Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Resource allocation for network profit maximization in NOMA-based F-RANs: a game-theoretic approach[*]

Xueyan CAO, Shi YAN[‡], Hongming ZHANG

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China

E-mail: 2013212868@bupt.edu.cn; yanshi01@bupt.edu.cn; zhanghm@bupt.edu.cn

Received July 9, 2021; Revision accepted Oct. 19, 2021; Crosschecked Nov. 9, 2021; Published online Mar. 4, 2022

Abstract: Non-orthogonal multiple access (NOMA) based fog radio access networks (F-RANs) offer high spectrum efficiency, ultra-low delay, and huge network throughput, and this is made possible by edge computing and communication functions of the fog access points (F-APs). Meanwhile, caching-enabled F-APs are responsible for edge caching and delivery of a large volume of multimedia files during the caching phase, which facilitates further reduction in the transmission energy and burden. The need of the prevailing situation in industry is that in NOMA-based F-RANs, energy-efficient resource allocation, which consists of cache placement (CP) and radio resource allocation (RRA), is crucial for network performance enhancement. To this end, in this paper, we first characterize an NOMA-based F-RAN in which F-APs of caching capabilities underlaid with the radio remote heads serve user equipments via the NOMA protocol. Then, we formulate a resource allocation problem for maximizing the defined performance indicator, namely network profit, which takes caching cost, revenue, and energy efficiency into consideration. The NP-hard problem is decomposed into two sub-problems, namely the CP sub-problem and RRA sub-problem. Finally, we propose an iterative method and a Stackelberg game based method to solve them, and numerical results show that the proposed solution can significantly improve network profit compared to some existing schemes in NOMA-based F-RANs.

Key words: Fog radio access network; Non-orthogonal multiple access; Game theory; Cache placement; Resource allocation

https://doi.org/10.1631/FITEE.2100341 **CLC number:** TN929.5

1 Introduction

With the proliferation of mobile services in the fifth-generation (5G) wireless communication system and interaction development of multiple domain resources in the six-generation (6G) wireless communication system, more and more applications offering a significantly enriched user experience have emerged in the domain of popular use; as examples, we may mention augmented reality (AR), virtual reality (VR), intelligent transportation, and environment protection. Furthermore, the integration of artificial intelligence (AI) and the next-generation networking techniques promotes the development of intellicise networks (Zhang P et al., 2022). However, there exists a gap between the large demands of user equipments (UEs) and the limited capabilities of network infrastructures. Specifically, the cloud server is responsible for global signal processing, management, and allocation for communication, caching, and computing resources, in a centralized way; ever-increasing demands and traffic aggravate the processing burden of the cloud server and the energy consumption from the cloud to the UEs,

[‡] Corresponding author

[*] Project supported in part by the National Natural Science Foundation of China (Nos. U21A20444 and 61901044) and Young Elite Scientist Sponsorship Program by China Institute of Communications

ORCID: Xueyan CAO, https://orcid.org/0000-0002-9996-6520; Shi YAN, https://orcid.org/0000-0002-5953-1979

which hinders the development of 6G low-energy network. To fill this gap, fog radio access networks (F-RANs) arise as a novel paradigm with multiple fog access points (F-APs), which have the abilities of resource management, distributed signal processing, edge computing, and caching (Peng et al., 2016, 2020; Dang and Peng, 2019). By sinking a part of the functions from the cloud server to the edge F-APs, much traffic can be processed locally and delivered by the F-APs to their serving UEs (FUEs) directly. The F-RANs greatly mitigate the burden of global processing in the cloud server, and reduce the energy consumption in processing and transmission via backhaul links (Park et al., 2016; Kong et al., 2018).

Non-orthogonal multiple access (NOMA) technique has been validated as a promising multiple access mechanism allowing multiple users to share the same resource block (RB), in power, code, and spatial domains rather than the conventional time and frequency domains. The network throughput and spectrum efficiency (SE) will be improved further when NOMA can be integrated with F-RANs. Specifically, multiple UEs located at the edge of the network can be served by F-APs where the requested files are delivered by F-APs with file copies. Additionally, the file delivery from F-APs to UEs obeys the NOMA protocol, and the UEs perform the successive interference cancellation (SIC) technique to extract the requested signal. Although characterized by distinct superiority, this approach also introduces severe challenges, as noted by Zhang HJ et al. (2018) and as the following content explains. First, it is doubtful that simultaneous transmissions by NOMA may consume more transmission energy and increase mutual interference, which is unexpected for the goal of energy-efficient resource allocation with energy-constrained devices. Second, the co-provision of radio resource, cache placement (CP), and user access mechanism in NOMA-based F-RANs increases the difficulty of resource allocation and performance enhancement.

Motivated by these observations, we integrate NOMA-based transmission and file caching in F-RANs and jointly allocate the caching and communication resources. Specifically, the file caching and user association, which rely on the file price, capacity of F-APs, and user requests, are considered to be done during the caching phase. In the caching phase, F-APs are responsible for caching a large volume of multimedia files according to the file price and user requests to further reduce the transmission energy and burden of backhaul links. In addition, the NOMA-based transmission, which is influenced by the RRA scheme, is considered to be done during the transmission phase. In this phase, F-APs serve multiple FUEs via the NOMA protocol in the F-AP mode, underlaid with the radio remote heads (RRHs) in the cellular mode. To achieve huge capacity, ultra-low delay, and high energy efficiency, we further formulate a resource allocation problem for maximizing the newly defined performance indicator. The problem is NP-hard and intractable, and we solve it by decomposing it into two sub-problems according to the two phases. An iterative programming algorithm and a Stakelberg framework are formulated for each sub-problem, where the CP scheme is attained by a simple and efficient heuristic algorithm, and the RRA scheme is attained by a game-theoretic method with sequential interaction relationship between players. Our contributions are summarized as follows:

1. We characterize an NOMA-based F-RAN in which the F-APs of caching capabilities are underlaid with RRHs serving multiple UEs via the power-domain NOMA protocol. We divide the whole file transmission into the caching phase (from the cloud to F-APs) and the transmission phase (from F-APs to FUEs). To mitigate the excessive cost of caching and improve network energy efficiency, we jointly optimize the CP, pricing of files, power and subchannel allocation, and formulate a resource allocation problem to maximize the newly defined performance indicator, namely network profit, which takes both cost of CP in the caching phase and energy efficiency in the transmission phase into consideration.

2. Due to the coexistence of binary, integer, and continuous variables, this non-convex problem is intractable with NP-hardness and we decompose it into two sub-problems, namely the CP sub-problem and RRA sub-problem. In the CP sub-problem, we propose an iterative programming algorithm based on simulated annealing (SA) to maximize the profit of CP in the caching phase. In the RRA sub-problem, we propose a hierarchical Stackelberg game approach consisting of a non-cooperative power allocation algorithm for the F-APs, and a one-to-many subchannel matching algorithm for the RRHs and F-APs to maximize network energy efficiency in the

transmission phase.

3. We examine the efficiency of our proposed iterative SA algorithm, non-cooperative power allocation algorithm, and stable subchannel matching algorithm. We also explore the impact of quantitative limits of FUEs, the number of F-APs, and the pricing of files on network performance, and find that there is no simple trend between caching and resource allocation strategies. Additionally, viewed in a comparative context against various caching approaches and resource allocation schemes, our proposed resource allocation approach is verified to achieve high network profit, especially when the essential constraints become more stringent.

2 Related works

Resource allocation is an effective approach for improving network throughput. Scholars have generated vast research output enumerating various suitable means to improve network throughput (Xu et al., 2016; Yang et al., 2018; Yao and Ansari, 2019). For example, Zhang JX et al. (2016) and Deng et al. (2016) focused on the storage allocation and content placement problem in hierarchical cache-enabled heterogeneous networks and F-RANs, respectively. Yu et al. (2019) investigated green fog computing by maximizing the network function considering energy efficiency with the constraints of power and interference. Sun et al. (2019) formulated a hierarchical RRA architecture in F-RANs, where network slicing was considered to relieve the heavy burden of centralized resource manager, and proposed a game-based resource allocation algorithm to solve the maximum SE problem efficiently. Zhou et al. (2019) introduced a machine learning based F-AP content placement method, where unsupervised learning was used to classify the popularity of requested contents and solve the content placement problem.

The NOMA protocol (Ding et al., 2016, 2019) functions as a valid technique to achieve high SE. However, NOMA-based F-RANs are faced with challenges such as simultaneity of various resource allocations and interaction influences among different resources (Zhai et al., 2018). To this end, Rai et al. (2021) proposed an NOMA-enabled fog-cloud structure in a novel density-aware F-RAN to tackle different aspects of high- and low-density regions, the objective being to meet the heterogeneous requirements of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) traffic. A framework of two different scenario configurations and performance analysis, consisting of independent caching association and transmission mode allocation, was considered to cater to the high-throughput and low-latency requirements in high- and low-density modes, respectively. In Cao et al. (2019), a communication resource allocation algorithm for energy efficiency maximization, consisting of subchannel reuse assignment and power allocation, was formulated for NOMA-based F-RANs and a game-based approach was proposed to solve it. In Liu et al. (2020), with the aim of maximizing the weighted sum rate while taking co-channel interference into consideration, a joint RB and power allocation problem was formulated. The authors applied monotonic optimization and proposed an outer polyblock approximation algorithm to obtain the global optimal solution. Bai et al. (2019) focused on dynamic power allocation of wireless subchannel in NOMA. By extending it to F-RANs, the authors proposed an improved fractional transmission power allocation (I-FTPA) algorithm. Yan et al. (2020) focused on the user access mode selection and content popularity prediction analysis without resource allocation in NOMA-based F-RANs.

However, the mentioned studies focus on communication resources such as power, subchannel, and integrated RB in NOMA-based F-RANs, without taking caching resource into consideration. Given the potential improvement of the integration of NOMA and F-RANs with edge caching ability, a joint resource allocation algorithm consisting of caching and radio resources in NOMA-based F-RANs is tackled in this study where the maximum network profit can be obtained by a tractable approach.

3 System model and problem formulation

In this section, we present a mathematical model of the NOMA-based F-RANs which we focus throughout the paper, and formulate an optimization problem in achieving the maximum network profit.

3.1 Network model

As depicted in Fig. 1, we consider the F-RANs comprising the cloud server, F-APs, RRHs, and UEs. The F-APs have the abilities of resource management, signal processing, and caching, while the RRHs have only the radio frequency function. Furthermore, we consider that the NOMA protocol is applied where one F-AP can serve multiple users in the same time and frequency RB. We denote the UEs served by the F-APs and RRHs as the FUEs and RUEs, respectively. Additionally, the communication mode where the FUEs are served by F-APs and receive the file directly from the F-AP is called the F-AP mode. Similarly, the communication mode where the RUEs are served by RRHs and receive the file from the cloud is called the RRH mode. We consider that the cloud server has all the desired files with the same size s_{f}, each of which is indexed by $\mathcal{F}=\{1,2,\ldots,F\}$. Denote the set of RRHs, F-APs, and FUEs by $\mathcal{R}=\{1,2,\ldots,R\}$, $\mathcal{N}=\{1,2,\ldots,N\}$, and $\mathcal{L}=\{1,2,\ldots,L\}$, respectively. We assume that R RRHs serve R RUEs by R orthogonal subchannels. Without loss of generality, we consider that subchannel r is allocated to RRH r for serving RUE r. On the other hand, with the aid of the NOMA technique, N F-APs can serve L FUEs, with each F-AP serving up to Q_n FUEs. We also assume that each F-AP can access one subchannel only, while each subchannel occupied by each RRH can accommodate at most M F-APs at one time slot. Unless otherwise specified, the notations used throughout the paper are the ones summarized in Table 1. The whole network operation process can be expressed as three sub-figures in Fig. 2.

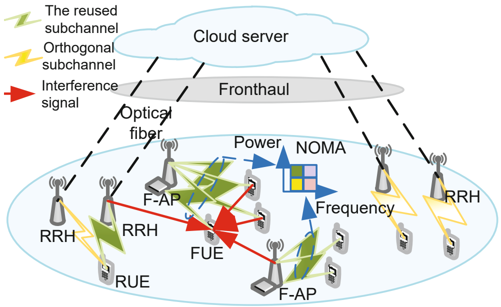

Fig. 1 System model of downlink transmission in an NOMA-based F-RAN

3.2 Caching model

We now present the caching model. The existing caching schemes consist of the random caching scheme (where all files are cached with an equal probability) and the popularity-based caching scheme (where the more popular file has the larger probability of being cached). In addition, some specialized caching schemes are proposed by certain rules. In our system, the popularity-based caching scheme is not feasible because the F-AP may not be willing to contribute to all of its limited space with additional

Table 1 Notations and main abbreviations used throughout the paper

Notation	Description	Notation	Description
F-AP	Fog access points	CP	Cache placement
RRA	Radio resource allocation	RRH	Radio remote head
FUE	User equipment served by F-APs	RUE	User equipment served by RRHs
$\mathcal{R}$, $\mathcal{N}$, $\mathcal{L}$	Sets of RRHs, F-APs, and FUEs, respectively	$\mathcal{F}$	Set of files
s_{f}	Size of each file	S_n^{C}	Limited storage of F-AP n
$c_{n\mathrm{c}}$	Backhaul cost from the cloud to F-AP n	$\eta_{n,f}$	Index of caching file f on F-AP n
$d_{n\mathrm{c}}$	Transmission delay from the cloud to F-AP n	$\xi_{n,l}$	Index of serving FUE l by F-AP n
$Z_{l,f}$	Index of requesting file f of FUE l	$\theta_{n,f}$	Price of caching file f of F-AP n
U	Profit of the cloud server	V	Profit of F-APs
$h_{r,n,q}$	Channel coefficient from RRH r to FUE q	$p_{n,q}$	Transmission power of F-AP n to FUE q
	served by F-AP n	p_n	Transmission power of F-AP n
$h_{n,q}$	Channel coefficient from F-AP n to FUE q	p_r	Transmission power of RRH r
h_r	Channel coefficient from RRH r to RUE r	$p_n^{\max}$	Maximum transmission power of F-AP n
$h_{i,n,q}$	Channel coefficient from F-AP i to FUE q	$\boldsymbol{p}$	Transmission power of F-APs
	served by F-AP n	ϕ_n	Spectrum efficiency of F-AP n
$x_{n,r}$	Matching index from F-AP n to RRH r	ϕ_r	Spectrum efficiency of RRH r
$y_{r,n}$	Matching index from RRH r to F-AP n	Q	Serving limitation of one F-AP
$\boldsymbol{x}$, $\boldsymbol{y}$	Matching indexes	M	Tolerant accommodation of each RRH
m_r	Number of F-APs matching RRH r		

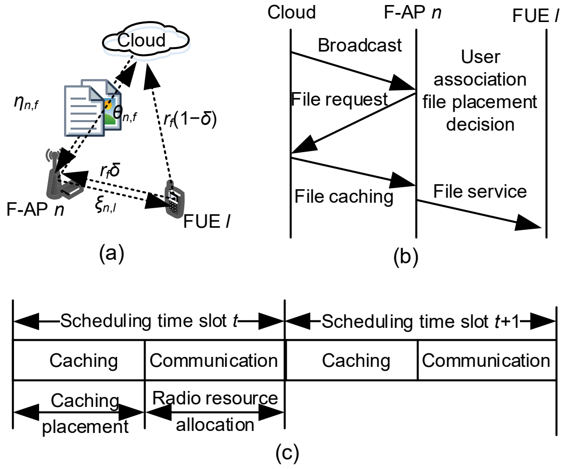

Fig. 2 Diagram of caching and communication phases: (a) relationship between variables and agents during the caching and communication phases; (b) scheduling procedure among the cloud, F-APs, and FUEs; (c) time sequence of the scheduling period

cost and the user requests are unpredictable. Thus, we propose an on-demand CP strategy which relies on the cost saved by caching, the file prices, and demands of UEs.

The overall framework of the proposed CP strategy is shown as Fig. 2b. All the available row files or the copyright is published in the cloud server and sold at different prices. The cloud server broadcasts the price to F-APs, which would buy some files via backhaul links to maximize own profit. In particular, the profit of F-AP is related to the file prices, incomes of users, and the transmission cost saved by local delivery. In turn, the cloud server desires to maximize its profit, which is related to the income for caching and user access. The objective of CP is to find the equilibrium among the optimal CP, pricing of files, and the user association scheme of F-APs and the cloud server. Upon reaching the optimal strategy, all F-APs cache the files and charge for FUEs. This being so, when FUE l accesses in F-AP n and requests for file f, if file f or its copyright has been cached in F-AP n, it can be delivered to FUE l directly with a low transmission delay and light backhaul link burden. Otherwise, F-AP n should send the first request for file f to the cloud server via the congested and weak backhaul link and transmit to FUE l with a large transmission delay.

To this end, we now describe the mathematical model mentioned throughout the paper and introduce some notations in detail. First, we introduce two binary indexes $\xi_{n,l} \in \{0,1\}$ and $\eta_{n,f} \in \{0,1\}$. When F-AP n serves FUE l, $\xi_{n,l} = 1$; otherwise, $\xi_{n,l} = 0$. Similarly, when file f is cached in F-AP n, $\eta_{n,f} = 1$; otherwise, $\eta_{n,f} = 0$. Note that $\sum_{n=1}^{N} \xi_{n,l} = 1,\ \forall l \in \mathcal{L}$, $Q_n = \sum_{l=1}^{L} \xi_{n,l} \le Q,\ \forall n \in \mathcal{N}$, and $\sum_{f=1}^{F} \eta_{n,f} s_{\mathrm{f}} \le S_n^{\mathrm{C}},\ \forall n \in \mathcal{N}$ should be satisfied. Q_n is the number of FUEs under F-AP n. For the F-APs, taking the CP into consideration, the profit of F-AP is defined as the difference between the revenue and the expenditure. The revenue comprises the backhaul transmission energy cost and delay cost saved by local caching, and the income for user association. The expenditure is the cost for local caching from the cloud. In particular, the revenue of F-AP n accessing FUE l which requests for file f can then be expressed as

$$\begin{aligned} V_n^{\mathrm{re}} = & \sum_{f=1}^{F} \eta_{n,f} \left(\kappa d_{nc} + c_{nc}\right) \\ & + \sum_{f=1}^{F} \sum_{l=1}^{L} \xi_{n,l} Z_{l,f} \eta_{n,f} r_f \delta, \end{aligned} \tag{1}$$

where κ represents the weight between transmission delay and delay cost, r_f represents the price of file f paid by FUE l to the cloud server, and δ represents the percentage of income which the cloud server shares with F-AP n. The expenditure of F-AP n accessing FUE l which requests for file f can then be expressed as

$$V_n^{\mathrm{ex}} = \sum_{f=1}^{F} \theta_{n,f} \eta_{n,f}. \tag{2}$$

So, the profit of F-AP n accessing FUE l which requests for file f can then be expressed as

$$V_n = V_n^{\mathrm{re}} - V_n^{\mathrm{ex}}, \tag{3}$$

and the profit of F-APs is expressed as

$$V = v_1 V_1 + v_2 V_2 + \ldots + v_N V_N, \tag{4}$$

where v_n $(n = 1, 2, \ldots, N)$ represents the weighted coefficient of each F-AP n. For the cloud server, the profit comes from the income for the files and user association due to the local CP. Thus, the profit of the cloud server can be expressed as

$$\begin{aligned} U = & \sum_{n=1}^{N} \sum_{f=1}^{F} \theta_{n,f} \eta_{n,f} \\ & + \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{f=1}^{F} \xi_{n,l} Z_{l,f} \eta_{n,f} r_f \left(1 - \delta\right). \end{aligned} \tag{5}$$

3.3 Communication model

We now focus on the communication model. A block fading channel is considered in our system, where the channel fading of each subchannel remains unchanged but varies independently across different subchannels. Recalling the definition of $h_{n,q}$ in Table 1, $h_{n,1} > h_{n,2} > ... > h_{n,Q_n}$ holds. Each F-AP decodes the signal with low ranking of the channel coefficient via the SIC technique. The transmission power of F-AP n to FUE q, i.e., $p_{n,q}$, satisfies $p_{n,q} \leq p_n^{\max}$.

In the transmission phase, the data transmission rate of FUE q accessing F-AP n is shown as

$$\begin{aligned}
&g_{n,q} = B\log_2\left(1+\frac{x_{n,r}y_{r,n}p_{n,q}|h_{n,q}|^2}{\sigma^2+I_1+I_2+I_3}\right),\\
&I_1 = \sum_{j=1}^{q-1} p_{n,j}|h_{n,q}|^2,\\
&I_2 = x_{n,r}y_{r,n}p_r|h_{r,n,q}|^2,\\
&I_3 = \sum_{i=1,i\neq n}^{m_r} x_{i,r}y_{r,i}p_{i,q}|h_{i,n,q}|^2,
\end{aligned} \tag{6}$$

where B is the bandwidth of each channel, I_1 is the interference from other FUEs under the same F-AP, I_2 is the interference from the RRH which reuses the same subchannel, and I_3 is the interference from the F-APs that reuse the same subchannel as F-AP n. When F-AP n (RRH r) matches RRH r (F-AP n), $x_{n,r} = 1$ ($y_{r,n} = 1$); otherwise, $x_{n,r} = 0$ ($y_{r,n} = 0$). Therefore, the data transmission rate of RUE r is

$$g_r = B\log_2\left(1+\frac{p_r|h_r|^2}{\sigma^2+\sum\limits_{i=1}^{m_r} x_{i,r}y_{r,i}p_{i,r}|h_{i,r}|^2}\right), \tag{7}$$

where σ^2 is the variance of the noise during the transmission phase, $p_{i,r}$ is the transmission power of F-AP i to RUE r, and $h_{i,r}$ is the channel coefficient from F-AP i to RUE r. To guarantee the quality of service of each UE, we also consider SE constraints as follows:

$$\phi_n(x_n, p_n) = \sum_{q=1}^{Q_n} g_{n,q},\ \phi_r(y_r) = g_r, \tag{8}$$

where $y_r = m_r = \sum_{n=1}^{N} y_{r,n}$, $x_n = \sum_{r=1}^{R} x_{n,r}$, $p_n = \sum_{q=1}^{Q_n} p_{n,q}$, and $Q_n = \sum_{l=1}^{L} \xi_{n,l}$. The network energy efficiency (EE), which is defined as the total average number of bit/(Hz·W) successfully delivered to the others (Ng et al., 2012), can be expressed as

$$\mathrm{EE} = \frac{\sum\limits_{n=1}^{N}\sum\limits_{q=1}^{Q_n} g_{n,q} + \sum\limits_{r=1}^{R} g_r}{\sum\limits_{n=1}^{N}\sum\limits_{q=1}^{Q_n} p_{n,q} + \sum\limits_{r=1}^{R} p_r + P_{\mathrm{cir}}^{\mathrm{F}} + P_{\mathrm{cir}}^{\mathrm{R}}}, \tag{9}$$

where $P_{\mathrm{cir}}^{\mathrm{F}}$ and $P_{\mathrm{cir}}^{\mathrm{R}}$ are the circuit power of F-APs and RRHs respectively which ensure the normal operation of the equipment.

3.4 Problem formulation

Our work aims to optimize the CP strategy to reduce the transmission burden and delay, and simultaneously achieve energy-efficient resource allocation. To this end, we define a new performance indicator, namely network profit $\mathcal{P}$, taking both profits of CP in the caching phase and energy efficiency of NOMA transmission in the transmission phase into consideration:

$$\mathcal{P} = \rho_1(V+U) + \rho_2\mathrm{EE}, \tag{10}$$

where $V+U$ and EE represent the profits in the caching phase and transmission phase respectively, and ρ_1 and ρ_2 are the weighted coefficients. Without loss of generality, we equate the influences of two phases and set $\rho_1 = \rho_2 = 1$. Then the optimization problem can be expressed as

$$\max_{\theta,\xi,\eta,\boldsymbol{p},\{\boldsymbol{x},\boldsymbol{y}\}} \mathcal{P} \tag{11}$$

$$\text{s.t. } \xi_{n,l} \in \{0,1\},\ \eta_{n,f} \in \{0,1\},\ \forall n \in \mathcal{N},\ \forall f \in \mathcal{F},\ \forall l \in \mathcal{L}, \tag{11a}$$

$$x_{n,r} \in \{0,1\},\ y_{r,n} \in \{0,1\},\ \forall n \in \mathcal{N},\ \forall r \in \mathcal{R}, \tag{11b}$$

$$0 \leq \theta_{n,f} \leq \theta_{\max}, \tag{11c}$$

$$\sum_{n=1}^{N} \xi_{n,l} = 1,\ \sum_{l=1}^{L} \xi_{n,l} \leq Q,\ \forall l \in \mathcal{L},\ \forall n \in \mathcal{N}, \tag{11d}$$

$$\sum_{f=1}^{F} \eta_{n,f}s_{\mathrm{f}} \leq S_n^{\mathrm{C}},\ \forall n \in \mathcal{N}, \tag{11e}$$

$$\sum_{q=1}^{Q_n} p_{n,q} = p_n^{\max},\ p_{n,q} > 0,\ \forall n \in \mathcal{N}, \tag{11f}$$

$$\phi_n(x_n, p_n) \geq \phi_n^{\min},\ \phi_r(y_r) \geq \phi_r^{\min}$$
$$\forall r \in \mathcal{R},\ \forall n \in \mathcal{N}, \tag{11g}$$
$$\sum_{r=1}^{R} x_{n,r} = 1,\ \sum_{n=1}^{N} y_{r,n} \leq M. \tag{11h}$$

In this optimization problem, θ, ξ, and η represent the price of caching, user association, and CP variables in the caching phase, respectively. Constraints (11a) and (11b) represent the limitations to the controllable variables in each phase. Constraints (11c)–(11e) are the limitations in the caching phase, where (11c) represents that the price of the file must not be too high due to the law of demand from the field of economics, (11d) represents that one FUE can access one F-AP only and one F-AP can accommodate Q FUEs at most within one time slot, and (11e) represents that the capacity of CP cannot exceed the storage of one F-AP. Constraints (11f)–(11h) are the limitations in the transmission phase, where (11f) suggests that the transmission power of Q_n FUEs under F-AP n should not exceed the upper limitation, (11g) represents the SE constraints of the FUEs and RUEs, and (11h) represents that one F-AP can match only one RRH to reuse the subchannel and that one RRH can match M F-APs at most within the interference tolerance. Note that the optimization problem is NP-hard because the 0–1 binary and the continuous variables are mixed. In the next section, we will decompose it into two sub-problems and propose two tractable algorithms.

4 Problem solution

Reviewing the expression of network profit and the optimization problem, it is evident that the CP, pricing of files, and user association are coupled with one another while working only in the caching phase, but function independent of the power and subchannel allocation; the energy-efficient resource allocation, consisting of power and subchannel allocation, works in the transmission phase. In this section, we first decompose problem (11) into two sub-problems, namely the CP sub-problem and RRA sub-problem; then we propose an iterative algorithm and a game-based algorithm to solve them, respectively. The descriptions of four algorithms are listed in Table 2 in detail.

4.1 CP sub-problem

In this subsection, we focus on the CP sub-problem by optimizing the CP strategy, pricing of files, and user association. Specifically, the cloud server desires to maximize its own profit through high pricing of files and high income from FUEs benefiting from local caching. Meanwhile, F-APs of caching capabilities desire to maximize profits which are related to the large backhaul link transmission cost and delay saved by local caching, high income from FUEs, and low expenditure to the cloud server for caching. We reformulate two optimization problems for the cloud server and the F-APs, respectively. First, the maximization optimization problem for the cloud server is shown as

$$\max_{\theta_{n,f}} U \tag{12}$$
$$\text{s.t. } 0 \leq \theta_{n,f} \leq \theta_{\max},\ \forall n \in \mathcal{N},\ \forall f \in \mathcal{F}, \tag{12a}$$

where constraint (12a) suggests that the prices of files must not be excessive since the demanded quantity of goods falls as the price of goods rises by the law of demand from the field of economics. Next, we can similarly show the maximization optimization problem for the F-APs as

$$\max_{\eta_{n,f},\xi_{n,l}} V \tag{13}$$
$$\text{s.t. } \xi_{n,l} \in \{0,1\},\ \forall n \in \mathcal{N},\ \forall l \in \mathcal{L}, \tag{13a}$$
$$\sum_{n=1}^{N} \xi_{n,l} = 1,\ \forall l \in \mathcal{L}, \tag{13b}$$
$$\sum_{l=1}^{L} \xi_{n,l} \leq Q,\ \forall n \in \mathcal{N}, \tag{13c}$$
$$\eta_{n,f} \in \{0,1\},\ \forall n \in \mathcal{N},\ \forall f \in \mathcal{F}, \tag{13d}$$

where constraint (13a) represents that the user association variable is a binary variable, constraints (13b) and (13c) suggest that each FUE can access only one F-AP and one F-AP can accommodate Q FUEs at most at one time slot, and constraint (13d) represents that the CP is a binary variable. Note that the optimization problem (13) is a multi-objective optimization problem. For simplification purposes, we define the weight $v_n = 1$ for each element ($\forall n \in \mathcal{N}$) of the weighted sum method $V = \sum_{n=1}^{N} v_n V_n$ in transferring from multiple objectives to a single objective $V = \sum_{n=1}^{N} V_n$.

Table 2 Description of the four algorithms

Algorithm	Description
Algorithm 1: cache placement (CP) algorithm	The CP strategy can be attained by the SA scheme, where the CP initialization, capacity threshold Q, and maximum price $\theta_{\max}$ are the input variables, and the optimum profit and strategies of the cloud and F-APs are the output solution.
Algorithm 2: non-cooperative game based power allocation algorithm	The power allocation strategy can be attained by the Dinkelbach scheme and sub-gradient based method, where the initial power constitutes input variables and the optimum power allocation of each F-AP constitutes the output solution.
Algorithm 3: one-to-many matching based subchannel allocation algorithm	Optimization can be attempted for the subchannel allocation solution between F-APs and RRHs until there is no point unmatched with any member of the opposite set.
Algorithm 4: radio resource allocation (RRA) algorithm	The optimum solution to multi-dimensional resource allocation can be attained iteratively.

It is clear that problem (12) is a linear programming problem for continuous variable θ but problem (13) is still a non-convex problem due to the coexistence of binary variables ξ and η. In particular, the optimization problem for the F-APs in problem (13) is a Knapsack problem and we propose an advanced iterative SA optimization algorithm to solve it within the acceptable latency. Based on this, an iterative dynamic programming algorithm for the CP sub-problem is proposed, as shown in Algorithm 1. Specifically, we denote a well-defined generation function to generate new solutions and calculate the increment of the evaluation function of two iterations, which is set as an optimization function generally. To facilitate the subsequent calculation and evaluation, and to reduce the algorithm's running time, the generation function is set through simple linear change based on the current solution. Moreover, we use the Metropolis guideline where the new solution is accepted if the increment (i.e., the difference, as ascertained by the evaluation function, between two solutions of two steps) is more than zero. Otherwise, it is accepted as a probability related to the increment. This method can significantly avoid getting trapped into local optimal solutions.

Algorithm 1 Cache placement (CP) algorithm

```
Initialize: the CP, user association, and the price randomly as $\eta_0$, $\xi_0$, $\theta_0$, which satisfy constraints (12a), (13a)–(13d), the initial profit of the cloud as $U(0)$, and that of F-APs as $V(0)$, $t = 1, k, T$
while $|V(t) - V(t-1)| > \zeta$ do
    Compute the difference of the evaluation function between two solutions, $\mathrm{d}E = V(\eta_{t+1}, \xi_{t+1}) - V(\eta_t, \xi_t)$
    if $\mathrm{d}E \geq 0$ then
        Update $\eta_{t+1} = \eta_t,\ \xi_{t+1} = \xi_t$
    else
        $\exp\left(\frac{\mathrm{d}E}{kT}\right) > \mathrm{rand}(1)$
        Update $\eta_{t+1} = \eta_t,\ \xi_{t+1} = \xi_t$
        Change the temperature, $T = \tau T$
        $t = t + 1$
        Return $\eta_{n,f}$ to the cloud and compute the profit function $U$ by expression (12)
        Update $\theta = \theta + \Delta_\theta$
        Return $\theta^*$ and output $U$
    end if
end while
```

In Algorithm 1, when the difference of two objective functions in two iterations ΔV is smaller than the threshold, the algorithm stops and the current strategy is the final optimum one. In addition, we choose the initial solution by fixed point variation to overcome the defect of getting trapped into local optimal solutions, and meanwhile speed up the convergence. Due to the finite number of players, Algorithm 1 is guaranteed to converge to a stable equilibrium, under which no player can improve its utility by unilaterally changing its own strategy without decreasing utilities of other players. As far as the application aspect is concerned, the user association and access, via multiple access techniques in time and frequency domains, can alleviate traffic burden and achieve efficient transmission in scenarios such as concerts and sports venues that involve relatively few base stations (BSs) and densely packed users, which typically leads to overloading in the operations of BSs or wireless access points.

4.2 RRA sub-problem

After obtaining the CP strategy ($\eta_{n,f}^*$, $\xi_{n,l}^*$, and $\theta_{n,f}^*$), we fix the optimal cost of caching, U, for the cloud server and the profit, V, for the F-APs in the caching phase and focus on the RRA strategies in the transmission phase. Specifically, one F-AP can serve more than one FUE via the NOMA protocol within the available transmission powers; for improving the efficiency of spectrum allocation, we assume that one subchannel can be occupied by more than one F-AP in the F-AP mode, in addition to RRHs in the cellular mode, under the condition in which the received co-channel interference does not exceed the acceptable threshold of the RRHs. Based on this, we reformulate the RRA sub-problem for the F-APs and RRHs under multiple constraints of SE, transmission power, and subchannel matching. In particular, the optimization problem is shown as

$$\max_{\{\boldsymbol{x},\boldsymbol{y}\},\boldsymbol{p}} \text{EE} \tag{14}$$

$$\text{s.t.} \sum_{q=1}^{Q_n} p_{n,q} = p_n^{\max},\ p_{n,q} > 0,\ \forall n \in \mathcal{N}, \tag{14a}$$

$$\phi_n(x_n, p_n) \geq \phi_n^{\min},\ \forall n \in \mathcal{N}, \tag{14b}$$

$$\phi_r(y_r) \geq \phi_r^{\min},\ \forall r \in \mathcal{R}, \tag{14c}$$

$$x_{n,r} \in \{0,1\},\ \forall r \in \mathcal{R}, \tag{14d}$$

$$y_{r,n} \in \{0,1\},\ \forall n \in \mathcal{N}, \tag{14e}$$

$$\sum_{r=1}^{R} x_{n,r} = 1,\ \sum_{n=1}^{N} y_{r,n} \leq M, \tag{14f}$$

where the solutions obtained are the multiple vectors expressed as $\boldsymbol{x} = [x_{n,r}]$, $\boldsymbol{y} = [y_{r,n}]$, $\boldsymbol{p} = [p_{n,q}]$, $\forall n \in \mathcal{N}$, $\forall r \in \mathcal{R}$, $\forall q \in \{1, 2, \ldots, Q_n\}$. Constraint (14a) suggests the power limitation of the F-APs via the NOMA protocol and the RRHs in the cellular mode, constraints (14b) and (14c) imply the SE lower-bound constraints for FUEs and RUEs respectively, and constraints (14d) and (14f) imply the matching rule for subchannel reuse matching. Similarly, due to the 0–1 binary variables $\{\boldsymbol{x},\boldsymbol{y}\}$ and continuous variable $\boldsymbol{p}$ being mixed, problem (14) is a mixed nonlinear integer programming problem with NP-hardness, which cannot be solved tractably by the conventional dynamic programming method (Peng et al., 2015). However, when $\{\boldsymbol{x},\boldsymbol{y}\}$ is fixed, the objective function is convex with respect to the transmission power vector $\boldsymbol{p}$, and when $\boldsymbol{p}$ is fixed, the function of $\{\boldsymbol{x},\boldsymbol{y}\}$ is a Knapsack problem about $\{\boldsymbol{x},\boldsymbol{y}\}$. To solve this problem, considering the property of orderly decision-making process and cyclic dependency between the F-APs and RRHs, we adopt a Stackelberg game based approach including an NOMA-based power allocation algorithm and a subchannel reuse assignment algorithm. In the Stackelberg game model, the follower decides the subchannel allocation strategy according to the observation of leader's behavior, and in turn, the leader makes its strategy according to the estimation result of the follower. Specifically, as the leaders, the F-APs decide the transmission powers by a non-cooperative power allocation algorithm to maximize the EE of the FUEs according to the anticipated subchannel allocation strategy and prior knowledge of the response function of the RRHs. As the followers, the RRHs assign the subchannel reuse allocation to minimize the interference by a one-to-many matching game algorithm according to the practical transmission power strategy. We will characterize the power allocation algorithm and subchannel allocation algorithm for the F-APs and RRHs in the forthcoming subsections.

4.2.1 Leader: NOMA-based power allocation algorithm

First, we introduce the NOMA-based power allocation optimization problem (15) for the F-APs, where there exists a competition relationship among the FUEs that access the same F-AP, due to the limited power of one F-AP and quality of service of data rate for each FUE. Specifically, the optimization problem for the F-APs is

$$\max_{\boldsymbol{p}} \left. \frac{\sum\limits_{n=1}^{N}\sum\limits_{q=1}^{Q_n} g_{n,q}}{\sum\limits_{n=1}^{N}\sum\limits_{q=1}^{Q_n} p_{n,q} + P_{\text{cir}}^{\text{F}}} \right|_{\Omega(n)=r} \tag{15}$$

$$\text{s.t.} \quad \sum_{q=1}^{Q_n} p_{n,q} = p_n^{\max},\ p_{n,q} > 0,\ \forall n \in \mathcal{N}, \tag{15a}$$

$$\phi_n(x_n, p_n) > \phi_n^{\min}, \tag{15b}$$

$$t_n \leq t_n^{\max},\ \forall n \in \mathcal{N}, \tag{15c}$$

where $\Omega(n) = r$ represents the subchannel matching strategy of F-AP n and RRH r, $t_n = t_0 \cdot \sum_{l=1}^{L}\sum_{f=1}^{F} \xi_{n,l} Z_{l,f} \left(1 - \eta_{n,f}\right)$ is the delay constraint, and t_0 is the delay per file transmission. A non-convex problem needs to be addressed in relation to transmission power vector $\boldsymbol{p}$, and to solve it efficiently, we propose a non-cooperative game theoretic framework where the Nash equilibrium (NE) theorem is guaranteed to converge in problem (15) according to the following theorem (Shi et al., 2009):

Theorem 1 Consider a non-cooperative game $\mathcal{G}$

where the n^{th} F-AP aims at maximizing the data transmission rates of all users and the system EE, with respect to the choice of the set of the RRH $\mathcal{R}$ and transmission power $\boldsymbol{p}$; the utility function Ψ_n will be shown in Definition 2. The best-response dynamics (BRD) of game $\mathcal{G}$ always converges to an NE.

Note that the objective function in problem (15) can be classified as a nonlinear fractional programming problem (Dinkelbach, 1967). To solve this problem, we first define the optimal value as $\gamma^* = \frac{G(\boldsymbol{p}^*)}{P(\boldsymbol{p}^*)}$, which is nonlinear. G is the total transmission rate and P represents the total power of all players. Then, we have Theorem 2 (Ng et al., 2012):

Theorem 2 (Sub-problem equivalence) γ^* is achieved if and only if

$$\max_{\boldsymbol{p}} G(\boldsymbol{p}) = P(\boldsymbol{p}) \cdot \gamma^* \tag{16}$$

holds, where $G(\boldsymbol{p}) = \sum_{n=1}^{N}\sum_{q=1}^{Q_n} g_{n,q}$ and $P(\boldsymbol{p}) = \sum_{q=1}^{Q_n} p_{n,q}$.

Based on Theorem 2, it is proved that the original problem can be transformed into the convex problem (17) by the Dinkelbach method, which is proved in Appendix A.

$$\begin{aligned} &\max_{\boldsymbol{p}} G(\boldsymbol{p}) - \gamma^* P(\boldsymbol{p}) \\ &\text{s.t. constraints (15a) and (15b).} \end{aligned} \tag{17}$$

Note that γ^* is any feasible solution to problem (15) that satisfies constraints (15a) and (15b).

The proof is given in Appendix B.

We define an equivalent function $F(\gamma) = \max_{\boldsymbol{p}} G(\boldsymbol{p}) - \gamma P(\boldsymbol{p})$, and for all feasible $\boldsymbol{p}$ and γ, $F(\gamma)$ is a strictly and monotonically decreasing function of γ, and $F(\gamma) > 0$. Thus, this transformed sub-problem can be solved by the Lagrange dual decomposition method. The Lagrange dual function is expressed in the form of Eqs. (18) and (19):

$$\begin{aligned} L(\boldsymbol{p},\boldsymbol{\beta},\boldsymbol{\lambda}) = &\left[\sum_{n=1}^{N}\sum_{q=1}^{Q_n} g_{n,q} - \gamma\left(\sum_{n=1}^{N}\sum_{q=1}^{Q_n} p_{n,q} + P_{\text{cir}}^{\text{F}}\right)\right] \\ &+ \sum_{n=1}^{N}\left(\sum_{q=1}^{Q_n}\beta_{n,q}g_{n,q} - \phi_n^{\min}\right) \\ &+ \sum_{n}\lambda_n\left(p_n^{\max} - \sum_{q=1}^{Q_n} p_{n,q}\right), \end{aligned} \tag{18}$$

$$\begin{aligned} d(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \max_{\boldsymbol{p}} L(\boldsymbol{p},\boldsymbol{\beta},\boldsymbol{\lambda}) \\ &= \max_{\boldsymbol{p}}\left\{\left[\sum_{n=1}^{N}\sum_{q=1}^{Q_n} g_{n,q} - \gamma\left(\sum_{n=1}^{N}\sum_{q=1}^{Q_n} p_{n,q} + P_{\text{cir}}^{\text{F}}\right)\right]\right. \\ &\quad + \sum_{n=1}^{N}\left(\sum_{q=1}^{Q_n}\beta_{n,q}g_{n,q} - \phi_n^{\min}\right) \\ &\quad \left. + \sum_{n=1}^{N}\lambda_n\left(p_n^{\max} - \sum_{q=1}^{Q_n} p_{n,q}\right)\right\}. \end{aligned} \tag{19}$$

$\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n]^{\text{T}}$, where $\boldsymbol{\beta}_n = [\beta_{n,1}, \beta_{n,2}, \ldots, \beta_{n,Q_n}]^{\text{T}}$. $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_n]$ is the Lagrange multiplier vector. The dual optimization problem is reformulated as follows:

$$\begin{aligned} &\min_{\{\boldsymbol{\beta},\boldsymbol{\lambda}\}} d(\boldsymbol{\beta},\boldsymbol{\lambda}) \\ &\text{s.t. } \boldsymbol{\beta} \geq 0,\ \boldsymbol{\lambda} \geq 0. \end{aligned} \tag{20}$$

With the Karush–Kuhn–Tucker (KKT) conditions, the optimal power allocation is derived by

$$p_{n,q}^* = \left[\omega_{n,q}^* - \frac{1}{\delta_{n,q}}\right]^+, \tag{21}$$

where $\delta_{n,q} = \frac{x_{n,r}y_{r,n}|h_{n,q}|^2}{\sigma^2 + I_1 + I_2 + I_3}$, $[x]^+ = \max\{x, 0\}$, and the optimal coefficient $\omega_{n,q}^*$ is obtained as

$$\omega_{n,q}^* = \frac{1 + \beta_{n,q}}{\ln\left(2(\gamma B P_{\text{cir}}^{\text{F}} + \lambda_n)\right)}. \tag{22}$$

After substituting the optimal power allocation into the decomposed problem (15) and according to the sub-gradient based method, the update equations for the dual variables are derived (Peng et al., 2015). Here, for brevity, we have omitted the associated detailed derivations and refer readers to Boyd and Vandenberghe (2004) for similar processes. Based on this method, we propose the noncooperative game based power allocation algorithm (Algorithm 2). The complexity of Algorithm 2 is $O(IN)$, where I refers to the number of iterations.

4.2.2 Follower: subchannel allocation algorithm

We now introduce the subchannel allocation problem for RRHs. As the followers, RRHs determine the subchannel matching strategy with the F-APs under the decided power allocation strategy. Taking available accommodation capacity of one subchannel and the acceptable co-channel interference into consideration, we model subchannel allocation as a one-to-many matching game among the F-APs and RRHs, and the key parameters are expressed as

1. Players: F-APs in $\mathcal{N}$ and RRHs in $\mathcal{R}$.

2. Strategies: The strategy set of the F-APs is constituted by all members in the set of RRH $\mathcal{R}$, and vice versa.

3. Utility: The utility of the F-APs is EE and the utility of the RRHs is the suffered interference I_2 in Eq. (6).

Before the matching starts, we should explore the formation rule of and solution to stable matching (SM). To this end, we first explain some definitions to facilitate the analysis.

Definition 1 (Two-side matching) Consider that $\mathcal{N}$ and $\mathcal{R}$ are two disjoint sets. A one-to-many two-side matching $\varOmega$ is a mapping from $\mathcal{N}$ into $\mathcal{R}$ satisfying:

(a) $\varOmega(n) \in \mathcal{R}$, $\varOmega(r) \in \mathcal{N}$,

(b) $\varOmega(n) = r \Leftrightarrow \varOmega(r) = n$,

(c) $|\varOmega(n)| = 1$, $|\varOmega(r)| \leq M$.

Condition (a) explains that each player can match any member of the opposite set, condition (b) explains that if F-AP n matches RRH r, RRH r matches F-AP n certainly, and condition (c) explains that one F-AP can match one RRH while one RRH can accommodate M F-APs at most.

Definition 2 (Stable matching) Given a matching $\varOmega$, $\varOmega(n) \neq r$, $\varOmega(r) \neq n$, denote the utility of each F-AP as the data rate in Eq. (6) and mark it as $\varPsi_n$. The utility of each RRH is the interference I_2 in Eq. (6) and it is marked as $\varPsi_r$. If $\varPsi_n(\varOmega(n)\backslash\bar{r} \cup r) > \varPsi_n(\varOmega)$ where $\varOmega(n)$ is the current partner of F-AP n, it means that F-AP n prefers RRH r to $\bar{r}$. If $\varPsi_r(\varOmega(r)\backslash\bar{n} \cup n) > \varPsi_r(\varOmega)$ where $\varOmega(r)$ is the current partner of RRH r, it means that RRH r prefers F-AP n to $\bar{n}$. As long as both conditions are met, F-AP n is matched with RRH r successfully, and it is marked as (n, r).

Algorithm 2 Non-cooperative game based power allocation algorithm

```
Initialize: $p^*_{n,q}(0)$ according to the average power distribution scheme, $i = 1$
while $p^*_{n,q}(i) \neq 0$ do
  for $n \in \mathcal{N}$ do
    Calculate the optimal power value according to Eq. (21)
    if $p^*_{n,q}(i) \neq p^*_{n,q}(i-1)$ then
      Repeat
    end if
  end for
end while
```

In general, before the matching starts, each player should build a preference list (PL) which consists of all the strategy members in descending order according to the utility. Note that if and only if $x_{n,r} = y_{r,n} = 1$, F-AP n matches RRH r successfully.

Then, based on the previous analysis, we propose a matching algorithm (Algorithm 3), and the best response of the game is guaranteed to converge to an NE due to the finite number of players. The complexity of Algorithm 3 is $O\,(INR)$. Available frequency bandwidth is close to boundaries in industry,

Algorithm 3 One-to-many matching based subchannel allocation algorithm

```
Initialize: the preference lists of all players. Denote $S_u = \{1, 2, \ldots, N\}$ as the set of F-APs which do not match any RRH yet and $S_m = \varnothing$ as the initial set of F-APs which match certain RRHs
while $S_u \neq \varnothing$ do
  for $n \in \mathcal{N}$ do
    Each F-AP sends the request to the RRH with the highest ranking from set $\mathcal{R}$
  end for
  for $r \in \mathcal{R}$ which receives a request from $n$ do
    if $m_r = 0$ then
      RRH $r$ accepts the request directly. $m_r \leftarrow m_r + 1$. Mark $x_{n,r} = y_{r,n} = 1$ and remove $n$ from $S_u$ to $S_m$; prefer $n$ to its current candidate $n'$, if there exists any feasible FAP $n$
      if $0 < m_r < M$ then
        $r$ holds $n$ as a partner. Mark $x_{n,r} = y_{r,n} = 1$ and remove $n$ from $S_u$ to $S_m$. $m_r \leftarrow m_r + 1$
        if $m_r = M$ then
          RRH $r$ accepts F-AP $n$ if it ranks higher than any current candidate and removes the F-AP item $n'$ which ranks the lowest at that time. Mark $x_{n,r} = y_{r,n} = 1$, $x_{n',r} = y_{r,n'} = 0$, remove $n$ from $S_u$ to $S_m$, and remove $n'$ from $S_m$ to $S_u$
        else
          RRH $r$ rejects the request directly and removes RRH $r$ from $\mathrm{PL}_{\mathrm{F-AP}n}$
        end if
      end if
    end if
  end for
end while
```

and two proposed approaches aim at improving the utilization ratio in terms of current limited resources. They work out in many problems for multiple users with few channels.

4.2.3 RRA problem

Finally, the RRA algorithm (Algorithm 4) is addressed by combining Algorithms 2 and 3. In Algorithm 4, j is the index of iterations and J is the maximum number of iterations. ε_0 is a fixed value. The optimal solution can be achieved after several iterations, and the performance gains obtained, in terms of complexity, convergence, and accuracy, are significant. The complexity of Algorithm 4 is $O(JNR)$. The final NE identifies the optimal NOMA-based power allocation strategy and subchannel reusing solution under the optimal CP and user association strategy.

Algorithm 4 Radio resource allocation (RRA) algorithm

```
Initialize: the proper transmission power of each F-AP according to Algorithm 2, $j = 1$
while $j < J$ and $\varepsilon \le \varepsilon_0$ do
    for $n \in \mathcal{N}$ and $r \in \mathcal{R}$ do
        Determine the subchannel reuse assignment according to the matching algorithm (Algorithm 3)
    end for
    for each F-AP with the determined subchannel state do
        Calculate the power allocation policy according to Algorithm 2
    end for
    $j \leftarrow j + 1$ and calculate $\varepsilon = \dfrac{p_{n,q}(j) - p_{n,q}(j-1)}{p_{n,q}(j-1)}$
end while
```

5 Numerical results

In this section, numerical results are provided to validate the performance of the proposed algorithms.

5.1 Simulation setup

In the simulations, we consider 8 F-APs and 10 RRHs. The RRHs are located randomly around the BS which covers the area with a radius of 1 km. We consider that each F-AP can provide coverage with a radius of 100 m. All the FUEs and RUEs are uniformly distributed around the F-APs and RRHs, respectively. The transmission power of each F-AP is 20 dBm (Li QP et al., 2019), and the number of simulation snapshots is 1000. Unless specified, all parameters are the ones summarized in Table 3.

5.2 Performance of the cache placement algorithm

In this subsection, we simulate the impact of the price and the number of FUEs on the profit of the cloud and F-APs. In Fig. 3, specifically, we set $c_{nc} = 1$ dollar/file and evaluate the impact of $\theta_{n,f}$ on the algorithm performance by setting S_n^{C} to 100 and 200. The vertical coordinates represent the profit U or V measuring the revenue and expenditure of CP. From Fig. 3, we can first see that the optimal price of one file gets smaller with the larger profit when the storage capacity of the F-AP increases to allow a larger actual caching quality and income from FUEs. However, the profit of the cloud server diminishes with the increase of price of the file. It can be explained by the law of demand, where the quantity of goods actually consumed decreases with the price increasing.

Table 3 Simulation parameters

Parameter	Value	Parameter	Value
Number of subchannels	10	R	10
S_n^{C}	100, 200	N	5–12
p_r	20 dBm	$p_n^{\max}$	20 dBm
$P_{\mathrm{cir}}^{\mathrm{R}}$	0.1 W	$P_{\mathrm{cir}}^{\mathrm{F}}$	0.12 W
κ	0.1	L	20–150
M	2	c_{nc}	1–3
ε_0	10^{-4}	ζ	0.4

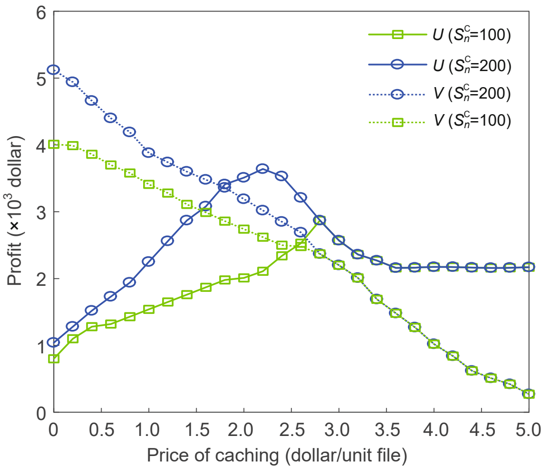

Fig. 3 Impact of the price on the profit of the cloud and F-APs

In Fig. 4, we set the storage capacity of the F-AP as $S_n^{\mathrm{C}} = 100$ under this simulation. Then we compare the maximum profit of the cloud server and the F-APs under the proposed algorithm versus the unit backhaul transmission cost, c_{nc}, with two other schemes: maximum hit rate (max) where for a given price $\theta_{n,f}$, each F-AP n chooses to cache the files that maximize the local cache hit rate, and random caching (random) where for a given price $\theta_{n,f}$, each F-AP n randomly selects the files to cache. It is clear that the profit of the cloud is the maximum with respect to other three algorithms. Furthermore, the profit of F-APs under our proposed scheme is larger than those under the max scheme and random scheme.

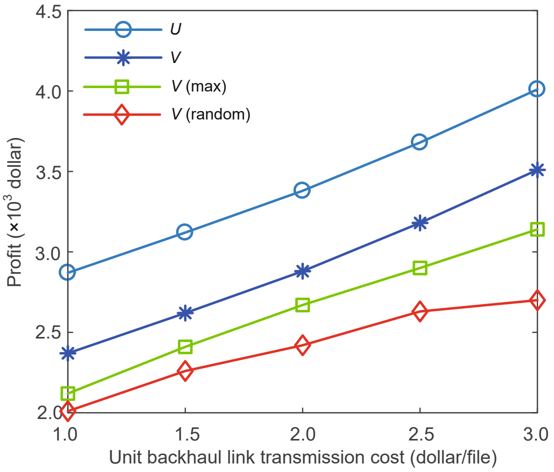

Fig. 4 Maximum profit vs. unit backhaul transmission cost

5.3 Performance of the radio resource allocation algorithm

In this subsection, the benefit of the SM algorithm is verified in Fig. 5 compared to the random matching (RM) scheme and exhaustive search matching (EM) scheme. Specifically, we can first see that for both the RM and SM schemes, the total latency decreases with the increase of the number of iterations. In addition, the SM scheme achieves better performance than RM because just the local channel state information is needed, and the RM scheme is worse than SM because there is a large probability to have a chaotic initial state, which leads to lower EE. The numerical results show that our proposed SM scheme can find the near-optimal value, which is very close to the optimal value obtained by the EM scheme with low complexity, and in particular, the difference of the SM and EM schemes for $Q = 10$ is only 2.6%.

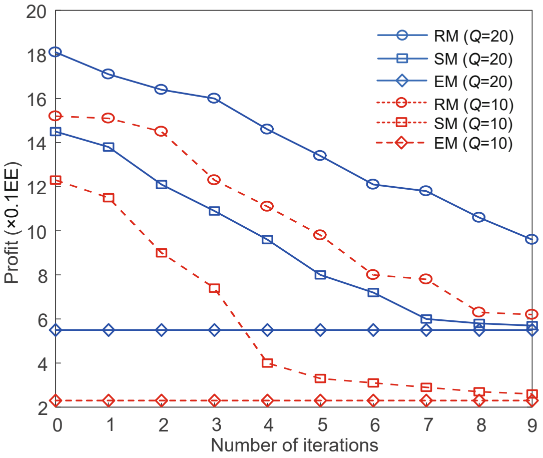

Fig. 5 Comparison among SM, RM, and EM

The benefit of the game-theoretic non-cooperative power allocation scheme is evaluated in Fig. 6. We can see that the network EE decreases with the increase of the maximum number of FUEs, suggesting tradeoff between the number of users in one NOMA group and the performance. We also see that the equal power allocation scheme achieves the worst performance because it does not consider dynamic channel conditions among different users, and that the greedy power allocation scheme ignores the efficiency but focuses on greedy throughput. The differences between these two schemes and our proposed scheme are distinct but become inconspicuous due to the increasing interference from subchannel reuse. However, our proposed NOMA-based power allocation scheme, which is dependent on the channel state, achieves the best performance gain.

Finally, we simulate the EE of the UEs in terms of the SE threshold of F-AP n. From Fig. 7, we can see that with the growth of the SE threshold of F-AP n, the EE decreases due to the decrease in the number of FUE users whose demand can be met by F-AP n. We also see that when more FUEs participate in transmission, EE increases largely due to the larger rate increment.

5.4 Performance of the network profit maximization algorithm

In this subsection, we adopt the CP, non-cooperative power allocation, and matching game

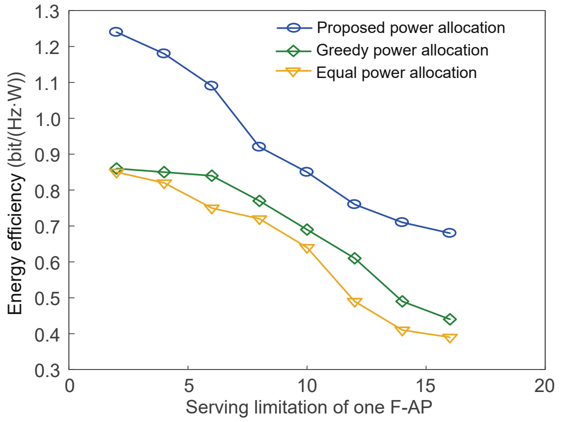

Fig. 6 Energy efficiency vs. serving limitation of one F-AP using NOMA

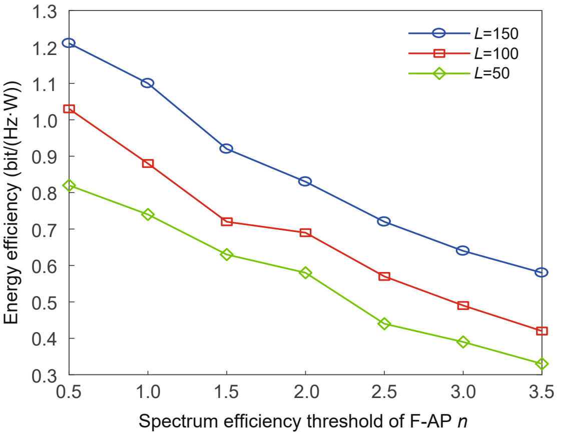

Fig. 7 Performance of energy-efficient power allocation vs. the spectrum efficiency threshold of F-AP n

schemes to verify the network profit performance. We set the price as $1, 3, \theta^*$, and the backhaul link cost is $c_{nc} = 1$ dollar/file. Two special schemes are simulated where 8 F-APs are with $S_n^{\mathrm{C}} = 100$ and $L = 50$, and 10 F-APs are with $S_n^{\mathrm{C}} = 100$ and $L = 100$. It is evident from Fig. 8 that with the increase of the number of F-APs, the network profit P increases significantly and the growth rate becomes larger. The most likely reason is that more F-APs participate in the file caching and obtain much income from FUEs and subchannel reuse. However, the different pricing schemes of files influence the profit under the special storage limitation due to the law of demand operative within the economics market.

6 Conclusions

In this study, we focus on the resource allocation in NOMA-based F-RAN with the F-APs of

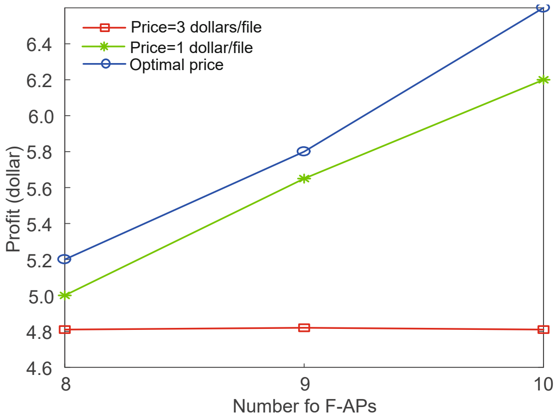

Fig. 8 Performance of network profit

edge caching, computing, and communication capabilities. Taking the heavy cloud burden and poor-quality backhaul link into consideration, we make edge caching for charging, transmission, and resource allocation in F-APs reasonably accessible for further network performance improvement. To this end, we define a new network performance indicator, namely network profit, which consists of the cost of caching and transmission energy efficiency during the file caching and transmission phases. Then we formulate an optimization problem for network profit maximization by jointly optimizing CP, pricing of files, and power and subchannel allocation. Two sub-problems are formulated to solve the NP-hard problem easily. Finally, we propose two algorithms based on iterative dynamic programming and game theory, and simulate them by comparison with some existing schemes in terms of efficiency and network performance.

Contributors

Xueyan CAO designed the research, processed the data, and drafted the paper. Xueyan CAO, Shi YAN, and Hongming ZHANG revised and finalized the paper.

Compliance with ethics guidelines

Xueyan CAO, Shi YAN, and Hongming ZHANG declare that they have no conflict of interest.

References

Bai WL, Yao T, Zhang HJ, et al., 2019. Research on channel power allocation of fog wireless access network based on NOMA. *IEEE Access*, 7:32867-32873. https://doi.org/10.1109/ACCESS.2019.2901740

Boyd S, Vandenberghe L, 2004. Convex Optimization. Cambridge University Press, Cambridge, UK.

Cao XY, Peng MG, Ding ZG, 2019. A game-theoretic approach of resource allocation in NOMA-based fog radio access networks. Proc 90th Vehicular Technology Conf, p.1-5.
https://doi.org/10.1109/VTCFall.2019.8891156

Dang T, Peng MG, 2019. Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks. *IEEE J Sel Areas Commun*, 37(7):1594-1607.
https://doi.org/10.1109/JSAC.2019.2916486

Deng RL, Lu RX, Lai CZ, et al., 2016. Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption. *IEEE Int Things J*, 3(6): 1171-1181.
https://doi.org/10.1109/JIOT.2016.2565516

Ding ZG, Fan PZ, Poor HV, 2016. Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. *IEEE Trans Veh Technol*, 65(8):6010-6023.
https://doi.org/10.1109/TVT.2015.2480766

Ding ZG, Fan PZ, Poor HV, 2019. Impact of non-orthogonal multiple access on the offloading of mobile edge computing. *IEEE Trans Commun*, 67(1):375-390.
https://doi.org/10.1109/TCOMM.2018.2870894

Dinkelbach W, 1967. On nonlinear fractional programming. *Manag Sci*, 13(7):492-498.
https://doi.org/10.1287/mnsc.13.7.492

Kong HB, Flint I, Wang P, et al., 2018. Fog radio access networks: Ginibre point process modeling and analysis. *IEEE Trans Wirel Commun*, 17(8):5564-5580.
https://doi.org/10.1109/TWC.2018.2846734

Li QP, Zhao JH, Gong Y, et al., 2019. Energy-efficient computation offloading and resource allocation in fog computing for Internet of Everything. *China Commun*, 16(3):32-41.

Li ZD, Wang Y, Liu M, et al., 2019. Energy efficient resource allocation for UAV-assisted space-air-ground Internet of remote things networks. *IEEE Access*, 7:145348-145362.
https://doi.org/10.1109/ACCESS.2019.2945478

Liu BH, Liu CX, Peng MG, et al., 2020. Resource allocation for non-orthogonal multiple access-enabled fog radio access networks. *IEEE Trans Wirel Commun*, 19(6):3867-3878.
https://doi.org/10.1109/TWC.2020.2978843

Ng DWK, Lo ES, Schober R, 2012. Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas. *IEEE Trans Wirel Commun*, 11(9):3292-3304.
https://doi.org/10.1109/TWC.2012.072512.111850

Park S, Simeone O, Shitz SS, 2016. Joint optimization of cloud and edge processing for fog radio access networks. *IEEE Trans Wirel Commun*, 15(11):7621-7632.
https://doi.org/10.1109/TWC.2016.2605104

Peng MG, Zhang KC, Jiang JM, et al., 2015. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Trans Veh Technol*, 64(11):5275-5287.
https://doi.org/10.1109/TVT.2014.2379922

Peng MG, Yan S, Zhang KC, et al., 2016. Fog-computing-based radio access networks: issues and challenges. *IEEE Netw*, 30(4):46-53.
https://doi.org/10.1109/MNET.2016.7513863

Peng MG, Quek TQS, Mao GQ, et al., 2020. Artificial-intelligence-driven fog radio access networks: recent advances and future trends. *IEEE Wirel Commun*, 27(2):12-13.
https://doi.org/10.1109/MWC.2020.9085257

Rai R, Zhu HL, Wang JZ, 2021. Performance analysis of NOMA enabled fog radio access networks. *IEEE Trans Commun*, 69(1):382-397.
https://doi.org/10.1109/TCOMM.2020.3028599

Shi Y, Wang JH, Letaief KB, et al., 2009. A game-theoretic approach for distributed power control in interference relay channels. *IEEE Trans Wirel Commun*, 8(6):3151-3161. https://doi.org/10.1109/TWC.2009.080831

Sun YH, Peng MG, Mao SW, et al., 2019. Hierarchical radio resource allocation for network slicing in fog radio access networks. *IEEE Trans Veh Technol*, 68(4):3866-3881.
https://doi.org/10.1109/TVT.2019.2896586

Xu C, Sheng M, Varma VS, et al., 2016. Wireless service provider selection and bandwidth resource allocation in multi-tier HCNs. *IEEE Trans Commun*, 64(12):5108-5124. https://doi.org/10.1109/TCOMM.2016.2613083

Yan S, Qi L, Zhou YC, et al., 2020. Joint user access mode selection and content popularity prediction in non-orthogonal multiple access-based F-RANs. *IEEE Trans Commun*, 68(1):645-666.
https://doi.org/10.1109/TCOMM.2019.2950215

Yang ZH, Xu W, Pan YJ, et al., 2018. Energy efficient resource allocation in machine-to-machine communications with multiple access and energy harvesting for IoT. *IEEE Int Things J*, 5(1):229-245.
https://doi.org/10.1109/JIOT.2017.2778766

Yao JJ, Ansari N, 2019. Joint content placement and storage allocation in C-RANs for IoT sensing service. *IEEE Int Things J*, 6(1):1060-1067.
https://doi.org/10.1109/JIOT.2018.2866947

Yu Y, Bu XY, Yang K, et al., 2019. Green large-scale fog computing resource allocation using joint benders decomposition, Dinkelbach algorithm, ADMM, and branch-and-bound. *IEEE Int Things J*, 6(3):4106-4117.
https://doi.org/10.1109/JIOT.2018.2875587

Zhai DS, Zhang RN, Cai L, et al., 2018. Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices. *IEEE Int Things J*, 5(3):1857-1868.
https://doi.org/10.1109/JIOT.2018.2816597

Zhang HJ, Qiu Y, Long KP, et al., 2018. Resource allocation in NOMA-based fog radio access networks. *IEEE Wirel Commun*, 25(3):110-115.
https://doi.org/10.1109/MWC.2018.1700326

Zhang JX, Zhang X, Wang WB, 2016. Cache-enabled software defined heterogeneous networks for green and flexible 5G networks. *IEEE Access*, 4:3591-3601.
https://doi.org/10.1109/ACCESS.2016.2588883

Zhang P, Peng MG, Cui SG, et al., 2022. Theory and techniques for "intellicise" wireless networks. *Front Inform Technol Electron Eng*, 23(1):1-4.
https://doi.org/10.1631/FITEE.2210000

Zhou YC, Yan S, Peng MG, 2019. Content placement with unknown popularity in fog radio access networks. Proc IEEE Int Conf on Industrial Internet, p.361-367.
https://doi.org/10.1109/ICII.2019.00068

Appendix A: Proof of the Dinkelbach solution

In Theorem 1, we use the Dinkelbach scheme to achieve the solution to the transformed problem, and prove its quasi-convexity as follows (Li ZD et al., 2019):

We define an equivalent function $Y(\boldsymbol{p}) = \max_{\boldsymbol{p}} G(\boldsymbol{p}) - \gamma P(\boldsymbol{p})$ with Theorem 2. Define two solutions as y_1 and y_2, and ascertain the function values for them to be $Y(y_1)$ and $Y(y_2)$, respectively. Then we define the derivative of function $\frac{\partial Y}{\partial \boldsymbol{p}}$ as G', and there are thus two situations in terms of two variables. If $y_1 \geq y_2$, $y_1 - y_2 \geq 0$, we just focus on whether Y' is positive or negative:

$$Y' = G' - \gamma^*, \tag{A1}$$

where γ^* is the assumed optimal solution. According to the proof in Appendix B, we can see that if $y_1 \geq y_2$, $Y(y_1) \leq Y(y_2)$, and $\gamma > y_2 \geq y_1$, then the derivative of function $Y'(y_2) \leq 0$ because the minimum value has not been achieved. Thus, the first-order condition of the quasi-convex function expression

$$\nabla Y(x)(y - x) \leq 0 \tag{A2}$$

holds where $x = y_1$, $y = y_2$. If $y_1 < y_2$, similar proofs can be obtained but are not mentioned here due to the space constraint for the paper.

Appendix B: Proof of problem equivalence

We prove the problem equivalence with two steps. First, we prove the sufficient condition and define the objective as $\gamma^* = \frac{G(\boldsymbol{p^*})}{P(\boldsymbol{p^*})}$, where $\boldsymbol{p^*}$ is the optimal power allocation policy. Then, it is evident that

$$\gamma^* = \frac{G(\boldsymbol{p}^*)}{P(\boldsymbol{p^*})} \geq \frac{G(\boldsymbol{p})}{P(\boldsymbol{p})}; \tag{B1}$$

based on this, we can derive

$$G(\boldsymbol{p}) - \gamma^* P(\boldsymbol{p}) \leq 0, \tag{B2}$$

$$G(\boldsymbol{p}^*) - \gamma^* P(\boldsymbol{p}^*) = 0. \tag{B3}$$

Therefore, $\max_{\boldsymbol{p}} G(\boldsymbol{p}) - \gamma^* P(\boldsymbol{p}) = 0$, and the sufficient condition is proved.

Second, the necessary condition should be proved. Suppose that $\tilde{\boldsymbol{p}}$ is the optimal policy and $G(\tilde{\boldsymbol{p}}) - \gamma^* P(\tilde{\boldsymbol{p}}) = 0$ is established. Thus, we have

$$G(\boldsymbol{p}) - \gamma^* P(\boldsymbol{p}) \leq G(\tilde{\boldsymbol{p}}) - \gamma^* P(\tilde{\boldsymbol{p}}) = 0. \tag{B4}$$

The above inequality can be derived as

$$\frac{G(\boldsymbol{p})}{P(\boldsymbol{p})} \leq \gamma^*, \ \frac{G(\tilde{\boldsymbol{p}})}{P(\tilde{\boldsymbol{p}})} = \gamma^*. \tag{B5}$$

Therefore, Theorem 2 is proved.