IEEE Access
The journal for rapid open access publishing

Received October 15, 2015, accepted November 19, 2015, date of publication December 17, 2015,
date of current version January 8, 2016.

Digital Object Identifier 10.1109/ACCESS.2015.2509638

Architecture Harmonization Between Cloud Radio Access Networks and Fog Networks

SHAO-CHOU HUNG[1], HSIANG HSU[1], SHAO-YU LIEN[2], AND KWANG-CHENG CHEN[1], (Fellow, IEEE)

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan
[2]Department of Electronic Engineering, National Formosa University, Yulin 632, Taiwan

Corresponding author: S.-C. Hung (d02942008@ntu.edu.tw)

This research was supported by the Ministry of Science and Technology, Taiwan ROC, and MediaTek Inc. under the contracts MOST 104-2622-8-002-002 and MOST 104-2221-E-002-082.

ABSTRACT To guarantee the ubiquitous and fully autonomous Internet connections in our daily life, the new technical challenges of mobile communications lie on the efficient utilization of resource and social information. To facilitate the innovation of the fifth generation (5G) networks, the cloud radio access network (RAN) and fog network have been proposed to respond newly emerging traffic demands. The cloud RAN functions more toward centralized resource management to achieve optimal transmissions. The fog network takes advantage of social information and edge computing to efficiently alleviate the end-to-end latency. In this paper, we conduct a comprehensive survey of these two network structures, and then investigate possible harmonization to integrate both for the diverse needs of 5G mobile communications. We analytically study the harmonization of cloud RAN and fog network from various points of view, including the cache of Internet contents, mobility management, and radio access control. The performance of transition between the cloud RAN and the fog network has been presented and the subsequent switching strategy has been proposed to ensure engineering flexibility and success.

INDEX TERMS 5G, fog network, cloud radio access network, RAN, heterogeneous network, edge computing, cloud computing, cache, radio resource management, mobility, mobile communications, vehicular network.

I. INTRODUCTION

Being deployed for several decades, mobile/cellular infrastructures successfully provide seamless and reliable streaming (voice/video) services for billions of mobile users. From GSM/GPRS, UMTS, to LTE/LTE-A, the transmission data rates have been enhanced a million-fold. The recent deployment of the heterogeneous networks (HetNets) [1]–[5] consisting of macrocells, small cells (femtocells, picocells), and/or further relay nodes, ubiquitously support basic multimedia and Internet browsing applications. In many occasions, it seems satisfactory to primitive human-to-human (H2H) communication applications using existing network architectures/technologies. However, to substantially facilitate human daily activities in addition to basic voice/video and Internet access services, achieving ''full automation'' and ''everything-to-everything'' (X2X) had been regarded as an ultimate goal not only for future information communication industry, but also for financial transactions, e-commerce, social communities, transportation, agriculture, and energy allocation [6]. ''Full automation'' implies a significant enhancement of human being's sensory and processing capabilities, which embraces unmanned or remotely controlled vehicles/robots/offices/factories/augmented/virtual reality, and immerse sensory human interactions of cyber-physical-social systems. The goal is to employ distributed-autonomous control to relieve/simplify the network control and evolutive, by which the resource utilization can be boosted in the dynamic complex networks, and be re-optimized after the major environmental changes [7]. On the other hand, X2X connection implies that diverse entities, including human and machines are able to form general-sense communities other than to H2H, such as social networks of human-to-machine (H2M) and machine-to-machine (M2M) facilitating the ultimate cyber-physical-social systems [8]–[10]. To name a few application scenarios include intelligent transportation systems (ITS) [11], volunteer information networks [12], Internet of Things (IoT) [13]–[15], smart grids [16], [17] and much more.

To enable these various applications, boosting the transmission data rates is just one of the diverse requirements. The performance in terms of end-to-end transmission latency [6], energy efficiency, reliability, scalability, cost efficiency as well as stability shall also be fundamentally enhanced to enable every aspect of mobile Internet services in an omni way. As the data traffic from Internet has gradually been dominating the traffic volume in mobile communication systems [18], in addition to the improvement of air-interface, migration to more efficient network architecture is definitely a must in technology development. Furthermore, a large portion of current traffic data is user-generated via social networks such as documents, pictures, videos, messages. Such data are circulated among the users' sides according to their social relationship, which precisely indicates the interplay between mobile communication networks and social network as shown in [19]. More precisely, most of the data are generated at the edge of networks, but stored and analyzed in the clouds. Consequently, the current Internet architecture which partitioning networks into layers will not be able to support these heterogeneous applications in an affordable cost [20]. All of these new technology opportunities suggest the need of evolving a state-of-the-art network architecture beyond ultra-efficient air-interface. This paper starts with the introduction of cloud radio access networks in Section II and fog networks in Section III, with their unique features to deal with future technology challenges, in spite of quite different philosophy behind. In Section IV, we propose harmonization H-CRAN with FogNet in architecture. With emerging cache technology into our harmonization, we analytically treat the network optimization as a whole in Section V. In addition to IoT/X2X and Internet content traffic, in light of automatic driving, unmanned vehicles, and service robots, we investigate mobility management and subsequent various handover design, to lay out the framework of this harmonization for 5G and future mobile communication networks in Section VI. The overall top-down design paradigm on top of resource access in this harmonization is presented in Section VII. Numerical results follow in Section VIII.

II. CLOUD RADIO ACCESS NETWORKS

The great success of mobile communications in past decades brings billions of user equipments and devices into networks to demand high bandwidth connections in the air. Following the breakthrough of multiple-input-multiple-output (MIMO) technology to approach the Shannon limit in past decades, the next generation of mobile communications can not solely rely on the enhancement from air-interface transmission. Introducing the HetNet architecture to facilitate the concept of small cells has been shown to further increase system capacity [3]. In the HetNet, in addition to the Macrocells formed by the existing eNodeBs, there are heterogeneous small cell networks (e.g. femto or pico cell) underlay or overlay the Macrocells. The motivation of such architecture is to increase the spatial spectrum reuse and increase the whole network efficiency. For this reason, the use of very dense and low-power-small-cells with highly spacial spectrum reuse is a promising way to allow handling such tremendous amount of devices [21], [22]. However, even though densely deployed small cells can provide shorter transmission distance and more efficient spatial reuse, it also introduces additional inter-cell interference problem and extra management issues. To compensate such those potential defects, heterogeneous cloud radio access network (H-CRAN) has been proposed on top of cloud radio access networks (C-RAN).

The C-RAN architecture can be traced back to the proposal by IBM [23] and further elaborated in [24]. The concept originates from the hierarchical network architecture of UMTS, in which each radio network controller (RNC) coordinates a number of NodeBs [25]. In UMTS, radio resource management (RRM) is conducted by each RNC, while NodeBs only perform physical signal transmissions/receptions. Although RRM is subsequently implemented to be performed by an eNodeB in LTE/LTE-A/EPC, this concept opens the designs of integrating the radio resource optimization in individual eNodeBs into a joint optimization. Coordinated multi-point (CoMP) transmissions/receptions are thus a practical paradigm of joint resource scheduling/optimization among multiple eNodeBs [26]–[28] and have been included in Release 11 of the specifications [29], [30]. The details of C-RAN were described in [24]. The baseband units pool (BBUs pool) and the radio head do not have to be collocated within an eNodeB. Instead, a number of BBUs and remote radio heads (RRHs) can be separated from an eNodeB to be massively deployed. Through the fiber-optic cables to connect an eNodeB and BBUs/RRHs, coverage of eNodeBs is therefore ubiquitously extended. This C-RAN architecture has attracted great research interests [31]–[34]. Later, Peng *et al.* [35] and Lei *et al.* [36] further revealed the conceptual realization of H-CRAN, as shown in Fig. 1. Compared to the C-RAN architecture, with the deployment of high power nodes (HPNs), the coordination of RRHs and HPN can be more efficient to alleviate the interference problem. Through the provided wired/wireless interfaces (i.e., S1, X2 and Un), not only BBUs/RRHs but also relay nodes, and HPNs are able to exchange information for joint resource scheduling/allocation.

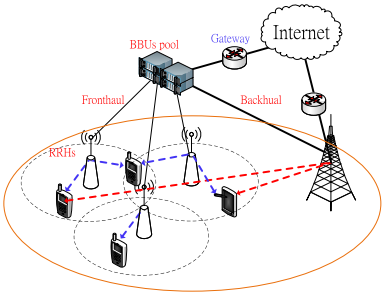


FIGURE 1. System architecture of the H-CRAN. RRHs can provide short-distance communication for UEs to improve transmission rate and HPN can provide ubiquitous connection to achieve seamless coverage.

Even though it has been analyzed that H-CRAN are more cost efficient [37], the entire cost structure behind H-CRAN remains worth further further investigation, due to potential IoT applications invoking significant growth in terms of the number of devices and traffic demand in the air. First, the number of heterogeneous devices (e.g., user equipments, sensors, vehicles, robots, etc.) is dramatically increasing in the future, particularly with the growth of IoT applications. The additional BBU pools are needed to support the high complexity of resource optimization in H-CRAN. Second, the traffic amount in the cloud radio increases explosively. More importantly, most data are generated by the users and is likely to be propagated via various social media platforms as interaction with friends (i.e. local in human networks to highlight the importance to treat mobile networks and social network together). With the treelike topology in the H-CRAN, all the information exchanged among edge users are not suitable for this characteristic. It will introduce additional amount of burden on the front-haul and back-haul links, especially in wireless type [38].

In the H-CRAN, three facts are generally ignored. (i) Traffic may be exchanged socially and locally. It is assumed that each packet from each edge entity may be delivered to another edge entity in the world under the H-CRAN. However, this assumption may not be generally practical, as more and more social applications only require data exchanges in close physical proximity. (ii) Each edge entity may exchange information with certain edge entities more frequently than other edge entities, while such entities of closer interaction can be considered as a social network in the general sense. Such a social network can be a set of webs, servers, individuals, or machines/devices. (iii) Downlink traffic to different edge entities or uplink traffic from different edge entities may be with strong correlations. For example, a large number of users may enjoy the same sport game streaming program simultaneously, and therefore the downlink traffic to these users is highly correlated. A group of densely deployed sensors measuring a common physical quantity may obtain a highly correlated result, and thus the uplink data to the cloud may also be correlated. It is reported to adopt in-network computation to significantly enhance spectrum utilization [39]. As a result, the technical merits of the H-CRAN simultaneously bring those engineering challenges to limit the performance of the H-CRAN at the edge side. This predicament thus motivates the concept of fog computing and thus fog networking.

III. FOG NETWORK

The Fog-Network (FogNet) was initiated by Cisco to enable the fog computing technology at the edge of the network [40]. The main characteristics of FogNet include ubiquity, decentralized management and cooperation [41]. FogNets are composed of a large amount of devices connecting to Internet like it device, wearable devices and self-driving vehicles, etc. These devices form many ''mini clouds'' at the edge of the network and manage themselves in a distributed way. On the contrast to the cloud computing, the fog computing facilitates processing/computing capabilities at edge entities, by which not all information for performance optimization should be delivered to the cloud. Only the tasks (and corresponding information for optimization) those cannot be well processed by edge entities are handled by the cloud. For example, the users in the FogNet can release some of their own computing/storage capacity to support their neighboring devices. The users need not download the data from the core network, instead, they just download the required data from their neighbors, like the adjacent small cell network, or other mobile devices through proximity direct links. Without complex routing in the core network, it can be expected that the reduction of end-to-end latency is reachable. The FogNet, therefore, may significantly alleviate the computing and routing burdens in the cloud-part of networks to achieve the scalability.

Aryafar *et al* pioneered the facilitation of fog computing into a new type of network architecture known as the FogNet [42]. Under the FogNet, each edge entity having social messages to be exchanged among other edge entities does not rely on traffic relay via the cloud, while edge entities in close physical proximity are able to locally share messages. This design leads to the concept of socially-aware traffic managements to significantly decrease the traffic amounts to be supported by the cloud. Recently, radio access technologies (RATs) such as device-to-device (D2D) communications [43]–[47] and using small cells [48] as smart data/traffic routing gateways are successful practices of the FogNet. When a group of edge entities have highly correlated traffic to be delivered to the cloud, each entity does not have to upload traffic individually. Instead, the common part of traffic is delivered once by a single edge entity. On the other hand, when the cloud has highly correlated traffic to be forwarded to multiple edge entities, the cloud does not forward traffic individually to each edge entity. Instead, the cloud only selects one edge entity to forward traffic, then the selected edge entity autonomously shares traffic with other edge entities. Consequently, the amounts of traffic supported by both the fronthaul and backhaul links can be largely alleviated.

Although the FogNet provides considerable technical virtues to potentially tackle the issues of complexity, scalability, and heavy traffic burdens in the H-CRAN, new challenges emerge at the same time. First, although traffic can be socially shared among edge entities, there is no guarantee that all edge entities needing this traffic are able to successfully receive this traffic. Therefore, reliability of data delivery turns out to be the primary concern. Second, the mobility management and service continuity may not be sufficiently supported in the FogNet. Third, for the H-CRAN, as the optimized resource scheduling/allocation is the key requirement, interference can be well rejected/mitigated. However, due to the lack of effective resource coordination among edge entities, interference may drastically impact on the performance of the FogNet.

IV. IMPACT OF TWO SYSTEMS WITHOUT HARMONIZATION

These two effective networking strategies of somewhat opposite approaches can create further challenges in their co-existence, which primarily include storage management, load balancing, and interference.

1) STORAGE MANAGEMENT

It is expected that the FogNet can help to reduce the data loading in the H-CRAN. Ideally, we may expect that the local information going around in a small set of entities in the social network are stored in the local FogNet and just those popular data/media would be uploaded into the cloud-part. However, without proper exchange of control signaling that takes time delay, the cloud-part (H-CRAN) and fog-part (FogNet) do not know whether another system stores such data or not. It may result in either both sides store or both sides do not. In this way, the storage resources cannot be utilized optimally and redundant traffic flows in the network. To solve this problem, a possible way is to identify the social relationship among the FogNet and to report this information back to the H-CRAN to help both appropriately allocate the storage resources and consequently reduce network traffic by popular data.

2) LOAD BALANCING

Under the coexistence of H-CRAN and FogNet, there are multiple types of devices ranged from the devices equipped with a large power transceiver like vehicles to the device relying only on energy harvesting (EH) such as sensor networks. The wireless links include cellular (like UMTS or LTE), WiFi, mmWave [49], etc. All differ in coverage, bandwidth, and capacity. To optimize the performance of the whole network, an important issue is the load balancing. Load balancing focuses on how to guarantee that all the network can optimally operate under their own way. In [50], it has been pointed out that the most intuitive way like always connecting to the network with the best signal-to-interference-plus-noise-ratio (SINR) may not be the most efficient scheme. Therefore, it is critical to design a handover scheme beyond SINR based for devices to choose between H-CRAN and FogNet.

3) INTERFERENCE

An indoor device under co-existence of H-CRAN and FogNet may obtain/send desirable data via D2D links, WiFi, and/or cellular systems. However, such a scenario may jeopardize the functions of a local FogNet. For example, the D2D links, underlaying in the cellular network, may suffer from the interference from the cellular network. The devices connected with WiFi may also be interrupted by the cellular network with Licensed-Assisted Access. It has been expected that there will be 80% of traffic is generated by the devices in the indoor environment. If the H-CRAN cannot provide a suitable way to control the interference to the FogNet, all these data can be only uploaded to the cloud-part and crash the whole systems.

TABLE 1. Comparison of H-CRAN and FogNet.

	H-CRAN	FogNet
Management	Centralized	Distributed
Deployment Cost	High	Low
Social Relation of Data	Less Utilization	More Utilization
Burden on Core Network	Large	Low
Resource Optimization	Global Optimization	Local Optimization
Mobility Management	Easy	Hard
Latency	High	Low
Reliability	High	Low

V. HARMONIZATION OF H-CRAN and FogNet

To compare the H-CRAN and the FogNet, the features of these two architectures are summarized in Table 1, while each of them is effective to deal with certain new traffic patterns. All in all, the H-CRAN focus on the global resource allocation/utilization optimization through a centralized way and the FogNet facilitates the information exchange and computation at the edge of the network. The design philosophy of the H-CRAN and FogNet appear opposite each other. However, instead of arguing preference, we note that different technical merits in the H-CRAN and in the FogNet may lead to a complementary harmonization. The necessity of harmonization includes the following:

1) For the devices of a small form factor like EH in FogNets, only very limited energy or power is available for transmission. Without coordination between cloud-part and fog-part, these small-size devices may suffer from severe interference and the FogNet may collapse.
2) Wireless fronthual (backhual) [51] is necessary in the region where the cost of building infrastructure is large. By offloading the burden from the H-CRAN, the FogNet can increase the feasibility of wireless fronthual (backhual) and therefore decrease the cost of the network simultaneously.
3) The flexibility of radio resource utilization and the latency performance can be improved via the FogNet. The BBU pools can allocate more radio resources to the hot spots; for example, the traffic storm in sensor networks due to some urgencies or emergent accidents like tsunami or earthquake. BBU pools can also broadcast the information to smart phones to save the time to respond.

To enable the harmonization, Peng *et al* bring the idea of FogNet into the H-CRAN architecture by taking the correlation among traffic to/from different edge users into account [52]. The hybrid architecture FogNet-HCRAN network (abbreviated F-CRAN in the following) are illustrated in Fig. 2. F-CRAN is composed of cloud-part and fog-part. In the cloud-part, there exists high power nodes (HPNs) to cover a wider geographical area and RRHs to provide the conventional functions in C-RAN. The devices can connect to Internet through the cloud-part or the fog-part. The fog-parts are composed of all kinds of devices that can provide services to other devices, or furthermore, including cloud-part.

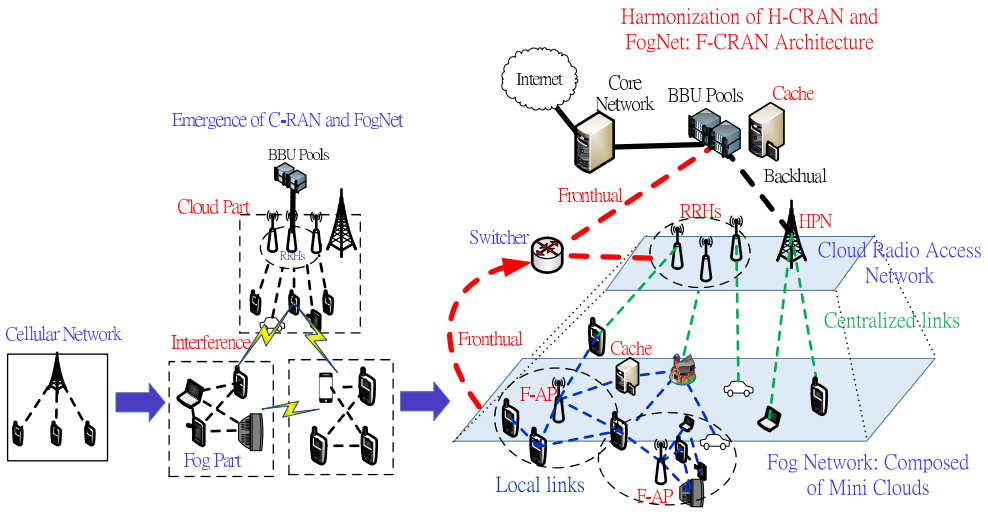


FIGURE 2. Evolution of the architecture of F-CRAN. The whole network is composed of cloud and fog-part., that is H-CRAN and FogNet. The cloud-part is composed of RRHs and HPN. The aim of cloud-part is to provide the ubiquitous connection service to all the device. The fog-part is composed of multiple mini clouds coexisting with the cloud-part. The devices can exchange data locally or connect to the cloud-part through F-APs.

These "powerful" devices form the multiple mini clouds at the edge of the network. F-APs may play the roles as the coordinators in each of these mini clouds. The F-APs can be interfaced into the BBU pool through the fronthual links. F-APs can provide services from physical layer management, like basic radio resource management, to the higher layer managements, like cache memory management. Since large amounts of data or control signals are processed in the FogNet, the burden of fronthauls and BBU pools are alleviated. If these services are removed from the F-APs, the F-APs degenerate into the RRHs and all the computing load and file exchange burdens are shifted to BBU pool or even to the core network and cloud, which may significantly increase the latency resulting in poor quality of user experience. To reduce the end-to-end latency and take advantages of social relationship among networks, the technique of cache can be implemented both in the cloud and fog-part to reduce the burden of the core network [53], [54].

However, whether these two eagles fuse into one of greater power or turn into a turkey remains unclear. In the following, we further investigate sufficient conditions to adequately integrate the H-CRAN and the FogNet. We analytically explore a unique top-down system design based on the proper allocation of all kinds of resources in the entire F-CRAN, according to the need of users or applications, particularly the caching for social media and virtual reality, and the resource/mobility management for radio access control of diverse service requirements. Our results establish the foundation toward a network architecture of 5G mobile communications.

VI. CACHING IN APPLICATION LAYER

A. CACHING IN WIRELESS NETWORKS

Caching mechanisms are originally a common methodology to reduce traffic volume and meanwhile access latency in computer systems like a CPU and database centers [55]. In late 1990s, caching has been implemented in the wired web application systems, called web cache systems (or HTTP cache systems), to store copies of heavily accessed documents in the networks, thereby reducing bandwidth usage, server load, and improving web retrieving stability, latency and quality of service (QoS) [56]–[58]. In web caching systems, a client could store web contents for later reuse, called a forward position system; moreover, a web server (*e.g.* a search engine) may also cache copies of web contents in content delivery networks (CDNs), called a reverse position system. These two caching mechanisms operate together in web caching systems as complements to each other, making a successful and efficient web content retrieving system. The reason for the efficacy behind caching mechanisms is that most of the traffic flows in the Internet are attributed to a relatively small part of the data or contents in the networks, a phenomenon which can be traced back to web requests and proxy traces [59]. In fact, the characteristics of these popular contents have been verified to follow mathematical forms as *power-law distributions*, meaning that the probability of attaining a certain content c_k of rank k (*i.e.* the k^{th} popular content) is proportional to $k^{-\zeta}$, with ζ greater than or equal to 1. That is $Pr\{c_k\} \propto k^{-\zeta}$. Power-law distribution explicitly implies that only a few higher ranks of data occupy most portion of the traffic volume in the Internet. Identifying the most popular Internet contents, and caching the contents in the networks therefore greatly reduce the traffic volume and latency, since for most Internet users, they need not to acquire the contents from remote data centers.

In the past decade, the communications industry has experienced a dramatic variation: mobile Internet traffic gradually dominates wireless networks [18], [60], with distinct features compared to traditional telephony traffic and short message service (SMS), which are usually transmitted to

and needed by a single user in the network. People start to access social media, streaming videos, and other contents that possess traffic characteristics of Internet contents via mobile devices. This fundamental dissimilitude on traffic characteristics gives designers opportunities to introduce caching in wireless networks and hence motivates us to redesign our wireless mobile network together with the core network to support caching mechanisms. Niesen *et al* and Maddah-Ali *et al* have developed information-theoretic frameworks to analyze the performance limit of caching gain by coded caching with respect to cache size available to mobile users [61], [62], revealing positive outcomes of caching in wireless networks. Consequently, the engineering implementation of caching becomes an urgent issue to network designers and researchers.

Even though caching mechanisms have brought extraordinary success in wired networks and theoretical foundation in computing has been well established, there still exist challenges to implement caching mechanisms in wireless networks, especially the cache in the radio access network to abate the traffic volume from remote data centers, with an aim of reducing latency as well. Different caching mechanisms have been proposed; *e.g.* in-network caching in the core network and base stations (BSs) to optimize data retrieving latency for mobile users [63]; in [64], the authors discuss caching at BSs and at mobile users as two cases of caching utilization. However, if caching is considered in a more general way as a sort of resource in the radio access network, rather than separate storage capacity at devices and infrastructures, then the utilization of caching turns into resource allocation problems in wireless networks [65]. The resource allocation perspective allows us to ruminate the harmonization of caching utilization in centralized H-CRANs (cloud-parts) and distributed FogNets (fog-parts), to result in a new cache utilization problem in F-CRAN.

B. CACHING UTILIZATION IN F-CRAN

Regarding caching as a general resource than wireless bandwidth facilitates the design of F-CRAN, which satisfactorily addresses three difficulties when implementing caching in wireless networks. First, considering caching at the infrastructures like BS, the backhaul traffic volume would indeed be greatly reduced, and thus enhances the latency since the transmission bottleneck on the backhaul is alleviated. However, the limited wireless cellular bandwidth still acts as another bottleneck for data acquisition when the number of mobile users increases [66]. An efficient way to tackle this problem is to allow D2D communications for direct content sharing, and hence abate the need of wireless cellular bandwidth. We expect similar methodologies for virtual reality (VR) traffic in the future. Introducing D2D communication in the networks is actually based on caching in the mobile devices; therefore FogNets are reasonably regarded as an auxiliary to reduce the burden of the air interface; making FogNet a promising design for the future realization of mobile networks. Second, in FogNets, the mobility of user devices might jeopardize the efficiency of D2D communication and hence the performance of caching due to the unstable nature of wireless communication. To optimally utilize device caching is the main challenge. However, with the aid of H-CRAN, the mobility and interference of devices could be managed, since H-CRAN integrates all the information and is equipped with mighty computational power. Furthermore, some coded caching protocols could also be implemented, rendering caching a more operative methodology. Third, the information collection and instantaneous monitoring of the characteristics of the contents becomes an indispensable part in CDNs. The highly-centralized H-CRAN provides an opportune solution for this problem, making caching at infrastructure and traffic monitoring viable. These three explanations cause H-CRAN and FogNet perfect complements to each other; the coordination of H-CRAN and Fog-Network in the F-CRAN is thus totally different from the early studies about D2D on relaying purpose [67], [68].

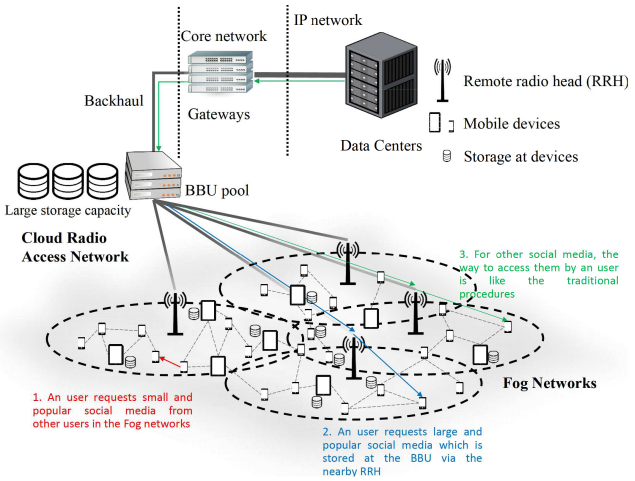


FIGURE 3. Under the F-CRAN architecture, there are three different paths for data retrieving: (1) retrieve the data via direct connections to other users (2) retrieve the data from cached copies at BBU pool (3) retrieve the data from the cloud network (traditional path). FogNet helps to reduce the burden of cloud-part by the first path.

The complete scenario of the F-CRAN system for caching as resource utilization is shown in Fig. 3. As H-CRAN and FofNets are combined via cahcing, three different main connections for the mobile users exist: (1) direct links to other devices (D2D communication) (2) devices to caching in BBUs pool (3) device to the cloud network (conventional link). To summarize, in order to integrate the H-CRAN and FogNets via these three different connections, the overall cache utilization can be generalized into the optimization of resources in a network, by incorporating infrastructure caching (H-CRAN cache) and device caching (D2D communication) into a new networking design scenario of joint optimization on the backhaul networking, storage, computing, and radio resource allocation in the air-interface.

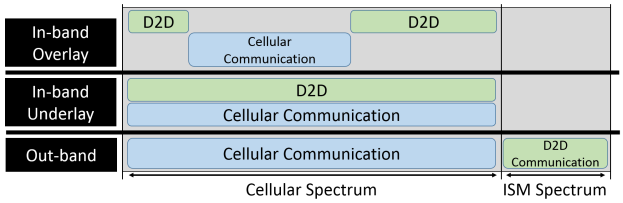


FIGURE 4. Schemes of in-band overlay, in-band underlay, and outband D2D. In the formulation, we adopt in-band overlay scheme and optimize the usage of D2D spectrum as caching in devices.

C. GRAPHICAL NETWORK MODEL AND CHARACTERISTICS OF INTERNET CONTENTS

Let us consider a straightforward scenario with one F-AP and mobile devices, where the mobile devices forms a FogNet, as illustrated in Fig. 3. This scenario could be extended to more than one F-AP under the management of the BBU pool. Since under the structure of H-CRAN, the cell size shrinks, leading to smaller mean distances among users, we adopt a random graph model by assuming that for any two users under a F-AP, there is a probability p that they can communicate and share contents with each other. Thus, we modelled the network topology of direct device communication by Erdös-Rényi (ER) random graph [69]. This random graph model offers a performance upper bound for device caching and D2D communication supported by the FogNets, since ER model overweights the connectivity of two mobile devices and ignore possible clustering structures formed by mobile users. Moreover, ER model also provides mathematical tractability for performance analysis.

The BBU pool has a given cache size M memory units (MU), and the storage sizes at the devices are finite. For an Internet content c_k of rank k, its file size is s_k, with $k = 1, 2, \cdots, K$, where K is the number of active Internet contents. As mentioned in Sec. VI-A, the popularities of Internet contents being requested in a network during a period of time can be characterized by power-law distributions [59], $p_k = Pr\{c_k\} = H_{K,\zeta}k^{-\zeta}$, with $\zeta > 0$ and $H_{K,\zeta}$ being the normalization factor. The popularity is reasonably assumed to be constants in the observed period. For the size of the Internet contents, it is assumed that s_k follows Log-Normal distribution with mean μ and variance σ^2 [70]. The powerful H-CRAN servers and the core network are able to acquire the information of popularities $\{p_k\}$ and sizes $\{s_k\}$ of the active Internet contents by traffic monitoring and statistics gathering. For instance, the H-CRAN server can record the type of data from the logic channels and transport channels and the uniform resource locator (URL) requested from mobile users.

Here, as suggested in [65], caching utilization among users in FogNets is actually spectrum utilization of D2D bandwidth. The spectrum utilization scheme for D2D communication can be categorized into three different types: in-band overlay, in-band underlay, and out-band (as visualized in Fig. 4).

1) In-Band Overlay: D2D devices utilize a reserved fraction of cellular spectrum. Therefore, the devices need not to perform spectrum sensing.
2) In-Band Underlay: D2D devices and cellular traffic share the same spectrum. However, D2D devices need spectrum sensing to control their interference to cellular network users under a certain threshold.
3) Out-Band: There is a part of unlicensed spectrum available for the D2D devices and all the D2D links operate in this band only.

In this paper, we consider the in-band overlay scheme as a conceptual discussion about spectrum utilization to avoid complicated spectrum sensing issues, as studied in [71]. Therefore, the total available spectrum in a F-AP cell is divided a fraction W_{D2D} for D2D communication and the other fraction W_C for original cellular downlink traffic.

D. FORMULATION OF CACHING UTILIZATION

To optimally utilize caching resource in F-CRAN, we need to decide which Internet content should be cached in the H-CRAN and which should be shared in FogNets through D2D links to minimize the total traffic volume, which consists of backhaul traffic volume F_B and downlink traffic volume F_D. Therefore, we introduce two binary variables with state space $\{0, 1\}$. δ_k^C decides whether the Internet content c_k should be cached at the BBU pool; δ_k^D decides whether c_k should be obtained using D2D communication in the FogNets. Therefore, the backhual traffic accounts from the requests of contents that are neither cached at the BBU pool, nor can be shared among users. Similarly, the downlink traffic volume comes from the requests of contents that are not cached at the BBU pool. We use scalarization method [72] to optimize the two traffic simultaneously, yielding the following optimization problem:

$$\underset{\{\delta_k^C, \delta_k^D\}}{\text{minimize}} \; F_{total} = \alpha F_B + \beta F_D \tag{1}$$

$$\text{subject to} \sum_{k=1}^{K} \delta_k^C s_k \leq M \tag{2}$$

$$F_D \leq W_C \tag{3}$$

$$\sum_{k=1}^{K} N p_k s_k - F_B - F_D \leq W_{D2D}, \tag{4}$$

where the constants $0 \leq \alpha \leq 1$ and $\beta = 1 - \alpha$ represent the importance of backhual traffic and downlink traffic respectively.

The constraints of the optimization problem in (1) should take the following issues into account. First, the utilization of caching resource at the BBU pool should not exceed its storage capacity M, as given in (2). Moreover, the volume of downlink traffic and D2D communication traffic in FogNet also should not exceed pre-allocated fraction W_C and W_{D2D}, as respectively described in (3) and (4), where $\sum_{k=1}^{K} N p_k s_k$ is the traffic volume of the active Internet Contents. In other words, we maximize caching utilization in F-CRAN by minimizing the traffic volume between FogNets and BBU pool, and between BBU pool and the data centers.

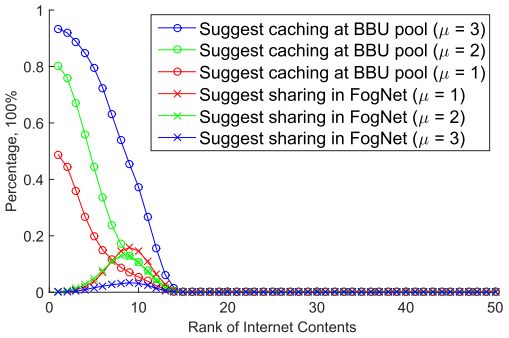


FIGURE 5. Optimal caching utilization in F-CRAN for $\alpha = \beta = 0.5$. It is suggested that H-CRAN and the FogNets should cooperatively store the Internet contents according to their characteristics.

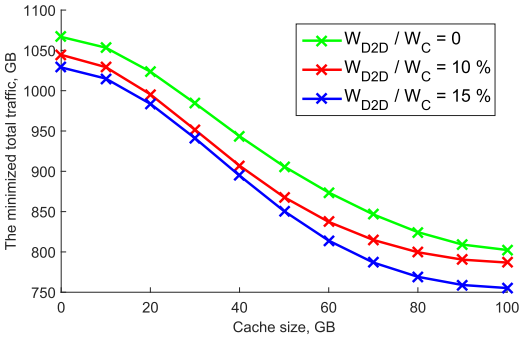


FIGURE 6. Minimized total traffic under optimal caching utilization for different ratios of W_{D2D} and W_C. It is clear that with the assistance of FogNets, the total traffic volume is further alleviated.

E. OPTIMAL HARMONIZATION OF CACHING UTILIZATION IN F-CRAN

The optimization problem in (1) is a binary integer programming problem. To carry out numerical results of optimal caching utilization, we assume 100 mobile users under a F-AP, and 50 Internet content actively requested by the users. The popularity of the contents follows power-law distribution with $\zeta = 2$. The caching capacity at the BBU pool is $M = 30$ GigaBytes (GB). For the existence of D2D sharing links, we assume that $p = 0.1$. We consider the representative case which the traffic load at backhaul and at downlink is equally heavy ($\alpha = \beta = 0.5$) and μ varies from 1 to 3 with $\sigma^2 = 1$. We ran the optimization problem for more than 10, 000 times and gathered statistics of the optimal caching utilization of Internet contents, *i.e.* to count the normalized sum of δ_k^C (number of Internet contents cached by the BBU pool) and the normalized sum of δ_k^D (number of Internet contents shared in the FogNet). The two statistics, shown as lines with circle markers and lines with cross markers respectively in Fig. 5, offer us the suggestions of optimal caching utilization regarding F-CRAN. In the figure, it is clear that for the higher rank Internet contents, caching them at the BBU pool is suggested to be optimal utilization of the caching capacity. Nonetheless, for the rest of the Internet contents and especially with small file sizes, it is suggested that the requests of them should be directly satisfied in the FogNet. The essence is that although FogNet indeed help caching and traffic load releasing, the unstable nature of FogNets should also be calculated; thus, for Internet contents with small file sizes, the chance for direct sharing is much higher, resulting in this optimal caching utilization. The minimized traffic in (1) under optimal caching utilization with respect to the different storage size at the BBU pool is shown in Fig. 6. As the spectrum of D2D communication increases (larger W_{D2D} to W_C ratio), the role of FogNets becomes more important; the benefit of introducing FogNets into H-CRAN also increases. The mitigation of traffic loads as well improves the retrieving latency for Internet contents, making another contribution from the optimal caching utilization in F-CRAN. The results of harmonization between H-CRAN and FogNets suggest that F-CRAN truly provides a more general solution to traffic offloading via caching implementation. Additionally, this harmonization philosophy of F-CRAN could be easily and effectively extended to other traffic types in the future than Internet contents, like VR traffic and popular road traffic information in vehicular networks.

VII. MOBILITY MANAGEMENT

A. VEHICULAR NETWORK IN FogNet

ITS emerges as an even more important feature in the 5G mobile communication systems than before. Over these years, ITS has been developed aiming at improving road traffic safety and at automated driving in the vehicle industry. Therefore, it is expected that the passengers have the same demand concerning connectivity performance in the vehicles as at home and work [73]. On the other hand, the global market is expected to reach 130 billions by 2019 [74]. These connected vehicles could provide alternatives to alleviate the vehicular traffic congestions via intelligent traffic control and managements [75]. To enable the vehicular communication, the Federal Communications Commission (FCC) in the United State has allocated 75MHz centered at 5.9GHz for ITS system [76]. Therefore, the success of vehicular network could be an essential part toward the success of the 5G mobile networks due to the ubiquitous deployment of cellular systems.

The goal of the vehicular networks is to provide human-safety services which include road safety information exchange, emergency alarms, traffic management, localization and navigation, and even unmanned intelligent driving [73]. All these applications involve a large amount of information change and extremely low end-to-end latency transmissions. To support these services, the early vehicular network protocol 802.11p [77] has been proposed. The 802.11p mainly focuses on the vehicle-to-vehicle (V2V) communication through wireless link based on the dedicated short range communication (DSRC). It is similar to the D2D links between vehicles. Easy deployment, low cost to construct and to maintain, and capability to accommodate the ad-hoc mode V2V scenario, are its technical merits [78]. However, it suffers from intermittent and short-live

connectivity between vehicles. Its scalability, geographic coverage, insufficient radio access to meet quality of services (QoS), and lacking overall network architecture to satisfy ultra-low latency to the cloud or to facilitate appropriate control functions of physical entities, become the major concerns. These challenges motivate us to consider other types of communication network, to achieve the safety and latency requirements. Therefore, the establishment of the heterogeneous vehicular networks and corresponding proper management of the resources in such networks arise as the new technology challenges regarding vehicular networks [79] and even networking for service robots. All these suggest new show-case opportunities for 5G cellular networks.

Nevertheless, the realizations of the HetNets are always a challenge for wireless network engineers. These challenges include interference coordination, radio resource allocation, cooperative radio signal processing and frequent handover problem. To overcome these obstacles, F-CRAN [52] may serve as an attractive solution. In the F-CRAN, not only the HPNs provide ubiquitous connections to the mobile devices, but also F-APs can provide short distance connections to the devices at the edge of the networks, by which, the performance of the vehicle network can be further improved. For example, the latency of traffic information can be achieved by downloading from the nearby F-APs instead of the remote cloud data center. Furthermore, the vehicles can offload or exchange the data with the FogNet to reduce the burden of the cloud network.

B. HANDOVER SCHEME IN F-CRAN

Handover and subsequent mobility management are of critical importance in the mobile communication networks, especially in the highly dynamic environment like vehicular networks. Particularly, with massive deployment of small cell networks like F-APs, the handovers happen more frequently and result in a heavy burden on fronthaul and core networks [80]. To achieve seamless services for the vehicles with high mobility, in [81], a survey of F-CRAN architecture is provided, which discusses how high mobility devices should be served by macro cell like HPN network and low mobility should be serviced by small cells like F-APs. With multiple access networks, traffic flows can be balanced to avoid congestion and performance degradation. Therefore, not only the switches between different access points in the same network (horizontal handover) but also between different networks (vertical handover) are urgently wanted.

The most common approach is designed based on the straightforward parameter, the received signal strength (RSS) [82]–[84]. By detecting the RSS through reference control signals, the mobile devices can access the best wireless network while entering new cells. More details can be found in the survey paper [85]. However, these RSS based handover algorithms might not be satisfactory under interference, say to cause unnecessary handover such as ping-pong effects, which is severer in the small coverage cell [86]. In addition, these RSS based algorithms are designed based on the assumption that the mobile devices can communicate with only one access point. However, with the facilitation of CoMP, the F-CRAN architecture can support a mobile device accessing multiple accessing points. The CoMP technique is presented to mitigate inter-cell interference and QoS improvement in highly density network. Through feeding all the information to the centralized processing server, the interference among different small cells like F-APs can be mitigated. In [87], a comprehensive introduction and performance evaluation are provided, and the evolution of the vehicular networks is illustrated in Fig. 7. By utilizing CoMP techniques, all mobile devices can access the F-APs at the fof-part of F-CRAN in the same frequency without suffering from the interference, which implies that no need for hard-lined cell definition in the F-CRAN architecture. Therefore, RSS based approach may not be the best choice for the F-CRAN architecture to tackle the handover problem.

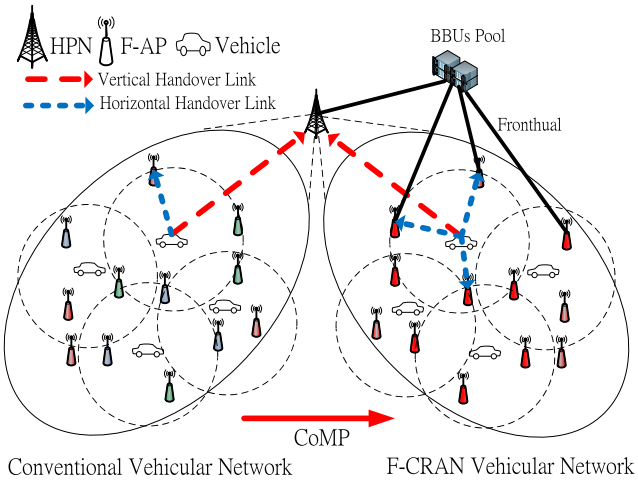


FIGURE 7. The evolution of the vehicular network under the F-CRAN architecture. We use different colors of access points to illustrate different frequency bands. In the conventional vehicular network, all the vehicles can connect to one access point with a certain frequency band to avoid interference. In the F-CRAN architecture, CoMP allows all the vehicles can connect to multiple ones in the same frequency without interference. Therefore, the cell concept is alleviated.

Please recall that the small access points are also limited resources in the wireless networks. Because the performance gain of the CoMP highly depends on the perfect knowledge of the channel state information and frequent control signal exchanges, if all the mobile devices utilize all the small access points, it may increase the complexity of interference cancellation algorithms [81]. On the other hand, for the mobile devices of urgent operation like having some emergent accidents, it is reasonable to allocate more accessing points for them to guarantee the emergency information can be transmitted successfully. Therefore, accessing all available small cell networks may not always be a good strategy throughout the entire network. A different thinking to design the mobile networks under such scenario may be needed, If we regard the access points of the mobile network as another kind of resource for mobile devices, then the handover can be

generalized to a dynamic resource allocation problem. In this new perspective, the resources to be allocated are not radio resource units in the spectrum anymore, but the access points like the RRHs or F-APs.

Therefore, we stand on the viewpoint of resource allocation to innovatively reformulate the vertical and horizontal handover in the F-CRAN architecture. With the realization of CoMP, the vehicles or mobile machines like robots can request more resource at the fog-part, like roadside access points (RAPs) or other small network access points F-APs in F-CRAN. While it is necessary, the vehicles can ask for more resource from the cloud-part of the F-CRAN system.

C. TIME DYNAMIC RESOURCE ALLOCATION FOR HANDOVER SCHEME

As previous description, the handover can be commonly considered as a detection mechanism. The devices need to make a decision whether the RSS is over the threshold or not. The handover problem and resource allocation are usually studied separately. Nevertheless, the direct impact of the handover problem of the resource allocation is the mechanism to allocate the resource for the newly arriving devices. It is possible that the network with the best SINR may not have remaining resource for the upcoming devices. To solve these two problems in a unified framework, Stevens-Navarro *et al.* design the vertical handover algorithm based on the Markov decision process (MDP) to dynamically allocate network resource for the mobile devices [88]. Qin *et al.* integrates the handover scheme into the time allocation scheme which simultaneously reach fairness and optimal link gains [89]. Generally speaking, little attention has been paid in the literatures to jointly optimize the radio resource allocation, allocation of access points in mobile network, and handover together. We will be working on this emerging technology challenge in the following.

D. STOCHASTIC NETWORK MODEL

We consider the scenario that the mobile machines like vehicles or robots connected to the Internet through the F-RAN as shown in the Fig. 7. The networks are composed of mobile devices, F-APs serving locally and HPN covers the whole network. These mobile devices can access the F-APs or the HPN through total M channels. To describe the connection between the vehicles and F-APs, a common approach is to model the transmission region as a circle. This is the reason that the we only utilize the wireless link with good long-term performance, which is dominated by the distance to the receiver. In this model, the candidate F-APs for vehicles to handover are only in the circle with the radius R. To find the general performance, we assume that all the vehicles and F-APs are uniformly distributed in the infinite flat area. We also assume that all the vehicles are moving straightly with velocity $v(meters/s)$ and choosing different direction randomly.

Under this model, the quality of wireless links depends on not only channel fading, but also the interference from other vehicles in the same channels. To connect the quality of wireless links and the interference, a popular approach is to describe the quality of wireless links as outage probability by stochastic geometry [90]. In the stochastic geometry, the power of the interference from a single source depending on the distance to the receiver and follows $d^{-\alpha}$, where α is a path loss effect coefficient. With this assumption, we can see that the interference may go to infinity while the distance d is close to zero, which is not possible in the real world. Therefore, the stochastic geometry analysis shall supply the performance lower bound of the outage probability.

E. HANDOVER PROBLEM FORMULATION

We define the queue length $U(t)$ as data in the queue at tth time slot. To guarantee the system stability, that is, $\overline{U} \triangleq \lim_{T\to\infty} 1/T \sum_{t=1}^{\infty} U(t) < \infty$, the system should allocate more resource to increase service rate $u(t)$ while it is necessary. At each time slot t, the service rate $u(t)$ of the mobile devices is determined by the number of connected communication links with F-APs and the decision space is denoted as $\mathbb{D}_t$:

$$\mathbb{D}_t = \begin{cases} \{1, \ldots, N(t)\}, & \text{if } N(t) \neq 0 \\ \{0\}, & \text{if } N(t) = 0. \end{cases} \tag{5}$$

We denote $n(t) \in \mathbb{D}_t$ as the number of connected F-APs at the time slot t. Therefore, the number of serviced packets at each time slot $u(t)$ can be expressed as

$$u(t) = \sum_{i=1}^{n(t)} \mathbf{1}_i, \tag{6}$$

where $\mathbf{1}_i$ is the index function of the wireless link corresponding to the ith F-AP. $\mathbf{1}_i = 1$ if the SIR of the ith link larger than the threshold θ.

From here we can note that the *horizontal handover* is the process that the vehicles or mobile devices ask more wireless link resource from the local F-APs. Once more resources (F-APs) are allocated to the vehicles, it has faster service rate or higher bandwidth of transmission. However, other vehicles may not get the necessary resource to stabilize its own queue. Therefore, a suitable solution is that all the vehicles minimize its utilization of APs resources while simultaneously they still can stabilize its own traffic queue. On the other hand, it may be possible that even though all the available resources are utilized by vehicles and mobile users, it is still possible that certain queue(s) cannot be stabilized. For example, such a situation becomes more likely for the data traffic flow from significantly increasing vehicles and mobile users in rush hours. A possible alternative to resolve this dilemma is to execute *vertical handover*, that is, the vehicles borrow more resources from cloud-part (*i.e.* HPNs of F-CRAN) to offload data traffic. Nevertheless, the vertical handover is much more complicated than horizontal handover. To avoid the additional burden of the cloud side and the complexity of vertical handover, the vehicles should not access the cloud as possible as they can. To describe

the *horizontal handover* and *vertical handover* problem more precisely, we formulate the mathematical problems as follows.

Horizontal Handover Problem:

$$\min_{n(t)\in\mathbb{D}(t)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} n(t)$$
$$\text{subject to } \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(U(t)) < \infty, \tag{7}$$

Vertical Handover Problem:

$$\max_{n(t)\in\mathbb{D}(t)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \lambda_p(t)$$
$$\text{subject to } \overline{n} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} n(t) \le N_{av}$$
$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(U(t)) \le \infty, \tag{8}$$

In (7), we try to minimize the time average utilization of $n(t)$, the *horizontal handover* time and simultaneously keep the data stable, that is, the constraints in (7). In (8), $\lambda_p(t)$ is the data flowing through F-APs networks at time t. We try to maximize time average of the data flowing through F-APs networks, which equally means that minimizing the utilization of vertical handover. The first constraint in (8) comes from the fact that the number of the utilized F-APs cannot more than the existing number of F-APs.

F. LYAPUNOV OPTIMIZATION

In general, the most popular way to solve the problem like (7) and (8) is to formulate by MDP. MDP can help us to find the best tradeoff between transmission delay and resource utilization [91]. Even though MDP can approach the best tradeoff between resource utilization and delay, it may take lots of time to find the optimal solution, especially in the scenario that the number of the states are large [92].

In fact, there is always an intuitive way to solve this type dynamic resource allocation problem, for example, prior allocating the resource to the queue suffering severe delay. However, there is always a tradeoff between resource utilization and the network performance. Though more resources suggest better network performance, it is difficult to intuitively conclude the tradeoff. Please recall that *Lyapunov optimization* [93], originating from *Lyapunov drift theory*, is used to develop dynamic control algorithms. It introduces the *drift-plus-penalty theorem* concept into the control algorithm. That is, it gives a cost weighting V to the network utility and tries to optimize *drift-plus-penalty* function subject to the queue stability. For example, we give the cost weighting to the utilization of $n(t)$ to maximize the service rate with the *horizontal handover*. That is,

$$\max_{n(t)} \mathbb{E}(2U(t)u(t) - Vn(t)). \tag{9}$$

In the *vertical handover*, the problem can also be converted into the similar form as following.

$$\max_{n(t)} \mathbb{E}(U(t)n(t) - X(t)n(t))$$
$$\min_{\lambda_p(t)} \mathbb{E}\left(2U(t)\lambda_p(t) - V\lambda_p(t)\right). \tag{10}$$

The results of the first equation give us the threshold $X(t)/p$ of the fully-utilizing APs. The second equation can be arranged as $\mathbb{E}((2U(t)-V)\lambda_p(t))$, therefore, all the arriving data should be switched through vertical handover if $U(t) > V/2$ and through the APs network if $U(t) < V/2$.

The advantage of *Lyapunov optimization* is that it provides the delay upper bound of the system with $O(V)$ and the utilization of resource will reach the optimal utilization within $O(1/V)$ [93]. On the other hand, *Lyapunov optimization* also introduces an interesting concept called *virtual queue*. The concept *virtual queue* can convert the constraint problem, like in (8), into the stability problems. It makes the *Lyapunov optimization* be able to solve more general dynamic optimization problems.

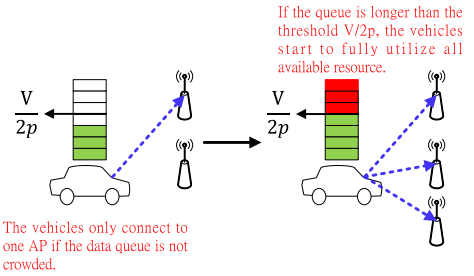


FIGURE 8. Illustration of the autonomous horizontal handover scheme for the vehicular networks. The vehicles access all the available F-APs only if the queuing delay is larger than the threshold $V/2p$, which comes from the results of (9).

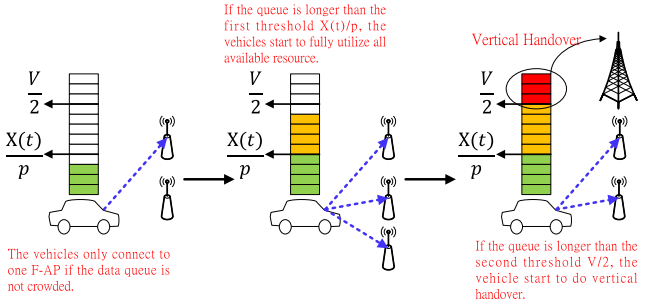


FIGURE 9. Illustration of the autonomous vertical handover scheme for the vehicular networks. From the figure, we can find that we can set different threshold via V to adjust the utilization of the vertical handover.

The Fig. 8 and Fig. 9 illustrates the resulting horizontal and vertical handover schemes based on the solutions of (9) and (10). Fig. 8 illustrates that the vehicles need to access only one F-AP if the data queue is smaller than the threshold $V/2p$. If the data queue exceeds the threshold $V/2p$, which means that the queuing delay has been intolerable, and the vehicles start to access all the available F-APs to decrease the queuing delay. While considering the vertical

handover into the design, there needs two thresholds in the final handover scheme, as shown in Fig. 9. The first one threshold is to determine whether to utilize all the F-APs or not. The second one is determined by the virtual queue $X(t)$, which can be interpreted as the indicator whether the constraint in the original optimization is satisfied or not. After queue size is larger than the second threshold, the vehicles just execute the vertical handover.

We consider the environment that the distribution density of F-AP $\lambda_{ap} = 5 \times 10^{-5}/m^2$ corresponding to different vehicle density λ_v. The velocity of the vehicles is $15m/s$ which is about $55km/hr$. The mean number of arrival packet in each time slot t is $\mathbb{E}(\lambda_p) = 200$ follows a Poisson distribution. The length of each time slot is $0.01s$. The total iteration are 1000 times, and the length of simulation time T for each iteration is 25000 time slots.

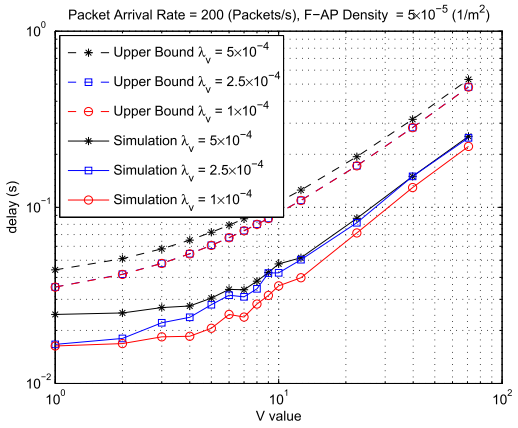


FIGURE 10. The upper bound of delay of the proposed handover scheme. If the density of vehicles λ_V is small enough the to satisfy the stability condition, the delay can be successfully upper bounded without the help of vertical handover.

Fig. 10 illustrates the mean delay corresponding to different values of V without the help of the vertical handover. The upper bounds in Fig. 10 is derived from the *Lyapunov optimization* and can be converted to mean delay via Little's Theorem. In practical operation, we can adjust the value of V to achieve the different delay requirements. Fig. 11 illustrates the probability of vertical handover corresponding to different value V. From the figure, we can find that the probability approaches to 0 while the value of V increases if the density of the vehicles is small. If the density of the vehicles increases, however, the probability of vertical handover cannot decreases to 0 with the increment of V. This is the reason that the density of F-APs is large enough to support all the data flows from the vehicles. However, when the density of the vehicles is large, there is no enough F-APs resources to support the data flows from the vehicles. In such case, the vertical handover scheme becomes necessary.

VIII. RESOURCE ACCESS CONTROL IN H-CRAN AND FogNet

With limited spectrum resource, it is inevitable that FogNet (fog-part) may be underlay or overlay under the

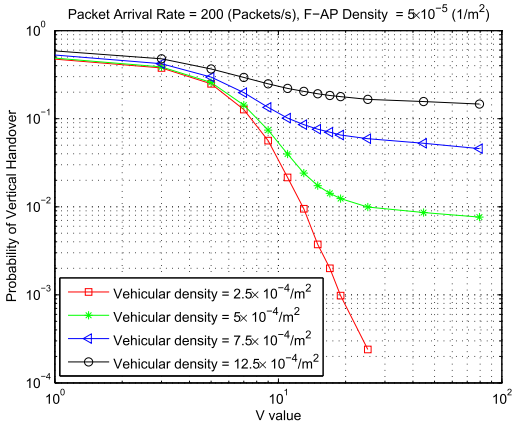


FIGURE 11. The probability of vertical handover of the proposed scheme. The probability of vertical handover does not decrease as V increasing if the value of λ_V is large. It is the reason there is no enough access resource F-APs and the vertical handover becomes the necessary alternative ways to service data flows.

H-CRAN (cloud-part). In such case, the successful coexistence of fog-part and cloud-part may rely on the resource access control. It raises a fundamental question "is the controlling signals necessary to organize the resource access between fog and cloud-part?". In this section, we show that, in some situation, the FogNet relying on random access control may outperform the fully centralized control like H-CRAN.

A. RESOURCE UTILIZATION SCHEME IN H-CRAN AND FogNet of F-CRAN

As aforementioned, H-CRAN is a fully centralized radio access architecture. To reach the radio resource optimization, H-CRAN can schedule all the radio resources to the users' needs. On the contrast, FogNet has a distributed architecture in which all the users may share a common pool of allocated radio resources. For a long run, we consider the systems with the centralized resource management or the distributed scheme such as random access in the form of carrier sense multiple access (CSMA) separately.

While we consider scheduling in H-CRAN, it needs to allocate the amount of the necessary radio resources to the users. In this way, all the users can individually utilize the radio resources such that there is no interference among users or statistically under an acceptable interference level. Nevertheless, H-CRAN must allocate more radio resources than the required amount to against deep fading in the channels. This scheme is further necessary while the probability of deep fading channels is large. Even though the performance of individual user can be improved through allocating more amount of radio resource than a user's request, but the performance of the entire network may degrade. Please recall that there may be no centralized coordination in the FogNet and all the users compete for the radio resources through random access like CSMA. Though random access may result in interference among users, but the fading effects in different

radio resources are varying for different users at different locations. Such a situation creates an effect similar diversity communication. From the viewpoint of frequency reuse in the network, FogNet may enjoy better spectrum efficiency than the H-CRAN. Therefore, the existence of FogNet in the H-CRAN invokes a fundamental question: when and how to efficiently switch between H-CRAN and FogNet for a mobile UE to get the best performance.

B. H-CRAN MODEL AND PERFORMANCE OF SCHEDULE

We consider the scenario that the UEs adopt a single frequency band for uplink transmission, and radio resources are allocated in the basic unit of a resource block (RB). As each uplink transmission may involve a number of RBs, these RBs used for one batch of resources in uplink transmission is referred as one batch of resources. A UE requests for one resource batch for uplink transmission, then a HPN may allocate exactly one resource batch for this UE through RRHs. Due to the mobility and distribution of UEs and RRHs channel conditions among H-CRAN/FogNet are stochastic. While the deep fading occurs, UEs cannot successfully transmit data with this radio resource. The net throughput contributed from this UE depends on the probability of deep fading occurrence.

Considering that there are totally M resource blocks in the frequency domain. We denote the probability of deep fading as p. To guarantee the successful transmission, the most naive approach is to allocate the m resource blocks to a UE. In this way, the throughput of a UE is $1-p^m$. Obviously, even though such approach increase the individual performance, but the total network throughput is severely degraded. To discuss the throughput of the whole system, we can define the entire network throughput υ as

$$\upsilon = \frac{\mathbb{P}(\text{Successfully Transmission})}{m} = \frac{1-p^m}{m}. \quad (11)$$

It can be interpreted as the probability of successful transmission per resource block.

C. FogNet MODEL AND PERFORMANCE OF RANDOM ACCESS

FogNet is a group of UEs and may be without the coordination of the H-CRAN in the F-CRAN system. To further enhance the throughput of such system, it prefers that more users can simultaneously utilize the same resource blocks. Due to the random location of UEs, different UEs may suffer from different level of deep fading at the same resource batch. Therefore, it gives the room for multiple UEs accessing the same resource blocks.

Among the total M resource blocks, each UE performs channel estimation at all the M resource blocks and selects one without deep fading condition. Then these resource blocks can be fully utilized to enhance the throughput of the entire network. However, if some of UEs unfortunately select the same unoccupied resource batch to transmit data, then a collision occurs and the throughput of UEs degrades, which is the issue in the FogNet adopting random access.

Grouping is an effective approach to alleviate above dilemma [94] and later adopted in LTE. While grouping UEs to form a FogNet, it is desirable to identify a proper size for a group. A large group can introduce severe competition for resource blocks, but a small group may result in low utilization efficiency of the network. It is necessary to find the number of UEs that can achieve the best system performance given the limited resource batches and deep fading probability. Considering that there are totally M resource batches indexed by $m = 1 \ldots M$ to be shared by N UEs. Before transmitting, the UEs will sense all M resource batches without deep fading and choose one to transmit data. Let

$$\mathbf{I}_{m,i} = \begin{cases} 1, & i\text{th UE selects the } m\text{th resource batch,} \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

be an indication function. Then the probability that ith UE utilizes mth resource block can be expressed as

$$q \triangleq \mathbb{P}(\mathbf{I}_{m,i} = 1) = \frac{1-p^M}{M}. \quad (13)$$

Then the throughput of the mth resource block is

$$\upsilon = Nq(1-q)^{N-1}. \quad (14)$$

To find the most suitable number of UEs N^* in the FogNet, we need to maximize the equation above.

$$N^* = \arg \max_N Nq(1-q)^{N-1} \quad (15)$$

We can substitute N^* into (14) and get the best performance of the FogNet.

D. SWITCH POINT BETWEEN FogNet AND H-CRAN

The centralized scheduling based approach is regarded as an effective scheme to enhance the throughput of the network. However, the performance of the throughput of the entire network may degrade if the deep fading or other interference (from other underlay network) is severe. If we allow all the UEs compete the resource randomly, a particular resource block that may be under deep fading for one UE but has a good channel condition for another UE in a different location. If this happens, the utilization efficiency of the resource blocks can be increased without interfering other UEs. This performance enhancement is boosted by *statistical multiplexing* of multiple UEs' channel access. According to this argument, it is necessary to find the switch point, which depends on the value of p, between the H-CRAN and FogNet.

In (11), we can find that the entire system reaches the largest throughput while $m = 1$, that is, each UE can be allocated with one resource block. Therefore, the performance of the H-CRAN can be expressed as $1-p$. Then we can compare the performance with the appropriate grouping number N^* with the best performance of the FogNet.

The performance of network switching point between centralized control and random access is illustrated in Fig. 12.

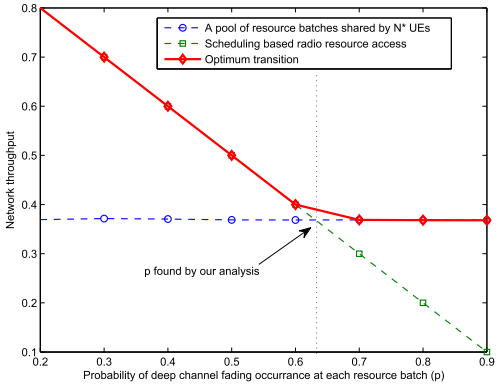


FIGURE 12. The throughput of the FogNet with scheduling based radio access, a pool of resource shared scheme and the optimal transition scheme.

We consider the environment with 20MHz bandwidth with $m = 100$ RBs in the frequency domain. The length of each RB is regarded as 4ms. The transmission power of each UE is 20dBm. The scheduling based radio access provides a better system throughput υ at $p = 1 - (1 - \frac{1}{m})^{m-1}$ is about 0.63, which confirms our arguments.

IX. CONCLUSION

In recent years, H-CRAN and FogNet were proposed to tackle the Internet contents, vehicular network and large amount of devices. The former one focuses on the centralized control to optimize the whole network and reach the best resource utilization. The latter one takes advantage of the characteristics of Internet contents to simplify the architecture of the network in a decentralized way. For the 5G wireless system, the big question may not to select from these two network architectures, but the way for devices to properly select between H-CRAN and FogNet for radio access, which has been overlooked in the literatures. Appropriate coordination between these two network architectures to fully utilize the advantages from each architecture starts with this paper, but definitely not ends here. There are quite a few works discussing about H-CRAN and FogNet separately but little attention has been paid about the coordination of them. Further coordination between these two network architectures to complement each other and to fully utilize network and radio resources remains a subject worth further pursuing. In this research, we illustrate the conditions and mechanism to switch between H-CRAN and FogNet from the viewpoint of cache in wireless network, mobility management and access control, and thus pave a new avenue to various research directions. Future research opportunities may include control and signaling between two network architectures, integration with short-range communication like millimeter-wave (mmWave), licensed assisted access (LAA) and energy harvesting (EH) devices into the H-CRAN. Due to the limited transmission power and computation ability, these devices may hard to access the H-CRAN and the only way is toward the FogNet. On the other hand, to tackle the dynamic environment, sensing as cognitive radios to coordinate the utilization of radio resources [95] may be a good candidate to further improve the FogNet without interfering H-CRAN system. Last but not the least, this harmonization suggests a good balance to standardize state-of-the-art mobile communications and more detailed design and precise analysis remains very much wanted.

REFERENCES

[1] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[2] Q. Li, R. Q. Hu, Y. Qian, and G. Wu, "Intracell cooperation and resource allocation in a heterogeneous network with relays," *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1770–1784, May 2013.

[3] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.

[4] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous LTE/LTE-A networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2142–2180, Fourthquarter 2014.

[5] M. Wildemeersch, T. Q. S. Quek, M. Kountouris, A. Rabbachin, and C. H. Slump, "Successive interference cancellation in heterogeneous networks," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4440–4453, Dec. 2014.

[6] S.-Y. Lien, S.-C. Hung, K.-C. Chen, and Y.-C. Liang, "Ultra-low-latency ubiquitous connections in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 22–31, Jun. 2015.

[7] J. M. Chapin and V. W. S. Chan, "Architecture concepts for a future heterogeneous, survivable tactical Internet," in *Proc. IEEE Military Commun. Conf.*, Nov. 2013, pp. 1874–1879.

[8] K.-C. Chen and S.-Y. Lien, "Machine-to-machine communications: Technologies and challenges," *Ad Hoc Netw.*, vol. 18, pp. 3–23, Jul. 2014.

[9] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[10] A. Rajandekar and B. Sikdar, "A survey of MAC layer issues and protocols for machine-to-machine communications," *IEEE Internet Things J.*, vol. 2, no. 2, pp. 175–186, Apr. 2015.

[11] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Veh. Technol. Mag.*, vol. 5, no. 1, pp. 77–84, Mar. 2010.

[12] C. Chakrabarti and S. Roy, "Adapting mobility of observers for quick reputation assignment in a sparse post-disaster communication network," in *Proc. Appl. Innov. Mobile Comput. (AIMoC)*, Feb. 2015, pp. 29–35.

[13] R. Pozza, M. Nati, S. Georgoulas, K. Moessner, and A. Gluhak, "Neighbor discovery for opportunistic networking in internet of things scenarios: A survey," *IEEE Access*, vol. 3, pp. 1101–1131, Jul. 2015.

[14] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 414–454, Firstquarter 2014.

[15] M. R. Palattella *et al.*, "Standardized protocol stack for the Internet of (important) Things," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1389–1406, Thirdquareter 2013.

[16] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 60–65, Apr. 2011.

[17] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[18] Erisson, "Erisson mobility report," Erisson, Stockholm, Sweden, Tech. Rep., Jun. 2015.

[19] K.-C. Chen, M. Chiang, and H. V. Poor, "From technological networks to social networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 548–572, Sep. 2013.

[20] V. Chan *et al.*, "Future heterogeneous networks," Nat. Sci. Found., Arlington, VA, USA, Tech. Rep., 2011.

[21] M. Dohler, R. W. Heath, A. Lozano, C. B. Papadias, and R. B. Valenzuela, "Is the PHY layer dead?" *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[22] N. Bhushan *et al.*, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[23] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM J. Res. Develop.*, vol. 54, no. 1, pp. 4:1–4:12, Jan./Feb. 2010.

[24] C. Mobile, *C-RAN: The Road Towards Green RAN*, China Mobile Res. Inst., Beijing, China, 2011.

[25] A. Samukic, "UMTS universal mobile telecommunications system: Development of standards for the third generation," *IEEE Trans. Veh. Commun.*, vol. 47, no. 4, pp. 1099–1104, Nov. 1998.

[26] Q. Cui *et al.*, "Evolution of limited-feedback CoMP systems from 4G to 5G: CoMP features and limited-feedback approaches," *IEEE Veh. Technol. Mag.*, vol. 9, no. 3, pp. 94–103, Sep. 2014.

[27] S. Sun, Q. Gao, Y. Peng, Y. Wang, and L. Song, "Interference management through CoMP in 3GPP LTE-advanced networks," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 59–66, Feb. 2013.

[28] O. Onireti, F. Heliot, and M. A. Imran, "On the energy efficiency-spectral efficiency trade-off in the uplink of CoMP system," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 556–561, Feb. 2012.

[29] P. Marsch and G. P. Fettweis, Eds., *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[30] J. Lee *et al.*, "Coordinated multipoint transmission and reception in LTE-advanced systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 44–50, Nov. 2012.

[31] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 365–368, Aug. 2014.

[32] F. A. Khan, H. He, J. Xue, and T. Ratnarajah, "Performance analysis of cloud radio access networks with distributed multiple antenna remote radio heads," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4784–4799, Sep. 2015.

[33] S.-N. Hong and J. Kim, "Joint coding and stochastic data transmission for uplink cloud radio access networks," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1619–1622, Sep. 2014.

[34] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[35] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2014.

[36] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 34–44, Jun. 2013.

[37] V. Suryaprakash, P. Rost, and G. Fettweis, "Are heterogeneous cloud-based radio access networks cost effective?" *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2239–2251, Oct. 2015.

[38] D. C. Chen, T. Q. S. Quek, and M. Kountouris, "Wireless backhaul in small cell networks: Modelling and analysis," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, May 2014, pp. 1–6.

[39] S.-C. Lin and K.-C. Chen, "Improving spectrum efficiency via in-network computations in cognitive radio sensor networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1222–1234, Mar. 2014.

[40] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st Ed. MCC Workshop Mobile Cloud Comput. (MCC)*, 2012, pp. 13–16.

[41] 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.

[42] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 998–1006.

[43] G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu, and N. Brahmi, "Device-to-device communications for national security and public safety," *IEEE Access*, vol. 2, pp. 1510–1520, Dec. 2014.

[44] Y. Li, D. Jin, J. Yuan, and Z. Han, "Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3965–3977, Jul. 2014.

[45] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 96–104, Jun. 2012.

[46] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 441–454, Feb. 2015.

[47] F.-M. T. Shao-Yu Lien, C.-C. Chien, and T.-C. Ho, "3GPP device-to-device communications for beyond 4G cellular networks," *IEEE Commun. Mag.*, 2015

[48] J. Oueis, E. C. Strinati, and S. Barbarossa, "The fog balancing: Load distribution for small cell cloud computing," in *Proc. Veh. Technol. Conf.*, May 2015, pp. 1–6.

[49] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[50] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless. Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[51] J. Zhao, T. Q. S. Quek, and Z. Lei, "Heterogeneous cellular networks using wireless backhaul: Fast admission control and large system analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2128–2143, Oct. 2015.

[52] M. Peng, S. Yan, K. Zhang, and C. Wang. (2015). "Fog computing based radio access networks: Issues and challenges." [Online]. Available: http://arxiv.org/abs/1506.04233

[53] E. Bastug, M. Bennis, and M. Debbah, "Social and spatial proactive caching for mobile data offloading," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 581–586.

[54] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[55] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*. Amsterdam, The Netherlands: Elsevier, 2011.

[56] A. Balamash and M. Krunz, "An overview of Web caching replacement algorithms," *IEEE Commun. Surveys Tuts.*, vol. 6, no. 2, pp. 44–56, Secondquarter 2004.

[57] J. Wang, "A survey of Web caching schemes for the Internet," *SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 5, pp. 36–46, Oct. 1999.

[58] M. Rabinovich and O. Spatscheck, *Web Caching and Replication*. Boston, MA, USA: Addison-Wesley, 2002.

[59] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, vol. 1. Mar. 1999, pp. 126–134.

[60] *Global Mobile Data Traffic Forecast Update 2014–2019 White Paper*, Cisco Syst., Inc., San Jose, CA, USA, 2015.

[61] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[62] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[63] Z. Ming, M. Xu, and D. Wang, "InCan: In-network cache assisted eNodeB caching mechanism in 4G LTE networks," *Comput. Netw.*, vol. 75, pp. 367–380, Dec. 2014.

[64] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[65] H. Hsu and K.-C. Chen, "A resource allocation perspective on caching to achieve low latency," *IEEE Commun. Lett.*, to be published.

[66] H. Chen and Y. Xiao, "Cache access and replacement for future wireless Internet," *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 113–123, May 2006.

[67] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: A new architecture for wireless communications," in *Proc. 19th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 3. Mar. 2000, pp. 1273–1282.

[68] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: ICAR," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2105–2115, Oct. 2001.

[69] M. E. J. Newman. (2002). "Random graphs as models of networks." [Online]. Available: http://arxiv.org/abs/cond-mat/0202208

[70] P. Sobkowicz, M. Thelwall, K. Buckley, G. Paltoglou, and A. Sobkowicz, "Lognormal distributions of user post lengths in Internet discussions—A consequence of the Weber–Fechner law?" *EPJ Data Sci.*, vol. 2, no. 1, p. 2, 2013.

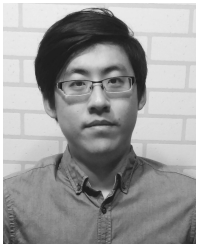
[71] M. G. Khoshkholgh, Y. Zhang, K.-C. Chen, K. G. Shin, and S. Gjessing, "Connectivity of cognitive device-to-device communications underlying cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 81–99, Jan. 2015.

[72] D. G. Luenberger and Y. Ye, *Introduction to Linear and Nonlinear Programming*, 3rd ed. Reading, MA, USA: Addison-Wesley, 1973.

[73] 5GPPP. (Oct. 2015). *5G Automotive Vision*. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf

[74] "Connected car market—Global industry analysis, size, share, growth, trends and forecast, 2013–2019," Transparency Market Res., New York, NY, USA, Tech. Rep., 2013.
[75] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014.
[76] *Standard Specification for Telecommunications and Information Exchange Between Roadside and Vehicle Systems—5 GHz Band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, ASTM Standard E2213-02, 2003.
[77] *Amendment 6: Wireless Access in Vehicular Environment*, IEEE Standard 802.11p, Jul. 2010.
[78] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: A survey," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, May 2013.
[79] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, pp. 1–1, Fourthquarter 2015.
[80] H. Zhang, C. Jiang, and J. Cheng, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 92–99, Jun. 2015.
[81] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.
[82] T. Ali and M. Saquib, "Performance evaluation of WLAN/cellular media access for mobile voice users under random mobility models," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3241–3255, Oct. 2011.
[83] N. W. Sung, N.-T. Pham, T. Huynh, and W.-J. Hwang, "Predictive association control for frequent handover avoidance in femtocell networks," *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 924–927, May 2013.
[84] J.-M. Moon and D.-H. Cho, "Efficient handoff algorithm for inbound mobility in hierarchical macro/femto cell networks," *IEEE Commun. Lett.*, vol. 13, no. 10, pp. 755–757, Oct. 2009.
[85] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 64–91, Firstquarter 2014.
[86] H. Kalbkhani, S. Yousefi, and M. G. Shayesteh, "Adaptive handover algorithm in heterogeneous femtocellular networks based on received signal strength and signal-to-interference-plus-noise ratio prediction," *IET Commun.*, vol. 8, no. 17, pp. 3061–3071, 2014.
[87] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.
[88] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 1243–1254, Mar. 2008.
[89] L. Qin and D. Zhao, "Channel time allocations and handoff management for fair throughput in wireless mesh networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 315–326, Jan. 2015.
[90] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.
[91] K. T. Phan, T. Le-Ngoc, M. van der Schaar, and F. Fu, "Optimal scheduling over time-varying channels with traffic admission control: Structural results and online learning algorithms," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4434–4444, Sep. 2013.
[92] Y. Cui, V. K. N. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems—Large deviation theory, stochastic Lyapunov drift, and distributed stochastic learning," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1677–1701, Mar. 2012.
[93] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Found. Trends Netw.*, vol. 1, no. 1, pp. 1–144, 2006.
[94] K.-C. Chen, "Medium access control of wireless LANs for mobile computing," *IEEE Netw.*, vol. 8, no. 5, pp. 50–63, Sep. 1994.
[95] S.-Y. Lien, K.-C. Chen, Y.-C. Liang, and Y. Lin, "Cognitive radio resource management for future cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 70–79, Feb. 2014.

SHAO-CHOU HUNG received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree with the Graduate Institute of Communication Engineering. His research interests include 5G network architecture, cognitive radio networks, and dynamic optimal control in wireless network.



HSIANG HSU received the B.S. degrees in electrical engineering and mathematics from National Taiwan University, in 2014, where he is currently pursuing the M.S. degree with the Graduate Institute of Communication Engineering. His research interests include resource allocation, optimization, and machine learning in complex networks.

SHAO-YU LIEN is currently an Assistant Professor with the Department of Electronic Engineering, National Formosa University, Taiwan. Recently, his focuses are particularly on cyber-physical systems and 5G communication networks. His research interests include optimization techniques for networks and communication systems. He received a number of prestigious recognitions, including the IEEE Communications Society Asia-Pacific Outstanding Paper Award in 2014, the Scopus Young Researcher Award (issued by Elsevier) in 2014, the URSI AP-RASC 2013 Young Scientist Award, and the IEEE ICC 2010 Best Paper Award.

KWANG-CHENG CHEN (M'89–SM'94–F'07) received the B.S. degree from the National Taiwan University in 1983, and the M.S. and Ph.D. degrees from the University of Maryland, College Park, USA, in 1987 and 1989, all in electrical engineering. From 1987 to 1998, he worked with SSE, COMSAT, IBM Thomas J. Watson Research Center, and National Tsing Hua University, working on the mobile communications and networks. Since 1998, he has been with the National Taiwan University, Taipei, Taiwan. After serving as the Director, Graduate Institute of Communication Engineering, Communication Research Center, and the Associate Dean for Academic Affairs, he is currently a Distinguished Professor with National Taiwan University and is visiting the Massachusetts Institute of Technology from 2015 to 2016. He has been actively involving in the organization of various IEEE conferences as General/TPC Chair/Co-Chair, and has served in editorships with a few IEEE journals. He also actively participates in and has contributed essential technology to various IEEE 802, Bluetooth, and LTE and LTE-A wireless standards. He is an IEEE Fellow and has received a number of awards, such as the 2011 IEEE COMSOC WTC Recognition Award, 2014 IEEE Jack Neubauer Memorial Award, and 2014 IEEE COMSOC AP Outstanding Paper Award. His recent research interests include wireless communications, network science, and data science.