Quantum-Assisted Joint Caching and Power Allocation for Integrated Satellite-Terrestrial Networks

Yu Zhang, *Student Member, IEEE,* Yanmin Gong, *Senior Member, IEEE,* Lei Fan, *Senior Member, IEEE,* Yu Wang, *Fellow, IEEE,* Zhu Han, *Fellow, IEEE*, and Yuanxiong Guo, *Senior Member, IEEE*

***Abstract*—Low Earth orbit (LEO) satellite network can complement terrestrial networks for achieving global wireless coverage and improving delay-sensitive Internet services. This paper proposes an integrated satellite-terrestrial network (ISTN) architecture to provide ground users with seamless and reliable content delivery services. For optimal service provisioning in this architecture, we formulate an optimization model to maximize the network throughput by jointly optimizing content delivery policy, cache placement, and transmission power allocation. The resulting optimization model is a large-scale mixed-integer nonlinear program (MINLP) that is intractable for classical computer solvers. Inspired by quantum computing techniques, we propose a hybrid quantum-classical generalized Benders' decomposition (HQCGBD) algorithm to address this challenge. Specifically, we first exploit the generalized Benders' decomposition (GBD) to decompose the problem into a master problem and a subproblem and then leverage the state-of-the-art quantum annealer to solve the challenging master problem. Furthermore, a multi-cut strategy is designed in HQCGBD to accelerate the solution process by leveraging the quantum advantages in parallel computing. Simulation results demonstrate the superiority of the proposed HQCGBD algorithm and validate the effectiveness of the proposed cache-enabled ISTN architecture.**

***Index Terms*—Integrated satellite-terrestrial network, cache placement, content delivery, quantum computing, generalized Benders' decomposition.**

I. Introduction

With the explosion of Internet-of-Things (IoT) devices, it is estimated that the global mobile data traffic will grow by a factor of nearly 4, reaching 325 EB per month in 2028 [1]. These data will cause an enormous burden on networks. Furthermore, emerging content-centric communications, such as full-motion video streaming and online live shows, typically require vast network throughput [2]. It is challenging for traditional terrestrial networks to meet these demands due to their limited backhaul capacity. Content caching at the network edge has been considered as a promising solution that could significantly alleviate backhaul pressure. Popular contents are proactively cached at the base stations (BSs), which are close to ground users, and delivered to the users directly without fetching from the content server via backhaul link [3]. However, the traditional cache-enabled terrestrial network is still infeasible in some signal-blocked or shadowed environments due to the immobility of infrastructure-based BSs [4].

In recent years, low earth orbit (LEO) satellite communication has attracted growing attention. Since LEO satellites operate from 500 m to 2000 km [5], LEO satellites can break through geographical restrictions. Consequently, several major LEO projects have been launched. For example, SpaceX Starlink plans to launch more than 12,000 LEO satellites to provide seamless global service for terrestrial users [6]. However, there are two major challenges in the satellite networks. One challenge is that satellite connections might experience interruptions due to various factors like atmospheric conditions or solar interference. The other challenge is that transmitting data over satellite links is costly. To address these challenges, we not only integrate caching into satellite networks but also integrate the satellite networks with terrestrial networks. This cache-enabled integrated satellite-terrestrial network (ISTN) is envisioned to play an essential role in 6G mobile communication systems [7]. Resource management becomes a critical problem in ISTNs due to the limited resources of satellites and edge base stations compared to remote clouds. There is a rich literature on resource management in ISTN with the goal of optimizing network throughput [8]–[10], energy efficiency [11]–[13], and system latency [14], [15]. Different from the terrestrial networks, ISTNs comprise not only terrestrial but also satellite segments. Specifically, satellites serve as the space base stations to complement the terrestrial edge networks, which are capable of attaining full coverage. These result in a higher level of collaboration and coordination among the various components in ISTNs. Thus, resource management optimization in the area of ISTN needs to be formulated as large-scale mixed-integer nonlinear programs (MINLPs), which are NP-hard. It is difficult to utilize classical computing techniques to solve them.

Quantum computing (QC) techniques provide a new promising approach to tackling this challenge. QC differs from classical computing in that it encodes information using qubits, representing a superposition of states, rather than binary bits [16].

Y. Zhang and Y. Gong are with the Department of Electrical and Computer Engineering, University of Texas at San Antonio, Texas, 78249, USA. (e-mail: {yu.zhang@my., yanmin.gong@}utsa.edu).

L. Fan is with the Department of Engineering Technology, University of Houston, Houston, TX 77204 USA (e-mail: lfan8@central.uh.edu).

Y. Wang is with Department of Computer and Information Sciences, Temple University, Philadelphia, Pennsylvania 19122, USA. (e-mail: wangyu@temple.edu).

Z. Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701. (e-mail: zhan2@uh.edu).

Y. Guo is with the Department of Information Systems and Cyber Security, University of Texas at San Antonio, Texas, 78249, USA. (e-mail: yuanxiong.guo@utsa.edu).

The work is partially supported by the US NSF (Grant NO. CNS-2106761, CNS-2318663, CNS-2047761, ECCS-2302469, CNS-2006604, CNS-2128378, CNS-2107216, CNS-2128368, CMMI-2222810, and ECCS-2302469), US Department of Transportation, Toyota, Amazon, and Japan Science and Technology Agency (JST) Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) under JPMJAP2326.

Besides, QC utilizes entanglement and quantum tunneling to find the solution. These features enable quantum computers to simultaneously explore exponentially large combinations of states, which can solve large-scale real-world optimization problems more efficiently and faster [16]. There are two major paradigms in quantum computing: gate-based quantum computing and adiabatic quantum computing. The gate-based quantum computing utilizes discrete quantum gate operations to evolve the qubits so that the final state of the qubits is the desired result. In gate-based quantum computing, we need to map the problem to an effective quantum algorithm and convert the instructions into a sequence of low-level quantum gates to operate the qubits. The main drawback of gate-based quantum computing is that the number of provided qubits is insufficient for real-world applications. We can only obtain less than 150 qubits for gate-based quantum computing [17], [18]. Unlike gate-based quantum computing, we encode the problem into the Hamiltonian of the quantum system in adiabatic quantum computing. After evolution, we can obtain the desired solution through the ground state of the quantum system. Implementing adiabatic quantum computing is hard since the quantum physical systems are susceptible to non-ideal conditions. Quantum annealing (QA) can be regarded as a relaxed adiabatic quantum computing that does not necessarily require universality or adiabaticity [19]. Currently, we can obtain about 5000 qubits for QA from D-wave [20]. With these large numbers of qubits, QA has the potential to enhance practical applications such as car manufacturing scheduling [21], RNA folding [22]–[24], and portfolio optimization [25].

In this paper, we propose the first QC approach for improving the network performance in cache-enabled ISTNs. Specifically, we jointly optimize the content delivery policy, cache placement, and transmission power allocation aiming at maximizing the network throughput. The formulated optimization problem is a large-scale MINLP that is NP-hard and difficult for classical computers to solve. To solve the problem efficiently, we propose a hybrid quantum-classical generalized Benders' decomposition (HQCGBD) algorithm. In particular, we first utilize the generalized Benders' decomposition (GBD) algorithm to decompose the MINLP into a master problem with the binary decision variables (i.e., content delivery policy and cache placement) and a subproblem with the continuous decision variables (i.e., transmission power). The subproblem is a convex problem that can be efficiently solved by classical computers. However, the master problem becomes more time- and memory-consuming with more Benders' cuts added to the constraints over iterations for classical computers [26]. Thus, we further convert the master problem into a quadratic unconstrained binary optimization (QUBO) formulation that is solvable by QA. Then, these two problems are iteratively solved until their solutions converge. Inspired by the parallel processing capability of QC, we further design a specialized quantum multi-cut strategy to accelerate HQCGBD. Finally, we evaluate the performance of the HQCGBD algorithm for the cache-enabled ISTN on the D-Wave's real-world quantum annealer computer [20].

The main contributions of this paper are stated as follows.

- We formulate an optimization problem to maximize the network throughput for a cache-enabled ISTN system.
- As the formulated model is a large-scale MINLP, which is generally intractable to classical computers, we propose an HQCGBD algorithm to solve the problem efficiently by leveraging GBD and QA.
- Motivated by the parallel processing capability of QC, we design a multi-cut strategy to accelerate the convergence of the proposed HQCGBD algorithm.
- We conduct extensive experiments to demonstrate the superiority of the proposed solution algorithm compared with the classical computing algorithm and show the proposed cache-enabled ISTN scheme largely outperforms the baseline schemes.

The rest of this paper is organized as follows. The related works are discussed in Section II. In Section III, we introduce the system model and describe a network throughput maximization formulation for the cache-enabled ISTN. In Section IV, we present the HQCGBD algorithm to solve the formulated problem. Section V shows the simulation results. Finally, the conclusion is given in Section VI.

II. Related Works

In this section, we discuss the most related prior works from two aspects: ISTN system and quantum annealing.

A. ISTN System

ISTN is considered as a compelling technology for extending the coverage area of the existing terrestrial networks, which has attracted increasing attention from both academia and industry. There is various literature on resource management in ISTN that aims at optimizing network throughput [8]–[10], energy efficiency [11]–[13], and system latency [14], [15]. Han et al. [8] studied the joint cache placement, LEO satellite and BS clustering, and multicast beamforming problem aiming at maximizing the network utility. Tran et al. [9] investigated cache placement, the UAV's resource allocation, and trajectory problem in LEO satellite- and cache-assisted unmanned aerial vehicle (UAV) communications aiming at maximizing the minimum achievable throughput per ground user. Alsharoa et al. [10] jointly optimized the resource allocations and the locations of high-altitude platforms (HAPs) to maximize the users' throughput. Li et al. [11] formulated a block placement, power allocation, and cache-sharing decision optimization problem to optimize the ISTN system energy efficiency. Tang et al. [12] optimized the computation offloading decisions to minimize the sum energy consumption of ground users. Ding et al. [13] investigated a joint user association, multi-user multiple input and multiple output (MU-MIMO) transmit precoding, computation task assignment, and resource allocation optimization problem to minimize the weighted sum energy consumption of the ISTN system. Han et al. [14] proposed a two-timescale learning-based scheme to minimize the overall task offloading delay by jointly optimizing offloading link selection and bandwidth allocation decisions for BSs and users. Cui et al. [15] studied a joint task offloading, communication, and computing resources allocation problem aiming at minimizing the latency of task offloading and processing.

However, most of these studies in the ISTN systems usually formulate their problems as MINLPs, which are hard to solve. Thus, they either use heuristics or complex optimization techniques to tackle them. Inspired by QC, we propose an HQCGBD algorithm to solve these MINLPs efficiently.

B. Quantum Annealing

Extensive research efforts have been made recently to utilize QA for solving practical optimization problems [27]. However, QA only accepts a quadratic polynomial over binary variables, and the cost of using a large number of ancillary qubits to represent the discretized variables is expensive. Most prior works formulate the real-world application as a binary quadratic model (BQM) problem [21]–[25] or mixed-integer linear programming (MILP) problem [28], [29] that is easily converted to the QUBO formulation for QA to optimize. For instance, Yarkoni et al. [21] minimized the number of color switches between cars in a paint shop queue by formulating a BQM problem. Fox et al. [22] optimized codon through a BQM. Mulligan et al. [23] found amino acid side chain identities and conformations to stabilize a fixed protein backbone. Fox et al. [24] formulated a BQM to maximize the number of consecutive base pairs and the average length of the stems. Mugel et al. [25] proposed a hybrid quantum-classical algorithm for dynamic portfolio optimization with a minimum holding period. Doan et al. [28] studied the MILP problem in robust fitting by leveraging QA. Dinh et al. [29] formulated the beam placement in satellite communication as an MILP and designed an efficient Hamiltonian Reduction method for QA to address this problem efficiently. However, the BQM and MILP formulations can not capture the complexity of the actual problem in the ISTN field. To the best of our knowledge, this is the first work to formulate a MINLP to optimize the network throughput in the cache-enabled ISTN system and solve it using QC technologies.

III. System Model And Problem Formulation

In this section, we first introduce the system models. After that, we formulate an optimization model to jointly optimize the optimal content delivery policy, cache placement, and transmission power allocation in the cache-enabled ISTN.

A. System Architecture

As illustrated in Fig. 1, we consider a cache-enabled ISTN that consists of a content server, a set of single-antenna users $n \in \mathcal{N} = \{1, \ldots, N\}$, and a set of base stations (BSs). We denote the set of BSs as $\mathcal{M} = \mathcal{S} \cup \mathcal{B}$, in which the index $m \in \mathcal{S} = \{1, \ldots, S\}$ represents a satellite, and $m \in \mathcal{B} = \{S+1, \ldots, M\}$ represents a terrestrial base station (TBS). The satellites and TBSs, which have limited cache storage onboard, cooperatively provide content delivery services to the ground users through satellite-to-user (S2U) and TBS-to-user (T2U) links, respectively. The content server contains a set of potential user-desired content $f \in \mathcal{F} = \{1, \ldots, F\}$. For simplicity, we assume the required content can be successfully downloaded for each user during the satellite visibility period.

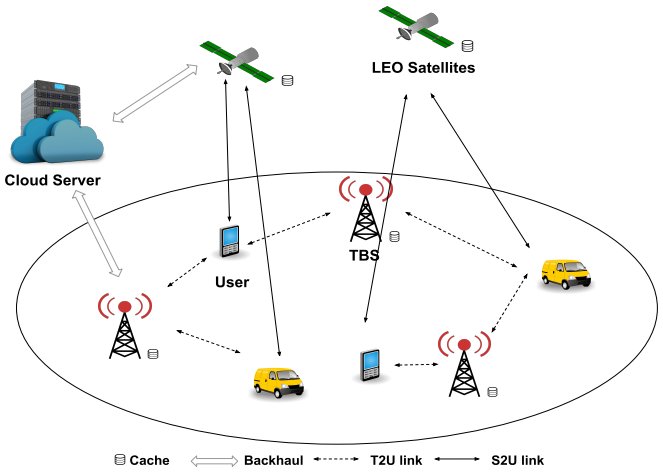


Fig. 1: The integrated satellite-terrestrial network system

The BSs are connected to the content server via the limited-capacity backhaul links and proactively cache a portion of the popular content from the content server and then efficiently serve the ground users.

B. Cache Model

When each user $n \in \mathcal{N}$ submits its content request $f_n \in \mathcal{F}$ to the BS, there are two ways for it to download its desired content f_n: if the requested content f_n is available in the cache, the associated BS m will deliver it to user n directly; otherwise, the associated BS m will fetch the requested content f_n from the content server via the backhaul link and then deliver it to user n.

We denote the cache placement matrix for the BSs as $\mathbf{X} = \{x_{f,m}\}^{F \times M}$, where $x_{f,m} \in \{0, 1\}$. $x_{f,m} = 1$ indicates the content f is placed at the m-th BS; otherwise $x_{f,m} = 0$. Considering the limited cache storage capability of each BS, we have

$$\sum_{f \in \mathcal{F}} x_{f,m} d_f \leq F_m, \quad \forall m \in \mathcal{M}, \tag{1}$$

where F_m is the cache storage capability of BS m, and d_f is the data size of content f.

C. Communication Model

Let $\mathbf{Z} = \{z_{n,m}\} \in \{0, 1\}^{N \times M}$ denote the BS service matrix in which $z_{n,m} = 1$ indicates that BS m serves user n and $z_{n,m} = 0$ otherwise. Thanks to the software-defined networking (SDN) and cloud radio access network technologies, each user can be cooperatively served by multiple BSs simultaneously [8], which is described as

$$\sum_{m \in \mathcal{M}} z_{n,m} \geq 1, \quad \forall n \in \mathcal{N}. \tag{2}$$

Moreover, we make the assumption that the OFDMA technique is adopted. The available spectrum of BSs is divided into multiple resource blocks. Each resource block for BS m has a bandwidth of B_m [10], [30]. We also assume that a user uses only one resource block of an associated BS, and ignore intra-cell and inter-cell interference since users associated with

the same BS use a different set of orthogonal resource blocks. The maximum number of associated users for BS m should satisfy

$$\sum_{n\in\mathcal{N}} z_{n,m} \leq I_m^{\max}. \quad \forall m \in \mathcal{M}, \tag{3}$$

where $I_m^{\max}$ is the maximum number of resource blocks for BS m.

The transmission power matrix from BSs to users is denoted as $\mathbf{P} = [p_{n,m}]_{n\in\mathcal{N},m\in\mathcal{M}}$. When BS m does not serve user n, the corresponding transmission power should be 0. This can be modeled as the following constraint:

$$z_{n,m}P_n^{\min} \leq p_{n,m} \leq z_{n,m}P_n^{\max}, \quad \forall n \in \mathcal{N}, m \in \mathcal{M}, \tag{4}$$

where $P_n^{\max}$ and $P_n^{\min}$ are the maximum and minimum transmission power of BS n for each connected device, respectively.

T2U channel model: We model the T2U channel as a Rayleigh channel with shadowing effect. Let $h_{n,m}$ denote the channel coefficients of the T2U link from TBS $m \in \mathcal{B}$ to user $n \in \mathcal{N}$. According to [31], $h_{n,m}$ can be modeled as

$$h_{n,m} = \gamma_{n,m}\beta_{n,b}(d_{n,m})^{-\alpha}, \quad \forall n \in \mathcal{N}, m \in \mathcal{B}, \tag{5}$$

where $\gamma_{n,m} \sim \mathcal{CN}(0,1)$ is a complex Gaussian variable representing Rayleigh fading, $\beta_{n,m}$ follows log-normal distribution representing shadowing fading, $d_{n,m}$ is the distance between the TBS and user, and α represents the path-loss exponent.

S2U channel model: We consider both the large-scale fading and the shadowed-Rician fading for the S2U channel [32]. We denote the coefficient of S2U channel connecting satellite $m \in \mathcal{S}$ and user $n \in \mathcal{N}$ as $h_{n,m}$, which is modeled as

$$h_{n,m} = A\exp(j\psi_{n,m}) + Z\exp(j\phi_{n,m}), \forall n \in \mathcal{N}, m \in \mathcal{S}, \tag{6}$$

where $\psi_{n,m} \in [0, 2\pi)$ indicates the stationary random phase and $\phi_{n,m}$ is the deterministic phase of the LOS component. Here A and Z denote the amplitudes of the scattering and the LOS components, which are independent stationary random processes following Rayleigh and Nakagami-m distributions, respectively.

Based on Shannon's theorem, the achievable data transmission rate from BS m to user n can be expressed as

$$R_{n,m} = z_{n,m}B_m \log_2\left(1+\frac{p_{n,m}G_{n,m}|h_{n,m}|^2}{\sigma^2}\right), \quad \forall n \in \mathcal{N}, m \in \mathcal{M}, \tag{7}$$

where $G_{n,m}$ is the channel gain of BS m towards user n, and σ^2 represents the additive white Gaussian noise (AWGN). The instantaneous transmission rate for BS m to deliver content f_n to user n via cache can be calculated as

$$Q_{n,m}^C = x_{f_n,m}R_{n,m}, \quad \forall n \in \mathcal{B}, m \in \mathcal{M}. \tag{8}$$

On the other hand, the instantaneous transmission rate for BS m to deliver content f_n to user n via backhaul can be written as

$$Q_{n,m}^B = (1 - x_{f_n,m})R_{n,m}, \quad \forall n \in \mathcal{B}, m \in \mathcal{M}. \tag{9}$$

As a result, the total network throughput can be modeled as

$$Q^T = \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} \left(Q_{n,m}^C + \lambda Q_{n,m}^B\right), \tag{10}$$

where λ is the trade-off parameter to balance the priorities of maximizing the total cache and backhaul throughput. Particularly, a small λ implies that maximizing total cache throughput has more priority than maximizing the total backhaul throughput. In practice, we set $0 < \lambda < 1$ to reduce the backhaul pressure.

D. Problem Formulation

In order to maximize the total network throughput that takes into account both cache and backhaul throughput, we jointly optimize the content delivery policy, cache placement, and transmission power allocation as

$$\begin{aligned}
\max_{\mathbf{X},\mathbf{Z},\mathbf{P}} \quad & Q^T \\
\text{s.t.} \quad & (\mathrm{C}_1): \sum_{f\in\mathcal{F}} x_{f,m}d_f \leq F_m, \quad \forall m \in \mathcal{M}, \\
& (\mathrm{C}_2): \sum_{n\in\mathcal{N}} p_{n,m} \leq P_m^{\max}, \quad \forall m \in \mathcal{M}, \\
& (\mathrm{C}_3): p_{n,m} \leq z_{n,m}P_n^{\max}, \quad \forall n \in \mathcal{N}, m \in \mathcal{M}, \\
& (\mathrm{C}_4): z_{n,m}P_n^{\min} \leq p_{n,m}, \quad \forall n \in \mathcal{N}, m \in \mathcal{M}, \\
& (\mathrm{C}_5): \sum_{m\in\mathcal{M}} z_{n,m} \geq 1, \quad \forall n \in \mathcal{N}, \\
& (\mathrm{C}_6): \sum_{n\in\mathcal{N}} z_{n,m} \leq I_m^{\max}, \quad \forall m \in \mathcal{M}, \\
& (\mathrm{C}_7): p_{n,m} \geq 0, \quad \forall n \in \mathcal{N}, m \in \mathcal{M}, \\
& (\mathrm{C}_8): x_{f,m} \in \{0,1\}, \quad \forall f \in \mathcal{F}, m \in \mathcal{M}, \\
& (\mathrm{C}_9): z_{n,m} \in \{0,1\}, \quad \forall n \in \mathcal{N}, m \in \mathcal{M}.
\end{aligned} \tag{11}$$

Constraint (C_1) ensures that the placed caches for each BS can not exceed the maximum cache storage. Constraints (C_2), (C_3), (C_4), and (C_7) guarantee the valid power allocation. Constraint (C_5) ensures each user is served by at least one BS. Constraint (C_6) represents the number of associated users that can not exceed the maximum number of resource blocks for each BS. Finally, constraints (C_8) and (C_9) show $\mathbf{X}$ and $\mathbf{Z}$ are binary matrices.

IV. Solution Methodology

The formulated problem (11) is a large-scale MINLP, which is NP-hard [33]. To tackle this problem, we propose a novel HQCGBD algorithm. Specifically in HQCGBD, we first utilize GBD to decompose the problem into a master problem and a subproblem. On the one hand, the subproblem is a convex problem, which can be solved efficiently on the classical computer. On the other hand, the master problem is a large-scale MILP that is intractable for classical computers. Thus, we convert the master problem into the QUBO formulation, which can be efficiently solved by the state-of-the-art quantum annealer. Moreover, we design a specialized quantum multi-cut strategy to accelerate the optimization process.

A. Hybrid Quantum-classical Generalized Benders' Decomposition

Problem (11) is hard to solve due to the coupling of continuous decision variables $\mathbf{P}$ with integer decision variables $\mathbf{X}$ and $\mathbf{Z}$. We first decompose (11) into a master problem and a subproblem. The master problem only involves the binary optimization decision variables, while the subproblem only involves the continuous optimization decision variables. We can obtain a performance upper bound of problem (11) by solving the subproblem and a performance lower bound of problem (11) by solving the master problem. Then, problem (11) is iteratively solved until the lower and upper bounds converge. In order to apply HQCGBD, we first reformulate problem (11) as

$$\begin{aligned} \min_{\mathbf{X},\mathbf{Z},\mathbf{P}} \quad & -Q^T \\ \text{s.t.} \quad & (\mathrm{C}_1)-(\mathrm{C}_9). \end{aligned} \tag{12}$$

We describe the details of the subproblem and master problem in the following.

1) Classical Optimization for Subproblem: For the given and fixed binary variables $\mathbf{X}^{(l)}$ and $\mathbf{Z}^{(l)}$ generated by the master problem at the $(l-1)$-th iteration, the subproblem can be written as

$$\begin{aligned} \text{(Subproblem)}: \min_{\mathbf{P}} \quad & -Q^T \\ \text{s.t.} \quad & (\mathrm{C}_2)-(\mathrm{C}_4),(\mathrm{C}_7). \end{aligned} \tag{13}$$

As we can see above, the subproblem has a convex objective function and linear constraints, so it is convex. Besides, it also satisfies Slater's conditions [34], strong duality holds between the subproblem and its dual problem. Thus, solving its dual problem is equivalent to solving the subproblem. We formulate its dual problem as

$$\begin{aligned} \max_{\boldsymbol{\delta}} \ \min_{\mathbf{P}} \mathcal{L}(\mathbf{P},\boldsymbol{\delta},\mathbf{X}^{(l)},\mathbf{Z}^{(l)}) = & -Q^T - \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} \delta_{1,n,m} p_{n,m} \\ & + \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} \delta_{2,n,m}\big(p_{n,m} - z^{(l)}_{n,m} P^{\max}_n\big) \\ & + \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} \delta_{3,n,m}\big(z^{(l)}_{n,m} P^{\min}_n - p_{n,m}\big) \\ & + \sum_{m\in\mathcal{M}} \delta_{4,m}\big(\sum_{n\in\mathcal{N}} p_{n,m} - P^{\max}_m\big), \end{aligned} \tag{14}$$

where $\mathcal{L}$ is the Largangian function of problem (13), $\mathbf{P}$ is the primal variables, and $\boldsymbol{\delta}$ is the dual variables associated with constraints $(\mathrm{C}_2)-(\mathrm{C}_4),(\mathrm{C}_7)$. This problem can be efficiently solved by the classical convex programming numerical solvers such as Mosek [35]. The optimal solutions of the primal variables $\mathbf{P}$ and dual variables $\boldsymbol{\delta}$ are then used as input to the master problem. Besides, the optimal objective value of problem (13) can be regarded as the performance upper bound of problem (12).

2) Quantum Optimization for Master Problem: With obtained primary variables $\mathbf{P}^{(l)}$ and dual variables $\boldsymbol{\delta}^{(l)}$ of problem (14) at the l-th iteration, we can formulate the master as

$$\begin{aligned} \text{(Master problem)}: \min_{\mathbf{X},\mathbf{Z},\phi} \quad & \phi \\ \text{s.t.} \quad & (\mathrm{C}_1),(\mathrm{C}_5),(\mathrm{C}_6),(\mathrm{C}_8),(\mathrm{C}_9), \\ & (\mathrm{C}_{10}):\ \phi \geq \mathcal{L}(\mathbf{P}^{(l)},\boldsymbol{\delta}^{(l)},\mathbf{X},\mathbf{Z}). \end{aligned} \tag{15}$$

Here, (C_{10}) is the Benders' cut [36]. By adding the Benders' cut at each iteration, the search region for the globally optimal solution is gradually reduced, which accelerates the searching speed [37]. Besides, the objective value of problem (15) is the performance lower bound of problem (12) at the l-th iteration.

The constructed master problem is a large-scale MILP, which is difficult for the classical computer to solve. QA technology is a promising approach to address this challenge. A quantum annealer solves optimization problems through energy minimization of a physical system. The energy profile of a quantum system is defined by its Hamiltonian. The system's initial state is set to the lowest energy state of the initial Hamiltonian and then annealed until the system's final state corresponds to the desired solution. The Hamiltonian at the end of the annealing process can be derived from the following model:

$$f_Q(\mathbf{x}) = \sum_{i=1}^{K} Q_{ii}x_i + \sum_{i=1}^{K}\sum_{j=1}^{K} Q_{ij}x_i x_j = \mathbf{x}^{\intercal}\mathbf{Q}\mathbf{x}, \tag{16}$$

where $f_Q : \{0,1\}^k \to \mathbb{R}$ represents a quadratic polynomial over binary variables, $\mathbf{x} = [x_1,\ldots,x_k]$ are binary variables, and $\mathbf{Q} \in \mathbb{R}^{K\times K}$ is an upper triangular matrix representing the corresponding cost coefficient of $x_i x_j$ [38].

A well-known approach for representing optimization problems in quantum annealing is to use the QUBO, which solves a problem of the formulation:

$$\min_{\mathbf{x}\in\{0,1\}^K} f_Q(\mathbf{x}). \tag{17}$$

Note that the master problem (15) has a structure similar to an integer linear program (ILP) but is not in the QUBO format. In order to leverage the state-of-art QA, we convert the master problem into the QUBO formulation as follows [39]:

Objective function reformulation: Since the QA only accepts a quadratic polynomial over binary variables, we first approximate the continuous variable ϕ using a binary vector $\mathbf{w}$ with the length of U bits. The QUBO formulation of objective function (15) can be written as

$$\begin{aligned} \bar{\phi} = & \sum_{i=-\underline{u}_+}^{\overline{u}_+} w_{(i+\underline{u}_+)} 2^i w_{(i+\underline{u}_+)} - \\ & \sum_{j=0}^{\overline{u}_-} w_{(j+1+\underline{u}_+ +\overline{u}_+)} 2^j w_{(j+1+\underline{u}_+ +\overline{u}_+)} = \bar{\phi}(\mathbf{w}), \end{aligned} \tag{18}$$

where $U = 1 + \underline{u}_+ + \overline{u}_+ + \overline{u}_-$. Here, $\overline{u}_-, \overline{u}_+, \underline{u}_+$ represent the number of bits that are assigned to represent the negative integer, positive integer, and positive decimal part of ϕ, respectively.

Constraints reformulation: After reformulating the objective function in (15), we can obtain a constraint ILP master problem, which is still not the QUBO formulation. We need to further reformulate the constrained ILP master problem as unconstrained QUBO by using penalties. According to the constraint-penalty pair principle in [39], we convert the constraints (C_1),(C_5),(C_6), and (C_{10}) as follows:

$$f_Q^{(\mathrm{C}_1)}(\mathbf{X},\mathbf{s}) = \sum_{m\in\mathcal{M}} \xi_{1,m}\Big(\sum_{f\in\mathcal{F}} x_{f,m}d_f - F_m + \sum_{l=0}^{\bar{l}_{1,m}} 2^l s_{1,m}\Big)^2, \tag{19}$$

where $\bar{l}_{1,m} = \lceil \log_2\big(\min_{\mathbf{X}}(F_m - \sum_{f\in\mathcal{F}} x_{f,m}d_f)\big)\rceil$.

$$f_Q^{(\mathrm{C}_5)}(\mathbf{Z},\mathbf{s}) = \sum_{n\in\mathcal{N}} \xi_{2,n}\Big(1 - \sum_{m\in\mathcal{M}} z_{n,m} + \sum_{l=0}^{\bar{l}_{2,n}} 2^l s_{2,n}\Big)^2, \tag{20}$$

where $\bar{l}_{2,n} = \lceil \log_2\big(\min_{\mathbf{Z}}(\sum_{m\in\mathcal{M}} z_{n,m} - 1)\big)\rceil$.

$$f_Q^{(\mathrm{C}_6)}(\mathbf{Z},\mathbf{s}) = \sum_{m\in\mathcal{M}} \xi_{3,m}\Big(\sum_{n\in\mathcal{N}} z_{n,m} - I_m^{\max} + \sum_{l=0}^{\bar{l}_{3,m}} 2^l s_{3,m}\Big)^2, \tag{21}$$

where $\bar{l}_{3,m} = \lceil \log_2\big(\min_{\mathbf{Z}}(I_m^{\max} - \sum_{n\in\mathcal{N}} z_{n,m})\big)\rceil$.

$$\begin{aligned} f_Q^{(\mathrm{C}_{16})}(\mathbf{X},\mathbf{Z},\mathbf{w},\mathbf{s}) =& \xi_4\big(\mathcal{L}(\mathbf{P}^{(l)},\boldsymbol{\delta}^{(l)},\mathbf{X},\mathbf{Z}) - \bar{\phi}(\mathbf{w}) \\ &+ \sum_{l=0}^{\bar{l}_4} 2^l s_4\big)^2, \end{aligned} \tag{22}$$

where $\bar{l}_4 = \lceil \log_2\big(\min_{\mathbf{w},\mathbf{Z}}\big(\bar{\phi}(\mathbf{w}) - \mathcal{L}(\mathbf{P}^{(l)},\boldsymbol{\delta}^{(l)})\big)\big)\rceil$.

Here, $\boldsymbol{s}$ is the binary slack variables, $\bar{\boldsymbol{l}}$ is the upper bound of the number of slack variables, and $\boldsymbol{\xi}$ is the penalty parameters which are defined according to [40]. Then, the master problem can be stated in the QUBO formulation as

$$\begin{aligned} \min_{\mathbf{X},\mathbf{Z},\mathbf{w},\mathbf{s}} \quad & \bar{\phi}(\mathbf{w}) + f_Q^{(\mathrm{C}_1)}(\mathbf{X},\mathbf{s}) + f_Q^{(\mathrm{C}_5)}(\mathbf{Z},\mathbf{s}) \\ & + f_Q^{(\mathrm{C}_6)}(\mathbf{Z},\mathbf{s}) + f_Q^{(\mathrm{C}_{16})}(\mathbf{X},\mathbf{Z},\mathbf{w},\mathbf{s}). \end{aligned} \tag{23}$$

3) Overall Algorithm: The overall HQCGBD algorithm is summarized in Algorithm 1. The algorithm contains an iterative procedure. First of all, we set the iteration index l to 1 and the maximum number of iterations as $L^{\max}$ and initialize the binary variables $\mathbf{X}$ and $\mathbf{Z}$. In the l-th iteration, we fix the binary variables $\mathbf{X}^{(l)}$ and $\mathbf{Z}^{(l)}$ and solve the subproblem using the classical computer (Line 2). We add the obtained Benders' cut to the master problem and update the performance upper bound $\mathrm{UB}^{(l)}$ by the optimal solution of the subproblem (Line 3-6). Then, we set the appropriate penalties and reformulate the master problem into the QUBO formulation (Line 7-8). We utilize the quantum computer to solve the QUBO master problem and update the performance lower bound $\mathrm{LB}^{(l)}$ (Line 9-11). This iteration procedure stops until the approximation gap $|\frac{\mathrm{UB}^{(l)}-\mathrm{LB}^{(l)}}{\mathrm{UB}^{(l)}}|$ is within a preset threshold ϵ or the maximal iteration index $L^{\max}$ is reached

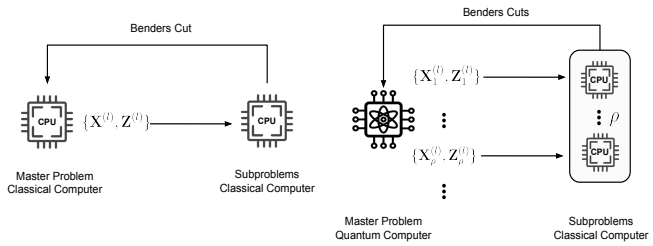


Fig. 2: An overview of (a) CBD and (b) Multi-Cut HQCGBD

B. Multi-cut Strategy

The main bottleneck of classical Benders' decomposition (CBD) is the time consumed by solving the master problems, which occupies over 90% total optimization time [41]. As shown in Fig. 2(a), classical computers generate just one Benders cut at an iteration. However, QA utilizes qubits that can explore many combinations of quantum states simultaneously by leveraging the superposition of quantum states [16]. This endows quantum computers the powerful parallel computing capacity. Thanks to this feature, we can observe multiple feasible solutions at the end of QA. Thus, quantum computers can potentially accelerate the convergence of the GBD compared to classical computers. In Fig. 2(b), we demonstrate the basic idea of multi-cut HQCGBD. In multi-cut HQCGBD, solving a master problem can obtain multiple feasible solutions. We select the top ρ feasible solutions for the classical computers to generate multiple Benders' cuts in parallel.

The details of the multi-cut HQCGBD are summarized in Algorithm 2. The main difference between Algorithm 1 and Algorithm 2 is that a set of feasible binary variables $\mathcal{X}$ is generated by solving the master problem on the quantum computer. For each feasible solution, we can solve a subproblem on the classical computer to generate a Benders' cut. Then, all Benders' cuts are added to the master problem for the quantum computer to solve. This iteration procedure continues until the upper bound and lower bound converge.

V. Numerical Experiments

In this section, we evaluate the performance of our proposed HQCGBD algorithm and cache-enabled ISTN architecture. Since quantum computing resources is still extremely expensive, we only conduct a small-scale setting that could be solved in fewer than 1000 iterations, but our results show how promising this technology could be in the future. Both HQCGBD and CBD are implemented in Python 3.7. Specifically, we utilize the Mosek [35] to solve the classical convex problems and Gurobi [42] to solve classical MILP problems on a desktop computer with a 3.2 GHz Intel$^{\mathrm{R}}$ Core$^{\mathrm{TM}}$ i7-8700 CPU and 16 GB of RAM. The HQCGBD master problems are solved by the D-Wave hybrid quantum computer, which has over 5,000 physical qubits, and 35,000 couplers based on the Pegasus topology [20].

Algorithm 1 The Proposed HQCGBD Algorithm

Input: Iteration index $l = 1$, maximum iteration number $L^{\max}$, iteration threshold ϵ, $UB^{(0)} = +\infty$, and $LB^{(0)} = -\infty$. Initialize $\mathbf{X}^{(0)}, \mathbf{Z}^{(0)}$.

while $|\frac{\mathrm{UB}^{(l-1)} - \mathrm{LB}^{(l-1)}}{\mathrm{UB}^{(l-1)}}| > \epsilon$ or $l < L^{\max}$ **do**
  Fix $\mathbf{X}$, $\mathbf{Z}$ as $\mathbf{X}^{(l-1)}, \mathbf{Z}^{(l-1)}$, and solve the subproblem using a standard convex solver in the classical computer
  Obtain the optimal solution $\mathbf{P}^{(l)}$ and $-Q^{\mathrm{T}(l)}$
  Obtain the Benders' cut C
  Update $\mathrm{UB}^{(l)} = \min\{\mathrm{UB}^{(l-1)}, -Q^{\mathrm{T},(l)}\}$
  Add the Benders cut C to the master problem
  Set appropriate penalty numbers or arrays $\boldsymbol{\xi}$
  Reformulate both objective and constraints in the master problem and construct the QUBO formation by using corresponding rules, i.e., (19)-(23)
  Solve the master problem by the quantum computer
  Obtain optimal solution $\mathbf{X}^{(l)}$, $\mathbf{Z}^{(l)}$, and $\phi^{(l)}$
  Update $\mathrm{LB}^{(l)} = \phi^{(l)}$
  Set l = l + 1
end while

Output: Optimal $\mathbf{X}^*, \mathbf{Z}^*, \mathbf{P}^*$.

Algorithm 2 The Proposed Multi-Cut HQCGBD Algorithm

Input: Iteration index $l = 1$, maximum iteration number $L^{\max}$, iteration threshold ϵ, $UB^{(0)} = +\infty$, and $LB^{(0)} = -\infty$. Initialize ρ feasible values as $\mathcal{X}^{(0)} = \{\mathbf{X}_i^{(0)}, \mathbf{Z}_i^{(0)}\}_{i=1}^{\rho}$.

while $|\frac{\mathrm{UB}^{(l-1)} - \mathrm{LB}^{(l-1)}}{\mathrm{UB}^{(l-1)}}| > \epsilon$ or $l < L^{\max}$ **do**
  for $\{\mathbf{X}, \mathbf{Z}\} \in \mathcal{X}^{(l-1)}$ **do**
    Fix $\mathbf{X}$ and $\mathbf{Z}$, and solve the subproblem using a standard convex solver in the classical computer
    Obtain the optimal solution $\mathbf{P}^{(l)}$ and $-Q^{\mathrm{T},(l)}$
    Obtain the Benders' cut C
    Add the Benders' cut C to the master problem
    Update $\mathrm{UB}^{(l)} = \min\{\mathrm{UB}^{(l-1)}, -Q^{\mathrm{T},(l)}\}$
  end for
  Set appropriate penalty numbers or arrays $\boldsymbol{\xi}$
  Reformulate both objective and constraints in the master problem and construct the QUBO formation by using corresponding rules, i.e., (19)-(23)
  Solve the master problem by the quantum computer
  Obtain ρ feasible solutions $\mathcal{X}^{(l)} = \{\mathbf{X}_i^{(l)}, \mathbf{Z}_i^{(l)}\}_{i=1}^{\rho}$ and $\{\phi_i^{(l)}\}_{i=1}^{\rho}$
  Update $\mathrm{LB}^{(l)} = \min\{\phi_i^{(l)}\}_{i=1}^{\rho}$
  Set l = l + 1
end while

Output: Optimal $\mathbf{X}^*, \mathbf{Z}^*, \mathbf{P}^*$.

A. Simulation Setup

In our simulation, we consider a cache-enabled ISTN system with one TBS placed in the center and 10 ground users that are uniformly distributed within an area of 1000×1000 m^2, excluding an inner circle of 50 m radius around the TBS. In order to meet the quality of service (QoS) requirement, one LEO satellite cooperatively provides the content delivery service to the ground users. The ISTN system operates at 2GHz with a bandwidth of 10MHz. For terrestrial communication, the path-loss exponent α is set to 3.7 [43], and the log-normal shadowing parameter is set to be 8dB. The small-scale fading over the terrestrial channel is modeled as the normalized Rayleigh fading. For satellite communication, the path-loss is modeled by $L_p(dB) = 92.44 + 20\log_{10} d + 20\log_{10}(f)$, where f is the operating frequency in GHz, and d is the distance between the satellite and user in kilometers, which is determined by the satellite altitude and elevation angle [44]. We set the total attenuation caused by rain, gas, clouds, and scintillation as 5.2dB [45], and the additional losses from polarization and antenna misalignment as 0.1dB and 0.35dB, respectively [46]. The small-scale fading over the satellite channel is modeled as Rician fading.

TABLE I: Simulation Parameters

Parameters	Values
Satellite altitude H_s	1000km
Transmission power of TBS $P_b^{\max}$	42dBm
Transmission power of satellite $P_s^{\max}$	43dBm
The antenna gain of TBS	10dBi
The antenna gain of satellite	35dBi
Number of content	5
Size of each content	30Mb
The cache storage capability of BS	30Mb
Zipf parameter κ	0.7
The spectral density of noise	-174dBm/Hz
Elevation angle θ	60°

We assume the popularity of F content follows a Zipf distribution [47]. Therefore, the probability of user n to require f-th content is given by $p_{f_n} = \frac{f_n^{-\kappa}}{\sum_{j=1}^{F} j^{-\kappa}}$, where κ is the skewness parameter. In general, large κ means more user requests are concentrated on less popular content. Moreover, we set the penalty parameter λ as 0.6 by default. The rest of our simulation parameters, unless otherwise stated, are given in Table I.

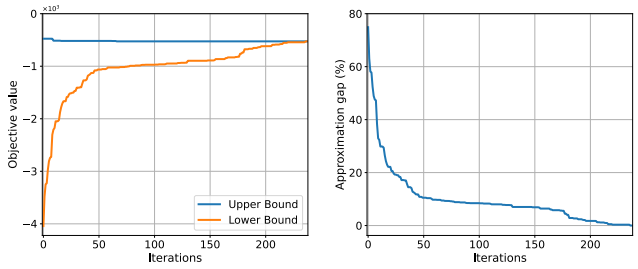
To evaluate the proposed schemes' performance, we consider two prevalent caching strategies in ISTNs as baselines [8], [48]–[52]. Besides, we also design a non-cooperative ISTN scheme as Baseline 3 to show the advantages of cooperation in the cache-enabled ISTN. The three baselines are as follows: 1) *Baseline 1 - Cooperative random caching: Similar to [8], [48], [49]*, TBSs and satellites cache the content randomly with equal probabilities regardless of content popularity distribution. 2) *Baseline 2 - Cooperative popularity-aware caching:* TBSs and satellites cache the most popular content until their storage is full [50]–[52]. In this caching scheme, the cached content in the TBS and satellite are the same since their cache sizes are the same. 3) *Baseline 3 - Non-cooperative caching*: In this scheme, TBSs and satellites provide the content delivery service to users without cooperation. Each user is only associated with one BS.

B. Comparsion of Proposed HQCGBD and CBD Algorithm

In this part, we first evaluate the convergence process of the proposed HQCGBD algorithm. Fig. 3 shows the convergence of our 5-cut HQCGBD algorithm. Fig. 3(a) shows that the

TABLE II: Performance comparison between CBD with different multi-cut HQCGBD

Algorithm	Iterations	Solver accessing time (ms)				
		Max.	Min.	Mean.	Std.	Total
CBD	828	92.21	0.59	35.87	18.58	29699.74
Single-cut HQCGBD	768	3.21	1.58	2.26	0.79	1738.65
3-cut HQCGBD	351	3.21	1.58	2.24	0.78	787.35
5-cut HQCGBD	237	3.21	1.58	2.12	0.75	502.41



(a) The upper and lower bounds of each iteration. (b) Approximation gap of each iteration.

Fig. 3: Convergence of the proposed HQCGBD algorithm.

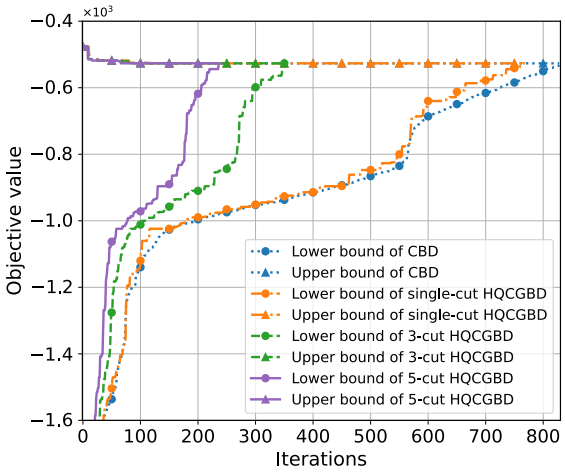


Fig. 4: The convergence performances of CBD and different multi-cut HQCGBD strategies.

difference between UB and LB decreases and finally converges after 237 iterations. Fig. 3(b) illustrates the corresponding approximation gap decreases w.r.t. the iteration number. In practice, to speed up the process, we can terminate the algorithm earlier to obtain an approximately optimal solution when the approximation gap is within a threshold ϵ.

In Fig. 4, we further compare the convergence of CBD and different multi-cut HQCGBD strategies. We can observe that the values of the lower bounds (i.e., the objective values of master problems) keep increasing while the upper bounds (i.e., the objective values of subproblems) keep decreasing until convergence. Specifically, the single-cut HQCGBD and CBD need 768 and 828 iterations to converge, respectively. This result verifies that our proposed single-cut HQCGBD algorithm is mathematically consistent with the CBD algorithm. Essentially, if the CBD algorithm can solve a problem, our proposed single-cut HQCGBD algorithm can reach the same result at least. Furthermore, we can see that the lower

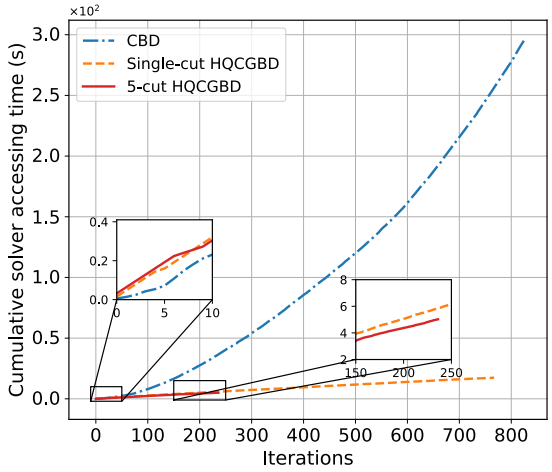
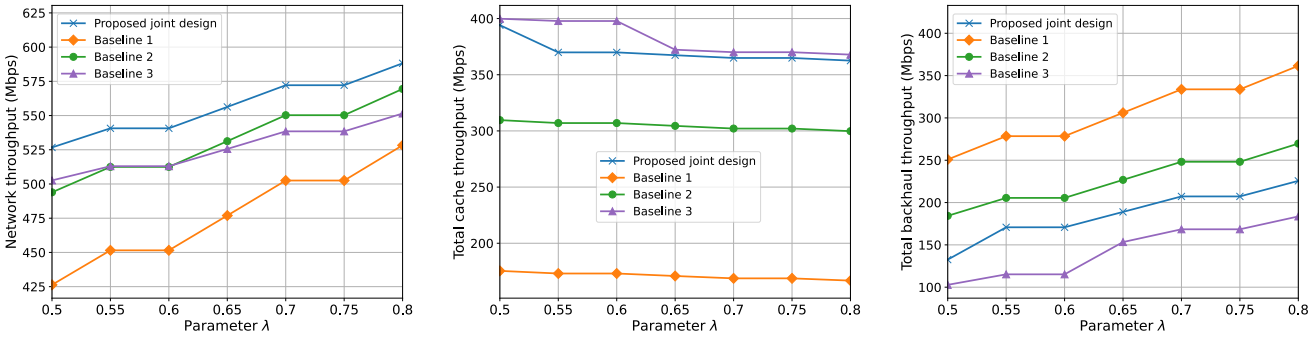


Fig. 5: The cumulative solver accessing time of master problems for CBD and different multi-cut HQCGBD strategies.

bounds of CBD and single-cut HQCGBD increase very slowly compared to the multi-cut HQCGBD strategies. The reason is that we can improve the obtained lower bounds through the multi-cut strategies. This figure also shows the superiority of the 3-cut and 5-cut HQCGBD strategies, which can reduce the number of required iterations by 57.60% and 71.37%, respectively.

Next, we compare the running time of our proposed algorithm and CBD. Although each iteration of HQCGBD requires processing multiple subproblems, the complexity of each subproblem is the same as that of the subproblem in CBD, and they can be executed in parallel. Therefore, we only compare the performances of CBD and different multi-cut HQCGBD strategies regarding the real solver accessing time of the master problems[1]. As illustrated in Fig. 5, the single-cut HQCGBD master problem's cumulative solver accessing time increases linearly w.r.t. the iteration number, while the CBD master problems' cumulative solver accessing time increases quadratically w.r.t. the iteration number. Before the 10-th iteration, the CBD outperforms the single-cut HQCGBD. However, the single-cut HQCGBD performs better and better when the master problem becomes more and more computation-intensive as we keep adding Benders' cuts to the master problem in each iteration. The computational time of the master problems on the quantum computers is less than that spend on the classical computers. This result demonstrates that quantum computers outperform classical computers in solving large-scale MILP problems. Specifically, the proposed

[1]The solver accessing time is the accessing time of QPU solver and local solver without considering other overheads, such as variables setup latency, network transmission latency, etc.



(a) Network throughput (b) Total cache throughput (c) Total backhaul throughput

Fig. 6: The impact of λ on the system performance.

single-cut, 3-cut, and 5-cut HQCGBD can save up to 94.14%, 97.34%, and 98.31% solver accessing time of the master problem compared with CBD, respectively.

We also show the performance details of CBD and multi-cut HQCGBD strategies in Table II. We can observe that the multi-cut HQCGBD strategies have a stable computation performance as they have a much smaller standard deviation of the master problem's solver accessing time than the CBD. In summary, the proposed multi-cut HQCGBD outperforms CBD in terms of both convergence speed of iterations and cumulative master problems' solver accessing time.

C. Impact of Parameter λ

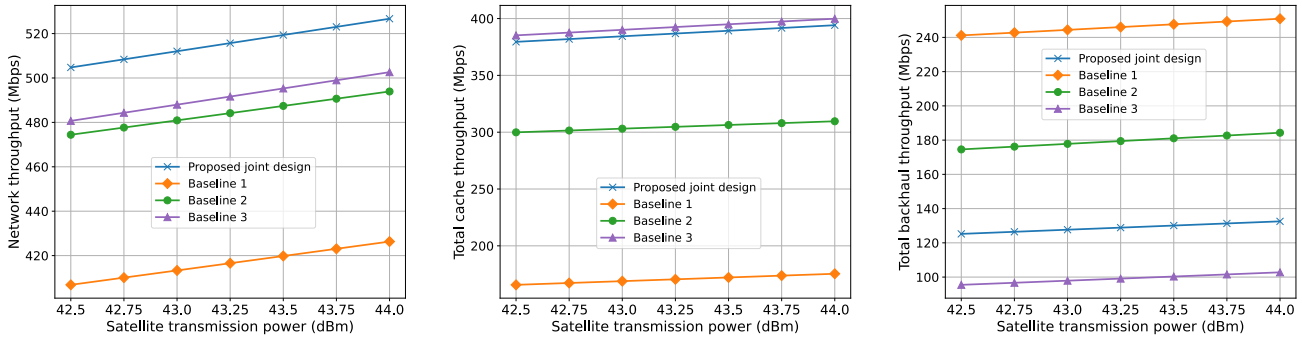
In Fig. 6, we study how the trade-off parameter λ affects the system performance in terms of the network throughput Q^T, total cache throughput $\sum_{f\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} Q^C_{n,m,f}$, and total backhaul throughput $\sum_{f\in\mathcal{F}}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} \lambda Q^B_{n,m,f}$. To that end, we increase the penalty parameter λ from 0.5 to 0.8. From Fig. 6(a), we can observe that the proposed ISTN scheme performs better than the other three baselines. Meanwhile, as λ increases, the network throughput of the non-cooperative caching ISTN scheme is first larger and then smaller than the cooperative popularity-aware caching ISTN scheme. Fig. 6(b) and Fig. 6(c) further demonstrate the details. Specifically, when λ is below 0.6, the weight of the total cache throughput is relatively large. Due to the cache optimization, the cooperative popularity-aware caching ISTN scheme outperforms the non-cooperative caching ISTN scheme. However, the total backhaul throughput becomes dominant when λ is larger. The non-cooperative caching ISTN scheme decreases the total cache throughput to increase the total backhaul throughput. The cooperative popularity-aware caching ISTN scheme sacrifices less total cache throughput to improve the total backhaul throughput due to the cooperative ISTN. Fig. 6(b) also shows that our proposed ISTN scheme outperforms the cooperative random and popularity-aware caching ISTN schemes in the total cache throughput. Although the non-cooperative caching ISTN scheme has a slightly larger total cache throughput, the proposed ISTN scheme has a much higher total backhaul throughput as shown in Fig. 6(c).

D. Impact of Satellite Transmission Power

In this part, we investigate how the network throughput, total cache, and total backhaul throughput change as the satellite transmission power increases. As demonstrated in Fig. 7(a), the network throughput increases as the satellite transmission power increases. It also shows that our proposed ISTN scheme achieves the highest network throughput than the other three baselines. The network throughput of the cooperative random caching ISTN scheme performs the worst. The reason is that the cooperative random caching ISTN scheme can not fully exploit the satellite cache capacity. As shown in Fig. 7(b) and Fig. 7(c), the proposed ISTN scheme can better utilize the cache storages of the TBSs and satellites and efficiently reduce the backhaul pressure in this case. Without cache placement optimization, the cooperative random caching and cooperative popularity-aware caching ISTN schemes push more throughput to backhaul traffic to achieve better network throughput performance. Although our proposed ISTN scheme has a similar total cache throughput to the non-cooperative caching ISTN scheme, our proposed ISTN scheme has a much higher total backhaul throughput. In general, the proposed ISTN has the ability to effectively leverage the power of satellite transmission by balancing the total cache and backhaul throughput.

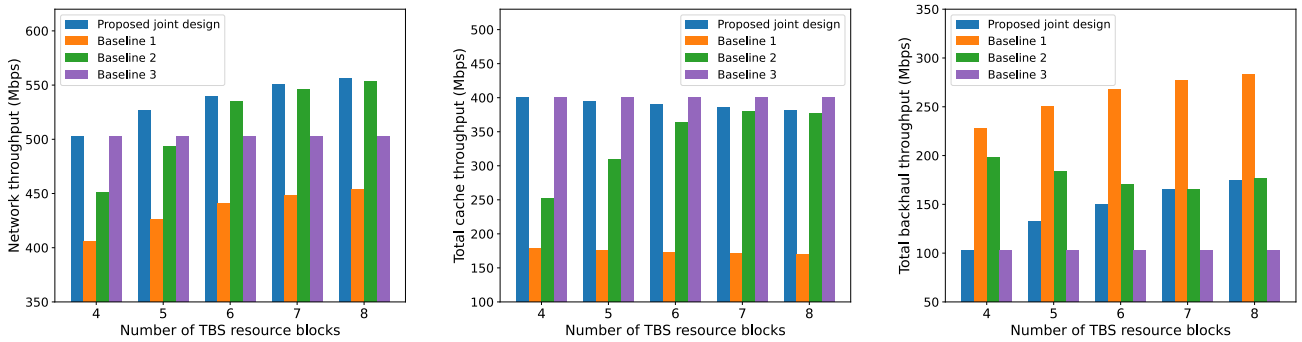
E. Impact of Resource Blocks

In this part, we compare the network throughput, total cache, and backhaul throughput of our proposed ISTN scheme with the three baseline schemes regarding the TBS resource blocks. As shown in Fig. 8(a), our proposed ISTN scheme has the highest network throughput than the other three baselines in the different numbers of TBS resource blocks. Besides, the network throughput of all schemes increases as the number of resource blocks at TBS increases, except that the network throughput of the non-cooperative caching ISTN scheme remains stable. The reason is that the transmission power of TBS is optimally allocated to each associated user. With more resource blocks available, TBSs can allocate their transmission powers more flexibly to increase the network throughput. However, the non-cooperative caching ISTN scheme does not have this flexibility since each user is served by only one



(a) Network throughput (b) Total cache throughput (c) Total backhaul throughput

Fig. 7: The impact of satellite transmission power on the system performance.



(a) Network throughput (b) Total cache throughput (c) Total backhaul throughput

Fig. 8: The impact of TBS resource blocks on the system performance.

BS. From Fig. 8(b), we can observe that the total cache throughput of the popularity-aware caching ISTN scheme increases as the number of resource blocks increases, and the cooperative random caching ISTN scheme performs the worst. Fig. 8(c) illustrates that the total backhaul throughput of the proposed ISTN scheme and the cooperative random caching ISTN scheme increases when the number of TBS resource blocks increases while the total backhaul throughput of the cooperative popularity-aware caching ISTN scheme keeps decreasing. This is because the cooperative popularity-aware caching ISTN scheme can not fully exploit the BS caching capacity with a small number of TBS resource blocks. Since the total cache throughput has a higher weight, the cooperative popularity-aware caching ISTN scheme needs to shift the network throughput from the backhaul link to the cache to increase the network throughput. Therefore, considering the trade-off between the total cache and backhaul throughput, our proposed ISTN scheme has the best performance in different numbers of TBS resource blocks.

VI. Conclusion

In this paper, we have investigated the problem of optimizing content delivery services to terrestrial users in ISTN. We have formulated a MINLP to jointly optimize the content delivery policy, cache placement, and transmission power allocation for maximizing the network throughput and proposed a hybrid quantum-classical solution method called HQCGBD to solve it. Furthermore, we have designed a specialized quantum multi-cut strategy to accelerate the convergence speed of HQCGBD. Due to the limited number of qubits in the current D-wave system, we conduct the experiments in a small-scale setting as a proof of concept. However, even in this setting, numerical results demonstrate that our proposed multi-cut HQCGBD can reduce both the required iteration numbers to achieve convergence and solver accessing time without losing optimality. This work is our first attempt to leverage quantum computing techniques for optimizing cache placement in ISTNs. In the future, we will extend our work to a setting of multi-timescale systems and investigate cache placement and resource allocation in ISTNs. Since the proposed algorithm can efficiently solve large-scale MINLPs, it holds great promise for various ISTN applications, e.g., routing and scheduling optimization problems. With the rapid development of quantum computers and an increasing number of qubits [53], we believe that the proposed quantum-assisted optimization can play a crucial role in the ISTN field.

References

[1] "Mobile data traffic outlook," Nov. 2022. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-traffic-forecast

[2] H. Li, K. Ota, and M. Dong, "ECCN: orchestration of edge-centric computing and content-centric networking in the 5G radio access network," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 88–93, Jun. 2018.

[3] Z. Piao, M. Peng, Y. Liu, and M. Daneshmand, "Recent advances of edge cache in radio access networks for internet of things: techniques, performances, and challenges," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 1010–1028, Aug. 2018.

[4] K. An, Y. Li, X. Yan, and T. Liang, "On the performance of cache-enabled hybrid satellite-terrestrial relay networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1506–1509, Jun. 2019.

[5] 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.

[6] A. U. Chaudhry and H. Yanikomeroglu, "Free space optics for next-generation satellite networks," *IEEE Consum. Electron. Mag.*, vol. 10, no. 6, pp. 21–31, Oct. 2020.

[7] X. Zhu and C. Jiang, "Integrated satellite-terrestrial networks toward 6G: Architectures, applications, and challenges," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 437–461, Nov. 2021.

[8] D. Han, W. Liao, H. Peng, H. Wu, W. Wu, and X. Shen, "Joint cache placement and cooperative multicast beamforming in integrated satellite-terrestrial networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3131–3143, Dec. 2021.

[9] D.-H. Tran, S. Chatzinotas, and B. Ottersten, "Satellite-and cache-assisted UAV: A joint cache placement, resource allocation, and trajectory optimization for 6G aerial networks," *IEEE Open J. Veh. Technol.*, vol. 3, pp. 40–54, Jan. 2022.

[10] A. Alsharoa and M.-S. Alouini, "Improvement of the global connectivity using integrated satellite-airborne-terrestrial networks with resource optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5088–5100, Apr. 2020.

[11] J. Li, K. Xue, D. S. Wei, J. Liu, and Y. Zhang, "Energy efficiency and traffic offloading optimization in integrated satellite/terrestrial radio access networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2367–2381, Jan. 2020.

[12] Q. Tang, Z. Fei, B. Li, and Z. Han, "Computation offloading in LEO satellite networks with hybrid cloud and edge computing," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9164–9176, Feb. 2021.

[13] C. Ding, J.-B. Wang, H. Zhang, M. Lin, and G. Y. Li, "Joint optimization of transmission and computation resources for satellite and high altitude platform assisted edge computing," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1362–1377, Aug. 2021.

[14] D. Han, Q. Ye, H. Peng, W. Wu, H. Wu, W. Liao, and X. Shen, "Two-timescale learning-based task offloading for remote IoT in integrated satellite-terrestrial networks," *IEEE Internet of Things J.*, vol. 10, no. 12, pp. 10 131–10 145, Jan. 2023.

[15] G. Cui, P. Duan, L. Xu, and W. Wang, "Latency optimization for hybrid GEO–LEO satellite-assisted iot networks," *IEEE Internet of Things J.*, vol. 10, no. 7, pp. 6286–6297, Nov. 2022.

[16] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke *et al.*, "Noisy intermediate-scale quantum algorithms," *Rev. Mod. Phys.*, vol. 94, no. 1, p. 015004, Jan. 2022.

[17] "Highlights of the ibm quantum summit 2022," Apr. 2023. [Online]. Available: https://www.ibm.com/quantum

[18] "Amazon braket - accelerate quantum computing research," Apr. 2023. [Online]. Available: https://https://aws.amazon.com/braket/

[19] T. Kadowaki and H. Nishimori, "Quantum annealing in the transverse ising model," *Phys. Rev. E*, vol. 58, no. 5, p. 5355, Nov. 1998.

[20] "D-wave hybrid solver service: An overview," Apr. 2023. [Online]. Available: https://www.dwavesys.com/resources/white-paper/d-wave-hybrid-solver-service-an-overview/

[21] S. Yarkoni, A. Alekseyenko, M. Streif, D. Von Dollen, F. Neukart, and T. Bäck, "Multi-car paint shop optimization with quantum annealing," in *Proc. IEEE Int. Conf. Quantum Comput. Eng*, Oct. 2021, pp. 35–41.

[22] D. M. Fox, K. M. Branson, and R. C. Walker, "mRNA codon optimization with quantum computers," *PloS one*, vol. 16, no. 10, p. e0259101, Oct. 2021.

[23] V. K. Mulligan, H. Melo, H. I. Merritt, S. Slocum, B. D. Weitzner, A. M. Watkins, P. D. Renfrew, C. Pelissier, P. S. Arora, and R. Bonneau, "Designing peptides on a quantum computer," *BioRxiv*, p. 752485, Sep. 2019.

[24] D. M. Fox, C. M. MacDermaid, A. M. Schreij, M. Zwierzyna, and R. C. Walker, "RNA folding using quantum computers," *PLoS Comput. Biol.*, vol. 18, no. 4, p. e1010032, Apr. 2022.

[25] S. Mugel, M. Abad, M. Bermejo, J. Sánchez, E. Lizaso, and R. Orús, "Hybrid quantum investment optimization with minimal holding period," *Sci. Rep.*, vol. 11, no. 1, p. 19587, Oct. 2021.

[26] M. Lee, N. Ma, G. Yu, and H. Dai, "Accelerating generalized benders decomposition for wireless resource allocation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1233–1247, Oct. 2020.

[27] L. Fan and Z. Han, "Hybrid quantum-classical computing for future network optimization," *IEEE Netw.*, vol. 36, no. 5, pp. 72–76, Nov. 2022.

[28] A.-D. Doan, M. Sasdelli, D. Suter, and T.-J. Chin, "A hybrid quantum-classical algorithm for robust fitting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 417–427.

[29] T. Q. Dinh, S. H. Dau, E. Lagunas, and S. Chatzinotas, "Efficient hamiltonian reduction for quantum annealing on satcom beam placement problem," in *Proc. IEEE Int. Conf. Commun.* IEEE, May 2023, pp. 2668–2673.

[30] T. He, H. Khamfroush, S. Wang, T. La Porta, and S. Stein, "It's hard to share: Joint service placement and request scheduling in edge clouds with sharable and non-sharable resources," in *IEEE Int. Conf. Distrib. Comput. Syst.*, Jul. 2018, pp. 365–375.

[31] 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, 2018.

[32] A. Abdi, W. C. Lau, M.-S. Alouini, and M. Kaveh, "A new simple model for land mobile satellite channels: First-and second-order statistics," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 519–528, May 2003.

[33] S. Burer and A. N. Letchford, "Non-convex mixed-integer nonlinear programming: A survey," *Surv. Oper. Res. Manag. Sci.*, vol. 17, no. 2, pp. 97–106, Jul. 2012.

[34] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, Mar. 2004.

[35] M. ApS, "Mosek optimizer API for python," *Version*, vol. 9, no. 17, pp. 6–4, Apr. 2022.

[36] A. M. Geoffrion, "Generalized benders decomposition," *J. Optim. Theory Appl.*, vol. 10, pp. 237–260, Oct. 1972.

[37] D. Li, X. Sun *et al.*, *Nonlinear integer programming*. Springer, Aug. 2006.

[38] W. Scherer, *Mathematics of quantum computing*. Springer, Aug. 2019.

[39] Z. Zhao, L. Fan, and Z. Han, "Hybrid quantum Benders' decomposition for mixed-integer linear programming," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2022, pp. 2536–2540.

[40] G. Kochenberger, J.-K. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang, "The unconstrained binary quadratic programming problem: a survey," *J. Comb. Optim.*, vol. 28, pp. 58–81, Jul. 2014.

[41] T. L. Magnanti and R. T. Wong, "Accelerating Benders' decomposition: Algorithmic enhancement and model selection criteria," *Oper. Res.*, vol. 29, no. 3, pp. 464–484, Jun. 1981.

[42] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2021.

[43] X. Xu, Z. Sun, X. Dai, T. Svensson, and X. Tao, "Modeling and analyzing the cross-tier handover in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7859–7869, Sep. 2017.

[44] S. Fu, J. Gao, and L. Zhao, "Integrated resource management for terrestrial-satellite systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3256–3266, Jan. 2020.

[45] J. Shi, W. Yu, Q. Ni, W. Liang, Z. Li, and P. Xiao, "Energy efficient resource allocation in hybrid non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3496–3511, Jan. 2019.

[46] N. Saeed, A. Elzanaty, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "CubeSat communications: Recent advances and future challenges," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 1839–1862, Apr. 2020.

[47] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Dec. 2015.

[48] Q. T. Ngo, K. T. Phan, W. Xiang, A. Mahmood, and J. Slay, "Two-tier cache-aided full-duplex hybrid satellite–terrestrial communication networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 3, pp. 1753–1765, Oct. 2021.

[49] C. Jiang and Z. Li, "Decreasing big data application latency in satellite link by caching and peer selection," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2555–2565, May 2020.

[50] D. Han, H. Peng, H. Wu, W. Liao, and X. S. Shen, "Joint cache placement and content delivery in satellite-terrestrial integrated C-RANs," in *IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[51] H. Wu, J. Li, H. Lu, and P. Hong, "A two-layer caching model for content delivery services in satellite-terrestrial networks," in *IEEE Glob. Commun. Conf.*, Dec. 2016, pp. 1–6.

[52] S. Gu, X. Sun, Z. Yang, T. Huang, W. Xiang, and K. Yu, "Energy-aware coded caching strategy design with resource optimization for satellite-UAV-vehicle-integrated networks," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5799–5811, Mar. 2021.
[53] "D-wave demonstrates large-scale coherent quantum annealing," Oct. 2023. [Online]. Available: https://www.dwavesys.com/company/newsroom/

Yu Zhang (S'19) received the B.Eng. degree in navigation technology from Wuhan University of Technology, Wuhan, China, in 2013, and the M.S. degree in information technology and management from University of Texas at Dallas, Dallas, Tx, USA, in 2017, respectively. He is pursuing the Ph.D. degree in electrical and computer engineering at University of Texas at San Antonio, San Antonio, TX, USA, where he works on resource management for satellite-terrestrial networks. He is a student member of IEEE.

Yanmin Gong (M'16, SM'21) received the B.Eng. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2012, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2016. She is currently an Associate Professor of Electrical and Computer Engineering with The University of Texas at San Antonio, San Antonio, TX, USA. Her research interests lie at the intersection of machine learning, cybersecurity, and networking systems. Dr. Gong is a recipient of the NSF CAREER Award, Cisco Research Award, the NSF CRII Award, the IEEE Computer Society TCSC Early Career Researchers Award for Excellence in Scalable Computing, the Rising Star in Networking and Communications Award by IEEE ComSoc N2Women, and the Best Paper Award at IEEE GLOBECOM. She has been serving on the editorial boards of ACM Computing Surveys, IEEE Transactions on Dependable and Secure Computing, and IEEE Wireless Communications.

Lei Fan (M'15, SM'20) is an Assistant Professor in the Department of Engineering Technology and holds a joint appointment with the Department of Electrical and Computer Engineering at the University of Houston. Prior to this position, he worked in the power industry for several years. He earned his Ph.D. in operations research from the Department of Industrial and Systems Engineering at the University of Florida. His research interests include power system operations and planning, electricity markets, optimization algorithms, complex network systems, and quantum computing.

Yu Wang (S'02-M'04-SM'10-F'18) is a Professor and Chair of the Department of Computer and Information Sciences at Temple University. He holds a Ph.D. from Illinois Institute of Technology, an MEng and a BEng from Tsinghua University, all in Computer Science. His research interest includes wireless networks, smart sensing, and distributed computing. He has published over 300 papers in peer reviewed journals and conferences. He is a recipient of *Sigma Xi Award* from Illinois Institute of Technology (2004), *Ralph E. Powe Junior Faculty Enhancement Awards* from Oak Ridge Associated Universities (2006), *Outstanding Faculty Research Award* from College of Computing and Informatics at the University of North Carolina at Charlotte (2008), *Fellow of IEEE* (2018), *ACM Distinguished Member* (2020), and *IEEE Benjamin Franklin Key Award* (2024). He has served as Associate Editor for *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Cloud Computing*, among others.

Zhu Han (S'01–M'04-SM'09-F'14)) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. Dr. Han's main research targets on the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing, data science, smart grid, carbon neutralization, security and privacy. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, IEEE Vehicular Technology Society 2022 Best Land Transportation Paper Award, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018 and ACM Distinguished Speaker from 2022 to 2025, AAAS fellow since 2019, and ACM Fellow since 2024. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award (an IEEE Field Award), for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks.

Yuanxiong Guo (M'14, SM'20) received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2012 and 2014, respectively. He is currently an Associate Professor in the Department of Information Systems and Cyber Security at the University of Texas at San Antonio, San Antonio, TX, USA. His current research interests include distributed machine learning, applied data science, and trustworthy AI with application to digital health, energy sustainability, and human-robot collaboration. He is on the Editorial Board of IEEE Transactions on Vehicular Technology and servers as the track co-chair for IEEE VTC 2021-Fall. He is a recipient of the Best Paper Award in the IEEE GLBOECOM 2011.