Resource Allocation in NOMA-Based Fog Radio Access Networks

Haijun Zhang, Yu Qiu, Keping Long, George K. Karagiannidis, Xianbin Wang, and Arumugam Nallanathan

Abstract

In the wake of growth in intelligent mobile devices and wide usage of bandwidth-hungry applications in mobile Internet, the demand for wireless data traffic and ubiquitous mobile broadband is rapidly increasing. Because of these developments, the research on next generation networks presents an accelerative tendency on a global scale. Edge computing is drawing lots of attention for reducing the time delay and improving the quality of service for networks. F-RAN is an emergent architecture that makes full use of edge computing and distributed storing capabilities in edge devices. In this article, we propose an architecture of NOMA-based F-RANs that has strong capability of edge computing and can meet the heterogeneous requirements in 5G/6G systems. NOMA with SIC is regarded as a critical multi-user access technology. In NOMA, more than one user can access the same time, code domain, and frequency resources. By assigning different power levels to multiple users and implementing SIC, multi-user detection can be achieved. In this article, we provide a description of the NOMA-based F-RANs architecture, and discuss the resource allocation in it. We focus on the power and subchannel allocation in consideration of using NOMA and edge caching. Simulation results show that the proposed NOMA-based F-RANs architecture and the resource management mechanisms can achieve high net utility for the RANs.

Introduction

With the increasing popularity of new applications and billions of mobile users, the past few decades have witnessed enormous growth of mobile and wireless networks. More intelligent phones, connected vehicles, and other smart Internet of Things devices require seamless and stable network connectivity, which results in a huge amount of traffic data. Meanwhile, this development presents a growing demand for broadband spectrum, high scalability, ultra-low latency, and less power consumption. In order to satisfy these requirements, many global research centers have proposed some novel technologies. During the past few years, cloud radio access networks (C-RANs) have drawn much attention and been regarded as a promising approach to handle the tremendous amount of devices [1]. In C-RANs, a crowd of remote radio heads (RRHs) are distributed within a particular geographical region, which are connected to a concentrated broadband unit (BBU) pool via high-bandwidth fronthaul links. Coordination among BBU pools improves resource utilization and reduces power consumption. Meanwhile, the fronthaul is commonly capacity and latency constrained in actuality, leading to dramatic reductions in spectral efficiency and energy efficiency performance of networks.

To compensate for the drawbacks of C-RANs, a new paradigm called fog computing has been introduced [2]. Cisco has proposed the concept of F-RANs, which integrates fog computing with RANs. Unlike the apparent centrality of C-RANs, F-RANs extend a high proportion of functions to the edge of the network, such as computing, storage, control, management, and application services [3]. Specifically, a great deal of signal processing and computing is performed in a distributed manner, and some data can be stored in the edge devices. Moreover, the real-time collaboration of radio signal processing and flexible cooperative radio resource management can be executed at the edge of the network. In F-RANs, the fog computing access points (F-APs) and fog user equipments (F-UEs) are equipped with caching. Besides, the F-APs can execute radio signal processing locally using their adequate computing capabilities and can manage their caching memories flexibly. These distinct features of F-RANs make a contribution to provide the superior user experience and improve network performance, including high mobility and ultra low latency.

Meanwhile, non-orthogonal multiple access (NOMA) has been validated as a promising multiple access mechanism for future RANs to meet the heterogeneous demands for low latency, massive connectivity, and high throughput [4]. By utilizing the power domain rather than the conventional time and frequency domains, NOMA can significantly improve network throughput. The essential idea of NOMA is that multiple users can share the same frequency resources and use different power levels [5]. Meanwhile, there is the inevitable expense of inter-user interference. Selective interference cancellation (SIC) is applied as an effective multi-user detection technology. Using NOMA with SIC, users with better channel conditions first successively subtract the messages of users with worse channel conditions and remove the interference before decoding their own messages [6]. Therefore, NOMA with SIC can be employed in F-RANs to ensure multiple users

Digital Object Identifier: 10.1109/MWC.2018.1700326

Haijun Zhang, Yu Qiu, and Keping Long are with the University of Science and Technology Beijing; George K. Karagiannidis is with Aristotle University of Thessaloniki; Xianbin Wang is with the University of Western Ontario; Arumugam Nallanathan is with Queen Mary University of London. Haijun Zhang and Keping Long are the corresponding authors.

1536-1284/18/$25.00 © 2018 IEEE

with simultaneous transmissions in complex wireless environments. A many-to-many subchannel matching algorithm was proposed and approached the performance of the upper bound and greatly outperformed orthogonal frequency-division multiple access (OFDMA) networks. A tree search scheme was introduced as a low computational complexity power assignment method for NOMA with a SIC receiver in [7]. From the standpoint of energy aspects, a cooperative NOMA transmission protocol was proposed in [8], in which nearby NOMA users were regarded as energy harvesting user relays for forwarding messages to far-off NOMA users. To the best of the authors' knowledge, it is the first attempt to investigate the application of NOMA in an F-RAN environment and resource allocation with interference management in the literature.

In this article, we investigate a solution of resource allocation to improve the network performance of NOMA-based F-RANs. First, we propose the network architecture of NOMA-based F-RANs. Then we study the resource management, which includes the power and subchannel allocation. The power allocation problem is modeled as a non-cooperative game. For subchannel allocation, we study the many-to-many two-sided matching algorithm. Moreover, we show that the proposed resource management mechanisms enhance the net utility of NOMA-based F-RANs.

System Architecture

In Fig. 1, we present the architecture of NOMA-based F-RANs. In this proposed network, the functions of the control plane are completed in the macro remote radio heads (MRRHs), rather than the BBU pool in C-RANs. The BBU pool will mainly provide centralized storage and communications in NOMA-based F-RANs. The MRRHs are interfaced to a BBU pool through backhaul links. All F-APs are connected to the BBU pool by fronthaul links. The F-APs are evolved from traditional RRHs, which play an important role in NOMA-based F-RANs. Unlike the centralized data storage in C-RANs, a substantial part of data are distributed in F-APs and some F-UEs as edge caching. Not only is a certain amount of caching assigned to F-APs, but also radio signal processing and radio resource management can be performed locally in F-APs. It is quite possible that F-UEs can directly connect to F-APs to get the desired content instead of establishing complex transmission links with the core network. Similarly, F-UEs can download data from neighboring F-UEs through device-to-device (D2D) technology. And when the distance of two possible paired F-UEs is so far that they cannot directly connect with each other, a third F-UE will play the role of relay to forward the information.

Edge Caching Utilization

As above, there are substantial data cached at the edge of the network, such as F-APs and F-UEs. Note that what is stored in the F-APs and F-UEs is not arbitrarily assigned. How to choose appropriate content for edge caching has a profound effect on network performance, especially for time delay and energy consumption. With the upward popular tendency of location-based mobile applications, enormous amounts of information may emerge anytime and anywhere. If all these surging data streams are transmitted to the centralized BBU pool, it will push the fronthaul links to their capacity limits. It is found that there is great relevance of these data flows among F-UEs in close physical proximity. At the same time, some social applications make the close-range F-UEs exchange data traffic more frequently than other distant F-UEs. Besides, F-UEs from the same social network or enjoying the same social activity can require the same data over the downlink. The state of network applications inspires us to effectively set cache placement and use the finite capacity of edge caching. Actually, F-APs and F-UEs store the most popular or relevant contents as edge caching until there is no space for storage instead of the contents being randomly assigned with equal probabilities. In these cases, the requested services can be completed locally by edge caching of the popular contents. Consequently, F-UEs have no need to interface with the BBU pool every time they request data. Similarly, if a crowd of F-UEs have highly relevant data to be recorded in the cloud, there is no need for all F-UEs to upload traffic individually. Then the contents of their common interests will be stored as edge caching in only a single F-UE. Benefitting from edge caching, the amount of system data supported by both the fronthaul and backhaul links can be significantly decreased. This caching utilization in NOMA-based F-RANs can reduce the long transmission time latency and heavy burden on the fronthaul and the BBU pool, which are the two hardest issues in C-RANs.

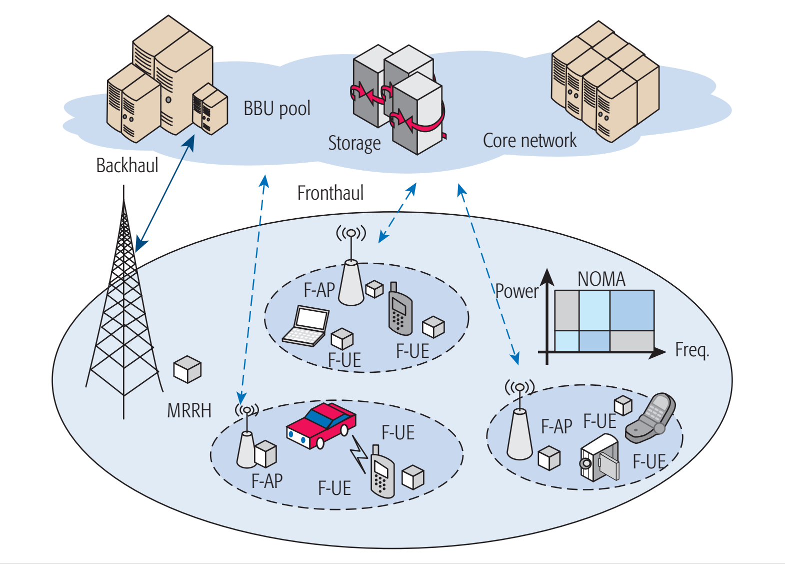

FIGURE 1. A NOMA-based F-RANs architecture.

NOMA Protocol

Figure 2 shows the typical NOMA scheme for a scenario of three users. There are three users who are allocated the same subchannel: $x_1(t)$, $x_2(t)$, and $x_3(t)$. Moreover, $x_1(t) > x_2(t) > x_3(t)$. The superposition signal at the SIC receiver can be denoted as X, including all users' signals about channel interference and the additive white Gaussian noise (AWGN). Before the first stage of user detection, the SIC receiver will put these users in increasing order of the channel response normalized by noise (CRNN). Consider that three users share the same subchannel with CRNN order $CRNN_1 \leq CRNN_2 \leq CRNN_3$. Depending on this order, user 1 with the

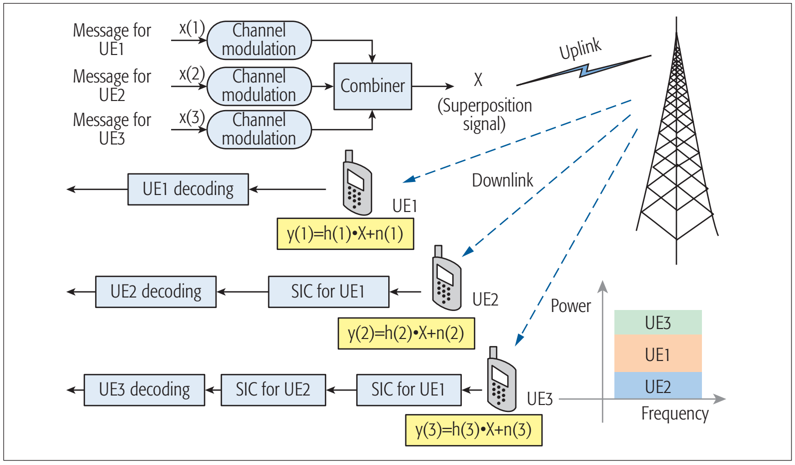

FIGURE 2. Basic NOMA scheme for three users. In downlink, SIC technology is applied to the user receivers. Three users are allocated to the same subchannel with the channel response normalized by noise (CRNN) order $CNRR_3 \geq CNRR_2 \geq CNRR_1$.

smallest CRNN can first decode from the superposition signal. Then the interference from user 1 in the poorest channel condition can be correctly removed by user 2 in better channel condition. Similarly, user 3 can successively subtract the interference from users 1 and 2, and then obtain its own message. Based on the increasing order of CRNNs, users 1, 2, and 3 can successfully decode and cancel the interference symbols in turn.

Furthermore, the performance of networks can be further improved by using NOMA technology in the D2D communication scenario. If a group of F-UEs in D2D mode transmit various contents, multiple bandwidth channels would be occupied in general. However, NOMA makes it possible that these receivers are served simultaneously in only one channel. For example, three F-UEs are considered in a D2D group, and what they require are video, audio, and text messages, respectively. Suppose the video and audio F-UEs have good channel conditions; they can receive accurate signals and achieve high data rates by SIC technology. Specifically, the impact of the partners in a D2D group can be completely eliminated after two or three times. As to an F-UE requiring text, the service can be successfully provided just with low data rates. Hence, the strong co-channel interference can be safely ignored. The integration with NOMA significantly enhances the performance of systems, especially in both SE and massive connectivity.

Resource Allocation in NOMA-Based F-RANs

With the growing popularity of big data and the Internet of Things in fog computing, the provided services and resources are becoming more complicated than ever before. What we should consider is not only the unique resource heterogeneity, but also the resource limitations, locality restrictions [9], and dynamic nature of resource demand. Accordingly, to efficiently allocate the heterogeneous resources to achieve the optimal network utility, edge caching [10] should be taken into account.

We focus on resource allocation in terms of maximizing the sum data rate of the system. In our consideration, all the F-APs have full knowledge of channel state information and share spectrum resources with the central MRRH. Notice that the spectrum sharing framework will bring about both co-tier and cross-tier interference. With adopting the NOMA protocol [11, 12], F-UE multiplexing in the power domain is exploited, which means different F-UEs are served with different power values. Meanwhile, due to one subchannel being occupied by a subset of F-UEs, the signal of an F-UE will cause interference to the other F-UEs allocated to the same subchannel. Therefore, the received signal at the F-UE includes the desired signal, the AWGN, co-tier interference (from the F-UEs in other F-APs reusing the same subchannel), cross-tier interference (from the MRRH), and the interference from the NOMA F-UEs served by the same F-AP.

In this article, power allocation and subchannel allocation are coupled with each other in terms of maximizing the sum data rate. First, F-APs allocate the transmission power to the F-UEs over each subchannel, depending on the non-cooperation game scheme. Then the subchannel allocation can be modeled as a many-to-many two-sided matching problem, and solved by matching theory, as discussed in detail in the next subsection.

Power Allocation in NOMA-Based F-RANs

In the existing literature, game theory has been extensively used for interference mitigation in networks [13]. We define the maximum achievable data rate of the F-UE as the optimization function of each F-UE in the system. The optimization of resource management can be formulated as a non-cooperative resource allocation game. Accordingly, all F-UEs in each F-AP only depend on their own interests to achieve their own greatest benefit, acting as selfish players in competition to satisfy themselves. There is huge power consumption when every F-UE adopts its maximal transmit power. Thus, the optimization function of overall F-APs can be modeled as the maximization of object function under the constraint of the maximal transmit power of each F-UE.

Pricing Function of Interference: In consideration of the fundamental role of MRRH in providing ubiquitous connection to achieve seamless coverage, the MRRH should be strictly protected from the impact by F-AP deployment. An interference threshold will be proposed to represent the maximum tolerable interference level for MRRH caused by massive numbers of F-UEs. At the same time, this density of F-UEs would interfere with MRRH, naturally leading to a significant decrease in total network throughput and affecting the capability of MRRH. It is essential to propose the pricing function to constrain the interference to MRRH. This pricing function can be applied to manage the interference created by F-UEs to MRRH, which is proportional to the transmit power of each F-UE.

Reward Function of Edge Caching: Similarly, the reward function corresponding to a certain edge caching strategy in NOMA-based F-RANs should be introduced to represent the benefit from the storage resources at the edge of networks. When the contents stored in F-APs or F-UEs is requested, the use of edge caching can alleviatie bandwidth requirements and reduce delay. In this article, we choose the alleviation of backhaul bandwidth as the reward of edge caching. By means of the coefficient, which is about the utilization of edge caching, we can compute a certain degree of compensation from the finite edge caching in NOMA-based F-RANs.

Utility Function: According to the assumption above, we propose an interference-aware resource power allocation scheme for the downlink of co-channel deployed F-APs. The optimization function under the proposed resource management scheme can be achieved by maximizing the net utility function of each F-UE, which is defined as the maximum achievable data rate of the F-UE minus the pricing function of the interference and adds the reward function of the edge caching offered from networks.

At fixed points, Nash equilibrium (NE) has been proven unique and existing on each individual subchannel in the non-cooperative resource allocation game. The power allocated for each F-UE is restricted between zero and the maximum power. We know that there are serious cross-tier interference and co-tier interference, when an F-UE wants to optimize its utility by being equipped with the maximum transmission power, making the strategy far from Pareto-optimality. Starting with the smallest available power value, we can update the transmission power of each F-UE successively through the designed iterative algorithm to find the optimal value.

Subchannel Matching in NOMA-Based F-RANs

To deal with the allocation issue between the F-UEs and the subchannels, we propose an approach depending on matching theory. First, we assume a set of *M* F-UEs from *K* F-APs and the set of *N* subchannels. The players in these two disjoint sets are rational and selfish, and intend to obtain the peak value of their own benefit. When SC_n is assigned to $F-UE_{k,m}$, it means that SC_n and $F-UE_{k,m}$ are matched with each other.

For a given subchannel, F-UE *i* desires to decode and remove the interference from F-UE *j* from the superimposed signals by the SIC technique. It is noticed that the application of SIC technology may lead to a certain complexity. The complexity at the receiver will increase along with the growing number of F-UEs reusing each subchannel. Hence, we should make a reasonable constraint to limit the maximum number of F-UEs occupying the same subchannel at the same time. Thus, the decoding complexity at the receiver can be tolerated. Also, there is significant improvement in the hardware complexity and processing delay.

About each player, the player of the other set has different preference. Note that the preferences of each player do not rely on the other players' behaviors. Then, to better demonstrate the dynamic matching process of all players in competing with each other and making decisions, we propose the preference lists of the F-UEs $Pre(F-UE)$ and subchannels $Pre(SC)$, respectively. The solution of the matching game is an allocation strategy between F-UEs and subchannels, which is based on the preference lists to satisfy each player's preference.

If $F-UE_{k,m}$ has a preference for SC_i rather than SC_j, it implies that $F-UE_{k,m}$ can achieve higher channel gain when assigned to SC_i than to SC_j. Similarly, we define *p* and *p*′ as the pair of F-UEs. When SC_n prefers the F-UEs in *q* to F-UEs in *p*′, the F-UEs in set *p* can provide higher profit than F-UEs in set *p*′ on SC_n. Also, the preference of each subchannel over different subsets of the set of F-UEs can be proposed. On the basis of that list, it is clearly which set of F-UEs works with better performance than the other one. Some matching schemes of different characteristics of preferences have been studied for various scenarios in [14, 15]. In this article, we formulate the subchannel assignment problem as a many-to-many matching problem, depending on the the preference lists.

As to many-to-many two-sided matching, there are some key definition:

- As to each user, it can match a subset of a subchannel.
- As to each subchannel, a subset of F-UEs can be matched.
- The number of F-UEs can be assigned on each subchannel is limited to *q*.
- Matching μ is a mapping that assigns at least two different F-UEs (but less than or equal to *q*) to each subchannel and multiple subchannels to each F-UE.

It is noted that the formulated matching model is more complicated than the traditional matching models. As above, in our model any F-UEs of each set can be matched with any subset of the subchannel sets. Although the finite F-UEs can be multiplexed on the same subchannel, the quantity of potential matching pairs may be very large when the F-UE set involves many members. Besides, we should consider the correlative dependence of F-UE combinations that are allocated to the same subchannel. Each subchannel should choose an ideal pair of F-UEs to match with, resulting in the matching process being more complicated even if we do not take the power allocation into account. Thus, a suboptimal matching scheme is proposed for subchannel assignment as follows for reducing the complexity.

In the matching procedure, at each round, each F-UE will send the matching request to its most preferred subchannel. Basing the list of subchannels ordered by decreasing channel gains, $F-UE_{k,m}$ will find the first non-zero entry. Then $F-UE_{k,m}$ will send a matching request to the corresponding subchannel. When the number of assigned F-UEs on the subchannel is less than *q*, the subchannel can accept the matching request directly. However, if there are *q* F-UEs allocated to the subchannel, we should compare the value of net utility provided by the different subsets of F-UEs. This subchannel will accept the subset of F-UEs that satisfies maximum net utility, or else the request will be rejected. The whole matching process is repeated until no available F-UE is left to be allocated. Then the assigned F-UEs and the corresponding subchannels in the preference list are set to zero.

Simulation Results

This section evaluates the performance of the proposed resource allocation algorithms for NOMA-based F-RANs. In the simulations, the F-APs and F-UEs are randomly distributed in the central MRRH coverage area as illustrated in Fig. 3. We set the radius *R* of MRRH and each F-AP to be 500 m and 10 m, respectively. The minimum distance from MRRH to F-AP and F-UEs belonging to MRRH is 300 m and 50 m, respectively, and the minimum distance between F-APs is 40 m. The bandwidth is limited to 5 MHz. We set the peak power of each F-AP as 41 dBm and noise power spectral density as –174 dBm/Hz. In the NOMA scheme, we restrict that only *q* F-UEs are allocated to each subchannel for the sake of reducing demodulating complexity at the SIC receiver side.

In Fig. 4, the performance of NOMA-based F-RANs is good with increasing number of F-APs,

We know that there are serious cross-tier interference and co-tier interference,when an F-UE wants to optimize its utility by being equipped with the maximum transmission power, making the strategy far from Pareto-optimality. Starting with the smallest available power value, we can update the transmission power of each F-UE successively through the designed iterative algorithm to find the optimal value.

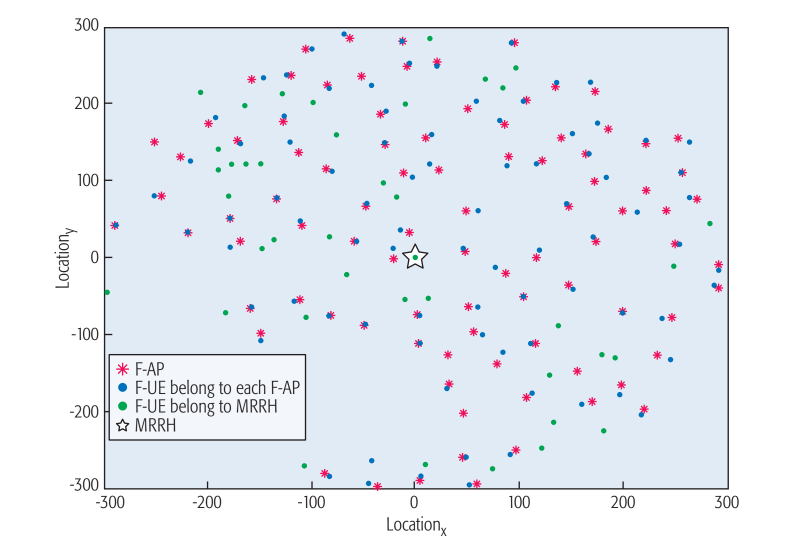

FIGURE 3. The system design of distance setting.

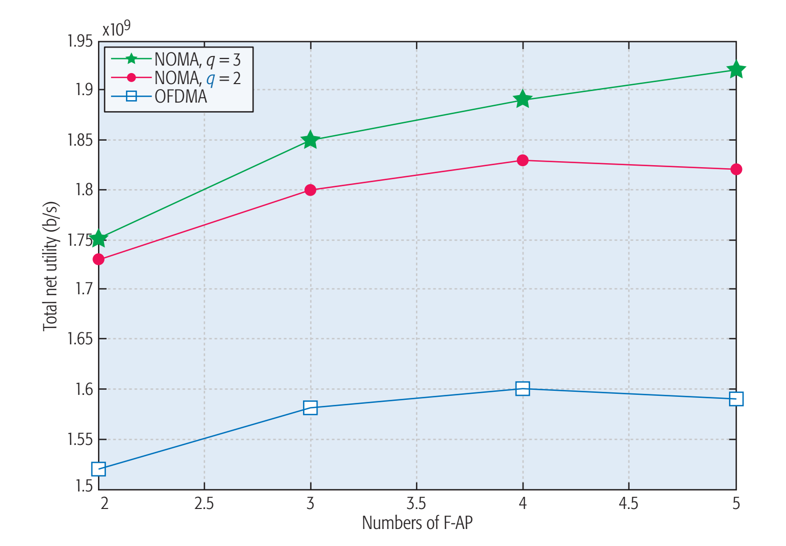

FIGURE 4. Total net utility of NOMA-based FRAN vs. different numbers of F-APs.

and much higher net utility of the system is achieved than that in the OFDMA scheme. In other words, the scheme we propose can guarantee good service even in dense deployed networks. But there is little reduction when the number of F-APs goes from 4 to 5. It motivates us to find the optimal number of F-APs in the NOMA-based F-RANs to achieve the best net utility. In Fig. 5, we compare the performance of the proposed scheme for NOMA-based F-RANs with the trantional schemes of OFDMA, where q is set as 2 and 3. As the number of F-UEs per F-AP grows, the net utility of NOMA is much higher than that of OFDMA due to the multiuser diversity gain. For example, when the number of F-UEs per F-AP is 30 and q = 2, the net utility gain of the proposed resource allocation scheme is 29 percent more than that of a conventional OFDMA scheme. This is because each subchannel can only be allocated to one F-UE in OFDMA systems. In other words, the F-APs do not make the best of the finite frequency resources. We find that when q = 3, the proposed scheme performs better than the case in which q = 2. This can be attributed to the fact that an F-AP has more possible selections over the set of F-UEs to be allocated to each subchannel.

Future Works

There is still some room for further investigation in NOMA-based F-RANs. Edge cache placement strategy should be seriously designed to provide the more flexible transmission opportunities to F-UEs. The edge caching reward can be widely exploited in the content delivery process, while the edge caching space at each F-AP and F-UE are restricted and practically very small. At the same time, there are many crucial factors that should be taken into account and jointly optimized. For example, the cost of caching hardware makes a request to the hardware device about the storage capacity, which also increases the economic cost. Besides, the cache hit ratio and the number of data requests play important roles in optimal radio resource usage.

Trustworthiness and security are other key aspects we should investigate further. F-RANs can support more flexible network operations in distributed manners. In consideration of the proximity to F-UEs and locality on the edge of networks, the massive nodes in F-RANs can often operate as the first node of access control. Thus, the network's edge nodes must be equipped with the right to monitori and supervise privacy-sensitive data. It is essential that we should take advantage of F-RANs to enhance network security instead of being restricted by its distributed characteristic. Meanwhile, multiple-input multiple-output, millimeter-wave communications, network functions virtualization, and other key technologies can be combined with NOMA-based F-RANs to achieve better performance. In the future we plan to work on solving some of these challenges.

Conclusion

In this article, the F-RANs architecture for 5G networks is proposed enhanced with NOMA. The NOMA-based F-RANs architecture takes full advantage of the edge of networks. The power and subchannel allocation problems in NOMA-based F-RANs are studied, with the aim of maximizing the net utility while considering the co-channel interference. For solving the power allocation problem, resource optimization mechanisms are proposed under the non-cooperation framework. We present a suboptimal algorithm based on formulating the subchannel allocation problem as a many-to-many two-side matching game. The numerical results demonstrate that the NOMA-based F-RANs architecture has more potential benefits in terms of net utility compared to conventional OFDMA networks.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61471025, 61771044), the Young Elite Scientist Sponsorship Program by CAST (2016QNRC001), the Research Foundation of Ministry of Education of China & China Mobile (MCM20170108), Beijing Natural Science Foundation (L172025), and the Fundamental Research Funds for the Central Universities(FRF-GF-17-A6, RC1631).

References

[1] S. C. Hung *et al.*, "Architecture Harmonization between Cloud Radio Access Networks and Fog Networks," *IEEE Access*, vol. 3, 2015, pp. 3019–34.

[2] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *MCC Wksp. Mobile Cloud Comp.*, 2012, pp. 13–16.

[3] H. Zhang *et al.*, "Fog Radio Access Networks: Mobility Management, Interference Mitigation and Resource Optimization," *IEEE Wireless Commun.*, 2017.

[4] P. Xu *et al*, "NOMA: An Information Theoretic Perspective," 2015; http://arxiv.org/abs/1504.07751.

[5] F. Fang *et al.*, "Joint User Scheduling and Power Allocation Optimization for Energy Efficient NOMA Systems with Imperfect CSI," *IEEE JSAC*, 2017.

[6] F. Fang *et al.*, "Energy-Efficient Resource Allocation for Downlink Non-Orthogonal Multiple Access Network," *IEEE Trans. Commun.*, vol. 64, no. 9, Sept. 2016, pp. 3722–32/

[7] Z. Ding *et al.*, "On the Performance of Nonorthogonal Multiple Access in 5G Systems with Randomly Deployed Users," *IEEE Signal Processing Lett.*, vol. 21, no. 12, Nov. 2014, pp. 1501–05.

[8] H. Zhang *et al.*, "Secure Communications in NOMA System: Subcarrier Assignment and Power Allocation," *IEEE JSAC*, 2017.

[9] X. Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014.

[10] X. Wang *et al.*, "D2D Big Data: Content Deliveries over Wireless Device-to-Device Sharing in Realistic Large-Scale Mobile Networks," *IEEE Wireless Commun.*, vol. 25, no.1, Feb. 2018, pp. 1–10.

[11] Y. Saito *et al.*, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," *IEEE 77th VTC*, Dresden, Germany, June. 2013, pp. 1–5.

[12] A. Li, A. Harada, and H. Kayama, "A Novel Low Computational Complexity Power Assignment Method for Non-Orthogonal Multiple Access Systems," *IEICE Trans. Fundam. Electron. Commun. Comp. Sci.*, vol. E97-A, no .1, Jan. 2014, pp. 57–68.

[13] H. Zhang *et al.*, "Incomplete CSI Based Resource Optimization in SWIPT Enabled Heterogeneous Networks: A Non-Cooperative Game Theoretic Approach," *IEEE Trans. Wireless Commun.*, 2017.

[14] A. Roth and M. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling And Analysis*, Cambridge Univ. Press., 1992.

[15] S. Bayat *et al.*, "Distributed User Association and Femtocell Allocation In Heterogeneous Wireless Networks," *IEEE Trans. Commun.*, vol. 62, no. 8, Aug. 2014.

Biographies

Haijun Zhang [M'13, SM'17] is currently a full professor at the University of Science and Technology Beijing, China. He was a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of British Columbia (UBC), Vancouver Campus, Canada. He serves as an Editor of *IEEE Transactions on Communications* and *IEEE 5G Tech Focus*, and serves/has served as a Leading Guest Editor for *IEEE Communications Magazine* and *IEEE Transactions on Emerging Topics in Computing*. He serves/has served as General Co-Chair of GameNets '16, Symposium Chair of IEEE GLOBECOM '19, TPC Co-Chair of INFOCOM '18 Workshop IECCO, General Co-Chair of ICC '18/ICC '17/GLOBECOM '17 Workshop on UDN, and General Co-Chair of GLOBECOM '17 Workshop on LTE-U. He received the IEEE ComSoc Young Author Best Paper Award in 2017.

Yu Qiu received her B.S. and M.S. degrees from Beijing University of Chemical Technology, China, in 2015 and 2018, respectively. She is currently pursing a Ph.D. degree at the University of Science and Technology Beijing, China. Her research interests include B5G networks, mobility management, and resource allocation.

Keping Long [SM'06] received his M.S. and Ph.D. degrees from UESTC in 1995 and 1998, respectively. He worked as an associate professor at BUPT. From July 2001 to November 2002, he was a research fellow in the ARC Special Research Centre for Ultra Broadband Information Networks (CUBIN) at the University of Melbourne, Australia. He is now a professor and dean at the School of Computer and Communication Engineering (CCE), USTB. He is a member of the Editorial Committee of *Sciences in China Series F* and *China Communications*. He is also a TPC and ISC member for COIN, IEEE IWCN, ICON, and APOC, and Organizing Co-Chair of of IWCMC '06, TPC Chair of COIN '05/'08, and TPC Co-Chair of COIN '08/'10, He was awarded the National Science Fund Award for Distinguished Young Scholars of China in 2007 and selected as the Chang Jiang Scholars Program Professor of China in 2008. He has published over 200 papers, 20 keynotes, and invited talks.

George K. Karagiannidis [M'96, SM'03, F'14] is currently a professor in the Electrical Computer Engineering Department and director of the Digital Telecommunications Systems and Networks Laboratory, Aristotle University of Thessaloniki, Greece. He is an author or co-author of more than 450 technical papers published in scientific journals and presented at international conferences. He is also an author of two books. He has been involved as General Chair, Technical Program Chair, and member of Technical Program Committees for several IEEE and non-IEEE conferences. In the past, he was an Editor of *IEEE Transactions on Communications*, a Senior Editor of *IEEE Communications Letters*, an Editor of the *EURASIP Journal of Wireless Communications Networks*, and several times a Guest Editor of *IEEE Selected Areas in Communications*. From 2012 to 2015 he was the Editor-in Chief of *IEEE Communications Letters*. He is one of the most highly cited authors across all areas of electrical engineering, recognized as a 2015, 2016, and 2017 Web-of-Science Highly Cited Researcher.

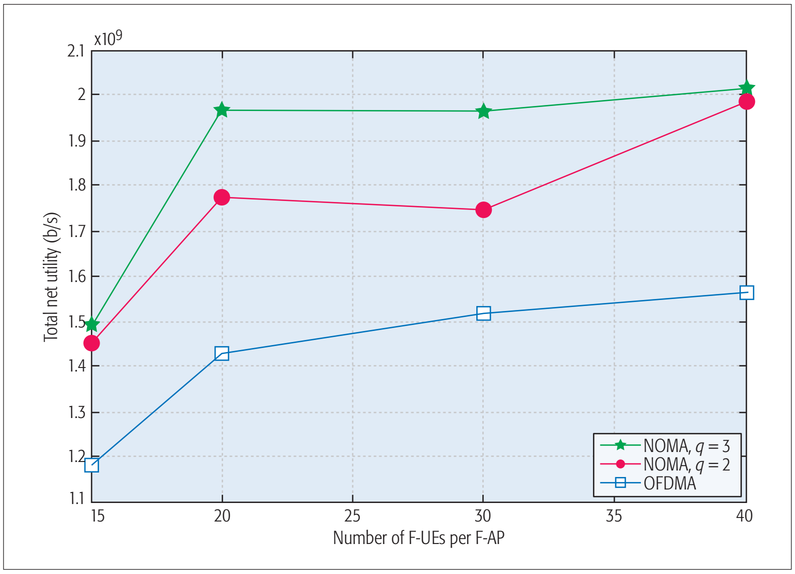

FIGURE 5. Total net utility of NOMA-based F-RAN vs. different numbers of F-UEs per F-AP.

Xianbin Wang [S'98, M'99, SM'06, F'17] is a professor and Canada Research Chair at Western University, Canada. He received his Ph.D. degree in electrical and computer engineering from National University of Singapore in 2001. He has over 300 peer-reviewed journal and conference papers, in addition to 26 granted and pending patents and several standard contributions. He is an IEEE Distinguished Lecturer. He has received many awards and recognitions, including the CRC President's Excellence Award, Canadian Federal Government Public Service Award, Ontario Early Researcher Award and five IEEE Best Paper Awards. He currently serves as an Editor/Associate Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Broadcasting*, and *IEEE Transactions on Vehicular Technology*. He was also an Associate Editor for *IEEE Transactions on Wireless Communications* between 2007 and 2011, and *IEEE Wireless Communications Letters* between 2011 and 2016. He has been involved in a number of IEEE conferences including GLOBECOM, ICC, VTC, PIMRC, WCNC, and CWIT, in different roles such as Symposium Chair, tutorial instructor, Track Chair, Session Chair, and TPC Co-Chair.

Arumugam Nallanathan [S'97, M'00, SM'05, F'17] has been a professor of wireless communications in the School of Electronic Engineering and Computer Science at Queen Mary University of London since September 2017. He was with the Department of Informatics at King's College London from December 2007 to August 2017, where he was a professor of wireless communications from April 2013 to August 2017. He has published more than 350 technical papers in scientific journals and international conferences. He is a co-recipient of the Best Paper Award presented at IEEE ICC 2016 and IEEE ICUWB 2007. He is an IEEE Distinguished Lecturer. He was selected as a Web of Science (ISI) Highly Cited Researcher in 2016. He is an Editor for *IEEE Transactions on Communicaitons* and *IEEE Transactions on Vehicular Technology*. He was an Editor for *IEEE Transactions on Wirelss Communcations* (2006–2011), *IEEE Wireless Communications Letters*, and *IEEE Signal Processling Letters*. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of IEEE Communications Society and Technical Program Chair and member of Technical Program Committees for numerous IEEE conferences. He received the IEEE Communications Society SPCE outstanding service award 2012 and IEEE Communications Society RCC outstanding service award 2014.