le \frac{1}{2\|\mathbf{W}_\perp\|} \left(\Delta L(\mathbf{W}) - \Delta L_\parallel(\mathbf{W})\right),
\end{aligned} \tag{16}$$

where the last step is true for $L(\mathbf{W}) \triangleq \|\mathbf{W}\|^2, L_\parallel(\mathbf{W}) \triangleq \|\mathbf{W}_\parallel\|^2$ and their drifts are defined by $\Delta L(\mathbf{W}) \triangleq \left(L(\mathbf{W}[k+1]) - L(\mathbf{W}[k])\right) \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}}$ and $\Delta L_\parallel(\mathbf{W}) \triangleq \left(L_\parallel(\mathbf{W}[k+1]) - L_\parallel(\mathbf{W}[k])\right) \mathbb{1}_{\{\mathbf{W}[k]=\mathbf{W}\}}$, and follows from the concavity of the function $f(x) \triangleq \sqrt{x}$, i.e., $f(x_2) - f(x_1) \le (x_2 - x_1) f'(x_1) = (x_2 - x_1)/2\sqrt{x_1}$ with $x_1 = \|\mathbf{W}_\perp[k]\|^2$ and $x_2 = \|\mathbf{W}_\perp[k+1]\|^2$.

Then we study the conditional expected drifts $\Delta L(\mathbf{W})$ and $\Delta L_\parallel(\mathbf{W})$, respectively.

$$\begin{aligned}
&\mathbb{E}\left[\Delta L(\mathbf{W})|\mathbf{W}\right] \\
&= \mathbb{E}\left[\|\mathbf{W}[k+1]\|^2 - \|\mathbf{W}[k]\|^2 \Big| \mathbf{W}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\|\mathbf{W} + \boldsymbol{\nu} - \mathbf{S}^* \otimes \boldsymbol{\mu}\|^2 - \|\mathbf{W}\|^2 \Big| \mathbf{W}\right] \\
&= \mathbb{E}\left[2\langle \mathbf{W}, \boldsymbol{\nu} - \mathbf{S}^* \otimes \boldsymbol{\mu}\rangle + \|\boldsymbol{\nu} - \mathbf{S}^* \otimes \boldsymbol{\mu}\|^2 \Big| \mathbf{W}\right] \\
&\overset{(b)}{\le} 2\mathbb{E}\left[\langle \mathbf{W}, \boldsymbol{\nu}\rangle | \mathbf{W}\right] - 2\mathbb{E}\left[\langle \mathbf{W}, \mathbf{S}^* \otimes \boldsymbol{\mu}\rangle | \mathbf{W}\right] + B_1 \\
&= 2T\left(\langle \mathbf{W}, \boldsymbol{\rho}\rangle - \mathbb{E}\left[\langle \mathbf{W}, \mathbf{S}^*\rangle | \mathbf{W}\right]\right) \\
&\quad + 2\mathbb{E}\left[\langle \mathbf{W}, \mathbf{S}^* \otimes (T\mathbf{1} - \boldsymbol{\mu})\rangle | \mathbf{W}\right] + B_1,
\end{aligned} \tag{17}$$

where step (a) is true for $\otimes$ denoting elementary multiplication between two vectors and (b) is true for $B_1 \triangleq N(\nu_{\max}^2 + T^2)$.

Now we provide the upper bound on each term in the RHS of (17).

$$\begin{aligned}
&\langle \mathbf{W}, \boldsymbol{\rho}\rangle - \mathbb{E}\left[\langle \mathbf{W}, \mathbf{S}^*\rangle | \mathbf{W}\right] \\
&\overset{(a)}{=} \langle \mathbf{W}, \boldsymbol{\rho}^{(0)} - \epsilon \mathbf{d}\rangle - \langle \mathbf{W}, \mathbb{E}\left[\mathbf{S}^* | \mathbf{W}\right]\rangle \\
&= -\epsilon \|\mathbf{W}_\parallel\| + \langle \mathbf{W}, \boldsymbol{\rho}^{(0)} - \mathbb{E}\left[\mathbf{S}^* | \mathbf{W}\right]\rangle \\
&\overset{(b)}{=} -\epsilon \|\mathbf{W}_\parallel\| + \min_{\mathbf{r} \in \Lambda} \langle \mathbf{W}, \boldsymbol{\rho}^{(0)} - \mathbf{r}\rangle \\
&\overset{(c)}{\le} -\epsilon \|\mathbf{W}_\parallel\| + \min_{\mathbf{r} \in \mathcal{B}_\delta} \langle \mathbf{W}, \boldsymbol{\rho}^{(0)} - \mathbf{r}\rangle \\
&\overset{(d)}{=} -\epsilon \|\mathbf{W}_\parallel\| + \min_{\mathbf{r} \in \mathcal{B}_\delta} \langle \mathbf{W}_\perp, \boldsymbol{\rho}^{(0)} - \mathbf{r}\rangle \\
&\overset{(e)}{=} -\epsilon \|\mathbf{W}_\parallel\| - \delta \|\mathbf{W}_\perp\|,
\end{aligned} \tag{18}$$

where step (a) follows from the definition of $\boldsymbol{\rho}^{(0)}$; (b) uses the definition of the UOS Algorithm, i.e., $\langle \mathbf{W}, \mathbb{E}[\mathbf{S}^*|\mathbf{W}]\rangle = \max_{\mathbf{r} \in \Lambda} \langle \mathbf{W}, \mathbf{r}\rangle$; (c) is true for $\mathcal{B}_\delta \triangleq \{\mathbf{r} \in \mathcal{H}^{(\mathbf{d})} \cap \Lambda : \|\mathbf{r} - \boldsymbol{\rho}^{(0)}\| \le \delta\}$ since we assume that $\boldsymbol{\rho}^{(0)}$ is the relative interior point of $\mathcal{H}^{(\mathbf{d})} \cap \Lambda$; (d) is true because $\mathbf{W}_\parallel$ is perpendicular to the vector $\boldsymbol{\rho}^{(0)} - \mathbf{r}$, i.e., $\langle \mathbf{W}_\parallel, \boldsymbol{\rho}^{(0)} - \mathbf{r}\rangle = 0$; and (e) follows from the definition of $\mathcal{B}_\delta$.

For the term $\mathbb{E}\left[\langle \mathbf{W}, \mathbf{S}^* \otimes (T\mathbf{1} - \boldsymbol{\mu})\rangle | \mathbf{W}\right]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{W}, \mathbf{S}^* \otimes (T\mathbf{1} - \boldsymbol{\mu})\rangle | \mathbf{W}\right] \\
&= \mathbb{E}\left[\sum_{n=1}^N W_n S_n^*(T - \mu_n) \Big| \mathbf{W}\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{=} \mathbb{E}\left[\sum_{n=1}^{N} W_n S_n^* \sum_{t=1}^{T}(1-\mu_n[k;t])\middle|\mathbf{W}\right]\\
&\overset{(b)}{\leq} \mathbb{E}\left[\sum_{n=1}^{N} W_n S_n^* \sum_{t=1}^{T}(1-p_n^{\max})^{|\mathcal{N}_n[k;t]|}\middle|\mathbf{W}\right]\\
&\overset{(c)}{\leq} \mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=1}^{T} W_n(1-p_n^{\max})^{(W_n-T)/w_{\max}}\middle|\mathbf{W}\right]\\
&\overset{(d)}{=} \mathbb{E}\Bigg[\sum_{n=1}^{N}\sum_{t=1}^{T} W_n(1-p_n^{\max})^{(W_n-T)/w_{\max}}\mathbb{1}_{\{W_n\leq\overline{w}_n\}}\\
&\quad+\sum_{n=1}^{N}\sum_{t=1}^{T} W_n(1-p_n^{\max})^{(W_n-T)/w_{\max}}\mathbb{1}_{\{W_n>\overline{w}_n\}}\middle|\mathbf{W}\Bigg]\\
&\overset{(e)}{\leq} \sum_{n=1}^{N}\sum_{t=1}^{T}\overline{w}_n(1-p_n^{\max})^{-T/w_{\max}}+\sum_{n=1}^{N}\sum_{t=1}^{T}1 \triangleq B_3, \quad (19)
\end{aligned}$$

where step (a) uses the fact that $\mu_n[k]=\sum_{t=1}^{T}\mu_n[k;t]$; (b) follows from the fact that $\mu_n[k;t]$ equals to one when a UAV is serving region n and at least one flow in region n has the maximum channel rate in time slot t; (c) is true for $w_{\max}\triangleq\lceil F_{\max}/c_{\max}\rceil$ denoting the maximum workload of a single arriving flow and follows from the fact that $S_n^*\leq 1$ and the facts that $|\mathcal{N}_n[k;t]|\geq W_n[k;t]/w_{\max}$ and $W_n[k;t]\geq W_n[k]-T,\forall t=1,2,\ldots,T$; (d) is true for some constants $\overline{w}_n$ such that $W_n(1-p_n^{\max})^{(W_n-T)/w_{\max}}\leq 1$ when $W_n>\overline{w}_n$ due to the fact that $\lim_{x\to\infty}x\beta^{hx}=0$ for $x>0, 0<\beta<1$ and h is some constant; and (e) uses definition of $\overline{w}_n$.

Substituting (18) and (19) into (17), we have

$$\mathbb{E}[\Delta L(\mathbf{W})|\mathbf{W}]\leq -2T\epsilon\|\mathbf{W}_{\|}\|-2T\delta\|\mathbf{W}_{\perp}\|+B_1+2B_3. \quad (20)$$

Next, we consider the conditional expected drift $\mathbb{E}[\Delta L_{\|}(\mathbf{W})|\mathbf{W}]$.

$$\begin{aligned}
&\mathbb{E}\left[\Delta L_{\|}(\mathbf{W})\middle|\mathbf{W}\right]=\mathbb{E}\left[\langle\mathbf{d},\mathbf{W}[k+1]\rangle^2-\langle\mathbf{d},\mathbf{W}[k]\rangle^2\middle|\mathbf{W}\right]\\
&=\mathbb{E}\left[\langle\mathbf{d},\mathbf{W}+\boldsymbol{\nu}-\mathbf{S}^*\otimes\boldsymbol{\mu}\rangle^2-\langle\mathbf{d},\mathbf{W}\rangle^2\middle|\mathbf{W}\right]\\
&=\mathbb{E}\left[2\langle\mathbf{d},\mathbf{W}\rangle\langle\mathbf{d},\boldsymbol{\nu}-\mathbf{S}^*\otimes\boldsymbol{\mu}\rangle+\langle\boldsymbol{\nu}-\mathbf{S}^*\otimes\boldsymbol{\mu}\rangle^2\middle|\mathbf{W}\right]\\
&\overset{(a)}{\geq}\mathbb{E}\left[2\langle\mathbf{d},\mathbf{W}\rangle\langle\mathbf{d},\boldsymbol{\nu}-\mathbf{S}^*\otimes\boldsymbol{\mu}\rangle\middle|\mathbf{W}\right]\\
&\overset{(b)}{=}2\|\mathbf{W}_{\|}\|\langle\mathbf{d},T\boldsymbol{\rho}-\mathbb{E}[\mathbf{S}^*\otimes\boldsymbol{\mu}|\mathbf{W}]\rangle\\
&\overset{(c)}{\geq}-2T\epsilon\|\mathbf{W}_{\|}\|+2T\|\mathbf{W}_{\|}\|\left(\langle\mathbf{d},\boldsymbol{\rho}^{(0)}\rangle-\langle\mathbf{d},\mathbb{E}[\mathbf{S}^*|\mathbf{W}]\rangle\right)\\
&\overset{(d)}{\geq}-2T\epsilon\|\mathbf{W}_{\|}\|, \quad (21)
\end{aligned}$$

where step (a) is true because $\langle\boldsymbol{\nu}-\mathbf{S}^*\otimes\boldsymbol{\mu}\rangle^2\geq 0$; (b) follows from the fact that the workload arrival process $\boldsymbol{\nu}$ is independent from current workload; (c) uses the definition of $\rho^{(0)}$ and the fact that $\mu_n\leq T$ for all n; (d) follows from the fact that all feasible schedules are inside the capacity region Λ and should be below the hyperplane $\mathcal{H}^{(\mathbf{d})}$.

Finally, apply the result in (20) and (21) to (16), we have

$$\mathbb{E}\left[\Delta V_{\perp}(\mathbf{W})|\mathbf{W}\right]\leq -T\delta+\frac{B_1+2B_3}{2\|\mathbf{W}_{\perp}\|}. \quad (22)$$

Thus when $V_{\perp}(\mathbf{W})=\|\mathbf{W}_{\perp}\|$ is sufficiently large, the expectation of its drift is strictly negative, independent of heavy traffic parameter ϵ. □

7.2 Upper Bound Analysis

Having established the state-space collapse result under our proposed UOS Algorithm, we are ready to analyze the mean workload performance in the heavy-traffic regime. To facilitate our proof, we introduce $U_n[k;t]\triangleq S_n^*[k](1-\mu_n[k;t])$ and $U_n[k]\triangleq\sum_{t=1}^{T}U_n[k;t]$, and thus the evolution of the workload under the UOS Algorithm can be rewritten as follows:

$$\mathbf{W}[k+1]=\mathbf{W}[k]+\boldsymbol{\nu}[k]-T\mathbf{S}^*[k]+\mathbf{U}[k], \quad (23)$$

where $\mathbf{U}[k]\triangleq(U_n[k])_{n=1}^{N}$. Note that $\mathbf{U}[k]$ is quite different from the unused service in the traditional queueing system (e.g., [9], [15]). In fact, if the flow served by the UAV in region n in time slot t of frame k has a file size larger than the maximum channel rate $c_n^{\max}$ and the service rate is less than $c_n^{\max}$, then $\mu_n[k;t]$ equals to 0 and thus $U_n[k;t]$ is 1. On the other hand, this flow indeed receives the full amount of service from the UAV in time slot t and does not incur any unused service. Therefore, the analytical method developed in [9] dealing with the unused service is no longer suitable in our considered flow-level dynamic model, and hence new techniques are required for analyzing heavy-traffic performance of our proposed UOS Algorithm in the presence of dynamic flows.

In the rest of the proof, we drop the heavy-traffic parameter ϵ associated with the workload process for simplicity. In order to derive an upper bound for $\mathbb{E}[\langle\mathbf{d},\widetilde{\mathbf{W}}\rangle]$, we use the following identity:

$$\begin{aligned}
\mathbb{E}[\langle\mathbf{d},\widetilde{\mathbf{W}}\rangle\langle\mathbf{d},T\widetilde{\mathbf{S}}^*-\boldsymbol{\nu}\rangle]&=\frac{1}{2}\mathbb{E}[\langle\mathbf{d},\boldsymbol{\nu}-T\widetilde{\mathbf{S}}^*\rangle^2]+\frac{1}{2}\mathbb{E}[\langle\mathbf{d},\widetilde{\mathbf{U}}\rangle^2]\\
&\quad+\mathbb{E}[\langle\mathbf{d},\widetilde{\mathbf{W}}+\boldsymbol{\nu}-T\widetilde{\mathbf{S}}^*\rangle\langle\mathbf{d},\widetilde{\mathbf{U}}\rangle], \quad (24)
\end{aligned}$$

which is derived by setting the expected drift of $\langle\mathbf{d},\widetilde{\mathbf{W}}\rangle^2$ to zero (also see [9, Lemma 8]). This is doable since the second moment of $\widetilde{\mathbf{W}}$ is bounded according to the Proposition 1.

Next, we provide a claim:

Claim 1. *Let* $q_{\mathbf{d}}\triangleq\Pr\left\{\langle\mathbf{d},\widetilde{\mathbf{S}}^*\rangle=b_{\mathbf{d}}\right\}$ *and* $\gamma_{\mathbf{d}}\triangleq\min\left\{b_{\mathbf{d}}-\langle\mathbf{d},\mathbf{r}\rangle:\ \textit{for all}\ \mathbf{r}\in\mathcal{S},\mathbf{r}\notin\mathcal{H}^{(\mathbf{d})}\right\}$ *is strictly positive. Then, for any* $\epsilon\in(0,\gamma_{\mathbf{d}})$, *we have*

$$1-q_{\mathbf{d}}\leq\frac{\epsilon}{\gamma_{\mathbf{d}}}. \quad (25)$$

The proof of Claim 1 uses the stability condition, i.e., $\mathbb{E}\left[\langle\mathbf{d},\widetilde{\mathbf{S}}^*\rangle\right]\geq\langle\mathbf{d},\boldsymbol{\rho}\rangle=b_{\mathbf{d}}-\epsilon$, and follows from the same line of arguments in [9, Claim 1]. We skip the proof for brevity. Claim 1 immediately implies that

$$\begin{aligned}
&\mathbb{E}\left[\left(b_{\mathbf{d}}-\langle\mathbf{d},\widetilde{\mathbf{S}}^*\rangle\right)^2\right]\\
=&\mathbb{E}\left[\left(b_{\mathbf{d}}-\langle\mathbf{d},\widetilde{\mathbf{S}}^*\rangle\right)^2\middle|\langle\mathbf{d},\widetilde{\mathbf{S}}^*\rangle\neq b_{\mathbf{d}}\right](1-q_{\mathbf{d}})\\
\leq&\frac{\epsilon}{\gamma_{\mathbf{d}}}\left(b_{\mathbf{d}}^2+\|\mathbf{d}\|_1^2\right)\\
\leq&\frac{\epsilon}{\gamma_{\mathbf{d}}}\left(b_{\mathbf{d}}^2+N\right), \quad (26)
\end{aligned}$$

where the last step uses the Cauchy-Schwarz inequality that $\|\mathbf{d}\|_1^2 = (\sum_{n=1}^N d_n \cdot 1)^2 \le (\sum_{n=1}^N 1^2)(\sum_{n=1}^N d_n^2) = N\|\mathbf{d}\|^2$ and the fact that $\|\mathbf{d}\| = 1$.

Then we provide a lower bound on the left-hand-side of the identity (24).

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{d}, \widetilde{\mathbf{W}}\rangle \langle \mathbf{d}, T\widetilde{\mathbf{S}}^* - \boldsymbol{\nu}\rangle\right] \\
&= \mathbb{E}\left[\|\widetilde{\mathbf{W}}_{\|}\|(Tb_{\mathbf{d}} - \langle \mathbf{d}, \boldsymbol{\nu}\rangle)\right] - \mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|(Tb_{\mathbf{d}} - \langle \mathbf{d}, T\widetilde{\mathbf{S}}^*\rangle)] \\
&\overset{(a)}{=} T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - T\mathbb{E}[\|\widetilde{\mathbf{W}}\|\cos(\theta)(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)] \\
&\overset{(b)}{=} T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - T\mathbb{E}[\|\widetilde{\mathbf{W}}\|\cos(\theta)\mathbb{1}_{\{\theta>\overline{\theta}\}}(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)] \\
&\overset{(c)}{=} T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - T\mathbb{E}[\|\widetilde{\mathbf{W}}_{\perp}\|\cot(\theta)\mathbb{1}_{\{\theta>\overline{\theta}\}}(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)] \\
&\overset{(d)}{\ge} T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - \cot(\overline{\theta})T\mathbb{E}[\|\widetilde{\mathbf{W}}_{\perp}\|\mathbb{1}_{\{\theta>\overline{\theta}\}}(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)]) \\
&\ge T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - \cot(\overline{\theta})T\mathbb{E}[\|\widetilde{\mathbf{W}}_{\perp}\|(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)] \\
&\overset{(e)}{\ge} T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - \cot(\overline{\theta})T\sqrt{\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}_{\perp}\right\|^2\right]\mathbb{E}[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)^2]} \\
&\overset{(f)}{\ge} T\epsilon\mathbb{E}[\|\widetilde{\mathbf{W}}_{\|}\|] - \cot(\overline{\theta})T\sqrt{\frac{K_2\epsilon}{\gamma_{\mathbf{d}}}(b_{\mathbf{d}}^2 + N)},
\end{aligned} \tag{27}$$

where step (a) is true because the definition of the hyperplane $\mathcal{H}^{(\mathbf{d})}$ and $\theta \triangleq \arccos(\|\widetilde{\mathbf{W}}_{\|}\|/\|\widetilde{\mathbf{W}}\|)$ denotes the angle between $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}_{\|}$; (b) is true for some $\overline{\theta} \in (0, \pi/2]$ satisfying $\langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle = b_{\mathbf{d}}$ for all $\widetilde{\mathbf{W}}$ such that $\|\widetilde{\mathbf{W}}_{\|}\|/\|\widetilde{\mathbf{W}}\| \ge \cos(\overline{\theta})$ (due to our proposed UOS Algorithm); (c) follows from the fact that $\|\widetilde{\mathbf{W}}\|\sin(\theta) = \|\widetilde{\mathbf{W}}_{\perp}\|$; (d) uses the fact that $\cot(\theta)$ is strictly decreasing in $\theta \in (0, \pi/2]$; (e) uses Cauchy-Swartz inequality; (f) uses Proposition 4 and (26).

Next, we provide upper bounds of each term on the RHS of the identity (24).

For the term $\mathbb{E}\left[\langle \mathbf{d}, \boldsymbol{\nu} - T\widetilde{\mathbf{S}}^*\rangle^2\right]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{d}, \boldsymbol{\nu} - T\widetilde{\mathbf{S}}^*\rangle^2\right] \\
=&\mathbb{E}\left[\left(\langle \mathbf{d}, \boldsymbol{\nu}\rangle - Tb_{\mathbf{d}} + Tb_{\mathbf{d}} - \langle \mathbf{d}, T\widetilde{\mathbf{S}}^*\rangle\right)^2\right] \\
=&\mathbb{E}\left[(\langle \mathbf{d}, \boldsymbol{\nu}\rangle - Tb_{\mathbf{d}})^2\right] + T^2\mathbb{E}\left[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)^2\right] \\
&+ 2T^2(\langle \mathbf{d}, \boldsymbol{\rho}\rangle - b_{\mathbf{d}})\mathbb{E}[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)] \\
\overset{(a)}{=}&\mathbb{E}[(\langle \mathbf{d}, \boldsymbol{\nu}\rangle - Tb_{\mathbf{d}})^2] + T^2\mathbb{E}\left[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)^2\right] \\
&- 2\epsilon T^2\mathbb{E}[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)] \\
\overset{(b)}{\le}&\mathbb{E}\left[(\langle \mathbf{d}, \boldsymbol{\nu} - T\boldsymbol{\rho}\rangle + \langle \mathbf{d}, T\boldsymbol{\rho}\rangle - Tb_{\mathbf{d}})^2\right] \\
&+ T^2\mathbb{E}\left[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)^2\right] \\
\overset{(c)}{=}&\mathbb{E}\left[\langle \mathbf{d}, \boldsymbol{\nu} - T\boldsymbol{\rho}\rangle^2\right] + T^2\epsilon^2 + T^2\mathbb{E}\left[(b_{\mathbf{d}} - \langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle)^2\right] \\
\overset{(d)}{\le}&T\langle \mathbf{d}^2, \boldsymbol{\sigma}^2\rangle + T^2\epsilon^2 + T^2\frac{\epsilon}{\gamma_{\mathbf{d}}}\left(b_{\mathbf{d}}^2 + N\right),
\end{aligned} \tag{28}$$

where step (a) is true because the definition of $b_{\mathbf{d}}$ (c.f. Section 3); (b) stands because for any feasible scheduling $\widetilde{\mathbf{S}}$, its projection on the normal vector $\mathbf{d}$ should be less than $b_{\mathbf{d}}$, i.e., $\langle \mathbf{d}, \widetilde{\mathbf{S}}\rangle \le b_{\mathbf{d}}$; (c) again uses the definition of $b_{\mathbf{d}}$; (d) uses (26) and the facts that $\mathbb{E}\left[\langle \mathbf{d}, \boldsymbol{\nu} - T\boldsymbol{\rho}\rangle^2\right] = \sum_{n=1}^N \sum_{t=1}^T d_n^2\sigma_n^2$ since both number of flows and their file sizes are i.i.d. over time and independently across regions.

To get the upper bound of $\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle^2]$, we first show that

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle] &= \mathbb{E}[\langle \mathbf{d}, T\widetilde{\mathbf{S}}^* - \boldsymbol{\nu}\rangle] = \mathbb{E}[\langle \mathbf{d}, T\widetilde{\mathbf{S}}^*\rangle] - \langle \mathbf{d}, T\boldsymbol{\rho}\rangle \\
&\overset{(a)}{=} T(\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{S}}^*\rangle] - (b_{\mathbf{d}} - \epsilon)) \overset{(b)}{\le} T\epsilon,
\end{aligned} \tag{29}$$

where step (a) follows from the definition of $b_{\mathbf{d}}$; (b) is true any feasible schedule should lie below the hyperplane $\mathcal{H}^{(\mathbf{d})}$, i.e., $\langle \mathbf{d}, \mathbf{S}\rangle \le b_{\mathbf{d}}$ for any feasible schedule $\mathbf{S}$. Then the upper bound of $\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle^2]$ is given by

$$\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle^2] \le \langle \mathbf{d}, T\mathbf{1}\rangle\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle] \le \epsilon T^2\|\mathbf{d}\|_1 \le \epsilon T^2\sqrt{N}, \tag{30}$$

where the first step holds because $\widetilde{U}_n$ is less than T by its definition and the second step uses (29).

Next, we show the upper bound of $\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{W}} + \boldsymbol{\nu} - T\widetilde{\mathbf{S}}^*\rangle\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle]$ in (24), and we define the following terms:

$\widehat{\mathbf{d}} \triangleq (d_n)_{n\in\mathcal{D}_+}$, $\widehat{\mathbf{W}} \triangleq (\widetilde{W}_n)_{n\in\mathcal{D}_+}$, and $\widehat{\mathbf{U}} \triangleq (\widetilde{U}_n)_{n\in\mathcal{D}_+}$, where $\mathcal{D}_+ \triangleq \{n \in \{1, 2, \ldots, N\} : d_n > 0\}$,

then we have

$$\begin{aligned}
&\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{W}} + \boldsymbol{\nu} - T\widetilde{\mathbf{S}}^*\rangle\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle] \\
\overset{(a)}{=}&\mathbb{E}\left[\langle \mathbf{d}, \widetilde{\mathbf{W}}^+\rangle\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle\right] - \mathbb{E}\left[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle^2\right] \\
\le&\mathbb{E}\left[\langle \mathbf{d}, \widetilde{\mathbf{W}}^+\rangle\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle\right] \\
\overset{(b)}{=}&\mathbb{E}\left[\langle \widehat{\mathbf{W}}_{\|}^+, \widehat{\mathbf{U}}_{\|}\rangle\right] = \mathbb{E}\left[\langle \widehat{\mathbf{W}}_{\|}^+, \widehat{\mathbf{U}} - \widehat{\mathbf{U}}_{\perp}\rangle\right] \\
\overset{(c)}{=}&\mathbb{E}\left[\langle \widehat{\mathbf{W}}_{\|}^+, \widehat{\mathbf{U}}\rangle\right] = \mathbb{E}\left[\langle \widehat{\mathbf{W}}^+ - \widehat{\mathbf{W}}_{\perp}^+, \widehat{\mathbf{U}}\rangle\right] \\
=&\mathbb{E}\left[\langle \widehat{\mathbf{W}}^+, \widehat{\mathbf{U}}\rangle\right] - \mathbb{E}\left[\langle \widehat{\mathbf{W}}_{\perp}^+, \widehat{\mathbf{U}}\rangle\right],
\end{aligned} \tag{31}$$

where step (a) is true for $\widetilde{\mathbf{W}}^+$ denoting $\widetilde{\mathbf{W}}[k+1]$; (b) uses the definition of $\widehat{\mathbf{d}}$, $\widehat{\mathbf{W}}_{\|}^+ \triangleq \langle \widehat{\mathbf{d}}, \widehat{\mathbf{W}}^+\rangle$ and $\widehat{\mathbf{U}}_{\|} \triangleq \langle \widehat{\mathbf{d}}, \widehat{\mathbf{U}}\rangle$; (c) is true because $\langle \mathbf{x}_{\|}, \mathbf{x}_{\perp}\rangle = 0$ for all vector $\mathbf{x}$.

Besides, based on (29), we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{U}_n\right] \le \frac{1}{d_{\min}}\mathbb{E}\left[\sum_{n\in\mathcal{D}_+} d_n\widetilde{U}_n\right] &= \frac{1}{d_{\min}}\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle] \\
&\le \frac{T\epsilon}{d_{\min}},
\end{aligned} \tag{32}$$

where the first step is true because $d_{\min} \triangleq \min_{\{n\in\mathcal{D}_+\}} d_n$ is strictly positive, and the last step uses (29).

Now we focus on each term on the RHS of (31).

$$\begin{aligned}
&\mathbb{E}\left[\langle \widehat{\mathbf{W}}^+, \widehat{\mathbf{U}}\rangle\right] = \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n[k+1]\widetilde{U}_n[k]\right] \\
&\le \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\left(\widetilde{W}_n[k] + \nu_n[k]\right)\widetilde{U}_n[k]\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n\widetilde{U}_n\right] + \sqrt{\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\nu_n^2\right]\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{U}_n^2\right]} \\
&\overset{(b)}{\le} \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n\widetilde{U}_n\right] + \sqrt{N\nu_{\max}^2\frac{\epsilon T^2}{d_{\min}}},
\end{aligned} \tag{33}$$

where step (a) uses Cauchy-Schwartz inequality; (b) follows from the fact that $\nu_n \leq \nu_{\max}$, $|\mathcal{D}_+| \leq N$, $\widetilde{U}_n \leq T$, and inequality (32).

For the first term in (33), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n\widetilde{U}_n\right] = \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n\widetilde{S}_n^*\sum_{t=1}^{T}(1-\widetilde{\mu}_n[k;t])\right]\\
&\overset{(a)}{\leq} \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n\widetilde{S}_n^*\sum_{t=1}^{T}(1-p_n^{\max})^{|\widetilde{\mathcal{N}}_n[k;t]|}\right]\\
&= \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{W}_n\widetilde{S}_n^*\sum_{t=1}^{T}(1-p_n^{\max})^{\frac{|\widetilde{\mathcal{N}}_n[k;t]|}{2}}(1-p_n^{\max})^{\frac{|\widetilde{\mathcal{N}}_n[k;t]|}{2}}\right]\\
&\overset{(b)}{\leq} \mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\sum_{t=1}^{T}\widetilde{W}_n\widetilde{S}_n^*(1-p_n^{\max})^{\frac{\widetilde{W}_n-T}{2w_{\max}}}(1-p_n^{\max})^{\frac{|\widetilde{\mathcal{N}}_n[k;t]|}{2}}\right]\\
&\overset{(c)}{\leq} \hat{w}\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\sum_{t=1}^{T}\widetilde{S}_n^*(1-p_n^{\max})^{\frac{|\widetilde{\mathcal{N}}_n[k;t]|}{2}}\right]\\
&\overset{(d)}{\leq} \hat{w}\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\sum_{t=1}^{T}\widetilde{S}_n^*(1-p_n^{\max})^{l|\widetilde{\mathcal{N}}_n[k;t]|}\right]^{\frac{1}{2l}}(T|\mathcal{D}_+|)^{\frac{2l-1}{2l}}\\
&\overset{(e)}{\leq} \hat{w}\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\sum_{t=1}^{T}\widetilde{S}_n^*(p_n^{\min})^{|\widetilde{\mathcal{N}}_n[k;t]|}\right]^{\frac{1}{2l}}(TN)^{\frac{2l-1}{2l}}\\
&\overset{(f)}{\leq} \hat{w}\mathbb{E}\left[\sum_{n\in\mathcal{D}_+}\widetilde{U}_n\right]^{\frac{1}{2l}}(TN)^{\frac{2l-1}{2l}}\\
&\overset{(g)}{\leq} \hat{w}\Big(\frac{T\epsilon}{d_{\min}}\Big)^{\frac{1}{2l}}(TN)^{\frac{2l-1}{2l}}\\
&= \hat{w}T\Big(\frac{\epsilon}{d_{\min}}\Big)^{\frac{1}{2l}}N^{\frac{2l-1}{2l}} \triangleq B_4^{(\epsilon)},
\end{aligned} \tag{34}$$

where step (a) is true because $\mu_n[k;t] = 1$ whenever at least one flow in region n in t^{th} time slot of time frame k achieves maximum channel rate; (b) uses the facts that the minimum number of flows in system given workload is $\widetilde{W}_n/w_{\max}$ and $\widetilde{W}_n[k;t] \geq \widetilde{W}[k] - T, \forall t = 1,2,3,\ldots,T$; (c) is true for some positive constant $\hat{w}$ due to the fact that $\lim_{\widetilde{W}_n\to\infty}\widetilde{W}_n(1-p_n^{\max})^{(\widetilde{W}_n-T)/2w_{\max}} = 0$; (d) uses Hölder's inequality for some number $l > 1$ such that $(1-p_n^{\max})^l \leq p_n^{\min}$; (e) uses the definition of l; (f) uses the fact that if all flows in region n do not have an available channel, then $\mathbb{E}[U_n[k;t]] \geq \mathbb{E}[S_n^*(p_n^{\min})^{\widetilde{\mathcal{N}}[k;t]}]$, and the fact that $U_n[k] \triangleq \sum_{t=1}^{T}U_n[k;t]$; (g) uses (32).

For the second term on the RHS of the identity (31), we have

$$\begin{aligned}
\mathbb{E}\left[\langle -\widehat{\mathbf{W}}_\perp, \widehat{\mathbf{U}}\rangle\right] &\overset{(a)}{\leq} \sqrt{\mathbb{E}\left[\|\widehat{\mathbf{W}}_\perp\|^2\right]\mathbb{E}\left[\|\widehat{\mathbf{U}}\|^2\right]}\\
&= \sqrt{\mathbb{E}\left[\|\widehat{\mathbf{W}}_\perp\|^2\right]\mathbb{E}\Big[\sum_{n\in\mathcal{D}_+}\widetilde{U}_n^2\Big]}\\
&\overset{(b)}{\leq} \sqrt{K_2\frac{\epsilon T^2}{d_{\min}}}
\end{aligned} \tag{35}$$

where step (a) uses Cauchy-Swartz inequality; (b) uses $\widetilde{U}_n \leq T$, inequality (32) and Proposition 4.

Substituting (33), (34), and (35) into (31), we have

$$\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{W}} + \boldsymbol{\nu} - T\widetilde{\mathbf{S}}^*\rangle\langle \mathbf{d}, \widetilde{\mathbf{U}}\rangle] \leq B_5^{(\epsilon)}, \tag{36}$$

where $B_5^{(\epsilon)} \triangleq B_4^{(\epsilon)} + \sqrt{N\nu_{\max}^2\frac{\epsilon T^2}{d_{\min}}} + \sqrt{K_2\frac{\epsilon T^2}{d_{\min}}}$.

In the end, we substitute (27), (28), (30) and (36) into (24) and bring the heavy-traffic parameter ϵ back to reinforce the result for proving Proposition 2 as follows:

$$\epsilon\mathbb{E}[\langle \mathbf{d}, \widetilde{\mathbf{W}}^{(\epsilon)}\rangle] \leq \frac{\langle \mathbf{d}^2, \boldsymbol{\sigma}^2\rangle}{2} + B^{(\epsilon)}, \tag{37}$$

where

$$\begin{aligned}
B^{(\epsilon)} \triangleq\ &\frac{1}{2}T\epsilon^2 + T\frac{\epsilon}{2\gamma_{\mathbf{d}}}\left(b_{\mathbf{d}}^2 + N\right) + \frac{1}{2}\epsilon T\sqrt{N} + \frac{1}{T}B_5^{(\epsilon)}\\
&+ \cot(\overline{\theta})\sqrt{\frac{K_2\epsilon}{\gamma_{\mathbf{d}}}\left(b_{\mathbf{d}}^2 + N\right)}.
\end{aligned} \tag{38}$$

As $\epsilon \downarrow 0$, we obtain the desired result.

8 Conclusion

In this paper, we considered the optimal scheduling design for wireless UAV networks in the presence of dynamic traffic flows, where each flow with a certain amount of service demand dynamically arrives at the system and departs once it completes its service. We developed a MaxWeight-type scheduling algorithm taking into account the sharp flow-level dynamics and showed that it not only achieves the maximum system throughput but also minimizes the mean workload in the heavy-traffic regime. Finally, extensive simulation results were provided to validate both the throughput and heavy-traffic optimality of our proposed algorithm. For future work, it would be interesting to jointly consider power control and scheduling, scheduling and caching.

Acknowledgments

We would like to thank anonymous reviewers' valuable comments. This work is supported in part by NSF grants: CNS-1717108 and CNS-1815563, and Natural Sciences and Engineering Research Council of Canada Discovery Grant: RGPIN-2017-05578.

References

[1] S. Aldhaher, P. D. Mitcheson, J. M. Arteaga, G. Kkelis, and D. C. Yates. Light-weight wireless power transfer for mid-air charging of drones. In *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pages 336–340. IEEE, 2017.

[2] A. Alsharoa, H. Ghazzai, A. Kadri, and A. E. Kamal. Energy management in cellular hetnets assisted by solar powered drone small cells. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2017.

[3] T. Bonald and M. Feuillet. On the stability of flow-aware CSMA. *Performance Evaluation*, 67(11):1219–1229, 2010.

[4] I. Bor-Yaliniz and H. Yanikomeroglu. The new frontier in ran heterogeneity: Multi-tier drone-cells. *IEEE Communications Magazine*, 54(11):48–55, 2016.

[5] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu. Efficient 3-d placement of an aerial base station in next generation cellular networks. In *Proc. IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016.

[6] S. Chandrasekharan, K. Gomez, A. Al-Hourani, S. Kandeepan, T. Rasheed, L. Goratti, L. Reynaud, D. Grace, I. Bucaille, T. Wirth, et al. Designing and implementing future aerial communication networks. *IEEE Communications Magazine*, 54(5):26–34, 2016.

[7] S.-F. Chou, Y.-J. Yu, and A.-C. Pang. Mobile small cell deployment for service time maximization over next-generation cellular networks. *IEEE Transactions on Vehicular Technology*, 66(6):5398–5408, 2017.

[8] K. L. Chung. *A course in probability theory*. Academic press, 2001.

[9] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72:311–359, 2012.

[10] Z. M. Fadlullah, D. Takaishi, H. Nishiyama, N. Kato, and R. Miura. A dynamic trajectory control algorithm for improving the communication throughput and delay in uav-aided networks. *IEEE Network*, 30(1):100–105, 2016.

[11] A. Fotouhi, M. Ding, and M. Hassan. Dynamic base station repositioning to improve spectral efficiency of drone small cells. In *Proc. IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Macao, China, December 2017.

[12] B. Galkin, J. Kibilda, and L. A. DaSilva. Deployment of uav-mounted access points according to spatial user locations in two-tier cellular networks. In *Proc. Wireless Days (WD)*, Toulouse, France, March 2016.

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

[14] S. Koulali, E. Sabir, T. Taleb, and M. Azizi. A green strategic activity scheduling for uav networks: A sub-modular game perspective. *IEEE Communications Magazine*, 54(5):58–64, 2016.

[15] B. Li, A. Eryilmaz, and R. Srikant. On the universality of age-based scheduling in wireless networks. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Hong Kong, China, April 2015.

[16] B. Li, A. Eryilmaz, and R. Srikant. Emulating round-robin in wireless networks. In *Proc. ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC)*, Chennai, India, 2017.

[17] B. Li, X. Kong, and L. Wang. Optimal load-balancing for high-density wireless networks with flow-level dynamics. In *Proc. ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC)*, Los Angeles, CA, 2018.

[18] 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.

[19] 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.

[20] J. Lu, S. Wan, X. Chen, Z. Chen, P. Fan, and K. B. Letaief. Beyond empirical models: Pattern formation driven placement of uav base stations. *IEEE Transactions on Wireless Communications*, 17(6):3641–3655, 2018.

[21] N. Lu, Y. Zhou, C. Shi, N. Cheng, L. Cai, and B. Li. Planning while flying: A measurement-aided dynamic planning of drone small cells. *IEEE Internet of Things Journal*, 6(2):2693–2705, 2018.

[22] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim. Placement optimization of uav-mounted mobile base stations. *IEEE Communications Letters*, 21(3):604–607, 2017.

[23] T. J. Nugent and J. T. Kare. Laser power for uavs. *Laser Motive White Paper-Power Beaming for UAVs, NWEN*, 2010.

[24] F. Tang, Z. M. Fadlullah, N. Kato, F. Ono, and R. Miura. Acpoca: Anticoordination game based partially overlapping channels assignment in combined uav and d2d-based networks. *IEEE Transactions on Vehicular Technology*, 67(2):1672–1683, 2018.

[25] L. Tassiulas. Scheduling and performance limits of networks with constantly varying topology. *IEEE Transactions on Information Theory*, 43:1067–1073, May 1997.

[26] 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.

[27] 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.

[28] 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.

[29] P. M. van de Ven, S. C. Borst, and L. Ying. Inefficiency of maxweight scheduling in spatial wireless networks. *Computer Communications*, 36(12):1350–1359, 2013.

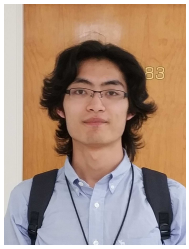
[30] Q. Wu, Y. Zeng, and R. Zhang. Joint trajectory and communication design for multi-uav enabled wireless networks. *IEEE Transactions on Wireless Communications*, 17(3):2109–2121, 2018.

[31] P. Yang, X. Cao, C. Yin, Z. Xiao, X. Xi, and D. Wu. Proactive drone-cell deployment: Overload relief for a cellular network under flash crowd traffic. *IEEE Transactions on Intelligent Transportation Systems*, 18(10):2877–2892, 2017.

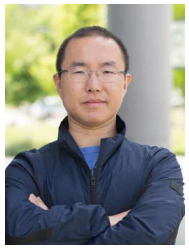
[32] S. Zhang, Y. Zeng, and R. Zhang. Cellular-enabled uav communication: A connectivity-constrained trajectory optimization perspective. *IEEE Transactions on Communications*, 67(3):2580–2604, 2018.

[33] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song. Joint trajectory and power optimization for uav relay networks. *IEEE Communications Letters*, 22(1):161–164, 2017.

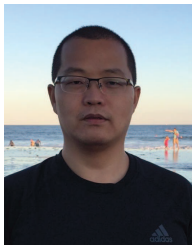
[34] Y. Zhou, N. Cheng, N. Lu, and X. S. Shen. Multi-uav-aided networks: aerial-ground cooperative vehicular networking architecture. *IEEE Vehicular Technology Magazine*, 10(4):36–44, 2015.



Xiangqi Kong (S'18) received his B.S. degree in Network Engineering in 2017, from Dalian University of Technology. He is currently a Ph.D. student in the Department of Electrical, Computer and Biomedical Engineering at the University of Rhode Island. His main research interest in wireless networks, virtual/augmented reality, edge computing, and system implementations.



Ning Lu (S'12-M'15) received the B.Eng. (2007) and M.Eng. (2010) degrees from Tongji University, Shanghai, China, and Ph.D. degree (2015) from the University of Waterloo, Waterloo, ON, Canada, all in electrical engineering. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Queen's University. Prior to joining Queen's University, he was an Assistant Professor in the Department of Computing Science at Thompson Rivers University, Kamloops, BC, Canada. From 2015 to 2016, he was a postdoctoral fellow with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. He also spent the summer of 2009 as an intern in the National Institute of Informatics, Tokyo, Japan. His current research interests include real-time scheduling, distributed algorithms, and reinforcement learning for wireless communication networks. He has published more than 40 papers in top IEEE journals and conferences, including IEEE/ACM Transactions on Networking, IEEE Journal on Selected Areas in Communications, ACM MobiHoc, and IEEE INFOCOM, etc.



Bin Li (S'12 - M'16) received his B.S. degree in Electronic and Information Engineering in 2005, M.S. degree in Communication and Information Engineering in 2008, both from Xiamen University, China, and Ph.D. degree in Electrical and Computer Engineering from The Ohio State University in 2014. Between 2014 and 2016, he worked as a Postdoctoral Researcher in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign. He is currently an Assistant Professor in the Department of Electrical, Computer, and Biomedical Engineering at the University of Rhode Island. His research interests include communication networks, virtual/augmented reality, fog computing, data centers, resource allocation and management, distributed algorithm design, queueing theory, and optimization theory.