Improving Handover of 5G Networks by Network Function Virtualization and Fog Computing

(Invited Paper)

Yu Qiu*†, Haijun Zhang†, Keping Long†, Hongjian Sun‡, Xuebin Li*, Victor C.M. Leung§
*Beijing University of Chemical Technology, Beijing, China
† University of Science and Technology Beijing, Beijing, China
‡ University of Durham, Durham, UK
§ The University of British Columbia, Vancouver, Canada
Email: †haijunzhang@ieee.org, ‡hongjian.sun@durham.ac.uk, §vleung@ece.ubc.ca

***Abstract*—In Fifth Generation (5G) cellular networks, it is necessary to meet a number of requirements, such as high scalability, ultra-low latency, reduced energy consumption, and high energy efficiency. Particularly in the high mobility scenario, the optimization of handover through managing signalling overhead and delay is of primarily importance. In this paper, the idea of integrating Network Function Virtualization (NFV) and Fog Computing is explored. NFV has the advantage of improving network flexibility whilst reducing overall overhead. The Fog-Computing Access Points (F-APs) are then employed with certain caches in the edge of networks. Moreover, a direct-X2 based handover scheme is proposed. Taking advantages of both edge caching and Virtual Machines (VMs), this proposed handover scheme has superior performance: the signalling cost of handovers can be as little as 65% of that of a conventional LTE network.**

***Index Terms*—5G cellular networks, handover, signalling overhead, ultra-low latency, virtual machines, fog computing.**

I. Introduction

Over the past decade, due to the rapid development of terminal equipments and increased demand for mobile broadband services, many research and industrial initiatives have focused on researching next generation cellular networks, i.e., 5G. It is expected that 5G can provide enormous mobile data capacity, 1,000 times greater than those of today, approximately several gigabits per second [1], and ultra-low latency as less as a few milliseconds. Meanwhile, a growing amount and variety of access devices and powerful terminal equipments have already emerged, e.g., smart phones, sensors, connected vehicles, and roadside units [2]. Moreover, they will consume and produce huge data. 5G would become an enabling technology in this new era of Internet of everything [3].

To deal with massive number of connected devices and explosive growth of data traffic, the cloud radio access network (CRAN) was proposed. The functions of control, computing, and storage are assigned to a centralized cloud [4]. Additionally, heterogeneous cloud radio access network (HCRAN) was studied to overcome the weaknesses of CRANs. In the HCRAN, high power nodes can support ubiquitous coverage and are connected to the baseband unit pool via backhaul links for interference management. Remote radio heads provide high-speed data rate to transmit the packet traffic [5]. But HCRAN may cause additional burden on the fronthaul and backhaul links.

Cisco Systems Inc. introduced a Fog Radio Access Network (FRAN) to shift the tasks of control, communication, data storage, and management to the edge of networks. In 5G networks, fog computing can provide low service delay, improved location awareness, and good Quality of Service (QoS), leading to superior user experience. The characteristics of FRAN include seamless coverage, distributed management and cooperation [6]. In the FRAN, data storage is near the end user equipment (UE), instead of only in remote data centers. Applying FRAN to 5G networks could allow massive number of devices connected to Internet, including sensors, devices and self-driving vehicles. These connected devices can establish multiple mini-clouds at the edge of the network, that can exchange data locally or connect to the core network through F-APs. With the deployment of F-APs, handovers however will cause huge energy consumption. In addition, these F-APs will result in a large neighbor cell list and cause interference [7]. On the other hand, with the commercial deployment of 5G networks, the mobile network users look forward to having much faster connection in 5G networks. Service providers thus are facing a significant challenge: how to meet the expectation of users with reasonable financial investments. Additionally, the maintenance and configuration of numerous different types of services running on various devices will aggravate the current difficult management situation of FRANs. Hence, it is eagerly needed to develop new methods of managing heterogeneous devices and their running services in the FRAN.

Telecom operators proposed to use Network Function Virtualization (NFV) for dealing with the shortage of business agility and meeting the continuous requirement of reliable infrastructures. The expenditures of network mainly depend on network infrastructure. The huge expense of any new service release or network enhancement upgrade will inevitably reduce the economic revenue of the service provider. The challenges of network management include: not only the rising energy costs, but also the spending of various expensive hardware devices and the competitive market for high qualified talents with the skills. At the same time, how to manage the network

The corresponding authors are Dr Haijun Zhang and Dr Hongjian Sun.

infrastructure is another main issue of service providers. These problems have serious implications for the scale of the fiscal revenues at operators, and create obstacles on the road of nourishing innovations for telecom enterprises. It is of great importance for network operators to minimize or even remove their dependencies on expensive proprietary hardware. Consequently, telecom operators have formed an industry specification team for NFV under the European Telecommunications Standards Institute. In October 2012, the group proclaimed their relevant policy. NFV is a novel network architecture concept that the network functions should migrate from custom hardware devices to virtualized software appliances. There are some network entities in the Evolved Packet Core (EPC), including Mobility Management Entity (MME), Policy and Charging Rules Function (PCRF), Serving gateway (SGW) and Packet data network gateway (PGW). According to the NFV, these network entities can be virtualized. Making the network functions run on clusters of Virtual Machines (VMs), NFV can achieve significant reduction in the energy consumption and complexity of realizing network functions [8]. In [9], the authors evaluated a commercial NFV-based EPC. Authors in [10] made central MME core node dispersed to multiple kinds of replicas and placed it closer to the access edge of networks to enable reduction in the delay and better handover performance. However, to date few studies focus on the combination NFV technology with future FRAN architectures.

In order to reduce latency, signalling overhead, and efficiently alleviate the burden on the fronthaul, backhaul, and backbone networks, this paper studies a new network architecture that integrates both fog computing and NFV. The rest of the paper is organized as follows. Section II presents the FRAN architecture using NFV. In Section III, handover procedures for F-APs are introduced. In Section IV, a signalling analysis model is introduced and the performance of F-AP mobility schemes is evaluated in Section V. Finally, Section VI concludes the paper.

II. Proposed FRAN Architecture

The overview of our proposed NFV-based FRAN architecture is shown in Figure 1. 5G network means reliable connection with various things, such as smartphones, connected vehicles, sensors, and embedded artificial intelligence [11, 12]. In this case, we denote the smart user equipment as F-UE, and there are four possible transmission modes for them to connect with each other in the FRAN. The F-APs, that are unique to FRANs, can implement collaborative radio signal processing locally by using their adequate computing capabilities, and can manage their caching memories flexibly. Certain caches are equipped with F-APs. The contents stored in F-APs are highly and locally popular or relevant. With the increasing popularity of location-based mobile applications, a lot of information may be generated adding to the surging data traffic over the network links, and push the links to their capacity limits. Some social applications would only generate data traffic between F-UEs in close physical proximity. When users have the same social interest or come from the same social group, they may request almost the same contents over the downlink. In these cases, the requested services can be locally supported by F-APs with their caches of highly popular content. As a result, users do not need to be connected to the core network every time when they require data or contents. The caches in F-APs simplify a substantial part of handover procedure, and can thus reduce those handover overhead and delay.

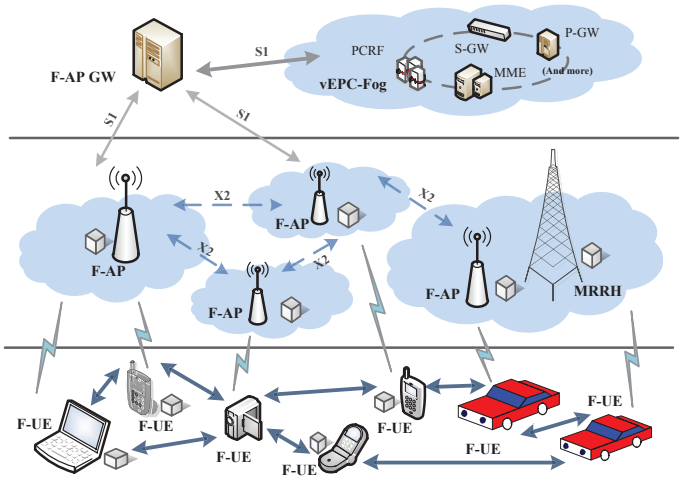

Fig. 1. A NFV-based FRAN Architecture.

The traditional EPC consists of MME, PCRF, SGW and PGW, with a lot of complex functionalities packed into a single box. MME plays a role of managing the session status, authenticating and tracking users. The responsibility of SGW is to route and transmit user data packets from and to the base station. The stable connection between the user data plane and external networks is provided by PGW. The PCRF supports the detection of service data flow, policy enforcement and charge in real time for each service and user. When traffic demand is becoming much higher, this makes huge cost to procure, maintain, and upgrade them. With the integration of various network functions with the NFV technology, the core network functions of the EPC (MME, SGW, PGW and PCRF) are deployed as virtualized network function, and are run on VMs in data centers. The virtualized EPC (vEPC) is used to denote these VMs. The corresponding data and control traffic flow across the vEPC under unified rules, policies, and definitions, rather than the strict constraint in fixed hardware. In the present network architecture, the F-APs connect to the vEPC with a set of S1 interfaces by F-AP Gateway (F-AP GW). The F-APs are also interconnected with each other by the direct X2 interface. But, with the increasing number of management data of the virtualized environment, the delay in understanding an alarm, performance degradation, configuration change, and other anomaly situations will emerge. While the distributed way of processing management data in the FRAN may contribute to make the understand anomaly situations faster and finer. Generally speaking, the introduction of NFV technologies simplifies the processing procedure of handover in traditional EPC, while the distributed management framework of FRAN makes the NFV performing effectively when facing explosive data. Thus this 5G NFV-

based FRAN architecture, that integrates fog computing and virtualization, can support high mobility, low latency and energy consumption.

III. Proposed Handover Procedure

The handover procedure among F-APs is different from existing ones. The handover call flow of the NFV-based FRAN architecture is shown in Fig. 2. A lot of X2 interfaces are deployed among the F-APs in the NFV-based FRAN. The handover signalling flow between the source F-AP and the target F-AP via the direct X2 interface is shown in Fig. 2, where the network is deployed by the combination NFV with fog computing. Taking advantage of the efficiently use of the edge of networks, the transmission of data from F-AP to F-UE does not need to go through the core network. Source F-AP transmits handover request signalling to the target F-AP through direct X2 interface. Admission control happens in the target F-AP, and then handover request Ack signalling would be transmitted back to the source F-AP. What follows is the data transmission between the source F-AP and the target F-AP, as shown in Fig. 2. In the traditional network, when the MME accepts a message, the path switches from target F-AP through the F-AP GW. Then, the modified bearer request is forwarded from the MME to the S-GW and P-GW. Next is the session modification in the PCRF. These modules implement their specified functions respectively in the designated order until the handover completion. With the introduction of NFV technology, changes have been made in the EPC. The vEPC as a single NFV equipment implements multiple functions in a safe and efficient working condition. The vEPC works independently to manage the session status, track users, forward data packets, and make some charging policies replace conventional handover procedure by the software running on commercial servers. Thus, a substantial part of handover procedure is eliminated.

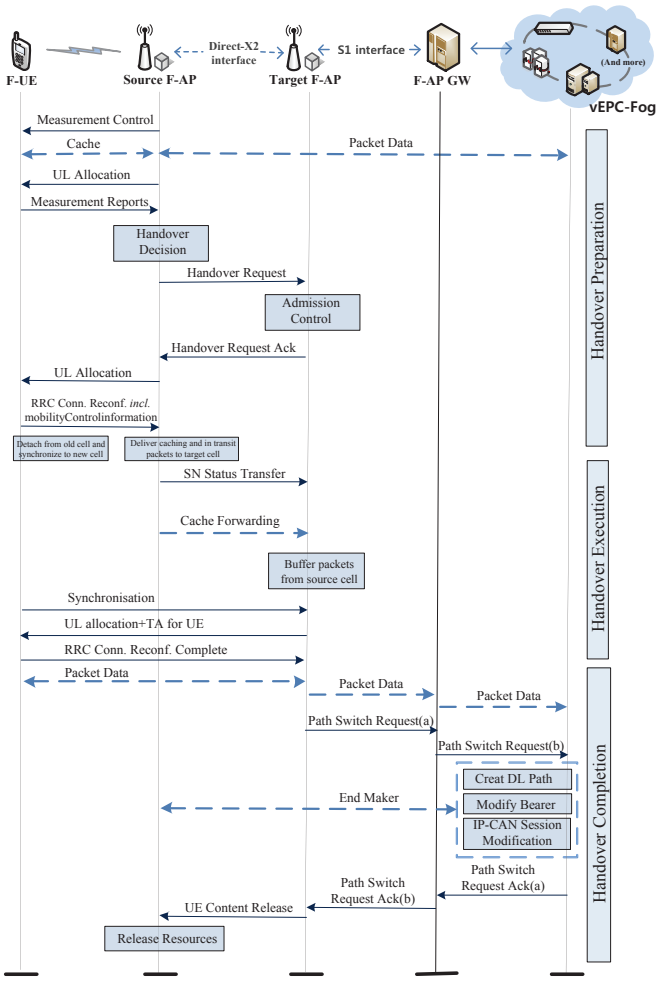

Fig. 2. Handover From F-AP To F-AP.

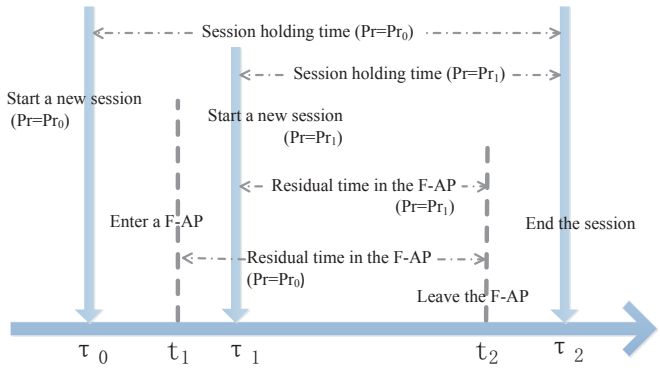

Fig. 3. Timing Diagram For Mobility Model.

IV. Signalling Analysis Model

This section will present a simple analytical model that investigates the signalling overhead of the F-AP based on the work in [13]. In this analytical model, if a single user moves across the border of two F-APs in an active state, it requires a handover from one F-AP to another, and generates handover signaling messages. That F-UE moves to the F-AP with equal probability is the crucial assumption for the mobility model. It is assumed that the call may occur at any moment. There are two scenarios that should be considered in the handover between source F-AP and target F-AP. Fig. 3 shows the timing diagram for the analytical model.

In this paper, we assume that communication session arrivals to a F-UE following the Poisson process with an average rate λ. In Fig. 3, there are two scenarios that call arrivals to the F-UE at two points (τ_0 and τ_1). Then, the session ends at τ_2. Pr_{τ_0} and Pr_{τ_1} are denoted as the probabilities of these two scenarios, respectively. t_1 is the moment when F-UE enters the range of a F-AP. In scenario 1, the F-UE has a call at the moment τ_0 before entering the range of the F-AP. By contrast, the F-UE has a call at the moment τ_1 after entering the range of the F-AP in scenario 2. t_2 is the moment when F-UE leaves the range of the F-AP and enters the range of another F-AP. Moreover, the call is going on at the moment t_2. The handover messages of two cases generate at the moment t_2. The probability of the handover happens on the border of the F-AP is defined as the sum probabilities of scenario 1 (i.e., Pr_{τ_0}) and scenario 2 (i.e, Pr_{τ_1}):

$$\mathrm{Pr} = \mathrm{Pr}_{\tau_0} + \mathrm{Pr}_{\tau_1}. \tag{1}$$

We also assume that the session holding time T_D follows an

exponential distribution with a mean $1/\alpha$. $f_{T_H}(t)$ is denoted as the probability density function (PDF) of T_H, given by:

$$f_{T_H}(t) = \alpha e^{-\alpha t}. \tag{2}$$

T_R is the UE/F-UE residence time in the range of a HeNB/F-AP and is exponentially distributed with a mean $1/\beta$. $f_{T_R}(t)$ is denoted as the PDF of T_R, given by:

$$f_{T_R}(t) = \beta e^{-\beta t}. \tag{3}$$

The session holding time and residence time are independent random variables.

Let T_{Hr} denote the residual time for session holding. T_{Rr} is denoted as the residual time for UE/F-UE residence time. T_{Hr} and T_H are exponentially distributed with the same mean $1/\alpha$. Similarly, T_{Rr} and T_R are exponentially distributed with the same mean $1/\beta$.

Building on the assumptions mentioned above, the probability for case 1 can be derived as follows [13]

$$\begin{aligned}\Pr_{\tau_0} =& P(\tau_0 < t_1 < \tau_0 + T_H) \times P(T_{Hr} > T_R) \\ =& \int_0^\infty \int_0^\infty \lambda t e^{-\lambda t} f_{T_H}(y) dy dt \times \\ & (1 - \int_0^\infty \int_t^\infty \beta e^{-\beta x} f_{T_{Hr}}(t) dx dy) \\ =& \frac{1}{(\alpha+\beta)} \times \frac{\lambda\beta}{(\lambda+\alpha)^2};\end{aligned} \tag{4}$$

the probability for case 2 can be derived as follows [13]

$$\begin{aligned}\Pr_{\tau_1} =& P(t_1 < \tau_1 < t_0 + T_{Rr}) \times P(T_H > T_{Rr}) \\ =& \int_0^\infty \lambda t e^{-\lambda t} f_{T_{Rr}}(t) dt \times \\ & \int_0^\infty \int_t^\infty \alpha e^{-\alpha y} f_{T_{Rr}}(t) dy dt) \\ =& \frac{1}{(\alpha+\beta)} \times \frac{\lambda\beta^2}{(\lambda+\beta)^2}.\end{aligned} \tag{5}$$

Transmission overhead is the cost of transmitting handover message between two nodes and signalling processing overhead is the cost of processing messages at each node in the network [14, 15]. The sum of transmission overhead and processing overhead is the handover signalling overhead. The signalling overhead in the F-AP related handover are denoted in Table I and Table II.

TABLE I
TRANSMISSION COST PARAMETERS

Cost	Transmission Cost
T_{F-UE}^{F-AP}	From F-UE to F-AP
T_{F-AP}^{F-AP}	From F-AP to F-AP
$T_{F-AP}^{F-AP\_GW}$	From F-AP to F-AP GW
$T_{F-AP\_GW}^{vEPC}$	From F-AP GW to vEPC

Signalling overhead of the F-AP related handover in each scenario is given by:

$$O = \Pr \times (\sum T_n^m + \sum P_k), \tag{6}$$

TABLE II
PROCESSING COST PARAMETERS

Cost	Processing Cost	Cost	Processing Cost
P_{F-UE}	At F-UE	P_{F-AP}	At F-AP
P_{F-AP_GW}	At F-AP GW	P_{vEPC}	At vEPC

where Pr is the probability of the handover in a scenario, $(\sum T_n^m + \sum P_k)$ is the signalling overhead in the scenario.

V. PERFORMANCE EVALUATION

In order to verify the performance of the proposed network architecture, we compare this proposed NFV-based FRAN architecture with the traditional architecture with respect to their system signalling overhead in different scenarios by using the signalling analysis model presented in Section IV.

The transmission overhead and processing overhead can be defined to be proportional to the time required for delivering and processing signaling messages. We assumed the overhead parameters have no unit. Other measurements for the overhead parameters are possible. Here we used the parameters in the following Table:

TABLE III
PARAMETERS CONFIGURATION

Parameters	Value	Parameters	Value
T_{F-UE}^{F-AP}	2	P_{F-UE}	32
T_{F-AP}^{F-AP}	2	P_{F-AP}	2
$T_{F-AP}^{F-AP\_GW}$	2	P_{F-AP_GW}	2
$T_{F-AP\_GW}^{vEPC}$	8	P_{vEPC}	5

We performed computer simulations to compare the performance of the proposed NFV-based FRAN handover procedure with that of a conventional LTE network in terms of system signalling overhead. Fig. 4 shows the signalling overhead versus average session arrival rate λ (session/minute) with $1/\alpha$ = 2 minutes, $1/\beta$ = 2 minutes. From this simulation graph, we can find that as the average session arrival rate increases, the signalling overhead continues to rise in both HeNB-HeNB (traditional) and F-AP-F-AP handover (proposed). This is due to that more handovers are generated as the increase of the session arrivals. With the increasing number of data, the proposed distributed FRAN performs well in handling the data traffic. It works effectively through a single NFV entity. These evolutions contribute to substantial reductions in the signalling overhead. With the average call arrival rate 0.5, the signalling cost of the proposed handover scheme can be as little as 65% (17/26) of that of a conventional LTE network.

Fig. 5 shows the signalling overhead versus the average holding time with the value of λ (session/minute) set as 0.1. As shown in the graph, the total signalling overhead increases as the average session holding time increases. This is because that the long session holding time introduces a higher probability

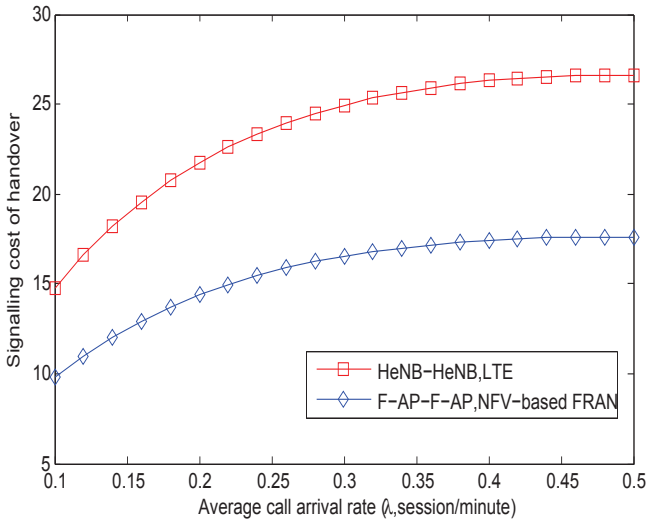

Fig. 4. Signalling Cost Versus Average Call Arrival Rate λ.

of cell-boundary crossings and handovers. Similar to the Fig. 4, the handover overhead in the proposed NFV-based FRAN is smaller than that of a conventional LTE network. Moreover, it can be seen that the transmission overhead and the processing overhead are closely related to the transmission delay and the processing delay, respectively. In other words, the delay performance of handovers using both NFV technology and fog computing is better than that of the conventional LTE networks.

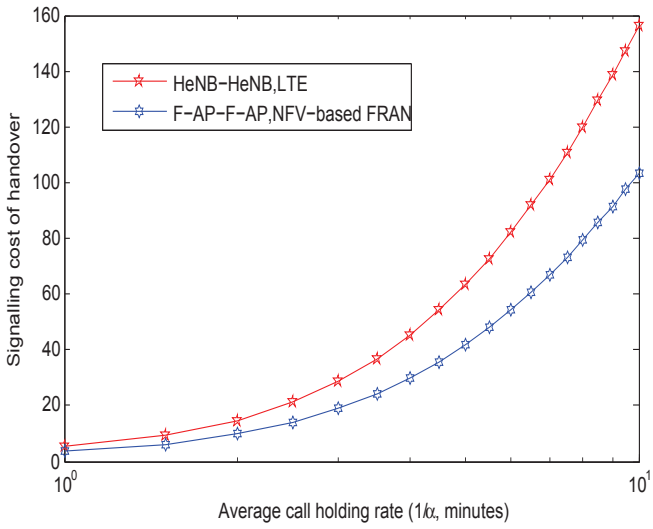

Fig. 5. Signalling Cost Versus The Average Call Holding Time $1/\alpha$.

VI. Conclusions

In this paper, we have proposed a new network architecture that integrates the fog computing and the NFV. A new signalling procedure of handovers has also been designed as a 5G NFV-based FRAN. According to the handover procedure building on direct X2 interface, this signalling overhead has been evaluated using an analysis model in Section IV. In comparison with a conventional LTE network, the proposed handover scheme has superior performance, i.e., its signalling cost is as little as 65% of that of LTE networks. Future research will focus on the hardware implementation of the proposed scheme, and laboratory tests in a more practical environment, for instance in connected vehicles.

Acknowledgments

This work was supported by the UK EPSRC (grant no. EP/P005950/1), and the European commissions horizon 2020 framework programme (H2020/2014-2020) under grant agreement no. 734325 TESTBED project (http://testbed-rise.com/). The authors would also acknowledge the organizing committee of the sixth IEEE/CIC International Conference on Communications in China (ICCC2017) for their kind invitation sent to us for preparing this invited paper and also for their recognition of our scientific contributions in this research field.

References

[1] *5G Vision and Requirements*, White paper, IMT-2020 (5G) Promotion Group, May 2014.
[2] H. Zhang, Y. Dong, J. Cheng, Md. J. Hossain, and V. C. M. Leung, "Fronthauling for 5G LTE-U Ultra Dense Cloud Small Cell Networks," IEEE Wireless Communications," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 48-53, Dec. 2016.
[3] D. Soldaniz and A. Mazalini, "5G: The Nervous System of the True Digital Society," *IEEE COMSOC MMTC E-Letter*, vol. 9, no. 5, pp. 5–9, Sep. 2014.
[4] H. Zhang, J. Du, J. Cheng, and V. C. M. Leung, "Resource Allocation in SWIPT Enabled Heterogeneous Cloud Small Cell Networks with Incomplete CSI," *IEEE GLOBECOM*, Washington, DC, Dec. 4-8, 2016.
[5] S. Hung, H. Hsu, K. Chen, "Architecture harmonization between cloud radio access networks and Fog networks," *IEEE Access*, vol. 3, pp. 3019–3034, Dec. 2015.
[6] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, pp. 27–32, Oct. 2014.
[7] H. Zhang, X. Wen, B. Wang, W. Zheng, and Y. Sun, "A novel handover mechanism between femtocell and macrocell for LTE based networks," *IEEE ICCSN10*, pp. 228–231.
[8] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, Nov. 2014.
[9] Brent Hirschman, Pranav Mehta, Kannan Babu Ramia, Ashok Sunder Rajan, Edwin Dylag, Ajaypal Singh, and Martin McDonald, "High-performance evolved packet core signaling and bearer processing on general-purpose processors," *IEEE Network*, vol. 29, no. 3, 2015.
[10] X. An, F. Pianese, I. Widjaja, and U. Gunay Acer, "DMME: A Distributed LTE Mobility Management Entity," *Bell Labs Technol.*, vol. 17, no. 2, 2012.
[11] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. Vincent Poor, "Energy Efficient User Association and Power Allocation in Millimeter Wave Based Ultra Dense Networks with Energy Harvesting Base Stations," *IEEE J. Sel. Areas Commun.* , accepted, 2017.
[12] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Commun. Mag.*, 2017.
[13] H. Zhang, W. Zheng, X. Wen and C. Jiang, "Signalling Overhead Evaluation of HeNB Mobility Enhanced Schemes in 3GPP LTE-Advanced," *IEEE VTC*, pp. 1–5, Budapest, May. 2011.
[14] X. Jiang and N. Uday, "Performance analysis of mobility support in IPv4/IPv6 mixed wireless networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 962–973, Dec. 2010.
[15] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri and M. S. Alouini, "Handover Management in 5G and Beyond: A Topology Aware Skipping Approach," in *IEEE Access*, vol. 4, no. , pp. 9073-9081, 2016.