Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(1): 1573–1589.
DOI: 10.3934/mbe.2024068
Received: 31 October 2023
Revised: 11 December 2023
Accepted: 12 December 2023
Published: 02 January 2024

Mini review

QoS-driven resource allocation in fog radio access network: A VR service perspective

Wenjing Lv[1], Jue Chen[1,*], Songlin Cheng[2], Xihe Qiu[1] and Dongmei Li[1]

[1] College of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

[2] School of Electronic and Information Engineering, Shanghai Dianji University, Shanghai 201306, China

* **Correspondence:** Email: jadeschen@sues.edu.cn.

Abstract: While immersive media services represented by virtual reality (VR) are booming, They are facing fundamental challenges, i.e., soaring multimedia applications, large operation costs and scarce spectrum resources. It is difficult to simultaneously address these service challenges in a conventional radio access network (RAN) system. These problems motivated us to explore a quality-of-service (QoS)-driven resource allocation framework from VR service perspective based on the fog radio access network (F-RAN) architecture. We elaborated details of deployment on the caching allocation, dynamic base station (BS) clustering, statistical beamforming and cost strategy under the QoS constraints in the F-RAN architecture. The key solutions aimed to break through the bottleneck of the network design and to deep integrate the network-computing resources from different perspectives of cloud, network, edge, terminal and use of collaboration and integration. Accordingly, we provided a tailored algorithm to solve the corresponding formulation problem. This is the first design of VR services based on caching and statistical beamforming under the F-RAN. A case study provided to demonstrate the advantage of our proposed framework compared with existing schemes. Finally, we concluded the article and discussed possible open research problems.

Keywords: virtual reality (VR); fog radio access network architecture (F-RAN); edge caching; fog computing; statistical beamforming

1. Introduction

With the rapid development of wireless communication, the widespread application of artificial intelligence (AI) technology and the explosive growth of edge intelligence, the consumer electronics

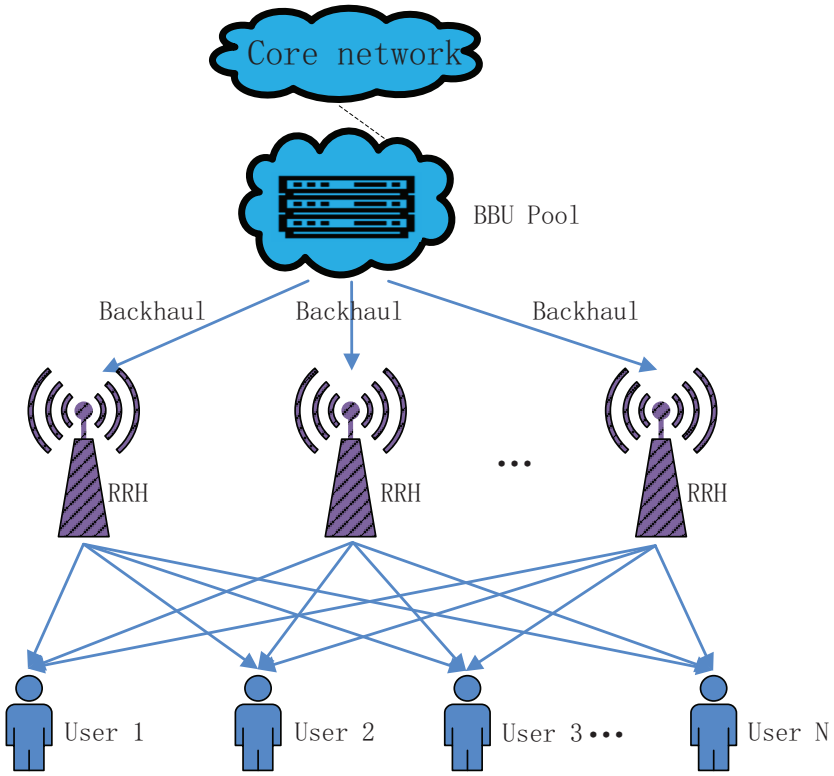

Figure 1. The C-RAN architecture.

services are getting more and more popular. New consumer services represented by the immersive extended reality (XR), metaverse, man-machine interaction and holographic communication put forward stricter requirements on performance and function [1]. These services need not only the fundamental connection assurance, but also the high-precision environment, network AI service and edge computing resource. Meanwhile, they are facing fundamental challenges. Spectrum resources are very scarce, especially for the low-frequency spectrum. The costs of operation are increasing, which is mainly due to the high energy consumption of large consumer electronics products. Consumers' electronic applications are proliferating to meet billions of personalized user demands in the internet of everything (IoE) [2]. In short, the mainstream communication mode such as the metaverse and immersive XR can bring users a multidimensional and immersive experience. Further, better services raise stricter requirements on the network bandwidth, latency, reliability, quality and efficiency. The above-mentioned characteristics have promoted the industry and the academia to research the novel architecture and technology for supporting diverse application scenarios [3]. As we know, the requirements for network characteristics such as latency, rate, number of connections and energy consumption are usually different in different scenarios [4]. Several requirements are even contradictory [5]. For example, when you drive automatically, you care about the delay; if the delay exceeds 10 millisecond, it will seriously affect safety [6]. While you watch a live broadcast of a high-definition concert, you focus on the timeliness and picture quality and usually ignore even if the whole concert is delayed by a few seconds or even a dozen seconds [7]. Oriented toward these requirements, the future network should be more service-based and flexible [8]. Especially for different customized network requirements, it is a prerequisite to build a more open ecosystem and introduce more participants [9].

The network architecture, especially the radio access network architecture (RAN), is one of the very important issues for the realization of the future communication vision [10]. In fact, the network architecture started with the one generation (G), then went through the 2G, 3G, 4G and continued to evolve to the 5G and the advanced 6G. The rapid evolution process of the network is essentially a process of the continuous separation for base station functions, which aims to meet the diversified needs of different scenarios [11]. The industry has successively proposed several RAN solutions, such as the cloud radio access network (C-RAN) and fog radio access network (F-RAN).

Specifically, the literature [12] first proposed the C-RAN architecture, which involves the cloud computing and cloud storage technology as shown in Figure 1. Different from the traditional distributed base stations, the base station function is virtualized in the C-RAN environment. Specifically speaking, the base station of the C-RAN is decomposed into two parts: One part is the remote radio head (RRH) and the other is the baseband unit (BBU). Multiple BBUs constitute a centralized cloud server. The cloud server is mainly used for high-capacity transmission and to support seamless coverage scenarios. Each RRH performs signal processing via a virtual baseband base station, which is supported by the processor. The BBUs and RRHs are connected through high-capacity backhaul links. So far, the C-RAN has been applied to the multiple mobile communication operators and the communication equipment providers [13]. It has also been applied to the experiment and update for the existing mobile communication networks. The C-RAN has the advantages of reducing the consumption cost and improving the resource utilization and flexibility of network deployment [14]. However, it also has some potential inherent flaws. On the one hand, this is mainly due to the RRH node, which can only perform simple processing for the signal such as radio-frequency signal amplification, frequency conversion and digital-to-analog conversion. On the other hand, both the uplink and downlink data are transmitted over a limited backhaul link; as a result, the burden of the backhaul link will become larger and larger when the number of users increases. Finally, the bottleneck effect is bound to become more prominent, which further reduces the spectrum and energy efficiency of the entire C-RAN network. It also increases the data transmission delay from the cloud server to the client.

Further, the F-RAN architecture was put forward based on the C-RAN architecture. In fact, the fog concept was first introduced in F-RAN [15], which is similar to the phenomenon of the "fog" nearer to the ground "cloud" in the physical world. It can fully utilize different resources of the edge facilities. In the F-RAN, the edge nodes, which have functions of storage and cooperative radio resource management (CRRM), are collectively evolved to fog access points (F-APs). The F-APs can solve the inherent defects of the C-RAN mainly due to the reuse characteristic of the content requested by the existing network users. Based on this characteristic, the F-APs nodes can selectively cache part of the data content with high popularity in advance by reasonably avoiding network busy hours [16]. As a result, the popular content can be sent to multiple request terminals simultaneously based on the same resource block. Moreover, combined with the multiplexing characteristics of content, multiple transmission modes can be deployed on the F-RAN network architecture for different practical scenarios. The F-RAN can provide a new resource dimension for improving the transmission efficiency. Compared with the traditional C-RAN architecture, the F-RAN can make full use of the remote radio frequency unit, fog access point and edge devices, which can collaborate on signal processing, radio resource management, caching, and computing [17].

Consider virtual reality (VR) application in the F-RAN architecture. As we know, the emerging VR video can provide an immersive experience by interactions of body movements [18]. This immersive

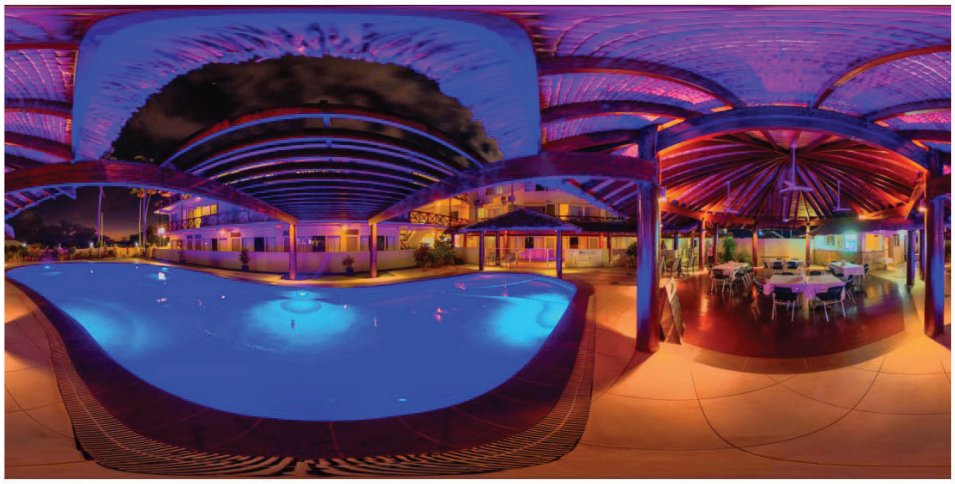

Figure 2. The virtual reality service scenario.

experience is equivalent to a subjective feeling in a comprehensive and real experience, as shown in Figure 2. In order to ensure the high quality experience for VR users, it is inevitable to adopt 360° video streaming in VR applications [19]. The 360° VR videos belong to immersive spherical videos. It was usually mapped into a 3D geometry in which the users can look around during playback by wearing a VR head-mounted display device [20]. However, the 360° VR videos need to take up a large amount of bandwidth resources [21]. The large amount of data from 360° video streaming makes the content storage and transmission more difficult. Therefore, how to efficiently transmit these large 360° VR videos in the limited bandwidth at acceptable quality levels is challenging for the current RAN architecture [22].

Naturally, the F-RAN architecture is discovered as a promising RAN architecture for VR video delivery. Due to only parts of the 360° VR video needed to be viewed in the application, the requested three-dimensional stereoscopic video (SV) and two-dimensional monocular video (MV), but not the whole stereoscopic panoramic video, should be transmitted in the downlink F-RAN [23]. The F-APs in the F-RAN, which are with caching capacity, can just be utilized to cache MVs and SVs of some viewpoints. That is to say, the F-APs, which are located between the cloud layer and the terminal layer, only need to deliver the requested SVs and MVs, which can further help improve efficiency by avoiding additional transmission and computation. Based on this idea, it is predictable that the VR video can be deployed on the novel F-RAN architecture. However, the caching capacity in the F-RAN is generally limited. It is difficult for the user to get all the requests. In addition, to promote the total resource utilization, the transmission policy also needs to be designed carefully in such a network architecture. Thus, a joint optimization design of the fog layer caching and wireless link beamforming is hopefully proposed to effectively reduce the total network cost as well as ensure the user experience.

The main work of our research is summarized as follows.

- First, we made an overview on the existing VR and RAN architecture research. we then proposed a mobile VR video delivery framework based on F-RAN architecture as shown in Figure 3. This framework introduced the caching and beamforming strategies for the VR video delivery under the F-RAN environment. In fact, this is an innovation and creative design on VR video services

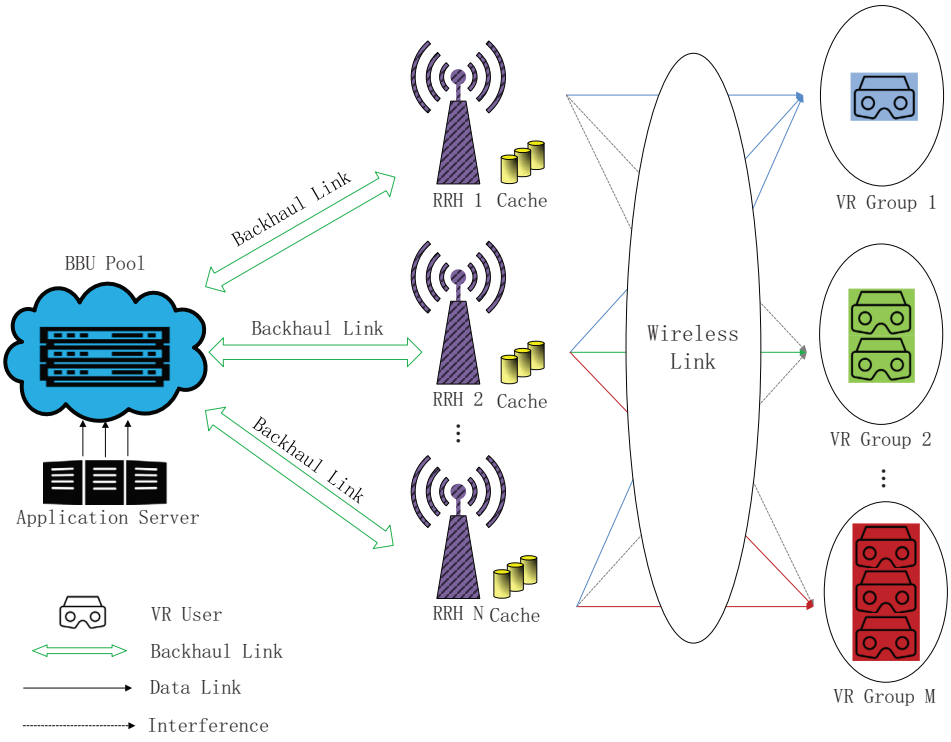

Figure 3. The VR delivery based on F-RAN architecture.

based on caching and beamforming.

- Next, we focused on a cost model aimed at trading off the fronthaul cost and the transmission power cost at the BSs. According to the cost model, an multi-objective optimization problem was derived to get an elaborate caching allocation and beamforming design. It is worth it to emphasize that the optimizations of the caching and the beamforming strategy occur in two timescales. Cache allocation is usually deployed in a long timescale, while the beamforming can dynamically adapt to the instantaneous channel realization in a short timescale. We present an innovative framework based on statistical channel state information (SCSI) to solve the different two-timescale problem based in F-RAN.
- Finally, a tailored algorithm was proposed by utilizing the branch-bound (BB) method and other relaxation techniques. Simulation demonstrates that our algorithm can obtain better performance. We test the framework for VR video delivery experiments in a given F-RAN environment. We also compare our proposal with the single solution.

To sum up, oriented toward various consumers' requirements, the novel network architecture and advanced technologies need to be further discussed. We also discuss several research challenges and opportunities, hoping that our work can give more inspiration on the upcoming design of caching and delivery scheme, aiming to support various F-RAN-based service scenarios, especially in the direction of VR video technology.

In section two, we introduce the F-RAN-based VR delivery scheme. In section three, we elaborate the corresponding algorithm and implementation procedure for the two-stage caching and beamforming design based on SCSI. In section four, a case study is provided to demonstrate the improvement.

At the end, we conclude the article, and we put forward several issues for the future research in section five.

2. F-RAN-based VR delivery framework

2.1. The F-RAN architecture

In the C-RAN, the storage layer, the control layer, as well as the communication processing functions are all integrated into a cloud computing network layer. Limited fronthaul link capacity and the increased demand on signal processing have become a bottleneck of the network performance gain. In addition, the C-RAN cannot achieve a smooth transition and compatibility in the existing mobile communication networks [24].

Later, the F-RAN architecture was successively proposed to overcome some drawbacks of the C-RAN. The basic idea of the F-RAN is to take the advantage of the F-APs, which can achieve the local distributed content service, the distributed signal processing and the distributed resource management. Since the F-APs are closer to the user terminal, the services of low-latency are becoming convenient. Specifically, the upgraded F-APs can be fully utilized to store part of the content, thereby avoiding the high real-time or the large-scale requirements of wireless signal processing in the BBU pool, effectively alleviating the pressure of large data transmission in the fronthaul link.

The F-RAN architecture has the features of ultra-low latency, proximity and network status awareness. In addition, some advanced technologies, such as millimeter wave communication and massive multiple input multiple output (MIMO) can also be well deployed in the F-RAN [25].

2.2. Characteristic of VR video transmission

The VR service is booming and has been rapidly commercialized in recent years, forming an extremely large potential market. The popularity of VR video can be attributed to two distinctive characteristics: Innovative interaction mode and immersive experience [26]. In the actual application scenario, the VR presents panoramic videos which simulate a virtual environment around users by a spherical canvas. The data of the orientations and positions can be recorded quickly by the sensors in the specific equipment when the user performs a moving action. Consider that the requirement for the seamless VR video experiences is less than 10 ms. It is necessary to deploy high resolution and high frame rate in VR video, which will affect the quality of viewing experience. Meanwhile it is difficult to meet the deployment requirements of the increasing data size and the efficiency of transmission at the same time. Thus, the VR video services, which require vltra-high transmission rate and looth vltra-low latency are becoming an insurmountable challenge.

The existing research mainly focuses on either the single caching-level design or single VR video's delivery-level design. A new transmission scheme, which is proposed in the literature [27], can reduce video transmission delay. In order to minimize consumption, on the one hand, a communications-constrained mobile edge computing (MEC) framework is proposed for VR video; on the other hand, a task scheduling framework is proposed aiming to analyze the tradeoffs between the caching and the computing in [28]. However, some resource allocation strategies are based on a centralized mode assumption. Heavy traffic also requires larger cache capacity, which will bring further burden.

The future VR video experience paradigm requires the consistent space-time of key parameters be-

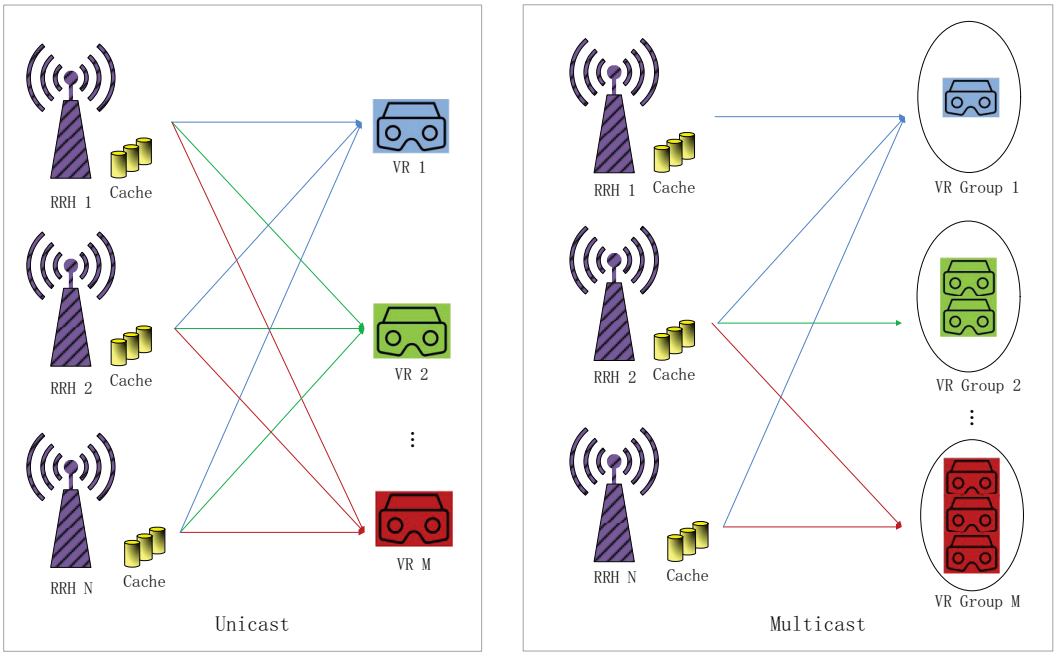

Figure 4. Different transmission mode on VR scenario.

tween physical space and virtual space, which puts forward ultra-high-precision timing and positioning requirements for current wireless access and transmission. At the same time, the real-time perception of the large dimensions of the physical space has generated massive heterogeneous data. These need higher transmission bandwidth and large-capacity networks to support the deterministic transmission of multidimensional information. However, it is not easy to fulfill these requirements simultaneously in a traditional RAN system. As known, the fog computing as well as the fog caching are both important on reducing the burden. Specifically, a careful design of the caching policy, i.e., how to allocate capacity and how to deploy appropriate caching strategies is needed in F-RAN architecture.

Considering that only parts of 360° VR videos need to be viewed for the terminals, rather than the whole stereoscopic panoramic video, only the requested MV and SV should be transmitted in the downlink. In the downlink F-RAN, the F-APs can be utilized to cache MVs and SVs of some viewpoints [29]. That is, the F-APs, located in the middle layer between the cloud and the terminal end, only need to deliver the requested MVs and SVs, helping to reduce both latency and energy consumption by avoiding transmission and computation.

Based on the above idea, it is anticipated that the emerging VR video service can be effectively deployed in the F-RAN. The VR users can be satisfied in the F-RAN at a less powerful consumption but a faster speed by deploying more functions at the edge F-APs. Therefore, the performance is expected to be further improved through designing joint caching and delivery strategies in advanced F-RAN architecture.

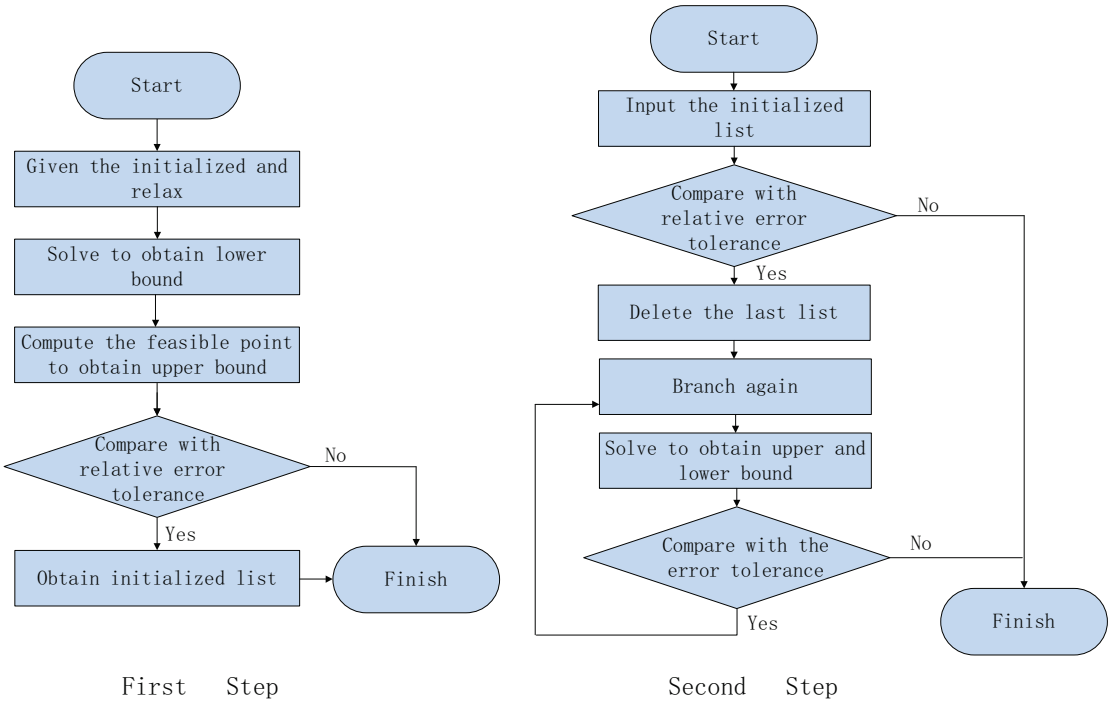

Figure 5. The flowchart of the designed algorithm.

2.3. Proposed solutions

Based on the theoretical analysis and transmission characteristic of VR video, we explore a tailored quality of service (QoS)-driven resource allocation solution in the F-RAN architecture. The detailed process is described below.

Framework: This is an innovation and creative design on VR video services based on caching and beamforming in the F-RAN. The initial idea is to utilize the functions of F-APs and advanced beamforming technology to solve the VR video delivery problem. Consider a downlink transmission scenario with BSs and VR user group. The users are finally classified in a cluster if they have the same requirement, and will be served by the multicast mode. This scenario also covers a special case when each group has only one user, which is equivalent to a unicast transmission mode in this case. Different transmission modes are illustrated in Figure 4. The F-APs of the F-RAN architecture can be used to cache SVs and MVs of the VR video. When the users have a request, the requested MVs and SVs of VR videos can be delivered by the F-APs, which will help reduce the consumption and resources.

Two-timescale deployment problem: The design is mainly divided into two parts; One part is the design for the caching and the other is the design for the transmission level. It is worth it to emphasize that the tradeoff between the cache size allocation and the beamforming strategy is usually happening in two different timescales. To be specific, the beamforming is adapted to the instantaneous channel information dynamically, while the cache size is optimized in a long-term statistical channel. Based on the SCSI, we aim to jointly optimize the two parts to better serve the network transmission and user experience. We then presented a compromise scheme based on SCSI to solve multiple-timescale deployment challenge in F-RAN. In addition, we have taken into account the difference of the file popularities to facilitate cache larger portions of more popular VR files.

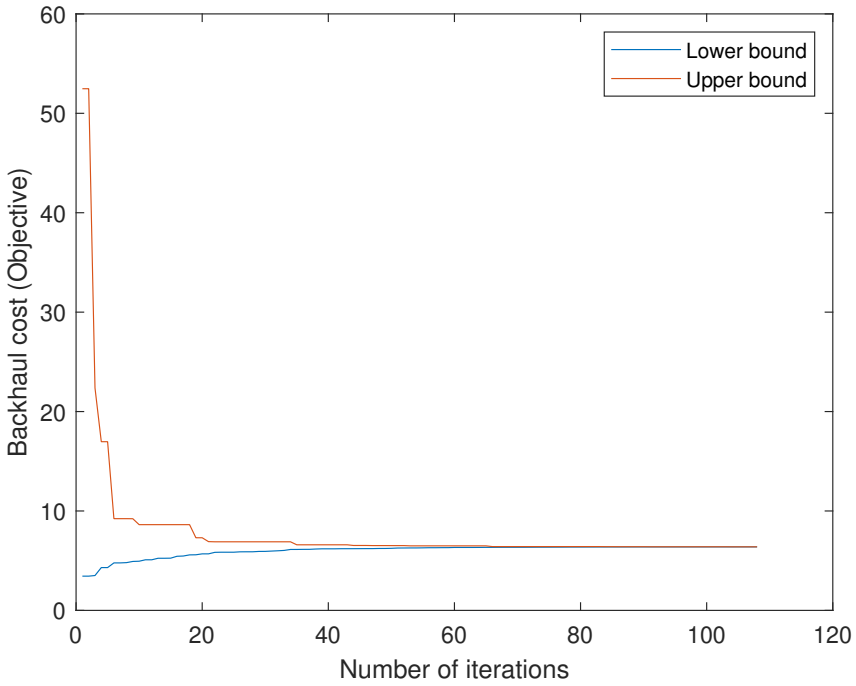

Figure 6. Convergence behavior of the proposed BB-based algorithm.

Beamforming based on SCSI: As we know, the beamforming can match the instantaneous channel real-time in a short time scale [30], but the cache allocation time slot is in a long-term in our framework. This means that the conventional instantaneous channel information is not suitable in the proposed architecture due to our joint design. Thus, we present a novel strategy based on the SCSI to solve the above multiple-timescale deployment problem in the above considered F-RAN architecture. We finally formulate a two-timescale beamforming and joint cache size allocation design problem, in which the cache and the beamforming is optimized based on the SCSI. Our final goal is to substantially reduce the total network cost for supporting better VR users' experiences. By leveraging the statistical channel information, we also develop a novel algorithm for optimizing the cache allocation and beamforming design, as well as quantify how much cache size should be allocated to each BS.

3. Analysis and algorithm

The cost consumption, which is based on a joint cache size allocation and statistical beamforming design, can be formulated as an optimization problem. However, the problem is not easy to solve. On the one hand, the objective is nonsmooth and nonconvex, especially the signal to interference plus noise ratio (SINR) constraints are also nonconvex [31]. On the other hand, it is necessary to emphasize that we also consider the base station (BS) cluster, which can be dynamically optimized.

Our tailored algorithm is based on a novel argument cut technique, which can provide effective relaxations for the nonconvex SINR constraints. As a result, the original nonconvex set can be reasonably translated into the convex envelope. In other words, with the help of argument cut theory [32], we can develop efficient convex relaxations, particularly for SINR constraints [33]. After argument cut

processing, the problem can be solved globally by using the interior-point method [34], which is a mature and complete solution for convex optimization. A specific flowchart of the algorithm is provided in Figure 5. We can demonstrate that the argument cuts strategy is highly effective in this case.

Next, we describe the main steps of the proposed specific BB algorithm as follows, which will be applied in our case study in the next section.

Initialization: We first initialize all intervals, then the problem can be reduced into a simpler solvable convex problem. We obtain an initial value of the algorithm, which is the optimal caching allocation and statistical beamforming accordingly.

Termination: We preselect a relative error tolerance as a termination algorithm threshold. We terminate the algorithm if the criterion is less than the threshold. Otherwise, we branch it into some intervals by the preset rule. The lower bound as well as the upper bound are the keys in avoiding unnecessary branches and further enumerations. According to theoretical analysis, an appropriate lower bound and upper bound can improve the efficiency of the algorithm significantly, and with that, we can branch one by one.

Branch: If the problem has the least lower bound, and it does not meet the stopping cycle condition, then the interval will be selected which leads the largest gap to be branched to smaller sub-intervals. Subsequently, we partition the winning interval into two sets, which have equal size and all the others are unchanged. Accordingly, we obtain two subproblems, which are tighter than the last. Once the interval has been broken down into two parts, the corresponding original problem will be cut from the original list, meanwhile two subproblems will be added into the library again when their optimal objective values are equal to or less than the current bound.

The Lower Bound: The smallest bound of all lower bounds is a lower bound of the original problem.

The Upper Bound: We obtain an upper bound by appropriately scaling the obtained solution within the feasible region. An upper bound is the best values from all of the known feasible solutions in each iteration.

Based on the above steps, we have shown the main procedures of the proposed scheme. We emphasize again that both the upper bound and the lower bound are key in improving the efficiency of the proposed BB algorithm, because the best upper and lower bound help detect inactive children problems, which further avoid unnecessary enumerations and branches. A specific convergence process is demonstrated in Figure 6.

4. Case study

We provide a case study to demonstrate our proposed framework in the downlink. In this case, we only consider the multicast transmission mode. The VR users requesting the same file are formed into a VR group served by the multicast transmission mode. Accordingly, a BB-based optimal algorithm for multicast transmission is shown in this section.

Suppose that each transmission time interval has enough number of frames to complete the content delivery. We consider the transmission of VR video in F-RAN architecture with K single-antenna mobile VR users and N multiple-antenna BSs. The total number of group is M. An independent group obtains the service with the help of a group of cooperative BSs in one frame. The limited-capacity fronthaul link acts as a bridge, which connects the BBU pool and the BS as shown in Figure 3. The BBU is connected to a database that contains Z contents in which each content has equal size. Users

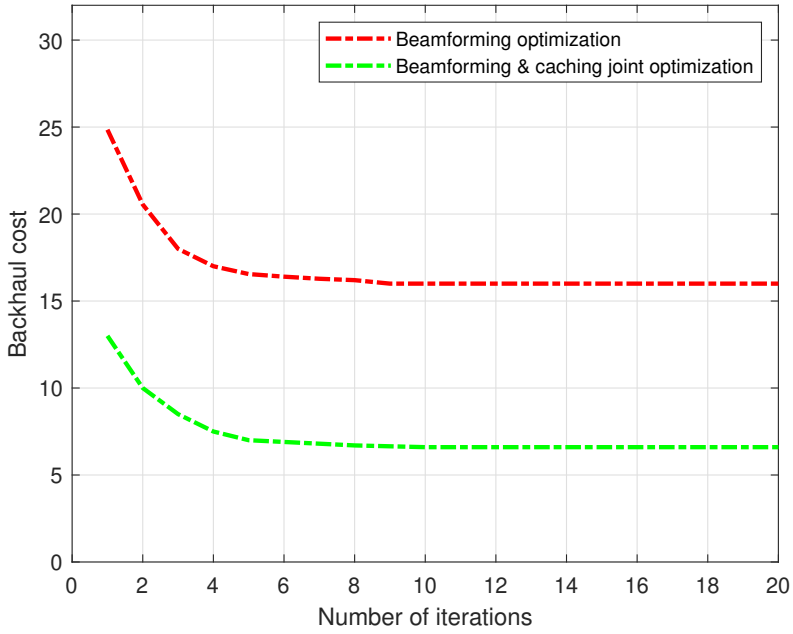

Figure 7. Backhaul network cost.

can request one content in each time slot according to certain demand probabilities. For each BS, if the requested VR content has been cached, it can access the VR content directly without backhaul cost. Otherwise, it needs to request this content from the central processing (CP) through the backhaul link.

For convenience to the description, we have denoted as follows:

1) Denote $\gamma = [\gamma_1, \gamma_2, ..., \gamma_M]$ as the SINR vector, where γ_m is the minimum required by user m.

2) Set the transmission rate of user m as $D_m = b \cdot \log(1 + \gamma_m)$, where b is the channel bandwidth.

3)The k-th user's SINR can be expressed as follows:

$$SINR_k^M = \frac{\left|\mathbf{h}_k^H \mathbf{w}_m\right|^2}{\sum\limits_{i=1, i\neq m}^{M} \left|\mathbf{h}_k^H \mathbf{w}_i\right|^2 + \sum\limits_{j=1}^{K} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2}, \forall k \in \Phi_m.$$

We make an average operation on the channel state information, then we can get the following expression:

$$E\left(\mathrm{SINR}_k^M\right) \geq \gamma_m, \ \forall k \in \Phi_m.$$

Note $\mathbb{E}\left(\mathbf{h}_k \mathbf{h}_k^H\right) = \Sigma_k$, and it can be written approximately as follows:

$$\frac{\mathbf{w}_m^H \Sigma_k \mathbf{w}_m}{\sum_{i=1, i\neq m}^{M} \mathbf{w}_i^H \Sigma_k \mathbf{w}_i + \sum_{j=1}^{K} \mathbf{v}_j^H \Sigma_k \mathbf{v}_j + \sigma_k^2} \geq \gamma_m.$$

There are two parts for the system cost: The backhaul cost and the transmission power consumption. The backhaul cost is generally proportional to its transmission capacity and related to the cache, user's request and matching service between RRHs and the user. It is worth remarking that if the VR file

has been cached in advance in the BS, it can access the VR file directly and will not produce the backhaul link cost. Otherwise, it will produce extra backhaul cost. The backhaul cost can be derived as a function about the transmission rate, which is expressed as:

$$C_B = \sum_{m=1}^{M} \sum_{n=1}^{N} s_{m,n}(1 - C_{z,m,n})R_m.$$

The total transmission power cost is expressed below:

$$C_p = \sum_{m=1}^{M} \sum_{n=1}^{N} \left\|\mathbf{w}_{m,n}\right\|_2^2.$$

$\mathbf{w}_{m,n}$ is the beamforming vector for user m from BS n. The $C_{z,m,n} = 1$ indicates that the z-th video content requested by the m-th VR user is cached in the n-th BS and zero otherwise. The original problem can be finally formulated as:

$$\begin{aligned} &\min_{\{\mathbf{w},s,C\}} \quad C_B \\ &s.t. \quad C_p \le P_f \\ &\qquad E(SINR_k) \ge \gamma_m \\ &\qquad \sum_{(z,m)} \hat{C}_{(z,m),n} \le Q_n \ . \end{aligned}$$

It is worth emphasizing that the multi-group multicast solution procedure is essentially the same to single-group multicast. For the convenience of elaboration, we provide a single-group multicast expression and solution procedure. Specifically, the original problem can be finally formulated as:

$$P^M : \min_{\{\mathbf{w}_{0,n}\},\{s_{0,n}\}} \sum_{z=1}^{Z} \frac{z^{-\alpha}}{\sum_{j=1}^{Z} j^{-\alpha}} \sum_{n=1}^{N} s_{0,n}(1 - C_{z,0,n})R_0 \tag{4.1a}$$

$$s.t. \sum_{n=}^{N} \left\|\mathbf{w}_{0,n}\right\|_2^2 \le P_f \tag{4.1b}$$

$$\mathbf{w}_0^H E(\tilde{\mathbf{h}}_k \tilde{\mathbf{h}}_k^H)\mathbf{w}_0 \ge 1, \quad \forall\ k \in \Phi_0 \tag{4.1c}$$

$$\sum_{(z,0)} C_{(z,0),n} \le Q_n, \forall n \in \mathcal{N}. \tag{4.1d}$$

- We aim to jointly optimize the edge caching $C_{z,m,n}$, the statistical beamforming $\mathbf{w}_{m,n}$ and the BS clustering $\mathbf{s}_{m,n}$ to minimize the backhaul cost objective, under the transmission power cost constraints (4.1b), the QoS (4.1c) and the cache size (4.1d). We also consider the difference in file popularities in caching files.

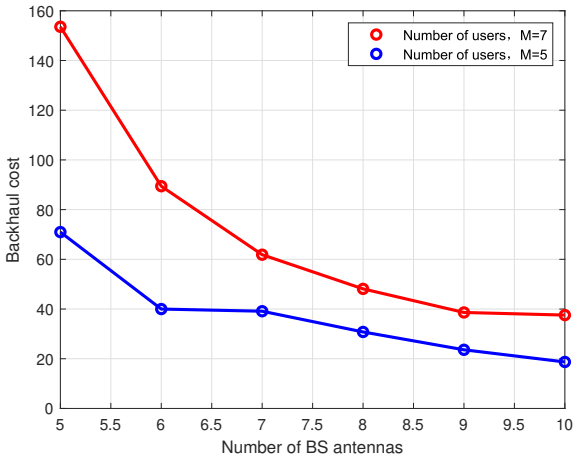

Figure 8. Backhaul network cost.

- Based on the SCSI, the 0,1 variable can be replaced by a continuous function. By adopting smooth l_0-norm approximation, the original problem can be written as an equivalent problem solved with the BB method for multicast transmission. Compared with the local optimal solution in the previous literature [35], we obtain the global optimal solution by using our tailored algorithms.

We provide a global optimal solution corresponding to our framework.

Figure 6 shows convergence behavior. It shows that the gap between the upper bound and the lower bound is reduced rapidly during the iterations. The explanation for this phenomenon is that many infeasible subregions are effectively removed. Finally, the upper bound is nonincreasing, while the lower bound is nondecreasing. The gap becomes increasingly smaller when the iteration number increases.

Figure 7 illustrates the performance gain. A large performance gap between the single beamforming strategy and the joint optimization is observed, further demonstrating effectiveness.

Figure 8 demonstrates the performance with setups $(M, N) = (7, 3)$, $(M, N) = (5, 3)$ and $L \in \{5, 6, 7, 8, 9, 10\}$. This provided the trend of the performance for different numbers of VR users. We can observe from Figure 8 that the cost consumption decreases as the antennas number increases, also illustrating the reality and feasibility of our proposed scheme.

5. Conclusions and future topics

In this paper, we first reviewed some relevant topics and then focused our attention on the challenges of VR video delivery. First, an F-RAN-based framework was proposed for VR service in which the F-APs can be utilized to cache some requests. Second, we provided a design describing how the requests can be delivered more effectively based on SCSI. A joint statistical beamforming and edge caching problem was formulated and solved under the QoS constraints in the downlink F-RAN. In addition, we gave a case study, which further confirmed the advantages of our framework. At the end, the numerical results demonstrated the considerable performance of the optimized scheme.

Based on the current research and combined with our research thinking, for one thing, network

architecture is the basis for the realization of the immersive media services and is also a key issue for 6G research. For another thing, the immersive scenarios represented by VR require not only the connection and high-precision service, but also the caching technology and network AI services. In addition, the immersive multimedia services have stricter requirements for network bandwidth, latency and reliability, especially resource allocation. Finally, it is worth emphasizing that the design of the algorithm is also crucial in the process of solving the optimization problem.

Correspondingly, we provide several research challenges and topics as follows, and hope that our work can provide some inspiration in the the multimedia field.

Recommendation 1: Network architecture

The basic idea on the 6G evolution is to improve capabilities and user experience by focusing on requirements. Thus, the new network architecture is based on the continuation of service-oriented design, and it aims to achieve a simplified network architecture by these approaches: Architecture, procedure and interface simplification. From this point of view, our proposed framework for VR delivery is worth further research, especially for customized network requirements in different XR application scenarios. How to optimize and make up for the shortcomings of the existing network architecture is still an interesting study.

Recommendation 2: Caching technology

The immersive scenarios represented by VR require not only the connection service, but also high-precision environment, caching resource and network AI services.

In our research, an F-RAN-based framework was proposed for VR service in which the F-APs with caching capacity can be utilized to cache some requests. We believe that making full use of caching capabilities to improve the users' experience will be an inevitable trend. Our framework provides a direction for the further development of the caching strategy. Caching issue is an advanced research point at present. Thus, it's a worthy study focusing on the the content-centric caching mechanisms and cooperation caching mechanisms.

Recommendation 3: Resource allocation

The flexible 6G requires that the networks should support on-demand deployment and traffic scheduling, as well as service loading. Especially for different customized requirements, agility and flexibility for resource scheduling are not only important but also prerequisites for building a more open ecosystem.

Resource allocation is a permanent topic in the communication system. A QoS-driven resource allocation in F-RAN architecture is a study in our research. We explore a tradeoff between the BS cache size allocation and the beamforming strategy allocation. It is meaningful to further explore how to trade off the cost consumption in an effective manner in the entire network, particularly in the VR video delivery problem.

Recommendation 4: Design of global algorithm

The 6G will introduce a more flexible resource cooperation scheme to satisfy users' various requirements. The traffic adaptation between networks and services can be implemented, which aims to obtain the best service experience of users. By introducing the idea of optimization theory, we innovatively proposed two global optimization algorithms to obtain the optimal resource allocation scheme. In fact, how to choose a reasonable optimization method is still critical in solving different resource allocation models. Thus, the design of the global algorithm is worth exploring, especially in new services represented by immersive XR and Metaverse.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62102241) and Shanghai Committee of Science and Technology "Science and Technology Innovation Action Plan" Natural Science Foundation of Shanghai (23ZR1425400).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Z. Yang, M. Chen, K. K. Wong, H. V. Poor, S. Cui, Federated learning for 6g: Applications, challenges, and opportunities, *Engineering*, **8** (2022), 33–41. https://doi.org/10.1016/j.eng.2021.12.0022095-8099
2. W. Qi, H. Su, A cybertwin based multimodal network for ecg patterns monitoring using deep learning, *IEEE Trans. Ind. Inf.*, **18** (2022), 6663–6670. https://doi.org/10.1109/TII.2022.3159583
3. W. Qi, S. E. Ovur, Z. J. Li, A. Marzullo, R. Song, Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network, *IEEE Rob. Autom. Lett.*, **6** (2021), 6039–6045. https://doi.org/10.1109/LRA.2021.3089999
4. Y. Ren, Y. Leng, J. Qi, P. K. Sharma, J. Wang, Z. Almakhadmeh, et al., Multiple cloud storage mechanism based on blockchain in smart homes, *Future Gener. Comput. Syst.*, **115** (2021), 304–313. https://doi.org/10.1016/j.future.2020.09.019
5. W. Qi, H. Fan, H. R. Karimi, H. Su, An adaptive reinforcement learning-based multimodal data fusion framework for human-robot confrontation gaming, *Neural Networks*, **164** (2023), 489–496.
6. J. Zhao, Y. F. Lv, Output-feedback robust tracking control of uncertain systems via adaptive learning, *Int. J. Control Autom. Syst.*, **21** (2023), 1108–1118. https://doi.org/10.1007/s12555-021-0882-6
7. Y. Y. Goh, D. J. Jung, G. Y. Hwang, J. M. Chung, Consumer electronics product manufacturing time reduction and optimization using AI-based PCB and VLSI circuit designing, *IEEE Trans. Consum. Electron.*, **69** (2023), 240–249. https://doi.org/10.1109/TCE.2023.3240249
8. Q. Yu, J. Ren, H. Zhou, W. Zhang, A cybertwin based network architecture for 6G, in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020. https://doi.org10.1109/6GSUMMIT49458.2020.9083808
9. Y. Wang, Z. Liu, J. Xu, W. Yan, Heterogeneous network representation learning approach for ethereum identity identification, *IEEE Trans. Comput. Soc. Syst.*, 2021. https://doi.org/10.1109/TCSS.2022.3164719

10. F. Tonini, C. Raffaelli, L. Wosinska, P. Monti, Cost-optimal deployment of a C-RAN with hybrid fiber/FSO fronthaul, *J. Opt. Commun. Networking*, **11** (2019), 397–408. https://doi.org/10.1364/JOCN.11.000397
11. C. Yoon, D. Cho, Energy efficient beamforming and power allocation in dynamic TDD based C-RAN system, *IEEE Commun. Lett.*, **19** (2015), 1806–1809. https://doi.org/10.1109/LCOMM.2015.2469294
12. M. S. Al-Abiad, M. Z. Hassan, M. J. Hossain, A joint reinforcement-learning enabled caching and cross-layer network code in F-RAN with D2D communications, *IEEE Trans. Commun.*, **70** (2022), 4400–4416. https://doi.org/10.1109/TCOMM.2022.3168058
13. Y. Zhang, J. Chen, C. Zhong, H. Peng, W. Lu, Active IRS-assisted integrated sensing and communication in C-RAN, *IEEE Wireless Commun. Lett.*, **12** (2023), 1295–1315. https://doi.org/10.1109/LWC.2022.3228405
14. J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, et al., An overview of signal processing techniques for joint communication and radar sensing, *IEEE J. Sel. Top. Signal Process.*, **15** (2021), 1295–1315. https://doi.org/10.1109/JSTSP.2021.3113120
15. D. Ngabo, D. Wang, C. Iwendi, J. H. Anajemba, L. A. Ajao, C. Biamba, Blockchain-based security mechanism for the medical data at fog computing architecture of internet of things, *Electronics*, **10** (2021), 2–17. https://doi.org/10.3390/electronics10172110
16. S. T. Chen, X. H. Qiu, X. Y. Tan, Z. J. Fang, Y. C. Jin, A model-based hybrid soft actor-critic deep reinforcement learning algorithm for optimal ventilator settings, *Inf. Sci.*, **611** (2022), 47–64. https://doi.org/10.1016/j.ins.2022.08.028
17. Y. Ren, Y. Leng, Y. Cheng, J. Wang, Secure data storage based on blockchain and coding in edge computing, *Math. Biosci. Eng.*, **16** (2021), 1874–1892. https://doi.org/10.3934/mbe.2019091
18. C. Iwendi, Innovative augmented and virtual reality applications for disease diagnosis based on integrated genetic algorithms, *Int. J. Cognit. Comput. Eng.*, **4** (2023), 266–276. https://doi.org/10.1016/j.ijcce.2023.07.004
19. Z. Yu, J. Liu, S. Liu, Q. Yang, Co-optimizing latency and energy with learning based 360 video edge caching policy, in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022. https://doi.org/10.1109/WCNC51071.2022.9771944
20. H. H. Gao, W. Q. Huang, T. Liu, Y. Yin, Y. Li, pp02: Location privacy-oriented task offloading to edge computing using reinforcement learning for intelligent autonomous transport systems, *IEEE Trans. Intell. Transp. Syst.*, 2022. https://doi.org/10.1109/TITS.2022.3169421
21. G. Gao, Y. Wen, J. Cai, Cache: Supporting cost-efficient adaptive bitrate streaming, *IEEE MultiMedia*, **24** (2017), 19–27. https://doi.org/10.1109/MMUL.2017.265091759
22. H. H. Gao, J. D. Huang, Y. Tao, W. Hussain, Y. Z. Huang, The joint method of triple attention and novel loss function for entity relation extraction in small data-driven computational social systems, *IEEE Trans. Comput. Soc. Syst.*, 2022. https://doi.org/10.1109/TCSS.2022.3178416
23. T. Dang, M. Peng, Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks, *IEEE J. Sel. Areas Commun.*, **37** (2019), 1594–1607. https://doi.org/10.1109/JSAC.2019.2916486

24. Y. Sun, M. Peng, S. Mao, Deep reinforcement learning-based mode selection and resource management for green fog radio access networks, *IEEE Internet Things J.*, **6** (2018), 1960–1971. https://doi.org/10.1109/JIOT.2018.2871020
25. A. Helmy, A. Nayak, Energy-efficient decentralized framework for the integration of fog with optical access networks, *IEEE Trans. Green Commun. Networking*, **4** (2020), 927–938. https://doi.org/10.1109/TGCN.2020.2974820
26. M. S. Elbamby, C. Perfecto, M. Bennis, K. Doppler, Toward lowlatency and ultra-reliable virtual reality, *IEEE Network*, **32** (2018), 78–84. https://doi.org/10.1109/MNET.2018.1700268
27. J. Tian, H. Zhang, D. Wu, D. Yuan, Interference-aware cross-layer design for distributed video transmission in wireless networks, *IEEE Trans. Circuits Syst. Video Technol.*, **26** (2015), 978–991. https://doi.org/10.1109/TCSVT.2015.2430611
28. X. Peng, Y. Shi, J. Zhang, K. B. Letaief, Layered group sparse beamforming for cache-enabled green wireless networks, *IEEE Trans. Commun.*, **65** (2017), 5589–5603. https://doi.org/10.1109/TCOMM.2017.2745579
29. H. Zhang, Y. Qiu, X. Chu, K. Long, V. C. M. Leung, Fog radio access networks: Mobility management, interference mitigation, and resource optimization, *IEEE Wireless Commun.*, **24** (2017), 120–127. https://doi.org/10.1109/MWC.2017.1700007
30. Y. Li, M. Xia, Y. Wu, First-order algorithm for content-centric sparse multicast beamforming in large-scale C-RAN *IEEE Trans. Wireless Commun.*, **17** (2018), 5959–5974. https://doi.org/10.1109/TWC.2018.2852300
31. W. J. Lv, R. Wang, J. Wu, J. Xu, P. Li, J. W. Dou, Degrees of freedom of the circular multirelay MIMO interference channel in IoT networks, *IEEE Internet Things J.*, **5** (2018), 1957–1966. https://doi.org/10.1109/JIOT.2018.2817580
32. C. Lu, Y. Liu, An efficient global algorithm for single-group multicast beamforming, *IEEE Trans. Signal Process.*, **65** (2017), 3761–3774. https://doi.org/10.1109/TSP.2017.2699640
33. B. Dai, W. Yu, A semiblind digital-domain calibration of pipelined A/D converters via convex optimization, *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, **23** (2015), 1375–1379. https://doi.org/10.1109/TVLSI.2014.2336472
34. L. Huang, D. Liu, Y. Fang, Convergence of an SDP hierarchy and optimality of robust convex polynomial optimization problems, *Ann. Oper. Res.*, **23** (2022), 33–59. https://doi.org/10.1007/s10479-022-05103-6
35. Y. Li, Y. Gong, S. Xiao, Synthesis of modular subarrayed phased-array withn shaped-beams by means of sequential convex optimization, *IEEE Antennas Wireless Propag. Lett.*, **21** (2022), 1168–1172. https://doi.org/10.1109/LAWP.2022.3160733