Joint Optimization of Transmission and Computation Resources for Satellite and High Altitude Platform Assisted Edge Computing

Changfeng Ding, *Student Member, IEEE*, Jun-Bo Wang, *Member, IEEE*, Hua Zhang, *Member, IEEE*, Min Lin, *Member, IEEE*, and Geoffrey Ye Li, *Fellow, IEEE*

***Abstract*— In this paper, we investigate a satellite-aerial integrated edge computing network (SAIECN) to combine a low-earth-orbit (LEO) satellite and aerial high altitude platforms (HAPs) to provide edge computing services for ground user equipment (GUE). In the SAIECN, GUE's computing tasks can be offloaded to HAP(s) or LEO satellite. In this paper, we minimize the weighted sum energy consumption of SAIECN via joint GUE association, multi-user multiple input and multiple output (MU-MIMO) transmit precoding, computation task assignment, and resource allocation. To solve the nonconvex problem, we decompose the optimization problem into four subproblems and solve each one iteratively. For the GUE association subproblem, quadratic transform based fractional programming (QTFP) and difference of convex function are utilized. The MU-MIMO transmit precoding subproblem is solved via QTFP and the weighted minimum mean-squared method. The computation task assignment is addressed using the classic interior point method while the computation resource allocation is derived in closed form. The numerical results show that the proposed SAIECN and the corresponding algorithm can solve the satellite based edge computing quite well and the energy cost is maintained at a relative low level.**

***Index Terms*— Computation offloading, edge computing, high altitude platform, LEO satellite, MU-MIMO, precoding, resource allocation.**

I. Introduction

ALTHOUGH the computing power of user equipment has experienced dramatic advancement in recent years, user equipment is facing higher computing demand applications, such as interactive gaming, augmented reality, and 3D modelling. Mobile edge computing (MEC) is a promising technology to satisfy the computation demand of resource limited mobile terminal, where the computation task is offloaded to the edge server, e.g., a mobile base station or a router, thus the computation-intensive and latency-critical applications can be enabled at user equipment [1], [2].

The most economical and easy way to implement MEC is to install edge servers in cellular base stations at fixed locations. However, over 50 percent of the world, especially the rural areas, remote mountain areas, and isolated islands, still lack Internet access [3]. In addition, cellular facilities are vulnerable to natural disasters, such as earthquakes and floods. To satisfy the demand of Internet access services in these cases, satellite communications have been paid great attention in recent years in both academic and industry circles [4]. In order to meet the demanding 5G requirements in terms of both large throughput and global connectivity, satellite communications provide valuable resources to extend and complement terrestrial networks [5]. In recent years, the satellite industry has witnessed great advancements in terms of low-cost satellite manufacturing, high-gain antenna, and laser transmission, which makes satellites, especially low-earth-orbit satellites (LEO SATs), a promising choice. Several companies have launched their commercial satellite communication systems to provide global broadband Internet access services, such as OneWeb [6] and SpaceX Starlink [7]. Meanwhile, the 3rd Generation Partnership Project (3GPP) has founded a work group to integrate LEO SAT system into 5G network, either as a stand-alone solution or as an integration to terrestrial network [8].

With the development of satellite-terrestrial communication systems, satellite communication systems are also facing the growing demands of users, such as quality-of-service (QoS), high data rates, low communication latency, and task computing services. Until now, most of the existing works have generally focused on the MEC under the scenario of cellular networks, where the edge server is coupled with the base station [9]–[13]. As an indispensable component in the future communication network, satellite can also exploit the advantage of edge computing and many works have discussed this topic. The architecture and application scenarios of satellite-terrestrial networks (STNs) have been introduced and several

Manuscript received November 26, 2020; revised March 22, 2021 and June 16, 2021; accepted August 5, 2021. Date of publication August 17, 2021; date of current version February 14, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1807803, in part by Jiangsu Province Basic Research Project under Grant BK20192002, and in part by the Key International Cooperation Research Project under Grant 61720106003. This article was presented in part at the IEEE International Conference on Communications (ICC), Montreal, QC, Canada, in June 2021. The associate editor coordinating the review of this article and approving it for publication was X. Gong. *(Corresponding author: Jun-Bo Wang.)*

Changfeng Ding, Jun-Bo Wang, and Hua Zhang are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 211111, China (e-mail: cfding@seu.edu.cn; jbwang@seu.edu.cn; huazhang@seu.edu.cn).

Min Lin is with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: linmin@njupt.edu.cn).

Geoffrey Ye Li is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: geoffrey.li@imperial.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3103764.

Digital Object Identifier 10.1109/TWC.2021.3103764

possible ways to implement MEC techniques in STNs have been investigated [14]. MEC has been applied into the satellite network and a novel architecture, named satellite-terrestrial integrated edge computing network, has been proposed [15]. A satellite aerial integrated network edge/cloud computing architecture has been developed for IoT systems [16]. Based on these studies, deploying MEC on satellites in STN can improve the quality-of-experience (QoE) of users and provide computing services for remote areas, such as fire monitoring and environmental monitoring.

Although integrating edge computing on satellite is attractive, the severe pathloss between the ground user equipment (GUE) and the LEO SAT poses significant challenge. Moreover, there are still some challenges in using LEO satellites to provide high-throughput Internet access services, such as latency-sensitive, prohibitive feedback load specially for a large number of users connected to the satellite. The unmanned aerial vehicle (UAV) has been used as a relay to upload the collected data from the ground devices to the satellite network [15], [16]. A satellite and UAV integrated network has been proposed to realize multi-cast communication with rate-splitting multiple access, where beamforming techniques are used for spectrum sharing between satellite served earth stations and internet-of-things devices [17]. Although UAV has the advantages of cost-effective and quick deployment, it only has short endurance due to limited battery energy or fuel cells, its coverage area is only within hundreds of meters, and it is vulnerable to unpredictable interference, which makes it hard to provide long-term and continuous wireless coverage. Recently, high-altitude platform (HAP) attracts the attention of researchers. HAPs are high altitude unmanned flying platforms situated in the stratosphere (from 17 to 22 km) and are a promising wireless solution to augment and support the existing satellite and terrestrial networks [18], which can provide multipurpose communications without requiring ground based infrastructure, especially in the remote mountain areas and disaster areas [19]. Compared with low-altitude UAV, the payload of a HAP can be over 100 kg [19] and HAP based communication has a wider coverage, longer endurance, and licensed spectrum. Therefore, HAPs can act as aerial base stations to improve the communication links between satellite and ground users and improve the overall network throughput. The HAP has been integrated into satellite communication system to share the millimeter wave spectrum, where a robust beamforming scheme has been proposed to maximize sum rate and minimize total power consumption [20]. In addition, satellite and HAPs are incorporated into terrestrial network to accommodate various services and applications in the future communication network, and software defined networking (SDN) based technologies have been utilized to realize a centralized control and exploit heterogeneous resources in an agile and flexible manner [21], [22].

In this paper, we propose a satellite-aerial integrated architecture to incorporate the LEO SAT and the aerial HAPs for edge computing, which can be conveniently and quickly deployed to provide on-demand computing services for GUEs. We will minimize the weighted sum energy consumption of GUEs, HAPs, and satellite, including both communication energy consumption and computing energy consumption. Moreover, the existing studies [9]–[13], [16], [23]–[27] of MEC generally use single antenna communication method for computation offloading. A MU-MIMO computation offloading scheme has been investigated to improve the offloading efficiency [28]. Secure beamforming schemes have been proposed to improve physical layer security for cognitive satellite-terrestrial networks in the presence of eavesdroppers [29], [30]. Advanced precoding techniques also have been discussed for multibeam satellites [31]. To improve link quality, a massive MIMO transmission scheme with full frequency reuse scheme has been proposed for the LEO SAT communication system [32]. In this paper, we adopt MIMO communication for GUE task offloading and HAP-satellite communication to combat pathloss and increase link capacity. We will optimize GUE-HAP association, and MU-MIMO transmit precoding for both GUE-HAP and HAP-satellite communication links. Moreover, the computation task assignment and computing resource allocation are also considered for satellite-aerial integrated edge computing network (SAIECN).

The main contributions of this paper are listed as below:

- We propose a SAIECN by utilizing the aerial HAPs and the LEO SAT to provide edge computing services for GUEs. To evaluate the performance of SAIECN, we formulate a weighted sum energy consumption minimization problem, where the GUE-HAP association, MU-MIMO transmit precoding, computation task assignment, and computation resource allocation are jointly optimized.
- To solve the mixed integer nonconvex weighted sum energy consumption minimization problem, we decompose the optimization problem into four subproblems and solve each one iteratively. For the GUE-HAP association subproblem, we first use the weighted minimum mean-squared error (WMMSE) to the rate expressions and decouple the binary variable. Then, we apply compressive sensing method to convert l_0 norm into a convex form. Difference of convex function algorithm (DCA) method is adopted to transform the association subproblem into a solvable convex problem.
- For the MU-MIMO transmit precoding subproblems, the QTFP method is used to deal with the fractional terms and convert the precoding problem into a convex one. Furthermore, we derive the upper bound and lower bound for the task assignment, which is solved by the interior point method. Moreover, the computation resource allocation subproblems are derived in closed form.

The rest of the paper is organized as follows. In Section II, the system model and the weighted sum energy minimization problem are introduced. The weighted sum energy consumption minimization problem is addressed in Section III. Simulation results are provided in Section IV and the paper is concluded in Section V.

$Notations$: Lower case boldface letters denote vectors while upper case boldface letters denote matrices. The trace of a matrix $\mathbf{A}$ is denoted by $\mathrm{Tr}\,(\mathbf{A})$. The transpose and conjugate-transpose of matrix $\mathbf{A}$ is denoted by $\mathbf{A}^T$ and $\mathbf{A}^H$, respectively. A positive semi-definite matrix $\mathbf{A}$ is denoted as

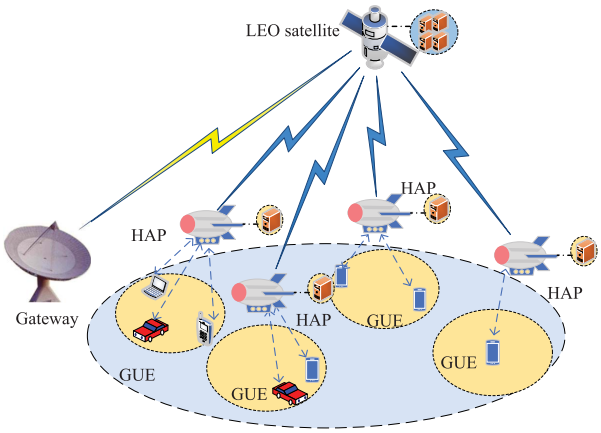


Fig. 1. System model of SAIECN.

$\mathbf{A} \succeq \mathbf{0}$. $\mathbb{E}[x]$ denotes the expectation of x. We use diag($\mathbf{a}$) to represent an diagonal matrix with its main diagonal elements from $\mathbf{a}$. The symbol, $\odot$, is the Hadamard product, i.e., the element-wise multiplication of two matrices. $\|\cdot\|_F^2$ denotes the square of Frobenius norm. $\|\cdot\|$ denotes Euclidean norm. $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. $\mathbb{C}^{M\times N}$ denotes the complex space of $M \times N$.

II. System Model and Problem Formulation

The considered satellite-aerial integrated edge computing network is shown in Fig.1. On the ground, I GUEs, denoted as $\mathcal{I} = \{1, 2, \ldots, I\}$, are randomly distributed. Each GUE supports the transmission of d data streams over N antennas ($d \leq N$) and has a computing task to handle. However, the task cannot be executed at local due to limited computing power and stored energy at GUE. We assume there is no available terrestrial cellular communication system for GUEs, such as a remote area. In the aerial layer, K high-altitude platforms (HAPs) powered by solar panels and solar cells, denoted as $k \in \mathcal{K} = \{1, 2, \ldots, K\}$, are deployed as relay nodes to provide communication/computing services for GUEs. Each HAP is equipped with an on-board computing processor and has M antennas, and floats at stratosphere in a quasi-static manner. Moreover, a LEO SAT equipped with computing server and M_S antennas is placed at an orbital altitude of H_S. The LEO SAT can communicate with HAPs and process the computing tasks offloaded by HAPs. Similar to [21], [22], we adopt SDN and network function virtualization technologies to integrate massive entities in the SAIECN and realize centralized control, where the centralized SDN controller can be deployed on the LEO satellite or the gateway station. Similar to [32], the doppler effects in satellite communication are assumed to be perfectly compensated.

A. GUE-HAP Channel Model

In this paper, the GUE and HAPs use MIMO communication to combat pathloss and channel fading. Moreover, we assume that the position of HAPs are quasi-fixed (e.g., when hovering). According to [33] and [34], the channel model between GUE i and HAP k can be modelled as the Rician fading channel model, which can be written as

$$\mathbf{H}_{k,i} = \sqrt{\frac{\psi_0}{L_{k,i}^2}}\left(\sqrt{\frac{\zeta}{\zeta+1}}\bar{\mathbf{H}}_{k,i} + \sqrt{\frac{1}{\zeta+1}}\hat{\mathbf{H}}_{k,i}\right) \in \mathbb{C}^{M\times N}, \quad (1)$$

where ψ_0 denotes the channel power gain at the reference distance, $L_{k,i}$ is the distance from GUE i to HAP k, $\bar{\mathbf{H}}_{k,i} \in \mathbb{C}^{M\times N}$ represents the LoS channel component with entry $\bar{\mathbf{H}}_{k,i}(m,n) = \exp(-j2\pi r_{m,n}/\lambda_g)$ with $r_{m,n}$ being the length of the direct path between the n_{th} transmit antenna and the m_{th} receive antenna and λ_g being the wavelength of the transmitted signals, $\hat{\mathbf{H}}_{k,i} \in \mathbb{C}^{M\times N} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ denotes the Rayleigh fading component, and $\zeta \geq 0$ is the Rician factor specifying the power ratio between the LoS and Rayleigh fading channel components in $\mathbf{H}_{k,i}$.

B. HAP-Satellite Channel Model

When the HAP receives the computing tasks from GUEs, it may offload certain amount of the tasks to the LEO SAT to release computing burden. Due to the relative long distance between the HAP and the LEO SAT, the HAP also uses MIMO communication with the LEO SAT, where the LEO SAT usually adopts high frequency bands above 10 GHz for communication, such as Ka or Qu band. Since such high frequency bands suffer from relative larger pathloss, high-gain and directive antennas are required at both the LEO SAT and the HAP to obtain sufficient link gain. Hence, the multipath component, if there is any, is suppressed by directional antenna [35]. For the air-to-space communication, the effect of pathloss, atmospheric impairment, and satellite antenna gain should be considered to properly model the satellite MIMO channel.

Denote $\mathbf{P}_{S,k} \in \mathbb{C}^{M_S\times M}$ as the LoS channel coefficient matrix between HAP k and the LEO SAT. It's (s,m)-th entry, denoted as $p_{sm}^{(k)}$, corresponds to the channel coefficient between the m-th HAP antenna and the s-th LEO SAT antenna can be expressed as [35]

$$p_{sm}^{(k)} = a_{sm}^{(k)} e^{-j\frac{2\pi}{\lambda_c} r_{sm}^{(k)}}, \quad (2)$$

where $\lambda_c = c_0/f_c$ denotes the wavelength of carrier frequency f_c, c_0 is the speed of light, and $a_{sm}^{(k)} = \lambda_c \Big/ \left(4\pi r_{sm}^{(k)}\right)$ is the free-space propagation loss and the parameter, r_{sm}, denotes the distance between the m-th transmit antenna and the s-th receive antenna. The last term in (2), $e^{-j\frac{2\pi}{\lambda_c} r_{sm}^{(k)}}$, represents the phase rotation of the pure LoS link between the m-th transmit antenna and the s-th receive antenna.

The air-to-space radio wave propagation is affected by the atmospheric impairments from troposphere, which include attenuation effect as well as phase disturbance [36]. The atmospheric impairments for the m-th HAP antenna can be formulated as $\varsigma_m = |\varsigma_m| e^{-j\xi_m}$, where $|\varsigma_m| \in [0,1]$ and $\xi_m \in [-\pi, \pi]$ denote the amplitude attenuation and phase shift, respectively. Accordingly, the atmospheric impairments matrix $\mathbf{D}_{S,k} \in \mathbb{C}^{M\times M}$ between the HAP antenna and the SAT is given by

$$\mathbf{D}_{S,k} = \mathrm{diag}\{\varsigma_1, \ldots, \varsigma_M\}. \quad (3)$$

The normalized radiation pattern of the satellite receive antennas is denoted as matrix $\mathbf{G}_{S,k} \in \mathbb{C}^{M_S \times M}$, its elements can be calculated as [35],

$$g_{zm} = J_1(u_{zm})/2u_{zm} + 36J_3(u_{zm})/u_{zm}^3, \tag{4}$$

where $u_{zm} = \pi D_s/\lambda_c \sin(\vartheta_{zm})$, and $J_1(u_{zm})$ and $J_3(u_{zm})$ are the Bessel functions of first kind and order one and three, respectively, D_s denotes the diameter of circular antenna array on the satellite, and ϑ_{zm} is the off-axis of the z-th beam's boresight to HAP's antenna m.

According to the above modelling of the path loss, the atmospheric impairments, and the satellite antenna gain, the overall HAP-satellite channel between the LEO SAT and HAP k can be given as

$$\mathbf{H}_{S,k} = \mathbf{P}_{S,k}\mathbf{D}_{S,k} \odot \mathbf{G}_{S,k}. \tag{5}$$

C. Transmission Model Between GUE and HAP

On the ground, each GUE is associated with one of the HAPs and can offload its computing task to the associated HAP, which can be expressed mathematically as follows

$$\sum_{k=1}^{K} a_{i,k} = 1, \quad \forall i \in \mathcal{I}, \tag{6}$$

where $a_{i,k} = \{0,1\}$, $\forall k \in \mathcal{K}$, is the offloading indicator of GUE i. Note that $a_{i,k} = 1$ indicates that GUE i is associated with HAP k, and $a_{i,k} = 0$ otherwise.

Moreover, the load of each HAP is limited and should satisfy

$$\sum_{i=1}^{I} a_{i,k} \le N_H, \quad \forall k \in \mathcal{K}, \tag{7}$$

where N_H is the maximum associated number of GUEs for HAP.

From the above discussion, the received signal, $\mathbf{y}_{U,k,i} \in \mathbb{C}^{M\times 1}$, of GUE i at HAP k can be expressed as

$$\mathbf{y}_{U,k,i} = \mathbf{H}_{k,i}\mathbf{Q}_{i,k}\mathbf{x}_i + \sum_{j=1,j\neq i}^{I}\sum_{m=1}^{K} a_{j,m}\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{x}_j + \mathbf{n}_k, \tag{8}$$

where $\mathbf{Q}_{i,k} \in \mathbb{C}^{N\times d}$ is the precoding matrix that GUE i used to transmit the signal $\mathbf{x}_i \in \mathbb{C}^{d\times 1}$ to HAP k and $\mathbf{n}_k \in \mathbb{C}^{M\times 1}$ is the additive white Gaussian noise (AWGN) with distribution $\mathcal{CN}\left(0, \sigma_k^2\mathbf{I}_M\right)$. We assume that the signals from different users are independent from each other and are also independent of the received noises. The achievable data rate between GUE i and HAP k can be given by

$$R_{i,k} = B_g \log\det\left(\mathbf{I}_M + \mathbf{H}_{k,i}\mathbf{Q}_{i,k}\mathbf{Q}_{i,k}^H\mathbf{H}_{k,i}^H\mathbf{N}_{i,k}^{-1}\right), \tag{9}$$

where B_g is the available bandwidth for GUE and $\mathbf{N}_{i,k}$ is the interference plus AWGN and is denoted by

$$\mathbf{N}_{i,k} = \sum_{j=1,j\neq i}^{I}\sum_{m=1}^{K} a_{j,m}\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{Q}_{j,m}^H\mathbf{H}_{k,j}^H + \sigma_k^2\mathbf{I}_M. \tag{10}$$

Accordingly, the transmission latency and energy consumption for GUE i to offload computing task to HAP k can be given respectively as

$$T_{tran,i,k} = \frac{\beta B_i}{R_{i,k}}, \tag{11}$$

and

$$E_{tran,i,k} = \frac{\beta B_i}{R_{i,k}}\|\mathbf{Q}_{i,k}\|_F^2, \tag{12}$$

where β denotes the ratio of the transmitted data size to the original task data size due to the transmission overhead [37] and B_i (in bits) is the size of the input data. $\|\mathbf{Q}_{i,k}\|_F^2$ is the transmit power from GUE i to HAP k and is constrained by the maximum transmit power, $p_{\max,i}$, which can be written as

$$\|\mathbf{Q}_{i,k}\|_F^2 \le p_{\max,i}, \quad \forall k \in \mathcal{K},\ i \in \mathcal{I}. \tag{13}$$

In this paper, partial offloading protocol is adopted, and GUE's input bits are bit-wise independent and can be arbitrarily divided to facilitate parallel trade-offs between computation offloading to the HAP or further to the LEO SAT with the assistance of HAP. Therefore, the HAP and the LEO satellite process different portions of each GUE's computation task. GUE i's ($i \in I$) computing task can be divided into two parts: $l_i \ge 0$ bits are processed at the HAP while $B_i - l_i \ge 0$ bits are offloaded to the LEO SAT for computation via the HAP. Hence, the time delay and energy consumption for computing GUE i's task at HAP k can be given respectively as

$$T_{com,k,i} = \frac{\alpha_i l_i}{f_{k,i}}, \tag{14}$$

and

$$E_{com,k,i} = \kappa_{U,k}\alpha_i l_i f_{k,i}^2, \tag{15}$$

where α_i (cycles/bit) is the processing density of GUE i's computing task, $f_{k,i}$ is the computation resource allocated for GUE i and $\kappa_{U,k}$ is a constant relative to the hardware architecture of HAP k.

D. Transmission Model Between HAP and LEO SAT

After dividing the offloaded tasks from the GUEs, the HAP can further offload the remaining tasks to the LEO SAT for computing. To overcome the large pathloss and enhance the received signal quality, the HAP also uses MIMO transmission to the LEO SAT. Moreover, we assume that each HAP can transmit b data streams over M antennas ($b \le M$). With MIMO transmission, the received signal, $\mathbf{y}_{S,k} \in \mathbb{C}^{M_S\times 1}$, of HAP k at the LEO SAT can be written as

$$\mathbf{y}_{S,k} = \mathbf{H}_{S,k}\mathbf{W}_k\mathbf{s}_k + \sum_{j=1,j\neq k}^{K}\left\|\sum_{i=1}^{I} a_{i,j}\right\|_0 \mathbf{H}_{S,j}\mathbf{W}_j\mathbf{s}_j + \mathbf{n}_S, \tag{16}$$

where $\mathbf{W}_k \in \mathbb{C}^{M\times b}$ is the precoding matrix that HAP k used to transmit signal vector $\mathbf{s}_k \in \mathbb{C}^{b\times 1}$ to the LEO SAT, $\mathbf{s}_k$ satisfies $E\left[\mathbf{s}_k\mathbf{s}_k^H\right] = \mathbf{I}_b$ and $E\left[\mathbf{s}_k\mathbf{s}_l^H\right] = \mathbf{0}$ for $k \neq l$, $\mathbf{H}_{S,k} \in \mathbb{C}^{M_S\times M}$ is the channel matrix from HAP k to the LEO SAT, and $\mathbf{n}_S \in C^{M_S\times 1}$ denotes the AWGN vector received

at the LEO SAT and $E\left[\mathbf{n}_S\mathbf{n}_S^H\right] = \sigma_S^2\mathbf{I}_{M_S}$. The l_0-norm of $\left\|\sum_{i=1}^{I} a_{i,j}\right\|_0$ is to indicate whether HAP j is chosen by GUEs as an offloading node. The achievable data rate of HAP k can be formulated as

$$R_k = B_h \log\det\left(\mathbf{I}_{M_S} + \mathbf{H}_{S,k}\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_{S,k}^H\mathbf{N}_{S,k}^{-1}\right), \quad (17)$$

where B_h is the available bandwidth for the HAP and $\mathbf{N}_{S,k}$ is the interference plus AWGN and is expressed by

$$\mathbf{N}_{S,k} = \sigma_S^2\mathbf{I}_{M_S} + \sum_{j=1,j\neq k}^{K}\left\|\sum_{i=1}^{I} a_{i,j}\right\|_0 \mathbf{H}_{S,j}\mathbf{W}_j\mathbf{W}_j^H\mathbf{H}_{S,j}^H. \quad (18)$$

Then, the time delay and energy consumption for offloading the remaining $B_i - l_i$ task bits from HAP k to the LEO SAT can be denoted, respectively, as

$$T_{o,s,k,i} = \frac{\beta\left(B_i - l_i\right)}{R_k} + T_p(k,s), \quad (19)$$

and

$$E_{o,s,k,i} = \frac{\beta\left(B_i - l_i\right)}{R_k}\|\mathbf{W}_k\|_F^2, \quad (20)$$

where $T_p(k,s)$ is the radio propagation latency from HAP k to the LEO SAT, $\|\mathbf{W}_k\|_F^2$ is the transmit power of HAP k and is constrained by the maximum transmit power $p_{\max,k}$,

$$\|\mathbf{W}_k\|_F^2 \leq p_{\max,k}, \quad \forall k \in \mathcal{K}. \quad (21)$$

Besides, for the LEO SAT, the computation time delay and energy consumption for executing the $B_i - l_i$ bits can be expressed respectively as

$$T_{com,s,i} = \frac{\alpha_i\left(B_i - l_i\right)}{f_{s,i}}, \quad (22)$$

and

$$E_{com,s,i} = \kappa_S\alpha_i\left(B_i - l_i\right)f_{s,i}^2, \quad (23)$$

where $f_{s,i}$ is the satellite computation resource allocated to GUE i, κ_S is the effective capacitance coefficient of the LEO SAT. Since the computing result is small, the time delay and energy consumption for returning the computing result are omitted, similar to most of the existing studies. At last, the roundtrip time delay between HAP k and the LEO SAT can be written as

$$\begin{aligned} T_{r,k,s} &= T_{o,s,k,i} + T_{com,s,i} + T_p(k,s) \\ &= \frac{\beta\left(B_i - l_i\right)}{R_k} + \frac{\alpha_i\left(B_i - l_i\right)}{f_{s,i}} + 2T_p(k,s). \end{aligned} \quad (24)$$

Notably, HAP can perform transmission and computing functions simultaneously. Hence, $T_{com,k,i}$ and $T_{r,k,s}$ are independent and can be overlapped in the timeline.

E. Overall Offloading Time Delay and Energy Consumption Model

According to the above analysis, the computation offloading time delay of GUE consists of two parts. The first part is the offloading time delay at the HAP, and the second part is the offloading time delay at the LEO SAT, which can be formulated respectively as

$$T_{H,i} = \sum_{k=1}^{K} a_{i,k}\left(\frac{\beta B_i}{R_{i,k}} + \frac{\alpha_i l_i}{f_{k,i}}\right), \quad (25)$$

and

$$\begin{aligned} T_{S,i} = \sum_{k=1}^{K} a_{i,k}\left(\frac{\beta B_i}{R_{i,k}} + \frac{\beta\left(B_i - l_i\right)}{R_k}\right) + \frac{\alpha_i\left(B_i - l_i\right)}{f_{s,i}} \\ + 2T_p(k,s). \end{aligned} \quad (26)$$

$T_{H,i}$ and $T_{S,i}$ represent the offloading time delays of independent computation tasks and both should satisfy the maximum tolerable latency of GUE i, $T_{\max,i}$.

For GUE i, the energy consumption for offloading the computation task to the HAP can be written as

$$E_{tran,i} = \sum_{k=1}^{K} a_{i,k}\frac{\beta B_i}{R_{i,k}}\|\mathbf{Q}_{i,k}\|_F^2. \quad (27)$$

Meanwhile, the energy consumption of the HAP for computing and relaying GUE i's computation task can be formulated respectively as

$$E_{com,i} = \sum_{k=1}^{K} a_{i,k}\kappa_{U,k}\alpha_i l_i f_{k,i}^2, \quad (28)$$

and

$$E_{o,i} = \sum_{k=1}^{K} a_{i,k}\frac{\beta\left(B_i - l_i\right)}{R_k}\|\mathbf{W}_k\|_F^2. \quad (29)$$

At last, the total energy consumption of the LEO SAT for computing GUEs' tasks can be given as

$$E_{com,s} = \sum_{i=1}^{I}\kappa_S\alpha_i\left(B_i - l_i\right)f_{s,i}^2. \quad (30)$$

F. Problem Formulation

The purpose of this paper is to minimize the overall energy consumption of the GUEs, HAPs and the LEO SAT under the maximum latency constraint. According to (27), (28), (29), and (30), the total weighted sum of energy consumption of the system can be expressed as

$$\begin{aligned} E = {} & w_G\sum_{i=1}^{I}\left(\sum_{k=1}^{K} a_{i,k}\frac{\beta B_i}{R_{i,k}}\|\mathbf{Q}_{i,k}\|_F^2\right) \\ & + w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K} a_{i,k}\kappa_{U,k}\alpha_i l_i f_{k,i}^2\right) \\ & + w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K} a_{i,k}\frac{\beta\left(B_i - l_i\right)}{R_k}\|\mathbf{W}_k\|_F^2\right) \\ & + w_S\sum_{i=1}^{I}\kappa_S\alpha_i\left(B_i - l_i\right)f_{s,i}^2, \end{aligned} \quad (31)$$

where w_G, w_H, and w_S are constant positive weights for GUEs' energy consumption, HAPs' energy consumption and LEO SAT's energy consumption, respectively. Mathematically, the optimization problem can be expressed as

$$\min_{\mathbf{A},\mathbb{Q},\mathbb{W},\mathcal{L},\mathcal{F}_U,\mathcal{F}_S} E \tag{32}$$

$$s.t. \sum_{k=1}^{K} a_{i,k} = 1, \quad \forall i \in \mathcal{I}, \tag{32a}$$

$$\sum_{i=1}^{I} a_{i,k} \le N_H, \quad \forall k \in \mathcal{K}, \tag{32b}$$

$$\|\mathbf{Q}_{i,k}\|_F^2 \le p_{\max,i}, \quad \forall k \in \mathcal{K},\ i \in \mathcal{I}, \tag{32c}$$

$$\|\mathbf{W}_k\|_F^2 \le p_{\max,k}, \quad \forall k \in \mathcal{K}, \tag{32d}$$

$$\sum_{i=1}^{I} a_{i,k} f_{k,i} \le f_{total,k}, \quad \forall k \in \mathcal{K}, \tag{32e}$$

$$\sum_{i=1}^{I} f_{s,i} \le f_{total,s}, \tag{32f}$$

$$0 \le l_i \le B_i, \quad \forall i \in \mathcal{I}, \tag{32g}$$

$$T_{H,i} \le T_{\max,i}, \quad \forall i \in \mathcal{I}, \tag{32h}$$

$$T_{S,i} \le T_{\max,i}, \quad \forall i \in \mathcal{I}, \tag{32i}$$

$$a_{i,k} \in \{0,1\}, \quad \forall i \in \mathcal{I},\ \forall k \in \mathcal{K}, \tag{32j}$$

where $\mathbf{A} = \{a_{i,k}\}_{i\in\mathcal{I},k\in\mathcal{K}}$, $\mathbb{Q} = \{\mathbf{Q}_{i,k}\}_{k\in\mathcal{K},i\in\mathcal{I}}$, $\mathbb{W} = \{\mathbf{W}_k\}_{k\in\mathcal{K}}$, $\mathcal{L} = \{l_i\}_{i\in\mathcal{I}}$, $\mathcal{F}_U = \{f_{k,i}\}_{k\in\mathcal{K},i\in\mathcal{I}}$, and $\mathcal{F}_S = \{f_{s,i}\}_{i\in\mathcal{I}}$. $f_{total,k}$ and $f_{total,s}$ are the total computation resource of HAP k and the LEO SAT, respectively.

Constraints (32a) and (32b) state that each GUE can only associate to one of the HAPs and the maximum number of associated GUEs for HAP, respectively. Constraints (32c) and (32d) are the maximum transmit power for the GUE and the HAP, respectively. (32e) and (32f) denote that the sum of allocated computation resource of the HAP and the LEO SAT should be no larger than their total computation resource, respectively. Constraint (32g) represents that the allocated computation bits on HAP should be no larger than B_i. Constraints (32h) and (32i) set up the maximum tolerable latency.

III. Algorithm Design for Satellite-Aerial Integrated Edge Computing

From (31), the objective function is non-convex due to the sum of ratios, the coupled optimization variables and the discrete binary constraints. Therefore, problem (32) is non-convex. To obtain a global optimal solution for this non-convex problem is almost impossible. Alternatively, we solve problem (32) iteratively and obtain a suboptimal solution. Subsequently, the association variable optimization is first tackled. After obtaining the optimized association variables, GUEs can be divided into different HAP-GUE sets and MU-MIMO transmit precoding optimization can be implemented for the GUE-HAP and the HAP-LEO SAT communication links, respectively. Then, the subproblem of computation task assignment is solved. At last, the computation resource allocation strategies for both the HAP and the LEO SAT are obtained.

A. Association Variable Optimization

Problem (32) is difficult to be handled due to the binary association variable and the nonlinear sum of ratios. Meanwhile, the binary association variables are coupled with the precoding matrices inside of the rate function of (9) and (17). To facilitate analysis, we apply the WMMSE method [38] to reformulate the rate expression. By adopting linear receive precoding strategy, the estimated signal vector of GUE i at HAP k can be given by

$$\hat{\mathbf{x}}_{k,i} = \mathbf{V}_{k,i}^H \mathbf{y}_{U,k,i}, \tag{33}$$

where $\mathbf{V}_{k,i} \in \mathbb{C}^{M\times d}$ is the receive precoding matrix for GUE i associated with HAP k. With the independence of signal vector $\mathbf{x}_i$ and noise vector $\mathbf{n}_k$, the MSE matrix of GUE i can be formulated as

$$\begin{aligned}\mathbf{E}_{i,k} &= E_{\mathbf{x},\mathbf{n}}\left[\left(\hat{\mathbf{x}}_{k,i} - \mathbf{x}_i\right)\left(\hat{\mathbf{x}}_{k,i} - \mathbf{x}_i\right)^H\right] \\ &= \left(\mathbf{V}_{k,i}^H \mathbf{H}_{k,i}\mathbf{Q}_{i,k} - \mathbf{I}_d\right)\left(\mathbf{V}_{k,i}^H \mathbf{H}_{k,i}\mathbf{Q}_{i,k} - \mathbf{I}_d\right)^H \\ &\quad + \sum_{j=1,j\neq i}^{I}\sum_{m=1}^{K} a_{j,m}^2 \mathbf{V}_{k,i}^H \mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{Q}_{j,m}^H \mathbf{H}_{k,j}^H \mathbf{V}_{k,i} \\ &\quad + \sigma_k^2 \mathbf{V}_{k,i}^H \mathbf{V}_{k,i}.\end{aligned} \tag{34}$$

Since $a_{j,m}$ is a binary variable, $a_{j,m}^2$ can be replaced with $a_{j,m}$ in the following analysis.

Based on Appendix A, the following theorem can be proved, which is given as

Theorem 1: Let $\{\mathbf{P}_{i,k} \succeq 0 | k \in \mathcal{K}, i \in \mathcal{I}\}$ be the set of auxiliary matrices, the rate function of $R_{i,k}$ in (9) can be given as

$$\tilde{R}_{i,k}(\mathbf{P}_{i,k}, \mathbb{Q}, \mathbf{V}_{k,i}) = \log\det\left(\mathbf{P}_{i,k}\right) - \mathrm{Tr}\left(\mathbf{P}_{i,k}\mathbf{E}_{i,k}\right) + d. \tag{35}$$

According to [38] and [39], the relationship between $\tilde{R}_{i,k}$ and $R_{i,k}$ can be established in the following lemma.

Lemma 1: $\tilde{R}_{i,k}$ is a concave function for each set of matrices $\mathbf{P}_{i,k}$, $\mathbb{Q}$ and $\mathbf{V}_{k,i}$ when the other two are given [39]. The optimal $\mathbf{V}_{k,i}$ and $\mathbf{P}_{i,k}$ for $\tilde{R}_{i,k}\left(\mathbf{P}_{i,k}, \mathbb{Q}, \mathbf{V}_{k,i}\right)$ to achieve the data rate is given by

$$\mathbf{V}_{k,i}^* = \mathbf{\Pi}_{k,i}^{-1}\mathbf{H}_{k,i}\mathbf{Q}_{i,k}, \tag{36}$$

$$\mathbf{P}_{i,k}^* = \mathbf{E}_{i,k}^{*-1}, \tag{37}$$

and

$$\mathbf{E}_{i,k}^* = \mathbf{I}_d - \mathbf{Q}_{i,k}^H \mathbf{H}_{k,i}^H \Pi_{k,i}^{-1} \mathbf{H}_{k,i}\mathbf{Q}_{i,k}, \tag{38}$$

where $\mathbf{\Pi}_{k,i}$ is expressed as

$$\begin{aligned}\mathbf{\Pi}_{k,i} &= \mathbf{H}_{k,i}\mathbf{Q}_{i,k}\mathbf{Q}_{i,k}^H\mathbf{H}_{k,i}^H + \sigma_k^2\mathbf{I}_M \\ &\quad + \sum\nolimits_{j=1,j\neq i}^{I}\sum\nolimits_{m=1}^{K} a_{j,m}\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{Q}_{j,m}^H\mathbf{H}_{k,j}^H\end{aligned} \tag{39}$$

and $\mathbf{E}_{i,k}^*$ is achieved by taking the formulation of $\mathbf{V}_{k,i}^*$ into equation (34).

Proof: By taking the MSE expression of (34) into expression (35), we have

$$\begin{aligned}&\tilde{R}_{i,k}\left(\mathbf{P}_{i,k},\mathbf{A},\mathbb{Q},\mathbf{V}_{k,i}\right)\\&=\log\det\left(\mathbf{P}_{i,k}\right)-\sigma_k^2\mathrm{Tr}\left(\mathbf{V}_{k,i}^H\mathbf{V}_{k,i}\mathbf{P}_{i,k}\right)+d\\&\quad-\mathrm{Tr}\left(\left(\mathbf{V}_{k,i}^H\mathbf{H}_{k,i}\mathbf{Q}_{i,k}-\mathbf{I}_d\right)\mathbf{P}_{i,k}\left(\mathbf{V}_{k,i}^H\mathbf{H}_{k,i}\mathbf{Q}_{i,k}-\mathbf{I}_d\right)^H\right)\\&\quad-\sum_{j=1,j\neq i}^{I}\sum_{m=1}^{K}a_{j,m}\mathrm{Tr}\left(\mathbf{V}_{k,i}^H\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{P}_{i,k}\mathbf{Q}_{j,m}^H\mathbf{H}_{k,j}^H\mathbf{V}_{k,i}\right).\end{aligned}\tag{40}$$

(40) can be equivalently reformulated as

$$\begin{aligned}&\tilde{R}_{i,k}\left(\mathbf{P}_{i,k},\mathbf{A},\mathbb{Q},\mathbf{V}_{k,i}\right)\\&=\log\det\left(\mathbf{P}_{i,k}\right)-\sigma_k^2\mathrm{Tr}\left(\mathbf{V}_{k,i}^H\mathbf{V}_{k,i}\mathbf{P}_{i,k}\right)\\&\quad-\left\|\left(\mathbf{V}_{k,i}^H\mathbf{H}_{k,i}\mathbf{Q}_{i,k}-\mathbf{I}_d\right)\mathbf{P}_{i,k}^{1/2}\right\|_F^2\\&\quad-\sum_{j=1,j\neq i}^{I}\sum_{m=1}^{K}a_{j,m}\left\|\mathbf{V}_{k,i}^H\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{P}_{i,k}^{1/2}\right\|_F^2+d.\end{aligned}\tag{41}$$

Since $\|\mathbf{X}\|_F^2$ is convex with respect to (w.r.t) $\mathbf{X}$, $\tilde{R}_{i,k}\left(\mathbf{P}_{i,k},\mathbf{A},\mathbb{Q},\mathbf{V}_{k,i}\right)$ is a concave function for each set of matrices $\mathbf{A}$, $\mathbf{P}_{i,k}$, $\mathbb{Q}$ and $\mathbf{V}_{k,i}$ when the other three are given. ∎

From (18), the indication variable, $\left\|\sum_{i=1}^{I}a_{i,j}\right\|_0$, is not smooth due to the l_0-norm, which can be approximated via a sequence of weighted l_1-norm minimizations in compressive sensing [40]. As a result, the l_0-norm in (18) can be approximated as

$$\left\|\sum_{i=1}^{I}a_{i,j}\right\|_0=\delta_j^{(n)}\sum_{i=1}^{I}a_{i,j},\tag{42}$$

where $\delta_j^{(n)}$ is the weight factor of the j-th HAP at the n-th iteration. $\delta_j^{(n)}$ can be updated by

$$\delta_j^{(n)}=\frac{1}{\sum\limits_{i=1}^{I}a_{i,j}^{(n)}+\nu},\quad\forall j\in\mathcal{K},\tag{43}$$

where ν is a small constant regulation parameter and $a_{i,j}^{(n)}$ is the value of $a_{i,j}$ in the n-th iteration.

Similarly, the WMMSE method can also be applied to the LEO SAT for receiving signals from HAPs. The MSE matrix of HAP k can be given as

$$\begin{aligned}\mathbf{E}_k&=E_{\mathbf{s},\mathbf{n}}\left[\left(\hat{\mathbf{s}}_k-\mathbf{s}_k\right)\left(\hat{\mathbf{s}}_k-\mathbf{s}_k\right)^H\right]\\&=\left(\mathbf{U}_k^H\mathbf{H}_{S,k}\mathbf{W}_k-\mathbf{I}_b\right)\left(\mathbf{U}_k^H\mathbf{H}_{S,k}\mathbf{W}_k-\mathbf{I}_b\right)^H\\&\quad+\sum_{j=1,j\neq k}^{K}\left\|\sum_{i=1}^{I}a_{i,j}\right\|_0\mathbf{U}_j^H\mathbf{H}_{S,j}\mathbf{W}_j\mathbf{W}_j^H\mathbf{H}_{S,j}^H\mathbf{U}_j\\&\quad+\sigma_S^2\mathbf{U}_k^H\mathbf{U}_k,\end{aligned}\tag{44}$$

where $\mathbf{U}_k\in\mathbb{C}^{M_S\times b}$ is the receive precoding matrix for the kth HAP at the LEO SAT. Therefore, by using the results in (33)-(41), and the approximation in (42), the rate expression in (17) can be reformulated as

$$\begin{aligned}&\tilde{R}_k\left(\mathbf{G}_k,\mathbf{A},\mathbb{Q},\mathbf{U}_k\right)\\&=\log\det\left(\mathbf{G}_k\right)-\sigma_S^2\mathrm{Tr}\left(\mathbf{U}_k^H\mathbf{U}_k\mathbf{G}_k\right)\\&\quad-\left\|\left(\mathbf{U}_k^H\mathbf{H}_{S,k}\mathbf{W}_k-\mathbf{I}_b\right)\mathbf{G}_k^{1/2}\right\|_F^2+b\\&\quad-\sum_{m=1,m\neq k}^{K}\delta_m^{(n)}\left(\sum_{i=1}^{I}a_{i,m}\right)\left\|\mathbf{U}_k^H\mathbf{H}_{S,m}\mathbf{W}_m\mathbf{G}_k^{1/2}\right\|_F^2,\end{aligned}\tag{45}$$

where $\{\mathbf{G}_k\succeq 0|k\in\mathcal{K}\}$ is the set of auxiliary matrices.

Similarly, the optimal $\mathbf{U}_k$, $\mathbf{G}_k$ and $\mathbf{E}_k$ for $\tilde{R}_k\left(\mathbf{G}_k,\mathbb{W},\mathbf{U}_k\right)$ to achieve the maximum data rate are given by

$$\mathbf{U}_k^*=\mathbf{\Xi}_k^{-1}\mathbf{H}_{S,k}\mathbf{W}_k,\tag{46}$$

$$\mathbf{G}_k^*=\left(\mathbf{E}_k^*\right)^{-1},\tag{47}$$

and

$$\mathbf{E}_k^*=\mathbf{I}_b-\mathbf{W}_k^H\mathbf{H}_{S,k}^H\mathbf{\Xi}_k^{-1}\mathbf{H}_{S,k}\mathbf{W}_k,\tag{48}$$

where $\mathbf{\Xi}_k$ is denoted by

$$\begin{aligned}\mathbf{\Xi}_k&=\mathbf{H}_{S,k}\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_{S,k}^H+\sigma_S^2\mathbf{I}_{M_S}\\&\quad+\sum_{m=1,m\neq k}^{K}\delta_m^{(n)}\left(\sum_{i=1}^{I}a_{i,m}\right)\mathbf{H}_{S,m}\mathbf{W}_m\mathbf{W}_m^H\mathbf{H}_{S,m}^H.\end{aligned}\tag{49}$$

According to (41) and (45), and relaxing the integer constraints into $[0,1]$, the association optimization problem in (32) with fixed $(\mathbb{Q},\mathbb{W},\mathcal{L},\mathcal{F}_U,\mathcal{F}_S)$ can be formulated as

$$\begin{aligned}\min_{\mathbf{A}}\ E_W&=w_G\sum_{i=1}^{I}\left(\sum_{k=1}^{K}a_{i,k}\frac{\beta B_i}{\tilde{R}_{i,k}}\|\mathbf{Q}_{i,k}\|_F^2\right)\\&\quad+w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K}a_{i,k}\kappa_{U,k}\alpha_i l_i f_{k,i}^2\right)\\&\quad+w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K}a_{i,k}\frac{\beta\left(B_i-l_i\right)}{\tilde{R}_k}\|\mathbf{W}_k\|_F^2\right)\end{aligned}\tag{50}$$

$$s.t.\ \sum_{k=1}^{K}a_{i,k}\left(\frac{\beta B_i}{\tilde{R}_{i,k}}+\frac{\alpha_i l_i}{f_{k,i}}\right)\leq T_{\max,i},\quad\forall i\in\mathcal{I},\tag{50a}$$

$$\begin{aligned}&\sum_{k=1}^{K}a_{i,k}\left(\frac{\beta B_i}{\tilde{R}_{i,k}}+\frac{\beta\left(B_i-l_i\right)}{\tilde{R}_k}\right)+\frac{\alpha_i\left(B_i-l_i\right)}{f_{s,i}}\\&\quad+2T_p(k,s)\leq T_{\max,i},\quad\forall i\in\mathcal{I},\end{aligned}\tag{50b}$$

$$\begin{aligned}&0\leq a_{i,k}\leq 1,\quad\forall i\in\mathcal{I},\ \forall k\in\mathcal{K},\\&(32a),(32b),(32e).\end{aligned}\tag{50c}$$

Although the transformation of rate functions facilitates the analysis, the objective function in (50), the constraints of (50a) and (50b) still belong to the case of sum of ratios and are difficult to handle. Luckily, according to [28] and [41], the method of quadratic transform can be adopted to deal with the sum of fractional terms in problem (50). From [41], we have the following Theorem.

Theorem 2: By applying quadratic transform to the fractional terms, $a_{i,k}\Big/\tilde{R}_{i,k}$ and $a_{i,k}\Big/\tilde{R}_{k}$, problem (50) can be equivalently rewritten as

$$\min_{\mathbf{A},\mathbf{E},\mathbf{T}} \bar{E}_W \tag{51}$$

$$s.t.\ \sum_{k=1}^{K}\beta B_i\left(2e_{i,k}\sqrt{a_{i,k}}-e_{i,k}^2\tilde{R}_{i,k}\right)+\sum_{k=1}^{K}a_{i,k}\frac{\alpha_i l_i}{f_{k,i}}\le T_{\max,i},\quad \forall i\in\mathcal{I}, \tag{51a}$$

$$\sum_{k=1}^{K}\beta B_i\left(2e_{i,k}\sqrt{a_{i,k}}-e_{i,k}^2\tilde{R}_{i,k}\right)+\sum_{k=1}^{K}\beta\left(B_i-l_i\right)\left(2t_{i,k}\sqrt{a_{i,k}}-t_{i,k}^2\tilde{R}_{k}\right)+\frac{\alpha_i\left(B_i-l_i\right)}{f_{s,i}}+2T_p(k,s)\le T_{\max,i},\quad \forall i\in\mathcal{I}$$
$$(32a),(32b),(32e),(50c) \tag{51b}$$

where the objective function $\bar{E}_W$ is given in (52), shown at the bottom of the page, $e_{i,k}$ and $t_{i,k}$ are real auxiliary variables, $\mathbf{E}=\{e_{i,k}|\forall i\in\mathcal{I},\forall k\in\mathcal{K}\}$ and $\mathbf{T}=\{t_{i,k}|\forall i\in\mathcal{I},\forall k\in\mathcal{K}\}$.

Given the precoding matrices, $\mathbb{Q}$ and $\mathbb{W}$, and the association strategy, $\mathbf{A}^{(n)}$, at the n-th iteration, the optimal $e_{i,k}$ and $t_{i,k}$ at the n-th iteration can be updated respectively by

$$e_{i,k}^{(n)*}=\frac{\sqrt{a_{i,k}^n}}{\tilde{R}_{i,k}\left(\mathbf{P}_{i,k},\mathbf{A}^{(n)},\mathbb{Q},\mathbf{V}_{k,i}\right)}, \tag{53}$$

and

$$t_{i,k}^{(n)*}=\frac{\sqrt{a_{i,k}^n}}{\tilde{R}_{k}\left(\mathbf{G}_{k},\mathbf{A}^{(n)},\mathbb{W},\mathbf{U}_{k}\right)}. \tag{54}$$

It should be noted that $\left(2e_{i,k}\sqrt{a_{i,k}}-e_{i,k}^2\tilde{R}_{i,k}\right)$ and $\left(2t_{i,k}\sqrt{a_{i,k}}-t_{i,k}^2\tilde{R}_{k}\right)$ in problem (50) can be reformulated as $\left(\left(-e_{i,k}^2\tilde{R}_{i,k}\right)-\left(-2e_{i,k}\sqrt{a_{i,k}}\right)\right)$ and $\left(\left(-t_{i,k}^2\tilde{R}_{k}\right)-\left(-2t_{i,k}\sqrt{a_{i,k}}\right)\right)$, which is a form of subtraction of two convex functions. Meanwhile, the optimization problem involving difference of convex functions (DC) is called DC programming problem, and can be solved efficiently by the DCA method. For simplicity, let $f\left(a_{i,k}\right)=2e_{i,k}\sqrt{a_{i,k}}$ and $g\left(a_{i,k}\right)=2t_{i,k}\sqrt{a_{i,k}}$. Then, the first order Taylor expansion of $f\left(a_{i,k}\right)$ and $g\left(a_{i,k}\right)$ around a feasible point $a_{i,k}^n$ at the n-th iteration can be written as $f\left(a_{i,k},a_{i,k}^n\right)=e_{i,k}\left(\sqrt{a_{i,k}^n}+a_{i,k}\Big/\sqrt{a_{i,k}^n}\right)$ and $g\left(a_{i,k},a_{i,k}^n\right)=t_{i,k}\left(\sqrt{a_{i,k}^n}+a_{i,k}\Big/\sqrt{a_{i,k}^n}\right)$, respectively. Combining (53) and (54), $f\left(a_{i,k},a_{i,k}^n\right)$ and $g\left(a_{i,k},a_{i,k}^n\right)$ can be reformulated at the n-th iteration respectively as

$$f\left(a_{i,k},a_{i,k}^n\right)=\frac{a_{i,k}^n+a_{i,k}}{\tilde{R}_{i,k}\left(\mathbf{P}_{i,k},\mathbf{A}^{(n)},\mathbb{Q},\mathbf{V}_{k,i}\right)}, \tag{55}$$

and

$$g\left(a_{i,k},a_{i,k}^n\right)=\frac{a_{i,k}^n+a_{i,k}}{\tilde{R}_{k}\left(\mathbf{G}_{k},\mathbf{A}^{(n)},\mathbb{W},\mathbf{U}_{k}\right)}. \tag{56}$$

Therefore, problem (51) can be transformed to a new optimization problem at the $(n+1)$-th iteration as

$$\min_{\mathbf{A},\mathbf{E},\mathbf{T}} \tilde{E}_W \tag{57}$$

$$s.t.\ \sum_{k=1}^{K}\beta B_i\left(f\left(a_{i,k},a_{i,k}^n\right)-e_{i,k}^2\tilde{R}_{i,k}\right)+\sum_{k=1}^{K}\frac{a_{i,k}\alpha_i l_i}{f_{k,i}}\le T_{\max,i},\quad \forall i\in\mathcal{I}, \tag{57a}$$

$$\sum_{k=1}^{K}\beta B_i\left(f\left(a_{i,k},a_{i,k}^n\right)-e_{i,k}^2\tilde{R}_{i,k}\right)+\sum_{k=1}^{K}\beta\left(B_i-l_i\right)\left(g\left(a_{i,k},a_{i,k}^n\right)-t_{i,k}^2\tilde{R}_{k}\right)+\frac{\alpha_i\left(B_i-l_i\right)}{f_{s,i}}+2T_p(k,s)\le T_{\max,i},\quad \forall i\in\mathcal{I},$$
$$(32a),(32b),(32e),(50c) \tag{57b}$$

where $\tilde{E}_W$ is expressed as (58), shown at the bottom of the page.

It can be noted that problem (57) is convex with respect to $\mathbf{A}$. Therefore, problem (57) can be solved by using the interior point method. Accordingly, a DCA-method based optimization algorithm can solve problem (50), which is summarized in Algorithm 1.

$$\bar{E}_W=w_G\sum_{i=1}^{I}\left(\sum_{k=1}^{K}\beta B_i\left\|\mathbf{Q}_{i,k}\right\|_F^2\left(2e_{i,k}\sqrt{a_{i,k}}-e_{i,k}^2\tilde{R}_{i,k}\right)\right)+w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K}a_{i,k}\kappa_{U,k}\alpha_i l_i f_{k,i}^2\right)+w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K}\beta\left(B_i-l_i\right)\left\|\mathbf{W}_k\right\|_F^2\left(2t_{i,k}\sqrt{a_{i,k}}-t_{i,k}^2\tilde{R}_{k}\right)\right) \tag{52}$$

$$\tilde{E}_W=w_G\sum_{i=1}^{I}\left(\sum_{k=1}^{K}\beta B_i\left\|\mathbf{Q}_{i,k}\right\|_F^2\left(f\left(a_{i,k},a_{i,k}^n\right)-e_{i,k}^2\tilde{R}_{i,k}\right)\right)+w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K}a_{i,k}\kappa_{U,k}\alpha_i l_i f_{k,i}^2\right)+w_H\sum_{i=1}^{I}\left(\sum_{k=1}^{K}\beta\left(B_i-l_i\right)\left\|\mathbf{W}_k\right\|_F^2\left(g\left(a_{i,k},a_{i,k}^n\right)-t_{i,k}^2\tilde{R}_{k}\right)\right) \tag{58}$$

Algorithm 1 DCA Based Algorithm for Solving GUE Association Problem (50)

1: **Initialization**
Initialize a feasible $\mathbf{A}^{(0)}$ for problem (32) with fixed $\mathbb{Q}$, $\mathbb{W}$, $\mathcal{L}$, $\mathcal{F}_U$, $\mathcal{F}_S$, set $n=0$ and maximum number of iterations: N_{max}
2: **repeat**
3: Given $\mathbf{A}^{(n)}$, update $\delta_j^{(n)}$ with (43).
4: With $\mathbf{A}^{(n)}$ and $\delta_j^{(n)}$, calculate $\tilde{R}_{i,k}$ and $\tilde{R}_k$ as in (41) and (45), respectively.
5: Update $e_{i,k}^n$ and $t_{i,k}^n$ by (53) and (54).
6: With $e_{i,k}^n$ and $t_{i,k}^n$, update $f\left(a_{i,k}, a_{i,k}^n\right)$ and $g\left(a_{i,k}, a_{i,k}^n\right)$ by (55) and (56).
7: Obtain $\mathbf{A}^{(n+1)}$ by solving problem (57) using interior point method.
8: Set $n=n+1$.
9: **Until**: the objective function (50) converges or $n=N_{max}$ and output the optimal solution $\mathbf{A}^{(n)}$.

B. Transmit Precoding Design for MU-MIMO Computation Offloading

After obtaining the GUE association strategy $\mathbf{A}$, we can tackle the transmit precoding subproblem for GUE-HAP and HAP-LEO SAT communication links, respectively. Given association strategy $\mathbf{A}$ and fixed $\mathcal{L}$, $\mathcal{F}_U$, and $\mathcal{F}_S$, the transmit precoding optimization for MU-MIMO computation offloading in problem (32) can be reformulated as

$$\begin{aligned}
\min_{\mathbb{Q},\mathbb{W}}\ & w_G \sum_{k\in\bar{\mathcal{K}}}\sum_{i\in\mathcal{I}_k}\frac{\beta B_i}{R_{i,k}}\|\mathbf{Q}_{i,k}\|_F^2 \\
& + w_H \sum_{k\in\bar{\mathcal{K}}}\sum_{i\in\mathcal{I}_k}\frac{\beta\left(B_i-l_i\right)}{R_k}\|\mathbf{W}_k\|_F^2 \qquad (59)\\
s.t.\ & \|\mathbf{Q}_{i,k}\|_F^2 \le p_{\max,i}, \quad \forall i\in\mathcal{I}_k,\ k\in\bar{\mathcal{K}}, \qquad (59a)\\
& \|\mathbf{W}_k\|_F^2 \le p_{\max,k}, \quad \forall k\in\bar{\mathcal{K}}, \qquad (59b)\\
& \frac{\beta B_i}{R_{i,k}}+\frac{\alpha_i l_i}{f_{k,i}} \le T_{\max,i}, \quad \forall i\in\mathcal{I}_k,\ k\in\bar{\mathcal{K}}, \qquad (59c)\\
& \frac{\beta B_i}{R_{i,k}}+\frac{\beta\left(B_i-l_i\right)}{R_k}+\frac{\alpha_i\left(B_i-l_i\right)}{f_{s,i}} \\
& +\ 2T_p(k,s) \le T_{\max,i}, \quad \forall i\in\mathcal{I}_k,\ k\in\bar{\mathcal{K}}, \qquad (59d)
\end{aligned}$$

where $\mathcal{I}_k=\{i\in\mathcal{I}|a_{i,k}=1\}$ is the set of the GUEs associated with HAP k and $\bar{\mathcal{K}}=\left\{k\in\mathcal{K}\middle|\left\|\sum_{i=1}^{I}a_{i,j}\right\|_0=1\right\}$ is the set of the HAPs that has associated GUEs.

Obviously, problem (59) can be classified as the sum-of-ratios problem, which is non-convex and difficult to handle. Fortunately, the quadratic transform method proposed in [41] can be used here to tackle the sum-of-ratios problem. In this subsection, we solve the precoding matrices optimization iteratively.

According to Theorem 2, by using quadratic transform [41], problem (59) can be equivalently transformed into the following problem.

$$\begin{aligned}
\min_{\mathbb{Q},\mathbb{W},\mathbf{z},\mathbf{c}}\ & w_G \sum_{k\in\bar{\mathcal{K}}}\sum_{i\in\mathcal{I}_k}\beta B_i\left(2z_{i,k}\sqrt{\|\mathbf{Q}_{i,k}\|_F^2}-z_{i,k}^2R_{i,k}\right) \\
& + w_H\sum_{k\in\bar{\mathcal{K}}}\sum_{i\in\mathcal{I}_k}\beta(B_i-l_i)\left(2c_{i,k}\sqrt{\|\mathbf{W}_k\|_F^2}-c_{i,k}^2R_k\right) \qquad (60)
\end{aligned}$$

subject to constraints (59a)-(59d), where $\mathbf{z}=\{z_{i,k}\}_{i\in\mathcal{I}_k,k\in\bar{\mathcal{K}}}$ and $\mathbf{c}=\{c_{i,k}\}_{i\in\mathcal{I}_k,k\in\bar{\mathcal{K}}}$ are the collection of auxiliary variables. From (60), $\sqrt{\|\mathbf{Q}_{j,k}\|_F^2}$ and $\sqrt{\|\mathbf{W}_k\|_F^2}$ are convex. Meanwhile, constraints (59a) and (59b) are convex too. Therefore, the convexity of problem (60) depends on the concavity of $R_{i,k}$ and R_k. According to the WMMSE transformation to $R_{i,k}$ and R_k in subsection III-A, $R_{i,k}$ and R_k can be recast respectively as

$$\begin{aligned}
&\tilde{R}_{i,k}\left(\mathbf{P}_{i,k},\mathbb{Q},\mathbf{V}_{k,i}\right) \\
&=\log\det\left(\mathbf{P}_{i,k}\right)-\sigma_k^2\mathrm{Tr}\left(\mathbf{V}_{k,i}^H\mathbf{V}_{k,i}\mathbf{P}_{i,k}\right)+d \\
&\quad-\left\|\left(\mathbf{V}_{k,i}^H\mathbf{H}_{k,i}\mathbf{Q}_{i,k}-\mathbf{I}_d\right)\mathbf{P}_{i,k}^{1/2}\right\|_F^2 \\
&\quad-\sum_{(j,m)\neq(i,k)}\left\|\mathbf{V}_{k,i}^H\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{P}_{i,k}^{1/2}\right\|_F^2, \\
&\quad\forall i,j\in\mathcal{I}_k,\quad\forall k,m\in\bar{\mathcal{K}}, \qquad (61)
\end{aligned}$$

and

$$\begin{aligned}
&\tilde{R}_k\left(\mathbf{G}_k,\mathbb{W},\mathbf{U}_k\right) \\
&=\log\det\left(\mathbf{G}_k\right)-\sigma_S^2\mathrm{Tr}\left(\mathbf{U}_k^H\mathbf{U}_k\mathbf{G}_k\right) \\
&\quad-\left\|\left(\mathbf{U}_k^H\mathbf{H}_{S,k}\mathbf{W}_k-\mathbf{I}_b\right)\mathbf{G}_k^{1/2}\right\|_F^2 \\
&\quad-\sum_{m\in\bar{\mathcal{K}},m\neq k}\left\|\mathbf{U}_k^H\mathbf{H}_{S,m}\mathbf{W}_m\mathbf{G}_k^{1/2}\right\|_F^2+b,\quad\forall k\in\bar{\mathcal{K}}. \qquad (62)
\end{aligned}$$

$\tilde{R}_{i,k}$ in (61) and $\tilde{R}_k$ in (62) are concave w.r.t $\mathbb{Q}$ and $\mathbb{W}$, respectively. Accordingly, problem (60) can be transformed into

$$\begin{aligned}
\min_{\mathbb{Q},\mathbb{W},\mathbf{z},\mathbf{c}}\ & w_G \sum_{k\in\bar{\mathcal{K}}}\sum_{i\in\mathcal{I}_k}\beta B_i\left(2z_{i,k}\sqrt{\|\mathbf{Q}_{i,k}\|_F^2}-z_{i,k}^2\tilde{R}_{i,k}\right) \\
& + w_H\sum_{k\in\bar{\mathcal{K}}}\sum_{i\in\mathcal{I}_k}\beta\left(B_i-l_i\right)\left(2c_{i,k}\sqrt{\|\mathbf{W}_k\|_F^2}-c_{i,k}^2\tilde{R}_k\right) \qquad (63)\\
s.t.\ & \frac{\beta B_i}{\tilde{R}_{i,k}}+\frac{\alpha_i l_i}{f_{k,i}} \le T_{\max,i}, \quad \forall i\in\mathcal{I}_k,\ k\in\bar{\mathcal{K}}, \qquad (63a)\\
& \frac{\beta B_i}{\tilde{R}_{i,k}}+\frac{\beta\left(B_i-l_i\right)}{\tilde{R}_k}+\frac{\alpha_i\left(B_i-l_i\right)}{f_{s,i}} \\
& +\ 2T_p(k,s) \le T_{\max,i}, \quad \forall i\in\mathcal{I}_k,\ k\in\bar{\mathcal{K}} \\
& (59a),(59b). \qquad (63b)
\end{aligned}$$

Now, problem (59) is turned into a convex problem in (63). Accordingly, the optimal $\mathbb{Q}$ and $\mathbb{W}$ can be efficiently solved by using the standard numerical method [42]. When $\mathbb{Q}$ and $\mathbb{W}$ are fixed, the optimal $z_{i,k}$ and $c_{i,k}$ are given respectively by

$$z_{i,k}^*=\frac{\sqrt{\|\mathbf{Q}_{i,k}\|_F^2}}{\tilde{R}_{i,k}}, \qquad (64)$$

and

$$c_{i,k}^{*}=\frac{\sqrt{\|\mathbf{W}_k\|_F^2}}{\tilde{R}_k}. \tag{65}$$

The whole iterative optimization algorithm for solving problem (59) is summarized in Algorithm 2.

Algorithm 2 Iterative Optimization Algorithm to Solve Problem (59)

1: **Initialization**
Initialize $\mathbb{Q}^{(0)}$, $\mathbb{W}^{(0)}$ into feasible values and set $n=0$ and maximum number of iterations: N_{max}.
2: **repeat**
3: Given $\mathbb{Q}^{(n)}$ and $\mathbb{W}^{(n)}$, update $\mathbf{z}$ and $\mathbf{c}$ by (64) and (65), respectively.
4: Update $\mathbb{Q}^{(n+1)}$ and $\mathbb{W}^{(n+1)}$ by solving problem (63).
5: Set $n=n+1$.
6: **Until**: the objective function (59) converges or $n=N_{max}$, and output the optimal solution $\mathbb{Q}^{(n)}$ and $\mathbb{W}^{(n)}$.

C. Computation Task Assignment

For problem (32) with given GUE association strategy $\mathbf{A}$, precoding matrices $\mathbb{Q}$ and $\mathbb{W}$, and computation resource allocation, the computation task assignment subproblem can be formulated as

$$\begin{aligned}\min_{\mathcal{L}}\ & w_H\sum_{k\in\bar{\mathcal{K}}}\sum_{j\in\mathcal{I}_k}\kappa_{U,k}\alpha_j l_j f_{k,j}^2\\ &+w_H\sum_{k\in\bar{\mathcal{K}}}\sum_{j\in\mathcal{I}_k}\frac{\beta\left(B_j-l_j\right)}{R_k}\|\mathbf{W}_k\|_F^2\\ &+w_S\sum_{k\in\bar{\mathcal{K}}}\sum_{j\in\mathcal{I}_k}\kappa_S\alpha_j\left(B_j-l_j\right)f_{s,j}^2 \qquad (66)\\ s.t.\ & 0\le l_j\le B_j,\quad \forall j\in\mathcal{I}_k,\ \forall k\in\bar{\mathcal{K}},\\ & (59c),(59d). \qquad (66a)\end{aligned}$$

It can be observed that the objective function of (66) is linear w.r.t $\mathcal{L}$ and constraints (66a), (59c), and (59d) confine the feasible region of l_i. According to (59c), we have

$$l_j\le\frac{T_{\max,j}-\frac{\beta B_j}{R_{j,k}}}{\frac{\alpha_j}{f_{total,k}}}=l_j^U,\quad\forall j\in\mathcal{I}_k,\ \forall k\in\bar{\mathcal{K}} \tag{67}$$

Similarly, from (59d), we have

$$l_j\ge B_j-\frac{T_{\max,j}-2T_p(k,s)-\frac{\beta B_j}{R_{j,k}}}{\frac{\beta}{R_k}+\frac{\alpha_j}{f_{total,s}}}=l_j^L,\quad\forall j\in\mathcal{I}_k,\ \forall k\in\bar{\mathcal{K}} \tag{68}$$

Therefore, the upper bound and lower bound of $l_j,\forall j\in\mathcal{I}_k,\forall k\in\mathcal{K}$ can be given respectively as

$$l_{j,up}=\min\left\{B_j,l_j^U\right\}, \tag{69}$$

and

$$l_{j,low}=\max\left\{0,l_j^L\right\}. \tag{70}$$

Hence, problem (66) can be reformulated as

$$\begin{aligned}\min_{\mathcal{L}}\ & w_H\sum_{k=1}^{K}\sum_{j\in\mathcal{I}_k}\kappa_{U,k}\alpha_j l_j f_{k,j}^2\\ &+w_H\sum_{k=1}^{K}\sum_{j\in\mathcal{I}_k}\frac{\beta\left(B_j-l_j\right)}{R_k}\|\mathbf{W}_k\|_F^2\\ &+w_S\sum_{i=1}^{I}\kappa_S\alpha_i\left(B_i-l_i\right)f_{s,i}^2 \qquad (71a)\\ s.t.\ & l_{j,\min}\le l_j\le l_{j,\max},\quad\forall j\in\mathcal{I}_k,\ k\in\bar{\mathcal{K}}, \qquad (71b)\end{aligned}$$

Obviously, problem (71) is a convex problem and can be solved efficiently by the interior point method. The algorithm to solve problem (71) is summarized in Algorithm 3.

Algorithm 3 Iterative Optimization Algorithm to Solve Problem (71)

1: **Initialization**
Initialize $\mathcal{L}^{(0)}$ for problem (71) with fixed $\mathbf{A}$, $\mathcal{F}_U$, $\mathcal{F}_S$, $\mathbb{Q}$ and $\mathbb{W}$, and set $n=0$ and maximum number of iterations: N_{max}.
2: **repeat**
3: Update $l_{j,\min}$ and $l_{j,\max}$ according to (69) and (70).
4: Update $\mathcal{L}^{(n+1)}$ by solving problem (71) using interior point method.
5: Set $n=n+1$.
6: **Until**: the objective function (71) converges or $n=N_{max}$ and output the optimal solution $\mathcal{L}^{(n)}$.

D. Computation Resource Allocation

For problem (32) with fixed GUE association strategy $\mathbf{A}$, precoding matrices $\mathbb{Q}$ and $\mathbb{W}$, and computation task assignment $\mathcal{L}$. Problem (32) can be reformulated as

$$\begin{aligned}\min_{\mathcal{F}_U,\mathcal{F}_S}\ & w_H\sum_{k\in\bar{\mathcal{K}}}\sum_{j\in\mathcal{I}_k}\kappa_{U,k}\alpha_j l_j f_{k,j}^2\\ &+w_S\sum_{i=1}^{I}\kappa_S\alpha_i\left(B_i-l_i\right)f_{s,i}^2 \qquad (72)\\ s.t.\ & \sum_{j\in\mathcal{I}_k}f_{k,j}\le f_{total,k},\quad\forall k\in\bar{\mathcal{K}}, \qquad (72a)\\ & \sum_{i=1}^{I}f_{s,i}\le f_{total,s}, \qquad (72b)\\ & f_{k,j}\ge f_{k,j}^{\min},\quad\forall k\in\bar{\mathcal{K}},\ \forall j\in\mathcal{I}_k, \qquad (72c)\\ & f_{s,j}\ge f_{s,j}^{\min},\quad\forall j\in\mathcal{I}_k, \qquad (72d)\end{aligned}$$

where $f_{k,j}^{\min}=\alpha_j l_j/(T_{\max,j}-\beta B_j/R_{j,k})$ and $f_{s,j}^{\min}=\frac{\alpha_j(B_j-l_j)}{T_{\max,j}-2T_p(k,s)-\beta B_j/R_{j,k}-\beta(B_j-l_j)/R_k}$. Obviously, problem (72) is a convex problem w.r.t $\mathcal{F}_U$ and $\mathcal{F}_S$. In this section, we solve the computation resource allocation problem in two steps. Firstly, the computation resource allocation strategy in the HAPs will be studied. Then, we tackle the computation resource allocation problem for the LEO SAT.

1) Computation Resource Allocation for HAPs: Given $\mathcal{F}_S$, the objective function of problem (72) monotonically increases with $\{f_{k,j}\}_{j\in\mathcal{I}_k}$ and problem (72) can be decomposed into $\left|\bar{\mathcal{K}}\right|$ subproblems since both the objective function and constraints can be decoupled. For HAP k, the computation resource allocation problem can be written as

$$\min_{\{f_{k,j}\}_{j\in\mathcal{I}_k}} \quad w_H \sum_{j\in\mathcal{I}_k} \kappa_{U,k}\alpha_j l_j f_{k,j}^2 \tag{73}$$

subject to constraints (72a) and (72c). It can be observed that problem (73) is a quadratic convex problem and the objective function of problem (73) is a monotonic increasing function w.r.t $\{f_{k,j}\}_{j\in\mathcal{I}_k}$. The optimal $f_{k,j}^*$ can be obtained at its lower bound. Its closed-form solution is given by

$$f_{k,j}^* = f_{k,j}^{\min}, \quad \forall j \in \mathcal{I}_k \tag{74}$$

2) Computation Resource Allocation for SAT: Given $\mathcal{F}_U$, the computation resource allocation problem for the LEO SAT is also a quadratic convex problem. Similarly, the closed-form solution of $f_{s,j}$ can also be derived as

$$f_{s,j}^* = f_{s,j}^{\min}, \quad \forall j \in \mathcal{I}_k \tag{75}$$

E. Proposed Iterative Optimization Algorithm and Analysis

The overall algorithm for solving problem (32) is summarized in Algorithm 4, which iteratively optimizes GUE association, transmit precoding, and computation task assignment.

Algorithm 4 Iterative Optimization Algorithm for Solving Problem (32)

1: **Initialization**
2: Initialize feasible values of $\left(\mathbf{A}^{(0)}, \mathbb{Q}^{(0)}, \mathbb{W}^{(0)}, \mathcal{L}^{(0)}, \mathcal{F}_U^{(0)}, \mathcal{F}_S^{(0)}\right)$, tolerance $\varepsilon > 0$, set $n = 0$ and maximum iteration number $N_{\max}$.
3: Compute the objective function value of (32) as $V_{obj}^{(0)} = E\left(\mathbf{A}^{(0)}, \mathbb{Q}^{(0)}, \mathbb{W}^{(0)}, \mathcal{L}^{(0)}, \mathcal{F}_U^{(0)}, \mathcal{F}_S^{(0)}\right)$.
4: **repeat**
5: Given $\left(\mathbb{Q}^{(n)}, \mathbb{W}^{(n)}, \mathcal{L}^{(n)}, \mathcal{F}_U^{(n)}, \mathcal{F}_S^{(n)}\right)$, obtain the optimal $\mathbf{A}^{(n+1)}$ of problem (57) using Algorithm 1.
6: Given $\left(\mathbf{A}^{(n+1)}, \mathcal{L}^{(n)}, \mathcal{F}_U^{(n)}, \mathcal{F}_S^{(n)}\right)$, obtain the optimal $\mathbb{Q}^{(n+1)}, \mathbb{W}^{(n+1)}$ of problem (59) using Algorithm 2.
7: Given $\left(\mathbf{A}^{(n+1)}, \mathbb{Q}^{(n+1)}, \mathbb{W}^{(n+1)}, \mathcal{F}_U^{(n)}, \mathcal{F}_S^{(n)}\right)$, obtain the optimal $\mathcal{L}^{(n+1)}$ of problem (66) using Algorithm 3.
8: Given $\left(\mathbf{A}^{(n+1)}, \mathbb{Q}^{(n+1)}, \mathbb{W}^{(n+1)}, \mathcal{L}^{(n+1)}\right)$, obtain the optimal $\mathcal{F}_U^{(n+1)}$ and $\mathcal{F}_S^{(n+1)}$ according to (74) and (75), respectively.
9: Set $n = n + 1$.
10: **Until** $\left|V_{obj}^{(n)} - V_{obj}^{(n-1)}\right| < \varepsilon$ or $n > N_{\max}$.

Theorem 3: The convergence of overall iterative Algorithm 4 is summarized in Theorem 3 and proved in Appendix B.

From Algorithm 4, the complexity of each iteration is mainly affected by solving the GUE association problem (51), precoding optimization problem (52), and task allocation problem (59).

To solve the GUE association problem, DCA based Algorithm 1 is adopted. The computation complexity of Algorithm 1 mainly comes from Step 3 to Step 7 in the loop for L_1 iterations. The complexity of updating $\delta_j^{(n)}$ is K. The complexity of updating $\tilde{R}_{i,k}$ and $\tilde{R}_k$ in Step 4 are IK and K, respectively. The complexity of Steps 5 and 6 is the same, which is IK. Since the number of optimized variables in problem (51) is IK, the complexity of solving problem (51) by using the standard interior point method (IPM) is $\mathcal{O}\left((IK)^3\right)$. Therefore, the complexity of Algorithm 1 is $\mathcal{O}\left(L_1 I^3 K^3\right)$ [43]. For Algorithm 2, its complexity lies in Step 4. The number of optimized variables in problem (56) is $\left(I\left|\bar{\mathcal{K}}\right|Nd + \left|\bar{\mathcal{K}}\right|Mb\right)$. Since problem (56) is also solved by IPM, the complexity of Algorithm 2 is $\mathcal{O}\left(L_2\left(I\left|\bar{\mathcal{K}}\right|Nd + \left|\bar{\mathcal{K}}\right|Mb\right)^3\right)$, where L_2 is the number of iterations when solving problem (56). Similarly, the complexity of Algorithm 3 is $\mathcal{O}\left(L_3 I^3\right)$, where L_3 is the number of iterations in solving problem (64).

The total complexity of Algorithm 4 is $\mathcal{O}\left(L_0 L_1 I^3 K^3 + L_0 L_2\left(I\left|\bar{\mathcal{K}}\right|Nd + \left|\bar{\mathcal{K}}\right|Mb\right)^3 + L_0 L_3 I^3\right)$, where L_0 is the number of outer iterations of Algorithm 4.

IV. Simulation and Results

In this section, we present numerical results to evaluate the proposed offloading schemes and algorithms on a Matlab-based simulator. We consider a $1000\,m \times 1000\,m$ square area, where GUEs are randomly distributed on the ground. Above the ground, 4 HAPs are randomly deployed in the stratosphere at the altitude of 20 km. Moreover, a LEO SAT is placed at the orbital altitude of 200 km. According to the ITU-R spectrum regulation, the HAP adopts 31 GHz bands on Ka band to communicate with the GUE and the LEO SAT, respectively, and the bandwidth for both links is 100 MHz. The background noise density is set to be -175 dBm/Hz. We assume that each GUE has the same maximum transmit power 0.3 W and the HAPs also have same maximum transmit power as 1.5 W. The task data size for each GUE is randomly chosen from 0.8 MB to 2 MB. The processing density of computing task is randomly chosen from 180.5 cycles/bit to 283.5 cycles/bit [44]. Unless stated otherwise, the default system parameters are set as follows: $N = d = 2$, $M = 16$, $M_S = 32$, $\beta = 1.2$, $T_{max,i} = 3$s for $\forall i \in \mathcal{I}$ and $\varepsilon = 10^{-3}$. Since the LEO SAT can carry larger weight (e.g., like hundred kilograms) than the HAP and support more advanced applications, we assume that the computing capability and efficiency of the LEO SAT are superior to the HAP. Therefore, we set $\kappa_{U,1} = \cdots = \kappa_{U,K} = 10^{-25}$ and $\kappa_S = 10^{-28}$. The computing capacity of each HAP is set as $f_{total,1} = \cdots = f_{total,K} = 2.5$ GHz, and the computing capacity of the LEO SAT is 10 GHz. The weight coefficients for the GUEs, the HAPs, and the LEO SAT are set as $w_G = 1$, $w_H = 0.8$, and $w_S = 0.5$, respectively.

For comparison, we consider the following offloading methods: 1) The proposed MU-MIMO computation offloading in satellite-aerial integrated computation offloading scheme

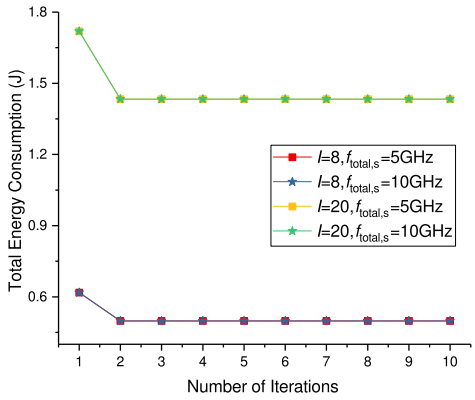


Fig. 2. Convergence of Algorithm 4.

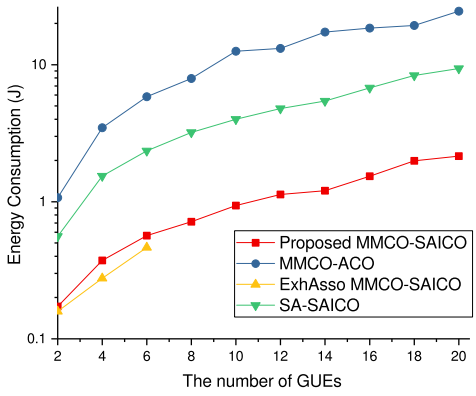


Fig. 4. Energy consumption versus the number of GUE.

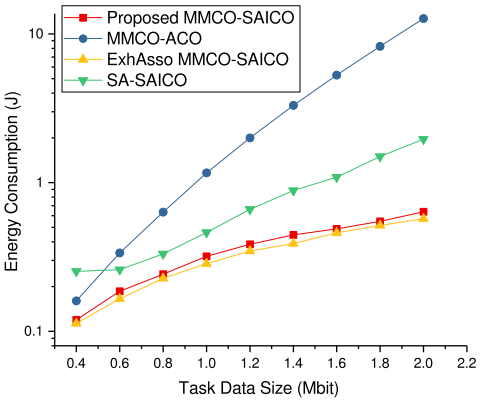


Fig. 3. Energy consumption versus the task data size.

(labelled as MMCO-SAICO). 2) The exhaustive search association based MMCO-SAICO (labelled as ExhAsso MMCO-SAICO), where all the possible GUE-HAP associations are tried. 3) The HAP only MU-MIMO computation offloading scheme (labelled as MMCO-ACO), where all the computation tasks are computed at the HAP. In this scheme, association variable, GUEs' precoding, and computation resource allocation are iteratively optimized like MMCO-SAICO. 4) The single antenna satellite-aerial integrated computation offloading scheme (labelled as SA-SAICO), where the GUE and the HAPs utilize single antenna for communication and the multiplexing scheme is FDMA.

Figure 2 plots the system's total energy consumption versus the number of iterations to verify the convergence of the proposed algorithm under different GUE numbers and satellite computing capacities. From the figure, the proposed algorithms for MMCO-SAICO converges quickly and only three iterations can ensure the convergence. From Figure 2, increasing the computing capacity of the LEO SAT does not affect the total energy consumption. The reason is that the proposed algorithm can minimize the energy consumption under the delay constraint without using additional computing resource.

In Figure 3, we simulate the energy consumption of different offloading schemes versus the task data size, where $I = 5$ and $N_H = 5$. From the figure, the energy consumption of all offloading schemes increases with the task data size. Since larger data size consumes more energy on data offloading and task computing, the performance of ExhAsso MMCO-SAICO and MMCO-SAICO is better than that of MMCO and SA-SAICO, which shows that the MU-MIMO computation offloading method and satellite-aerial integrated edge computing method can effectively reduce the energy consumption during computation offloading. Meanwhile, the performance gap between ExhAsso MMCO-SAICO and the proposed MMCO-SAICO is quite small, which illustrates the superiority of the proposed association strategy in MMCO-SAICO. Although MMCO-ACO utilizes MU-MIMO communication for task offloading, its energy consumption is greatly impacted by the low computing efficiency of the HAP and grows quickly as the task data size increases. The performance of SA-SAICO lies between MMCO-ACO and proposed MMCO-SAICO, which shows that SA-SAICO can exploit the advantage of satellite edge computing, but its performance is limited by the single antenna offloading method.

Figure 4 demonstrates the energy consumption versus the number of GUEs for different offloading schemes, where $N_H = 8$. From the figure, the energy consumption of all offloading schemes increases as the number of GUEs increases. Since the computing power of the HAP is weak, the MMCO-ACO method has the highest energy consumption. Meanwhile, the proposed MMCO-SAICO has the lowest energy consumption by using MU-MIMO computation offloading and strong computing power at satellite. Notably, the performance gap between the proposed MMCO-SAICO and ExhAsso MMCO-SAICO is quite small for GUE numbers of 2, 4 and 6 (Other values are omitted due to extremely huge search space). From the figure, the energy consumption of SA-SAICO increases quickly as the number of GUEs increases, because the available bandwidth for each GUE is correspondingly reduced and GUE/HAPs will cost more energy on task offloading as the number of GUEs increases.

Figure 5 plots the energy consumption (EC) versus the orbital altitude of satellite, where $I = 16$ and $N_H = 8$. From the figure, the proposed MMCO-SAICO has the lowest energy consumption. Moreover, the consumed energy is almost composed of GUE transmission energy consumption and remains almost unchanged versus the variation of the satellite altitude.

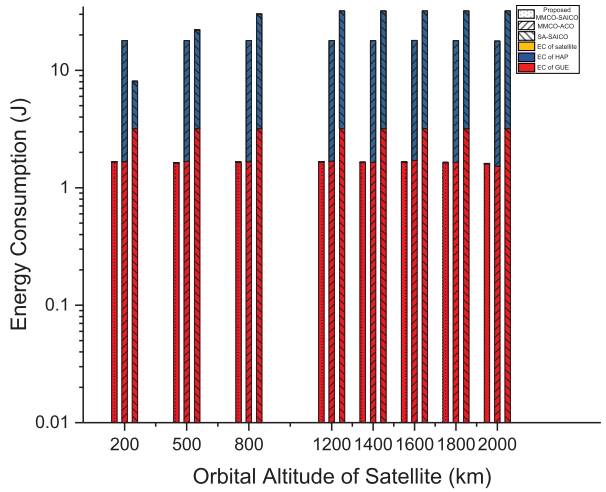


Fig. 5. Energy consumption versus the orbital altitude of satellite.

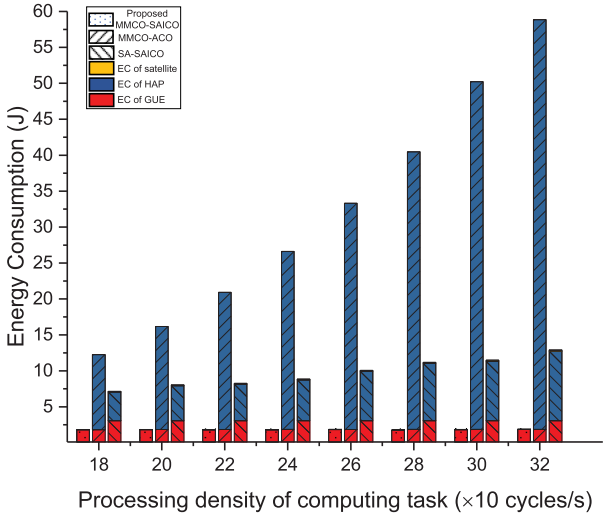


Fig. 7. Energy consumption versus the processing density of computing task.

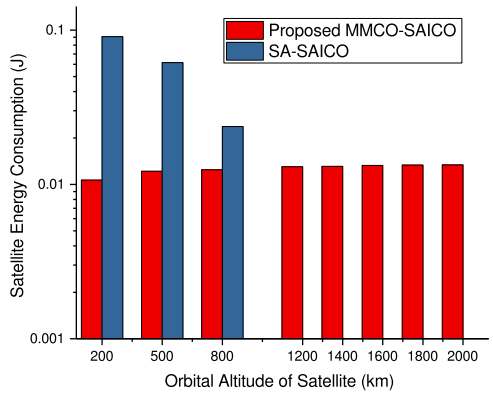


Fig. 6. Satellite energy consumption versus the orbital altitude of satellite.

Therefore, the application of MU-MIMO computation offloading for GUE/HAP and multi-antenna beam gain at satellite can provide high link capacity to combat severe pathloss, and the adoption of satellite edge computing can effectively reduce the computing energy consumption. Compared to MMCO-SAICO and MMCO-ACO, the GUE in SA-SAICO has larger transmission energy consumption. Moreover, as the orbital altitude of satellite increases, the energy consumption of SA-SAICO increases quickly and is larger than MMCO-ACO. From the picture, the increased energy consumption of SA-SAICO is mainly on the HAP because the single antenna HAP cannot combat severe HAP-satellite link pathloss and complete task offloading within the time delay constraint as the orbital altitude of satellite increases. Therefore, when the satellite altitude increases, the computing tasks gradually shift to the HAP in SA-SAICO. A clear illustration of the satellite energy consumption is shown in Figure 6. From Figure 6, the satellite energy consumption of SA-SAICO decreases with satellite altitude and reaches 0 when $H_S > 800$ km, because the pathloss increases with satellite altitude and the communication rate between the satellite and the HAP decreases quickly in SA-SAICO, and the offloading time delay constraint cannot be satisfied, which means the computation tasks will be shifted to the HAP to satisfy the delay constraint and the corresponding satellite energy consumption is 0. In contrast, the satellite energy consumption in the proposed MMCO-SAICO grows slowly and almost unchanged when $H_S > 1200$ km. Although the increased satellite altitude increases the pathloss, the HAP-satellite communication rate is still relatively high with MIMO communication and antenna gain in MMCO-SAICO. Moreover, compared with satellite computing, HAP computing has higher energy consumption. Hence, the computation tasks of GUEs are still mainly processed by the satellite. However, the increased satellite altitude increases the HAP-satellite offloading time delay. To make the overall offloading time delay satisfy the delay constraint, the satellite needs to allocate more computing resource to reduce computing time at the satellite side. Hence, the satellite energy consumption increases with satellite altitude and gradually reaches stability when all the computing resources are allocated.

In Figure 7, we plot the energy consumption versus the processing density of computing task, where $I = 16$ and $N_H = 8$. From the figure, the proposed MMCO-SAICO has the lowest energy consumption and maintains at a rather low level. Moreover, the energy is almost entirely consumed by the GUEs and the energy consumption of the HAP and the LEO SAT is very small. The HAP only computing method in MMCO-ACO has the highest energy consumption and grows exponentially with the processing density. Meanwhile, the performance of SA-SAICO is inferior to the proposed MMCO-SAICO and its HAP energy consumption grows as the processing density increases. A clear illustration of the satellite energy consumption is shown in Figure 8. The satellite energy consumption of SA-SAICO is apparently larger than the proposed MMCO-SAICO and grows quickly as the processing density increases. The satellite energy consumption of the proposed MMCO-SAICO grows slowly because the offloading delay of SA-SAICO is larger than that of the proposed MMCO-SAICO. To satisfy the time delay constraint, the satellite will use more computation resources in SA-SAICO to reduce the time delay, which causes higher energy consumption. The MMCO-SAICO has relative smaller offloading time delay by using MU-MIMO

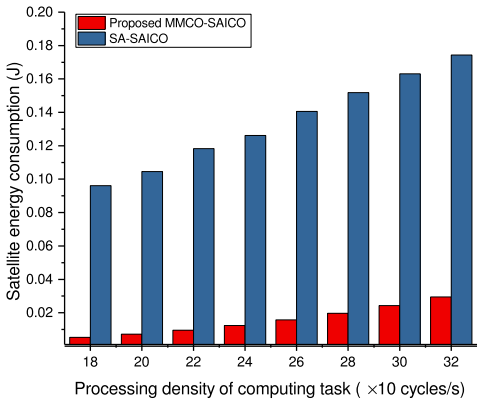


Fig. 8. Satellite energy consumption versus the processing density of computing task.

computation offloading, and the satellite uses less computation resource to reduce the computing energy consumption at the cost of longer processing delay.

V. Conclusion

In this paper, we have proposed the satellite-aerial integrated edge computing architecture to extend the application of MEC to space. We have formulated a weighted sum energy minimization problem by jointly considering GUE association, MU-MIMO precoding, computation task assignment, and computation resource allocation. To solve the complex non-convex optimization problem, we have decomposed the optimization problem into four subproblems and proposed an algorithm to solve the four subproblems iteratively. For the GUE association subproblem, we solve it via the quadratic transform based fractional programming and the DCA method, and the compressive sensing method is specially utilized to handle the l_0-norm. We optimize the MU-MIMO precoding by using the quadratic transform based fractional programming and WMMSE methods. The computation task assignment subproblem is solved via the interior point method and the computation resource allocation subproblem is derived in closed form. Simulation results show that the proposed satellite-aerial integrated edge computing architecture can provide computing services for GUE with relative low energy consumption and the MU-MIMO computation offloading method can increase offloading efficiency and outperform the single antenna transmission method. In this paper, we only consider a single satellite and static offloading scenario. However, the LEO SAT moves very fast and provides service for a coverage area within a few minutes. Hence, the cooperation among multiple satellites is required. How to exploit multiple satellites to further improve the system performance is an attractive direction for future research.

Appendix A

It can be observed that $\tilde{R}_{i,k}$ is a concave differential function over $\mathbf{P}_{i,k}$ when $\mathbb{Q}$ is fixed. Therefore, the optimal $\mathbf{P}_{i,k}$ can be obtained by setting $\partial \tilde{R}_{i,k}\left(\mathbf{P}_{i,k}, \mathbb{Q}, \mathbf{V}_{i,k}\right) \Big/ \partial \mathbf{P}_{i,k}$ to zero, which can be given as

$$\frac{\partial \tilde{R}_{i,k}\left(\mathbf{P}_{i,k}, Q, \mathbf{V}_{i,k}\right)}{\mathbf{P}_{i,k}} = \left(\mathbf{P}_{i,k}\right)^{-T} - \left(\mathbf{E}_{i,k}\right)^{T} \tag{76}$$

Therefore, the optimal $\mathbf{P}_{i,k}$ is $\mathbf{P}_{i,k}^{opt} = \mathbf{E}_{i,k}^{-1}$. By substituting the optimal $\mathbf{V}_{i,k}$ and $\mathbf{P}_{i,k}$ in (34), $\tilde{R}_{i,k}\left(\mathbf{P}_{i,k}^{opt}, \mathbb{Q}, \mathbf{V}_{i,k}^{opt}\right)$ can be written as $\tilde{R}_{i,k}\left(\mathbf{P}_{i,k}^{opt}, Q, \mathbf{V}_{i,k}^{opt}\right)$ in (77), shown at the bottom of the page, where Sherman-Morrison formula is applied in (Δ) and the property of $\det\left(\mathbf{I} + \mathbf{A}_1\mathbf{A}_2\right) = \det\left(\mathbf{I} + \mathbf{A}_2\mathbf{A}_1\right)$ is used in (Ξ).

Appendix B

To prove the convergence of Algorithm 4, we need to show that the sum energy consumption (31) is nonincreasing when the variable sets $(\mathbf{A}, \mathbb{Q}, \mathbb{W}, \mathcal{L}, \mathcal{F}_U, \mathcal{F}_S)$ is updated after each iteration. According to Algorithm 4, we have

$$\begin{aligned} E_{obj}^{(n-1)} &= E\left(\mathbf{A}^{(n-1)}, \mathbf{Q}^{(n-1)}, \mathbf{W}^{(n-1)}, L^{(n-1)}, F_U^{(n-1)}, F_S^{(n-1)}\right) \\ &\overset{(\mathrm{I})}{\geq} E\left(\mathbf{A}^{(n)}, \mathbf{Q}^{(n-1)}, \mathbf{W}^{(n-1)}, L^{(n-1)}, F_U^{(n-1)}, F_S^{(n-1)}\right) \\ &\overset{(\mathrm{II})}{\geq} E\left(\mathbf{A}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{W}^{(n)}, L^{(n-1)}, F_U^{(n-1)}, F_S^{(n-1)}\right) \end{aligned}$$

$$\begin{aligned} \tilde{R}_{i,k}\left(\mathbf{P}_{i,k}^{opt}, Q, \mathbf{V}_{i,k}^{opt}\right) &= \log\det\left(\mathbf{P}_{i,k}^{opt}\right) - \mathrm{Tr}\left(\mathbf{P}_{i,k}^{opt}\mathbf{E}_{i,k}\right) + d \\ &= \log\det\left(\left(\mathbf{E}_{i,k}^{*}\right)^{-1}\right) - \mathrm{Tr}\left(\mathbf{I}_d\right) + d \\ &= \log\det\left(\left(\mathbf{I}_d - \mathbf{Q}_{i,k}^{H}\mathbf{H}_{k,i}^{H}\Pi_{k,i}^{-1}\mathbf{H}_{k,i}\mathbf{Q}_{i,k}\right)^{-1}\right) \\ &\overset{(\Delta)}{=} \log\det\left(\mathbf{I}_d^{-1} + \mathbf{I}_d^{-1}\mathbf{Q}_{i,k}^{H}\mathbf{H}_{k,i}^{H}\left(\Pi_{k,i} - \mathbf{H}_{k,i}\mathbf{Q}_{i,k}\mathbf{I}_d^{-1}\mathbf{Q}_{i,k}^{H}\mathbf{H}_{k,i}^{H}\right)^{-1}\mathbf{H}_{k,i}\mathbf{Q}_{i,k}\right) \\ &= \log\det\left(\mathbf{I}_d + \mathbf{Q}_{i,k}^{H}\mathbf{H}_{k,i}^{H}\left(\sum_{j=1,j\neq i}^{I}\sum_{m=1}^{K} a_{j,m}\mathbf{H}_{k,j}\mathbf{Q}_{j,m}\mathbf{Q}_{j,m}^{H}\mathbf{H}_{k,j}^{H} + \sigma_k^2\mathbf{I}_M\right)^{-1}\mathbf{H}_{k,i}\mathbf{Q}_{i,k}\right) \\ &\overset{(\Xi)}{=} \log\det\left(\mathbf{I}_d + \mathbf{Q}_{i,k}^{H}\mathbf{H}_{k,i}^{H}\mathbf{H}_{k,i}\mathbf{Q}_{i,k}\mathbf{N}_{i,k}^{-1}\right) \\ &= R_{i,k} \end{aligned} \tag{77}$$

$$\overset{(III)}{\geq} E\left(\mathbf{A}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{W}^{(n)}, L^{(n)}, F_U^{(n-1)}, F_S^{(n-1)}\right)$$
$$\overset{(IV)}{\geq} E\left(\mathbf{A}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{W}^{(n)}, L^{(n)}, F_U^{(n)}, F_S^{(n)}\right)$$
$$\overset{(V)}{=} E_{obj}^{(n)} \tag{78}$$

Inequality (I) is obtained from that $\mathbf{A}^{(n)}$ is one suboptimal GUE association of problem (32) with fixed precoding $\mathbf{Q}^{(n-1)}$ and $\mathbf{W}^{(n-1)}$, task allocation $L^{(n-1)}$, computation resource allocation $F_U^{(n-1)}$ and $F_S^{(n-1)}$. Inequality (II) is from that $\mathbf{Q}^{(n)}$ and $\mathbf{W}^{(n)}$ are the suboptimal precoding matrices of problem (32) with fixed GUE association $\mathbf{A}^{(n)}$, task allocation $L^{(n-1)}$, computation resource allocation $F_U^{(n-1)}$ and $F_S^{(n-1)}$. Inequality (III) follows from that $L^{(n)}$ is the suboptimal solution of problem (32) with fixed GUE association $\mathbf{A}^{(n)}$, precoding matrices $\mathbf{Q}^{(n)}$ and $\mathbf{W}^{(n)}$, computation resource allocation $F_U^{(n-1)}$ and $F_S^{(n-1)}$. Inequality (IV) is derived from that $F_U^{(n)}$ and $F_S^{(n)}$ are the optimal solution of problem (32) with fixed GUE association $\mathbf{A}^{(n)}$, precoding matrices $\mathbf{Q}^{(n)}$ and $\mathbf{W}^{(n)}$, and task allocation $L^{(n)}$. From the last equation (V), we can know that the weighted sum energy consumption is non-increasing after optimizing the variable sets in each iteration, and the convergence of Algorithm 4 is proved.

REFERENCES

[1] M. Zhanikeev, "A cloud visitation platform to facilitate cloud federation and fog computing," *Computer*, vol. 48, no. 5, pp. 80–83, May 2015.

[2] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the suitability of fog computing in the context of Internet of Things," *IEEE Trans. Cloud Comput.*, vol. 6, no. 1, pp. 46–59, Oct. 2018.

[3] T. Azzarelli, "Oneweb global access," in *Proc. ITU Int. Satell. Symp.*, 2016.

[4] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband LEO satellite communications: Architectures and key technologies," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 55–61, Apr. 2019.

[5] L. Boero, R. Bruschi, F. Davoli, M. Marchese, and F. Patrone, "Satellite networking integration in the 5G ecosystem: Research trends and open challenges," *IEEE Netw.*, vol. 32, no. 5, pp. 9–15, Sep. 2018.

[6] FCC File Number: SAT-MOD-20180319-00022. *OneWeb Ka-Band NGSO Constellation Expansion*. [Online]. Available: http://licensing.fcc.gov/cgi-bin/ws.exe/prod/ib/forms/reports/swr031b.hts?q_set=V_SITE_ANTENNA_FREQ.file_numberC/File+Number/%30/SATMOD20180319000022&prepare=&column=V_SITE_ANTENNA_FREQ.file_numberC/File+Number

[7] S. E. Holdings, "LLC application for approval for orbital deployment and operating authority for the SpaceX V-band NGSO satellite system," FCC File Number: SAT-LOA-20170301-00027. [Online]. Available: http://licensing.fcc.gov/cgibin/ws.exe/prod/ib/forms/reports/swr031b.hts

[8] A. Guidotti *et al.*, "Architectures and key technical challenges for 5G systems incorporating satellites," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2624–2639, Mar. 2019.

[9] J. Du, L. Zhao, J. Feng, and X. Chu, "Computation offloading and resource allocation in mixed fog/cloud computing systems with min-max fairness guarantee," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1594–1608, Apr. 2018.

[10] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.

[11] S. Bi and Y. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Apr. 2018.

[12] Y. Mao, J. Zhang, Z. Chen, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[13] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Dec. 2018.

[14] Z. Zhang, W. Zhang, and F.-H. Tseng, "Satellite mobile edge computing: Improving QoS of high-speed satellite-terrestrial networks using edge computing techniques," *IEEE Netw.*, vol. 33, no. 1, pp. 70–76, Jan. 2019.

[15] R. Xie, Q. Tang, Q. Wang, X. Liu, F. R. Yu, and T. Huang, "Satellite-terrestrial integrated edge computing networks: Architecture, challenges, and open issues," *IEEE Netw.*, vol. 34, no. 3, pp. 224–231, May 2020.

[16] N. Cheng *et al.*, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.

[17] Z. Lin, M. Lin, T. de Cola, J.-B. Wang, W.-P. Zhu, and J. Cheng, "Supporting IoT with rate-splitting multiple access in satellite and aerial integrated networks," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11123–11134, Jul. 2021.

[18] X. Cao, P. Yang, M. Alzenad, X. Xi, D. Wu, and H. Yanikomeroglu, "Airborne communication networks: A survey," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 1907–1926, Sep. 2018.

[19] J. Qiu, D. Grace, G. Ding, M. D. Zakaria, and Q. Wu, "Air-ground heterogeneous networks for 5G and beyond via integrating high and low altitude platforms," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 140–148, Dec. 2019.

[20] Z. Lin, M. Lin, Y. Huang, T. D. Cola, and W.-P. Zhu, "Robust multi-objective beamforming for integrated satellite and high altitude platform network with imperfect channel state information," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6384–6396, Dec. 2019.

[21] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101–109, Jul. 2017.

[22] Y. Shi, Y. Cao, J. Liu, and N. Kato, "A cross-domain SDN architecture for multi-layered space-terrestrial integrated networks," *IEEE Netw.*, vol. 33, no. 1, pp. 29–35, Jan. 2019.

[23] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 47–62, Jan. 2019.

[24] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4576–4589, Sep. 2019.

[25] T. Bai, J. Wang, Y. Ren, and L. Hanzo, "Energy-efficient computation offloading for secure UAV-edge-computing systems," *IEEE Trans. Veh. Technol*, vol. 68, no. 6, pp. 6074–6087, Apr. 2019.

[26] X. Hu, K. K. Wong, K. Yang, and Z. Zheng, "UAV-assisted relaying and edge computing: Scheduling and trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4738–4752, Oct. 2019.

[27] Q. Hu, Y. Cai, G. Yu, Z. Qin, M. Zhao, and G. Y. Li, "Joint offloading and trajectory design for UAV-enabled mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1879–1892, Oct. 2019.

[28] C. Ding, J.-B. Wang, H. Zhang, M. Lin, and J. Wang, "Joint MU-MIMO precoding and resource allocation for mobile-edge computing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1639–1654, Mar. 2021.

[29] Z. Lin, M. Lin, B. Champagne, W.-P. Zhu, and N. Al-Dhahir, "Secure and energy efficient transmission for RSMA-based cognitive satellite-terrestrial networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 251–255, Feb. 2021.

[30] Z. Lin, M. Lin, B. Champagne, W.-P. Zhu, and N. Al-Dhahir, "Secure beamforming for cognitive satellite terrestrial networks with unknown eavesdroppers," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2186–2189, Jun. 2021.

[31] M. Á. Vázquez *et al.*, "Precoding in multibeam satellite communications: Present and future challenges," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 88–95, Dec. 2016.

[32] L. You, K.-X. Li, J. Wang, X. Gao, X.-G. Xia, and B. Ottersten, "Massive MIMO transmission for LEO satellite communications," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1851–1865, Aug. 2020.

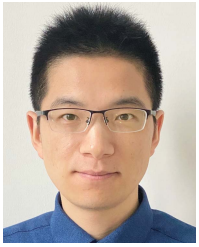
[33] L. Liu, S. Zhang, and R. Zhang, "Multi-beam UAV communication in cellular uplink: Cooperative interference cancellation and sum-rate maximization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4679–4691, Oct. 2019.

[34] C. Zhang, "Broadband air-to-ground communications with adaptive MIMO datalinks," in *Proc. IEEE/AIAA 30th Digit. Avionics Syst. Conf.*, Dec. 2011, p. 4D4-1.

[35] R. T. Schwarz, T. Delamotte, K.-U. Storek, and A. Knopp, "MIMO applications for multibeam satellites," *IEEE Trans. Broadcast.*, vol. 65, no. 4, pp. 664–681, Dec. 2019.

[36] J. E. Allnutt, "Satellite-to-ground radiowave propagation," *IET Digit. Library*, p. 696, 2011.

[37] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.
[38] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
[39] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.
[40] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.
[41] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.
[42] R. B. J. Stoer, *Introduction to Numerical Analysis*. Springer, 2013.
[43] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[44] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. 2nd USENIX Workshop Hot Topics Cloud Comput. (HotCloud)*, E. M. Nahum and D. Xu, Eds. Boston, MA, USA: USENIX Association, Jun. 2010.



Changfeng Ding (Student Member, IEEE) received the B.S. degree in communication engineering from Nanjing Tech University, Nanjing, China, in 2015, and the M.S. degree in information and communication engineering from Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the National Mobile Communications Research Laboratory, Southeast University, Nanjing. His current research interests include applications of MIMO communication, mobile edge computing, and satellite communication.

Jun-Bo Wang (Member, IEEE) received the B.S. degree in computer science from Hefei University of Technology, Hefei, China, in 2003, and the Ph.D. degree in communications engineering from Southeast University, Nanjing, China, in 2008. From October 2008 to August 2013, he was with Nanjing University of Aeronautics and Astronautics, China. From February 2011 to February 2013, he was a Post-Doctoral Fellow with the National Laboratory for Information Science and Technology, Tsinghua University, China. Since August 2013, he has been an Associate Professor with the National Mobile Communications Research Laboratory, Southeast University, China. From October 2016 to September 2018, he held the European Commission Marie Curie Fellowship and was a Research Fellow with the University of Kent, U.K. His current research interests include cloud radio access networks, mmWave communications, and wireless optical communications.

Hua Zhang (Member, IEEE) received the B.S. and M.S. degrees from the Department of Radio Engineering, Southeast University, Nanjing, China, in 1998 and 2001, respectively, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology in 2004. From 2001 to 2004, he was a Graduate Research Assistant with Georgia Institute of Technology. From 2004 to 2005, he worked as a Senior System Engineer with Skyworks Solutions Inc., Irvine, CA, USA. From 2005 to 2006, he worked as a Staff Engineer with MaxLinear Inc., Carlsbad, CA, USA. Since August 2006, he has been an Associate Professor with the School of Information Science and Engineering, Southeast University. His current research interests include massive MIMO, software defined radio, and cooperative communications. He received the Best Paper Awards of IEEE MAPE in 2013 and IEEE Globecom in 2014.

Min Lin (Member, IEEE) received the B.S. degree from the National University of Defense Technology, Changsha, China, in 1993, the M.S. degree from Nanjing Institute of Communication Engineering, Nanjing, China, in 2000, and the Ph.D. degree from Southeast University, Nanjing, in 2008, all in electrical engineering. From April 2015 to October 2015, he has visited the University of California at Irvine, Irvine, as a Senior Research Fellow. He is currently a Professor and a Supervisor of Ph.D. and graduate students with Nanjing University of Posts and Telecommunications, Nanjing. He has authored or coauthored over 130 papers. His current research interests include wireless communications and array signal processing. He has served as the Track Chair for Satellite and Space Communications (SSC) of IEEE ICC 2019 and Globecom 2021, and a TPC member for many IEEE sponsored conferences.

Geoffrey Ye Li (Fellow, IEEE) was with Georgia Institute of Technology, GA, USA, for 20 years, as a Professor, and AT&T Labs—Research, NJ, USA, for five years, as the Principal Technical Staff Member. He has been a Chair Professor with Imperial College London, U.K., since 2020. His publications have been cited over 47 000 times and he has been recognized as a Highly Cited Researcher by Thomson Reuters almost every year. His general research interests include statistical signal processing and machine learning for wireless communications. In the related areas, he has published over 600 journal articles and conference papers in addition to over 40 granted patents.

Dr. Ye Li was awarded as an IEEE Fellow for his contributions to signal processing for wireless communications in 2005. He won several prestigious awards from the IEEE Signal Processing Society, such as Donald G. Fink Overview Paper Award in 2017, the IEEE Vehicular Technology Society, such as James Evans Avant Garde Award in 2013 and Jack Neubauer Memorial Award in 2014, and the IEEE Communications Society, such as Stephen O. Rice Prize Paper Award in 2013, the Award for Advances in Communication in 2017, and Edwin Howard Armstrong Achievement Award in 2019. He also received the 2015 Distinguished ECE Faculty Achievement Award from Georgia Tech. He has organized and chaired many international conferences, including the Technical Program Vice Chair of the IEEE ICC03, the General Co-Chair of the IEEE GlobalSIP14, the IEEE VTC19 (Fall), and the IEEE SPAWC20. He has been involved in editorial activities for over 20 technical journals, including the Founding Editor-in-Chief of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) Special Series on ML in Communications and Networking.