An Optimal Cache Resource Allocation in Fog Radio Access Networks

Sovit Bhandari[1], Hong Ping Zhao[1], Hoon Kim[1], John M. Cioffi[2]

[1]IoT and Big-data Research Center, Department of Electronics Engineering,
Incheon National University, South Korea
[2]Department of Electrical Engineering, Stanford University, USA
sovit198@gmail.com, {hongpieng614, hoon}@inu.ac.kr, cioffi@stanford.edu

Abstract

Fog radio access networks (F-RANs) are the key promising architecture for next-generation wireless networks. This paper addresses the cache resource allocation problem in multi fog access points (F-APs) of F-RANs system. A linear programming (LP)-based problem is formulated for optimal allocation of the cache resource of multiple F-APs to the user equipments (UEs) to maximize the delivered contents while providing the minimum user data volume request and transmission delay constraints. Simulation results are shown to validate the proposed scheme.

Keywords: Cache resource allocation, Fog access point, Optimal, Linear programming

1 Introduction

With the rapid advancement of wireless technology, mostly smartphones, we are able to witness the abrupt increment of internet traffic over the internet. It is reported that the global mobile data traffic (mainly video) will increase by sevenfold between 2016 and 2021 [1]. To mitigate the problem associated with the increasing demand for mobile data traffic much research has addressed the design of 5th-generation (5G) cellular communication systems. Among many research activities related to the development of 5G infrastructure, F-RANs have been considered as a promising network architecture for improving spectral efficiency (SE) [2-4].

F-RANs, which is evolved from cloud radio access network (C-RAN), differs from the conventional cellular architectures, as it pays more attention towards the improvement of user experience by minimizing latency and transmission delay occurring in the backhaul link [5-6]. Unlike C-RAN, F-RANs can alleviate the problem associated with the backhaul load by enabling remote radio heads (RRHs) with edge cache [7]. RRHs with certain caching and signal processing capabilities are referred to as F-APs [8]. The edge cache in the F-APs has the ability to store some content that is more likely to be accessed by the users, thereby reducing backhaul load and latency. Nonetheless, the fronthaul load is still the key limiting factor to the performance of F-RANs [9]. Many studies have been done to design the best algorithm to enhance the performance of F-RANs in terms of bandwidth consumption in content diffusion [10].

There are some related works worth mentioning. The work in [11] focused on designing cache enabled RRHs to store popular contents to maximize the received data rate under fronthaul capacity and power constraints. The paper in [12] presented an information-theoretic model of F-RANs to provide a latency-centric understanding of the degrees of freedom in the F-RANs. Similarly, the work in [13] focused on reducing delivery latency and the burden on fronthaul links by fully utilizing caching and signal processing capabilities of enhanced remote radio heads (eRRHs). In addition, the works in [14-16] focused on providing higher data rate and the best signal quality (BSQ) leading majority of UEs to associate to the interesting F-APs which have the better signal to interference ratio (SIR). Although these approaches obtain higher data rate, the preferred F-APs is most likely to be suffered which causes unstable cache resource allocation between the F-APs.

In this paper, to lessen the burden of most commonly used F-APs, fully exploit the caching capabilities of multiple F-APs and maximize the total delivered data content, an optimal LP based method is proposed which allocates proper cache resource to UEs from F-APs.

The remainder of this paper is organized as follows. The system model is described in section 2. Section 3 formulates the cache resource optimization problem of the multi F-APs F-RANs system. Matlab-based simulations are provided in Section 4. Finally, Section 5 concludes the paper.

*Corresponding Author: Hoon Kim; E-mail: hoon@inu.ac.kr
DOI: 10.3966/160792642019122007005

2 System Model

This section models a downlink $N \times H$ F-RANs system composed of N multi-antenna UEs devices, $U = (1, 2, \ldots, N)$, H multi-antenna F-APs, $F = (1, 2, \ldots, H)$, 1 base-band unit (BBU) pool, and 1 centralized cloud as shown in the Figure 1. In the system model diagram, fronthaul links are denoted by the solid lines whereas air interface links are indicated by dashed lines. Each UEs are served by different F-APs that are connected to a BBU pool in the cloud via common public radio interface (CPRI) cables.

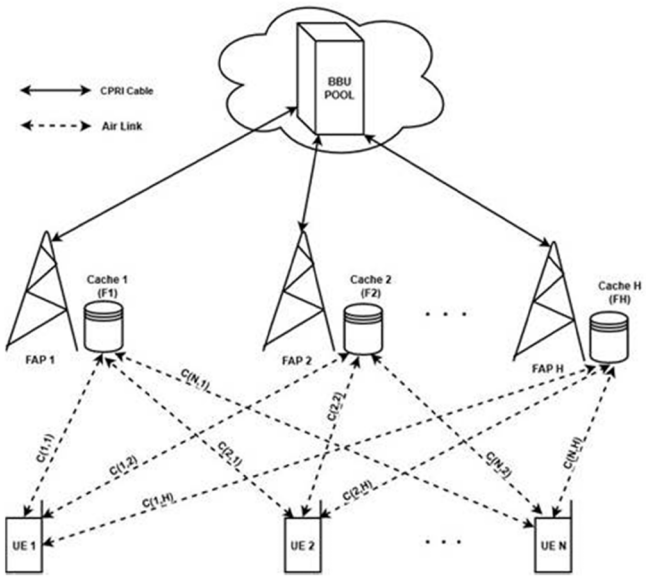

Figure 1. A system model for F-RANs

The model deals only with F-RANs cache delivery phase, so for a full caching case, all the requested files are assumed to be stored in the local cache such that it can be retrieved directly from the F-AP without downloading it from BBU. Since the single F-AP with edge cache is not enough to store all the information, the model assumes that the portion of the information requested by the users is present in many F-APs [11]. The single UE can be connected to multiple F-APs at the same time based on the concept of global cloud radio access network mode [8]. The link capacity between UEs and F-APs is determined by Shannon capacity limit i.e., $B * log_2(1 + SIR)$, where B is the channel bandwidth. The channel distribution of the air interface link is assumed to be a Rayleigh distribution, while the fronthaul link capacity is fixed.

Let $C_{i,j}^{Ai}$ be the air interface link capacity between UE $i (i \in U)$ and access point j $(j \in F)$, $C_{j,1}^{FH}$ be the fronthaul link capacity between access point j and cloud 1, and F_j be the cache capacity of F-AP j. Similarly, let R_i be the data to be received by the UE i, V_i be the minimum data volume requested by UE i and $\varnothing_{i,j}$ be the optimal cache resource allocation parameter and can be defined as the portion user i requested content in F-AP j.

3 Problem Formulation

This section characterizes the maximum total data that can be delivered in a multi F-APs F-RANs system by solving the cache resource allocation problem presented in section 2 with the arbitrary number of users, N. The total data obtained by all the users at any timeslot t is given by $R_{total} = \sum_{i=1}^{N} R_i$, where R_i is $\sum_{j=1}^{H} \varnothing_{i,j} F_j$. In cache-level transmission, the maximization of the total sum of receivable data contents of all users can be done by realizing the solution of following problem:

$$\text{(P1)} \quad \max \left[\sum_{i=1}^{N} \sum_{j=1}^{H} \varnothing_{i,j} F_j \right] \tag{1}$$

$$\text{s. t.} \ \frac{\varnothing_{i,j} F_j}{\tau} \le C_{i,j}^{Ai} \qquad \forall i, j \tag{2}$$

$$\sum_{j=1}^{H} \varnothing_{i,j} F_j \ge V_i \qquad \forall i \tag{3}$$

$$\sum_{i=1}^{N} \varnothing_{i,j} \le 1 \qquad \forall j \tag{4}$$

$$\varnothing_{i,j} \ge 0 \qquad \forall i, j \tag{5}$$

where transmission delay τ is a summation of the signal processing and information exchange time between F-AP j and the BBU. For simplicity, τ is assumed to be same for all F-APs. In (P1), constraint (2) shows that the portion of user data delivery rate $\frac{\varnothing_{i,j} F_j}{\tau}$ obtained from specific F-AP is bounded by the air interface link capacity connecting user to that F-AP i.e., $C_{i,j}^{Ai}$. Likewise, constraint (3) denotes that the total data volume obtained by the user must be greater than or equal to the volume of the data requested by them. Similarly, constraint (4) means that the sum of the portion of contents that can be accessed by all the users from any particular F-AP cannot be more than 100% of its total capacity. Furthermore, constraint (5) ensures that the portion of content that any user i can access from F-AP j is non-negative.

The problem (P1) is an optimization problem that can be easily solved with a LP server. To find the optimal solution for (P1) all the constraints representing this problem need to be satisfied. The known values of F_j, τ, V_i and $C_{i,j}^{Ai}$ are used to compute the optimal value of $\varnothing_{i,j}$, such that the cache resource allocation problem is solved as well as the

total data delivered i.e. R_{total} is maximized.

If any of the constraints presented in (P1) is not satisfied, then the solution obtained will not be feasible and is referred to as an outage. Outage is a condition in which data will not be transferred from F-APs to UEs. This can be represented as:

$$\text{Outage} = \begin{cases} 1, & \text{if (P1) is not feasible} \\ 0, & \text{otherwise} \end{cases}$$

3.1 Delivery Phase of Multi F-APs F-RANs System

In this part, problem formulation scenario is graphically represented to enhance the readability of this paper.

Figure 2 illustrates an example of delivery phase of $N \times H$ F-RANs system with an optimal allocation of cache resource. As shown in the figure, the buffer zone is regarded as transition phase before UEs receives data. The total cache resource of any F-APs is mapped to buffer zone as per the user data volume request and number of users requesting data at the same time. The portion of cache resource allocated by F-APs to any UEs depends on the strength of air link capacity with which UEs communicates to F-APs. The data delivered from the multiple F-APs is at the same time as the UE is embedded with multiple antennas. Thus, the total data received by any UEs is the summation of portion of requested data obtained from connected F-APs.

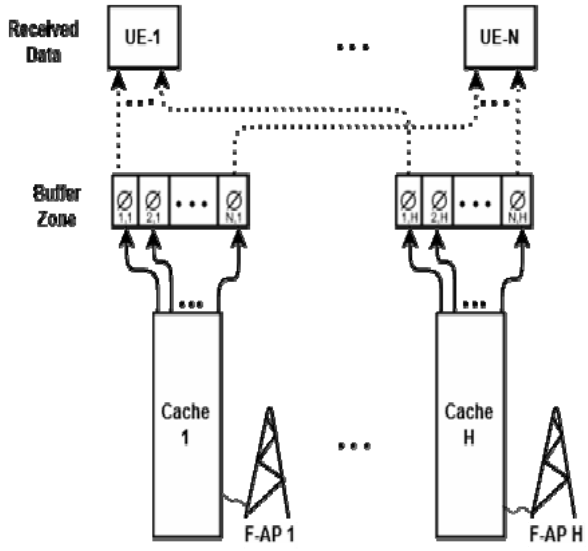

Figure 2. Illustration of delivery phase of multi F-AP F-RANs system

4 Performance Evaluation

In this section, MATLAB-based simulations are conducted to evaluate the performance of proposed LP-based method. The nature of total data delivered and outage are analyzed for different simulation parameters such as, F-AP caching capacity, transmission delay, minimum user data volume requirement, and average SIR.

4.1 Simulation Parameters

Simulations consider a scenario with 6 mobile users and 3 F-APs. For simplicity, the channel capacity between 6 mobile users and 3 F-APs is considered to be dependent on the same average value of SIR. Similarly, the cache capacity of 3 F-APs is assumed to be the same at any particular timeslot. The channel bandwidth of the sub-carriers is assumed to be 20 MHz. The minimum data volume requirement is assumed to be same for all the users at any particular timeslot. All the simulation results are obtained by averaging over 1,000 random channel realization. The fixed parameter settings are listed in Table 1.

Table 1. Fixed simulation parameters

Parameters	Value
Number of UEs	6
Number of F-APs	3
Channel variation distribution	Rayleigh distribution
Channel capacity	$B * log_2(1 + SIR)$
Channel bandwidth	20 MHz
Timeslot	1,000

4.2 Simulation Results

4.2.1 F-APs Cache Capacity

In this part, the results of total data delivered maximization problem are displayed for the different cache capacity of F-APs. The value of τ, average SIR, and V_i is taken as 10 ms, 2.5 dB, and 1 Mb respectively.

Figure 3 shows the behavior of total data that is delivered for different cache capacity F-APs, provided that the minimum data volume requirement for each UE is set to 1Mb. There is an outage up to 2 Mb cache capacity of F-AP as the total cache capacity of 3 F-APs doesn't satisfy the demand of 6 UEs. After 2 Mb cache capacity of each individual F-AP, the total data delivered drastically increases till 4 Mb cache capacity of F-AP. The total data delivered saturates after 4 Mb cache capacity of F-AP due to limiting nature of air interface link capacity. As shown in Figure 2, (2-4) Mb cache capacity of the F-APs is said to be an optimal capacity of F-APs, for which total data that can be delivered is maximum.

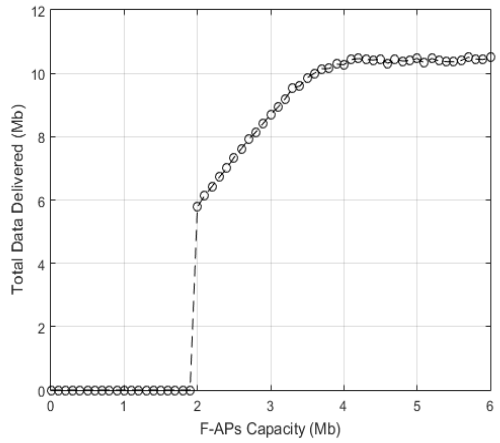

Figure 3. Total data delivered vs. F-APs capacity

Similarly, Figure 4 shows the probability of an outage occurrence for different values of F-APs cache capacity. Here, the y-axis representing the probability of an outage occurrence is in logarithmic scale to show the trends of highest and lowest data types in the same graph, while the values used are linear. As illustrated in the figure, the probability of an outage occurrence is 1 up to 2 Mb cache capacity of 3 F-APs which is caused by the greater data volume requirement than the total cache capacity of F-APs. After 2 Mb cache capacity of the F-APs, the probability of an outage occurrence drops to less than 0.01, as the total F-APs capacity overcomes the total data request. The probability of an outage occurrence is 'not zero', not even after F-APs capacity becomes 2 Mb. It is due to the fluctuating nature of air interface link capacity.

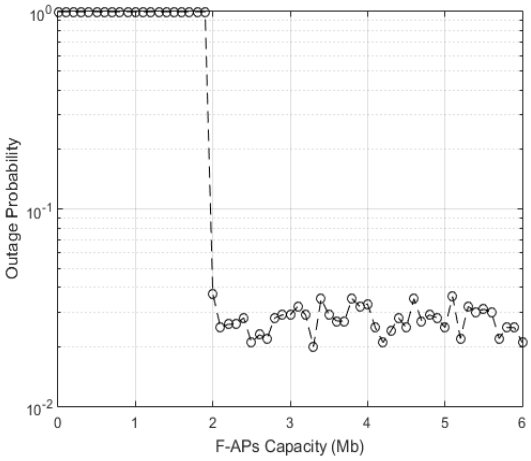

Figure 4. Outage probability vs. F-APs capacity

With the graphs illustrated above, we can easily determine the capacity of F-APs that is sufficient to meet the requirements of the users such that the number of users, the air interface link capacity, transmission delay and the user behavior regarding minimum data volume requirements are predictable.

4.2.2 Transmission Delay

In this part, the results of total data delivered maximization problem are displayed for the different transmission delay. The value of F_j, average SIR, and V_i is taken as 6 Mb, 2.5 dB, and 1 Mb respectively.

Figure 5 illustrates the volume of data that is delivered for different values of the transmission delay τ when the minimum user data volume requirement is set to 1 Mb. For the first few milliseconds, the volume of data transmitted is 'zero' as the minimum data requested by the users is not attained. After the fulfillment of minimum user requirement, the volume of total data that is delivered increases as the delay constraint is relaxed. However, the volume of total data that is delivered comes to saturation level after a certain transmission delay due to the optimum utilization of available F-APs cache resource.

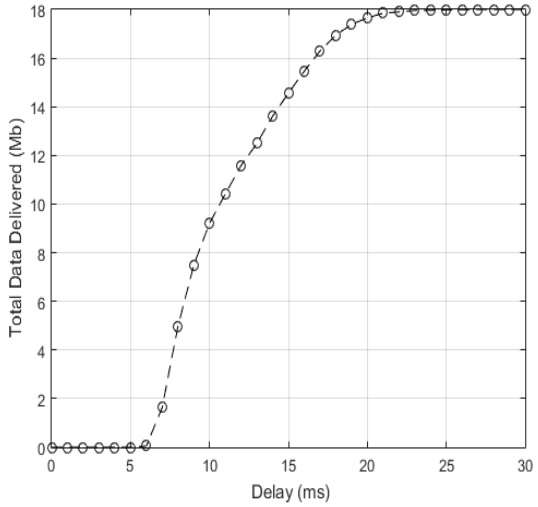

Figure 5. Total data delivered vs. transmission delay

Similarly, Figure 6 portraits the probability of an outage when the value of transmission delay is varied keeping other parameters same. As shown in the figure, the probability of an outage is 1 when the transmission delay is low, i.e., below 6 ms. Likewise, when the transmission delay is increased beyond 6 ms, outage probability exponentially decreases and drops to 'zero' after fulfilling minimum user data request.

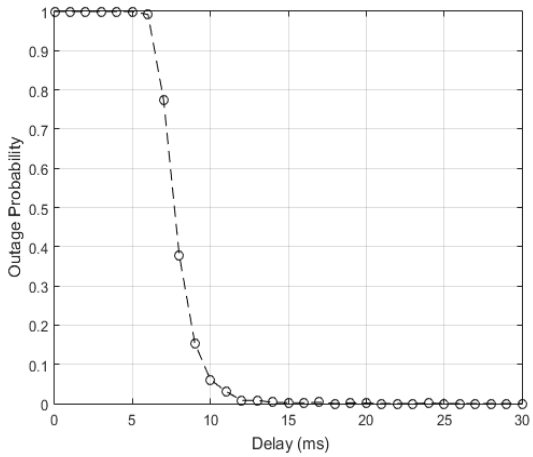

Figure 6. Outage probability vs. transmission delay

With the graphs presented in Figure 5 and Figure 6, we can easily determine the value of transmission delay that meets the minimum requirements of the users such that total data delivered is maximized.

4.2.3 Minimum Data Volume Requirement

In this part, for variable minimum data volume requirement V_i the outputs of (P1) are analyzed. The value of F_j, average SIR, τ and is taken as 6 Mb, 2.5 dB, and 30 ms respectively.

Figure 7 shows the volume of data that is delivered for varying minimum user data volume requirement. When the user data volume requirement is low, the data received is high. Likewise, when the user data volume requirement is set to high, the total volume of data obtained drops to 'zero' as the total cache capacity of multi F-APs F-RANs system cannot address the minimum requirement of each user.

Similarly, Figure 8 shows the outage probability for varying user data volume requirement. When the user data volume requirement is low, the outage probability is low. Likewise, when the user data volume requirement is set to high, the outage probability soars to 100 % as the total cache capacity of multi F-APs of F-RANs system cannot address the minimum requirement of each user.

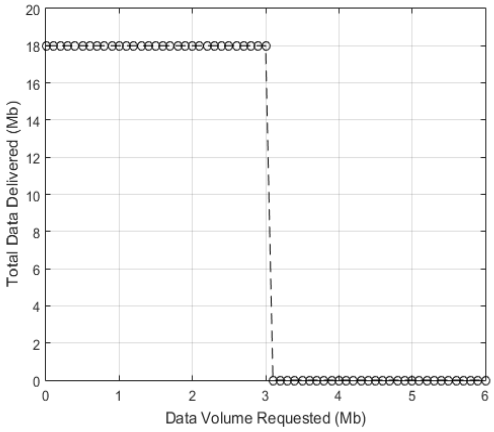

Figure 7. Total data delivered vs. minimum data volume request

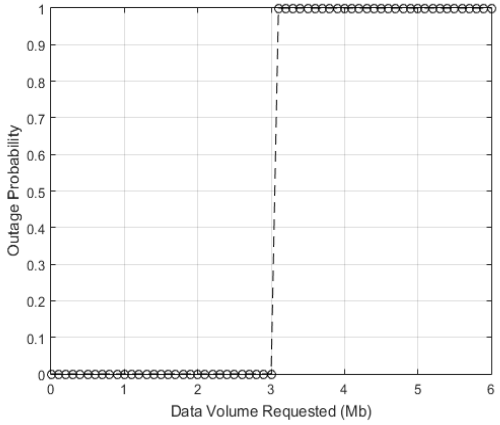

Figure 8. Outage vs. minimum data volume request

4.2.4 Average SIR

In this part, for variable SIR the outputs of (P1) are analyzed. The value of F_j, V_i, and τ is taken as 6 Mb, 2 Mb, and 30 ms respectively.

Figure 9 shows the volume of data delivered for different SIR. When the average value of SIR is poor, the total data received by all the users is 'zero' as the minimum data volume requirement of each user cannot be fulfilled with that SIR. Similarly, when the average SIR value is increased such that it satisfies minimum user requirement, then the data obtained also increases. However, the total data received saturates after a certain limit due to the fixed cache capacity of F-APs irrespective of good strength of the radio signal used.

Similarly, Figure 10 illustrates the probability of an outage for different values of average SIR such that other parameters are not changed. As shown in the figure, the probability of an outage is 1 when the average SIR is low, i.e., below 8. Likewise, when the average SIR is increased beyond 8, outage probability decreases exponentially and reaches near marginal line.

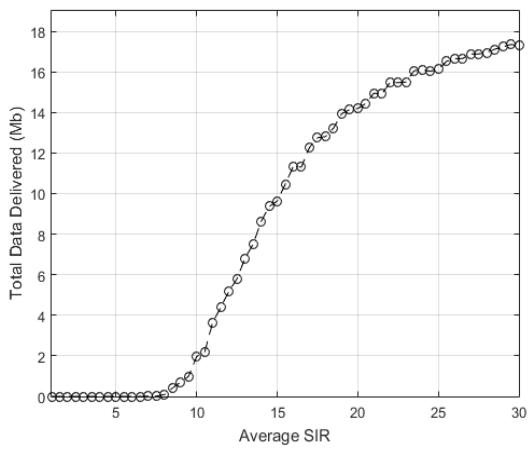

Figure 9. Total data delivered vs. average SIR

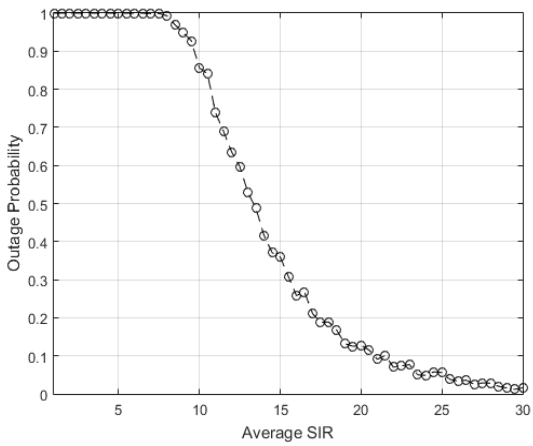

Figure 10. Outage probability vs. average SIR

With the graphs presented in Figure 9 and Figure 10, the optimal value of average SIR can be determined, that meets the minimum requirements of the users as well as maximizes total data delivered.

5 Conclusion

This paper studied the optimal cache resource allocation process in multi F-APs of F-RANs system. In this paper, LP based method has been proposed to allocate the cache resource of multiple F-APs such that it is fully utilized. The feasibility of the proposed method has been studied under computer simulations by setting up the values of system parameters. Simulation results validated by our theoretical model indicated that the proposed method can maximize the total content which can be delivered in F-RANs system (up to the total cache capacity of F-APs) and limit the occurrence of an outage, provided that; the user minimum data volume requirement, average SIR, transmission delay and F-APs cache capacity are not the limiting factors. For future works, more general simulation environments need to be considered.

6 Acknowledgement

This work was supported by Incheon National University Research Grant in 2015.

References

[1] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper,* https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] X. Huang, G. Xue, R. Yu, S. Leng, Joint Scheduling and Beamforming Coordination in Cloud Radio Access Networks With QoS Guarantees, *IEEE Transactions on Vehicular Technology*, Vol. 65, No. 7, pp. 5449-5460, July, 2016.

[3] S. Kitanov, T. Janevski, Fog Computing Service Orchestration Mechanisms for 5G Networks, *Journal of Internet Technology*, Vol. 19, No. 1, pp. 297-305, January, 2018.

[4] Q. Jia, R. Xie, T. Huang, J. Liu, Y. Liu, Efficient Caching Resource Allocation for Network Slicing in 5G Core Network, *IET Communications*, Vol. 11, No. 18, pp. 2792-2799, December, 2017.

[5] H. Dahrouj, A. Douik, O. Dhifallah, T. Al-Naffouri, M. Alouini, Resource Allocation in Heterogeneous Cloud Radio Access Networks: Advances and Challenges, *IEEE Wireless Communications*, Vol. 22, No. 3, pp. 66-73, June, 2015.

[6] M. Peng, Y. Li, Z. Zhao, C. Wang, System Architecture and Key Technologies for 5G Heterogeneous Cloud Radio Access Networks, *IEEE Network*, Vol. 29, No. 2, pp. 6-14, Mar.-Apr., 2015.

[7] Q. Li, H. Niu, A. Papathanassiou, G. Wu, Edge Cloud and Underlay Networks: Empowering 5G Cell-Less Wireless Architecture, *European Wireless 2014; 20th European Wireless Conference*, Barcelona, Spain, 2014, pp. 1-6.

[8] M. Peng, S. Yan, K. Zhang, C. Wang, Fog-computing-based Radio Access Networks: Issues and Challenges, *IEEE Network*, Vol. 30, No. 4, pp. 46-53, Jul-Aug., 2016.

[9] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, G. Caire, FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers, *IEEE Transactions on Information Theory*, Vol. 59, No. 12, pp. 8402-8413, December, 2013.

[10] X. Wang, S. Leng, K. Yang, Social-Aware Edge Caching in Fog Radio Access Networks, *IEEE Access*, Vol. 5, pp. 8492-8501, April, 2017.

[11] S. Park, O. Simeone, S. Shamai Shitz, Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks, *IEEE Transactions on Wireless Communications*, Vol. 15, No. 11, pp. 7621-7632, November, 2016.

[12] R. Tandon, O. Simeone, Cloud-aided Wireless Networks with Edge Caching: Fundamental Latency Trade-offs in Fog Radio Access Networks*, IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, 2016, pp. 2029-2033.

[13] S. He, C. Qi, Y. Huang, Q. Hou, A. Nallanathan, Two-Level Transmission Scheme for Cache-Enabled Fog Radio Access Networks, *IEEE Transactions on Communications*, Vol. 67, No. 1, pp. 445-456, January, 2019.

[14] B. Cheng, X. Mi, X. Xu, Z. Xu, X. Xu, M. Zhao, A Real-time Implementation of Comp Transmission Based on Cloud-ran Infrastructure, *IEEE International Wireless Communications and Mobile Computing Conference (IWCMC)*, Nicosia, Cyprus, 2014, pp. 1033-1038.

[15] M. Mezzavilla, K. Somasundaram, M. Zorzi, Joint User Association and Resource Allocation in Uerelay Assisted Heterogeneous Networks, *IEEE International Conference on Communications Workshops (ICC),* Sydney, Australia, 2014, pp. 628-634.

[16] J. G. Andrews, Seven Ways that Hetnets Are a Cellular Paradigm Shift, *IEEE Communications Magazine*, Vol. 51, No. 3, pp. 136-144, March, 2013.

Biographies

Sovit Bhandari received his Bachelor's degree from Kathmandu University, Nepal, in 2016. He is currently working as a Graduate Research Assistant at Incheon National University, South Korea. His research interests include networking, 5G mobile communication, machine learning and internet of things.

Hong Ping Zhao received his Bachelor's degree from Jilin Jianzhu University, China, in 2016. He is currently working as a Graduate Research Assistant at Incheon National University, South Korea. His research interests include machine learning, big data and networking.

Hoon Kim has been working as an Associate Professor at Department of Electronics Engineering, Incheon National University, South Korea, since 2008. His research interests include radio resource management, optimization techniques, 5G mobile communication systems, machine learning, and internet of things. He is a Member of KICS, IEIE, IEEE, and IEICE.

John M. Cioffi BSEE, 1978, Illinois; PhDEE, 1984, Stanford; Bell Laboratories, 1978-1984; IBM Research, 1984-1986; EE Prof., Stanford, 1986-present, now emeritus.Cioffi has published over 600 papers and holds over 100 patents, of which many are heavily licensed including key necessary patents for the international standards in ADSL, VDSL, vectored VDSL, G.fast, DSM, LTE, Massive-MIMO, and various Wi-Fi methodologies.