Localized Mobility Management for 5G Ultra Dense Network

Hucheng Wang, Shanzhi Chen, *Senior Member, IEEE*, Ming Ai, and Hui Xu

***Abstract*—It is commonly agreed that the ultra dense network (UDN) will be a key technology to face extremely dense traffic and high-speed data rate in the fifth-generation (5G) network. However, due to its new characteristics such as high dense small cells, fast and flexible deployment of small cell access points, and flexible backhaul connectivity, how to enable mobility support is becoming a great challenge. In this paper, based on newly proposed network architectures for UDN, we present two efficient localized mobility management schemes considering small cell deployments and backhaul topology. The first one centralizes mobility management control from small cell access points into a local access server (LAS) closing to radio access network. Another one allows individual small cell access points to handle mobility events, but still requires the LAS to act as mobility anchor. According to the performance evaluation results of the proposed schemes by using numerical analysis and simulation, respectively, including average handover signaling cost, average packet delivery cost, average handover latency, and average signaling load to the core network, the localized mobility management with centralized control scheme has the best performance, and the other one has less handover signaling cost, but higher handover latency than the third-generation partnership project (3GPP) scheme.**

***Index Terms*—Fifth-generation (5G), handover, mobility management, ultra dense network (UDN).**

I. Introduction

DRIVEN largely by new mobile internet services such as ultra-high definition video, 3-D-video, virtual reality, and augmented reality, mobile data traffic will grow rapidly in near future. According to the most recent Cisco VNI forecast (May 2015) [1], traffic from wireless and mobile devices will exceed traffic from wired devices by 2019, and the mobile data traffic will increase 10-fold than that in 2014. This huge capacity demands place stringent requirements to the future mobile network, including high speed data rate and extremely dense traffic and connections. For example, the researches in China [2] forecast that the peak data rate in fifth generation (5G) will reach tens of Gbps, the traffic volume density will reach tens of Gbps per square kilometer and the connection density will reach 1 million connections per square kilometer. Such strict requirements cannot be met merely by the advanced radio access technologies because the capacity of single BS (base station) is still very limited, thus providing connection(s) to the mobile terminal via additional SC (small cell) is deemed as a viable solution. Just as "dual connectivity" technology developed by 3GPP (3rd Generation Partner Project) [3], a UE (user equipment) is allowed to simultaneously connect to macro eNB (eNodeB) and SeNB (secondary eNodeB), which provides additional radio resources in small coverage area. By overlaying macro cell with a number of SCs, the overall system capacity can be enhanced significantly. Based on the 5G requirements, it is speculated that ultra dense SCs will be deployed in 5G mobile networks, i.e. UDN (Ultra Dense Network) would be supported. In fact, the UDN has been promoted as one of key technologies to meet the high throughput requirement in 5G by ITU-R [4]. This new technology requires not only new mechanisms on the air interface, including interference management [5], [6], radio resource management [7], and coverage optimization [8], but also new network architecture with new mobility management mechanism, for efficient network resource utilization and good user experience.

Considering new characteristics of UDN, e.g. extremely high dense SC APs (access points), fast and flexible deployment of SC APs etc., it is not suitable to still use traditional mobile network architectures and mobility management schemes, such as those in LTE (long term evolution) network because i) the traditional network architectures place centralized network control in core network, thus location management and handover control are always performed by the core network, which would lead to high signaling overhead in UDN; ii) crossing cells in UDN causes lots of inter-small cell handovers, which would lead to high signaling overhead and frequent session interruption if any mobility management scheme being used in current mobile operators' networks, i.e. that developed by 3GPP, was applied; iii) handover decision in the traditional network is made by radio access point without considering the deployment of BSs and backhaul topology, but if similar mechanism is applied to UDN with flexibly deployed SC APs, the most suitable target SC(s)

Manuscript received January 17, 2016; revised May 31, 2016, October 30, 2016, and April 4, 2017; accepted April 10, 2017. Date of publication April 19, 2017; date of current version September 15, 2017. This work was supported by the National Science and Technology Major Project of China under Grant 2016ZX03001_002_004 and Grant 2017ZX03001014, the National Natural Science Foundation of China for Distinguished Young Scholar under Grant 61425012, and the National High-Technology Program (863) of China under Grant 2015AA01A706. The review of this paper was coordinated by C. Assi. *(Corresponding author: Hucheng Wang.)*

H. Wang is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100088, China, and State Key Laboratory of Wireless Mobile Communications, China Academy of Telecommunications Technology, Beijing 100083, China (e-mail: huchengwang@gmail.com).

S. Chen, M. Ai, and H. Xu are with the State Key Laboratory of Wireless Mobile Communications, China Academy of Telecommunications Technology, Beijing 100083, China, and also with Datang Telecom Technology & Industry Group, Beijing 100191, China (e-mail: chensz@datanggroup.cn; aiming@catt.cn; xuhui@catt.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2017.2695799

would not be selected. Therefore, new network architecture and mobility management for UDN need to be researched.

Existing research [9]–[12] proposes new network architecture to UDN. [9] and [10] propose to apply C-RAN (cloud RAN) architecture to UDN for efficient radio resource management or mobility management. However, C-RAN requires ideal fronthaul for each RRH (remote radio head), thus it may not be applicable to dynamically deployed SCs with non-ideal fronthauls. The research in [11] proposes to dynamically partition operator's network into a number of districts, so that extremely dense wireless networks on each district can be controlled by a CLC (CROWD Local Controller) and served by distributed Gateways. The proposed network architecture and corresponding mobility management scheme aim to avoid the issues about sub-optimal routing, scalability and reliability etc. However, coordination among heterogeneous wireless access points, including legacy 3G BS, LTE BS, WiFi AP, etc., is not supported; thus, there is still heavy signaling overhead in the wireless network. [12] proposes a C/U (control plane and user plane) split architecture for 5G UDN, where the macro cell maintains "always-on" coverage, the SCs provide high throughput capacity, and each SC is connected to a macro eNB through Xn interface. By removing public system information and cell level control signaling from SC, faster access and higher spectrum efficiency can be achieved. However, when a UE crosses SCs served by different macro eNBs, additional signaling is required for inter-macro eNB handover.

With regard to mobility management mechanism in UDN, most of existing researches propose to locally handle mobility events, for example, in [13], the authors propose a local anchor based mobility management scheme for HetNet (heterogeneous network). The solution defines the local anchor as the SC maintaining links with other SCs in a cluster. Relying on the coordination with other SCs in the cluster, the local anchor can concentrate the uplink and downlink traffic, relay LTE S1 messages, and act as local mobility anchor for handover between SCs. Compared to existing 3GPP schemes, this mobility management scheme can minimize signaling load in core network, and save the total handover costs and handover interruption time, whereas, it is difficult to group SCs if backhaul topology is not available, especially for dynamically deployed SCs. Authors in [14] not only propose to apply local anchor based mobility management to UDN, but also suggest integrating SDN (software defined networking) into backhaul network for providing the shortest forwarding chain between current serving SC and anchor SC. When the length of forwarding chain exceeds the pre-defined threshold, the SDN controller needs to setup a new forwarding path between serving SC and Serving GW (gateway). Although the solution can promise the shortest forwarding chain between serving SC and anchor SC, it cannot promise the shortest forwarding chain between serving SC and Serving GW, because the shortest path between the anchor SC and the Serving GW may be longer. The researches in [15] provide a handover scheme utilizing cooperation based cell clustering, which can reduce handover signaling overhead in core network, but the interactions between an anchor cell and its neighboring cells still require lots of control signaling. The authors of [16] focus on identifying potential handover candidates in small cell networks, and then propose a distance based neighboring cells scanning algorithm to minimize the number of small cells scanning, whereas no specific handover procedure is touched.

In this paper, by analyzing the deployment of UDN, we propose new network architectures and corresponding localized mobility management (LMM) schemes. Based on our previous works [17] and [18], we extend the functions of Local Access Server (LAS) consisting of Local Service Center (LSC) and Local Data Center (LDC), so that the LSC can collect topology status of backhaul network and maintain the backhaul connections; the LDC responsible for data forwarding can act as user plane mobility anchor. Under this network architecture, we propose to apply LMM with centralized control scheme to UDN with heterogeneous deployment of macro cells overlaid with many SCs, and LMM with distributed control scheme to UDN with independent SCs. The LMM with centralized control scheme requires the LSC to handle mobility events, and in contrast, the LMM with distributed control scheme allows each SC AP to handle mobility events itself. Under the developed 3D network deployment model, relying on the numerical analysis using Markov Chain model and simulation experiment, we demonstrate that the proposed LMM with centralized control scheme has the least handover signaling cost and the lowest handover latency, and the LMM with distributed control scheme has less handover signaling cost but higher handover latency than 3GPP scheme. Moreover, we demonstrate that the network with centralized backhaul topology management has minimized packet delivery cost. Finally, the evaluation of signaling load to the core network shows that both LMM schemes have much less signaling cost than 3GPP scheme and local anchor schemes introduced in [13].

The rest of this paper is organized as follows. In Section II, the deployment characteristics of UDN are analyzed and the corresponding network architecture is illustrated. In Section III, we propose two LMM schemes under the network architecture depicted in Section II. In Section IV, the analytical model is developed for the proposed LMM schemes. Furthermore, performance metrics are introduced. In Section V, by using current 3GPP scheme as the baseline, the performance of each proposed scheme is evaluated. Finally, the article is concluded in Section VI.

II. Deployment Scenario and Network Architecture

It is necessary to investigate the network deployment for designing a reasonable architecture. Compared with current cellular mobile network, the deployment of UDN has the following characteristics.

First, the deployment of SC APs is flexible for different purposes, such as operator-deployed picocells for traffic offloading, user-deployed femtocells for indoor coverage, and RF (radio frequency) units for extending the coverage of a central BS introduced in [19].

Second, due to high dense SC APs in the hotspot area such as shopping mall, airports, large office buildings or auditoriums, it is unnecessary even impossible to connect all SC APs to network directly. Let "planned SC APs" denote the SC APs connecting

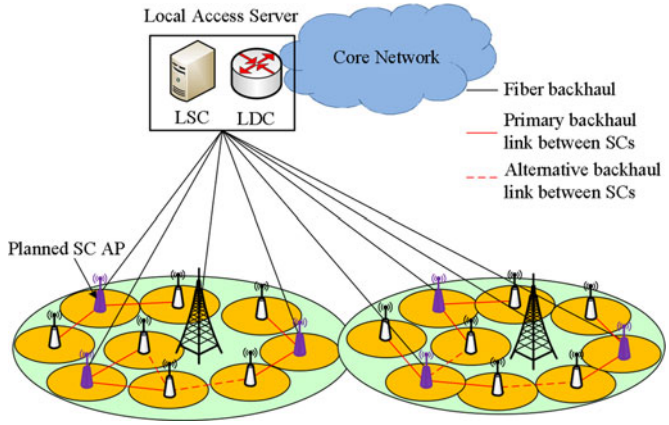

Fig. 1. General network architecture for UDN.

to the network directly, other SC APs have to connect to the network via the planned SC APs.

Third, there are both wired and wireless backhaul in UDN. The deployment of SC APs subjects to many geographical and physical constrains, such as road, building etc., it is impossible to lay wired backhauls for all SC APs. Therefore, just as introduce in [20] and [21], wireless backhaul will play an important role in the UDN.

Based on the characteristics of UDN analyzed above, we propose to apply the network architecture introduced in [17] to the UDN. Due to the flexible deployment of SC APs and the change of backhaul conditions, especially for SC APs with wireless backhaul, the backhaul topology may be changed dynamically. In order to efficiently calculate the optimal path for each SC AP for establishing backhaul connection in such network, a feasible solution is to centrally maintain backhaul topology. In our architecture, as shown in Fig. 1, the LSC is used to collect backhaul network status and manage the backhaul topology, so that the optimal path between LDC and SC AP can be easily calculated and established, for example, when the LSC detects that a primary backhaul link which connects a SC AP to the LDC is broken, it can rapidly update the backhaul connection to go through an alternative backhaul link.

It is proved that internet traffic accounts for large proportion in mobile network, it is wisdom to route such traffic to internet as soon as possible, but it is cost ineffective to deploy gateway function in the SC APs in UDN because of high dense deployment of SC APs. Therefore, we need a traffic concentrator for traffic offloading, i.e. the LDC.

III. Localized Mobility Management Schemes

While implementing LMM mechanism in the architecture introduced in Section II, it is a good option to use the LDC as the user plane local mobility anchor. However, depending on specific mobility management scheme, the control plane mobility management functions may be deployed on different network entities. The Table I shows the mobility management schemes choosing different network entities to deploy control plane mobility management functions, i.e. LMM with centralized control and LMM with distributed control.

TABLE I
Comparison of LMM Schemes

	LMM with centralized control	LMM with distributed control
Control plane functional entities	LSC, macro BS	LSC, SC AP
Handover initiation	LSC	SC AP
Target SC selection	LSC	SC AP, LSC
Handover decision	LSC	SC AP
Handover control	LSC	SC AP, LSC
Backhaul topology management	LSC	LSC
User plane mobility anchor	LDC	LDC

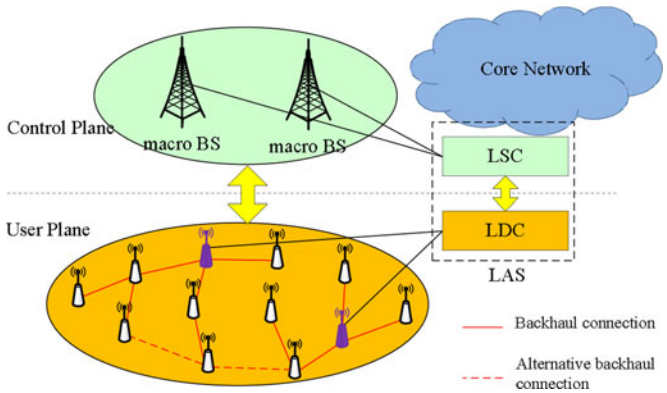

Fig. 2. Logical network architecture with centralized control.

A. LMM With Centralized Control

Learning from SDN principles [22], we separate the network described in Fig. 1 into control plane and user plane. The control plane consists of LSC responsible for mobility management control and macro BSs responsible for signaling transmission. The user plane consists of LDC acting as gateway/router and SC APs responsible for data transmission. In addition, the user plane functional entities are centrally controlled by the LSC for efficient mobility management.

The Fig. 2 shows the network architecture employing C/U split and centralized network control. In order to apply LMM scheme to this network, the LSC needs to support following functions:

1) Handover decision: Before performing inter-small cell handover, the LSC needs to collect information about mobile terminal and candidate target SCs, including velocity and moving direction of the mobile terminal, and signal strength, backhaul conditions and load status of the candidate SC APs, to determine the target SC AP(s).
2) Configuring data forwarding path: Once the target SC(s) is determined, the LSC needs to establish new data forwarding path(s) by sending configuration information to target SC AP(s) and LDC respectively.
3) Candidate SCs discovery assistance: As the LSC has the knowledge of network deployment topology, it can also be used to assist mobile terminal to perform SC discovery by providing candidate SCs information, e.g. configuring targeted measurements to mobile terminal as introduced in [23].

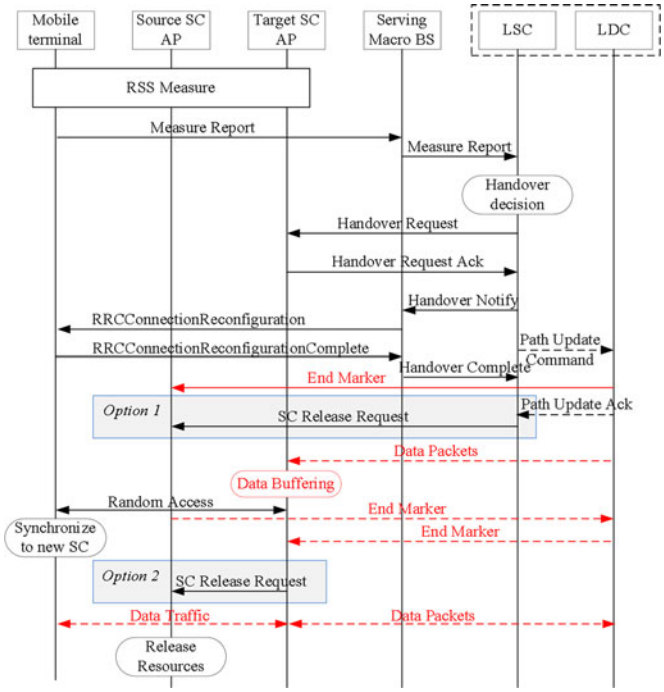

Fig. 3. Flow chart of LMM with centralized control scheme.

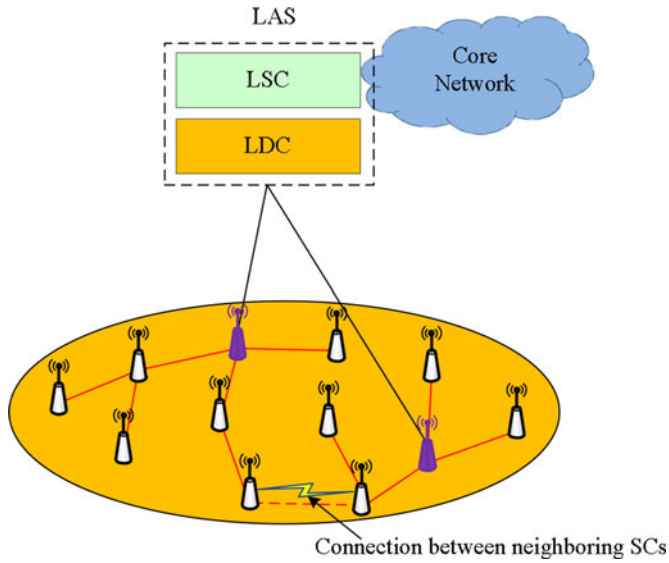

Fig. 4 Logical network architecture with distributed control.

The handover procedure of the proposed LMM scheme is shown in Fig. 3. Before the handover, a mobile terminal established a control plane connection with LSC via macro BS, and user plane connection(s) with LDC via a SC AP. When the mobile terminal moves across neighboring SCs, it measures the signal strength of all candidate SCs, and sends the measure report to the LSC via macro BS. The LSC makes handover decision, e.g. by checking the signal strength of each candidate SC AP contained in the measure report, and selects the target SC AP. It should be noticed that, the criteria of selecting target SC AP should include the backhaul conditions, such as the length of backhaul connection, the backhaul capacity and so on, because these factors have significant impacts on quality of services, e.g. end-to-end latency.

After the target SC AP is determined, the LSC sends "Handover Request" message to the target SC AP for configuring new data forwarding paths which are calculated by the LSC based on the topology of backhaul network. The message includes both radio and route configuration information, so that the target SC AP can reserve the radio resources and establish data forwarding paths for the mobile terminal. As long as the LSC receives the acknowledgement for "Handover Request," it notifies the macro BS that inter-small cell handover is performing, hence the macro BS can instruct the mobile terminal to associate with the target SC AP by sending "RRCConnectionReconfiguration" message. Meanwhile, the LSC sends a "Path Update Command" message containing route configuration to the LDC to update downlink data forwarding path. Before the LDC updates the data forwarding path by applying the new configuration, it has to send an END MARKER to the source SC AP on the old path, which indicates no downlink data on this path any longer. The END MARKER will be forwarded to the target SC AP via LDC if downlink data received by source SC AP needs to be forwarded to the target SC AP.

After RRC (radio resource control) connection reconfiguration procedure is completed, the macro BS sends "Handover Complete" message to confirm with the LSC that the handover is completed. Then the LSC can instruct the source SC AP to release the radio resources and data forwarding path. The source SC AP that received the END MARKER and "SC Release Request" message will release the radio connections established for the mobile terminal after completing the transmission of buffered downlink data. However, if the SC APs are enhanced with support of handover control, as described in the Fig. 3, there is another option to send the "SC Release Request" message, i.e., when the target SC AP establishes connections with the mobile terminal, it sends the message to the source SC AP instead of the LSC.

Similar procedure can also be applied to other target SCs simultaneously with optimized signaling flows, if more than one target SC AP is selected.

B. LMM with Distributed Control

In the case that SC APs support both control plane and user plane functions, the network architecture can be re-drawn as shown in Fig. 4, where the mobility management functions are distributed to SC APs for support of LMM with distributed control scheme. In this network, coordination between neighboring SC APs may be needed for candidate SCs discovery, e.g. each SC AP is configured with neighboring SCs list information, so that mobile terminals served by this SC AP can be provided target measurements.

The Fig. 5 shows the handover procedure of LMM with distributed control scheme. A mobile terminal with active radio connection(s) reports the signal strength of all candidate SCs to the serving SC AP, so that the serving SC AP can decide whether the relocation of SC AP is needed, if the relocation is needed, the source SC AP sends the signaling strengths of all

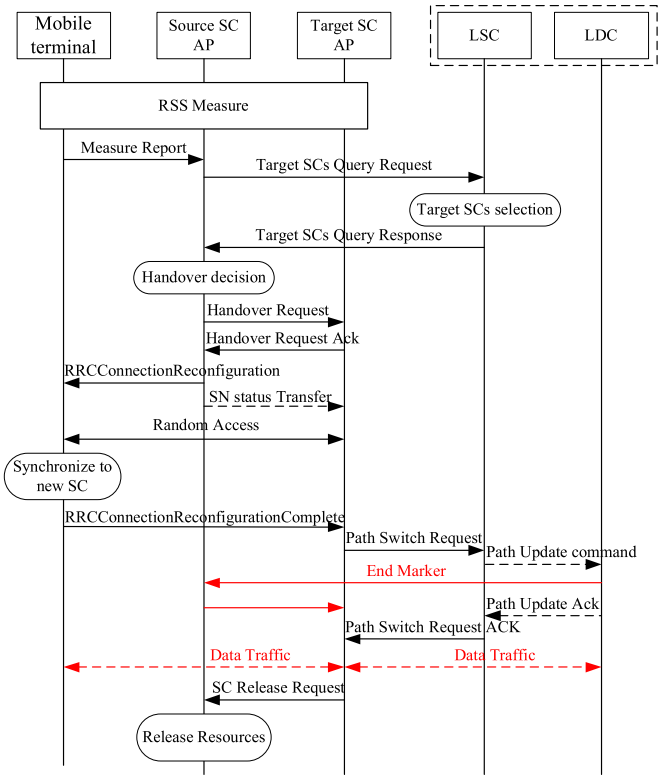

Fig. 5. Flow chart of LMM with distributed control scheme.

candidate SC APs to the LSC for selecting target SC AP(s). The LSC determines the target SC AP(s) by checking the signaling strength, the backhaul conditions, velocity and direction of the mobile terminal etc., and responses to the source SC AP. After making a handover decision based on the response of the LSC, the source SC AP initiates an X2-based handover [24], but the path switch request will be sent to the LSC instead of MME (mobility management entity). As shown in Fig. 4, the LMM scheme requires direct connection between any two neighboring SCs to minimize signaling cost on X2 interface, especially for those connecting to the LAS via different planned SC APs, where even wireless connection between of them may be used.

Thereafter, the LSC sends "Path update command" to LDC to configure new data forwarding path. Before the LDC updates the new path, it also sends an END MARKER on the old path to the source SC AP, which indicates no further downlink data packets will be sent on this path. By receiving both the END MARKER and "SC release request" message, the source SC AP can release the radio connections after completing the transmission of buffered downlink data.

IV. Analytical Model

In order to evaluate the performance of the proposed mobility management schemes, in this section, we take the simulation guidance published by METIS (Mobile and wireless communications Enablers for the Twenty-twenty Information Society) [25] as the reference to develop a model of UDN network deployment. Thereafter, we use Markov Chain to develop an analytical model, and quantify the signaling cost, the packet

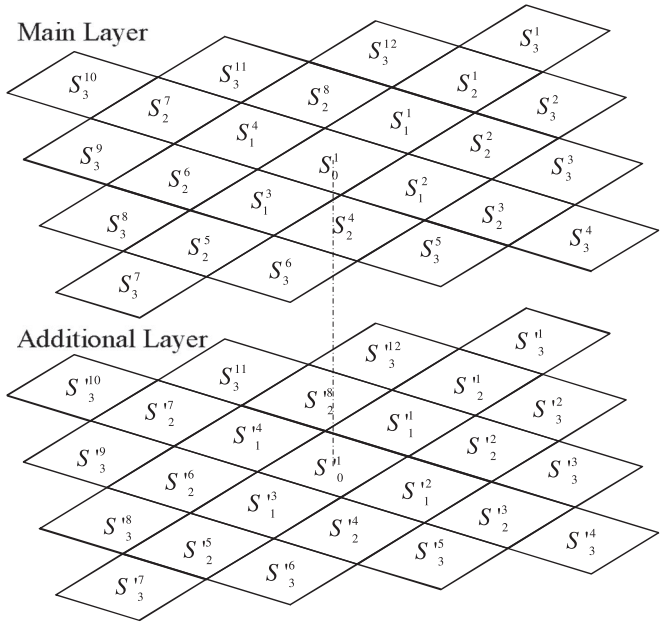

Fig. 6. Three-dimensional grid topology for small cell deployment.

delivery cost and the handover latency of each scheme while the developed deployment model being applied.

A. Model Description

The network deployment in [25] is proposed for the test cases including dense urban information society, virtual reality office and so on, it consists of several layers of grid topology networks. Based on this network deployment, we build a 3D grid network deployment model with following assumptions or differences. First, we assume only the neighboring layers of networks can provide services to mobile terminals, in other words, a mobile terminal can only select three layers of networks at most. Second, all layers of the networks have same grid topology. At last, the network coverage on each layer is partitioned into several areas each of which deploys a planned SC AP, and all partitioned areas have same deployment. With such assumptions, we can model the whole UDN by using two layers of grid networks shown in Fig. 6. According to the figure, on each layer, each block represents a SC AP, the block centrally located in the grid corresponds to a planned SC AP, and the surrounding blocks corresponds to other SC APs which have direct or indirect connections with the planned SC AP. For identifying the location of these surrounding blocks, based on the grid topology, each block is marked by index (i, j), where i denotes ring label of the block and j denotes its position at the ring. As the two layers of networks have overlapped radio coverage but with different signaling strength, a mobile terminal should have different probabilities to access each layer of network. Let's assume that, the mobile terminal has a probability α to select block S_i^j and probability $1-\alpha$ to select block $S_i^{'j}$; the mobile terminal in the block S_i^j can randomly move to any neighboring blocks in same layer with same probability $p = \alpha/4$ and to any neighboring blocks in same layer with same probability $p = (1-\alpha)/4$; the rate of

Algorithm 1: Algorithm for state aggregation.

For $i = 1\ to\ k-1$

$$S_i^1 = \bigcup_{0\le n\le 3} S_i^{n*i+1};$$

For $m = 2\ to\ \lceil (i+1)/2 \rceil$

$$S_i^m = \bigcup_{0\le n\le 3} S_i^{n*i+m} + \bigcup_{1\le n\le 4} S_i^{n*i-m+2}$$

session arrival in each block follows Poisson distribution with a mean λ; the session duration follows an exponential distribution with a mean $1/\mu$; the residence time of the mobile terminal in each overlapped blocks (e.g. S_i^j and $S_i'^j$) follows an exponential distribution with a mean $1/\gamma$; and, the state of mobile terminal is only changed at the end of each time slot τ and only one change is allowed at a time. Then, following the modeling theory introduced in [13]–[26], the traffic and mobility behavior of the mobile terminal can be modeled by a Discrete-Time Markov Chain model.

In the developed model, state space is defined to $S = \{S_{idle}, S_i^j (0 \le i \le K, 1 \le j \le 4i), S_i'^j (0 \le i \le K, 1 \le j \le 4i)\}$. The state S_{idle} represents that a mobile terminal has no data forwarding path towards LAS, i.e. without any active session; the state $S_i^j (0 \le i \le K, 1 \le j \le 4i)$ represents the state at which the mobile terminal locating at the block (i, j) on main layer has the shortest data forwarding path towards LAS, and the state $S_i'^j (0 \le i \le K, 1 \le j \le 4i)$ has similar meaning but for the mobile terminal locating at the block (i, j) on additional layer. The length of the shortest data forwarding path equals $i + H_s$ in terms of hops, where i implies the hops between serving SC AP and the planed SC AP, and H_s indicates the hops between the planed SC AP and the LAS. The maximum number of hops between serving SC AP and the planed SC AP is K, which depends on the range of partitioned area that served by a planned SC AP.

Just as introduced in [27], this model has state-space explosion problem when the size of partitioned area increases. Therefore, based on the grid symmetry, we aggregate the states where the mobile terminal has same behavior. The state aggregation algorithm for the grid topology is shown in Algorithm 1.

Observing the developed analytical model, the following characteristics are identified:

When a mobile terminal at the state S_{idle} has a session arrival, depending on its location, it may enter any states S_i^j. Due to state aggregation, the probability of the mobile terminal entering the state S_i^j or $S_i'^j$ tightly depends on the number of original states that are aggregated to this state, i.e. the probability equals $\frac{N_{(i,j)}}{N} \times P_\lambda$, where $N_{(i,j)}$ is the number of states which are aggregated to S_i^j or $S_i'^j$, N indicates the total number of states of S_i^j or $S_i'^j$, and P_λ is the probability of session arriving during a time slot τ and equals $\lambda\tau$. According to the distribution of session duration, an active session departs with probability P_μ in each time slot τ, where P_μ equals $\mu\tau$, thus the mobile terminal at any state S_i^j or $S_i'^j$ can move into the state S_{idle} with the same probability at the end of the time slot.

Based on the distribution of mobile terminal residence time, the mobile terminal with an active session can move to another neighboring block with probability αP_γ or $(1-\alpha)P_\gamma$ in each time slot, where P_γ equals $\gamma\tau$. In addition, due to overlapped radio coverage, the mobile terminal can also dynamically select the blocks on main layer or additional layer, even it is stationary, thus in each time slot, this mobile terminal is allowed to move from block (i, j) on main layer to corresponding block on additional layer with probability $(1-\alpha)(1-P_\gamma)$, or vice versa. The above mobility behaviors lead to state change of the mobile terminal as shown in the Fig. 7. As all partitioned areas are assumed to have same deployment, while the mobile terminal is moving out of the current area, its state can be modeled as S_{K-1}^j or $S_{K-1}'^j$ if K is even, or S_K^j or $S_K'^j$ if K is odd.

Based on the state transition diagram shown in Fig. 7, we can easily obtain the balance (1) to (8), shown at the top of the next two pages, where the stationary probabilities of the mobile terminal at the state S_{idle} S_i^j and S'^j_i are represented by the variables π_{idle}, $\pi_i^j (0 \le i \le K, 1 \le j \le \lceil (i+1)/2 \rceil)$ and $\pi_i'^j (0 \le i \le K, 1 \le j \le \lceil (i+1)/2 \rceil)$, respectively.

B. Performance Metrics

In order to numerically analyze the performance of proposed schemes, the handover performance metrics need to be determined. Based on the developed Markov Chain model, there are three key performance metrics to be analyzed, i.e. average handover signaling cost for each time slot, average packet delivery cost of a mobile terminal, average handover latency, and average signaling load to the core network in each time slot.

1) Average Handover Signaling Cost: Based on the definition in [28], the signaling cost is the product of the length of signaling message and the weighted distance (hops), i.e. $C = L * H$, where L denotes the average length of signaling messages, and H denotes the hops of data forwarding path on which the signaling message is transmitted.

According to the developed model, average handover signaling cost per each time slot is determined by the states and mobility behaviors of mobile terminals. We first define $C_{fw}(i)$ and $C_{bw}(i)$ to represent the handover signaling cost for a mobile terminal at the state $S_i^j (S'^j_i)$ $(0 \le i < K$ and $1 \le j \le \lceil (i+1)/2 \rceil)$ moving forward to the state S_{i+1}^j or $S_{i+1}^{j+1} (S'^j_{i+1}$ or $S'^{j+1}_{i+1})$, and that for the mobile terminal at the state S_i^j (S'^j_i) $(0 < i \le K$ and $1 \le j \le \lceil (i+1)/2 \rceil)$ moving backward to the state S_{i-1}^j or $S_{i-1}^{j-1} (S'^j_{i-1}$ or $S'^{j-1}_{i-1})$, respectively. Then, considering the mobile terminal can move to a state in different layer of network, we define $C_{fw}^{cha}(i), C_{bw}^{cha}(i)$ and $C_{sam}^{cha}(i)$ to represent the handover signaling cost for the mobile terminal at the state $S_i^j (S'^j_i)$ moving to the state S'^j_{i+1} or $S'^{j+1}_{i+1} (S^j_{i+1}$ or $S^{j+1}_{i+1})$, S'^j_{i-1} or $S'^{j-1}_{i-1} (S^j_{i-1}$ or $S^{j-1}_{i-1})$, and $S'^j_i (S^j_i)$ respectively. After that, we obtain the average handover signaling cost by taking the stationary probability of the mobile terminal at each

$$\pi_{idle} + \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j + \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i'^j = 1 \tag{1}$$

$$\pi_{idle} = (1 - P_\lambda)\pi_{idle} + P_\mu \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j + P_\mu \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i'^j \tag{2}$$

$$\begin{cases}
\pi_0^1 = \alpha \left[\frac{1}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_0^1 + \pi_0'^1) + (1 - P_\mu) P_\gamma \left(\frac{1}{4}\pi_1^1 + \frac{1}{4}\pi_1'^1 \right) \right] \\
\pi_1^1 = \alpha \left[\frac{4}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_1^1 + \pi_1'^1) \right. \\
\qquad \left. + (1 - P_\mu) P_\gamma \left(\pi_0^1 + \frac{1}{4}\pi_2^1 + \frac{1}{2}\pi_2^2 + \pi_0'^1 + \frac{1}{4}\pi_2'^1 + \frac{1}{2}\pi_2'^2 \right) \right] \\
\pi_i^1 = \alpha \Big[\frac{4}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_i^1 + \pi_i'^1) + (1 - P_\mu) P_\gamma \\
\qquad \left(\frac{1}{4}\pi_{i-1}^1 + \frac{1}{4}\pi_{i+1}^1 + \frac{1}{4}\pi_{i+1}^2 + \frac{1}{4}\pi'^1_{i-1} + \frac{1}{4}\pi'^1_{i+1} + \frac{1}{4}\pi'^2_{i+1} \right) \Big] \quad \forall 1 < i < K - 1 \\
\pi_{K-1}^1 = \alpha \left[\frac{4}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_{K-1}^1 + \pi'^1_{K-1}) + (1 - P_\mu) P_\gamma \right. \\
\qquad \left. \left(\frac{1}{4}\pi_{K-2}^1 + \pi_k^1 + \frac{1}{2}\pi_K^2 + \frac{1}{4}\pi'^1_{K-2} + \pi_k'^1 + \frac{1}{2}\pi_K'^2 \right) \right] \\
\pi_K^1 = \alpha \left[\frac{4}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_K^1 + \pi_K'^1) + (1 - P_\mu) P_\gamma \left(\frac{1}{4}\pi_{K-1}^1 + \frac{1}{4}\pi'^1_{K-1} \right) \right]
\end{cases} \tag{3}$$

$$\begin{cases}
\pi_2^2 = \alpha \left[\frac{4}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_2^2 + \pi_2'^2) + (1 - P_\mu) P_\gamma \left(\frac{1}{2}\pi_1^1 + \frac{1}{4}\pi_3^2 + \frac{1}{2}\pi_1'^1 + \frac{1}{4}\pi_3'^2 \right) \right] \\
\pi_3^2 = \alpha \left[\frac{8}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_3^2 + \pi_3'^2) + (1 - P_\mu) P_\gamma \right. \\
\qquad \left. \left(\frac{1}{2}\pi_2^1 + \frac{1}{2}\pi_2^2 + \frac{1}{4}\pi_4^2 + \frac{1}{2}\pi_4^3 + \frac{1}{2}\pi_2'^1 + \frac{1}{2}\pi_2'^2 + \frac{1}{4}\pi_4'^2 + \frac{1}{2}\pi_4'^3 \right) \right] \\
\pi_i^2 = \alpha \Big[\frac{8}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_i^2 + \pi_i'^2) + (1 - P_\mu) P_\gamma \\
\qquad \left(\frac{1}{2}\pi_{i-1}^1 + \frac{1}{4}\pi_{i-1}^2 + \frac{1}{4}\pi_{i+1}^2 + \frac{1}{4}\pi_{i+1}^3 + \frac{1}{2}\pi'^1_{i-1} + \frac{1}{4}\pi'^2_{i-1} + \frac{1}{4}\pi'^2_{i+1} + \frac{1}{4}\pi'^3_{i+1} \right) \Big] \\
\qquad \forall 3 < i < K - 1 \\
\pi_{K-1}^2 = \alpha \left[\frac{8}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_{K-1}^2 + \pi'^2_{K-1}) + (1 - P_\mu) P_\gamma \right. \\
\qquad \left. \left(\frac{1}{2}\pi_{K-2}^1 + \frac{1}{4}\pi_{K-2}^2 + \frac{1}{2}\pi_K^2 + \frac{1}{2}\pi_K^3 + \frac{1}{2}\pi'^1_{K-2} + \frac{1}{4}\pi'^2_{K-2} + \frac{1}{2}\pi_K'^2 + \frac{1}{2}\pi_K'^3 \right) \right] \\
\pi_K^2 = \alpha \left[\frac{8}{N} P_\lambda \pi_{idle} + (1 - P_\mu)(1 - P_\gamma)(\pi_K^2 + \pi_K'^2) + (1 - P_\mu) P_\gamma \left(\frac{1}{2}\pi_{K-1}^1 + \frac{1}{4}\pi_{K-1}^2 + \frac{1}{2}\pi'^1_{K-1} + \frac{1}{4}\pi'^2_{K-1} \right) \right]
\end{cases} \tag{4}$$

state into account, and derive (9), shown at the top of the page after the next page.

2) Average Packet Delivery Cost: The average packet delivery cost means the average cost for transmitting a PDU (protocol data unit). It is also described by the product of the length of the PDU (protocol data unit) and the weighted distance (hops), i.e., $U = L * H$, where L and H indicates average length of PDUs and the hops of data forwarding path for transmitting a

$$\begin{cases} \pi_{2i}^{i+1} = \alpha\left[\frac{4}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{2i}^{i+1} + \pi'^{i+1}_{2i}) + (1-P_\mu)P_\gamma\left(\frac{1}{4}\pi_{2i-1}^{i} + \frac{1}{4}\pi_{2i+1}^{i+1} + \frac{1}{4}\pi'^{i}_{2i-1} + \frac{1}{4}\pi'^{i+1}_{2i+1}\right)\right] \\ \qquad \forall 1 < i < \frac{K-1}{2} \\ \pi_{2i+1}^{i+1} = \alpha\left[\frac{8}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{2i+1}^{i+1} + \pi'^{i+1}_{2i+1}) + (1-P_\mu)P_\gamma\right. \\ \qquad \left.\left(\frac{1}{4}\pi_{2i}^{i} + \frac{1}{2}\pi_{2i}^{i+1} + \frac{1}{4}\pi_{2i+2}^{i+1} + \frac{1}{2}\pi_{2i+2}^{i+2} + \frac{1}{4}\pi'^{i}_{2i} + \frac{1}{2}\pi'^{i+1}_{2i} + \frac{1}{4}\pi'^{i+1}_{2i+2} + \frac{1}{2}\pi'^{i+2}_{2i+2}\right)\right] \forall 1 < i < \frac{K-2}{2} \end{cases} \tag{5}$$

$$\begin{cases} \pi_{i+2}^{j} = \alpha\left[\frac{8}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{i+2}^{j} + \pi'^{j}_{i+2}) + (1-P_\mu)P_\gamma\right. \\ \qquad \left.\left(\frac{1}{4}\pi_{i+1}^{j-1} + \frac{1}{4}\pi_{i+1}^{j} + \frac{1}{4}\pi_{i+3}^{j} + \frac{1}{4}\pi_{i+3}^{j+1} + \frac{1}{4}\pi'^{j-1}_{i+1} + \frac{1}{4}\pi'^{j}_{i+1} + \frac{1}{4}\pi'^{j}_{i+3} + \frac{1}{4}\pi'^{j+1}_{i+3}\right)\right] \\ \qquad \forall 3 < i < K-3 \text{ and } 2 < j < \left\lceil\frac{i+1}{2}\right\rceil \\ \pi_{K-1}^{j} = \alpha\left[\frac{8}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{K-1}^{j} + \pi'^{j}_{K-1}) + (1-P_\mu)P_\gamma\right. \\ \qquad \left.\left(\frac{1}{4}\pi_{K-2}^{j-1} + \frac{1}{4}\pi_{K-2}^{j} + \frac{1}{2}\pi_{K}^{j} + \frac{1}{2}\pi_{K}^{j+1} + \frac{1}{4}\pi'^{j-1}_{K-2} + \frac{1}{4}\pi'^{j}_{K-2} + \frac{1}{2}\pi'^{j}_{K} + \frac{1}{2}\pi'^{j+1}_{K}\right)\right] \\ \qquad \forall 2 < j < \left\lceil\frac{i+1}{2}\right\rceil \\ \pi_{K}^{j} = \alpha\left[\frac{8}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{K}^{j} + \pi'^{j}_{K}) + (1-P_\mu)P_\gamma\left(\frac{1}{4}\pi_{K-1}^{j-1} + \frac{1}{4}\pi_{K-1}^{j} + \frac{1}{4}\pi'^{j-1}_{K-1} + \frac{1}{4}\pi'^{j}_{K-1}\right)\right] \\ \qquad \forall 2 < j < \left\lceil\frac{i+1}{2}\right\rceil \end{cases} \tag{6}$$

$$\begin{cases} \text{If K is even:} \\ \pi_{K-1}^{\frac{K}{2}} = \alpha\left[\frac{8}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{K-1}^{\frac{K}{2}} + \pi'^{\frac{K}{2}}_{K-1}) + (1-P_\mu)P_\gamma\right. \\ \qquad \left.\left(\frac{1}{4}\pi_{K-2}^{\frac{K}{2}-1} + \frac{1}{2}\pi_{K-2}^{\frac{K}{2}} + \frac{1}{2}\pi_{K}^{\frac{K}{2}} + \pi_{K}^{\frac{K}{2}+1} + \frac{1}{4}\pi'^{\frac{K}{2}-1}_{K-2} + \frac{1}{2}\pi'^{\frac{K}{2}}_{K-2} + \frac{1}{2}\pi'^{\frac{K}{2}}_{K} + \pi'^{\frac{K}{2}+1}_{K}\right)\right] \\ \pi_{K}^{\frac{K}{2}+1} = \alpha\left[\frac{4}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{K}^{\frac{K}{2}+1} + \pi'^{\frac{K}{2}+1}_{K}) + (1-P_\mu)P_\gamma\left(\frac{1}{4}\pi_{K-1}^{\frac{K}{2}} + \frac{1}{4}\pi'^{\frac{K}{2}}_{K-1}\right)\right] \\ \text{If K is odd:} \end{cases} \tag{7}$$

$$\begin{cases} \pi_{K-1}^{\frac{K+1}{2}} = \alpha\left[\frac{4}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{K-1}^{\frac{K+1}{2}} + \pi'^{\frac{K+1}{2}}_{K-1}) + (1-P_\mu)P_\gamma\right. \\ \qquad \left.\left(\frac{1}{4}\pi_{K-1}^{\frac{K-1}{2}} + \frac{1}{2}\pi_{K}^{\frac{K+1}{2}} + \frac{1}{4}\pi'^{\frac{K-1}{2}}_{K-1} + \frac{1}{2}\pi'^{\frac{K+1}{2}}_{K}\right)\right] \\ \pi_{K}^{\frac{K+1}{2}} = \alpha\left[\frac{8}{N}P_\lambda\pi_{idle} + (1-P_\mu)(1-P_\gamma)(\pi_{K}^{\frac{K+1}{2}} + \pi'^{\frac{K+1}{2}}_{K}) + (1-P_\mu)P_\gamma\right. \\ \qquad \left.\left(\frac{1}{4}\pi_{K-1}^{\frac{K-1}{2}} + \frac{1}{4}\pi_{K-1}^{\frac{K+1}{2}} + \frac{1}{4}\pi'^{\frac{K-1}{2}}_{K-1} + \frac{1}{4}\pi'^{\frac{K+1}{2}}_{K-1}\right)\right] \end{cases} \tag{8}$$

$$\pi_i'^{j} = (1-\alpha)\pi_i^j/\alpha$$

$$
\begin{aligned}
C_{avg} &= \pi_0^1[\alpha(1-P_\mu)P_\gamma C_{fw}(0) + (1-\alpha)(1-P_\mu)P_\gamma C_{fw}^{cha}(0) + (1-\alpha)(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(0)] \\
&+ \sum_{i=1}^{K-1} \pi_i^1 \left[\alpha(1-P_\mu)P_\gamma \left(\frac{3}{4}C_{fw}(i) + \frac{1}{4}C_{bw}(i)\right) + (1-\alpha)(1-P_\mu)P_\gamma \left(\frac{3}{4}C_{fw}^{cha}(i) + \frac{1}{4}C_{bw}^{cha}(i)\right)\right. \\
&\left. + (1-\alpha)(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(i)\right] \\
&+ \sum_{i=2}^{K-1} \sum_{j=2}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j \left[\alpha(1-P_\mu)P_\gamma \left(\frac{1}{2}C_{fw}(i) + \frac{1}{2}C_{bw}(i)\right) + (1-\alpha)(1-P_\mu)P_\gamma \left(\frac{1}{2}C_{fw}^{cha}(i) + \frac{1}{2}C_{bw}^{cha}(i)\right)\right. \\
&\left. + (1-\alpha)(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(i)\right] \\
&+ \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \pi_K^j \left[(1-P_\mu)P_\gamma(\alpha C_{bw}(K) + (1-\alpha)C_{bw}^{cha}(K)) + (1-\alpha)(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(K)\right] \\
&+ \pi_0'^1 \left[(1-P_\mu)P_\gamma(\alpha C_{fw}^{cha}(0) + (1-\alpha)C_{fw}(0)) + \alpha(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(0)\right] \\
&+ \sum_{i=1}^{K-1} \pi_i'^1 \left[(1-\alpha)(1-P_\mu)P_\gamma \left(\frac{3}{4}C_{fw}(i) + \frac{1}{4}C_{bw}(i)\right) + \alpha(1-P_\mu)P_\gamma \left(\frac{3}{4}C_{fw}^{cha}(i) + \frac{1}{4}C_{bw}^{cha}(i)\right) + \alpha(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(i)\right] \\
&+ \sum_{i=2}^{K-1} \sum_{j=2}^{\lceil \frac{i+1}{2} \rceil} \pi_i'^j \left[(1-\alpha)(1-P_\mu)P_\gamma \left(\frac{1}{2}C_{fw}(i) + \frac{1}{2}C_{bw}(i)\right) + \alpha(1-P_\mu)P_\gamma \left(\frac{1}{2}C_{fw}^{cha}(i)\frac{1}{2}C_{bw}^{cha}(i)\right)\right. \\
&\left. + \alpha(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(i)\right] \\
&+ \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \pi_K'^j [(1-\alpha)(1-P_\mu)P_\gamma C_{bw}(K) + \alpha(1-P_\mu)P_\gamma C_{bw}^{cha}(K) + \alpha(1-P_\mu)(1-P_\gamma)C_{sam}^{cha}(K)] \qquad (9)
\end{aligned}
$$

PDU respectively. According to the developed model, the hops of data forwarding path is determined by the states of mobile terminals.Thus we use $U(i)$ to represent the packet delivery cost for a mobile terminal at the state S_i^j or S'^j_i ($0 \le i \le K$ and $1 \le j \le \lceil (i+1)/2 \rceil$).

Considering the stationary probability of the mobile terminal at each state, the average packet delivery cost under the analytical model is expressed by

$$U_{avg} = \frac{1}{1-\pi_{idle}} \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} (\pi_i^j + \pi_i'^j) U(i). \qquad (10)$$

3) Average Handover Latency: As defined in [29], Handover latency is calculated as the duration between the time when the measurement procedure is initiated and the time when the handover procedure is completed. For estimating the handover latency of each LMM scheme, we divide the handover procedure into a set of signaling flows, so that the handover latency can be calculated by the aggregated time delay of processing each signaling flows.

According to [30], the time delay for transmitting a signaling message depends on the length of data forwarding path, thus mobility behavior and state of mobile terminals need to be considered for determining the handover latency. Let $D_{fw}(i)$ and $D_{bw}(i)$ indicate the handover latency for a mobile terminal at the state S_i^j or S'^j_i ($0 \le i < K$ and $1 \le j \le \lceil (i+1)/2 \rceil$) moves forward to the state S_{i+1}^j or S_{i+1}^{j+1}(S'^j_{i+1} or S'^{j+1}_{i+1}), and that for the mobile terminal at the state S_i^j or S'^j_i ($0 < i \le K$, and $1 \le j \le \lceil (i+1)/2 \rceil$) moves backward to the state S_{i-1}^j or S_{i-1}^{j-1}(S'^j_{i-1} or S'^{j-1}_{i-1}), respectively, and $D_{fw}^{cha}(i)$, $D_{bw}^{cha}(i)$ and $D_{sam}^{cha}(i)$ respectively represent the handover latencies for the mobile terminal at the state $S_i^j(S'^j_i)$ moving to the state S'^j_{i+1} or S'^{j+1}_{i+1}(S_{i+1}^j or S_{i+1}^{j+1}), S'^j_{i-1} or S'^{j-1}_{i-1}(S_{i-1}^j or S_{i-1}^{j-1}), and S'^j_i(S_i^j). Then the average handover latency per each handover can be expressed by (11) and (12), shown at the top of the page after the next page.

4) Average Signaling Load to the Core Network: Based on the introduction of the proposed LMM schemes, when a mobile terminal moves within an area served by a LAS, there is no handover signaling to the core network. However, in order to use the signaling cost of LMM schemes as the baseline to obtain scalarization signal cost, it is assumed that, a LAS can only serve two groups of SC APs which locates on main layer and additional layer respectively, and there is only one planned SC AP in each group of SC APs, with this assumption, as long

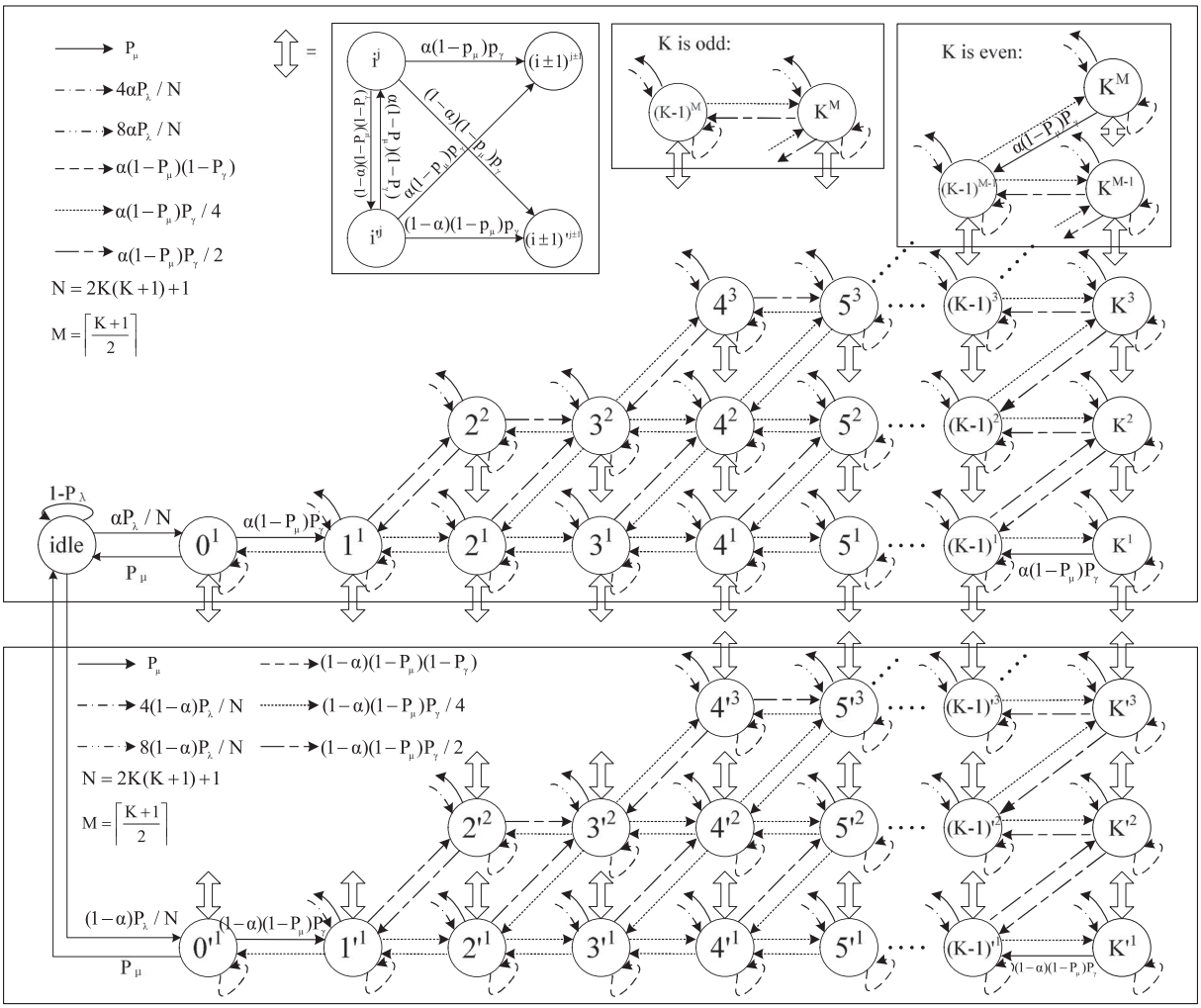

Fig. 7. State transition in the analytical model.

as the mobile terminal moves out of the coverage of these SC APs, a signaling procedure to the core network will be triggered to update the serving LAS. Let the signaling cost for a signaling procedure to the core network is C^{core}, we can derive the expression of average signaling load to the core network for each time slot (see (11) and (12)).

V. Performance Evaluation

For comparing the proposed LMM schemes with 3GPP scheme, we assume mobility management scheme designed for LTE network is applied to UDN, and choose it as the baseline scheme for analysis.

By deriving the expressions of average handover signaling cost, average packet delivery cost and average handover latency of the proposed LMM schemes and the baseline scheme respectively, we conduct numerical and simulation analysis for performance comparison.

A. Baseline Mechanism

For supporting dual connectivity, 3GPP introduced a new mobility management scheme for LTE network with SC deployment [25], where the handover of SCs can be handled either by a macro eNB only, or by both macro eNB and MME. However, due to flexible and dense deployment of SC APs in UDN, it is not feasible to connect all SC APs to the eNB, hence we choose later one as the baseline.

In the mobility management scheme for LTE network supporting SCs, there is no centralized backhaul topology management for SC APs, thus the length of data forwarding path between each SC AP and the planned SC AP can't be modeled based on network deployment even the SCs are deployed to grid topology. In order to model such network, we can only assume that the network has the optimal backhaul topology, i.e. macro eNBs act as planed SC APs, and the data forwarding path between each SC AP and the planned SC AP is the shortest path. With such assumption, the analytical model developed in Section IV can be applied.

B. Cost Functions of Performance Metrics

1) Handover Signaling Cost: Based on the flow chart of LMM with centralized control scheme in the Fig. 3, the signaling

$$D_{avg} = \frac{1}{1-\pi_{idle}} * \frac{1}{(1-\alpha)(1-P_\gamma)+P_\gamma} * \{\pi_0^1[\alpha P_\gamma D_{fw}(0) + (1-\alpha)P_\gamma D_{fw}^{cha}(0) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(0)]$$
$$+ \sum_{i=1}^{K-1} \pi_i^1 \left[\alpha P_\gamma \left(\frac{3}{4}D_{fw}(i) + \frac{1}{4}D_{bw}(i)\right) + (1-\alpha)P_\gamma \left(\frac{3}{4}D_{fw}^{cha}(i) + \frac{1}{4}D_{bw}^{cha}(i)\right) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(i)\right]$$
$$+ \sum_{i=2}^{K-1} \sum_{j=2}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j \left[\alpha P_\gamma \left(\frac{1}{2}D_{fw}(i) + \frac{1}{2}D_{bw}(i)\right) + (1-\alpha)P_\gamma \left(\frac{1}{2}D_{fw}^{cha}(i) + \frac{1}{2}D_{bw}^{cha}(i)\right) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(i)\right]$$
$$+ \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \pi_K^j \left[\alpha P_\gamma D_{bw}(K) + (1-\alpha)P_\gamma D_{bw}^{cha}(K) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(K)\right]$$
$$+ \pi_0'^1[(1-\alpha)P_\gamma D_{fw}(0) + \alpha P_\gamma D_{fw}^{cha}(0) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(0)]$$
$$+ \sum_{i=1}^{K-1} \pi_i'^1 \left[(1-\alpha)P_\gamma \left(\frac{3}{4}D_{fw}(i) + \frac{1}{4}D_{bw}(i)\right) + \alpha P_\gamma \left(\frac{3}{4}D_{fw}^{cha}(i) + \frac{1}{4}D_{bw}^{cha}(i)\right) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(i)\right]$$
$$+ \sum_{i=2}^{K-1} \sum_{j=2}^{\lceil \frac{i+1}{2} \rceil} \pi_i'^j \left[(1-\alpha)P_\gamma \left(\frac{1}{2}D_{fw}(i) + \frac{1}{2}D_{bw}(i)\right) + \alpha P_\gamma \left(\frac{1}{2}D_{fw}^{cha}(i) + \frac{1}{2}D_{bw}^{cha}(i)\right) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(i)\right]$$
$$+ \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \pi_K'^j [(1-\alpha)P_\gamma D_{bw}(K) + \alpha P_\gamma D_{bw}^{cha}(K) + (1-\alpha)(1-P_\gamma)D_{sam}^{cha}(K)]\} \quad (11)$$

$$C_{avg}^{core} = [\frac{3}{4}(\pi_K^1 + \pi_K'^1) + \frac{1}{2} \sum_{j=2}^{\lceil \frac{K+1}{2} \rceil} (\pi_K^j + \pi_K'^j)](1-P_\mu)P_\gamma C^{core} \quad (12)$$

cost for a handover procedure can be expressed by

$$C^{CenC} = 3C_{mc} + 2C_{tsc} + C_{ssc} + 2C_l + 3C_{air} \quad (13)$$

where C_{mc} means the signaling cost for transferring a message between macro BS and LSC, C_{tsc} represents the signaling cost for transferring a message between target SC AP and LSC, C_{ssc} represents the signaling cost for transferring a message between source SC AP and LSC, C_l represents the signaling cost for transferring a message between LSC and LDC, and the C_{air} represents the signaling cost for transferring a message over the air interface.

By assuming the message size to unit length, only the hops of data forwarding path needs to be considered to determine C^{CenC}. We further assume that the LSC and the LDC are collocated in LAS, then the cost C_l becomes a constant value.

Based on the analytical model and the (13), we derive functions of the signaling cost C_{fw}^{CenC}, C_{bw}^{CenC},$C_{fw}^{CenCcha}$, $C_{bw}^{CenCcha}$ and $C_{sam}^{CenCcha}$:

$$C_{fw}^{CenC}(i) = 3H_m + 2(H_s + i + 1) + (H_s + i) + 2 + 3C_{air} \quad (14)$$

$$C_{bw}^{CenC}(i) = 3H_m + 2(H_s + i - 1) + (H_s + i) + 2 + 3C_{air} \quad (15)$$

$$C_{fw}^{CenCcha}(i) = C_{fw}^{CenC}(i), \quad (16)$$

$$C_{bw}^{CenCcha}(i) = C_{bw}^{CenC}(i) \quad (17)$$

$$C_{sam}^{CenCcha}(i) = 3H_m + 2(H_s + i) + (H_s + i) + 2 + 3C_{air} \quad (18)$$

where H_m is the hops of data forwarding path between macro BS and LSC, and H_s is the hops of data forwarding path between SC AP and LSC.

However, if SC APs are enhanced with support of handover control, then the signaling cost for a handover procedure becomes

$$C^{CenC2} = 3C_{mc} + 2C_{tsc} + C_{bsc} + 2C_l + 3C_{air} \quad (19)$$

where C_{bsc} represents the signaling cost for transferring a message between two neighboring SC APs in same or different layer of network. When a message is transferred from a SC AP on main layer to a SC AP on additional layer, it is assumed that the message needs to be routed via the planned SC APs on both layers. Considering the mobility behaviors of the mobile terminal, the cost functions can be expressed by

$$C_{fw}^{CenC2}(i) = 3H_m + 2(H_s + i + 1) + 3 + 3C_{air} \quad (20)$$

$$C_{bw}^{CenC2}(i) = 3H_m + 2(H_s + i - 1) + 3 + 3C_{air} \quad (21)$$

$$C_{fw}^{CenC2cha}(i) = 3H_m + 2(H_s + i + 1) + 2i + 2 + 2 + 3C_{air} \quad (22)$$

$$C_{bw}^{CenC2cha}(i) = 3H_m + 2(H_s + i - 1) + 2i + 2 + 3C_{air} \tag{23}$$

$$C_{sam}^{CenC2cha}(i) = 3H_m + 2(H_s + i) + 2i + 1 + 2 + 3C_{air}. \tag{24}$$

Finally, by submitting above equations into expression (9), the average handover signaling cost C_{avg}^{CenC} or C_{avg}^{CenC2} can be obtained.

Similarly, we can derive the handover signaling cost for LMM with distributed control scheme according to the flow chart shown in Fig. 5:

$$C^{DisC} = 2C_{ssc} + 3C_{bsc} + 2C_{tsc} + 2C_l + 3C_{air}. \tag{25}$$

Considering the mobility behaviors and states of the mobile terminal, following cost functions can be determined:

$$C_{fw}^{DisC}(i) = 2(H_s + i) + 3 + 2(H_s + i + 1) + 2 + 3C_{air} \tag{26}$$

$$C_{bw}^{DisC}(i) = 2(H_s + i) + 3 + 2(H_s + i - 1) + 2 + 3C_{air} \tag{27}$$

$$C_{fw}^{DisCcha}(i) = 2(H_s + i) + 3(2i + 2) + 2(H_s + i + 1) + 2 + 3C_{air} \tag{28}$$

$$C_{bw}^{DisCcha}(i) = 2(H_s + i) + 6i + 2(H_s + i - 1) + 2 + 3C_{air} \tag{29}$$

$$C_{sam}^{DisCcha}(i) = 2(H_s + i) + 3(2i + 1) + 2(H_s + i) + 2 + 3C_{air}. \tag{30}$$

Thus we can get the average handover signaling cost of LMM with distributed control scheme by submitting (26)–(30) into (9).

Finally, we derive the cost functions for 3GPP scheme:

$$C_{fw}^{3GPP}(i) = 3(i + 1) + 2i + 2H_e + 2 + 3C_{air} \tag{31}$$

$$C_{bw}^{3GPP}(i) = 3(i - 1) + 2i + 2H_e + 2 + 3C_{air} \tag{32}$$

$$C_{fw}^{3GPPcha}(i) = C_{fw}^{3GPP}(i) \tag{33}$$

$$C_{bw}^{3GPPcha}(i) = C_{bw}^{3GPP}(i) \tag{34}$$

$$C_{sam}^{3GPPcha}(i) = 3i + 2i + 2H_e + 2 + 3C_{air} \tag{35}$$

where H_e indicates the hops of signaling path between eNB and MME. Finally, the average handover signaling cost of the 3GPP scheme is also obtained.

2) Packet Delivery Cost: As the proposed LMM scheme have no impacts on air interface and the cost of data transmission over radio channel is not modeled, we only consider the packet delivery cost from the network to the SC AP serving mobile terminals.

Let us assume the payload size of each PDU to be a unit length, the packet delivery cost for a mobile terminal located at the ring i of the grid can be expressed by $U(i) = H_g + i$, where H_g equals the length of data forwarding path between a user plane anchor, e.g. PGW, and a planned SC AP. Thus the average packet delivery cost given in expression (10) becomes

$$U_{avg} = \frac{1}{1 - \pi_{idle}} \left(\sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j (H_g + i) + \sum_{i=1}^{K} \sum_{j=1}^{\lceil \frac{i}{2} \rceil} \pi_i'^j (H_g + i) \right). \tag{36}$$

Since user plane packet delivery cost depends on backhaul topology and is irrelevant with the control plane signaling flows, both proposed LMM schemes have same average packet delivery cost. However, the average packet delivery cost of 3GPP scheme is more than U_{avg} due to lack of centralized backhaul topology management.

3) Average Handover Latency: As introduced in [29], the handover latency can be divided into the time delay for individual signaling flows. Therefore, the handover latencies of the proposed LMM schemes and LTE mobility management scheme can be described as

$$D^{CenC} = D_{air} + D_{mc} + D^{proc} + 2D_{tsc} + D_{mc} + 2D_{air} + \max\{D^{acs}, D_{mc} + D_{ssc}\} \tag{37}$$

where D_{air} indicates the time for transferring a signaling message between mobile terminal and SC AP over the air interface, D_{mc} indicates the time for message transmission between macro BS and LSC, D^{proc} indicates the time for network entity processing received measure report, D_{ssc} and D_{tsc} indicate the time for message transmission between source SC AP and LSC, and the time for message transmission between target SC AP and LSC, respectively, and D^{acs} is the time for random access on the air interface

$$D^{DisC} = D_{air} + D^{proc} + 2D_{ssc} + D^{proc} + 2D_{bsc} + D_{air} + D^{acs} + D_{air} + 2D_{tsc} + 2D_{ldc} + D_{bsc} \tag{38}$$

where D_{bsc} indicates the time for transferring a signaling message between two neighboring SC APs in same or different layer of network, and D_{ldc} is the time for message transmission between LSC and LDC.

$$D^{3GPP} = D_{air} + D^{proc} + 2D_{mc-tsc} + D_{mc-ssc} + 2D_{air} + \max\{D^{acs}, D_{mc-tsc} + 2D_{mme} + 2D_{sgw} + D_{mc-ssc}\} \tag{39}$$

where D_{mc-ssc} and D_{mc-tsc} are the time for message transmission between macro eNB and source SC AP, and the time for message transmission between macro eNB and target SC AP, D_{mme} indicates the time for message transmission between macro eNB and MME, and D_{sgw} indicates the time for message transmission between MME and Serving GW.

For LMM with centralized control, if SC APs are enhanced with support of handover control, then the handover latency expressed in the (23) needs to be revised to

$$D^{CenC2} = D_{air} + D_{mc} + D^{proc} + 2D_{tsc} + D_{mc} + 2D_{air} + \max\{D^{acs}, D_{mc}\} + D_{bsc}. \tag{40}$$

In order to obtain exact handover latencies, we need following notations:

t_{radio} indicates the time required for one hop packet transmission through radio link;

t_{wired} indicates the time required for one hop packet transmission through wired link.

Then, the latencies introduced in Section IV can be determined:

For LMM with centralized control scheme

$$D_{fw}^{CenC}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i + 1)t_{wired} + D^{proc} + \max\{D^{acs}, (H_m + H_s + i)t_{wired}\} \quad (41)$$

$$D_{bw}^{CenC}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i - 1)t_{wired} + D^{proc} + \max\{D^{acs}, (H_m + H_s + i)t_{wired}\} \quad (42)$$

$$D_{fw}^{CenCcha}(i) = D_{fw}^{CenC}(i) \quad (43)$$

$$D_{bw}^{CenCcha}(i) = D_{bw}^{CenC}(i) \quad (44)$$

$$D_{sam}^{CenCcha}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i)t_{wired} + D^{proc} + \max\{D^{acs}, (H_m + H_s + i)t_{wired}\}. \quad (45)$$

If SC APs in this scheme support handover control:

$$D_{fw}^{CenC2}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i + 1)t_{wired} + D^{proc} + \max\{D^{acs}, H_m t_{wired}\} + t_{wired} \quad (46)$$

$$D_{bw}^{CenC2}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i - 1)t_{wired} + D^{proc} + \max\{D^{acs}, H_m t_{wired}\} + t_{wired} \quad (47)$$

$$D_{fw}^{CenC2cha}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i + 1)t_{wired} + D^{proc} + \max\{D^{acs}, H_m t_{wired}\} + (2i + 2)t_{wired} \quad (48)$$

$$D_{bw}^{CenC2cha}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i - 1)t_{wired} + D^{proc} + \max\{D^{acs}, H_m t_{wired}\} + 2it_{wired} \quad (49)$$

$$D_{sam}^{CenC2cha}(i) = 3t_{radio} + 2H_m t_{wired} + 2(H_s + i)t_{wired} + D^{proc} + \max\{D^{acs}, H_m t_{wired}\} + (2i + 1)t_{wired}. \quad (50)$$

For LMM with distributed control scheme

$$D_{fw}^{DisC}(i) = 3t_{radio} + 2(H_s + i)t_{wired} + 3t_{wired} + 2(H_s + i + 1)t_{wired} + 2t_{wired} + 2D^{proc} + D^{acs} \quad (51)$$

$$D_{bw}^{DisC}(i) = 3t_{radio} + 2(H_s + i - 1)t_{wired} + 3t_{wired} + 2(H_s + i)t_{wired} + 2t_{wired} + 2D^{proc} + D^{acs} \quad (52)$$

$$D_{fw}^{DisCcha}(i) = 3t_{radio} + 2(H_s + i)t_{wired} + 3(2i+2)t_{wired} + 2(H_s + i + 1)t_{wired} + 2t_{wired} + 2D^{proc} + D^{acs} \quad (53)$$

$$D_{bw}^{DisCcha}(i) = 3t_{radio} + 2(H_s + i - 1)t_{wired} + 6it_{wired} + 2(H_s + i)t_{wired} + 2t_{wired} + 2D^{proc} + D^{acs} \quad (54)$$

$$D_{sam}^{DisCcha}(i) = 3t_{radio} + 2(H_s + i)t_{wired} + 3(2i + 1)t_{wired} + 2(H_s + i)t_{wired} + 2t_{wired} + 2D^{proc} + D^{acs}. \quad (55)$$

For 3GPP mobility management scheme

$$D_{fw}^{3GPP}(i) = 3t_{radio} + D^{proc} + 2(i + 1)t_{wired} + it_{wired} + \max\{D^{acs}, (i + 1)t_{wired} + (2H_g + 2 + i)t_{wired}\} \quad (56)$$

$$D_{bw}^{3GPP}(i) = 3t_{radio} + D^{proc} + 2(i - 1)t_{wired} + it_{wired} + \max\{D^{acs}, (i - 1)t_{wired} + (2H_g + 2 + i)t_{wired}\} \quad (57)$$

$$D_{fw}^{3GPPcha}(i) = D_{fw}^{3GPP}(i) \quad (58)$$

$$D_{bw}^{3GPPcha}(i) = D_{bw}^{3GPP}(i) \quad (59)$$

$$D_{sam}^{3GPPcha}(i) = 3t_{radio} + D^{proc} + 2it_{wired} + it_{wired} + \max\{D^{acs}, it_{wired} + (2H_g + 2 + i)t_{wired}\}. \quad (60)$$

By submitting expressions (41)–(45), (46)–(50), (51)–(55), and (56)–(60) into (11) separately, the average handover latency for the proposed LMM schemes and 3GPP scheme can be obtained.

4) Average Signaling Load to the Core Network: Based on the introduction in the previous section, for all LMM schemes proposed in this paper, the same average signaling load to the core network is shown in the (12). For 3GPP scheme, the average signaling load to the core network for each time slot is shown in (61), shown at the bottom of the page.

In order to compare the signaling cost of the proposed LMM schemes with those of other local anchor based mobility management schemes, e.g., local anchor schemes proposed in [13], we further calculate the signaling to the core network generated by those schemes. It is still assumed that, with those local anchor schemes, the signaling cost for a signaling procedure to the core network is C^{core}, and then, the average signaling load caused by those schemes in each time slot is shown in (62), shown at the bottom of the page.

C. Performance Analysis

We first numerically analyze the performance metrics of proposed LMM schemes and 3GPP scheme based on developed mathematical model, then we setup simulation environment to validate the analysis.

MATLAB is used for discrete-event simulations. We assume the 3D grid network deployment shown in Fig. 6 is used in the simulation as well, where a mobile terminal residing in a block can re-select to corresponding block in different layer of network with probability α or $1 - \alpha$, or move to neighboring blocks in same or different layer of network with different probabilities, at the end of a time slot. If the mobile terminal moves to any of neighboring blocks in main layer with probability $\alpha/4$, it moves to any of neighboring blocks in additional layer with probability $(1 - \alpha)/4$. However, if the mobile terminal locates at the Kth ring, where K is the maximum ring label, it can only move

backward. The mobile terminal state is assumed to be changed after each time slot, and 5 million time slots are designated to each simulation.

Based on Poisson distribution with mean λ, we use MATLAB to identify the time slots that have sessions arriving. We further generate each session duration based on exponential distribution with mean $1/\mu$, then we obtain start time and end time of each session, which can be used to determine when the mobile terminal has active session(s). Residence time of the mobile terminal staying in each SC is also generated by MATLAB based on exponential distribution with mean $1/\gamma$. When the mobile terminal has active session, i.e., before the session duration time expires, and changes its serving SC AP, a handover procedure is performed.

The values of parameters used in the simulations and evaluations are given in the Table II.

TABLE II
PARAMETERS USED IN THE EVALUATION

Parameter	Description	Value
λ	Session arrival rate	0.001/sec
μ	Session duration ratio	0.01/sec
γ	cell residence ratio	0.1/sec
K	Range of the area that a planned SC AP serves	1, 2, 3, 4, 5, 6
α	Probability of a mobile terminal select blocks on the main layer	1, 0.95, 0.9, 0.85, 0.8, 0.75
τ	Time slot	0.01sec
H_m	Distance between macro BS and LAS	1 hop
H_s	Distance between SC AP and LAS	2 hops
H_e	Distance between eNB and MME	10 hops
H_g	Distance between planned SC AP and GW in core network	10 hops
T_{radio}	Average time for one hop packet transmission through radio link	5 ms
T_{wired}	Average time for one hop packet transmission through wired link	1 ms
D^{proc}	Average time for making handover decision or selecting target SC AP	5 ms
D^{acs}	Average time for air interface random access	10 ms

1) Average Handover Signaling Cost: In order to analyze the average handover signaling cost of the proposed schemes, and compare them with that of 3GPP scheme, we use ratio of average signaling cost and normalized signaling cost to characterize the performance. The ratio of signaling cost in Fig. 8(a) or (b) is defined as the average signaling cost of each proposed handover scheme to that of 3GPP handover scheme under the same network conditions; the normalized signaling cost is the signaling cost of each scheme to the maximum signaling cost among them. As introduced in Section IV, the signaling cost is in terms of signaling message length and the weighted distance (hops), thus the length of forwarding path determines the overall signaling cost. Equation (61), (62) as shown at the bottom of the page.

Fig. 8(a) shows the changes of average signaling cost ratio for each proposed LMM scheme against the range of area that a planned SC AP serves when a mobile terminal in a small cell is not allowed to select other small cells than neighboring ones on main layer, where the range is indicated by K, and the probability of the mobile terminal selecting neighboring small cells on main layer is indicated by α.According to the figure, we observe that, significant signaling cost saving is archived by the proposed LMM schemes. As the handover procedure is locally handled by the LMM schemes, the cost for transmitting signaling message between RAN and core network can be saved. Another observation is that, LMM with centralized control scheme (abbreviated to CC in the figures) saves more signaling cost than LMM with distributed control scheme (abbreviated to DC in the figures), especially when SC APs are enhanced with handover control function (abbreviated to CC2 in the figures). Benefit from the separation of control and user planes, the LMM with centralized control scheme can transmit signaling messages via macro BS, which prevents the cost of transmitting these signaling being affected by the increased K. In contrast, the LMM

$$C_{avg}^{3gppCore} = \left\{ \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j [(1-P_\mu)P_\gamma + (1-\alpha)(1-P_\mu)(1-P_\gamma)] + \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i'^j [(1-P_\mu)P_\gamma + \alpha(1-P_\mu)(1-P_\gamma)] \right\} C^{core} \tag{61}$$

$$\begin{aligned} C_{avg}^{LAnchCore} = & \left\{ \sum_{i=0}^{K-1} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j (1-\alpha)(1-P_\mu) + \pi_K^1 \left[\frac{3}{4}\alpha(1-P_\mu)P_\gamma + (1-\alpha)(1-P_\mu) \right] \right. \\ & + \pi_K'^1 \left[\frac{3}{4}(1-\alpha)(1-P_\mu)P_\gamma + \alpha(1-P_\mu) \right] \\ & + \sum_{i=0}^{K-1} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i'^j \alpha(1-P_\mu) + \sum_{j=2}^{\lceil \frac{K+1}{2} \rceil} \pi_K^j \left[\frac{1}{2}\alpha(1-P_\mu)P_\gamma + (1-\alpha)(1-P_\mu) \right] \\ & \left. + \sum_{j=2}^{\lceil \frac{K+1}{2} \rceil} \pi_K'^j \left[\frac{1}{2}(1-\alpha)(1-P_\mu)P_\gamma + \alpha(1-P_\mu) \right] \right\} C^{core}. \end{aligned} \tag{62}$$

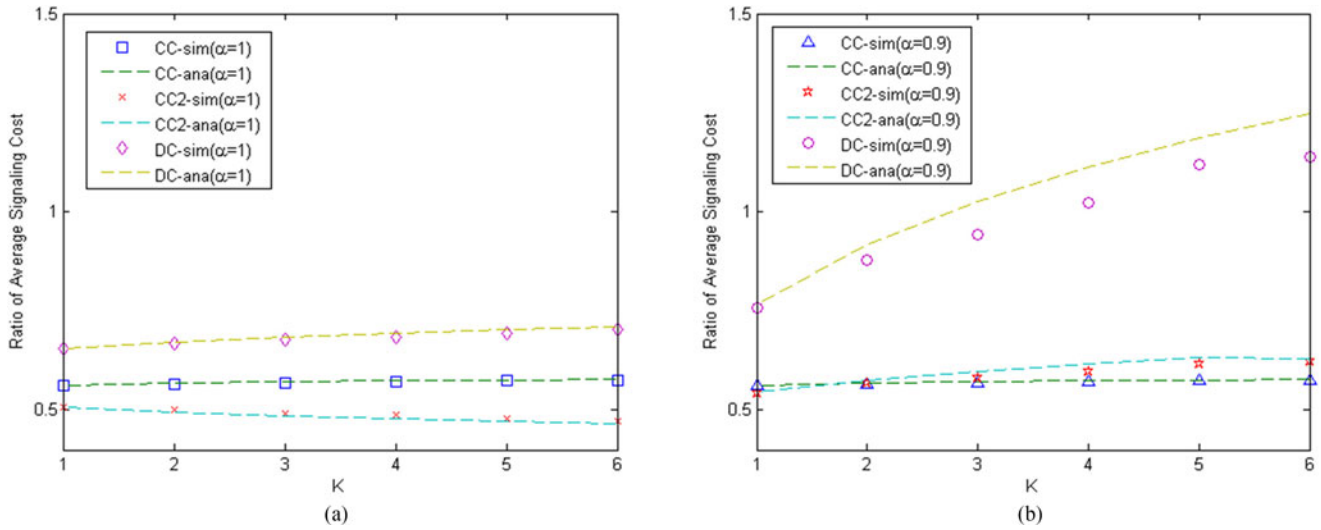

Fig. 8. Change of signaling cost ratio against maximum length of signaling path.

with distributed control scheme completely relies on data forwarding path between SC AP and LSC to transmit signaling messages. The last observation is that, the ratio of average signaling cost for each proposed scheme increases with larger K, except from LMM with centralized control and enhanced SC APs. Larger K implies longer data forwarding path, which leads to higher cost for transmitting related signaling, and ultimately increases overall signaling cost during the handover procedure. Compared with 3GPP scheme, the LMM with centralized control and enhanced SC APs sends less signaling messages via that data forwarding path, thus the ratio of average signaling cost for this scheme decreases with larger K. Fig. 8(b) also shows the changes of average signaling cost ratio for each proposed LMM scheme against the range of area that a planned SC AP serves, when the mobile terminal has the probability $(1-\alpha) = 10\%$ to select small cells on additional layer. Based on the figure, it is observed that, allowing the mobile terminal to select small cells on additional layer brings significant increase of handover signaling cost for the LMM with distributed control scheme, but has little impact on the LMM with centralized control scheme. That is because messages exchanged between source and target SC APs within different layers of networks need to be routed by the planned SC APs on both layers. When target small cell on different layer is selected, the LMM with distributed control scheme which requires direct communication between source and target SC APs will bring more signaling cost. In contrast, LMM with centralized control scheme, especially when direct communication between SC APs is not required, doesn't rely on direct communications between source and target SC APs, thus it is not impacted by α.

Fig. 9(a) shows the change of normalized signaling cost of each scheme against the probability of selecting main layer of network. Based on the figure, all schemes have more signaling cost with the decrease of α, especially for LMM with distributed control scheme and 3GPP scheme. For the LMM with distributed control scheme, the reason is quite obvious, the decreasing α brings the mobile terminal more chance to select a target small cell within additional layer of network, thus more and more direct communications between source and target SC APs in different layers of networks are introduced. Whereas, for 3GPP scheme, although direct communication between source and target SC APs is not required, it can produce lots of signaling cost in each handover procedure, when the mobile terminal has greater possibility to select the target small cell within additional layer of network, handover procedures can be triggered even it keeps not moving, thus the 3GPP scheme also have largely increased signaling cost. As more handover procedures are triggered, the LMM with centralized control scheme have increased signaling cost as well.

Fig. 9(b) shows the change of normalized signaling cost of the handover schemes as a function of the residential time $1/\gamma$, where the mobile terminal only selects small cells on main layer. According to the figure, with the smallest cell residence time, all schemes have the highest signaling cost. That is because smaller cell residence time triggers more frequent handover. As the proposed LMM schemes have signaling cost saving gains in the handover procedure, when the most frequent handover happens, the most signaling cost saving is obtained, i.e. with the smallest cell residence time, the proposed schemes achieve the most signaling cost saving. With the increase of cell residence time, the handover frequency falls, and then the signaling cost of all schemes decreases.

Fig. 9(c) shows the change of normalized signaling cost ratio against the session duration time $1/\mu$, where the mobile terminal only selects small cells on main layer. In the simulation, the maximum signaling cost of the 3GPP mechanism is obtained when the session duration is set to 400 seconds. The curve in the figure shows that the signaling cost ratio of each scheme increases with the longer session duration. That is because longer session duration increases the number of handover.

2) Average Packet Delivery Cost: The concept of average packet delivery cost per mobile terminal is defined in Section IV. According to the previous numerical analysis, the packet delivery cost only depends on backhaul topology, hence

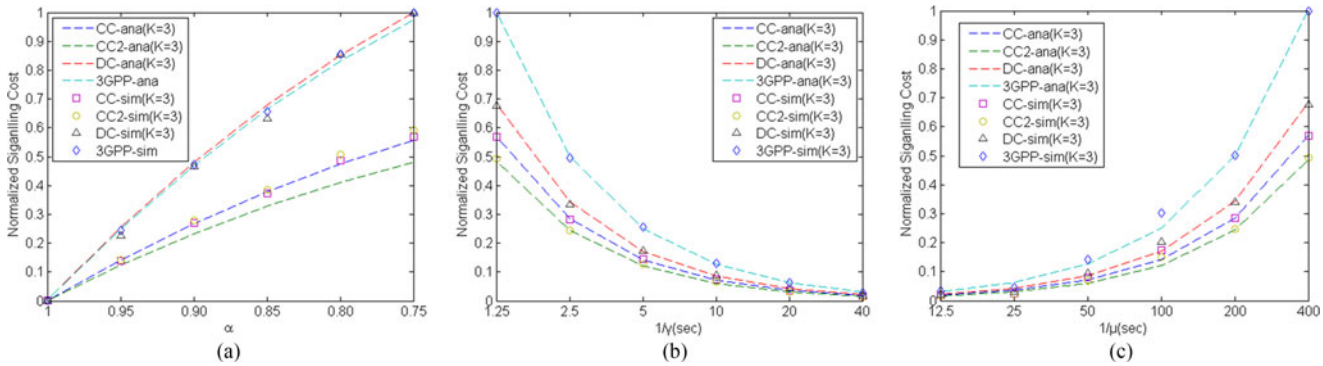

Fig. 9. Change of normalized signaling cost against the probability of selecting main layer of network, cell residence time, and session duration.

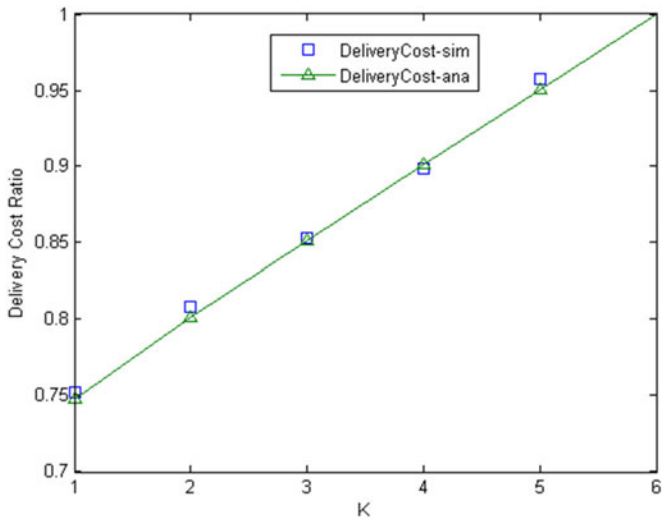

Fig. 10. Change of packet delivery cost ratio against maximum length of data forwarding path.

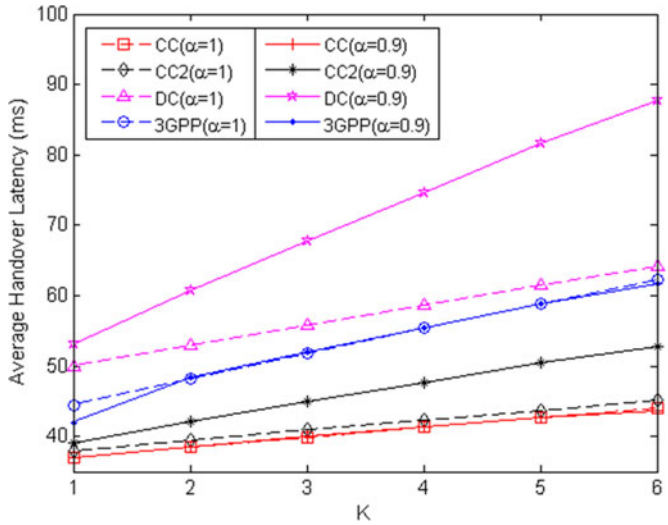

Fig. 11. Change of average handover latency against maximum length of data forwarding path.

both proposed LMM schemes share same packet delivery cost. With same backhaul topology, 3GPP scheme can also have same packet delivery cost. However, such optimal backhaul topology is not always obtained by the 3GPP scheme lacking in centralized backhaul topology management, thus the delivery cost in the real network employing 3GPP scheme might be higher.

We uses delivery cost ratio to illustrate the relation between packet delivery cost and the range of the area that a planned SC AP serves, where the delivery cost ratio is defined as the delivery cost of each scheme to the maximum delivery cost that the schemes achieved. Fig. 10 which is depicted based on numerical analysis and demonstrated by simulation, shows that the delivery cost ratio grows linearly with the increase of K.

3) Average Handover Latency: According to the definitions introduced in Section IV and the values of parameters given in Table II, the average handover latency of each scheme with the change of K is depicted in Fig. 11. Moreover, the handover latency of each scheme is calculated by taking the probability of selecting main layer of network into account, i.e. the value of α is set to 1 and 0.9 respectively.

According to the Fig. 11, we first observed that, the handover latency of each scheme grows with the increase of K, because the handover latency is determined by the time for network nodes processing requests and the time for transmitting signaling messages, and the time for transmitting a signaling message increases with longer signaling path. We further observed that, the LMM with centralized control scheme has the lowest handover latency, especially when SC APs have no handover control function, and the LMM with distributed control scheme on the contrary has the highest handover latency. That is because, the LMM with centralized control scheme allows random access procedure on the air interface and path switch procedure in the LAS to be performed concurrently, which saves some time during the handover; Moreover, in the LMM with centralized control scheme, the LSC can make handover decision and choose the target SC AP, but in the LMM with distributed control scheme, the cooperation between source SC AP and LSC is required for initiating handover procedure, thus additional processing time on the SC AP can be saved. The third observation is that, even the 3GPP scheme has to send request to core network, it still has lower handover latency than the LMM with distributed control scheme. The main reason is that, 3GPP scheme supports concurrently handling signaling flows during the handover procedure as well, and the second reason is that the

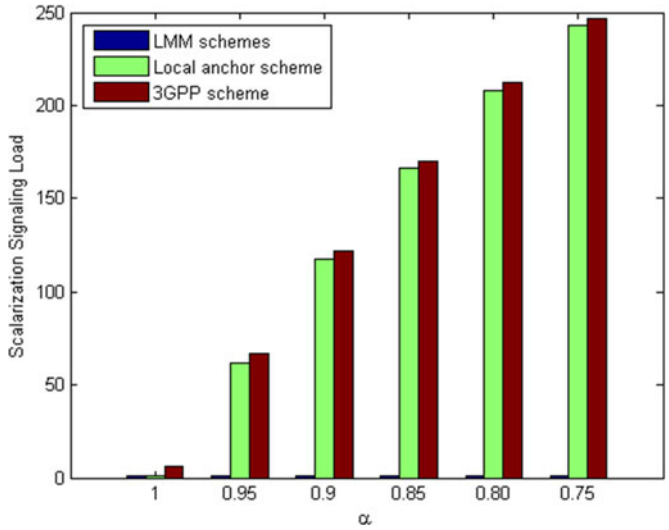

Fig. 12. Scalarization signaling load against the probability of selecting main layer of network.

handover decision is only made by eNBs acting as planned SC APs, which reduces the time for signaling transmission. Finally, handover latencies of the schemes relying on direct communication between source and target SC APs are seriously impacted by the probability α, e.g., the LMM with distributed control scheme, because more handovers between SC APs in different layers of networks, more signaling cost will be introduced.

4) Average Signaling Load to the Core Network: As introduced in Section IV regarding average signaling load to the core network, we use the scalarization signaling load to characterize this cost, so that the performance difference can be more easily captured. The scalarization signaling load here is defined as the average signaling cost of each proposed handover scheme to the minimum signaling cost among them. The values of parameters are set according to Table II except for the cell residence ratio γ being set to 1/sec, since its value in the table is too small to differentiate the performances of local anchor schemes and 3GPP scheme.

According to the Fig. 12, the LMM schemes always have the least signaling load to the core network, since any mobility management signaling can be locally processed a LAS. With the local anchor schemes, the signaling to the core network will be introduced during the handover as long as the local anchor serving the mobile terminal is changed, thus the decreased α which brings the mobile terminal more chance to leave the current layer of network will lead to more signaling to the core network. 3GPP scheme always introduces handover signaling to the core network no matter whether the target and source SC APs are on the same layer or not. Hence the 3GPP scheme has the most signaling load to the core network. Besides these, we further observed that, when the mobile terminal is allowed to select small cells on the additional layer, the signaling cost of the local anchor schemes and 3GPP scheme are significantly increased. The reason we analyzed is that, most of the time, the mobile terminal keeps not moving, then the handovers between two overlapped small cells takes a great proportion in the total handovers. Therefore, even local anchor schemes will not generate signaling to the core network during the intra-local anchor handover, of the total signaling load to the core network is still high. According to this observation, if the average cell residence time for the mobile terminal is decreased (i.e. higher mobility ratio), the difference of scalarization signaling load between the 3GPP scheme and the local anchor schemes will be enlarged.

VI. Conclusion

Although UDN is a well known key technology for 5G networks, it still faces challenges on providing efficient mobility management. In this paper, we first analyze the deployment characteristics of UDN, such as high dense SCs, flexible deployment of SC APs, and multiple types of backhaul links. Then we discuss possible network architectures, as well as network capabilities that need to be supported. By taking those characteristics into account, we proposed that localized mobility management mechanism should be applied to the UDN, and presented LMM with centralized control scheme and LMM with distributed control scheme. For evaluating the performance of the proposed schemes, we developed a mathematical model based on Markov Chain, and determined the performance metrics to be evaluated, i.e. average handover signaling cost, average packet delivery cost, average handover latency and average signaling load to the core network. Based on the model, we derived the expressions of such costs for both proposed schemes and 3GPP scheme, where the 3GPP scheme is used as the baseline for comparative analysis. While evaluating average signaling load to the core network, we further calculated the cost for local anchor schemes. With the help of analytical analysis and simulation experiment, we demonstrated that, the LMM with centralized control scheme has the lowest handover signaling cost, the lowest handover latency, and lowest signaling load to the core network; the LMM with distributed control scheme has lower handover signaling cost, but more handover latency; packet delivery costs of both LMM schemes are same and no more than that of 3GPP scheme.

References

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2014-2019 white paper," May 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

[2] IMT-2020(5G) Promotion Group, "White paper on 5G vision and requirements v1.0," May 2014.

[3] "Study on small cell enhancements for E-UTRA and E-UTRAN; higher layer aspects (Release 12)," 3GPP TR 36.842 v12.0.0, Dec. 2013.

[4] ITU-R report M.2320: Future technology trends of terrestrial IMT systems, Nov. 2014. [Online]. Available: http://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2320-2014-PDF-E.pdf

[5] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility enhancements for heterogeneous networks through interference coordination," in *Proc IEEE Wireless Commun. Netw. Conf. Workshops*, Paris, France, 2012, pp. 69–74.

[6] H. L. Zhang *et al.*, "Interference management for heterogeneous network with spectral efficiency improvement," *IEEE Wireless Commun. Mag.*, vol. 22, no. 2, pp. 101–107, Apr. 2015.

[7] J. Bartelt, A. Fehske, H. Klessig, G. Fettweis, and J. Voigt, "Joint bandwidth allocation and small cell switching in heterogeneous networks," in *Proc. IEEE 78th Veh. Technol. Conf.,* Las Vegas, NV, USA, 2013, pp. 1–5.

[8] A. J. Fehske, I. Viering, J. Voigt, C. Sartori, S. Redana, and G. P. Fettweis, "Small-cell self-organizing wireless networks," *Proc. IEEE*, vol. 102, no. 3, pp. 334–350, Mar. 2014.
[9] K. Sakaguchi, S. Sampei, H. Shimodaira, R. Rezagah, G. K. Tran, and K. Araki, "Cloud cooperated heterogeneous cellular networks," in *Proc. 2013 Int. Symp. Intell. Signal Process. Commun. Syst.*, Naha, Japan, 2013, pp. 787–791.
[10] 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.
[11] H. Ali-Ahmad *et al.*, "An Sdn-based network architecture for extremely dense wireless networks," in *Proc. IEEE SDN Future Netw. Serv.*, Trento, Italy, 2013, pp. 1–7.
[12] J. Zhang, J. Feng, C. Liu, X. Hong, X. Zhang, and W. Wang, "Mobility enhancement and performance evaluation for 5G Ultra dense Networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, New Orleans, LA, USA, 2015, pp. 1793–1798.
[13] R. Balakrishnan and I. F. Akyildiz, "Local mobility anchoring for seamless handover in coordinated small cells," in *Proc. IEEE Global Commun. Conf.*, Atlanta, GA, USA, 2013, pp. 4489–4494.
[14] D. Wang, L. Zhang, Y. Qi, and A. U. Quddus, "Localized mobility management for SDN-integrated LTE backhaul networks," in *Proc. IEEE 81st Veh. Technol. Conf.*, Glasgow, U.K., 2015, pp. 1–6.
[15] H. Leem, J. Y. Kim, K. S. Dan, and Y. Yi, "A novel handover scheme to support small-cell users in a HetNet environment," in *Proc. IEEE Wirel. Commun. Netw. Conf.*, New Orleans, LA, USA, 2015, pp. 1978–1983.
[16] M. Vondra and Z. Becvar, "Distance-based neighborhood scanning for handover purposes in network with small cells," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 883–895, Feb. 2016.
[17] H. Wang, S. Chen, H. Xu, and M. Ai, "SoftNet: A software defined decentralized mobile network architecture toward 5G," *IEEE Netw*., vol. 29, no. 2, pp. 16–22, Mar. 2015.
[18] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks (UUDN) for 5G: Challenges, methodologies and directions," *IEEE Wireless Commun*., vol. 23, no. 2, pp. 78–85, Jul. 2015.
[19] K. Mahmoud, H. Walaa, and Y. Amr, "Ultra-dense networks: A survey," *IEEE Commun. Surv. Tuts.*, vol. 18, no. 4, pp. 2522–2545, Oct.–Dec. 2016.
[20] Third-Generation Partnership Project, "Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN (Release 12)," 3GPP TR 36.932 v12.1.0, Mar. 2013.
[21] C. J. Bernardos *et al.*, "An architecture for software defined wireless networking," *IEEE Wireless Commun*., vol. 21, no. 3, pp. 52–61, Jun. 2014.
[22] Open Networking Foundation, "Software-defined networking: The new norm for networks," ONF White Paper, Apr. 2012. [Online]. Available: https://www.opennetworking.org
[23] EU FP7 INFSO-ICT-317669 METIS, Deliverable D4.3 Version 1: Final report on network-level solutions, Feb. 2015. [Online]. Available: https://www.metis2020.com/documents/deliverables/
[24] Third-Generation Partnership Project, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 13), 3GPP TS 36.300 v13.1.0, Sep. 2015.
[25] EU FP7 INFSO-ICT-317669 METIS, Deliverable D6.1 Version 1: Simulation guidelines, Oct. 2013. [Online]. Available: https://www.metis2020.com/documents/deliverables/
[26] T. Guo, A. Ul Quddus, N. Wang, and R. Tafazolli, "Local mobility management for networked femtocells based on x2 traffic forwarding," *IEEE Trans. Veh. Technol.*, vol. 62, no. 1, pp. 326–340. Jan. 2013.
[27] R. Langar, N. Bouabdallah, and R. Boutaba, "A comprehensive analysis of mobility management in MPLS-based wireless access networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 918–931, Aug. 2008.
[28] J.-H. Lee, T. Ernst, and T.-M. Chung, "Cost analysis of IP mobility management protocols for consumer mobile devices," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 1010–1017, May 2010.
[29] S. H. Cho, E. W. Jang, and J. M. Cioffi, "Handover in multihop cellular networks," *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 64–73, Jul. 2009.
[30] L. Yi, H. Zhou, D. Huang, and H. Zhang, "D-pmipv6: A distributed mobility management scheme supported by data and control plane separation," *Math. Comput. Modelling*, vol. 58, nos. 5/6, pp. 1415–1426, Sep. 2013.

Hucheng Wang received the M.S. degree in 2008 from Beijing University of Posts and Telecommunications, Beijing, China, where he is currently working toward the Ph.D. degree with Communication and Information Systems. He is also a Senior Standard Engineer in China Academy of Telecommunication Technology. His research interests include architectures, networking, and protocols for cellular mobile networks.

Shanzhi Chen (SM'04) received the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1997. He joined the Datang Telecom Technology and Industry Group in 1994, where he has been serving as CTO since 2008. He was a member of the steering expert group on information technology of the 863 Program of China from 1999 to 2011. He is a member of Advisory Committee of Experts on the Internet of Things (IoT) development of China, the Director of State Key Laboratory of Wireless Mobile Communications, and the board member of Semiconductor Manufacturing International Corporation. He has made outstanding contributions to the development from TD-SCDMA 3G to TD-LTE-advanced 4G. He received the State Science and Technology Progress Award of China in 2001 and 2012. His current research interests include network architecture, wireless mobile communication, IoT, and emergency communication.

Ming Ai joined the Datang Telecom Technology and Industry Group in 1998. Since 2008, he has been participating in 3GPP CT1 and SA2 meetings as a standard delegate and a Coordinator of Datang. Before 2008, he worked as a Software Engineer and as the R&D Manager for telecommunication equipment. His research interests include mobile communication technology, Internet technologies, and standard activities.

Hui Xu received the Ph.D. degree from Xian Jiaotong University, Xi'an, China, in 1999. She is currently the Manager of the Ubiquitous Network Department, Datang Wireless Mobile Innovation Center, and is involved in research of key technologies in Internet of Things and Machine to Machine Communications.