IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 29 August 2022, accepted 23 September 2022, date of publication 4 October 2022, date of current version 11 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211940

RESEARCH ARTICLE

Joint Caching and User Association Optimization for Adaptive Bitrate Video Streaming in UAV-Assisted Cellular Networks

JUNFENG XIE[1,2], **ZHAOBA WANG**[1], **AND YOUXING CHEN**[1]

[1]School of Information and Communication Engineering, North University of China, Taiyuan 030051, China
[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Junfeng Xie (xiejunfeng@nuc.edu.cn)

This work was supported in part by the Research Project Supported by Shanxi Scholarship Council of China under Grant 2022-147; in part by the Open Foundation of State key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, under Grant SKLNST-2020-1-12; and in part by the Research Project Supported by Fundamental Research Program of Shanxi Province.

ABSTRACT Edge caching and adaptive bitrate video streaming are two promising techniques to ensure a good video viewing experience. Edge caching can bring contents closer to users to alleviate redundant content transmissions, reduce user-perceived delay and improve transmission capability. Adaptive bitrate video streaming is able to adaptively adjust video quality based on time-varying network conditions and different users' preference. Due to the strong coupled relationship between caching and user association, in this article, we focus on the issue of joint caching and user association optimization for adaptive bitrate video streaming in UAV-assisted cellular networks. First, we formulate the optimization problem as a non-linear integer programming (NLIP) to minimize the content delivery delay. To solve this challenging NP-hard problem, a heuristic algorithm based on quantum-inspired evolutionary algorithm (QEA) is proposed to obtain the best caching and user association solutions iteratively. Finally, simulations are conducted to demonstrate that compared with three benchmark algorithms without joint optimization of caching and user association, the proposed algorithm can greatly improve users' video viewing experience and achieve better system performance in terms of reducing the total content delivery delay.

INDEX TERMS Edge caching, adaptive bitrate video streaming, user association, unmanned aerial vehicle, quantum-inspired evolutionary algorithm.

I. INTRODUCTION

With the rapid development of mobile video technologies, such as 4K/8K, 3D, augmented reality (AR)/virtual reality (VR)/mixed reality(MR), high frame rate (HFR) and high dynamic range (HDR), global data traffic has increased dramatically in recent years and is estimated to keep growing for the next decades. According to a recent report from Ericsson [1], global mobile data traffic is estimated to reach 282EB per month in 2027, and nearly 79% of the all traffic is video traffic. Thus, providing a good video viewing experience for users is very important.

Recently, unmanned aerial vehicle (UAV), which is also known as drone, has become an emerging research filed because of its small size, low price and high flexibility. Compared to conventional terrestrial cellular networks, UAVs transmit videos to ground users over air-to-ground (A2G) channels, which enhance the probability of line-of-sight (LoS) radio access links [2], [3]. With these characteristics, such as flexible deployment and high probability of LoS links, UAVs can enhance throughput, reduce transmission latency and improve the quality of video viewing experience. Because of its promising performance, UAV-assisted cellular

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa Zaman Chowdhury.

networks have received extensive attention from both industry and academia [4], [5], [6], [7].

Edge caching [8], [9], [10], [11] is a key enabling technique to improve content transmission and enhance content delivery efficiency by caching video contents closer to users. In UAV-assisted cellular networks, by deploying storage resources at UAVs, a part of popular video contents can be prestored at UAVs during off-peak hours. If the requested video content has be cached by UAVs, it will be transmitted to users directly, thereby alleviating the traffic pressure of ground base stations (BSs) during peak hours, relieving the backhaul bottleneck, reducing network congestion and content delivery latency, as well as improving users' quality of experience (QoE).

Adaptive bitrate video streaming is another promising technique to provide a good video viewing experience for users. Because of the time-varying wireless network conditions and the heterogeneity of devices' processing capabilities, users' preference for the quality of a video might be different. For example, when network connection is good and devices are highly capable, users usually prefer high quality videos. Adaptive bitrate video streaming technology [12], [13], [14] is able to adaptively adjust video quality according to the high dynamics of network conditions and different users' preference. The basic principle of adaptive bitrate video streaming is to divide a video into multiple chunks, each of which is encoded into multiple bitrates. Different bitrates of a video chunk have different quality levels. Thus, the appropriate bitrates of video chunks should be selected and transmitted to users.

In UAV-assisted cellular networks, combining edge caching and adaptive bitrate video streaming is able to further improve users' video viewing experience. However, due to caching capacity limitation, it is impossible for UAVs to cache all bitrates of video contents. Thus, designing an effective caching strategy for adaptive bitrate video streaming is important. However, caching strategy is strongly coupled with user association strategy. On the one hand, user association strategy can be affected by both wireless channel conditions and UAVs' cache state, i.e., whether UAVs have cached the requested contents or not, so caching strategy has a direct influence on user association strategy. On the other hand, although caching video contents in UAVs can reduce the content delivery latency, an inappropriate user association will negatively affect the performance gain of edge caching, so user association strategy also has an effect on the design of caching strategy. Thus, there is a strong coupled relationship between caching strategy and user association strategy. In this case, to improve the user-perceived network performance and fully utilize the system resources, the caching strategy and user association strategy should be considered and optimized jointly.

In this article, we investigate the issue of joint caching and user association optimization for adaptive bitrate video streaming in UAV-assisted cellular networks. Our optimization target is to minimize the delay of video streaming service by designing user association strategy and caching strategy (caching appropriate bitrates of video chunks at UAVs). More specifically, the main contributions of this article are summarized as follows:

- By deploying caching and computing resources at UAVs, we consider the UAV-assisted cellular networks and study the issue of joint caching and user association optimization for adaptive bitrate video streaming. Three different cases for processing the video chunk requests and responds are considered, including *direct hit case*, *transcoding hit case* and *cache miss case*. Then, the joint optimization problem is formulated as a non-linear integer programming (NLIP), aiming at minimizing the total content delivery delay.
- The formulated problem is a NP-hard problem. To reduce the computation complexity, a heuristic algorithm based on quantum-inspired evolutionary algorithm (QEA) is proposed to iteratively reach the best caching and user association solutions of our formulated problem.
- Finally, we conduct extensive simulations to evaluate the convergence and performance of the proposed QEA-based caching and user association algorithm. Simulation results show that compared with the other three benchmark algorithms, the proposed algorithm can significantly enhance content delivery efficiency and achieve better system performance in terms of reducing the total content delivery delay.

The rest of this article is organized as follows. In Section II, we review the related works. In Section III, we present the system model and formulate the joint caching and user association optimization problem. Then, to solve the formulated problem, a QEA-based heuristic algorithm is proposed in Section IV. In Section V, we present the simulation results. Finally, this article is concluded in Section VI.

II. RELATED WORK

A. UAV CACHING

Edge caching can improve content transmission and enhance content delivery efficiency. Thus, the optimization of caching strategy in cache-enabled UAV networks has been widely studied. In order to maximize users' QoE and reduce transmission power of UAVs, the authors in [15] use random waypoint user mobility model to predict the location of UAVs, and present a dueling reinforcement learning-based algorithm to optimize the content caching strategy. In [16], the authors improve the content download delay and energy consumption by optimizing the caching and dynamic flight strategy. Literature [17] divides contents into a popular subset and a less popular subset, and proposes a hybrid caching strategy to enhance the overall spectral efficiency. Taking into account the vehicle mobility, request frequency, cache size and content popularity, an energy-aware coded caching strategy is presented in [18] to improve energy efficient performance. Another notable work is [19], which proposes an alternating iterative algorithm to maximize the minimum throughput by jointly optimizing cache placement, UAV trajectory and transmission power. The authors in [20] focus on a UAV-relaying network and investigate the secure transmission problem, and optimize the user association, UAV trajectory, UAV scheduling and transmission power jointly aiming at maximizing the minimum secrecy rate. In [21], the content delivery delay is minimized by optimizing the content caching strategy and UAVs' positions. However, these works mainly focus on the video contents with one bitrate, but rarely consider the video contents with multiple bitrates.

B. VIDEO CACHING FOR ADAPTIVE BITRATE VIDEO STREAMING

Given the benefits of adaptive bitrate video streaming in providing a good viewing experience for users, there are some studies focusing on video caching for adaptive bitrate video streaming. Considering the limited caching capacity, literature [22] proposes a convex optimization-based content caching scheme for HTTP adaptive bit rate (ABR) streaming aiming at maximizing users' QoE. In [23], the authors propose a bitrate version selection and resource allocation scheme called EdgeDASH, which exploits edge caching to improve the buffer stalls and cache hits. The authors in [24] optimize the QoE-driven caching placement aiming at maximizing the aggregate average video distortion reduction. In [25], to improve the system throughput and cache hit ratio, a Stackelberg game is used to formulate the issue of joint caching and radio resource allocation. The authors in [26] propose a polynomial complexity algorithm to solve the caching problem of video contents with multiple bitrates. Another notable work is [27], which investigates the caching and prefetching problem for adaptive bitrate video streaming, and proposes a QoE-oriented deep neural network model called CoLEAP aiming at improving users' QoE. However, these works do not consider utilizing the computing resources of edge nodes to transcode a video content from a higher bitrate version to a lower bitrate version. Therefore, the potential benefits of adaptive bitrate video streaming have not been fully exploited.

C. VIDEO CACHING AND TRANSCODING FOR ADAPTIVE BITRATE VIDEO STREAMING

A few related works have been conducted to consider caching and transcoding for adaptive bitrate video streaming. Taking into account the video caching and transcoding capability in Mobile Edge Computing (MEC) servers, literature [28] provides low-latency adaptive bitrate video streaming services by optimizing the video caching and transcoding strategy. In [29], the authors propose a heuristic divide-and-conquer algorithm to trade off the video caching and transcoding cost. Another notable work is [30], which proposes a multicast-enabled virtualized heterogeneous network framework to improve the utilization of communication, caching and computing resources. In [31], online iterative greedy-base adaptation (OIGA) algorithm is presented aiming at maximizing the average video bitrate. The authors in [12] aim at minimizing the video retrieval delay by optimizing the caching placement strategy and video request scheduling strategy. Literature [32] improves the U-video mean opinion score (U-vMOS) by considering video bitrate adaptation, computing resource scheduling and traffic engineering jointly. However, these works do not consider the effect of user association on the content delivery delay. Meanwhile, these works mainly focus on the terrestrial wireless network scenarios, but ignore the UAV network scenario. Motivated by this, in this article, we consider video caching and transcoding for adaptive bitrate video streaming, and study the issue of joint caching and user association optimization in UAV-assisted cellular networks aiming at minimizing the total content delivery delay.

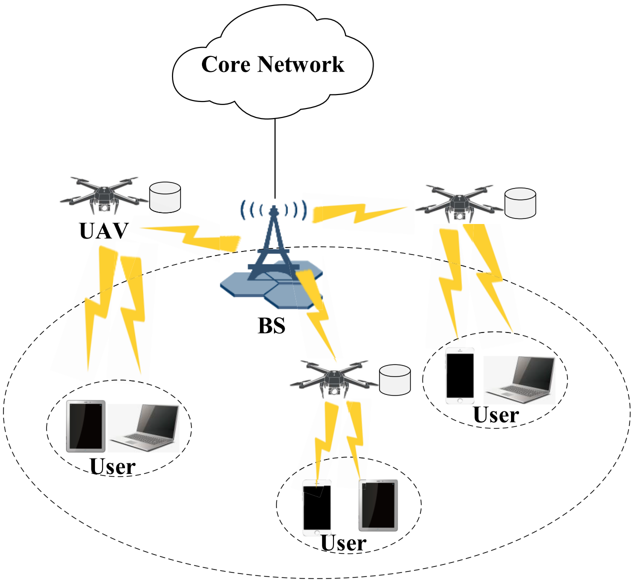

FIGURE 1. System model of a cache-enabled UAV-assisted cellular network.

III. SYSTEM DESCRIPTION

In this section, we first present the system model, and then formulate the problem of joint caching and user association optimization for adaptive bitrate video streaming.

A. SYSTEM MODEL

As shown in Figure 1, we focus on the downlink transmission of a cache-enabled UAV-assisted cellular network, which consists of one ground BS, L UAVs and K mobile users. Let $\mathcal{L} = \{1, 2, \cdots, l, \cdots, L\}$ and $\mathcal{K} = \{1, 2, \cdots, k, \cdots, K\}$ denote the set of L UAVs and the set of K mobile users, respectively. UAVs act as relay nodes to communicate with mobile users and the ground BS. The ground BS is connected to the core network through high-capacity wired optical fiber links while UAVs communicate with the ground BS through wireless backhaul links of limited capacity. Caching and computing resources are deployed at each UAV. The caching capacity and computing capacity of UAV l are denoted as C_l^{cache} and C_l^{comp}, respectively. Besides, the radio access links and wireless backhaul links are allocated with orthogonal spectrum bands, such that there is no interference between the radio access links from UAVs to users and the wireless backhaul links from the ground BS to UAVs. We assume that the bandwidth of radio access links is B, which is reused by all UAVs to communicate with users. The bandwidth of wireless backhaul links is B_h, which is equally allocated to L UAVs.

We assume that there are total F video chunks in the network. The set of F video chunks is denoted as $\mathcal{F} = \{1, 2, \cdots, f, \cdots, F\}$. Each video chunk has Z bitrate versions and the set of bitrate versions is denoted as $\mathcal{Z} = \{1, 2, \cdots, z, \cdots, Z\}$, which is indexed by an ascending order ranging from the smallest bitrate version 1 to the highest bitrate version Z. The video chunk f with bitrate version z is denoted as $v_{f,z}$. Without loss of generality, we assume that all video chunks have the same playtime duration. Thus, the size of each video chunk is proportional to its bitrate.

1) TRANSMISSION MODEL

Here, we will describe transmission channel models between UAVs and users, as well as the ground BS and UAVs. We assume that UAVs and users remain static during the process of data transmission and the hovering altitude of UAVs is H. Considering the complexity of environment and the height characteristic of UAVs, we model the radio access links from UAVs to users and the wireless backhaul links from the ground BS to UAVs as the probabilistic line-of-sight (LoS) and non-line-of-sight (NLoS) links. Thus, if user k is associated with UAV l, the LoS and NLoS path loss from UAV l to user k can be expressed as (in dB) [33], [34]:

$$g_{l,k}^{LoS} = 20\log(4\pi f_c d_0/c) + 10 n_{LoS}\log(d_{l,k}) + \varsigma_{LoS} \quad (1)$$

$$g_{l,k}^{NLoS} = 20\log(4\pi f_c d_0/c) + 10 n_{NLoS}\log(d_{l,k}) + \varsigma_{NLoS} \quad (2)$$

where $20\log(4\pi f_c d_0/c)$ denotes the free space path loss, f_c is the carrier frequency, d_0 is the free space reference distance and c is the speed of light. $d_{l,k} = \sqrt{(X_l - X_k)^2 + (Y_l - Y_k)^2 + H^2}$ represents the distance between UAV l located at (X_l, Y_l, H) and user k located at $(X_k, Y_k, 0)$. n_{LoS} and n_{NLoS} represent the path loss exponents for LoS and NLoS links, respectively. ς_{LoS} and ς_{NLoS} represent the shadowing random variables for LoS and NLoS links, respectively.

In general, the probability of LoS links is impacted by many factors such as the density and height of buildings, the location of users and UAVs, as well as the elevation angle between users and UAVs [35]. The probability of LoS links can be calculated by

$$P_{l,k}^{LoS} = \frac{1}{1 + \kappa \exp[-\zeta(\theta_{l,k} - \kappa)]} \quad (3)$$

where κ and ζ are constant values depending on the environment. $\theta_{l,k} = \sin^{-1}(H/d_{l,k})$ represents the elevation angle between user k and UAV l. Therefore, the average path loss is expressed as: $g_{l,k} = P_{l,k}^{LoS} g_{l,k}^{LoS} + P_{l,k}^{NLoS} g_{l,k}^{NLoS}$, where $P_{l,k}^{NLoS} = 1 - P_{l,k}^{LoS}$ is the probability of NLoS links.

Based on the transmission channel model of radio access links, the received signal-to-interference-plus-noise-ratio (SINR) of user k from UAV l is calculated by

$$SINR_{l,k} = \frac{P_{UAV} 10^{-g_{l,k}/10}}{\sum\limits_{l' \in \mathcal{L}, l' \neq l} P_{UAV} 10^{-g_{l',k}/10} + \sigma^2} \quad (4)$$

where P_{UAV} is the transmission power of UAVs and σ^2 is the variance of additive white Gaussian noise (AWGN). Considering the energy constraint of UAVs, we assume that the number of users associated with each UAV is limited and the maximum number is M. Thus, the downlink transmission rate from UAV l to user k can be calculated by

$$r_{l,k} = \frac{B}{M}\log_2\left(1 + SINR_{l,k}\right) \quad (5)$$

Using the similar analysis method, we can model the transmission channel model of wireless backhaul links. Because there are more obstructions between the ground BS and UAVs, the LoS and NLoS path loss from the ground BS to UAV l are expressed as (in dB) [36]:

$$g_{BS,l}^{LoS} = d_{BS,l}{}^{-\gamma} \quad (6)$$

$$g_{BS,l}^{NLoS} = \eta d_{BS,l}{}^{-\gamma} \quad (7)$$

where $d_{BS,l}$ represents the distance between the ground BS and UAV l. γ represents the path loss exponent and η represents the excessive path loss coefficient for NLoS links. Thus, the received SINR of UAV l from the ground BS can be calculated by

$$SINR_{BS,l} = \frac{P_{BS}}{10^{g_{BS,l}/10}\sigma^2} \quad (8)$$

where P_{BS} represents the transmission power of the ground BS and $g_{BS,l} = P_{BS,l}^{LoS} g_{BS,l}^{LoS} + P_{BS,l}^{NLoS} g_{BS,l}^{NLoS}$ is the average path loss between the ground BS and UAV l. Then, the downlink transmission rate from the ground BS to UAV l can be calculated by

$$r_{BS,l} = \frac{B_h}{L}\log_2\left(1 + SINR_{BS,l}\right) \quad (9)$$

2) CONTENT DELIVERY DELAY MODEL

In this article, let $\mathcal{X} = \left\{x_{l,k} \,| l \in \mathcal{L}, k \in \mathcal{K}\right\}$ denote the user association strategy and its element $x_{l,k} \in \{0, 1\}$ indicate whether user k is associated with UAV l. If user k is associated with UAV l, $x_{l,k} = 1$; otherwise, $x_{l,k} = 0$. We assume that one user can only associate with one UAV, but one UAV can be associated with several users. Let $\mathcal{Y} = \left\{y_{l,f,z} \,| l \in \mathcal{L}, f \in \mathcal{F}, z \in \mathcal{Z}\right\}$ denote the caching strategy and its element $y_{l,f,z} \in \{0, 1\}$ indicate whether $v_{f,z}$ is cached by UAV l. If $v_{f,z}$ is cached by UAV l, $y_{l,f,z} = 1$; otherwise, $y_{l,f,z} = 0$. In addition, let $p_{k,f,z} \in \{0, 1\}$ indicate whether user k requests $v_{f,z}$. If user k requests $v_{f,z}$, $p_{k,f,z} = 1$; otherwise, $p_{k,f,z} = 0$. Users' requests for video chunks are affected by the popularity of video chunks. In this article, we assume the popularity of video chunks follows a Zipf-like distribution [37].

Depending on the caching and computing capabilities, UAVs can not only cache a part of popular video chunks with different bitrates but also transcode a video chunk from a higher bitrate to a lower bitrate. According to whether UAVs cache the requested bitrate of a video chunk or a higher bitrate, there are three possible cases to process the video chunk requests and responds, including *direct hit case*, *transcoding hit case* and *cache miss case*. In the following, we will analyze the content delivery delay from UAV l to user k denoted as $D_{l,k}$ of these three cases.

Direct hit case: When UAV l has cached the requested $v_{f,z}$ (i.e., $y_{l,f,z} = 1$), user k can directly get the requested $v_{f,z}$ from UAV l. In this case, the content delivery delay only contains the downlink radio transmission delay, which can be calculated by

$$D^1_{l,k} = \sum_{f=1}^{F}\sum_{z=1}^{Z} p_{k,f,z} y_{l,f,z} \frac{s_{f,z}}{r_{l,k}} \tag{10}$$

where $s_{f,z}$ represents the size of $v_{f,z}$.

Transcoding hit case: When UAV l does not cache the requested $v_{f,z}$ (i.e., $y_{l,f,z} = 0$), but caches a higher bitrate (i.e., $h_{l,f,z} = min\left\{\sum\limits_{z'=z+1}^{Z} y_{l,f,z'}, 1\right\} = 1$), it will first transcode the cached higher bitrate to the requested bitrate, and then transmit the requested bitrate to user k. In this case, the content delivery delay contains the downlink radio transmission delay and the transcoding delay, which can be calculated by

$$\begin{aligned} D^2_{l,k} = &\sum_{f=1}^{F}\sum_{z=1}^{Z} p_{k,f,z}\left(1 - y_{l,f,z}\right) h_{l,f,z} \\ &\times \left(\frac{s_{f,z}}{r_{l,k}} + \frac{w_0(s_{f,z^+} - s_{f,z})}{c_{l,k}}\right) \end{aligned} \tag{11}$$

where $h_{l,f,z} \in \{0, 1\}$ indicates whether UAV l caches the video chunk f with at least one bitrate version higher than z. If UAV l caches the video chunk f with at least one bitrate version higher than z, $h_{l,f,z} = 1$; otherwise, $h_{l,f,z} = 0$. $z^+ = \left\{z' \left| min(z' - z), z < z', y_{l,f,z'} = 1\right.\right\}$ represents the minimum cached bitrate version of video chunk f higher than z. w_0 represents the computing resources required to process one bit input. $c_{l,k}$ represents the computing resources allocated to user k by UAV l. Thus, the transcoding delay can be expressed as $\frac{w_0(s_{f,z^+}-s_{f,z})}{c_{l,k}}$ [38]. Because the computing capacity of UAV l is C^{comp}_l and the maximum number of users associated with each UAV is M, we assume $c_{l,k} = \frac{C^{comp}_l}{M}$.

Cache miss case: In this article, we assume all video chunks with different bitrates are stored in the core network. Because the ground BS is connected to the core network through high-capacity wired optical fiber links, all video chunks with different bitrates are available at the ground BS. When UAV l caches neither the requested $v_{f,z}$ (i.e., $y_{l,f,z} = 0$) nor a higher bitrate (i.e., $h_{l,f,z} = 0$), it will first obtain the requested $v_{f,z}$ from the core network via wireless backhaul links with the ground BS, and then transmit the requested $v_{f,z}$ to user k. In this case, the content delivery delay contains the downlink radio transmission delay and the backhaul link transmission delay, which can be calculated by

$$D^3_{l,k} = \sum_{f=1}^{F}\sum_{z=1}^{Z} p_{k,f,z}\left(1 - \sum_{z'=z}^{Z} y_{l,f,z'}\right)\left(\frac{s_{f,z}}{r_{l,k}} + \frac{s_{f,z}}{r_{BS,l}}\right) \tag{12}$$

Therefore, the content delivery delay from UAV l to user k is given by

$$D_{l,k} = D^1_{l,k} + D^2_{l,k} + D^3_{l,k} \tag{13}$$

B. PROBLEM FORMULATION

Based on the above analysis, the total content delivery delay of all users is expressed as

$$D_{tot} = \sum_{l=1}^{L}\sum_{k=1}^{K} x_{l,k} D_{l,k} \tag{14}$$

Consequently, we formulate the joint optimization problem of caching and user association for adaptive bitrate video streaming as follows.

$$\min_{\mathcal{X},\mathcal{Y}} D_{tot} \tag{15}$$

$$\text{s.t. } \sum_{f=1}^{F}\sum_{z=1}^{Z} y_{l,f,z} s_{f,z} \leq C^{cache}_l, \quad \forall l \in \mathcal{L} \tag{15a}$$

$$\sum_{l=1}^{L} x_{l,k} = 1, \quad \forall k \in \mathcal{K} \tag{15b}$$

$$\sum_{k=1}^{K} x_{l,k} \leq M, \quad \forall l \in \mathcal{L} \tag{15c}$$

$$x_{l,k} \in \{0, 1\}, \quad \forall l \in \mathcal{L}, \ k \in \mathcal{K} \tag{15d}$$

$$y_{l,f,z} \in \{0, 1\}, \quad \forall l \in \mathcal{L}, \ f \in \mathcal{F}, \ z \in \mathcal{Z} \tag{15e}$$

The target of the optimization problem is to minimize the total content delivery delay of all users. The constraints of the optimization problem are specified in (15a)-(15e). Constraint (15a) indicates that the total size of video chunks cached at each UAV is less than its caching capacity C^{cache}_l. Constraint (15b) indicates that each user can only be associated with one UAV. Constraint (15c) indicates that the number of users being associated with each UAV is not greater than M. Constraints (15d) and (15e) indicates that the two variables ($x_{l,k}$ and $y_{l,f,z}$) that need to be optimized are both binary variables.

So far, the joint caching and user association optimization problem for adaptive bitrate video streaming has been formulated and the optimal solution will be described in Section IV.

IV. PROPOSED ALGORITHM

As mentioned in Section III, the joint caching and user association optimization problem for adaptive bitrate video streaming is formulated as (15). Apparently, the variables in this optimization problem are both binary discrete variables.

Thus, the optimization problem is a non-linear integer programming (NLIP), which is NP-hard and difficult to search the best solution within polynomial time [39]. In order to reduce the computation complexity, a heuristic algorithm based on quantum-inspired evolutionary algorithm (QEA) is proposed to iteratively reach the best caching and user association solutions. In this section, we first introduce the basic principle of QEA, and then present the joint caching and user association optimization algorithm based on QEA.

A. THE BASIC PRINCIPLE OF QEA

Inspired by quantum computing theory and evolutionary theory, QEA [40] is proposed as an iterative algorithm to solve various combinatorial optimization problems. The characteristics of QEA mainly include the representation of individuals, evaluation functions and population dynamics. In QEA, each population consists of a group of individuals, and each individual consists of a string of Q-bits. A Q-bit, the smallest unit of information in QEA, may be in the "0" state, "1" state, or a linear superposition of the two states. The probabilistic representation of a Q-bit is defined as

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \tag{16}$$

where $|0\rangle$ represents the "0" state, $|1\rangle$ represents the "1" state. $|\alpha|^2$ and $|\beta|^2$ denote the probability of observing the Q-bit in "0" state and "1" state respectively. Meanwhile, to meet the probability normalization condition, $|\alpha|^2 + |\beta|^2 = 1$.

To facilitate the analysis, the probabilistic representation of a Q-bit is simplified as

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{17}$$

Based on the definition of a Q-bit, an individual with a string of m Q-bits is expressed as

$$\left[\begin{array}{c|c|c|c|c} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_m \\ \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_m \end{array}\right] \tag{18}$$

where $|\alpha_i|^2 + |\beta_i|^2 = 1$, $i = 1, 2, \ldots, m$. QEA is a generational population-based iterative algorithm. During the iterative process, quantum gates (Q-gates), such as NOT gate, Hadamard gate and rotation gate, are employed as a variation operator to update the individuals, driving the individuals towards better solutions. The mechanism of QEA, such as Q-bit probabilistic representation and Q-gate operation, allows it to effectively explore and exploit the search space of a specific optimization problem.

B. THE JOINT CACHING AND USER ASSOCIATION OPTIMIZATION ALGORITHM BASED ON QEA

According to the basic principle of QEA, we present a joint caching and user association optimization algorithm based on QEA, which consists of nine steps and is summarized in Algorithm 1. Next, we will introduce the details of each step.

- At step 1, set the generation counter $t = 0$.
- At step 2, initialize $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$.

Algorithm 1 The Joint Caching and User Association Optimization Algorithm Based on QEA

```
Step 1: Set the generation counter $t = 0$.
Step 2: Initialize $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$.
Step 3: Determine $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ by observing the states of $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$ respectively.
Step 4: Repair the obtained $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ by Algorithm 2.

Step 5: Evaluate $\mathcal{X}(t)$ and $\mathcal{Y}(t)$.
Step 6: Store the best binary solution into $B(t)_{\mathcal{X}}$ and $B(t)_{\mathcal{Y}}$.

Step 7: Update $Q(t+1)_{\mathcal{X}}$ and $Q(t+1)_{\mathcal{Y}}$ by employing Q-gates.
Step 8: Set $t = t + 1$.
Step 9: Check whether the termination condition is met. If $t > t_{MAX}$ where $t_{MAX}$ represents the maximum number of generations, return $B(t)_{\mathcal{X}}$ and $B(t)_{\mathcal{Y}}$; otherwise, go to Step 3.
```

QEA is an iterative algorithm based on generational population. $Q(t) = \left\{q_1^t, q_2^t, \ldots, q_n^t\right\}$ is defined as a population with n individuals at generation t. Here n is the population size. Because the optimization problem (15) has two binary variables $\mathcal{X}$ and $\mathcal{Y}$, to adopt QEA to the optimization problem, we have to define two population $Q(t)_{\mathcal{X}} = \left\{q_{1\mathcal{X}}^t, q_{2\mathcal{X}}^t, \ldots, q_{n\mathcal{X}}^t\right\}$ and $Q(t)_{\mathcal{Y}} = \left\{q_{1\mathcal{Y}}^t, q_{2\mathcal{Y}}^t, \ldots, q_{n\mathcal{Y}}^t\right\}$. Each individual in $Q(t)_{\mathcal{X}}$ has $L \times K$ Q-bits, while each individual in $Q(t)_{\mathcal{Y}}$ has $L \times F \times Z$ Q-bits. Thus, the ith individual in $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$ are expressed as

$$q_{i\mathcal{X}}^t = \left[\begin{array}{c|c|c|c} \alpha_{i11}^t & \alpha_{i12}^t & \cdots & \alpha_{iLK}^t \\ \beta_{i11}^t & \beta_{i12}^t & \cdots & \beta_{iLK}^t \end{array}\right] \tag{19}$$

$$q_{i\mathcal{Y}}^t = \left[\begin{array}{c|c|c|c} \alpha_{i111}^t & \alpha_{i112}^t & \cdots & \alpha_{iLFZ}^t \\ \beta_{i111}^t & \beta_{i112}^t & \cdots & \beta_{iLFZ}^t \end{array}\right] \tag{20}$$

At generation 0, we initialize $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$ by setting all α and β to $1\big/\sqrt{2}$, which means that the probability of observing each Q-bit in "0" state and "1" state is the same at the beginning.

- At step 3, determine $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ by observing the states of $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$ respectively.

$\mathcal{X}(t) = \left\{\mathcal{X}_1^t, \mathcal{X}_2^t, \ldots, \mathcal{X}_n^t\right\}$ and $\mathcal{Y}(t) = \left\{\mathcal{Y}_1^t, \mathcal{Y}_2^t, \ldots, \mathcal{Y}_n^t\right\}$ are both a group of binary solutions and are determined by observing $Q(t)_{\mathcal{X}}$ and $Q(t)_{\mathcal{Y}}$ respectively. The ith binary solution $\mathcal{X}_i^t$ and $\mathcal{Y}_i^t$ are denoted as $\mathcal{X}_i^t = \left[x_{i11}^t, x_{i12}^t, \ldots, x_{iLK}^t\right]$ and $\mathcal{Y}_i^t = \left[y_{i111}^t, y_{i112}^t, \ldots, y_{iLFZ}^t\right]$. The binary value x_{ilk}^t in $\mathcal{X}_i^t$ and y_{ilfz}^t in $\mathcal{Y}_i^t$ are obtained by

$$x_{ilk}^t = \begin{cases} 0, & random\,[0,1] > \left|\beta_{ilk}^t\right|^2 \\ 1, & random\,[0,1] < \left|\beta_{ilk}^t\right|^2 \end{cases} \tag{21}$$

$$y_{ilfz}^t = \begin{cases} 0, & random\,[0,1] > \left|\beta_{ilfz}^t\right|^2 \\ 1, & random\,[0,1] < \left|\beta_{ilfz}^t\right|^2 \end{cases} \tag{22}$$

where *random* [0, 1] is a uniformly distributed random number within [0, 1].

- At step 4, repair the obtained $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ by Algorithm 2.

Binary solutions $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ determined at step 3 may not satisfy constraints (15a)-(15c) of the formulated optimization problem in Section III-B. When $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ violate constraints (15a)-(15c), Algorithm 2 is employed to repair $\mathcal{X}(t)$ and $\mathcal{Y}(t)$, in order to make them satisfy constraints (15a)-(15c).

Algorithm 2 Binary Solution Repair Algorithm

```
for each $l \in \mathcal{L}$ do
    while $\sum_{f=1}^{F}\sum_{z=1}^{Z} y_{l,f,z}^{t} s_{f,z} > C_l^{cache}$ do
        Select an item $f$ and $z$ from the sets $\mathcal{F}$ and $\mathcal{Z}$ satisfying $y_{l,f,z}^{t} = 1$
        Set $y_{l,f,z}^{t} = 0$
    end while
    while $\sum_{f=1}^{F}\sum_{z=1}^{Z} y_{l,f,z}^{t} s_{f,z} \leq C_l^{cache}$ do
        Select an item $f$ and $z$ from the sets $\mathcal{F}$ and $\mathcal{Z}$ satisfying $y_{l,f,z}^{t} = 0$
        Set $y_{l,f,z}^{t} = 1$
    end while
end for
for each $k \in \mathcal{K}$ do
    if $\sum_{l=1}^{L} x_{l,k}^{t} > 1$ then
        Select an item $l$ from the set $\mathcal{L}$ satisfying $x_{l,k}^{t} = 1$
        Set $x_{l',k}^{t} = 0, \forall l' \in \mathcal{L}, l' \neq l$
    end if
end for
for each $l \in \mathcal{L}$ do
    while $\sum_{k=1}^{K} x_{l,k}^{t} > M$ do
        Select an item $k$ from the set $\mathcal{K}$ satisfying $x_{l,k}^{t} = 1$
        Set $x_{l,k}^{t} = 0$
    end while
end for
for each $k \in \mathcal{K}$ do
    if $\sum_{l=1}^{L} x_{l,k}^{t} = 0$ then
        Select an item $l$ from the set $\mathcal{L}$ satisfying $\sum_{k=1}^{K} x_{l,k}^{t} < M$
        Set $x_{l,k}^{t} = 1$
    end if
end for
```

- At step 5, evaluate $\mathcal{X}(t)$ and $\mathcal{Y}(t)$.

Since our formulated optimization problem aims at minimizing the total content delivery delay, the objective function in (15) is used as the evaluation function to evaluate binary solutions $\mathcal{X}(t)$ and $\mathcal{Y}(t)$.

TABLE 1. Lookup table of $\Delta\theta_{ilk}^{t}$.

x_{ilk}^{t}	b_{lk}^{t}	$f\left(\mathcal{X}_i^t\right) \geq f\left(B(t)_{\mathcal{X}}\right)$	$\Delta\theta_{ilk}^{t}$
0	0	False	0
0	0	True	0
0	1	False	0.01π
0	1	True	0
1	0	False	-0.01π
1	0	True	0
1	1	False	0
1	1	True	0

- At step 6, store the best binary solution into $B(t)_{\mathcal{X}}$ and $B(t)_{\mathcal{Y}}$.

If the generation counter $t = 0$, the best binary solution among $\mathcal{X}(t)$ is selected and stored into $B(t)_{\mathcal{X}}$, and the best binary solution among $\mathcal{Y}(t)$ is selected and stored into $B(t)_{\mathcal{Y}}$; otherwise, the best binary solution among $\mathcal{X}(t)$ and $B(t-1)_{\mathcal{X}}$ is selected and stored into $B(t)_{\mathcal{X}}$, and the best binary solution among $\mathcal{Y}(t)$ and $B(t-1)_{\mathcal{Y}}$ is selected and stored into $B(t)_{\mathcal{Y}}$.

- At step 7, update $Q(t+1)_{\mathcal{X}}$ and $Q(t+1)_{\mathcal{Y}}$ by employing Q-gates.

In this article, the rotation gate is employed to update the individuals and produce the next generation $Q(t+1)_{\mathcal{X}}$ and $Q(t+1)_{\mathcal{Y}}$. Specifically, the Q-bit of the ith individual in $Q(t+1)_{\mathcal{X}}$ is updated by

$$\begin{bmatrix} \alpha_{ilk}^{t+1} \\ \beta_{ilk}^{t+1} \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_{ilk}^{t}) & -\sin(\Delta\theta_{ilk}^{t}) \\ \sin(\Delta\theta_{ilk}^{t}) & \cos(\Delta\theta_{ilk}^{t}) \end{bmatrix} \begin{bmatrix} \alpha_{ilk}^{t} \\ \beta_{ilk}^{t} \end{bmatrix} \tag{23}$$

where $\Delta\theta_{ilk}^{t}$ denotes a rotation angle which affects the speed of convergence. At generation t, the value of $\Delta\theta_{ilk}^{t}$ is determined according to Table 1, which is recommended in [40]. In Table 1, x_{ilk}^{t} represents a value in the binary solution $\mathcal{X}_i^t$, b_{lk}^{t} represents a value in the best binary solution $B(t)_{\mathcal{X}}$ denoted as $B(t)_{\mathcal{X}} = \left[b_{11}^{t}, b_{12}^{t}, \ldots, b_{LK}^{t}\right]$, $f\left(\mathcal{X}_i^t\right)$ and $f\left(B(t)_{\mathcal{X}}\right)$ represent the objective function values of $\mathcal{X}_i^t$ and $B(t)_{\mathcal{X}}$ respectively.

Similarly, the Q-bit of the ith individual in $Q(t+1)_{\mathcal{Y}}$ can be updated by

$$\begin{bmatrix} \alpha_{ilfz}^{t+1} \\ \beta_{ilfz}^{t+1} \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_{ilfz}^{t}) & -\sin(\Delta\theta_{ilfz}^{t}) \\ \sin(\Delta\theta_{ilfz}^{t}) & \cos(\Delta\theta_{ilfz}^{t}) \end{bmatrix} \begin{bmatrix} \alpha_{ilfz}^{t} \\ \beta_{ilfz}^{t} \end{bmatrix} \tag{24}$$

- At step 8, set $t = t + 1$.
- At step 9, check whether the termination condition is met.

If $t > t_{MAX}$ where t_{MAX} represents the maximum number of generations, return $B(t)_{\mathcal{X}}$ and $B(t)_{\mathcal{Y}}$; otherwise, go to step 3.

V. SIMULATION RESULTS AND DISCUSSIONS

In this section, we use computation simulation method to evaluate the performance of our proposed algorithm (i.e., QEA-based joint caching and user association optimization algorithm for adaptive bitrate video streaming), which is called as "QEA-based CA and AU" for simplicity in the simulations. For the performance comparison, we consider the other three algorithms, including "QEA-based CA only",

TABLE 2. Simulation parameters.

Parameter	Value	Parameter	Value
L	5	$\varsigma_{LoS}, \varsigma_{NLoS}$	5.3, 5.27
K	100	κ, ζ	11.9, 0.13
F	100	P_{UAV}	$20dBm$
Z	4	P_{BS}	$30dBm$
C_l^{cache}	$6Gbits$	M	30
C_l^{comp}	$3.4GHz$	B	$20MHz$
H	$100m$	B_h	$10MHz$
f_c	$38GHz$	γ	2
d_0	$5m$	η	100
n_{LoS}, n_{NLoS}	2, 2.4	σ^2	$-95dBm$

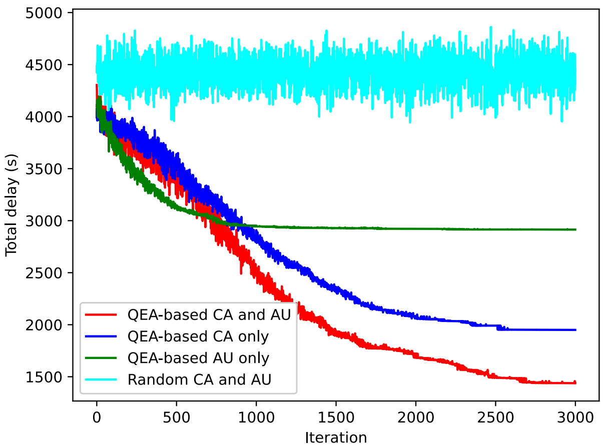

FIGURE 2. Convergence of the proposed algorithm.

"QEA-based AU only" and "Random CA and AU". "QEA-based CA only" algorithm associates users with UAVs randomly and only optimizes video caching strategy by using QEA. "QEA-based AU only" algorithm caches video chunks randomly and only optimizes user association strategy by using QEA. "Random CA and AU" algorithm not only caches video chunks randomly but also associates users with UAVs randomly.

In our simulation, we assume that there are one ground BS and L=5 UAVs in the UAV-assisted cellular network. K=100 users are randomly distributed in the network. The hovering altitude of UAVs is set to 100m. The transmission power of UAVs and the ground BS are set to P_{UAV}= 20dBm and P_{BS}= 30dBm respectively. The coverage radius of each UAV is 1km. Besides, we assume that there are F=100 video chunks, each of which has Z=4 bitrate versions. Each user requests a video chunk at a time based on the Zipf-like distribution. The detailed simulation parameters are summarized in Table 2. Note that some of these parameters are adjusted based on the evaluation scenarios. In the following, the numerical results and discussions are given.

Figure 2 shows the convergence of all the algorithms. As we can see, the total content delivery delay of "QEA-based CA only", "QEA-based AU only" and "QEA-based CA and AU" are all gradually converge with the number of iterations increasing. "QEA-based CA and AU" takes a longer time to converge than "QEA-based CA only". "QEA-based CA only" takes a longer time to converge than "QEA-based AU only". This is because "QEA-based CA and AU" needs to optimize both video caching strategy and user association strategy, "QEA-based CA only" only needs to optimize video caching strategy, and "QEA-based AU only" only needs to optimize user association strategy. The number of elements in user association strategy $\mathcal{X} = \{x_{l,k} | l \in \mathcal{L}, k \in \mathcal{K}\}$ (i.e., $L \times K = 500$) is less than the number of elements in video caching strategy $\mathcal{Y} = \{y_{l,f,z} | l \in \mathcal{L}, f \in \mathcal{F}, z \in \mathcal{Z}\}$ (i.e., $L \times F \times Z = 2000$), resulting in a faster convergence speed. In addition, as shown in Figure 2, the obtained total content delivery delay of "QEA-based CA only" is much lower than "Random CA and AU", which indicates that optimizing video caching strategy can improve system performance. Similarly, the obtained total content delivery delay of "QEA-based AU only" is much lower than "Random CA and AU", which indicates that optimizing user association strategy can improve system performance. The obtained total content delivery delay of "QEA-based CA and AU" is much lower than "QEA-based CA only" and "QEA-based AU only", which demonstrates the necessity and advantage of jointly optimizing video caching strategy and user association strategy.

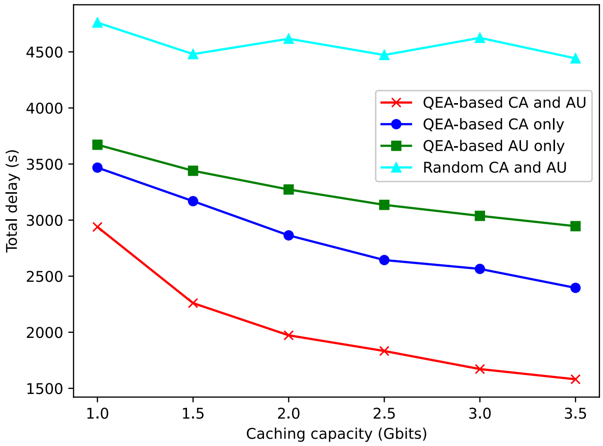

FIGURE 3. Total content delivery delay under different values of the caching capacity of each UAV.

In Figure 3, we show the impact of caching capacity of each UAV on total content delivery delay. From the figure, it can be observed that the total content delivery delay decreases with the caching capacity of each UAV increasing, which means that as the caching capacity increases, more bitrate versions of video chunks requested by users can be cached in UAVs and are not necessary to be transcoded from a higher bitrate version or be fetched from the ground BS through wireless backhaul links. In addition, as shown in Figure 3, "QEA-based CA and AU" achieves the lowest total content delivery delay compared with the other three algorithms. The obtained total content delivery delay of "QEA-based CA only" is lower than "QEA-based AU only". This is because the caching capacity of each UAV mainly influences video caching strategy, thus optimizing video caching strategy can

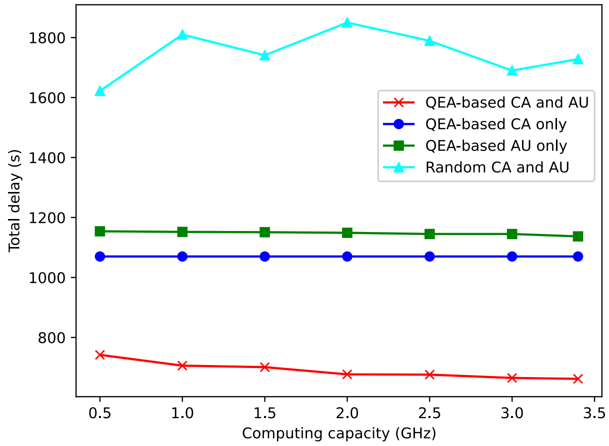

FIGURE 4. **Total content delivery delay under different values of the computing capacity of each UAV.**

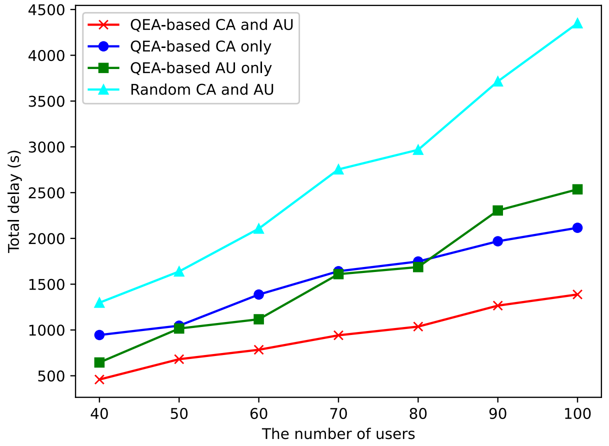

FIGURE 6. **Total content delivery delay under different values of the number of users.**

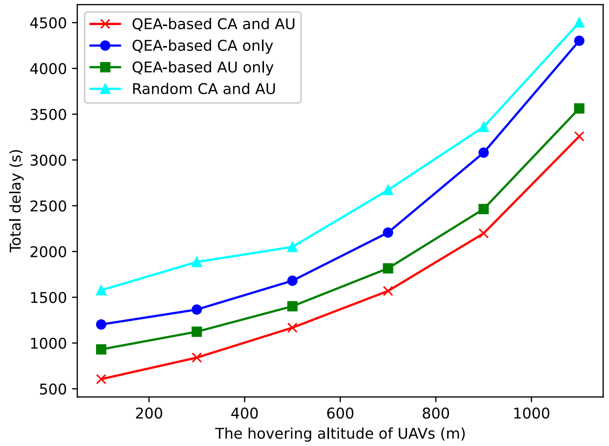

FIGURE 5. **Total content delivery delay under different values of the hovering altitude of UAVs.**

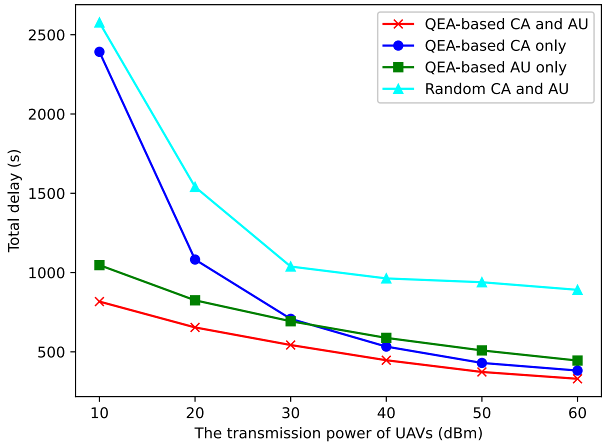

FIGURE 7. **Total content delivery delay under different values of the transmission power of UAVs.**

achieve better system performance than only optimizing user association strategy.

In Figure 4, we illustrate the relationship between the total content delivery delay and the computing capacity of each UAV. From the figure, it can be observed that the total content delivery delay of ‘‘QEA-based CA only’’, ‘‘QEA-based AU only’’ and ‘‘QEA-based CA and AU’’ decrease slowly as the computing capacity of each UAV increases. This can be explained as follows. Because each UAV can cache more than one bitrate version for a video chunk, the number of video requests requiring transcoding is reduced. Meantime, with the computing capacity of each UAV increasing, transcoding a video chunk from a higher bitrate version to the requested bitrate version requires lower delay. As a result, the transcoding delay is relatively small, and the total content delivery delay is mainly contributed by the downlink radio transmission delay and the backhaul link transmission delay.

In Figure 5, we illustrate the impact of the hovering altitude of UAVs on total content delivery delay. From the figure, it can be observed that the total content delivery delay increases with the hovering altitude of UAVs increasing. This can be explained as follows. According to Eq. (1) and (2), both the LoS and NLoS path loss from UAVs to users increase when the hovering altitude of UAVs increases, leading to the decrease of the downlink transmission rate from UAVs to users and the increase of the downlink radio transmission delay. In this case, the total content delivery delay increases. Besides, as shown in Figure 5, the obtained total content delivery delay of ‘‘QEA-based AU only’’ is lower than ‘‘QEA-based CA only’’. This is because the hovering altitude of UAVs has impact on the downlink transmission rate from UAVs to users, which mainly influences user association strategy, thus optimizing user association strategy can achieve better system performance than only optimizing video caching strategy.

Figure 6 investigates the impact of the number of users on total content delivery delay. From the figure, it can be observed that the increase of the number of users leads to the increase of the total content delivery delay. This is because the number of video requests increases when the number of users increases, resulting in the increase of the total downlink radio transmission delay, the total transcoding delay, as well as the total backhaul link transmission delay. In this case, the total content delivery delay increases. Besides, because

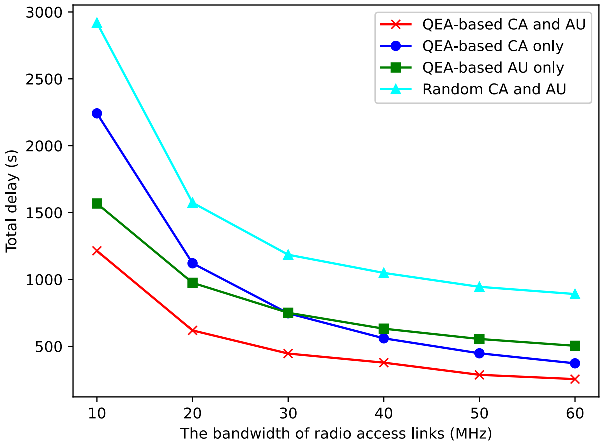

FIGURE 8. **Total content delivery delay under different values of the bandwidth of radio access links.**

the proposed "QEA-based CA and AU" can associate users with UAVs and cache video chunks smartly, the total content delivery delay of it increases with the smoothest speed, which yields better system performance than the other three algorithms.

Figure 7 shows how the total content delivery delay changes as the transmission power of UAVs varies. From the figure, it can be observed that the total content delivery delay decreases significantly as the transmission power of UAVs increases. This is because higher transmission power of UAVs leads to the increase of the downlink transmission rate from UAVs to users and the decrease of the downlink radio transmission delay.

Figure 8 illustrates the relationship between the total content delivery delay and the bandwidth of radio access links. Similar to Figure 7, the total content delivery delay decreases significantly with the bandwidth of radio access links increasing. This is because the downlink transmission rate from UAVs to users increases when the bandwidth of radio access links increases, resulting in the decrease of the downlink radio transmission delay.

VI. CONCLUSION

In this article, we have investigated the issue of joint caching and user association optimization for adaptive bitrate video streaming in UAV-assisted cellular networks. The problem was formulated as a non-linear integer programming aiming at minimizing the total content delivery delay of adaptive bitrate video streaming service. In order to reduce the computation complexity of the NP-hard problem, we presented a QEA-based heuristic algorithm to iteratively reach the best caching and user association solution, which caches the proper bitrate versions of video chunks at UAVs and selects the appropriate user-UAV association relationship. Simulation results have been given to demonstrate the convergence of the proposed algorithm and its better performance than three benchmark algorithms in terms of reducing content delivery delay and enhancing content delivery efficiency. In future works, it would be interesting to jointly consider UAV deployment, resource allocation, content caching and user association for adaptive bitrate video streaming to enhance system performance and improve users' QoE.

REFERENCES

[1] Ericsson. (Jun. 2022). *Ericsson Mobility Report*. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report

[2] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, and L. Wang, "Power control in UAV-supported ultra dense networks: Communications, caching, and energy transfer," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 28–34, Jun. 2018.

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

[4] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart., 2019.

[5] R. Shahzadi, M. Ali, H. Z. Khan, and M. Naeem, "UAV assisted 5G and beyond wireless networks: A survey," *J. Netw. Comput. Appl.*, vol. 189, Sep. 2021, Art. no. 103114.

[6] B. Alzahrani, O. S. Oubbati, A. Barnawi, M. Atiquzzaman, and D. Alghazzawi, "UAV assistance paradigm: State-of-the-art in applications and challenges," *J. Netw. Comput. Appl.*, vol. 166, Sep. 2020, Art. no. 102706.

[7] M. Basharat, M. Naeem, Z. Qadir, and A. Anpalagan, "Resource optimization in UAV-assisted wireless networks: A comprehensive survey," *Trans. Emerg. Telecommun. Technol.*, vol. 33, p. e4464, Feb. 2022.

[8] B. Jedari, G. Premsankar, G. Illahi, M. D. Francesco, A. Mehrabi, and A. Yla-Jaaski, "Video caching, analytics, and delivery at the wireless edge: A survey and future directions," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 431–471, 1st Quart., 2021.

[9] J. Yao, T. Han, and N. Ansari, "On mobile edge caching," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2525–2553, 3rd Quart., 2019.

[10] H. S. Goian, O. Y. Al-Jarrah, S. Muhaidat, Y. Al-Hammadi, P. Yoo, and M. Dianati, "Popularity-based video caching techniques for cache-enabled networks: A survey," *IEEE Access*, vol. 7, pp. 27699–27719, 2019.

[11] 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, Feb. 2019.

[12] T. X. Tran and D. Pompili, "Adaptive bitrate video caching and processing in mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 18, no. 9, pp. 1965–1978, Sep. 2019.

[13] J. Zhang, H. Wu, X. Tao, and X. Zhang, "Adaptive bitrate video streaming in non-orthogonal multiple access networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3980–3993, Apr. 2020.

[14] Y. Sani, A. Mauthe, and C. Edwards, "Adaptive bitrate selection: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2985–3014, 4th Quart., 2017.

[15] S. Anokye, D. Ayepah-Mensah, A. M. Seid, G. O. Boateng, and G. Sun, "Deep reinforcement learning-based mobility-aware UAV content caching and placement in mobile edge networks," *IEEE Syst. J.*, vol. 16, no. 1, pp. 275–286, Mar. 2022.

[16] L. Li, M. Wang, K. Xue, Q. Cheng, D. Wang, W. Chen, M. Pan, and Z. Han, "Delay optimization in multi-UAV edge caching networks: A robust mean field game," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 808–819, Jan. 2021.

[17] F. Zhou, N. Wang, G. Luo, L. Fan, and W. Chen, "Edge caching in multi-UAV-enabled radio access networks: 3D modeling and spectral efficiency optimization," *IEEE Trans. Signal Inf. Process. over Netw.*, vol. 6, pp. 329–341, 2020.

[18] 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, Apr. 2022.

[19] J. Ji, K. Zhu, D. Niyato, and R. Wang, "Joint cache placement, flight trajectory, and transmission power optimization for multi-UAV assisted wireless networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5389–5403, Aug. 2020.

[20] J. Ji, K. Zhu, D. Niyato, and R. Wang, "Joint trajectory design and resource allocation for secure transmission in cache-enabled UAV-relaying networks with D2D communications," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1557–1571, Feb. 2021.

[21] A. Bera, S. Misra, and C. Chatterjee, "QoE analysis in cache-enabled multi-UAV networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6680–6687, Jun. 2020.

[22] W. Zhang, Y. Wen, Z. Chen, and A. Khisti, "QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1431–1445, Oct. 2013.

[23] S. Bayhan, S. Maghsudi, and A. Zubow, "EdgeDASH: Exploiting network-assisted adaptive video streaming for edge caching," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1732–1745, Jun. 2021.

[24] C. Li, L. Toni, J. Zou, H. Xiong, and P. Frossard, "QoE-driven mobile edge caching placement for adaptive video streaming," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 965–984, Apr. 2018.

[25] X. Xu, J. Liu, and X. Tao, "Mobile edge computing enhanced adaptive bitrate video delivery with joint cache and radio resource allocation," *IEEE Access*, vol. 5, pp. 16406–16415, 2017.

[26] Z. Qu, B. Ye, B. Tang, S. Guo, S. Lu, and W. Zhuang, "Cooperative caching for multiple bitrate videos in small cell edges," *IEEE Trans. Mobile Comput.*, vol. 19, no. 2, pp. 288–299, Feb. 2020.

[27] W. Shi, C. Wang, Y. Jiang, Q. Li, G. Shen, and G.-M. Muntean, "CoLEAP: Cooperative learning-based edge scheme with caching and prefetching for DASH video delivery," *IEEE Trans. Multimedia*, vol. 23, pp. 3631–3645, 2021.

[28] C. Liu, H. Zhang, H. Ji, and X. Li, "MEC-assisted flexible transcoding strategy for adaptive bitrate video streaming in small cell networks," *China Commun.*, vol. 18, no. 2, pp. 200–214, Feb. 2021.

[29] H. Zhao, Q. Zheng, W. Zhang, B. Du, and H. Li, "A segment-based storage and transcoding trade-off strategy for multi-version VoD systems in the cloud," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 149–159, Jan. 2017.

[30] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Video transcoding, caching, and multicast for heterogeneous networks over wireless network virtualization," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 141–144, Jan. 2018.

[31] A.-T. Tran, N.-N. Dao, and S. Cho, "Bitrate adaptation for video streaming services in edge caching systems," *IEEE Access*, vol. 8, pp. 135844–135852, 2020.

[32] C. Liang, Y. He, F. R. Yu, and N. Zhao, "Enhancing video rate adaptation with mobile edge computing and caching in software-defined mobile networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 7013–7026, Oct. 2018.

[33] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[34] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.

[35] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[36] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Jun. 2016.

[37] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, New York, NY, USA, Mar. 1999, pp. 126–134.

[38] P. Zhao, H. Tian, C. Qin, and G. Nie, "Energy-saving offloading by jointly allocating radio and computational resources for mobile edge computing," *IEEE Access*, vol. 5, pp. 11255–11268, 2017.

[39] M. Garey and D. Johnson, *Computers Intracdtability: A Guide to Theory NP-Completeness*. San, Francisco, CA, USA: Freeman, Jan. 1979.

[40] K.-H. Han and J.-H. Kim, "Quantum-inspired evolutionary algorithm for a class of combinatorial optimization," *IEEE Trans. Evol. Comput.*, vol. 6, no. 6, pp. 580–593, Dec. 2002.

JUNFENG XIE received the B.S. degree in communication engineering from the University of Science and Technology Beijing, in 2013, and the Ph.D. degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, in 2019. From September 2017 to September 2018, he visited Carleton University, Ottawa, ON, Canada, as a Visiting Ph.D. student. He is currently an Assistant Professor with the North University of China. His research interests include machine learning, content delivery networks, resource management, and wireless networks.

ZHAOBA WANG received the Ph.D. degree in instruments science and technology from the Nanjing University of Science and Technology, in 2002. He is currently a Professor with the School of Information and Communication Engineering, North University of China. His research interests include information processing and reconstruction, resource management, and wireless networks.

YOUXING CHEN received the Ph.D. degree from the North University of China, in 2010. He is currently a Professor with the School of Information and Communication Engineering, North University of China. His research interests include machine learning, signal processing, and resource management.