Local Anchor Schemes for Seamless and Low-Cost Handover in Coordinated Small Cells

Ravikumar Balakrishnan, *Member, IEEE,* and Ian Akyildiz, *Fellow, IEEE*

Abstract—Cellular systems are rapidly evolving from a homogeneous macrocell deployment to a heterogeneous deployment of macrocells overlaid with many small cells. The striking features of small cells include small coverage area, ad-hoc deployment, and flexible backhaul connectivity. These features call for a profound rethinking of traditional cellular concepts including mobility management and interference management among others. Owing to the unique features, coordinated small cells or commonly referred to as network of small cells promises several benefits including efficient mobility management in a rapid and scalable fashion. The problem of handover in a high-density small cell deployment is studied in this work. A novel local anchor-based architecture for a static cluster of small cells is proposed using which new handover schemes are presented. Such clusters are prevalent in the evolving cellular systems involving high-density small cell deployments in urban, residential, and enterprise environments. A mathematical framework is developed using discrete-time-Markov-models to evaluate the proposed schemes. Using this, closed-form expressions for key handover metrics including handover cost and interruption time are derived. Extensive numerical and simulation studies indicate significant savings of over 50 percent in the handover costs and, more importantly, up to 80 percent in the handover interruption time compared to the existing 3GPP scheme for coordinated small cells.

Index Terms—Small cells, local anchor, handover, performance evaluation

1 Introduction

We are witnessing a gigantic need for ubiquitous and mobile wireless data and services. A 33-fold mobile traffic increase (total of about 127 Exabytes) by 2020 compared to 2010 is predicted in [1]. In addition, studies reveal that over 50 percent of voice traffic and over 70 percent of data traffic originate from within indoor environments [2]. Such growing demands have resulted in tremendous advances in mobile networks both in the radio access and the network management technologies. In particular, cellular networks are undergoing a major transformation with the existing macrocell coverage area overlaid with a number of low-powered small cell base stations. Such a heterogeneous network (HetNet) of cells, each with its unique transmission power, carrier frequency and backhaul technology, is termed as a heterogeneous network. HetNets offer several key benefits including over three orders of magnitude increase in overall capacity and cost-effective coverage extension. However, the growing prominence of this multilayered HetNet architecture leads to several challenges and calls for a profound rethinking of several existing approaches for mobility management, interference management and self-organization among others [3].

Mobility management is a major challenge faced in the large-scale adoption of the HetNet architecture. For this reason, a 3GPP work item for HetNet mobility improvements in LTE has been laid out in [4]. The 3GPP technical report in [5] provides a summary of the mobility performance study and the proposed mobility enhancements under HetNets. Under this, it is shown that the handover performance for users deteriorates with increase in small cell density. This is mainly owing to the numerous cell edges resulting from the HetNet architecture. However, a dense deployment of small cells (SCs), especially in indoor environments, is deemed essential to achieve superior capacity and ubiquitous coverage.

In the case of large-scale small cell deployments such as in airports, large office buildings or auditoriums, coordination among small cells is beneficial in achieving optimal performance. In fact, recent standardization and industry efforts target enterprise small cells for the above scenarios [6] where coordination plays a key role. It enables to achieve improved mobility management, interference management as well as self-organizing (SON) functions by utilizing the underlying network infrastructure [7]. Specifically for mobility management, coordination plays a significant role in the following ways. First, it can facilitate scalable small cell deployments by potentially minimizing the core network (CN) load during handovers. Second, since the small cells incorporate different backhaul technologies to connect to the operator's network, it can help overcome potential backhaul issues of long latency and handover related signaling load during handovers. However, the use of coordinated small cells also place some constraints such as requiring a network infrastructure with high-speed links between small cells.

Handover management in HetNets has been well studied in recent years with a number of solutions aimed to enhance mobility performance. In the case of small cells, since the number of cell crossings are frequent, it is necessary to minimize the impact of the handover on the network and user. This includes the cost incurred during a handover in the form of signaling and data forwarding costs as well

- *The authors are with the Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: {rbalakrishnan6, ian}@ece.gatech.edu.*

Manuscript received 13 Feb. 2014; revised 24 June 2015; accepted 2 July 2015. Date of publication 16 July 2015; date of current version 31 Mar. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2015.2457414

as handover interruption time for the user. Analogous to the earlier studies of fast handovers in [8], a fast handover scheme is proposed in [9] where the higher layer data is buffered to all the small cells in the proximity of the current cell. Although these approaches reduce the handover interruption time, the signaling load at the core network is still large which can negatively impact the large scale deployment of small cells. In [10], it is proposed to move the mobility anchor point from the mobility management entity (MME) to the femtocell gateway. Such an approach does not minimize the interruption time or the handover-related costs since the femtocell gateway is also located at the core network.

In [11], a distributed implementation of the MME, which is the anchor for handover signaling is proposed. Nevertheless, this type of architecture can raise major issues of security, failure handling and synchronization. For instance, upon an incoming connection request for a mobile user (MU), the MME must be aware of the user contexts including location information, associated base station, etc. If the MME is implemented in a distributed fashion, synchronization is required across other distributed entities to update the user information. A lack of synchronization or delay can result in connection failure or failed detection.

In [12], the handover signaling costs are compared for “direct X2” based and “X2-Gateway” based approaches where the direct X2 interface based handover scheme shows significant reduction in core network load and signaling cost. However, these schemes involve “path switching” procedure with the core network causing additional signaling load on the core network if the small cells have internet backhaul. The path switching procedure is a key factor in determining the downlink handover interruption time and delay jitter especially when the number of in-transit packets forwarded from source base station’s buffer is not large [13].

Local mobility anchoring is a promising way to achieve overall mobility performance improvement. Local anchor (LA)-based mobility management schemes were studied in [14] for optimizing paging and registration updates. Similary, cellular IP was proposed in [15] where some of the mobility management functions were moved to the base stations. In [16], new architectures were proposed to move the mobility anchor point closer to the base stations. However, the proposed approach requires redefining the security key mechanisms and signaling flow among other major changes. In order to overcome the delay due to path switching, a local anchor based handover was proposed in [17] using X2 data forwarding analogous to the pointer forwarding technique originally proposed in [18]. In addition to increasing the link utilization on the neighbour cells, it is not clear how the intermediate small cells will participate in establishing and maintaining the X2 forwarding chain to enable data forwarding for the user after multiple handovers.

In this work, we consider a static cluster of small cells, such as in high-density small cell deployment scenarios including enterprise and residential buildings. Under this setup, we propose to utilize a local anchor small cell that anchors the handover for users within the small cell cluster. The local anchor small cell, in a way, operates as a Home eNodeB gateway in LTE-A systems by terminating the signaling and data plane between small cells and the core network. Utilizing such an architecture allows the local anchor to handle user mobility locally. However, our approach also allows the execution of a full core network handover whenever necessary. Therefore, the local anchor based architecture is a tradeoff between the existing centralized MME architecture and the totally radical distributed implementation of the MME.

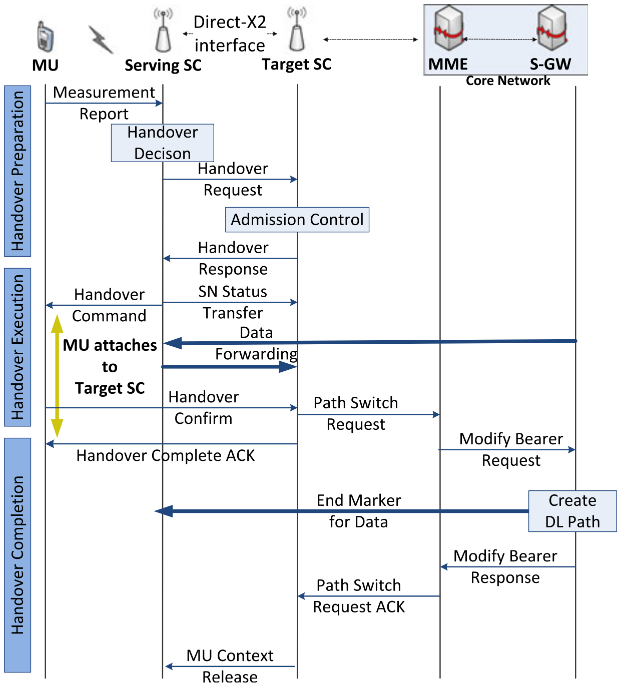


Fig. 1. 3GPP LTE-A inter-small cell handover using direct X2 interface.

Being the extended work of our previous study in [19], the contributions of the paper are summarized as follows:

- We propose novel handover mechanisms using a local anchor-based architecture for coordinated small cells.
- We develop analytical models for a cluster of small cells to study the handover performance.
- We provide closed-form expressions for handover performance parameters under the proposed handover schemes.
- We present numerical and simulation results highlighting the performance gains of the proposed handover schemes.

The rest of this paper is organized as follows. The local anchor-based architecture and the proposed handover mechanisms are presented in Section 2. In Section 3, the analytical model developed to study the mobility behavior of users is described. In Section 4, the closed-form expressions for several handover performance metrics are derived. Numerical and simulation results are presented in Section 5. Finally, the main conclusions are summarized in Section 6.

2 Local Anchor-Based Handover Management

2.1 Motivation

The handover procedure for 3GPP LTE-A systems utilizing the direct interface (X2 interface) between small cells is illustrated in Fig. 1 [20]. The handover procedure is divided into three phases: (i) handover preparation, (ii) handover

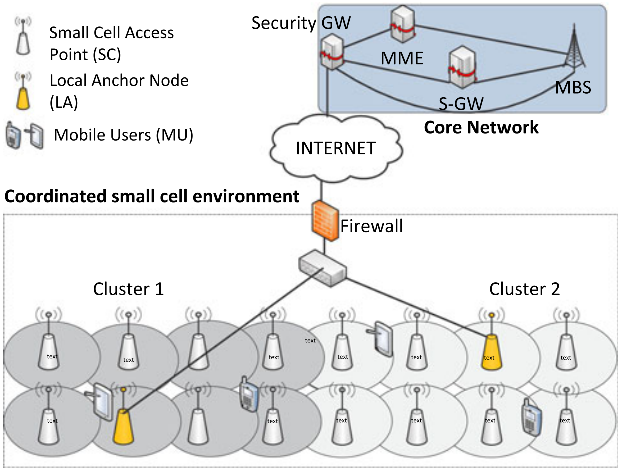


Fig. 2. Local anchor based handover architecture.

execution, and (iii) handover completion. During the *handover preparation* phase, the mobile user sends a measurement report of its neighboring SCs to the serving SC. The serving SC determines the target SC based on the measurement report and issues a handover request. The target SC, in turn, executes the admission control algorithm to determine if the user can be admitted with the requested resources. If the user is admitted, the target SC sends a handover response message to the serving SC. In addition, an uplink data path is established between the target SC and the core network (Serving gateway or S-GW).

During the *handover execution* phase, the serving SC issues a handover command message to the MU to initiate handover execution. The MU now starts attaching to the target SC which includes the establishment of a radio connection. During the same time, the serving SC will forward all the "in-transit" data in its buffer to the target SC. The completion of the radio connection setup is indicated by a handover confirm message from the MU to the target SC.

In the *handover completion* phase, to complete the downlink data path setup from the core network to the target SC, the target SC sends a path switch request message to the core network (MME) after receiving the handover confirm message. During this time, the buffered data from serving SC is continued to be forwarded to the target SC. Once the S-GW has completed the downlink path setup with target SC, a path switch request acknowledgement is sent to the target SC from the MME. In addition, an end marker is sent to the serving SC which is forwarded to the target SC. At this point, the downlink data from the core network for the MU is directly received by the target SC. This also marks the completion of the handover procedure and the serving SC releases all the resources for the MU.

The key observation from the above handover procedure is that the core network (MME) anchors the user mobility. Therefore, intuitively, the backhaul for the small cells must have the key characteristics of low-latency and high reliability for achieving good handover performance. This is given by the handover interruption time computation [13]. The total downlink interruption due to handover τ is determined as

$$\tau = \max(T_r, T_p - T_f), \tag{1}$$

EPS Bearer ID	SC Info			CN Info		
	SC S1AP ID	Transport Layer Address	GTP TEID of SC	CN S1AP ID	Transport Layer Address	GTP TEID of CN

Fig. 3. Local anchor registration table.

where T_r is the time required for the MU to establish radio connection with the target SC where $T_r \approx 18.5ms$ [13]. T_p is the time required by the target SC to perform path switch and depends on the backhaul latency. T_f is the time taken to forward the "in-transit" packets from the serving SC to the target SC. Therefore, to obtain low handover interruption time, the backhaul latency for the small cell must not exceed few tens of milliseconds. However, the nature of the small cell deployments makes it difficult to achieve this stringent backhaul latency requirements. Therefore, we utilize local mobility anchoring to satisfy the joint objectives of minimizing total handover costs, handover interruption time and core network load.

2.2 Local Anchor-Based Architecture

To achieve the above core objectives, we propose a local anchor-based (LA-based) architecture for handover in coordinated small cells. This is shown in Fig. 2. A large array of small cells in a hotspot area is divided into several clusters where each cluster contains a subset of small cells. The cluster is formed based on a group of neighboring small cells which can form a local network. In the local network, we consider that there is one small cell that maintains a link with each of the small cells in the cluster. This small cell will be referred to as the local anchor for the cluster. Coordination within the cluster is enabled using the local network.

The local anchor is networked to the IP gateway of the local network which interfaces with the core network through a firewall and public internet which is one of the commonly adopted backhaul solutions for small cells. The major benefit of the LA-based architecture is that inter-small cell handover mechanisms can be proposed which utilize the **LA as the mobility anchor** therefore minimizing the handover interruption as well as the handover-related costs.

2.2.1 Functions of the Local Anchor

The key functions of the local anchor are described as follows:

a) *Concentrator of traffic between SCs and CN*: The LA performs proxy or anchor function for the data and signaling traffic flows between SCs in its cluster and the CN. This includes proxy function for the S1-AP tunnel, which is a per-user signaling connection between an SC and CN, as well as proxy for the GPRS tunnelling protocol (GTP), which is a per-user data connection between an SC and CN. Such a proxy function is performed by Donor eNodeBs in LTE-A systems to support relays [21].

   To achieve this proxy function, the LA utilizes a local anchor registration table (LART) as shown in Fig. 3. The LART maintains a mapping of the data and signaling plane end point addresses for SC ⇔

EPS Bearer ID	Original SC Info			New SC Info			CN Info		
	Original SC S1AP ID	Transport Layer Address	GTP TEID of original SC	New SC S1AP ID	Transport Layer Address	GTP TEID of new SC	CN S1AP ID	Transport Layer Address	GTP TEID of CN

Fig. 4. Local anchor registration table update after handover.

LA and LA ⇔ CN links corresponding to each MU session indicated by EPS Bearer ID. The proxy function is supported by maintaining S1 security over two hops between CN ⇔ LA and LA ⇔ SC links. Further, to support handover to cells outside the cluster including macrocells, the LA also supports proxy X2 functions. However, we focus our study on intra-cluster handovers in this work.

b) *Setup initial user attach to the network*: The S1 messages carrying the initial attach request goes through the proxy functionality of the LA such that the LA is explicity aware of the user attaching via the SC. Similarly, for the user bearer setup procedures, the S1 messages carrying the bearer setup request arrive at the LA directly from the serving SC. These messages are relayed between the serving SC and the core network. Thus, two GTP tunnels will be established corresponding to the S1-U bearer with the same QoS identifiers between the two hops, first, between the SC and the LA and then between the LA and the CN.

The overheads due to initial user attach on the local anchor will be upper-bounded by the number of SCs within the cluster. However, these overheads do not necessarily contribute to the handover related signaling and interruption delay. Therefore, this is not discussed elaborately in this work.

c) *Local mobility anchor for handover between SCs*: The LA acts as a local mobility anchor for the users performing handover between SCs within its cluster. The LA provides means to avoid the initiation of path switch procedure whenever possible without affecting the established procedures.

Using the local anchor based-architecture, the proposed handover mechanisms are described in the next sections.

2.3 Local Path Switching-Based Handover Mechanism

In the existing X2-based handover process, the destination SC sends a path switch request message to the core network to initiate the switching of the downlink path for the MU towards the target SC. Besides the switching of the downlink path, the path switch request message also initiates the generation of new vertical keys for the target SC ⟺ MU interface. The newly generated key also enables forward key separation, i.e., once the target SC receives the new key (through vertical key derivation) in path switch request ack from the MME, the previous SC cannot decipher the data from the target SC. A detailed explanation of the key management is provided in [22]. In essence, if the path switching is not performed, then the target SC will continue to use the key derived by the previous SC (horizontal key derivation).

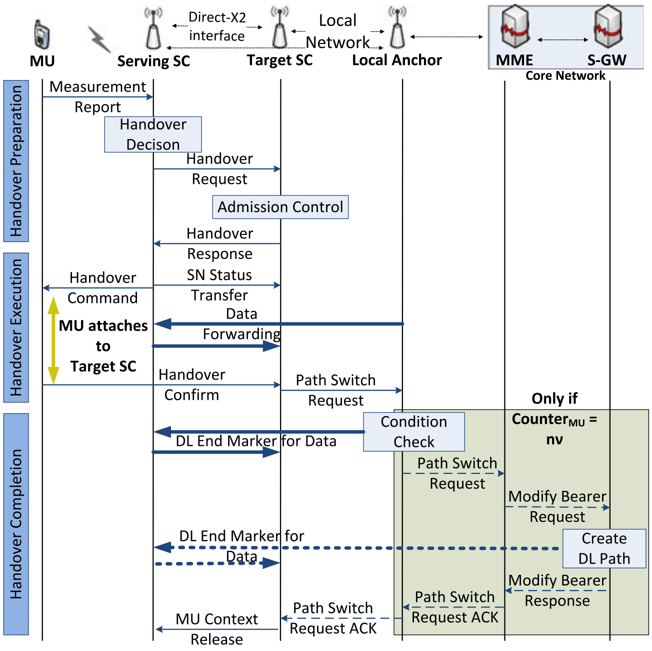


Fig. 5. LP-based handover mechanism for coordinated small cells.

However, we argue that this raises several issues. Due to the large backhaul latency and frequent handovers in small cells, the **path switch request ack** from the MME is not always received in time at the new small cell. *On the one hand*, it can be argued that the two-hop forward key separation is difficult to achieve anyway for small cells that experience poor backhaul conditions since the user can undergo several handovers before the new small cell receives the necessary information to generate a new vertical key. *On the other hand*, the long delays associated with the completion of handover may also increase the likelihood of the MU leaving the target small cell to another cell before the handover is completed. This could lead to higher handover failure rates in small cells compared to the macrocells.

In this work, we propose a *local path switching-based* (LP-based) handover mechanism for SCs belonging to the same cluster for up to ν number of handovers for a given MU as shown in Fig. 5. Due to our proposed local-anchor based architecture, the network still perceives that the user is still attached to the local anchor. Therefore, the handover process can be completed without experiencing handover failures or delay jitters.

To achieve this, the LA maintains a local counter per MU $Count_{MU}$ for the session duration. Path switching is performed with the CN only when $Count_{MU} = n\nu$ where n is a positive integer, i.e., in the *handover completion* phase, the path switch message from the target SC is not forwarded to the CN by the LA. Instead, a new S1 path is created between the target SC and the LA. To support this, the MU session in the LART is updated with the new small cell endpoint address as shown in Fig. 4 until the next handover takes place. As a result of this approach, all future downlink data and signaling for the MU is forwarded from the LA to the target SC. For subsequent handovers, the new target SC information will be updated in the LART. Once the target SC receives the "end marker" from the serving SC over the

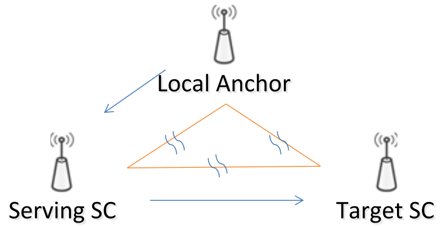


Fig. 6. Triangular routing of in-transit data during handover.

X2 interface, the new DL path between LA and target SC will be utilized for future communication.

For the case of a handover when $Count_{MU} = n\nu$, the actual path switch from the new target SC is forwarded to the core network and the default handover procedure is performed and the backward key separation restored. The counter design, i.e., the choice of ν is based on how often the full path switching is required by the system. We also show through our results how relaxing this constraint can improve the handover performance. As highlighted before, the constraint of two-hop key separation is difficult to achieve for small cell systems with high-latency backhaul and more often leads to handover failures or delay jitters. The benefit of our proposed scheme is that we propose an alternative handover architecture in order to deal with the practical aspects of small cells systems until a major revamp for creating a more distributed MME architecture is adopted.

The proposed LP-based handover mechanism offers the major benefit of minimizing the handover interruption time τ by minimizing the path switching delay T_p during handover. In addition, it offers key benefits of minimizing the handover-related signaling costs. These benefits will be shown quantitatively through numerical and simulation results in later sections.

2.4 Route Optimization-Enhanced (RO-Enhanced) Handover

Although the LP-based handover mechanism can potentially offer key handover performance benefits, the handover cost can be further minimized for the LA-based architecture. During an inter-small cell handover, the "in transit" downlink data forwarded from the serving SC's buffer to the target SC follows the path LA ⇒ Serving SC ⇒ Target SC. This, albeit only for the handover duration, is analogous to the **triangular routing** occuring in mobile IP networks and results in increased handover cost (specifically, data forwarding cost). This is indicated in Fig. 6.

In this approach, the triangular routing can be overcome by sending the path switch request message from the target SC before the MU performs radio connection setup at the target SC. Using this modified approach, the LA will be able to establish a new S1 path for the target SC and switch the local data path to the target SC. This means that the downlink data will be directly forwarded to the target SC instead of being forwarded through the X2 link setup between serving SC and target SC. To avoid loss of in-transit data packets, the sequence number (SN) of the downlink data packet (delivered over X2 control path over the **SN status transfer message**) is indicated in the path switch request message. However, it must be noted that the uplink data of the user will continue to be forwarded from serving SC's buffer to the target SC until the buffer is empty. We term this approach as route optimization-enhanced handover mechanism. Similar to the case of LP-based handover, the RO-enhanced mechanism involves sending the path switch request message to the core network whenever $Count_{MU} = n\nu$. The RO-enhanced handover scheme is illustrated in Fig. 7.

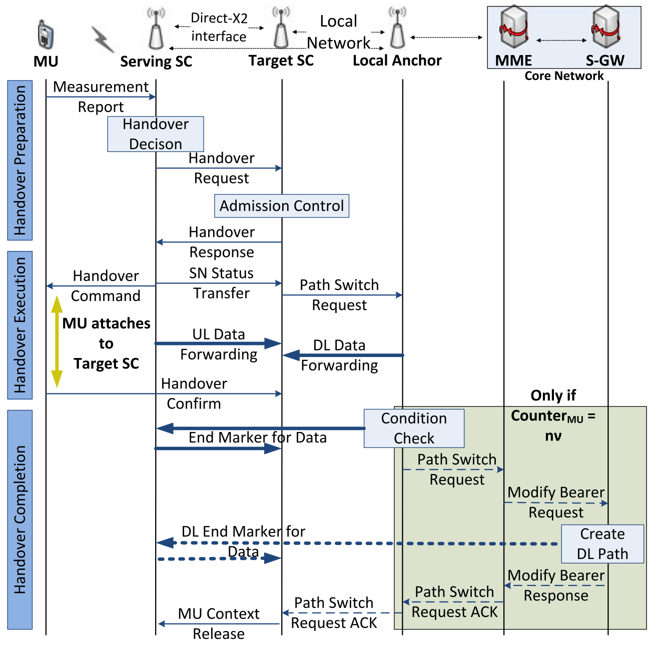


Fig. 7. RO-enhanced handover mechanism for coordinated small cells.

2.5 Data Forwarding-Enhanced Handover (DF-Enhanced)

We further investigate a special handover scenario when the MU handovers from a LA to a neighboring SC in the cluster. In this case, it is possible to entirely eliminate the path switching operation. The X2 link created between LA and target SC during the handover preparation phase will be utilized is generally released after the handover procedure. However, in this data forwarding-enhanced approach, the X2 link will be continued to be used to forward data packets from the LA to the target SC in a similar way as proposed in [17]. Without the path switch procedure, the serving SC does not receive an end marker for data transfer and hence the X2 link is continued to be used for data forwarding. Nevertheless, when the MU moves out of the target SC to a new SC, the RO-enhanced handover mechanism will be applied to obtain maximum savings in terms of handover costs. The DF-enhanced handover mechanism is illustrated in Fig. 8.

2.6 Impacts of Our Proposed Handover Schemes

Even though the concept of locally anchoring mobility has been proposed in the literature, there is not sufficient work that provides a scalable and practical architecture for supporting handover in high-density small cells. In our proposed schemes, the network still perceives that the mobile user is attached to the local anchor even if it is attached to any small cell in the cluster. This way, there is

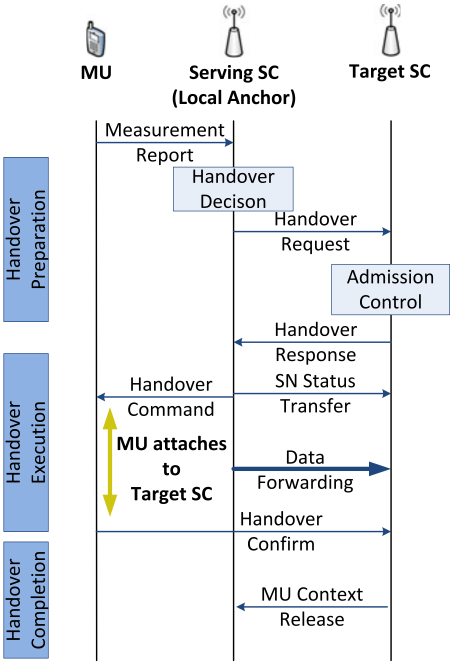


Fig. 8. DF-enhanced handover mechanism for LA to SC handover.

a single point of control (MME) removing the need for synchronization.

In terms of security key management, the only impact we foresee is that small cell keys will not be refreshed after handover resulting in horizontal key derivation instead of vertical key derivation. At the same time, our argument is that the forward key separation is difficult to achieve anyway for small cells that experience poor backhaul conditions since the users can undergo several handovers before the new small cell receives the necessary information to generate a new virtual key after the user handovers.

In another example, the local anchor maybe equipped with the capability to create its own master key. In such a case, the key separation requirement will never be violated. As soon as the handover of a user is completed from one small cell to another within the cluster, the local anchor will be able to generate a new vertical key K_{eNB} for the new air link between the target small cell and the user.

3 Analytical Model

To evaluate the performance of the proposed handover mechanisms, we need to study the evolution of the user's traffic and mobility behavior in the coordinated small cell network. To achieve this, we utilize a discrete-time Markov model to capture the user behavior in a cluster of small cells containing a local anchor small cell. For the developed analytical model, the stationary probabilities for a MU in each small cell are derived. This result will be later utilized to obtain closed-form expressions for handover performance parameters.

3.1 Model Description

We consider a two-dimensional grid topology to model a cluster of small cells as proposed in [23], [24]. This model has also been utilized in [17] mentioned in the prior art. This is shown in Fig. 9. Each block in the grid corresponds to a small cell. The local anchor is located centrally in the grid and the other smalls cells within the cluster are deployed surrounding the local anchor in tiers. The topology consists of up to K tiers of small cells around the LA. It is constructed such that each SC has four neighboring SCs except for the SCs in the Kth tier. The number of SCs in tier i equals $4i$. State variables are used to indicate the presence of a user with an active session within the area of a small cell in the model. The state S_0^1 indicates that the MU is associated with the LA. In general, the state variable S_i^j indicates that the MU is associated with the SC in tier i and cell j within the tier i. An additional state variable S_{idle} is used to indicate that the user currently has no active session independent of which cell the user is associated with.

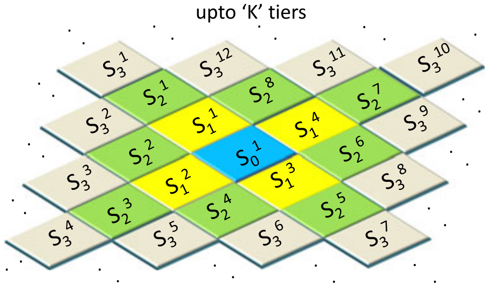


Fig. 9. Two-dimensional grid topology for small cell cluster.

We consider that the MU changes state only at the end of a discrete time slot Δt. The user session is represented using the session arrival and session duration parameters. In this work, we consider that the session arrival follows Poisson distribution with rate λ. Therefore, the session arrival probability is given by $P_\lambda = \lambda\Delta t$. The session duration follows exponential distribution with a mean of $1/\mu$. Therefore, the corresponding probability $P_\mu = \mu\Delta t$ indicates the probability of a session terminating. For the user mobility, a random mobility model is considered where the users can move from an SC to any of its neighbors with equal probability. The cell residence time is modeled using exponential distribution with a mean of $1/r$ and the corresponding probability of a user leaving the current cell given by $P_r = r\Delta t$. Due to the memoryless properties considered, the evolution of user's traffic and mobility behavior can be modeled using a discrete-time Markov model.

Algorithm 1. Algorithm to perform state aggregation for the Markov model.

```
for $m = 1 \to K$ do
    for $b = 1 \to \lceil (m+1)/2 \rceil$ do
        for $a = 0 \to 3$ do
            $S_m^b = S_m^b \cup S_m^{am+b} \cup S_m^{(a+1)m+(2-b)}$
        end
    end
end
```

In the Markov model, each cell would be represented by a state variable. However, this will result in state space explosion as the number of SCs in the cluster increases. Therefore, we apply state aggregation for the Markov model making use of location symmetry arising from the considered mobility model. The state aggregation algorithm is given in Algorithm 1. After state aggregation, we have K tiers with M

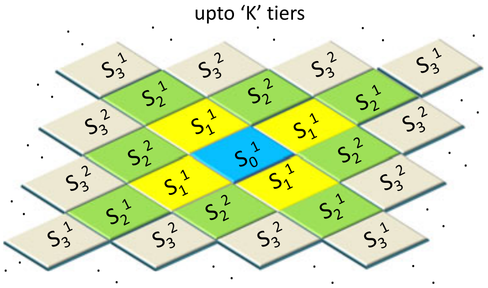


Fig. 10. Grid topology representing aggregated states.

states in each tier such that $M = \lceil (K+1)/2 \rceil$. The topology after performing state aggregation is shown in Fig. 10.

The discrete-time Markov model for the aggregated states is represented in Fig. 11. In the Markov model, the MU remains in state S_{idle} with a probability of $1 - P_\lambda$. After session arrival, the MU wakes up into any of the S_i^j states based on the cell density in each state. The MU returns to the S_{idle} state from any of the S_i^j states with a probability of P_μ. With a probability of $(1-P_\mu)(1-P_r)$, the MU remains in the current cell while having an active session. Also, $P_r(1-P_\mu)P_{s_{ij}s_{\bar{i}\bar{j}}}$ indicates the transition probability that the MU handovers from state S_i^j to state $S_{\bar{i}}^{\bar{j}}$. Upon the completion of a session, the MU returns to the S_{idle} state. In the model, N represents the total number of small cells in the model. The probability $P_{s_{ij}s_{\bar{i}\bar{j}}}$ is obtained based on the user mobility characteristics and the number of tiers in the model. For example,

$$P_{s_{31}s_{21}} = \begin{cases} \frac{1}{4} & \text{if } K > 3, \\ 1 & \text{if } K = 3. \end{cases} \tag{2}$$

Using the transition probability matrix of the Markov model, the normalization and balance equations are determined. Before providing the balance and normalization equations, we define the following parameters:

$$\{\alpha_i, \beta_i\} = \begin{cases} \{\frac{1}{2}, \frac{1}{4}\} & \text{if } i < K-1, \\ \{1, 1\} & \text{if } i = K-1, \\ \{0, 0\} & \text{if } i = K. \end{cases} \tag{3}$$

$$\pi_{idle} = (1 - P_\lambda)\pi_{idle} + P_\mu \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j, \tag{4}$$

$$\pi_0^1 = \frac{P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_0^1 + P_r(1-P_\mu)\beta_0\pi_1^1, \tag{5}$$

$$\begin{aligned} \pi_1^1 = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_1^1 \\ + P_r(1-P_\mu)\{\pi_0^1 + \beta_1\pi_2^1 + \alpha_1\pi_2^2\}, \end{aligned} \tag{6}$$

$$\begin{aligned} \pi_i^1 = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_i^1 + P_r(1-P_\mu) \\ \times \left\{\frac{1}{4}\pi_{i-1}^1 + \beta_i\pi_{i+1}^1 + \frac{1}{2}\alpha_i\pi_{i+1}^2\right\} ; \ \forall\, i > 1, \end{aligned} \tag{7}$$

$$\begin{aligned} \pi_2^2 = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_2^2 \\ + P_r(1-P_\mu)\left\{\frac{1}{2}\pi_1^1 + \frac{1}{2}\alpha_2\pi_3^2\right\}, \end{aligned} \tag{8}$$

$$\begin{aligned} \pi_3^2 = \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_3^2 \\ + P_r(1-P_\mu)\left\{\frac{1}{2}\pi_2^1 + \frac{1}{2}\pi_2^2 + \frac{1}{2}\alpha_3\pi_4^2 + \alpha_3\pi_4^3\right\}, \end{aligned} \tag{9}$$

$$\begin{aligned} \pi_i^2 = \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_i^2 + P_r(1-P_\mu) \\ \times \left\{\frac{1}{2}\pi_{i-1}^1 + \frac{1}{4}\pi_{i-1}^2 + \frac{1}{2}\alpha_i\pi_{i+1}^2 + \frac{1}{2}\alpha_i\pi_{i+1}^3\right\}; \ \forall i > 3, \end{aligned} \tag{10}$$

$$\begin{aligned} \pi_{2j-2}^j = \frac{4P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_{2j-2}^j \\ + P_r(1-P_\mu)\left\{\frac{1}{4}\pi_{2j-3}^{j-1} + \frac{1}{2}\alpha_{2j-2}\pi_{2j-1}^j\right\} ; \ \forall j > 2, \end{aligned} \tag{11}$$

$$\begin{aligned} \pi_{2j-1}^j = \frac{8P_\lambda}{N}\pi_{idle} + P_r(1-P_\mu) + (1-P_\mu)(1-P_r)\pi_{2j-1}^j \\ \times \left\{\frac{1}{4}\pi_{2j-2}^{j-1} + \frac{1}{2}\pi_{2j-2}^j + \frac{1}{2}\alpha_{2j-1}\pi_{2j}^j + \alpha_{2j-1}\pi_{2j}^{j+1}\right\}; \ \forall j > 2, \end{aligned} \tag{12}$$

$$\begin{aligned} \pi_i^j = \frac{8P_\lambda}{N}\pi_{idle} + (1-P_\mu)(1-P_r)\pi_i^j + P_r(1-P_\mu) \\ \times \left\{\frac{1}{4}\pi_{i-1}^{j-1} + \frac{1}{4}\pi_{i-1}^j + \frac{1}{2}\alpha_i\pi_{i+1}^j + \frac{1}{2}\alpha_i\pi_{i+1}^{j+1}\right\}; \ \forall j > 2, i > 2j, \end{aligned} \tag{13}$$

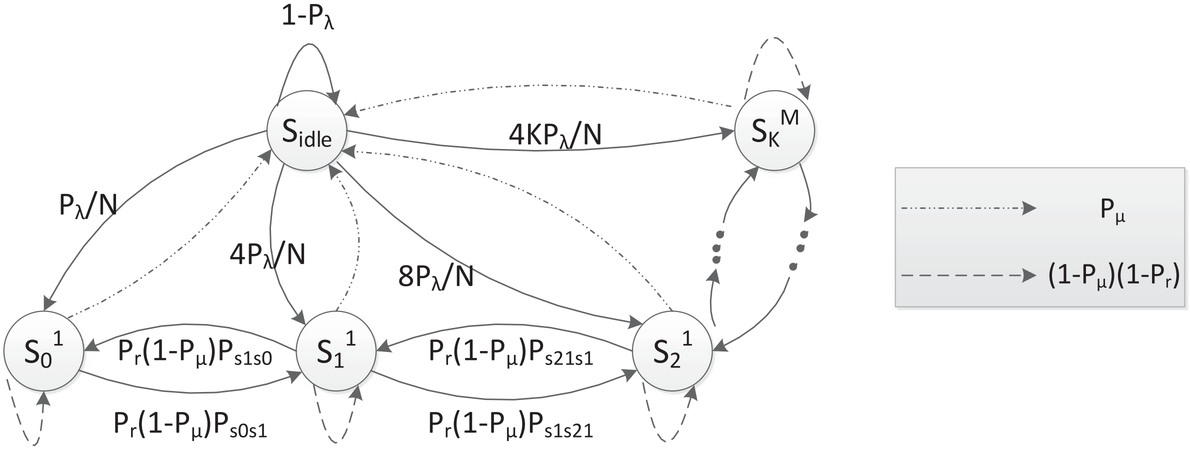


Fig. 11. Discrete-time Markov model for the aggregated states.

$$\pi_{idle} + \sum_{i=0}^{K} \sum_{j=1}^{\lceil \frac{i+1}{2} \rceil} \pi_i^j = 1. \tag{14}$$

The balance equations are given in equations (4)–(13). The normalization equation is given in equation (14). Using equations (4)-(14) and using iterative method, we obtain the stationary probability distribution of the Markov model. Here, π_i^j is of the form $\pi_i^j = x_i^j \pi_{idle} + y_i^j \pi_i^j + P_r(1- P_\mu) z_i^j$; $\forall i, j$. For simplicity, in the future sections, we use the following notations to represent the stationary probabilities:

$$15\pi_i^j = \begin{cases} \Psi_i^j + \Omega_i^j & \forall i \neq 1, j \neq 1, \\ \Psi_i^j + \Theta_i^j + \Omega_i^j & \forall i = 1, j = 1, \end{cases} \tag{15}$$

where $\Psi_i^j = x_i^j \pi_{idle} + y_i^j \pi_i^j$, $\Theta_1^1 = P_r(1 - P_\mu)\pi_0^1$, $\Omega_i^j = P_r(1 - P_\mu) z_i^j$ and $\Omega_1^1 = P_r(1 - P_\mu)(z_i^j - \pi_0^1)$. In other words, the terms Ψ, Θ and Ω are parts of the stationary probability term π_i^j each indicating a specific action that the user performs. The term Ψ_i^j indicates the probability that the user moves to a cell classified under the same state S_i^j from the previous time while a session is active or simply wakes up into state S_i^j from an idle state S_{idle} upon the arrival of a session. Ω_i^j indicates the probability that the user undergoes handover from any state other than its current state S_i^j or S_{idle}. For the special case when the user's current state is S_1^1, the term Θ_1^1 indicates the probability with which the user handovers from the local anchor to a cell in state S_1^1.

We utilize the stationary probability distribution π_i^j and π_{idle} to obtain several handover performance metrics.

4 Handover Performance Metrics

Our objective is to utilize the mathematical framework developed above to obtain the closed-form expressions for different handover performance metrics. The key parameters we consider in this work include average handover cost and average handover interruption time.

4.1 Average Handover Cost

The average handover cost C^{HO} is a key performance metric for handover schemes and we define this parameter as the mean of the total handover-related costs required per handover in the network. The cost functions are expressed in terms of the link latency and processing time involved for a handover.

$$\begin{aligned} 16C^{HO} = & \frac{1}{\xi} \left\lfloor \frac{\xi}{\nu} \right\rfloor \Big\{ \dot{C}_{01}\Omega_0^1 + \dot{C}_{11}\Theta_1^1 + \sum_{i=1}^{K} \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \dot{C}_{ij}\Omega_i^j \Big\} \\ & + \frac{1}{\xi} \Big(\xi - \left\lfloor \frac{\xi}{\nu} \right\rfloor \Big) \Big\{ \ddot{C}_{01}\Omega_0^1 + \ddot{C}_{11}\Theta_1^1 \\ & + \sum_{i=1}^{K} \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \ddot{C}_{ij}\Omega_i^j \Big\}, \end{aligned} \tag{16}$$

where

- ξ : Mean number of handovers per MU per session.
- ν : Number of handovers before full path switching.
- $\dot{C}_{ij}$: Handover cost to an SC in state S_i^j when $Counter_{MU} = n\nu$.
- $\ddot{C}_{ij}$: Handover cost to an SC in state S_i^j with local path switching when $Counter_{MU} \neq n\nu$.

Equation (16) is used to compute both the signaling cost C_{ij}^s and the data forwarding cost C_{ij}^D incurred during a handover by using the appropriate cost functions for $\dot{C}_{ij}$ and $\ddot{C}_{ij}$. For example, to obtain the average signaling cost, the cost functions $\dot{C}_{ij}$ and $\ddot{C}_{ij}$ in equation (16) are replaced by the signaling cost functions $\dot{C}_{ij}^s$ and $\ddot{C}_{ij}^s$ defined in Section 4.4 for each of the proposed handover schemes. Similarly, the average data forwarding cost is obtained by utilizing the data forwarding cost functions $\dot{C}_{ij}^d$ and $\ddot{C}_{ij}^d$ defined in Section 4.4.

4.2 Average Handover Interruption Time

As we discussed before, handover interruption time for a single MU handover is $\tau = \max(T_r, T_p)$ where T_r is the time required for the MU to establish radio connection with the target SC and T_p is the time required to perform path switch at the target SC. Therefore, τ is a measure of the seamlessness of the handover which is a very important metric for QoS performance.

The mean handover interruption time is obtained as follows. Let $\dot{\tau}_{ij}$ and $\ddot{\tau}_{ij}$ indicate the interruption time incurred when a user handovers to a cell in state S_i^j with and without full path switching. Since the interruption times depends on the link delay as shown in equations (18)-(20), different cases of handovers lead to different interruption times. Therefore, we multiply these interruptions by the probability of the user undergoing each handover case. In addition, the percentage of the user undergoing a handover within the local anchor and with the core network is also considered while computing the expression for the mean handover interruption time. For the proposed model, we obtain the mean handover interruption time τ^{HO} as

$$\begin{aligned} \tau^{HO} = & \frac{1}{\xi} \left\lfloor \frac{\xi}{\nu} \right\rfloor \Big\{ \dot{\tau}_{01}\Omega_0^1 + \dot{\tau}_{11}\Theta_1^1 + \sum_{i=1}^{K} \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \dot{\tau}_{ij}\Omega_i^j \Big\} \\ & + \frac{1}{\xi} \Big(\xi - \left\lfloor \frac{\xi}{\nu} \right\rfloor \Big) \Big\{ \ddot{\tau}_{01}\Omega_0^1 + \ddot{\tau}_{11}\Theta_1^1 \\ & + \sum_{i=1}^{K} \sum_{j=1}^{\lceil \frac{K+1}{2} \rceil} \ddot{\tau}_{ij}\Omega_i^j \Big\}, \end{aligned} \tag{17}$$

where

- $\dot{\tau}_{ij}$: Interruption times for handover to an SC in state S_i^j when $Counter_{MU} = n\nu$.
- $\ddot{\tau}_{ij}$: Interruption times for handover to an SC in state S_i^j with local path switching when $Counter_{MU} \neq n\nu$ respectively.

The interruption times are given by $\dot{\tau}_{ij} = max(T_r, \dot{T}_p)$ and $\ddot{\tau}_{ij} = max(T_r, \ddot{T}_p)$ where $\ddot{T}_p$ is given by

$$\ddot{T}_p^{\,LP} = \ddot{T}_p^{\,RO} = \begin{cases} C_{la}; & \text{if } i = 0, j = 1, \\ C_{sc} + C_{la} + 2 * C_{s1}; & \text{if } i = 1, j = 1, \\ C_{sc} + C_{la} + 2 * C_{s1}; & \textit{otherwise}, \end{cases} \tag{18}$$

for the LP-based and RO-enhanced schemes. Here, C_{sc}, C_{la} are the processing costs at small cell and local anchor respectively in milliseconds. C_{s1} represents the S1 link cost with local anchor. For the DF-enhanced scheme, it is given by

$$\ddot{T}_p^{DF} = \begin{cases} C_{la}; & \text{if } i=0, j=1, \\ 0; & \text{if } i=1, j=1, \\ C_{sc}+C_{la}+2*C_{s1}; & otherwise, \end{cases} \tag{19}$$

$\dot{T}_p$ for all the three schemes is given by

$$\dot{T}_p = \begin{cases} C_{la}+C_{scgw}+2C_{s1*}+C_{ps}; & \text{if } i=0, j=1, \\ C_{sc}+2C_{la}+C_{scgw}+2C_{s1*} & \\ +C_{ps}; & otherwise, \end{cases} \tag{20}$$

where C_{s1*} is the link cost for the S1 link between the SC and the core network. In addition, C_{scgw} represents the processing cost at the small cell gateway and C_{ps} represents the path switching cost. In order to compare our solution to the 3GPP LTE-A handover approach, we apply $\ddot{\tau}_{ij} = \dot{\tau}_{ij}$.

4.3 Average Core Network Load

The average core network load is defined as the average number of signaling messages that involve the core network (both MME and S-GW) resulting from a MU handover. The expression for the average core network load can be obtained from equation (16) by replacing $\dot{C}_{ij}$ and $\ddot{C}_{ij}$ with the per-handover core network load l_{CN}.

4.4 Cost Functions

The cost functions are two-fold: (i) Signaling cost C^s_{ij}, and (ii) Data forwarding cost C^d_{ij}. The signaling cost represents the total delay associated with executing handover related signaling messages during the entire handover procedure. The data forwarding cost represents the total delay associated with forwarding the data from the serving SC to the target SC.

Therefore, C^s_{ij} and C^d_{ij} account for the processing and transmission of all the signaling and data required in the successful completion of a handover process. Processing cost of a signaling or data packet is determined by the time required at the protocol layer to process the signaling/data packet. Transmission delay indicates the amount of time required to transmit the signaling or data from the transmitting node to the receiving node. We compute the cost functions as the sum of the link delays due to the transmission of the signaling messages and the processing delay for these signaling messages at different nodes during a handover process.

In our topology, the SC in a cluster can be connected to the LA through multi-hop using the local network. However, since the intermediary SCs only act as IP routers for the SC $\Leftrightarrow$ LA communication, the processing cost at the intermediate SCs mainly originate from the router processing cost. In this case, the intermediate SCs do not require the additional processing cost at the X2 layer performing GPRS tunneling functions and S1 application protocol functions. With this in mind, we have considered that the routing processing cost to be negligible in the computation of the cost function.

4.4.1 LP-Based Handover

We provide the signaling and data forwarding costs for the LP-based handover mechanism when local path-switching is applied under conditions that $Counter_{MU} \neq n\nu$. For the case of $Counter_{MU} = n\nu$, the 3GPP LTE-A handover cost will be applied which will be presented later in this section. The handover costs can vary depending on whether the handover occurs from LA to SC or SC to SC.

In relation to Fig. 5, the total signaling cost $\ddot{C}^s_{ij}$ in the LP-based handover scheme involves the cost of processing and transmitting all the signaling messages between the serving small cell, target small cell and the local anchor during the handover procedure. This includes the handover request, handover response, SN status transfer, path switch request and the end marker messages. The total signaling cost for the case of $Counter_{MU} = n\nu$ can be determined in a similar manner by accounting for the processing and transmission delays of all the signaling messages following the Fig. 5.

The total signaling cost $\ddot{C}^s_{ij}$ is given as

$$\ddot{C}^s_{ij} = \begin{cases} 4C_{sc}+2C_{la}+4C_{X2}+C_{s1}; & \text{if } i=0, j=1, \\ 5C_{sc}+2C_{la}+4C_{X2}+2C_{s1}; & \text{if } i=1, j=1, \\ 5C_{sc}+2C_{la}+4C_{X2}+3C_{s1}; & otherwise, \end{cases} \tag{21}$$

where C_{sc}, C_{la} are the processing costs at small cell and local anchor respectively. C_{X2} and C_{s1} represent the X2 and S1 link costs respectively.

The data forwarding cost $\ddot{C}^d_{ij}$ is given as

$$\ddot{C}^d_{ij} = \begin{cases} C_{sc}+C_{la}+C_{X2}+C_{s1}; & \text{if } i=0, j=1, \\ C_{sc}+C_{X2}; & \text{if } i=1, j=1, \\ C_{sc}+C_{la}+C_{X2}+C_{s1}; & otherwise. \end{cases} \tag{22}$$

4.4.2 RO-Enhanced Handover

The signaling cost for this approach is the same as in equation (21). The data forwarding cost is given as

$$\ddot{C}^d_{ij} = \begin{cases} q(C_{sc}+C_{X2})+C_{la}; & \text{if } i=0, j=1, \\ C_{la}+C_{s1}; & \text{if } i=j=1, \\ q(C_{sc}+C_{X2})+ & \\ \ (1-q)(C_{la}+C_{s1}); & otherwise, \end{cases} \tag{23}$$

where q and $1-q$ are fractions of uplink and downlink data respectively.

4.4.3 DF-Enhanced Handover

The signaling cost and data forwarding cost under the DF-enhanced handover approach differ from the RO-enhanced handover only for the case when an MU experiences handover from LA to SC. These are given as $\ddot{C}^s_{11} = 4C_{sc}+4C_{X2}$ and $\ddot{C}^d_{11} = C_{sc}+C_{X2}$.

For the proposed handover approaches, when path switching is applied, the data forwarding cost will still be $\dot{C}^d_{ij} = \ddot{C}^d_{ij}$. However, the signaling cost is given by

$$\dot{C}^s_{ij} = 5C_{sc}+4C_{X2}+3C_{s1*}+2C_{scgw}+C_{ps}. \tag{24}$$

We compare our solution for the case when no local mobility anchoring is utilized. In this case, we apply $\ddot{C}^s_{ij} = \dot{C}^s_{ij}$ to determine the signaling cost. Similarly the data forwarding cost is given by $\dot{C}^d_{ij} = \ddot{C}^d_{ij} = C_{sc}+C_{X2}+C_{la}+C_{s1}$.

TABLE 1
System Parameters

Parameter	Value	Parameter	Value
Session arrival rate λ	$0.001/s$	C_{sc}	5ms
Session duration parameter μ	$0.01/s$	C_{la}	10ms
Cell residence parameter r	$0.1/s$	C_{s1}	10ms
Number of tiers K	$1, 2, 3, 4$	C_{X2}	5ms
Number of small cells N	$5, 13, 25, 41$	C_{s1*}	50ms
Discrete-time slot duration Δt	$0.01s$	C_{scgw}	10ms
–	–	C_{ps}	5ms

5 Analytical and Simulation Results

5.1 Numerical Evaluation

First, we numerically evaluate the proposed schemes based on the performance metrics derived in Section 4. The system parameter values used for the numerical evaluation are provided in Table 1 as recommended in [13]. The handover performance metrics are provided as ratio of the values achieved by the proposed scheme to the ones achieved by the current 3GPP handover scheme. This means that a smaller ratio corresponds to improved handover performance.

In Fig. 12a, the average signaling cost ratio under the proposed handover schemes is plotted as a function of the number of tiers (K) in the cluster and the path switching threshold (ν). We observe that all three proposed schemes vary with K and ν in a similar manner. The signaling cost ratio is large for small values of ν. This is expected as low values of ν will trigger more frequent path switching for each handover. However, as ν is increased, we observe that the signaling cost ratio decreases rapidly and reaches to around 40 percent of the maximum ratio. This corresponds to about 60 percent of signaling cost savings. This validates our claims that reducing the frequency of full path switching with the network results in signaling cost savings. Among the three mechanisms, the DF-enhanced handover offers the highest savings in signaling cost of about 70 percent for $K \leq 2$ and $\nu \geq 8$. Even for $\nu = 6$, all the three proposed schemes are able to offer almost 60 percent signaling cost savings. It is also interesting to note that increase in tier size (K) does not have a major effect on the signaling cost performance. This can be attributed on the one hand to the single-hop links considered in our cost functions. On the other hand, this is also due to the overall reduction in the signaling delay by minimizing the signaling towards to the core network.

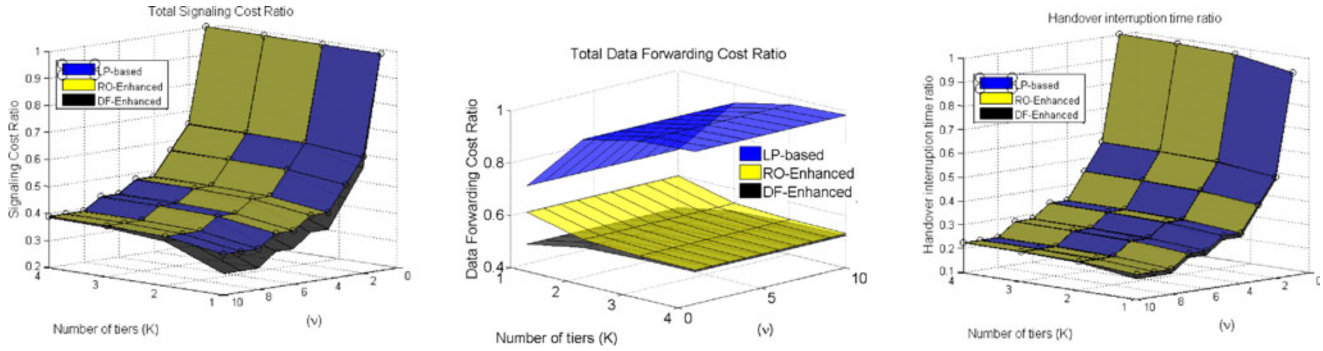
The average data forwarding cost is plotted as a function of ν and K in Fig. 12b. We observe that the LP-based handover does not offer significant performance gain when $K \geq 3$. This is expected as there is a triangular routing of the "in-transit" data from the serving SC to the target SC during a handover. However, with the support of route-optimization for the "in-transit" data packets, the RO-enhanced and DF-enhanced schemes are able to achieve about 50 percent gain in the average data forwarding cost. It is also important to note that both these schemes do not vary significantly with ν since the route optimization is independent of the frequency of path switching. It is worthy of observing that the RO-enhanced scheme has a unique behavior compared to other schemes with increasing K. This is caused due to the high cost involved for a MU handover from LA to a neighboring SC. However, as K increases, the probability of this handover occuring decreases and hence the data forwarding cost becomes low.

Fig. 12c highlights the average handover interruption time ratio plotted against ν and K. The parameter affecting the handover interruption time is mainly the path switching threshold. As ν increases, the path switching takes place less frequently and hence we can observe that the handover interruption time decreases and reaches a minimum of up to 20 percent of the maximum ratio. This corresponds to about 80 percent reduction in the interruption time. This is a key benefit of our proposed schemes as lower interruption time reduces handover failures and enable seamless mobility for users.

5.2 Simulation Setup

In order to validate the performance of the proposed handover schemes, we conduct Monte-Carlo simulation of a small cell cluster with several mobile users using MATLAB. A two-tier small cell cluster is considered consisting of N = 13 small cells (one of which is a local anchor) and 50 mobile users in a 100×100m area. This can be visualized as a deployment in a typical office building with a uniform and coordinated deployment. The topology is shown in Fig. 13. In the figure, each small cell is considered to provide coverage over a 10×10 m grid. The users are deployed randomly in the grid indicated by the thick dots under the small cell coverage.

Session arrival for users is modeled using Poisson distribution with mean arrival rate λ. Session duration and cell residence time are modeled using exponential distribution



(a) (b) (c)

Fig. 12. Numerical evaluation of handover performance versus number of tiers K and path switching threshold ν.

with parameters μ and r respectively. As considered in the analytical model, the users can move to each of their neighboring cells with equal probability at the end of the slot duration. A slot duration of 10 ms is considered and the simulation is performed for 10,000 time slots.

The state evolution is generated for each of the users based on the discrete-time Markov model for the entire simulation duration. Handover occurs in cases when the MUs change their states between any of the S_i^j states in successive time slots. At each time step, the previous and current states are identified to determine if a handover for the MU has occurred. When the MU experiences a handover, the handover performance metrics are computed for the proposed schemes and the current LTE-A handover scheme. The entire simulation is conducted for 10^4 Monte-Carlo iterations. The ratio of the handover costs and handover interruption times of the proposed schemes and the existing LTE-A handover scheme are determined.

First, the above simulation is performed for different values of ν and the results are plotted against the path switching threshold ν in comparison to the analytical results obtained earlier. Further, we also study the effect of the session arrival and user mobility on the handover performance metrics. In the latter case, the path switching threshold is fixed but the session arrival parameters and mobility parameters are varied for which the results are obtained.

5.3 Simulation of a Large Network with Gauss-Markov Mobility Model

In order to further validate our results, we simulate a large network with several clusters. For this, we constructed a scenario with a rectangular grid containing 100 × 100 small cells with each having a coverage area of 10 m. The intersection of every third row and column contains a local anchor. This is chosen to resemble a cluster of tier size $K = 2$. The network consists of 500 users with the initial locations of users randomly selected in the 2-D grid. The users experience mobility based on the well-known Gauss-Markov mobility model [25].

Under the Gauss-Markov mobility model, each user chooses a velocity and direction to follow at every discrete time step. The velocity V and direction D of the users at time step t are computed based on the values at time step $t-1$ and a random variable as

$$V_t = \alpha V_{t-1} + (1-\alpha)\bar{V} + \sigma_V \sqrt{1-\alpha^2} w_{V_{t-1}}, \tag{25}$$

$$D_t = \alpha D_{t-1} + (1-\alpha)\bar{D} + \sigma_D \sqrt{1-\alpha^2} w_{D_{t-1}}, \tag{26}$$

where $0 < \alpha < 1$ is a tuning parameter to determine the randomness of the movement. $\bar{V}$ and σ_V indicate the asymptotic mean and standard deviation of the velocity. Similarly $\bar{D}$ and σ_D indicate the asymptotic mean and standard deviation of the direction. $w_{V_{t-1}}$ and $w_{D_{t-1}}$ are uncorrelated Gaussian processes with zero mean and unit variances and are independent of V_{t-1} and D_{t-1}.

Then the (x_t, y_t) coordinates of the user in 2D space at time t is given by

$$x_t = x_{t-1} + V_{t-1} cos D_{t-1}, \tag{27}$$

$$y_t = y_{t-1} + V_{t-1} cos D_{t-1}. \tag{28}$$

The simulation is run for 1,000 time steps such that each time step has a duration of 1s. The session parameters are retained from Section 5.2. For the Gauss-Markov mobility model, we adopt $\alpha = 0.5$, $\bar{V} = 6m/s$ and $\bar{D} = \pi/2$. The asymptotic variances are set as $\sigma_V = 3$ and $\sigma_D = \pi/8$. $w_{V_{t-1}}$ and $w_{V_{t-1}}$ determined from a standard normal distribution.

Users experience handover if they cross cell boundaries while a session is active between successive time steps. The handover cost functions are computed based on the type of the handover, i.e., if the user handovers from small cell to another small cell or local anchor and so on. For inter-cluster handover, we consider the cost to be equivalent to a SC-SC handover with full path-switching. The simulation is conducted for 10,000 Monte Carlo trials and the average handover costs are computed.

5.3.1 Handover Failure Probability Computation

We also compute the handover failure probability for the proposed schemes under the Gauss-Markov mobility model. Our analysis will indicate how often users experience handover failures with the existing and the proposed handover schemes. A greater handover failure rate causes radio link failures for the users and requires robust handover recovery solutions. However, in this work, we do not focus on how to deal with the handover failure, i.e., the handover failure recovery process.

The handover failure probability P_f is defined as the probability that a user leaves the coverage of the target SC before handover is completed. If a user leaves the coverage of the target small cell before handover is completed, the handover is considered to have failed. In this case, the user will have to reattach to the serving SC or require a new handover to a different SC. The time to complete the handover is determined as the sum of the handover signaling cost and the data forwarding cost. The probability of user leaving the coverage of the target SC depends on the user velocity, direction and the cell radii.

5.4 Comparison of Numerical and Simulation Results

We provide a comparison of the results from the analytical and the simulation models described above.

5.4.1 Impact of Path Switch Threshold on Handover Performance

The handover costs and interruption times are plotted against ν for a fixed $K = 2$. Fig. 14 plots the average signaling cost ratio against the path switching threshold ν for the proposed schemes. All three proposed handover schemes show consistent behavior in the numerical and simulation evaluation where the average signaling cost decreases sharply at the beginning. However, beyond $\nu = 4$, there is not a major impact on the signaling cost. This can be attributed to the session duration of the MU. In the proposed handover schemes, the $Counter_{MU}$ is incremented as long as the user session remains active. However, if the session terminates, the $Counter_{MU}$ value is reset. It is very likely that the counter value gets reset before it reaches high values of ν. Therefore, the increase in ν does not seem to significantly impact the handover performance beyond a certain point.

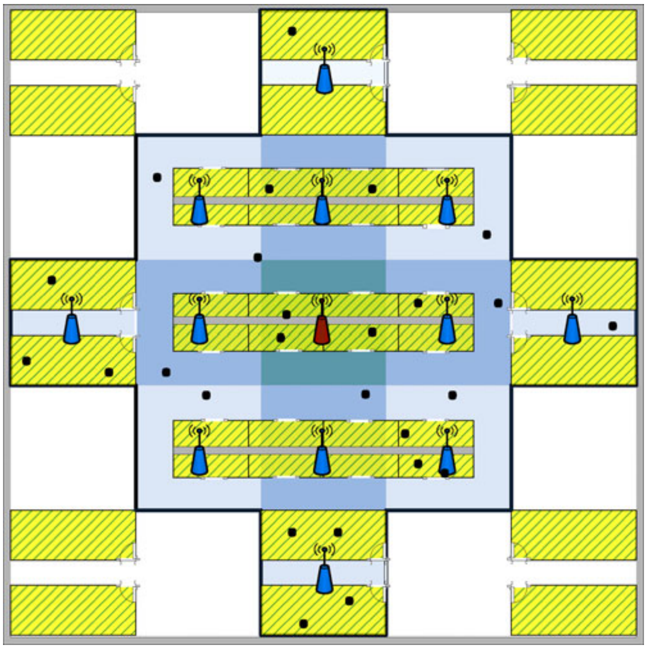


Fig. 13. Simulation of a small cell cluster with K = 2.

It is also interesting to observe that there is greater signaling cost savings through simulations compared to the numerical results achieving cost saving of over 80 percent for the LP-based and RO-enhanced schemes. The DF-enhanced scheme is able to achieve over 90 percent cost savings for $\nu \geq 4$. It is highly likely that the difference could be due to some of the artifacts in the simluation. One possibility could be the initial location of the users within the clusters in the simulations. It is also possible that this difference is a result of the simulation assumption that the users could leave the cluster area after which the users are considered to be released and new users spawned in the cluster area.

In Fig. 15, the average data forwarding cost is plotted. As expected, the data forwarding cost is independent of ν, which is in agreement with the numerical results. In addition, the cost values through simulation and numerical analysis follow each other very closely reaching up to 50 percent cost savings for the DF-enhanced scheme. A similar performance is observed for the handover interruption time as illustrated in Fig. 16. All three schemes show similar performance with over 80 percent reduction in average interruption time as seen from the simulation results.

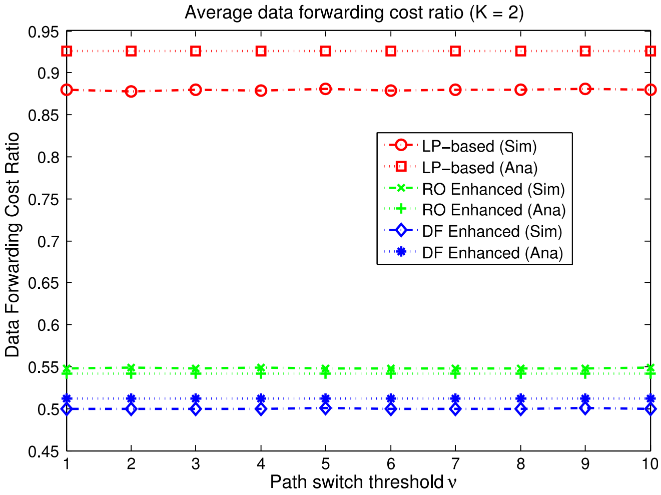


Fig. 15. Impact of ν on average data forwarding cost ratio.

5.4.2 Impact of Call-to-Mobility Ratio on Handover Performance

We study the impact of the session arrival rate and cell residence time on the handover performance. To this end, we utilize the call-to-mobility ratio parameter $\rho = \frac{\lambda}{r}$ where λ is the mean session arrival rate and r is the cell residence parameter. A low value of call-to-mobility ratio indicates the user is highly mobile relative to the mean time between two session arrivals. This implies that the probability of a user undergoing a handover increases leading to high average handover costs. On the contrary, very large values of ρ indicate that the user is less mobile relative to the session arrival. This must lead to a lower probability of user handover leading to low average handover related costs.

The average handover costs for the LP-based scheme are plotted against the call-to-mobility ratio in Fig. 17. For low values of ρ, the signaling and data forwarding costs remain very high due to high mobility of users. As ρ increases, the

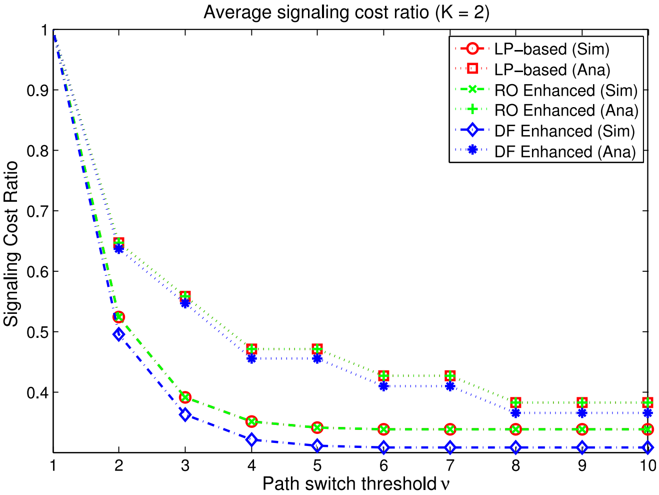


Fig. 14. Impact of ν on average signaling cost ratio.

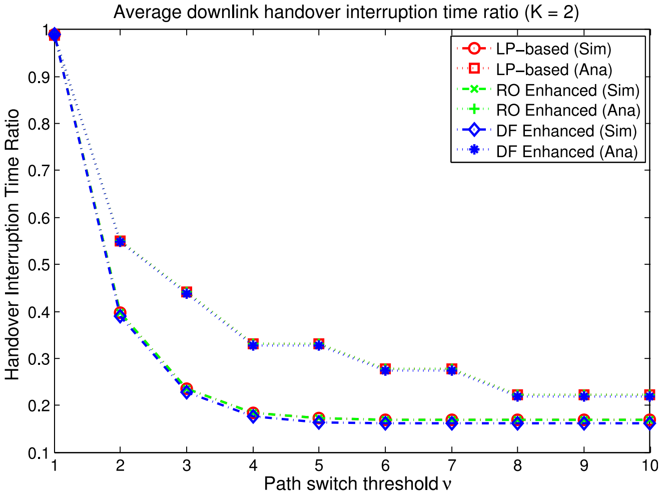


Fig. 16. Impact of ν on handover interruption time ratio.

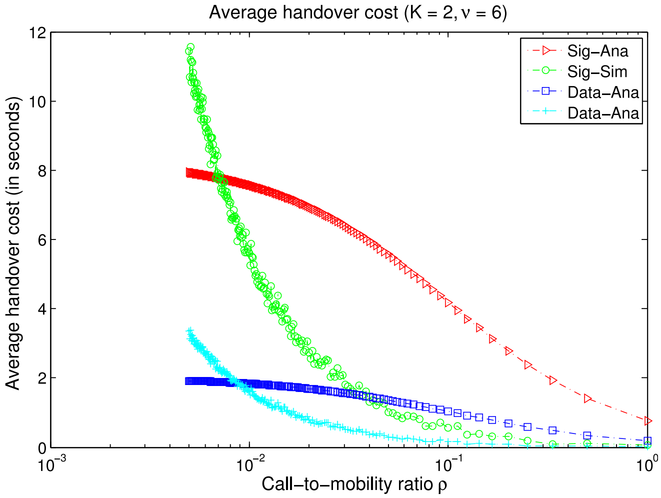


Fig. 17. Impact of call-to-mobility ratio on signaling cost.

users become less mobile thereby experiencing lower handover probability. This is reflected by the low signaling and data forwarding costs for large values of ρ. Although, both the simulation and analytical results show the handover cost as a decreasing function of ρ, the handover cost experiences a steep fall much earlier in the simulation results than in the case of the analytical results. A similar behavior to the handover cost is observed for the handover interruption time when plotted against ρ as shown in Fig. 18. Again, the difference in the performance could be attributed to the artifacts of the simulation study mentioned earlier. It is later shown in Section 5.4.3 in the simulation of multiple clusters, the analytical and simulation results are more closer than observed in this section.

5.4.3 Impact of Different Mobility Models on Handover Performance

The handover costs and interruption times are computed for the different mobility models considered. These results are plotted in Figs. 19 and 20. Our results show that the handover performance both in terms of handover costs and interruption time do not vary significantly if the users follow different mobility models. Especially, the Gaussian Markov mobility model, which provides a tradeoff between the randomness and memory follows closely the performance results from our analytical model. This further shows that the proposed analytical model is fairly accurate in capturing the behavior of the handover dynamics in the system.

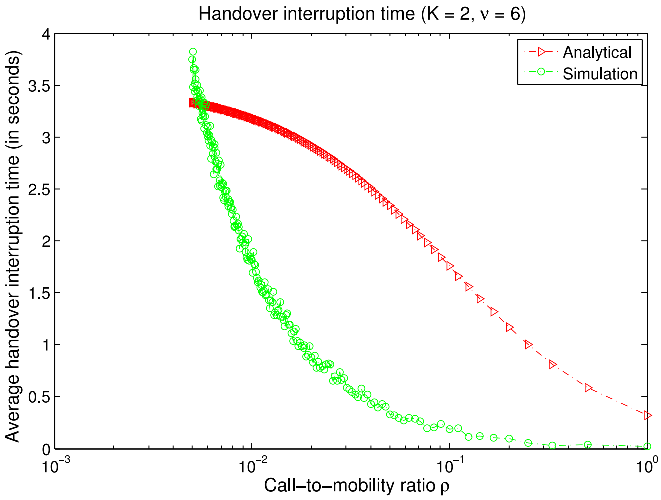


Fig. 18. Impact of call-to-mobility ratio on handover interruption time.

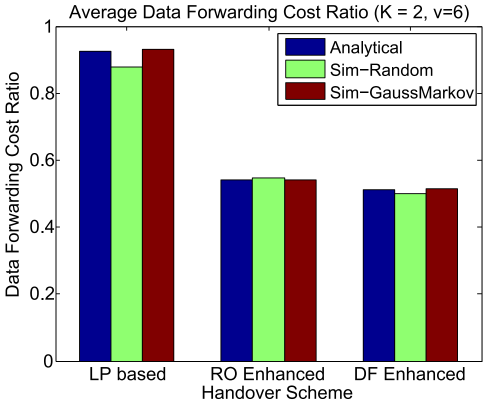


Fig. 19. Data forwarding cost under different mobility models.

5.4.4 Handover Failure Probability

For the simulation of a large network with Gauss Markov mobility model, we determine the handover failure probability P_f for different mean user velocities $\bar{V}$. This is plotted in Fig. 21. We compare the handover failure probability for the baseline LP-based handover scheme and the existing X2 based 3GPP handover scheme. The mean user velocity can be divided into two categories: 1) low velocity $(< 5m/s)$, and 2) medium velocity $(5m/s < 12m/s)$.

For low velocities of users, both the LP-based and the X2-based schemes result in low P_f (under 0.2). However, the difference in P_f for both the schemes increases steadily as the mean velocity is increased. For medium velocities, it is observed that the LP-based scheme leads to about 25 percent reduction in the handover failure probability. As anticipated, the reduction in the core network signaling in the LP-based scheme directly contributes to the reduced handover failure rates. However, as the user velocity increases for the given small cell radii, the probability that the users invariably leave the small cell area increases. Therefore, we observe high handover failure rates for large velocities.

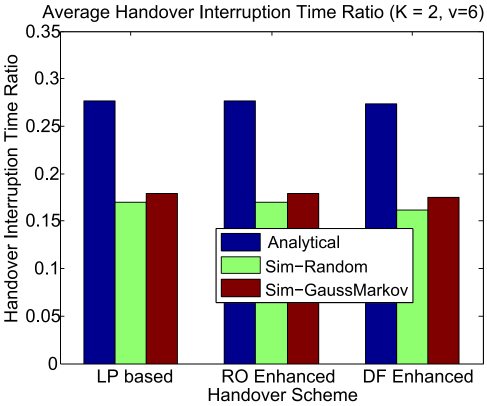


Fig. 20. Handover interruption time under different mobility models.

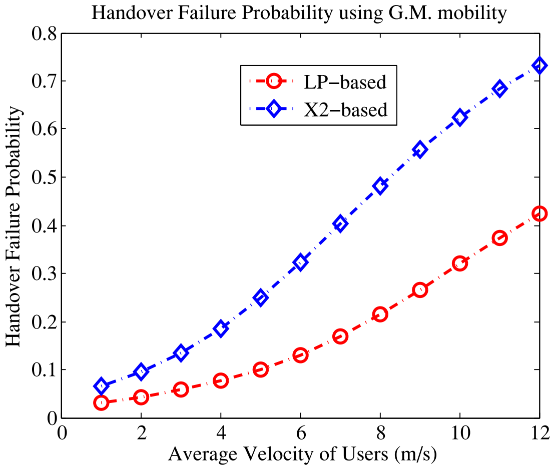


Fig. 21. Handover failure probability using Gauss-Markov mobility model.

6 Conclusion

The emergence of small cells requires a rethinking of the traditional cellular system concepts. Coordination is expected to play a key role in smalls cells to achieve improved mobility management and SON features among others. In this work, we utilized coordination among small cells to propose a local anchor-based handover architecture. Based on this, we proposed novel handover schemes employing a local mobility anchor. To evaluate the performance of the proposed schemes, we developed a mathematical framework based on Markov models. Using this mathematical framework, we derived closed-form expressions for the key handover performance parameters including average handover cost and average handover interruption time. The performance is evaluated using numerical analysis and simulations which indicate savings of over 50 percent in the handover cost and up to 80 percent reduction in the handover interruption over existing schemes. In addition, our results from simulations also indicate a reduction in the handover failure rate for medium velocity users.

With the expected large-scale adoption of small cells in large offices and hot-spot areas, handover management approaches such as our proposed method will be key to facilitate seamless service for mobile users in a scalable and efficient manner. The proposed mathematical framework have been shown to closely follow more realistic models through simulation and hence can be utilized to analyze the performance of new handover schemes that employ a local anchor.

Acknowledgments

This material was based upon work supported by the Georgia Research Alliance Phase 1 award.

References

[1] "Mobile Traffic Forecasts: 2010-2020," UMTS Forum, Tech. Rep. 44, Jan. 2011.

[2] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[3] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); Further advancements for E-UTRA physical layer aspects," TR 36.814, Mar. 2010.

[4] "RP 110438: HetNet mobility improvements for LTE," Nokia Siemens Networks, Nokia Corporation, Alcatel-Lucent, Uusimaa, Finland, TSG-RAN Meeting #51, Mar. 2011.

[5] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); Mobility enhancements in heterogeneous networks," TR 36.839, Dec. 2012.

[6] "Release Two-Enterprise: Overview," Small cell Forum, Nov. 2013.

[7] J. Ferragut and J. Mangues-Bafalluy, "A self-organized tracking area list mechanism for large-scale networks of femtocells," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2012, pp. 5129–5134.

[8] E. Ha, Y. Choi, and C. Kim, "A new pre-handoff scheme for Picocellular networks," in *Proc. IEEE Int. Conf. Personal Wireless Commun.*, Feb. 1996, pp. 140–143.

[9] A. Rath and S. Panwar, "Fast handover in cellular networks with Femtocells," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2012, pp. 2752–2757.

[10] L. Wang, Y. Zhang, and Z. Wei, "Mobility management schemes at radio network layer for LTE Femtocells," in *Proc. IEEE 69th Veh. Technol. Conf.*, Apr. 2009, pp. 1–5.

[11] X. An, F. Pianese, I. Widjaja, and U. Acer, "dMME: Virtualizing LTE mobility management," in *Proc. IEEE 36th Conf. Local Comput. Netw.*, Oct. 2011, pp. 528–536.

[12] H. Zhang, W. Zheng, X. Wen, and C. Jiang, "Signalling overhead evaluation of HeNB mobility enhanced schemes in 3GPP LTE-advanced," in *Proc. IEEE 73rd Veh. Technol. Conf.*, May 2011, pp. 1–5.

[13] 3GPP, "Feasibility study for evolved universal terrestrial radio access (UTRA) and universal terrestrial radio access network (UTRAN)," TR 25.912, Sep. 2012.

[14] J. Ho and I. Akyildiz, "Local anchor scheme for reducing signaling costs in personal communications networks," *IEEE/ACM Trans. Netw.*, vol. 4, no. 5, pp. 709–725, Oct. 1996.

[15] A. Campbell, J. Gomez, and A. Valko, "An overview of cellular IP," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 1999, vol. 2, pp. 606–610.

[16] F. Zdarsky, A. Maeder, S. Al-Sabea, and S. Schmid, "Localization of data and control plane traffic in enterprise Femtocell networks," in *Proc. IEEE 73rd Veh. Technol. Conf.*, May 2011, pp. 1–5.

[17] 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. 2012.

[18] R. Jain and Y.-B. Lin, "An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS," *Wireless Netw.*, vol. 1, no. 2, pp. 197–210, 1995.

[19] R. Balakrishnan and I. F. Akyildiz, "Local mobility anchoring for seamless handover in coordinated small cells," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 4489–4494.

[20] 3GPP, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2," 36.300, Mar. 2013.

[21] 3GPP, "Relay architectures for E-UTRA (LTE- Advanced) (Release 9)," TR 36.806 v9.0.0, 2010.

[22] 3GPP, "3GPP system architecture evolution (SAE); Security architecture," TR 33.401, Mar. 2013.

[23] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); TDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis," TR 36.922, Sep. 2012.

[24] S. C. Forum, "Enterprise Femtocell deployment guidelines," A small cell forum whitepaper, Tech. Rep. 32, Feb. 2012.

[25] B. Liang and Z. Haas, "Predictive distance-based mobility management for PCS networks," in *Proc. IEEE INFOCOM*, Mar. 1999, vol. 3, pp. 1377–1384.

Ravikumar Balakrishnan received the MS and PhD degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, in 2011 and 2015, respectively. He is currently a research scientist with Intel Labs focusing on internet of things for 5G. He has held several internship positions with Intel Corporation and ORB Analytics Inc. His current research interests are in next generation cellular networks, cognitive radio networks, and the Internet of Things. He is a member of the IEEE.

Ian Akyildiz (M‘86-SM‘89-F‘96) received the BS, MS, and PhD degrees in computer engineering from the University of Erlangen-Nurnberg, Germany, in 1978, 1981, and 1984, respectively. He is currently the Ken Byers Chair professor in telecommunications with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, the director of the Broadband Wireless Networking Laboratory and chair of the Telecommunication Group at Georgia Tech. He is an honorary professor with the School of Electrical Engineering, Universitat Polite‘cnica de Catalunya (UPC), in Barcelona, Catalunya, Spain, and the founder of N3Cat (NaNoNetworking Center in Catalunya). He is also an honorary professor with the Department of Electrical, Electronic, and Computer Engineering, University of Pretoria, South Africa, and the founder of the Advanced Sensor Networks Lab. Since 2011, he has been a Consulting Chair professor in the Department of Information Technology, King Abdulaziz University (KAU), in Jeddah, Saudi Arabia. Since January 2013, he has also been a FiDiPro professor (Finland Distinguished Professor Program (FiDiPro) supported by the Academy of Finland) at Tampere University of Technology, Department of Communications Engineering, Finland. He is the editor-in-chief of the *Computer Networks (Elsevier) Journal*, and the founding editor-in-chief of the *Ad Hoc Networks (Elsevier) Journal*, the *Physical Communication (Elsevier) Journal,* and the *Nano Communication Networks (Elsevier) Journal*. He is a fellow of the IEEE (1996) and the ACM (1997). He received numerous awards from the IEEE and ACM. According to Google Scholar, as of June 2015, his h-index is 93 and the total number of citations he received is 70+K. His current research interests are in molecular communication, nano-scale machine communication, 5G cellular systems, software-defined networking, and underground wireless sensor networks.