INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2017, VOL. 63, NO. 1, PP. 51-64
Manuscript received October 30, 2016; revised December, 2016. DOI: 10.1515/eletel-2017-0008

Cloud Cooperated Heterogeneous Cellular Networks for Delayed Offloading using Millimeter Wave Gates

Ehab Mahmoud Mohamed

***Abstract*—Increasing the capacity of wireless cellular network is one of the major challenges for the coming years. A lot of research works have been done to exploit the ultra-wide band of millimeter wave (mmWave) and integrate it into future cellular networks. In this paper, to efficiently utilize the mmWave band while reducing the total deployment cost, we propose to deploy the mmWave access in the form of ultra-high capacity mmWave gates distributed in the coverage area of the macro basestation (Macro BS). Delayed offloading is also proposed to proficiently exploit the gates and relax the demand of deploying a large number of them. Furthermore, a mobility-aware weighted proportional fair (WPF) user scheduling is proposed to maximize the intra-gate offloading efficiency while maintaining the long-term offloading fairness among the users inside the gate. To efficiently link the mmWave gates with the Macro BS in a unified cellular network structure, a cloud cooperated heterogeneous cellular network (CC-HetNet) is proposed. In which, the gates and the Macro BS are linked to the centralized radio access network (C-RAN) via high-speed backhaul links. Using the concept of control/user (C/U) plane splitting, signaling information is sent to the UEs through the wide coverage Macro BS, and most of users' delayed traffic is offloaded through the ultra-high capacity mmWave gates. An enhanced access network discovery and selection function (eANDSF) based on a network wide proportional fair criterion is proposed to discover and select an optimal mmWave gate to associate a user with delayed traffic. It is interesting to find out that a mmWave gate consisting of only 4 mmWave access points (APs) can offload up to 70 GB of delayed traffic within 25 sec, which reduces the energy consumption of a user equipment (UE) by 99.6 % compared to the case of only using Macro BS without gate offloading. Also, more than a double increase in total gates offloaded bytes is obtained using the proposed eANDSF over using the conventional ANDSF proposed by 3GPP due to the optimality in selecting the associating gate.**

***Keywords*—millimeter wave, delayed offloading, CC-HetNet, ANDSF, C/U splitting**

I. Introduction

DUE to the huge proliferation of smart phone and tablet users, the cellular network capacity comes at the bottleneck. Many studies claim that by the end of the year 2019, the global mobile data traffic will exceed 20 Exabytes (EB) per month, with an expected exponential increase of 2-fold yearly [1] [2]. Techniques such as massive MIMO [3], heterogeneous networks (HetNets) using small cells [4], and bandwidth expansion using millimeter wave (mmWave) technology [5] are all introduced to address this issue.

Considering bandwidth expansion, mmWave is the most attractive because up to 7 GHz of continuous spectrum is available worldwide in the 60 GHz unlicensed frequency band [5]. Therefore, multi-Gbps rate can be achieved using mmWave communications. However, mmWave transmissions suffer from high propagation loss, oxygen absorption, and path blocking [6] [7], which limit mmWave coverage to be few meters around a mmWave small cell (access point (AP)) [6] [7]. Currently, many research works have been done to integrate the mmWave access into future 5G cellular networks [5] - [13]. Nevertheless, due to its short-range transmission, researchers argued that thousands of mmWave APs should be deployed to provide Gbps for all users inside typical Macro BS area [5] [9] [10]. However, deploying thousands of mmWave APs not only increases the deployment cost of future 5G cellular networks, but also complicates the control required for radio resource management (RRM) inside the network.

In this paper, to efficiently integrate the mmWave access into future cellular networks, we propose to deploy it in the form of ultra-high capacity mmWave gates distributed inside the coverage area of legacy cellular network such as LTE and located at the entrance of buildings, train stations, airports, enterprise, shopping mall, etc. The proposed mmWave gate consists of a number of mmWave APs operating at 60 GHz with 2.16 GHz bandwidth as standardized by IEEE 802.11ad [14] and controlled by an AP controller (APC) installed inside the gate [8]. Compared to the case of a single mmWave AP, the mmWave gate is capable of offloading a massive amount of users' traffic in a very short time using mmWave concurrent transmissions empowered by mmWave spatial diversity inherent in its directional transmissions [8]. To efficiently utilize the proposed mmWave gates and relax the demand of deploying a high number of them, we also propose to use the concept of delayed offloading in conjunction with the proposed gates. Currently, most of smart phones support the on-the-spot offloading, where user equipment (UE) gives more priority to Wi-Fi interface than cellular interface [15] whenever it gets into a Wi-Fi coverage. Recently, many researchers highlighted the concept of delayed offloading based on the fact that the massive part of mobile data is indeed non-real time, which can be delayed for some predefined time without affecting user satisfaction such as video/game

This work is partially supported by National Telecom Regulatory Authority (NTRA) Egypt under project title "LTE/WiFi/WiGig internetworking for future 5G cellular networks".

Ehab Mahmoud Mohamed is with Electrical Engineering Dept., Aswan University, Aswan, Egypt. (e-mail: ehab_mahmoud@aswu.edu.eg).

downloading, software update, mobile backup, etc. [15] - [19]. In this scheme, each data transfer is associated with a fixed deadline, and the data transfer is resumed whenever the user is being under the coverage of a Wi-Fi hotspot until the transfer is completed. If the transfer is not completed within its deadline, Macro BS (e.g. LTE) finally completes the transfer [15] - [19]. Therefore, there is an inherent win-win relationship between mmWave gates and delayed offloading concept. On one hand, the gates act as ultra-high capacity offloading zones capable of offloading a massive amount of users' delayed traffic in a very short time saving users' time and energy. On the other hand, the concept of delayed offloading effectively overcomes the short-range transmissions of mmWave and reduces the necessity of deploying a high number of gates to gain a high offloading efficiency.

To efficiently implement, fully monitor and control the delayed offloading process over the gates, we propose a cloud cooperated heterogeneous network (CC-HetNet). In which, the gates and the Macro BS are tightly coupled in a centralized manner to the centralized radio access network (C-RAN) via high speed backhaul links, e.g. optical fiber links. The concept of control / user (C/U) plane splitting [20] [21] is used to control the delayed offloading process inside the CC-HetNet. In which, user's context information such as location and delayed files information (titles, sizes and deadlines) are conveyed by the control channel of the Macro BS to the C-RAN. Navigating a user with delayed traffic to an optimal nearby gate is also controlled via Macro BS signaling. A combination of Macro BS and gates opportunities is used to deliver the users' delayed traffic using the U plane of the mmWave gates or the Macro BS. Beside proposing the CC-HetNet, we also give the protocol that organizes the delayed offloading process inside it. Two important challenges are considered in designing the CC-HetNet through this paper. The first challenge is the RRM inside the gate, which is designed to offload users' delayed traffic while they are passing through the gate. Thus, the RRM entity in the APC should maximize the gate offloading efficiency while maintaining the long-term offloading fairness among users with un-equal stay times inside the gate coverage. To cope with this challenge, a mobility-aware weighted proportional fair (WPF) user scheduling algorithm is proposed. The second challenge is the gates discovery with optimal gate selection. To efficiently cope with this challenge, an enhanced access network discovery and selection function (eANDSF) is also proposed.

In this paper, we are only focusing on investigating the effectiveness of the proposed delayed offloading CC-HetNet in boosting the capacity of the current cellular network and relaxing the traffic demand on the Macro BS via optimizing the RRM inside the mmWave gates and throughout the CC-HetNet. Other issues related to the detailed CC-HetNet internetworking architecture based on the current 3GPP [22] and IEEE 802.11ad [14] standards including the required interfaces, the required integrating protocol stack, and the detailed architecture of the C-RAN and the gate APC are out of the scope of this paper, and it will be the motivation of our future work. Some research works considering LTE/ mmWave internetworking architectures can be found in [9] [23].

The rest of this paper is organized as follows; Section II gives the literature review related to the paper work. Section III introduces the proposed delayed offloading CC-HetNet using mmWave gates. The intra-gate RRM is given in Section IV. Section V presents the proposed adaptive process for gate association. Section VI gives the simulation analysis followed by the conclusion in Section VII.

II. Literature Review

Recently, mmWave band attracts researchers from academia and industry as a key enabler of future 5G cellular networks. IEEE 802.11ad standard is ratified for 60 GHz communication [14]. In [10], capacity and coverage of mmWave cellular networks are theoretically investigated. Outdoor and indoor mmWave propagation measurements are given in [6] [7] [11]. The authors in [5] [9] gave the detailed architecture and required technologies to overlay the mmWave small cells (APs) to the legacy cellular networks in a unified network using the concept of multi-band HetNets. At the best of our knowledge, no study investigating the efficient utilization of mmWave access using the concept of delayed offloading is given.

Also, there have been various studies focusing on the concept of delayed offloading for pressing the need of additional cellular network capacity. In [15], using daily Wi-Fi usage traces of users who generate 7 GB per month as mobile traffic, the authors proved that a 29 % increase in the offloading gain can be achieved over the on-the-spot offloading if the traffic transmission can be delayed by 1 hour. Also, in [16], the authors reported that, if the response to a user traffic request can be delayed by only 100 sec, about 20 % ~ 30 % additional offloading gains over the on-the-spot offloading can be accomplished using high speed users traces. Several studies are also conducted to show the potential of delayed offloading using Wi-Fi APs in reducing the demand of cellular network traffic [24] - [28]. Despite the potential of delayed offloading studies using Wi-Fi APs, they lacked a complete network framework of how we can efficiently implant the delayed offloading process to be a part of future cellular networks [25] - [27]. Although the authors in [27] explored this important issue, they proposed to use the conventional ANDSF [29] to manage the delayed offloading process inside the network. The conventional ANDSF is proposed by 3GPP as a loosely coupled internetworking

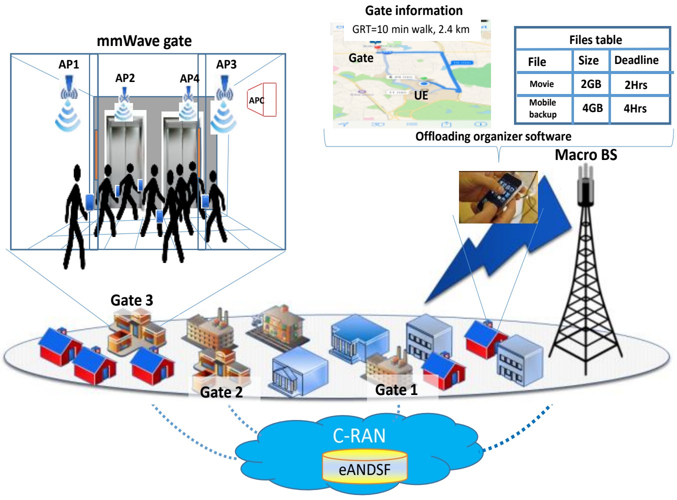

Fig. 1. Proposed delayed offloading CC-HetNet structure using mmWave gates.

mechanism to mainly assist UE for discovering nearby non-3GPP access networks (WLANs) using UE scanning [29]. Then, the UE selects one of the available offloading zones to connect with it, which is usually the nearest available zone. This user based selected zone might not have a remaining offloading capacity suitable for its delayed traffic. Also, frequent UE scanning highly consumes the UE energy especially if we consider the small coverage of mmWave access. Thus, conventional ANDSF cannot deal with delayed offloading especially assuming the small coverage mmWave APs.

III. Proposed Delayed Offloading CC-HetNet Structure and Protocol

Fig. 1 shows the detailed structure of the proposed CC-HetNet including the mmWave gates. A mmWave gate consists of a number of IEEE 802.11ad based mmWave APs operating at 60 GHz with 2.16 GHz bandwidth [14], which are controlled by an APC installed inside the gate. The APC is responsible for the RRM inside the gate, such as APs associations, re-associations, mmWave beamforming and joint user scheduling, etc. Hence, it is responsible for maximizing the gate offloading efficiency while maintaining the long-term offloading fairness among the users inside the gate. Besides, it works as a Gateway to connect the gate with the cellular network. In Fig. 1, gate 3 consists of 4 mmWave APs controlled by an APC installed inside the gate.

To fully monitor and control the delayed offloading process, the gates and the Macro BS are tightly coupled to the C-RAN via ultra-high speed backhaul links, e.g., fiber links. Also, the concept of C/U plane splitting [20] [21] is used to facilitate the management of the delayed offloading process. In which, the wide coverage Macro BS is used to do the signaling and a combination of Macro BS and mmWave gates opportunities to deliver the data. Through Macro BS control channel, C-RAN sends (picks up) the signaling information to (from) UEs respectively. Moreover, C-RAN performs UE location estimation, which can be effectively done using the observed time difference of arrival (OTDOA) or by performing location estimation using small cells [30] [31]. The proposed eANDSF is performed by an eANDSF entity installed inside the C-RAN, as shown in Fig.1. Based on the estimated UE location and the registered gates positions, the eANDSF entity discovers the available gates around the UE. From these discovered gates, it selects an optimal gate for user association. The optimal gate selection is performed through a proposed network-wide online algorithm given in the following sections, in which new users are linked up with the gates that maximize the total gates offloaded bytes and users offloading experiences (OFEs) without changing the selected gates of the existing users.

On the UE side, as the UE generates a delay tolerant file such as movie downloading or mobile backup, it will be registered in a delayed files data base, including the file necessary information such as its title, total size and assigned deadline. The user or an application program typically assigns the file deadline. As a user interface, the user can see the information of the delayed files and the information related to the discovered gates including the optimally selected one using an offloading organizer software, as shown in Fig. 1. This software can be easily implemented using typical smart phone platforms, e.g., iOS and Android software. A Google map like interface built into the organizer software is used for navigating the user to the optimal gate, as shown in Fig. 1. Hence, the user can be navigated easily to the designated gate within the estimated gate reaching time (*GRT*). Moreover, the organizer software can trigger the delayed files data base to change the deadlines of some specific files based on user's desires. Also, the user can select a gate other than the optimally selected one, for example the user urgently wants to go to a certain place, e.g., train station. In such cases, the C-RAN takes into account the specific selections done by the user when selecting the optimal gate for the other users. The delayed files are scheduled for transmission in shortest remaining time first (SRTF).

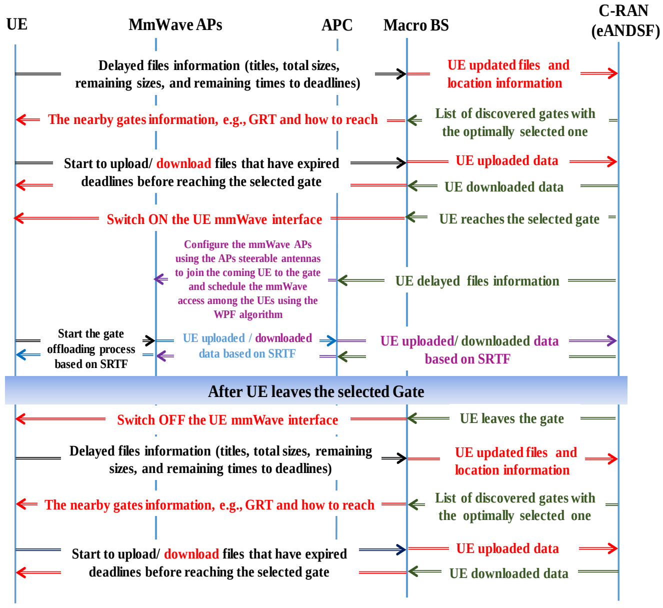
UE
MmWave APs
APC
Macro BS
C-RAN
(eANDSF)
Delayed files information (titles, total sizes, remaining sizes, and remaining times to deadlines)
UE updated files and location information
The nearby gates information, e.g., GRT and how to reach
List of discovered gates with the optimally selected one
Start to upload/ download files that have expired deadlines before reaching the selected gate
UE uploaded data
UE downloaded data
Switch ON the UE mmWave interface
UE reaches the selected gate
Configure the mmWave APs using the APs steerable antennas to join the coming UE to the gate and schedule the mmWave access among the UEs using the WPF algorithm
UE delayed files information
Start the gate offloading process based on SRTF
UE uploaded / downloaded data based on SRTF
UE uploaded/ downloaded data based on SRTF
After UE leaves the selected Gate
Switch OFF the UE mmWave interface
UE leaves the gate
Delayed files information (titles, total sizes, remaining sizes, and remaining times to deadlines)
UE updated files and location information
The nearby gates information, e.g., GRT and how to reach
List of discovered gates with the optimally selected one
Start to upload/ download files that have expired deadlines before reaching the selected gate
UE uploaded data
UE downloaded data

Fig. 2. Proposed protocol for the delayed offloading CC-HetNet.

Fig. 2 summarizes the proposed simple protocol organizing the operation among the basic elements of the CC-HetNet, i.e., the UE, the mmWave APs, the gate APC, the Macro BS, and the C-RAN (eANDSF entity). In this protocol, the C-RAN is always collecting the delayed files information from the UEs registered in their offloading organizer software including their titles, total sizes, remaining sizes and remaining times to deadlines using Macro BS control channel. Based on this information and the estimated UE location, the eANDSF entity in the C-RAN discovers the mmWave gates located nearby the UE. Based on the gates statuses at the expected user arrival, the eANDSF entity selects the optimal gate for associating the UE using the proposed online algorithm explained in the following sections. After discovering the nearby gates and obtaining the optimal one, the C-RAN sends this information to the UE via Macro BS control channel with an information about the estimated *GRTs* and how to reach for the purpose of user navigation. While UE is moving towards the optimal gate or his specially selected gate, delayed files with expired deadlines before the user reaching the gate will be transmitted using the Macro BS data channel. Before the user entering the gate, a switch ON signal is transmitted to the UE from the C-

RAN using Macro BS control channel to switch on its mmWave module if it is in a sleep mode. This is can be done by the C-RAN simply by monitoring the distance between the updated UE location and the gate location. If this distance is less than or equal to a certain threshold value and decreasing, this will be interpreted by the C-RAN as the UE is near to enter the gate, then it switches on its mmWave module. As the UE enters the gate, the offloading process through mmWave data channel starts. To speed up the offloading process, the C-RAN prefetches the UE delayed files and sends them to the gate at the user arrival [28]. Inside the gate a WPF scheduling algorithm is used for user scheduling. It is used to maximize the gate offloading efficiency while maintaining the long-term offloading fairness among the scheduled UEs. After the UE leaves the mmWave gate, the C-RAN sends a switch OFF signal to the UE via the Macro BS control channel to switch off its mmWave module. This is also can be done by the C-RAN, simply by monitoring the distance between the updated UE location and the gate location. If this distance is larger than the threshold value and increasing, this means that the UE leaves the gate and is moving outwards it, then the C-RAN switches off its mmWave module. This switching ON/OFF functionality highly reduces UE energy consumption by switching on/off its mmWave module based on the usage eliminating UE frequent scanning required by conventional ANDSF. After UE leaves the gate, the C-RAN keeps collecting its context information including its updated location and the updated information of its delayed files. Hence, the eANDSF entity can always list its nearby gates and select the optimal gate for its association. Till the UE reaching the second mmWave gate, its delayed files with expired deadlines before reaching the gate will be transmitted using the Macro BS data channel, and so on.

As the proposed delayed offloading CC-HetNet is on the top of currently proposed, standardized and prototyped technologies like mmWave access [14], C-RAN [32], ANDSF [29], C/U plan splitting [20] [21], indoor and outdoor localizations [33] [34], and smartphone and tablets software platforms, it can be realized and being practical. However, some technical challenges still exist and need to be tackled. One of the main technical challenges facing the proposed delayed offloading CC-HetNet is the need of frequently collecting the network updated information including UEs updated locations and their delayed files information and the updated statuses of the mmWave gates. To do that, the UEs need to be always synchronized with the network through their offloading organizer softwares for collecting their updated locations, updated delayed files information and their special gate selections. At the same time, the network should always monitor the updated gates statuses, i.e., capacities and loads via backhaul links for making an optimal gate selection. This kind of technical challenges should be investigated for making delayed offloading CC-HetNet a practical enabler of 5G cellular networks. By only focusing on studying the offloading efficiency of the proposed delayed offloading CC-HetNet in the current paper, these kind of technical challenges will be the motivation of our future work by designing a high efficient network synchronization and monitoring protocol suitable for the proposed delayed offloading mechanism.

IV. Proposed Intra-Gate Radio Resource Management

The intra-gate RRM is one of the most challengeable problems in the gate design because we assume that the offloading process is done while the users are passing through the gate with their different normal speeds. Therefore, maximizing the gate offloading efficiency while maintaining the long-term offloading fairness among users with different mobilities (un-equal stay times) is one of the most challengeable tasks for the APC. In this paper, a mobility-aware WPF joint user scheduling algorithm is proposed to efficiently cope with this challenge.

A. Link Model

The received signal to interference pulse noise ratio (SINR) at time slot n for user k from mmWave AP m using beam identification (ID) b_m is:

$$SINR_{mk}^{b_m}(n) = \frac{\left|g_{mk}^{b_m}(n)\right|^2 P_m}{\sum_{j\neq m}^{M(n)} \left|g_{jk}^{b_j}(n)\right|^2 P_j + \sigma^2}, \tag{1}$$

where beam ID is the identification of beam direction results from using certain antenna pattern. P_m is the transmit power of AP m, where all APs are assumed to have the same value of the transmit power, and σ^2 is the noise power. $g_{mk}^{b_m}(n)$ is the channel impulse response including antenna gains between AP m and user k using beam ID b_m at time slot n, $M(n)$ is the total number of operated APs at time slot n. In this paper, we assume that communications between APs and UEs are done only using the beam IDs corresponding to the highest link quality, which can be determined using the exhaustive search beamforming [12] or using advanced mmWave beamforming techniques as given in [13]. Therefore, $SINR_{mk}^{b_m}(n)$ and $g_{mk}^{b_m}(n)$ can be simplified to $SINR_{mk}(n)$ and $g_{mk}(n)$, respectively. Considering single carrier (SC) transmissions and by following the channel model given by IEEE 802.11ad, $g_{mk}(n)$ is expressed as [35]:

$$g_{mk}(n) = \int_0^{2\pi}\int_0^{\pi} G(\theta,\emptyset) h_{mk}(\theta,\emptyset,n)\sin(\theta) d\theta d\emptyset, \tag{2}$$

where $G(\theta,\emptyset)$ is the antenna gain function (antenna pattern) of AP m, $h_{mk}(\theta,\emptyset,n)$ is the multi-path channel impulse response from AP m to user k without antenna gain, and θ and $\emptyset$ are the elevation and azimuth directions. Therefore, the instantaneous achievable transmission rate in bps for user k from AP m is [36]:

$$r_{mk}(n) = \eta_B BW log_2(1 + \eta_s\, SINR_{mk}(n)), \tag{3}$$

where η_B and η_s are the bandwidth and SNR efficiencies [36], and BW is the used bandwidth in Hz. In this paper, to cope with the different sizes of the users' delayed traffic inside the gate, we define the offloading rate in bps for user k from AP m as:

$$c_{mk}(n) = \min\left(r_{mk}(n), \frac{l_k(n)}{T_{\mathrm{sl}}}\right), \tag{4}$$

where $c_{mk}(n)$ is the instantaneous achievable offloading rate at time slot *n* for user *k* from AP *m*, $l_k(n)$ is the total delayed traffic in bits registered in the files data base of user *k* at time slot *n*, and T_{sl} is the time slot duration in seconds. Through using this definition, during a one-time slot period T_{sl}, user *k* cannot offload more data bits than $r_{mk}(n) \times T_{\text{sl}}$, and it cannot offload more data bits than it is indexed in its offloading organizer, which enables the APC to maintain offloading fairness among users with delayed traffic. The value of $l_k(n)$ can be estimated by the APC by knowing the initial size of user *k* delayed traffic $l_k(0)$ when it enters the gate. This information is sent by the C-RAN to the gate APC at the time of user arrival.

B. Intra-gate AP-UE Assignment Problem

The objective of the intra-gate user assignment problem is to maximize the gate aggregate utility over the long-term users' average offloading rates $\bar{C}_k$ during their total stay inside the gate coverage, which can be formulated as:

$$\max \sum_{k\in K} U_k(\bar{C}_k) \qquad \text{subject to} \quad \bar{C}_k \in \mathcal{C}^+, \tag{5}$$

where $U_k(.)$ is an increasing, strictly concave and continuously differentiable utility function for user *k*, $\mathcal{C}^+$ is the set of all achievable offloading rate vectors, and *K* is the total number of users passed through the gate. Usually, the logarithmic utility function is used for all users, i.e., $U_k(\bar{C}_k) = \log(\bar{C}_k)$, to maximize the users' average rates while maintaining the long-term fairness among them [37] [38].

Thanks to spatial diversity inherent in mmWave directional transmissions, the optimization problem in (5) should be solved jointly by the APC by enabling mmWave concurrent transmissions using the $M(n)$ operated APs at every time slot *n*. Therefore, the optimization problem in (5) can be converted into a per time slot AP-UE assignment problem by defining $\mathbf{I}(n) = \left(I_{mk}(n): m \in M(n), k \in K(n)\right)$ as the AP-UE assignment indicator matrix of size $M(n) \times K(n)$ at time slot *n*, where, $K(n)$ is the total number of users inside the gate at time slot *n* and generally we assume $M(n) < K(n)$. $I_{mk}(n) = 1$, when user *k* is assigned to AP *m* at time slot *n*, and 0 elsewhere. Since only one user can be assigned to AP *m* at time slot *n*, we have $\sum_{k\in K(n)} I_{mk}(n) = 1$ for all $m \in M(n)$. Using $\mathbf{I}(n)$ and the log utility function, the optimization problem in (5) can be formulated as a per time slot AP-UE assignment problem as follows:

$$\textbf{Assignment}: \max_{\mathbf{I}(n)} \left(\sum_{k\in K(n)} \log\left(\bar{C}_k(n)\right) \right) \qquad \text{Subject to} \sum_{k\in K(n)} I_{mk}(n) = 1, \ \forall m \in M(n). \tag{6}$$

This means that the APC maximizes $\sum_{k\in K(n)} \log(\bar{C}_k(n))$ over all possible AP-UE assignment matrices $\mathbf{I}(n)$ at every time slot *n*. $\bar{C}_k(n)$ is the average offloading rate given to user *k* up to time slot *n*, which can be expressed as [37] [38]:

$$\bar{C}_k(n) = \left(1 - \frac{1}{n}\right) \bar{C}_k(n-1) + \frac{1}{n} C_k(n), \tag{7}$$

where $C_k(n) = \sum_{m\in M(n)} I_{mk}(n)\, c_{mk}(n)$ is the instantaneous offloading rate assigned to user *k* at time slot *n*, $C_k(n) \in [0,\, c_{mk}(n)]$.

From Taylor series expansion:

$$\log\left(\bar{C}_k(n)\right) \approx \log\left(\bar{C}_k(n-1)\right) + \frac{1}{\bar{C}_k(n-1)} \left(\bar{C}_k(n) - \bar{C}_k(n-1)\right). \tag{8}$$

From (7), (8) can be re-written as:

$$\log(\bar{C}_k(n)) \approx \log(\bar{C}_k(n-1)) + \frac{1}{n}\left(\frac{C_k(n)}{\bar{C}_k(n-1)} - 1\right). \tag{9}$$

Thus, maximizing $\log(\bar{C}_k(n))$ is equivalent to maximizing $\frac{C_k(n)}{\bar{C}_k(n-1)}$. Accordingly, the solution of the optimization problem in (6) becomes:

$$I_{mk}(n) = \begin{cases} 1, \text{ if } \forall k \in K_{M(n)} = \underset{\forall K_{M(n)} \subset K(n)}{\arg\max} \left\{ \sum_{k\in K_{M(n)}} \left(\frac{C_k(n)}{\bar{C}_k(n-1)} \right) \right\}, \\ 0, \qquad \text{elsewhere.} \end{cases} \tag{10}$$

where $\forall K_{M(n)} \subset K(n)$ indicates all available users subsets of length $M(n)$ of the users domain $K(n)$. The solution in (10) means that the APC picks up the user subset $K_{M(n)}$ of length $M(n)$ with the largest $\sum_{k\in K_{M(n)}} \left(\frac{C_k(n)}{\bar{C}_k(n-1)}\right)$ to be scheduled in time slot *n*, and $I_{mk}(n)$ will equal 1 for $\forall k \in K_{M(n)}$. In the case of $M(n) = 1$, the solution in (10) becomes:

$$I_{mk}(n) = \begin{cases} 1, & \text{if } k = \underset{\forall k\in K(n)}{\arg\max} \left(\frac{C_k(n)}{\bar{C}_k(n-1)} \right), \\ 0, & \text{elsewhere.} \end{cases} \tag{11}$$

C. Proposed Mobility-aware Weighted Proportional Fair User Scheduling

The solution given in (10) and (11) is a proportional fair (PF) user scheduling in which users are scheduled based on the ratio between their instantaneous achievable offloading rates and their average offloading rates. Based on the analysis given in [39] and [40], the average offloading rate of user *k* up to time slot *n* using the PF scheduling given in (10) and (11) can be approximated as:

$$\bar{C}_k(n) = \frac{n_k(n)}{n_{tot}(n)} \mathrm{E}_n\big(c_k(n)\big), \tag{12}$$

where $n_k(n)$ is the number of scheduling periods given to user k up to time slot n and $n_{tot}(n)$ is the total number of scheduling periods provided by the PF scheduler up to time slot *n*. $\mathrm{E}_n\big(c_k(n)\big)$ is the expected value of user *k* offloading rate up to time slot *n*. As it is shown in (12), the average offloading rate of user *k* using PF scheduling highly depends on the number of scheduling periods $n_k(n)$ given to it, which is directly related to user *k* channel conditions and the time span it will stay inside the gate coverage. Almost all analysis of PF scheduling assume that users will stay infinite time under the BS coverage, which is not the case in the gate assumption. By considering users' stay times, the longer user *k* stays inside the gate coverage, the higher probability it may be scheduled. Therefore, using PF scheduling, the gate offloading efficiency and offloading fairness among the users will be degraded if we assume that users will have un-equal stay times inside the gate coverage.

In this paper, to overcome this problem, we propose to use a mobility-aware weighted proportional fair (WPF) user scheduling, in which, more priority is given to the users with shorter stay times inside the gate coverage, as follows:

$$I_{mk}^{WPF}(n) =$$

$$\begin{cases} 1, \text{ if } \forall k \in K_{M(n)} = \underset{\forall K_{M(n)} \subset K(n)}{\arg\max} \left\{ \sum_{k \in K_{M(n)}} \left(w_k(n) \, \dfrac{C_k(n)}{\bar{C}_k(n-1)} \right) \right\}, \\ \\ 0\,, \qquad \text{elsewhere.} \end{cases} \tag{13}$$

where $w_k(n) \geq 1$ is the user k priority factor at time slot *n*. $w_k(n)$ is defined as:

$$w_k(n) = \left(\frac{TS_k(n)}{TS_h(n)}\right)^{-1}, \tag{14}$$

where $TS_k(n)$ is the expected remaining stay time of user k from time slot *n* until it leaves the gate, and $TS_h(n)$ is that of the user with the expected longest stay. $TS_k(n)$ can be expected as:

$$TS_k(n) = \frac{d_k(n)}{\overline{v_{k(n)}}}, \tag{15}$$

where $d_k(n)$ is the distance between the gate exit position and the current position of user *k* at time slot *n*, and $\overline{v_{k(n)}}$ is the average velocity of user *k* at time slot *n* projected to the vector in the direction of the gate exit position. The user *k* initial velocity will be considered as its velocity when it enters the gate. Estimating the current position of user *k* can be effectively done using one of the well-known indoor positioning techniques such as Wi-Fi fingerprints, Wi-Fi trilateration, mobile positioning sensors, etc. [34]. In this paper, for the sake of simplicity, we assume accurate UE location estimation. The effect of the location estimation error on the performance of the proposed scheme and how to compensate it will be a motivation of our future work. Moreover, since the proposed mmWave gate contains a small number of mmWave APs, reducing the complexity of the WPF scheduling algorithm is not considered in this paper and is left as our future work when considering gates with dense mmWave APs deployments.

Lemma *The proposed mobility-aware WPF achieves higher gate offloading efficiency than conventional PF scheduling.*

Proof

To simplify the theoretical analysis, let us assume that there are K users simultaneously enter the gate, and they move toward the gate exit with different velocities. Also, we assume that there are no users inside the gate at their arrival, and no user will enter the gate until they leave. Also, we assume that each user has an infinite number of delayed bytes to be offloaded during its stay inside the gate coverage, i.e., $c_{mk}(n) = r_{mk}(n)$. Therefore, on the average, the total gate offloading rate results from using PF scheduling can be approximated as:

$$C_g^{PF} = \sum_{k=1}^{K} \left(\frac{N_k}{N_{tot}}\right) \mathrm{E}(r_k). \tag{16}$$

where $\mathrm{E}(r_k)$ is the expected value of user k offloading rate during its total stay inside the gate coverage. N_{tot} is the total number of scheduling periods provided by the PF scheduler, which can be expressed as: $N_{tot} = TS_h/T_{\mathrm{sl}}$, where TS_h is the total stay time of the user with the longest stay time inside the gate coverage. N_k is the total number of scheduling periods given to user k during its total stay inside the gate coverage, which can be expressed as: $N_k = \beta_k TS_k/T_{\mathrm{sl}}$, where TS_k is the total stay time of user k inside the gate, and β_k is the scheduling ratio. Definitely, using PF scheduling, user k will not be scheduled in all its stay time if there are other users inside the gate competing for the channel and their number is much larger than the number of operated APs, i.e., $\beta_k < 1$. So, (16) can be re-written as:

$$C_g^{PF} = \sum_{k=1}^{K} \left(\frac{\beta_k TS_k}{TS_h}\right) \mathrm{E}(r_k). \tag{17}$$

In (17), β_k can be considered as the term related to user k channel conditions, and (TS_k/TS_h) is the term related to its stay time inside the gate coverage. Almost all conventional PF analysis neglect the effect of TS_k on C_g^{PF} by assuming $TS_k = TS_h$ [33] [34].

On the other hand, using the proposed mobility-aware WPF, the total gate offloading rate can be approximated as:

$$C_g^{WPF} = \sum_{k=1}^{K} \left(\frac{TS_k}{TS_h}\right)^{-1} \left(\frac{N_k}{N_{tot}}\right) \mathrm{E}(\grave{r}_k), \tag{18}$$

$$= \sum_{k=1}^{K} \left(\frac{TS_k}{TS_h}\right)^{-1} \left(\frac{\grave{\beta}_k TS_k}{TS_h}\right) \mathrm{E}(\grave{r}_k), \tag{19}$$

$$= \sum_{k=1}^{K} \grave{\beta}_k \mathrm{E}(\grave{r}_k), \tag{20}$$

where $\mathrm{E}(\grave{r}_k)$ is the expected value of user k offloading rate during its total stay inside the gate coverage using the proposed WPF scheduling, and $\grave{\beta}_k$ is the user k scheduling ratio using the proposed WPF. If we assume stationary channel conditions, it is obvious that the gate offloading rate given in (20) is higher than that given in (17) by compensating the differences in users stay times. Also, offloading fairness can be attained among the different mobility users.

V. Proposed Adaptive Gate Association for Delayed Offloading eANDSF

In this section, we propose the adaptive process needed to associate a user with delayed traffic to an optimal mmWave gate from its nearby discovered gates. This adaptive process will be implemented in the eANDSF entity inside the C-RAN. Fig. 3 shows the details of the proposed adaptive process. In this process, as the user generates a new delayed file with a new assigned deadline, the process of gates discovery and optimal gate selection is re-initiated. Using the proposed eANDSF, the optimal gate selection is done using a proposed online algorithm given in the following subsection, and it is executed after the eANDSF entity discovers the available gates around the UE based on its estimated location. If the user selects a gate, other than that optimally selected by the eANDSF entity, the UE organizer software will inform the C-RAN about its selection. Hence, the eANDSF entity considers its choice when it selects the optimal gates for the other users. If the user moves away from the selected gate, and it does not select any other gate by itself, the process of gates discovery and optimal gate selection will be re-initiated.

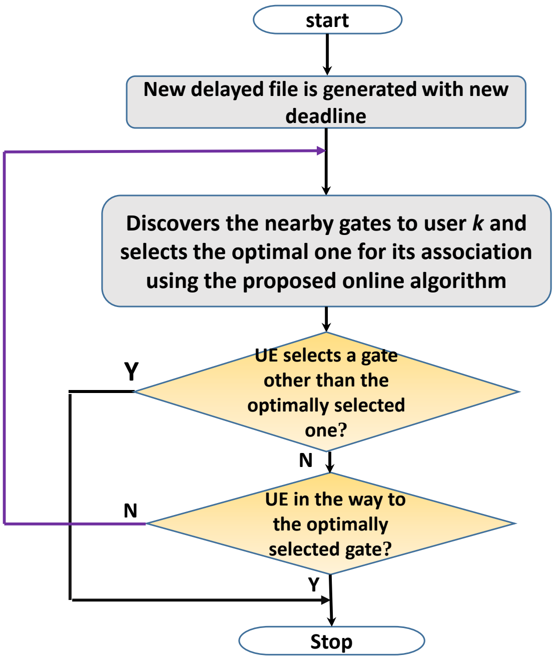
start
New delayed file is generated with new deadline
Discovers the nearby gates to user k and selects the optimal one for its association using the proposed online algorithm
UE selects a gate other than the optimally selected one?
Y
N
UE in the way to the optimally selected gate?
N
Y
Stop

Fig. 3. Proposed adaptive process for associating a user with delayed traffic to an optimal mmWave gate.

A. Optimal MmWave Gate Selection

In selecting an optimal gate for a user with delayed traffic, from its nearby discovered gates, the eANDSF should maximize the total gates offloaded bytes and users OFEs. Increasing user OFE will increase its satisfaction because it postpones files transmission until entering a gate at the aim of reducing its UE energy consumption. The user OFE is defined as:

$$OFE = \frac{\sum User\ bytes\ successfully\ offloaded\ by\ the\ gates}{User\ total\ delayed\ bytes}. \tag{21}$$

1) Problem Formulation

Suppose that some mobile users having different velocities arrive at the area of different capacity mmWave gates, i.e., each gate uses a different number of mmWave APs. Each user has a number of delayed files with different remaining sizes and delay times. At the gate selection time t, let $l_k(t)$ indicates the total number of delayed bytes indexed in the delayed files data base of user k. Also, user k needs different times to reach the available nearby discovered gates, $GRT_{gk}(\mathrm{t})$, where $GRT_{gk}(\mathrm{t})$ is the estimated time required by user k to reach gate g, in seconds. $GRT_{gk}(\mathrm{t})$ mainly depends on the location of registered gate g and the estimated location of user k and its estimated velocity. Based on $GRT_{gk}(\mathrm{t})$ and the remaining delay times of user k delayed files, the C-RAN can expect the load in bytes of user k on gate g if it joins, $l_{gk}(t)$, as follows:

$$l_{gk}(t) = l_k(t) - \sum_{f=1}^{F_{gk}(t)} \left(\pi_{gk}^{f}(t)\right), \tag{22}$$

$$\pi_{gk}^{f}(t) = \mu\left(GRT_{gk}(t) - \Gamma_k^f(t)\right), \quad \Gamma_k^f(t) < GRT_{gk}(\mathrm{t}), \tag{23}$$

where $F_{gk}(t)$ is the total number of user k delayed files expected to have expired deadlines before user k reaches gate g. $\pi_{gk}^{f}(t)$ is the total number of delayed bytes that will be transmitted from the expired deadline file f through the Macro BS before user k reaches gate g, and μ is the Macro BS transmission rate in bytes per second. Thanks to user context information collected by Macro BS control channel, $F_{gk}(t)$ can be easily estimated by the C-RAN via estimating $GRT_{gk}(\mathrm{t})$ and knowing the remaining delay time of file f $\Gamma_k^f(t)$, i.e., $F_{gk}(t) = \left|\forall_f\left(\Gamma_k^f(t) < GRT_{gk}(\mathrm{t})\right)\right|$. Based on these expectations, the eANDSF entity can select an optimal gate for associating user k from its nearby discovered gates. This optimal gate selection should maximize the total gates offloaded bytes and users OFEs. The trivial solution is to select the gate that maximizes (22) by minimizing the total amount of delayed bytes transmitted by the Macro BS. This can be done by associating all users to their nearest available gates with the smallest $GRT_{gk}(\mathrm{t})$. This solution is mainly used by the conventional ANDSF, in which the user always joins the nearest available Wi-Fi hotspot after Wi-Fi scanning. Although this solution highly decreases the amount of delayed bytes transferred by the Macro BS, it does not care either the gate's offloading capacity or its load at the expected user arrival that highly affect the total offloading efficiency and the user OFE. The user may join the nearest available gate at the aim of reducing the number of delayed bytes transmitted by the Macro BS, but unfortunately this gate has a remaining capacity un-relevant to the size of its delayed traffic. To sum up, in solving the problem of optimal gate selection, remaining sizes and delay times of user's delayed files as well as capacities and loads of nearby offloading gates at the expected user arrival, which highly affects the number of scheduling opportunities given to the user, should be jointly considered.

2) *A Network-wide WPF Algorithm for Optimal MmWave Gate Selection*

In this paper, to solve the problem of optimal gate selection in delayed offloading networks based on the aforementioned problem formulation, we consider it as a network-wide WPF problem by maximizing the logarithmic utility function of the gates' average loads, which maximizes the total gates offloaded bytes while maintaining load balance among the gates. Moreover, to consider the different offloading capacities of the distributed gates, a weight parameter corresponding to the gate's offloading capacity is added to the WPF optimization problem. Therefore, the gates WPF optimization problem for optimal gate selection can be written as:

$$\max\left(\sum_{g=1}^{G}\chi_g \log\left(L_g(t)\right)\right) \qquad (24)$$
$$\text{Subject to } \sum_{g\in G} a_{gk}(t)=1, \qquad \forall k\in K,$$

where *t* is the time of gate selection decision done by the eANDSF entity, and $a_{gk}(t)$ is the gate selection index at time *t*. $a_{gk}(t)=1$ if gate g is selected for user *k* at time *t*, and 0 elsewhere. The constraints $\sum_{g\in G} a_{gk}(t)=1$, means that only one gate is selected for user k at the time *t*. χ_g is the weight factor of gate g. Because the delayed traffic load assigned to gate g should be related to its offloading capacity, as it is explained in the above problem formulation, we define χ_g as the offloading capacity of gate g normalized to lowest offloading capacity of the nearby gates; thus $\chi_g \geq 1$. $L_g(t)$ is the expected accumulated average load inside gate g at time *t*, which is given as [41]:

$$L_g(t)=\left(1-\tfrac{1}{T}\right)L_g(t-1)+\tfrac{1}{T}l_{gk}(t)a_{gk}(t), \qquad (25)$$

where $L_g(t-1)$ is the expected average load inside gate g just before the gate selection time *t*. *T* is the averaging low pass filter window size, and $l_{gk}(t)$ given in (22) is the expected load of user *k* on gate *g* if it joins it. Thus, at the time of gate selection *t*, if gate *g* is selected for user *k*, i.e., $a_{gk}(t)=1$, its average load $L_g(t)$ will equal to $\left(1-\frac{1}{T}\right)L_g(t-1)+\frac{1}{T}l_{gk}(t)$. Otherwise, it will equal to $\left(1-\frac{1}{T}\right)L_g(t-1)$.

Following the same procedure given in Section IV using (22) and (25), the solution of the optimization problem in (24) becomes:

$$a_{gk}(t)=$$
$$\begin{cases}1, & \text{if} \quad g=\arg\max\limits_{\forall g\in G}\left(\chi_g\left(\dfrac{l_k(t)-\sum_{f=1}^{F_{gk}(t)}\pi_{gk}^{f}(t)}{L_g(t-1)}\right)\right). \\ 0, & \text{elsewhere.}\end{cases} \qquad (26)$$

As it is given in (26), the proposed solution of optimal gate selection contains all the requirements stated in the problem formulation. At the gate selection time *t*, the proposed solution simultaneously considers the remaining sizes of user *k* delayed files $l_k(t)$ in addition to their remaining delay times relative to GRT_{gk}(t) in the term $\sum_{f=1}^{F_{gk}(t)}\pi_{gk}^{f}(t)$. As well, it considers the gates' offloading capacities χ_g and their expected loads $L_g(t-1)$ at the expected user arrival, which indicates the expected scheduling opportunities given to user *k*. Therefore, if there are two available gates for user *k* with equal expected average loads at its expected arrival, i.e., $L_1(t-1)=L_2(t-1)$, and equal offloading capacities, i.e., $\chi_1=\chi_2$, nearer gate to the user (with smaller GRT_{gk}(t)), will be selected by the eANDSF entity for associating the user. This will highly minimize the number of delayed bytes transmitted by the Macro BS in the user way to the selected gate. At the same time, if the two gates have equal offloading capacities, i.e., $\chi_1=\chi_2$, and user *k* needs the same time to reach them, i.e., GRT_{1k}(t) =GRT_{2k}(t), the gate with a lower average load at the expected user arrival will be selected by the eANDSF entity for associating the user. Finally, if they have equal average loads at the expected user arrival, i.e., $L_1(t-1)=L_2(t-1)$, and user *k* can reach them within the same time, i.e., GRT_{1k}(t) = GRT_{2k}(t), the eANDSF entity selects the gate with a higher offloading capacity for associating the user. Thus, the proposed WPF gate selection algorithm always maximizes the total gates' offloaded bytes and users OFEs.

B. *Proposed Online Algorithm for Gates Discover and Optimal Gate Selection.*

Fig. 4 shows the online algorithm used by the eANDSF entity to discover the nearby gates to user k and select the optimal gate for its association. $GRT_{\max}$ is defined as the maximum allowable gate reaching time, which is a design parameter used for nearby gates discovery. User location, velocity, and its delayed traffic information (remaining sizes and delay times) at the time of optimal gate selection as well as the χ_g values of the gates are also given to the online algorithm.

```
Algorithm Online mmWave gates discovery and optimal gate selection
$t$ is the time of optimal gate selection for user $k$, $t$-1 is the time immediately before the selection decision.
Input: Current delayed load $l_k(t)$ of user $k$, remaining times to the files deadlines $\Gamma_k^f(t)$, estimated UE location and velocity, $GRT_{\max}$, and gates $\chi_g$ values.
    1. Estimate $GRT_{gk}(t), \quad \forall g \in \{GRT_{gk}(t) \le GRT_{\max}\}$
    2. Estimate $F_{gk}$(t), $\sum_{f=1}^{F_{gk}(t)}\left(\pi_{gk}^f(t)\right)$, and $l_{gk}(t)$.
    3. Estimate $L_g(t-1)$.
    4. Select an optimal mmWave gate as:
       $g^*(t) = \underset{\forall g \in \{GRT_{gk}(t) \le GRT_{\max}\}}{\arg\max}\left(\chi_g\left(\frac{l_{gk}(t)}{L_g(t-1)}\right)\right)$
```

Fig. 4. Proposed online algorithm for gates discovery and optimal gate selection.

VI. Simulation Analysis

In this section, we give the simulation analysis that prove the effectiveness of the proposed delayed offloading CC-HetNet under different scenarios. In the first part of the simulation analysis, we prove the effectiveness of using mmWave gates in delayed offloading using different values of delayed files’ deadlines and different numbers of mmWave APs. By this simulation, we estimate the offloading capacity of a mmWave gate, which will be used in the second part of the simulation analysis. Also, simulation analysis that prove the effectiveness of the proposed intra-gate user scheduling will be given. In the second part of the simulation analysis using the pre-evaluated offloading capacity of the gate, we give the simulations that prove the effectiveness of the proposed eANDSF compared to the conventional ANDSF.

A. *MmWave Gate Simulation Analysis*

In this part of simulation analysis, we give the simulation analysis that prove the high impact of using mmWave gates as ultra-high-capacity offloading zones for delayed offloading using the proposed mobility–aware WPF joint user scheduling.

1) *Simulation Scenario and Simulation Area*

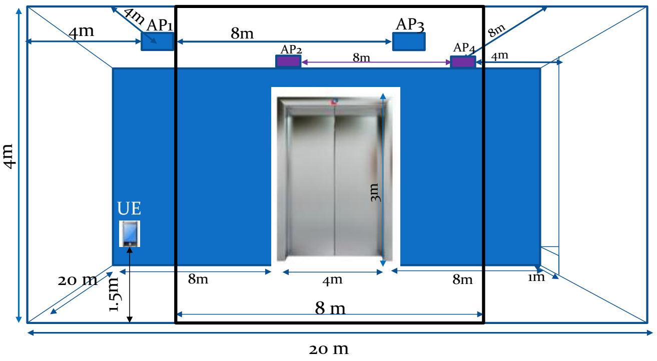

Fig. 5. Ray tracing simulation area of the mmWave gate.

Table I

Intra-mmWave Gate Simulation Parameters

Parameter	Setting
Num. of mmWave APs	4
Num. of UEs inside the gate	14
AP bandwidth	2.16 GHz
MmWave Tx power	10 dBm
MmWave antenna gain	25 dBi
User traffic model	The file sizes are generated from exponential random distribution with an average of 1.67 GB, and the files inter-arrival time (IAT) is generated from exponential random distribution with an average of 10 minutes (the expected mobile traffic density in the year 2024)
UE mobility model inside the gate	Random walk with an average speed of 5 km/hr moving toward the gate exit position
Bandwidth and SNR efficiencies η_B and η_s	0.6 and 1
Time slot period	3 msec
Macro BS transmission rate	100 Mbps
UE LTE module Tx power	23 dBm
Probability of mmWave direct path (LOS) blocking due to human shadowing	0.5
Files deadlines	0, 10, 20 and 30 min

To evaluate the offloading capacity of a mmWave gate, we consider one mmWave gate with an APC, the APC and the Macro BS are tightly coupled to the C-RAN via high-speed backhaul links. We assume that all users reach the gate within the same *GRT* value of 30 min. Each UE has a random number of delayed files to be uploaded/ downloaded, indexed in its delayed files data base. The files are generated with an average traffic density of 1.67 GB/10 min (the expected traffic density of the year 2024 [2]). Each file is associated with a deadline typically assigned by the user or an application program, and the files that have expired deadlines, before the user reaches the gate, will start to be transmitted via Macro BS. As soon as, the user reaches the gate, the ongoing file transmission is switched to be transmitted using the gate through seamless handover controlled by the C-RAN. The mobility model of the user inside the gate is a random walk with an average speed of 5 km/hr moving toward the gate exit position. In the simulation assumption, after the user leaves the gate, the Macro BS switches OFF the UE mmWave module and all its remaining delayed traffic will be transmitted using Macro BS. Also, we consider the use of LTE Macro BS with 100 Mbps net transmission speed, regardless of the UE location as an ideal case of LTE. Fig. 5 shows the ray tracing simulation area of the mmWave gate. In Fig. 5, we disperse the locations of the mmWave APs to reduce the mutual interference as much as possible, since we assume that all APs are using the same communication channel. In addition, to reduce the human blocking effect as much as possible, we attached the APs into the ceiling. All gate materials are made from concrete except the gate entrance door and the elevator are made from glass and metal, respectively. Other simulation parameters are given in Table I.

The steering antenna model defined in IEEE 802.11ad [35] is used as the transmit antenna directivity for mmWave AP, in which, the 3D beam gain in dB is given as [35]:

$$G(\varphi,\theta)[dB] = G_0[dB] - \min[-(G_H(\varphi) + G_V(\theta)), A], \tag{27}$$

$$A[dB] = 12 + G_0[dB], \tag{28}$$

$$G_0[dB] = 20\log_{10}\left(\frac{1.6162}{sin\left(\frac{\theta_{-3dB}}{2}\right)}\right), \tag{29}$$

where φ, θ are the azimuth and elevation angles. $G_H(\varphi) = -\min\left[12\left(\frac{\varphi-\varphi_{\text{beam}}}{\varphi_{-3dB}}\right)^2, A\right]$ and $G_V(\theta) = -\min\left[12\left(\frac{\theta-\theta_{\text{tilt}}}{\theta_{-3dB}}\right)^2, A\right]$ are the beam gains in horizontal and vertical directions, φ_{beam} and θ_{tilt} are the angles of beam center, and φ_{-3dB} and θ_{-3dB} are the half power beamwidths. To fully cover the simulation area, 3D beamforming is considered.

2) Performance Metrics

In the analysis, we concern in measuring the gate offloading efficiency (GOFE), which is defined as:

$$GOFE = \frac{Total\ bytes\ successfully\ transferred\ by\ the\ gate}{Total\ users\ delayed\ bytes}. \tag{30}$$

Also, we measure the gate offloading time (T_{gate}), which is the total time in seconds taken by the gate to accomplish the offloading process. The gate's offloading capacity as a function of the used number of APs can is also measured using the following equation:

$$C_g(M) = \frac{Total\ bits\ successfully\ transferred\ by\ the\ gate(M)}{T_{\text{gate}}(M)}. \tag{31}$$

where M is the total number of operated APs. We also measure the energy efficiency of using delayed offloading by calculating the average value of UE normalized energy consumption (E_{Norm}) results from using delayed offloading, which can be defined as:

$$\text{E}_{\text{Norm}} = \frac{1}{K}\sum_{k=1}^{K} \text{E}_{\text{Norm}}^{k}\ , \tag{32}$$

$$\text{E}_{\text{Norm}}^{k} = \frac{\left(P_{\text{LTE}} \times T_{\text{LTE}}^{k(\text{del})} + P_{\text{mmWave}} \times T_{\text{mmWave}}^{k(\text{del})}\right)}{\left(P_{\text{LTE}} \times T_{LTE}^{k(\text{no})}\right)}. \tag{33}$$

where K is the total number of simulated users, P_{LTE} is the transmit power of user k using the LTE module, and P_{mmWave} is its transmit power using the mmWave module. $T_{\text{LTE}}^{k(\text{del})}$ is the total time that user k uses its LTE module in case of delayed offloading, and $T_{\text{LTE}}^{k(\text{no})}$ is the total time in case of no offloading. $T_{\text{mmWave}}^{k(\text{del})}$ is the total time that user k uses its mmWave module in case of delayed offloading.

The effectiveness of the proposed mobility-aware WPF joint user scheduling in comparison to the conventional PF and round robin (RR) is measured using GOFE and byte offloading fairness $F_{BO}(\Delta t)$, which is defined as:

$$F_{BO}(\Delta t) = \frac{\left(\sum_{k=1}^{K} BO_k(\Delta t)\right)^2}{\left(K\sum_{k=1}^{K}\left(BO_k(\Delta t)\right)^2\right)}, \tag{34}$$

where Δt is the total simulation time which is equal to TS_h. Hence, $BO_k(\Delta t)$ denotes the total number of bytes offloaded by user k, during its total stay inside the gate coverage. $F_{BO}(\Delta t)$ = 1 indicates that all users offload the same amount of bytes.

3) Simulation Results

Fig. 6 shows GOFE and T_{gate} using different number of APs and files' deadlines, with *GRT* of 30 min. From Fig. 6, as the files' deadlines are increased, a higher number of delayed bytes will reach the gate. Thus, GOFE is highly increased, especially if the gate uses a high number of APs.

Using files' deadlines of 30 min, 100 % GOFE can be obtained using 4 APs within T_{gate} of 25 sec, which means that the gate successfully offloads an average delayed traffic of 70 GB within multiple seconds. It is interesting to compare this result with the case of the on-the-spot offloading using distributed mmWave APs (existing proposals [5] [9]). By assuming mmWave coverage of 10 m around the AP and user average speed of 5 km/hr (Table I), all users will stay about 14 sec under one AP coverage. Therefore, to obtain an

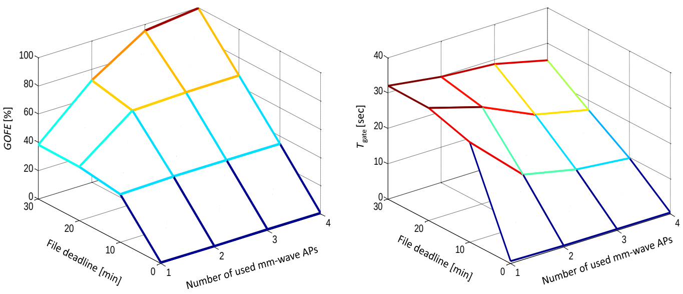

Fig. 6. *GOFE* and T_{gate}.

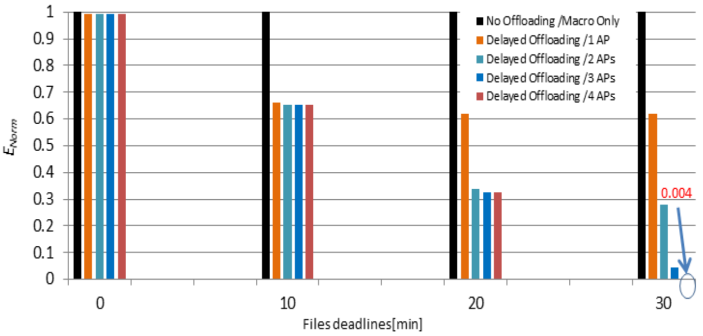

Fig. 7. Average value of the normalized UE energy consumption.

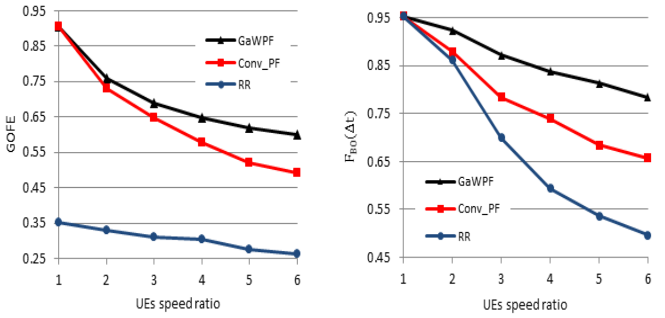

Fig. 8. GOFE and byte offloading fairness comparisons.

offloading efficiency of 100 % and fully cover all users in a time span of *GRT* + 25 sec (~30.5 min), the number of required APs will equal to $\frac{(30.5\times60\,)}{14}\approx 131$ APs. This high number of APs highly increases the deployment cost and complicates the required control compared to the proposed delayed offloading CC-HetNet. Using the data given in Fig. 6 and using (31), the gate's offloading capacity $C_g(M)$ becomes 6.5, 12, 18 and 22 Gbps for 1-AP, 2-APs, 3-APs and 4- APs, respectively.

Fig. 7 shows the average value of the normalized UE energy consumption using different number of APs and files' deadlines. It is interesting to find out that a mmWave gate consists of 4 APs, with files deadlines' of 30 min, can reduce the UE energy consumption by 99.6 % compared to the no offloading case (Macro BS only). By these results, mmWave gates are much better candidates than Wi-Fi APs for delayed offloading. This is because massive delayed traffic can be offloaded within multiple seconds, which saves user's time and energy.

Fig. 8 confirms the effectiveness of the proposed gate WPF (GaWPF) user scheduling over the conventional PF (Conv_PF) and the round robin (RR) scheduling. The scheduling algorithms are tested using different UEs speed ratios over the total time span of the gate coverage, where a speed ratio is the ratio between the highest speed user and the lowest speed one. UEs are assumed to move in the same direction toward the gate exist position. Hence, they have un-equal stay times inside the gate. A gate with three operating APs is used in the simulation, and all users have infinite delayed traffic to be offloaded. From Fig. 8, the proposed GaWPF has the best performance among the tested schemes in GOFE and $F_{BO}(\Delta t)$, because it takes into account the different stay times (mobilities) of the users when it makes a joint user scheduling decision.

B. *Simulation Analysis of MmWave Gate Association*

In this part of the simulation analysis, we give the simulations that prove the effectiveness of the proposed eANDSF for delayed offloading using mmWave gates. The gates' offloading scapacities $C_g(M)$ evaluated in (31) are used through the simulations.

1) Simulation Scenario and Simulation Area

In the simulation scenario, we assume a number of randomly distributed UEs inside the simulation area of 2 km x 2 km. Likewise, a number of mmWave gates with different number of APs and different offloading capacities $C_g(M)$ are randomly distributed within the same area. Each UE has a random velocity in the range of [5, 50] km/hr. Hence, $GRT_{\max}$ will be in the range of $\frac{2\sqrt{2}}{5}\approx 30$ min, which is our definition of the farthest nearby gate. Although we assume ideal UE location estimation, practical UE location estimation can be easily considered by introducing some errors in the estimated $GRT_{gk}(t)$. Also, we assume that each user has a random number of delayed files with different remaining sizes and delay times. This delayed traffic can be a remaining traffic after the user left a mmWave gate, or it can be initially generated by the user for the first time. The total size of a user's delayed traffic is generated using a uniform random distribution in range of [1, 15] GB. Without loss of generality and for the purpose of fair comparisons between the compared schemes, we assume that no user generates delayed files until reaching its associating gate. In addition, the users always move toward their selected gates, although the proposed adaptive process can handle any case. A delay time, generated using a uniform random distribution in the range of [0, 90] min, is assigned to each delayed file. The zero deadline file is a file with an expired deadline. The files with expired deadlines before the user reaches the selected gate will start to be transmitted using the Macro BS. As soon as, the user reaches the gate, the ongoing file transmission is switched to be offloaded using the gate through seamless handover controlled by the C-RAN. In the simulation assumption, we assume that the user average stay inside the gate coverage is about 30 sec, which corresponds to an average walking speed of 2.5 km/hr with the gate dimensions given in Fig. 5. After the user leaves the gate, the Macro BS switches OFF the UE mmWave module and all user's remaining files will be transmitted using Macro BS. Table II summarizes the used simulation parameters.

2) Compared Gate Association Schemes

In simulation comparisons, we compare the performance of two gate association algorithms. The first one is the conventional ANDSF with the user always selects the shortest *GRT* gate, which is the currently existing approach in literature [27]. The second scheme is the proposed eANDSF

TABLE II

SIMULATION PARAMETERS OF MMWAVE GATE ASSOCIATION

Parameter	Setting
Simulation area	2 km x 2 km
Num. of UEs	500 ~ 4000
Num. of mmWave gates	2 ~ 30
The user delayed files.	Random number of delayed files with total traffic size in the range of [1, 15] GB.
Gate offloading capacity ($C_g(M)$)	[6.5 (1-AP) 12 (2-APs) 18 (3-APs) 22 (4-APs)] Gbps. χ_g = [1 1.85 2.77 3.4]
Files remaining delay time model	Uniform random distribution within the range of [0, 90] min
Average user stay time inside the gate coverage	30 sec
$GRT_{\max}$	30 min

using the network-based online algorithm for gates discovery and optimal gate selection.

3) Performance Metrics

In the conducted simulations, we concern in measuring the followings:

- *Total Gates Offloaded Bytes (G_{BOF})*

$$G_{BOF} = \sum_{t=\text{int}}^{\text{finl}} \sum_{g=1}^{G} \min(l_g(t), C_g(M(t))), \tag{35}$$

$$l_g(t) = \sum_{k=1}^{K_g(t)} l_{gk}(t), \tag{36}$$

where, $t = \text{int}$ is the simulation starting time, and $t = \text{finl}$ is the simulation ending time. For calculating the total number of bytes offloaded by a gate g, at every second t, we count the total number of users inside the gate $K_g(t)$. $K_g(t)$ is used by the APC to assign the number of mmWave APs $M(t)$ required for mmWave concurrent transmissions with total gate offloading capacity of $C_g(M(t))$, as given in Table II.

$$M(t) = \begin{cases} K_g(t) & K_g(t) < M \\ M & K_g(t) \geq M \end{cases}, \tag{37}$$

where M is the total number of mmWave APs inside the gate. Then, the total traffic load inside the gate $l_g(t)$ is calculated using (36) and compared with $C_g(M(t))$ using (35) to find out the total number of bytes a gate g can offload at every second t.

- *Average User OFE*

The average user OFE can be calculated as:

$$OFE = \frac{1}{K} \sum_{k=1}^{K} OFE^k, \tag{38}$$

where K is the total number of simulated users, and OFE^k is the offloading experience of user k as defined in (21).

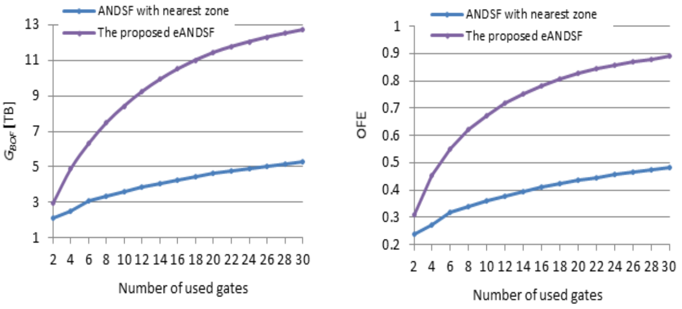

Fig. 9. Total gates offloaded bytes and average users OFE using 2000 users.

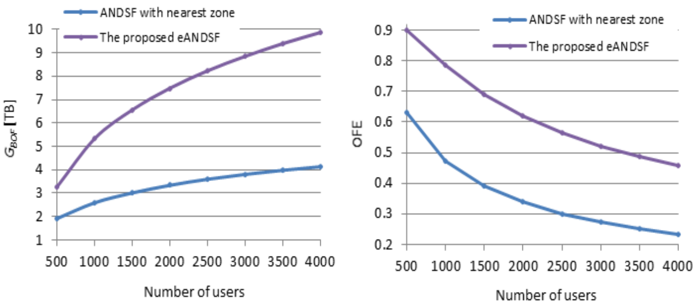

Fig. 10. Total gates offloaded bytes and average users OFE using 8 gates.

4) Simulation Results

Fig. 9 shows G_{BOF} in Tera Byte [TB] and the average users OFE using a fixed number of users (2000 users) and different number of gates (2 ~ 30 gates), while Fig. 10 gives G_{BOF} and the average users OFE using a fixed number of gates (8 gates) and different number of users (500 ~ 4000 users). These figures prove the significance of the proposed eANDSF in maximizing the total gates offloaded bytes compared to the conventional ANDSF. The proposed eANDSF has more than double the performance of the conventional ANDSF using 30 gates / 2000 users and 8 gates / 4000 users. Also, the proposed eANDSF always maximizes the average user OFE compared to the conventional ANDSF. This is because, the proposed eANDSF considers the user's delayed traffic relative to the gates' remaining capacities at the expected user arrival when it makes a gate selection decision. Instead, in the conventional ANDSF, the user itself who selects the associating gate from its nearby discovered gates. Therefore, no load balance relative to gates' offloading capacities is guaranteed. Consequently, the gates with low offloading capacities may be heavily loaded while the high capacity ones are lightly loaded. As a result, total gates offloaded bytes and users OFE are not maximized.

VII. Conclusion

In this paper, as a contribution to solve the capacity problem of future cellular networks, we proposed to implant the ultra-high speed mmWave access into cellular networks in the form of mmWave gates consisting of many coordinated mmWave APs. The concept of delayed offloading is investigated in conjunction with the proposed gates to relax the demand of deploying a high number of them, which greatly reduces the deployment cost and the required control compared to the current proposals. To efficiently control the delayed offloading process over the gates, we proposed to tightly couple the gates and the Macro BS through the C-RAN using the proposed delayed offloading CC-HetNet. We gave the comprehensive architecture of the proposed CC-HetNet in addition to the protocol that organizes its operation.

Then, we highlighted the challenge of intra-gate user scheduling comes from the un-equal stay times of the users passing through the gate. An efficient mobility-aware WPF joint user scheduling was proposed to cope with this challenge. By theoretical means, we proved the high potential of the proposed WPF compared to the conventional PF. Another highlighted challenge inherent in delayed offloading concept was the design of an eANDSF to efficiently discover and select an optimal offloading zone for associating a user with delayed traffic. Thanks to the tightly coupled CC-HetNet, an adaptive gate association scheme including an online algorithm for gates discovery and optimal gate selection was proposed as eANDSF for delayed offloading networks.

We proved the high impact of the proposed mmWave gate in terms of GOFE, gate offloading time, gate offloading capacity, and normalized UE energy consumption. Also, we proved the high efficiency of the proposed mobility–aware WPF joint user scheduling compared to the conventional PF and round robin schemes. Also, we proved the high impact of the proposed eANDSF over the conventional ANDSF in optimizing the performance of delayed offloading networks in terms of total gates offloaded bytes and average user OFE. A mmWave gate consisting of only 4 mmWave access points (APs) can offload up to 70 GB of delayed traffic within 25 sec, which highly relaxing the traffic demand of the Macro BS and reduces the energy consumption of a user equipment (UE) by 99.6 % compared to the case of only using Macro BS without gate offloading. Also, more than a double increase in total gates offloaded bytes is obtained using the proposed eANDSF over using the conventional ANDSF proposed by 3GPP due to the optimality in selecting the associating gate.

Although the proposed delayed offloading CC-HetNet has a promising impact in solving the capacity problem of future cellular networks, further investigations are need to be done. One of these investigations is the possibility of automatic adjustments of the files deadlines based on user location and velocity and nearby gates conditions. Also, the detailed network architecture of delayed offloading CC-HetNet including the required interfaces, gateways, protocol stacks and synchronization should be considered.

References

[1] K. Sakaguchi et al., “Cloud cooperated heterogeneous cellular networks”, in Proc. IEEE ISPACS, pp. 787-791, Nov. 2013.

[2] Cisco, “Cisco visual networking index: global mobile data traffic forecast update, 2014-2019,” Feb. 2015 [online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

[3] Lu Lu, G.Y. Li, A.L. Swindlehurst, A. Ashikhmin and Rui Zhang, “An overview of massive MIMO: Benefits and challenges,” IEEE J. of Sel. Topics in Sig. Proces., vol. 8, no. 5, pp. 742 – 758, Apr. 2014.

[4] R. Q. Hu, Y. Qian, S. Kota and G. Giambence, “HetNets - a New paradigm for increasing cellular capacity and coverage [Guest Editorial],” IEEE Wireless Commun., vol. 18, no. 3, pp.8-9, Jun. 2011.

[5] K. Sakaguchi et al “Millimeter-wave evolution for 5G cellular networks,” IEICE Trans. Commun, vol. E98-B, no. 3, pp.338-402, Mar. 2015.

[6] T. S. Rappaport et al., “Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications,” IEEE Trans. On Antennas and Propagation, vol. 61, No. 4, pp. 1850-1859, Apr. 2013.

[7] T. S. Rappaport et al., “Millimeter wave mobile communications for 5G cellular: it will work!,” IEEE Access, vol. 1, pp. 335-349, May 2013.

[8] E. M. Mohamed, K. Sakaguchi and S. Sampei, “Delayed offloading using cloud cooperated millimeter wave gates,” in Proc. IEEE PIMRC, pp. 1852- 1856, Sept. 2014.

[9] H. Peng, T. Yamamoto, and Y. Suegara, “Extended user/control plane architectures for tightly coupled LTE/WiGig interworking in millimeter-wave heterogeneous networks,” in Proc. IEEE WCNC, pp. 1566-1571, Mar. 2015.

[10] T. Bai, A. Alkhateeb, and R. W. Heath Jr, “Coverage and capacity of millimeter-wave cellular networks,” IEEE Commun. Magaz., vol.52, no.9, pp.70-77, September 2014.

[11] R. J. Weiler et al., “Enabling 5G backhaul and access with millimeter-waves,” In Proc. of EuCNC 2014, pp. 1-5, Jun. 2014.

[12] K. Hosoya et al., “Multiple sector ID capture (MIDC): a novel beamforming technique for 60 GHz band multi-Gbps WLAN/PAN systems,” IEEE Trans. Antennas Propagat., vol .63, no. 1, pp. 81-96, Jan. 2015.

[13] E. M. Mohamed, K. Sakaguchi and S. Sampei, “Millimeter wave beamforming based on WiFi fingerprinting in indoor environment,” in Proc. IEEE ICC Workshops, Jun. 2015.

[14] IEEE 802.11ad standard: “Enhancements for Very High Throughput in the 60 GHz Band,” 2012.

[15] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, “Mobile data offloading: how much can wifi deliver?” IEEE/ACM Trans. On Net., vol. 21, no. 2, pp. 536-550, Apr. 2013.

[16] A. Balasubramanian, R. Mahajan, and A. Venkataramani, “Augmenting mobile 3G using wifi,” in Proc. ACM MobiSys, pp. 209-222, Jun. 2010.

[17] J. Less, Y. Yi, S. chong, and Y. Jin, “Economics of wifi offloading: trading delay for cellular capacity,” IEEE Trans. on wireless commun., vol. 13, no. 3, pp. 1540-1554, Mar. 2014.

[18] Y. Im, C. Joe-Wong, S. Sen, T. T. Kwon, and M. Chiang, “AMUSE: Empowering users for cost-aware offloading with throughput-delay tradeoffs,” in Proc. INFOCOM, pp. 435-439, Apr. 2013.

[19] A. J. Nicholson and B. D. Noble, “Breadcrumbs: forecasting mobile connectivity,” in Proc. ACM MOBICOM, pp. 46-57, Sept. 2008.

[20] T. Nakamura et al., “Trends in small cell enhancements in LTE advanced,” IEEE Commun. Magaz., vol. 51, no. 2, pp. 98-105, Feb. 2013.

[21] C. Hoymann, D. Larsson, H. Koorapaty, and C. Jung-Fu, “A lean carrier for LTE,” IEEE Commun. Magaz., vol. 51, no. 2, pp. 74-80, Feb. 2013.

[22] http://www.3gpp.org/

[23] K. Sakaguchi et al “Millimeter-wave wireless LAN and its extension toward 5G heterogeneous networks,” IEICE Trans. Commun, vol. E98-B, no. 10, pp.1932-1948, Oct. 2015.

[24] X. Zhuo, W. Gao, G. Cao, and S. Hua, “An incentive framework for cellular traffic offloading,” IEEE Trans. on Mobile Computing, vol. 13, no. 3, pp. 541-555, Mar. 2014.

[25] M. H. Cheung and J. Huang, “Optimal delayed offloading,” in Proc. IEEE WiOpt, pp. 564-571, May 2013.

[26] V.A. Siris and M. Anagnostopoulou, “Performance and energy efficiency of mobile data offloading with mobility prediction and prefetching,” in Proc. IEEE WoWMoM 2013, pp. 1-6, Jun. 2013.

[27] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, “Cellular traffic offloading through wifi networks,” in Proc. IEEE MASS, pp. 192-201, Oct. 2011.

[28] V.A. Siris and M. Anagnostopoulou, “Performance and energy efficiency of mobile data offloading with mobility prediction and prefetching,” in Proc. IEEE WoWMoM 2013, pp. 1-6, Jun. 2013.

[29] 3GPP TS 24.312V12.7.0: “Access network discovery and selection function (ANDSF) management object (MO)”.

[30] Y. Chan, W. Tsui, H. So and P. Ching, “Time-of-arrival based localization under NLOS conditions,” IEEE Vehic. Techn. Trans., vol. 55, no. 1, pp. 17-24, Jan. 2006.

[31] R. Mondal, J. Turkka and T. Ristaniemi, “An efficient grid-based RF fingerprint positioning algorithm for user location estimation in heterogeneous small cell networks,” In Proc. IEEE ICL-GNSS, pp. 1-5, Jun. 2014.

[32] K. Chen et al., “C-RAN the road towards green RAN,” China Mobile Research Institute, white paper, 2011.

[33] Yilin Zhao , “Standardization of mobile phone positioning for 3G systems,” IEEE Commun. Magaz., vol. 40, no. 7, pp. 108 – 116, Jul. 2002.

[34] H. Liu, H. Darabi, P. Banerjee and J. Liu, “Survey of wireless indoor positioning techniques and systems,” IEEE Transc. Syst., Man and Cybernet., Part C (Applications and Reviews), vol. 37, no. 6, pp. 1067-1080, Nov. 2007.

[35] doc.: IEEE 802.11-09/0334r8, “Channel models for 60 GHz WLAN systems,” May 2010.

[36] P. Mogensen et al, “LTE capacity compared to the shannon bound,” In Proc. IEEE VTC 2007, pp. 1234 – 1238, Apr. 2007.

[37] K. Son, S. Chong, and G. Veciana, “Dynamic association for load balancing and interference avoidance in multi-cell networks,” IEEE Trans. On Wireless Commun., vol. 8, no. 7, pp. 3566-3576, Jul. 2009.

[38] H. Kim, K. Kim, Y. Han, and S. Yun, “A proportional fair scheduling for multicarrier transmission systems,” IEEE Commun. Letters, vol. 9, no. 3, pp. 210-2012, Mar. 2005.

[39] H. J. Kushner and P. A. Whiting, “Convergence of proportional fair sharing algorithms under general conditions,” IEEE Trans. Wireless Commun., vol. 3, no. 4, pp. 1250-1259, Jul. 2004.

[40] C. Yang, W. Wang, Y. Qian, and X. Zhang, “A weighted proportional fair scheduling to maximize best-effort service utility in multicell network,” in Proc. IEEE PIMRC 2008, pp.1-5, Sept. 2008.

[41] E. M. Mohamed, K. Sakaguchi and S. Sampei, “Delayed offloading zone associations using cloud cooperated heterogeneous networks,” in Proc. IEEE WCNC workshops, pp. 374- 379, Mar. 2015.