Integrating LEO Satellites and Multi-UAV Reinforcement Learning for Hybrid FSO/RF Non-Terrestrial Networks

Ju-Hyung Lee, *Student Member, IEEE,* Jihong Park, *Member, IEEE,*
Mehdi Bennis, *Senior Member, IEEE,* and Young-Chai Ko, *Senior Member, IEEE*

***Abstract*—A mega-constellation of low-altitude earth orbit (LEO) satellites (SATs) and burgeoning unmanned aerial vehicles (UAVs) are promising enablers for high-speed and long-distance communications in beyond fifth-generation (5G) systems. Integrating SATs and UAVs within a non-terrestrial network (NTN), in this article we investigate the problem of forwarding packets between two faraway ground terminals through SAT and UAV relays using either millimeter-wave (mmWave) radio-frequency (RF) or free-space optical (FSO) link. Towards maximizing the communication efficiency, the real-time associations with orbiting SATs and the moving trajectories of UAVs should be optimized with suitable FSO/RF links, which is challenging due to the time-varying network topology and a huge number of possible control actions. To overcome the difficulty, we lift this problem to multi-agent deep reinforcement learning (MARL) with a novel action dimensionality reduction technique. Simulation results corroborate that our proposed SAT-UAV integrated scheme achieves 1.99x higher end-to-end sum throughput compared to a benchmark scheme with fixed ground relays. While improving the throughput, our proposed scheme also aims to reduce the UAV control energy, yielding 2.25x higher energy efficiency than a baseline method only maximizing the throughput. Lastly, thanks to utilizing hybrid FSO/RF links, the proposed scheme achieves up to 62.56x higher peak throughput and 21.09x higher worst-case throughput than the cases utilizing either RF or FSO links, highlighting the importance of co-designing SAT-UAV associations, UAV trajectories, and hybrid FSO/RF links in beyond-5G NTNs.**

***Index Terms*—LEO satellite, UAV, non-terrestrial network, hybrid FSO/RF, multi-agent deep reinforcement learning.**

I. Introduction

Witnessing the recent developments in non-terrestrial network (NTN) [1], [2], we are at the cusp of a communication revolution where space is envisaged to meet the ground through the sky [3]–[5]. On the one hand, space is emerging as the new frontier in beyond fifth-generation (5G) communication, through which the high-speed wireless coverage is extended to remote areas even in the ocean [6]. This trend has been encouraged by the recent launches of low-altitude earth orbit (LEO) satellite (SAT) mega-constellations [7]–[9]. In particular, SpaceX's 713 Starlink SATs have already been provisioning $>$ 100 Mbps data rate with $<$ 20 ms latency for ground terminals [10] while orbiting at an altitude of 550 km [11]. The data rate is comparably fast, and the latency is even lower than wired connections thanks to the direct links between ground terminals and overflying SATs. Such SAT-assisted systems have great potential particularly in enabling low-latency and long-range communications [12], [13]. The key challenge is to overcome the time-varying SAT network topology for improving communication efficiency.

On the other hand, the sky is a burgeoning frontier wherein unmanned aerial vehicles (UAVs), ranging from small drones to high altitude platforms (HAPs), can assist terrestrial communication systems in 5G and beyond [14]–[16]. There are a variety of UAV-assisted communication applications such as offloading ground terminal traffic in urban areas [17], providing emergency networks for disaster sites [18], and extending the wireless coverage to rural areas (e.g., Google's *Loon* [19], Facebook's *Aquila* [20], and HAPSMobile's *HAWK30* [21]). The benefits of these applications are rooted in the flexible maneuvering capability of UAVs. However, the coverage per UAV is smaller than SATs due to the low altitude, limiting the scalability and applicability within relatively short-range communications.

Motivated by the aforementioned complementary advantages of SATs and UAVs, in this article we study the problem of forwarding packets between two faraway ground terminals through multiple SATs and UAVs, as depicted in Fig. 1. The long-distance communications therein are enabled by the multiple orbital lanes of SATs while the communication efficiency is improved by controlling UAVs relaying the inter-orbit traffic. To this end, for the given SAT orbiting dynamics, UAVs optimize their locations and determine the associations with SATs and ground terminals in real time, so as to maximize the long-term average end-to-end data rate while minimizing the UAV control energy. To cope with the time-varying SAT topology, a novel multi-agent deep reinforcement learning (MARL) framework is applied for UAV control and associations. Furthermore, to take into account different SAT and UAV altitudes and channel characteristics, free-space optical (FSO) and millimeter wave (mmWave) radio frequency (RF) communication links are considered for maximizing the data rates. More details on the backgrounds and contributions of this work are elucidated in the following subsections.

A. Related Works

1) Non-Terrestrial Network: Ground-based wireless connectivity has already been extending towards the sky by integrating UAVs. As opposed to fixed ground base stations, these

J.-H. Lee, and Y.-C. Ko are with the School of Electrical and Computer Engineering, Korea University, Seoul, Korea (Email: leejuhyung@korea.ac.kr; koyc@korea.ac.kr).

†J. Park was with the University of Oulu, Finland, and is now with the School of Information Technology, Deakin University, Geelong, VIC 3220, Australia (Email: jihong.park@deakin.edu.au).

‡M. Bennis is with the Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland (Email: mehdi.bennis@oulu.fi).

Part of this work is to be presented in 2020 IEEE Global Communications Conference (GLOBECOM).

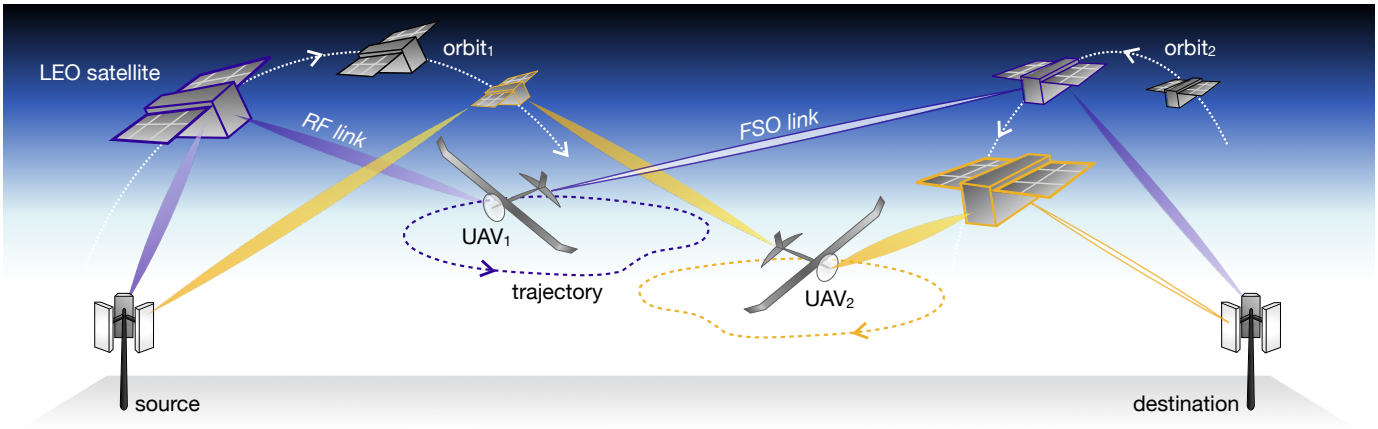


Figure 1: An illustration of an SAT-UAV integrated NTN wherein packets are forwarded between ground source and destination terminals through the two (purple and yellow) paths formed by SAT and UAV relayins utilizing hybrid FSO/RF (filled/void) links.

UAV terminals can be mobile and flexibly deployed, provisioning large-scale three-dimensional (3D) wireless connectivity. Many studies, for instance [22]–[24], have carried out the optimization for multi-UAV communication. For instance in [22], network throughput has been maximized by optimizing the association and power for UAVs allocation. Besides in [23], latency has been minimized by optimizing the cell association and multi-UAV location.

Likewise, towards provisioning high-speed global Internet access, SAT constellation is currently under heavy investigation by the industry (such as SpaceX's *Starlink* [11], *OneWeb* [7], Amazon's *Kuiper* [8], and *Telesat* [9]) and academia. Recent research works [12], [13] have advocated that exploiting SAT relays can achieve faster communication for long distances $>$ 3000 km compared to terrestrial optical fiber links. It is noted that in [12], lower latency performance can be achieved through additional fixed ground relays in-between the SAT relays.

NTN, including UAVs and SATs, has been studied in academia and formally organized by 3GPP [25]. However, existing works have focused only on NTN with UAVs or NTN with SATs, even though NTN deployment scenarios include both space and airborne platforms. These two aerial platforms have been separately investigated, in contrast to our work in which SAT and UAV relays are jointly optimized.

2) Deep Reinforcement Learning: Towards improving the communication efficiency of NTNs, a variety of optimization techniques have been applied, such as successive convex approximation (SCA) method [26], [27], block coordinate descent (BCD) method [28], [29], and graph theoretical [30] method. These approaches, however, require the global knowledge of the system parameters, which is often not feasible in practice. Furthermore, these methods commonly incur huge computing overhead and/or rely on multiple approximations to tackle the non-convex nature of the problems.

To overcome the aforementioned limitations, several recent works have proposed deep reinforcement learning (DRL) based solutions [31], [32], a machine learning based decision-making framework. In [32], DRL for multi-agent, so-called multi-agent deep reinforcement learning (MARL), is proposed for the distributed trajectory design of cooperative multi-UAV. While DRL has attracted growing attention in designing NTNs, to our best knowledge, the problem of jointly designing SAT associations and UAV path-planning (See Fig. 1) has not been studied in this literature.

3) FSO/RF Link: Unlike terrestrial networks, wireless communication systems operating at very high frequencies, such as mmWave and FSO communications, are widely considered in NTN. For instance, *Aquila* [20] considered FSO for ground-to-air and air-to-air links, and *Starlink* [11] considered both RF and FSO for the inter-satellite links (ISLs). In addition to industrial projects, academia have expressed a significant interest in FSO-based NTN, by leveraging the inherent characteristics of FSO communication, e.g., unlicensed broad-spectrum, immunity to electromagnetic interference, and security [33]–[35]. Nevertheless, few studies [36], [37] have presented an analysis of the possible link types for NTN, and those few studies have not considered time-varying scenarios of SATs.

Motivated by the potential of LEO SAT-based internet constellation, our previous works of [5] investigated SAT supported by high-altitude platform (HAP) relaying with RF links. The previous work, however, only considered a single UAV for only one selected SAT. Also, there was no consideration of other link options of SAT-UAV communication. Accordingly, several questions still remain regarding whether integrating multi space-borne vehicles and multi airborne vehicle is feasible and efficient, and whether RF or other link candidates are effective in this scenario. In this work, we answer these challenging questions by jointly optimizing the SAT associations and UAV locations in time-varying multi-topologies, using a novel MARL approach of distributed-Actor and centralized-Critic. Besides, we present key insights concerning which relaying scheme is efficient and which link candidate is a better option for the NTN scenario.

B. Contributions and Paper Organization

The main contributions of this work are summarized as follows.

- A novel problem has been formulated (P1 in Sec. III), which jointly optimizes multi-SAT association and multi-

UAV trajectory under a time-varying network topology, so as to maximize the system throughput of long-range non-terrestrial communication while minimizing the energy consumption of the system. Considering RF and FSO links, we examine the performance of both link candidates for NTN. To the best of our knowledge, this is the first work taking into account the joint association of SATs in multi-orbit, and the mobility management for multi-UAV, in the context of non-terrestrial communication.

- A MARL based solution has been proposed (see Sec. IV). Dealing with the scenario of distributed multi-agents (e.g., multi-SAT and multi-UAV), we propose a network structure of distributed-Actor and centralized-Critic. To cope with a large action space due to time-varying network topology, a novel action dimension reduction technique has been proposed, which focuses only on a couple of SATs proximal to Src and Dst.
- In the optimization problem, we examine the feasibility and the performance of the LEO SAT mega-constellation based NTN, with the following perspectives; Firstly, considering the end-to-end (E2E) throughput, the SAT-based NTN is investigated and propose possible improvements. Besides, the possible link types for NTN (e.g., RF, FSO links) are analyzed, and the E2E throughput of each candidate link is compared in our NTN scenario.
- Numerical results corroborate that the ground-relaying achieves an average system throughput 2.883x higher than a baseline of SAT-Only (which uses only SAT-based relaying). Furthermore, the mobile UAV relaying obtains 1.988x higher than the case of fixed ground relaying, while the case in which the association is optimized acquires 1.988x higher throughput than otherwise. This suggests using additional relays in-between SATs to provide high-throughput wide area networking. Furthermore, this highlights the importance of SAT association and UAV mobility management in enabling high-throughput non-terrestrial communication.

The remainder of this article is organized as follows. In Sec. II, the SATs-UAVs assisted NTN architecture and hybrid FSO/RF channel model are presented. In Sec. III, the energy efficiency maximization problem, in which the E2E system throughput is maximized while the energy consumption of UAV relay is minimized, is formulated. In Sec. IV, a MARL based solution is proposed. In Sec. V, simulation results are provided, followed by concluding remarks in Sec. VI.

II. System Model

The network under study consists of two SAT constellations rotating around a given orbital lane and additional UAV relay terminal in-between SATs, as illustrated in Fig. 1. In particular, we consider a multi-hop communication link forwarding packets from source terminal (Src) to destination terminal (Dst), via SAT and UAV relays. For the multi-hop communication, hybrid FSO/RF communication links are considered for the ground-to-air channel as well as the air-to-air channel.

We denote a set of SAT constellation network $\mathcal{I}_k = \{i_k = 1, 2, \ldots, I\}, k = 1, 2, \ldots K$ in an orbital lane k. A set of UAVs, which relay between the SAT constellation network, is denoted as $\mathcal{J} = \{j = 1, 2, \ldots, J\}$. Two terrestrial terminals, such that one transmits while another receives, are denoted by $\mathcal{S}$ and $\mathcal{D}$, respectively. Particularly, $\mathcal{S}$ refers to a Src and $\mathcal{D}$ refers to a Dst.

Based on three dimensional Cartesian coordinates for the location of the terminals, we assume that the backhaul terminal and the terrestrial base station are located at position $\mathbf{q}_{\mathcal{S}}$ and $\mathbf{q}_{\mathcal{D}}$, respectively, while SAT i orbits at the fixed altitude of H_{L} with constant speed $\mathbf{v}_{\mathrm{L}}^{k}$ following a predetermined orbital lane k. On the orbital lane, SATs are spaced at equal intervals and circulated. Particularly, we consider 22 SATs in the orbital lane and the orbital lane circumference. In addition, we consider that UAV j flies horizontally in the xy-plane with a fixed altitude H_{U}. The time-varying coordinate of the UAV-relay node can be denoted in [km] as $\mathbf{q}_{\mathrm{U}}^{j}(t) = [x(t), y(t), H_{\mathrm{U}}]^T \in \mathbb{R}^3$ for $0 \leq t \leq T$. Here, the UAVs serve in-between the two orbital lanes of SATs as additional relay terminals.

To obtain a more tractable optimization problem, we apply the discrete linear state-space approximation similarly to [26]. Based on the time step size δ_t, time T (or t) and time slot N (or n) can be determined according to $T = \delta_t \cdot N$ (or $t = \delta_t \cdot n$). Accordingly, the position of UAV j, $\mathbf{q}_{\mathrm{U}}^{j}(t)$ and velocity $\mathbf{v}_{\mathrm{U}}^{j}(t)$ can be well characterized by the discrete-time UAV position vector $\mathbf{q}_{\mathrm{U}}^{j}[n] = \mathbf{q}_{\mathrm{U}}^{j}(n\delta_t)$ as well as the velocity vector $\mathbf{v}_{\mathrm{U}}^{j}[n] = \mathbf{v}_{\mathrm{U}}^{j}(n\delta_t)$ and the acceleration vector $\mathbf{a}_{\mathrm{U}}^{j}[n] = \mathbf{a}_{\mathrm{U}}^{j}(n\delta_t)$ for $n = 0, 1, \cdots, N+1$. Therefore, the discrete state of UAV j can be expressed as

$$\mathbf{v}_{\mathrm{U}}^{j}[n+1] = \mathbf{v}_{\mathrm{U}}^{j}[n] + \mathbf{a}_{\mathrm{U}}^{j}[n]\delta_t, \ n = 0, \cdots, N, \tag{1}$$

$$\mathbf{q}_{\mathrm{U}}^{j}[n+1] = \mathbf{q}_{\mathrm{U}}^{j}[n] + \mathbf{v}_{\mathrm{U}}^{j}[n]\delta_t + \frac{1}{2}\mathbf{a}_{\mathrm{U}}^{j}[n]\delta_t^2, n = 0, \cdots, N, \tag{2}$$

where $\mathbf{q}_{\mathrm{U}}^{j}[0] = \mathbf{q}_{\mathrm{U,I}}^{j}$ is the initial positions of UAV j, and $\mathbf{v}_{\mathrm{U}}^{j}[0] = \mathbf{v}_{\mathrm{U,I}}^{j}$ is the initial velocity of UAV j.

On the other hand, the discrete state of SAT i in a certain constellation k can be expressed as

$$\mathbf{q}_{\mathrm{L}}^{i_k}[n+1] \ = \ \mathbf{q}_{\mathrm{L}}^{i_k}[n] + \mathbf{v}_{\mathrm{L}}^{k}\delta_t, \ n = 0, \cdots, N, \tag{3}$$

where $\mathbf{q}_{\mathrm{L}}^{i_k}[0] = \mathbf{q}_{\mathrm{L,I}}^{i_k}$ is the initial position of SAT i in an orbit plane k, and $\mathbf{v}_{\mathrm{L,I}}^{k}$ is the orbital speed in an orbit plane k which is calculated by orbital period. Note that SAT i in orbit k returns to the original position after the orbital period, i.e., $\mathbf{q}_{\mathrm{L}}^{i_k}[N+1] = \mathbf{q}_{\mathrm{L}}^{i_k}[0]$, since SAT circulates following the predetermined orbit plane k.

A. UAV Energy Consumption

In the UAV system, the energy consumption model is key for optimizing the service time, as it is not able to recharge propulsion fuel or electric power during a flight. We consider the propulsion energy for a flight as E_f and the communication energy for a signal processing as E_c. Note that in the energy consumption model, E_c is known to be much smaller than E_f in practical scenarios (i.e., $E_f \gg E_c$) [26]. Thus, we simply use total energy as $E_T \simeq E_f$. Following [26], [38], the power consumption for a fixed-wing type UAV j in [Watt], $P_j[n]$, is expressed as (4). Note that c_1 and c_2 are two parameters related to the effect of aircraft weight, its

$$P_j(\mathbf{v}_j[n],\mathbf{a}_j[n]) = c_1\|\mathbf{v}_j[n]\|^3 + \frac{c_2}{\|\mathbf{v}_j[n]\|}\left(1+\frac{\|\mathbf{a}_j[n]\|^2 - \frac{(\mathbf{a}^T[n]\mathbf{v}_j[n])^2}{\|\mathbf{v}_j[n]\|^2}}{g^2}\right) + \frac{m}{2\delta_t}\left(\|\mathbf{v}[N+1]\|^2 - \|\mathbf{v}[0]\|^2\right), \text{ [Watt]}. \quad (4)$$

wing area, and the air density. Here, g is the acceleration due to gravity (9.8 m/s^2) and m is the mass of the UAV.

Based on the power model, the energy consumption of UAV j for time slot n with given time step size δ_t is expressed as

$$E_j[n] = P_j[n]\delta_t \text{ [Joule]}. \quad (5)$$

B. FSO/RF Channel

1) RF link: Since we consider channel characteristics of ground-to-air link and air-to-air link, we assume LoS links without Doppler effect as in [26]. Therefore, the deterministic propagation models are adopted under the position of UAV and attenuation conditions. The channel gain of RF link h_{RF} between each terminal at a link distance d can be expressed as $h_{\mathrm{RF}} = \sqrt{\beta_0/d^2}$, where β_0 represents the received power at the reference distance $d_0 = 1$ [m]. Accordingly, the transmission rate in [bps] for the slot n can be expressed as

$$C_{\mathrm{RF}} = B_{\mathrm{RF}}\log_2\left(1+\frac{\gamma_0}{\|d\|^2}\right) \text{ [bps]}, \quad (6)$$

where B_{RF} represents the RF bandwidth, and $\gamma_0 = \frac{\beta_0 \cdot P}{\sigma^2_{\mathrm{RF}}}$ indicates the reference SNR with constant transmission power P and noise variance σ^2_{RF}.

2) FSO link: Based on the Beer-Lambert law, which represents the signal attenuation of optics, the channel gain for a FSO link at a link distance d can be expressed as [39]

$$h_{\mathrm{FSO}} = e^{-\beta\cdot\|d\|}, \quad (7)$$

where $\beta_{\mathrm{dB}} = \frac{3.91}{V}\left(\frac{\lambda}{550\text{ nm}}\right)^{-p}[\mathrm{dB/km}]$ value depends on the wavelength λ assumed to be 1550 [nm] in this paper, V is the visibility in [km], and the size distribution coefficient p determines by Kim model [40]. Note that $\beta = \frac{\beta_{\mathrm{dB}}}{10^4\log_{10}e}[\mathrm{m}^{-1}]$.

While the capacity of FSO is unknown in closed-form, the capacity bounds of FSO have been proposed in several papers. In this paper, the lower bound of FSO capacity introduced in [39] is used for describing the data rate of FSO link. The average optical SNR (ASNR) is denoted as $\gamma^2_{\mathrm{FSO}} = \frac{\varepsilon^2}{\sigma^2_{\mathrm{FSO}}}$ in which ε and σ^2_{FSO} are the average optical power and noise variance for FSO, respectively. The parameters related to ASNR, k_1 and attenuation condition, k_2 are formulated as [39]

$$k_1 = \begin{cases} \frac{e^{2\alpha\mu^*}}{2\pi e}\left(\frac{1-e^{-\mu^*}}{\mu^*}\right)^2\frac{\gamma^2_{\mathrm{FSO}}}{\alpha^2} & \text{if } 0<\alpha<\frac{1}{2}, \\ \frac{\gamma^2_{\mathrm{FSO}}}{2\pi e\alpha^2} & \text{if } \frac{1}{2}<\alpha<1, \end{cases} \quad (8)$$

$$k_2 = 2\beta. \quad (9)$$

Here, μ is the free parameter, which refers to the solution of $\alpha = \frac{1}{\mu^*} - \frac{e^{-\mu^*}}{(1-e^{-\mu^*})}$ in which the average-to-peak ratio (APR) is set to $\alpha = \frac{\varepsilon}{\Lambda}$ where Λ is peak optical power.

With the channel gain for a FSO link in (7), the FSO achievable rate in bits/second [bps] for the time slot n can be expressed as

$$C_{\mathrm{FSO}} = \frac{B_{\mathrm{FSO}}}{2}\log_2\left(1+k_1e^{-k_2\cdot\|d\|}\right) \text{ [bps]}. \quad (10)$$

Note that B_{FSO} is the bandwidth of FSO, and $\overline{\gamma}^2_{\mathrm{FSO}} = h^2_{\mathrm{FSO}}\cdot\gamma^2_{\mathrm{FSO}}$ represents the received ASNR.

3) Hybrid FSO/RF link: In the multi-hop communication scenario, we employ hybrid FSO/RF communication for each links. The hybrid FSO/RF method switches to a link with a better channel condition among the RF link and the FSO link. Accordingly, the achievable rate of the hybrid FSO/RF method can be expressed as [41]

$$C = \max\{C_{\mathrm{FSO}}, C_{\mathrm{RF}}\} \text{ [bps]}. \quad (11)$$

We point out here that the reasons for considering the hybrid FSO/RF link are as follows; i) The existing NFP projects, such as Starlink [11] and Aquila [20], consider both RF and FSO communication for the air-to-ground and air-to-air links, such as inter SAT-UAV link. However, only few studies focus on the achievable rate of RF and FSO link in the NTN scenario. We consider the hybrid FSO/RF link to answer the question of which link (RF link or FSO link) is more advantageous. Adopting the hybrid FSO/RF link enables the comparison study of these two link options which will be given in our numerical results. ii) Unlike FSO, mm-waves are not attenuated by fog, while mm-waves are highly attenuated by water molecules, such as rainfall [41]. To combat the advantages of both mm-waves and FSO, the hybrid FSO/RF is considered. The system switches to FSO during rainy conditions when mm-wave transmission is not favorable due to high rain attenuation and switches to mm-waves during foggy conditions. A hybrid FSO/RF link itself can be considered as a potential alternative solution to overcome the link degradation under bad weather conditions in the air-to-ground and air-to-air links [42]. We examine the hybrid FSO/RF link as a promising vertical backhauling/fronthauling link configuration.

III. Energy-Efficiency Maximization of a Satellite-UAV Integrated NTN

This section addresses the energy-efficiency (EE) maximization of the UAV-assisted LEO SAT constellation network. The connected LEO SATs forward information between the source and destination as a relay node, during SATs circulate rapidly in the predetermined orbital lane. For efficient relaying, appropriate node-to-node association design are necessary. Besides, as we consider the UAVs as additional mobile relay nodes in-between the LEO SATs, proper movement control for UAV is also necessary. Therefore, we optimize the association for SATs and the trajectory of UAVs.

A. Multi-Hop Communication in NTN

As shown in Fig. 1, multiple SAT and UAV configure the multi-hop communication. In our NTN scenario, information is transmitted from $\mathcal{S}$ to $\mathcal{D}$ via SAT relay nodes and via UAV relay nodes. We consider that two distant terminals $\mathcal{S}$ and $\mathcal{D}$ are connected through as many links as the number of UAV relays, J. That is, multiple links support the E2E communication between Src and Dst, if $J \neq 1$. In addition, we consider two orbital lane, i.e., $K = 2$.

According to the network scenario, the instantaneous achievable rates of each link for UAV j in time slot n in bps is given by

$$R_{\mathcal{S},i_1^j}[n] \leq C_{\mathcal{S},i_1^j}[n], \tag{12}$$

$$R_{i_1^j,j}[n] \leq \min\{C_{i_1^j,j}[n],\ R_{\mathcal{S},i_1^j}[n]\}, \tag{13}$$

$$R_{j,i_2^j}[n] \leq \min\{C_{j,i_2^j}[n],\ R_{i,j}[n]\}, \tag{14}$$

$$R_{i_2^j,\mathcal{D}}[n] \leq \min\{C_{i_2^j,\mathcal{D}}[n],\ R_{j,i_2^j}[n]\}, \tag{15}$$

where $R_{\mathcal{S},i_1^j}[n]$, $R_{i_1^j,j}[n]$, $R_{j,i_2^j}[n]$, and $R_{i_2^j,\mathcal{D}}[n]$ denote the achievable rate of $\mathcal{S}-i_1^j$, i_1^j-j, $j-i_2^j$, and $i_2^j-\mathcal{D}$, respectively, for time slot n. Note that to clarify which UAV j is associated with SAT i on orbit k, the superscript j is used in $i_k^j \in \mathcal{I}_k$, from now on. In the above equations, the transmission rate of each link for UAV j in bps are represented, respectively, as

$$C_{\mathcal{S},i_1^j}[n] = w_{\mathcal{S},i_1^j}[n]C(d_{\mathcal{S},i_1^j}[n]), \tag{16}$$

$$C_{i_1^j,j}[n] = w_{i_1^j,j}[n]C(d_{i_1^j,j}[n]), \tag{17}$$

$$C_{j,i_2^j}[n] = w_{j,i_2^j}[n]C(d_{j,i_2^j}[n]), \tag{18}$$

$$C_{i_2^j,\mathcal{D},j}[n] = w_{i_2^j,\mathcal{D}}[n]C(d_{i_2^j,\mathcal{D}}[n]), \tag{19}$$

where $w_{\mathcal{S},i_1^j}[n]$, $w_{i_1^j,j}[n]$, $w_{j,i_2^j}[n]$, and $w_{i_2^j,\mathcal{D}}[n]$ represents the association of $\mathcal{S}-i_1^j$, i_1^j-j, $j-i_2^j$, and $i_2^j-\mathcal{D}$, respectively. Note that we consider the only one SAT is selected between i_1^j-j and between $j-i_2^j$ in each orbit k for time slot n, so that it holds $w_{\mathcal{S},i_1^j} = w_{i_1^j,j}$ and $w_{j,i_2^j} = w_{i_2^j,\mathcal{D}}$. In the above operations, the link distance for each link for UAV j is expressed as

$$d_{\mathcal{S},i_1^j}[n] = \|\mathbf{q}_{\mathrm{L}}^{i_1^j}[n] - \mathbf{q}_{\mathcal{S}}\|, \tag{20}$$

$$d_{i_1^j,j}[n] = \|\mathbf{q}_{\mathrm{U}}^{j}[n] - \mathbf{q}_{\mathrm{L}}^{i_1^j}[n]\|, \tag{21}$$

$$d_{j,i_2^j}[n] = \|\mathbf{q}_{\mathrm{L}}^{i_2^j}[n] - \mathbf{q}_{\mathrm{U}}^{j}[n]\|, \tag{22}$$

$$d_{i_2^j,\mathcal{D}}[n] = \|\mathbf{q}_{\mathcal{D}} - \mathbf{q}_{\mathrm{L}}^{i_2^j}[n]\|. \tag{23}$$

Here, the position of SATs ($\mathbf{q}_{\mathrm{L}}^{i_1^j}[n]$ and $\mathbf{q}_{\mathrm{L}}^{i_2^j}[n]$) and the position of UAV ($\mathbf{q}_{\mathrm{U}}^{j}[n]$) are time-varying.

1) Practical Considerations for Multi-Hop Communication: In the multi-hop communication, there are two practical considerations. Firstly, it is necessary to consider association overlap, since multiple links by UAV relays are considered in the E2E multi-hop communication. The association overlap happens when multiple UAV associate to the same SAT. When occurring association overlap, the achievable rate of link should be decreasing, as the resource (e.g., bandwidth, transmission power) is limited for each link and the interference issue may arise. Thus, we assume that each node can use limited bandwidth and the limited bandwidth is equally divided for association overlapped links. For instance, if $w_{i_1^1,1}[n] = w_{i_1^2,2}[n]$, $C_{i_1^1,1}[n] = \frac{1}{2}w_{i_1^j,1}[n]C(d_{i_1^1,1}[n])$ and $C_{i_1^1,2}[n] = \frac{1}{2}w_{i_1^2,2}[n]C(d_{i_1^2,2}[n])$.

Secondly, it is necessary to consider communication overhead between UAVs. The communication between UAVs is needed for avoiding a collision between UAVs and for designing the actions of multi UAV in a cooperative manner which guarantee better network performance. To deal with the communication overhead issue, we consider the latency between UAVs, τ, which is derived by propagation delay.

2) End-to-End summarized Throughput for Multi-Hop Communication: Without losing generality, assuming decode-and-forward (DF) relaying protocol without buffer, the constraints of (12)-(15) evolve to $\min\{R_{\mathcal{S},i_1^j}[n], R_{i_1^j,j}[n], R_{j,i_2^j}[n], R_{i_2^j,\mathcal{D}}[n]\}$. Accordingly, the instantaneous E2E sum throughput for the multi-hop communication in time slot n can be approximated as

$$R_{\mathcal{S},\mathcal{D}}^{j}[n] \approx \min\{R_{\mathcal{S},i_1^j}[n], R_{i_1^j,j}[n], R_{j,i_2^j}[n], R_{i_2^j,\mathcal{D}}[n]\}. \tag{24}$$

B. Problem Formulation

For our problem's objective, we consider EE, which includes both the E2E network throughput and the UAVs' energy consumption. The following problem, P1, corresponds to the EE maximization under the constraints related to the practical condition of SAT and UAV. For mathematical convenience, we define the set of association as $\mathcal{W} = \{w_{i_1^j,j}[n], w_{j,i_2^j}[n], \forall n, i_k \in \mathcal{I}_k, k = 1, 2, \ldots, K, j \in \mathcal{J}\}$, and the set of UAV acceleration as $\mathcal{A} = \{\mathbf{a}_{\mathrm{U}}^{j}[n], \forall n, j\}$.

$$\begin{aligned}
(\mathrm{P1}) \quad \max_{\mathcal{W},\mathcal{A}} \quad & \mathrm{EE} \triangleq \frac{\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} R_{\mathcal{S},\mathcal{D}}^{j}[n]}{\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} E_j[n]} \\
\text{s.t.} \quad & (1)-(3), (12)-(15), \\
& w_{i_1^j,j}[n] \in \{0,1\},\ \forall n, i_1^j, j, && (25)\\
& w_{j,i_2^j}[n] \in \{0,1\},\ \forall n, i_2^j, j, && (26)\\
& \textstyle\sum_{i_1^j\in\mathcal{I}_1} w_{i_1^j,j}[n] = 1,\ \forall n, j, && (27)\\
& \textstyle\sum_{i_2^j\in\mathcal{I}_2} w_{j,i_2^j}[n] = 1,\ \forall n, j, && (28)\\
& \|\mathbf{a}_{\mathrm{U}}^{j}[n]\| \leq A_{\max},\ \forall n, j, && (29)\\
& \tau[n] \leq \tau_{\max},\ \forall n, && (30)
\end{aligned}$$

where N is the one orbital cycle of SAT, $A_{\max}$ denotes the maximum acceleration of UAV, and $\tau_{\max}$ indicates the maximum allowable delay between UAVs.

For the maneuverability of UAV, the equality constraints in (1) and (2) characterize the discrete-time model of UAV, i.e., the position of UAV $\mathbf{q}_{\mathrm{U}}^{j}[n]$, the velocity of UAV $\mathbf{v}_{\mathrm{U}}^{j}[n]$, as well as the acceleration of UAV $\mathbf{a}_{\mathrm{U}}^{j}[n]$. For the orbit of SATs, the equality constraint in (3) characterizes the discrete-time model of SAT, i.e., the position of SAT. In addition, the constraints in (12)-(15) represent the information-causality constraint in the multi-hop communication. The information-causality, that relay node can only forward the information which has been previously received from the source, is a inherent constraint of relay node. For the association of SAT-UAV, the constraints in

(27)-(28) represent that UAV j can be linked by at most one SAT in each orbit, while the constraints of (25)-(26) compose the association variables as integer variables. The constraint in (29) ensures that the acceleration of UAV is within acceptable acceleration. Besides, the constraint in (30) limits the latency for link between UAVs, as UAV should share the information of itself with other UAVs within a finite time.

For solving the optimization problem (P1), however, traditional convex optimization methods (such as SCA, BCD method) face several challenges. Firstly, the formulated problem of (P1), which designs the SAT-UAV association and finds the optimal trajectory of UAVs, is obviously non-convex and thus difficult to be directly solved. Furthermore, the variables of association and trajectory, which depends on the time-varying network topology (e.g., (12)-(15)), are challenging to be efficiently relaxed to convex. Secondly, traditional off-line optimization-based methods require the global knowledge of the system parameters, which is non-trivial to acquire in practice. Even given global knowledge of the system parameters, it is difficult to deal with the association overlap issue, such as $w_{i_1^1,1}[n] = w_{i_1^2,2}[n]$.

IV. Multi-Agent Reinforcement Learning for an Energy-Efficient Satellite-UAV Integrated NTN

To address the challenges for association and trajectory design in the time-varying NTN topology, we propose a novel MARL method, which only requires observable information of UAVs along with its flight as the input. We leverage the multi-agent Actor-Critic method, and further introduce the action dimension reduction technique to improve the learning convergence and performance. For using the MARL method, the objective, variables, and constraints of (P1) is needed to be mapped into a state, action, and reward function in the environment. However, the mapping of (P1) is non-trivial due to problematic learning convergence of the MARL method.

A. Problem Reformulation

Before applying a MARL method directly into (P1), we slightly change the original problem for the following fact; The value of the fractional objective in (P1) fluctuates easily according to the denominator that changes according to the dynamic state of UAV (e.g., $\mathbf{v}_j[n], \mathbf{a}_j[n]$), such that it may cause the learning convergence problem.

To this end, motivated by the Dinkelbach's algorithm, a widely-known method to solve large-scale fractional programming for which its optimality and convergence properties are established [43], we rewrite (P1) as

$$\begin{aligned}(\text{P1}^\star) \quad \max_{\mathcal{W},\mathcal{A}} \quad & \sigma_{\text{R}} \sum_{n=1}^{N} \sum_{j\in\mathcal{J}} R^j_{\mathcal{S},\mathcal{D}}[n] - \sigma_{\text{E}} \sum_{n=1}^{N} \sum_{j\in\mathcal{J}} E_j[n] \\ \text{s.t.} \quad & (1)-(3), (12)-(15), (25)-(29), \\ & d_{j,j'}[n] \le d_{\max}, \ \forall n, \end{aligned} \tag{31}$$

where σ_{R} and σ_{E} are the coefficient for the sum of E2E throughput and the sum of energy consumption of UAVs, respectively, and the distance between UAVs is denoted as

$$d_{j,j'}[n] = \|\mathbf{q}^j_{\text{U}}[n] - \mathbf{q}^{j'}_{\text{U}}[n]\|, \ j \in \mathcal{J}, j' \in \mathcal{J}/j, \tag{32}$$

which is limited by the maximum allowable distance between UAVs, $d_{\max}$.

Now, we can handle the weights for the E2E system throughput and the energy consumption of UAVs by adjusting σ_{R} and σ_{E} in (P1*). For example, the throughput performance, $\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} R^j_{\mathcal{S},\mathcal{D}}[n]$, can be mainly optimized by relatively increasing σ_{R}, or the energy (related to the service-time), $\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} E_j[n]$, can be mainly optimized by relatively increasing σ_{E}. The latency constraint in (30) is rewritten as (31). Note that the latency is derived by the propagation delay, as the distance between UAVs is quite far and the amount of information exchanged between UAVs is not considerable. We now describe how the formulated problem (P1*) is matched to the MARL approach in detail.

B. Markov Decision Process Modeling

In this subsection, we design our optimization problem of (P1*) into a Markov decision process (MDP) model. We note that MDP is a discrete-time stochastic control process, which is provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. In the MDP model, the policy corresponds to the probability of choosing an action according to the current state. The optimal policy π^* is the policy that contributes to the maximal long-term system reward. Our goal is to find π^* to maximize the average long-term system reward.

By the dynamic states of UAV j in (1)-(2) (e.g., $\mathbf{q}^j_{\text{U}}[n]$ and $\mathbf{v}^j_{\text{U}}[n]$) and the dynamic states of UAV j in (3) (e.g., $\mathbf{q}^{i_k}_{\text{L}}[n]$), it holds a Markov characteristics. As such, we can map the optimization problem (P1*) as MDP. Briefly, mapping of (P1*) into MDP model, each system utility in the objective function (such as E2E system throughput, consumption energy of UAVs) corresponds to reward function, while the optimization variables (e.g., $\mathcal{W}, \mathcal{A}$) corresponds to action space. A mapping of (P1*) into the environment, state, action, and reward functions in MDP model is presented in detail in the following subsections.

1) Environment: As shown in Fig. 2, the MARL framework for joint SAT-UAV multi-hop communication is based on the MDP model of (P1*), wherein multi-agents, corresponding to multiple UAVs, interact with an environment. In this scenario, the environment includes everything related to the multi-hop communication. At each time n, each UAV j, as a agent, observes a state $s_j[n]$ from the state space S, and accordingly takes an action $a_j[n]$ from the action space A selecting the association action set and the acceleration action set based on the policy π. Following the action, the state of the environment transitions to a new state $s_j[n+1]$ and the agent j receives a reward $r_j[n]$ which is determined by the E2E system throughput in the multi-hop communication and the energy consumption of UAVs.

2) State: In our system, the state observed by each UAV node for characterizing the environment consists of several parts: the position of SAT in the two orbital lane $\mathbf{q}^{i_k}_{\text{L}}[n] \in \mathbb{R}^{I\times 3}, i \in \mathcal{I}, k = 1,2$, the position of UAV j $\mathbf{q}^j_{\text{U}}[n] \in \mathbb{R}^{J\times 3}$, the link distance for each link $d_j[n] =$

$\{d_{\mathcal{S},i'_1}[n], d_{i'_1,j}[n], d_{j,i'_2}[n], d_{i'_2,\mathcal{D}}[n]\} \in \mathbb{R}^4$, the energy consumption of UAV j for the predetermined time $P_j[n]\delta_t$, and the time slot n. Thus, the state of agent j is given by

$$s_j[n] = \{\mathbf{q}_{\mathrm{L}}^{i_k}[n], \mathbf{q}_{\mathrm{U}}^{j}[n], \mathbf{v}_{\mathrm{U}}^{j}[n], d_j[n], P_j[n]\delta_t, n\}. \quad (33)$$

In the state space, n is used as a fingerprint, considering two methods proposed in [44] for stabilizing experience replay in the MARL; i) using a multi-agent variant of importance sampling to naturally decay obsolete data, and ii) conditioning each agent's value function on a fingerprint that disambiguates the age of the data sampled from the replay memory.

3) Action: The action space in our system includes two actions, i.e., associations $\mathcal{W}$ and accelerations $\mathcal{A}$. Firstly, with $w_{i_1^j,j}[n]$ and $w_{j,i_2^j}[n]$, the agent j choose one SAT to associate with among SATs on the orbital lane $k = 1$ and $k = 2$, respectively, for the association of $i_1^j - j$ and $j - i_2^j$. Note that we consider that the decision on $w_{\mathcal{S},i_1^j}[n]$ follows $w_{i_1^j,j}[n]$, and the decision on $w_{i_2^j,\mathcal{D}}[n]$ follows $w_{j,i_2^j}[n]$. To design the action set of association, $a_j^{\mathcal{W}}[n] \in \mathbb{R}^{I\times 2}, \forall j$, we use the one-hot encoding.

Secondly, for trajectory design of UAV, UAV chooses the acceleration action set in each time slot. While there are various ways to efficiently handle a continuous-state action, the most straightforward approach is discretizing it to form a finite-state. Hence, we consider the action set of acceleration, $a_j^{\mathcal{A}}[n] \in \mathbb{R}^{2\times(2D+1)}$, by uniformly discretizing between $-A_{\max}$ and $A_{\max}$ as follows.

$$\begin{aligned} a_j^{\mathcal{A}}[n] \in \{&-A_{\max}, \cdots, \frac{-A_{\max}}{D}, 0, \frac{A_{\max}}{D}, \cdots, A_{\max}; \\ &-A_{\max}, \cdots, \frac{-A_{\max}}{D}, 0, \frac{A_{\max}}{D}, \cdots, A_{\max}\}. \end{aligned} \quad (34)$$

Here, the acceleration action space can be managed by a discretization level $2D+1$ with a positive integer D. Accordingly, the action space is given by

$$a_j[n] = \{a_j^{\mathcal{W}}[n], a_j^{\mathcal{A}}[n]\}. \quad (35)$$

4) Reward: We design the reward function that will drive UAV to find optimal actions, which maximize the E2E system throughput while minimizing the energy consumption and the distance between UAVs. In the problem, system utilities are closely related to the E2E system throughput, the energy consumption of UAVs, and the distance between UAVs, in each time slot.

To reflect the expressions in reward function, we adopt the normalization function for x, $g(x) = \frac{x-\mu}{\sigma}$. The mean value for the E2E system throughput μ_{R} is the E2E system throughput of the baseline (i.e., SAT-Ground in Sec. V), the mean value for the energy consumption μ_{E} is the minimum energy for hovering [26], and the mean value for distance between UAVs μ_{D} is $d_{\max}$. The standard deviations for $g(R_{\mathcal{S},\mathcal{D}}[n])$, $g(\sum_{j\in\mathcal{J}} P_j[n]\delta_t)$, and $g(d_{j,j'}[n])$ are denoted as σ_{R}, σ_{E}, and σ_{D}, respectively, in which the output value of each function yields between -1 and 1. Note that σ_{R} and σ_{E} accord with c_{R} and c_{E}, respectively. In addition, we further use a *Relu* function of $h(x) = \max\{0, g(x)\}$, for the distance constraint in (31), such that the agent gets penalty when exceeding a threshold.

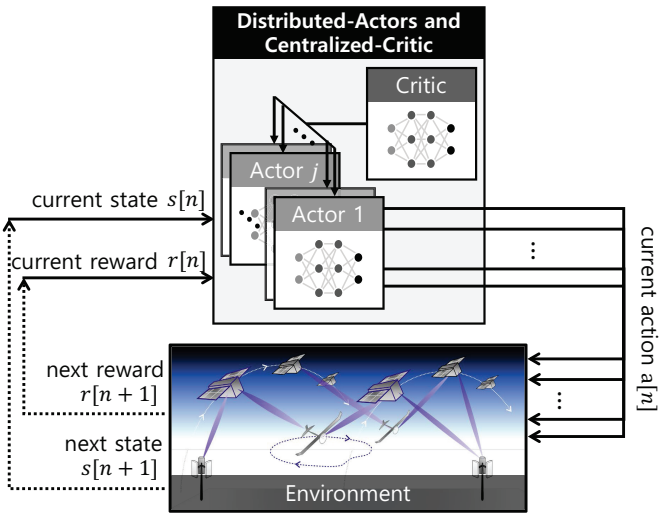


Figure 2: Proposed multi-agent Actor-Critic structure for vertical multi-hop communication in LEO SAT constellation network.

As such, the system reward in the n-th time slot induced by the current state $s[n]$ and action $a[n]$ is defined as

$$\begin{aligned} r[n] = g(\textstyle\sum_{j\in\mathcal{J}} R_{\mathcal{S},\mathcal{D}}[n]) - g(\textstyle\sum_{j\in\mathcal{J}} E_j[n]) \\ - h(\textstyle\sum_{j\in\mathcal{J}, j'\in\mathcal{J}/j} d_{j,j'}[n]). \end{aligned} \quad (36)$$

We note that the reward function remains positive if our network model outperforms a baseline; it will be a penalty, a negative reward, if not.

C. Centralized-Critic MARL

Following the standard RL settings, we consider an environment that is interacted with RL agents for a given number of discrete time steps. At each time step n, the agent j receives a state $s_j[n]$ and selects an action $a_j[n]$ from some set of possible actions A according to its policy π_θ, where π_θ is a mapping from states $s_j[n]$ to actions $a_j[n]$. In return, the agent receives the next state $s_j[n+1]$ and receives a scalar reward $r[n]$. The process continues until the agent reaches a terminal state after which the process restarts. The return $R[n] = \sum_{k=0}^{\infty} \gamma^k r[n+k]$ is the total accumulated return from time step n with discount factor $\gamma \in (0, 1]$. Multiple agents act in environments with the goal of maximizing their shared utility in a cooperative manner.

In our prior work [31], we considered a single UAV agent operating a deep-Q network (DQN) framework wherein each action of the agent is evaluated as the Q-value that is approximated using the output of a neural network (NN). While effective, DQN may not guarantee the convergence due to its off-policy learning nature, particularly under complicated tasks with large state and action dimensions [45]. By contrast, there exist policy gradient methods in which an NN's output approximates each policy [46], yet the training convergence is often too slow due to its ignoring the effort to find better actions associated with higher values. Alternatively, in this work we consider the Actor-Critic RL framework [47], which combines the benefits of both policy gradient and value-based

methods. In what follows, we first explain the basic Actor-Critic RL operations for a single agent, and then describe how to extend this to MARL settings, followed by presenting an action dimensionality reduction technique.

1) Single-Agent Actor-Critic RL: Actor-Critic framework comprises a pair of two NNs: an Actor NN seeking to take better actions to obtain higher rewards based on the policy gradient method; and its paired Critic NN aiming at approximating the value functions of the actions more accurately via the value-based method. We follow the training operations of the synchronous advantage Actor-Critic (A2C) framework [48], in which the Critic NN updates its model parameters ϕ according to the policy π_θ given by the Actor NN. Meanwhile, the Actor NN updates its model parameters θ according to the value functions $V^{\pi_\theta}(s[n];\phi)$ approximated by the Critic NN. Specifically, the Critic NN aims to minimize the loss function

$$L_{\text{Critic}}(\phi) = \kappa[n]^2, \tag{37}$$

where $\kappa[n] = r[n+1] + \gamma V^{\pi_\theta}(s[n+1];\phi) - V^{\pi_\theta}(s[n];\phi)$ is referred to as the advantage of an action [49]. The Critic NN model parameters are then updated as

$$\phi \leftarrow \phi + \beta_C \kappa[n] \nabla_\phi V^{\pi_\theta}(s[n];\phi), \tag{38}$$

where β_C is the learning rate of the Critic NN. Meanwhile, the Actor NN aims to minimize the loss function

$$L^j_{\text{Actor}}(\theta_j) = -\kappa[n] \log \pi(a_j[n]|s_j[n];\theta_j). \tag{39}$$

Hereafter, the index j identifies different actors for multi-agent scenarios, which can be omitted for a single agent case. Consequently, the Actor NN model parameters are updated as

$$\theta_j \leftarrow \theta_j + \beta_A \kappa[n] \nabla_{\theta_j} \log \pi(a_j[n]|s_j[n];\theta_j), \tag{40}$$

where β_A is the learning rate of the Actor NN.

2) Centralized-Critic MARL: MARL frameworks are broadly categorized into cooperative and non-cooperative methods [50]. In non-cooperative methods such as independent Q-learning [51], each agent is independently and simultaneously trained without sharing its learned knowledge. Following this principle, the aforementioned single-agent Actor-Critic framework can be extended to its non-cooperative MARL version wherein each agent runs a pair of Actor and Critic NNs without exchanging any information. However, from our experimental observations, such a non-cooperative MARL extension does not guarantee the convergence, probably due to the dynamic nature of orbiting SATs and moving UAV agents. Alternatively, based on a cooperative MARL principle in [52]–[54], we consider the centralized-Critic MARL framework wherein a single Critic NN guides the training of distributed Actor NNs that determine individual agents' policies, as visualized in Fig. 3. In this approach, while interacting with the environment, each agent j storing an Actor NN uploads its learned policy $\pi(a_j[n]|s_j[n];\theta_j)$, i.e., a pair of its action $a_j[n]$ and state $s_j[n]$, to the Critic NN located at a randomly chosen agent. The centralized-Critic NN takes the entire J agents' actions $a[n] = \{a_1[n], \ldots, a_J[n]\}$ and states $s[n] = \{s_1[n], \ldots, s_J[n]\}$ as the input, and produces the advantage $\kappa[n]$ as the output that is a function of the values $V^{\pi_\theta}(s[n], a[n])$ obtained via the temporal difference method [46]. Each agent downloads $\kappa[n]$, and thereby updates its Actor NN.

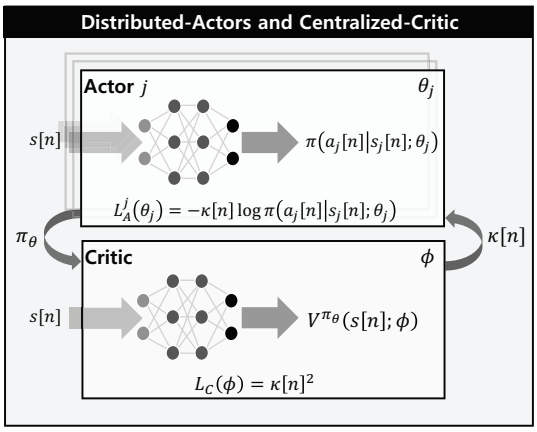


Figure 3: A schematic illustration of a centralized-Critic MARL wherein distributed Actor NNs exchange their learned policies with a single Critic NN.

3) Action Dimensionality Reduction: As identified in [45], a task with large numbers of discrete actions is intractable or even often impossible to train. Unfortunately, in our MARL settings, the space of association actions $a_{\mathcal{W}}$ is large due to the time-varying network topology with many orbiting SATs. To clarify how large the action space is, in what follows we describe the orbiting dynamics of SATs in an orbital lane. Since each SAT periodically completes an orbit, the position $\mathbf{q}_{\text{L}}^{i_k}[n] \in \mathbb{R}^{I\times 3}$ of an SAT in the k-th orbital lane is updated as

$$\mathbf{q}_{\text{L}}^{i_k}[n] = \mathbf{q}_{\text{L}}^{i_k}[n] \bmod c_{\text{E}}, \tag{41}$$

where mod represents the modulo operation, $c_{\text{E}} = 2\pi r_{\text{E}}$ denotes the circumstance of orbital lane (such as circumstance of Earth), and r_{E} denotes the radius of orbit. Note that by the operation in (41), SAT i in orbit k returns to the original position after the orbital period.

To ameliorate the problem incurred by the large action space in our MARL environment, we devise the following action dimension reduction method. The key idea is to utilize the fact that for maximizing the E2E sum throughput, only the SATs sufficiently close to Src or Dst are the candidates of optimal associations $a_j^{\mathcal{W}}[n]$. Thereby we can filter out distant SATs from Src or Dst, and focus on only a few SATs in the action space. Particularly, we focus only on a certain orbital line segment c_{C} which is located proximal to Src or Dst, instead of the entire orbital lane circumference c_{E}. The position of the selected SAT with c_{C} is thereby rewritten as follows

$$\mathbf{q}_{\text{L}}^{i_k^*}[n] = \mathbf{q}_{\text{L}}^{i_k^*}[n] \bmod c_{\text{C}}. \tag{42}$$

Here, c_{C} depends on the size of considered coordinate space. Since c_E is based on an Earth scale while c_C is based on a distance of Src-Dst, it is clear that $c_C \ll c_E$. As a result, the dimension of action space as well as the state space related to the number of SATs reduces from I to I', since we now only consider the position of selected SAT. Note that it holds $I' < I$, where I' denotes the maximum number of SATs that can be included in a given coordinate space which depends on

Table I: Simulation parameters.

Parameter	Value
Time slot size	$\delta_t = 10$
Number of time slots per episode	$N = 572$
Location of Src	$\mathbf{q}_{\mathcal{S}} = [0, 0, 0]^T$
Location of Dst	$\mathbf{q}_{\mathcal{D}} = [4000, 4000, 0]^T$
Altitude of SAT	$H_{\mathrm{L}} = 550$ [km] (from [11])
Velocity (Speed) of SAT1 ($k = 1$)	$\mathbf{v}^1_{\mathrm{L,I}} = [0, 7.59, 0]^T$ (7.59 [km/s])
Velocity of SAT2 ($k = 2$)	$\mathbf{v}^2_{\mathrm{L,I}} = [0, -7.59, 0]^T$
Radius of an orbit	$r_{\mathrm{E}} = 6921$ [km]
Circumference of an orbital lane	$c_{\mathrm{E}} = 43486$ [km]
Number of SATs per orbital lane	$I = 22$
Inter-SAT distance	1977 [km]
Length an orbital line segment	$c_{\mathrm{C}} = 6000$ [km]
Number of SATs per orbital line segment	$I' = 3$
Altitude of UAV	$H_{\mathrm{U}} = 50$ [km] (from [25])
Maximum acceleration of UAV	$A_{\max} = 5$ [m/s^2]
Discretization level for $a^{\mathcal{A}}_j[n]$	$D = 5$
Bandwidth for RF and FSO links	$B_{\mathrm{RF}} = B_{\mathrm{FSO}} = 10^9$ [Hz]
Reference SNR ($d = 1$ [m])	$\gamma_0 = 10^9$
Visibility	$V = 15$ [km]
ASNR	$\gamma_{\mathrm{FSO}} = 25$ [dB] ($\alpha = 0.1$)

c_{C}. For instance, the action space of $a^{\mathcal{W}}[n]$ reduces to $\mathbb{R}^{I' \times 2}$ and the state space of $\mathbf{q}^{i_k}_{\mathrm{L}}[n] \in I$ shrinks to $\mathbf{q}^{i^*_k}_{\mathrm{L}}[n], \in I'$.

V. Numerical Evaluations

In this section, we present the simulation results to demonstrate the performance of the proposed MARL method in terms of EE, E2E throughout, SAT associations, and UAV trajectories. To this end, we consider an area of $4000 \times 6000 \times 550$ [km^3] wherein the ground Src and Dst located at $[0, 0, 0]$ and $[4000, 4000, 0]$ are connected through two LEO orbital lanes of SATs between which UAVs are flying at an altitude of 50 [km] [25]. Each orbital lane consists of 22 SATs with the orbital lane circumference of 43486 [km], resulting in an inter-SAT distance of 1977 [km]. Since the orbital lane circumference is much larger than the area of interest, each orbital lane is approximated as a line segment at an altitude of 550 [km] [11], in which an infinite number of moving SATs are separated with the inter-SAT distance 1977 [km] and re-numbered at an interval of 22. In our MDP model, the objective function of (P1*) corresponds to the un-discounted accumulated rewards over an episode up to time slot N. The length of each episode is set as one orbital period, i.e., $N = 572$, during which an SAT with the orbital speed 7.59 [km/s] completes one orbit. Here, the orbital period and speed are calculated using the relations $4\pi^2(r_{\mathrm{E}})^3 = T^2GM$ and $V^2 r_{\mathrm{E}} = GM$, respectively, where r_{E}, radius of orbit in metres; T, orbital period in seconds; V, orbital speed in [m/s]; G, gravitational constant, approximately 6.673×10^{-11} [m^3/kg^1/s^2]; M, mass of Earth, approximately 5.98×1024 [kg].

Given the aforementioned MARL environment, we consider up to 3 UAVs that are agents. Following the centralized-Critic MARL framework [52]–[54], each agent has an Actor NN, while only one randomly selected agent stores a Critic NN. For all Actor and Critic NNs, we identically consider a 4-layer fully-connected multi-layer perceptron (MLP) NN architecture. Each MLP has 2 hidden layers, each of which has

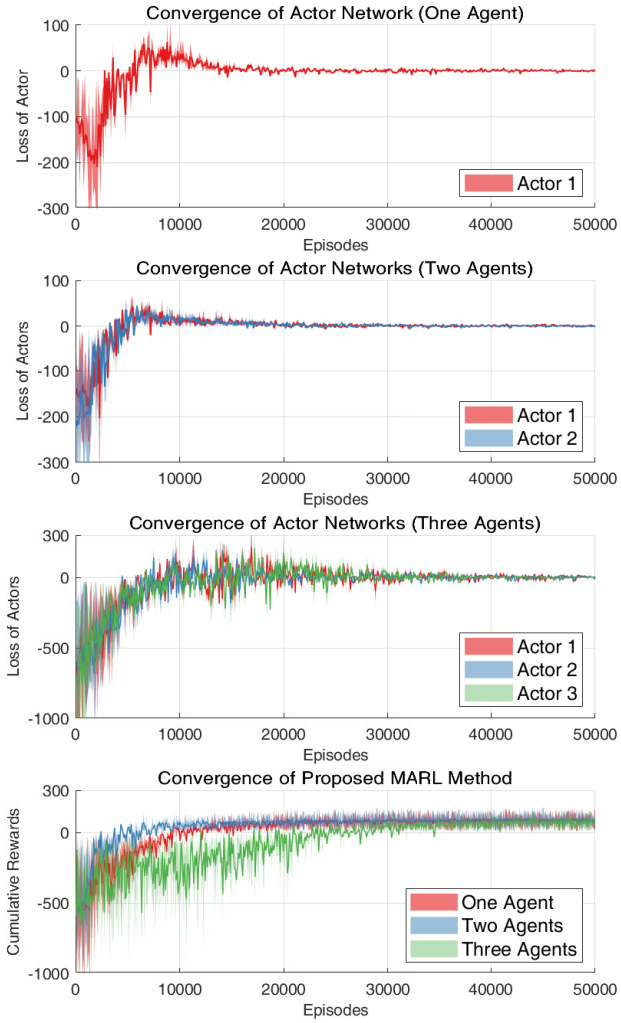


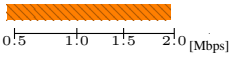
Figure 4: Learning curves of the proposed centralized-Critic MARL algorithm ($J = 1$-3, $K = 2$).

the dimension 128×128 with the rectified linear unit (ReLU) activation functions. Both Actor and Critic NNs are trained using RMSprop with the learning rate 0.0001, batch size 1072, training episodes 50000, and 1072 iterations per update. The simulations are implemented using TensorFlow Version 2.1. More details on simulation settings are provided in Table I. Throughout this section, our proposed scheme is compared with three benchmark schemes as follows.

1) **(Proposed) SAT-UAV** is a scenario with the proposed framework, in which Src and Dst are connected through SATs and UAVs. Unless otherwise specified, by default we consider 2 SAT orbital line segments having the opposite orbiting directions, between which 2 UAVs are flying, i.e., $K = J = 2$. The initial UAV positions are chosen as $[2000, 2667, 50]$ and $[2000, 1333, 50]$.
2) **SAT-Ground** implies a scenario wherein Src and Dst are connected through SATs and ground relay (GRs). The number of GRs and their initial positions are set as the same values of the UAV configurations, only except for their zero altitudes. GRs do not have any constraints and energy consumption incurred by UAV movements.

Table II: End-to-end throughput with or without relaying ($J = 1$).

Network Configuration	E2E Sum Txpt. [Mbps]
Direct Transmission	0.4507
SAT-Only, $K = 1$	0.6488
SAT-Only, $K = 2$	0.6839
SAT-Only, $K = 3$	1.1202
SAT-Ground, $K = 2$	**1.9721**



Accordingly, in this case (P1*) is reduced to

$$\text{(P2)} \quad \max_{\mathcal{W}} \sum_{j\in\mathcal{J}} \sum_{n=1}^{N} R^j_{\mathcal{S},\mathcal{D}}[n]$$
$$\text{s.t.} \quad (3), (12)-(15), (25)-(28).$$

In (P2) for SAT-Ground, $\mathcal{J} = \{j = 1, 2, \ldots, J\}$ represents a set of ground terminals, which relay between the SAT constellation network.

3) **SAT-Only** is the case where Src and Dst are connected only through SATs without any UAVs or GRs.
4) **Direct Transmission** refers to a scenario where Src directly communicates with Dst without any relays. For simplicity we assume that the LoS condition is always guaranteed, yielding the throughout upper bound performance.

A. MARL Training Convergence

In the proposed centralized-Critic MARL framework, the training convergence of the Actor and Critic NNs is illustrated in Fig. 4, where the solid curves denote average values and the shared areas correspond to the maximum deviations during 3 simulation runs. The first three figures show the loss values of the Actor NNs when the number J of agents (i.e., the number of Actor NNs) increases from 1 to 3, respectively, validating the convergence of the Actor NNs. The last figure illustrates the cumulative rewards at the Critic NN when $J = 3$, which implies the convergence of the overall MARL given the centralized-Critic architecture under study.

The results show that by up to 2 agents, each Actor NN and the overall MARL converge within about 10000 episodes, whereas for 3 agents, the convergence requires around 50000 episodes. For large-scale systems, it is thus necessary to accelerate the training convergence. In this regard, periodically averaging the Actor NN parameters (i.e., federated learning [55]) or outputs (i.e., federated distillation [56]) across agents could be an interesting topic for future study. Note that the cumulative reward does not directly represent the system utilities, i.e., E2E system throughput, energy, and EE, which are evaluated in the following subsections.

B. Throughput and Energy Efficiency

In our proposed NTN framework, UAVs and SATs play key roles for improving the system throughput and EE. In Tables II-IV, we provide ablation studies to clarify their contributions in relation to long-distance, inter-orbit, and mobile relaying under different objectives, as elaborated next.

Table III: End-to-end throughput with GRs or UAVs ($J = K = 2$).

Network Configuration	E2E Sum Txpt. [Mbps]
SAT-Ground, Non-cooperative	17.9519
SAT-Ground, Cooperative	29.3159
SAT-UAV, Cooperative	**58.2921**

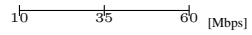


Table IV: Comparison of proposed SAT-UAV over different objectives ($J = K = 2$).

Objective	E2E Sum Txpt. [Mbps]	Avg. Power [W]	EE [Kbits/J]
Rate-max	**73.231**	1614.2	45.367
Energy-min	18.223	**314.24**	57.991
EE-max	58.292	570.62	**102.16**

To identify the effectiveness of relaying in E2E throughput, Table II compares SAT-Only relaying, SAT-Ground relaying, and Direct Transmission without relaying. With $K = 2$, the result shows that SAT-Only achieves 1.51x higher throughput compared to Direct Transmission. Next, it is remarkable that SAT-Ground achieves 4.38x higher throughput than Direction Transmission. This throughput is 2.9x higher than SAT-Only with the same $K = 2$, which is even 1.88x higher compared to SAT-Only with $K = 3$. The result highlights the effectiveness of additional relaying in long-distance communication throughput.

The previous result in Table II demonstrates that even a single fixed GR can provide a significant throughput gain. The gain can be further improved by multiple-UAV mobile relaying. In this regard, Table III illustrates the effectiveness of multi-agent relaying with or without cooperation across agents (i.e., UAVs or GRs). With multiple relaying agents, there exist multiple E2E communication paths by up to the number of agents. Then, cooperative scenarios correspond to (P1*) with $\sigma_{\mathrm{R}} = 1 \times 10^9$ and $\sigma_{\mathrm{E}} = 3 \times 10^4$ and (P2) for SAT-UAV and SAT-Ground, respectively, in which multiple agents aim to maximize their sum-path E2E throughput. By contrast, non-cooperative scenarios are the cases where each agent aims to maximize its own-path E2E throughput. The result shows that cooperative SAT-Ground yields 1.63x higher E2E sum throughput than its non-cooperative counterpart. This is mainly thanks to avoiding the overlapping associations with the same SAT that orthogonally splits the total bandwidth, significantly degrading the E2E sum throughput. Next, cooperative SAT-UAV achieves 1.99x higher E2E sum throughput than cooperative SAT-Ground, corroborating the effectiveness of the mobile UAV relaying compared to the fixed GRs.

Achieving the throughput gain in Table III comes at the cost of UAV movement energy consumption. To carefully identify its impact, Table IV focuses on SAT-UAV with $J = 2$ for different objective functions, and evaluates the E2E sum throughput, the power consumption averaged across UAVs, and EE, where the numbers are average values and the error bars indicate the maximum deviations during 3 simulation runs. Here, EE-max corresponds to considering the original objective function of (P1*) in (31), while Rate-max and Energy-min consider only the first term or the second term of the objective function in (31), respectively. The result shows that EE-max is well-designed, in that EE-max yields

Table V: Best-Effort: SAT associations and UAV trajectories during time slots $n = 1$ - 572 under (P1*) ($J = K = 2$).

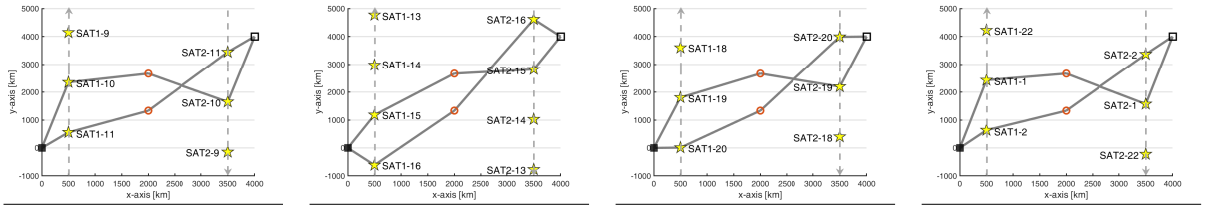
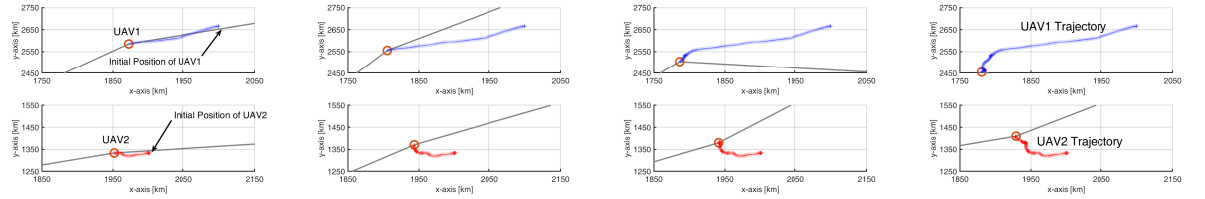
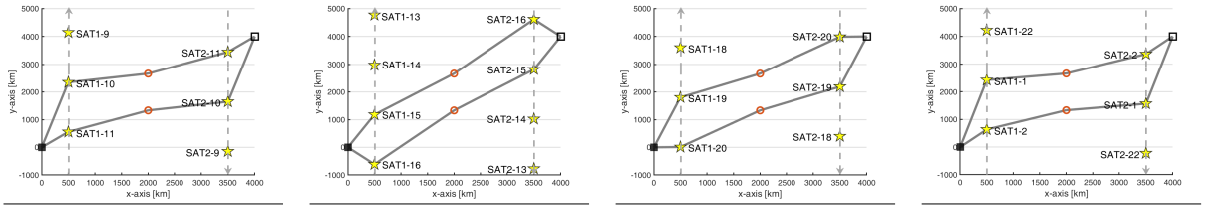
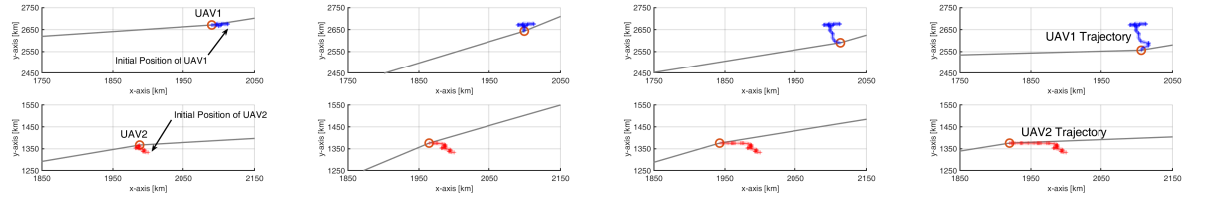
	$n = 200$ ($t = 2000$ [sec])	$n = 300$	$n = 400$	$n = 500$
Association (Best-Effort)				
Trajectory (Best-Effort)				

Table VI: Fairness-Oriented: SAT associations and UAV trajectories during time slots $n = 1$ - 572 under the objective function in (44) ($J = K = 2$).

	$n = 200$ ($t = 2000$ [sec])	$n = 300$	$n = 400$	$n = 500$
Association (Fairness)				
Trajectory (Fairness)				

the highest EE while achieving about 80% of Rate-max's E2E sum throughput and consuming less than 2x average power compared to Energy-min. As observed by these results, the proposed framework well excludes undesirable extreme outcomes of MARL, e.g., motionless UAVs to maximize the EE (i.e., sum throughput divided by total energy consumption) by making the denominator zero, which also advocates the usefulness of the objective reformulation from (P1) to (P1*) avoiding such extreme cases.

C. Best-Effort Associations and Trajectories

In Table V, we present the results of association and trajectory in SAT-UAV during one orbital period $n = 1 - 572$. In this table, the association for Src-SAT1-UAV1(or UAV2)-SAT2-Dst and the trajectory of UAV1 and UAV2 is shown over time slots, for Best-Effort scenario which refers to the proposed SAT-UAV (EE-max).

Over time slots, the association for the two UAV relays in-between SAT1 and SAT2 changes according to time-varying SATs position. Accordingly, in the optimization result, the handover occurs at a certain period as SATs circulate on an orbital lane at a constant velocity. Overall, UAV1 selects the association with SATs relatively close to itself, while UAV2 selects the association with SATs relatively close to Src and Dst. Specifically, in the cooperative system of SAT-UAV, one UAV occupies the optimal link, and the other UAV occupies the remaining link, which is more advantageous in terms of the E2E system throughput. This is because the optimal policy decides not to optimize specific links, but to improve the weakest link according to (24).

The optimized trajectory for UAV1 and UAV2 is shown in the bottom of this table. Note that the blue dots represent the traces of the UAV1 trajectory, and the red dots represent the traces of the UAV2 trajectory. The optimal trajectories draw like a straight line. It seems likely that each UAV goes to better location for connecting with moving SATs. We note that the E2E system throughput of the SAT-based NTN is efficiently improved with mobile UAV relaying, as identified in Table III.

The best-effort scenario focuses on the E2E system throughput, rather than fairness. If one mainly considers the fairness or link loss case, the policy should take account of proportional fairness. In the following subsection, we further investigate the fairness-oriented scenario.

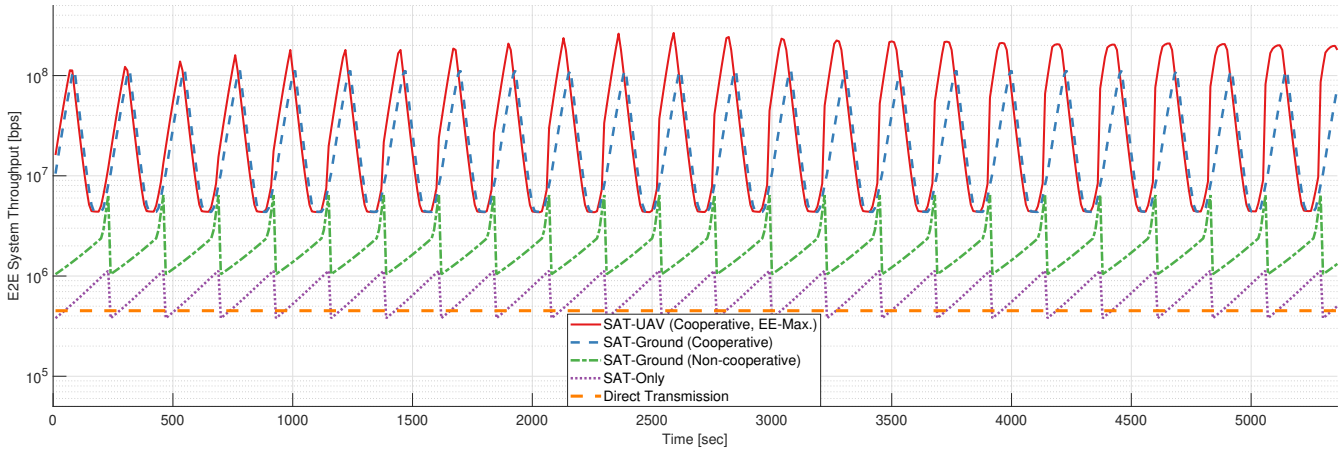


Figure 5: End-to-end system throughput of a time-varying network topology for different NTN configurations ($J = K = 2$, Best-Effort).

D. Fairness-Oriented Associtions and Trajectories

LEO SAT communication, which propagates through quite long link distance, may suffer severe link deterioration. Hence, consideration for link loss can be another significant issue. Accordingly, we further present the simulation results of the fairness-oriented scenario. In this scenario, the objective concerns fairness, such that the policy aims to the throughput of each agent, not the entire system.

Particularly, the fairness-oriented optimization is implemented with only minor modification to the proposed reward function. We re-design the reward function to improve the performance of each link fairly rather than maximizing one superior link to improve the E2E system throughput, with a *Sigmoid* function. Adopting the non-linear *Sigmoid* function of $f(x) = 1/(1+e^{-g(x)})$, the reward function in (36) slightly changes as follows.

$$r'[n] = f(\textstyle\sum_{j=1}^{2} R_{\mathcal{S},\mathcal{D}}[n]) - g(\sum_{j=1}^{2} P_j[n]\delta_t) \quad (43)$$
$$- h(d_{j=1,j'=2}[n]). \quad (44)$$

Note that in the optimization for a multi-agent system, the *Sigmoid* function, which saturates at large values, allows each agent to achieve the reward evenly, rather than a particular agent obtaining a large reward alone.

Table VI shows the association and trajectory as a result of the fairness-oriented optimization. Two agents of UAV1 and UAV2 are connected to the SATs close to them, considering their own E2E throughput. Both links with UAV1 and UAV2 achieves almost the same E2E throughput. In the worst case of losing one link, the fairness scenario achieves a 9.4372x higher E2E throughput than the best-effort scenario.

E. Impact of Time-Varying Network Topology

Fig. 5 compares the E2E system throughput over time slots with the baselines. As shown in this figure, the E2E system throughput result of relaying schemes fluctuates over time slots due to handover between SATs, e.g., $w_{i_1,j}[n] \neq w_{i_1,j}[n+1]$. Direct Transmission achieves stable throughput but relatively very low throughput, since no relay terminal supports its link. SAT-Only, which leverage LEO SAT constellation between Src and Dst, outperforms the Direct Transmission. SAT-Ground and SAT-UAV achieve the significantly more E2E system throughput than SAT-Only. For all baselines except Direct Transmission, the gap between high and low peak rates is huge overall, suggesting that it may require frequent and fast handover.

It is interesting to note that the low peak rate of SAT-Ground and SAT-UAV are similar, while SAT-UAV is superior at the high peak rate. In other words, the E2E system throughput of mobile UAV relaying is higher but more fluctuating than that of the fixed ground relaying. We investigate this fact in detail in the following subsection.

F. Impact of Hybrid RF/FSO Links

In order to investigate the candidates of the inter SAT-UAV link and the ground-to-SAT link, i.e., RF and FSO, we present simulation results for the hybrid FSO/RF.

In Fig. 7, to compare RF and FSO links in hybrid FSO/RF, we firstly show the spectral efficiency over link distance for RF and FSO links. As shown, the spectral efficiency of FSO and RF show different trends over link distance. The spectral efficiency of FSO follows the equation (10), and the spectral efficiency of RF follows (6). Comparing the FSO link of $\gamma_{\text{FSO}} = 25$ [dB] and the RF link of $\gamma_0 = 100$ [dB], the FSO link achieves a higher rate up to around 2400 [km] while the RF link achieves a higher rate after around 2400 [km]. Note that the crossing point of 2400 [km] is highlighted in Fig. 7.

Narrowing the scope to our network scenario, we compare the system throughput of RF and FSO links in more detail. In Fig. 6, we show the E2E system throughput of RF, FSO, and hybrid FSO/RF in our proposed scenario of SAT-UAV. Note that the result of SAT-UAV (Hybrid FSO/RF) in Fig. 6 is equivalent result of SAT-UAV (Cooperative, EE-max) in Fig. 5. In the proposed scenario, the hybrid FSO/RF link achieves up to 62.56x higher peak throughput and 21.09x higher worst-case throughput, compared to RF and FSO links, respectively. As shown in Fig. 6, the FSO link shows a high rate performance overall, but a very low peak rate at a specific point. On the other hand, the RF link shows a low rate performance overall, but the difference between the high peak

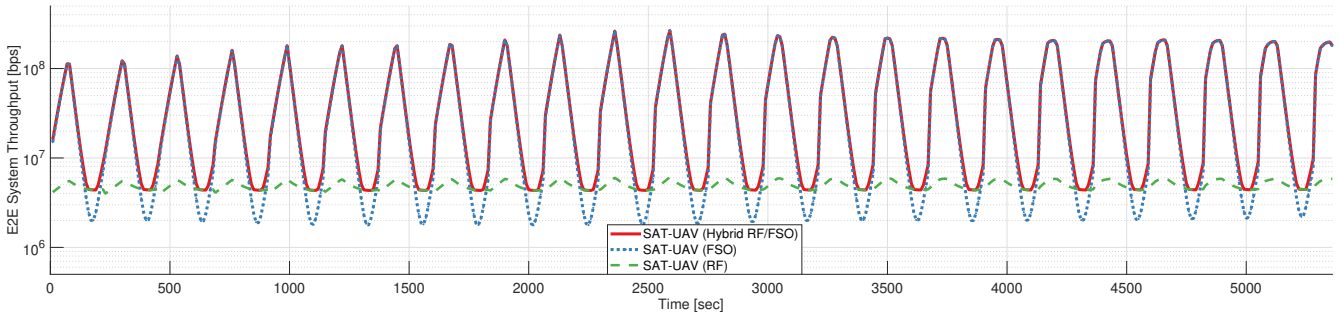


Figure 6: End-to-end system throughput of proposed SAT-UAV for different FSO/RF link types ($J = K = 2$, Best-Effort).

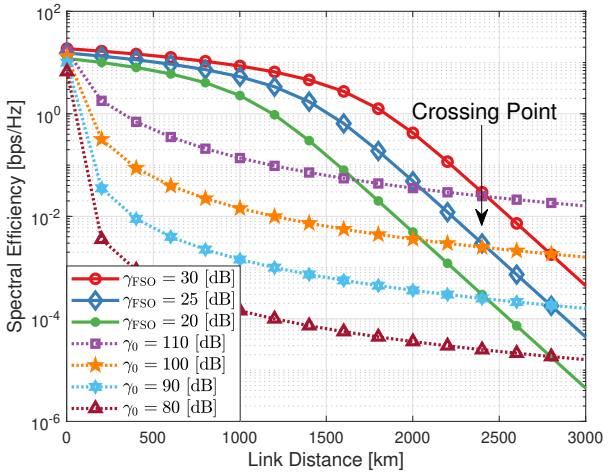


Figure 7: Spectral efficiency over distance for a given RF/FSO link.

Table VII: Selected link types ($J = K = 2$, Best-Effort).

Path with UAV1	**Src-LEO1**	**LEO1-UAV1**	**UAV1-LEO2**	**LEO2-Dst**
FSO:RF Ratio [%]	69.9 : 30.1	100 : 0	100 : 0	69.9 : 30.1
Avg. Link Distance [km]	2099.4	1643.6	1842.5	2099.4
Link Distance Range				
Path with UAV2	**Src-LEO1**	**LEO1-UAV2**	**UAV2-LEO2**	**LEO2-Dst**
FSO:RF Link Ratio [%]	100 : 0	88.1 : 11.9	13.1 : 86.9	100 : 0
Avg. Link Distance [km]	907.6	2027.1	2949.3	907.6
Link Distance Range				

rate and the low peak rate is relatively small. These facts explain that of Fig. 5, an RF link is dominant on the low peak rate of SAT-Ground and SAT-UAV and an FSO link is dominant on the high peak rate.

In Table VII, we show which link type is selected in each link. Note that the result of this table is based on three replications of the trained algorithm for SAT-UAV, and more favorable link (RF or FSO) is selected in each links. In our network scenario which involves mobile topology (such as SAT and UAV), no certain link type has an absolute dominance since the link distance for each links is time-varying.

VI. Conclusion

In this article, we investigated a satellite-UAV integrated hybrid FSO/RF NTN, and proposed a novel centralized-Critic MARL solution to maximize the E2E sum throughput while minimizing the UAV energy consumption, i.e., EE maximization. Numerical results corroborated the effectiveness of jointly controlling the UAV movements and associations with SATs. Not only the EE performance but also the UAV trajectories depend highly on the objective function choices. Optimizing the objective function for a specific task could thus be an interesting topic for future research. Besides, the MARL simulations revealed that the training convergence time increases with the number of agents. To scale up our proposed framework particularly under dynamic environments requiring frequent retraining, it is promising to apply federated learning methods that accelerate the training convergence. Last but not least, our simulations showed that FSO links are effective in improving the maximum throughput of SAT-LEO integrated NTNs, while RF links are useful for improving the worst-case throughput. Depending on application-specific and spatial channel characteristics, optimizing FSO/RF link types is therefore mandated for enabling fast and reliable NTNs in beyond 5G.

References

[1] Loon LLC. (2020) The stratosphere: High altitude, higher ambitions. [Online]. Available: https://loon.com/resources/content-library/

[2] S. Batir, N. Humann, C. Ortiz, D. Bettinger, S. Heger, and B. Vorster, "Low earth orbit satellites provide continuous enterprise data connectivity," 2020.

[3] B. Di, L. Song, Y. Li, and H. V. Poor, "Ultra-Dense LEO: Integration of satellite access networks into 5G and beyond," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 62–69, 2019.

[4] N. Saeed, A. Elzanaty, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini, "Cubesat communications: Recent advances and future challenges," *IEEE Commun. Surveys Tuts.*, 2020.

[5] J.-H. Lee, J. Park, M. Bennis, and Y.-C. Ko, "Integrating LEO satellite and UAV relaying via reinforcement learning for Non-Terrestrial networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, 2020, pp. 1–6.

[6] L. Singh, W. Whittecar, M. DiPrinzio, and et al., "Human-level control through deep reinforcement learning," *Nature Communication*, vol. 11, no. 200, 2020.

[7] Oneweb, Accessed: May- 2020. [Online]. Available: https://www.oneweb.world/

[8] Project Kuiper, Accessed: May- 2020. [Online]. Available: https://www.amazon.jobs/en/teams/projectkuiper

[9] Telesat, Accessed: May- 2020. [Online]. Available: https://www.telesat.com/

[10] Federal Communication Commission (FCC). (2018, Mar.) FCC authorizes SpaceX to provide broadband satellite services. [Online]. Available: https://www.fcc.gov/document/fcc-authorizes-spacex-provide-broadband-satellite-services

[11] Starlink, Accessed: May- 2020. [Online]. Available: https://www.starlink.com/

[12] M. Handley, "Using ground relays for low-latency wide-area routing in megaconstellations," in *Proc. ACM Workshop on Hot Topics in Networks*, 2019, pp. 125–132.

[13] ——, "Delay is not an option: Low latency routing in space," in *Proc. ACM Workshop on Hot Topics in Networks*, 2018, pp. 85–91.

[14] M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, Mar. 2019.

[15] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May. 2016.

[16] E. Yaacoub and M. Alouini, "A key 6G challenge and Opportunity—Connecting the base of the pyramid: A survey on rural connectivity," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 533–582, Apr. 2020.

[17] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proceedings of the IEEE*, vol. 107, no. 12, pp. 2327–2375, 2019.

[18] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 24–32, 2017.

[19] Loon LLC. (2020) Project Loon. [Online]. Available: https://loon.co/

[20] J. Marriott, B. Tezel, Z. Liu, and N. Stier. (2018) Trajectory optimization of solar-powered high-altitude long endurance aircraft. [Online]. Available: https://research.fb.com/publications/trajectory-optimization-of-solar-powered-high-altitude-long-endurance-aircraft/

[21] HAPSMobile. (2020) HAWK30. [Online]. Available: https://www.hapsmobile.com/

[22] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected uav: Uplink association, power control and interference coordination," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5380–5393, 2019.

[23] M. Mozaffari, A. T. Z. Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5G with UAVs: Foundations of a 3d wireless cellular network," *IEEE Trans on Wireless Commun.*, vol. 18, no. 1, pp. 357–372, 2019.

[24] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-Trajectory design and resource allocation for Solar-Powered UAV communication systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4281–4298, 2019.

[25] 3GPP TR38.811, "Study on new radio (NR) to support non-terrestrial networks," Jul. 2020.

[26] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[27] J. Lee, K. Park, M. Alouini, and Y. Ko, "Free space optical communication on UAV-assisted backhaul networks: Optimization for service time," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, 2019, pp. 1–6.

[28] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, March 2018.

[29] J. Lee, K. Park, M. Alouini, and Y. Ko, "Optimal resource allocation and placement for terrestrial and aerial base stations in mixed RF/FSO backhaul networks," in *Proc. IEEE Vehicular Technology Conf. (VTC-Spring)*, May 2020, pp. 1–5.

[30] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A Connectivity-Constrained trajectory optimization perspective," *IEEE Trans. on Commun.*, vol. 67, no. 3, pp. 2580–2604, 2019.

[31] J. Hu, H. Zhang, L. Song, Z. Han, and H. V. Poor, "Reinforcement learning for a cellular internet of UAVs: Protocol design, trajectory control, and resource management," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 116–123, 2020.

[32] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of UAVs: Distributed trajectory design by Multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, pp. 1–1, 2020.

[33] M. Alzenad, M. Z. Shakir, H. Yanikomeroglu, and M.-S. Alouini, "FSO-based vertical backhaul/fronthaul framework for 5G+ wireless networks," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 218–224, Jan. 2018.

[34] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini, "Hybrid Radio/Free-Space optical design for next generation backhaul systems," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2563–2577, 2016.

[35] V. Jamali, D. S. Michalopoulos, M. Uysal, and R. Schober, "Link allocation for multiuser systems with hybrid RF/FSO backhaul: Delay-limited and delay-tolerant designs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3281–3295, May 2016.

[36] M. Najafi, H. Ajam, V. Jamali, P. D. Diamantoulakis, G. K. Karagiannidis, and R. Schober, "Statistical modeling of the FSO fronthaul channel for UAV-Based communications," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3720–3736, 2020.

[37] H. Ajam, M. Najafi, V. Jamali, and R. Schober, "Ergodic sum rate analysis of UAV-Based relay networks with mixed RF-FSO channels," *IEEE Open Journal of the Communication Society*, vol. 1, pp. 164–178, 2020.

[38] A. Filippone, *Flight Performance of Fixed and Rotary Wing Aircraft*. Amsterdam, Netherlands: Elsevier, 2006.

[39] A. Lapidoth, S. M. Moser, and M. A. Wigger, "On the capacity of free-space optical intensity channels," *IEEE Trans. Inf. Theory.*, vol. 55, no. 10, pp. 4449–4461, Oct. 2009.

[40] I.-R. P.1814, "Prediction methods required for the design of terrestrial free-space optical links," in *International Telecommunication Union*, Geneva, Switzerland, 2007, pp. 1–12.

[41] F. Nadeem, V. Kvicera, M. S. Awan, E. Leitgeb, S. S. Muhammad, and G. Kandus, "Weather effects on hybrid FSO/RF communication link," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 9, pp. 1687–1697, 2009.

[42] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: challenges and solution approaches," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 22–31, 2015.

[43] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. 492–498, 1967.

[44] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep Multi-Agent reinforcement learning," ser. Proc. of International Conf. on Machine Learning (ICML), vol. 70, Sydney, Australia, 06–11 Aug 2017, pp. 1146–1155.

[45] G. Dulac-Arnold, R. Evans, H. V. Hasselt, and et al., "Reinforcement learning in large discrete action spaces," *arXiv:1512.07679 [cs.AI]*, 12 2015.

[46] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. MIT Press, 1998.

[47] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," ser. Proc. of International Conf. on Neural Information Processing (NIPS), 1999, pp. 1008–1014.

[48] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. of International Conf. on Machine Learning (ICML)*, vol. 48, New York, New York, USA, 20–22 Jun 2016, pp. 1928–1937.

[49] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," 2017.

[50] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proc. International Conference on Machine Learning (ICML)*, 1993.

[51] A. Tampuu, T. Matiisen, and et al., "Multiagent cooperation and competition with deep reinforcement learning," *CoRR*, 2015.

[52] J. Foerster, Y. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. of International Conf. on Neural Information Processing (NIPS)*, 2016, pp. 2145–2153.

[53] R. Lowe, Y. WU, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-Agent Actor-Critic for mixed Cooperative-Competitive environments," in *Proc. of International Conf. on Neural Information Processing (NIPS)*, 2017, pp. 6379–6390.

[54] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual Multi-Agent policy gradients," in *AAAI Conf. on Artificial Intelligence*, 2018.

[55] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis *et al.*, "Advances and open problems in federated learning," *ArXiv preprint, arXiv:1912.04977*, 2019.

[56] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *presented at Advances in Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning on the Phone and other Consumer Devices (MLPCD), Montreal, Canada*, 2018.